PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

الجمهوريــة الجزائريــة الديمقراطيــة الشعبيــة

Ministry of Higher Education and Scientific Research

وزارة التعليم العالي والبحث العلمي

University of Mohamed Khider Biskra

Civil and Hydraulic Engineering Department

Faculty of Science and Technology

Ref: …

جامعة محمد خيضر- بسكرة

قسم الهندسة المدنية و الري

كلية العلوم و التكنولوجيا

المرجع :…

Dissertation submitted for the **degree** of

Doctor of Philosophy 3rd cycle LMD

In : Hydraulic

Option : Water resources

Theme

Classification Of Irrigation Water Based On Machine Learning Approach

Presented by :

Aymen ZEGAAR

Publicly Defended, the 27/02/2025, before the jury composed of:

Ms.	Leila Youcef	Professor	President
Ms.	Samira Ounoki	Professor	Superviser
Mr.	Abdelmoutie Telli	MCA	Co-Superviser
Mr.	Djeddou Messaoud	Professor	Examiner
Mr.	Soheyb Ayad	MCA	Examiner

To my Parents,

Your unwavering love, encouragement, and enduring sacrifices have been the bedrock of my journey. This achievement is a testament to your profound impact on my life and aspirations. I dedicate this work to you with deep gratitude.

To my Brothers and Sisters,

In every challenge and triumph, you've been my steadfast companions. Your unwavering support and shared joys have made this academic pursuit a richer experience. I dedicate this work to our collective bond.

To the Wider Family,

For your understanding, encouragement, and shared pride in my accomplishments. Your support, spanning generations, has been a source of strength. I dedicate this work to the extended family that stands as a pillar of support.

To my Supervisors,

Your guidance, wisdom, and mentorship have been instrumental in shaping the trajectory of my PhD journey. I am grateful for your expertise and unwavering support. I dedicate this work to your invaluable contributions.

To All Those Who Contributed Directly or Indirectly,

To friends, colleagues, and mentors who have played a role, seen and unseen, in the creation of this thesis. Your varied contributions have enriched my academic experience. I dedicate this work to the collaborative spirit that made it possible.

This work is dedicated to those who have been my constants, my pillars of strength, and my sources of inspiration throughout the journey of my Ph.D. Your collective presence has made this accomplishment all the more meaningful.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Aymen ZEGAAR
March 2025

Acknowledgements

The authors express gratitude to the anonymous referees for their valuable remarks and comments. Special thanks are extended to the water analysis laboratory of the Algerian Water Company (ADE) in Msila for providing access to the necessary data for this work. The authors acknowledge the support received from the Directorate-General for Scientific Research and Technological Development (DGRSDT) in Algeria, as well as the LARHYSS laboratory of Biskra University.

ملخص

هذه الدراسة تهدف لاستغلال نماذج التعلم الآلي المتقدمة في تصنيف مياه الري. بداية من تقييم جودة المياه الجوفية من خلال IWQI وتصنيفها، ثم يتطور البحث للاستفادة من نماذج التعلم الآلي القابلة للتفسير من اجل التنبؤ. يمثل هذا البحث تحولا جذريا في منهجيات تقييم جودة المياه، مبرزا تحقيق كفاءة أكبر. تطبيق التعلم الآلي يضمن محاكاة دقيقة لمؤشر جودة مياه الري IWQI وطريقة اقتصادية محسنة لادارته. يحمل هذا العمل آثاراً كبيرة على إدارة موارد المياه، ويعود بالفائدة بشكل خاص على الفلاحين واصحاب القرار، مما يسهم في تطور ممارسات إدارة المياه المستدامة، ويقدم وجهة نظر محورية في تقييم جودة مياه الري باستخدام التعلم الآلي.

الكلمات المفتاحية : الري، نوعية المياه الجوفية، التصنيف، تعلم الآلة، النموذج الاقتصادي، مؤشراة جودة المياه.

Abstract

This thesis pioneers the integration of advanced machine learning models into irrigation water classification. Starting from groundwater quality assessment through IWQI, and groundwater classification, the research evolves to leverage ML model interpretability for predictions. It marks a paradigm shift in water quality assessment methodologies, emphasizing potential efficiency gains. The application of machine learning assures accurate simulation of the Irrigation Water Quality Index (IWQI) and streamlined economic monitoring approach. This work carries substantial implications for water resource management, particularly benefiting farmers and decision-makers. The findings contribute to the advancement of sustainable water management practices, providing a transformative perspective at the intersection of machine learning and irrigation water quality assessment.

Keywords: Irrigation, Groundwater quality, classification, Machine learning, Economic model, Water quality indices.

Résumé

Cette thèse inaugure l'intégration de modèles avancés d'apprentissage automatique dans la classification de l'eau d'irrigation. Partant de l'évaluation de la qualité de l'eau souterraine par l'IWQI et la classification de l'eau souterraine, la recherche évolue pour exploiter l'interprétabilité des modèles d'apprentissage automatique pour les prédictions. Cela marque un changement de paradigme dans les méthodologies d'évaluation de la qualité de l'eau, mettant l'accent sur des gains d'efficacité potentiels. L'application de l'apprentissage automatique assure une simulation précise de l'indice de qualité de l'eau d'irrigation (IWQI) et une approche économique rationalisée pour la surveillance. Ce travail a des implications substantielles pour la gestion des ressources en eau, bénéficiant particulièrement aux agriculteurs et aux décideurs. Les résultats contribuent à l'avancement des pratiques durables de gestion de l'eau, offrant une perspective transformative à l'intersection de l'apprentissage automatique et de l'évaluation de la qualité de l'eau d'irrigation.

mots clés : Irrigation, Qualité des eaux souterraines, Classification, Apprentissage automatique, Modèle économique, Indices de qualité de l'eau.

Table of contents

List of figures **xv**

List of tables **xvii**

Chapter 1 General introduction **1**
1.1 Study background 1
1.2 Research Gap 2
1.3 Research problem 2
1.4 The rationale of the study 3
1.5 Study objectives 3
1.6 Thesis structure 4

Chapter 2 Literature review **7**
2.1 Introduction 7
2.2 Water quality indices for quality assessment 7
2.3 Machine learning elevating for water quality classification and prediction . 11
2.4 Conclusion 14

Chapter 3 Methodology **17**
3.1 Introduction 17
3.2 Study area and Data description 17
3.3 Water quality parameters 20
3.4 Water quality indices 24
3.4.1 The irrigation water quality index (IWQI) 24
3.4.2 Sodium Adsorption Ratio (SAR) 25
3.4.3 Soluble Sodium Percent (Na%) 25
3.4.4 Potential Salinity (PS) 26
3.4.5 Permeability Index (PI) 26
3.4.6 Magnesium Adsorption Ratio (MAR) 26

3.4.7 Kelly's Ratio (KR) . . . 26
3.5 Hydrochemical Facies Characterization using Piper Diagram . . . 27
3.6 Data Preprocessing . . . 27
3.6.1 Feature Relabeling . . . 28
3.6.2 Missing Value Imputation . . . 28
3.6.3 Outliers Detection and Handling . . . 29
3.6.4 Data Partitioning . . . 29
3.6.5 Data Standardization . . . 30
3.7 Correlation analysis . . . 30
3.8 Feature Engineering . . . 31
3.8.1 Feature Generation . . . 31
3.8.2 Recursive Feature Elimination with Cross-Validation (RFECV) . . . 31
3.8.3 Permutation Importance (PI) . . . 32
3.8.4 Mutual Information (MI) . . . 32
3.9 ML models . . . 32
3.9.1 Random forest . . . 32
3.9.2 Extra Trees . . . 34
3.9.3 Gradient Boosting Classifier . . . 34
3.9.4 XGBoost . . . 36
3.9.5 Categorical Boosting (Catboost) . . . 37
3.9.6 LightGBM (LGBM) . . . 38
3.9.7 Support Vector Machine (SVM) . . . 38
3.10 Performance evaluation metrics . . . 40
3.10.1 Root Mean Squared Error (RMSE) . . . 41
3.10.2 Mean Absolute Error (MAE) . . . 41
3.10.3 Accuracy . . . 42
3.10.4 Precision and Recall . . . 42
3.10.4.1 Precision . . . 43
3.10.4.2 Recall . . . 43
3.10.5 F1 Score . . . 43
3.10.6 ROC Curve and AUC-ROC . . . 44
3.10.6.1 ROC Curve Plotting . . . 44
3.10.6.2 Area Under the Curve (AUC-ROC) . . . 44
3.10.6.3 Interpretation and Significance . . . 44
3.11 SHAP (SHapley Additive exPlanations) . . . 45
3.11.1 Philosophy and Foundation . . . 45

3.11.2 Elucidating Predictive Outputs . . . 45
3.11.3 Interpretability Across Models . . . 45
3.11.4 Contributions Visualized . . . 46
3.12 Adopted Methodology . . . 46
3.13 Conclusion . . . 46

Chapter 4 Water Quality Assessment For Irrigation Purposes 49
4.1 Introduction . . . 49
4.2 Methods . . . 50
4.3 Results and discussion . . . 50
4.3.1 Salinity hazard . . . 51
4.3.2 pH . . . 53
4.3.3 Total Hardness . . . 53
4.3.4 Ions Concentration . . . 54
4.3.5 Total Alkalinity . . . 55
4.3.6 Water quality indices . . . 56
4.3.6.1 Irrigation Water Quality Index (IWQI) . . . 56
4.3.6.2 Sodium Adsorption Ratio (SAR) . . . 57
4.3.6.3 Sodium Percent (Na%) . . . 58
4.3.6.4 Kelly's Ratio (KR) . . . 59
4.3.6.5 Permeability Index (PI) . . . 59
4.3.6.6 Magnesium Adsorption Ratio (MAR) . . . 60
4.3.6.7 Potential Salinity (PS) . . . 60
4.3.7 Hydrochemical Characterization of the Water Samples . . . 61
4.4 Discussion . . . 62
4.5 Conclusion . . . 63

Chapter 5 ML-Based Irrigation Water Quality Classification 65
5.1 Introduction . . . 65
5.2 Data Preprocessing . . . 66
5.3 Correlation Analysis of Parameters . . . 66
5.4 Methodology . . . 68
5.5 Model Evaluation Metrics . . . 69
5.6 Results And Discussion . . . 70
5.6.1 First Scenario . . . 70
5.6.2 Second Scenario . . . 72
5.7 Discussion . . . 74

5.8 Conclusion 75

Chapter 6 Interpretable ML for irrigation water Quality Prediction 77

6.1 Introduction 77
6.2 Correlation Analysis 78
6.3 Feature selection 79
6.4 Results and discussion 81
6.4.0.1 Models performances 81
6.4.0.2 SHAP values interpretation 84
6.5 The novelty of the study 88
6.6 Conclusion 89

Chapter 7 General Conclusion 91

References 95

List of figures

Fig. 2.1: WQI historical development (Uddin et al., 2021a) 8

Fig. 3.1: The geographical location of the study area 18
Fig. 3.2: Land cover distribution of the study area 19
Fig. 3.3: Train test split and Cross validation 30
Fig. 3.4: The architecture of the random forest model 33
Fig. 3.5: The architecture of the gradient boosting trees model (Deng et al., 2021) 35
Fig. 3.6: The architecture of the extreme gradient boosting trees model 36
Fig. 3.7: The optimal hyperplane of SVM model 39

Fig. 4.1: Stiff diagram: a) Stiff diagram of 2019, b) Stiff diagram of 2020; c) Stiff diagram of 2021, d) Stiff diagram of 2022 55
Fig. 4.2: Pipper diagram 61

Fig. 5.1: Correlation Heatmap of the parameters used as inputs 67
Fig. 5.2: Features importance ranking based on MI 68
Fig. 5.3: Pairwise Relationship Plots of Input Variables in Two Scenarios . . . 69
Fig. 5.4: Performance of the ML models in the 2 scenarios: a) First scenario, b) Second scenario 72
Fig. 5.5: Confusion matrices of the ML models for the first scenario: a) CatBoost classifier, b) Extra Trees Classier, c) Multilayer perceptrons classifier, d) Gradient Boosting classifier, e) LGBM classifier, f) Random Forest classifier, g)KNN classifier, h)SVM classifier 73
Fig. 5.6: Confusion matrices of the ML models for the second scenario: a) CatBoost classifier, b) Extra Trees Classier, c) Multilayer perceptrons classifier, d) Gradient Boosting classifier, e) LGBM classifier, f) Random Forest classifier, g)KNN classifier, h)SVM classifier 74

Fig. 6.1: Heatmap of correlation values 78

Fig. 6.2: Recursive feature elimination with cross-validation plot 80
Fig. 6.3: Residual plots of the employed ML models 83
Fig. 6.4: Scatter plots of the employed ML models 84
Fig. 6.5: Feature importance based on SHAP values 85
Fig. 6.6: Waterfall plot of 4 random samples 86
Fig. 6.7: Beeswarm plot of SHAP values 86

List of tables

Table 3.1 Descriptive Statistics of Water Quality Parameters 20
Table 3.2 Weights (Wi) of IWQI parameters 25

Table 4.1 Statistical characteristics of water quality parameters 51
Table 4.2 Standard values of water quality parameters with samples percentage and classes (Ayers et al., 1985). 52
Table 4.3 Classification of irrigation water according to IWQI 56
Table 4.4 Groundwater classification based on WQI 57
Table 4.5 Statistical characteristics of water quality indices 58

Table 5.1 The optimal hyperparameters for the ML models. 70
Table 5.2 Performance Evaluation of Machine Learning Models under both scenarios 71

Table 6.1 Performances of the regressors 81
Table 6.2 Models hyperparameters 82

Chapter 1

General introduction

1.1 Study background

Water, a vital component for ecosystems, life, and agriculture, plays a pivotal role in biochemical reactions, nutrient transport, and sustaining aquatic life. However, global freshwater scarcity, with only 2.5% available as freshwater, poses a challenge, particularly for irrigation, where water quality is paramount. Groundwater, constituting 30.8% of freshwater, becomes crucial for agriculture, supplying 43% of irrigation water used globally (Siebert et al., 2010). This study focuses on groundwater's significance in irrigation, emphasizing the need for a robust classification system utilizing advanced machine learning for nuanced groundwater quality assessment.

Urbanization, industry, and climate change threaten groundwater quality, demanding urgent attention (Ouhamdouch et al., 2019; Vrba, 1983). The global context, marked by deteriorating groundwater quality and water scarcity challenges (Sinha Ray and Elango, 2019; Konikow and Kendy, 2005), necessitates immediate action. To secure food production and economic stability, maintaining groundwater quality aligned with global standards becomes imperative (Mancosu et al., 2015; Irfeey et al., 2023). In regions like M'sila, heavily reliant on groundwater for irrigation due to arid conditions (Siebert et al., 2010), safeguarding water quality is crucial for agricultural sustenance.

In this scenario, machine learning emerges as a powerful tool for predicting and monitoring groundwater quality (Javaid et al., 2023). AI-driven models offer innovative approaches to understanding complex water quality dynamics, facilitating real-time monitoring and early contamination detection (Krishnan et al., 2022). Integrating AI into groundwater studies holds promise for sustainable water management solutions, addressing critical challenges and ensuring judicious water resource utilization. This study navigates the complexities of

groundwater quality assessment, contributing to both scientific understanding and practical agricultural management in the quest for safe and sustainable irrigation water.

1.2 Research Gap

Our research is motivated by a discernible gap in the existing water quality literature, marked by the conspicuous absence of economic considerations in prior studies. This omission underscores a critical lacuna, as economic factors play a pivotal role in shaping the feasibility and practicality of water quality management strategies.

Moreover, the scarcity of studies employing explainable artificial intelligence (AI) techniques within the realm of water resources management highlights another notable gap. In response, our research endeavors to pioneer the implementation of these transparent and interpretable methods to address crucial water-related concerns. This represents a departure from conventional approaches, as the integration of explainable AI in water quality studies has been conspicuously underexplored.

Central to our focus is a fundamental research gap: the pursuit of heightened accuracy in water quality assessment using a minimal set of input parameters, all while adhering to an economic and practical framework. This specific aspect has received limited attention in prior research endeavors, necessitating our concerted efforts to bridge this scholarly gap.

1.3 Research problem

Within the overarching discourse on the pivotal role of water, particularly groundwater, in sustaining agricultural practices, a critical juncture emerges—a juncture delineated by the rationalization of this thesis. As we traverse the landscape of water's ecological and agricultural significance, two fundamental lacunae in existing studies come into focus. The first gap pertains to the economic considerations that have been conspicuously absent in prior investigations concerning irrigation water quality. While the importance of water quality is indisputable, the economic implications of adopting specific measures or technologies for water quality assessment have been notably overlooked. This oversight raises questions about the pragmatic feasibility of implementing water quality management strategies, especially in contexts where financial constraints play a pivotal role in shaping agricultural practices. Simultaneously, a second void becomes apparent—a gap in the comprehensive assessment of a diverse array of machine learning models within the specific domain of irrigation water quality classification. Prior research may have delved into the application of machine learning, but a holistic evaluation of various models in the context of irrigation water

quality remains elusive. This deficiency underscores the need for a nuanced exploration of machine learning methodologies, discerning their efficacy, limitations, and applicability to the intricate task of classifying irrigation water quality. The transition from these lacunae in prior studies to the rationale behind this thesis seamlessly aligns with the exigencies of the agricultural landscape, particularly in regions where groundwater serves as the lifeblood for irrigation. However, as we embark on the scientific journey outlined in this thesis, it becomes imperative to acknowledge the practical challenges faced by local farmers. The laborious and expensive nature of traditional water quality assessment processes poses a significant hurdle, often rendering these approaches unaffordable for the very stakeholders who rely most on sustainable irrigation practices. Bridging this divide between scientific inquiry and on-the-ground realities forms a pivotal aspect of the thesis's rationale.

1.4 The rationale of the study

In essence, this thesis crystallizes around the imperative to bridge these critical gaps—integrating economic considerations into the discourse on irrigation water quality, comprehensively evaluating machine learning models, and addressing the pragmatic challenges faced by local farmers. As we traverse the forthcoming chapters, this scientific inquiry endeavors not only to contribute to the academic discourse but also to offer tangible, economically viable solutions that resonate with the lived experiences of those intricately connected to the agricultural landscape.

1.5 Study objectives

As we pivot from the identified gap in prior studies and the rationale behind this thesis, the focus seamlessly transitions to the objectives that underscore the scientific pursuits of this investigation.

The study is designed to address the nexus of machine learning abilities, economic feasibility, and practical challenges faced by local farmers involved in the complex field of managing irrigation water quality. The stated goals can be summarized as follows:

1. Development of Robust Machine Learning Models
    (a) Development of sophisticated machine learning models suited to the complexities of the dynamics of irrigation water quality.
    (b) Constructing models with robustness, accuracy, and condition adaptability to strengthen their usefulness in practical applications.

2. Streamlining Irrigation Water Quality Assessment

    (a) Proposing strategies to streamline the intricate process of irrigation water quality assessment.

    (b) Attempting to improve the efficacy and efficiency of the evaluation procedures in order to guarantee accurate and timely outcomes.

3. Interpretable Machine Learning Models Through SHAP

    (a) Leveraging the SHAP (SHapley Additive exPlanations) methodology to instill interpretability into machine learning models.

    (b) Enabling a clear understanding of the key variables influencing irrigation water quality.

4. Economic Optimization of Assessment Process

    (a) Innovating strategies for economically optimizing the irrigation water quality assessment process.

    (b) Striking a balance between precision and cost-effectiveness, ensuring that the assessment protocols align with economic considerations.

5. Assessment of Study Area's Suitability for Irrigation

    (a) Conducting a comprehensive evaluation of the study area's suitability for irrigation practices.

6. Identification of Paramount Parameters

    (a) Systematically identifying and prioritizing the paramount parameters influencing irrigation water quality.

These objectives coalesce into a concerted scientific endeavor, driven by the imperative to harmonize agricultural practices with water quality considerations, laying the groundwork for sustainable and informed irrigation practices in the study area.

1.6 Thesis structure

The progression from these defined objectives naturally segues into the overarching structure of the thesis. The narrative unfolds across six meticulously crafted chapters, each contributing

uniquely to the elucidation of irrigation water quality classification. The initial chapter serves as an introduction, casting light on the context, rationale, and objectives of the study. This sets the stage for an exhaustive exploration in the second chapter, delving into the existing literature to provide a robust foundation for the subsequent scientific inquiry.

Chapter three, an integral section of the thesis, elucidates the intricacies of data collection and the machine learning models strategically employed to achieve the endeavors that incentivized this study. Chapter four unfurls the implementation and results, navigating through the facets of Water Quality Assessment for Irrigation Purposes, Groundwater Quality Classification Using ML, and Groundwater Quality Prediction Using interpretable ML.

The subsequent chapter, chapter five, offers a reflective interlude, delineating the limitations of the study and charting the course for future directions in this scientific endeavor. This contemplative pause sets the stage for the concluding chapter, wherein the culmination of findings and insights converges into a cohesive narrative.

Chapter 2

Literature review

2.1 Introduction

The literature review in water resources serves as a crucial foundation, providing a nuanced understanding of past and present research endeavors. This introduction emphasizes the significance of comprehending prior scholarship, particularly in water quality assessment for irrigation. Historical milestones, such as the establishment of water quality indices and the integration of artificial intelligence (AI) and machine learning (ML), have shaped the evolution of methodologies in environmental studies. ML's application in groundwater quality classification marks a significant departure from conventional approaches, highlighting its potential for innovative predictive models. As we delve into the literature review, historical developments serve as guiding beacons, illuminating the trajectory of scientific efforts in water quality assessment.

2.2 Water quality indices for quality assessment

Landwehr and Deininger (1976) undertook a meticulous exploration, presenting five distinct water quality indices with the objective of comparison. These indices, including the arithmetic (WQIA), multiplicative (WQIM), unweighted arithmetic (WQIU), and unweighted multiplicative (WQIMU) indices proposed by Brown et al. (1970,1973) and an index formulated by Harkins (1974) based on Kendall's nonparametric multivariate ranking procedure, were subjected to scrutiny. The evaluation, involving the mean ratings (on a 0 to 100 scale) provided by 100 water experts for 20 samples from diverse U.S. rivers, demonstrated commendable performance by all five indices. Notably, the unweighted multiplicative index (WQIMU) garnered favor among experts, underscoring its efficacy (Fig. 2.1).

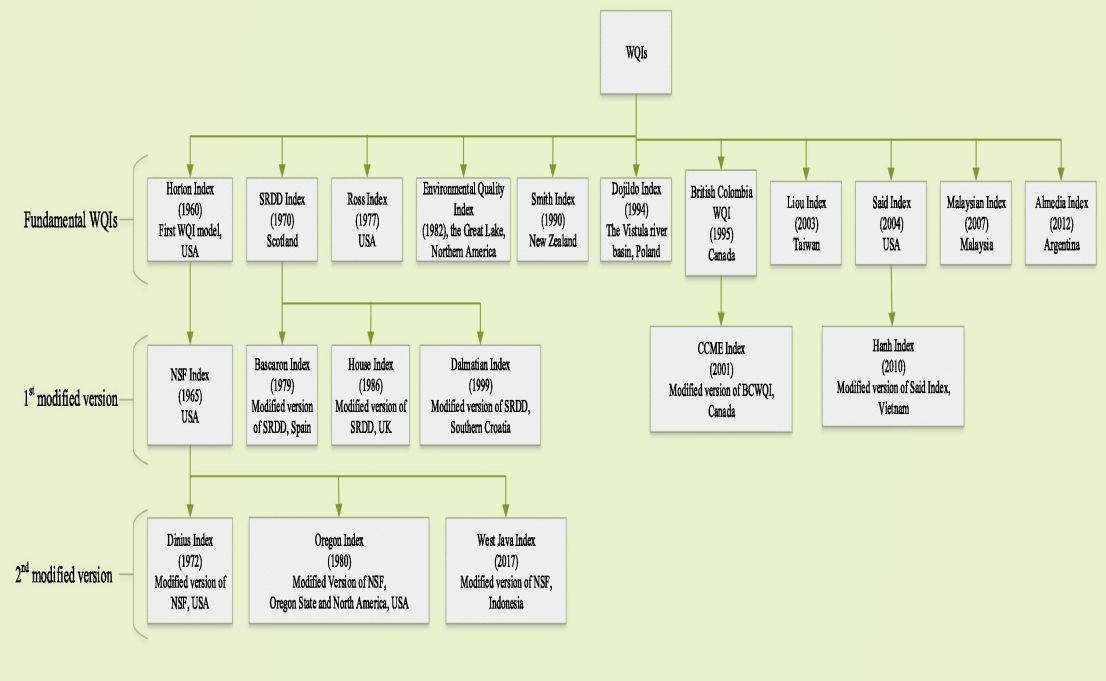

Fig. 2.1 WQI historical development (Uddin et al., 2021a)

In a divergent geographic context, Chandra et al. (2017) contributed to the discourse by assessing the water quality of Vijayawada in the Krishna district of Andhra Pradesh, India. Employing the weighted arithmetic water quality index method, their investigation encompassed a substantial dataset comprising approximately 380 samples collected during pre-monsoon and post-monsoon seasons in 2014. The discerning selection of nineteen data points from each season revealed distinctive water quality dynamics, indicating suitability for drinking purposes in pre-monsoon conditions and a notable increase in pollution post-monsoon.

Further enriching the spectrum of water quality indices, Singh et al. (2018) introduced the concept of the IWQI for the Indian context. This comprehensive index, based on 12 parameters aligned with local standards, including those set by the Central Pollution Control Board (CPCB), the Central Ground Water Board (CGWB), and FAO guidelines, categorized water quality into five classes. Employing Saaty's Analytic Hierarchy Process (AHP), a multiple criteria decision analysis (MCDA) tool, the authors sought to mitigate subjectivity in parameter weight assignment, thereby enhancing the objectivity of the assessment process.

Turning attention to Egypt's Kafr El-Sheikh Governorate, Jahin et al. (2020) engaged in an in-depth examination of surface water quality for irrigation. Employing multivariate statistical tools, principal component analysis (PCA), and factor analysis (FA), the authors developed two indices (IWQI-1 and IWQI-2) with season-based weights. This approach, aligned with

Food and Agriculture Organization (FAO) recommendations and the National Sanitation Foundation Water Quality Index (NSFWQI) formula, streamlined the dimensionality of the dataset. Seven key parameters, including pH, Na^+, HCO_3^-, Zn, As, NO_3^-, and B, emerged as robust indicators, exhibiting enhanced performance when weights were determined through PCA.

In a parallel vein, Ewaid et al. (2019) contributed a guide for irrigation water quality (IWQG) along with user-friendly software, fashioned using Visual Basic 6, to assess water appropriateness in Al-Gharraf Canal, southern Iraq. Grounded in FAO recommendations and Meireles' irrigation water quality index Meireles et al. (2010), their evaluative framework, appraised through a dataset of 612 samples, classified the water quality of the study area as moderately restricted, emphasizing the pragmatic utility of their developed tools.

In the pursuit of comprehending water suitability, Tyagi et al. (2020) delved into an exploration of widely employed water quality indices. Their study encapsulated indices such as the National Sanitation Foundation Water Quality Index (NSFWQI), Oregon Water Quality Index (OWQI), Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI), and the weighted arithmetic Water Quality Index. This comprehensive examination elucidated the mathematical structures of these indices, delineated the quality parameters integral to the evaluation process, and presented a nuanced analysis of their respective merits and demerits.

Shifting focus to the Netravati River basin in Karnataka state, India,Sudhakaran et al. (2020) conducted a meticulous investigation into the appropriateness of river and well water for both drinking and irrigation purposes. Their study, spanning pre-monsoon, monsoon, and post-monsoon seasons in 2017 across sixteen monitoring stations, employed the water quality index proposed by Brown et al. (1970) for evaluating water adequacy for drinking. Notably, the study revealed discrepancies in the water quality for downstream drinking water, surpassing permissible values set by the World Health Organization (WHO). The application of the Water Quality Index (WQI) indicated variations across seasons, attributed to diverse factors like salt deposits, sewage, industrial waste, anthropogenic activities, and seasonal fluctuations. Additionally, the study extended its analysis to well water, showcasing its resilience to seasonal fluctuations. Multivariate statistical analyses, including principal components' analysis (PCA) and Pearson correlation, were employed to identify key pollutant sources. The findings underscored the influence of anthropogenic activities on water quality, emphasizing the seasonal nuances in the study area.

In Egypt's El-Sharkia Governorate, Abdel-Fattah et al. (2020) undertook a robust assessment of the Bahr Mouise canal's water quality. Leveraging the IWQI proposed by Meireles Meireles et al. (2010) and employing multivariate analysis techniques, including principal

components analysis (PCA) and agglomerative hierarchical clustering (AHC), the authors evaluated six distinguished locations along the canal across four seasons in 2019. The IWQI classification revealed low restrictions (class II) during summer and no restrictions (class I) during other seasons. Concurrently, statistical-based classification demonstrated consistent results across seasons, except for September, where IWQI and PCA methods yielded different classifications. The study delved into the correlation between land use and land cover (LULC) changes, normalized difference vegetation index (NDVI), and water quality. Agricultural activities emerged as the dominant land use, with variations in NDVI attributed to changes in crop types and growth stages. Chakravarty and Gupta (2021) conducted an exhaustive assessment of the water quality status of the river Jatinga in south Assam, northeast India. Employing the water quality index rating system alongside multivariate statistical analysis, the study encompassed five sites along the river, collecting samples throughout all seasons of 2018-2019. Thirteen physiochemical parameters were scrutinized, most of which adhered to permissible limits set by local Indian standards and World Health Organization recommendations. Utilizing the weighted arithmetic index method proposed by Brown et al. (1970), the authors concluded that the average seasonal status of the river was excellent. Factor analysis, facilitated by Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity, unveiled three significant pollution sources representing 91.13% of the total variance, namely anthropogenic and organic sources (37.32%), natural sources (30.95%), and agricultural runoff (22.86%).

