PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

Ministry of Higher Education and Scientific Research

University Hassiba Benbouali of Chlef

Faculty of Civil Engineering and Architecture

Department of Hydraulics

END OF STUDY Dissertation

To obtain the diploma of

Master

Field: Hydraulics.
Specialty: Urban Hydraulics

By

Mounir HABBEL

Groundwater quality classification using machine learning and cluster analysis ;case of Upper and Middle Cheliff plains

Dissertation defense :	28/06/2025/	in front of the jury composed of	
Hamid Benaouda	Doctor	Hassiba Benbouali University of Chlef	President
Sofiane Abaidia	Doctor	Hassiba Benbouali University of Chlef	Examiner
Yamina Elmeddahi	Professor	Hassiba Benbouali University of Chlef	Supervisor

Academic year: 2024/2025

Acknowledgement

Completing this thesis, a product of several years work, I feel deeply indebted to a great Many people who have greatly inspired and supported me during my study.

First, I express my fully gratitude to Allah the almighty for protecting me and for giving me the ability to do this work

My sincere gratitude my supervisor, professor, Elmeddahi yamina, who not only encouraged me but also guided me through the journey , I will always remember your encouragement to go ahead, when I was hesitant to move forward during writing due to certain immature ideas. I take full responsibility for any shortcomings in this thesis.

Furthermore, I would like to express my gratitude to my co-directors; all my professors. For their securities. Insights and contributions have greatly enriched this research.

Words cannot express the feelings I have for my parents, my Dad ; Mr. Mustapha and my MAMA ;Mrs. Batache Aicha , who formed part of my vision and taught me good things that really matter in life. Their infallible love and support has always been my strength

Finally, I thank my family in Chlef as well as the entire Habbel and Batache families, for their encouragement and prayers during the course of my research.

Best regards,

mounir

Mounir Habbel

Dedication

To my mom and Dad , family and friends

Not enough word or work could give you or repay you , i will just stick with a big thank you and everything else is in my heart .

i hope i was the son and the friend and student that you ever wanted me to be .

this work is dedicated to you all .

Best regards,

Habbel Mounir

Abstract

الملخص:

تلعب مراقبة جودة المياه الجوفية دورًا حيويًا في الإدارة المستدامة للموارد المائية واتخاذ القرارات البيئية المستنيرة. تستكشف هذه الدراسة استخدام تقنيات التعلم الآلي والتجزئة لتصنيف جودة المياه الجوفية في سهل الشلف الأعلى والأوسط، وذلك استنادًا إلى عشرة معايير وفقًا لمعايير منظمة الصحة العالمية (2017). تم استخدام خوارزميات آلة الدعم الناقل(SVM) ، والتعزيز التدريجي المتطرف(XGBoost) ، بالإضافة إلى خوارزمية التجميع-K means.شملت المنهجية معالجة البيانات المسبقة وتوحيدها، ثم تصنيف العينات إلى خمس فئات نوعية (ممتازة، جيدة، متوسطة، ضعيفة، وغير مقبولة) بناءً على مؤشر جودة المياه .(WQI) تم تدريب نماذج SVM وXGBoostوتقييمها باستخدام تقنية التحقق المتقاطع الطبقي، وتم قياس أدائها من خلال مؤشرات الدقة، والتذكّر، والدقة التنبؤية، ومعامل.F1 أظهرت النتائج أن نموذجXGBoostتفوق على SVM خلال فترة الوفرة المائية، حيث حقق دقة تحقق بلغت 82.98% مقارنة بـ 66.86% لـSVM ، ويُعزى هذا التفوق إلى قدرةXGBoostعلى نمذجة العلاقات غير الخطية وتحديد أهمية المتغيرات. في المقابل، تفوق نموذج SVM خلال فترة الشح المائي محققًا دقة بلغت 82.79% مقابل 66.73% لـ. XGBoostتُبرز نتائج هذه الدراسة فعالية المنهج المعتمد على التعلم الآلي كحل قابل للتوسيع في المناطق القاحلة وشبه القاحلة التي تعاني من تدهور جودة المياه الجوفية.

الكلمات المفتاحية: جودة المياه الجوفية، التعلم الآلي، مؤشر جودة المياه(WQI) ، (SVM) ، XGBoost، التجميع، K-means

Abstract:

Monitoring water quality is essential for resource protection and management. This study examines the application of machine learning methods, particularly Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost), alongside K-means clustering, to assess groundwater quality in the Upper and Middle Cheliff plains based on WHO (2017) standards. The methodology included data pre-processing and standardization, followed by classifying samples into quality categories for Water Quality Index (WQI) computation. SVM and XGBoost models underwent training and evaluation through stratified cross-validation, utilizing performance metrics such as accuracy, precision, recall, and F1-score, along with K-means for clustering. Results showed that XGBoost outperformed SVM with 82.98% validation accuracy during high water periods, attributed to its capability in modelling non-linear relationships and variable importance. Nitrates, chlorides, and EC were identified as pivotal parameters influencing classification. In contrast, during low water periods, SVM outperformed XGBoost with an accuracy of 82.79% compared to 66.73%. The proposed machine learning strategy offers a scalable framework for similar arid and semi-arid regions facing groundwater challenges.

Keywords: Groundwater quality, Machine learning, WQI, SVM, XGBoost, Cluster, K-means.

Résumé :

Le contrôle de la qualité de l'eau est crucial pour la gestion durable des ressources hydriques. Cette étude applique des techniques d'apprentissage automatique (SVM, XGBoost) et de regroupement (K-means) pour classifier la qualité des eaux souterraines selon dix paramètres, conformément aux normes OMS (2017), dans les plaines du Haut et du Moyen Cheliff. Après normalisation, les échantillons ont été classés en cinq catégories de qualité selon l'indice IQE. Les modèles d'apprentissage automatique ont été évalués par validation croisée avec des métriques comme l'exactitude, la précision et le score F1. XGBoost a surpassé SVM en période de hautes eaux (82,98 % vs 66,86 %), tandis que SVM a mieux performé en basses eaux (82,79 % vs 66,73 %). La conductivité électrique, les nitrates et les chlorures étaient les paramètres les plus influents. Cette approche offre une solution efficace pour surveiller la qualité de l'eau en zones arides.

Mots-clés : Qualité des eaux souterraines, Apprentissage automatique, WQI, SVM, XGBoost, Clustering, K-means.

List of Abbreviations

ABH-CZ: Cheliff-Zahrez Water Basin Agency

AI: Artificial Intelligence

ANOVA: Analysis of Variance

ANRH: National Agency for Water Resources (Algeria)

Ca^{2+}: Calcium ion

Cl^{-}: Chloride ion

CT: Complex Terminal (aquifer)

CV: Coefficient of Variation

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

DL: Deep Learning

EC: Electrical Conductivity (µS/cm)

F1-Score: Harmonic mean of Precision and Recall

FN: False Negatives

FP: False Positives

GBDT: Gradient Boosting Decision Tree

GPUs: Graphics Processing Units

GW: Groundwater

HCO_3^{-}: Bicarbonate ion

Hm^3/year: Hectometer cubed per year

IC: Intercalary Continental (aquifer)

K^{+}: Potassium ion

K-means: K-means Clustering Algorithm

L1: Lasso Regularization

List of Abbreviations

L2: Ridge Regularization

Mg^{2+}: Magnesium ion

ML: Machine Learning

Na^{+}: Sodium ion

NO_3^{-}: Nitrate ion

NPK: Nitrogen-Phosphorus-Potassium (fertilizer)

ONM: National Meteorological Office (Office National de la Météorologie, Algeria)

PCA: Principal Component Analysis

RBF: Radial Basis Function (kernel in SVM)

RFE: Recursive Feature Elimination

SO_4^{2-}: Sulfate ion

SSL: Semi-Supervised Learning

SVM: Support Vector Machine

TN: True Negatives

TP: True Positives

WHO: World Health Organization

XGBoost: Extreme Gradient Boosting

WQI: Water Quality Index

Table of content

Table of content

Table of content

Chapter IV: Results and Discussion

Table of content

List of tables

Chapter I : Bibliographic of groundwater and its quality

Chapter II: Presentation of the study area

Chapter IV: Results and Discussion

List of Figures

Chapter I : Bibliographic of groundwater and its quality

Chapter II: Presentation of the study area

Chapter III: Materials and Methods

Chapter IV: Results and Discussion

General Introduction

General introduction:

Groundwater quality monitoring plays an important role in water resources management, especially in arid and semi arid countries, such as Algeria, and is considered one of the most important natural resources for domestic, industrial and agricultural uses due to limited surface water and reduced rainfall which exacerbated by climate change (Elmeddahi et al. 2016). Generally, natural and anthropogenic factors affect groundwater quality (climatic conditions, soil type, rock interactions, interconnection with other aquifers and their influences, agriculture, industry, urbanisation, overexploitation of groundwater, and transport of pollutants in the soil) (Sargazi et al. 2021). Given the importance of groundwater, its protection has direct or indirect impacts on economic development, human health and population growth.

The large number of physical, chemical and biological parameters of water makes water quality (WQ) monitoring difficult. In addition, there is no general factor to determine WQ. For this reason, researchers have employed water quality indices (WQI), which summarise these parameters into a single value for easy interpretation. The assessments of ground and surface water quality have been studied extensively by various researchers, proposing a quality index (WQI) with different definitions and presentations (El Baba et al. 2020; Judran and Kumar 2020).The WQI uses a statistical methodology to translate water quality parameters into a single unitless value consistent with the WHO standards.

In solving complex groundwater quality monitoring problems, classical methodologies such as direct in situ measurements or laboratory analysis of physicochemical parameters are difficult to conduct. This is mainly due to the limitations of these methods. Handling the large and high dimensional datasets is time-consuming and costly. Other problems one would face include the incomplete data records and the difficulty to access and download data sets from some websites. In such situations, predictive and classifier models offer an alternative solution to this problem. These models are effective estimation tools that reduce computation time and analysis costs. Various techniques have been used to develop classification models that can classify groundwater quality (Islam Khan 2021).

Recently, several techniques based on machine learning have extensively been developed and applied to solve various water resources engineering problems, e.g., in streamflow (Kumar et al. 2021), groundwater level prediction (Azizpor et al. 2021) and water quality prediction (Elmeddahi et al., 2022). Various models are proposed to predict the

water quality index in the literature. In recent years, several machine learning models have been proposed and applied. However, the results differed from one author to another.

This study proposed two machine learning (ML) approaches for classifying groundwater quality. Support vector machine (SVM) and XGBoost were used, while K-means was used for clustering. The goal is to transition from the traditional method of classifying ground water quality (WQ) to an advanced approach using these ML algorithms and clustering. To achieve this, the work has been divided into four chapters, beginning with an introduction and concluding with a conclusion. The first chapter provides an overview of groundwater quality and its bibliography. The study area is presented in the second chapter. In the third chapter, the methods and materials that were used are discussed. The last chapter presents the results of the groundwater classification, offering an interpretation and comparing the approaches used.

Chapter I : Bibliographic of groundwater and its quality

Chapter I : Bibliographic of groundwater and its quality

I.1. Introduction:

Water accounts for 71% of space on the planet, but 3.5% of this blue gold is fresh, and only 0.7% is accessible for human consumption. Over the last few decades, problems relating to the protection and use of water resources have increased worldwide. Water problems affect developing countries, with limited economic resources, as well as developed countries.

The chemical composition of groundwater is largely determined by the types of formations it passes through, as well as by possible inter-communications between different aquifers. Water quality can be impacted by various pollutants, with human activity being the primary source of contamination.

I.2. Natural water sources:

The available reserves of natural water include groundwater (infiltration and aquifers), stagnant surface water (lakes and reservoirs), flowing surface water (rivers) and seawater.

I.3. Groundwater:

Water that circulates underground, filling rock fractures and pores in granular environments such as sand and gravel, is known as groundwater. Unlike surface water, it moves deep into geological formations.The origin of groundwater (GW) is attributed to the accumulation of infiltrations into the soil, a process that varies depending on the porosity and geological structure of the soil, which formed over time. Aquifers play a vital role in supplying fresh water and represent the largest reservoirs of potable water. GW is typically shielded from sources of pollution and is therefore of an excellent physicochemical and microbiological quality compared to surface water. This kind of water is extracted through well drilling and emerges at surface springs (Touahria., 2013; Jean bontoux., 1993).Groundwater accounts for around 30% of the worlds freshwater. Of the remaining 70%, almost 69% is captured in ice caps and mountain snow and glaciers, while only 1% is found in rivers and lakes. Figure I.1 gives an overview of the distribution of the Earth's water.

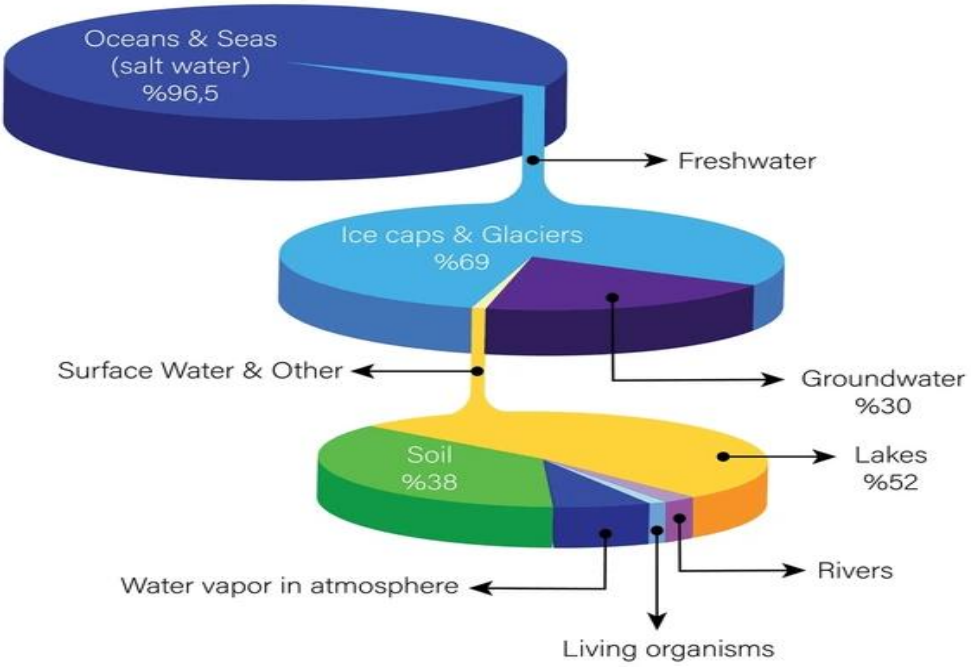

Figure I. 1: Distribution of the earth's water (USGS, 2022)

Groundwater is a vital natural resource that plays a significant role in the economy. It is the primary source of water for irrigation and the food industry.

I.4. Types of water table:

A groundwater table is a water-saturated zone in the subsoil, contained in permeable rocks known as aquifers. These aquifers are fed by precipitation that infiltrates the ground and is recharged mainly in autumn and winter. The water then flows underground, feeding springs, rivers and wells.

There are (03) three types of water table:

- Unconfined groundwater.
- Confined groundwater
- Semi-confined aquifers.

I .4.1.Unconfined groundwater:

It communicates with the surface because it is covered by a permeable layer. The pores in the rock are partially filled with water and the soil is not saturated, meaning that rainwater can permeate the groundwater over its entire surface. Its level rises or falls in response to precipitation

and it renews itself rapidly. Groundwater belongs to this category.

I .4.2.Confined aquifers:

These are aquifers located between the piezometric surface and the bedrock (the base of the aquifer). The groundwater is trapped within a permeable hydrogeological formation, which is sandwiched between two impermeable formations: the bedrock at the base and the roof at the top. The depth is greater than 30 metres. The groundwater is said to be ascending. The borehole may be artesian.

I .4.3.Semi-confined aquifers:

The roof or bedrock (or both) of the aquifer is often formed by favourable hydrodynamics. There are differences in loads that favour the desiccation of water (or pressure) in the overlying or underlying aquifer, known as the draining aquifer. This formation is then incorporated into a multilayer aquifer.

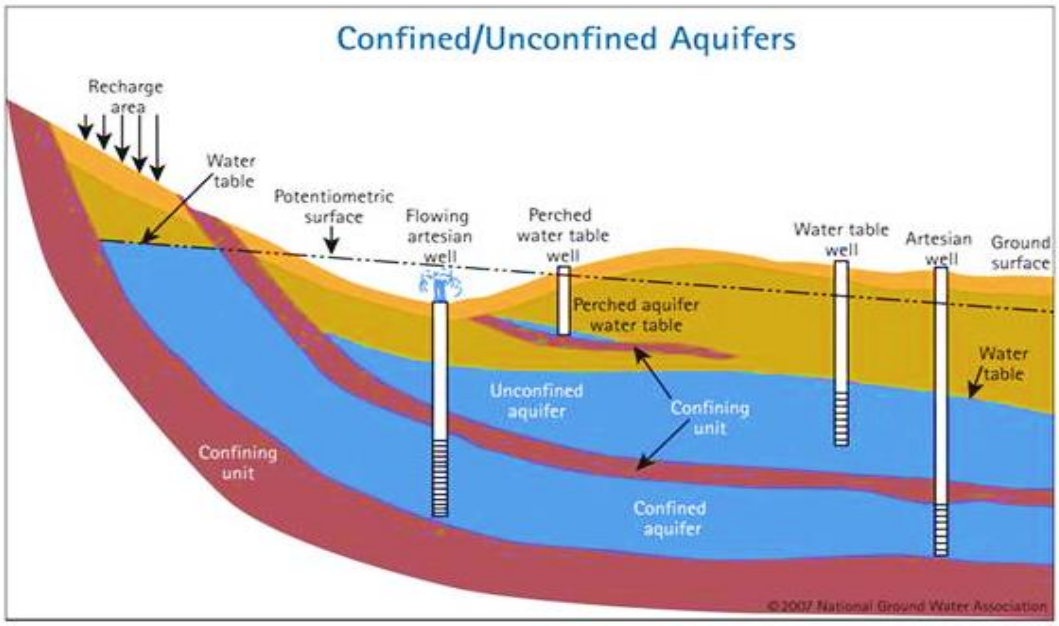

Figure I. 2: Subsurface view of various aquifers (NGWA, 2007)

I.5.Groundwater quality:

The methodology employed to assess diverse characteristics relating to groundwater quality is of paramount importance in environmental science. Despite the complexity of the physical and chemical parameters involved, groundwater quality can be defined and evaluated

through established analytical techniques.The chemical composition of groundwater originating from the natural environment can vary significantly. This depends on the geological nature of the soil from which it originates, as well as any reactive substances it may have encountered during flow.

The quality of groundwater is determined by the quantitative and qualitative composition of suspended and dissolved matter, whether mineral or organic. This quality can be altered when the aquifer comes into contact with external substances. This is the case with undesirable or toxic substances, which render groundwater unsuitable for various uses, particularly for drinking. The intensive use of natural resources and the increase in human activity have led to serious problems for groundwater quality.

I.6.Principal water quality parameters:

I.6.1.Physical parameters:

I.6.1.1.The Potential of Hydrogen (pH):

It ranks among the most important parameters evaluating water quality. This factor characterizes a numerous physicochemical balances and depends on many factors including the origin of the water. This Parameter is associated with the concentration of hydrogen ions (H^+) Present in the water. In simple terms, it determines the acidity or alkalinity and neutrality of water, pH is a parameter that defines whether water is acidic or basic; the pH of Pure water is 7 at 25°, this value that was chosen as the reference for neutral water; while water with a pH below 7 is considered acidic, and water with a pH higher than 7 is basic. The pH is a factor that depends on natural environmental conditions, such as vegetation cover and the nature of rocks and soil and soil substrate.

I.6.1.2. Temperature:

Water temperature plays an important role in the solubility of salts and gases. Temperature also increases the speed of chemical and biochemical reactions. The increase in temperature is accompanied by a decrease in density, a reduction in viscosity, an increase in saturation vapour pressure at the surface and a decrease in gas solubility. Groundwater generally remains cool, but the temperature of surface water varies according to a number of factors, including climatic conditions.

I.6.1.3. Electrical Conductivity:

Electrical conductivity measures the ability of water to conduct an electric current; it is

expressed in µS/cm. Conductivity can be used to quickly assess water mineralisation and monitor its development. Since most substances dissolved in water are electrically charged ions, conductivity indicates the total mineral content in a solution. Soft water generally has low conductivity, whereas hard water has high conductivity. Conductivity is also a function of water temperature: it increases as the temperature rises and is directly related to mineral concentration. High conductivity indicates a normal pH level or, more commonly, high salinity.

I.6.1.4. Total Water Hardness (TH):

Total water hardness, also known as hydratimetric title, is a general measure of water mineralisation, representing the total concentration of metallic cations excluding alkali metals (Na^+, K^+) and hydrogen ions. Hardness is primarily caused by calcium and magnesium ions, though occasionally by iron, aluminium, manganese and strontium ions.

I.6.1.5.Alkalinity:

It is related to the pH of the water, and is linked to the presence of strong bases (carbonates and alkalis) and weak bases (bicarbonates). We can distinguish two types of alkalinity which Correspond to two pH limits:

- Alkalinity titre (TA): which represents the quantity of strong bases .

- The complete alkalimetric titre (TAC), which corresponds to weak bases and strong bases.

I.6.2.Chemical Parameters:

The mineralisation of most waters is dominated by eight ions, often referred to as 'major ions'. These include the cations calcium, magnesium, sodium and potassium, as well as the anions: chloride, sulfate, nitrate and bicarbonate. Their concentrations in water vary independently and primarily depend on their solubility.

