الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
Tasdawit Akli Muḥend Ulḥağ - Tubirett -

وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
-البويرة-

Faculté des Sciences et des Sciences Appliquées

كلية العلوم والعلوم التطبيقية

Référence : ...../MM/2021

المرجع:...../م م/ 2021

Mémoire de Master

Présenté au

Département : Génie Électrique
Domaine : Sciences et Technologies
Filière : Electronique
Spécialité : Electronique des systèmes embarqués

Réalisé par :

DAOUD Lylia

Et

SARRI Asma

Thème

Employing an artificial neural network for the prediction of water quality parameters

Soutenu le: **2 / 7 / 2024**

Devant la commission composée de :

Mr :	ALIMOHAD Abdennour	M.C.B	Univ. Bouira	Président
	REZKI Mohamed	M.C.A	Univ. Bouira	Rapporteur
	LADJOUZI Samir	M.C.B	Univ. Bouira	Examinateur

Aknowledgement

We would thank God for giving us the health and the courage to finish this work and without his reconciliation this project will not see daylight.

We would also like to express our gratitude to Mr.REZKI Mohamed, who agreed to supervise this dissertation and offered us invaluable advice. Your expertise, enthusiasm and commitment to us success have been a source of inspiration throughout this process.

We would also like to express our gratitude to our colleagues and friends who have supported and encouraged our efforts throughout this academic adventure. Your moral support and enriching discussions have been invaluable to us.

Finally, we would like to express our gratitude to our families for their constant support and unwavering encouragement throughout our studies. Your love, trust and encouragement have been our deepest inspiration.

Without the help and support of each and every one of you, the completion of this Master's thesis would not have been possible. We are grateful for all the opportunities for learning and growth that we have had thanks to you.

Once again, thank you from the bottom of my heart for your invaluable support.

Without forgettind our propre efforts and the all-nighters that we pulled together to achieve the best we could give to this dissertation.

Abstract

Water is an essential factor for human life, and any dysfunction may seriously threaten the environment and health, so the World Health Organization has set water standards, yet there is the problem of poor water quality.

And that brings us to our goal, which is to use synthetic neural networks to anticipate water quality (in our case is pH), where it involves using a computational model and that is through a processing data. And doing her training, so that it contains inputs (such as temperature,turbidity… etc) and outputs (such as pH, potability). Then it is done validating the ANN model and splitting the available data into training and testing sets, which are techniques used to evaluate the predictive accuracy of the model.

Once the ANN model has been trained and validated, it can then be used to predict water quality parameters in real-time or for future scenarios. These predictive capabilities provided by the ANN model are valuable for environmental monitoring, informed decision-making in water resource management and early detection of potential pollution events.

Keywords : Water quality, parameters, prediction, ANN, PH.

Résumé

L’eau est un facteur essentiel pour la vie humaine et tout dysfonctionnement peut sérieusement menacer l’environnement et la santé, de sorte que l’Organisation mondiale de la santé a fixé des normes d’eau, mais il y a le problème de la mauvaise qualité de l’eau.

Et cela nous amène à notre objectif, qui est d’utiliser des réseaux neuronaux synthétiques pour anticiper la qualité de l’eau (dans notre cas, le pH), où il s’agit d’utiliser un modèle informatique et c’est à travers des données de traitement. Et faire sa formation, afin qu’elle contienne des entrées (comme la température, la turbidité…ect) et des sorties (comme le pH, la potabilité). Ensuite, il a fallu valider le modèle ANN et diviser les données disponibles en ensembles de formation et de test, qui sont des techniques utilisées pour évaluer la précision prédictive du modèle.

Une fois que le modèle ANN a été formé et validé, il peut ensuite être utilisé pour prédire les paramètres de la qualité de l’eau en temps réel ou pour des scénarios futurs. Ces capacités prédictives fournies par le modèle ANN sont précieuses pour la surveillance environnementale,

la prise de décisions éclairées en matière de gestion des ressources en eau et la détection précoce des événements de pollution potentiels.

Mots-clés : Qualité de l'eau, paramètres, prédiction, ANN, PH.

الملخص

الماء عامل أساسي لحياة الإنسان وأي خلل وظيفي قد يهدد البيئة والصحة بشكل خطير، لذلك وضعت منظمة الصحة العالمية معايير للمياه، ومع ذلك هناك مشكلة رداءة نوعية المياه.

