

https://dspace.org/

Bayesian Belief Networks and Score-Based Network Construction for Nitrate and Nitrite Level Prediction in Water Quality Testing of the Vaal River in South Africa

Stylianou, Nicholas

2024-09

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/46725

Stylianou, Nicholas. (2024). Bayesian Belief Networks and Score-Based Network Construction for Nitrate and Nitrite Level Prediction in Water Quality Testing of the Vaal River in South Africa. [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/46725

Downloaded from a DSpace repository. https://dspace.org/

Bayesian Belief Networks and Score-Based Network Construction for Nitrate and Nitrite Level Prediction in Water Quality Testing of the Vaal River in South Africa

---

Nicholas Stylianou

Student Number: 503815

Supervisors: Prof Mary Evans, Dr Khuliso Masindi and Prof Ritesh Ajoodha

MSc by dissertation



A dissertation submitted in fulfilment of the requirements for the degree in Master of Science in Geography from the School of Geography, Archaeology and Environmental Studies.

Johannesburg, 2024

DECLARATION

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in Geography at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Nicholas Stylianou

(Signature of candidate)

30 day of September 2024 at Johannesburg.

ABSTRACT

This research addresses the essential task of monitoring water quality, particularly in accurately predicting nitrate and nitrite levels in river water. These parameters are crucial indicators of water's suitability for human consumption and ecological health. Traditional monitoring methods, such as manual sampling and expensive sensors, are often prohibitive due to their high costs and labour-intensive nature. In response to these challenges, this research leverages machine learning (ML) algorithms to more efficiently predict nitrate and nitrite levels, utilising a dataset from one water monitoring station on the Vaal River comprising 1071 data points gathered over 21 years (1995-2016). This dataset includes a variety of water quality variables such as calcium (Ca), chloride (Cl), dissolved major salts (DMS), electrical conductivity (EC), fluoride (F), potassium (K), magnesium (Mg), ammonium (NH_4), sodium (Na), phosphate (PO_4), sulphate (SO_4), silicon (Si), total alkalinity (TAL), and temperature. This research employs a series of Score-Based Constructed Bayesian Belief Network models, including Tree-Based All-Features (AF), Hill Climb (HC) with various criteria like Bayesian Information Criterion (BIC) and Random Features (RF), and a Random Forest model with All-Features with varying number of trees, including optimisations. These models are evaluated for their ability to predict nitrate and nitrite values, aiming to provide a cost-effective and efficient alternative to traditional monitoring techniques. The research assesses each model's prediction accuracy, striving for reliable predictions of nitrate and nitrite levels with the Akaike Information Criterion and Enhanced Bayesian Information Criterion Score-Based models performing with the best accuracy and capable of predicting nitrate and nitrite with Root Mean Square Error (RMSE) on the test data of 0.356 and 0.360 respectively, Mean Square Error (MSE) on the test data of 0.127 and 0.130 respectively, Mean Absolute Error (MAE) on the test data of 0.197 and 0.197 respectively and R-squared of 0.7940 and 0.7944 respectively. The study's findings have significant implications for monitoring and managing water resources, potentially shifting the paradigm for ensuring water quality sustainably.

To my beloved wife, Claudia, for her unwavering support and love and for helping me set aside the time needed to pursue this research. Thank you to my friends and family for their support and encouragement throughout this journey.

ACKNOWLEDGEMENTS

I wish to thank the individuals who have made this degree possible. I want to acknowledge Prof Mary Evans for her support, guidance, and knowledge, but most importantly, for believing in me and accepting me as a student under her supervision, and Dr Khuliso Masindi for providing me with the data used in this research. Additionally, I want to thank Prof Ritesh Adjoodha for his support and guidance on the topics of, Bayesian Belief Networks and machine learning.

TABLE OF CONTENTS

DECLARATION ……..... i

ABSTRACT …….…….….. ii

DEDICATION …….…... iii

ACKNOWLEDGEMENTS ……..…. iv

TABLE OF CONTENTS …….…….….. v

ABBREVIATIONS …….. viii

LIST OF FIGURES ix

LIST OF EQUATIONS xii

LIST OF TABLES xiii

Chapter 1: 1

INTRODUCTION 1

1.1 Background 1

1.2 Aim 3

1.3 Objectives 4

1.4 Problem Statement and Rationale 4

1.5 Structure of Dissertation 5

Chapter 2: 6

LITERATURE REVIEW 6

2.1 Introduction 6

2.2 Water quality parameters 8

2.3 What is Machine Learning? 11

2.3.1 Supervised and unsupervised learning 12

2.3.2 How is machine learning being used for water quality 13

2.4 Explainable Artificial Intelligence 19

2.5 Summary of algorithms in water quality 21
Chapter 3: 22
METHODOLOGY 22
3.1 Introduction 22
3.2 General setting and study site 23
3.3 Machine Learning Life Cycle 25
3.4 Collecting and accessing the data for processing 26
3.5 Preparing and exploring the data for processing 28
3.6 Preprocessing of the modelling data 29
3.6.1 Outlier removal 29
3.6.2 Splitting of the data 29
3.6.3 Data summaries 30
3.6.3.1 Imputation of missing data 31
3.6.3.2 Imputation data summaries 32
3.7 Prepare the models for training and testing 33
3.7.1 Data set distribution evaluation 34
3.7.2 Kernel Density Estimation 34
3.7.3 Boxplots 35
3.8 Bayesian Score-Based structured learning models 35
3.8.1 Tree-Based All-Features (Tree-Based-AF) Model 37
3.8.2 Hill Climb – Bayesian Information Criterion – Random Features (HC-BIC-RF) Model 37
3.8.3 Hill Climb – Akaike Information Criterion – Random Features (HC-AIC-RF) Model 38
3.8.4 Hill Climb - Loglikelihood - All Features (HC-Loglik-AF) Model 38
3.8.5 Hill Climb - Akaike Information Criterion - All Features (HC-AIC-AF) Model 39
3.8.6 Hill Climb - Bayesian Information Criterion - All Features (HC-BIC-AF) 39
3.8.7 Hill Climb - Extended Bayesian Information Criterion - All Features (HC-EBIC-AF) Model 39
3.8.8 Hill Climb - Bayesian General Equilibrium - All Features (HC-BGE-AF) Model 40

3.8.9 Hill Climb - Bayesian Information Criterion - Correlated Features (HC-BIC-CR) 40
3.8.10 Random Forest - All Features with 100 trees (RF-AF-100) 40
3.8.11 Random Forest - All Features with 500 trees (RF-AF-500) 41
3.8.12 Random Forest - All Features with 1000 trees (RF-AF-1000) 41
3.8.13 Random Forest - All Features with 1500 trees (RF-AF-1500) 41
3.8.14 Random Forest - All Features with optimisation (RF-AF-Opt) 41
3.9 Evaluation of the different models 42
3.9.1 Actual vs Predicted plots 45
3.9.2 Prediction error per sample size 46
Chapter 4: 47
RESULTS 47
4.1 Introduction 47
4.2 Kernel Density Estimation 47
4.3 Boxplot results 51
4.4 Model structures (Directed Acyclic Graph) 55
4.5 Model Cross-Validation performance 59
4.6 Model Actual vs Predicted results 64
4.7 Error vs sample size performance 67
4.8 Model execution time per sample size 69
4.9 Evaluation of model performance 70
Chapter 5: 77
DISCUSSION AND CONCLUSION..... 77
5.1 Introduction 77
5.2 Analysis of results 77
5.3 Overview of research 82
5.4 Limitations and recommendations of research 83
REFERENCES 85

ABBREVIATIONS

WQI Water Quality Index

BIC Bayesian Information Criterion

EBIC Enhanced Bayesian Information Criterion

AIC Akaike Information Criterion

BGE Bayesian General Equilibrium

AF All Features

RF Random Forest

CR Correlated Features

ANN Artificial Neural Network

DNN Deep Neural Network

RMSE Root Mean Square Error

MSE Mean Square Error

MAPE Mean Average Precision

MAE Mean Absolute Error

DO Dissolved Oxygen

TDS Total Dissolved Solids

LIST OF FIGURES

Figure 1 An example of a decision tree, designed with an explainable structure. By starting at the top and going down, level by level the final decisioning can be determined 19

Figure 2 A) Monitoring sample station (C2H018) in South Africa. B) The approximate location of the site. C) The satellite image of the station on the Vaal River..... 24

Figure 3 Flow chart of the data and modelling process as described in the research methodology, summarising the steps as taken from the machine-learning life cycle framework but include the modelling process using the Bnlearn R package..... 26

Figure 4 A code snippet changing the column headings for calcium and chloride..... 28

Figure 5 The kernel density estimation plot for nitrate and nitrite..... 34

Figure 6 The boxplot for nitrate and nitrite..... 35

Figure 7 The actual vs predicted plot of the Tree-Based All-Features Model. The true values are represented by the blue points while the predicted are represented by the red points. For each record pair of actual/predicted at an index in the test set, the smaller the distance the closer the predictions..... 45

Figure 8 The Kernel Density Estimation plots for A) calcium, B) chloride, C) conductivity and D) DMS..... 48

Figure 9 The Kernel Density Estimation plots for E) fluoride, F) magnesium, G) nitrate and nitrite and H) phosphate..... 49

Figure 10 The Kernel Density Estimation plots for I) potassium, J) silicon, K) TAL, L) temperature and M) ammonium..... 50

Figure 11 The boxplots for A) ammonium, B) calcium, C) chloride and D) conductivity..... 51

Figure 12 The boxplots for E) DMS, F) fluoride, G) magnesium and H) nitrate and nitrite..... 52

Figure 13 The boxplots for I) phosphate, J) sodium, K) potassium and L) silicon..... 53

Figure 14 The boxplots for M) sulphate, N) temperature, and O) TAL..... 54

Figure 15 Directed Acyclic Graph (DAG) for A) Tree-Based-AF, B) HC-BIC-RF, C) HC-AIC-RF and D) HC-Loglik-AF..... 56

Figure 16 Directed Acyclic Graph (DAG) for E) HC-AIC-AF, F) HC-BIC-AF, G) HC_EBIC-AF, H) HC-BGE-AF and I) HC-BIC-CR..... 57

Figure 17 Cross validation estimation; five models are built with a different subset of the data and validated on a percentage not included in the training process. The error estimation is averaged..... 59

Figure 18 The cross-validation performance of the per fold for the A) Tree-Based-AF, B) HC-BIC-RF, C) HC-AIC-RF and D) HC-Loglik-AF models..... 61

Figure 19 The CV performance of the per fold for the E) HC-AIC-AF, F) HC-BIC-AF, G) HC-EBIC-AF, H) HC-BGE-AF and I) HC-BIC-CR models..... 63

Figure 20 The actual versus predicted plots per sample for A) Tree-Based-AF, B) HC-BIC-RF, C) HC-AIC-RF, and D) HC-Loglik-AF models..... 65

Figure 21 The actual versus predicted plots per sample for E) HC-AIC-AF, F) HC-BIC-AF, G) HC-EBIC-AF, H) HC-BGE-AF and I) HC-BIC-CR models..... 66

Figure 22 The prediction error for the models for A) RMSE, B) MSE, C) MAPE and D) MAE evaluation metrics..... 68

Figure 23 Execution Time per Sample Size for each model. As the sample size increases so does the time for the models to complete a training run with selected sample size..... 70

Figure 24 The RMSE evaluation metrics values per model for the A) validation set and B) test set..... 71

Figure 25 The MSE evaluation metric values per model for the C) validation set and D) test set..... 72

Figure 26 The MAE evaluation metric values per model for the E) validation set and F) test set..... 73

Figure 27 The MAPE evaluation metric values per model for the G) validation set and H) test set..... 74

Figure 28 The R-squared evaluation metric values per model for the I) validation set and J) test set.....75

Figure 29 The R-squared evaluation metric values for all the Random Forest models for the K) validation set and L) test set.....76

Figure 30 Features that influence nitrate and nitrite predictions, collated from all Bayesian Belief Networks..... 78

Figure 31 The residual plots for the A) HC-AIC-AF and B) HC-EBIC-AF models..... 81

LIST OF EQUATIONS

Equation 1 The original Water Quality Index (WQI) formula. Wi is the specific weight of each variable (low (1) to high (4)). Pi is the assigned value based on its tolerance range. K is level of contamination when sampling has taken place (1; 0.75; 05) where 1 represents clear water and 0.5 representing a high level of turbidity..... 8

Equation 2 The MDL/BIC scoring function..... 37

Equation 3 The AIC scoring function..... 38

Equation 4 The equation for RMSE. The best achievable value = 0; worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data..... 43

Equation 5 The equation for MSE. The best achievable value = 0; worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data..... 43

Equation 6 The equation for MAPE. The best achievable value = 0; worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data..... 44

Equation 7 The equation for MAE. The best achievable value = 0; worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data..... 44

Equation 8 Equation 8: The equation for R-squared. The worst value = -inf; the best achievable value = +1. R-squared indicates the percentage of variability in the dependent variable that can be predicted by the independent variables..... 44

LIST OF TABLES

Table 1 The water parameters that are considered an acute health risk..... 2

Table 2 A summary of the six most popular WQIs globally and the water quality parameters used to calculate them..... 9

Table 3 A summary of the algorithms and water quality parameters in the literature..... 17

Table 4 A sample of the data once loaded for monitoring station C2H018..... 27

Table 5 A sample of the data for monitoring station C2H018 after water parameter selection..... 28

Table 6 Statistical summary for the train set..... 30

Table 7 Statistical summary of the validation set..... 31

Table 8 Statistical summary of the test set..... 31

Table 9 Statistical summary of the imputed train set..... 32

Table 10 Statistical summary of the imputed validation set..... 33

Table 11 Statistical summary of the imputed test set..... 33

Table 12 A summary of the features which have a direct effect on the target variable with calcium in all but nine models (calcium excluded from feature list) being the most common feature present..... 58

Table 13 A summary of the libraries and associated version used in R and Python.....69

Table 14 The final correlated features as extracted from the nine Bayesian models..... 77

Table 15 A sample of the test set with the predictions and residuals of the HC-AIC-AF and HC-EBIC-AF models..... 82

CHAPTER 1

INTRODUCTION

1.1 Background

Water is considered a critical resource globally; therefore, it is crucial to keep it clean for the survival of all things living on Earth. Humans rely on surface water sources being of suitable quality for consumption Bansal & Geetha, (2020), as well as for hygiene, sanitation, and cooking. Approximately 56% of the world's population's wastewater was treated safely, relying on working infrastructure such as sewage systems, municipal piping systems and treatment facilities United Nations, (2021); therefore, the primary objective in protecting our health is to continue to ensure access to sufficient, safe quantities of water (Sengupta, 2013). As a result, water quality standards play an essential part in determining if the water has degraded and compromised or needs to be treated to a certain level to be fit for human consumption, often referring to the biological, physical and chemical characteristics of the water (Bansal & Geetha, 2020). The approach of using hydrometry stations which measure dissolved oxygen (DO), chemical oxygen demand (COD), the potential of Hydrogen (pH), electrical conductivity (EC), temperature and dissolved solids such as potassium (K), sodium (Na) and magnesium (Mg) are used to determine water quality when being measured in an area (Haghiabi *et al*., 2018). By combining machine learning algorithms and water monitoring sensors, sensor measurements, water quality, and system parameters can be adjusted to make more informed decisions (Anupama *et al.*, 2020; Bansal & Geetha, 2020). According to the World Health Organization, (2017), global water quality standards for drinking water quality cover the following guideline aspects: chemical (lead, arsenic and nitrate), microbial (animal and human faeces), physical (colour and conductivity), and radiological (naturally occurring or human-made radiation such as potassium-40 and uranium-234) parameters to ensure water safety. These guidelines are used as a basis for water quality standards in Africa, with each country adapting or adopting them, since none are unified, to set their guidelines at a national level depending on environmental and economic conditions and priorities of the respective country's health sector (World Health Organization, 2021). In

South Africa, the standards for potable drinking water are aligned with the World Health Organization (WHO) guidelines and are tailored to the country's needs; these are in the form of a national standard: South African National Standards (SANS) 241, (World Health Organization, 2021).

Nitrates and nitrites in water quality are of particular interest due to the acute health effects on human health; in the short-term, nitrates and nitrates can interfere with the body's ability to carry oxygen, leading to a fatal condition in infants called blue baby syndrome, if not treated (Water Quality Association, 2023; World Health Organization, 2017), in the long-term exposure to high levels have been associated with cancer increases, in particular colon and nitrites can react with amides and amines, responsible for intestinal function regulation, forming nitrosamines responsible for stomach, oesophagus and nasopharynx cancer (Water Quality Association, 2023). In the aquatic environment, nitrate is essential for algae, some bacteria and macrophyte growth. Still, high concentration levels can lead to eutrophication and lower dissolved oxygen, negatively impacting fish and invertebrates (Kermorvant *et al*., 2023).

Nitrates and nitrites are not the only water quality parameters that can affect human health, according to SANS 241. By understanding which aspects of water quality parameters, Table 1 are of the most concern, the relationships between them and water quality can be better understood.

Table 1: The water parameters that are considered an acute health risk (Source: https://integrallabs.co.za/wp-content/uploads/2022/04/SANS241-Drinking-Water-Specifications.pdf)

Water Quality Parameters associated with Acute Health Risks			
Parameter	Unit	Risk	Standard Limit
E. coli / faecal coliforms	(count per 100 mL)	Acute health	Not Detected

Cryptosporidium spp	(count per 10 Litre)	Acute health	Not Detected
Giardia spp	(count per 10 Litre)	Acute health	Not Detected
Nitrate	(mg / L as N)	Acute health	≤ 11
Nitrite	(mg / L as N)	Acute health	≤ 0.9
Combined nitrate plus nitrite	(mg / L as N)	Acute health	≤ 1
Sulphate	(mg / L as SO4)	Acute health	≤ 500
Cyanide (Recoverable)	(µg / L as CN)	Acute health	≤ 200

This study will use the water quality variables of a hydrometry station on the Vaal River and the following algorithms: Nine Score-Based Constructed Bayesian Belief Network (BBN) models to predict nitrate and nitrite values.

1.2 Aim

The aim of this research is, firstly, to determine whether the levels of nitrate and nitrite measurements can be predicted using the combined nitrate and nitrite sensor using Bayesian Belief Networks (BBN) with a learned structure from the data of a water monitoring station on the Vaal River. Secondly, to determine whether replacing the original sensor with a model or models with acceptable prediction error is possible.

1.3 Objectives

1. Prepare the data for use in the machine learning model using existing data collected from the Vaal River water monitoring station, consisting of 1071 data points over 21 years (1995 to 2016).

2. Create models (14 in total) using BBN with various scoring functions and varying feature selections; a Tree-based model, Hill Climb Structured learned models with different scoring functions (Bayesian Information Criterion, Akaike Information Criterion, Log-Likelihood, Enhanced Bayesian Information Criterion and Bayesian General Equilibrium) and a selection of Random Forest (RF) machine learning models to predict the values of the nitrate and nitrate sensor using the remaining features.

3. Apply the model to test data or independent data to compare the measured values to the predicted values and evaluate the models' ability to generalise. Evaluate the model's outputs to determine which are best suited to predict nitrate and nitrite values and if there are features that are correlated and influence the nitrate and nitrite in the data.

1.4 Problem Statement and Rationale

Monitoring water quality can be expensive, depending on the measurements taken, the distance between monitoring stations and how frequently data is collected; these costs can increase further due to the need for more sampling or transportation (Crocker & Bartram, 2014). Because of potential sampling errors, manual sample collection can be more costly and less accurate (Nexsens Technology, 2018).

Predictive modelling using statistical and machine learning approaches provide a monitoring alternative (Barzegar *et al.*, 2020). The basis for this research is replacing the nitrate and nitrite sensors with a model equivalent that provides a similar prediction value with the lowest error and best accuracy at a lower cost in terms of data and equipment needed.

1.5 Structure of Dissertation

Chapter 1 introduces the research, aims, objectives and rationale for the study, while chapter 2 presents the literature review focusing on water quality indexes and parameters, machine learning models including traditional and Bayesian, explainable artificial intelligence and how it is used in water quality and lastly which machine learning models are used to predict different features and which have been successful or not.

The methodology employed in this study is explained in chapter 3, including loading and processing the data, preprocessing the data to remove outliers, correcting formats, visualising the data and distributions, creating the final datasets suitable for use in modelling.

Chapter 4 focusses on the results achieved by examining the kernel density plots, box plots for outliers, determine the correlations of features based on the model structures and directed acyclic graph's, the evaluation of error metrics, error values per samples and modelling runtime.

The significance of the results are discussed in chapter 5, determining if the models are capable of predicting nitrate and nitrite levels to an accurate and acceptable level with the end summarising the study with conclusions and recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The Sustainable Development Goals (SDG), as set by the United Nations, are 17 goals which call for all countries in partnership to take action for the prosperity of the planet and all who inhabit her, with strategies created that tackle climate change, water sustainability, inequality, improvement in health and education, preserving and protecting the forests and oceans, all while ensuring economic growth by 2030 (United Nations, 2023).

Ensuring availability and sustainable management of water and sanitation for all is the objective of SDG Goal 6; with billions of people lacking safely managed drinking water, sanitation and hygiene facilities, ensuring that water is potable is essential now more than ever United Nations, (2023), with appropriate health policies that integrate hygiene and sanitation, to control and prevent World Health Organization, (2023), the transmission of diseases such as viruses: E. coli, Salmonella, Shigella, Vibrio, Entamoeba hystolitica, and helminths such as Ascaris (Chaudhary *et al*., 2023). Potable, serviceable and drinkable water should be free of contaminants, discolouration, undesired tastes and odours and free from bacteria, protozoa, toxic metals, herbicides and pesticides, chemical and radioactive compounds, petroleum-based products and sewage (Shittu *et al*., 2023).

With the increase in the global human population, economic development and an increase in water use, usage up more than seven times in the last century will reduce water availability per person and create more stress on the global ecosystem (Kılıç, 2020). Monitoring the quality of water has become more critical to prevent and control pollution, with research into real-time monitoring with sensors becoming more prevalent (Bansal & Geetha 2020; Zhou *et al*., 2021). Pollutants in water can include asbestos, lead, mercury, nitrates, phosphates, sulphur, oils and petrochemicals Ohm, (2021), including calcium and magnesium. In particular,

alkaline ions such as calcium and magnesium in high concentrations, these dissolved ions contribute to water being classified as hard. Conversely, low concentrations lead to water being classified as soft. Hard water may influence cardiovascular mortality, lime scales in plumbed pipes, water tanks and machinery Orellana *et al*., (2023); however, the water may be left untreated to supplement populations deficient in these minerals (Sengupta, 2013).

Decision-making in water policy and industries utilising water is changing due to advanced technologies such as real-time sensors monitoring water quality, which enable analysts, technicians, and scientists to see the data without sampling and testing it physically. Sensors which are commonly used to monitor water quality parameters include but are not limited to flow, temperature, conductivity, pH, turbidity, oxidation-reduction potential (ORP), dissolved oxygen (DO), total dissolved solids (TDS), nitrate, biological oxygen demand (BOD) and Chlorophyll-a (Chl-a) (Cloete *et al*., 2016; Pujar *et al*., 2020).

Monitoring systems collect large amounts of data which are not adequately analysed due to unfamiliarity with analysis tools, storage, processing and accessibility, and this increases the complexity of the maintenance, monitoring, operation and management of water supplies Velimirovic *et al.*, (2021), with efficient water quality and contamination detection being aided by combining network theory, optimisation and a strategic framework in sensor positioning to improve contamination response capabilities whilst controlling water pollution resulting in a more effective way to manage water quality (Nazempour *et al*., 2018; Dhruba *et al*., 2023).

A solution to solving the maintenance and monitoring of wastewater assets and equipment will require a decision-making platform making extensive use of a variety of machine learning algorithms such as Extreme Gradient Boosting and Logistic Regression will be able to forecast, monitor and predict system failures in advance, identify factors or features that are shown to be predictors of failure such as pipe geometry and pipe age, allow for more accurate decisions from data analysis, energy and cost efficiency and less water wastage by applying modelling and optimisation, to alert management and plant operators of issues before they

happen and make changes proactively and not reactively (Velimirovic *et al.*, 2021; Rahbaralam *et al*., 2020).

2.2 Water quality parameters

The Water Quality Index (WQI) (Equation 1) provides a simple numeric value to express the quality of water in a system at a particular time: a low value (0) represents low quality, and a high value (100) represents high quality. It uses water quality parameters that are available and selectable based on their importance to the system (Sutadian *et al*., 2016; Rubio-Arias *et al*., 2013).

$$WQI = \frac{\sum_{i=1}^{n} Pi * Wi}{\sum_{i=1}^{n} Pi} K$$

Equation 1: The original Water Quality Index (WQI) formula. Wi is the specific weight of each variable (low (1) to high (4)). Pi is the assigned value based on its tolerance range. K is level of contamination when sampling has taken place (1; 0.75; 05) where 1 represents clear water and 0.5 representing a high level of turbidity. *(*Rubio-Arias *et al., 2013)*.

Sutadian *et al*., (2016), discuss 30 different WQIs and the water quality parameters used in each. By identifying the parameters used in the seven most popular WQIs the following are deemed significant (Table 1): dissolved oxygen (DO), faecal coliform (FC), pH, biochemical oxygen demand-5 days (BOD_5), temperature, total phosphorus (TP), nitrates (NO_3), turbidity, total solids (TS) pesticides, toxic compounds, ammonia, total coliform (TC), colour, detergents, hardness, oil and grease, free CO_2, sulphates (SO_4), cyanides, sodium (Na), chloride (Cl), conductivity, magnesium (Mg), calcium (Ca), apparent aspect, ammoniacal

nitrogen, suspended solids (SS), free and saline ammonia and Escherichia coli (E. coli).

The six most popular WQI's , Table 2 are 1) Canadian Council of Ministers of the Environment (CCME, 2017) WQI, used to assess the overall water quality in Canada and countries such as New Zealand, Brazil, India, Vietnam and Iraq CCME, (2017), and aggregating various parameters into a single easy-to-communicate score successful in reporting results to the general public; 2) National Sanitation Foundation (NSF) Index, developed in the 1970s for use in the United States provides a score in the range of 0 to 100, the higher the better, to monitor water quality, Marselina *et al*., 2022); 3) Oregon Index, developed for use in Oregon's rivers by the Oregon Department of Environmental Quality utilised in other countries such as Indonesia, Marselina *et al*., (2022); 4) Bascarón Index, created by Dr. Victor Bascarón relying on weightings of the parameters and used in India, Spain, Korea, Brazil and Argentina, Kumar *et al*., (2024); 5) House Index, a simple method for assessing water quality that does need expert knowledge and used in Spain and the United Kingdom, Kumar *et al*., (2024); and 6) Scottish Research Development Department (SSRD) Index, providing a comprehensive evaluation based on multiple parameters but may not be applicable to other countries but are still used in Iran, Spain, Portugal and Thailand (Kumar *et al*., 2024).

Table 2: A summary of the six most popular WQIs globally and the water quality parameters used to calculate them *(*Sutadian, *et al., 2016)*.

WQI Name	Parameters Selected
Canadian Council of Ministers of the Environment (CCME) Water Quality Index	Select at least 4 parameters
National Sanitation Foundation (NSF) Index	dissolved oxygen (DO), faecal coliform (FC), pH, biochemical oxygen demand-5 days (BOD5),

	temperature, total phosphorus (TP), nitrates (NO3), turbidity, total solids (TS), pesticides, toxic compounds
Oregon Index	Dissolved oxygen (DO), pH, faecal coliform (FC), total solids (TS), nitrates (NO3), ammonia
Bascarón index	pH, biochemical oxygen demand-5 days (BOD5), dissolved oxygen (DO), temperature, total coliform (TC), colour, turbidity, permanganate reduction, detergents, hardness, dissolved oxygen (DO), pesticides, oil and grease, sulphates (SO4), nitrates (NO3), cyanides, sodium (Na), ammonia nitrogen, chloride (Cl), conductivity, magnesium (Mg), phosphorus (P), nitrites (NO2), calcium (Ca), apparent aspect
House Index	Dissolved oxygen (DO), ammoniacal nitrogen, biochemical oxygen demand-5 days (BOD5), suspended solids (SS), nitrate (NO3), pH, temperature, chlorides (Cl), total coliform (TC)
Scottish Research Development Department (SRDD) Index	Dissolved oxygen (DO), ammoniacal nitrogen, biochemical oxygen demand-5 days (BOD5), free and saline ammonia, pH, total oxidized (TO), nitrogen (N), phosphate (P), suspended solids (SS), temperature, conductivity and Escherichia Coli (E. Coli)

In South Africa, Banda & Kumarasamy, (2020), has developed a Universal Water Quality Index which yielded 13 significant water quality parameters: NH_3, Ca, Cl, Chl-a, EC, F, $CaCO_3$, Mg, Mn, NO_3, pH, SO_4 and turbidity based on data collected by six hydrometry stations provided by the Umgeni Water Board (UWB) in a similar format and available parameters as the Vaal River data used in this study. The importance of having a universal index; it is the ability to generalise, allow for more straightforward comparability, less complicated and allow for standardisation as mentioned by, Kumar *et al*., (2024), the challenge with creating a standard WQI lies in the complexity of incorporating different parameters that account for

different locations, parameters of interest and an index that simplifies the communication to policymakers, water managers and the general public.