I.6.2.1.The cations:

❖ Calcium(Ca^{2+}):

Calcium exists in all natural waters and typically dominates the composition of drinking water, with its content varying significantly depending on the geological nature of the terrain through which it flows. Calcium is an alkaline earth metal that is found abundantly in nature, particularly in limestone rocks in the form of carbonates. It is an important component of overall water hardness. It mainly exists in its original natural form, either as the dissolution of

carbonate formations ($CaCO_3 = Ca^{2+} + CO_3^{2-}$), or as the dissolution of gypsum formations ($CaSO_4 \cdot 2H_2O = Ca^{2+} + SO_4^{2-} + 2H_2O$), which easily dissolve. Low calcium content indicates either base exchange with sodium, or the absence of calcium-rich minerals that are easily altered. Conversely, high calcium content results from the dissolution of gypsum or anhydrite. Water of high quality contains 250–350 mg/l. Water exceeding 500 mg/l presents serious drawbacks for domestic use and for supplying boilers. However, WHO standards recommend a maximum concentration of 100 mg/L. Calcium does not otherwise cause problems in drinking water; the only domestic inconvenience associated with high hardness is scaling (Gaujour .,1995 ; Athamena.,2006 ; Sedrati ., 2011 and Khellili and Lazali ., 2015).

❖ Magnesium(Mg^{2+}):

Magnesium is an essential substance that is crucial for life and often accompanies calcium. It plays a vital role in respiration. It is also considered one of the most widespread elements in nature and gives water a bitter flavour. According to Nowayti et al. (2015), magnesium's origins appear to be related to water interacting with limestone and dolomite rocks, which are carbonate formations rich in magnesium. Due to its geological abundance, high solubility and wide industrial use, its content in water can potentially reach high levels. The WHO recommends a maximum magnesium concentration limit in water of 150 mg/L.

❖ Sodium (Na^+):

Sodium is a stable element in water, but its concentration can be extremely variable. In simpler terms, sodium is an element found in all waters because the solubility of its salts is very high. Regardless of the leaching of geological formations containing sodium chloride, the salt can come from a numerous source; for instance, the decomposition of mineral salts such as sodium and aluminium silicates, from marine-origin fallout, from intrusion of saltwater into groundwater, various industrial activities… etc. While in high-quality water, the WHO suggests a maximum sodium concentration of 200 mg/L.

❖ Potassium(K+) :

Potassium is a chemical element that is mainly found in igneous rocks, particularly volcanic rocks and clays. In silicate rocks, it is mostly found as orthoclase ($KALSi_3O_8$), micas, and feldspathoids such as leucite ($KAlSi_2O_6$). Typically, groundwater contains a potassium level below 10 mg/l (BRGM. 2007). The concentration of potassium in water is usually low because vegetation absorbs this element. Nevertheless, anthropogenic factors,

especially those relating to human activity, can cause a significant increase in concentrations in water.

I.6.2.2.TheAnions:

❖ Chlorides(Cl^-):

Chlorides are constantly present in natural waters in varying proportions their occurrence in groundwater is due to the dissolution of natural salts, such as sylvite (kcl) and halite (NaCl). In drinking water, the acceptable limit is 250mg/1 (WHO standard). Exceeding this value can affect the taste of the water. Elevated chloride concentration can cause eczema and erythema. Also, water with high chloride content is laxative and corrosive.

❖ Bicarbonate (HCO_3^-):

Bicarbonates are found in natural waters. Their presence results from the dissolution of limestone and dolomite rocks, or from the release of magma from deep within the Earth. The atmospheric contribution to this is negligible (Athamena.,2006).The primary origin of bicarbonates is often the dissolution of carbonate minerals and the influence of CO_2 from meteoric water, sand and soil .The WHO does not set a standard for this parameter because the level of bicarbonate in drinking water does not affect its suitability for consumption.

❖ Nitrates(NO_3^-):

Nitrates are present in soil, surface water and groundwater. It is available in water through the leaching of nitrogen products in the soil, the decomposition of organic matter or synthetic or natural fertilisers. Nitrates are used primarily as fertilisers, with many other nitrogen fertilisers passing as nitrates in the soil. Contamination of groundwater with nitrates is one of the significant problems in water research (Phogat et al., 2014). The concentration limit of nitrate in water, established by the WHO, is 50 mg/l.

❖ Sulfate (SO_4^{2-}):

Sulfur is a non-metallic element that occurs naturally in soils and rocks in both organic and mineral forms: organic forms include protein sulfur, while mineral forms include sulfides, sulfates and elemental sulfur. When it combines with oxygen, it forms the sulfate ion, which is present in various minerals, such as gypsum and barite, and is the predominant form of sulfate found in groundwater. Sulfates (SO_4^{2-}) originate from runoff or infiltration in gypsum areas. They are also produced by certain bacteria, such as chlorothiobacteria and

rhodothiobacteria. This process can oxidise toxic hydrogen sulfide (H_2S) into sulfate .The WHO has set the maximum allowable concentration of sulfate in drinking water at 250 mg/L.

I.7.Groundwater pollution:

Groundwater is generally of a higher quality than surface water because it is less directly exposed to pollution. However, in most developing countries, groundwater is polluted as a result of increased human activity, as well as possible natural pollutants. Therefore, pollution can be seen as both a consequence of human activity and a natural phenomenon. Groundwater pollution occurs when foreign or natural substances are found in underground water at levels that pose a threat to human and/or plant health.

There are two principal pollution sources:

I.7.1. Depending on the origin of the pollution:

In general, this type of pollution caused by human activity can lead to the release of substances that can contaminate the soil and seep into the groundwater.

I.7.1.1.Domestic pollution:

The water is used by families for household and leisure needs. In the event of faulty sanitation systems (whether collective or individual), undesirable substances contained in wastewater can contaminate the water table, including organic matter, detergents, solvents, antibiotics and microorganisms.

I.7.1.2.Industrial pollution:

Pollution of groundwater is the second biggest environmental problem, and it can have a variety of causes. The most common are waste dumps and tanks used for washing or treating water by the mining, metallurgical or chemical industries. These tanks are often placed directly on the ground on excessively permeable soils without any precautions being taken. Other sources include facilities for storing or transporting products and waste from chemical or industrial accidents (Gaujour, 1995). This pollution is characterised by the presence of organic matter and fats (from the agri-food industry), hydrocarbons (from refineries), metals (from metallurgical and surface treatment industries), acids and bases (from the chemical industry), hot water from the cooling circuits of thermal power stations, and radioactive materials (from nuclear power stations).

I.7.1.3.Agricultural pollution:

The massive use of fertilisers and intensive livestock farming leads to excessive inputs of nitrogen-based fertilisers into the soil, resulting in nitrate pollution of water. This input is the main cause of nitrates being carried into underground aquifers, especially as these aquifers lie beneath highly permeable agricultural land.

I.7.1.4.Urban pollution:

Urban wastewater carries substances in suspension and solution, such as household products. possible contamination of groundwater by wastewater, due to incomplete or faulty connections, poor condition of networks, overloading or poor operation of treatment plants. (Chekroud, 2007). The sealing of surfaces (roads, streets, car parks, and roofs) produces large quantities of run-off water loaded with various pollutants (hydrocarbons, animal droppings, etc.). This polluted rainwater must never be transferred to groundwater.

I.7.2.Depending on the nature of the pollution:

This type of pollution is usually of natural origin. Contamination from natural sources is specific because it is linked to the geological context. Depending on the mineralogical context, problems may arise due to metals in the groundwater. Groundwater contains numerous microorganisms, including viruses, bacteria, protozoa, fungi and algae. The anaerobic conditions generally found in groundwater limit their diversity. The bacteria, viruses and other pathogens found in groundwater originate from septic tanks, landfill sites, sewage spills, livestock farming, fermented matter, cemeteries and surface water discharges. Pollution can also be caused by leaks in pipes and sewers or by the infiltration of surface water.

I.8.Groundwater resources in Algeria:

I.8.1.Groundwater in northern Algeria:

Groundwater resources in northern Algeria are estimated at over 2 billion m^3 (Figure I.3). They are more than 90% exploited, i.e. 1.9 billion m^3, and many aquifers are currently over-exploited. This assessment was carried out by ANRH in 1993, 2004 and 2009. Table I.1 gives estimates of groundwater resources in northern Algeria.

Table I. 1 : Groundwater potential in northern Algeria

Hydrographic regions	Cheliff-Zahrez	Algérois - Hodna-Soummam	Constantinois-Seybouse-Mellegue	Oranie-Chott-Chergui
Groundwater resources (Hm^3/year)	245	775	580	400
Available groundwater resources (Hm^3/year)	230	745	550	375

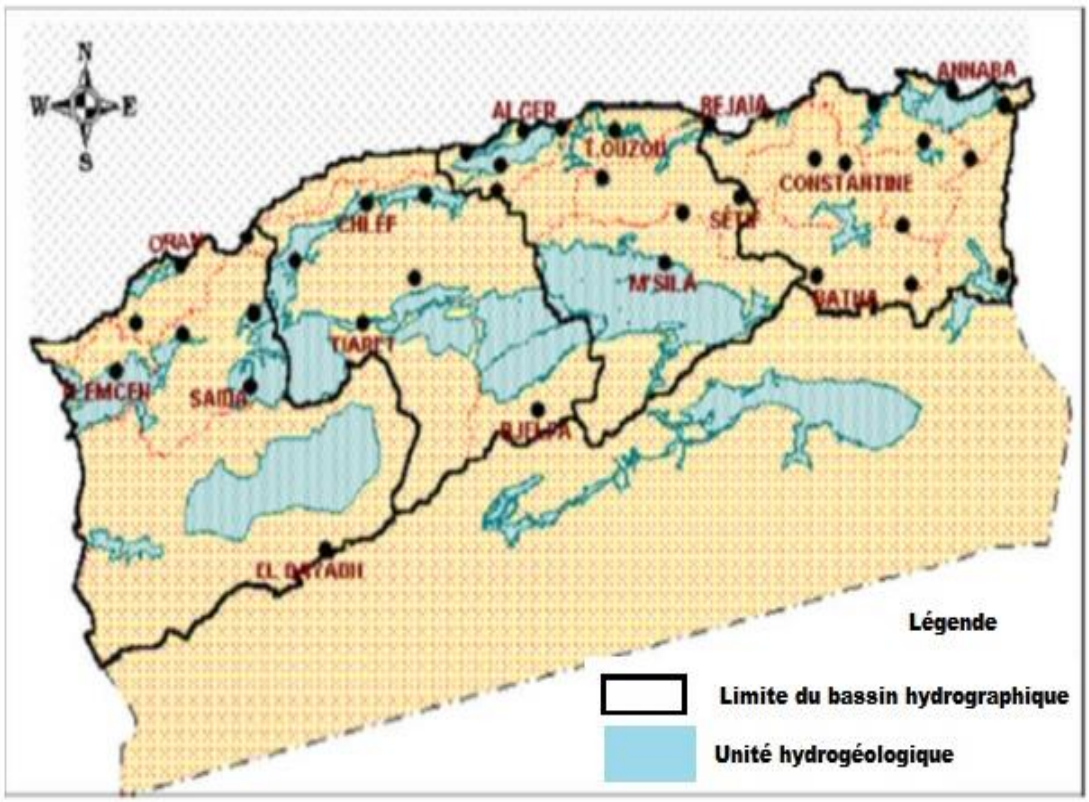

Figure I. 3: Groundwater resources in northern Algeria (ANRH)

I.8.2. Groundwater in southern Algeria:

The south is characterised by the existence of considerable groundwater resources (Figures I.4). They come from the Intercalary Continental (IC) and Complex Terminal (CT) aquifers. The reserves that can be exploited without risk of hydrodynamic imbalance are estimated at 5 billion m^3/year.

The Intercalary Continental aquifer, which is more extensive and deeper than that of the Terminal Complex, covers an area of more than 10 million km^2, spread over three countries (Algeria, Tunisia and Libya). It is a fossil water table, also known as the Albian water table, estimated at 60,000 billion m^3, but its waters are not renewable, or only to a very limited extent (Rémini, 2005).

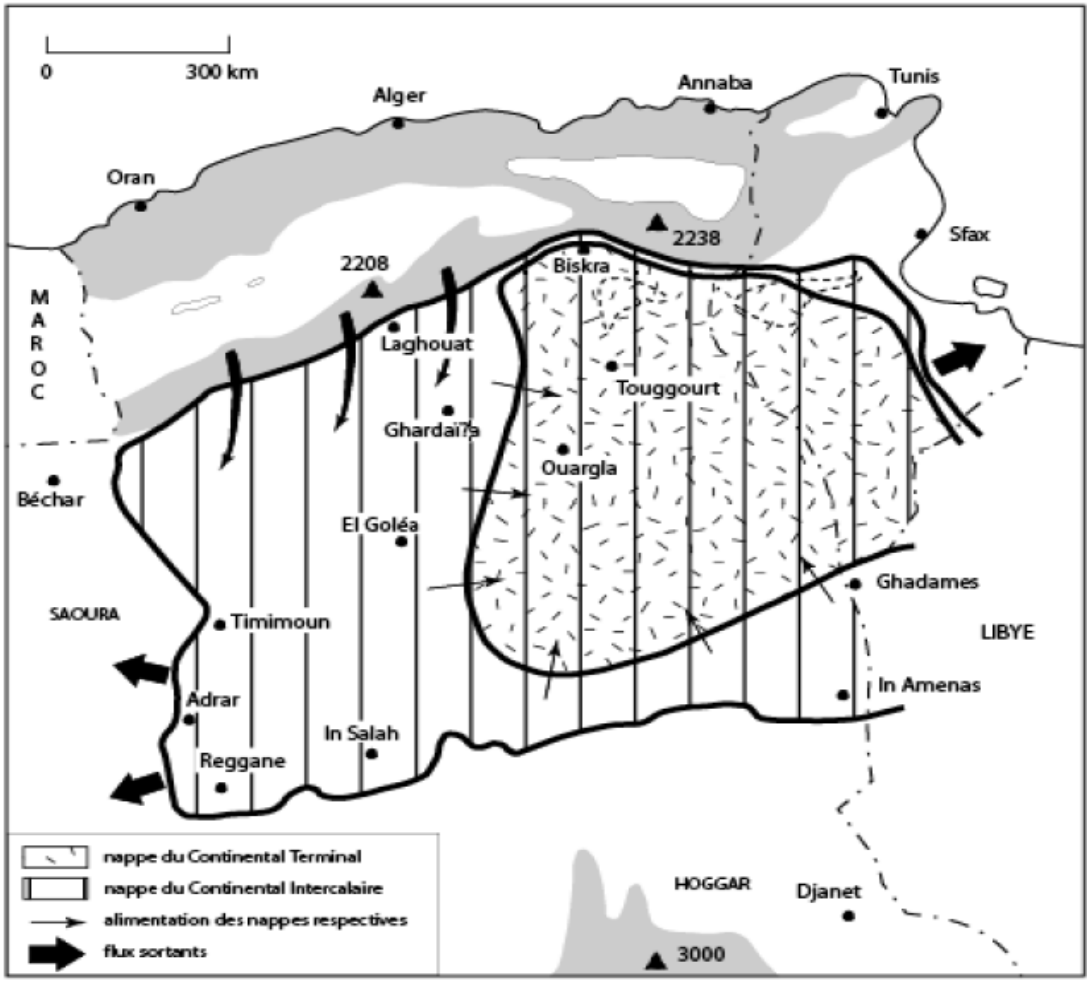

Figure I. 4: Fossil aquifers in the Sahara (Mutin, 2009)

The high level of water consumption by the three countries has been accelerated by the increase in the number of boreholes and the way they are exploited, which has led to a

significant drop in groundwater levels, the disappearance of artesianism in some parts of the Sahara and the drying up of several foggaras. The water potential of the two aquifers is estimated at 4,935 Hm^3/year, of which the volume exploited is estimated at 1,296.5 Hm^3/year.

There are other aquifers in the south, with smaller capacities than those of the terminal complex and the intercalary continental, such as the Combro-Ordovician aquifer and the Lower Devonian aquifer in Adrar, the Eocene limestone aquifer in Biskra and the Lower Devonian sandstone aquifer in Illizi, to name but a few.

The water tables, often brackish, located in or near wadi beds (OuedGhir, M'zab, Saoura, etc.) cover small areas (Rémini, 2005).

I.8.3. Groundwater in the Cheliff- Zahrez region :

The Cheliff- Zahrez region is characterised by 42 main hydrographical units. Numerous geological formations have petro-physical characteristics that are conducive to groundwater storage. The oldest of these are attributed to the Jurassic and the most recent to the Quaternary alluvium.

In the northern zone, the two Tellian ranges have poor resources that cannot be exploited. They are generally poorly developed and embedded in powerful formations with very low permeability.

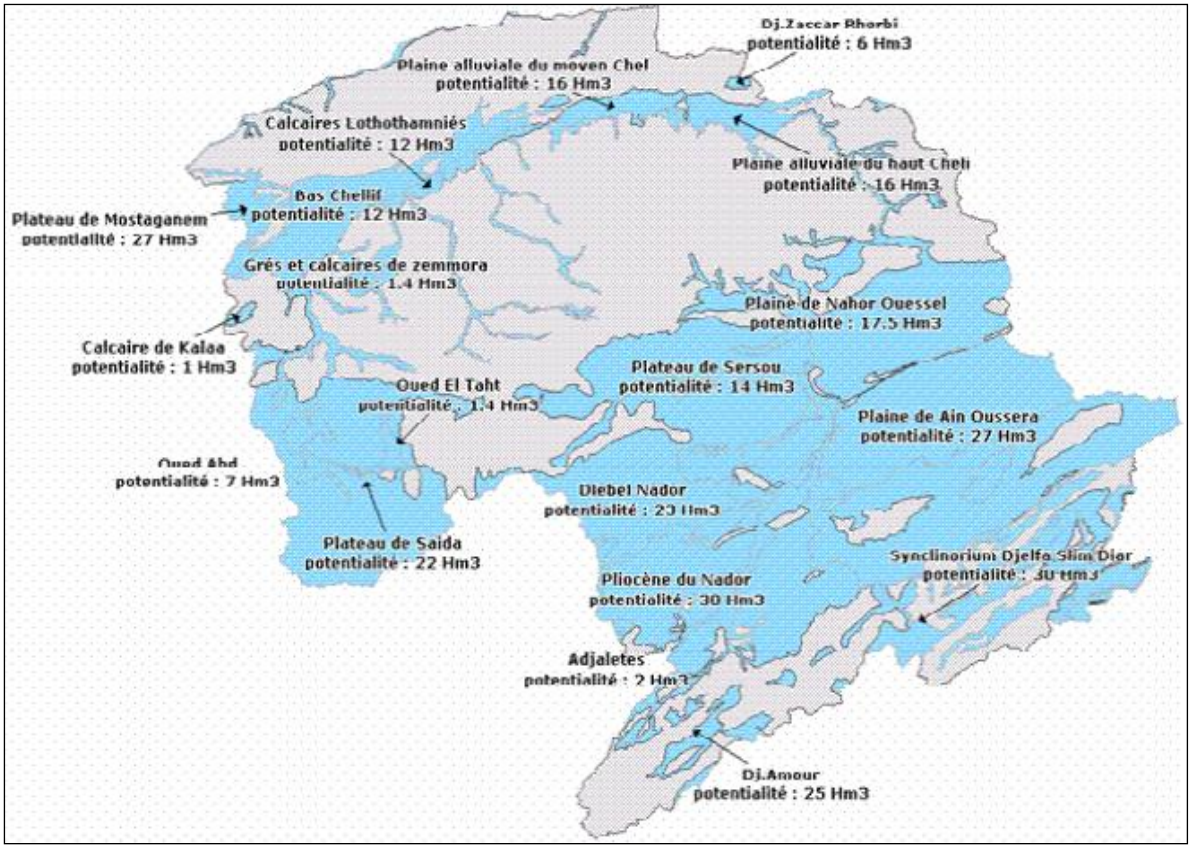

Figure I. 5: Potential of the Cheliff- Zahrez region (ABH-CZ)

The Cheliff furrow is subdivided into three basins (Upper, Middle and Lower Cheliff) separated by two sills, the Ain Defla sill and the Oum Drou sill. The aquifer formations are limited in extent. The potential of these aquifers in the Cheliff basin is estimated at 237.5 Hm^3/year.

I.9.Conclusion:

Water pollution has become one of the most significant environmental issues. Proper management of water resources requires that the quality of groundwater is not significantly deteriorated in terms of its chemical or biological properties. A decline in the quality of groundwater can impact both human health and the health of ecosystems. Although groundwater is naturally protected by soil and vegetation, it must still be safeguarded against contamination.

Chapter II:

Presentation of the study area

II.1. Introduction:

Groundwater quality is influenced by various natural and anthropogenic activities such as climate, geological factors and agricultural practices. In this chapter, we present a general overview of the Upper and Middle Cheliff plains.

II.2. Description of Case Study Area:

The Upper and Middle Cheliff plains are part of the Cheliff-Zahrez watershed which covers more than 22% of the area of northern Algeria. The study area lies between 36° 01′and 36° 20′ north latitude and 0° 58′and 02° 30′ east longitude (Figure II.1).

The water shed of the Upper and Middle Cheliff includes a total of 11 sub-basins. Its area is approximately 10,701 km^2. This basin is drained by the Cheliff Wadi, which crosses over 349 km.