وهذا يقودنا إلى هدفنا وهو استخدام الشبكات العصبية الاصطناعية لتوقع جودة المياه (في حالتنا الأس الهيدروجيني)، حيث يتضمن استخدام نموذج حسابي، وذلك من خلال قاعدة بيانات معالجة . و القيام بتدريبها ، بحيث تحتوي على مدخلات(مثل درجة الحرارة، العكارة...الخ) والمخرجات (مثل الأس الهيدروجيني، قابلية الشرب). ثم يتم التحقق من صحة نموذج ANN وتقسيم البيانات المتاحة إلى مجموعات تدريب واختبار، وهي تقنيات تستخدم لتقييم الدقة التنبؤية للنموذج.

بمجرد تدريب نموذج ANN والتحقق من صحته، يمكن استخدامه بعد ذلك للتنبؤ بمعلمات جودة المياه في الوقت الفعلي أو للسيناريوهات المستقبلية. هذه القدرات التنبؤية التي يوفرها نموذج ANN ذات قيمة للرصد البيئي، واتخاذ القرارات المستنيرة في إدارة موارد المياه، والكشف المبكر عن أحداث التلوث المحتملة.

الكلمات الرئيسية : جودة المياه، المعلمات، التنبؤ، ANN، PH.

Table of contents

Chapter2: Artificial Neural Networks (ANNs)

List of figures

List of Tables

List of Acronyms and Symbols

▪ Acronyms

ACC	Accuracy
AI	Artificial intelligence
ANN	Artificial Neural Networks
°C	Celsius
Ca^{2+}	Calcium
Cl^-	Clorides
CNN	Convolutional Neural Network
CO_2	Carbon dioxide
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DL	Deep learning
DNN	Deep neural networks
EC	Electrical conductivity
ECG	Electrocardiogram
EM	Expectation Maximization
FP	False positive
F1	Score
GMM	Gaussian Mixture Models
HCO^-	Bicarbonates
$H3O^+$	Hydronium
JORADP	Journal officiel de la République algérienne démocratique et populaire
K^+	Potassium
KCl	Potassium chloride
KNN	K-Nearest Neighbors
MADs	Maximum allowable doses

MAE	Mean Absolute Error
Mg^{2+}	Magnesium
Mg/L	Milligram per liter
ML	Machine learning
MLP	Multilayer perceptron
Mmhos/cm	Millimhos per centimeter
MSE	Mean Squared Error
Ms/cm	Millisiemens per centimeter
Na^{+}	Sodium
NaCl	Sodium salts
NO^{3-}	Nitrates
NTU	Nephelometric Turbidity Unit
P	Precision
PH	Hydrogen potential
PPM	Parts per million
R	Recall
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks
S.cm	Meter siemens per centimeter
S.m	Siemens per meter
SO_4^{2-}	Sulfate
SVC	Support Vector Classifier
SVM	Support vector machine
SVR	Support Vector Regression
TDs	Total dissolved solids
TH	Total hardness
TN	True negative

TP	True positive
Umhos/c	Micromhos per centimeter
uS/cm	Microsiemens per centimeter
WHO	World Health Organization
XOR	Exclusive OR

- ## Symbols

Ke	Constant of proportionality
n	Number of observations
$\mathbf{y_i}$	Real value
$\mathbf{\hat{y}_i}$	Predicted value

General Introduction

General Introduction

Water is a crucial resource for all life on Earth, and its quality is of utmost importance for human health, environmental preservation, and sustainable development.

In today's world, water is involved in most daily activities, including domestic, industrial and agricultural, making it a receiving element exposed to all types of pollution. Water quality control is essential for monitoring and maintaining appropriate water standards.

Water supply is currently a major need in different areas of life, due to the increase in the population and its standard of living.

In the world, surface water is the main source of drinking water, but increasingly, individuals and municipalities are also turning to groundwater that contains a huge volume of exploitable water.

In order to avoid any disturbance in the water that may threaten human health, the World Health Organization has set an International water standards.

However, water quality can be influenced by a wide range of natural phenomena and human activities. Natural processes can affect the characteristics of fresh water elements and chemical compounds. In addition, several anthropogenic impacts such as industrial activity, agricultural use or river engineering projects can also degrade water quality, Therefore water quality must be controlled. But traditional methods can be costly and time-consuming. This leads us to search for new technologies to control and improve water quality. [1]

In our current era, we find that the most important technologies are artificial intelligence, which itself includes many advanced technologies. Among these technologies we mention artificial neural networks (ANNs).

Artificial neural networks (ANNs) offer a promising alternative for predicting water quality parameters. These artificial intelligence systems efficiently analyze large amounts of water data and identify complex patterns and trends that affect water quality.

This is what we will talk about in our research topic :

In the first chapter, we talked generally about water and the different water quality parameters.

The second chapter, included machine learning and deep learning, and We delved deeper in artificial neural networks (ANNs), and Evaluation.