Reducing the complicated process of calculating the WQI, building models using the available water quality parameters, including biochemical oxygen demand (BOD), dissolved oxygen (DO), electrical conductivity (EC), nitrate, nitrite, phosphate, pH, temperature and turbidity provide an improvement in performance and parameter reduction using the Extra Tree Regressor (ETR) algorithm (Asadollah *et al*., 2021). However, with different water quality variables and machine learning algorithms, more parameters can be utilised simultaneously without varying or complex equations or expert knowledge to provide a simple communicative score solution.

2.3 What is Machine Learning?

Machine learning, a subset of artificial intelligence, allows computers to perform actions without explicitly being programmed to imitate how humans learn (IBM, 2020. An algorithm improves gradually and automatically through experience through data and automatically adjusts its parameters and logic. By gradually making changes with the experience learned in each task and the identified patterns, the algorithm will handle the task without any human intervention – an artificial thinking machine (Anupama *et al.*, 2020). This type of learning is also defined as a machine learning algorithm (Heller, 2023).

Machine learning can be applied to various fields and applications including image processing, email filtering, Natural Language Processing (NLP), Computer Vision (CV), speech recognition and medical diagnosis, and is most useful in areas difficult or infeasible to develop conventional algorithms or systems (Bansal & Geetha, 2020; Heller, 2023), these include the understanding of speech, new medicine discoveries, cancer detection, facial recognition and fraud detection (Tufail *et al*., 2023).

A machine learning model is a computer program optimised on a performance metric using training data that includes parameters, and the learning thereof lies in the execution of this program. These models may be predictive or descriptive if predictions or knowledge gain are the desired outcomes, respectively, with the core task of making an inference on a sample using the theory of statistics in the model-building process (Alpaydin, 2020).

2.3.1 Supervised and unsupervised Learning

Models built using supervised learning algorithms do so by using data that have both the inputs or features and the output or the target, and through an iterative process of optimisation of a loss function, these models' associated patterns are uncovered and applied to new inputs or data to predict a value learned from the training data, Tufail *et al*., (2023), these algorithms can be used to solve classification or regression tasks depending on the problem. Regression is used when the output or target variables are in a numerical range, and classification is when the outputs are fixed or are classed into fixed buckets (Tufail *et al*., 2023)

Models built using unsupervised learning algorithms use data that contains only the inputs and identify patterns in the data to produce groups or clusters as outputs to create modular models capable of predicting algal blooms without human intervention (Ananias *et al*., 2022). Test data are also used in the training process, and these models are considered efficient. At the same time, a machine learning algorithm best performs the predicted output variables from available data. It has been used to predict weather conditions and variables such as temperature without being in direct contact (Yashaswini *et al*., 2021). These models can be both complex or simple depending on the purpose, as complex neural networks require large amounts of data and processing power and might not always be the best fit for real-time predictions (Czerwinski *et al.*, 2021; James *et al*., 2021).

Supervised and unsupervised models are, therefore, two fundamental approaches to learning from data with supervised models such as; 1) random forests (RF) have been successful in achieving good performance and robustness when predicting

water bloom, Liu & Wu, (2017); and 2) Artificial Neural Networks (ANN) and hybrids thereof are capable of dealing with a wide range of water quality parameters with accurate predictions, Chen *et al*., (2020), while unsupervised models such a k-means, hierarchical clustering perform well without relying on large amounts of labelled data (Brill, 2014).

2.3.2 How is machine learning being used for water quality?

Supervised machine learning has become a powerful prediction tool for reliable forecast modelling in recent years. These models have assisted in determining environmental changes in rainfall, water flow, and water levels (Voza, 2020). The collection, measurement, and analysis of water samples provide quantitative information to assist water resource management and alert decision-makers to emerging environmental issues. DO and Chl-a are two water variables used to monitor the water quality in lakes, which help to determine the processes that produce and use oxygen and the concentration of algae in the Small Prespa Lake in Greece (Barzegar *et al.*, 2020).

Low-cost sensors will replace high-cost equipment used for water quality monitoring with better performance and greater mobility, in turn, will allow for areas not previously able to be monitored and reduce the cost while doing so, Liu, *et al*., (2020) with Random Forests machine learning algorithms providing a lower error rate versus linear traditional modelling improving the time and cost of traditional sampling methods (Castrillo & García, 2020).

Wastewater treatment monitoring has been restricted to water quality. While machine learning methods have improved efficiency in these systems, they focus on individual operational aspects and do not use the algorithm's power to determine the relationships between sensor measurements even if, at first, they seem unrelated (Velimirovic *et al*., 2021).

Temperature estimation of electric motors has become efficient using machine learning and deep learning algorithms. However, predicting any value, provided sufficient data, displays an unquestionable advantage in mapping the relationships

between the input and output variables (Velimirovic *et al*., 2021). Neural networks, decision trees, AdaBoost, support vector machines, stochastic gradient descent regressor and ElasticNet algorithms have proven effective in predicting the temperature estimate of motors, allowing the torque to be adjusted to reduce overheating without a direct temperature measuring device (Czerwinski *et al.*, 2021).

Algorithms such as Artificial Neural Networks (ANN) and long short-term memory (LSTM) have been used successfully to predict dissolved oxygen, with both well and reasonably well predictions being made receptively for data from the Burnet River in Australia, Dheda & Cheng, (2020), and neural networks or Seasonal Auto-Regressive Integrated Moving Average (SARIMA) to predict Chlorophyll-a (Chl-a) at three samples sites in the Arabian Gulf, with SARIMA performing poorly and the neural networks performing favourably (Al Shehhi & Kaya, 2020).

Yu *et al*., (2020) proposed a data-driven approach to predict dissolved oxygen (DO) and hypoxia in Chesapeake Bay, USA, using empirical orthogonal functions, feature transformation and automatic selection as inputs for the Neural Network model, showing potential applicability to other coastal systems. At the same time Lu *et al*., (2020) described using an Ensemble Back-Propagation Neural Network (BP-NN) to predict sulphate and nitrate over East Asia, with the estimated values being well observed. A scalable and cluster-based model, as proposed by Agarwal & Xue, (2020), demonstrates an approach to analysing sulphate pollution in rivers in the Ohio watershed in Pennsylvania, USA, with excellent, robust, scalable performance.

With the use of Sentinel-2 satellite data, Chlorophyll-a (chl-a) in the Menor Sea, Spain, was predicted with good model performance using Random Forest (RF), Support Vector Machine (SVM) and Deep Neural Network (DNN) providing the best accuracy while offering a cost-effective near real-time monitoring of a water quality parameter (Gomez *et al*., 2021). Wei *et al*., (2019) investigated the use of traditional modelling techniques (Extreme Learning Machine (ELM), Neural Network (NN)) to predict Chlorophyll-a (chl-a) and found that Particle Swarm

Optimization-Extreme Learning Machine (PSO-ELM) had the best overall predictability.

Keller *et al*., (2018) found that combining multi-sensor hyperspectral data of the Elbe River in Germany with extremely randomised trees (ET), support vector machines (SVM), and artificial neural networks achieved the best results when estimating coloured dissolved organic matter (CDOM), Chlorophyll-a (chl-a), turbidity, diatoms, and green algae concentrations. Enhancing the predictive capabilities of water quality modelling when focussing on bacteria levels in surface waters, Panidhapu *et al*., (2019) show that incorporating various features, pH, turbidity, precipitation, incoming solar radiation and temperature faecal coliforms, and E. coli can be utilised in Bayesian Belief Networks to achieve highly accuracy predictions for E. coli and coliform levels even with incomplete or missing water quality or weather data.

Identifying high-risk areas for nitrate pollution in Lenjanat Plain, Iran, was possible by creating a novel framework of assessed nitrate groundwater contamination combined with the following: Boosted Regression Trees (BRT), Multivariate Discriminant Analysis (MDA) and Support Vector Machine (SVM) (Sajedi-Hosseini *et al*., 2019). Source contamination in water networks is discussed by Ortega *et al*., (2017); by utilising Mixed Integer Linear Programming, an effective mathematical modelling and optimisation solution, an approach to identify source contamination which is both practical and effective accurately is demonstrated.

Lu & Ma, (2020) proposed two hybrid tree-based machine learning models with Random Forest (RF) and XGBoost with Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) denoising. The Random Forest (RF) provided the best predicted results for temperature, dissolved oxygen (DO), conductance while XGBoost was best suited for pH, turbidity, fluorescent dissolved organic matter (FDOM). Successful high accuracy of pH prediction is discussed by Memon *et al*., (2012), by using Artificial Neural Networks (ANN) and 17 different

parameters; the effectiveness of these models is shown and highlighted in assisting with monitoring the quality of the different phases of water in urban systems.

Li *et al*., (2023) evaluate the efficacy of Random Forest, XGBoost, MLP and LightGBM in predicting pH values in Georgia, USA, with results that suggest machine learning algorithms can be practical in water quality parameter prediction and the potential for data-driven approaches in environmental management. General regression neural networks (GRNN) have successfully been shown to be able to predict dissolved oxygen concentration with other algorithms, including domain knowledge modelling systems, fuzzy predictors, neural networks and genetic algorithms, from samples collected from 20 reservoirs in Taiwan (Chou *et al*., 2018).

Water quality parameters, which include pH, temperature, dissolved oxygen, ammonia, nitrate and nitrite and gradient boosting algorithms to create models capable of early detection of fish diseases; identifying parameters that are the most correlated with certain diseases to reduce financial losses and preventing outbreaks to enable a more effective fish farming management framework (Nayan *et al*., 2021).

Wang *et al*., (2021) discussed a stacking ensemble machine learning approach using five different machine learning models (Multiple Linear Regression (MLR), Partial Least Squares (PLS), Random Forest (RF), Bayesian Belief Network (BBN) and Sparse Partial Least Squares (SPLS) with the estimates from each being used at the input for the ensemble using data for three beaches along Lake Erie, New York State, USA with the performance of the ensemble performing the best.

Machine learning algorithms have been used to predict various water quality parameters (Table 3) in varying environments and data, with great success, demonstrating the algorithm's strength in finding patterns and complex relationships in the data. Models such as random forest, support vector machine,

artificial neural network, deep neural network, multiple linear regression, multivariate quadratic regression, back-propagation neural network, extreme learning machine, and genetic algorithms successfully predicting Chlorophyll-a levels (Yu *et al*., 2020; Wei *et al*., 2019. Nitrates and sulphates levels are accurately predicted using back-propagation neural networks, boosted regression trees, multivariable discriminant analysis, support vector machine (Sadjedi-Hosseini *et al*., 2019; Lu *et al*., 2020) while dissolved oxygen, with neural networks, random forest and xgboost denoised models, Yu *et al*., (2020), Lu & Ma, (2020), and lastly ensemble models have been shown to predict accurately, the Feacal Indicator Bacteria parameter (Wang *et al*., 2021.)

Table 3: A summary of the algorithms and water quality parameters in the literature.

ML Algorithms Used	**Parameters Used**	**Target Variable**	**Author/s**
Neural network (NN)	Streamflow	Dissolved Oxygen	Yu *et al.*, (2020)
Random Forest (RF), Support Vector Machine (svmRadial), Artificial Neural Network (ANN), Deep Neural Network (DNN)	Sentinel-2 satellite data, spectral indices	Chlorophyll-a (Chl-a)	Gomez *et al.*, (2021)
Boosted Regression Trees (BRT), Multivariate Discriminant Analysis (MDA), Support Vector Machine (SVM)	Nitrate concentration data from 102 wells	Nitrate	Sadjedi-Hosseini *et al.*, (2019)
Multiple Linear Regression (MLR), Multivariate Quadratic Regression	Salinity, Temperature, Depth, Dissolved Oxygen (DO), pH	Chlorophyll-a (Chl-a)	Wei *et al.*, (2019)

(MQR), Backpropagation-Neural Network (BP-NN), Extreme Learning Machine (ELM), Genetic Algorithm-ELM (GA-ELM), Particle Swarm Optimization-ELM (PSO-ELM), Ant Colony Optimization-ELM (ACO-ELM)			
Ensemble back-propagation neural network (BP-NN)	N/A	Wet depositions of non-sea-salt sulphate and nitrate	Lu *et al.*, (2020)
Random Forest (RF), XGBoost with Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) denoising	N/A	dissolved oxygen, pH, water temperature, specific conductance, turbidity, fluorescent dissolved oxygen demand	Lu & Ma, 2020
Ensemble Models	N/A	Faecal Indicator Bacteria	Wang *et al.*, (2021)

Accurately predicting water quality parameters is not enough to get a full understanding of what the models are doing, the relationships between the variables, and whether they interact in some way that has never been identified before (Xu *et al*., 2019). Therefore, understanding the decisions for a prediction has

become increasingly important for transparency, human understanding, and credibility.

2.4 Explainable Artificial Intelligence

Explainable AI (XAI) is a growing field that aims to address the understanding of the ever-growing-complexity of artificial intelligence (AI) models; the outputs of Deep Neural Networks (DNNs) cannot be explained either by the network itself or by an external component, Xu *et al*., (2019), and these include the CNN, LSTM and RNN architectures which are considered black box and difficult to explain; which are crucial for trust, fairness and compliance. Humans are curious and seek explanations for events, coupled with the reason for a model decision can be as important as the prediction itself, Molnar, (2023), lending to the use of decision trees (Figure 1) as an example which contributes to our understanding of how the model or models made the decision it made.

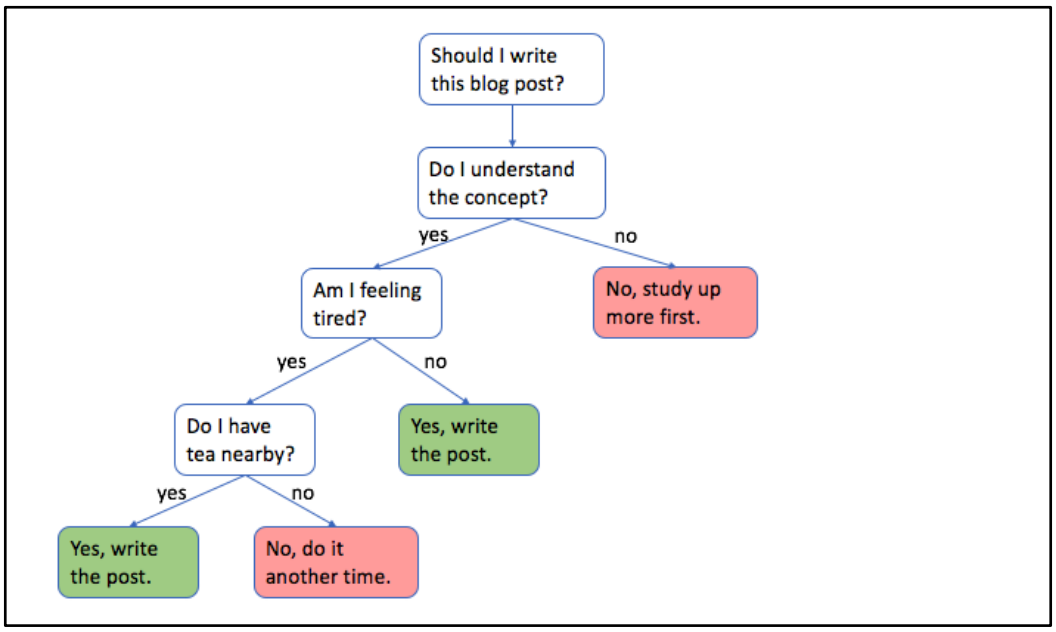


Figure 1: An example of a decision tree, designed with an explainable structure. By starting at the top and going down, level by level the final decisioning can be determined (source: (Xu *et al*., 2019).

A technique now widely used in XAI and machine learning is SHAP (Shapley Additive exPlanations), a model-agnostic framework derived from cooperative game theory, a branch of mathematics originally widely used for economics and now this framework has contributed to areas in medicine, energy and the environment (da Costa, 2023). SHAP fairly distributes the payoff to players in a game depending on a player's contribution, and by relating this concept to machine learning, the following is defined: 1) Each feature or variable is a player, 2) The prediction is the payoff to a player and 3) The Shapley value shows the feature contribution amongst features (players). By breaking down the predictions into the sums of each feature, individual predictions and predictions across the entire dataset can be provided (Lundberg & Su-in, 2017; da Costa, 2023).

Madni *et al*., (2023) propose an AutoML (Automatic Machine Learning) and XAI techniques for modelling water quality predictions using a freely available dataset containing pH, hardness, solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity and the target variable potability. Missing data are imputed and used in the following models: RF, GBM, Logistic Regression (LR) and a stacked ensemble with a 97% accuracy in determining safe and harmful drinking water. Madni et al. (2023) demonstrated that using SHAP, sulphate, and pH contribute significantly to predicting if water is unsafe for drinking. Thrun, Ultsch, & Breuer, (2021), provide an XAI framework for time series data consisting of variables: soil temperature, groundwater level, soil moisture, rainfall, discharge, electrical conductivity, solar radiation, air temperature stream water temperature and nitrate. This framework is independent of previous knowledge of the data and provides meaningful explanations of the relationships within the data.

Challenges remain despite the advancements in XAI. An example is the computational cost of calculating SHAP values, which can be prohibitive for large datasets and models. The more complex the model, the lower the level of transparency; it is, therefore, essential to balance interpretability, computational cost, and model performance (Rudin, 2019) while ensuring there is no sacrifice to model performance and accuracy utilising BBNs. These can be an alternative to

SHAP values since they inherently encode relationships that users can easily interpret (Afrabandpey *et al*., 2020).

2.5 Summary of algorithms in water quality

While machine learning has been successful in predicting various water quality variables through the use of different algorithms and varying data sets, there is still a need for explainability; traditional models with XAI frameworks are a start and an ever-growing field for contribution, the BBN models fundamentally allow the examining of the effects on the target variable to the rest of the features at a low computational cost and by learning the underlying structure of data applied to predictions and contributions or the correlations of these variables have at a low computation cost and high explainability in its simplicity.

Chapter 3

METHODOLOGY

3.1 Introduction

This research uses Bayesian and Frequentist machine-learning algorithms to predict nitrate sensor values from existing water monitoring variables such as calcium (Ca), chloride (Cl), dissolved major salts (DMS), electrical conductivity (EC), fluoride (F), potassium (K), magnesium (Mg), ammonium (NH_4), nitrate (NO_3) and nitrite (NO_2), sodium (Na), phosphate (PO_4), sulphate (SO_4), silicon (Si), total alkalinity (TAL) and temperature (Celsius). The machine-learning life cycle, the Python programming language, the R language and selected structured learning Bayesian machine-learning algorithms with a combination of the columns of available data, and as such, the naming conventions indicating their scoring functions, scoring functions as discussed by Liu *et al*., (2012), in the importance of selecting appropriate scoring functions during model selection to account for different data sizes and conditions including smaller datasets. These models will be tested to determine their ability to predict the concentration of nitrate and nitrites in the absence of the sensor.

In the experiments, Score-Based structure learning was explored to construct a model to predict nitrate levels and determine water quality. The methods included Tree-Based and Random Forest models, which used the following features: calcium, chloride, dissolved major salts, electrical conductivity, fluoride, potassium, magnesium, nitrate and nitrite, sodium, phosphate, sulphate, silicon, and total alkalinity, with the latter employing varying number of trees. Hill Climb (HC) algorithms were extensively tested with different selection criteria: Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), Log-Likelihood (Loglik), Enhanced Bayesian Information Criterion (EBIC), and Bayesian General Equilibrium (BGE), both with all features and specific feature subsets like Random-Features (RF) and Correlated-Features (CR).

3.2 General Setting and Study Site

The sample monitoring station (C2H018) at which the data were collected is at coordinates S26.970278^{o}, E27.211111^{o} on the Vaal River, where the borders of North West and Free State provinces meet (Figure 2). The area's climate is considered warm and temperate, with temperatures averaging 17 degrees Celsius and yearly rainfall of 669mm (Climate Data, 2021). The Vaal River, a tributary of the Orange River, start near the town of Breyten, Mpumalanga, flowing 1210km in a southwest direction, with the middle section forming most of the Free State's northern provincial boundary (Britannica, 2024).

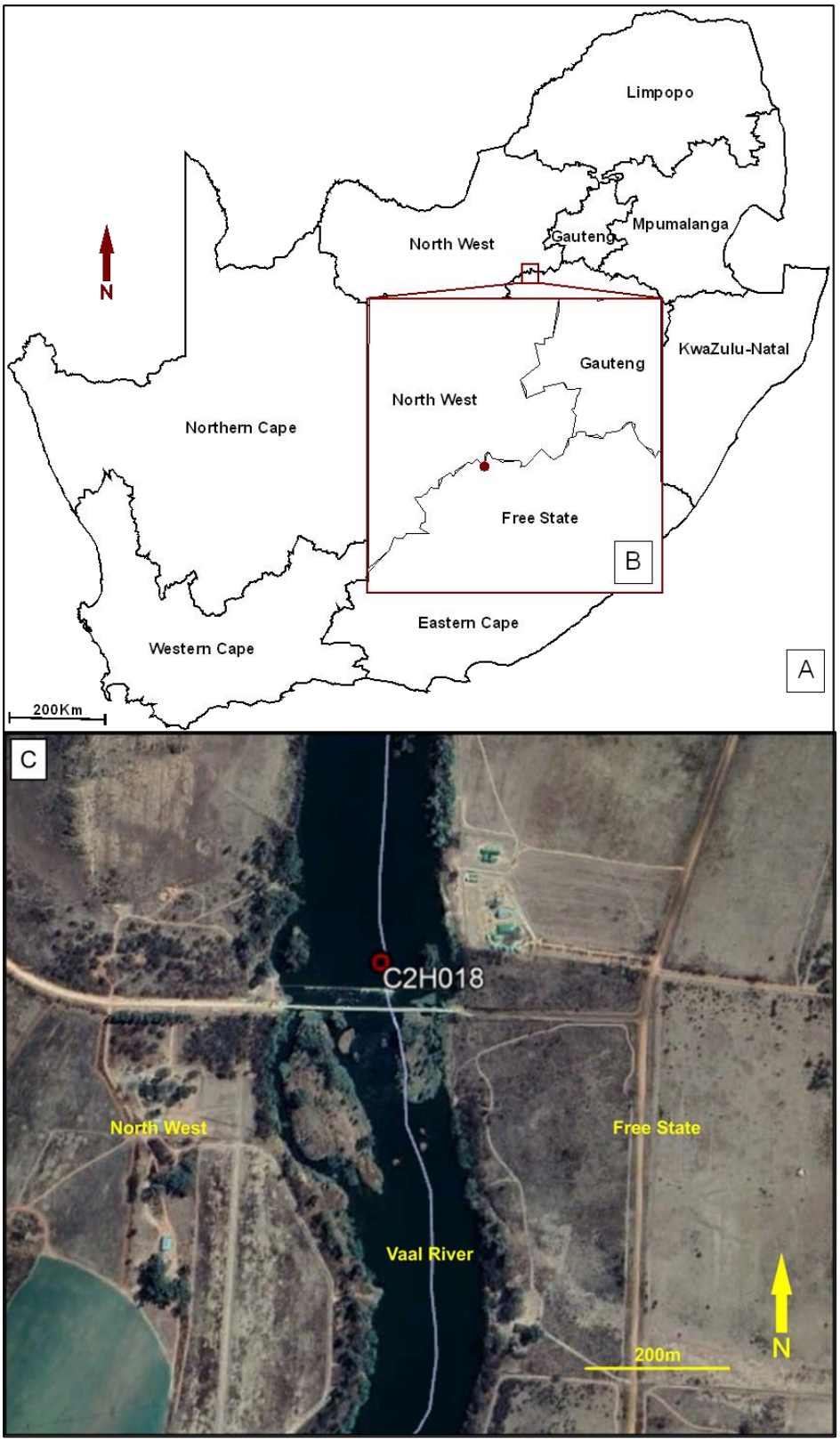


Figure 2: A) Monitoring sample station (C2H018) in South Africa. B) The approximate location of the site. C) The satellite image of the station on the Vaal River. (Data Sources: https://data.humdata.org and Google Earth).

3.3 Machine-learning life cycle

When building machine-learning models, the process is iterative or cyclical, meaning the data processing, modelling and optimisations will need to be done as often as necessary until the model's error metric or difference between actual and predicted is as small as possible. For this reason, the machine-learning life cycle framework is used to build a model to determine if nitrate can be predicted (Yip, 2020; Machine Learning Life Cycle, 2021).

The processes of the machine-learning life cycle (MLLC) are broken down into the following six steps:

1. Data access and collection
2. Data preparation and exploration
3. Model, build and train
4. Model evaluation
5. Model deployment
6. Model monitoring

This study utilises steps one to four as the framework to prepare the collected data and train and evaluate the models. In contrast steps five and six are excluded since deploying a working model using new data and monitoring the performance are out of scope. The framework can be broken down to explain the steps (Figure 3) further; starting at the top and working down, the dataset is processed into a suitable format for modelling, with the data split into random records that will be used to train, validate and test the structural learning of the models and the performance respectively. The bnlearn library for Bayesian Network learning is used in the structural learning and cross-validation (CV) steps in the flow chart. Further evaluation metrics determine which model is the most accurate and best for the study.

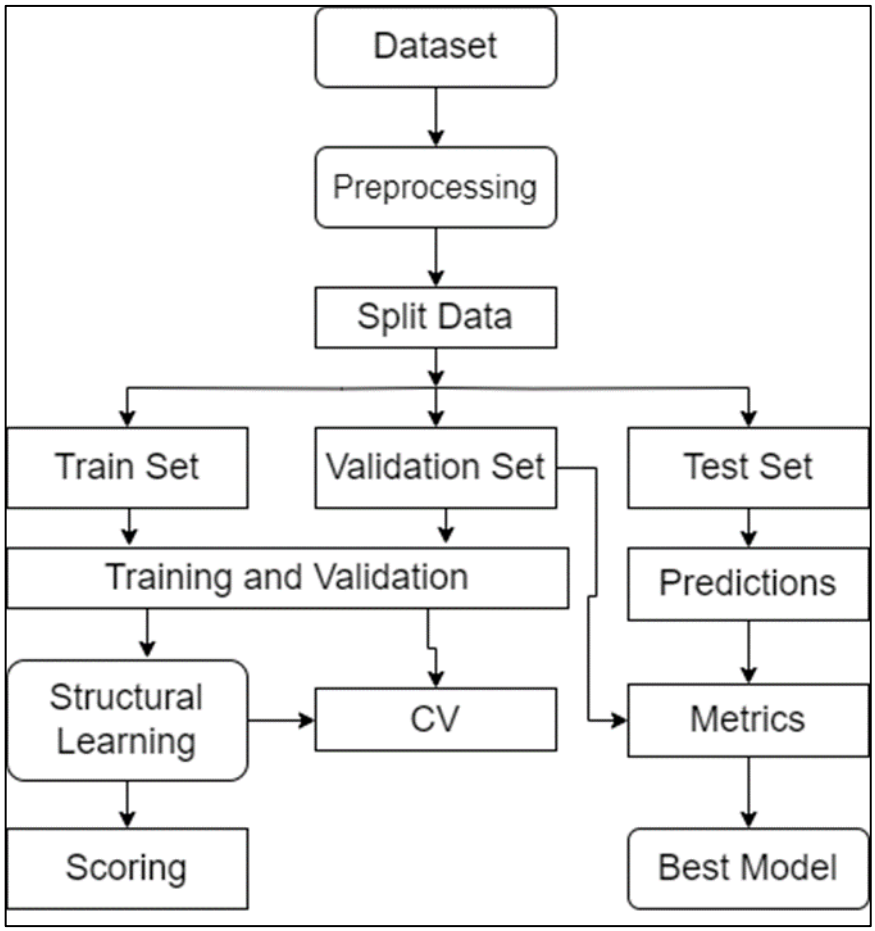


Figure 3: Flow chart of the data and modelling process as described in the research methodology, summarising the steps as taken from the machine-learning life cycle framework but include the modelling process using the Bnlearn R package.