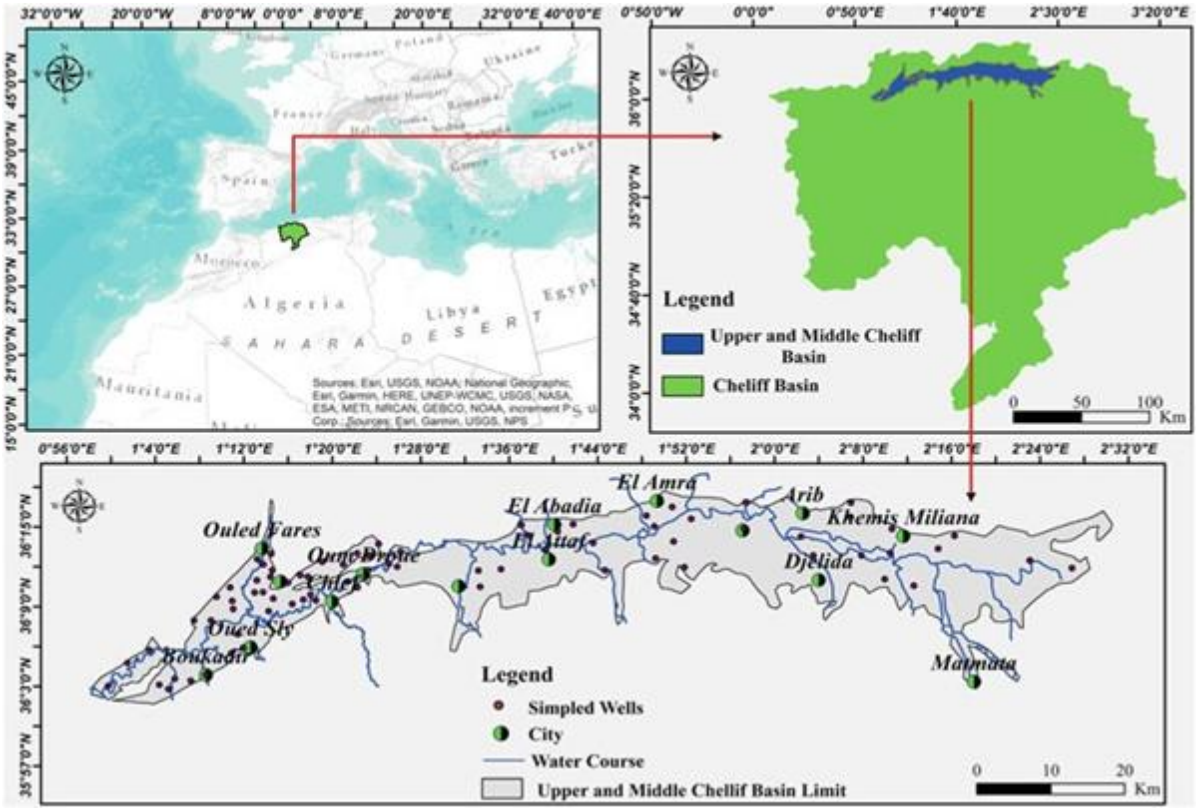

Figure II. 1: Location map of the study area

The Upper and Middle Cheliff basin is bounded by:

- ✓ The north by the coastal Dahra basin;
- ✓ The south, by the Upstream Cheliff Basin;

- ✓ The east by the Algiers hydrographic basin;
- ✓ The west, by the Lower Cheliff basin.

II.3. Geology:

The upper and middle Cheliff watershed is located in the Tell Atlas in Algeria and corresponds to an intra-mountainous subsidence furrow. According to Perrodon (1957) and Mattaeur (1958), it is located between the Boumaâd Massif and the Beni Naceur Massif to the north, and the Ouarsenis strongholds to the south (Figure II. 2).

According to lithostratigraphic information, the Upper and Middle Cheliff depression is entirely composed of Mio-Pliocene Quaternary formations (Achour and Bouzelboudjen, 1998).

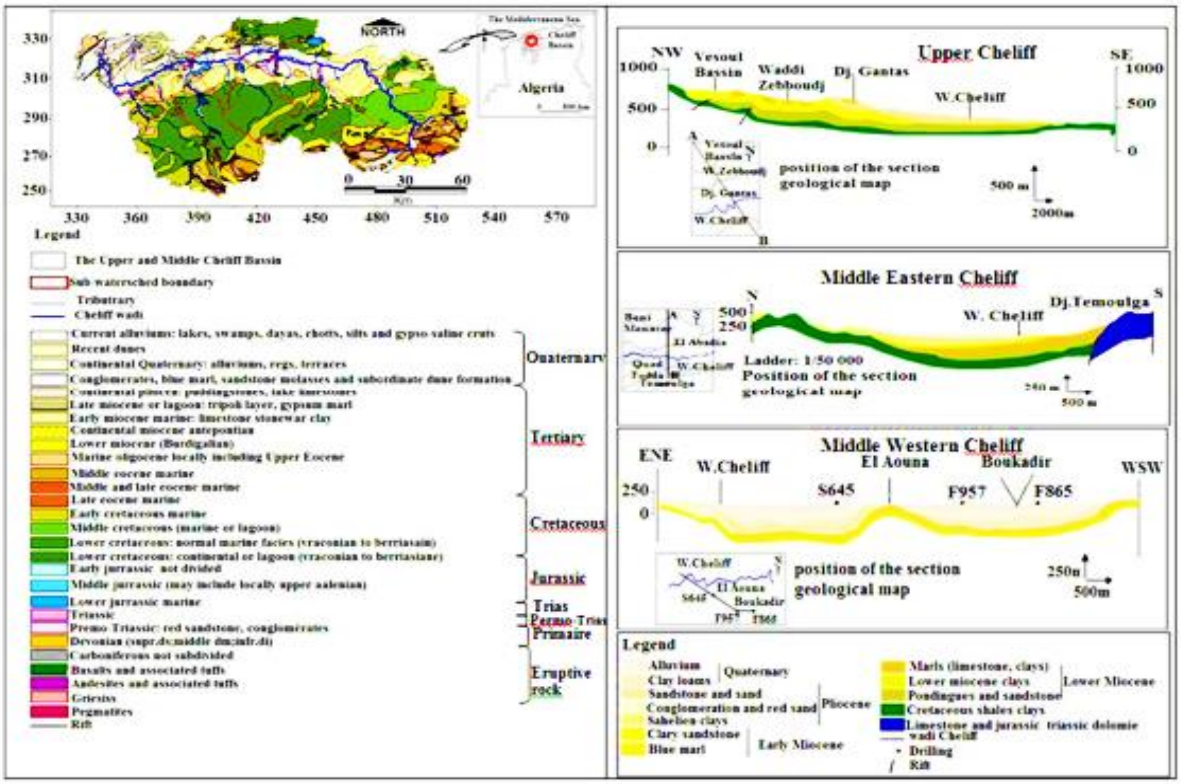

Figure II. 2: Geology of the study area and geological sections across the Upper and Middle Cheliff basin

The Cheliff Wadi crosses the plain from east to west. It enters the Upper Cheliff plain through the Djendel threshold and leaves it through the Doui threshold (Djada, 1987). It enters the Middle Eastern Cheliff plain through the Doui threshold and leaves it through the Oum Drou threshold (Pontéba) (Bouzelboudjen, 1987). It enters the Middle Western Cheliff

plain through the Oum Drou threshold (Pontéba) and leaves it through the Boukadir threshold (Charon).

These hydraulic thresholds correspond to the upwelling of the impermeable clay-marl substratum, through which any underground flow is practically excluded.

The intra-mountainous furrow corresponding to the Upper and Middle Cheliff plains was filled with Neogene deposits, upon which Quaternary, Pliocene and Miocene sediments accumulated (Mattauer, 1958; Perrodon, 1957). The Neogene marine formations can reach a thickness of 3,000 m (Meghraoui, 1982; Perrodon, 1957) and form the base of the Quaternary deposits.

They are predominant in the plains where they are consisted of coarse alluvium (ancient Quaternary) and silt(Late Quaternary) placed on the Upper Pliocene, which is formed by sandstone and limestone elements as well as sand, and the Lower Pliocene (Marine Pliocene) beginning with a transgression on the gypsum series of the Late Miocene, to end with the Astian regression (Figure II. 2).

They are predominant in the plains, consisting of coarse alluvium (ancient Quaternary) and silt (Late Quaternary), which is placed on the Upper Pliocene. The Upper Pliocene is formed from sandstone and limestone, as well as sand. The Lower Pliocene (Marine Pliocene) begins with a transgression on the gypsum series of the Late Miocene and ends with the Astian regression (Figure II. 2).

The Upper Miocene (Vindobonian) and Lower Miocene (Burdigalian) periods represent the final significant phase of tangential tectonics, resulting in the formation of a marly series. The Neogene deposits fill the basin and include three large plains, which can be distinguished from east to west as follows:

The plain of El Khemis or the plain of Upper Cheliff; the plain of El Abadia-Amra or the plain of Middle Eastern Cheliff; the plain of Chlef or the plain of Middle Western Cheliff.

This basin is characterised by intense neotectonics (Meghraoui et al., 1986), as evidenced by the seismic rift of Oued Fodda (the earthquake of 10 October 1980), which caused significant changes to surface and groundwater flows, raising the groundwater level by nearly 2 metres.

II.4. Hydro-geological context:

A study of the stratigraphic series, including its lithological and structural characteristics (Figures II. 2 and 3), reveals the main aquifer formations.

- The main aquifer of the Upper and Midlle Cheliff plain is made up of alluvial deposits, including pebbles, gravels, sands and clay formations, which are between 50 and 150 m thick. It is a confined aquifer as it is covered by 5–20 m of silt and clay at the surface. The alluvial aquifer is overlain by Mio-Pliocene formations, with sandstones present to the north-east at Gontas and to the south-east at Ain-Lechiekh. Several wells have been drilled in these two areas, producing high-quality groundwater.
- The Mio-Pliocene formations can be up to 200 metres thick. The Mio-Plio-Quaternary formations are therefore considered to be a multi-layer aquifer system. Hydraulic continuity between these formations indicates continuity in certain areas, while clay lenses separate the formations in other parts of the zone. This study only considers the Quaternary aquifer.
- The Pliocene formation takes the form of yellowish Astian sandstones, which are topped by helix dune sands, with an average thickness of 100m.
- Another aquifer attributed to the Miocene is the Zaccar limestone aquifer.

The water table is fed by the impluvium and by run-off from the Cheliff, Ebda and El Arch wadis.

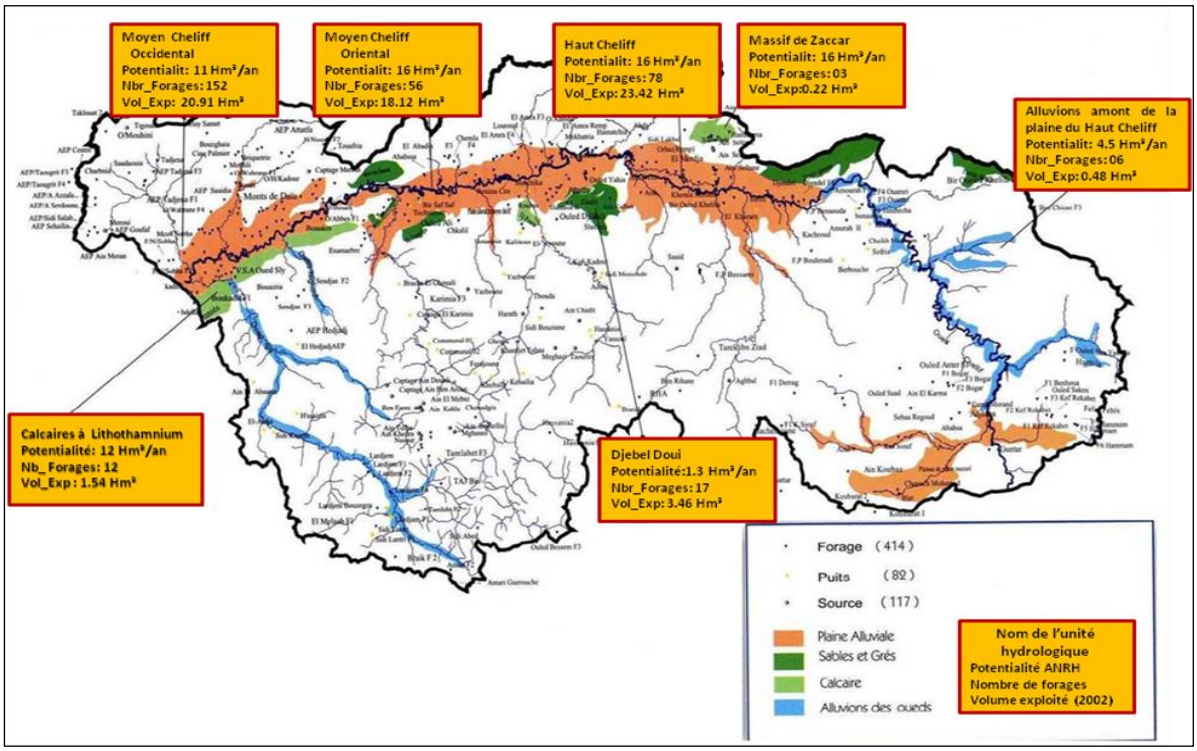

Figure II. 3: Potential of the Upper and Middle Cheliff (ABH-CZ, 2004).

Groundwater potential in the Upper and Middle Cheliff basin is estimated at 67.5 hm3 /year according to the ANRH. Depending on the lithology of the study area, it is concentrated in alluvial geological formations and Lithothamnium limestone formations (Figure II.3).

II.5. Soil:

The study area has great agricultural potential. This agricultural region is characterised by two irrigation schemes, which were built in the 1930s: the Middle Cheliff scheme, commissioned in 1936, and the Upper Cheliff scheme, commissioned in 1937.

The region's soils fall into five categories according to the classification proposed by ANRH based on physico-chemical properties and various natural factors (geomorphology, topography, etc.) (P.D.A.R.E, 2011).

- **Category I** soils do not present any major development problems and should be developed as a priority; they are suitable for all crops.
- **Category II** soils present minor development problems (stone removal or surface reclamation). These soils are particularly suitable for industrial crops.
- **Category III** soils should be reserved for crop rotation, where the main management problems are drainage after irrigation and desalination.

Table II. 1: Irrigable areas by soil category (hectare: ha)

Basin	soil category					Irrigable soils (I+ II+ III) ha
	I	II	III	IV	V	
Upper and Middle Cheliff	26851	24956	30156	140962	18913	81963

- **Category IV** soils are sometimes saline or hydromorphic with a shallow water table. They present major development problems. Cultivation potential is often limited to cereal, forage and market garden crops. Dry farming is recommended.
- **Category V** soils correspond to soils that are unsuitable for irrigation for various reasons: the presence of limestone crusts at shallow depths, halomorphy, very pronounced hydromorphy and unfavourable topography.

II.6.Climate context:

II.6.1.Climate characteristics:

The study area has a Mediterranean climate with continental characteristics, featuring hot, dry summers and cold, fairly rainy winters (Legoupil, 1974). The region itself has a distinctive climate characterised by long, hot, dry summers and short, cold, rainy winters.

II.6.1.1.Precipitation:

Precipitation is an important climatic factor because of its influence on the repair of plant species. The data consists of annual rainfall totals for the period from 1980 to 2016 for three representative stations (Figure II.4), and monthly temperatures obtained from the National Water Resources Agency (ANRH) and the National Meteorological Office (ONM).

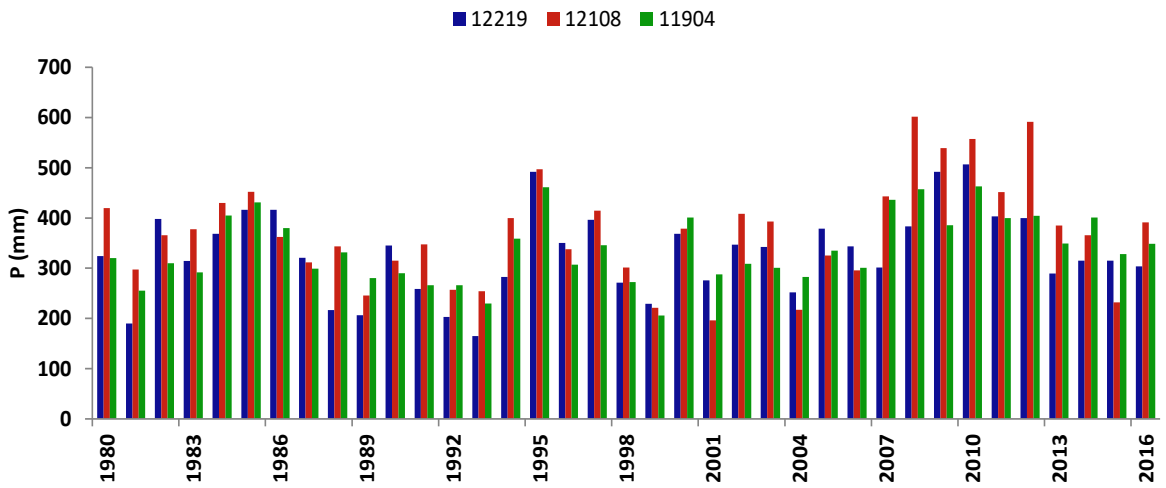

Figure II. 4: Histogram of mean annual precipitation (1980-2016).

Figure II.4 shows annual variations in rainfall over the period from 1980 to 2016 for three stations representative of the study area: The ANRH CHLEF station (012219), the FODDA BARRAGE station (012108), and the ROUINA MAIRIE station (011904). These variations are highly irregular.

At the ANRH CHLEF station, the average annual rainfall varies from 165.0 mm to 507.0 mm. The maximum value was recorded in 2010, while 1993 was the driest year.

The average annual rainfall at the FODDA BARRAGE station ranges from 196.4 mm to 601.9 mm, and from 205.7 mm to 463.0 mm at the ROUINA MAIRIE station. The maximum value was recorded at the FODDA BARRAGE station in 2008, while the maximum value was recorded at the ROUINA MAIRIE station in September 2010.

II.6.1.2.Temperature:

The climate in the study area is semi-arid Mediterranean, characterised by hot, dry summers and mild, wet winters (Troll, 1964, cited in Gomer, 1994).

Figure II.5 shows estimated minimum and maximum annual temperatures of 9.6°C and 28.8°C, respectively.

Examination of the data presented in (Figure II.5) shows that mean inter-annual temperatures differs slightly between stations. The average maximum temperature is 29.6°C at the Harreza station and 29.8°C at the ONM Chlef station. The minimum temperature is around 9.6°C at both stations.

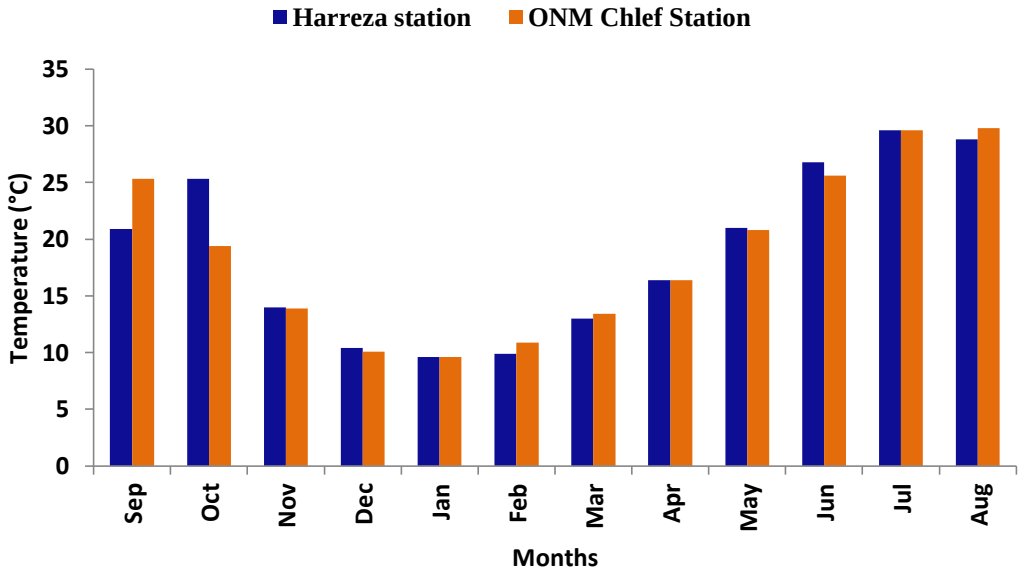

Figure II. 5: Histogram of average monthly temperature (1995-2014).

An analysis of the histogram showing the average monthly temperatures reveals:

- A cold period from November to March;

- A hot period from April to October.

II.7 Conclusion:

The study area is characterised by highly diverse geology and a predominantly flat landscape, corresponding to the valleys adjoining the Wadi Cheliff and its tributaries. Analysis of various climatological parameters indicates that the study area has a semi-arid climate with highly irregular rainfall.

Chapter III:

Materials and Methods

Chapter III: Materials and Methods

III.1. Overview:

In the recent decades that have transpired, there has been a remarkable and unprecedented surge in the growth of the field of machine learning, which can be attributed largely to groundbreaking advancements and innovations in both computational power and storage technologies that have revolutionized the capabilities of data processing and analysis, thereby contributing significantly to the development of pioneering technologies across a multitude of disciplines, particularly in the realms of science and engineering. The domain of machine learning is undeniably expansive and complex, encompassing a wide variety of methodologies and approaches, and this chapter aims to furnish a comprehensive overview of the fundamental principles underlying the concept of learning, as well as to highlight some of the most widely recognized and utilized machine learning techniques that serve as the foundational bedrock for the innovative technologies employed in diverse applications across numerous sectors. A crucial and indispensable element of any given technique, particularly those related to machine learning, is the necessity to rigorously assess and evaluate its effectiveness in accomplishing the specific tasks for which it is designed, as this assessment is paramount to understanding the practical utility and applicability of the technique within real-world contexts. This principle holds true for machine learning methodologies as well, where thorough evaluation is essential to validate their performance metrics and ensure they meet the required standards of efficacy. Consequently, this chapter will also delve into the typical and widely accepted methodologies that are systematically employed to evaluate various learning techniques in a general sense, thereby providing insights into the metrics and benchmarks used to gauge their success and applicability. By outlining these evaluation strategies, the chapter aims to furnish readers with a deeper understanding of how machine learning techniques are scrutinized and assessed, thus laying the groundwork for further exploration into their practical implications and potential for future development. Furthermore, this discourse on evaluation methodologies will encompass both qualitative and quantitative approaches, highlighting the importance of a balanced perspective in measuring the performance of machine learning systems. Ultimately, the exploration of these concepts and methodologies will serve to enhance the reader's comprehension of the intricate relationship between machine learning techniques and their effectiveness in addressing complex problems across various disciplines. In summary, this chapter endeavors to provide an in-depth examination of the landscape of machine learning, encompassing both

foundational concepts and practical evaluation methodologies, thereby equipping the reader with the knowledge necessary to navigate the ever-evolving field of innovative technologies.