Rahman et al. (2021) delved into the water quality of Lake Gulshan in Bangladesh, employing the water quality index (WQI) and multivariate statistical analysis to discern prime pollution sources. Their study, spanning dry and wet seasons in 2018 and 2019, adhered to the USEPA (2013) standard for surface water sampling. The National Sanitation Foundation water quality index (NSF-WQI) and the index proposed by the Canadian Council of Ministers of the Environment (CCME-WQI) were applied, revealing desirable water quality ratings for both seasons. Multivariate statistical analyses, namely principal component analysis (PCA) and positive matrix factorization (PMF), pinpointed four factors. PCA elucidated point sources of pollution (42.32% and 38.49% in dry and wet seasons, respectively), while PMF revealed nutrient-rich point sources and industrial waste as significant contributors. The study emphasized water quality and diverse pollution sources, offering valuable insights for regulatory measures.

In a broader perspective, Uddin et al. (2021b) provided a comprehensive review of various water quality indices, offering a comparative analysis based on model structures, components, and applications. Their work, grounded in a review of 110 studies, identified 21 commonly utilized WQI models, primarily derived from seven essential models. The authors highlighted four fundamental processes intrinsic to WQI models: the selection of

water quality parameters, determination of parameter sub-indices, attribution of parameter weightings, and the aggregate function leading to the overall WQI. The conclusions drawn emphasized the limited regional applicability, dependency on variable conditions, variability in structural design, hindrances in result comparisons, and inherent challenges of accuracy due to eclipsing and uncertainty. This comprehensive review provides valuable insights for researchers and practitioners grappling with the intricacies of water quality assessment.

2.3 Machine learning elevating for water quality classification and prediction

Dezfooli et al. (2018) addressed the challenges associated with the time and cost involved in water sampling and laboratory analysis by proposing an innovative approach for water quality classification in Iran's Karoon River. Operating across eight stations and collecting 172 water samples, the study introduced machine learning models as a potential substitute for the National Sanitation Foundation Water Quality Index (NSFWQI). Probabilistic neural network (PNN), k-nearest neighbor (KNN), and support vector machine (SVM) were evaluated, utilizing performance metrics such as error rate (ER), error value (EV), and accuracy (Acc). The PNN model emerged as the most accurate, achieving 94.57% and 90.70% accuracy at the training and testing stages, respectively. Notably, the study underscored the significant role of fecal coliform in the water quality classification process, positioning PNN as an efficient alternative to NSFWQI.

Ewaid et al. (2018) proposed a model leveraging multiple linear regression (MLR) analysis to predict the water quality of the Tigris River within Baghdad. Using water quality index values as dependent variables and twenty-three monitored water quality parameters as independent variables, the model showcased a robust ability to forecast water quality variations across seasons. Monthly data from ten monitoring stations were utilized, resulting in a mean water quality index of 266 and a high model accuracy ($r=0.987$, $r^2=0.974$, $p<0.01$). The MLR model's superior performance was particularly evident when incorporating water quality index values as inputs, providing a more comprehensive understanding of water quality dynamics.

Ahmed et al. (2019) explored the effectiveness of machine learning models for water quality prediction in the Johor River Basin, Malaysia. Employing Adaptive Neuro-Fuzzy Inference System (ANFIS), Multi-Layer Perceptron Neural Networks (MLP-ANN), and Radial Basis Function Neural Networks (RBF-ANN), the study recommended an enhanced ANFIS model integrating a wavelet de-noising technique (WDT-ANFIS). This model surpassed oth-

ers, exhibiting high prediction accuracy (R^2 values greater than or equal to 0.9) for all assessed water quality parameters. The study employed various assessment approaches, including the segmentation of neural network connection weights to investigate individual inputs' preponderance and scenarios considering spatial variations. The proposed model demonstrated its applicability in predicting water quality parameters at each station and showed enhanced performance when incorporating input values predicted at upstream stations, emphasizing the importance of considering spatial dynamics in water quality modeling.

Bilali and Taleb (2020) conducted a comprehensive investigation into the efficiency of eight Machine Learning (ML) models, namely Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Random Forest (RF), Decision Tree, Support Vector Regression (SVR), Stochastic Gradient Descent (SGD), k-Nearest Neighbour (kNN), and Adaptive Boosting (AdaBoost). Their study aimed to predict eight water quality parameters in the Bouregreg watershed in Morocco, including Sodium Absorption Ratio and adjusted SAR (SAR, SARa), Percentage of Exchangeable Sodium (ESP), Residual Sodium Carbonate (RSC), percentage of Sodium (%Na), Kelly Ratio (KR), Chloride (Cl^-), Magnesium Absorption Ratio (MAR), Permeability Index (PI), and total dissolved solid (TDS). To streamline the water quality assessment process, the authors utilized only two physical parameters, EC and pH, as inputs. The models exhibited high accuracy in predicting most parameters, with correlation coefficients ranging from 0.56 to 0.99 during the training and validation phases. Notably, some models faced challenges in predicting MAR and PI parameters. Generalization attempts to Cherrate and Nfifikh watersheds highlighted the models' effectiveness for specific parameters in different geographical contexts.

In the work titled "A Machine Learning Approach towards Automatic Water Quality Monitoring" by Bansal and Geetha (2020), the authors explored the effectiveness of machine learning algorithms for water quality assessment and classification. Adopting the decision tree algorithm and following the guidelines of the World Health Organization (WHO) as the standard for water quality parameters, the study demonstrated the superiority of decision trees over traditional assessment methods. With an accuracy of 98.3%, decision trees outperformed standard water quality index formulae, which achieved an accuracy of 80.02

Bedi et al. (2020) delved into the use of three machine learning models—Artificial Neural Network (ANN), Support Vector Machine (SVM), and XGBoost (XGB)—to predict groundwater contamination levels from pesticides and nitrate, considering sparse data and non-linear relationships. The dataset comprised 303 wells across 12 Midwestern states in the USA, incorporating multiple hydrogeologic, water quality, and land use features as independent variables. The study assessed classification performance under various scenarios, comparing these machine learning models with regression models. Additionally,

class imbalance mitigation techniques were tested, and game-theoretic Shapley values were employed for model interpretability through feature importance analysis.

Lu and Ma (2020) conducted a study focusing on Gales Creek water quality, employing hybrid decision tree-based machine learning models and establishing their efficacy in water quality assessment. The hybrid models showcased notable effectiveness, contributing to the broader understanding of water quality dynamics.

Singha et al. (2021) employed a deep learning (DL) model to predict groundwater quality and conducted a comparative analysis with three additional machine learning (ML) models: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). The study involved the collection of 226 groundwater samples from an agriculturally dense region in Chhattisgarh, India. Various physicochemical parameters were utilized to compute the entropy weight-based groundwater quality index (EWQI). The DL model exhibited superior performance compared to the other ML models, establishing itself as the most realistic and accurate method for predicting groundwater quality in the examined area.

Agrawal et al. (2021) conducted a study assessing the performance of artificial intelligence techniques, including Particle Swarm Optimization (PSO), Naive Bayes Classifier (NBC), and Support Vector Machine (SVM), in predicting the Water Quality Index (WQI). The authors applied PSO for optimization and utilized SVM and NBC for prediction, utilizing groundwater quality data from Chhattisgarh, India. Among the ensemble machine learning algorithms, PSO-NBC outperformed PSO-SVM, demonstrating high prediction accuracies in the evaluation of water quality.

Bilali et al. (2021) explored the application of artificial intelligence models for predicting irrigation water quality indexes in aquifer systems, utilizing physical parameters as features. The study evaluated four models—Adaptive Boosting (AdaBoost), Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Regression (SVR)—using data from the Berrechid aquifer in Morocco. Adaboost and RF models exhibited superior prediction performances, while ANN and SVR models demonstrated enhanced generalization ability and sensitivity to inputs.

Ravindran et al. (2021) delved into the utilization of deep neural networks (DNN) for forecasting daily reference evapotranspiration (ETo) with a single input parameter. The study emphasized the significance of feature relevance scores derived from machine learning techniques such as random forest (RF) and extreme gradient boosting (XGBoost). The investigation explored the feasibility of utilizing SHapley Additive exPlanations (SHAP) to elucidate and validate feature selection approaches. Solar radiation emerged as a prominent feature in three California Irrigation Management System (CIMIS) weather station datasets, leading to the construction of three ETo models (DNN-Ret, XGB-Ret, and RF-Ret) with solar

radiation as the primary input. DNN-Ret demonstrated superior performance, establishing its efficacy in single input parameter-based ETo modeling across diverse climatic zones.

Raheja et al. (2022) conducted an investigation into the performance of three machine learning algorithms—Deep Neural Network (DNN), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost)—for evaluating groundwater indices in Haryana state, India. The study focused on two water quality indices, namely Entropy Water Quality Index (EWQI) and Water Quality Index (WQI). Results indicated that DNN outperformed the other models, exhibiting lower error values and better predictive capabilities for both EWQI and WQI. The analysis identified Electrical Conductivity (EC) as the most significant input parameter for predictions, with 'pH' holding the least significance.

Shrivastava et al. (2022) conducted a comparative assessment of Extra Trees and Random Forest ensemble learning techniques for groundwater quality assessment in Chhattisgarh. The study revealed the effectiveness of both techniques in groundwater classification, emphasizing their utility in evaluating and categorizing groundwater quality.

Nasir et al. (2022) developed seven individual classifiers to predict the Water Quality Index (WQI), with the CATBOOST approach yielding the most favorable predictive results. This highlights the efficacy of CATBOOST in achieving comprehensive water quality assessment.

Abuzir and Abuzir (2022) employed J48, Naïve Bayes, and Multi-Layer Perceptron (MLP) algorithms for predicting Water Quality Classes. Despite working with a 10-feature dataset, MLP demonstrated the highest accuracy among the algorithms, showcasing its effectiveness in water quality class prediction. Xia et al. (2022) scrutinized the application of Long Short-Term Memory (LSTM) and XGBoost for predicting dichloroethene (DCE) concentrations in a pesticide-contaminated groundwater site undergoing natural attenuation. XGBoost exhibited greater effectiveness in capturing DCE variations and performed well, particularly with high concentration values, while LSTM demonstrated superior overall accuracy. SHAP values provided explanations consistent with biodegradation rules in real environmental conditions. Both LSTM and XGBoost successfully predicted DCE concentrations using water quality variables, with LSTM displaying better overall performance compared to XGBoost.

Gupta and Mishra (2023) introduced an entropy-based river water quality index using machine learning models, with logistic regression identified as the top-performing model in their study.

2.4 Conclusion

In conclusion, the comprehensive exploration of existing literature has illuminated key aspects of water quality assessment and the integration of machine learning methodologies. The synthesis of knowledge from various studies has provided valuable insights into the complexities and challenges associated with groundwater quality evaluation for irrigation purposes. The literature has underscored the significance of considering multiple parameters, including conductivity, chloride (Cl^-), bicarbonate ($HCO_3{}^-$), sodium (Na^+), calcium (Ca^{2+}), and magnesium (Mg^{2+}), in the context of IWQI calculations. Moreover, it has revealed the diverse array of machine learning algorithms such as LGBM, CatBoost, Extra Trees, Random Forest, Gradient Boosting classifiers, Support Vector Machines, Multi-Layer Perceptrons, and the K-Nearest Neighbors Algorithm that have been employed for water quality classification in various studies. As we transition into the methods chapter, this literature review sets the stage for the application of machine learning techniques in the assessment of groundwater quality for irrigation. The insights gained from existing research will inform our approach in developing a streamlined and economically viable model. The next chapter will delve into the details of the methodology, encompassing data preparation, data preprocessing, feature selection, model development, and evaluation. It is within this framework that we aim to contribute to the evolving landscape of efficient and effective water quality assessment methodologies, driven by advancements in machine learning techniques.

Chapter 3

Methodology

3.1 Introduction

The methodology chapter begins with a firm recognition of the critical research gaps identified in existing literature, notably the absence of economic considerations and the pausity of leveraging artificial intelligence (AI) techniques in water quality studies. Our research aims to develop a methodological paradigm that combines accuracy, cost-effectiveness, and a thorough analytical framework. The quantitative approach is chosen for its empirical nature, allowing for objective measurement and analysis of numerical data to explore complex water quality dynamics and economic factors. The research design follows a sequential framework, beginning with theoretical underpinnings and progressing to variable identification, data analysis, statistical methods, machine learning, and result interpretation. This deliberate alignment with study objectives ensures the methodology's capability to extract knowledge and address identified research gaps effectively.

3.2 Study area and Data description

The research unfolds in the M'sila region of north-central Algeria, extending across 17,927 km², delineated spatially between longitudes 3° 22' 13" E and 5° 9' 9" E, and latitudes 34° 12' 58" N and 36° 2' 9" N, as illustrated in Figure 3.1. The region exhibits a semi-arid climate, characterized by scorching, arid summers and frigid winters, with an annual precipitation ranging from 200 mm to 480 mm, demonstrating a partial desert influence. Land cover distribution, depicted in Figure 6.1, delineates 70,012% as rangelands, 23.83% as bare ground, 4.71% as crop areas, and 1.291% as construction, with the remainder comprising trees and various water bodies.

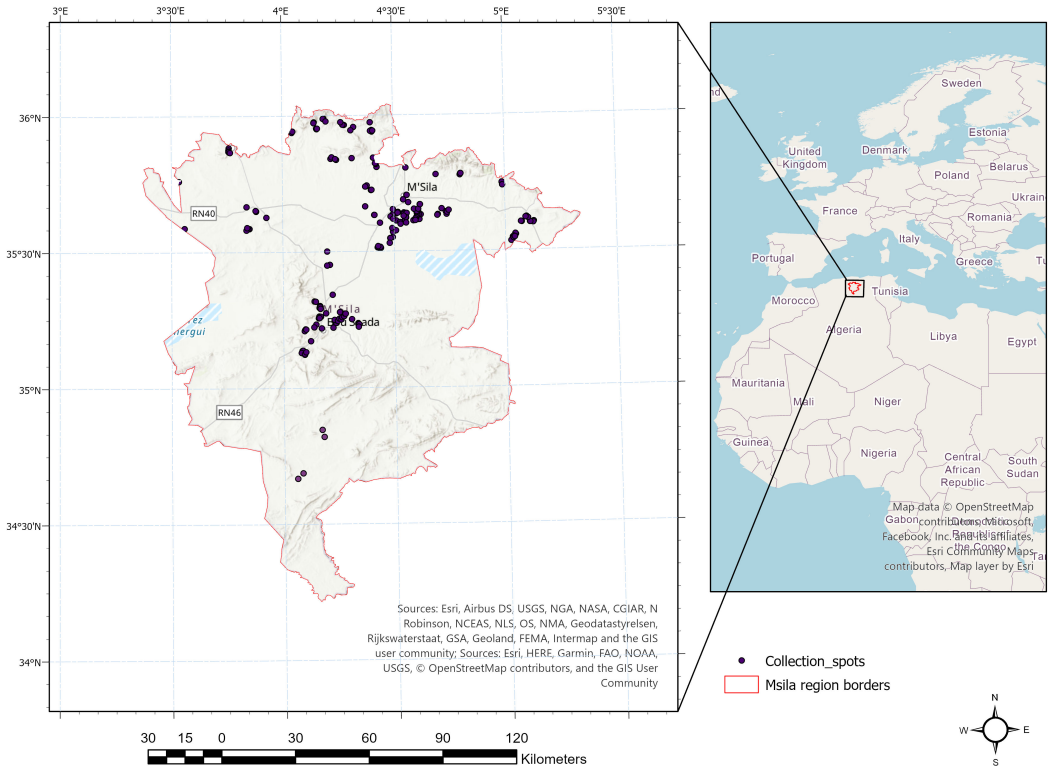

Fig. 3.1 The geographical location of the study area

In pursuit of a holistic understanding of groundwater quality, the Algerian Water Company's water analysis laboratory meticulously monitored water quality over four years, from 2018 to 2022. This extensive evaluation encompassed 210 wells strategically positioned across the study area, with a comprehensive analysis of 19 physicochemical parameters conducted in adherence to ISO 5663 standards. Rigorous quality control standards, following ISO 5663 guidelines for groundwater sampling and transportation, were applied during data collection to ensure the precision and accuracy of results.

The laboratory employed cutting-edge equipment and methodologies for analysis. Spectrophotometry determined elemental composition, while instruments such as ADWA AD1020 pH meter, Hach HQ14D EC meter, and HACH-TL2300 turbidity meter gauged physicochemical properties. Further, titration methods were used for Total Hardness (TH) and chloride, bicarbonate, magnesium, and calcium ion concentrations. Sodium and potassium measurements utilized a flame photometer (Jenway PFP7), and sulfate ions were quantified using a spectrophotometer (HASH/Dr/4000). Adhering to strict quality assurance and control

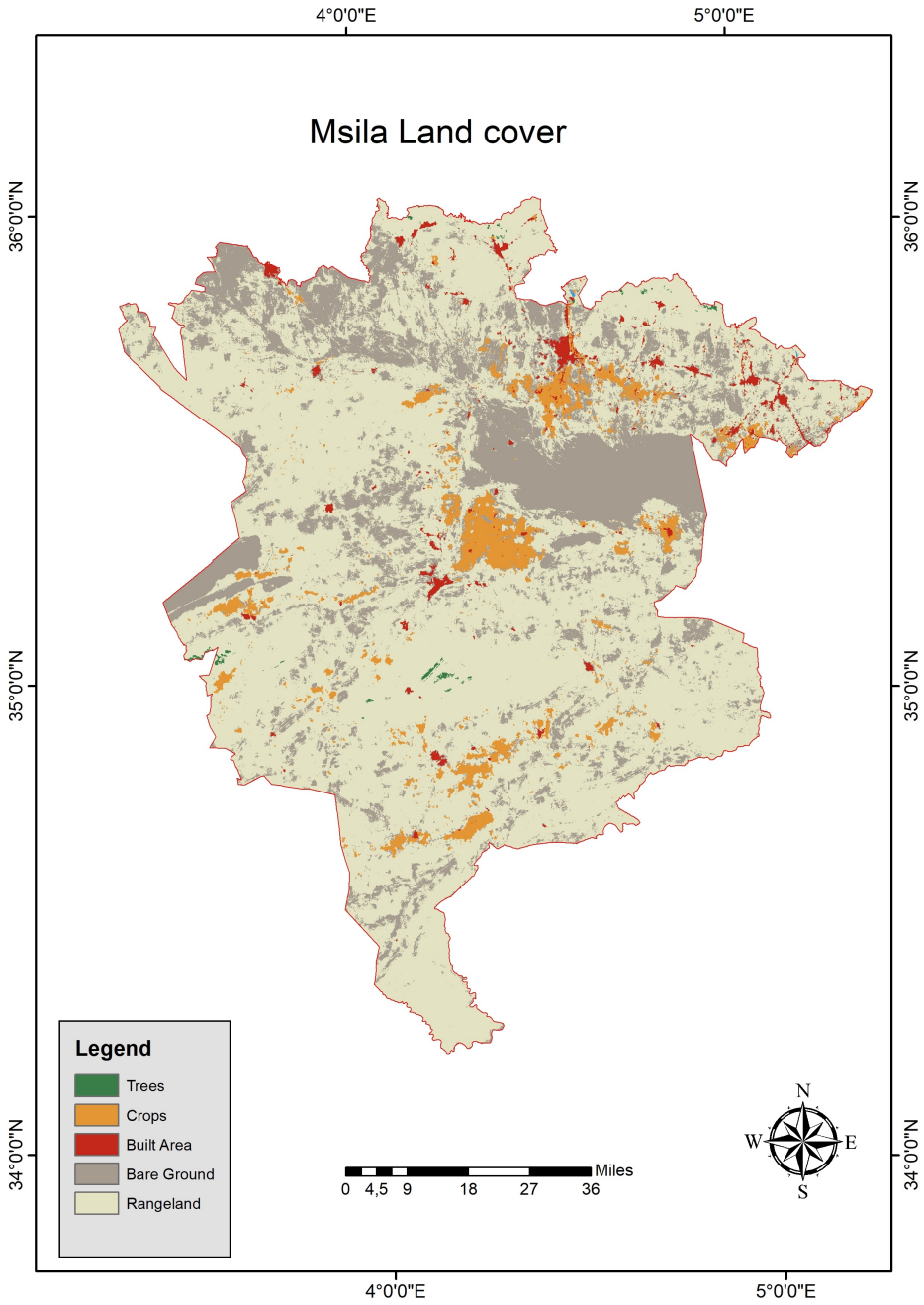

Fig. 3.2 Land cover distribution of the study area

protocols, procedural blank measurements, sample spiking, and duplicate observations were employed to detect and rectify potential errors.

Table 3.1 provides a statistical overview of the integrated water quality parameters. The assessment employed various indices, including the Irrigation Water Quality Index Meireles et al. (2010), Sodium Adsorption Ratio (SAR), Magnesium Adsorption Ratio (MAR), Soluble

Sodium Percent (SSP), Permeability Index (PI), and Kelly's Ratio (KR), to comprehensively evaluate water quality for irrigation purposes.

This rich dataset, derived from meticulous monitoring and analysis, serves as a linchpin for unraveling the intricacies of groundwater conditions in the M'sila region. It assumes pivotal importance for future land-use planning and effective resource management strategies.

Table 3.1 Descriptive Statistics of Water Quality Parameters

	mean	std	min	max
pH	7.36	0.37	6.08	9.30
Na^+ (mg/l)	116.74	72.51	3	450
K^+ (mg/l)	5.24	2.73	0.80	18
Ca^{2+} (mg/l)	177.32	86.09	0	544
Mg^{2+} (mg/l)	90.52	44.82	0	238.14
Turbidity (NTU)	6.19	15.54	0.01	147
TDS (mg/l)	829.77	344.72	193	1895
T (°C)	21.14	4.97	8.30	34.40
TAC (°F)	107.14	98.80	3.20	400
Conductivity (μs/cm)	1979.97	989.72	426	8970
Total Hardness (°F)	349.87	376.83	0	1800
Cl^- (mg/l)	186.50	148.20	0	942.88
$HCO_3{}^-$ (mg/l)	337.78	109.85	0	976
$NH_4{}^+$ (mg/l)	0.06	0.33	0	5
$SO_4{}^{-2}$ (mg/l)	629.90	255.71	100	1600
$NO_3{}^-$ (mg/l)	33.32	111.50	0	2379
$NO_2{}^-$ (mg/l)	0.04	0.23	0	5
Fe^{+2} (mg/l)	0.03	0.09	0	0.72
$PO_4{}^{-3}$ (mg/l)	0.03	0.05	0	0.10

3.3 Water quality parameters

Groundwater pH

pH is a crucial parameter indicating the acidity or alkalinity of water, ranging from 0 to 14, with 7 as neutral. Groundwater pH influences the solubility and mobility of nutrients and metals, affecting both water quality and its suitability for irrigation and consumption. Low pH can increase the dissolution of metals such as aluminum, iron, and manganese, potentially leading to toxicity and contamination. Conversely, high pH can cause the precipitation of

essential nutrients like phosphorus, iron, and zinc, reducing their availability and leading to deficiencies. Additionally, groundwater pH affects microbial activity and geochemical processes, influencing carbonate equilibria, mineral dissolution, and overall water chemistry.

Conductivity (EC)

Conductivity in groundwater refers to its ability to conduct electric current, influenced by dissolved solids like salts, minerals, and metals. It is measured in microsiemens per centimeter (µS/cm) or millisiemens per centimeter (mS/cm). Conductivity indicates water salinity and dissolved solids content, with high conductivity suggesting high salinity or dissolved solids, potentially impacting water taste and quality. Conversely, low conductivity may signify purity and freshness. Soil conductivity mirrors salt accumulation and leaching, affecting fertility and structure. High soil conductivity can reduce fertility, while low conductivity may impact moisture and nutrient retention. It also affects osmotic pressure and water balance in plants, with high conductivity reducing water availability and low conductivity increasing it.

Phosphate (PO_4^{3-})

Phosphate is a vital nutrient present in groundwater that can lead to eutrophication, characterized by excessive algal growth. This phenomenon diminishes dissolved oxygen levels, reduces light penetration, and decreases biodiversity in water bodies. Moreover, it heightens the risk of harmful algal blooms (HABs) and cyanotoxins. Phosphorus also influences the soil phosphorus cycle, impacting its adsorption, desorption, and availability for plants. Essential for plant energy metabolism and nucleic acid synthesis, phosphorus deficiency can significantly impact plant growth, yield, and quality.

Hardness (TH)

Hardness in groundwater refers to the concentration of calcium and magnesium ions, impacting its suitability for irrigation. High levels of hardness can lead to scaling, corrosion, and reduced efficiency of irrigation systems. It also affects soil texture, aggregation, and cation exchange capacity, influencing soil structure, fertility, and water retention crucial for irrigation. Furthermore, hardness influences plant cell wall stability, enzyme activity, and nutrient availability, essential for plant growth, photosynthesis, and stress response.

Alkalinity (TAC)

Alkalinity in groundwater refers to its ability to neutralize acids. It impacts the pH, buffering capacity, and stability of water, as well as the solubility and toxicity of metals and nutrients. High alkalinity can elevate the pH and decrease the solubility of metals and phosphorus, affecting aquatic life and plant growth. Moreover, it influences soil pH, buffering capacity, and nutrient availability, impacting soil microbial activity, nutrient cycling, and plant uptake. Elevated alkalinity may raise soil pH and decrease the availability of essential nutrients like iron, zinc, and manganese, leading to plant deficiency and chlorosis.

Sodium (Na^+)

Sodium, a prevalent cation in groundwater, can impact its taste and salinity, influencing its suitability for irrigation purposes. Elevated sodium levels can increase water salinity and osmotic pressure, affecting water balance and potentially causing toxicity in plants and animals. Moreover, sodium can alter soil dispersion, permeability, and sodicity, thereby impacting soil physical, chemical, and biological characteristics. High sodium content may lead to soil particle dispersion, reducing soil porosity, infiltration, and aeration. Additionally, sodium affects osmotic potential, water uptake, and sodium toxicity in plants, potentially reducing water availability and causing sodium accumulation and tissue injury in plants.

Nitrate (NO_3^-) and Nitrite (NO_2^-)

Nitrate and nitrite, forms of nitrogen, are indicators of pollution from fertilizers, sewage, or animal waste in groundwater used for irrigation. Their presence can impact water quality and safety for both drinking and agricultural purposes. Elevated levels of nitrate and nitrite pose health risks such as methemoglobinemia in infants and cancer risk in adults. Moreover, they contribute to eutrophication, hypoxia, and harmful algal blooms in water bodies. In agricultural contexts, they affect the soil nitrogen cycle, leading to increased nitrification and leaching, which in turn reduces soil nitrogen retention and enhances groundwater contamination. Furthermore, they influence plant nitrogen metabolism and protein synthesis, potentially causing nitrate accumulation and toxicity in plant tissues.

Potassium (K^+)

Potassium is a vital nutrient for plant growth but can contribute to salinity issues in groundwater used for irrigation. It may come from rock weathering, fertilizer leaching, organic matter, or seawater intrusion. Elevated potassium levels can raise groundwater salinity and

osmotic pressure, impacting plant and animal water balance and toxicity. Additionally, it affects soil cation balance, fertility, and structure. Removal methods include reverse osmosis, ion exchange, or dilution techniques.

Magnesium (Mg^{2+})

Magnesium is a vital mineral for human and animal health, yet it can lead to hardness and scaling issues in groundwater used for irrigation. It occurs naturally through the dissolution of minerals like dolomite and magnesite or can be introduced by human activities such as mining and industrial processes. Elevated levels of magnesium contribute to increased hardness and scaling in groundwater, impacting the corrosion, clogging, and efficiency of pipes and appliances. Moreover, it influences the soil magnesium cycle, affecting plant growth, yield, and quality. Techniques like softening, ion exchange, or chemical precipitation can be employed to mitigate magnesium levels in groundwater.

Calcium (Ca^{2+})

Calcium, a vital mineral for human and animal health, can pose challenges in groundwater used for irrigation due to its potential to cause hardness and scaling. It naturally occurs in groundwater through the dissolution of minerals like calcite and gypsum, or it can be introduced by human activities such as mining and industrial processes. Elevated calcium levels can exacerbate groundwater hardness and scaling issues, impacting the efficiency and durability of pipes, boilers, and appliances. Additionally, calcium influences the soil calcium cycle and availability, affecting plant growth, yield, and quality. Methods like softening, ion exchange, or chemical precipitation can be employed to mitigate calcium-related issues in groundwater used for irrigation.

Bicarbonates (HCO_3^-)

These anions play a crucial role in groundwater equilibrium and buffering. They stem from various sources including the dissolution of carbon dioxide, carbonates, and bicarbonates, as well as biological processes like photosynthesis and decomposition. Bicarbonates impact groundwater pH, alkalinity, and salinity, along with the solubility and availability of certain metals and nutrients. Moreover, they affect soil carbonate equilibrium, pH, and salinity, influencing soil microbial activity, nutrient cycling, and plant uptake. Bicarbonates can be mitigated through acidification, aeration, or reverse osmosis techniques.

Sulfate (SO_4^{2-})

Sulfate, an anion involved in groundwater oxidation processes, can originate from various sources such as sulfates, sulfides, and biological processes like sulfate reduction and denitrification. Its presence affects groundwater quality for drinking, irrigation, and industrial uses, causing a bitter taste, pipe corrosion, and impacting soil sulfur availability and plant metabolism. Sulfate removal methods include biological denitrification, ion exchange, reverse osmosis, or dilution.