3.4 Collecting and accessing the data for processing

The previously collected water monitoring data for station C2H018 is stored as a Microsoft Excel file and was accessed and edited for use in the steps following in the MLLC. The data were accessed and pre-processed, including splitting into a train set, validation set and test set (70%, 20%, and 10%, respectively), using the Python programming language and Jupyter Notebooks. At the same time, the modelling, graphs and evaluations were done using the R language and R Studio. The Python programming and R language provide the base for the methodology, since R as well as the use of libraries allow for the loading of data, analysis, modelling, optimising, and evaluation. The Python programming language is open-source, allowing for easy readability, reproducibility, cross-system and hardware

compatibility (Windows, Linux, MacOS, x86, x64 and ARM) and access to an extensive function library list (Python Software Foundation, 2022). The R language is widely used for statistical tasks, modelling, and the ability to produce graphical outputs that can include complex mathematical symbols and formulas (The R Foundation, 2023).

The use of Jupyter Notebooks and R Studio allows for all the steps and coding required for this research, such as data preparation and loading, analysis, modelling, optimisation, and evaluation, to be in multiple scripts for reproducibility, error or logic checking, and version control (Project Jupyter, 2023; Posit Software, 2023).

The previously collected water monitoring data for station C2H018, contains 66 columns and 1078 rows of data with the first seven rows included data from a different monitoring station and were ignored when loaded; once loaded there are 1071 rows of data. The data are in a familiar tabular format (Table 4).

Table 4: A sample of the data once loaded for monitoring station C2H018.

Monitoring Point ID	Monitoring Point Name	Latitude	Longitude	Located on Feature Name	Located on Type	Drainage Region Name	Feature Reference Code	Monitoring Active	Sample Start Date	...	Si-Diss-Water (SILICON) (mg/L) Detection Limit
90626	C2H018Q01 VAAL RIVER AT DE VAAL/SCHOEMANSDRIFT	-26.970278	27.211111	VAALRIVIER C11A TO C92C	Rivers	C23L	NaN	Yes	1995-01-03	...	0.4
90626	C2H018Q01 VAAL RIVER AT DE VAAL/SCHOEMANSDRIFT	-26.970278	27.211111	VAALRIVIER C11A TO C92C	Rivers	C23L	NaN	Yes	1995-01-10	...	0.4
90626	C2H018Q01 VAAL RIVER AT DE VAAL/SCHOEMANSDRIFT	-26.970278	27.211111	VAALRIVIER C11A TO C92C	Rivers	C23L	NaN	Yes	1995-01-17	...	0.4

3.5 Preparing and exploring the data for processing

The water quality parameters are a subset of the 66 columns in the data and were selected separately along with any time and identification fields. The resulting data consisted of 17 columns with 1071 rows (Table 5). Data for the target variable nitrate and nitrite were missing in 61 records of the data and were removed, creating a final dataset before further processing of 1010 rows with 17 columns.

Table 5: A sample of the data for monitoring station C2H018 after water parameter selection.

Sample Start Date	Sample Start Time	Ca-Diss-Water (CALCIUM) (mg/L) Result	Cl-Diss-Water (CHLORIDE) (mg/L) Result	DMS-Tot-Water (DISSOLVED MAJOR SALTS) (mg/L) Result	EC-Phys-Water (ELECTRICAL CONDUCTIVITY) (mS/m) Result	F-Diss-Water (FLUORIDE) (mg/L) Result	K-Diss-Water (POTASSIUM) (mg/L) Result	Mg-Diss-Water (MAGNESIUM) (mg/L) Result
1995-01-03	13:59:00	32.1	21.8	278.0	41.3	0.44	6.68	10.6
1995-01-10	12:25:00	30.7	24.7	268.0	39.3	0.37	6.99	10.5
1995-01-17	13:36:00	37.4	25.9	319.0	44.8	0.36	7.85	13.2
1995-01-17	14:40:00	37.1	27.9	322.0	44.4	0.39	8.00	13.7
1995-01-24	13:30:00	34.9	23.6	287.0	42.2	0.47	6.24	12.0

The data was prepared in a way that is suitable for model consumption and included several steps: 1) The column headings were renamed to remove characters such as brackets, forward-slash or to reduce the length (Figure 4), 2) Features with missing data were identified since these features were imputed with the average value for that column.

```
'Ca-Diss-Water (CALCIUM) (mg/L) Result': 'calcium'
,'Cl-Diss-Water (CHLORIDE) (mg/L) Result': 'chloride'
```

Figure 4: A code snippet changing the column headings for calcium and chloride.

Visualising the data was handled using the skimr library and R's built-in summary function within R Studio. Visualising provides a quick view of the distribution of the data, identifying outliers or values that are substantially varied from the rest of the data.

3.6 Preprocessing of the Modelling Data

3.6.1 Outlier Removal

An initial outlier removal process was done on the data using a technique called the z-score method. By calculating the standard deviations of the data points in relation to the mean of the dataset and selecting a cut-off of two standard deviations, data points that fall outside of this range are removed, and so this is a widely used tool to manage data quality (Zar, 2010). Removal of these points is important to the models' stability and generalised learning ability and in predicting points in new and unseen data. The result of the outlier removal reduced the overall dataset further to 970 rows and 17 columns.

3.6.2 Splitting of the data

The data are split randomly into three subsets of data in the following percentages: train set (70%), validation set (20%) and test set (10%), equating to 682, 192 and 96 rows of data, respectively. These splits help assess if the models can generalise on new and unseen data by learning the patterns rather than memorising only the data in the train set. The validation set assists with model parameter tuning and stability, while the test set provides an unbiased view of performance (James *et al*, 2019).

3.6.3 Data Summaries

The summaries for the datasets provide a view of data distributions, the number of missing data per feature, the mean value per feature, the standard deviation (sd) and the range of values per feature for the training, validation and test sets, respectively.

In Table 6, the train set, there were missing data for each feature except for nitrate and nitrite (no3andno2), with temperature, DMS, potassium and phosphate missing the most information. The histograms display that ammonium's values are small and skewed to one side; this was a problem for the model since no pattern can be learned.

In Table 7, the validation set, the proportions of missing data were the same as in the test set, with temperature, DMS, and fluoride missing the most information. The histograms are similar to the train set, but silicon had ranges that were not in the same distribution. Too much deviation in the data sets introduces noise and hampers the learning process.

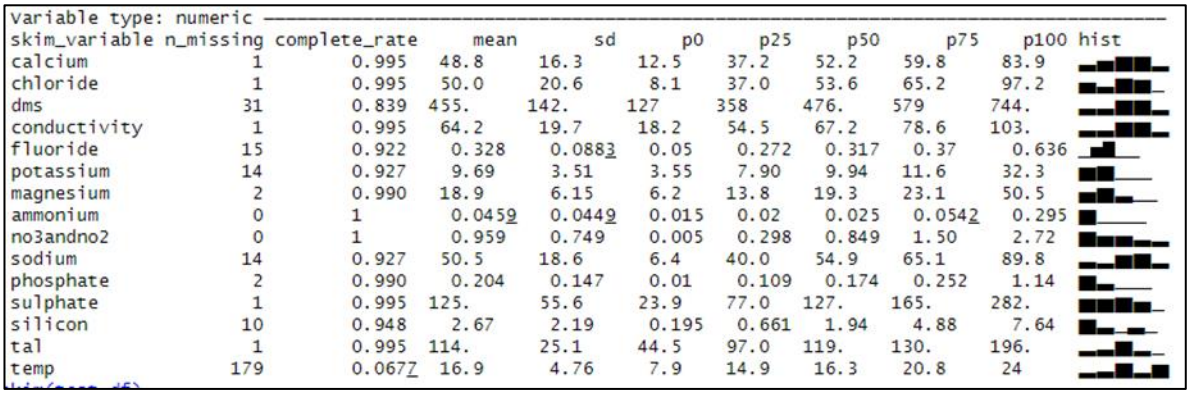
In Table 8, the test set, the proportions of missing data were different when compared to the validation and train sets, with temperature and DMS missing the most information. The histograms are similar to the train set, but TAL and sodium had ranges that are not in the same distribution; too much deviation in the data sets introduces noise and hampers the learning process.

Table 6: Statistical summary for the train set.

— Variable type: numeric ——

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	calcium	6	0.991	50.2	17.1	11.9	38.1	53.1	63.4	88.5	
2	chloride	4	0.994	51.5	21.4	8.1	36.4	55.4	68.2	97.2	
3	dms	92	0.865	462.	149.	125	357.	491.	581.	744.	
4	conductivity	3	0.996	65.4	20.4	18.2	52	68.6	82	105.	
5	fluoride	42	0.938	0.336	0.0931	0.025	0.28	0.33	0.38	0.8	
6	potassium	46	0.933	9.82	3.39	3.26	7.57	10.1	11.9	32.3	
7	magnesium	10	0.985	19.1	6.20	2.35	14.4	19.8	23.6	50.5	
8	ammonium	1	0.999	0.0575	0.346	0.015	0.02	0.025	0.051	9	
9	no3andno2	0	1	0.962	0.759	0.005	0.292	0.842	1.52	2.80	
10	sodium	46	0.933	51.0	19.9	6.4	35.2	55.5	67.0	90.1	
11	phosphate	7	0.990	0.205	0.144	0.005	0.111	0.177	0.256	1.39	
12	sulphate	6	0.991	128.	56.8	20.7	81.7	132.	175.	282.	
13	silicon	29	0.957	2.73	2.22	0.2	0.728	1.94	4.88	13.3	
14	tal	9	0.987	116.	26.9	43.4	95.4	122.	133.	248.	
15	temp	632	0.0733	18.8	5.23	6.5	15.9	20.0	22.7	27.7	

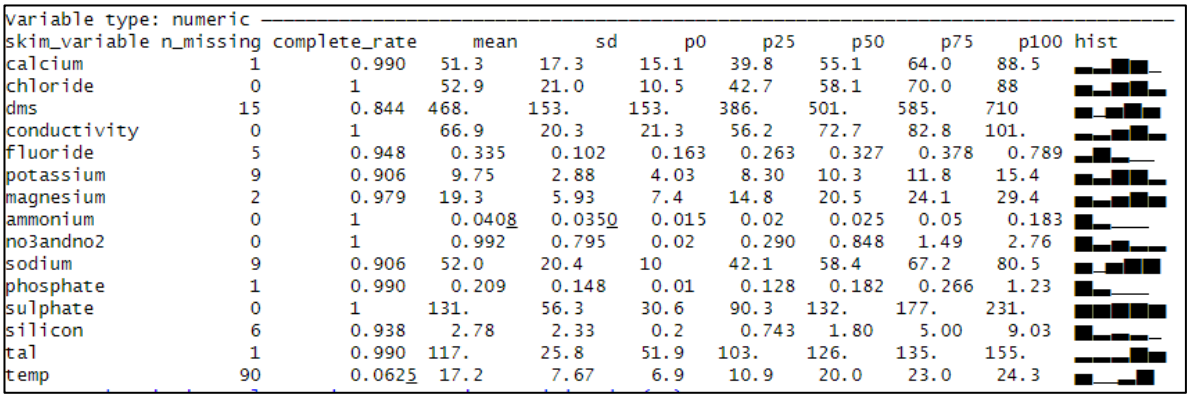
Table 7: Statistical summary of the validation set.



Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
calcium	1	0.995	48.8	16.3	12.5	37.2	52.2	59.8	83.9	
chloride	1	0.995	50.0	20.6	8.1	37.0	53.6	65.2	97.2	
dms	31	0.839	455.	142.	127	358	476.	579	744.	
conductivity	1	0.995	64.2	19.7	18.2	54.5	67.2	78.6	103.	
fluoride	15	0.922	0.328	0.0883	0.05	0.272	0.317	0.37	0.636	
potassium	14	0.927	9.69	3.51	3.55	7.90	9.94	11.6	32.3	
magnesium	2	0.990	18.9	6.15	6.2	13.8	19.3	23.1	50.5	
ammonium	0	1	0.0459	0.0449	0.015	0.02	0.025	0.0542	0.295	
no3andno2	0	1	0.959	0.749	0.005	0.298	0.849	1.50	2.72	
sodium	14	0.927	50.5	18.6	6.4	40.0	54.9	65.1	89.8	
phosphate	2	0.990	0.204	0.147	0.01	0.109	0.174	0.252	1.14	
sulphate	1	0.995	125.	55.6	23.9	77.0	127.	165.	282.	
silicon	10	0.948	2.67	2.19	0.195	0.661	1.94	4.88	7.64	
tal	1	0.995	114.	25.1	44.5	97.0	119.	130.	196.	
temp	179	0.0677	16.9	4.76	7.9	14.9	16.3	20.8	24	

Table 8: Statistical summary of the test set.



Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
calcium	1	0.990	51.3	17.3	15.1	39.8	55.1	64.0	88.5	
chloride	0	1	52.9	21.0	10.5	42.7	58.1	70.0	88	
dms	15	0.844	468.	153.	153.	386.	501.	585.	710	
conductivity	0	1	66.9	20.3	21.3	56.2	72.7	82.8	101.	
fluoride	5	0.948	0.335	0.102	0.163	0.263	0.327	0.378	0.789	
potassium	9	0.906	9.75	2.88	4.03	8.30	10.3	11.8	15.4	
magnesium	2	0.979	19.3	5.93	7.4	14.8	20.5	24.1	29.4	
ammonium	0	1	0.0408	0.0350	0.015	0.02	0.025	0.05	0.183	
no3andno2	0	1	0.992	0.795	0.02	0.290	0.848	1.49	2.76	
sodium	9	0.906	52.0	20.4	10	42.1	58.4	67.2	80.5	
phosphate	1	0.990	0.209	0.148	0.01	0.128	0.182	0.266	1.23	
sulphate	0	1	131.	56.3	30.6	90.3	132.	177.	231.	
silicon	6	0.938	2.78	2.33	0.2	0.743	1.80	5.00	9.03	
tal	1	0.990	117.	25.8	51.9	103.	126.	135.	155.	
temp	90	0.0625	17.2	7.67	6.9	10.9	20.0	23.0	24.3	

3.6.3.1 Imputation of Missing Data

Missing data can be present in the data for a variety of reasons: collection errors, no information, and intentional omissions. Handling these data points can impact the analysis or outcomes, such as model performance and accuracy. Approaches to dealing with missing data can include removing the records, removing the features, simple imputations (mean, mode or median), and more advanced imputations using algorithms (Emmanuel *et al.*, 2021).

The Expectation-Maximization (EM) algorithm better represents the data, providing an effective option for handling missing data. This algorithm utilises a two-step process: one for the expectation step and one for maximisation. Starting

from an initial estimate, the steps alternate until there is little change in values between successive iterations (Dempster *et al*., 1977).

The imputation process can introduce negative values as a replacement value, identified in the data by the negative red values in the summaries. Since the presence of a variable cannot be negative, the absolute values were taken to ensure the values for silicon were correct since there were no negative temperatures recorded originally in the data, and these were taken to be absolute as well.

3.6.3.2 Imputation Data Summaries

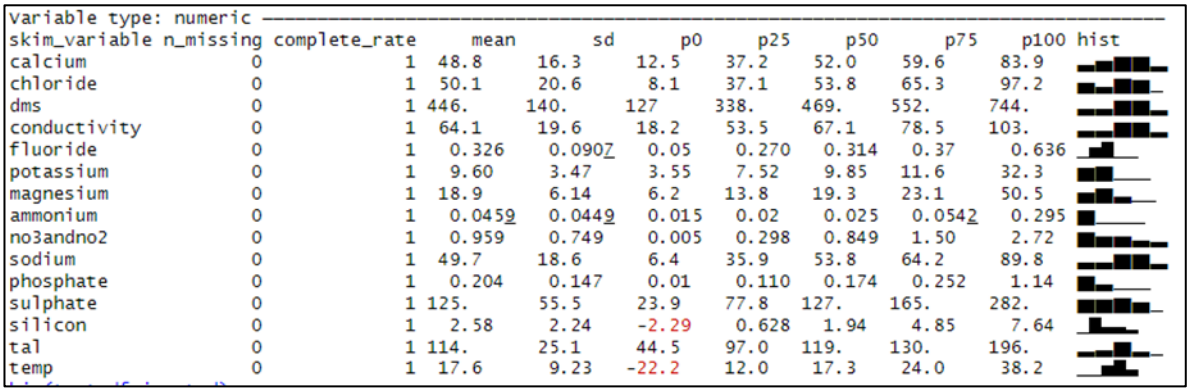
Table 9, the imputed train set, with no missing data; Table 10, the imputed validation set; and Table 11, the imputed test set, were the final data sets used for the model training step. Even though the histograms were similar, additional checks were made to ensure the split data sets did not include features or values that deviated too much from the rest of the data.

Table 9: Statistical summary of the imputed train set.

── Variable type: numeric ──

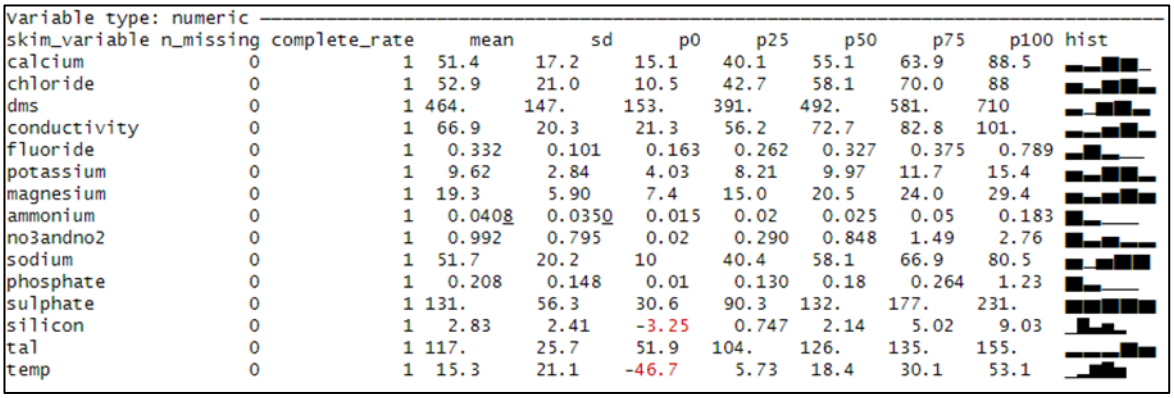
	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	calcium	0	1	50.3	17.0	11.9	38.2	53.2	63.3	88.5	
2	chloride	0	1	51.6	21.4	8.1	36.7	55.4	68.2	97.2	
3	dms	0	1	456.	147.	125	354.	482.	574.	744.	
4	conductivity	0	1	65.3	20.4	18.2	51.9	68.6	82	105.	
5	fluoride	0	1	0.333	0.0944	0.025	0.276	0.328	0.38	0.8	
6	potassium	0	1	9.73	3.35	2.34	7.51	10.0	11.8	32.3	
7	magnesium	0	1	19.1	6.21	2.35	14.5	19.8	23.6	50.5	
8	ammonium	0	1	0.0580	0.346	0.015	0.02	0.025	0.051	9	
9	no3andno2	0	1	0.962	0.759	0.005	0.292	0.842	1.52	2.80	
10	sodium	0	1	50.8	19.7	6.4	35.2	55.3	66.3	90.1	
11	phosphate	0	1	0.206	0.145	0.005	0.112	0.178	0.258	1.39	
12	sulphate	0	1	128.	56.7	20.7	81.3	132.	175.	282.	
13	silicon	0	1	2.72	2.21	-2.33	0.750	1.96	4.83	13.3	
14	tal	0	1	116.	26.8	43.4	95.5	122.	133.	248.	
15	temp	0	1	21.3	9.05	-1.38	16.3	21.3	25.9	173.	

Table 10: Statistical summary of the imputed validation set.



skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
calcium	0	1	48.8	16.3	12.5	37.2	52.0	59.6	83.9	
chloride	0	1	50.1	20.6	8.1	37.1	53.8	65.3	97.2	
dms	0	1	446.	140.	127	338.	469.	552.	744.	
conductivity	0	1	64.1	19.6	18.2	53.5	67.1	78.5	103.	
fluoride	0	1	0.326	0.0907	0.05	0.270	0.314	0.37	0.636	
potassium	0	1	9.60	3.47	3.55	7.52	9.85	11.6	32.3	
magnesium	0	1	18.9	6.14	6.2	13.8	19.3	23.1	50.5	
ammonium	0	1	0.0459	0.0449	0.015	0.02	0.025	0.0542	0.295	
no3andno2	0	1	0.959	0.749	0.005	0.298	0.849	1.50	2.72	
sodium	0	1	49.7	18.6	6.4	35.9	53.8	64.2	89.8	
phosphate	0	1	0.204	0.147	0.01	0.110	0.174	0.252	1.14	
sulphate	0	1	125.	55.5	23.9	77.8	127.	165.	282.	
silicon	0	1	2.58	2.24	-2.29	0.628	1.94	4.85	7.64	
tal	0	1	114.	25.1	44.5	97.0	119.	130.	196.	
temp	0	1	17.6	9.23	-22.2	12.0	17.3	24.0	38.2	

Table 11: Statistical summary of the imputed test set.



skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
calcium	0	1	51.4	17.2	15.1	40.1	55.1	63.9	88.5	
chloride	0	1	52.9	21.0	10.5	42.7	58.1	70.0	88	
dms	0	1	464.	147.	153.	391.	492.	581.	710	
conductivity	0	1	66.9	20.3	21.3	56.2	72.7	82.8	101.	
fluoride	0	1	0.332	0.101	0.163	0.262	0.327	0.375	0.789	
potassium	0	1	9.62	2.84	4.03	8.21	9.97	11.7	15.4	
magnesium	0	1	19.3	5.90	7.4	15.0	20.5	24.0	29.4	
ammonium	0	1	0.0408	0.0350	0.015	0.02	0.025	0.05	0.183	
no3andno2	0	1	0.992	0.795	0.02	0.290	0.848	1.49	2.76	
sodium	0	1	51.7	20.2	10	40.4	58.1	66.9	80.5	
phosphate	0	1	0.208	0.148	0.01	0.130	0.18	0.264	1.23	
sulphate	0	1	131.	56.3	30.6	90.3	132.	177.	231.	
silicon	0	1	2.83	2.41	-3.25	0.747	2.14	5.02	9.03	
tal	0	1	117.	25.7	51.9	104.	126.	135.	155.	
temp	0	1	15.3	21.1	-46.7	5.73	18.4	30.1	53.1	

3.7 Prepare the model for training and testing

Building models require the correct machine learning algorithms and features for a specific problem or use case. The data used in this study has the target feature available, nitrate. Therefore, a supervised approach will be used (Yip, 2020; Machine Learning Life Cycle, 2021).

The models chosen for this study are a series of advanced ML models, including Tree-Based All-Features (AF), Hill Climb (HC) with various criteria like Bayesian Information Criterion (BIC) and Random Features (RF), Random Forest models with All-Features and 100, 500, 1000, 1500 trees and an optimised Random Forest Model which selects the best parameters producing the lowest error based on the

following selection of parameters 1) Trees: 500, 1000, 500, 2) Minimum number of observations per node: 5, 10, 15 and 3) Max depth: 10, 20, 30).

3.7.1 Data Set Distribution Evaluation

To ensure reproducibility, proportional representation of the data, consistent model evaluation and generalisation of the models, and maintaining the same distribution across the datasets are essential steps in the data process (Encord, 2023).

3.7.2 Kernal Density Estimation

This is a non-parametric method for estimating the probability density of a variable without assuming the distribution of given data; it is able to estimate the probability distribution of the data (Jiang *et al*., 2019). Each feature in the train, validation, and test sets was plotted (Figure 5) to ensure the splits of data were as close as possible in distribution using the ggplot2 library in R.

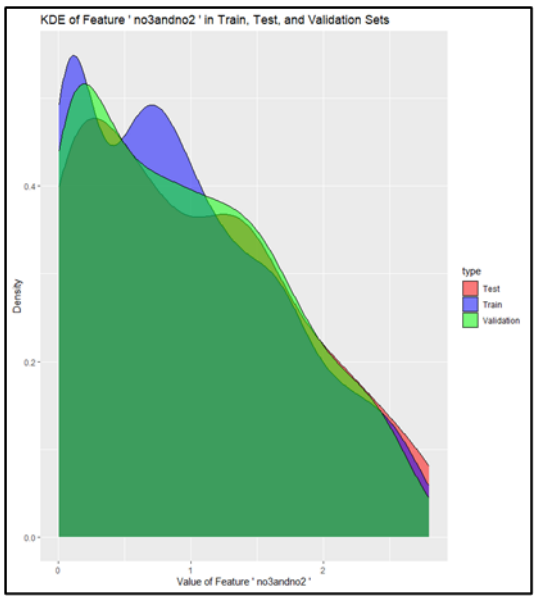


Figure 5: The kernel density estimation plot for nitrate and nitrite.

3.7.3 Box Plots

As part of the data processing, Box plots (Figure 6) offer another visual representation of the data distribution. The plots highlight the interquartile range (IQR) representing 50% of the data and a median line. Points are individually plotted if they are considered outliers, and these can easily be seen on the plots, enabling decisions that can later assist with improving model accuracy and further data processing (Brase & Brase, 2012).

Features with ranges of values vastly different from each data set and too many points considered outliers, outside of the initial outlier removal of two standard deviations per feature, were excluded from the modelling process and plotted using the built-in graphics library in R.

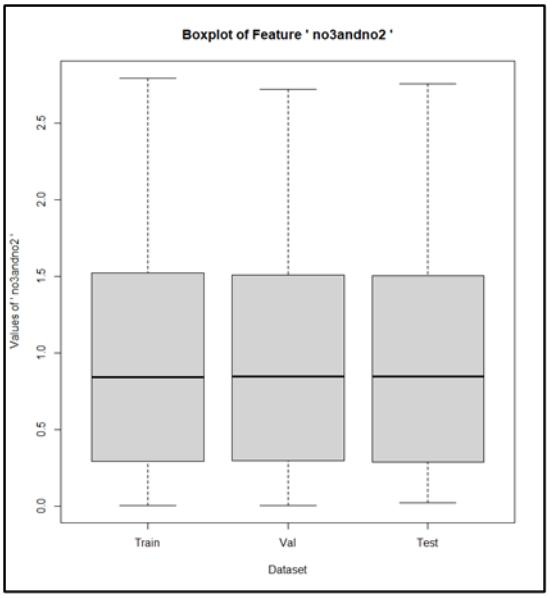


Figure 6: The boxplot for nitrate and nitrite.

3.8 Bayesian Score-based Structured Learning Models

Most models are built using structured learning algorithms with varying scoring functions suitable for continuous data using the bnlearn R package. First released in 2007, the library is a continuously updated tool for Bayesian network analysis. It

supports various Bayesian network types, including discrete, continuous, and mixed data. The package features multiple structure learning algorithms (constraint-based, score-based, hybrid) and local discovery algorithms. It offers diverse parameter learning methods, conditional independence tests, and network scoring options. bnlearn enables precise inference, model validation, and expert knowledge integration. It also provides advanced plotting and import and export capabilities across different file formats and R packages, making it a versatile and comprehensive solution for Bayesian network analysis (Scutari, 2023).

The modelling process for all models irrespective of their algorithm or scoring function were as follows:

1. A model was trained on the train set, previously split by a seed value for reproducibility.
2. A model's performance was calculated using cross-validation over ten folds, which assists in determining its stability.
3. The trained model's predictions on the test set were compared to the actual values

The Hill Climbing (HC) Score-based learning algorithm is used for the training phase of the process Hill Climbing Score-based Learning Algorithm which employs a greedy search with removal and reversals of the directed acyclic graphs by single-arc additions with random restarts to avoid local optima. When optimised, the function builds models faster by reducing duplicated tests (Scutari, 2023).