III.2. Machine learning:

Machine learning, commonly abbreviated as ML, represents an intricate and multifaceted domain within the broader realm of artificial intelligence, frequently referred to as AI, which is primarily concerned with the meticulous formulation and advancement of sophisticated algorithms and robust statistical models that empower computational systems to execute a plethora of tasks autonomously, without the need for explicitly defined programming directives. In this scholarly article, we aim to furnish a comprehensive overview of the expansive field of machine learning, encompassing not only its pivotal concepts but also the various techniques and practical applications that have emerged as a result of its rapid evolution. We delve into the diverse categories of machine learning methodologies (Figure III.1), which include, but are not limited to, supervised learning, unsupervised learning, and reinforcement learning, each accompanied by their distinct algorithms and real-world use cases that illustrate their utility and effectiveness in problem-solving. Furthermore, we engage in an in-depth exploration of foundational principles such as the processes of model training, rigorous evaluation, and subsequent deployment of machine learning systems, while also taking into account the burgeoning trends that are reshaping the landscape, including the transformative impacts of deep learning and the innovative strategies associated with transfer learning. Through this review, we aim to offer a comprehensive introduction to machine learning, catering to both beginners and seasoned practitioners, and highlight its significance in advancing AI-driven solutions across diverse domains.

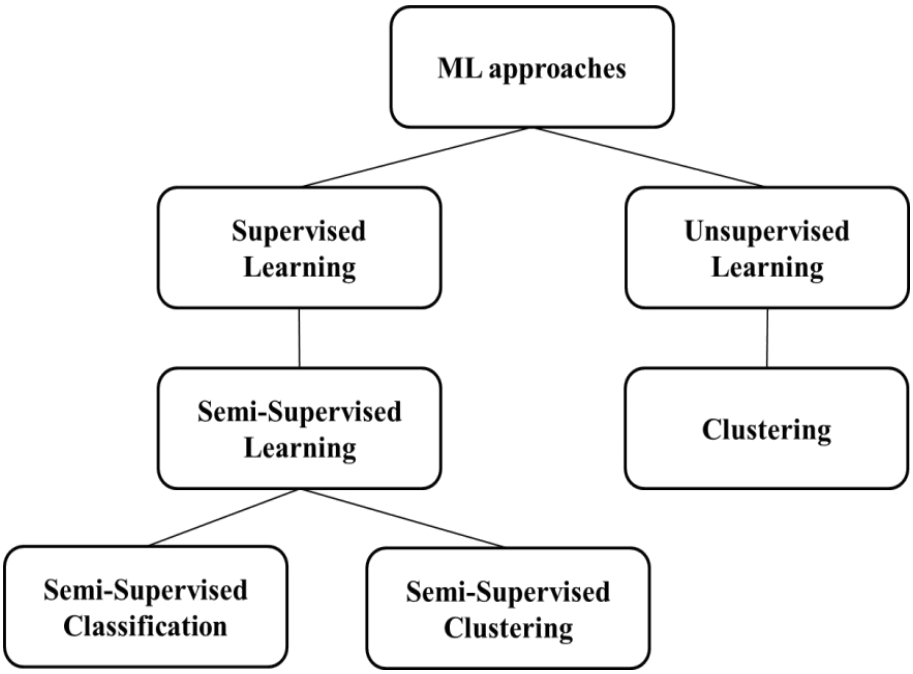

Figure III. 1: Machine Learning Approaches

III.3. Type of machine learning:

III.3.1.Supervised learning:

Supervised learning is a fundamental paradigm in machine learning where models are trained on labeled datasets to predict outcomes or classify data. This approach is widely utilized across various fields, including medical diagnosis, image recognition, and financial forecasting. The process typically involves two phases: training, where the model learns from input-output pairs, and testing, where it is evaluated on unseen data (Sharma, 2024). The effectiveness of supervised learning hinges on the quality of the labeled data and the choice of algorithms, which include linear regression, decision trees, and neural networks (Sharma, 2024)

III.3.2 unsupervised learning:

Unsupervised learning involves finding hidden patterns in unlabeled data without error signals. Techniques include clustering, dimensionality reduction, and neural network models like SOM and ART for pattern recognition tasks (Buhmann, et al., 1999).

Unsupervised learning is a machine learning paradigm focused on identifying hidden

structures in unlabeled data, distinguishing it from supervised and reinforcement learning. In this approach, algorithms analyse input data without explicit target outputs, relying on the inherent patterns within the data itself. This method is crucial for tasks such as clustering, dimensionality reduction, and feature extraction, which help summarize and explain key data characteristics.

III.3.2.1. Cluster:

In unsupervised learning, a cluster denotes a collection of data points that possess similar traits and are assembled together based on patterns or structures inherent in the data, without requiring any predefined labels. The primary objective of clustering is to divide the dataset into clear-cut groups where data points within the same cluster display a high degree of similarity, whilst those in disparate clusters are as different as possible. This methodology is extensively employed for exploratory data analysis, enabling researchers to reveal concealed structures, discern trends, or identify anomalies in unlabeled datasets. Well-known clustering algorithms encompass K-means, which segments data into a set number of clusters by minimising variance; hierarchical clustering, which creates nested clusters in a tree-like configuration.

clustering represents a significant unsupervised learning approach employed to categorise groundwater samples according to their physicochemical characteristics, including pH, electrical conductivity (EC), and the concentrations of principal ions (e.g., Na^{+}, K^{+}, Ca^{2+}, Mg^{2+}, Cl^{-}, SO_4^{2-}, HCO_3^{-}). Through the utilisation of clustering algorithms, researchers are able to discern distinct groups of groundwater exhibiting similar hydrochemical traits, which aids in comprehending aquifer systems, fluctuations in water quality, and potential sources of contamination. Frequently utilised clustering techniques in the realm of groundwater studies comprise K-means and hierarchical clustering.

III.3.3.semi-supervised learning:

Semi-supervised learning (SSL) is a type of Machine Learning (ML) technique. It is half-way between supervised and unsupervised learning i. e the dataset is partially labeled is shown in Figure III. 2. The main objective is to overcome the drawbacks of both supervised and unsupervised learning. Supervised learning requires huge amount of training data to classify the test data, which is cost effective and time consuming process. On the other hand, unsupervised learning doesn’t require any labeled data, which clusters the data based on similarity in the data points by using either clustering or maximum likelihood approach. The

main downfall of this approach, it can't cluster an unknown data accurately. To overcome these issues, SSL has been proposed by research community, which can learn with small amount of training data can label the unknown (or) test data. SSL builds a model with few labeled patterns as training data and treats the rest of the patterns as test data. The generic Semi-supervised learning Algorithm

III.3.4ReinforcementLearning:

Reinforcement learning is a very distinct learning system that is carried out without direct supervision, through observation with the environment. Generally, this type involves rewarding desired behaviours and/or penalizes undesirable ones and also interpreting environment, taking actions and learning via trial and error. The primary goal is to find the best behaviour strategy to maximize the overall reward. To do this, the machine needs basic in order to define the nature of its behaviour. This result feed-back is known as there enforcement signal. Although it is not obvious to manually create or program scenarios to achieve better results, machine learning is extremely useful when there are large numbers of conditions to forecast

III.4. Used models:

III.4.1 Cluster (unsupervised learning):

III.4.1.1. K-means:

K-means clustering is a powerful and widely used unsupervised machine learning technique for classifying groundwater samples based on their hydrochemical properties. The algorithm operates by partitioning a dataset into K distinct, non-overlapping clusters, where each sample belongs to the group with the nearest mean value. In groundwater studies, this method helps identify patterns in water quality parameters such as pH, electrical conductivity (EC), and concentrations of major ions like sodium (Na^+), calcium (Ca^{2+}), chloride (Cl^-), and bicarbonate (HCO_3^-). By applying K-means, hydrogeologists can categorize water samples into meaningful groups that reflect different hydrochemical facies, contamination sources, or aquifer conditions, providing valuable insights for water resource management.

The K-means algorithm follows an iterative process to optimize cluster assignments. First, the data is standardized to ensure all parameters contribute equally to the clustering. Next, the

number of clusters (K) is selected (Figure III.2), often using methods like the Elbow Method or Silhouette Analysis, which balance model complexity and clustering performance. The algorithm then initializes K centroids, either randomly or using an optimized approach like K-means++, and iteratively refines their positions by assigning each data point to the nearest centroid and recalculating the centroids based on the assigned points. This process continues until the centroids stabilize, resulting in well-defined clusters. For groundwater applications, the final clusters may represent distinct water types, such as freshwater, brackish water, or saline water, each with unique geochemical signatures.

K-means clustering has proven particularly useful in groundwater studies for tasks such as identifying contamination sources, mapping hydrochemical facies, and assessing spatial variations in water quality. For example, in coastal aquifers, K-means can help delineate zones affected by saltwater intrusion by clustering samples with elevated chloride and sodium concentrations. Similarly, in agricultural regions, the method can detect nitrate-enriched clusters linked to fertiliser use. However, K-means has limitations, including sensitivity to initial centroid placement and an assumption of spherical, equally sized clusters, which may not always hold true for complex groundwater datasets. To address these challenges, researchers frequently integrate K-means with dimensionality reduction methods such as Principal Component Analysis (PCA) or opt for alternative algorithms like hierarchical clustering or DBSCAN for more refined classifications. Notwithstanding these constraints, K-means continues to be an indispensable instrument for initial groundwater classification owing to its straightforwardness, efficiency, and clarity.

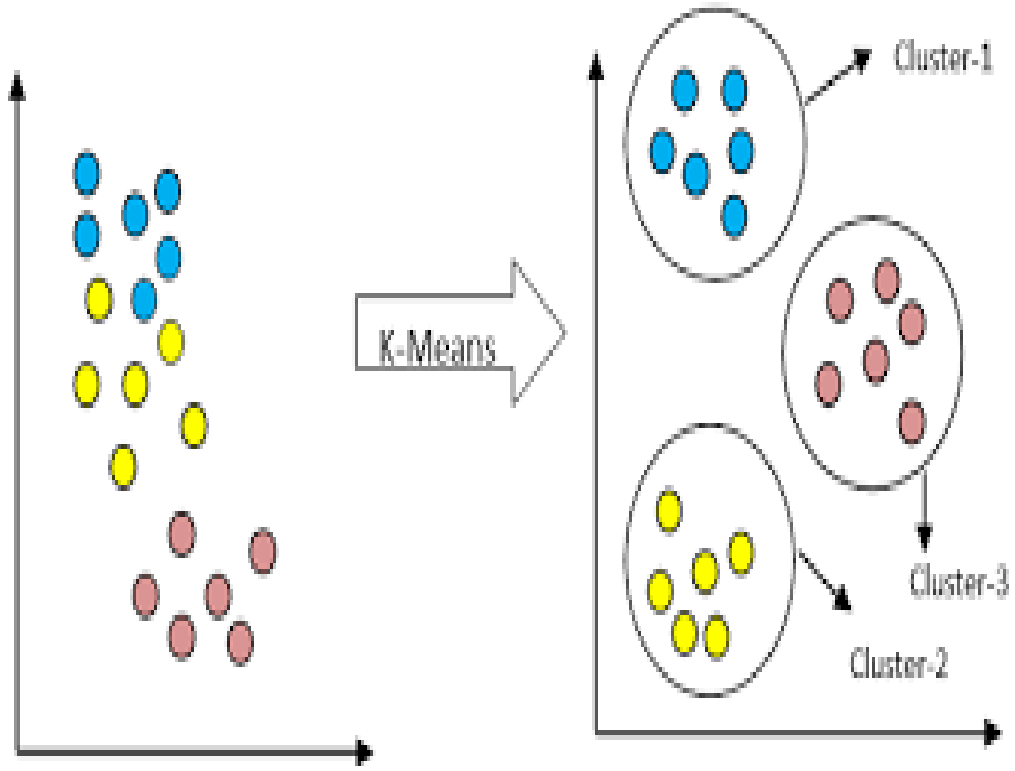

Figure III. 2: K-means clustering

III.4.2. Supervised Learning:

III.4.2.1. XGBoost (Extreme Gradient Boosting):

XGBoost, which is an abbreviation for Extreme Gradient Boosting, represents a meticulously optimized distributed library that employs gradient boosting techniques and is specifically engineered to demonstrate a high degree of efficiency, adaptability, and portability across various computational environments. This sophisticated library provides implementations of diverse machine learning algorithms that operate within the comprehensive framework of Gradient Boosting, thereby facilitating the development of robust predictive models. Initially introduced to the academic and professional communities, XGBoost has rapidly ascended to prominence, establishing itself as one of the most widely utilized and extraordinarily powerful tools available for machine learning applications, particularly in contexts involving structured or tabular data formats. The remarkable capabilities and versatility of XGBoost have led to its widespread adoption across various domains, further solidifying its reputation as an indispensable asset in the toolkit of data scientists and machine learning practitioners alike (Chen, and Guestrin, 2016).

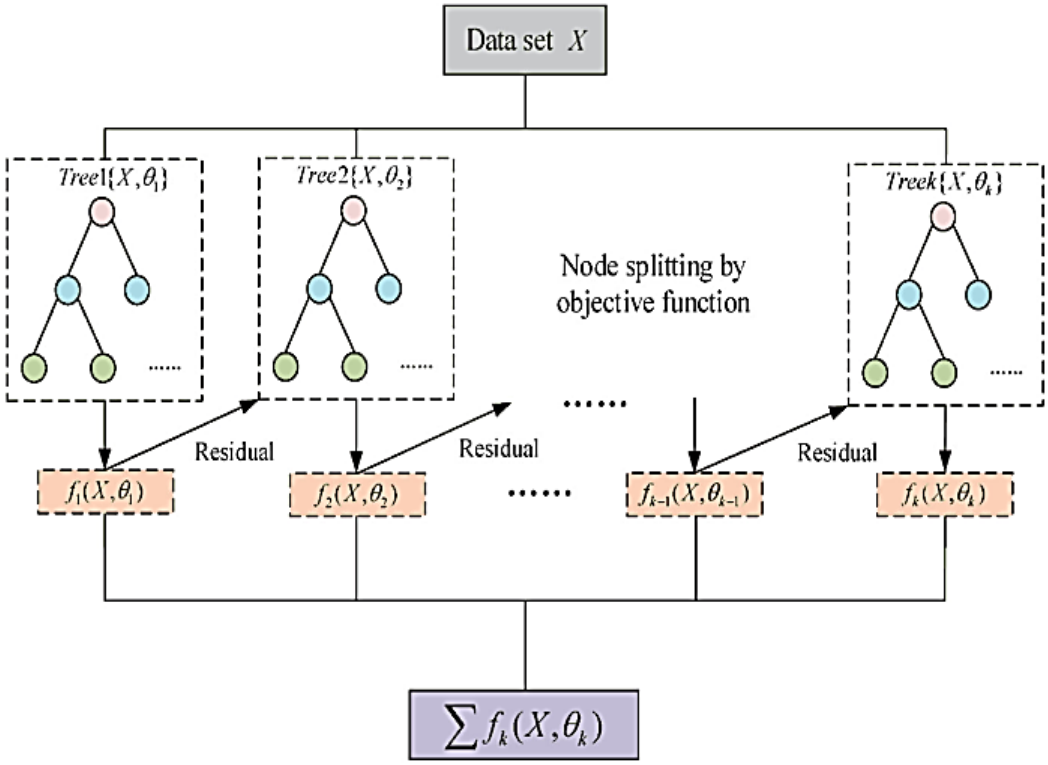

Figure III. 3: The structure of XGBoost

- Key Features of XGBoost :

- **Performance:**

XGBoost is designed to push the limits of computing efficiency. It can handle sparse data, missing values, and extremely large datasets with minimal computational cost.

- **Regularisation:**

Unlike classic gradient boosting, XGBoost includes both L1 (Lasso) and L2 (Ridge) regularisation, which helps to prevent overfitting — a common issue in machine learning models.

- **Parallel Processing:**

XGBoost leverages parallel and distributed computing during model training, which dramatically reduces computation time compared to traditional boosting algorithms.

- **Handling Missing Data:**

The algorithm can automatically learn the best direction to handle missing values during training without the need for pre-processing.

- **Scalability:**

XGBoost can efficiently scale to billions of data points and is designed to work with cloud services and multi-core CPUs/GPUs.

➢ How XGBoost Works :

XGBoost is based on the **Gradient Boosting Decision Tree (GBDT)** algorithm. It works by building an ensemble of weak prediction models (typically decision trees) in a stage-wise fashion. At each stage, it minimises a loss function using gradient descent.

➢ Use of Boost:

- Robust to over fitting when properly tuned.
- Superior accuracy in competitions and real-world tasks.
- Built-in handling of missing data.
- Efficient memory and computation usage.

- Wide adoption in Kaggle competitions, scientific research, and industry (e.g., finance, healthcare, and hydrology).

III.4.2.2. SVM (support vector machine):

Support Vector Machine (SVM) is powerful supervised learning algorithms primarily used for classification, regression, and outlier detection. SVM was first introduced by Vladimir Vapnik and Alexey Chervonenkis in the 1960s, and later developed further in the 1990s for practical machine learning applications.

The core idea behind SVM is to find an optimal hyperplane that best separates data points of different classes. This hyperplane maximizes the margin between two classes, meaning it creates the largest possible distance from the nearest data points (called support vectors) of each class (Cortes andVapnik, 1995).

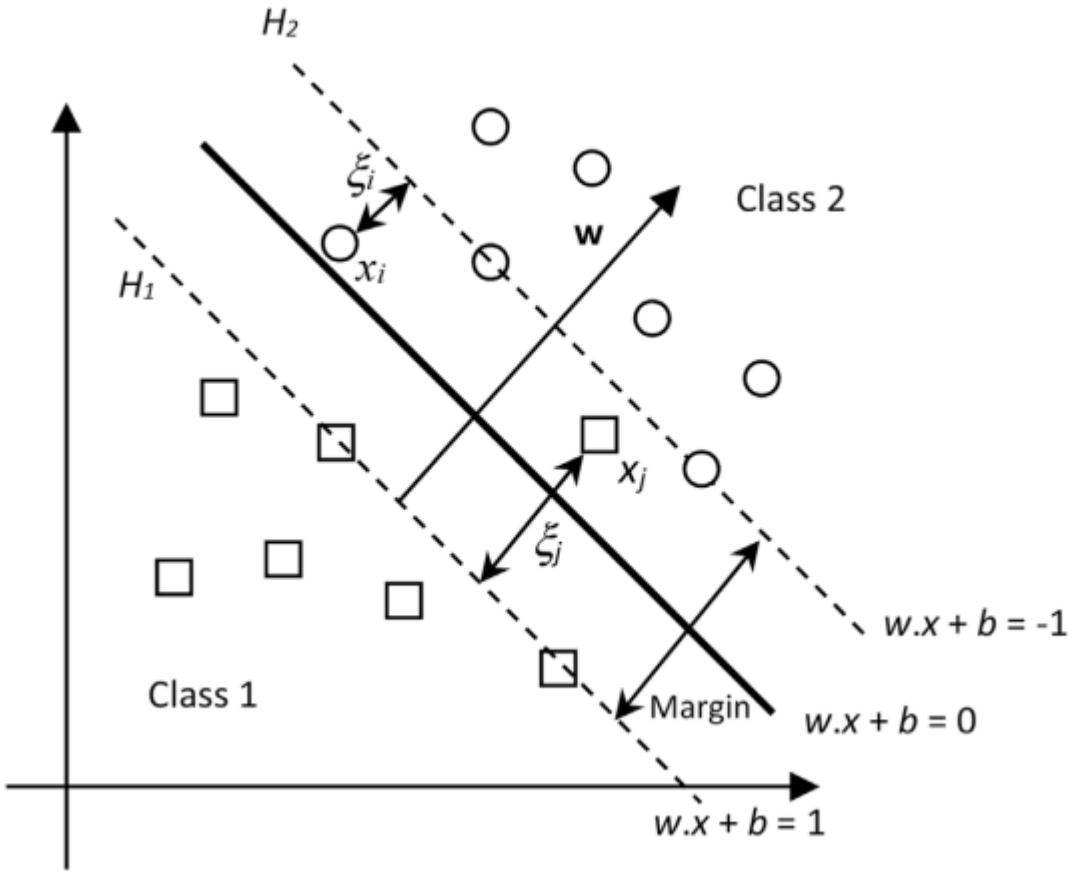

Figure III. 4: The structure of a basic SVM

- How SVM Works:

If the data is **not linearly separable**, SVM uses a **kernel trick** to project data into a higher-dimensional space where separation is possible. Popular kernels include:

- Linear kernel,
- Polynomial kernel,
- Radial Basis Function (RBF) kernel,
- Sigmoid kernel.

- Use of SVM:

- SVM are highly effective in **high-dimensional spaces**.
- They remain effective even if the number of dimensions exceeds the number of samples.
- Versatile through the use of different **kernel functions**.
- Strong theoretical foundation rooted in **statistical learning theory**.

III.5. Model Performance Evaluation:

The sole approach in evaluating a model's capability to generalize to novel instances is to put it into a test. This consists of deploying the model into production, evaluating its performance, validating the model, and checking the results based against expectations. For our research, two statistical criteria were considered:

III.5.1.Accuracy:

Accuracy is one of the most intuitive and commonly used metrics for evaluating classification models. It measures the proportion of correctly predicted instances relative to the total number of predictions:

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

Where:

TP: stands for True Positives,

TN: for True Negatives

FP: for False Positives

FN: for False Negatives.

While accuracy is useful for balanced datasets, it may be misleading when data are imbalanced, as it does not distinguish between the types of errors (Sokolova and Lapalme, 2009).