Chloride (Cl^-)

Chloride, an anion found in groundwater, impacts its salinity, conductivity, and taste. It originates from salt dissolution, including sodium chloride or calcium chloride, and human activities like road salt application, industrial processes, and wastewater discharge. Elevated chloride levels increase groundwater salinity and osmotic pressure, affecting water balance and toxicity in plants and animals. It also influences soil chloride cycle, leaching, and accumulation, impacting soil fertility and structure. Chloride removal methods include reverse osmosis, ion exchange, or distillation.

3.4 Water quality indices

3.4.1 The irrigation water quality index (IWQI)

The evaluation of water quality suitability for irrigation purposes in this study hinges on the adoption of the Irrigation Water Quality Index (IWQI). This index, as per Meireles et al. (2010), is a mathematical representation that consolidates multiple water quality parameters into a singular value, providing a comprehensive assessment. Key parameters influencing water quality, identified through factorial analysis and principal component analysis (PCA), include SAR, Electrical Conductivity (EC), Bicarbonates (HCO_3^-), Chloride (Cl^-), and Sodium (Na^+).

The IWQI calculation involves multiplying the quality measure parameter, qi, by the corresponding assigned weights, Wi, for each parameter. The determination of qi values adheres to Equation (3.1), incorporating limit values proposed by Ayers and Westcot (1985):

$$q_i = q_{i.max} - \left(\frac{(x_{ij} - x_{inf}) \times q_{i.amp}}{x_{amp}} \right) \tag{3.1}$$

Here, $q_{i.max}$ signifies the upper value for the relevant qi class, x_{ij} denotes the measured value of the corresponding parameter, x_{inf} represents the lower limit of the parameter's class, $q_{i.amp}$ is the amplitude of the class, and x_{amp} signifies the amplitude of the parameter's class. To derive the overall IWQI, Equation (3.2) is applied:

$$IWQI = \sum_{1}^{5} (q_i \times W_i) \tag{3.2}$$

The assigned weights, W_i, are outlined in Table 4.1.

Table 3.2 Weights (Wi) of IWQI parameters

Parameters	Wi
EC	0.211
Na^+	0.204
HCO_3^-	0.202
Cl^-	0.194
SAR	0.189

These weights (W_i) play a crucial role in the summation process, providing a nuanced and weighted evaluation of the water quality parameters to ascertain the overall IWQI.

3.4.2 Sodium Adsorption Ratio (SAR)

The Sodium Adsorption Ratio (SAR) serves as a pivotal index in delineating the impact of sodium ions on the soil, providing insights into potential sodium hazards Wilcox (1955). SAR, as computed through Equation (3.3), encapsulates the ratio of sodium ions (Na^+) to the square root of the average of calcium (Ca^{2+}) and magnesium (Mg^{2+}) concentrations, where concentration values are expressed in meq/l.

$$SAR = \frac{Na^+}{\sqrt{\frac{Ca^{2+}+Mg^{2+}}{2}}} \tag{3.3}$$

3.4.3 Soluble Sodium Percent (Na%)

The quantification of Soluble Sodium Percent (Na%) plays a pivotal role in the assessment of irrigation water quality, providing insights into soil permeability. Na%, computed through Equation (3.4), represents the percentage of sodium (Na^+) and potassium (K^+) relative to the total cation concentrations, including calcium (Ca^2+) and magnesium (Mg^2+), with concentrations expressed in meq/l.

$$Na\% = \frac{(Na^+ + K^+) \times 100}{Ca^{2+} + Mg^{2+} + Na^+ + K^+} \tag{3.4}$$

3.4.4 Potential Salinity (PS)

Potential Salinity (PS) provides a valuable metric in water quality assessment, calculated through Equation (3.5). This equation involves the summation of chloride (Cl^-) concentrations with half of sulfate (SO_4^{2-}) concentrations, with concentrations expressed in meq/l.

$$PS = Cl^- + \frac{SO_4^{2-}}{2} \tag{3.5}$$

3.4.5 Permeability Index (PI)

The Permeability Index (PI), pioneered by Doneen (1964), stands as a crucial metric in determining the suitability of irrigation water by assessing its impact on soil permeability. Computed through Equation (3.6), the PI involves the multiplication of the sum of sodium (Na^+) and the square root of bicarbonate (HCO_3^-) by 100, divided by the total cation concentrations, including calcium (Ca^2+), magnesium (Mg^{2+}), and sodium (Na^+), with concentrations expressed in meq/l.

$$PI = \frac{(Na^+ + \sqrt{HCO_3^-}) \times 100}{Ca^{2+} + Mg^{2+} + Na^+} \tag{3.6}$$

3.4.6 Magnesium Adsorption Ratio (MAR)

The Magnesium Adsorption Ratio (MAR), an influential indicator, is calculated considering the concentrations of magnesium (Mg^2+) and calcium (Ca^2+), as depicted in Equation (3.7). The MAR signifies the percentage of magnesium relative to the total of magnesium and calcium concentrations, with concentrations expressed in meq/l.

$$MAR = \left(\frac{Mg^{2+}}{Mg^{2+} + Ca^{2+}} \right) \times 100 \tag{3.7}$$

3.4.7 Kelly's Ratio (KR)

Kelly's Ratio (KR), a fundamental parameter in water quality assessment, is computed using Equation (3.8). The ratio involves dividing sodium (Na^+) concentrations by the sum of calcium (Ca^{2+}) and magnesium (Mg^{2+}) concentrations, with concentrations expressed in meq/l.

$$KR = \frac{Na^+}{Ca^{2+} + Mg^{2+}} \tag{3.8}$$

These indices collectively offer nuanced insights into diverse aspects of water quality, facilitating a comprehensive evaluation tailored for irrigation purposes.

3.5 Hydrochemical Facies Characterization using Piper Diagram

In the pursuit of comprehensively characterizing the hydrochemical facies of groundwater, this study employs the Piper diagram as a powerful tool. Originating from the seminal work of Piper (1944), the Piper diagram serves as a graphical representation facilitating the identification of hydrochemical facies and the elucidation of the predominant cations and ions within water samples.

The hydrochemical facies are discerned based on the abundance of key constituents, specifically calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^+), bicarbonates (HCO_3^-), carbonates (Cl^-), and sulfate (SO_4^{2-}) within the water samples Piper (1944). The conceptual framework involves plotting two triangles: one delineates calcium (Ca) and magnesium (Mg) as "alkaline earths," with sodium (Na) represented as "alkali"; the second triangle represents sulfate (SO_4^{2-}) and chloride (Cl^-) as "strong acids," and bicarbonates ($HCO3^-$) as "weak acid." Additionally, a diamond consolidates the outcomes from both triangles.

To ascertain the hydrochemical facies, a systematic approach is applied. For each sample, a perpendicular line is drawn from its point in each triangle towards the diamond. The intersection of these lines determines the position of ions in the diamond, unveiling the hydrochemical facies in alignment with the spatial distribution of the samples on the diagram. This method provides a nuanced understanding of the hydrochemical composition, offering valuable insights into the origins and characteristics of the groundwater under investigation.

3.6 Data Preprocessing

Data preparation is a critical step in our study that is intended to improve the caliber and performance of our machine learning models. This complex procedure begins with thorough data cleaning, which includes finding and fixing any anomalies or mistakes that could jeopardize the dataset's integrity. The numerical features are then standardized using normalization procedures, which guarantee consistent scales and lessen the impact of magnitude differences between variables. A careful imputation method is used to deal with missing

values, completing any gaps in the dataset while maintaining the statistical characteristics of the original distribution. Simultaneously, feature selection techniques are used to find and keep the most useful variables, eliminating unnecessary or redundant characteristics that could cause noise in the models. The main aim of our classification challenge is to classify each water sample according to the intervals that Meireles et al. (2010) recommends. The process involves calculating the IWQI values for every sample and then classifying them into five groups ranging from 0 to 5. The IWQI intervals of 0–40, 40–55, 55–70, 70–85, and 85–100 are defined by these classes, which enable a more detailed depiction of the water quality levels. Recognizing the potential ramifications of imbalanced data on the efficacy of machine learning algorithms, we strategically employ the Synthetic Minority Over-sampling Method (SMOTE). This technique addresses the disproportionality in class distribution by oversampling minor classes, thereby fostering a more equitable representation in the dataset. Such preprocessing endeavors collectively fortify the robustness and reliability of our subsequent machine learning models, laying a solid foundation for the forthcoming analytical phases (García et al., 2014).

The dataset collected for our study underwent a comprehensive array of preprocessing steps, meticulously orchestrated to ensure the requisite quality and appropriateness for subsequent modeling and analysis endeavors. The following methodologies were systematically applied:

3.6.1 Feature Relabeling

Erroneously written or mislabeled features were systematically identified and rectified to uphold consistency and precision in the dataset. This procedure involved a meticulous verification of feature names, with necessary relabeling implemented to guarantee accurate representation.

3.6.2 Missing Value Imputation

Addressing the issue of missing data in our dataset necessitated the application of sophisticated imputation methodologies. To this end, the K-nearest neighbors (KNN) imputer, tailored with a designated k value of 5, emerged as the method of choice. This advanced imputation technique draws upon the collective knowledge embedded in the five closest neighbors of each missing data point. By calculating the average based on the available local data patterns, the KNN imputer effectively imputes missing values, ensuring a data-driven and contextually sensitive approach to rectifying gaps in our dataset.

3.6.3 Outliers Detection and Handling

Outliers within the dataset were discerned by establishing lower and upper bounds using statistical quantiles, specifically the first quartile (Q1) and third quartile (Q3). Data points falling below (Q1 - 1.5 * Interquartile Range (IQR)) or above (Q3 + 1.5 * IQR) were classified as outliers. Addressing these outliers involved their removal from the dataset or the application of pertinent data transformation techniques to mitigate their potential influence on subsequent analyses.

3.6.4 Data Partitioning

A pivotal phase in our methodology involved the judicious splitting of the dataset into two fundamental components, denoted as X and y. This segregation was predicated on the inherent demarcation between independent variables, constituting the water quality parameters and designated as model inputs (X), and the dependent variable, epitomizing the IWQI and assuming the role of the model's output (y). Such a meticulous separation was instrumental in facilitating a clear distinction between the input features and the target variable throughout the ensuing modeling phase.

To execute this partitioning strategy, we employed the train_test_split functionality from the scikit-learn library. This methodological choice ensured a seamless and randomized allocation of data, attributing 80% for training purposes and reserving the remaining 20% for subsequent testing (Fig. 3.3). Such a division into training and testing subsets served as a robust foundation for training our machine learning models and subsequently evaluating their performance on unseen data.

Cross validation subsequently is a robust resampling technique used in machine learning to assess the predictive performance and generalizability of a model. By partitioning the available dataset into complementary subsets, the method systematically trains the model on one subset while validating it on another. This process, repeated across multiple folds, ensures that each data point is used for both training and evaluation. Consequently, cross validation provides a comprehensive estimate of model performance, mitigating issues related to overfitting and variance. Common implementations include K-fold cross validation, where the data is divided into K equally sized folds, and repeated K-fold cross validation.

This standardized practice aligns with established conventions in the field, ensuring the reliability and generalizability of our model outcomes.

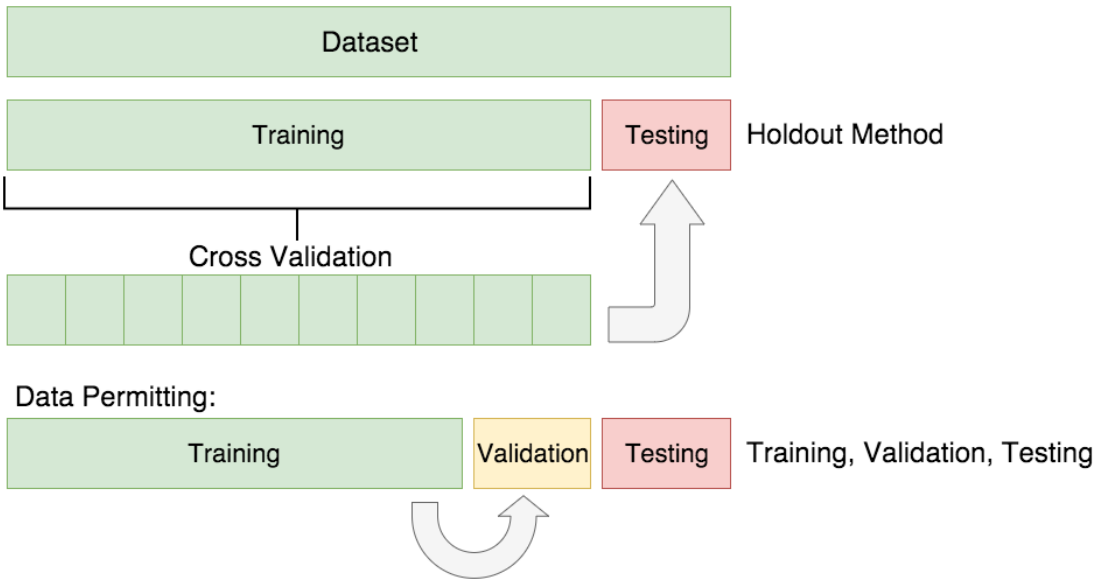

Fig. 3.3 Train test split and Cross validation

3.6.5 Data Standardization

To ensure uniformity and avert potential biases stemming from disparate scales, feature values in X underwent standardization using the StandardScaler from the sklearn library. This standardization process transformed the data to possess a mean of 0 and a standard deviation of 1, preserving the relative relationships between features while enhancing the comparability of their magnitudes.

Through these meticulously executed data pretreatment processes, the dataset was cleansed, missing values were imputed, outliers were addressed, and the data was primed for subsequent modeling and analysis. These preprocessing steps were integral to enhancing the reliability and accuracy of the results derived from our machine learning models.

3.7 Correlation analysis

Within the framework of this study, a thorough correlation analysis was conducted to examine the complex interactions between various water quality measures and the IWQI. Finding which water quality metrics showed strong correlations with the IWQI was the main goal, along with determining the direction and intensity of linear connections between variables. Pearson Correlation Coefficient: The Pearson correlation coefficient made it easier to assess linear correlations between continuous variable pairs. This statistical metric, which produces values between -1 and +1, is used to quantify the degree of linear correlation between two variables. A positive correlation is indicated by a positive coefficient, which also suggests

that both variables are increasing at the same time. A negative correlation, on the other hand, indicates a negative coefficient and implies an inverse relationship in which one variable tends to drop as the other rises. A coefficient of 0 signifies the absence of linear correlation.To ascertain the statistical significance of correlations, established standards articulated by Chan (2003) and Dancey and Reidy (2007) were rigorously applied. This analytical approach aligns with best practices in correlation analysis, ensuring the robustness and reliability of our findings.

Heatmap Visualization:In enhancing the interpretability of the correlation matrix and delineating the vigor of associations between water quality parameters and the IWQI, a heatmap emerged as a pivotal visualization tool. The heatmap, characterized by a color-coded matrix, imparted a visual representation where deeper hues denoted more robust positive correlations, while lighter tones signified either weaker or negative correlations. This graphical representation facilitated an expeditious and intuitive appraisal of the intricate relationships existing among various variables. The incorporation of a heatmap aligns with established practices in exploratory data analysis, contributing to the clarity and accessibility of our correlation findings.

3.8 Feature Engineering

3.8.1 Feature Generation

In the field of data engineering, one important aspect was creating additional features to strengthen the analytical base. Important water quality indicators were methodically calculated, including the SAR, KR, MAR, and PI. These indices were created by combining current water quality indices, which resulted in the creation of new characteristics. The purpose of this enhancement was to capture more complex relationships between the IWQI and the water quality parameters. It was intended for the inclusion of these derived indices to improve the dataset's level of detail, which in turn would improve the accuracy and predictive power of the machine learning models we used in our research (Duboue, 2020).

3.8.2 Recursive Feature Elimination with Cross-Validation (RFECV)

We implemented Recursive Feature Elimination with Cross-Validation (RFECV) in order to find the best possible feature set for our predictive models. This method creates a dynamic feature selection mechanism by combining cross-validation with Recursive Feature Elimination (RFE). Iteratively removing features from the dataset, utilizing the features that are kept, iteratively refines the model, and uses cross-validation to evaluate prediction

performance. What makes RFECV unique is its inherent ability to choose the ideal feature count that both maximizes prediction accuracy and preserves model parsimony. This method recognizes the fine balance between model complexity and performance and is in line with current feature optimization approaches.

3.8.3 Permutation Importance (PI)

To determine each feature's relative importance in our prediction models, we used Permutation Importance (PI). This approach, which was proposed by Breiman (2001), provides a quantitative assessment of the impact of shuffling a particular feature's values on the overall performance of the model. The relative importance of each feature is revealed by systematically permuting the feature values and evaluating the resulting drop in model accuracy or other specified assessment criteria. Features that exhibit a stronger influence on the model's performance under shuffling are given greater weight. By outlining the significance of features in predicting the irrigation water quality index and assisting in the thoughtful selection and interpretation of features, the use of PI fulfills its purposes.

3.8.4 Mutual Information (MI)

In our study, we employed Mutual Information (MI), a statistical method designed to quantify the extent of information shared between two variables. MI serves as a robust metric to ascertain the degree of interdependence and information exchange between distinct variables in our analytical framework.

3.9 ML models

The present study integrates various machine learning models, each tailored to fulfill distinct tasks encompassing prediction and classification. Both regression and classification aspects of these models are judiciously employed to address the overarching objectives of our investigation. The utilization of both classifier and regressor components within each model is a deliberate choice, strategically aligned with the multifaceted nature of our research inquiry.

3.9.1 Random forest

Introduced by Breiman (2001), Random Forest is a well-known ensemble learning technique that is regarded as a versatile and powerful algorithm that performs well in both classification

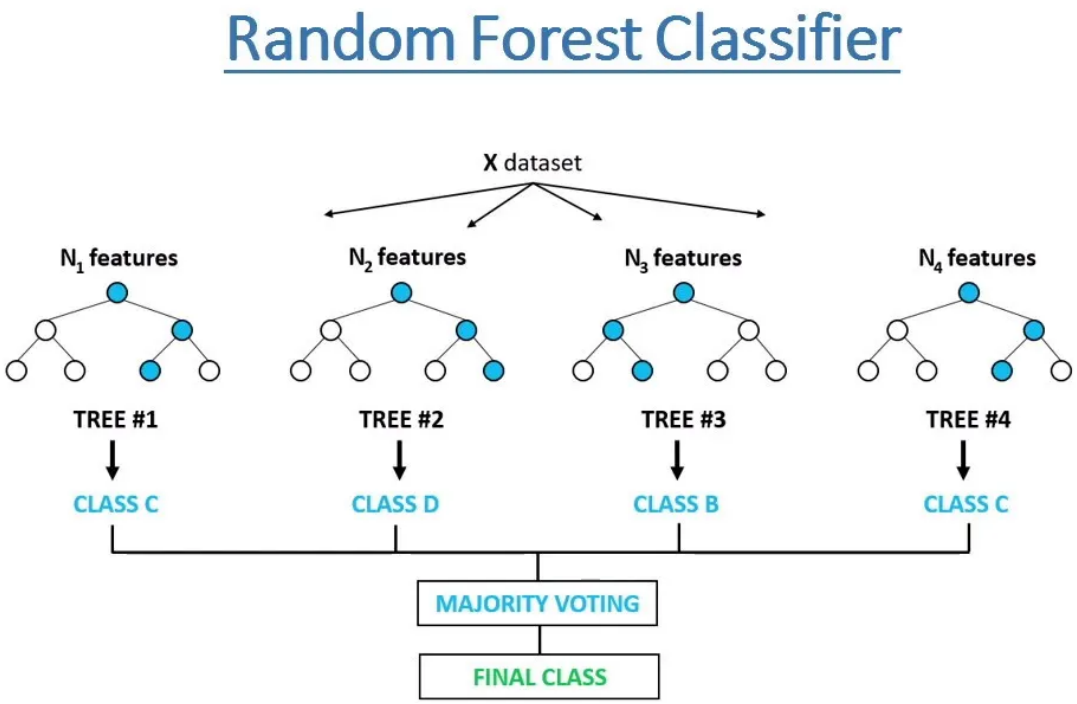

Fig. 3.4 The architecture of the random forest model

and prediction tasks. Random Forest is stands out in improving prediction accuracy and reduce overfitting in decision tree algorithms.

In the context of classification, Random Forest demonstrates its superiority, by building an ensemble of decision trees, each trained on a random selection of features and data instances (Fig. 6.2). Intentionally introducing diversity during training results in a diverse group of trees that together generate a strong classifier. The results from each decision tree are combined during prediction to get a final classification. Combining different decision trees allows Random Forest to manage intricate relationships in the data and be resilient to overfitting, which is a typical problem in machine learning. For regression tasks, the Random Forest regressor employs a similar ensemble approach. Every decision tree is trained on bootstrap samples, which are obtained from the original dataset. The various projections of these trees are combined by the regressor, usually by average, to forecast the target variable. This methodology works very well with non-linear correlations, therefore it can be applied to situations where there are complex interactions between the target variable and the input data. The capacity of the Random Forest regressor to offer a feature importance measure is one of its distinguishing characteristics (Amit and Geman, 1997). This characteristic makes feature selection easier and provides information about the relative impact of various input elements on the predictions made by the model. The technique is suitable for scalability and efficiency in real-world applications due to its parallelizability and robustness in handling huge and

complicated datasets. Hyperparameters like the number of trees in the forest, maximum tree depth, and features taken into account for splitting at each node must be carefully evaluated in order to maximize speed. Through this fine-tuning, the Random Forest model is able to function at its best, producing results that are dependable and precise in a variety of applications.

3.9.2 Extra Trees

Extra Trees is a machine learning ensemble learning method that is an enhanced form of the Random Forest paradigm. With unique feature selection and separation processes, the Extra Trees methodology differs from its predecessor, the Random Forest, and is specifically tailored for classification and prediction problems. In the context of classification, the Extra Trees Classifier distinguishes itself by using an original feature selection method. The Extra Trees Classifier goes one step farther than the Random Forest method, which chooses a random subset of features and determines the best split. To find the best split for every decision tree, it uses a random threshold in addition to selecting a random subset of characteristics. This divergence results in a significant decrease in computation time, rendering the Extra Trees Classifier a viable substitute that may attain superior accuracy in some situations in contrast to Random Forest.

The integration of Extra Trees in classification tasks involves the construction of an ensemble of decision trees, each employing the aforementioned feature selection and separation mechanisms. During prediction, the output of these individual trees is amalgamated to produce a final classification. The accelerated computation, stemming from the distinctive approach to feature selection, positions the Extra Trees Classifier as a compelling choice for scenarios where rapid and accurate classification is imperative. In prediction tasks, the Extra Trees regressor employs a comparable approach. It makes use of the same cutting-edge feature selection method, producing a collection of unique decision trees. The outputs of these trees are combined to produce the final forecast, guaranteeing a strong and trustworthy model that can handle complex relationships in the data. With its distinct feature selection strategy, the Extra Trees approach makes a significant contribution to the field of ensemble learning. It provides a sophisticated and effective substitute that is especially well-suited for situations in which computational speed is crucial without sacrificing predictive accuracy.

3.9.3 Gradient Boosting Classifier

The concept of gradient boosting, initially introduced by Friedman (2001), has evolved into a versatile and widely applicable machine learning algorithm capable of addressing both

classification and regression tasks. Rooted in the iterative minimization of a targeted loss function, gradient boosting operates by consecutively training additional estimators within the sequence. Decision trees function as the basic building blocks of the gradient boosting technique in the domains of classification and regression. A new estimator, usually a decision tree, is introduced to the growing sequence at each iteration (Fig. 6.3). The main goal of this iterative procedure is to gradually reduce the initial loss function, improving the predictive power of the model with each new addition.

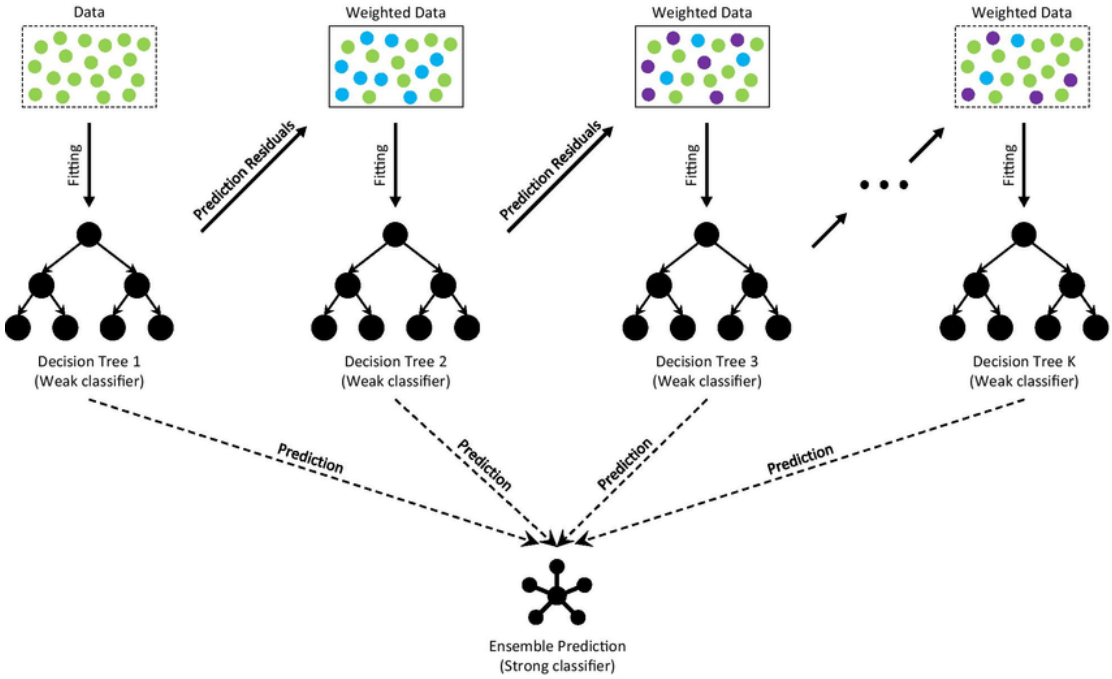

Fig. 3.5 The architecture of the gradient boosting trees model (Deng et al., 2021)

Gradient boosting integration in classification problems is contingent upon decision trees' collective contribution to the ensemble. Collectively, these carefully chosen trees that have been trained to maximize the classification objective improve the model's predictive accuracy. Gradient boosting's iterative structure guarantees a continuous improvement in the model's ability to discern patterns and relationships within the data. In the same way, gradient boosting in regression problems uses decision trees to create an ensemble that performs exceptionally well in continuous result prediction. Gradient boosting creates an intricate understanding of intricate linkages within the data by iteratively adding additional trees and fine-tuning the model depending on the underlying loss function. The utility of gradient boosting extends beyond its versatility; it is particularly effective in scenarios where intricate relationships, non-linear patterns, and nuanced dependencies are prevalent. As an ensemble learning method, gradient boosting continues to be a cornerstone in the machine learning

toolkit, offering a robust and adaptable solution for a diverse array of predictive modeling challenges.

3.9.4 XGBoost

XGBoost, an ensemble learning technique, has established its prominence in various machine learning problems, encompassing regression tasks (Chen and Guestrin, 2016). Operating within the gradient boosting framework, XGBoost amalgamates the predictions of multiple weak learners, often represented as decision trees, to craft a robust and accurate predictive model (Fig. 6.4).

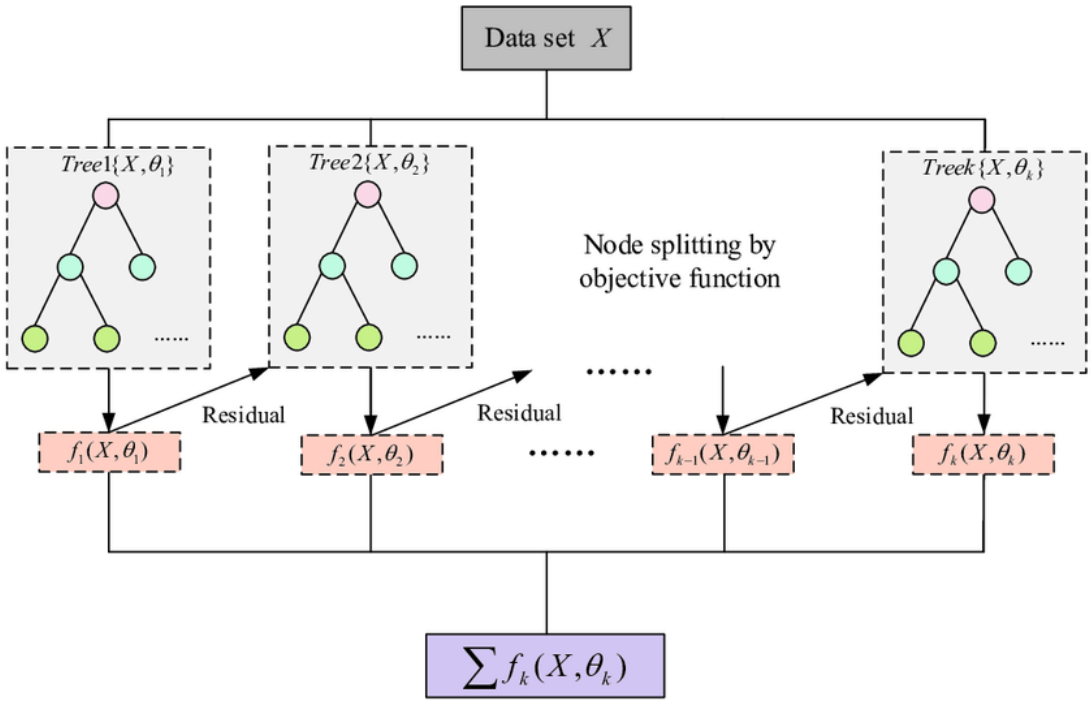

Fig. 3.6 The architecture of the extreme gradient boosting trees model

XGBoost is based on the progressive production of decision trees, each of which is designed to correct the mistakes made by its predecessors. One noteworthy aspect of its methodology is the inclusion of a unique regularization technique known as the "regularized learning objective." This method is deliberately used to reduce overfitting, improving the model's ability to generalize and its resistance to data-related noise. Scalability and efficiency are two of XGBoost's key characteristics. XGBoost is designed to manage large datasets with high-dimensional features with ease, making it a suitable option for solving challenging real-world problems. In this situation, using "approximate tree learning" is essential since it significantly speeds up training while maintaining predicted accuracy. Performance

adjustment of XGBoost is enabled by a collection of hyperparameters that are optimized using cross-validation. In order to balance model complexity with predictive accuracy, these parameters—which include the learning rate, number of trees (iterations), tree depth, and regularization parameters—are essential. A notable characteristic of XGBoost is that it offers a feature importance measure, which gives researchers information about how important each feature is in relation to the other within the prediction model.