The models were labelled to keep track of the learning algorithm, scoring function, and features or parameters used, and they followed this sequence: learning algorithm – scoring function – features used.

3.8.1 Tree-Based All-Features (Tree-Based-AF) Model

The AF model utilised all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

The modelling process was as described above but with the depth of the graphs set to equal one, the model represented a structure similar to a tradition decision tree.

3.8.2 Hill Climb - Bayesian Information Criterion - Random Features (HC-BIC-RF) Model

This RF model utilised features selected at random from the pre-processing of the data; chloride, sodium, silicon, calcium, sulphate, DMS, potassium, TAL, phosphate, conductivity and the target feature: nitrate and nitrite.

This model follows the process of using the Hill Climbing algorithm but makes use of the Bayesian Information Criterion (BIC) as the scoring function; The BIC (Equation 2) is equivalent to the Minimum Descript Length (MDL) (Scutari, 2023), which approaches the networks as an information theoretic task, encoding the data into two parts and applying penalty which is larger than that of most other scoring functions. This leads to optimal networks being sparser than other scoring functions (Liu *et al*., 2012).

$$PenaltyMDL(X_i, B, D) = \frac{\log N * p_i}{2},$$

Equation 2: The MDL/BIC scoring function (Liu *et al*., 2012).

3.8.3 Hill Climb - Akaike Information Criterion - Random Features (HC-AIC-RF) Model

This RF model utilised features selected at random from the pre-processing of the data; sodium, silicon, TAL, DMS, sulphate, calcium, chloride, potassium, phosphate, conductivity and the target feature: nitrate and nitrite.

This model follows the process of using the Hill Climbing algorithm but makes use of the Akaike's Information Criterion (AIC) as the scoring function; The AIC (Equation 3) is similar to BIC but with a lower penalty which favours more complex networks (Liu *et al*., 2012).

$$PenaltyAIC(X_i, B, D) = p_i.$$

Equation 3: The AIC scoring function (Liu *et al*., 2012).

3.8.4 Hill Climb - Log-Likelihood – All Features (HC-Loglik-AF) Model

This AF model utilised all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

This model follows the process of using the Hill Climbing algorithm but makes use of log-likelihood (Loglik) as the scoring function; Loglik is an equivalent metric to BIC but with a lower penalty which favours more complex networks (Liu *et al*., 2012).

3.8.5 Hill Climb - Akaike Information Criterion – All Features (HC-AIC-AF) Model

This AF model utilised all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

This model follows the process of using the Hill Climbing algorithm but makes use of AIC as the scoring function.

3.8.6 Hill Climb - Bayesian Information Criterion – All Features (HC-BIC-AF) Model

This AF model utilised all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

This model follows the process of using the Hill Climbing algorithm but makes use of BIC as the scoring function.

3.8.7 Hill Climb - Extended Bayesian Information Criterion – All Features (HC-EBIC-AF) Model

This AF model utilised all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

This model follows the process of using the Hill Climbing algorithm but makes use of EBIC as the scoring function; this includes a second penalty to BIC to further penalise dense networks (Scutari, 2023).

3.8.8 Hill Climb - Bayesian General Equilibrium – All Features (HC-BGE-AF) Model

This AF model utilised all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

This model follows the process of using the Hill Climbing algorithm but makes use of BGE as the scoring function; a score equivalent to the Bayesian Dirichlet score for continuous data (Scutari, 2023).

3.8.9 Hill Climb - Bayesian Information Criterion - Correlated Features (HC-BIC-CR)

This CR model utilised only the features that are not highly correlated with the target feature: no3andno2. Feature correlation is an aspect, especially in complex and high-dimensional (large numbers of features) datasets that affect the model's ability to create accurate predictions. Feature selection methods exist to remove redundant, irrelevant or non-informative features to enhance the effectiveness and generalisability of machine learning models (Pudjihartono *et al*., 2022).

The features that remain after the removal process are; TAL, sulphate, sodium, silicon, fluoride and the target feature: nitrate and nitrite.

3.8.10 Random Forest – All Features with 100 trees (RF-AF-100) Model

This model utilises all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

From the literature the Random Forest (RF) algorithm is a widely used ensemble machine learning algorithms that constructs a forest (many) of decision tree

classifiers. Each tree in the forest is trained on a different subset of the data and features with the predicted values being voted the best (Sipper & Moore, 2021).

3.8.11 Random Forest - All Features with 500 trees (RF-AF-500) Model

This model utilises all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

3.8.12 Random Forest - All Features with 1000 trees (RF-AF-1000) Model

This model utilises all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

3.8.13 Random Forest - All Features with 1500 trees (RF-AF-1500) Model

This model utilises all the features resulting from the pre-processing of the data; chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite.

3.8.14 Random Forest - All Features with optimisation (RF-AF-Opt) Model

This model utilises all the features resulting from the pre-processing of the data: chloride, DMS, sodium, TAL, silicon, fluoride, magnesium, conductivity, calcium, potassium, sulphate, phosphate and the target feature: nitrate and nitrite, as well as

the following selection of hyperparameters 1) Number of trees: 500, 1000, 500, 2) Minimum number of observations per node: 5, 10, 15 and 3) Max depth: 10, 20, 30) resulting in 27 combinations for model training to find the values which contribute to a best Random Forest model if possible.

3.9 Evaluation of the different models

Evaluating the performance metrics of the model is essential to understand if the model is working as intended and fit for purpose (Yip, 2020). The target feature nitrate and nitrite are numeric; therefore, the model was a regression, and the metrics that were used for evaluation were Root Mean Square Error (Equation 4), Mean Square Error (Equation 5), Mean Average Precision (Equation 6), Mean Average Error (Equation 7) and R-squared (Equation 8). The error metric values will be evaluated against one another, and the base model will test if the prediction error has improved.

Algorithms which have previously been used in water quality or water forecasting, such as decision tree regression and random forest, while LSTM is used for forecasting, which is beyond the scope of what this study aims to achieve (Kanyama *et al.*, 2020; Dheda & Cheng, 2020).

Root Mean Square Error (RMSE) is a standard metric used to evaluate the performance of models, particularly in fields like meteorology, air quality, and climate research. RMSE is the square root of the mean squared error (MSE), and it yields a metric with the same units as the observed values, representing the typical or standard error for normally distributed errors. RMSE and Mean Absolute Error (MAE) are often used interchangeably, but this needs to be more accurate. RMSE is optimal for Gaussian (normal) error distributions, while MAE is better suited for Laplacian error distributions (Hodson, 2022). R-squared, also known as the coefficient of determination, provides insight into how well a regression model

explains the variance and variability of the predictions, and thus, how well the model fits the data (Chicco *et al*., 2021).

Understanding and utilising this metric is key to ensuring the robustness of the models. evaluation metrics commonly used for the evaluation of regression models are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and can be used to detect outliers further, and Mean Average Precision (MAPE). Plotting the predicted values versus the known actual values will give a reasonable indication of how well the models are performing.

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2}$$

Equation 4: The equation for RMSE. The best achievable value = 0; the worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data (Chicco *et al*., 2021).

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2$$

Equation 5: The equation for MSE. The best achievable value = 0; the worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data (Chicco *et al*., 2021).

$$\mathrm{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{Y_i - X_i}{Y_i}\right|$$

Equation 6: The equation for MAPE. The best achievable value = 0; the worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data (Chicco *et al*., 2021).

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i|$$

Equation 7: The equation for MAE. The best achievable value = 0; the worst = +inf, with Xi is the predicted ith value, and the Yi element is the actual ith value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth data (Chicco *et al*., 2021).

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2}$$

Equation 8: The equation for R-squared. The worst value = -inf; the best achievable value = +1. R-squared indicates the percentage of variability in the dependent variable that can be predicted by the independent variables (Chicco *et al*., 2021)

3.9.1 Actual vs Predicted Plots

The actual vs predicted plot (Figure 7) compared each record pair of actual/predicted at an index in the test set; the smaller the distance between the prediction and actual value, the better the model was at understanding and generalising patterns from the train set to apply to the test set for evaluation. The closer the points are for all samples, the more accurate the model is (The Mathworks, Inc., 2024). These plots compare the absolute difference between predicted and actual values, allowing for direct comparison across different models and metrics.

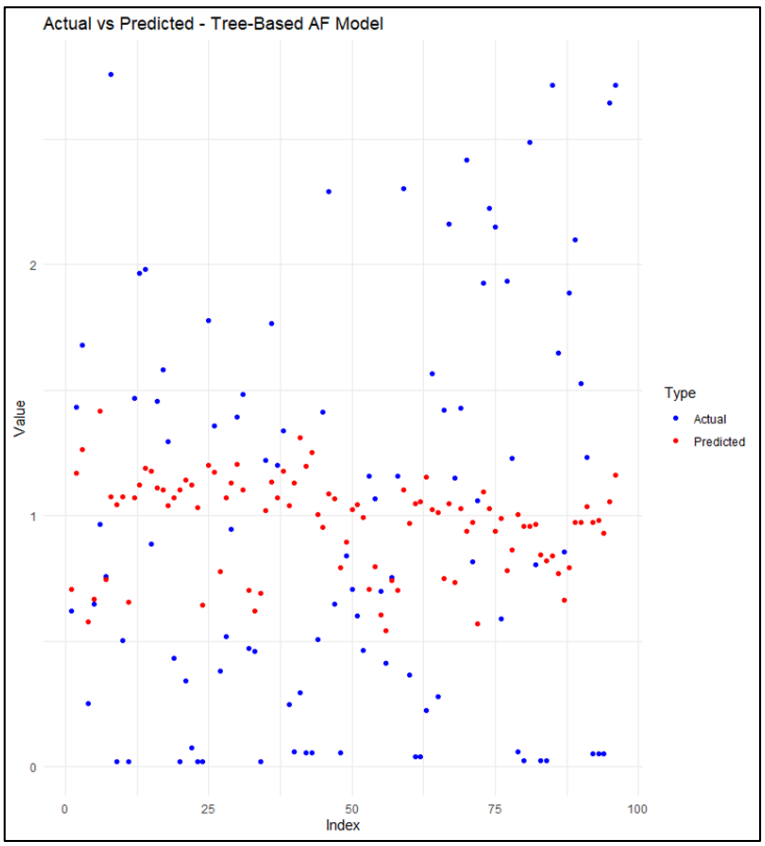


Figure 7: The actual vs predicted plot of the Tree-Based All-Features Model. The true values are represented by the blue points while the predicted are represented by the red points. For each record pair of actual/predicted at an index in the test set, the smaller the distance between one another, the closer the predictions are to the true values.

3.9.2 Prediction Error per Sample Size

All the models are trained on all the records within the train set, but does training the models on different random sample sizes affect the accuracy of the predictions? If so, is there a point where the evaluation metrics stop improving, or is there a need for more data or different features to be tested?

Chapter 4

RESULTS

4.1 Introduction

This chapter outlines the research results, beginning with data preprocessing, model structure, cross-validation evaluation and model performance. Processing of the data to remove outliers, splitting the data into train, validation and test sets, imputation of missing data, data set distribution and further outline detection are further expanded.

4.2 Kernel Density Estimation

The shapes of the distributions for the train, validation, and test sets per plotted feature should ideally follow a similar shape without the peaks or the size of the x-axis being too dissimilar.

The density plots for calcium (Figure 8A), chloride (Figure 8B), conductivity (Figure 8C) and DMS (Figure 8D) show similar distributions and these features can be used in the modelling process.

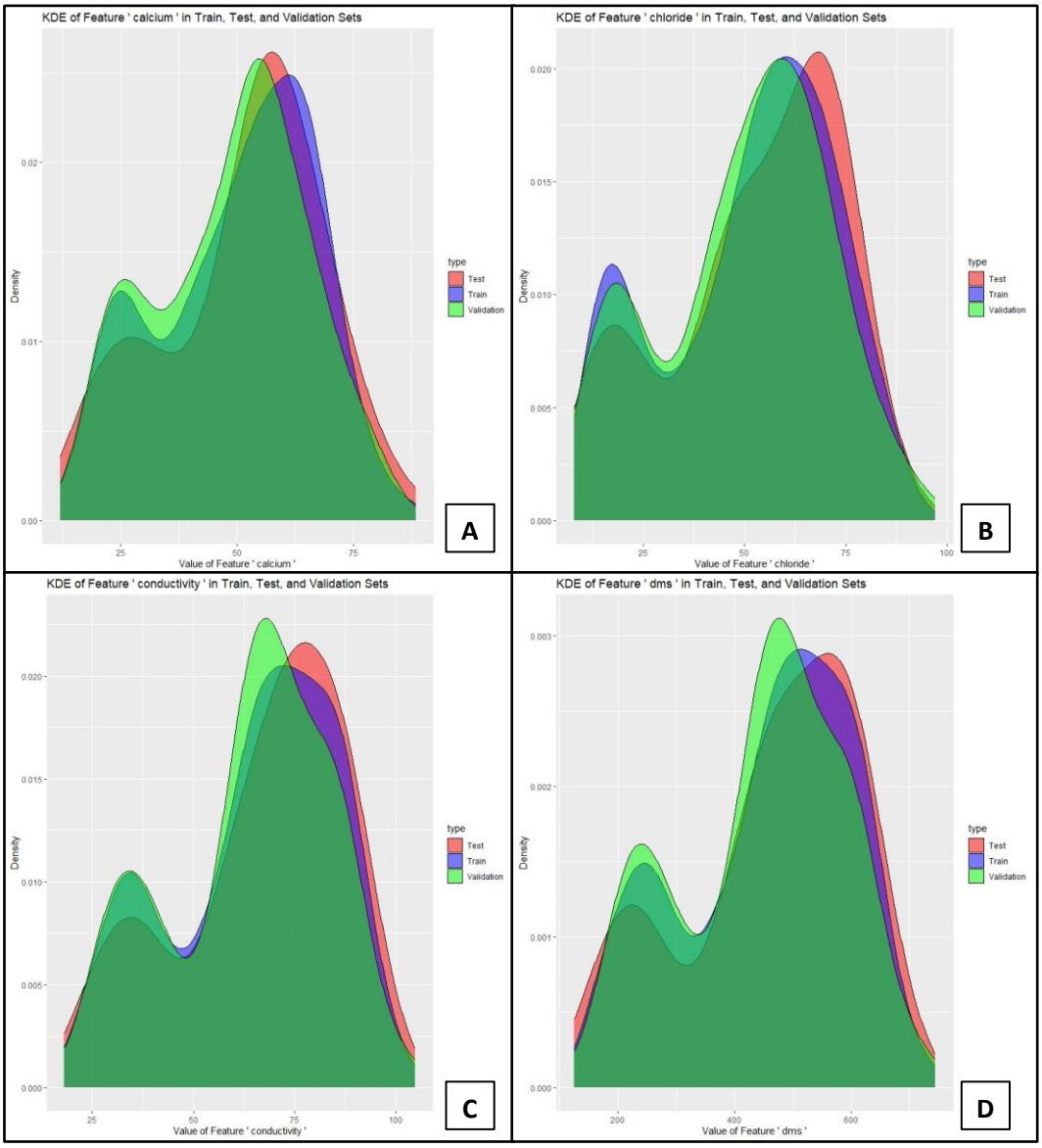


Figure 8: The Kernel Density Estimation plots for A) calcium, B) chloride, C) conductivity and D) DMS.

The density plots for fluoride (Figure 9E), magnesium (Figure 9F), and phosphate (Figure 9G) show similar distributions and these features can be used in the modelling process. The nitrate and nitrite plot (Figure 9H) show a similar range and size for the target variable, this is an important check to identify possible outliers or issues when used training the models.

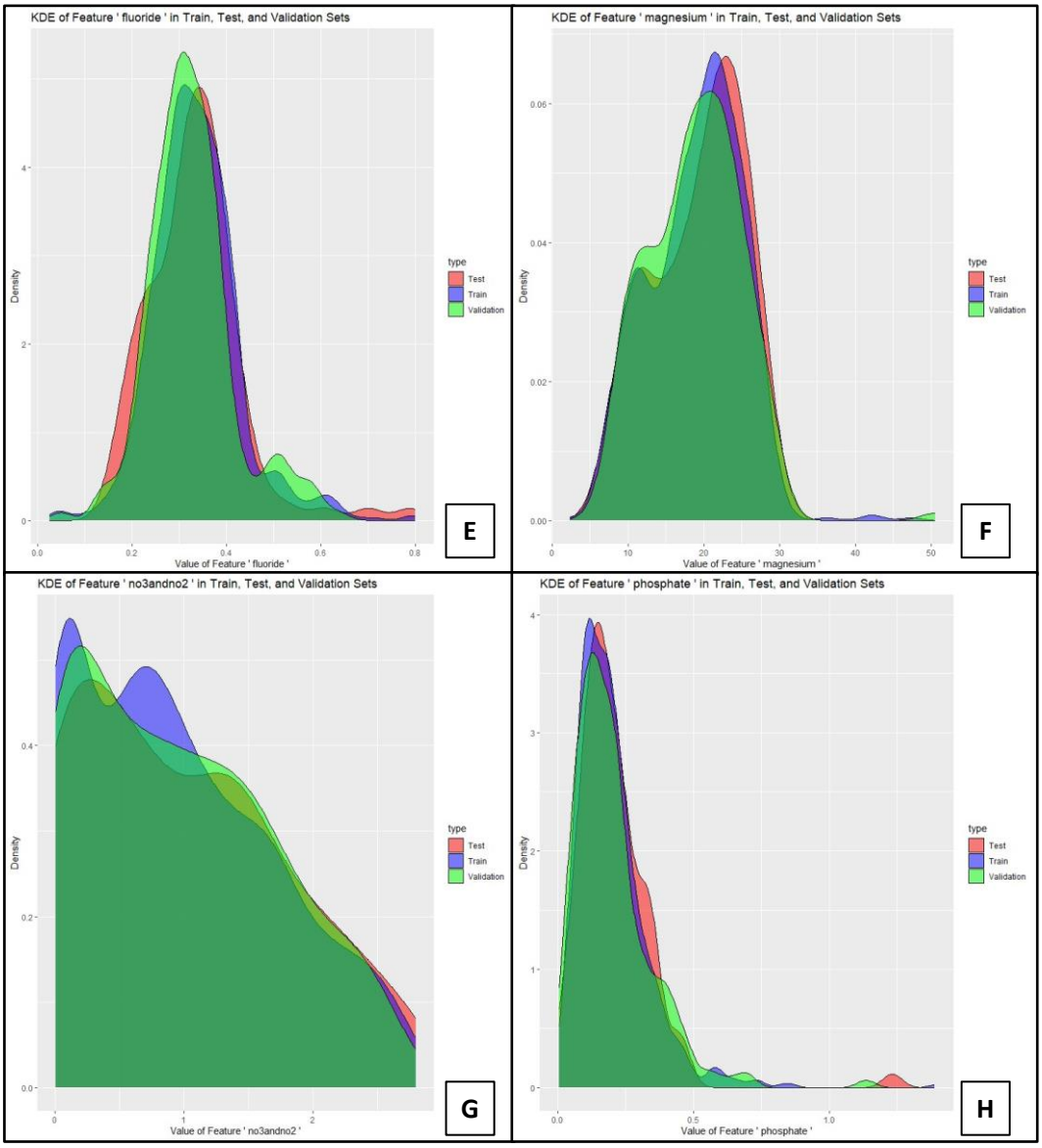


Figure 9: The Kernel Density Estimation plots for E) fluoride, F) magnesium, G) nitrate and nitrite and H) phosphate.

The density plots for potassium (Figure 10I), silicon (Figure 10J), and TAL (Figure 10K) show similar distributions, and these features can be used in the modelling process. The density plots for temperature (Figure 10L) are dissimilar from the test set, showing the most significant difference; this will affect the training and validation of the models and can be excluded. The density plots for ammonium (Figure 10M) are dissimilar with a very narrow range across the train, validation and test sets; this will affect the models' training and validation and can be excluded.

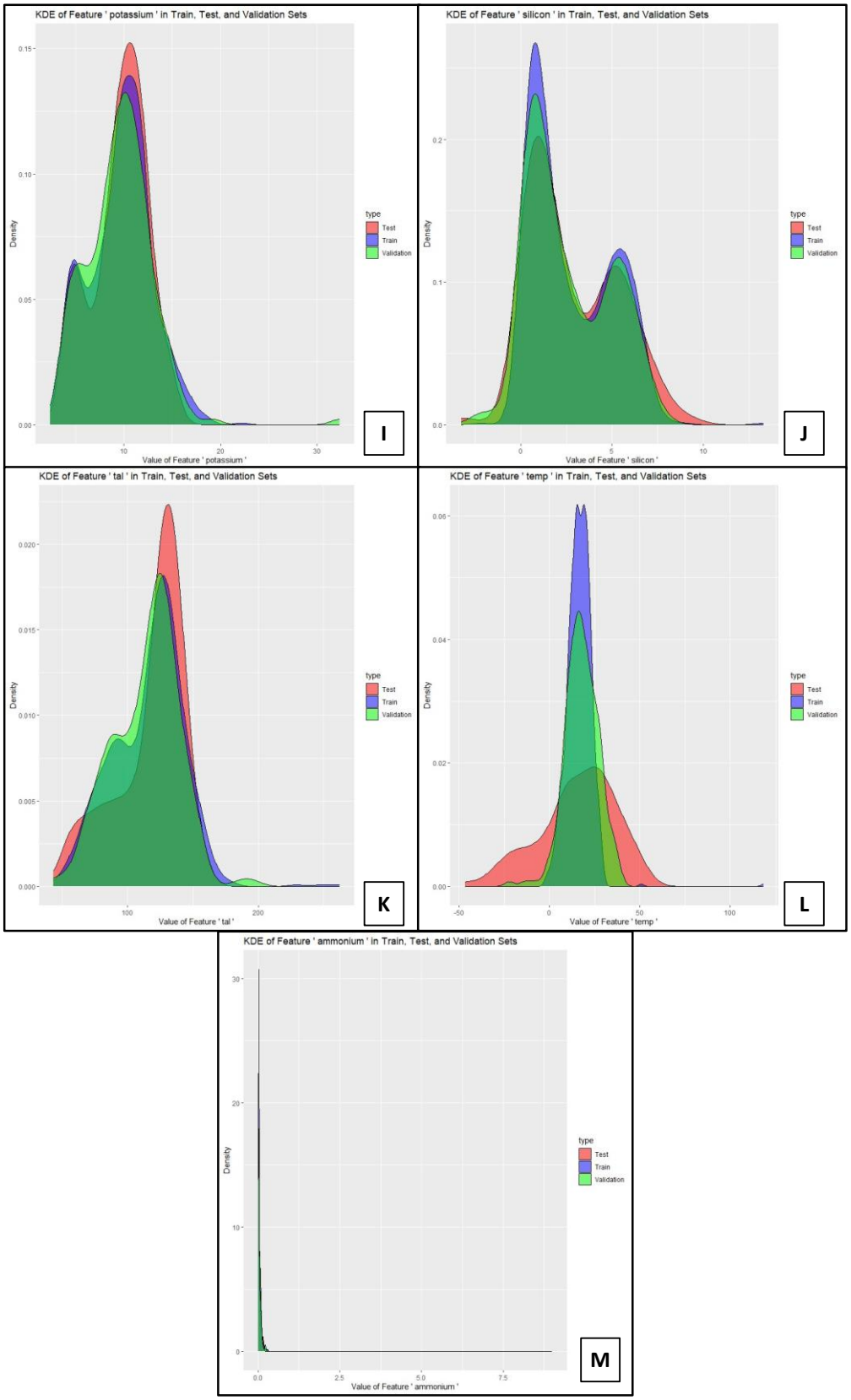


Figure 10: The Kernel Density Estimation plots for I) potassium, J) silicon, K) TAL, L) temperature and M) ammonium.

4.3 Boxplots

The boxplot for ammonium (Figure 11A) has small overall value ranges with large value outliers in the train set, which could represent no differentiation in the modelling process, and this feature can be excluded. The boxplots for calcium (Figure 11B), chloride (Figure 11C), and conductivity (Figure 11D) have no apparent outliers or values that differ too much from the mean, and these features can be used in the modelling process.

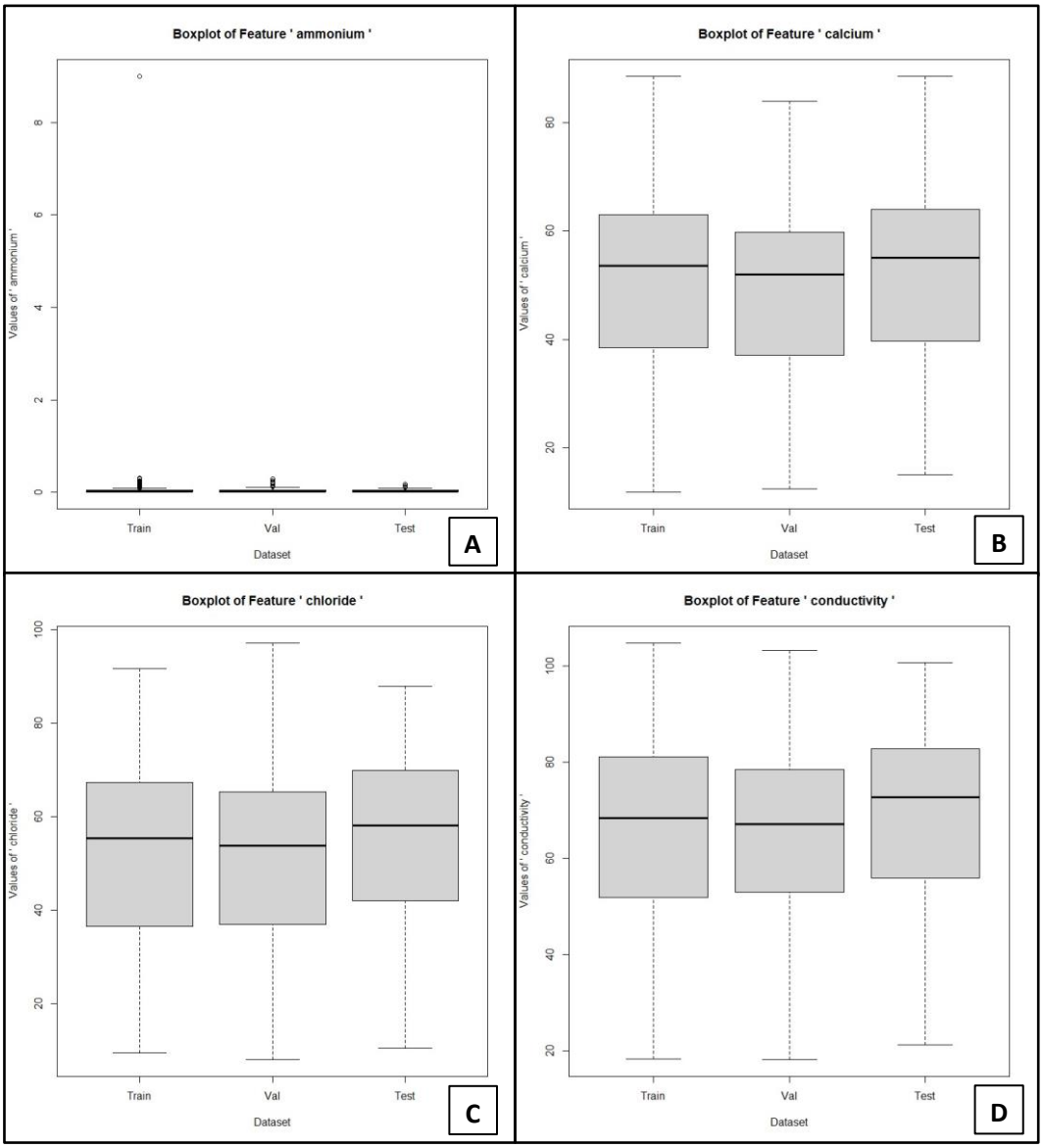


Figure 11: The boxplots for A) ammonium, B) calcium, C) chloride and D) conductivity.

The boxplots for DMS (Figure 12E) and nitrate and nitrite (Figure 12H) have no apparent outliers or values that differ too much from the mean, and these features can be used in the modelling process. The box plot for fluoride (Figure 12F) has a similar range of values, but there are many outliers present in all data sets, particularly the train set; this could pose an issue and can be excluded from the modelling process. Magnesium (Figure 12G) has a similar range of values, with some outliers present in the train set and can be used in the modelling process.