III.5.2. Mean Accuracy (Average Accuracy):

Mean Accuracy is often applied in multi-class classification problems. Unlike plain accuracy, which can be biased by the dominant class, Mean Accuracy calculates the classification accuracy for each class individually and averages them:

$$\textbf{Mean Accuracy} = \frac{\mathbf{1}}{\mathbf{C}} \sum_{\mathbf{i=1}}^{\mathbf{C}} \frac{\mathbf{TPi}}{\mathbf{TPi} + \mathbf{FNi}}$$

Where:

C is the number of classes.

This metric ensures fair treatment of all classes, especially in datasets with unequal class distributions **(**Huang et al., 2006).

III.5.3. Precision:

Precision measures the proportion of correctly predicted positive observations out of all observations predicted as positive:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

Precision is particularly useful when the cost of false positives is high, such as in spam detection or fault diagnosis systems (Powers, 2011).

III.5.4. Recall:

Recall, also known as Sensitivity or True Positive Rate, measures the model's ability to correctly identify all actual positive cases:

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}}$$

High recall is critical in applications where missing a positive case would lead to severe consequences, such as medical screenings (Davis and Goadrich, 2006).

III.5.5. F1 Score

The F1 Score combines Precision and Recall into a single metric using their harmonic mean

$$\textbf{F1} = 2 \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

It is especially relevant when there is an uneven class distribution or when both false positives and false negatives are equally costly (Chinchor, 1992).

III.6. Feature selection and input data:

The accuracy of machine learning models in water quality classification hinges on selecting the most discriminative input parameters while eliminating redundant or noisy features. This chapter systematically evaluates feature selection methodologies and input parameter combinations to optimize an XGBoostbased classifier. The study addresses the trade-off between model complexity and predictive performance, emphasizing computational efficiency and interpretability for environmental monitoring applications

III.6.1. Data Pre-processing:

The raw dataset comprised 10 hydro-chemical parameters (pH, conductivity, chlorides, etc.). Each parameter was standardised via Z-score normalisation to ensure equal weighting during feature selection. Class imbalance was mitigated by computing XGBoost's sample weights inversely proportional to class frequencies.

III.6.2.Feature Selection Strategy:

Feature selection constitutes an essential pre-processing phase in the domain of machine learning, significantly contributing to the enhancement of model performance, the mitigation of overfitting, and the augmentation of interpretability. In contrast to hybrid models that amalgamate various methodologies, non-hybrid models principally depend on singular feature selection techniques. Presented below are several of the most efficacious strategies.

- **Statistical Filter Methods :**

Statistical filter methodologies assess features by examining their correlation with the target variable through the application of statistical tests. These approaches are agnostic to specific models and exhibit computational efficiency. Common methodologies encompass Pearson correlation for linear associations, Chi-square tests for categorical variables, and ANOVA F-values for classification tasks. Mutual information serves as additional valuable metric, capturing non-linear inter-dependencies. These methodologies rank features according to their statistical significance, facilitating the selection of the most pertinent features while excluding redundant or irrelevant variables.

- **Variance-Based Feature Elimination :**

A straightforward yet potent technique involves the exclusion of features exhibiting low variance. When a feature demonstrates near-constant values across the dataset, it contributes minimal to no predictive capacity. The Variance Threshold technique discards such features by establishing a minimum variance criterion. This strategy is particularly advantageous in high-dimensional datasets where numerous features may display negligible variability. Nevertheless, it is advisable to integrate this approach with other methodologies, as certain low-variance features may retain informative value in specific contexts.

- **Recursive Feature Elimination (RFE) :**

Recursive Feature Elimination represents a wrapper methodology that systematically eliminates the least significant features based on the performance of the model. It commences with a complete set of features and sequentially discards those that contribute the least, as determined by coefficients (in linear models) or feature importance (in tree-based models). RFE demonstrates particular efficacy when employed with models that furnish feature importance scores, such as logistic regression, support vector machines (with linear kernels), or decision trees. Given that it evaluates subsets of features based on actual model performance, it frequently yields superior outcomes compared to filter methods in isolation.

- **Embedded Methods: L1 Regularization and Tree-Based Importance :**

Embedded methodologies conduct feature selection as an integral component of the model training process. L1 regularisation (Lasso) stands out as a prevalent technique in linear models whereby non-essential features are assigned zero coefficients, thereby effectively excluding them from the model. Additionally, tree-based algorithms such as Random Forest, XGBoost, and LightGBM provide intrinsic feature importance rankings based on metrics like Gini impurity or information gain. These methods are efficient due to their incorporation of feature selection directly into the model training, thereby obviating the necessity for separate pre-processing stages.

- **Permutation Feature Importance :**

Permutation importance is a model-agnostic approach that evaluates feature relevance by permuting each feature and observing the resultant impact on model performance. If the permutation of a feature significantly deteriorates model accuracy, it is deemed important. This methodology is particularly advantageous for black-box models in which traditional coefficient-based importance assessments are not feasible. It offers a robust mechanism for evaluating feature contributions without necessitating assumptions regarding feature distributions.

- **Dimensionality Reduction Techniques :**

Although not strictly classified as feature selection, dimensionality reduction methods such as Factor Analysis facilitate the transformation of features into a lower-dimensional space. These techniques prove beneficial when features exhibit high correlation; however, they compromise the retention of original feature interpretability. In cases where the preservation of feature significance is paramount, conventional selection methods are preferable.

III.7 Conclusion:

This particular chapter has comprehensively elucidated the methodological framework that has been devised for the precise classification of groundwater quality through the implementation of advanced machine learning techniques alongside sophisticated clustering methodologies. The investigation undertook a meticulous and systematic approach, which commenced with the diligent collection and thorough preprocessing of hydrochemical data, encompassing crucial parameters such as pH levels, electrical conductivity measurements, and concentrations of major ions found within the groundwater samples. Subsequently, the K-means clustering algorithm was adeptly employed in order to effectively categorise the various groundwater samples into clearly defined hydrochemical groups, with the optimal number of clusters being ascertained through the application of the elbow method in conjunction with silhouette analysis for enhanced accuracy. Furthermore, a comparative evaluation utilizing hierarchical clustering was conducted to ensure that the classification results achieved were robust and reliable, thus bolstering the integrity of the research outcomes. By seamlessly integrating these multifaceted methods, this study successfully

established a comprehensive data-driven strategy aimed at the assessment of groundwater quality, thereby laying a solid foundation for the subsequent spatial analysis and environmental interpretation that will be elaborated upon in the ensuing chapters

Chapter IV:

Results and Discussion

IV.1. Introduction:

The evaluation of machine learning models for groundwater classification in the Upper and Middle Cheliff plain requires a comprehensive analysis of classification performance across different water regimes. This chapter presents a rigorous examination of the classification results obtained from XGBoost and Support Vector Machine (SVM**)** models applied to the low water period and high water period subsets. The analysis incorporates multiple performance metrics, including accuracy, precision, recall, and F1-score, supported by confusion matrices to provide deeper insights into classification patterns. A comparative assessment of the two algorithms is presented, followed by a discussion of their respective strengths and limitations in the context of classification. Statistical validation and potential sources of error are also examined to contextualize the results.

IV.2.Groundwater quality study:

We have twenty-one (21) and twenty-three (23) analyses for high and low water that are relevant for the year 2022.

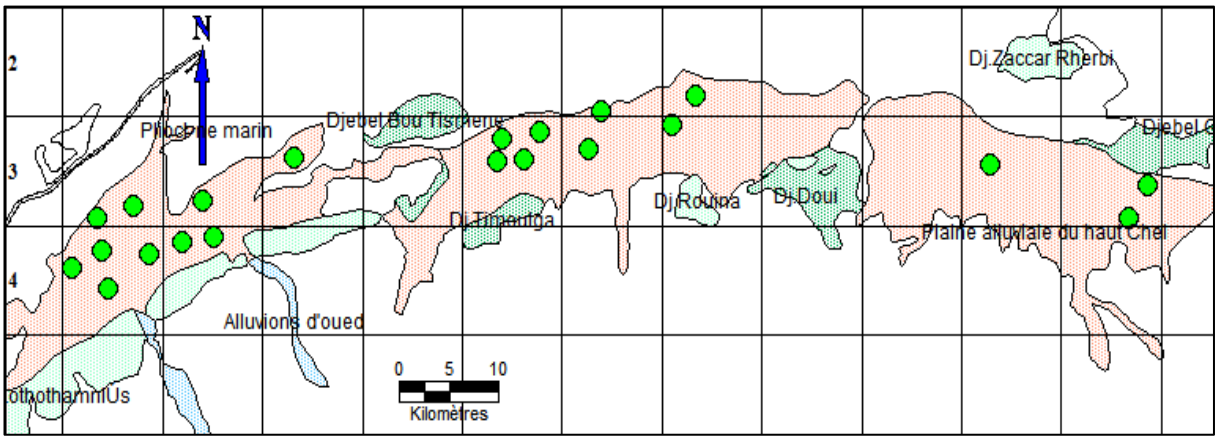

Figure IV. 1a: Sampling plan (High water period, 2022)

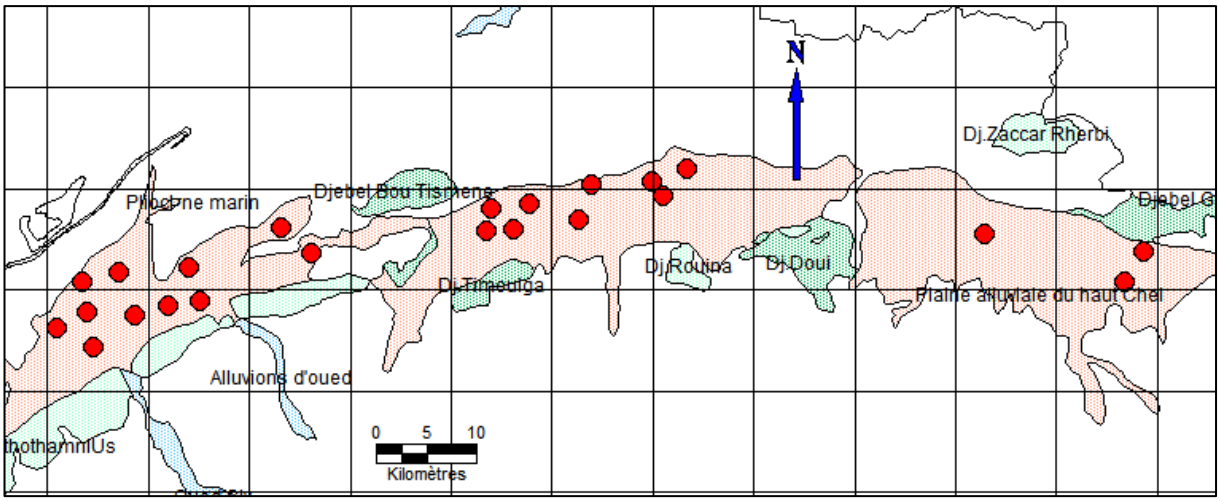

Figure IV. 1b: Sampling plan (Low water period, 2022)

IV.3.Analysis of Water Quality Parameters:

The groundwater hydro chemical properties of the study area are summarized in (Table IV.1) The World Health Organization (WHO,2017) drinking standards, the statistical database of physicochemical parameters, and their assigned weights (w_i) are also illustrated in this table. The results showed that the pH ranged from 6.4 to 9.1 in high water period (the wet season) and from 6.0 to 9.8 in low water period (the dry season). Of the samples, one and two are below the WHO limit for high and low water periods, respectively, while a single water point for each companion is above WHO standards. The concentration of calcium indicated that more than 33.3% and 39.1of samples were below the WHO's norm in both periods respectively, with a mean of about 187.8 and 184.4 mg/l in the wet and dry season, respectively. The highest values were observed in the plain with values ranging from 1048 to 1051 mg/L in low and highwater, respectively. It can be argued that the high value of Ca^{+2} could be due to either the dissolution of carbonate formations ($CaCO_3$), or the dissolution of gypsum formations ($CaSO_4$). For the magnesium, the values are comparable to those of calcium, because they come from the dissolution of carbonate formations with high magnesium contents (magnesite and dolomite) from the Triassic of Ouarsenis (Schrambach, and Mostefa, 1966).

Table IV. 1: Statistics description of water quality (WQ) parameters and assigned weights of each

Parameters	WHO (2017)	Descriptive statistics								Weight (w_i)
		High water				Low water				
		Max	Min	Mean	CV	Max	Min	Mean	CV	
pH	8.5	9.1	6.4	7.9	6.0	9.8	6.0	7.7	9.6	3
EC	1500	15210	980	4049.3	81.9	15300	1017	3876	81.6	5
Na^{+}	150	1600	102	471.3	107.8	1520	118	438.3	165.3	3
Ca^{2+}	100	1051	14	187.8	117.8	1048	8	184.4	114.2	2
Mg^{2+}	75	271	28	110.8	68.4	274	16	108.9	62.4	3
K^{+}	12	21	0	5.8	96.9	20	1	7.6	77.1	2
Cl^{-}	250	5080	184	902.2	125.8	5200	198	875.1	125.5	4
SO_4^{2-}	200	1700	0	411.3	117.7	1245	10	397.1	108.4	4
NO_3^{-}	50	83	2	23.9	90.9	125	0	21.1	77.9	5
HCO^{-}_3	300	860	31	267.3	72.2	769	15	259.9	67.9	1

All values are in mg/l except pH, and EC (μS/cm)

Many sampled water points (57.1% and 69.6 %) provided water with magnesium contents above the drinking water standard of 75 mg/L. The average value for Mg^{+2} was 110.8 and 108.9 mg/L in high and low water respectively. The concentration of sodium shows that about 19 % and 17.4% of the samples were below the standards for drinking water, with a wide variation from 102 to 1600 mg/L and 118 to 1520 mg/L with an average of 471.3 mg/L and 438.3 mg/L in wet and dry season, respectively. These values varied with a decrease in rainfall. In both periods, potassium concentrations exceeded 12 mg/L in two and four wells, respectively, in the plain. One can speculate that this came from the alteration of potassium clays and the dissolution of chemical fertilizers (NPK) which were used by farmers in agriculture. The presence of this may also be associated with wastewater effluents discharge. The mean value of k^{+} was 5.8 and 7.6 mg/L in wet and dry seasons, respectively.

The chloride concentrations showed that 9.5 % and 17.4% of sampled water points were below the WHO's norm with an average ranging from 902.2 to 875.1 mg/L. The highest value (5080 mg/L and 5200 mg/L) was recorded in the wet and dry season(Table IV.2).The elevated content of Cl was presumably due to discharge of chlorinated fertilizers, the dissolution of evaporate deposits (the dissolution of the halite) the permanent interaction of the groundwater with the marly substratum; the association of the marly Miocene soils **(**Elaid et al., 2022) and

the gypsiferous formation. This formation is responsible for the salinity of certain runoff water and consequently for the salinity of aquifers **(**Elaid et al., 2022).

Additionally, the sources of SO_4^{2-} in groundwater can be attributed to dissolution and oxidation of sulphate minerals, discharges from industrial and domestic sewers, as well as leaching of waste deposits. The average values of SO_4^{2-} varied from 397.1 to 411.3 mg/L in low and high water, respectively. Likewise, a few samples had nitrate levels above the standard limit of 50 mg/L.

The sources of nitrate in groundwater were presumably industrial wastewaters, nitrogenous fertilizers, infiltration of surface water, return sewage water, and agricultural sewage usage.

The average values of HCO_3 range from 259.9 to 267.3 mg/l in low and high water, respectively. The concentration of bicarbonate in water depending on the types of soil it crosses (infiltration) or its flow (runoff).

The outcomes showed that most samples (90.5% and 87%) had the highest values of EC and exceeded the WHO's norm (1500 μS/cm).

The high values of EC in the centre of the plain were likely due to the anthropological or man-made pollution of groundwater or due to the water–rock interaction (i.e., the geology of the aquifer). Also, the electrical conductivity was strongly dependent on the chemical composition of water and its temperature (Bouderbala, 2019).

IV.4. Calculation of the groundwater quality index (WQI):

The estimation of the water quality index aims to convert the selected water parameters into a single scaled value, based on the weighted arithmetic indicator to identify the quality of water and its suitability for drinking purposes. Ten parameters (EC, pH, Ca^2, Mg^2 , Na , K , Cl-, SO_4^{2-}, HCO_3^- and NO_3^-) from the high and low water periods in 2022 are chosen to calculate the quality index values for each water point and classify them generally into five categories (Table IV.2) excellent, good, poor, very poor and unsuitable for human consumption.

Table IV. 2: Groundwater classification based on the quality water index

Class	WQI value	Type of water quality
1	<50	Excellent
2	50-100	good,
3	100.1-200	Poor
4	200.1-300	very poor
5	>300	unsuitable for drinking

The computation of the WQI consists firstly in determining the weighted value (*wi*) of each parameter according to their relative influence on the overall quality of drinking water. A weighting value, between 1and5, has been assigned to each factor according to its importance (Bouderbala, 2019). The highest weight of 5was assigned to electrical conductivity (EC) and nitrate, because of their direct effect on drinking water quality. Bicarbonate was given a minimum weight of 1 because it has less importance in the assessment of water quality. The remaining parameters were weighted between 2 and 4 according to their influence on the drinking water quality assessment (Table IV.1).

The relative weight (*wi*) of each parameter was calculated, after determining the quality rating scale (*qi*) for each parameter.

Then, the WQI was calculated by summing the sub-index of all parameters by the following equations:

$$\mathbf{Wi} = \frac{\boldsymbol{wi}}{\sum_{i=1}^{n} \boldsymbol{wi}} \quad \textbf{(1)}$$

$$\mathbf{qi} = \frac{\boldsymbol{Ci}}{\boldsymbol{Si}} * \mathbf{100} \quad \textbf{(2)}$$

$$\boldsymbol{SI_i} = \boldsymbol{W_i} * \boldsymbol{q_i} \quad \textbf{(3)}$$

$$\mathbf{WQI} = \sum_{i=1}^{n} \boldsymbol{SI_i} \quad \textbf{(4)}$$

Where $W_i$is the relative weight, w_i is the parameter weight and n is number of parameters, (q_i) is the quality ranking, (C_i) is the concentration of each chemical parameter in the analysed sample in milligrams per litter, and (S_i) is the admissible limit of each parameter in drinking water according to WHO standard.

IV.5.Assessment of Water Quality using WQI:

In this investigation, WQI values of groundwater samples varied from 37 -616.7 for the high water period(Figure IV.2).The highest WQI values were observed at sample well 21, and the lowest values were observed at sample well 14.

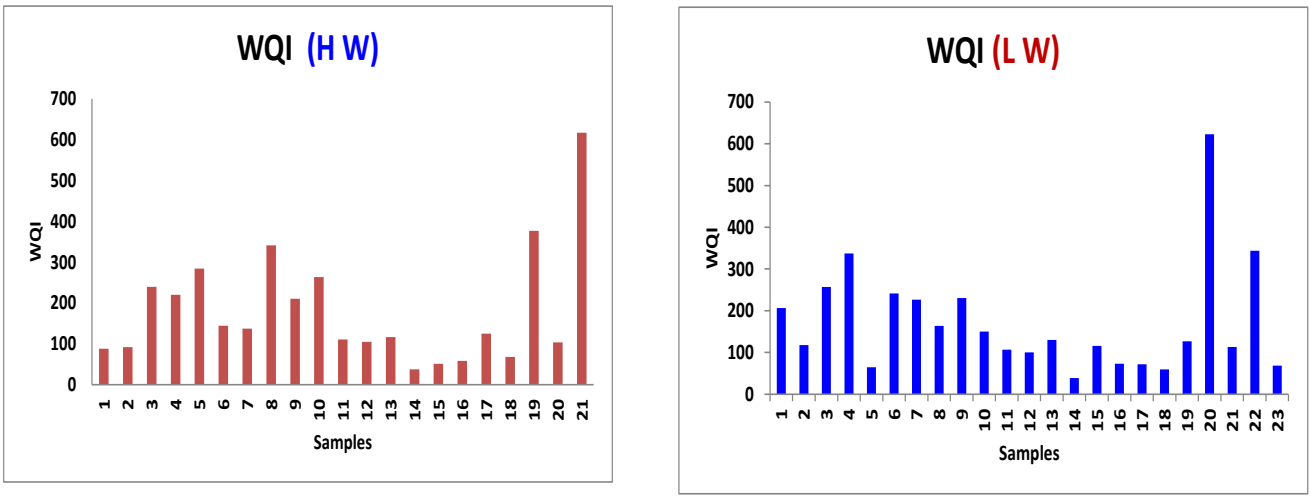

Figure IV. 2: WQI variation for the various sampling points for the High Water (HW) and Low Water (LW) periods.

For low water (FigureIV.2), WQI values ranged from 38.5 to 622.8. The same samples as for the high water had the highest and lowest WQI values.

Figure IV.3 summarizes the potable water quality variation which could be classified from excellent to unsuitable for drinking, with4.8% of water samples in the “excellent” category, 23.8 % as “good”, 33.3 % in the “poor” category , 23.8 % as “very poor”, and 14.3 % as unsuitable for drinking in high water.

In low water, the result was similar to that in high water, with five classes, 4.4% of water samples in the "excellent" category, 26.1% in the "good" category, 34.8% in the "poor" category, 21.7% in the "very poor" category and 13% in the "high" category.