3.9.5 Categorical Boosting (Catboost)

CatBoost, a state-of-the-art gradient boosting methodology introduced by Prokhorenkova et al. (2018), stands out as an advanced solution tailored for efficient handling of categorical data, rendering it particularly adept for classification and regression tasks. The nomenclature "Category Boosting" encapsulates its primary focus and has contributed to its widespread adoption, driven by its stellar performance and user-friendly attributes. A distinctive and pivotal trait of CatBoost lies in its innate ability to manage categorical variables seamlessly, obviating the need for explicit encoding or preprocessing. This is made possible through the implementation of the innovative "ordered boosting" method, a mechanism that inherently incorporates information from categorical features. This strategic approach negates the necessity for conventional techniques like one-hot encoding or label encoding. This feature proves invaluable when dealing with datasets exhibiting a mix of numerical and categorical attributes, a commonplace scenario in numerous real-world applications (Dorogush et al., 2018). Operating within the framework of gradient boosting, akin to other algorithms like XGBoost and LightGBM, CatBoost iteratively trains a sequence of decision trees. Each tree aims to rectify errors introduced by its predecessors, thereby enhancing predictive accuracy. Crucially, CatBoost integrates ordered boosting and feature importance estimations during training, effectively addressing the nuances associated with categorical features and augmenting model performance. In the context of regularization, CatBoost incorporates the "ordered boosting" technique, strategically implemented to mitigate the risk of overfitting. This regularization strategy ensures that the model learns from less significant features, reducing susceptibility to overfitting in the presence of noisy data. CatBoost's most notable characteristics is its computational efficiency, which is defined by memory economy and the capacity to manage massive datasets with millions of records and thousands of features without sacrificing processing speed. This feature makes CatBoost a compelling option for working with high-dimensional, real-world data relevant to environmental research and water resources.

3.9.6 LightGBM (LGBM)

A novel gradient boosting system called LightGBM offers a sophisticated way to manage large-scale machine learning workloads, especially when large datasets are involved. LightGBM was introduced with the intention of improving accuracy and efficiency. It does this by utilizing unique methodologies in conjunction with tree-based learning techniques. Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB): As a preventative precaution against overfitting, LightGBM combines the GOSS technique with EFB supplementary features. GOSS contributes to a more efficient learning process by maximizing the sampling process by keeping examples with significant gradients. By grouping exclusive characteristics together, EFB improves regularization and guards against relying too much on any one feature during model training. This further refines the technique. Histogram-based Gradient Boosting (HGB): LightGBM utilizes Histogram-based Gradient Boosting (HGB) to increase the productivity of tree-building procedures. By converting categorical variables into histograms, this method expedites the process of making decisions during the construction of trees. Adopting HGB results in significant gains in processing efficiency, which makes LightGBM especially good at managing big datasets. Effectiveness and Straightforward Assistance for Categorical Features: LightGBM has the unusual advantage of being able to examine large datasets faster than traditional tree-based learning algorithms. This accelerated performance is essential in situations when handling big amounts of data is necessary. Furthermore, LightGBM provides direct support for categorical attributes, eliminating the need for pre-processing steps such as one-hot encoding and other pre-processing procedures. This improves the model's overall efficiency in addition to streamlining the workflow.

3.9.7 Support Vector Machine (SVM)

Support Vector Machines (SVM), introduced by Vapnik et al. (1996), represent a powerful machine learning paradigm extensively applied in both classification and regression domains. At its core, SVM operates on the principle of identifying the optimal hyperplane to effectively separate two classes of data. This optimization is achieved by concurrently minimizing empirical classification errors and maximizing the geometric margin (Fig. 6.6).

The distinguishing feature of SVM lies in its efficacy, surpassing traditional methods, and its ability to address overfitting challenges. By emphasizing the geometric margin between classes, SVM not only facilitates accurate classification but also enhances generalization performance. This is particularly advantageous when dealing with complex datasets characterized by intricate decision boundaries.

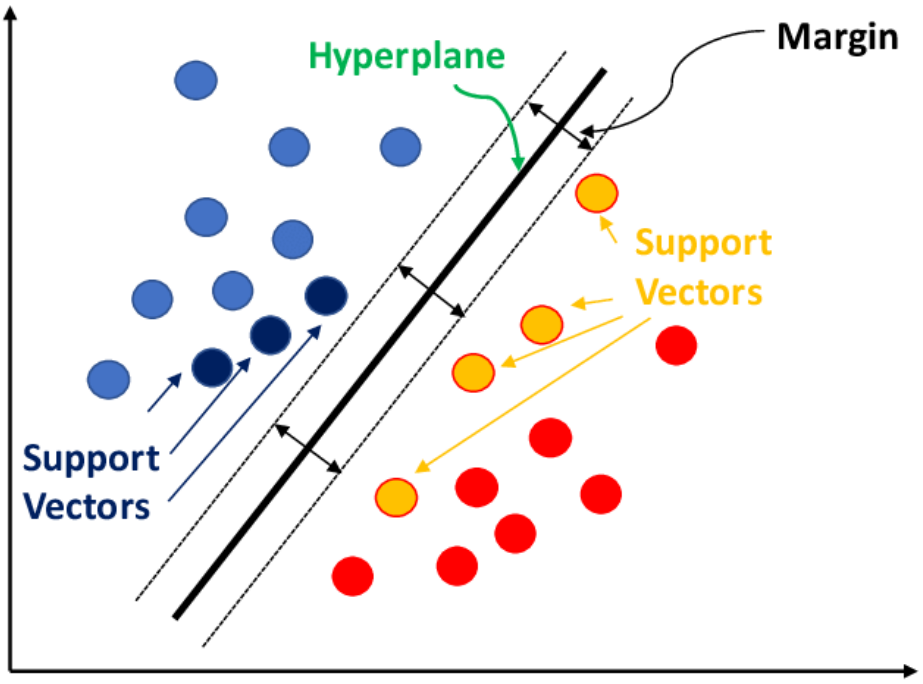

Fig. 3.7 The optimal hyperplane of SVM model

In classification tasks, SVM excels at discerning the optimal hyperplane that maximizes the margin between different classes, ensuring robust and accurate predictions. Additionally, SVM demonstrates notable resilience in scenarios where the dataset may exhibit noise or overlapping patterns. Its adaptability to nonlinear relationships is enhanced through the utilization of kernel functions, allowing SVM to operate effectively in high-dimensional spaces.

For regression applications, SVM leverages its inherent capacity to handle complex relationships by formulating a hyperplane that best captures the underlying structure of the data. This results in a predictive model that is not only accurate but also adept at accommodating intricate patterns in the dataset.

The optimum separation hyperplane is found based on solving the following optimization problem :

$$\text{Minimize:} \quad \frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i \tag{3.9}$$

$$\text{Subject to :} y_i(w^T\phi(x_i) + b \geq 1 - \xi_i),$$
$$\xi_i \geq 0, \forall i \in \{1, ..., n\}$$

$$\text{Maximize:} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) + \sum_{j=1}^{n} \alpha_j \tag{3.10}$$

$$\text{Subjet to:} \quad \alpha_i \geq 0, \quad \forall i \in \{1, ..., n\} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.11}$$

where : w is a normal vector, $\frac{1}{2}w^2$ is the regularization factor, C is the error penalty factor, b is a bias, ξ is the error function, x_i is the input vector, n is the number of elements in the training data set, $\phi(x_i)$ is a feature space, ξ_i are the training phase parameters that should be optimized $K(x_i, x_j)$: is known as the kernel function.

SVM has various kernel functions, some of which are listed below:

$$\text{Linear kernel:} \quad K(x_i, x_j) = x_i^T x_j \tag{3.12}$$

$$\text{Polynomial Kernel:} \quad K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \tag{3.13}$$

$$\text{Radial Basis (RBF):} \quad (x_i^T x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0 \tag{3.14}$$

$$\text{Sigmoid Kernel:} \quad K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \tag{3.15}$$

Where: x_i and x_j are the inputs, γ is the regularization factor. Selecting a suitable kernel and particular parameters such as γ, C, and ξ can increase the model's efficiency.

3.10 Performance evaluation metrics

Within the fields of data engineering and machine learning, model performance evaluation is a crucial aspect of determining how effective an algorithm is. In particular, when it comes to classification and prediction tasks, choosing the right metrics is crucial. Metrics for performance evaluation act as the quantitative criteria by which models' overall predictive power, accuracy, and robustness are measured. IIn this context, the judicious choice of metrics aligns with the inherent objectives of classification and prediction endeavors. From recall and precision to more comprehensive measurements like the F1 score and area under the Receiver Operating Characteristic (ROC) curve, these metrics cover a wide range of factors. Every metric provides a different perspective on how well the model differentiates the classes, shedding light on the complex interactions between true positives, false positives, true negatives, and false negatives. In addition to being a tool for evaluating model accuracy, the addition of performance evaluation metrics in classification and prediction tasks also acts as a roadmap for further model development and refinement. The development and modification

of these metrics play a crucial role in improving the dependability and interpretability of classification and prediction models as researchers delve deeper into the complexities of machine learning applications. The following sections will provide an in-depth exposition to the utilized metrics for both tasks, classification and prediction.

3.10.1 Root Mean Squared Error (RMSE)

One commonly used and important statistic in the field of regression model evaluation in machine learning and data engineering is the Root Mean Squared Error (RMSE). It functions as a measurable indicator, painstakingly documenting the differences between predicted and actual values to clarify the accuracy and dependability of regression models. The power of RMSE is its ability to extract the average residual magnitude and reveal the total effect of model predictions deviating from actual results. The vertical gaps between the model's projected values and the corresponding actual values are represented by these residuals, which stand for the residual errors. By combining these disparities into a single, all-encompassing metric, RMSE provides a logical illustration of the overall correctness of the model. Mathematically, the computation of RMSE involves the square root of the mean of the squared differences between predicted and actual values. Due to the methodological approach's natural emphasis on squared disparities, greater errors are magnified and made more noticeable during the evaluation process (Eq. 3.16). As such, RMSE not only measures the magnitude of prediction errors but also provides a subtle focus on the importance of larger errors in affecting the overall performance of the model. In its interpretive context, lower RMSE values indicate higher model accuracy since they reflect smaller prediction errors. Higher RMSE values, on the other hand, are associated with greater differences between expected and actual values, which denotes less model precision.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_p)^2} \tag{3.16}$$

3.10.2 Mean Absolute Error (MAE)

In essence, MAE shares a common objective with Root Mean Squared Error (RMSE) by encapsulating the model's predictive accuracy, albeit through a distinct computational lens. Unlike RMSE, MAE adopts an approach of calculating the absolute differences between predicted and actual values without squaring these discrepancies. This characteristic imparts a particular resilience to extreme outliers, as MAE treats all errors uniformly, prioritizing their magnitudes over directional considerations.

Mathematically, MAE manifests as the average magnitude of these absolute errors (Eq. 3.17), offering a concise representation of the model's precision. Lower MAE values are indicative of superior model performance, denoting a reduced average magnitude of prediction errors. On the contrary, higher MAE values reflect a larger average magnitude of errors, suggesting a diminished precision in the model's predictions.

MAE, akin to RMSE, emerges as a fundamental tool for researchers and practitioners engaged in refining regression models. By gauging the average magnitude of errors without the amplifying effect of squared differences, MAE complements the evaluation landscape, providing a well-rounded perspective on the predictive capabilities of regression models.

$$\text{MAE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}|y_i - y_p|} \tag{3.17}$$

3.10.3 Accuracy

Accuracy quantifies the proportion of correctly predicted instances among the total instances in the dataset. As a dimensionless quantity, accuracy provides a clear and intuitive measure of a model's ability to discern and classify data points accurately. It is particularly relevant in classification tasks, where the goal is to assign data points to specific categories or classes.

Mathematically, accuracy is computed as the ratio of correctly predicted instances to the total number of instances (Eq. 3.18), yielding a value between 0 and 1. A perfect predictive model attains an accuracy of 1, signifying that all predictions align precisely with the ground truth. Conversely, an accuracy score of 0 suggests that the model fails to make correct predictions.

Accuracy, while seemingly straightforward, is a crucial metric that demands attention, especially in scenarios where class imbalances exist. In such cases, a high accuracy score may not necessarily indicate a model's robustness, as it might disproportionately reflect the performance on the majority class while neglecting the minority class.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_p)^2}{\sum_{i=1}^{n}(y_i - y_m)^2} \tag{3.18}$$

3.10.4 Precision and Recall

Model performance is evaluated using precision and recall, two critical measures in machine learning, especially when it comes to classification tasks. These metrics are vital resources for researchers navigating the intricate world of data engineering because they offer subtle insights into a model's capacity to produce correct predictions within particular classes.

3.10.4.1 Precision

Precision, a metric with profound implications, delineates the accuracy of positive predictions made by a model. Specifically, it quantifies the ratio of true positives to the sum of true positives and false positives (Eq. 3.19). In essence, precision gauges the model's capability to precisely identify instances belonging to a positive class, minimizing the inclusion of false positives in its predictions. A precision score of 1 indicates a perfect precision, signifying that every positive prediction made by the model is indeed accurate.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.19}$$

3.10.4.2 Recall

Contrasting precision, recall, also known as sensitivity or true positive rate, captures the model's ability to identify all positive instances within the dataset. It represents the ratio of true positives to the sum of true positives and false negatives (Eq. 3.20), providing an understanding of the model's sensitivity to positive instances. A recall score of 1 indicates that the model successfully identifies all positive instances without missing any.

$$\text{Recall } = \frac{TP}{TP+FN} \tag{3.20}$$

Achieving a balanced assessment of classification models requires an understanding of the dynamic interaction between precision and recall.

3.10.5 F1 Score

The F1 score is mathematically formulated as the harmonic average of precision and recall (Eq. 3.21). This strategic choice of the harmonic mean, as opposed to the arithmetic mean, ensures that the F1 score adeptly considers both precision and recall, giving equal weight to their contributions. The F1 score ranges from 0 to 1, where a score of 1 signifies an ideal balance between precision and recall, and a score of 0 implies a lack thereof. The F1 score is especially useful in situations when finding a balance between recall and precision is crucial. For example, the F1 score offers a nuanced view of a model's efficacy in classification tasks with imbalanced datasets, when one class greatly dominates the other. It acts as a resort for researchers negotiating the fine balance needed to maximize recall and precision at the same time, guaranteeing the stability and dependability of machine learning models in real-world settings.

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.21}$$

3.10.6 ROC Curve and AUC-ROC

An informative tool that provides a nuanced view of the interaction between true positive rate (sensitivity) and false positive rate is the Receiver Operating Characteristic (ROC) curve. When investigating the model's performance across different threshold values and revealing the complex trade-offs prevalent in classification tasks, this graphical depiction becomes especially relevant.

3.10.6.1 ROC Curve Plotting

Over a range of threshold values, the true positive rate is carefully plotted against the false positive rate in the ROC curve (Eq. 3.22 and Eq. 3.23). This graphic representation offers a thorough summary of a model's discriminatory capacity and shows how well it can discern between positive and negative instances. The trajectory of the curve provides a dynamic representation of the model's performance and indicates how sensitive the model is to changes in the decision threshold.

3.10.6.2 Area Under the Curve (AUC-ROC)

The Area Under the Curve (AUC), or more particularly AUC-ROC, is integral to the ROC curve and measures the classification model's overall performance. The likelihood that the model will accurately distinguish between positive and negative cases is reflected in the AUC-ROC. A model that performs better at discriminating across a range of threshold values has a higher AUC-ROC score.

3.10.6.3 Interpretation and Significance

A thorough grasp of the underlying trade-offs in classification tasks is made possible by the ROC curve and AUC-ROC. A model that successfully strikes a balance between sensitivity and specificity and achieves an AUC-ROC score near 1 is considered to have robust performance. On the other hand, a model with an AUC-ROC value of about 0.5 indicates performance that is equal to random chance.

$$\text{True positive rate} = \frac{TP}{TP + FN} \tag{3.22}$$

$$\text{false positive rate} = \frac{FP}{FP+TN} \tag{3.23}$$

While:
TP: True positives.
TN: True negatives.
FP: False positives.
FN: False negatives.

3.11 SHAP (SHapley Additive exPlanations)

In the field of machine learning model interpretability, the SHAP (SHapley Additive exPlanations) technique is a cutting-edge tool that lays out a roadmap for removing the many layers of ambiguity that surround prediction algorithms (Lundberg and Lee, 2017).

3.11.1 Philosophy and Foundation

SHAP aims to discern the contributions of each feature to the prediction outcome. It is based on cooperative game theory. Taking cues from Shapley values—a notion that was first presented to allocate rewards equitably among participating players—the technique attempts to allocate fair values to every variable by assessing how each feature affects the model's predictions. Essentially, SHAP aims to clarify the opaque internal mechanisms of machine learning models, adding a level of clarity and understanding (Biecek and Burzykowski, 2021).

3.11.2 Elucidating Predictive Outputs

Essentially, SHAP assigns a portion of the result to each feature in order to deconstruct the intricate ensemble of variables affecting a prediction. It provides a sophisticated knowledge of the relative significance of distinct features in influencing predictions by closely examining the "black box" nature of machine learning models. This deconstruction is essential in understanding the individual contribution of the features utilized in the training phase.

3.11.3 Interpretability Across Models

One of SHAP's main advantages is that it is not dependent on any particular model (model-agnostic). This means that it may be used with a wide range of machine learning models,

such as support vector machines, decision trees, and neural networks. Because of its universality, SHAP is useful for a wide range of machine learning applications and promotes interpretability regardless of the model architecture.

3.11.4 Contributions Visualized

The method's prowes is further revealed by the sophisticated visuals it creates. SHAP results are converted into easily understood visuals that show how each feature influences the model's prediction. These visual aids give end users and data scientists alike a clear understanding of the variables influencing model outcomes.

3.12 Adopted Methodology

In response to this identified gap, our study introduces a methodologically robust framework that strategically incorporates considerations of water consumers' affordability. A key innovation lies in the utilization of Mutual Information (MI) and other techniques as a feature selection strategy. This approach ensures precision in water quality assessment while simultaneously aligning with economic considerations. By integrating affordability into the assessment process, our methodology enhances the real-world applicability of water quality studies.

The selection of LightGBM, Catboost, Extra Trees, and Random Forest classifiers is grounded in their proven accuracy and efficiency across diverse scenarios. This methodological choice is not arbitrary; rather, it is a deliberate step toward advancing the understanding of effective machine learning tools tailored for water quality assessment. The study's distinctive contribution lies in delving into the intersection of machine learning and economic considerations, thereby adding a novel dimension to the existing discourse in water quality research.

Having meticulously examined the existing literature, pinpointing crucial gaps in the current discourse on water quality assessment, our study is poised to transcend these limitations through a carefully designed methodology. As we navigate from the comprehensive review of prior research to the upcoming methodology chapter, it becomes evident that our research is uniquely positioned to address the identified gaps.

3.13 Conclusion

The methods employed in this study, ranging from the selection of machine learning models to the utilization of feature selection techniques and performance metrics, have been meticulously designed to achieve the overarching goals of groundwater quality assessment for irrigation purposes. The adoption of advanced machine learning algorithms, including LGBM, CatBoost, Extra Trees, Random Forest, Gradient Boosting classifiers, Support Vector Machines, Multi-Layer Perceptrons, and the K-Nearest Neighbors Algorithm, reflects a thoughtful consideration of diverse methodologies to ensure a comprehensive evaluation.

Feature selection techniques, such as Mutual Information, were strategically utilized to identify the most influential parameters contributing to the variability of the dataset. This approach not only optimized model efficiency but also ensured economic viability by reducing the number of input parameters without compromising classification accuracy.

Performance metrics, encompassing ROC-AUC, precision-recall, F1 score, and accuracy, were chosen with precision to evaluate the models rigorously. The incorporation of water quality indices, particularly the IWQI, provided a robust reference for classification, aligning the machine learning models with real-world applicability in the context of irrigation.

The next chapter will transition seamlessly from the methodologies employed to the comprehensive exploration of results and discussions. Through an in-depth analysis of the outcomes, we aim to unveil critical insights into groundwater quality dynamics, the efficacy of machine learning models, and their implications for water resource management. The results and discussions chapter will further contribute to the evolving understanding of efficient and economically viable water quality assessment methodologies, building upon the foundations laid in the literature review and methods chapters.

Chapter 4

Water Quality Assessment For Irrigation Purposes

4.1 Introduction

This section represents a comprehensive endeavor aimed at conducting a holistic evaluation of water quality specifically tailored for irrigation purposes. This preliminary assessment stands as a crucial prerequisite, laying the groundwork for more intricate phases of research, particularly the application of Machine Learning (ML) models for classification and prediction in subsequent stages. To attain the objectives of this water quality assessment for irrigation, a set of well-established water quality indices has been judiciously employed. The key water quality indices utilized in this assessment encompass the Irrigation Water Quality Index (IWQI), Sodium Adsorption Ratio (SAR), Soluble Sodium Percent (Na%), Potential Salinity (PS), Permeability Index (PI), Magnesium Adsorption Ratio (MAR), and Kelly's Ratio (KR). This diverse set of indices collectively enables a nuanced understanding of various facets of water quality, each index addressing specific parameters critical for irrigation suitability.

Additionally, this assessment integrates a hydrochemical characterization of the water samples within the study area. This complementary analysis provides insights into the inherent chemical composition of the water, enriching the understanding of the contextual factors influencing water quality. The amalgamation of these indices and hydrochemical characterizations forms the basis for a robust and multidimensional evaluation of water quality, essential for informing subsequent phases of research and decision-making in agricultural and environmental domains.

4.2 Methods

The holistic assessment of water quality for irrigation in this study is grounded in a multi-faceted approach. Central to this evaluation is the utilization of various indices, notably the Irrigation Water Quality Index, SAR, MAR, Na%, Permeability Index (PI), and KR. These indices collectively contribute to a comprehensive analysis, offering insights into different facets of water quality relevant to irrigation purposes.

Furthermore, to achieve a thorough characterization of the hydrochemical facies of groundwater, the study employs the Piper diagram as a robust tool. This diagram serves as a valuable instrument in elucidating the intricate relationships between different chemical components in groundwater, contributing to a nuanced understanding of the hydrochemical composition. Together, these methodologies establish a robust framework for the holistic assessment of water quality, encompassing both quantitative indices and graphical representations to provide a comprehensive overview for informed decision-making in water resource management and agricultural sustainability.

4.3 Results and discussion

The quality of irrigation water is inherently dynamic, intricately influenced by a myriad of constituents that emanate from the surrounding environment, with the soil type exerting a particularly profound impact. These constituents collectively constitute the quality parameters of irrigation water, meticulously cataloged in Table 4.1. The foundation of water quality evaluation rests upon a comprehensive understanding of prevalent soil-related challenges. In this context, the research delves into salinity hazards, water infiltration rates, ion toxicity, and various other issues of diverse typologies (Ayers and Westcot, 1985). These soil-related predicaments serve as pivotal metrics in the holistic assessment of water quality, shaping the criteria for quantitative analysis. To rigorously quantify these hazards, the research adopts a battery of criteria, each designed to unravel specific facets of the water quality landscape. Noteworthy indices employed for this purpose include the Irrigation Water Quality Index (IWQI), SAR, Na%, PS, Permeability Index (PI), MAR, and KR. Detailed statistical insights pertaining to each index are meticulously presented in Table 6.1. The synergistic application of these indices serves as the linchpin for a nuanced understanding of irrigation water quality, elucidating not only the presence of specific contaminants but also their potential impact on soil, crops, and overall agricultural sustainability.

Table 4.1 Statistical characteristics of water quality parameters

	mean	std	min	max
T (°C)	21.14	4.97	8.30	34.40
pH	7.36	0.37	6.08	9.30
Conductivity (μs/cm)	1979.97	989.72	426	8970
Turbidity (NTU)	6.19	15.54	0.01	147
TDS (mg/l)	829.77	344.72	193	1895
TAC (°F)	107.14	98.80	3.20	400
HCO_3^- (mg/l)	337.78	109.85	0	976
Total Hardness (°F)	349.87	376.83	0	1800
Ca^{2+} (mg/l)	177.32	86.09	0	544
Mg^{2+} (mg/l)	90.52	44.82	0	238.14
Cl^- (mg/l)	186.50	148.20	0	942.88
NO_2^- (mg/l)	0.04	0.23	0	5
NH_4^+ (mg/l)	0.06	0.33	0	5
SO_4^{-2} (mg/l)	629.90	255.71	100	1600
NO_3^- (mg/l)	33.32	111.50	0	2379
Fe^{+2} (mg/l)	0.03	0.09	0	0.72
PO_4^{-3} (mg/l)	0.03	0.05	0	0.10
Na^+ (mg/l)	116.74	72.51	3	450
K^+ (mg/l)	5.24	2.73	0.80	18

4.3.1 Salinity hazard

The salinity hazard emerges as a paramount concern in irrigated regions, as emphasized by De Paz et al. (2004). The deleterious consequences stemming from salt accumulation manifest in water scarcity for plants, inducing symptoms akin to those observed in plants facing drought conditions Ayers and Westcot (1985). In this intricate interplay between soil and water quality, electrical conductivity (EC) assumes significance, representing the sum of anions or cations and standing as a proxy for dissolved solids Wilcox (1955).

Within the ambit of this study, the assessment of the salinity hazard hinges on the meticulous evaluation of EC and Total Dissolved Solids (TDS). Table 4.1 presents a comprehensive overview, indicating that EC values ranged from 426 to 8970 µs/cm, while TDS values spanned from 193 to 1895 mg/l, with mean values of 1979.97 µs/cm and 829.77 mg/l, respectively. Employing the TDS classification proposed by Allison and Richards (1954), a predominant 82.98% of water samples are deemed fit for use with moderate restrictions, as delineated in Table 4.2. In alignment with the Wilcox standard based on EC, 85.04% of water samples are classified as severely saline (Table 4.2).

Table 4.2 Standard values of water quality parameters with samples percentage and classes (Ayers et al., 1985).

WQ parameter	Range	Parameter Class	Samples percentage
pH	6.5-8.4	NR	97.73
Conductivity	< 700	NR	3.49
	700-3000	MR	11.17
	> 3000	SR	85.04
TDS	< 450	NR	17.02
	450-2000	MR	82.98
	> 2000	SR	0.00
Total Alkalinity	< 150	NR	3.82
	150-300	MR	31.36
	> 300	SR	68.26
Na^{+}	< 70	NR	25.16
	70-200	MR	66.67
	> 200	SR	8.18
Mg^{2+}	< 140	NR	85.50
	140-355	MR	13.74
	> 355	SR	0.00
Ca^{2+}	0-400	NR	97.14
	> 400	R	2.86
$HCO_3{}^{-}$	< 90	NR	0.19
	90-500	MR	94.10
	> 500	SR	5.33
Cl^{-}	< 140	NR	44.93
	140-350	MR	40.34
	> 350	SR	14.34
Total Hardness	0-6	S	0.00
	6-12	MH	0.00
	12-18	H	0.00
	> 18	VH	99.62
$NO_3{}^{-}$	< 5	NR	24.29
	5-30	MR	47.14
	> 30	SR	28.37
$PO_4{}^{-3}$	0-2	NR	100.00
	> 2	R	0.00
Fe^{2+}	< 0.5	NR	99.62
	0.5-1.5	MR	0.37
	> 1.5	SR	0.00

Note: The acronyms in the Table stand for: NR:No Restriction, MR:Moderate Restriction, SR:Severe Restriction, R:Restriction, S:Soft, MH:Moderate Hard, H:Hard, VH:Very Hard.