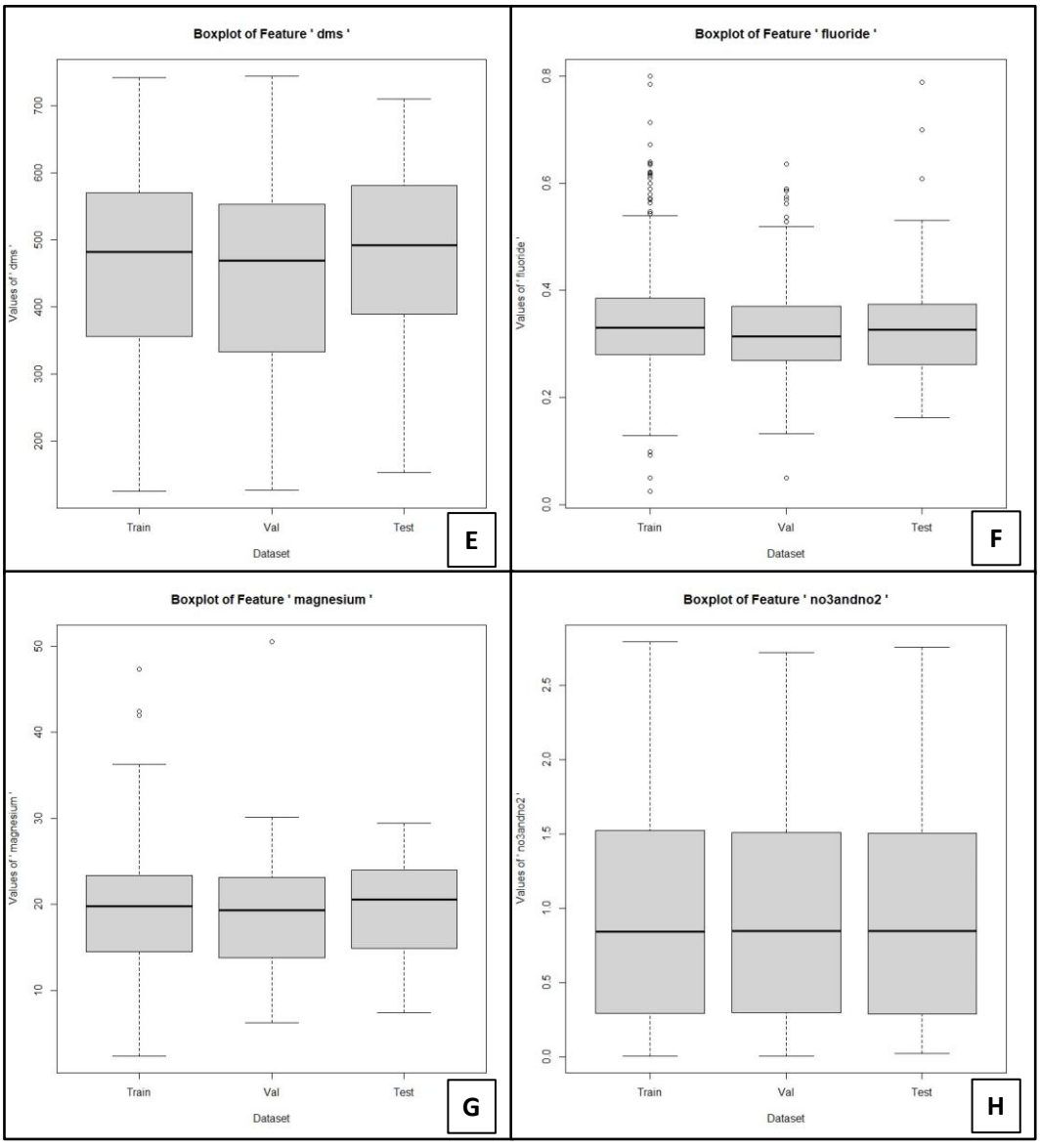


Figure 12: The boxplots for E) DMS, F) fluoride, G) magnesium and H) nitrate and nitrite.

The boxplots for phosphate (Figure 13I) show that the data sets have a similar range of values. However, many outliers are present in all data sets, but in particular, the train set; this could pose an issue and be excluded from the modelling process. The box plots for sodium (Figure 13J) have no outliers and similar value ranges. In contrast, silicon (Figure 13L) and potassium (Figure 13K) have some outliers present but are already close to the max range in the data sets and can be included in the modelling process.

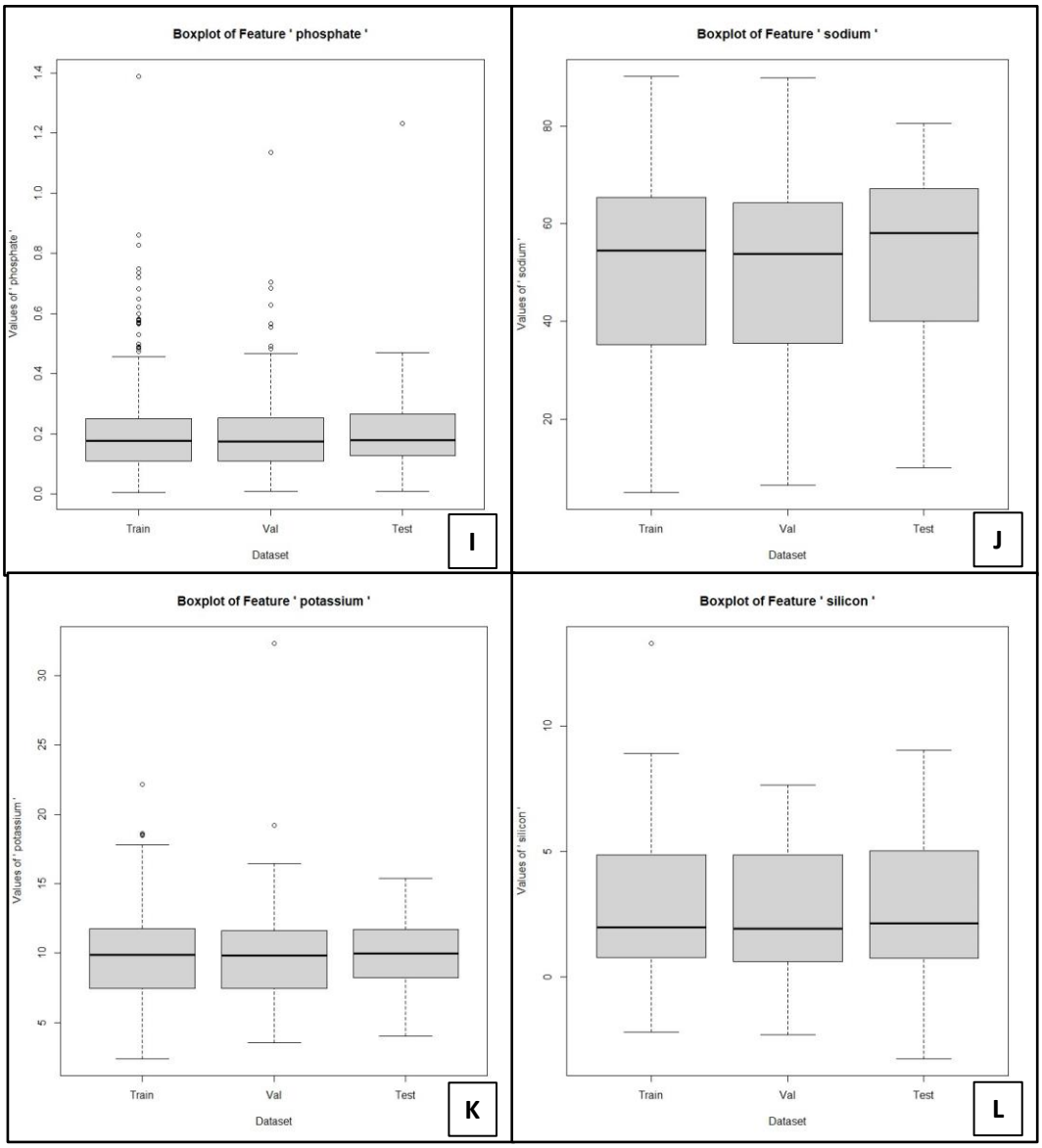


Figure 13: The boxplots for I) phosphate, J) sodium, K) potassium and L) silicon.

The box plots for sulphate (Figure 14M) have no outliers and similar value ranges and can be used in modelling. Temperature (Figure 14N) has different value ranges across the data sets, especially the test set, and this feature can be excluded from the modelling process. The box plots for TAL (Figure 14O) have some outliers present but can be included in the modelling process.

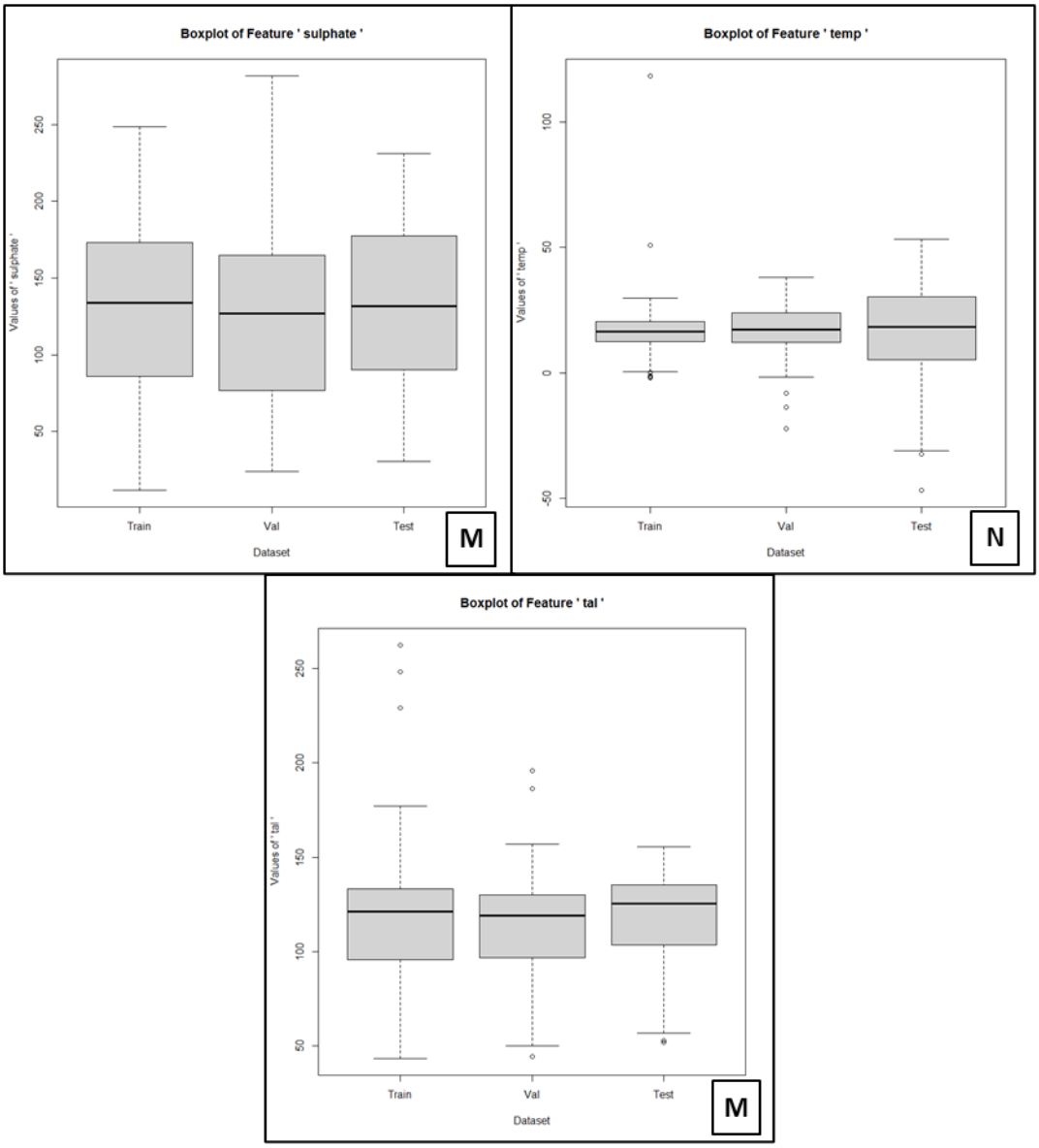


Figure 14: The boxplots for M) sulphate, N) temperature and O) TAL.

4.4 Model Structures (Directed Acyclic Graph)

A Directed Acyclic Graph (DAG) is a graph used to represent causal pathways in various research and analytical contexts. Its directed edges characterise it, and the absence of cycles means it is impossible to start at one node and return to it by following the directed edges (Watson *et al*., 2023). These nodes and edges model the overall dependence structure of multiple variables with the lines identifying the direction of effect (Moffa *et al*., 2017).

In Figure 15A, the structure is hierarchical, starting with the nodes at the top being the most independent features, with the nodes becoming more dependent the further down the tree they go. The directions of the arrows indicate dependencies, with the arrow pointing from the independent variable to the dependent variable, the direction of dependency. When focusing on the target variable, chloride directly influences DMS, DMS on calcium, and calcium influences nitrate and nitrite.

In Figure 15B, chloride affects calcium, and calcium directly influences nitrate and nitrite, which in turn influences potassium, sulphate and conductivity. In contrast, in Figure 15C, calcium and TAL directly influence nitrate and nitrite, which also influence potassium, conductivity, and sulphate. In Figure 15D, this model produces a more complex relationship between the features: magnesium, silicon, chloride, DMS, sodium, calcium, conductivity, phosphate, potassium, TAL, and fluoride have a direct influence on nitrate and nitrite.

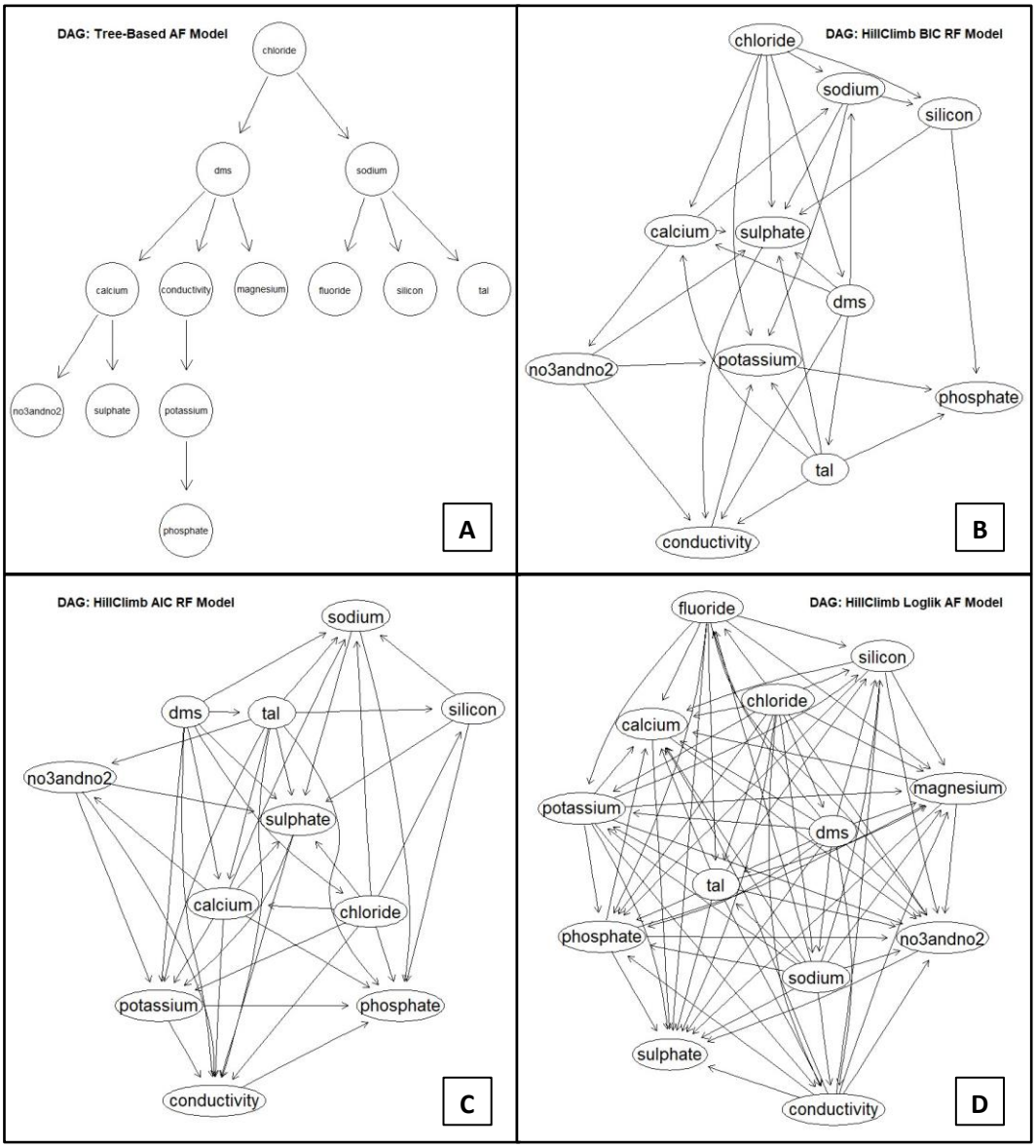


Figure 15: Directed Acyclic Graph (DAG) for A) Tree-Based-AF, B) HC-BIC-RF, C) HC-AIC-RF and D) HC-Loglik-AF.

In Figure 16E, calcium, magnesium, potassium, TAL, and conductivity directly influence nitrate and nitrite. In contrast, in Figure 16F and Figure 16G, only calcium directly influences nitrate and nitrite. In Figure 16H, phosphate and calcium directly influence nitrate and nitrite. In Figure 16I, sodium has a direct influence on nitrate and nitrite.

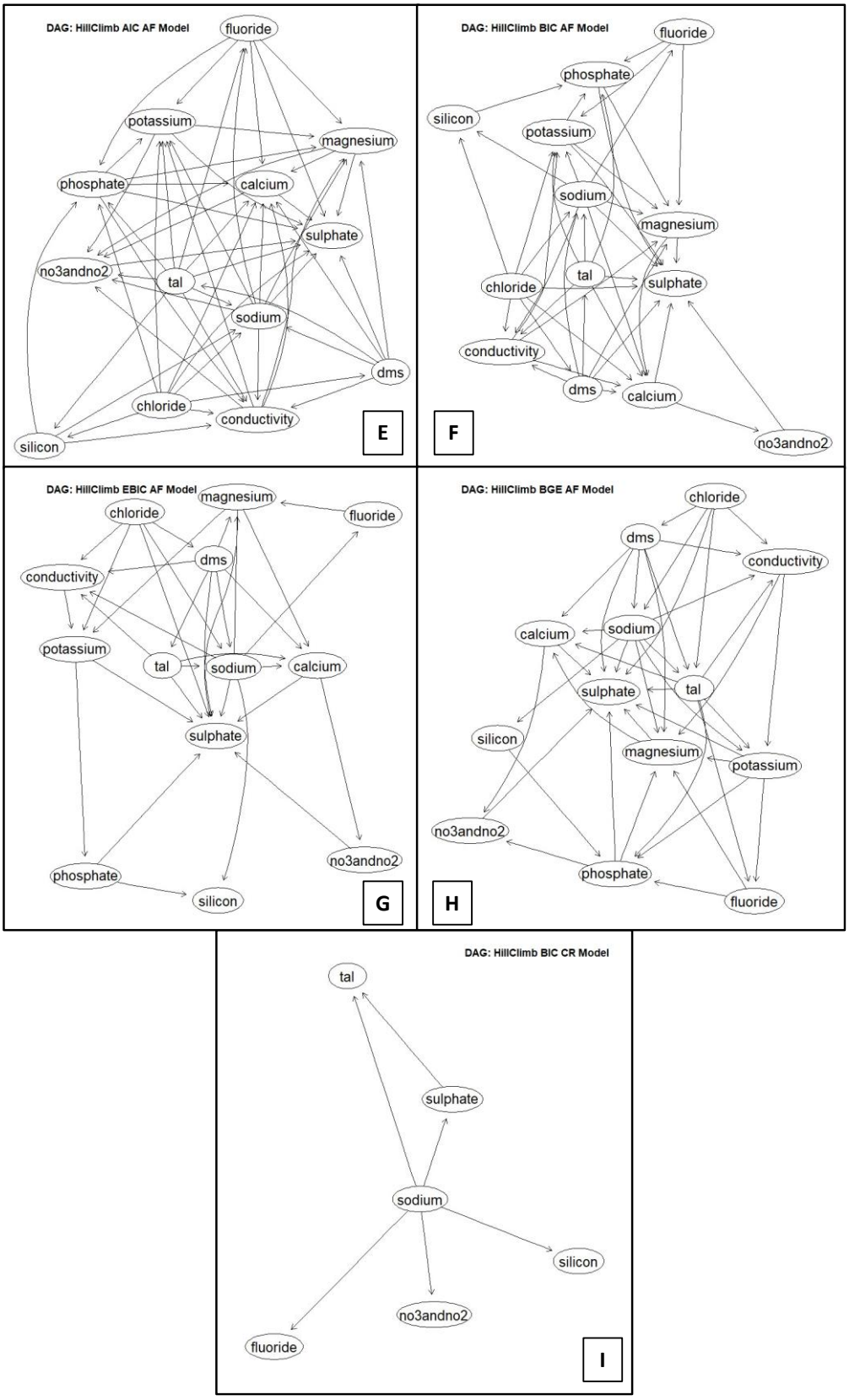


Figure 16: Directed Acyclic Graph (DAG) for E) HC-AIC-AF, F) HC-BIC-AF, G) HC_EBIC-AF, H) HC-BGE-AF and I) HC-BIC-CR.

Collating a view of the parameters (Table 12) that have an effect on nitrate and nitrite regardless of the model provides an overall understanding of how different scoring functions within the models learn the structures in the data to predict the target variable.

Table 12: A summary of the features which have a direct effect on the target variable with calcium in all but nine models (calcium excluded from feature list) being the most common feature present.

Model	Calcium	Chloride	DMS	Conductivity	Fluoride	Potassium	Magnessium	Sodium	Phosphate	Silicon	TAL	Total
Tree-Based AF	X											1
HC-BIC-RF	X											1
HC-AIC-RF	X										X	2
HC-Loglik-AF	X	X	X		X	X	X	X	X	X	X	10
HC-AIC-AF	X			X		X	X				X	5
HC-BIC-AF	X											1
HC-EBIC-AF	X											1
HC-BGE-AF	X								X			2
HC-BIC-CR								X				1

By collating the correlated features from each of the models, we have the following unique pairings of features:

1. Calcium, nitrate and nitrite, conductivity and phosphate.
2. Chloride, potassium, sulphate, TAL, sodium, silicon, phosphate, fluoride, calcium, conductivity and nitrate and nitrite.
3. Conductivity, calcium, phosphate, silicon, sulphate, potassium, nitrate and nitrite, magnesium.
4. DMS, sodium, silicon, calcium, sulphate, potassium, phosphate, conductivity, nitrate and nitrite, TAL and fluoride.
5. Fluoride, sulphate, silicon, TAL, phosphate, calcium, potassium, nitrate and nitrite, and magnesium.
6. Magnesium, nitrate and nitrite, sulphate, calcium and potassium.
7. Nitrate and nitrite, sulphate, potassium and conductivity.
8. Phosphate, potassium, sulphate, silicon, calcium, nitrate and nitrite and magnesium.

9. Silicon, sodium, sulphate, phosphate, conductivity, calcium, nitrate and nitrite and magnesium.
10. Sodium, silicon, sulphate, TAL, potassium, phosphate, fluoride, calcium, conductivity, nitrate and nitrite and magnesium.
11. Sulphate, TAL, potassium and conductivity.
12. TAL, silicon, sodium, sulphate, potassium, phosphate, fluoride, calcium, conductivity, nitrate and nitrite and magnesium.

4.5 Model Cross-Validation Fold Performance

Cross-validation assesses a model's performance during the training phase by partitioning the dataset and running multiple training runs (Figure 17) with validation to minimise the likelihood of overfitting and provide a more accurate estimate of the model's performance. The model's and data's stability can also be examined; if the error value is inconsistent over the fold, it may indicate an issue with the data or the model itself (scikit-learn developers, 2023).

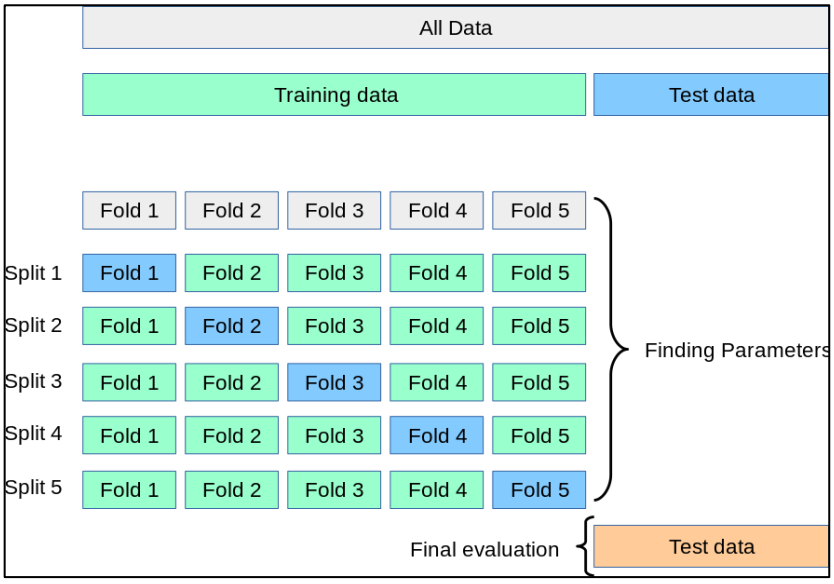


Figure 17: Cross validation estimation; five models are built with a different subset of the data and validated on a percentage not included in the training process. The error estimation is averaged (Source: *https://scikit-learn.org*).

Cross-validation was assessed on the validation set for each of the nine models to assess the stability of the prediction errors and evaluation using MSE, which penalises outliers and will magnify the size of the predictions.

The CV performance and stability of the models were evaluated using each model's mean MSE error across all fold's relative to each fold. The CV performance of the Tree-Based-AF model (Figure 18A) model, the HC-BIC-RF model (Figure 18B), the HC-AIC-RF model (Figure 18C) and the HC-Loglik-AF model (Figure 18D) have large variations in the MSE scores per fold when compared to the mean MSE respectively, displaying possible instability due to the lack of sufficient data or further data points, which are outliers.