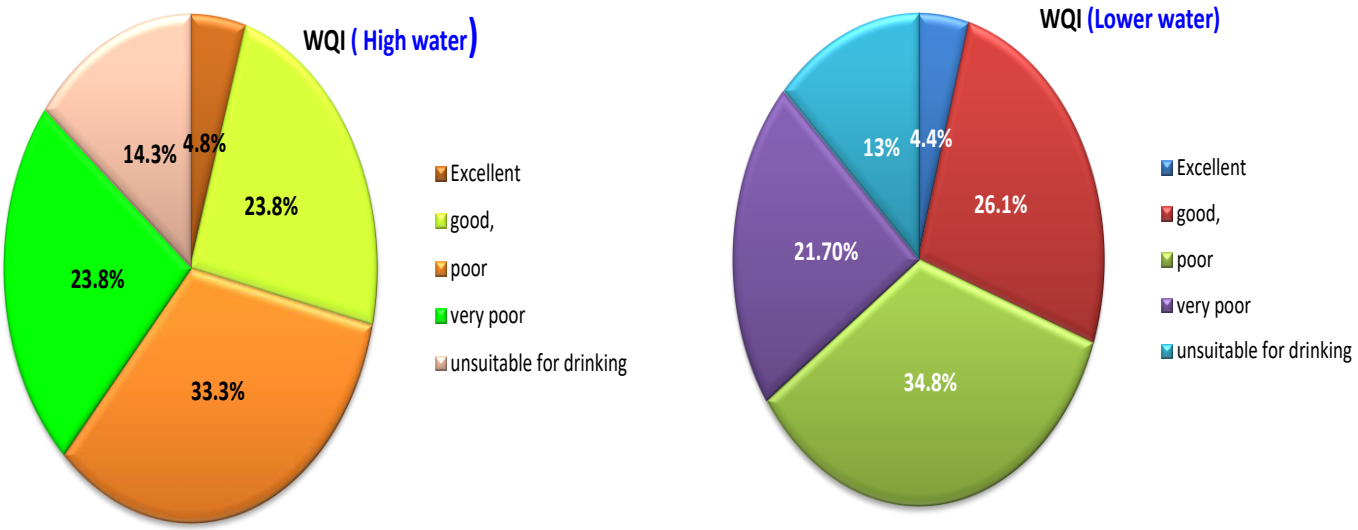

Figure IV. 3: Distribution of WQI by class or category (%) in the case study area (High and lower water periods)

IV.6.Water classification and graphical representations:

To identify hydrochemical facies and get an indication of the quality of groundwater, the graphical representation of analysis results is an essential tool. To achieve this objective, the Piper and Wilcox diagrams were used. These diagrams were produced using the diagram software.

IV.6.1.Piper water classification:

The Piper diagram is particularly suitable for studying changes in water facies as mineralization increases, or for comparing groups of samples and indicating the dominant types of cations and anions.

The representation of the physico-chemical data (21 and 23 samples of the water table for the two campaigns) on the Piper diagram (Figures. IV.4and IV.5) reveals several chemical facies, the origin of which is due to the lithological heterogeneity of the Plio-Quaternary formations forming the alluvial aquifer.

Two facies families have been identified in high water by plotting water chemical analyses on Piper diagrams (Figures IV.4 and IV.5). The calcic chloride facies remains predominant in the groundwater samples (53% of the total samples), The sodium potassium chloride or sodium sulphate facies represent 43% of the sample. One water point belongs to the sodium potassium bicarbonate facies.

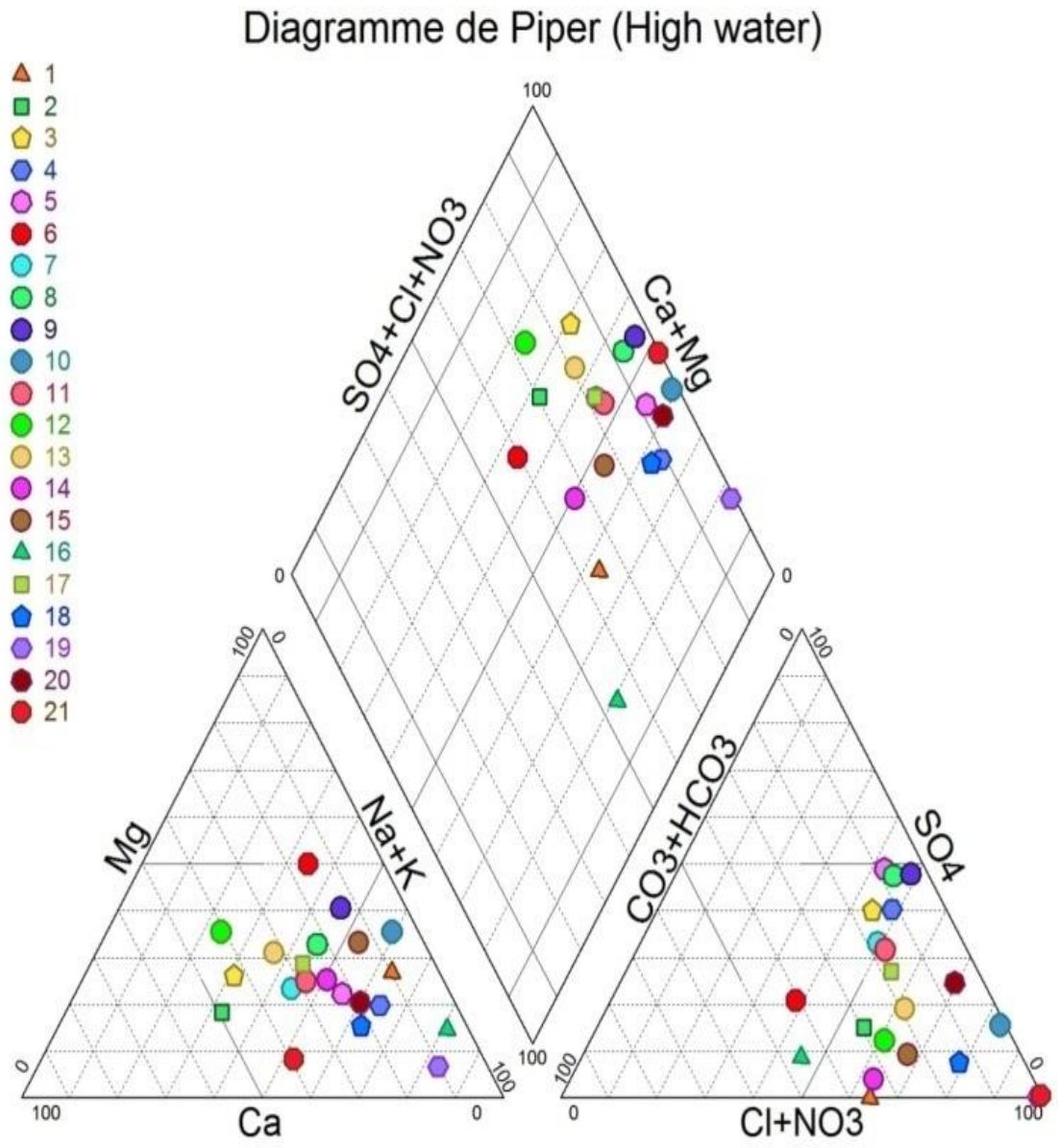

Figure IV. 4: Piper diagram for groundwater (high water period).

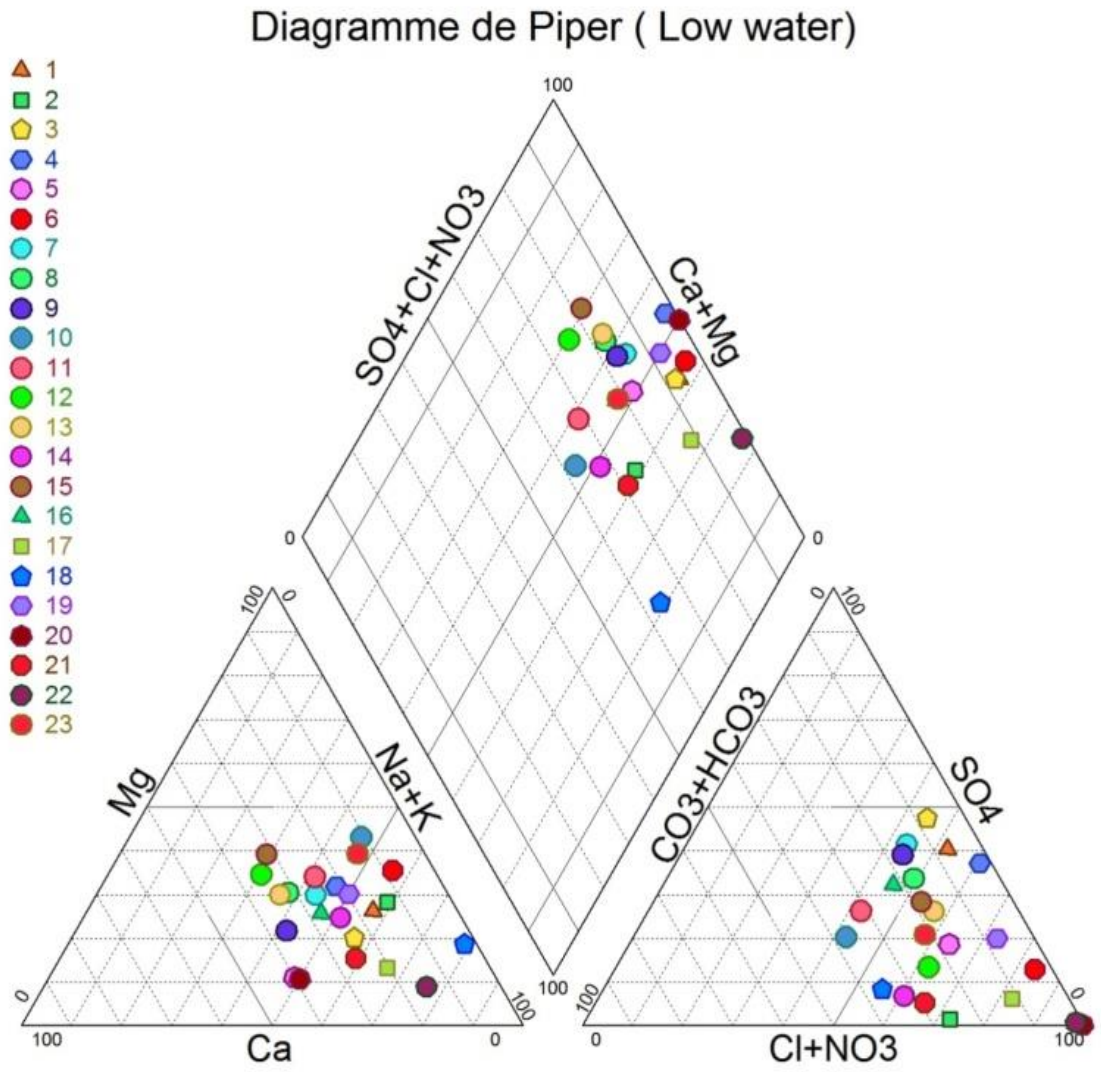

Figure IV. 5: Piper diagram for groundwater (low water period)

IV.6.2.Classification of water WILCOX method:

The Wilcox diagram is based on the Na+ percentage formula, which is given by the following formula:

$$Na\% = (Na /Ca + Mg + Na + k)*100.$$

The Wilcox classification is based on the combination of sodium content in water and electrical conductivity, generally known as %. This classification defines five classes: excellent, good, acceptable, mediocre and poor.

IV.6.2.1. High water period:

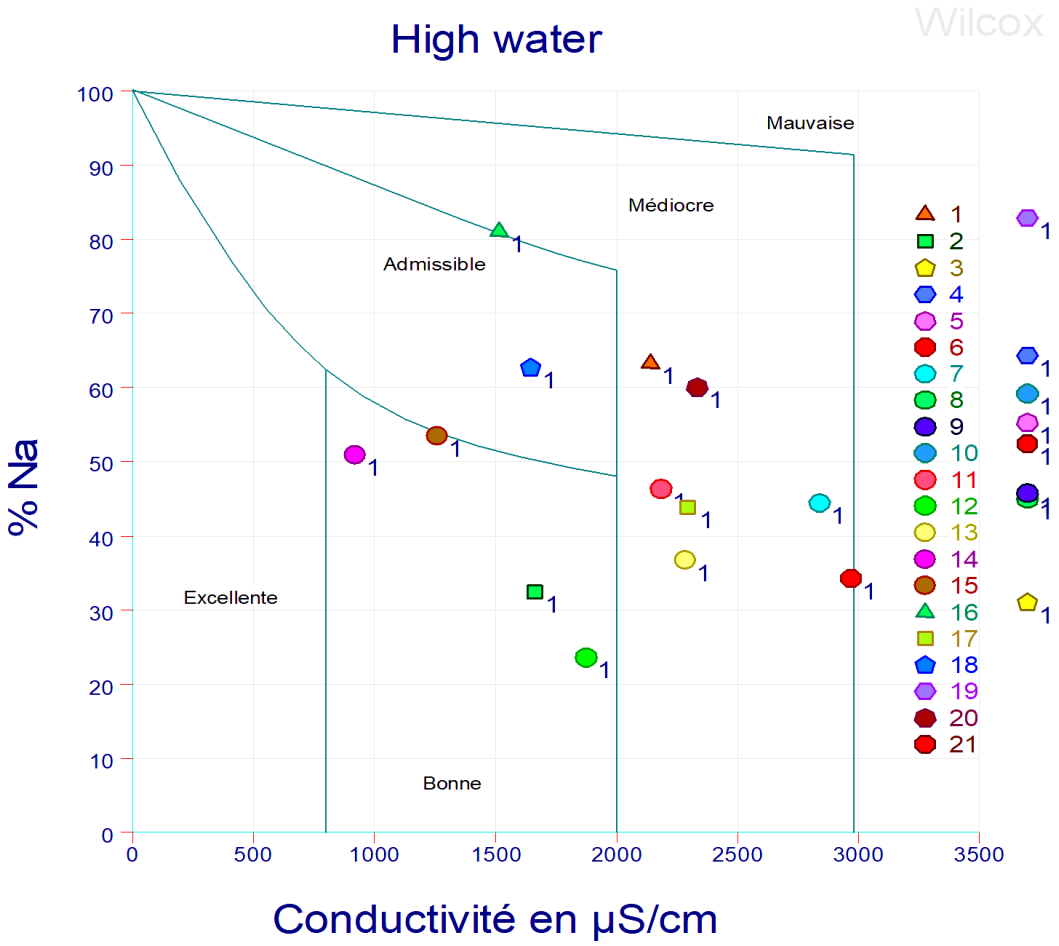

Figure IV. 6: Water quality according to Wilcox (High water period)

According to WILCOX, which classifies water according to its sodium content as a function of electrical conductivity, 15 water points in the aquifer have poor to mediocre quality. On the other hand, six water points are of good to admissible quality (Figure IV.6).

IV.6.2.2. Low water period:

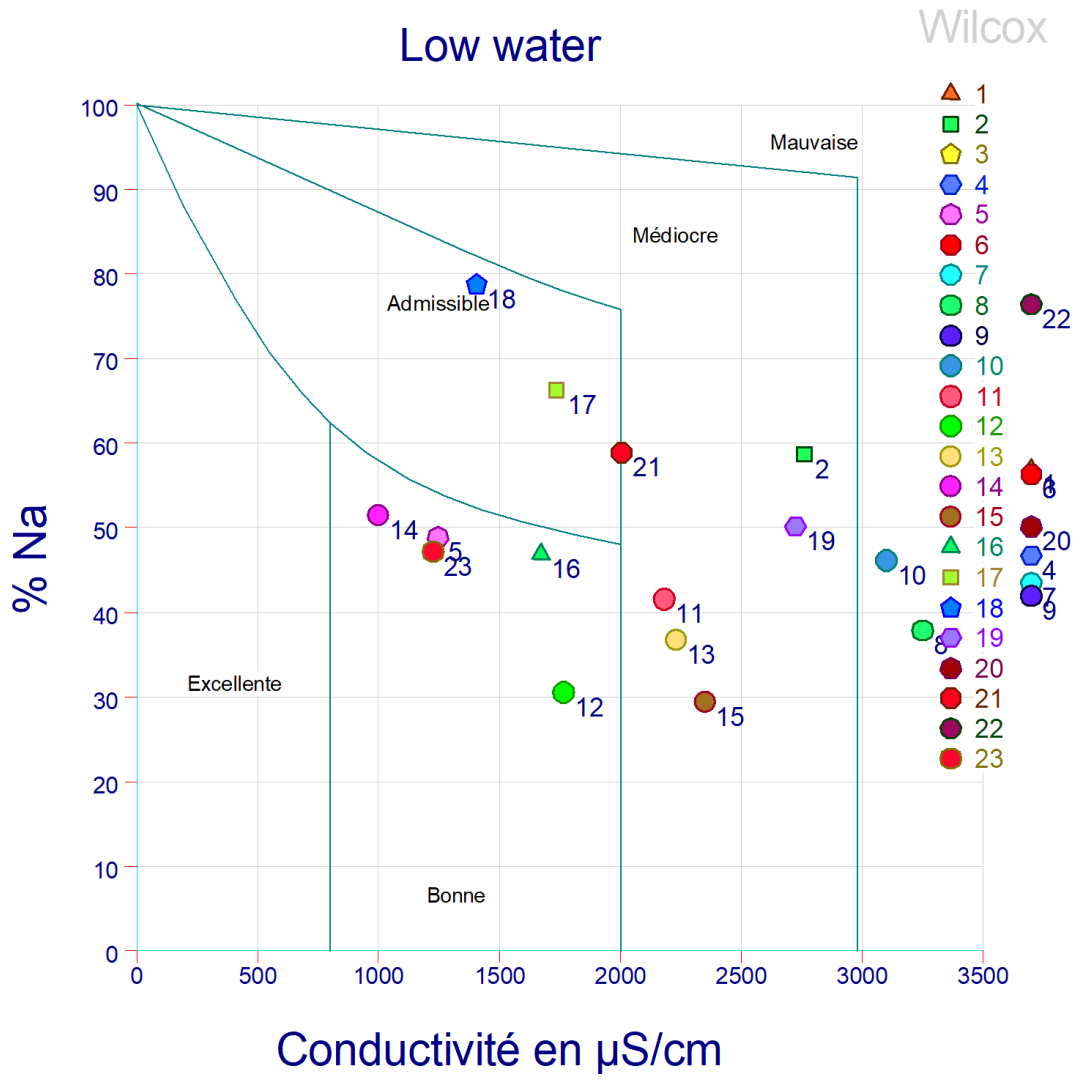

Figure IV. 7: Water quality according to Wilcox (Low water period)

Wilcox's classification system, which categories water according to its sodium content as a function of electrical conductivity, has identified 15 water points in the aquifer that exhibit substandard to merely mediocre quality. On the other hand, eight water points are of good to admissible quality (Figure IV.7) during the dry season.

The high values of conductivity and sodium in groundwater show large proportions of salinity in the water, which reduces its suitability for irrigation and leads to problems when used.

IV.7. Clustering Analysis:

IV.7.1. K-means:

IV.7.1.1 High water period:

Table IV. 3: K-means Contribution of High water period

Observation	DDL (Model)	Mean square (Model)	DDL (Error)	Mean square (Error)	F	Pr> F
pH	4	0.646	16	0.122	5.294	0.007
EC µ/cm	4	52264759.258	16	665664.227	78.515	< 0.0001
Calcium	4	224063.646	16	5120.702	43.756	< 0.0001
Magnesium	4	16997.928	16	2934.256	5.793	0.004
Sodium	4	787218.434	16	32415.308	24.285	< 0.0001
Potassium	4	60.832	16	24.369	2.496	0.084
Chlorides	4	6175266.503	16	66018.202	93.539	< 0.0001
Sulfates	4	984888.787	16	46660.446	21.108	< 0.0001
Nitrates	4	330.706	16	745.976	0.443	0.776
Bicarbonates	4	66126.473	16	29968.775	2.207	0.114

Table IV.3 shows the K-means clustering results, significant differences between clusters for most variables, indicating distinct groupings in the dataset. Electrical Conductivity (EC) shows the strongest separation ($F = 78.515$, $*p* < 0.0001$), suggesting clusters differ markedly in salinity or dissolved ions. Similarly, Calcium ($F = 43.756$), Sodium ($F = 24.285$), Chlorides ($F = 93.539$), and Sulfates ($F = 21.108$) all exhibit highly significant differences ($*p* < 0.0001$), highlighting these as key drivers of cluster divergence. Magnesium ($F = 5.793$, $*p* = 0.004$) and pH ($F = 5.294$, $*p* = 0.007$) also vary meaningfully across clusters, though with less extreme effects.

In contrast, Potassium ($*p* = 0.084$), Nitrates ($*p* = 0.776$), and Bicarbonates ($*p* = 0.114$) do not differ significantly between clusters, implying these variables are either homogeneous across groups or irrelevant to cluster separation. This suggests the K-means algorithm

primarily distinguishes clusters based on salinity-related metrics (EC, Na^+, Cl^-, Ca^{2+}, and SO_4^{2-}) and pH, while other ions (K^+, NO_3^-, HCO_3^-) play negligible roles in the partitioning.

Table IV. 4: K-means clustering classes High water period

Class	pH	EC µ/cm	Calcium	Magnesium	Sodium	Potassium	Chlorides	Sulfates	Nitrates	Bicarbonates
1	7.98	1896	96	57	210	4.9	369	139	29	252
2	7.95	5922	291	178	640	5	924	1255	19	368
3	8.07	3670	148	177	364	5	527	546	13	418
4	7.60	7135	117	158	1169	5	2263	211	31	46
5	6.40	15210	1051	136	1600	21	5080	31	2	31

The Table IV.4 K-means clustering analysis reveals five distinct groups, each characterized by unique water chemistry profiles. Cluster 1 represents samples with moderate mineralization, showing neutral pH (7.98), intermediate electrical conductivity (1896 µS/cm), and balanced ion concentrations, including calcium (95.5), magnesium (57.3), and chlorides (369.5). Cluster 2 stands out with elevated mineralization, evidenced by higher conductivity (5922 µS/cm), calcium (291), and sodium (640.3), along with notably high sulfates (1254.5), suggesting possible sulfate-rich water sources. Cluster 3 displays moderately high conductivity (3670 µS/cm) and magnesium levels (177.3) but with lower sodium (364) and chlorides (527) compared to Cluster 2, indicating a different ionic balance.