This result suggests that there is a demonstrable risk that using the water from the study area for irrigation may raise the salinity levels in agricultural areas. Because of the complex relationship between soil quality and water supplies, irrigation management must be done carefully, and salinity measures are important tools for making well-informed decisions

4.3.2 pH

The role of pH in water quality is pivotal, especially in the context of irrigation, where the physicochemical properties of water significantly influence soil health and plant growth. The pH values, as delineated in Table 4.1, exhibit a range from 6 to 9, with a mean value of 7.36. These values serve as fundamental indicators, guiding our understanding of the suitability of water for irrigation purposes. In adherence to the guidelines established by the FAO-UN Ayers and Westcot (1985), which provide a robust framework for water quality assessment, the majority of the samples, precisely 97.73%, are classified as safe for use without any restriction (Table 4.2). This classification aligns with the overarching goal of ensuring optimal conditions for agricultural productivity, where pH serves as a barometer of the water's acidity or alkalinity. The pH levels within this permissible range promote a soil condition that is favorable to microbial activity and nutrient availability, two critical components of plant growth. The research yielded subtle insights that go beyond numerical numbers, providing light on the practical implications for sustainable irrigation techniques. Using water with a pH that is suited to the area reduces the likelihood of soil erosion and prolongs the life of farming landscapes. As a result, the pH component of water quality assessment becomes essential to the comprehensive effort to maximize irrigation effectiveness while maintaining soil health.

4.3.3 Total Hardness

Total hardness, a parameter reflective of the concentration of calcium (Ca^{2+}) and magnesium (Mg^{2+}) ions in water samples, is a crucial determinant in irrigation water quality assessment. The comprehensive analysis of water samples, detailed in Table 4.1, reveals an average total hardness value of 349.87 French degrees, with a maximum value reaching 1800 French degrees. The classification provided by EPA Gold Book (1986) categorizes the majority of the water samples within the very hard class (Table 4.2).

The high total hardness values observed pose potential challenges for sustained agricultural practices. According to Ewaid et al. (2019), prolonged use of such water can lead to the clogging of irrigation equipment, a consequence of the precipitation of minerals within

the water. Additionally, foliar staining problems may arise, impacting the overall aesthetic quality of crops and potentially affecting market value.

Understanding the total hardness of irrigation water is integral to devising effective water management strategies. This knowledge allows for the implementation of preventive measures to mitigate equipment damage and optimize irrigation efficiency. The nuanced insights gained from assessing total hardness underscore its significance as a key parameter in the broader framework of irrigation water quality evaluation.

4.3.4 Ions Concentration

The presence of ions, encompassing sodium, calcium, magnesium, potassium, chloride, and bicarbonates, holds paramount significance in both soil and water, as their uptake in substantial quantities can detrimentally impact crop yields. Maintaining a delicate balance is imperative, as even at lower concentrations, these ions have the potential to induce crop-related issues. In this study, a meticulous analysis of ion concentrations in water samples was conducted, and the findings are meticulously presented in Table 4.1.

For cations such as Na^+, K^+, Ca^{2+}, Mg^{2+}, and Fe^{2+}, concentrations exhibited variations spanning from 3, 0.8, 0, 0, and 0 mg/l to 450, 18, 544, 238.14, and 0.72 mg/l, respectively, with mean values of 116.74, 5.24, 177.32, 90.53, and 0.03 mg/l, correspondingly. On the anionic front, Cl^-, $HCO_3{}^-$, $NO_3{}^-$, and $PO_4{}^{-3}$ ranged from 0 to 942.88, 976, 2379, and 0.1 mg/l, with average values of 186.5, 337.78, 33.32, and 0.03 mg/l, respectively.

To gauge the appropriateness of ion concentrations, this research employed standard values as outlined in Table 4.2. Concerning cations, a substantial majority of water samples were categorized as non-restricted for Fe^{2+}, Ca^{2+}, and Mg^{2+}, with percentages of 99.62%, 97.14%, and 85.50%, respectively. As for anions, Cl^-, $HCO_3{}^-$, and $NO_3{}^-$ witnessed 40.34%, 94.10%, and 47.14%, respectively, falling within the moderately restricted class (MR). Notably, $PO_4{}^{-3}$ demonstrated 100% adherence to the non-restricted class.

A comprehensive evaluation of water quality extends beyond a sole focus on ion concentrations, as this approach may fall short in unveiling issues arising from intricate interactions among water constituents. Recognizing this limitation, the study incorporates water quality indices to provide a more nuanced understanding of the overall water quality scenario. In the ensuing sections, the Water Quality Index (WQI) will be employed as a robust tool to undertake a thorough assessment of water quality.

In pursuit of a visual representation of the chemical composition of water samples, the study adopts the Stiff diagram, as illustrated in Figure 6.2. This diagram serves as a graphical representation of the mean values of major cations and anions observed over the span of four years (2019-2022). The Stiff diagram offers a swift and insightful overview of the

variations in mean values of the most prevalent ions. This visualization not only facilitates a comprehensive understanding of the water chemistry but also aids in identifying patterns and trends, contributing to a more holistic interpretation of irrigation water quality.

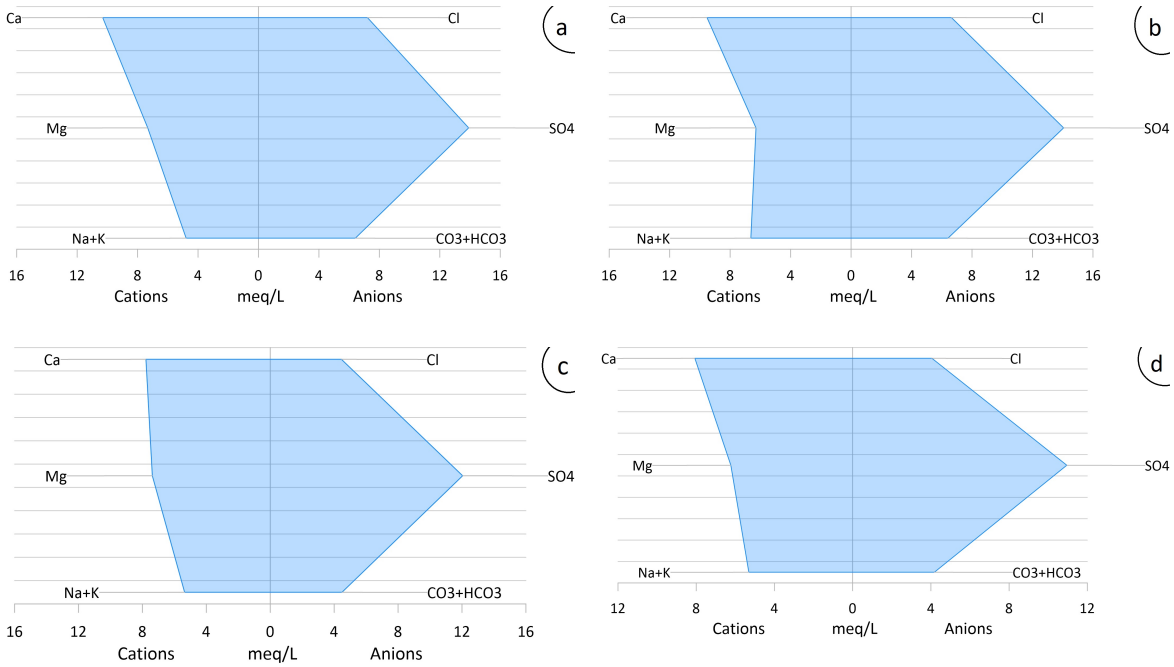

Fig. 4.1 Stiff diagram: a) Stiff diagram of 2019, b) Stiff diagram of 2020; c) Stiff diagram of 2021, d) Stiff diagram of 2022

4.3.5 Total Alkalinity

Prior to delving into the results of the Water Quality Index (WQI), it is imperative to underscore the significance of the total alkalinity parameter (TAC) in the context of irrigation water quality. TAC serves as a pivotal descriptor of water's capacity to stabilize pH levels, gauged by the presence of bicarbonates (HCO_3^-), carbonates (CO_3^{2-}), and hydroxides (OH^-). The influence of TAC on crops is multifaceted, manifesting as ion imbalances and nutritional complications at elevated levels (>300 °F), while at lower levels (<3 °F), the water may lack the capability to neutralize acidity effectively. As delineated in Table 4.1, TAC spans a range from 3.2 to 400 °F, with an average value of 107.14 °F. Strikingly, 68.26% of the examined samples fall into the category of severely restricted water for irrigation use (SR). This categorization implies an anticipation of nutrient deficiencies in areas where irrigation with such water is prevalent. An ameliorative measure proposed to address this

issue involves the injection of acids into the water. The ubiquity of HCO_3^-, as expounded in the preceding section, may offer insights into the origins of water alkalinity.

4.3.6 Water quality indices

The subsequent discussion entails the interpretation of the outputs generated by the water quality indices employed in this study.

4.3.6.1 Irrigation Water Quality Index (IWQI)

The Irrigation Water Quality Index (IWQI) has long been recognized as a potent instrument for succinctly characterizing the status of water quality, condensing extensive groundwater quality data into a singular representative value (Uddin et al., 2021b). The IWQI categorizes irrigation water into five classes based on its impact on both irrigated plants and soil, as detailed in Table 4.3.

Table 4.3 Classification of irrigation water according to IWQI

WQI	Class	Samples percentage (%)	Recommendations for soil	Recommendations for plant
85-100	No restriction (NR)	0.00	Can be employed for most type of soils with low salinity and sodicity problems	Safe for the majority of plants
70-85	Low restriction (LR)	8.24	Can be employed in light soil textures or moderate permeability	risky to salt sensitive plants
55-70	Moderate restriction (MR)	44.06	Can be employed in soils with moderate to high permeability	Plants with moderate salt sensitivity
40-55	High restriction (HR)	41.95	Can be employed in soil of high permeability	plants with moderate to high salt sensitivity
0-40	Severe restriction (SR)	5.75	Avoid its use	Plants with low salt sensitivity

In light of the classification criteria, the water samples underwent the following categorization: 44.06% were assigned to the Moderate Restriction class (MR), indicating limitations

on their use for irrigating plants with moderate salt sensitivity. Concurrently, 41.95% of the collected samples fell within the High Restriction range (HR), imposing constraints on groundwater utilization for plants with high to moderate salt sensitivity and high permeability. A mere 8.24% of the water samples attained classification as Low Restricted water (LR), permitting irrigation for a broad spectrum of plants except those sensitive to salt.

Conversely, 5.75% of the analyzed water samples found themselves in the Severe Restriction class (SR), severely limiting groundwater usage to plants with low salt sensitivity. Notably, no samples were categorized in the non-restricted category. This classification provides an insight into the suitability of water for irrigation purposes, taking into account both the specific salt sensitivity of plants and the permeability of the soil.

Table 4.4 Groundwater classification based on WQI

Index	Range	Class	Samples percentages
SAR (Richards, 1954)	< 10	excellent	100.00
	10-18	good	0.00
	18-26	doubtful	0.00
	> 26	unsuitable	0.00
Na% (Eaton, 1950)	< 20	excellent	25.95
	20-40	good	67.72
	40-60	permissible	6.01
	60-80	doubtful	0.32
	> 80	unsuitable	0.00
KR (Kelley, 1940)	< 1	suitable	98.73
	> 1	unsuitable	1.27
PI (Doneen, 1964)	> 75	suitable	0.32
	25-75	good	85.44
	< 25	unsuitable	14.24
MAR (Raghunath, 1987)	< 50	suitable	60.00
	> 50	unsuitable	40.00
PS (Doneen, 1954)	< 3	suitable	1.19
	> 3	unsuitable	98.81

4.3.6.2 Sodium Adsorption Ratio (SAR)

Table 4.5 Statistical characteristics of water quality indices

	mean	std	min	Q1	Q2	Q3	max
SSP	24.31	10.00	0.73	18.88	24.54	30.16	60.42
SAR	1.86	1.02	0.04	1.27	1.75	2.21	5.89
PI	36.84	11.08	9.93	30.45	36.95	43.78	77.46
KR	0.35	0.20	0.01	0.23	0.33	0.44	1.59
PS	11.57	5.74	2.04	7.17	10.90	14.05	34.68
Na%	25.00	10.07	0.98	19.57	25.32	30.86	61.93
MAR	45.84	10.45	2.57	42.43	47.66	51.27	76.03

The prevalence of sodium ions in the study area underscores the imperative need to scrutinize potential sodicity issues, leading to the application of the SAR index. SAR emerges as a pivotal diagnostic tool specifically designed to assess the hazard of sodicity within the water utilized for irrigation. The Sodium Adsorption Ratio is calculated by determining the ratio of sodium concentration to the combined concentrations of magnesium and calcium. This ratio serves as a fundamental indicator of the propensity for soil sodification, offering valuable insights into the potential challenges associated with sodium accumulation in the soil. Upon analyzing the results presented in Table 6.1, it is evident that SAR values within the study area range from 0.04 to 5.89 meq/l, with an average SAR value of 1.86 meq/l. This quantification unveils the sodium proportion relative to magnesium and calcium concentrations, providing a nuanced perspective on the sodicity hazard. Notably, based on the established standard values of SAR delineated in Table 4.4, the findings indicate that 100% of the water samples fall within the category of excellent water quality in terms of SAR. This classification implies that the water is well-suited for irrigation purposes without posing adverse effects on soil structure or compromising the growth of sodium-sensitive crops. The discerned excellence in SAR values underscores the suitability of the water for sustaining agricultural practices and emphasizes its compatibility with crops that exhibit sensitivity to sodium levels. This is based on the sole reliance on SAR values, albeit not satisfactory, to achieve the overarching goal of the study, i.e., a comprehensive evaluation of the water quality. This drives the research further to explore more water quality indices as they provide a multifaceted analysis.

4.3.6.3 Sodium Percent (Na%)

The Sodium Percent (Na%) index serves as a valuable metric for assessing the influence of sodium on the overall quality of irrigation water. This index provides insights into the relative concentration of sodium concerning magnesium, calcium, and potassium, offering a nuanced perspective on the water's suitability for irrigation purposes. The categorization of irrigation water quality based on Na% values is detailed into five classes, namely excellent, good, permissible, doubtful, and unsuitable, as outlined in Table 4.4. Within the study area, the Na% values exhibit a range from 0.98 to 61%, with a calculated mean value of 25% (refer to Table 6.1). The distribution of Na% values indicates that 25.95%, 67.72%, and 6.01% of the water samples fall into the categories of excellent, good, and permissible, respectively. A marginal 0.32% of groundwater samples are classified as doubtful, signifying a slight uncertainty in the water quality for irrigation. Importantly, no water sample is relegated to the unsuitable category, affirming a general appropriateness for agricultural use. Understanding the relationship between Sodium SAR and Na% is essential to comprehending the whole effect of water on soil structure. Combining the Na% and SAR assessments offers a synergistic viewpoint that show the absence of sodicity problems and supports the thoroughness of the water quality study.

4.3.6.4 Kelly's Ratio (KR)

Kelly's Ratio (KR) stands as a pivotal index in the comprehensive evaluation of water quality for irrigation purposes, providing valuable insights into the potential sodicity hazards associated with the water samples. This index categorizes water quality into two distinct classes based on its calculated value. Specifically, if the KR value is less than 1, the water is deemed suitable, whereas a value exceeding 1 renders the water unsuitable for irrigation (refer to Table 4.4). In the context of the study area, the analysis of KR values reveals a range from 0.01 to 1.59 meq/l, with a computed average value of 0.35 meq/l (refer to Table 6.1). The outcome of the KR index assessment underscores the favorable quality of the groundwater in the study area. Notably, a substantial 98.73% of the water samples fall within the suitable range, attesting to their compatibility with irrigation needs. Only a minimal 1.27% of the samples are categorized as unsuitable, signifying a minor proportion with potential concerns. This work's findings collectively affirm that the groundwater quality in the study area predominantly aligns with the criteria for suitability in terms of the sodicity hazard.

4.3.6.5 Permeability Index (PI)

The transition from assessing sodicity hazards through indices such as Na%, SAR, and KR is complemented by the incorporation of the Permeability Index (PI). This index serves as a valuable tool in furthering the comprehensive evaluation of water quality, offering insights into the water's impact on soil permeability, a critical aspect in irrigation water quality assessment.

The Permeability Index categorizes water quality into three distinct classes: suitable, good, and unsuitable, as outlined in Table 4.4. In the specific context of the study area, the PI values for the examined water samples span a range from 9.93 to 77.46%, with a calculated average value of 36.84% (refer to Table 6.1).

The results derived from the PI assessment indicate that a significant majority, precisely 85.44% of the water samples, fall within the classification of good water quality. This suggests a favorable soil permeability outcome associated with the majority of the groundwater in the study area. However, it is noteworthy that 14.24% of the water samples are categorized as unsuitable, signifying potential concerns regarding soil permeability for this subset of samples. Interestingly, a minor proportion, only 0.32% of the samples, are classified as suitable, indicating a limited subset with optimal soil permeability characteristics.

The inclusion of the Permeability Index (PI) enriches the water quality assessment framework, offering a nuanced understanding of soil permeability dynamics.

4.3.6.6 Magnesium Adsorption Ratio (MAR)

Magnesium Adsorption Ratio (MAR) is used to complement the the prior assessment of the sodicity and the infiltration hazards. This index plays a key role in further elucidating the intricacies of water quality concerning magnesium concentrations, contributing valuable insights into the suitability of irrigation water for the study area. Within the study area, MAR values exhibit a diverse spectrum, ranging from 2.57 to 76.03%, with a computed average value of 45.84% (see Table 6.1). The MAR classification schema, which stratifies water quality into suitable and unsuitable categories, reveals that 60% of the examined samples fall within the suitable classification, signifying optimal magnesium concentrations for irrigation purposes. In contrast, 40% of the samples are categorized as unsuitable, indicating a subset with magnesium concentrations that may pose challenges in the context of irrigation water quality. A more complex understanding of the consequences of magnesium for the interaction between soil and water is offered by the addition of the MAR to the array of assessment indicators. The MAR contributes to a better understanding of magnesium's involvement in irrigation water quality by distinguishing between appropriate and inappropriate categories.

4.3.6.7 Potential Salinity (PS)

Extending the evaluation beyond sodicity hazards and infiltration rate assessments, the focus now converges on the quantification of salinity hazards, a crucial facet in understanding the overall water quality for irrigation purposes. In this endeavor, the Potential Salinity (PS) parameter takes center stage, providing valuable insights into the salinity-related challenges that may be encountered in the study area.

The computation of Potential Salinity, as delineated by the equation (3.6), yields values ranging from 2.04 to 34.38 meq/L, encapsulating a diverse range of salinity levels within the examined water samples. The average PS value is calculated at 11.57 meq/L, indicative of the general salinity conditions prevalent in the study area.

Upon classification, the results underscore the salinity challenges faced, with an overwhelming majority of the water samples, amounting to 98.81%, falling within the unsuitable class. This classification highlights the potential impediments posed by salinity, emphasizing the need for targeted interventions and management strategies to mitigate the impact of salinity hazards on irrigation practices.

4.3.7 Hydrochemical Characterization of the Water Samples

As the investigation progresses from hazard quantification to a more nuanced exploration, an integral aspect involves the hydrochemical characterization of the groundwater samples. To unravel the intricate composition of the water and gain insights into its hydrochemical facies, the study adopts the Piper diagram as a powerful tool in this endeavor. The Piper diagram (Fig 6.3), pioneered by Piper (1944), stands as a graphical representation that unravels the hydrochemical facies and origins of predominant cations and ions in the water samples. The diagram categorizes key components, including calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^{+}), bicarbonates ($HCO_3{}^{-}$), Chloride (Cl^{-}), and sulfate ($SO_4{}^{-2}$), providing a comprehensive depiction of the hydrochemical composition. The methodology, as proposed by Piper (1944), involves plotting two triangles. The first triangle encapsulates calcium (Ca^{2+}) and magnesium (Mg^{2+}) as "alkaline earths" and sodium (Na^{+}) as "alkali." The second triangle encompasses sulfate ($SO_4{}^{-2}$) and chloride (Cl^{-}) as "strong acids" and bicarbonates ($HCO_3{}^{-}$) as "weak acid." A diamond synthesizes the outcomes of the two triangles, offering a holistic view of the hydrochemical facies. The positional assignment of ions within the diagram is pivotal for deciphering the hydrochemical facies. By drawing perpendicular lines from the sample point in each triangle towards the diamond, the intersection of these lines determines the ion's position within the diamond.

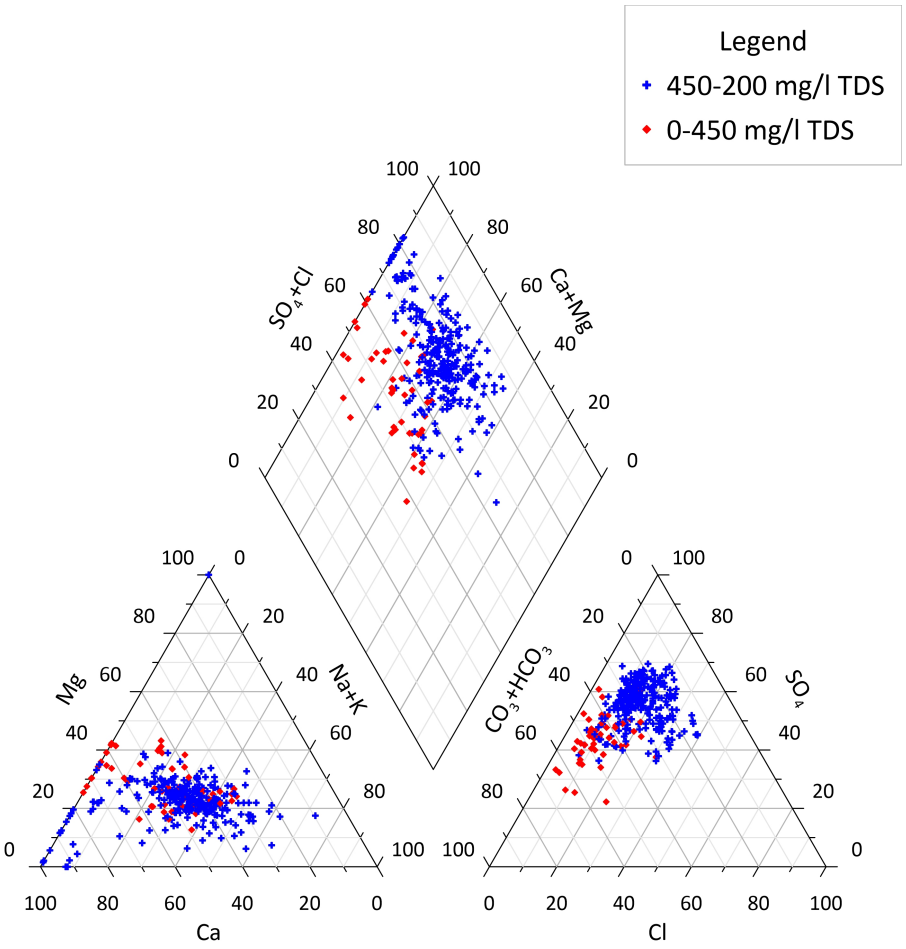

Fig. 4.2 Pipper diagram

Upon careful analysis of the Piper diagram output, the study reveals compelling insights. The prevalence of alkaline earths (Ca^{2+} and Mg^{2+}) surpasses that of alkalies (Na^{+} and K^{+}). Furthermore, the dominance of strong acids ($SO_4{}^{-}$ and Cl^{-}) over weak acids ($CO_3{}^{-2}$ and $HCO_3{}^{-}$) is evident in the majority of the water samples. The diamond plot underscores a notable trend, with no cation-anion pairs exceeding 50%. Consequently, the dominant hydrochemical type emerges as mixed, followed by the calcium chloride facies type (Ca^{2+}+Cl^{-}). Essentially, the Piper diagram's ability to allow hydrochemical characterisation adds a layer of sophistication to our comprehension of water composition. It offers important information about the dominant hydrochemical facies and directs future decisions for the management of water resources.

4.4 Discussion

Following the analysis of groundwater samples against established quality standards, several deviations from recommended values were observed. In this discussion, we delve into the underlying causes of these deviations, shedding light on the factors contributing to the observed results.

High Salinity

In the semi-arid study area, groundwater exhibits high salinity, as evidenced by elevated EC and TDS. This salinity is primarily attributed to evaporation, particularly in shallow aquifers, compounded by the presence of hypersaline lakes. Minerals like gypsum, halite, and carbonate rocks contribute to groundwater salinity due to the area's geological composition. Salinity is further increased by irrigation return flow, leaching salts from soil and fertilizers, and mixing with irrigation water, often higher in salinity than native groundwater. Industrial and agricultural wastewater also introduce various salts and chemicals, including chlorides, sulfates, nitrates, phosphates, and metals, into groundwater.

High Hardness, Total Alkalinity, and Bicarbonate

Elevated levels of hardness, total alkalinity, and bicarbonate in groundwater indicate the presence of calcium, magnesium, and carbonate ions. These ions originate from the dissolution of carbonate rocks, facilitated by increased soil carbon dioxide levels, leading to carbonic acid formation, pH reduction, and enhanced carbonate solubility. Organic matter degradation, exacerbated by livestock overgrazing, can further elevate groundwater hardness and alkalinity by introducing organic matter and nutrients, stimulating microbial activity, and carbonate precipitation. Industrial, agricultural, and livestock activities can also introduce calcium-magnesium-bicarbonate-rich sources, impacting the carbonate system with diverse organic and inorganic compounds.

4.5 Conclusion

In conclusion, this comprehensive study underscores the imperative of employing diverse classification indices to attain a holistic comprehension of irrigation water quality. The meticulous analysis of key parameters, including EC and TDS, has revealed that an alarming majority, exceeding 85% of the water samples, falls within the category of severe saline water. Consequently, prudent water management strategies, possibly involving treatment

processes, are recommended to avert salinity-related challenges, such as diminished crop yields. The investigation into water hardness indicates that a substantial proportion of the sampled water manifests as very hard, signaling a potential risk of clogging in water network equipment upon prolonged utilization of the groundwater from the study area. Regarding cations concentrations, iron (Fe), calcium (Ca), and magnesium (Mg) conform within the standard limits. However, stringent restrictions on water usage are imperative due to elevated concentrations of sodium (Na^+). Conversely, concerning anions, while phosphate ($PO_4{}^-$) concentrations adhere to permissible levels, chloride (Cl^-), bicarbonate ($HCO_3{}^-$), and nitrate ($NO_3{}^-$) surpass the prescribed limits, leading to a classification of water samples under the category of moderate restriction.

The assessment of water alkalinity through the Total Alkalinity parameter designates the water as severely restricted, necessitating mineral acid injection to counteract potential crop nutrient deficiencies arising from heightened alkalinity. IWQI highlights the severity of the water quality situation, with over 90 of the analyzed samples falling within the categories of moderately to highly restricted, signifying challenges for the normal growth of crops with moderate to low salt sensitivity.

Nevertheless, a nuanced interpretation emerges when considering additional indices such as SAR, Na%, KR, Permeability Index (PI), and MAR, which collectively categorize the majority of water samples as of good to excellent quality for irrigation purposes. Expectedly, PS classifies the water samples under the unsuitable category, necessitating careful consideration.

Furthermore, the hydrochemical facies, elucidated through the Piper diagram analysis, delineates a nuanced water composition. The prevalence of alkaline earths over alkalies and the dominance of strong acids over weak acids are discerned. The mixed type emerges as the dominant hydrochemical facies, followed by the calcium chloride facies type.

In light of these findings, it is imperative for relevant authorities and institutions to acknowledge the substantial implications of incessant groundwater utilization in the study area. Urgent attention to treatment measures before water application is crucial to circumvent potential agricultural and environmental repercussions. The synthesis of insights from diverse indices and hydrochemical characterization furnishes a comprehensive foundation for informed decision-making in water resource management.

Chapter 5

ML-Based Irrigation Water Quality Classification

5.1 Introduction

In this pivotal section of the thesis, the focus transitions towards pioneering advancements in the integration of machine learning methodologies dedicated to the nuanced classification of groundwater quality, specifically tailored for irrigation purposes. The overarching objective is to harness the transformative potential of machine learning techniques to distill actionable insights from groundwater data, facilitating informed decision-making in agriculture and water resource management. The imperative evaluation of water quality stands as a linchpin for safeguarding both environmental integrity and human well-being. Despite the considerable strides made in employing machine learning for assessing water quality, there exists a research gap concerning its application in the classification of groundwater devoted for irrigation. Particularly noteworthy is the scarce literature exploring the efficacy of utilizing machine learning with a reduced set of input parameters while still achieving satisfactory classification outcomes. This study embarks on addressing this research gap by meticulously investigating the feasibility of employing machine learning for the classification of groundwater designated for irrigation purposes. A distinctive facet of this research lies in its endeavor to achieve robust classification outcomes using a minimalistic set of input parameters. The methodology involves the development of machine learning models that simulate the Irrigation Water Quality Index (IWQI) and an economic model, with a deliberate emphasis on optimizing the number of inputs to maximize accuracy. To elucidate the diverse landscape of machine learning classifiers, eight algorithms were meticulously selected for evaluation. These include the LightGBM classifier, CatBoost, Extra Trees, Random Forest, Gradient

Boosting classifiers, Support Vector Machines, Multi-Layer Perceptrons, and the K-Nearest Neighbors Algorithm. Two distinct scenarios were contemplated to assess the classification performance. The first scenario utilized six inputs, encompassing conductivity, chloride (Cl^-), bicarbonate ($HCO_3{}^-$), sodium (Na^+), calcium (Ca^{2+}), and magnesium (Mg^{2+}). The second scenario reduced the input parameters to three, namely total hardness (TH), chloride (Cl^-), and sulfate ($SO_4{}^{2-}$), judiciously selected based on the Mutual Information (MI) results. This innovative investigation aims to advance the field of machine learning applications in water quality assessment while also providing a pragmatic framework for optimizing inputs, which will increase the effectiveness and scalability of irrigation-specific groundwater quality classification models.