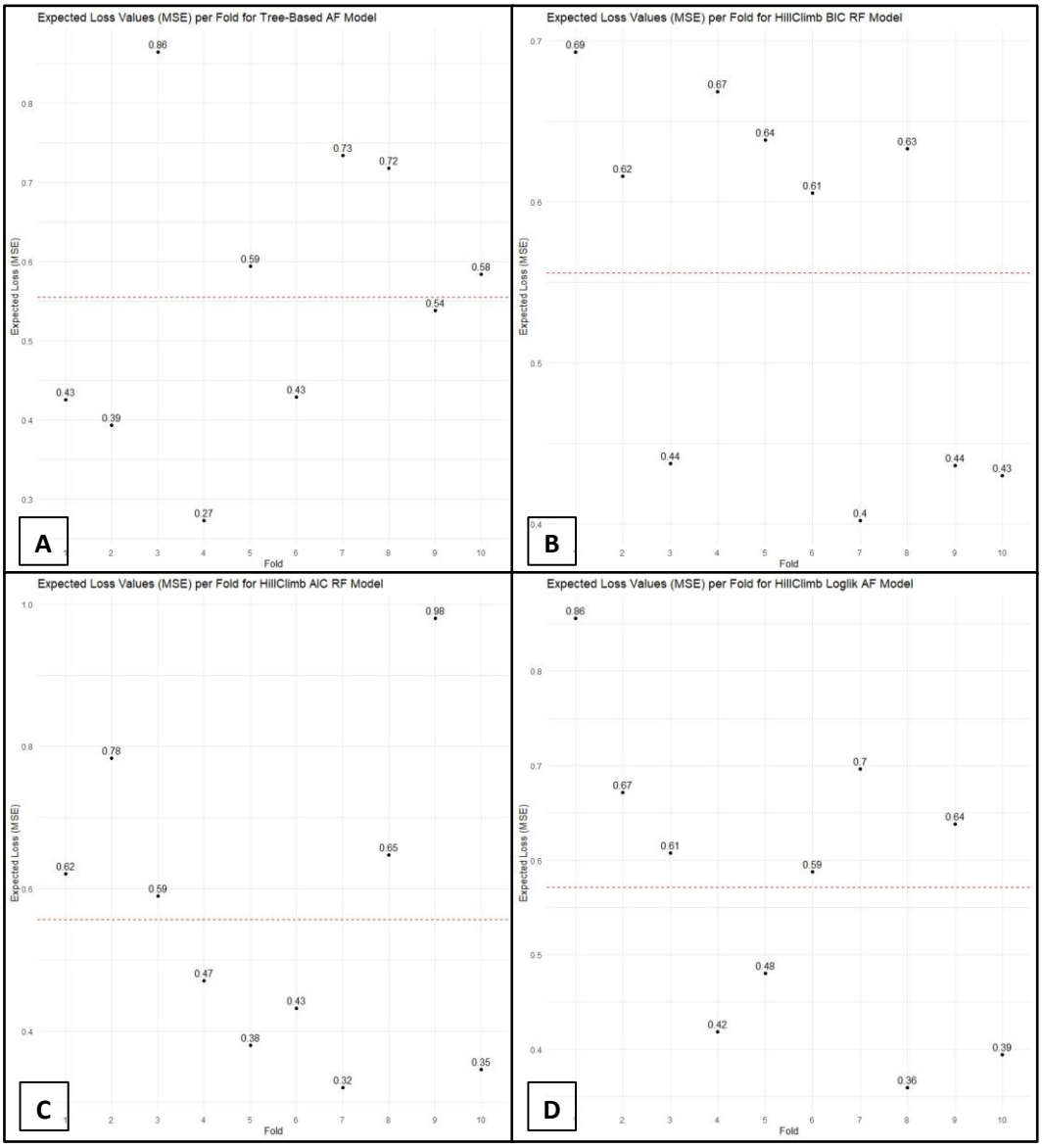


Figure 18: The cross-validation performance of the per fold for the A) Tree-Based-AF, B) HC-BIC-RF, C) HC-AIC-RF and D) HC-Loglik-AF models.

The CV performance of the HC-BIC-AF (Figure 19F) model and the HC-BGE-AF model (Figure 19H) have significant variations in the MSE scores per fold when compared to the mean MSE, respectively, displaying possible instability due to the lack of sufficient data or further data points, which are outliers. The HC-EBIC-AF model (Figure 19G) is relatively stable since most of the losses per fold are lower and closer to the mean across all folds. The HC-AIC-AF model (Figure 19E) is the most stable, with all but two folds having low losses.

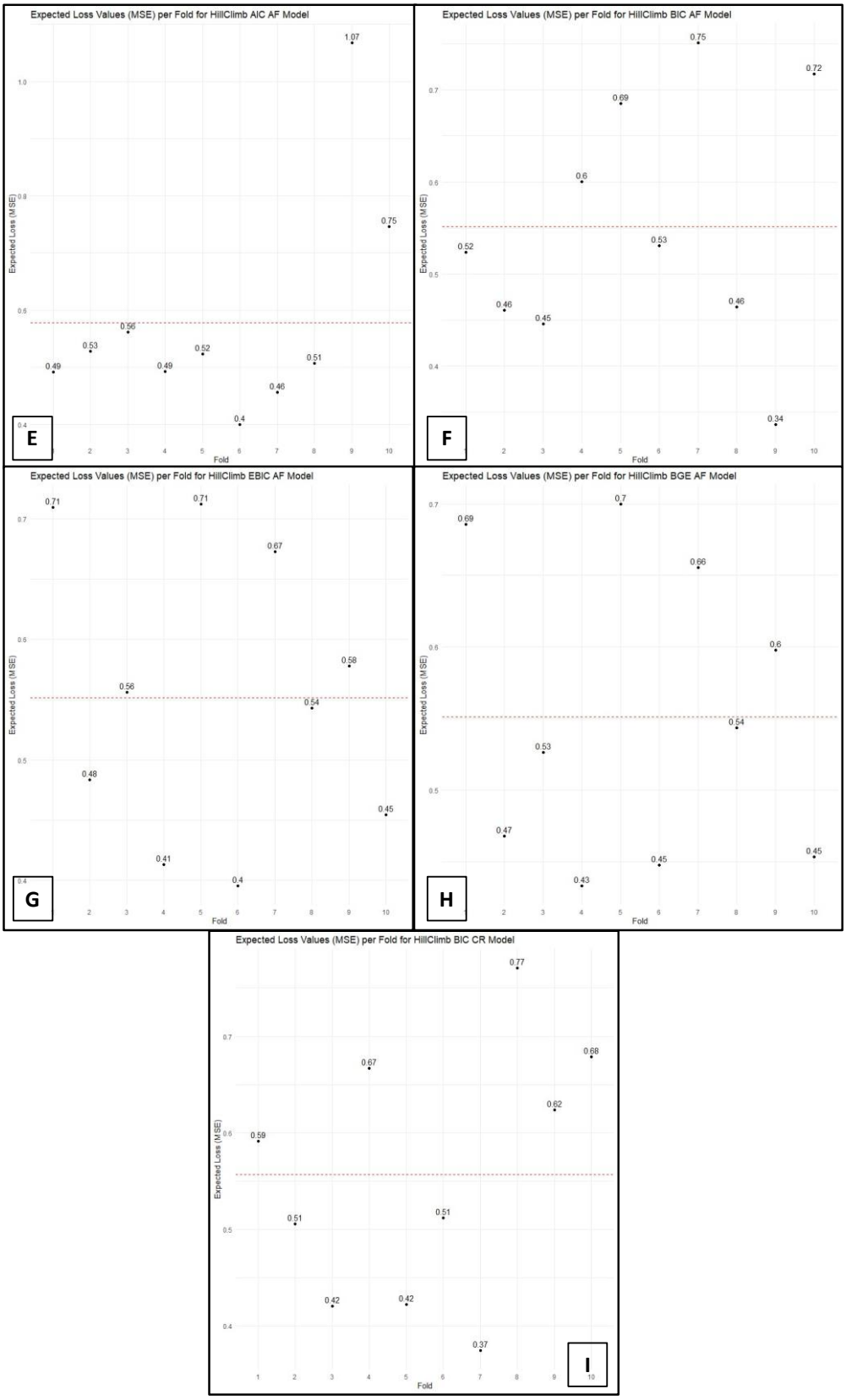


Figure 19: The CV performance of the per fold for the E) HC-AIC-AF, F) HC-BIC-AF, G) HC-EBIC-AF, H) HC-BGE-AF and I) HC-BIC-CR models.

4.6 Model Actual vs Predicted Performance

To better understand model performance, comparing the actual data to the machine learning model's predicted values can help assess how well the model can predict a real-world outcome. This information can later be used to fine-tune the model or select a different one that may improve its predictive ability. We can see how well the models fit the data by examining the absolute differences between these values. The larger the differences and the number of points further away from the zero difference could indicate poor fit, while conversely, the smaller the differences, the better the fit (Grant, 2022).

The predictions for the Tree-based-AF model (Figure 20A) are clustered together, with significant differences between the predicted points (red) and the actual points (blue) when comparing the points vertically; this indicates poor predictive performance. The predictions for the HC-BIC-RF model (Figure 20B) are closer to the actual points. However, the model cannot reduce the difference for the lower actual values by overestimating, resulting in poor performance. The predictions for the HC-AIC-RF model (Figure 20C), and HC-Loglik-AF model (Figure 20D) are closer to the actual values, indicating reasonable performance.

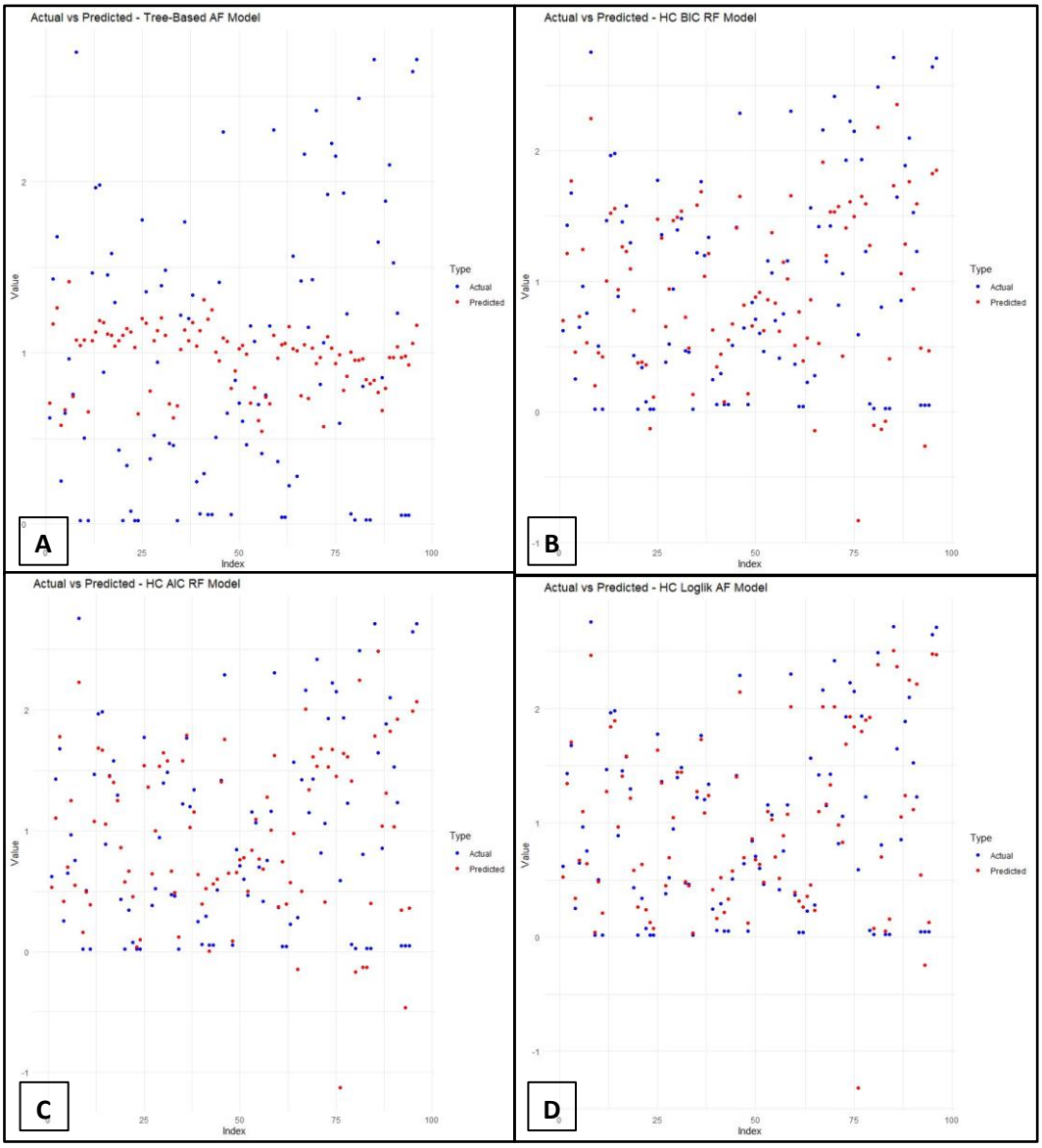


Figure 20: The actual versus predicted plots per sample for A) Tree-Based-AF, B) HC-BIC-RF, C) HC-AIC-RF, and D) HC-Loglik-AF models.

The predictions for the HC-AIC-AF (Figure 21E), HC-BIC-AF (Figure 21F), HC-EBIC-AF (Figure 21G), and HC-BGE-AF (Figure 21H) have the slightest difference between the predicted points and actual points, indicating good performance when compared the rest of the models. The predictions for the HC-BIC-CR model (Figure 21I) are clustered similarly to the Tree-Based-AF model,

with significant differences between the actual and predicted points; this model performs the worst overall.

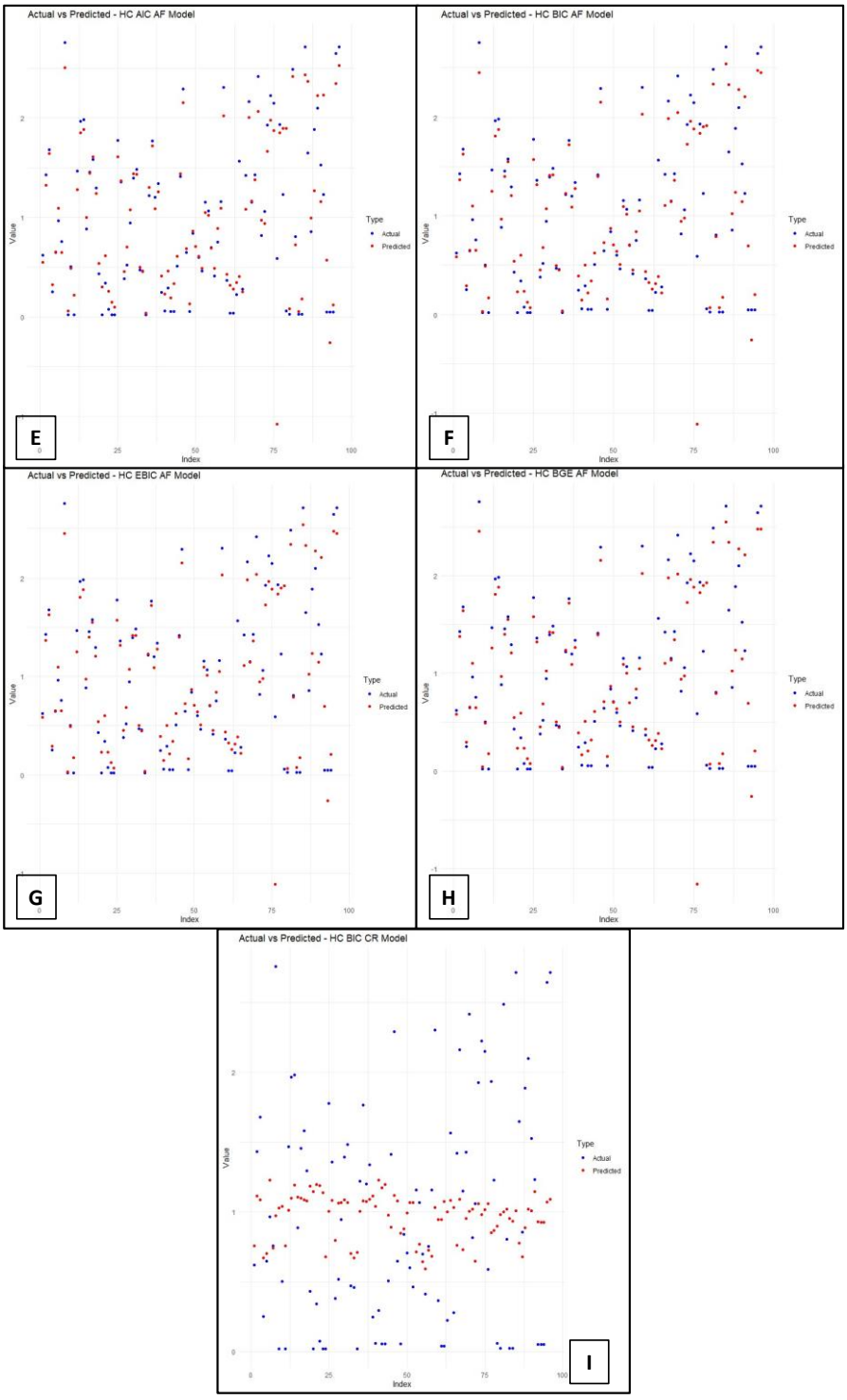


Figure 21: The actual versus predicted plots per sample for E) HC-AIC-AF, F) HC-BIC-AF, G) HC-EBIC-AF, H) HC-BGE-AF and I) HC-BIC-CR models.

4.7 Error vs Sample Size Performance

At what point, when building models, is the number of samples sufficient to produce a model with suitable predictive power and error rate to satisfy the problem statement and can building multiple models with the same scoring criteria assist with this?

Five Bayesian models were built using all features with the AIC, BIC, EBIC, BGE and Loglikelihood scoring functions, with ten runs per sample size to reduce sample bias with the average error per sample size plotted for the selected error metrics: RMSE, MSE, MAPE and MAE.

To better understand the performance of the models, each is plotted on a single graph per error metric starting at sample 30 to 230; by doing this, we can directly see how each model performs.

The overall model performance for RMSE (Figure 22A), the AIC, BIC, and Loglikelihood scoring functions converge the fastest to a lower error value until 100 samples; BGE and EBIC then continue to improve on lowering the error value at about the same rate per increase in sample size. The EBIC and AIC have the lowest error value at 230 samples.

The overall model performance for MSE (Figure 22B), the AIC, BIC and Loglikelihood scoring functions converge the fastest to a lower error value until 100 samples; BGE and EBIC then continue to improve on lowering the error value at about the same rate per increase in sample size. The EBIC and AIC have the lowest error value at 200 samples.

The overall model performance for MAPE (Figure 22C) shows that the AIC, BIC, and Loglikelihood scoring functions converge the fastest to a lower error value until 100 samples; BGE and EBIC then continue to improve on lowering the error value at about the same rate per increase in sample size. The EBIC and AIC have the lowest error value at 200 samples.

The overall model performance for MAE (Figure 22D), the AIC, BIC, and Loglikelihood scoring functions converge the fastest to a lower error value until

100 samples; BGE and EBIC then continue to improve on lowering the error value at about the same rate per increase in sample size. The EBIC and AIC have the lowest error value at 190 samples.

The overall model performance shows the same trend, regardless of the error metric used; AIC, BIC, and Loglikelihood learn some structure early on in the modelling process with fewer samples, but as more data are introduced to the models, the error values get lower. This is true for all models, but EBIC and AIC metrics go lower.

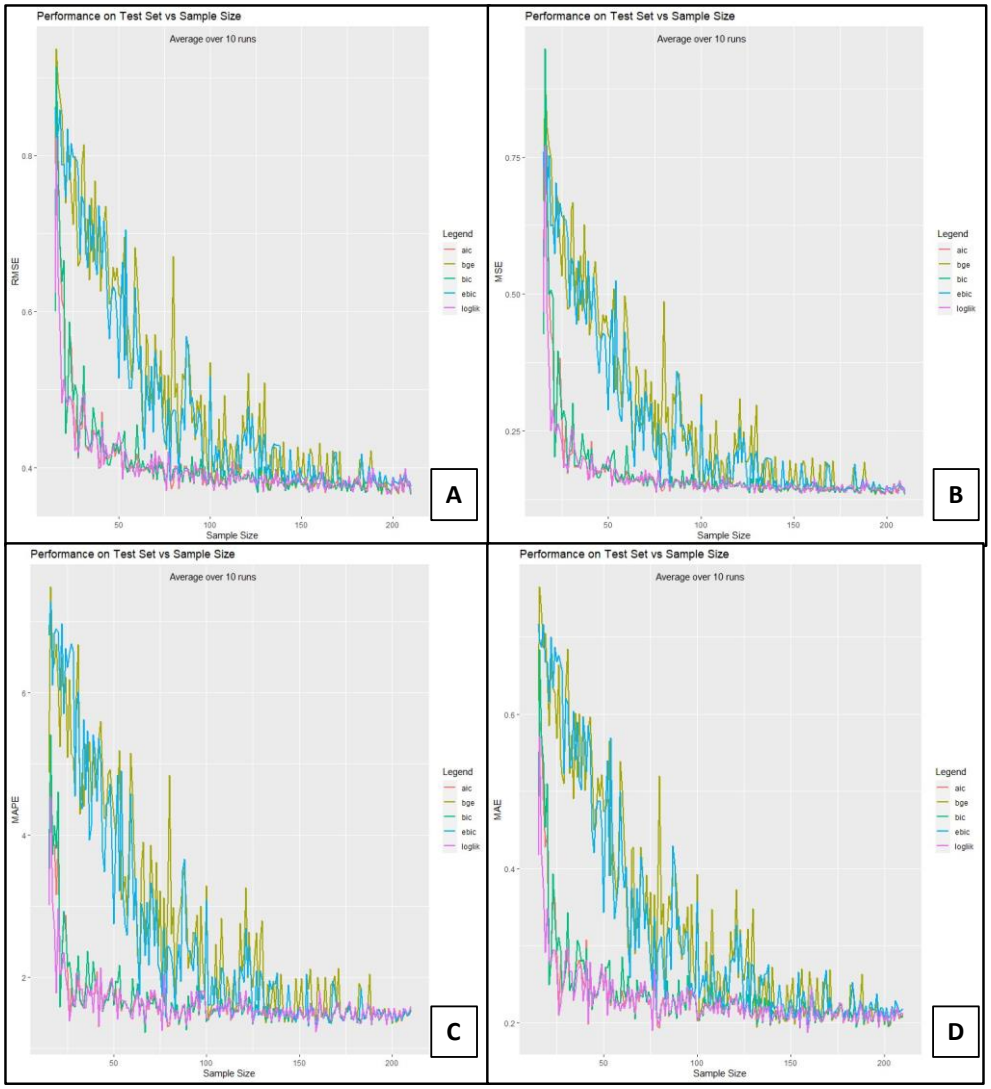


Figure 22: The prediction error for the models for A) RMSE, B) MSE, C) MAPE and D) MAE evaluation metrics.

4.8 Model Execution Time per Sample Size

As the sample size increases, the error value decreases, which is expected, but there is a time cost concerning the training of the models; as the size of the samples increases, so does the time it takes to train a model (Figure 23). The HC-EBIC-AF model completes the fastest initially but requires more data and time to learn the data structure. The HC-BGE-AF model requires the least training time using all the training data but does not have the lowest or best predictive performance overall.

The results for model execution time are based on models run on the following hardware; Intel i5-9500T with 6 cores, 8GB DDR4 memory and 120GB Solid State Drive (SSD) (Intel Corporation, 2024), with the required R and Python libraries installed, Table 13.

Table 13: A summary of the libraries and associated version used in R and Python.

The R Language		
Software/Library	**Version**	**Description**
R	3.4.0	The R language for statistical analysis and modelling.
BiocManager	1.30.21.1	Manages package installation for genomic packages.
graph	1.78.0	Handling of graph data structures often used in bioinformatics.
Rgraphviz	2.44.0	Plotting of graph objects.
dplyr	1.1.2	Data analysis and manipulation tool.
bnlearn	4.8.3	Used for learning Bayesian networks.
ggplot2	3.4.2	A data visualization package.
cvms	1.6.0	Cross-validation for model selection in machine learning workflows.
rsample	1.1.1	Allows for resampling of data.
readxl	1.4.3	Simplifies importing data from Excel files.
mice	3.16.0	Handling of missing data.
caret	6.0.94	Used for building, training, and optimsing ML models.
forecast	8.21	Handling of time series-based data.
Metrics	0.1.4	Offers functions and metrics to evaluate ML models.
skimr	2.1.5	Enhanced summary function in R.
randomForest	4.7.1.1	Implement Random Forest algorithm.
ranger	0.16.1	Optimise Random Forest algorithm.
Python		
Software/Library	**Version**	**Description**
Anaconda Individual Edition	2020.07	Includes the installation of Jupyter Notebook and Python 3.7
Pandas	1.4.1	Data analysis and manipulation tool

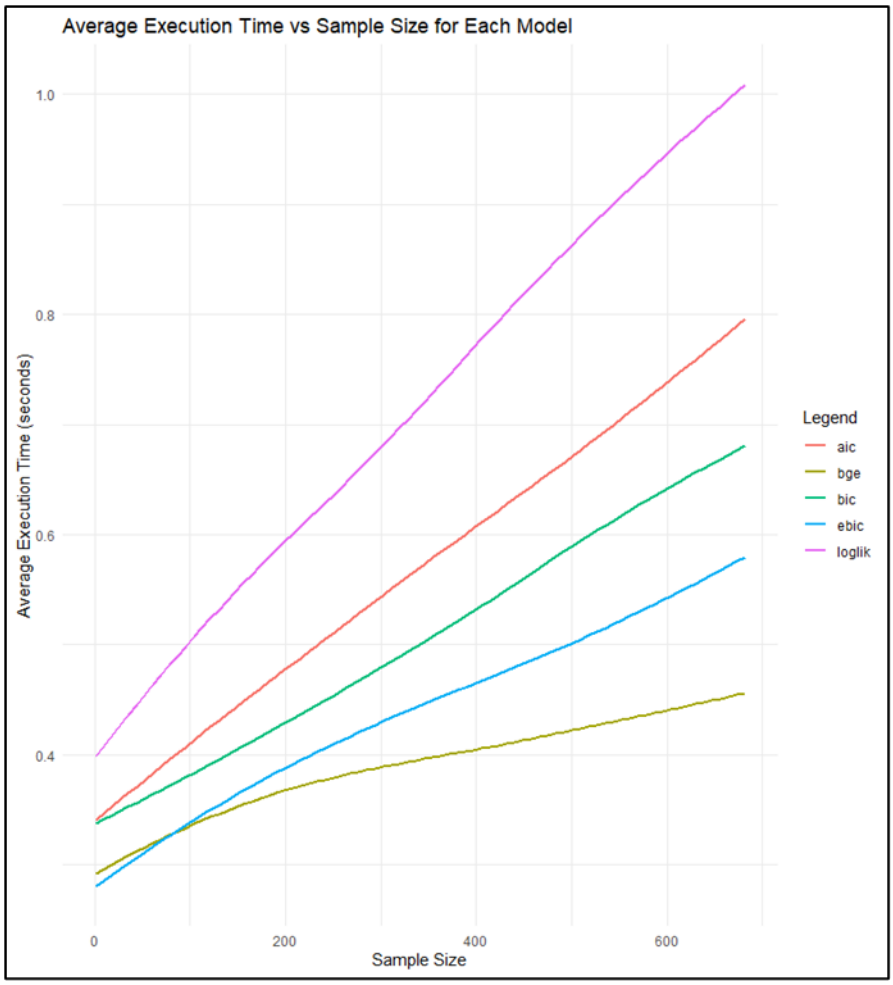


Figure 23: Execution Time per Sample Size for each model. As the sample size increases so does the time for the models to complete a training run with selected sample size.

4.9 Evaluation of Model Performance

Evaluating the predictive power of the models is an important step in determining whether the models can predict the nitrate and nitrite values with the lowest error and which model is better suited to it.

When the RMSE (Equation 4), values for the validation (Figure 24A) and test set (Figure 24B) are compared, the values for each model are similar, indicating stability and generalisation of the data during the training process with the selected training data. The HC-AIC-AF and HC-BGE-AF models show the best predictive performance when examining the unseen test set data.