Cluster 4 exhibits extreme sodium dominance (1169) and very high chlorides (2262.5), paired with low bicarbonates (46), which could point to saline or brackish water influenced by sodium chloride. Finally, Cluster 5 is the most mineralized group, with exceptionally high conductivity (15210 µS/cm), calcium (1051), and chlorides (5080), but very low sulfates (31) and nitrates (2). This suggests a calcium-chloride-dominated water type, possibly linked to deep groundwater or industrial contamination. The potassium levels are generally low across clusters, except in Cluster 5 (21), which may indicate specific anthropogenic inputs. These clusters effectively differentiate water types based on salinity, major ions, and pH, providing insights for targeted water management or further environmental investigation.

IV.7.1.2 Low water period:

Table IV. 5: K-means Contribution of low water period

Observation	DDL (Model)	Mean Square (Model)	DDL (Error)	Mean Square (Error)	F	Pr> F
pH	4	1.457	18	0.343	4.252	0.013
EC µ/cm	4	54167771.071	18	187130.651	289.465	< 0.0001
Calcium	4	222971.684	18	4709.273	47.347	< 0.0001
magnesium	4	16786.634	18	1919.961	8.743	0.000
Sodium	4	705301.860	18	18834.857	37.447	< 0.0001
Potassium	4	56.980	18	29.420	1.937	0.148
Chlorides	4	6314944.359	18	70150.022	90.021	< 0.0001
Sulfates	4	525112.644	18	76833.668	6.834	0.002
Nitrates	4	903.648	18	1284.846	0.703	0.600
Bicarbonates	4	66319.844	18	23291.513	2.847	0.054

Table IV.5 represents results for the K-means clustering during the low water period reveal significant differences between clusters for most water quality parameters, indicating distinct hydrochemical groupings. The most strongly discriminating variables are electrical conductivity (EC) (F = 289.465, *p* < 0.0001), calcium (F = 47.347, *p* < 0.0001), sodium (F = 37.447, *p* < 0.0001), and chlorides (F = 90.021, *p* < 0.0001), confirming that mineralization patterns (salinity, hardness) are key factors separating the clusters. Magnesium (F = 8.743, *p* = 0.000) and sulfates (F = 6.834, *p* = 0.002) also show significant but weaker differentiation, while pH (F = 4.252, *p* = 0.013) exhibits modest variation across clusters.

Notably, potassium (*p* = 0.148), nitrates (*p* = 0.600), and bicarbonates (*p* = 0.054) do not vary significantly between clusters, suggesting these parameters are either uniformly distributed or unrelated to the clustering logic during low-flow conditions. The dominance of EC and major ions (Ca^{2+}, Na^{+}, Cl^{-}) in defining clusters aligns with typical low-water-period

hydrochemistry, where evaporation and reduced dilution concentrate salts. The weaker role of nutrients (nitrates) and bicarbonates may reflect their seasonal stability or external inputs (e.g., agricultural runoff) affecting all clusters similarly.

Table IV. 6: K-means clustering classes low water period

Class	pH	EC µ/cm	Calcium	magnesium	Sodium	Potassium	Chlorides	Sulfates	Nitrates	Bicarbonates
1	7.8	5180	232	166	601	5.400	937	859	8	299.000
2	7.78	2871	130	115	275	8.86	533	299	22	382.143
3	6.60	7630	271	181	1078	10	2095	638	3.5	73.500
4	7.94	1574	73	42	185	5.75	290	97	36	205.625
5	6.00	15300	1048	171	1520	20	5200	10	0	15.000

Table IV.6 represent the K-means clustering analysis during the low-water period identified five distinct water quality profiles, each characterized by unique hydrochemical signatures. Cluster 1 represents moderately mineralized water with neutral pH (7.8) and elevated conductivity (5180 µS/cm), featuring high concentrations of sodium (601 mg/L), chlorides (937.2 mg/L), and particularly sulfates (858.8 mg/L), suggesting possible sulfate-rich geological influences or anthropogenic inputs. Cluster 2 shows less mineralized water (EC 2871 µS/cm) with relatively balanced ion concentrations, though it stands out with the highest nitrate levels (21.9 mg/L), potentially indicating agricultural runoff impacts. Cluster 3 presents brackish characteristics with very high conductivity (7630 µS/cm), extremely elevated sodium (1077.5 mg/L) and chloride (2095 mg/L) concentrations, and low bicarbonates (73.5 mg/L), likely reflecting saline intrusion or industrial contamination.

Cluster 4 comprises the freshest waters (EC 1573 µS/cm) with the lowest mineralization across all major ions, though it shows the highest nitrate content (35.6 mg/L), possibly from surface runoff. Finally, Cluster 5 represents hyper-mineralized water (EC 15300 µS/cm) with exceptional calcium (1048 mg/L) and chloride (5200 mg/L) levels but negligible sulfates and nitrates, characteristic of deep groundwater or extreme evaporative concentration. These clusters effectively capture the hydrochemical variability during low-flow conditions, highlighting the dominant influences of geological processes, anthropogenic activities, and water-rock interactions in shaping water quality.

IV.8. Machine learning classification results and model performance :

IV.8.1.XGBoost Classification Results

IV.8.1.1High Water period Subset

The XGBoost model exhibited improved generalization for the High water period subset, achieving **82.98% validation accuracy** compared to the Low water period results.

Table IV. 7: XGBoost Classification Report (High Water period – Validation Set)

Class	Precision	Recall	F1-Score	Support
Fair	0.75	1	0.86	**3**
Poor	1	0.67	0.8	**3**
Accuracy	0.83			6
Macro Avg	0.88	0.83	0.83	6
Weighted Avg	0.88	0.83	0.83	6

The confusion matrix (Figure IV.8) reveals perfect recall for the "Fair" class but shows one misclassified "Poor" instance. This pattern suggests the model prioritizes "Fair" classification at the expense of "Poor" detection.

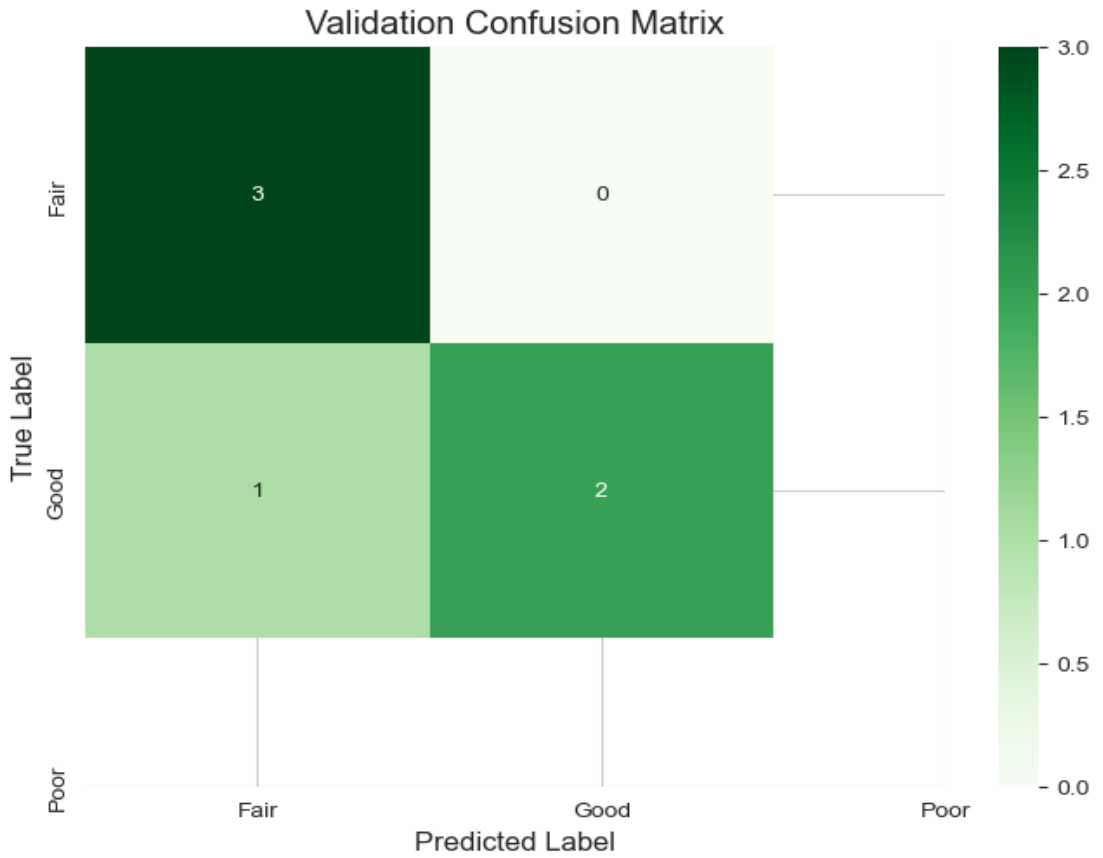

Figure IV. 8: XGBoost Confusion matrix High water validations

The improved performance compared to the High water period subset suggests that hydrological conditions during high-water periods may exhibit more distinct, machine-learnable patterns. However, the persistent class imbalance and small sample size remain limiting factors. The model's higher precision for the "Poor" class (1.00) but lower recall (0.67) indicates a conservative prediction strategy for this category.

IV.8.1.2 Low water period Subset:

The XGBoost model demonstrated exceptional performance on the training data, achieving perfect classification with **100% accuracy** across all categories (Fair, Good, Poor). This suggests strong learning capability and effective capture of training patterns. However, validation performance revealed significant degradation, with accuracy dropping to **66.73%**, indicating potential overfitting.

Table IV. 8: XGBoost Classification Report (Low water – Validation Set)

Class	Precision	Recall	F1-Score	Support
Fair	0.75	0.75	0.75	**4**
Poor	0.5	0.5	0.5	**2**
Accuracy	0.67			6
Macro Avg	0.62	0.62	0.62	6
Weighted Avg	0.67	0.67	0.67	6

The accompanying confusion matrix (Figure IV.9) illustrates that the model correctly classified **3 out of 4 "Fair" instances** and **1 out of 2 "Poor" instances**. The misclassification of one "Poor" instance as "Fair" suggests limited discriminative power for the minority class. The macro-average F1-score of **0.62** further underscores the model's struggle with class imbalance, particularly given the extremely small validation set (n=6).

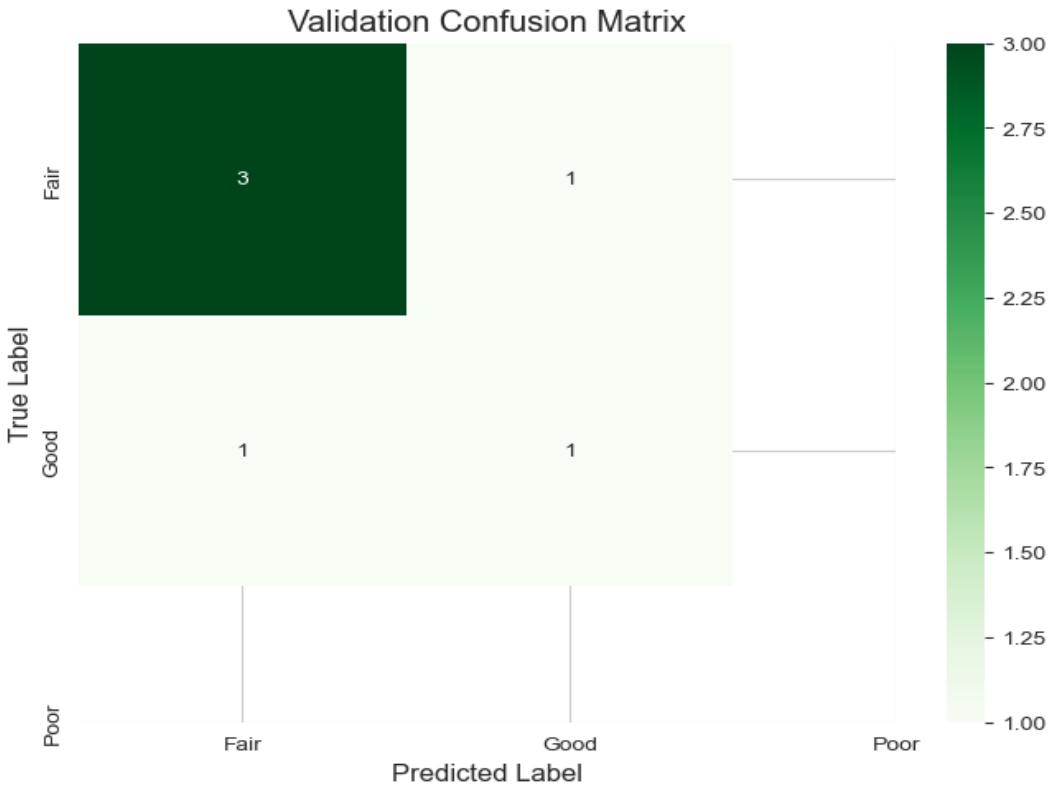

Figure IV. 9: XGBoost Confusion matrix Low water period validation

The substantial suggests **overfitting**, likely exacerbated by the limited dataset. The model appears to have memorized training patterns rather than learning generalizable features. Potential mitigation strategies include: discrepancy between training and validation accuracy strongly **Increasing training data** to improve generalization **Applying regularization techniques** (e.g., higher L2 regularization, early stopping) **Implementing cross-validation** to better estimate true performance.

IV.8.2.Feature importance XGboost:

IV.8.2.1. High water period:

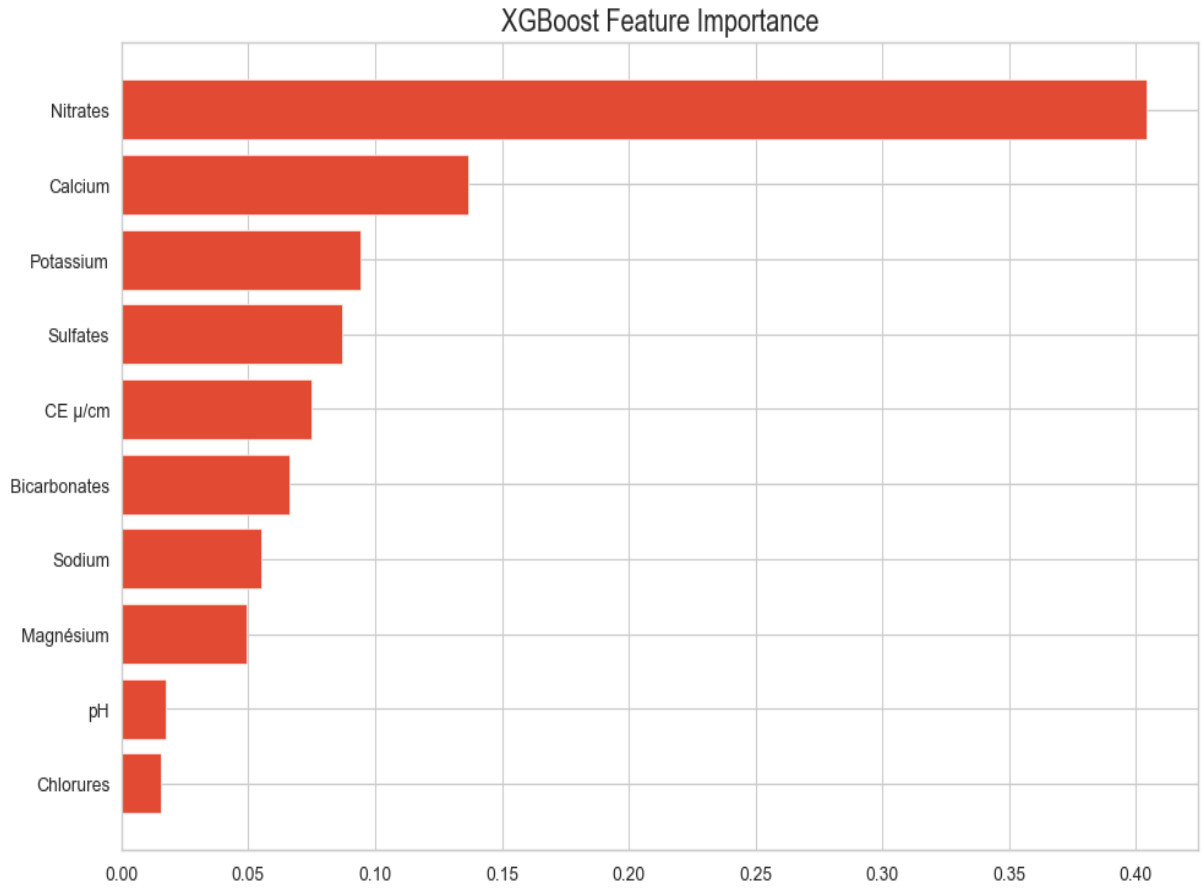

Figure IV. 10: XGBoost feature importance high water period

During the period of high water levels, the XGBoost feature importance analysis indicates that nitrates are the most significant factor, boasting a considerably higher importance score (approximately 0.35) in comparison to the other features. This implies that nitrate levels are

crucial in determining the target variable, likely owing to increased agricultural runoff or other seasonal elements that raise nitrate concentrations during times of high water. Calcium, potassium, and sulfates also demonstrate moderate importance, signifying their secondary yet still noteworthy influence on the model's predictions. Conversely, features such as bicarbonates, sodium, magnesium, pH, and chlorides exert minimal influence, with importance scores falling below 0.10. This suggests that these parameters may not fluctuate significantly during high water periods or possess a weaker correlation with the target variable. For the purposes of modelling or environmental management, concentrating on nitrates, calcium, potassium, and sulfates would prove to be the most beneficial, whilst the less significant features could

IV.8.2.2. Low water Period:

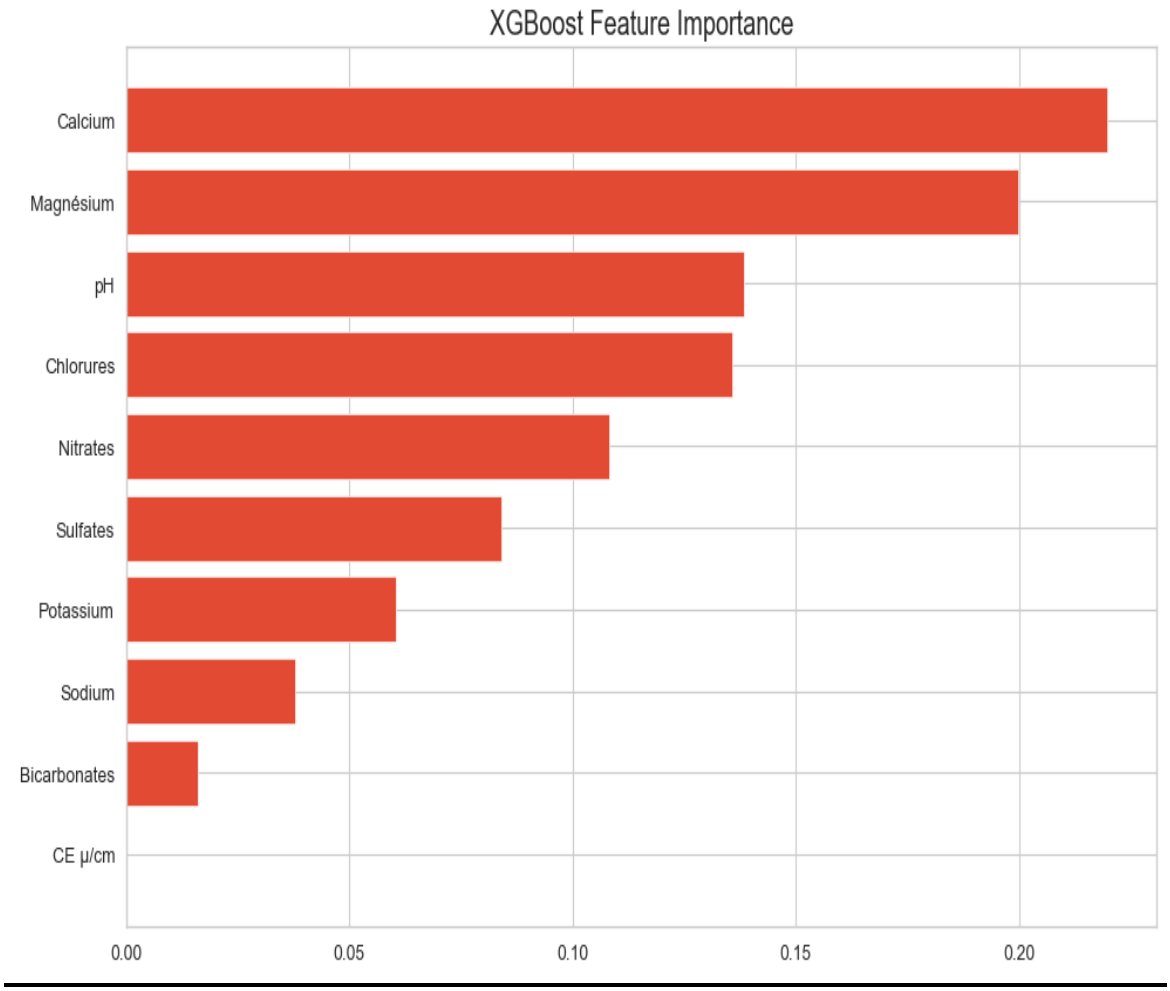

Figure IV. 11: XGBoost feature importance low water period

During the period of low water, the analysis of feature importance using XGBoost reveals that calcium is the most significant factor, attaining the highest importance score (approximately 0.20), indicating its vital role in affecting the target variable under these circumstances. This may be attributed to diminished dilution effects, thereby rendering calcium concentrations more discernible. Following closely are magnesium and pH, which exhibit moderate importance, signifying their secondary yet still pertinent contributions, potentially associated with mineral dissolution and the stability of water chemistry in conditions of low flow. Conversely, features such as nitrates, sulfates, and potassium display reduced importance, which may indicate less agricultural runoff or slower biogeochemical processes during dryer intervals. The features with the least influence sodium, bicarbonates, and EC µ/cm (electrical conductivity) have a negligible impact, likely due to their variability being less pronounced or less predictive in scenarios of low water. When it comes to modelling or assessing water quality during this period, placing emphasis on calcium, magnesium, and pH would prove to be the most efficacious, while other features could be relegated to a lower priority to simplify analyses without incurring a significant loss in accuracy. This alteration in feature importance in comparison to conditions of high water highlights how seasonal hydrology modifies the key drivers of water quality dynamics.