5.2 Data Preprocessing

In our study, the data preprocessing phase plays a pivotal role in refining the dataset to ensure the quality and effectiveness of subsequent machine learning models. A series of essential steps, including data cleansing, normalization, missing value imputation, and feature selection techniques, are systematically implemented. To address missing data values, the KNN imputer is employed, utilizing the K-nearest neighbors algorithm with a specified k value of 5. This imputation technique calculates the average values from the five nearest neighbors to fill missing data points, ensuring a robust and data-driven approach to handle incomplete records. The subsequent classification task assigns classes based on Meireles' recommended classes, wherein IWQI values for each water sample are determined and corresponding class values are assigned. These class values range from 0 to 5, representing IWQI intervals of 0–40, 40–55, 55–70, 70–85, and 85–100, respectively. To mitigate potential negative impacts associated with unbalanced data on machine learning algorithms, the Synthetic Minority Over-sampling Method (SMOTE) is employed to oversample the minor classes. SMOTE facilitates the generation of synthetic instances within the minority class, addressing imbalances and enhancing the overall robustness and generalization capabilities of the models. The careful execution of these preprocessing steps lays the foundation for subsequent machine learning analyses, ensuring the reliability and accuracy of the models.

5.3 Correlation Analysis of Parameters

Correlation analysis stands as a pivotal step in machine learning, facilitating the identification of key features and the development of robust predictive models. This analysis entails evaluating pairwise correlations among irrigation water quality parameters, assigning

values within the range of -1 to +1. The investigation discerns a high positive correlation between conductivity and several variables, including (SO_4^{2-}), (Na^+), (Mg^{2+}), and (Ca^{+2}) (Fig. 5.1). This observation implies a potential presence of multicollinearity, which can have ramifications on both model stability and interoperability.

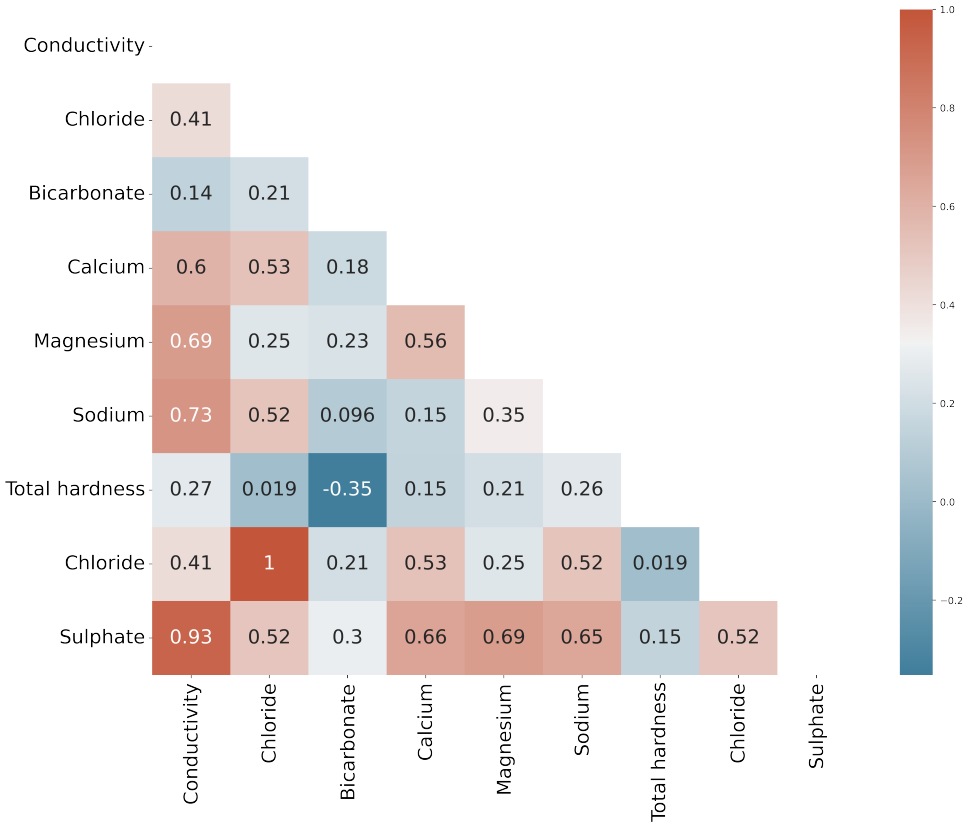

Fig. 5.1 Correlation Heatmap of the parameters used as inputs

Moreover, positive correlations are observed among (SO_4^{2-}), conductivity, (Ca^{+2}), (Mg^{2+}), and (Na^+), indicating interdependencies among these parameters. Conversely, a weak correlation is identified between (HCO_3^-) and other parameters. However, relying solely on these correlations may not be adequate to pinpoint the most influential parameters contributing to the dataset's variability. To address this limitation, mutual information analysis (MI) is employed to identify the most influential parameters systematically. This approach offers a more nuanced understanding of the dataset, ensuring a comprehensive evaluation of parameter significance beyond the scope of traditional correlation measures (Fig. 5.2). In tandem with correlation analysis, pairplots are utilized to visually comprehend the relationships between input parameters (refer to Fig. 5.3). This graphical representation aids in identifying patterns, trends, outliers, and potential non-linear relationships. The x-axis corresponds to parameters from the first scenario, while the y-axis represents parame-

ters from the second scenario, providing a comprehensive visual overview of the dataset's interparameter relationships.

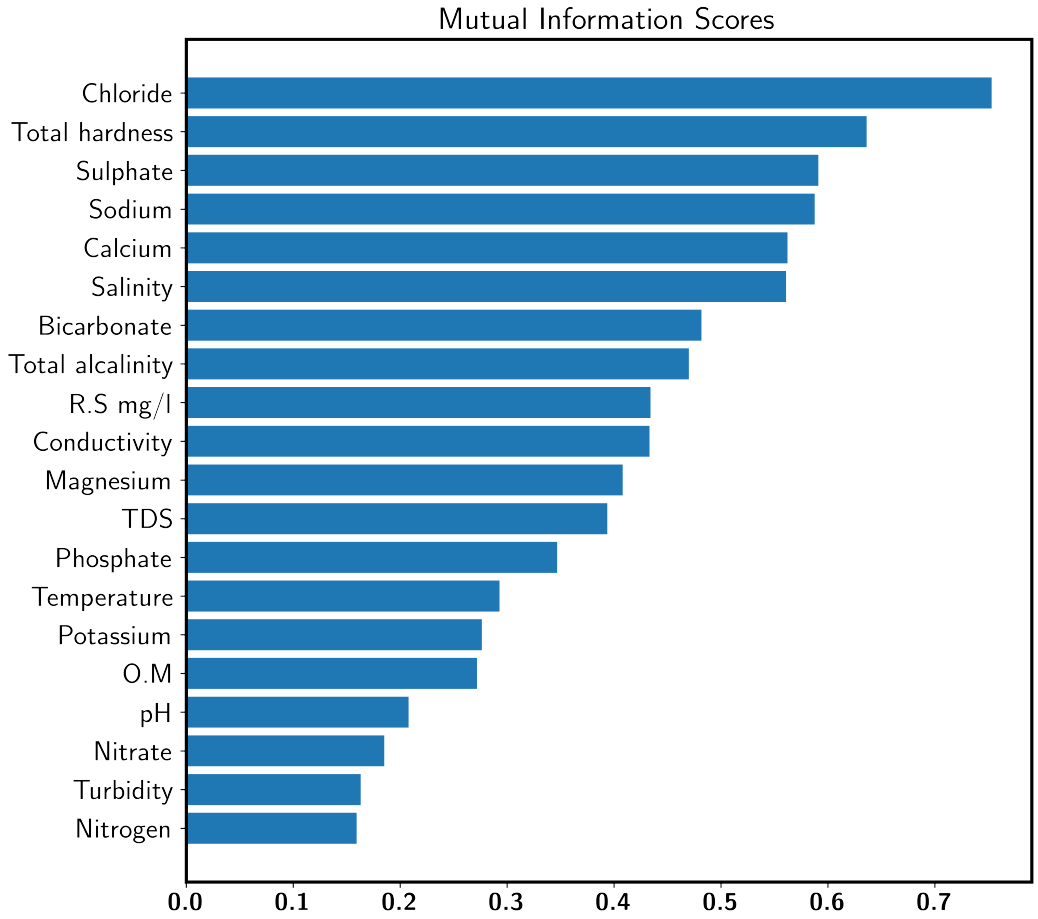

Fig. 5.2 Features importance ranking based on MI

5.4 Methodology

The methodology employed in this study represents a meticulous and systematic approach designed to achieve the overarching goal of developing robust machine learning models for the classification of groundwater quality tailored for irrigation purposes. Dataset Splitting and Parameter Selection The dataset was partitioned into training and testing sets utilizing an 80–20% split, ensuring a comprehensive yet independent assessment of model performance. The simulation of the Irrigation Water Quality Index (IWQI) was initially performed using six key parameters. Subsequently, a judicious reduction of input parameters to three—total hardness (TH), chloride (Cl^-), and sulfate ($SO_4{}^{2-}$)—was executed through the Mutual

Information approach (Fig. 6.1). This reduction aimed at optimizing model efficiency without compromising classification accuracy.

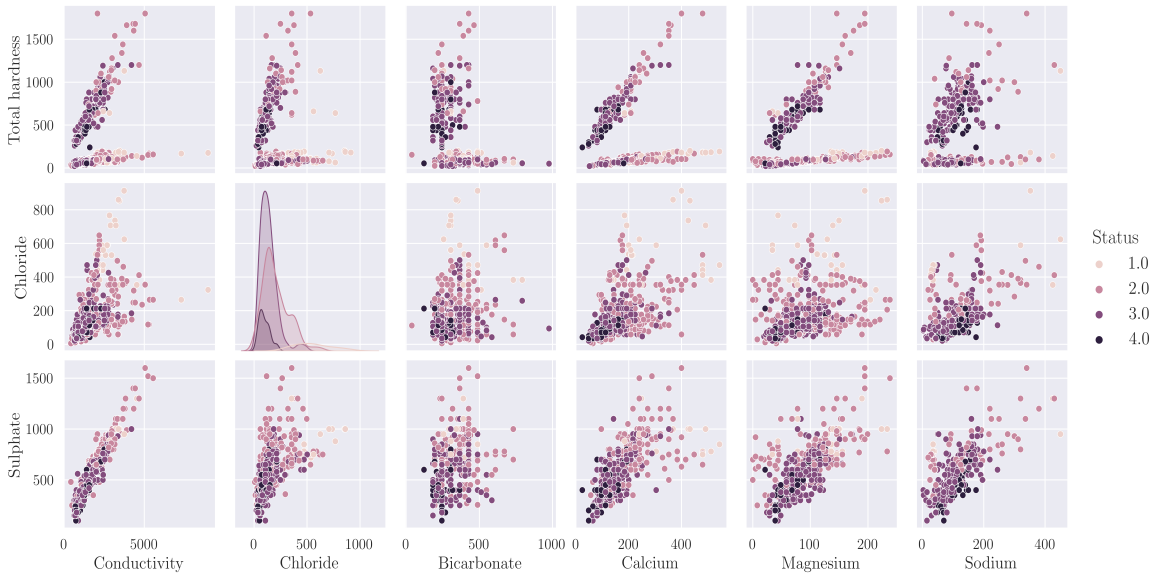

Fig. 5.3 Pairwise Relationship Plots of Input Variables in Two Scenarios

Cross-Validation for Robustness To fortify the robustness and generalization capabilities of the developed models, cross-validation was employed. Specifically, Repeated K-Fold Cross-Validation (cv=RepeatedKFold (n_splits=10, n_repeats=3)) was implemented. This technique ensures that the models are trained and tested across various subsets of the dataset, mitigating the risk of overfitting and enhancing their capacity to handle diverse data patterns. **Hyperparameter Tuning** GridSearch, a systematic and exhaustive hyperparameter tuning technique, was deployed to optimize the performance of the machine learning models. This involved an exhaustive search through a predefined hyperparameter grid to identify the most effective configuration for achieving superior classification results. The ultimate hyperparamters setting is highlighted in the Table 5.1.

5.5 Model Evaluation Metrics

The evaluation of model performance was conducted based on a comprehensive set of metrics, including accuracy, precision, recall, and F1-score. These metrics provide a nuanced understanding of the model's classification prowess, addressing aspects of correctness, completeness, and trade-offs between precision and recall. Benchmarking against Diverse Models for Robust Comparison The results obtained from the developed models were rigorously compared with those of various machine learning models. This comparative

Table 5.1 The optimal hyperparameters for the ML models.

Random Forest	Extra Trees	GBoost
n_estimators= 1030 criterion='gini' min_samples_split= 2 min_samples_leaf= 1	n_estimators=170 criterion='gini' min_samples_split=2 min_samples_leaf=1 max_features="sqrt"	n_estimators= 100 min_samples_split= 4 learning_rate= 0.1 min_samples_leaf= 2
XGBoost	**Catboost**	**LGBM**
n_estimators= learning_rate= gamma= reg_alpha= reg_lambda= base_score=	iterations=992 learning_rate=0.313	learning_rate= 0.1 n_estimators= 631

analysis aimed to discern the best-performing model for the classification of groundwater samples. The overarching goal is to contribute insights that inform the selection of an optimal machine learning approach for accurately and efficiently classifying groundwater quality for irrigation purposes developed in our study.

5.6 Results And Discussion

5.6.1 First Scenario

The evaluation of machine learning models in simulating the Irrigation Water Quality Index (IWQI) using six parameters—conductivity, chloride (Cl^-), bicarbonate ($HCO_3{}^-$), sodium (Na^+), calcium (Ca^{2+}), and magnesium (Mg^{2+})—revealed nuanced performances across various classifiers. The LightGBM (LGBM) classifier emerged as the top-performing model, achieving the highest accuracy of 91.16% in accurately classifying IWQI based on the specified parameters. Following closely, the CatBoost classifier demonstrated notable accuracy, recording a score of 90.91The Extra Trees, Random Forest, and Gradient Boosting classifiers also exhibited commendable performances, achieving accuracy scores of 90.62%, 89.72%, and 88.38%, respectively. Despite slightly reduced accuracy scores, the Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and K-Nearest Neighbors Algorithm (KNN) classifiers still showcased acceptable performances. To assess the overall value of the classifiers, a weighted average of precision and recall scores was computed. The results high-

lighted the superiority of the LGBM, CatBoost, and Extra Trees classifiers, with weighted precision scores of 91.41%, 91.16%, and 90.99%, respectively. Similarly, the weighted recall scores for these classifiers were 91.16%, 90.91%, and 90.73%, underscoring their robustness in minimizing false positives and false negatives. F1 scores, which provide a balanced measure of precision and recall, were computed for each model, further substantiating their overall performance. Additionally, the area under the ROC curve (ROC-AUC) was employed as a metric to assess the models' ability to distinguish between positive and negative classes. The results, summarized in Table 5.2, revealed high ROC-AUC values ranging from 89.33% to 98.64%, affirming the classifiers' accuracy in correctly classifying instances across both positive and negative classes.

Table 5.2 Performance Evaluation of Machine Learning Models under both scenarios

First scenario					
	ROC_AUC	precision	recall	F1	Accuracy
LightGBM Classifier	0.9864	0.9141	0.9116	0.9108	0.9116
CatBoost Classifier	0.9859	0.9116	0.9091	0.9075	0.9091
Extra Trees	0.9844	0.9099	0.9073	0.9020	0.9062
Random Forest	0.9836	0.8990	0.8972	0.8956	0.8972
Gradient Boosting	0.9775	0.8878	0.8838	0.8830	0.8838
SVM	0.9680	0.8717	0.8658	0.8647	0.8658
MLP	0.9590	0.8707	0.8642	0.8572	0.8629
KNN	0.8933	0.8413	0.8409	0.8368	0.8409
Second scenario					
Extra Trees	0.9627	0.8683	0.8647	0.8630	0.8647
Random Forest	0.9567	0.8442	0.8405	0.8393	0.8405
CatBoost Classifier	0.9567	0.8499	0.8471	0.8449	0.8405
SVM	0.9466	0.8408	0.8380	0.8353	0.8380
KNN	0.8919	0.8415	0.8377	0.8351	0.8377
LightGBM Classifier	0.9566	0.8394	0.8348	0.8346	0.8348
Gradient Boosting	0.9553	0.8354	0.8322	0.8283	0.8308
MLP	0.9394	0.8084	0.7952	0.7909	0.7902

Visual representation of the models' performance was provided through confusion matrices (Fig. 6.4), offering insights into accurate and inaccurate classifications across various classes. The matrices presented true positive, false positive, false negative, and true negative values, essential for determining precision, recall, and F1 scores for each class. A comprehensive evaluation, including precision, recall, F1 score, and accuracy values, is presented in

Table 5.2 and Figure (Fig. 6.3). These findings collectively emphasize the robust performance of the selected machine learning models in simulating the IWQI for groundwater quality assessment in the first scenario.

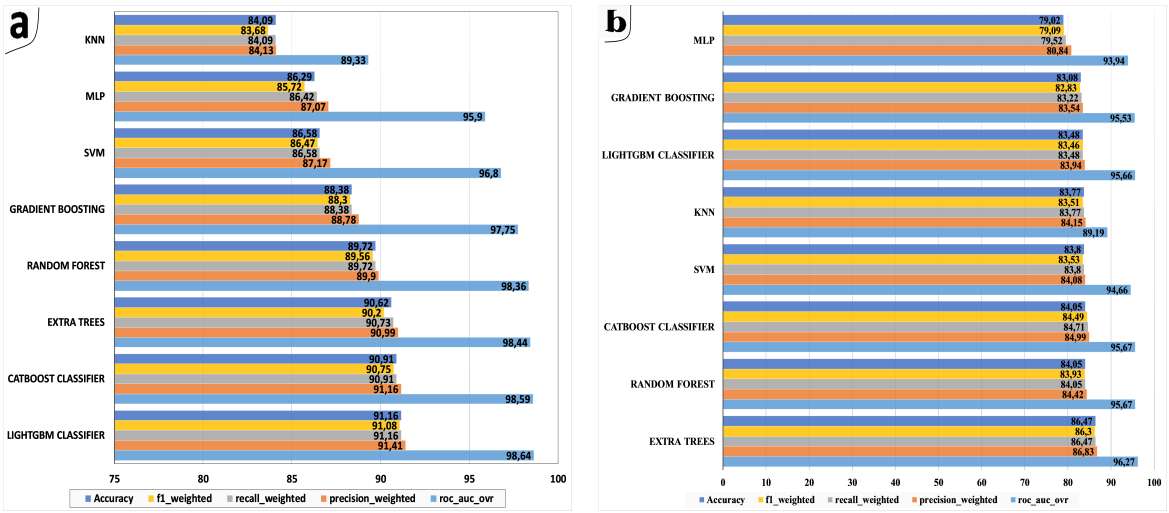

Fig. 5.4 Performance of the ML models in the 2 scenarios: a) First scenario, b) Second scenario

5.6.2 Second Scenario

In the second scenario, the study aimed to minimize model input parameters by employing mutual information (MI) to identify the most influential factors contributing to the dataset's variability. Following the MI analysis, a subset of three parameters—total hardness (TH), chloride (Cl^-), and sulfate (SO_4^{2-})—was selected as input for the machine learning models. The model's performance under these specified conditions underwent a comprehensive evaluation using various metrics, as detailed in Table (5.2), alongside corresponding values for each machine learning (ML) model. The Extra Trees classifier demonstrated superior performance, achieving an accuracy score of 86.47%, outperforming other models. Following closely were the Random Forest (RF) and CatBoost classifiers, each attaining an accuracy score of 84.05%. SVM, K-Nearest Neighbors (KNN), and LightGBM (LGBM) models exhibited accuracy scores of 83.80%, 83.77%, and 83.48%, respectively. The Gradient Boosting and Multi-Layer Perceptrons (MLP) classifiers concluded with accuracy scores of 83.08% and 79.02%, respectively. Assessing the F1 score, the Extra Trees classifier emerged as the top performer with a score of 86.30%, closely followed by CatBoost, RF, and SVM classifiers with F1 scores of 84.49%, 83.93%, and 83.53%, respectively. Other classifiers

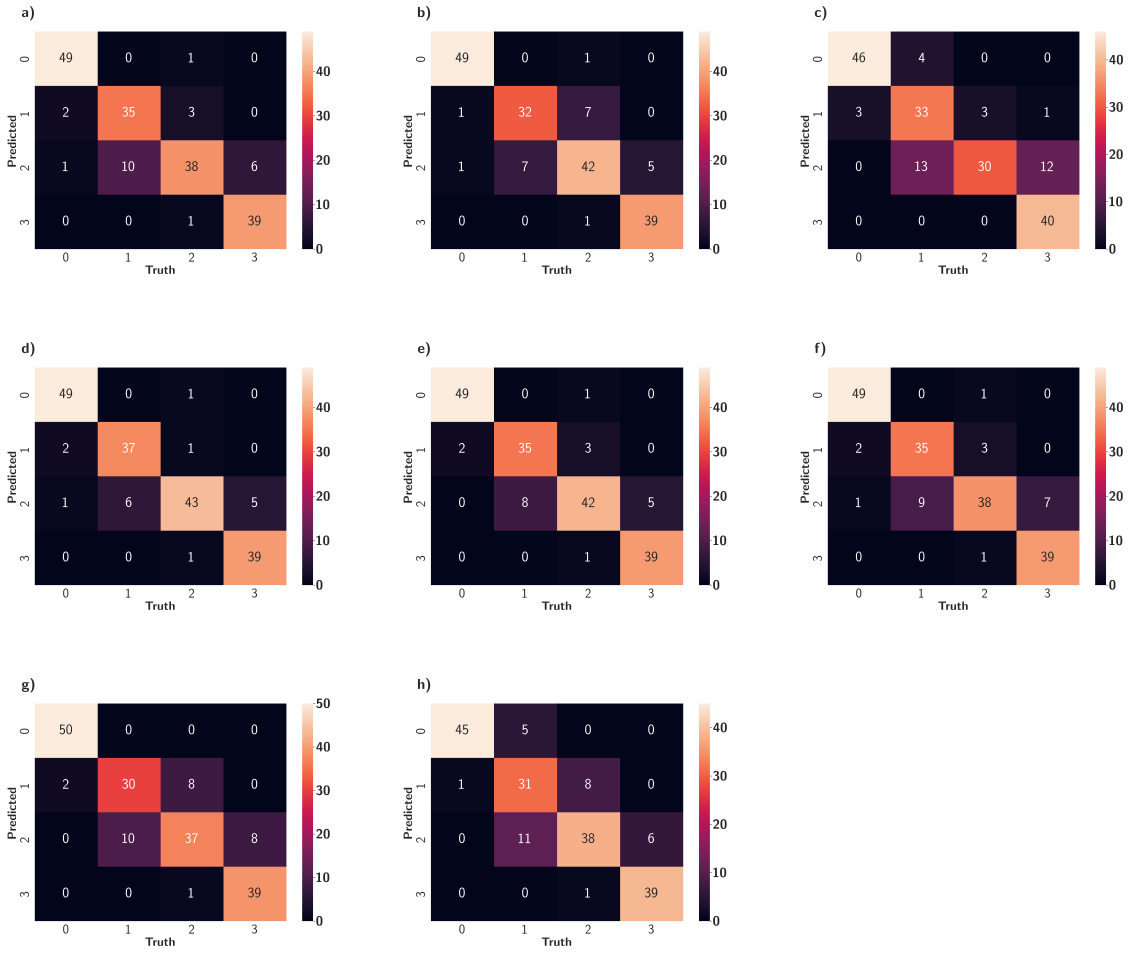

Fig. 5.5 Confusion matrices of the ML models for the first scenario: a) CatBoost classifier, b) Extra Trees Classier, c) Multilayer perceptrons classifier, d) Gradient Boosting classifier, e) LGBM classifier, f) Random Forest classifier, g)KNN classifier, h)SVM classifier

exhibited F1 scores ranging from 79.09% to 83.77%. Emphasizing the importance of the F1 score in assessing classification models due to its ability to account for both false positives and false negatives, the values are presented in Table 5.2 along with precision and recall values. ROC AUC values, ranging from 93.94% to 96.27%, were obtained, indicating the substantial ability of all employed models to accurately identify and categorize instances of both positive and negative classes. The confusion matrices for each classifier, illustrated in Fig.6.6, provide a visual representation of their performance in classifying instances across various classes. These results underscore the effectiveness of the Extra Trees classifier in achieving accurate groundwater quality classification for irrigation purposes with a reduced set of input parameters.

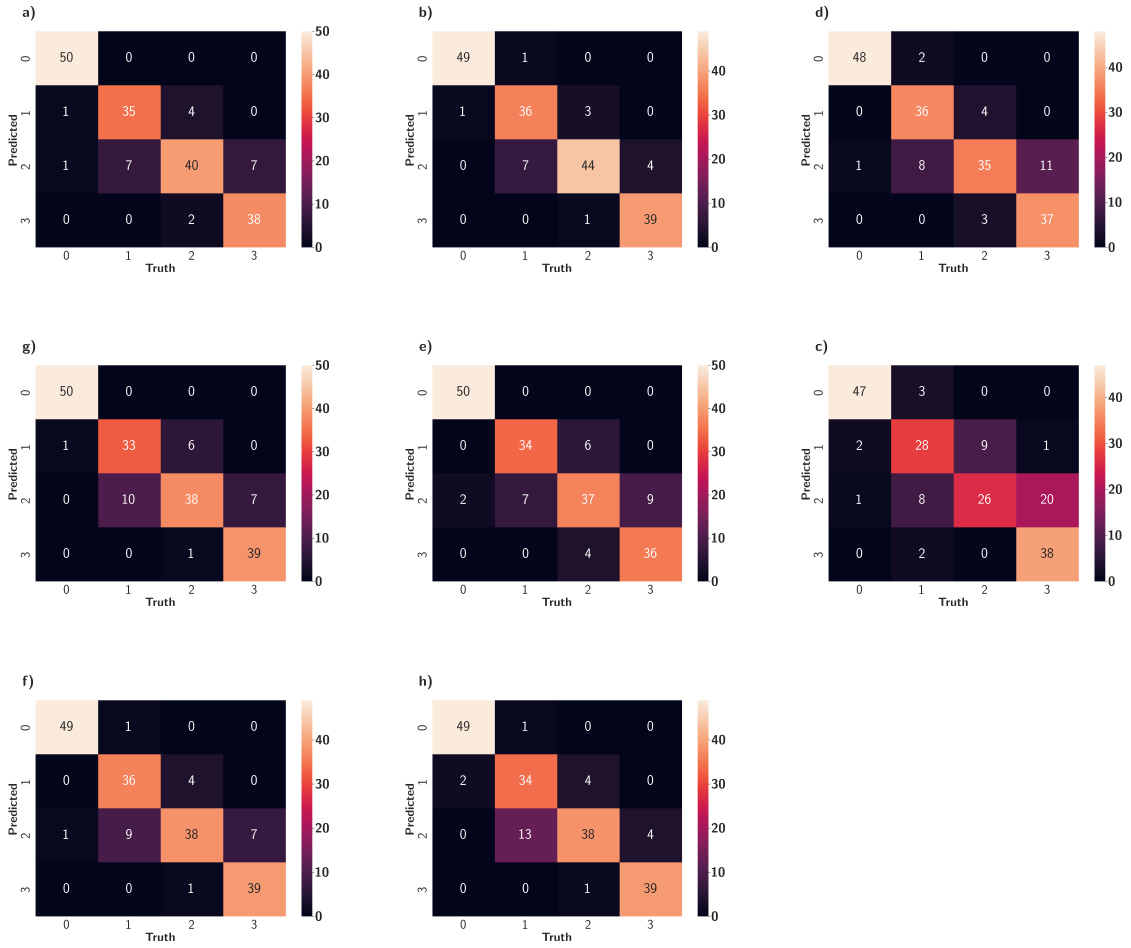

Fig. 5.6 Confusion matrices of the ML models for the second scenario: a) CatBoost classifier, b) Extra Trees Classier, c) Multilayer perceptrons classifier, d) Gradient Boosting classifier, e) LGBM classifier, f) Random Forest classifier, g)KNN classifier, h)SVM classifier

5.7 Discussion

Our study, exploring machine learning applications for water quality assessment, aligns with valuable insights derived from recent literature. The adopted feature selection strategy resonates with findings from studies such as Sadat-Noori et al. (2014) work , which underscores the preponderant role of chloride in driving water quality changes. Similarly, Nasir et al. (2022) emphasized total hardness (TH) as a significant contributor to water quality, corroborating our focus on this parameter. The utilization of LightGBM, CatBoost, Extra Trees, and Random Forest classifiers in our study showcased the efficiency and accuracy of these models in various contexts, echoing similar observations found in the literature. In the realm of groundwater quality prediction, Kumar (2022) geostatistical analyses, employing

gradient boosting and extra trees classifiers, align with our use of Extra Trees for multiclass classification. Shrivastava et al. (2022) comparative analysis further supports our choice of Extra Trees and Random Forest for efficient groundwater classification in multiclass scenarios. Furthermore, Nasir et al. (2022) findings, highlighting CATBOOST as highly predictive, reinforce our conclusion about the efficacy of CATBOOST in achieving high classification accuracy. These consistent observations across various studies strengthen the reliability and generalizability of our machine learning-based approach for groundwater quality assessment tailored specifically for irrigation purposes.