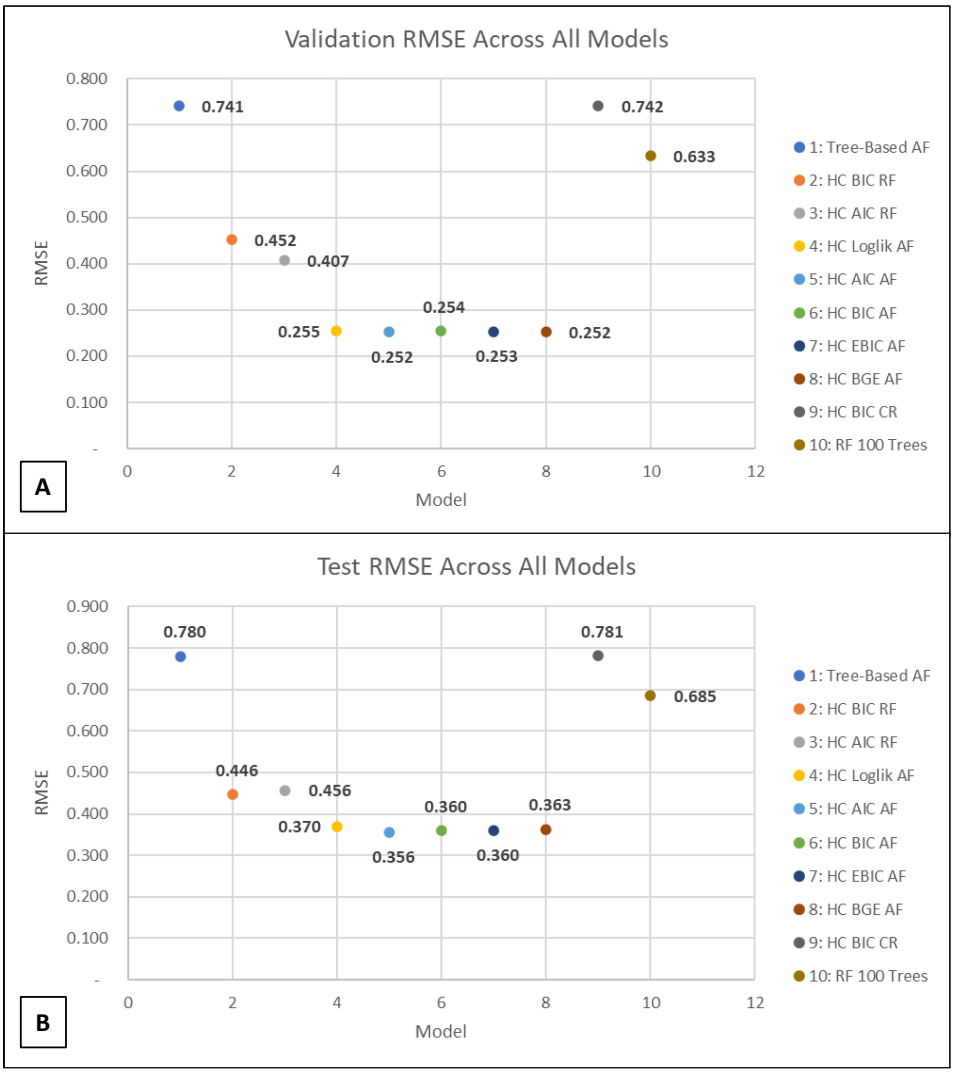


Figure 24: The RMSE evaluation metrics values per model for the A) validation set and B) test set.

When the MSE (Equation 5), values for the validation (Figure 25C) and test set (Figure 25D) are compared, the values for each model are similar, indicating stability and generalisation of the data during the training process with the selected training data. The HC-AIC-AF, HC-BIC-AF, and HC-EBIC-AF models show the best predictive performance when examining the unseen test set data.

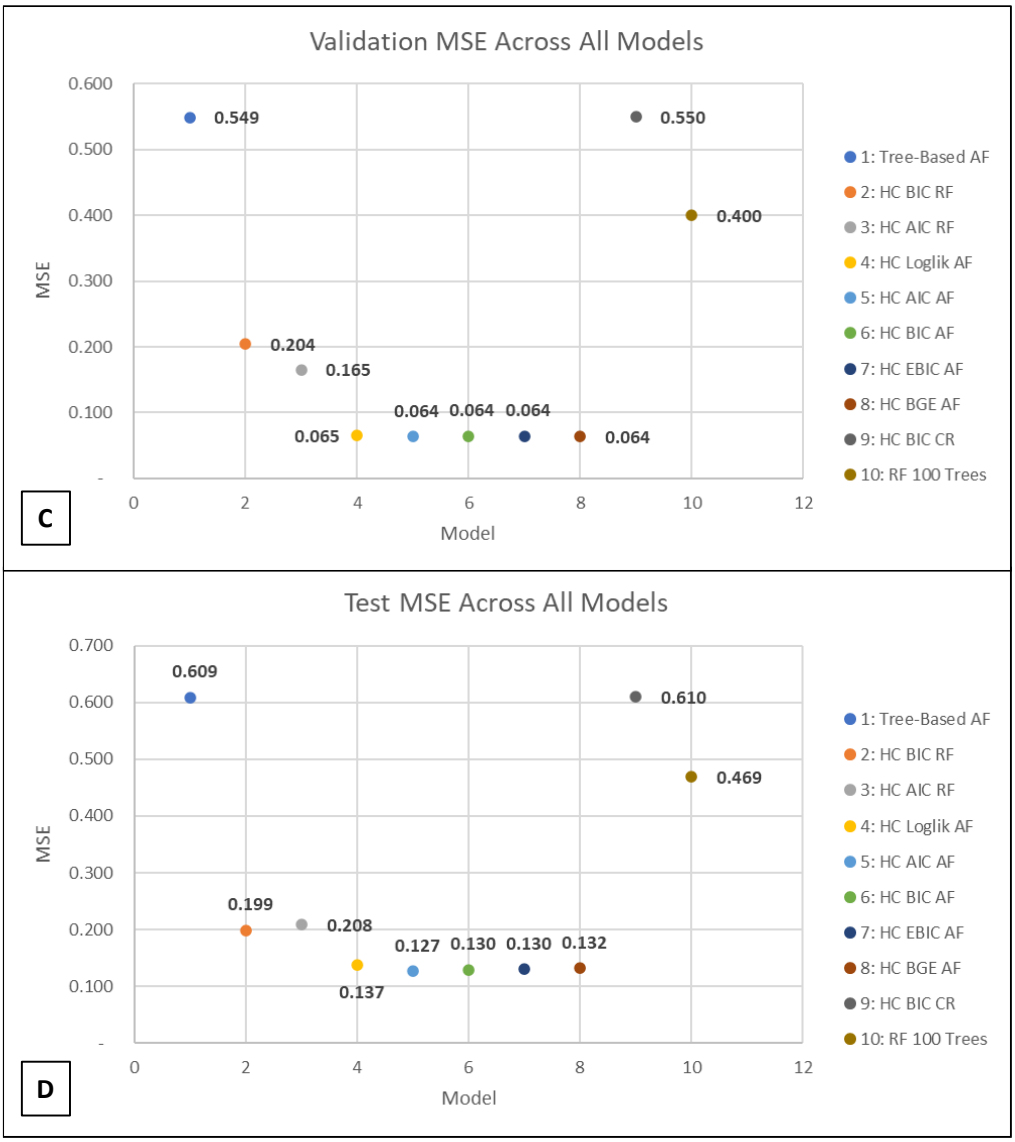


Figure 25: The MSE evaluation metric values per model for the C) validation set and D) test set.

When the MAE (Equation 6) values for the validation (Figure 26E) and test set (Figure 26F) are compared, the values for each model are similar, indicating stability and generalisation of the data during the training process with the selected training data. The HC-AIC-AF, HC-BIC-AF, HC-BGE-AF, and HC-EBIC-AF models show the best predictive performance when examining the unseen test set data.

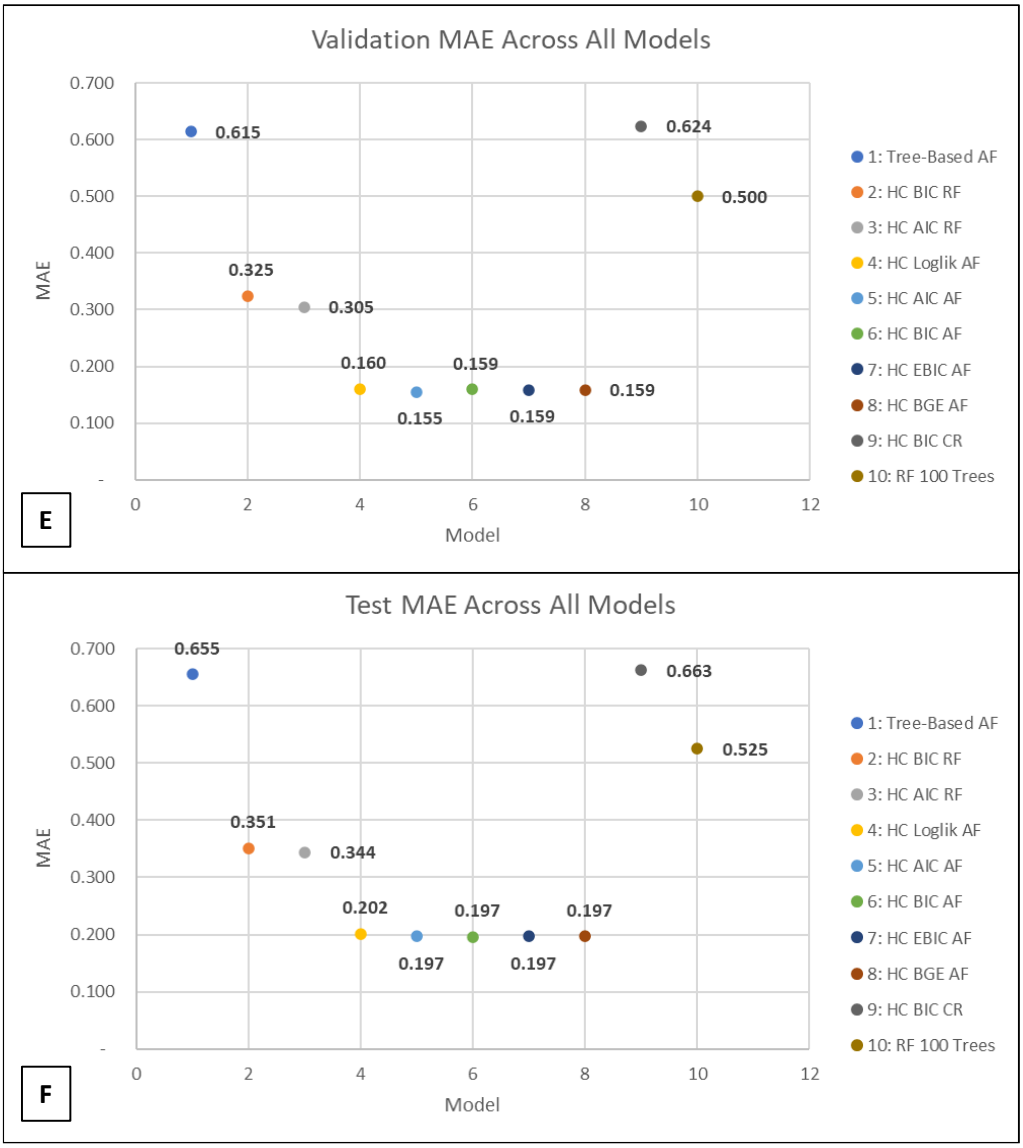


Figure 26: The MAE evaluation metric values per model for the E) validation set and F) test set.

When the MAPE (Equation 7) values for the validation (Figure 27G) and test set (Figure 27H) are compared, the values for each model are similar, indicating stability and generalisation of the data during the training process with the selected training data. The HC-Loglik-AF, HC-BIC-AF, and HC-EBIC-AF models show the best predictive performance when examining the unseen test set data.

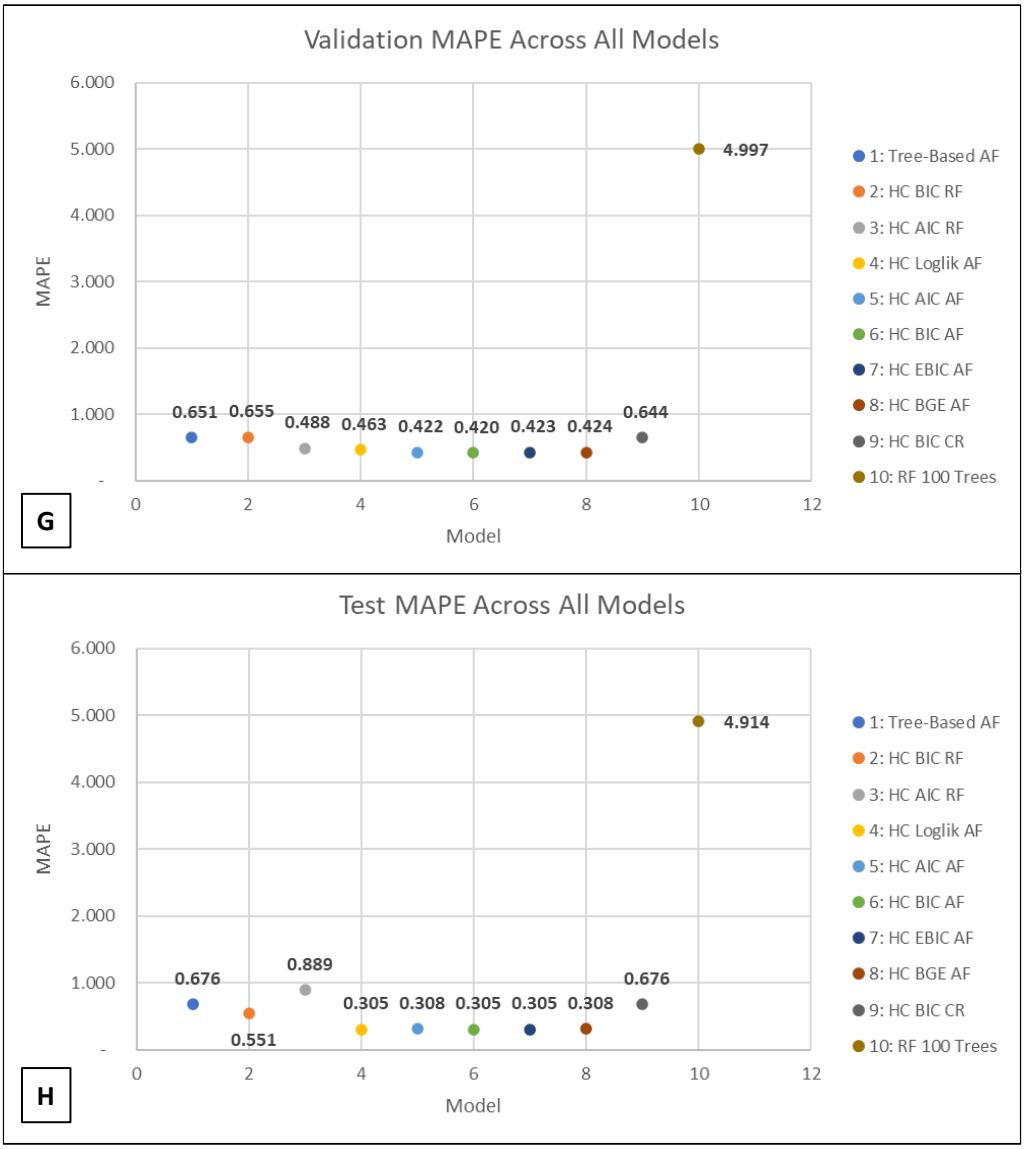


Figure 27: The MAPE evaluation metric values per model for the G) validation set and H) test set.

When the R-squared (Equation 8) values for the validation (Figure 28I) and test set (Figure 28J) are compared, the values for Bayesian models that use all the features are similar and more significant when compared to the other models, indicating that greater than 88% of the data can explain the variance in the data and thus represent models with good fit. The Tree-Based-AF, HC-BIC-CR and RF-100-trees have the lowest values when examining the unseen test set data. These values confirm the findings from the other evaluation metrics and, as mentioned by Chicco *et al*.

(2021), provide a more straightforward method of understanding how good a prediction is and the quality of the model predictions.

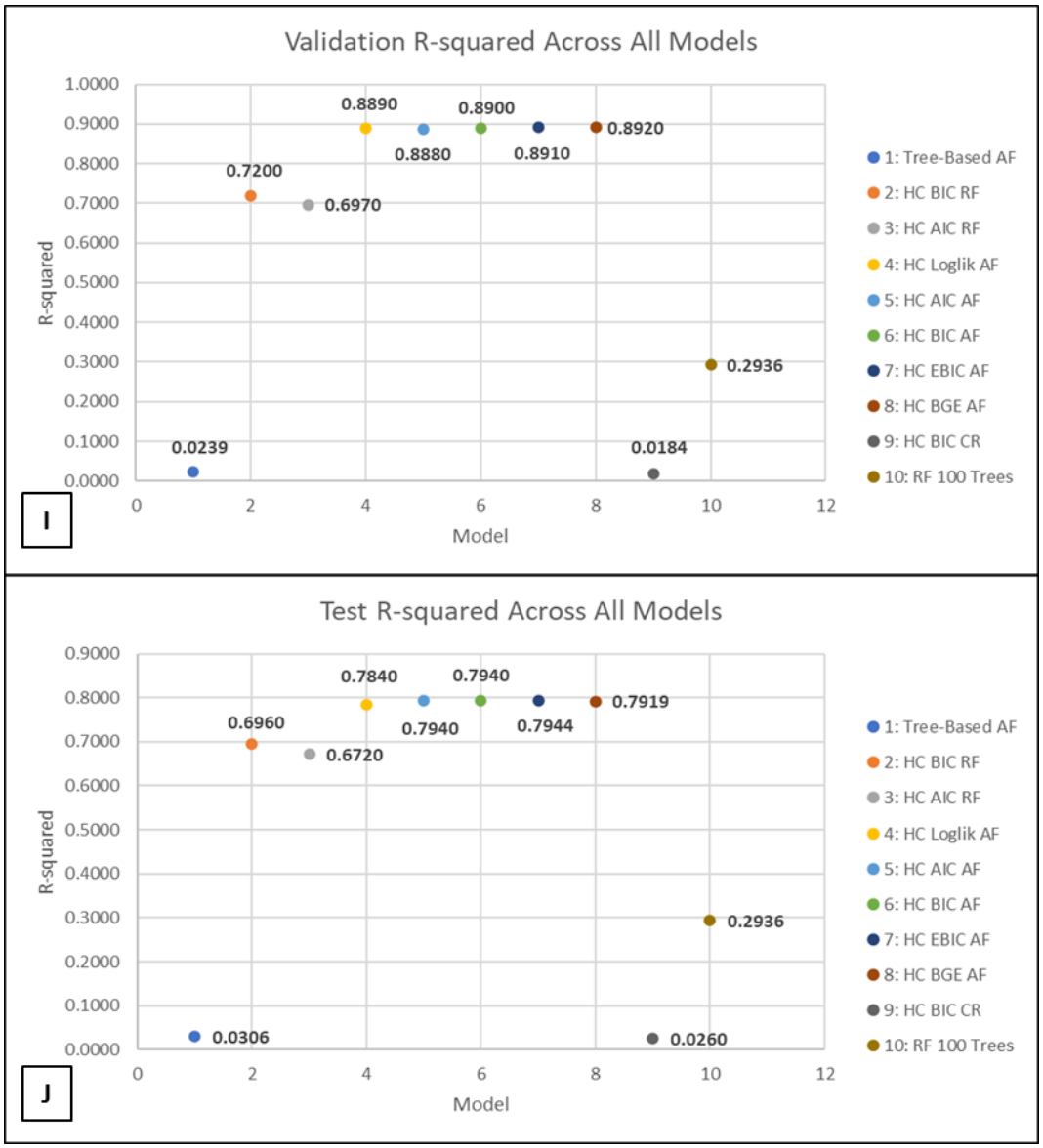


Figure 28: The R-squared evaluation metric values per model for the I) validation set and J) test set.

As an additional validation check to see whether different hyperparameters affect the fit of the Random Forest models, the R-squared (Equation 8) values for the validation (Figure 29K) and test set (Figure 29L) are compared; the larger the number of trees the larger the scores thus better fit and model performance. These

values are similar, indicating minimal improvement when just increasing the number of trees. The grid search to train the optimised Random Forest selected 500 trees, five minimum observations per node, and a maximum depth of 30 for the model, which performs similarly to the Random Forest with 100 trees.

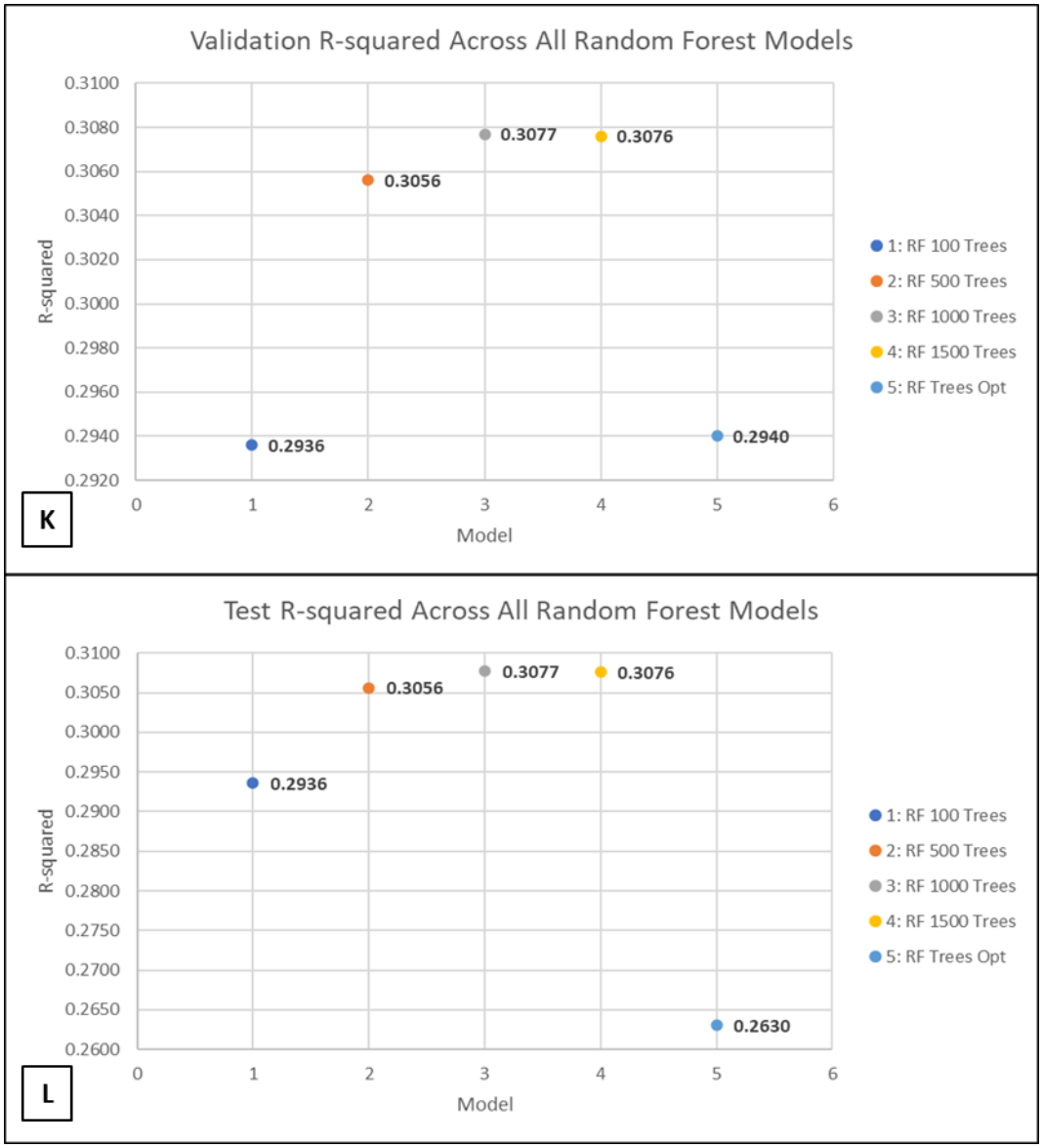


Figure 29: The R-squared evaluation metric values for all the Random Forest models for the K) validation set and L) test set.

Chapter 5

DISCUSSION AND CONCLUSION

5.1 Introduction

This chapter discusses the significance of the results from the Bayesian Belief Networks with learned structures: Are they able to predict the levels of nitrate and nitrite and correlated features? Are the models' prediction errors low enough to replace the original water monitoring station's nitrate and nitrite sensor?

5.2 Analysis of Results

The features that are correlated across all Bayesian models are summarised in Table 13, with the relationships between features, if any, expanded further. The correlated features are based on a base feature and follow the direction and number of the arrows to all other features if present from every BBN model.

Table 14: The final correlated features as extracted from the nine Bayesian models.

Final Correlated Features												
Calcium	Chloride	Conductivity	DMS	Fluoride	Magnesium	No3andno2	Phosphate	Potassium	Silicon	Sodium	Sulphate	Tal
conductivity	potassium	calcium	potassium	sulphate	no3andno2	sulphate	potassium	sulphate	sodium	silicon	tal	silicon
no3andno2	sulphate	phosphate	sulphate	silicon	sulphate	potassium	sulphate	phosphate	sulphate	sulphate	potassium	sodium
phosphate	tal	silicon	tal	tal	calcium	conductivity	silicon	silicon	phosphate	tal	conductivity	sulphate
sulphate	sodium	sulphate	sodium	phosphate	potassium		calcium	fluoride	conductivity	potassium		potassium
	silicon	potassium	silicon	calcium			no3andno2	calcium	calcium	phosphate		phosphate
	phosphate	no3andno2	phosphate	potassium			magnesium	no3andno2	no3andno2	fluoride		fluoride
	fluoride	magnesium	calcium	no3andno2				magnesium	magnesium	calcium		calcium
	calcium		conductivity	magnesium						conductivity		conductivity
	conductivity		no3andno2							no3andno2		no3andno2
	no3andno2		magnesium							magnesium		magnesium

This study focused on predicting the levels of nitrate and nitrite data and determining if the other features in the data have an effect or are correlated to a prediction. The features in the data that are linked directly to predictive performance (Figure 30) are magnesium, fluoride, DMS, conductivity, chloride, calcium, TAL, sodium, silicon, potassium and phosphate. There is no link between ammonium and

temperature, as they were excluded from the data, while sulphate has no link to nitrate and nitrite in any of the models.

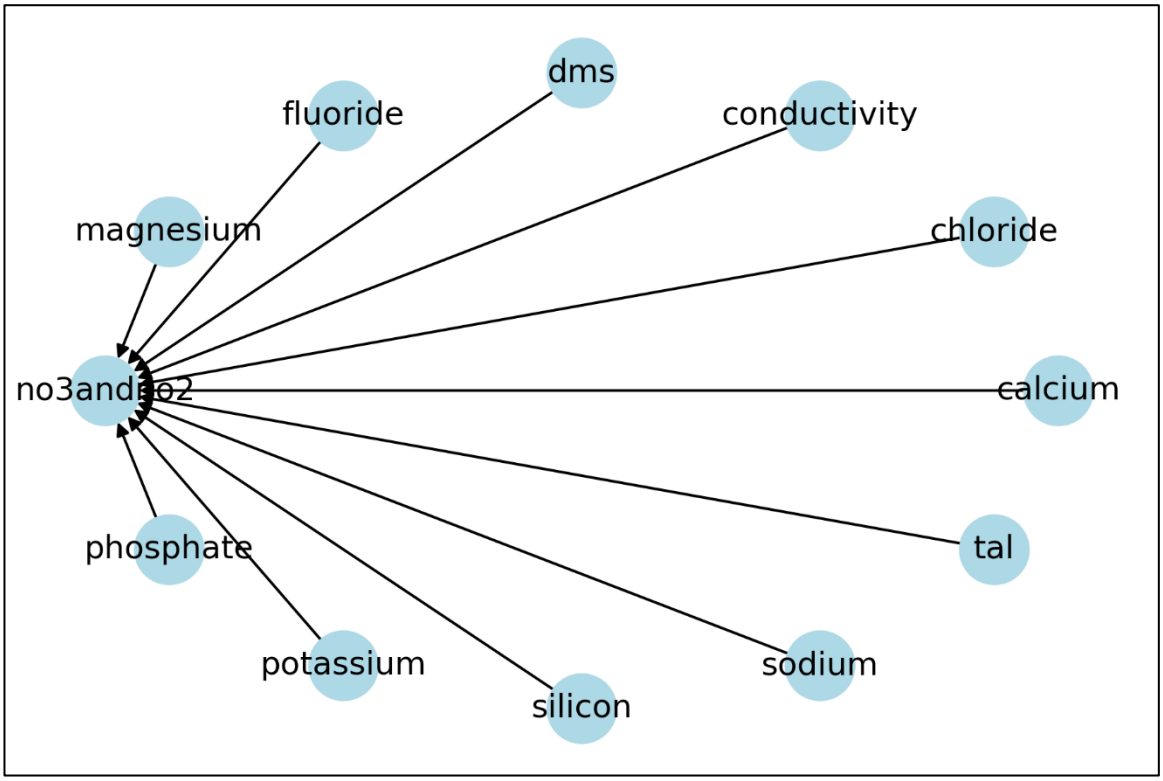


Figure 30: Features that influence nitrate and nitrite predictions, collated from all Bayesian Belief Networks.

The effects of treating contaminated sediment with in-situ calcium nitrate injection, as studied by Tongzhou *et al*., (2015), led to significant releases of nitrate, nitrite, and ammonia, with potential toxic effects on freshwater species. The research underscores the ecological risks of such treatments, particularly in stagnant water bodies, where accumulating these compounds can harm aquatic ecosystems, and this highlights the intricate relationship between calcium and nitrate/nitrite in water treatment and environmental impact.