IV.9. SVM Classification Results:

IV.9.1. High water period Subset:

SVM performance degraded significantly in the **High water period** subset, achieving only **66.86% validation accuracy**.

Table IV. 9: SVM Classification Report (High water period – Validation Set)

Class	Precision	Recall	F1-Score	Support
Fair	0.6	1	0.75	**3**
Poor	1	0.33	0.5	**3**
Accuracy	0.67			6
Macro Avg	0.8	0.67	0.62	6
WeightedAvg	0.8	0.67	0.62	6

The SVM's poorer performance compared to XGBoost in this subset suggests that the data's decision boundaries may be more complex than what a linear SVM can effectively capture. The perfect recall for "Fair" but poor "Poor" recall (0.33) again shows the model's tendency to favor the majority class.

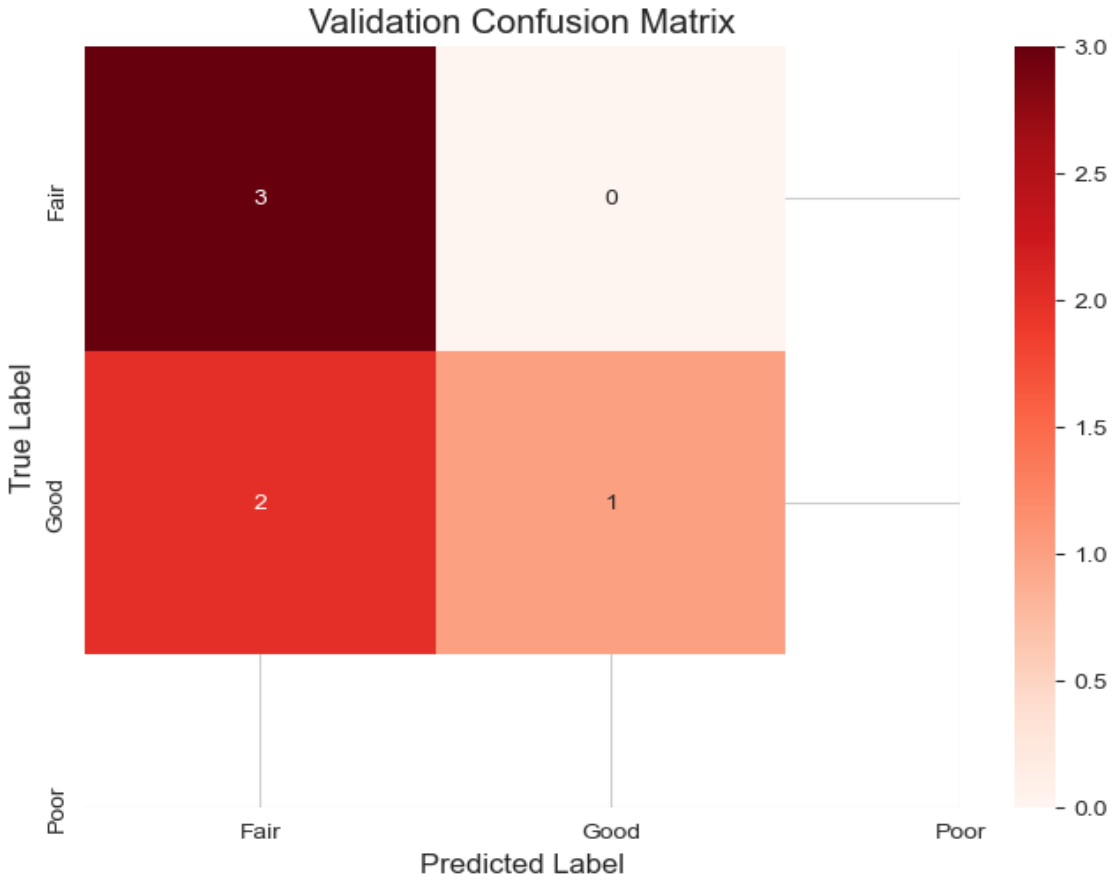

Figure IV. 12: SVM Confusion matrix High water period validation

IV.9.2. Low water period Subset:

The SVM model achieved **81.43% training accuracy**, demonstrating substantial variability across classes. Notably, the model completely failed to classify the "Good" category, which comprised only a single training instance.

Table IV. 10: SVM Classification Report (Low water period – Training Set)

Class	Precision	Recall	F1-Score	Support
Fair	0.8	0.89	0.84	**9**
Good	0	0	0	**1**
Poor	0.86	0.86	0.86	**7**
Accuracy	0.82			17
Macro Avg	0.55	0.58	0.57	17
Weighted Avg	0.78	0.82	0.8	17

Validation performance slightly improved to **82.79% accuracy**, an unusual case where validation exceeds training performance, potentially indicating a favourable data split as shows confusion matrix (Figure IV.13).

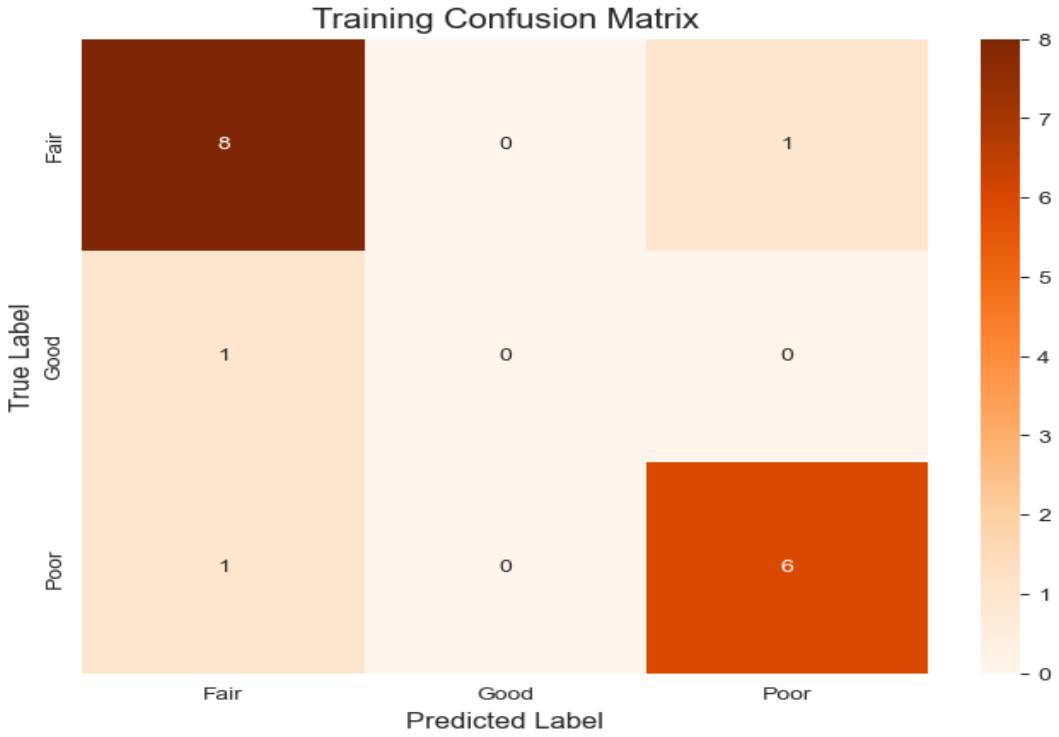

Figure IV. 13: SVM Confusion matrix Low water period Training

Table IV. 11: SVM Classification Report (Low water period – Validation Set)

Class	Precision	Recall	F1-Score	Support
Fair	0.8	1	0.89	**4**
Poor	1	0.5	0.67	**2**
Accuracy	0.83			6
Macro Avg	0.9	0.75	0.78	6
Weighted Avg	0.87	0.83	0.81	6

The SVM's relatively stable performance across training and validation sets suggests better generalization than XGBoost for this subset. However, its complete failure on the "Good" class highlights SVM's sensitivity to extreme class imbalance. The high recall for "Fair" (1.00) comes at the cost of reduced "Poor" recall (0.50), indicating the model defaults to predicting the majority class when uncertain as shown in confusion matrix (Figure IV.14).

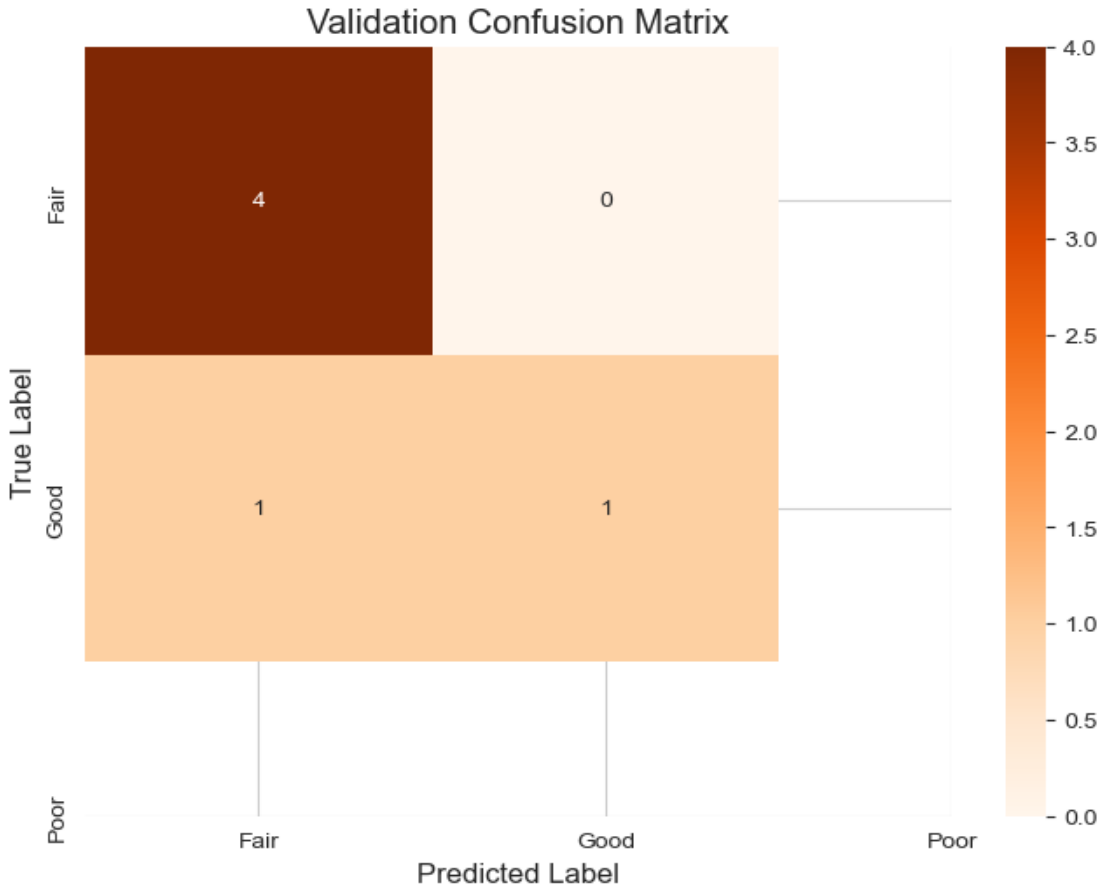

Figure IV. 14: SVM Confusion matrix Low water period validation

IV.10.Comparative Analysis:

IV.10.1. Algorithm Performance Comparison:

Table IV. 12: Model Comparison across Subsets

Metric	XGBoost (Low)	XGBoost (High)	SVM (Low)	SVM (Highs)
Val. Accuracy	66.73%	82.98%	82.79%	66.86%
Macro F1	0.62	0.83	0.78	0.62
Worst Class F1	0.50 (Poor)	0.80 (Poor)	0.67 (Poor)	0.50 (Poor)
Overfitting	Severe	Moderate	Minimal	Moderate

The comparative examination of XGBoost and SVM models during periods of high and low water unveiled clear performance patterns. XGBoost attained a higher validation accuracy of

82.98% in the high water phase, as opposed to its lower water performance of 66.73%, alongside a corresponding Macro F1-score of 0.83 and 0.62, respectively. Remarkably, its worst-class F1-score for "Poor" samples improved from 0.50 (low water) to 0.80 (high water), signifying enhanced recognition of minority classes under high water circumstances. Nevertheless, XGBoost displayed significant overfitting within the low water subset, presumably due to restricted training data. Conversely, SVM exhibited more consistent generalisation, achieving a validation accuracy of 82.79% during the low water period with minimal overfitting, although its performance in high water declined sharply to 66.86% accuracy (0.62 Macro F1). Both models encountered difficulties with the "Poor" class in high water (F1=0.50), indicating ongoing challenges in classifying imbalanced datasets. These findings underscore XGBoost's superior capability in managing intricate patterns in high water conditions, whereas SVM's resilience against overfitting renders it more dependable for low water situations. Future research ought to tackle data limitations and class imbalance to further improve model reliability.

IV.11. Conclusion:

This chapter presented a detailed evaluation of groundwater quality classification in the Upper and Middle Cheliff plain using **XGBoost** and **SVM** models across high and low water periods. The results demonstrated that **XGBoost** achieved higher validation accuracy (82.98%) during the high water period, with nitrates as the most influential feature, while calcium dominated in the low water phase. However, overfitting was observed in the low water subset, reducing its validation accuracy to 66.73%. In contrast, **SVM** exhibited better generalization in the low water period (82.79%) but struggled during high water (66.86%), suggesting limitations in capturing complex decision boundaries. The **Water Quality Index (WQI)** analysis revealed that most samples fell into "poor" to "unsuitable" categories, emphasizing seasonal variability and anthropogenic impacts. Graphical methods, including Piper and Wilcox diagrams, further clarified hydrochemical facies and quality trends. Despite their strengths, both models faced challenges due to class imbalance and limited data. Overall, the findings highlight the effectiveness of machine learning in groundwater classification while underscoring the need for larger datasets and refined modelling techniques to enhance predictive accuracy and generalizability.

General conclusion

General conclusion:

This work presents a comprehensive assessment of groundwater quality in the Upper and Middle Cheliff plains in Algeria, through an integrated approach that combines water quality indices (WQI), and machine learning techniques.

The hydrochemical study was carried out to characterise the physico-chemical quality of water, which is an essential step in determining its suitability for various uses. The calculation of WQI revealed five classes as 4.8% and 23.8 % of the samples were classified as excellent and good category, respectively. However, poor category is the most dominant with 33.3 % of the samples 23.8 % as "very poor", and 14.3 % as unsuitable for drinking in high water. A similar result was observed in low water. WQI analysis revealed that 21.7–34.8% of samples were categorised as 'poor' to 'unsuitable' for drinking, primarily due to high levels of electrical conductivity, nitrate and chloride. The analysis also revealed higher mineralisation during low-water periods due to reduced dilution.

Two dominant facies were identified by Piper diagrams: calcic chloride (53%) and sodium–potassium chloride/sulfate (43%), indicating significant water–rock interactions and anthropogenic influences.

The presence of salinity and sodium risks was the primary cause of unsuitability or marginality for irrigation in 65% of the samples, as demonstrated by Wilcox.

Machine learning analysis employed both supervised (XGBoost, SVM) and unsupervised (K-means) models. Clustering revealed three natural water quality groupings, which closely aligned with WQI categories. XGBoost performed best during high-water periods (82.98% accuracy), while SVM showed better generalization in low-water conditions (82.79%). Nitrate was the most influential predictor in high-water periods, while calcium dominated in low-water conditions, reflecting seasonal variations in contaminant mobilisation.

Key contributions include:

- Integrating water quality index with machine learning for classification .
- Demonstrated utility of ensemble models for capturing and classifying non-linear water quality patterns and data.

Bibliographic References

References

ABH-CZ (2004) : Cadastre Hydraulique du bassin hydrographie du Cheliff-Aval du barrage de Boughzoul – Première partie : Haut et moyen Cheliff, 62 p.

Achour, F., &Bouzelboudjen, M. (1998). Variabilité spatio-temporelle des ressources en eau en région semi-aride : application au bassin du Cheliff, Algérie*. In *Proceedings of the IAHS Abidjan'98 Conference (IAHS Publ. 252). Abidjan, Côte d'Ivoire.

Agence Nationale des Ressources Hydrauliques (ANRH). (2004). Annuaire hydro-géologique de la nappe alluviale du Haut et Moyen Cheliff [Hydrogeologicalyearbook]. Algiers, Algeria.

ATHAMENA.M,(2006).Etude des ressources thermales de l'ensemble Sud sétifien.Algerie.Mémoire de Magistère. Univ Batna. 131p

Azizpor A, Izadbakhsh MA, Shabanlou S, Yosefvand F, Rajabi A (2021) Estimation of water level fuctuations in groundwater through a Hybrid Learning Machine. Groundw Sustain Dev 15:100687. https://doi.org/10.1016/j.gsd.2021.100687.

BRGM,(2007).Suivi de la qualité des eaux souterraines de Martinique , compagne de saison despluies2006 : Résultats et interprétation.

Buhmann, J. M., Hofmann, T., &Buhmann, J. (1999).Unsupervised learning and exploratory data analysis. In M. Jordan, M. J. Kearns, & S. A. Solla (Eds.), Advances in Neural Information Processing Systems (Vol. 11, pp. 531–537). MIT Press.

Chen, T., &Guestrin, C. (2016).XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).ACM. https://doi.org/10.1145/2939672.2939785

Chinchor, N. (1992). MUC-4 evaluation metrics.In Proceedings of the 4th Message Understanding Conference (MUC-4) (pp. 22–29).DARPA.

Cortes, C., &Vapnik, V. (1995).Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Bibliographic References

Davis, J., &Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (pp. 233–240).ACM. https://doi.org/10.1145/1143844.1143874

El Baba M, Kayastha P, Huysmans M, De Smedt F (2020) Evaluation of the groundwater quality using the water quality index and geostatistical analysis in the dier al-balah governorate, gaza strip. Palestine Water 12:262. https://doi.org/10.3390/w12010262

Elmeddahi Y, Mahmoudi H, Issaadi A, Goosen GMFA, Ragab R (2016) Evaluating the efects of climate change and variability on water resources: a case study of the Chelif Basin in Algeria. AJEAS 9:835–845. https://doi.org/10.3844/ajeassp.2016.835.845;

ELMORHIT.M, (2009). Hydrochimie, éléments traces métalliques et incidences écotoxicologiques sur les différentes composantes d'un écosysthème estuarien(Basloukkos). Thèse de doctorat. Univ Mouhammed V. Agdal, Rabat,232p.

Gaujour D.,(1995).La pollution des milieux aquatiques :Aidemémoire .2émeédition ,Lavoisier, P49.

Huang, C. L., Chen, M. C., & Wang, C. J. (2006). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), 847–856. https://doi.org/10.1016/j.eswa.2006.07.018

Islam Khan SM, Islam N, Uddin J, Islam S, Nasir MK (2021) Water quality prediction and classifcation based on principal component regression and gradient boosting classifer approach. J King Saud Univ - ComputInf Sci. https://doi.org/10.1016/j.jksuci.2021.06.003

Jean bontoux.M,(1993). Introduction à l'étude des eaux douces: Eaux naturelles, Eaux usées, Eaux de boisson; Qualité et santé, Editions CEBEDOC, Liège, 1993, p 79.

Khellili R., Lazali D. (2015). Etude des propriétés physico-chimiques et bactériologiques de l'eau du barrage Harraza (Wilaya de Ain Defla) (2005).

Kumar A, Ramsankaran R, Brocca L, Munoz-Arriola F (2021) A simple machine learning approach to model real-time streamfow using satellite inputs: demonstration in a data scarce catchment. J Hydrol 595:126046. https://doi.org/10.1016/j.jhydrol.2021. 126046

Mania, J., &Djeda, F. (1990).Hydrogeological study of the Upper Cheliff alluvial plain (Khemis-Miliana, Algeria) Ministry of Water Resources Report, Algeria.

Mattauer, M. (1958).Geological contributions to the region of Zaccar and Jebel Doui [In French].Unpublished manuscript, Geological Survey Archives, Algeria.

Bibliographic References

Meghraoui, M., El HadjLarbi, A., & Co-authors.(1986). *Etude géologique et structurale de la région de Zaccar-Ouarsenis (Algérie)*. *Bulletin Centre de Recherches Géologiques et Minières, Série Géologie, Alger.

Mutin G (2009): Le Monde arabe face au défi de l'eau, Enjeux et Conflits. Institut d'Etudes Politiques de Lyon, 164.

P.D.A.R.E **(**2011) : Région hydrographique Cheliff- Zahrez, rapport de synthèse, Plan Directeur d'Aménagement des Ressources en Eaux, 193p.

Perrodon, A. (1957). Contribution à l'étude géologique du Cheliff (Algérie).,University of AlgiersGeological Archives.

Phogat,V.,Skewes,M.A.,Cox,J.W.,Sanderson,G.,Alam,J.,&Šimůnek,J.(2014).Seasonal simulation of water, salinity and nitrate dynamics under drip irrigated mandarin (Citrus reticulata) and assessing management options for drainage and nitrate leaching. Journal of Hydrology, 513, 504-516.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Sedrati N.,(2011).Origines et caractéristiques physico-chimiques des eaux de la wilaya de Biskra-sudest Algérien, thèse de doctorat en géologie, Hydrogéologie, faculté des sciences de la terre, département de géologie, Université Badji Mokhtar-Annaba, 252p.

Sharma, A. (2024). Foundations of supervised machine learning for engineering and science.Springer.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing& Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Touahria, K.(2013)Evaluation De La Qualite Des Eaux De Forages Par Comparaison De Leurs Caracteristiques Physico-Chimiques (Region De Tebessa), Master'sthesis, University of Souk Ahras.

Bibliographic References