While our study aligns with existing literature, there are notable distinctions that set us apart from prior research endeavors. Unlike the work conducted by Dezfooli et al. (2018), which achieved high accuracy (90.70%) with only three water quality parameters using a Probabilistic Neural Network (PNN), our study focuses on physiochemical parameters with an emphasis on an economically oriented approach. This deliberate choice of a subset of parameters aims to reduce costs and labor associated with water quality assessment. In contrast to studies that incorporate biological parameters, such as fecal coliform, in their comprehensive analyses, our research strategically prioritizes physiochemical parameters. Despite the potential for higher accuracy achieved through the inclusion of additional parameters, our emphasis on economic considerations sets us on a distinct path. Our approach prioritizes efficiency and cost-effectiveness in water quality suitability assessment, aligning with the practical constraints often encountered in real-world applications. This distinct facet of our research contributes to the feasibility and scalability of implementing machine learning techniques in groundwater quality assessment tailored for irrigation purposes (Zegaar et al., 2023).

5.8 Conclusion

In the exploration of groundwater quality assessment for irrigation purposes, this section delved into the comprehensive evaluation of machine learning models, utilizing a dataset spanning the years 2018 to 2022 from various locations within the Msila region. Anchored in the Irrigation Water Quality Index (IWQI), the benchmark for classification, a diverse ensemble of machine learning models underwent meticulous scrutiny, employing robust performance metrics including ROC-AUC, precision-recall, F1 score, and accuracy.

The array of machine learning models considered encompassed the LGBM classifier, CatBoost, Extra Trees, Random Forest, Gradient Boosting classifiers, Support Vector Machines, Multi-Layer Perceptrons, and the K-Nearest Neighbors Algorithm. The evaluation revealed that the models exhibited satisfactory performance, with the LGBM classifier emerging as

the top-performing model, achieving an impressive 91.08% F1 score. Notably, this model utilized six inputs—conductivity, chloride (Cl^-), bicarbonate ($HCO_3{}^-$), sodium (Na^+), calcium (Ca^{2+}), and magnesium (Mg^{2+})—to calculate IWQI.

In an appreciable stride towards model optimization, this research applied the mutual information technique, resulting in a reduction of input parameters to three crucial factors: total hardness (TH), chloride (Cl^-), and sulfate ($SO_4{}^{2-}$). Within this streamlined framework, the Extra Trees classifier emerged as the optimal model, achieving a notable 86.30% F1 score.

The implications of these findings are profound, introducing an ML model capable of accurately simulating IWQI while presenting a streamlined and economically viable approach to groundwater quality classification. The application of this model holds the potential to significantly reduce the time and effort invested in water quality assessment, making it a valuable tool for real-time monitoring. This is particularly beneficial for farmers and decision-makers engaged in water resource management. In essence, this study contributes to a paradigm shift towards efficient and economically feasible water quality assessment methodologies, promising widespread applicability and adoption across diverse environmental contexts.

Chapter 6

Interpretable ML for irrigation water Quality Prediction

6.1 Introduction

The study now pivots towards the subsequent prediction task. Having identified key interdependencies among irrigation water quality parameters and established a robust classification framework, the focus shifts toward predictive modeling. In addressing the challenges associated with predicting groundwater quality, this section proposes a sophisticated and interpretable machine learning approach. Our methodology incorporates state-of-the-art algorithms, including XGBoost, Random Forest, GradientBoost, and CatBoost regressors, to craft predictive models for assessing groundwater quality. The utilization of the Shapley Additive Explanations (SHAP) method provides valuable insights into the contributions of water quality parameters to the Irrigation Water Quality Index (IWQI). To rigorously assess the accuracy of our predictive models, key performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination ($\mathbb{R}^2$) will be employed. Additionally, feature engineering techniques, specifically Recursive Feature Elimination with Cross-Validation (RFECV) and Permutation Importance (PI), will be leveraged to optimize model performance and facilitate insightful feature selection. The primary objective is to contribute to sustainable irrigation management by providing data-driven insights. The outcomes of this investigation are expected to inform strategies for optimizing irrigation practices, ensuring regional food security, and promoting responsible water resource management. As such, the ensuing results hold the promise of aiding policymakers, agricultural stakeholders, and water resource managers in making informed decisions for the betterment of the region's water-related practices.

6.2 Correlation Analysis

The outcomes of the correlation analysis under the new settings unveiled diverse magnitudes of correlation among these parameters, offering insights into their respective impacts on groundwater quality tailored for irrigation purposes (Fig. 6.1). Noteworthy among these correlations is the moderately positive association (0.534) observed between Total Alkalinity (TAC) and IWQI. This correlation implies that maintaining TAC levels within recommended thresholds is linked to enhanced groundwater quality suitable for irrigation, as TAC plays a pivotal role in pH control, establishing a stable environment conducive to optimal crop growth.

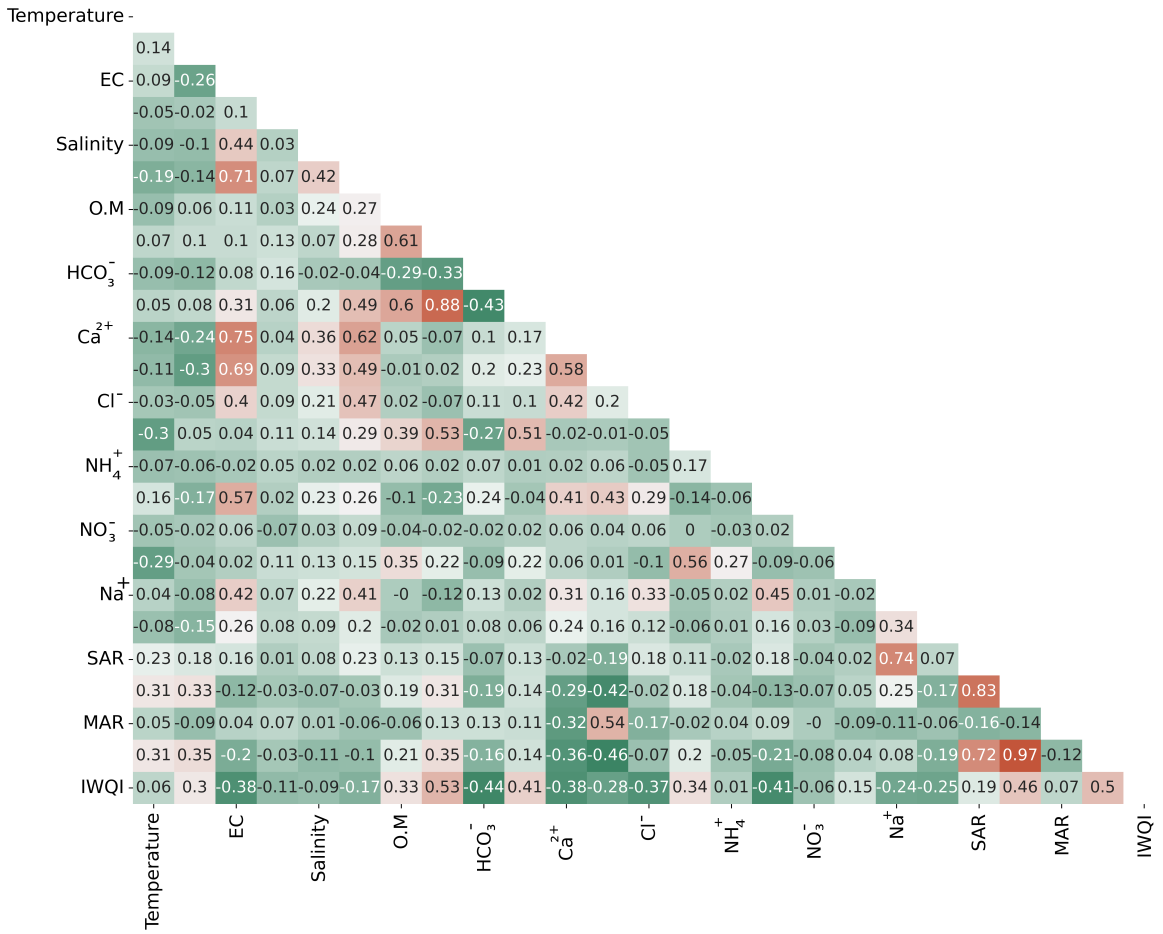

Fig. 6.1 Heatmap of correlation values

Similarly, Kelly's Ratio (KR) manifested a moderate positive correlation (0.462) with IWQI. This finding suggests that a higher KR corresponds to a favorable sodium-calcium ratio, positively influencing soil permeability and, consequently, overall water quality suitability for irrigation. Furthermore, the Permeability Index (PI) exhibited a moderate positive correlation

with IWQI (0.504), indicating that elevated PI values are associated with a reduced risk of water-induced soil degradation. This correlation underscores the importance of PI in enhancing groundwater quality for agricultural use. Conversely, Electrical Conductivity (EC) demonstrated a moderate negative correlation (-0.378) with IWQI. Lower EC values, indicative of reduced dissolved salt levels, prove advantageous for preserving soil structure and mitigating the risk of salinization. Additionally, bicarbonates ($HCO_3{}^-$) showcased a moderate negative correlation (-0.445) with IWQI, suggesting that lower bicarbonate levels contribute to improved irrigation water quality. Moreover, Sulfates ($SO_4{}^{2-}$) displayed a moderate negative correlation (-0.406) with IWQI, signifying that diminished sulfate concentrations lead to enhanced groundwater quality for irrigation, mitigating potential adverse effects on soil and crop health.

The weak positive correlation of 0.3046 with pH suggests that higher pH levels are somewhat correlated with an increase in IWQI. In contrast, weak negative correlations are evident with electrical conductivity (EC), turbidity, total dissolved solids (TDS), calcium ions (Ca^{2+}), magnesium (Mg^{2+}), and sodium (Na^+), implying that elevated values of these parameters may be linked to a decrease in IWQI. Notably, IWQI does not display a straightforward linear relationship with temperature, salinity, ammonium ($NH_4{}^+$), nitrate ($NO_3{}^-$), and magnesium adsorption ratio (MAR). These observations align with existing irrigation water quality knowledge, highlighting the significant role of specific water quality parameters in assessing groundwater suitability for irrigation. It is imperative to emphasize, however, that correlation analysis alone cannot serve as a standalone method for feature selection. Correlation measures solely account for linear connections between variables and overlook complex interactions or non-linear dependencies. To ensure robust feature selection and a comprehensive understanding of the relative importance of each water quality parameter, a broader analysis should incorporate additional feature selection methods such as Recursive Feature Elimination with Cross-Validation (RFECV) or Permutation Importance (PI). These supplementary methodologies contribute to the development of more accurate and reliable predictive models for groundwater quality assessment in irrigation systems by providing a thorough evaluation of feature relevance.

6.3 Feature selection

In the process of feature selection, the Recursive Feature Elimination with Cross-Validation (RFECV) technique was implemented to identify the optimal number of water quality parameters for optimizing the Random Forest model's performance. The RFECV analysis demonstrated that the Random Forest model achieved its peak performance with 7 parameters

(Fig. 6.2), signifying the initial consideration of a larger set of features for predicting the Irrigation Water Quality Index (IWQI).

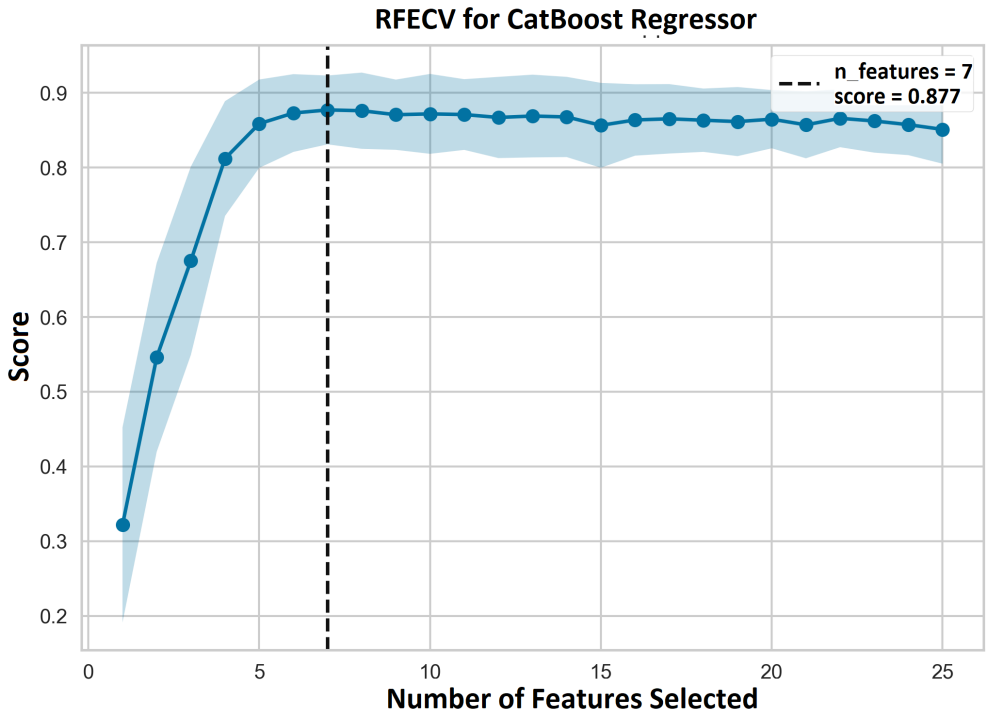

Fig. 6.2 Recursive feature elimination with cross-validation plot

To further refine feature selection and ensure the inclusion of the most relevant parameters, additional methods were employed. The permutation importance analysis results, insights from correlation analysis, and domain knowledge of irrigation water quality collectively played a crucial role in the final selection of the most influential features. Upon thorough consideration, 6 parameters were identified as the optimum inputs for the predictive model. These parameters were chosen based on their substantial positive or negative correlations with IWQI and their individual importance in the permutation importance analysis. Domain knowledge further provided insights into the practical relevance and impact of these features on groundwater quality for irrigation. The integration of RFECV, permutation importance, and correlation analysis with domain knowledge in the feature selection process resulted in a refined set of input parameters significantly contributing to the accuracy and interpretability of the predictive model. The utilization of these 6 key parameters enhances the model's capacity to capture essential factors influencing groundwater quality for irrigation. This optimized set of features ensures a more robust and reliable model, facilitating improved decision-

making in water resource management and promoting sustainable agricultural practices. It is crucial to emphasize that the incorporation of multiple feature selection techniques and domain knowledge is paramount for obtaining a comprehensive understanding of feature importance. By considering results from diverse methods, potential limitations inherent in individual approaches are overcome, leading to a more informed and well-rounded feature selection process. This approach ensures that the selected features align with both statistical significance and practical significance in the context of irrigation water quality assessment.

6.4 Results and discussion

6.4.0.1 Models performances

The evaluation of machine learning model performance in predicting the irrigation water quality index (IWQI) encompassed three key measures: R-squared (score), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), as detailed in Table 6.1. These metrics collectively offer a comprehensive assessment of the models' ability to capture variability in IWQI and the accuracy of their predictions.

Table 6.1 Performances of the regressors

Regressors	MAE	RMSE	R^2
Catboost	1.789	2.911	0.871
XGBoost	1.963	3.159	0.851
GBoost	2.085	3.209	0.852
Random forests	2.452	3.714	0.797

The optimal hyperparameters utilized to achieve the highest model performance are presented in Table 6.2.

Among the tested models, the CatBoostRegressor exhibited superior performance, attaining a high R-squared score of 0.871, accompanied by a standard deviation of 0.056. This outcome implies that approximately 87.1% of the variance in IWQI can be elucidated by the model's predictions. Furthermore, the CatBoostRegressor demonstrated the lowest MAE, featuring a mean value of 1.789 and a standard deviation of 0.326. This indicates that, on average, the model's predictions deviated by approximately 1.789 units from the actual IWQI values. Additionally, the CatBoostRegressor displayed the smallest RMSE, featuring a mean value of 2.911 and a standard deviation of 2.911. The reduced RMSE underscores the proximity of the model's predictions to the true IWQI values, thereby enhancing the accuracy

Table 6.2 Models hyperparameters

Random Forest	GBoost	XGBoost	Catboost
n_estimators= 1000	n_estimators= 1100	base_score=0.5	iterations=1000
max_depth= 50	min_samples_split= 2	grow_policy='depthwise'	learning_rate=0.05
min_samples_split= 2	learning_rate= 0.055	max_cat_threshold=64	
min_samples_leaf= 1	min_samples_leaf= 2	min_child_weight=1	
	max_depth=5	n_estimators=1100	
	loss= 'lad'		
	max_depth= 3		
		learning_rate=0.2	
		booster='gbtree'	
		gamma=0.0463	

The acronyms in the Table stand for: lad: least absolute deviation

of groundwater quality assessments. The XGBRegressor and GradientBoostingRegressor models exhibited commendable performance, both yielding equivalent R-squared scores of approximately 0.851. In terms of Mean Absolute Error (MAE), the XGBRegressor and GradientBoostingRegressor demonstrated closely aligned values, approximately 1.963 and 2.085, respectively. This proximity in average prediction deviation indicates a similar accuracy in predicting the irrigation water quality index (IWQI) values. Furthermore, the Root Mean Squared Error (RMSE) values for the XGBRegressor and GradientBoostingRegressor models were also comparable, with mean values of approximately 3.159 and 3.209, respectively. These findings underscore the models' efficacy in generating accurate predictions for groundwater quality assessment in the context of irrigation. Finally, the performance of the RandomForestRegressor model was marginally lower when compared to the other models, manifesting an R-squared score of approximately 0.797. The Mean Absolute Error (MAE) for the RandomForestRegressor stood at around 2.452, and the Root Mean Squared Error (RMSE) was approximately 3.714. While these metrics suggest a relatively satisfactory predictive capability, the model exhibited a slightly diminished performance when juxtaposed with the other three models under scrutiny. In summary, the CatBoostRegressor model demonstrated superior predictive prowess compared to the XGBRegressor and GradientBoostingRegressor models. Despite a noticeable decline in performance, the RandomForestRegressor still provided reasonably accurate estimations of groundwater quality for irrigation. These outcomes endorse the utilization of machine learning models for the assessment and prediction of irrigation water quality index values based on water quality parameters. The notable efficacy of these models enhances our capacity to make informed

decisions regarding water resource management and agricultural practices. However, it is crucial to consider these findings in the context of the models' interpretability, as accurate predictions are most valuable when the underlying processes influencing groundwater quality are comprehended. The residual and scatter plots are presented in Fig. 6.3 and Fig. 6.4, respectively, offering visual insights into the models' predictive performance and the distribution of residuals.

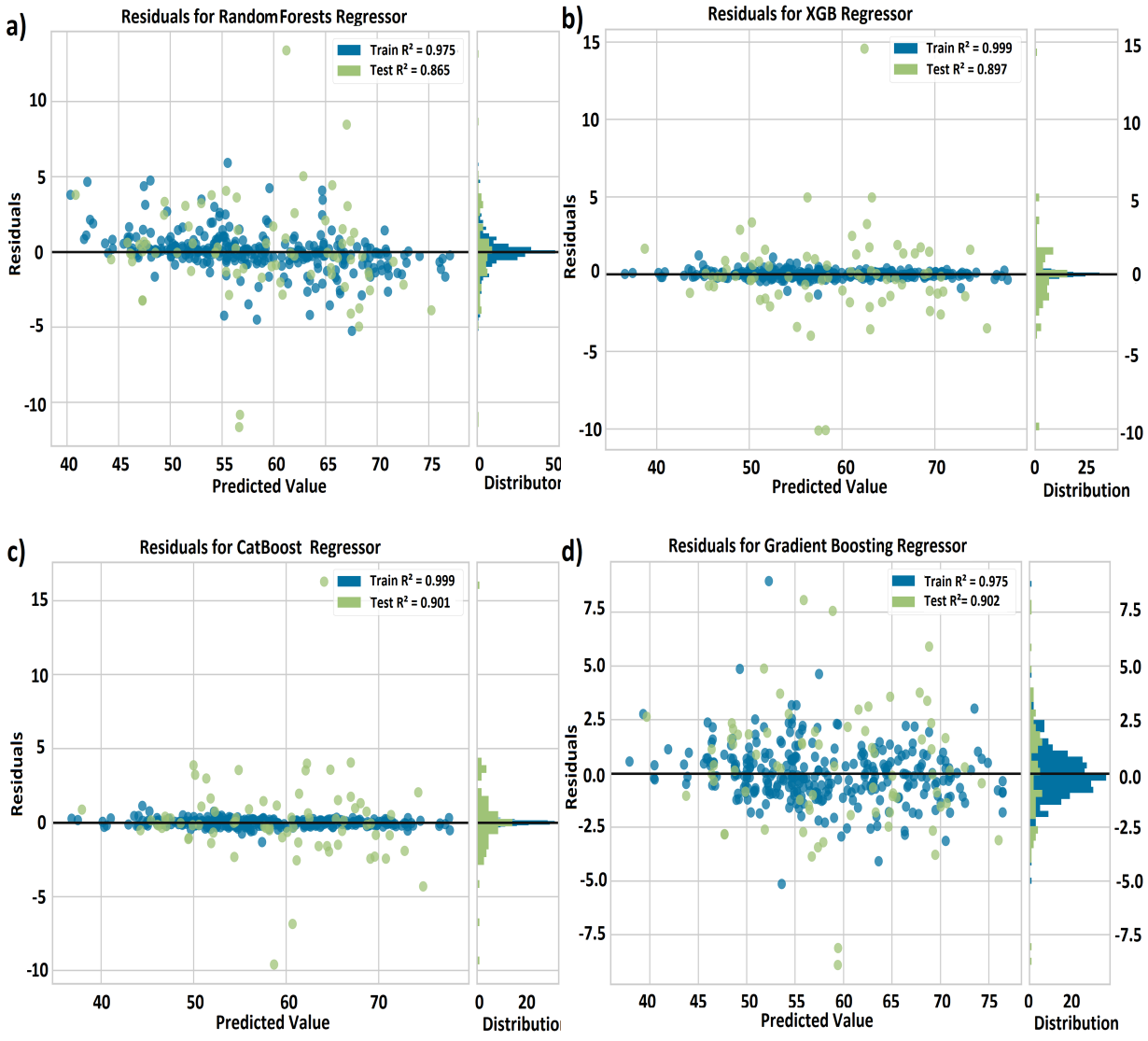

Fig. 6.3 Residual plots of the employed ML models

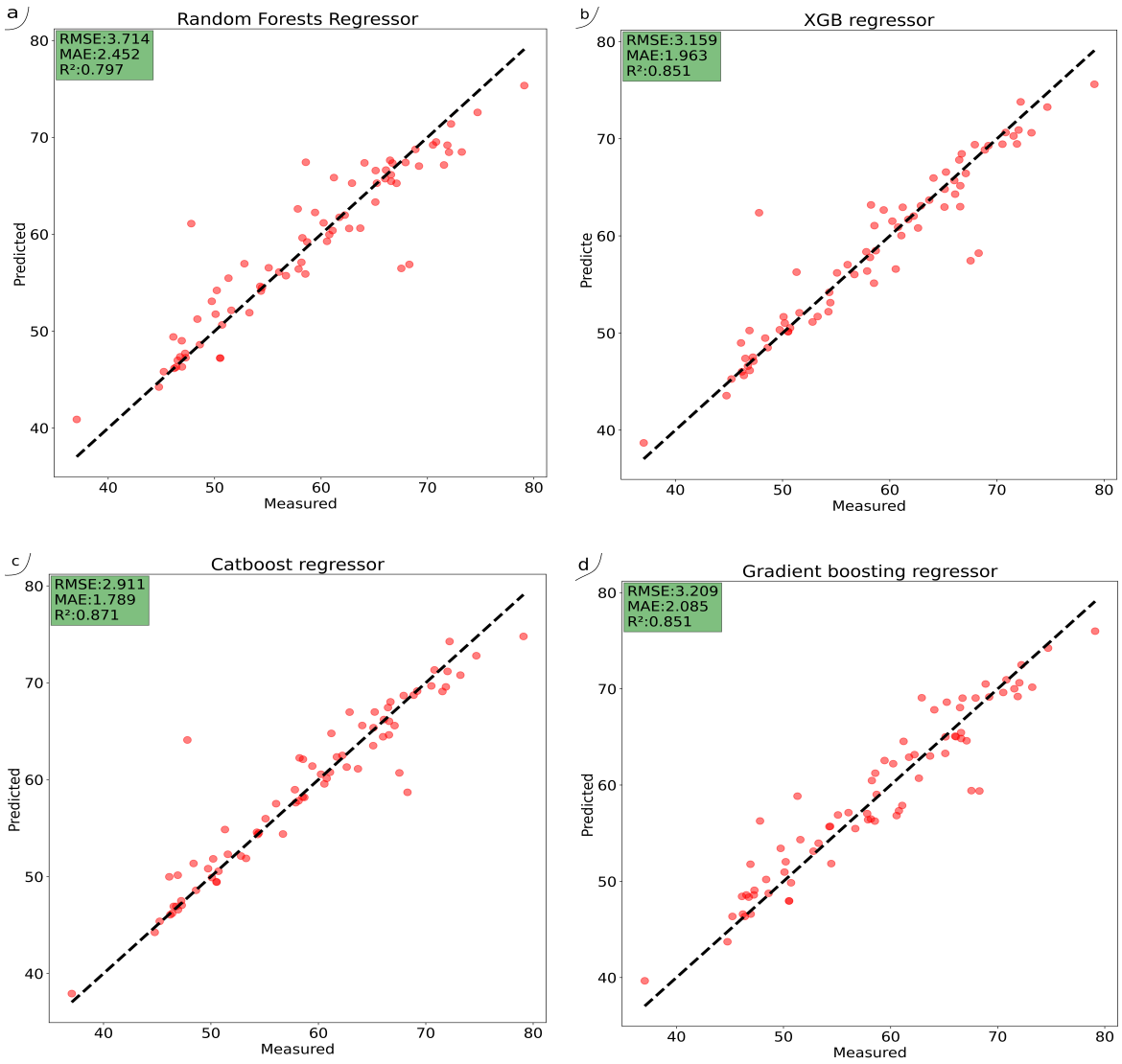

Fig. 6.4 Scatter plots of the employed ML models

6.4.0.2 SHAP values interpretation

This section facilitates a comparative analysis between the outcomes derived from the SHAP analysis and those obtained through correlation analysis. While correlation analysis provides a comprehensive overview of inter-variable relationships, the examination of SHAP values delves into more nuanced and model-specific insights. SHAP values furnish valuable indicators for assessing machine learning models, offering a detailed understanding of the influential water quality parameters in predicting the Irrigation Water Quality Index (IWQI) by analyzing the impact of distinct features on the model's predictions. The analysis of feature importance utilizing mean SHAP values elucidates the relative significance of water

quality parameters in predicting the IWQI (Fig. 6.5). Notably, Total Alkalinity (TAC) emerges as the most crucial feature, boasting a mean SHAP value of 2.7, underscoring its pronounced positive influence on the model's predictions. Electrical Conductivity (EC) and Chloride (Cl^-) closely follow, with mean SHAP values of 2.23 and 2.21, respectively, signifying their substantial contributions to the model's output. In contrast, Permeability Index (PI) and Sodium (Na^+) exhibit relatively lower importance, with mean SHAP values of 1.24 and 1.23, respectively.

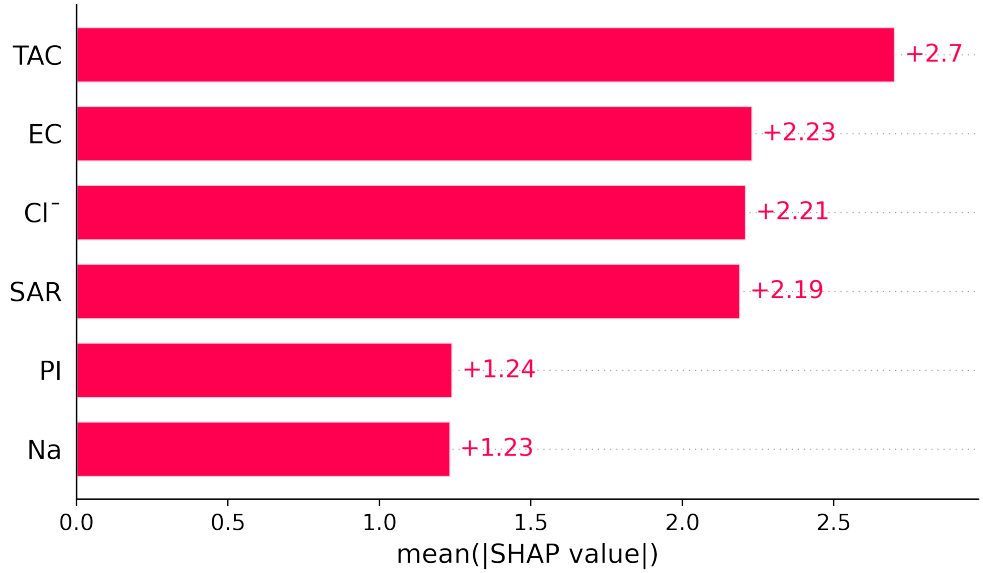

Fig. 6.5 Feature importance based on SHAP values

Subsequent to the SHAP analysis, the SHAP values for four randomly selected samples from the dataset through Waterfall SHAP plots (Fig. 6.6) were illustrated. These plots serve as visual representations, elucidating the contribution of each feature to the prediction for individual samples, offering a transparent depiction of the decision-making process for specific instances.

Furthermore, the Beeswarm SHAP plot is presented in (Fig. 6.7), which effectively illustrates the aggregate impact of features on the model's predictions across the entire dataset. This plot provides a holistic perspective on feature importance, accentuating the paramount role of specific water quality parameters in determining the Irrigation Water Quality Index (IWQI).