Phosphates influence nitrate and nitrites in the models and are therefore correlated to one another; in untreated water bodies, nitrates and phosphates are major nutrients for microorganisms, but when the concentration exceeds the recommended limit, they are pollutants degrading the water quality and accelerating algae growth (Singh, 2013).

Nitrates, conductivity and turbidity are connected but may also be due to the area's climate or proximity to urban areas. Missing data, models being utilised at different locations but built using data from one, predicting critically important water quality variables without other features, and cost-prohibitive sensors are essential, relevant considerations in managing water quality (Kermorvant *et al*., 2023).

Groundwater samples collected from Erode District, Tamil Nadu in India, show that fluoride and calcium contribute to the total hardness of the drinking water and would be safe to drink after treatment (Karthikeyan *et al*., 2008). Samples collected from Nagaon District, Assam in India, showed a poor positive correlation between fluoride, magnesium and calcium but a strong correlation with sodium and potassium (Saikia & Sarma, 2011).

Agarwal *et al*., (2014) collected groundwater samples to correlate the water quality parameters for samples collected from 10 dug wells and borewells from different locations. Phulambri Taluqa show strong correlations between electrical conductivity, pH, turbidity, total hardness, calcium, magnesium, sodium, potassium, iron, total alkalinity, carbonate, bicarbonate, chloride, nitrate and sulphate.

A correlation matrix constructed using pH, conductivity, TDS, turbidity, total hardness, carbonate, bicarbonate, sulphates, nitrates, fluoride, calcium, magnesium, sodium, potassium and iron monitored over two years shows a good correlation between electrical conductivity and TDS (Agarwal *et al*., 2014).

Water samples collected from the Birjand flood plain of eastern Iran were measured, with correlations between the water quality parameters: pH, temperature, electrical conductivity, TDS, total hardness, calcium, magnesium, sodium,

potassium, chloride, sulphate, phosphate, bicarbonate, carbonate, nitrite, nitrate, DO, BOD5 and COD (Mansouri *et al*., 2013).

Gorham, (1958) found that hydrogen, potassium, calcium, nitrate, and sulphate are strongly correlated. Shah *et al*., (2008) studied the physio-chemical parameters of bore well water samples in villages of Gandhinagar, India, and found that sodium, potassium, calcium, magnesium, chloride, and fluoride are correlated.

Nemati *et al*., (2015) proposed a data-driven approach: MLR, adaptive neural-based fuzzy inference system (ANFIS), and ANN models were investigated to predict the water quality parameter and dissolved oxygen, with the resulting ANN performing better when evaluated. The parameters contributing to the performance were chloride, temperature, nitrate, nitrite, ammonia nitrogen and total phosphorus.

All the models have been able to predict nitrate and nitrite levels, but the HC-AIC-AF and HC-EBIC-AF perform the best:

1. These models, Figure 21E, and, Figure 21G respectively have the slightest difference in predicted and actual data points.
2. They have the lowest errors for test set evaluation metrics, RMSE Figure 24B, MSE, Figure 25D and MAE, Figure 26F.
3. They have the most stable CV performance, Figure 19E and, Figure 19G.
4. They have varying structural complexity; complex for HC-AIC-AF, Figure 16E and simpler for HC-EBIC-AF, Figure 16G.

The prediction performance can be demonstrated for the two best-predicting models by plotting the residuals (actual values minus the predicted values) per sample; the closer to zero, the more accurate the predictions. For both HC-AIC-AF (Figure 31A) and HC-EBIC-AF (Figure 31B), the residuals are closer to zero, as expected, with some values having large predicted values.

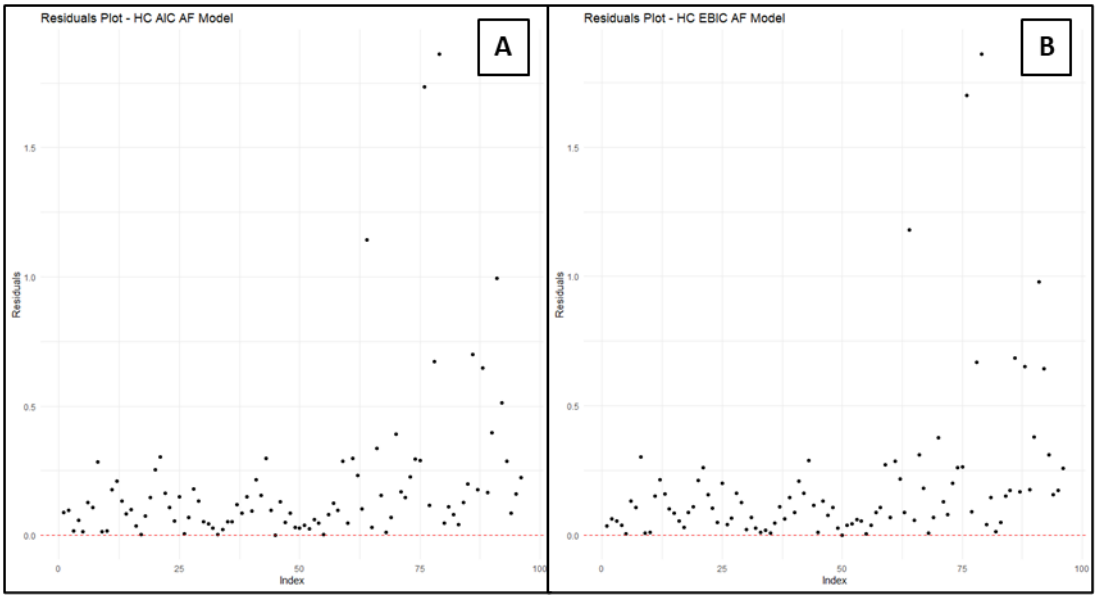


Figure 31: The residual plots for the A) HC-AIC-AF and B) HC-EBIC-AF models.

Suppose the residuals are compared directly to the data (Table 15). In that case, the performance scale for these models is demonstrated further by how close the predictions are to the actual values.

Table 15: A sample of the test set with the predictions and residuals of the HC-AIC-AF and HC-EBIC-AF models.

		HC-AIC-AF		HC-EBIC-AF	
Sample	**Actual**	**Predicted**	**Residuals**	**Predicted**	**Residuals**
1	0.622	0.535	0.087	0.585	0.037
2	1.43	1.334	0.096	1.367	0.063
3	1.679	1.663	0.016	1.624	0.055
4	0.251	0.310	0.059	0.291	0.040
5	0.647	0.660	0.013	0.653	0.006
6	0.964	1.092	0.128	1.098	0.134
7	0.757	0.649	0.108	0.649	0.108
8	2.757	2.473	0.284	2.453	0.304
9	0.02	0.035	0.015	0.030	0.010
10	0.504	0.487	0.017	0.491	0.013
11	0.02	0.197	0.177	0.173	0.153
12	1.466	1.258	0.208	1.251	0.215
13	1.965	1.833	0.132	1.804	0.161
14	1.982	1.900	0.082	1.880	0.102
15	0.886	0.985	0.099	0.971	0.085
16	1.457	1.420	0.037	1.400	0.057
17	1.579	1.583	0.004	1.548	0.031
18	1.296	1.221	0.075	1.207	0.089

5.3 Overview of Research

This research aimed to determine if using water monitoring station data to predict the levels of nitrate and nitrite in the Vaal River was possible using Bayesian Belief Networks, which learn the structure of the data automatically, using different scoring functions and features available in the Department of Water and Sanitation's data of 1071 data points collected between 1995 and 2016.

Firstly, the data is processed and removed, outliers and missing nitrate and nitrite data points are removed, any missing values are replaced instead of removed, and the data is split into train, validation, and test sets. Secondly, the Bayesian Belief Networks models were trained and evaluated in R with the bnlearn library. Finally, the best models were selected based on the evaluation performance of each model.

The main findings of this Research show that using prediction of nitrate and nitrite levels is possible with two models, the HC-AIC-AF and HC-EBIC-AF models, with RMSE on the validation data of 0.252 and 0.253, respectively, MSE on the test data of 0.127 and 0.130 and MAE on the test data of 0.197 and 0.197. These low error metric values and individual sample predictions (Table 15) demonstrate the success of using these techniques in water quality testing. The explainability of these networks also show that features in the data are correlated to the predicted nitrate and nitrite values, Figure 30.

The R-squared for these models (Figure 28) is high on both the validation data (0.888 and 0.7940) and the test data (0.7940 and 0.7944). The HC-EBIC-AF model represents the best fit across all models tested in this study, with the data explaining 79.44% of the model's variance. Running these models on hardware similar to what was used in the study to predict nitrate and nitrite levels using data with no physical sensor provides a suitable solution for cost-effective sensor replacement.

5.4 Limitations and Recommendations of Research

The data in this research are from one water monitoring station; the models' performance improves as the samples increase (Figure 22). The lack of additional data impacts how low and accurate the final predictions can be, but this presents an opportunity for enhancement by acquiring more data from more water monitoring stations and possibly more features, which is a recommendation for further study.

Log-transforming the data to see if there is improvement in the scores, underscoring the importance of data processing and model selection, will complete the thoroughness of the techniques used alongside traditional machine learning models like Random Forest, as used in this study, with a more exhaustive list of

hyperparameters, including column and sample selection, and implementing techniques like 'early stopping' (a method to prevent overfitting) could provide with lower error rates and greater prediction accuracy are recommendations for further study.

The best-performing models in the study can be paired with low-cost devices, such as Raspberry Pis or similar ARM devices, serving as replacements for various sensors, not only nitrite and nitrate, providing an innovative and easily deployable solution for water monitoring and data collection that water managers can use to reduce costs and positively affect environmental policy and management.

REFERENCES

Afrabandpey, H., Peltola, T., Piironen, J., Vehtari, A., & Kaski, S. (2020). A decision-theoretic approach for model interpretability in Bayesian framework. *Machine Learning*, 1855-1876.

Agarwal, A., & Xue, L. (2020). Model-Based Clustering of Nonparametric Weighted Networks with Application to Water Pollution Analysis. *Technometrics, 62*(2), 161-172.

Al Shehhi, M. R., & Kaya, A. (2020). Time series and machine learning to forecast the water quality from sateliite data. *arXiv*, 21.

Alpaydin, E. (2020). *Introduction to Machine Learning* (Fourth ed.). Cambridge, Massachusetts: The MIT Press.

Ananias, P. M., Negri, R. G., Dias, M. A., Silva, E. A., & Casaca, W. (2022). A Fully Unsupervised Machine Learning Framework for Algal Bloom Forecasting in Inland Waters using MODIS Time Series and Climatic Products. *remote sensing, 14*(4283), 1-22.

Anupama, K., Chalapathi Rao, Y., & Gurrall, V. K. (2020). A Machine Learning Approach to Monitor Water Quality in Aquaculture. *Internation Journal of Performability Engineering*, 1845-1852.

Asadollah, S. B., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering*, 1-14.

Banda, T. D., & Kumarasamy, M. (2020). Development of a Universal Water Quality Index (UWQI) for South African River Catchments. *Water, 12*(6), 1-20.

Bansal, S., & Geetha, G. (2020). A Machine Learning Approach towards Automatic Water Quality Monitoring. *Journal of Water Chemistry and Technology*, 321-328.

Barzegar, R., Aalami, M. T., & Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*, 415-433.

Brase, C. H., & Brase, C. P. (2012). *Understanding Basic Statistics* (Sixth ed.). Brooks/Cole.

Brill, E. (2014). Implementing Machine Learning Algorithms for Water Quality Event Detection: Theory and Practice. In *Securing Water and Wastewater systems* (pp. 107-122). Springer.

Britannica. (2024). *Vaal River*. Retrieved March 15, 2024, from https://www.britannica.com/place/Vaal-River

Castrillo, M., & García, A. L. (2020). Estimation of high frequency nutrient concentrations from water quailty surrogates using machine learning methods. *Water Research, 172*, 10.

CCME. (2017). *Canadian Water Quality Guidelines for the Protection of Aquatic Life*. Retrieved February 16, 2024, from https://ccme.ca/en/res/wqimanualen.pdf

Chaudhary, A., Rana, S., Gurian, P. L., Singh, R., Betancourt, W., Kumar, A., & Kumar, A. (2023). Non-potable water reuse and the public health risks from protozoa and helminths: a case study from a city with a semi-arid climate. *Journal of Water and Health, 21*, 13.

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., . . . Ren, H. (2020). Comparitative analyis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research, 171*, 1-9.

Cheng, L., & Dheda, D. (2020). A multivariate water quality parameter prediction model using recurrent nerual network. arXiv.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, 1-24.

Chou, J.-S., Ho, C.-C., & Hoang, H.-S. (2018). Determing quality of water in reservoir using machine learning. *Ecological Informatics, 44*, 57-75.

Climate Data. (2021). *Vredefort Climate*. Retrieved March 05, 2024, from https://en.climate-data.org/africa/south-africa/free-state/vredefort-27405/

Cloete, N. A., Malekian, R., & Nair, L. (2016). Design of Smart Sensors for Real-Time Water Quality Monitoring. *IEEE Access*, 1-14.

Crocker, J., & Bartram, J. (2014). Comparison and Cost Analysis of Drinking Water Quality Monitoring Requirements versus Practice in Seven Developing Countries. *Internation Journal of Environmental Research and Public Health*, 7333 - 7346.

Czerwinski, D., Geca, J., & Kolano, K. (2021). Machine Learning for Sensorless Temperature Estimation of a BLDC Motor. *Sensors*, *21*(14), 1-16.

da Costa, K. (2023, March 8). *SHAP Values: An Intersection Between Game Theory and Artificial Intelligence*. Retrieved from Holistic AI: https://www.holisticai.com/blog/shap-values-game-theory-and-ai

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.

Dhruba, A. R., Alam, K. N., Khan, M. S., Saha, S., Khan, M. M., Baz, M., . . . Alzain. (2022). IoT-Based Water Quality Assessment System for Industrial Waste WaterHealthcare Perspective. *Journal of Healthcare Engineering, 2023*, 1-4.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data, 8*(140), 37.

Gomez, D., Salvador, P., Sanz, J., & Casanova, J. L. (2021). A new approach to monitor water quality in the Menor sea (Spain) using satellite data and machine learning methods. *Environmental Pollution, 286*, 1-9.

Gorham, E. (1958). The influence and imprtance of daily wather conditions in the supply of chloride, sulphate and other ions to fresh waters from atmospheric precipitation. *Philosophical Transactions of the Royal Society of London.*, 147-178.

Grant, P. (2022). *How to Find Residuals in Regression Analysis*. Retrieved December 20, 2023, from https://builtin.com/data-science/how-to-find-residuals

Haghiabi, A. H., Nasrolahi, A. H., & Parsiae, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal, 53*(1), 3-12.

Heller, M. (2023). *The engines of AI: Machine learning algorithms explained*. Retrieved December 8, 2023, from IDG Communications, Inc.: https://www.infoworld.com/article/3394399/machine-learning-algorithms-explained.html

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development, 15*(14), 6.

IBM. (2020, July 15). *Machine Learning*. Retrieved from IBM.com: https://www.ibm.com/za-en/cloud/learn/machine-learning

Intel Corporation. (2024, August 3). *Intel® Core™ i5-9500T Processor*. Retrieved from www.intel.com: https://ark.intel.com/content/www/us/en/ark/products/191052/intel-core-i5-9500t-processor-9m-cache-up-to-3-70-ghz.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). New York: Springer.

Jiang, J., Larson, N. B., Prodduturi, N., Flotte, T. J., & Hart, S. N. (2019). Robust hierarchical density estimation and regression for re-stained histological whole slide image co-registration. *PLOS ONE, 14*(7), 1-11.

Kanyama, Y., Ajoodha, R., Seyler, H., Makondo, N., & Tutu, H. (2020). Application of Machine Learning Techniques In Forecasting Groudwater Levels in the Grootfontein Aquifer. *IMITEC.*

Karthikeyan, K., Nanthakumar, K., Velmurugan, P., Tamilarasi, S., & Lakshmanaperumalsamy, P. (2008). Prevalence of certain inorganic constituents in groundwater samples of Erode district, Tamilnadi, India, with special emphasis on fluoride, fluorosis and its remedial measures. *Environmental Monitoring and Assesment, 160*(2010), 141-155.

Keller, S., Maier, P. M., Riese, F. M., Norra, S., Holbach, A., Borsig, N., . . . Hinz, S. (2018). Hyperspectral Data and Machine Learning for Estimating CDOM, Chlorophyll-a, Diatoms, Green Algae and Turbidity. *Internation Journal of Environmental Research and Public Health, 15*(1881), 1-15.

Kermorvant, C., Liquet, B., Litt, G., Mengersen, K., Peterson, E. E., Hyndman, R. J., . . . Leigh, C. (2023). Understanding links between water-quaility variables and nitrate concentration in freshwater streams using high frequency sensor data. *PLOS ONE, 18*(6), 16.

Kılıç, Z. (2020). The importance of water and conscious use of water. *International Journal of Hydrology, 4*(5), 239-241.

Kumar, D., Kumar, R., Sharma, M., Awasthi, A., & Kumar, M. (2024). Global water quality indices: Development, implications, and limitations. *Total Environment Afvances, 9*, 1-15.

Li, Y., Mao, S., Yuan, Y., Wang, Z., Kang, Y., & Yao, Y. (2023). Beyond Times and Time: Machine Learning's Triumph in Water Quality Forecasting. arXiv.

Liu, C., Li, H., & Zhang, Q. (2020). Research on Sewage Monitoring and Water Quality Prediction Based on Wireless Sensors and Support Vector Machines. *Wireless Communications and Mobile Computing, 2020*(1), 1-8.

Liu, Y., & Wu, H. (2017). Water bloom warning model based on random forest. *International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, 45-48.

Liu, Z., Malone, B., & Yuan, C. (2012). Empirical evaluation of scoring functions for Bayesian netwok model selection. *Bioinformatics, 13*, 16.

Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere, 249*(2020), 2-10.

Lu, X., Yuan, D., Chen, Y., Fung, J. C., Li, W., & Lau, A. K. (2020). Estimations of Long-Term nss-SO42- and NO32- Wet Depositions over East Asia by Use of Ensemble Machine-Learning Method. *Environmental Science & Technology*, 11118-11126.

Lundberg, S. M., & Su-in, L. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 4765-4774.

Machine Learning Life Cycle. (2021). Retrieved from DataRobot, Inc.: https://www.datarobot.com/wiki/machine-learning-life-cycle/

Madni, H. A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., . . . Ashraf, I. (2023). Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques. *Water*, 475.

Mansouri, B., Salehi, J., Etebari, B., & Kardanmoghadam, H. (2013). Assesement of the Physico-Chemical Parameters of Surface Water Samples in Birjand Flood Plain, Iran. *Science Journal of University of Zakho, 1*, 420-426.

Marselina, M., Wibowo, F., & Mushfiroh, A. (2022). Water quality index assesment methods for surface water: A case study of the Citarum River in Indonesia. *Heliyon, 8*(7), 1-10.

Memon, N., Unar, M. A., & Ansari, A. K. (2012). pH Prediction by Artificial Neural Networks for the Drinking water of the Distribution System of Hyderabad City. *Mehran University Research Journal of Engineering & Technology, 31*(1), 10.

Moffa, G., Catone, G., Kuipers, J., Kuipers, E., Freeman, D., Marwaha, S., . . . Bebbington, P. (2017). Using Directed Acyclic Graphs in Epidemiological Research in Psychosis: An Analysis of the Rolse of Bullying in Psychosis. *Schizophrenia Bulletin, 43*, 6.

Molnar, C. (2023). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. In C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (p. 328).

Nayan, A.-A., Saha, J., Mozumder, A. N., Mahmud, K. R., Azad, A. K., & Kibria, M. G. (2021). A Machine Learning Approach for Early Detection of Fish Diseases by Analyzing Water Quality. *Trends in Sciences, 18*(21), 11.

Nazempour, R., Monfared, M. A., & Zio, E. (2018). A complex network theory approach for optimzing contamination warning sensor location in water distribution networks. *International Journal of Disaster Risk Reduction, 30*, 225-234.

Nemati, S., Fazelifard, M. H., & Terzi, O. (2015). Estimation of dissolve oxygen using data-driven techniques in the Tai Po River, Hong Kong. *Environmental Earth Sciences, 74*, 4065-4073.

Nexsens Technology, I. (2018, December 19). *Nutrient Monitoring - Nitrate*. Retrieved from Nexsens Technology: https://www.nexsens.com/blog/nutrient-monitoring-nitrate.htm

Ohm, C. (2021, September 28). *Industrial Water and Water Pollution*. Retrieved from water-pollution.org.uk: https://www.water-pollution.org.uk/industrial-water-pollution/

Orellana, G., del Mar Darder, M., & Quílez-Alburquerque, J. (2023). Luminescence-Based Sensors for Water Quality Analysis. *Encyclopedia of Sensors and Biosensors, 4*, 599-613.

Ortega, E., Lage-Castellanos, A., & Braunstein, A. (2017). Contamination source inference in water distribution networks. *Revista Cubana de Fisica, 34*, 12.

Panidhapu, A., Li, Z., Aliashrafi, A., & Peleato, N. M. (2019). Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. *Water Research, 170*, 11.

Posit Software. (2023). *RStudio Desktop*. Retrieved August 14, 2023, from https://posit.co/download/rstudio-desktop/

Project Jupyter. (2023). *Jupyter*. Retrieved 2023, from Project Jupyter: https://jupyter.org/

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O' Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics, 2*, 17.

Pujar, P. M., Kenchannavar, H. H., Kulkarni, R. M., & Kulkarni, U. P. (2020). Real-time water quality monitoring through Internet of Things and ANOVA-based analysis: a case study on river Krisha. *Applied Water Science*, 10-22.

Python Software Foundation. (2022). *Python*. Retrieved from https://www.python.org/: https://www.python.org/

Rahbaralam, M., Modesto, D., Cardús, J., Abdollahi, A., & Cucchietti, F. M. (2020). Predictive Analytics for Water Asset Management: Machine learning and Survival Analysis. arXiv.

Rubio-Arias, H., Ochoa-Rivero, J. M., Quintana, R. M., Saucedo-Teran, R., Ortiz-Delgado, R. C., Rey-Burciaga, N. I., & Espinoza-Prieto, J. R. (2013). Development of a Water Quality Index (WQI) of an Artifical Aquatic Ecosystem in Mexico. *Journal of Environmental Protection*, 1296-1306.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Model Instead. *Nature Machine Intelligence*, 206-215.

Saikia, M., & Sarma, H. P. (2011). Fluoride geochemistry of Kollong river basin, Assam, India. *Archives of Applied Science Research, 3*, 367-372.

Sajedi-Hosseini, F., Malekian, A., Choubin, B., Omid, R., Cipullo, S., Coulon, F., & Pradhan, B. (2019). A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the Total Environment*, 954-962.

scikit-learn developers. (2023). *3.1. Cross-validation: evaluating estimator performace*. Retrieved December 20, 2023, from https://scikit-learn.org/stable/modules/cross_validation.html

Scutari, M. (2023). *bnlearn - an R package for Bayesian network learning and inference*. Retrieved July 8, 2023, from https://www.bnlearn.com/

Scutari, M. (2023). *Structure learning algorithms*. Retrieved 07 15, 2023, from https://www.bnlearn.com/documentation/man/structure.learning.html

Sengupta, P. (2013). Potential health Impacts of Hard Water. *International Journal of Preventive Medicine*, 866-875.

Shah, M., Shilpkar, P., & Acharya, P. B. (2008). Ground Water Quality of Gandhinagar Taluka, Gujarat, India. *Journal of Chemistry, 5*, 435-446.

Shittu, E., Lakhanpaul, M., Vigurs, C., Sarkar, K., Koch, M., Parikh, P., & Campos, L. C. (2023). A rapid systematic scoping review of research on the impacts of water contaminated by chemicals on very young children. *Science of the Total Environment, 891*, 16.

Singh, A. L. (2013). Nitrate and Phosphate Contamination in Water and Possible Remedial Measures. In *Environmental Problems and Plant* (1st ed., pp. 44-56). Heidelberg, Germany: Springer Verlag GmbH.

Sipper, M., & Moore, J. H. (2021). Conservation machine learning: a case study of random forests. *Nature, 11*, 6.

Sutadian, A. D., Muttil, N., Yilmaz, A., & Perera, C. (2016). Development of River Water Quality Indices - A Review. *Environmental Monitoring and Assessment*, 188:58.

The Mathworks, Inc. (2024). *Visualize and Assess Model Performance in Regression Learner*. Retrieved March 15, 2024, from https://www.mathworks.com/help/stats/assess-model-performance-in-regression-learner.html

The R Foundation. (2023). *What is R?* Retrieved November 10, 2023, from https://www.r-project.org/about.html

Thrun, M. C., Ultsch, A., & Breuer, L. (2021). Explainable AI Framework for Multivariate Hydrochemical Time Series. *Machine Learning & Knowledge Extraction*, 170-204.

Tongzhou, L., Yuan, J., Dong, W., Wu, H., & Wang, H. (2015). Effects on inorganic nitrogen compounds release of contaminated sediment treatment with in situ calcium nitrate injection. *Environmental Science and Pollution Research International, Jan*(22), 10.

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications and Algorithms. *Electronics, 12*(1789), 1-43.

United Nations. (2021). *Progress on Wastewater Treatment - 2021 Update*. Retrieved Ferbruary 15, 2024, from https://www.unwater.org/publications/progress-wastewater-treatment-2021-update

Velimirovic, l. Z., Jankovic, R., Velimirovic, J. D., & Janjic, A. (2021). Wastewater Plant Reliability Predcition Using the Machine Learning Classification Algorithms. *Symmetry, 13*(1518), 1-17.

Voza, D. (2020). Comparative Analysis of Efficiency of Various Supervised Machine-Learning Algorithms in Water Quality Prediction. *International May Conference on Strategic Management.*

Wang, L., Zhu, Z., Sassoubre, L., Yu, G., Liao, C., Hu, Q., & Wang, Y. (2021). Improving the robustness of beach water quality modeling using an ensemble machine learning approach. *Science of the Total Environment*.

Water Quality Association. (2023). *Nitrites/Nitrates Fact Sheet.* Retrieved February 15, 2024, from https://wqa.org/resources/nitrites-nitrates/

Watson, M. L., Hickman, S. H., Dreesbeimdiek, K. M., Kohler, K., & Stubbs, D. J. (2023). Directed acyclic graphs in perioperative observational research-A systematic review and critique against best practice recommendations. *Plos One, 2*, 18.

Wei, Y., Huang, H., Chen, B., Zheng, B., & Wang, Y. (2019). Application of Extreme Learning Machine for Predicting Chlorophyll-a Concentration In Artifical Upwelling Processes. *Mathematical Problems in Engineering, 2019*(1), 1-11.

World Health Organization. (2017, April 24). *Guidelines for drinking-water quality, 4th edition, incorporating the 1st addendum.* Retrieved February 15, 2024, from https://www.who.int/publications/i/item/9789241549950

World Health Organization. (2021, December 16). *A global overview of national regulations and standards for drinking-water quality, 2nd ed.* Retrieved February 15, 2024, from https://www.who.int/publications/i/item/9789240023642

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. *Natural Language Processing and Chinese Computing*, 563-574.

Yashaswini, M., Sowmya, G., & Jai deep, P. (2021). Weather Clustering Using K-Means. *International Journal of Creative Research Thoughts (IJCRT), 9*(6), 1-4.

Yip, W. (2020). *Lifecycle of machine learning models.* California, USA: Oracle Corporation. Retrieved from oracle.com: https://www.oracle.com/a/ocom/docs/data-science-lifecycle-ebook.pdf

Yu, X., Shen, J., & Du, J. (2020). A Machine-Learning-Based Model for Water Quality in Coastal Waters, Taking Dissolved Oxygen and Hypoxia in Chesapeake Bay as an Example. *Water Resources Research, 56*(9), 1-16.

Zar, J. H. (2010). *Biostatistical Analysis.* New Jersey: Pearson.

Zhou, J., Qi, G., & Liu, C. (2021). A Chaotic Parallel Artifical Fish Swarm Algorithm for Water Quality Monitoring Sensor Networks 3D Coverage Optimization. *Journal of Sensors, 2021*(1), 1-10.