In the Beeswarm plot, each data point signifies a unique record within the dataset, and the vertical axis corresponds to individual water quality parameters. The SHAP value, serving as a metric for assessing the extent of contribution and influence, is utilized to measure

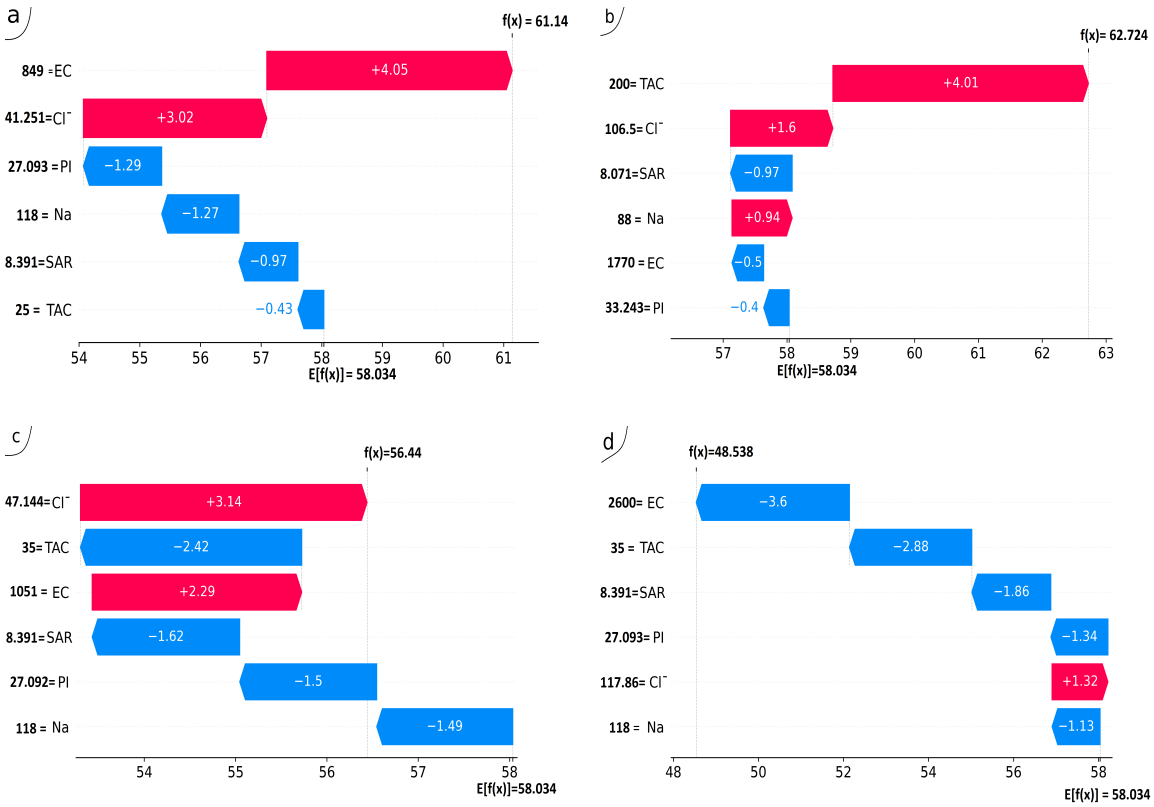

Fig. 6.6 Waterfall plot of 4 random samples

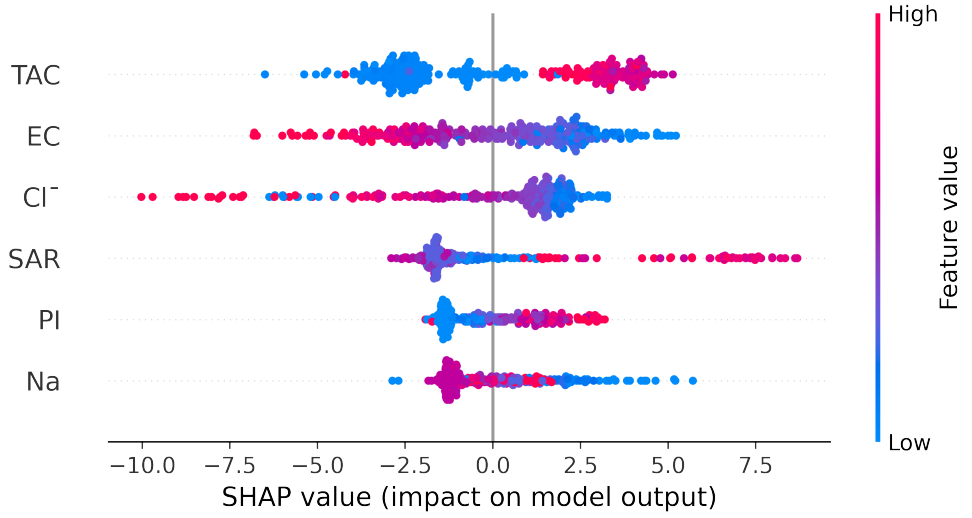

Fig. 6.7 Beeswarm plot of SHAP values

the impact of each feature on the model's prediction for each data point. The color-coded

representation in the Beeswarm plot facilitates the interpretation of results, where positive SHAP values are visually represented on the right side of the axis. This positioning signifies that increasing values of these features have a positive impact on the model's predictions. Conversely, negative SHAP values are depicted on the left side of the axis, indicating that increasing values of these features negatively influence the model's predictions. This visual representation aids in comprehending the directional impact of each water quality parameter on the model's predictions, contributing to a nuanced understanding of the model's decision-making process. The observations derived from the Beeswarm plot underscore the pivotal role of specific water quality parameters in delineating the irrigation water quality index. Particularly, Total Alkalinity (TAC) exhibits a positive influence on the model's predictions, signifying that elevated TAC values contribute to enhanced groundwater quality suitable for irrigation. Conversely, Electrical Conductivity (EC) is identified as having a negative impact on the model's outcomes, implying that heightened EC values may lead to a diminished suitability of groundwater quality for irrigation. Sodium Adsorption Ratio (SAR) emerges as a notably influential parameter with a positive effect on the model's predictions, suggesting that heightened SAR values are correlated with improved groundwater quality for irrigation. Likewise, Permeability Index (PI) manifests a positive influence, indicating that increased PI values are advantageous for maintaining suitable groundwater quality for irrigation purposes. In contrast, Chloride (Cl^-) is observed to exert a highly negative impact on the model's predictions, suggesting that elevated chloride concentrations may adversely affect groundwater quality for irrigation. Notably, Sodium (Na^+) is highlighted as a highly influential parameter with a positive impact, signifying that heightened sodium levels are associated with improved groundwater quality for irrigation purposes. The cautious interpretation of SHAP analysis outcomes is essential, as it may appear to diverge from the conventional understanding of water quality parameters and their impact on groundwater suitability for irrigation. An instance of such a discrepancy is observed in the case of higher Total Alkalinity (TAC) values, traditionally deemed detrimental to water quality due to potential scaling and soil pH issues, yet indicated by SHAP analysis to contribute to enhanced groundwater quality for irrigation. This incongruity can be reconciled by considering the specific contextual factors of the study area and the recommended water quality thresholds tailored for irrigation purposes (Wilcox, 1955). It is plausible that TAC levels within certain defined ranges offer benefits for pH buffering, creating a stable environment conducive to crop growth in the investigated region. Similarly, the findings of SHAP analysis may indicate that elevated Electrical Conductivity (EC) values result in diminished groundwater quality suitability for irrigation. However, this observation necessitates a comparison with established water quality guidelines and thresholds for EC in irrigation water. Depending

on the specific crop varieties and soil characteristics prevalent in the study area, certain EC levels might be deemed acceptable and even beneficial for irrigation purposes (Ayers and Westcot, 1985). This principle extends to other water quality parameters, including Sodium Adsorption Ratio (SAR), Permeability Index (PI), chloride concentrations, and sodium levels, where SHAP analysis may propose varying effects on groundwater quality compared to traditional interpretations. Contextual assessments should be grounded in regional water quality standards, crop-specific requirements, and environmental conditions to ensure judicious groundwater management for sustainable irrigation practices.

The integration of Waterfall and Beeswarm SHAP plots serves to augment the interpretability of the machine learning model, providing valuable insights into the relative importance of each water quality parameter and their collective contributions to the prediction of groundwater quality for irrigation purposes.

6.5 The novelty of the study

In contrast to previous investigations in the literature, our study exhibits several notable strengths that distinguish it from prior endeavors in predicting the Irrigation Water Quality Index (IWQI). Ibrahim et al. (2023) utilized Support Vector Machines (SVM) and Adaptive Neuro-Fuzzy Inference System (ANFIS), employing nine water quality parameters as inputs and achieving RMSE values of 12.45 and 4.54, respectively. Similarly, Gaagai et al. (2023) employed the Gradient Boosting Regressor (GBR) and Artificial Neural Networks (ANN) to predict IWQI, yielding RMSE values of 2.562 and 2.175, respectively. In another study, Yıldız and Karakuş (2020) employed various ANN structures and obtained RMSE values ranging from 1.634 to 5.231 in predicting IWQI. However, our research distinguishes itself by not only employing a diverse array of state-of-the-art machine learning algorithms, including XGBoost, Random Forest, GradientBoost, and CatBoost regressors, but also by incorporating the Shapley Additive Explanations (SHAP) method for enhanced model interpretability. The absence of model interpretability in previous studies represents a significant gap in the existing literature. Through the utilization of SHAP analysis, our research provides valuable insights into the contributions of individual water quality parameters in determining IWQI, thereby advancing our understanding of the intricate relationships between groundwater quality and irrigation suitability. Furthermore, our investigation offers an exhaustive examination of feature engineering methodologies, including Recursive Feature Elimination with Cross-Validation (RFECV) and Permutation Importance (PI), aimed at refining predictive models. This approach facilitates the identification of pertinent features, thereby augmenting model

interpretability and ultimately contributing to a more precise and dependable prediction of the Irrigation Water Quality Index (IWQI) (Zegaar et al., 2024a).

6.6 Conclusion

In summary, this research constitutes a thorough exploration into the prediction of groundwater quality tailored for irrigation purposes. The primary focus involves a nuanced understanding of the importance of water quality parameters facilitated by advanced machine learning methodologies and SHAP analysis. The outcomes divulge pivotal insights into the influence of diverse water quality parameters on the Irrigation Water Quality Index (IWQI), furnishing valuable indicators for the optimization of water resource management strategies and agricultural practices. Through meticulous correlation analysis, we unveiled significant positive and negative correlations between IWQI and specific water quality parameters. Total Alkalinity (TAC), Kelly's Ratio (KR), Permeability Index (PI), and Sodium Adsorption Ratio (SAR) surfaced as pivotal factors exerting positive influences on groundwater quality for irrigation. In contrast, Electrical Conductivity (EC), Chloride (Cl^-), and Sulfates ($SO_4{}^{2-}$) were identified as parameters with potential adverse impacts. This discernment holds promise for advancing our comprehension of the intricate interplay between groundwater quality and its suitability for irrigation, thereby contributing to informed decision-making in the realm of water resource management and sustainable agricultural practices. Our investigation showcases the interpretive capabilities of machine learning models through the application of SHAP analysis, thereby fostering a profound comprehension of the predictive mechanisms inherent in these models. The utilization of Beeswarm SHAP plots offers a visual representation of the impact that individual features exert on model predictions, elucidating the significance of Total Alkalinity (TAC), Electrical Conductivity (EC), Sodium Adsorption Ratio (SAR), Permeability Index (PI), Chloride (Cl^-), and Sodium (Na) in delineating the suitability of groundwater for irrigation. These discernments not only augment our scientific understanding but also furnish decision-makers with valuable insights to prioritize interventions aimed at enhancing water quality and, consequently, elevating agricultural productivity. Nevertheless, it is imperative to recognize the inherent limitations of this study, such as its dependence on a specific dataset and the omission of certain pertinent parameters. To advance the field, future research endeavors should broaden their scope by encompassing diverse geographical regions, integrating additional water quality variables, and undertaking prolonged monitoring initiatives to encapsulate temporal variations comprehensively.

Despite these acknowledged constraints, the study significantly contributes to the domain of groundwater quality prediction for irrigation, effectively bridging the nexus between water

resource management and agricultural sustainability. The insights derived from this research possess the potential to guide the formulation of evidence-based policies and practices, thereby ensuring judicious water utilization, heightened crop yields, and the preservation of invaluable water reservoirs.

In essence, the amalgamation of machine learning models and SHAP analysis propels the attainment of a holistic comprehension of water quality dynamics, empowering stakeholders to make enlightened decisions conducive to sustainable irrigation practices. Embracing these insights and persistently delving into novel frontiers in groundwater research pave the way for a future where agricultural productivity harmonizes seamlessly with conscientious water stewardship, yielding benefits for both human livelihoods and the environment.

Chapter 7

General Conclusion

In conclusion, this comprehensive thesis has navigated the intricate intersection of machine learning and water resources, offering a nuanced exploration into the classification and assessment of groundwater quality for irrigation purposes. The research journey embarked upon a multifaceted exploration, encompassing diverse aspects such as data preprocessing, correlation analysis, feature selection, model development, and performance evaluation. The overarching goal was to develop a robust framework that classifies the irrigation water based on machine learning.

The early chapters of the thesis laid the groundwork by delving into the intricacies of the dataset, spanning the years 2018 to 2022, sourced from various locations within the Msila region. Data preprocessing emerged as a critical step, involving cleansing, normalization, and imputation techniques, ensuring the trustworthiness and accuracy of subsequent modeling endeavors. The classification task, anchored in the Irrigation Water Quality Index (IWQI), became the focal point, where an array of state-of-the-art machine learning models underwent rigorous evaluation. The LGBM classifier emerged as the top-performing model, showcasing its prowess in accurately simulating IWQI.

A novel dimension unfolded with the optimization of model performance through the application of the mutual information technique, streamlining input parameters to three crucial factors: total hardness, chloride, and sulfate. This streamlined framework led to the emergence of the Extra Trees classifier as the optimal model, further emphasizing the adaptability of machine learning methodologies to enhance groundwater quality classification.

The correlation analysis shed light on the interplay of various water quality parameters, unraveling their intricate relationships and implications for groundwater quality. Subsequent feature selection techniques, including Recursive Feature Elimination with Cross-Validation (RFECV) and Permutation Importance (PI), were employed to distill the most influential parameters, enhancing model accuracy and interpretability.

Model performance evaluation was a pivotal aspect, with metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) providing a comprehensive assessment. In the prediction task, the CatBoostRegressor emerged as a standout performer, signifying its efficacy in capturing variability in the IWQI and producing accurate predictions.

The interpretability of machine learning models took center stage with the incorporation of Shapley Additive Explanations (SHAP) analysis. This not only enhanced our understanding of individual parameter contributions but also provided valuable insights into the complex decision-making processes of the models. The Waterfall and Beeswarm SHAP plots visually illustrated these contributions, offering a holistic view of the relative significance of each water quality parameter.

As the thesis advanced into its concluding chapter, the limitations inherent in the study were candidly addressed.

The primary limitations include:

- Geographical Extension: The study focuses mainly on the M'sila state region, limiting generalizability to other areas with potentially different water quality dynamics.
- Under-Representation of IWQI Classes: Sampling bias towards certain IWQI classes may affect classifier performance. Future research should prioritize collecting samples from all IWQI classes.
- Ensemble Learning Models: The study could benefit from exploring more complex models like CNN and RNN to improve classification accuracy, particularly regarding the fifth IWQI class.
- Limited Dataset Reach: Relying on data from 210 wells in M'sila state restricts generalizability. Broader data collection efforts are needed to encompass diverse water quality conditions.
- Incomplete Feature Set: The study's predefined water quality parameters may overlook other influential factors. Including additional parameters is essential for a comprehensive understanding.
- Model Selection: Exploring a broader spectrum of models beyond XGBoost, Gradient Boosting, and CatBoost could yield diverse results and improve accuracy.
- Limited SHAP Analysis: While SHAP analysis enhances interpretability, its efficacy depends on the underlying model's design. Complementary techniques should be explored for a nuanced understanding of model behavior (Kumar et al., 2020).

As we navigate the terrain of groundwater quality assessment with machine learning, recognizing limitations propels us towards future research directions.

The future directions outlined a roadmap for :

- Geographical Expansion: Extending the study beyond the Msila region will enrich understanding of water quality dynamics across diverse climates and geological conditions, enhancing model applicability.
- Inclusion of Additional Parameters: Expanding predictive models to include biological and ecological indicators will provide a more holistic insight into groundwater quality, refining accuracy and depth.
- Long-Term Monitoring Initiatives: Establishing long-term monitoring initiatives will capture temporal variations in groundwater quality, enabling proactive management strategies.
- Collaborative Research Opportunities: Collaborations with experts in hydrology, ecology, and environmental science will enrich groundwater quality assessment through diverse datasets and methodologies.
- Advanced Machine Learning Techniques: Exploring ensemble learning, deep learning, and hybrid models aims to elevate model accuracy and keep pace with evolving methodologies.
- Incorporating Emerging Technologies: Integrating IoT, remote sensing (Zegaar et al., 2024b), and UAVs enhances data acquisition, enabling real-time, high-frequency monitoring for dynamic model development.
- Continuous Model Refinement: Iterative model evaluation, validation, and adaptation ensure relevance and effectiveness in addressing evolving groundwater quality dynamics.

In totality, this thesis stands as a testament to the synergy between machine learning and water resources. It has transcended the boundaries of traditional approaches, offering a holistic framework that not only predicts groundwater quality but also interprets the intricate dynamics governing the process. The contributions made, coupled with the candid acknowledgment of limitations and a visionary outlook toward the future, position this research as a valuable cornerstone in the ongoing quest for sustainable water resource management and agricultural practices.

References

Abdel-Fattah, M. K., Abd-Elmabod, S. K., Aldosari, A. A., Elrys, A. S., and Mohamed, E. S. (2020). Multivariate analysis for assessing irrigation water quality: A case study of the bahr mouise canal, eastern nile delta. *Water*, 12:2537.

Abuzir, S. Y. and Abuzir, Y. S. (2022). Machine learning for water quality classification. *Water Quality Research Journal*, 57.

Agrawal, P., Sinha, A., Kumar, S., Agarwal, A., Banerjee, A., Villuri, V. G. K., Annavarapu, C. S. R., Dwivedi, R., Dera, V. V. R., Sinha, J., and Pasupuleti, S. (2021). Exploring artificial intelligence techniques for groundwater quality assessment. *Water (Switzerland)*, 13.

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M., and Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, 578:124084.

Allison, L. and Richards, L. A. (1954). *Diagnosis and improvement of saline and alkali soils*. Number 60. Soil and Water Conservative Research Branch, Agricultural Research Service ....

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, 9.

Ayers, R. and Westcot, D. (1985). Water quality for agriculture. *FAO of the UNITED NATIONS,Rome,italy*, page 97.

Ayers, R. S., Westcot, D. W., et al. (1985). *Water quality for agriculture*, volume 29. Food and agriculture organization of the United Nations Rome.

Bansal, S. and Geetha, G. (2020). A machine learning approach towards automatic water quality monitoring. *Journal of Water Chemistry and Technology*, 42:321–328.

Bedi, S., Samal, A., Ray, C., and Snow, D. (2020). Comparative evaluation of machine learning models for groundwater quality assessment. *Environmental Monitoring and Assessment*, 192.

Biecek, P. and Burzykowski, T. (2021). *Explanatory model analysis*.

Bilali, A. E. and Taleb, A. (2020). Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. *Journal of the Saudi Society of Agricultural Sciences*, 19:439–451.

Bilali, A. E., Taleb, A., and Brouziyne, Y. (2021). Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agricultural Water Management*, 245:106625.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brown, R. M., McClelland, N. I., Deininger, R. A., and Tozer, R. G. R. G. (1970). A water quality index-do we dare. *Water and sewage works*, 117:339–343.

Chakravarty, T. and Gupta, S. (2021). Assessment of water quality of a hilly river of south assam, north east india using water quality index and multivariate statistical analysis. *Environmental Challenges*, 5:100392.

Chan, Y. (2003). Biostatistics 104: correlational analysis. *Singapore Med J*, 44(12):614–619.

Chandra, D. S., Asadi, S. S., and Raju, M. V. (2017). Estimation of water quality index by weighted arithmetic water quality index method: A model study. *International Journal of Civil Engineering and Technology*, 8:1215–1222.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. volume 13-17-August-2016.

Dancey, C. P. and Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

De Paz, J., Visconti, F., Zapata, R., and Sánchez, J. (2004). Integration of two simple models in a geographical information system to evaluate salinization risk in irrigated land of the valencian community, spain. *Soil Use and Management*, 20:333–342.

Deng, H., Zhou, Y., Wang, L., and Zhang, C. (2021). Ensemble learning for the early prediction of neonatal jaundice with genetic features. *BMC Medical Informatics and Decision Making*, 21.

Dezfooli, D., Hosseini-Moghari, S.-M., Ebrahimi, K., and Araghinejad, S. (2018). Classification of water quality status based on minimum quality parameters: application of machine learning techniques. *Modeling Earth Systems and Environment*, 4:311–324.

Doneen, L. D. (1954). Salination of soil by salts in the irrigation water. *Eos, Transactions American Geophysical Union*, 35(6):943–950.

Doneen, L. D. (1964). Water quality for agriculture. *Department of Irrigation, University of California, California*, 48.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*.

Eaton, F. M. (1950). Significance of carbonates in irrigation waters. *Soil science*, 69:123–134.

EPA Gold Book (1986). Quality criteria for water. *American Fisheries Society: Bethesda, MD, USA*.

Ewaid, S. H., Abed, S. A., and Kadhum, S. A. (2018). Predicting the tigris river water quality within baghdad, iraq by using water quality index and regression analysis. *Environmental Technology & Innovation*, 11:390–398.

Ewaid, S. H., Kadhum, S. A., Abed, S. A., and Salih, R. M. (2019). Development and evaluation of irrigation water quality guide using iwqg v.1 software: A case study of al-gharraf canal, southern iraq. *Environmental Technology & Innovation*, 13:224–232.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gaagai, A., Aouissi, H. A., Bencedira, S., Hinge, G., Athamena, A., Heddam, S., Gad, M., Elsherbiny, O., Elsayed, S., and Eid, M. H. (2023). Application of water quality indices, machine learning approaches, and gis to identify groundwater quality for irrigation purposes: a case study of sahara aquifer, doucen plain, algeria. *Water*, 15:289.

García, S., Luengo, J., and Herrera, F. (2014). *Data preprocessing in data mining*.

Gupta, D. and Mishra, V. K. (2023). Development of entropy-river water quality index for predicting water quality classification through machine learning approach. *Stochastic Environmental Research and Risk Assessment*.

Ibrahim, H., Yaseen, Z. M., Scholz, M., Ali, M., Gad, M., Elsayed, S., Khadr, M., Hussein, H., Ibrahim, H. H., Eid, M. H., Kovács, A., Péter, S., and Khalifa, M. M. (2023). Evaluation and prediction of groundwater quality for irrigation using an integrated water quality indices, machine learning models and gis approaches: A representative case study. *Water (Switzerland)*, 15.

Irfeey, A. M. M., Najim, M. M., Alotaibi, B. A., and Traore, A. (2023). Groundwater pollution impact on food security. *Sustainability (Switzerland)*, 15.

Jahin, H. S., Abuzaid, A. S., and Abdellatif, A. D. (2020). Using multivariate analysis to develop irrigation water quality index for surface water in kafr el-sheikh governorate, egypt. *Environmental Technology & Innovation*, 17:100532.

Javaid, M., Haleem, A., Khan, I. H., and Suman, R. (2023). Understanding the potential applications of artificial intelligence in agriculture sector. *Advanced Agrochem*, 2.

Kelley, W. (1940). Permissible composition and concentration of irrigation water.

Konikow, L. F. and Kendy, E. (2005). Groundwater depletion: A global problem. *Hydrogeology Journal*, 13.

Krishnan, S. R., Nallakaruppan, M. K., Chengoden, R., Koppu, S., Iyapparaja, M., Sadhasivam, J., and Sethuraman, S. (2022). Smart water resource management using artificial intelligence—a review. *Sustainability (Switzerland)*, 14.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. (2020). Problems with shapley-value-based explanations as feature importance measures. volume PartF168147-8.

Kumar, M. J. (2022). Geostatistical analyses empowered with gradient boosting and extra trees classifier algorithms in the prediction of groundwater quality and geology-lithology attributes over ysr district, india. *International Journal of Hydrology Science and Technology*, 1.

Landwehr, J. M. and Deininger, R. A. (1976). A comparison of several water quality indexes. *Journal (Water Pollution Control Federation)*, 48:954–958.

Lu, H. and Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, 249:126169.

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. volume 2017-December.

Mancosu, N., Snyder, R. L., Kyriakakis, G., and Spano, D. (2015). Water scarcity and future challenges for food production. *Water (Switzerland)*, 7.

Meireles, A. C. M., de Andrade, E. M., Chaves, L. C. G., Frischkorn, H., and Crisostomo, L. A. (2010). A new proposal of the classification of irrigation water. *Revista Ciência Agronômica*, 41:349–357.

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

Ouhamdouch, S., Bahir, M., Ouazar, D., Carreira, P. M., and Zouari, K. (2019). Evaluation of climate change impact on groundwater from semi-arid environment (essaouira basin, morocco) using integrated approaches. *Environmental Earth Sciences*, 78.

Piper, A. M. (1944). A graphic procedure in the geochemical interpretation of water-analyses. *Eos, Transactions American Geophysical Union*, 25:914–928.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. volume 2018-December.

Raghunath, H. M. (1987). *Ground water: hydrogeology, ground water survey and pumping tests, rural water supply and irrigation systems*. New Age International.

Raheja, H., Goel, A., and Pal, M. (2022). Prediction of groundwater quality indices using machine learning algorithms. *Water Practice and Technology*, 17.

Rahman, K., Barua, S., and Imran, H. (2021). Assessment of water quality and apportionment of pollution sources of an urban lake using multivariate statistical analysis. *Cleaner Engineering and Technology*, 5:100309. this paper give a decent details about PCANSF CCME WQIPMFit's worth to revise later.

Ravindran, S. M., Bhaskaran, S. K. M., and Ambat, S. K. N. (2021). A deep neural network architecture to model reference evapotranspiration using a single input meteorological parameter. *Environmental Processes*, 8.

Richards, L. A. (1954). *Diagnosis and improvement of saline and alkali soils*. Number 60. US Government Printing Office.

Sadat-Noori, S. M., Ebrahimi, K., and Liaghat, A. M. (2014). Groundwater quality assessment using the water quality index and gis in saveh-nobaran aquifer, iran. *Environmental Earth Sciences*, 71.

Shrivastava, A., Sahu, M., and Jhariya, D. C. (2022). Comparative analysis on ensemble learning techniques for groundwater quality assessment of chhattisgarh region.

Siebert, S., Burke, J., Faures, J. M., Frenken, K., Hoogeveen, J., Döll, P., and Portmann, F. T. (2010). Groundwater use for irrigation - a global inventory. *Hydrology and Earth System Sciences*, 14.

Singh, S., Ghosh, N. C., Gurjar, S., Krishan, G., Kumar, S., and Berwal, P. (2018). Index-based assessment of suitability of water quality for irrigation purpose under indian conditions. *Environmental Monitoring and Assessment*, 190:29.

Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., and Kumar, S. (2021). Prediction of groundwater quality using efficient machine learning technique. *Chemosphere*, 276.

Sinha Ray, S. P. and Elango, L. (2019). *Deterioration of Groundwater Quality: Implications and Management*, pages 87–101. Springer Singapore, Singapore.

Sudhakaran, S., Mahadevan, H., Arun, V., Krishnakumar, A. P., and Krishnan, K. A. (2020). A multivariate statistical approach in assessing the quality of potable and irrigation water environs of the netravati river basin (india). *Groundwater for Sustainable Development*, 11:100462.

Tyagi, S., Sharma, B., Singh, P., and Dobhal, R. (2020). Water quality assessment in terms of water quality index. *American Journal of Water Resources*, 1:34–38.

Uddin, M. G., Nash, S., and Olbert, A. I. (2021a). A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, 122:107218.

Uddin, M. G., Nash, S., and Olbert, A. I. (2021b). A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, 122:107218.

Vapnik, V., Golowich, S., and Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. *Advances in neural information processing systems*, 9.

Vrba, J. (1983). The impact of human activities on groundwater systems programs of the international association of hydrogeologists. *Environmental Geology*, 5:9–9.

Wilcox, L. V. (1955). Classification and use of irrigation waters. *United States Department of Agriculture*, Circular N.

Xia, F., Jiang, D., Kong, L., Zhou, Y., Wei, J., Ding, D., Chen, Y., Wang, G., and Deng, S. (2022). Prediction of dichloroethene concentration in the groundwater of a contaminated site using xgboost and lstm. *International Journal of Environmental Research and Public Health*, 19.

Yıldız, S. and Karakuş, C. B. (2020). Estimation of irrigation water quality index with development of an optimum model: a case study. *Environment, Development and Sustainability*, 22:4771–4786.

Zegaar, A., Ounoki, S., and Telli, A. (2023). Machine Learning for Groundwater Quality Classification: A step towards Economic and sustainable groundwater quality Assessment Process. *Water Resources Management*, 38(2):621–637.

Zegaar, A., Telli, A., Ounoki, S., and Shahabi, H. (2024a). Interpretable machine learning models for irrigation sustainability: Groundwater quality Prediction in M'Sila, Algeria. *Environmental Modeling and Assessment*.

Zegaar, A., Telli, A., Ounoki, S., Shahabi, H., and Rueda, F. (2024b). Data-driven approach for land surface temperature retrieval with machine learning and sentinel-2 data. *Remote Sensing Applications: Society and Environment*, 36:101357.