The third and last chapter will cover the application of our research topic, which involves using artificial intelligence to forecast a characteristic (pH in this case) in the realm of water potability.

Chapter 1 : WATER GENERALITIES

Chapter 1 :
Water Generalities

1 Introduction

Water is the primary component of living organisms. In the majority of living things, the water content is approximately 70% or higher.

In this chapter, first we will introduce the definition of water and discuss its various properties. Thenwe shift our focus to water standards, covering both international standards and those specific to Algeria. Following this, We are going toexplore the different methods of water analysis and measurement of specific parameters , and finally, the potential difficulties that can arise during the process of water analysis.

2 Definition of water

According to its origin, the term “water” comes from the Latin aqua. This chemical, called H2O, is a dilute water solution. As a liquid, its presence is crucial to preserving life and all known living organisms. There are three forms of water: liquid, solid or gas. However, it is most often defined in its liquid form because it is the most common form of water on our planet. [2]

3 Drinking water

Drinking water is water that can be consumed or used for domestic and industrial purposes . It can be distributed in bottles (mineral water or spring water, still water or sparkling water), in running water (tap water) or in tanks for industrial use. However, pure water is not naturally present. When it reaches our taps, it is supplied with elements that are both essential to our health but can also be confronted with substances potentially toxic to our body. It is therefore necessary that it respects several criteria in order to guarantee water consumption without any danger to everyone's health. [3]

4 Characteristics of the water quality

Water is characterized by various aspects, including its organoleptic, physical, chemical and biological properties.

4.1 Organoleptic properties

They are concerned with the water's color, flavor and aroma. The water ought to be odorless, transparent and pleasant to drink. These characteristics have no direct bearing on health however are related to consumption comfort. [3]

4.1.1 Color

The colouring of a water is qualified as "true" or "real" when it comes only from dissolved substances, that is to say those that pass through a filter of porosity of 0.45 μm. It is considered "apparent" when suspended substances contribute to its coloring. The actual and apparent colors are generally similar in clear waters and low turbidity. [4]

4.1.2 Smell and taste

Water for drinking ought to smell and taste well.

Most waters, treated or not, have a flavor and fragrance that is rather perceptible. These two characteristics alone Organoleptic responses are highly subjective and lack a measurable instrument.

Physiologists maintain that there are just four primary flavors: bitter, sour, sweet and salty. [4]

4.1.3 Turbidity

Turbidity quantifies the loss of transparency of water due to suspended particles that absorb or scatter light. Water with more suspended solids will appear cloudier, thus exhibiting higher turbidity. Cloudy waters tend to warm as suspended particles absorb solar heat, which can lead to decreased oxygen levels. [5]

4.2 Physico-chemical characteristics

4.2.1 Temperature

Accurately determining the temperature of the water is crucial. In fact, it affects electrical conductivity, pH measurement, the dissociation of dissolved ions and the solubility of gases and salts. [6]

4.2.2 PH (hydrogen potential)

The PH of water is a measure of its acidity, representing the concentration of hydrogen ions (H+). On the pH scale, which generally ranges from 0 (very acidic) to 14 (very alkaline), a value of 7 corresponds to a neutral solution at 25°C. The pH of a natural water source can vary between 4 and 10 depending on the acidic or basic characteristics of the soils it crosses. It is important to note that pH has no direct health significance, but it is crucial for assessing water corrosivity. [7]

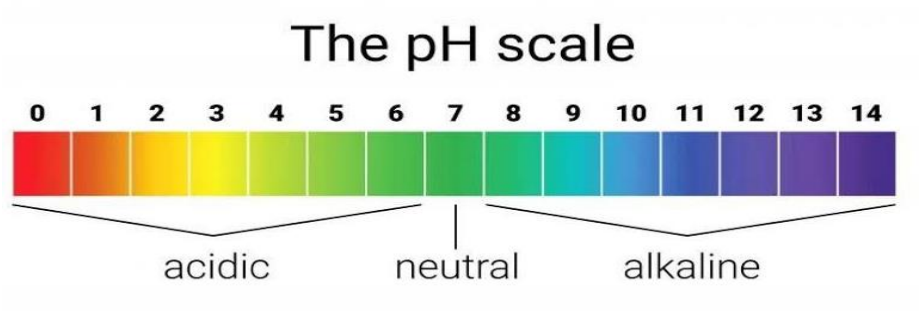


Figure I-1 : PH scale. [8]

4.2.3 Conductivity

Conductivity is a measure of the capacity of water to allow the passage of electric current. This capacity is directly related to the concentration of ions in water. These conductive ions come from dissolved salts and inorganic materials such as alkalis, chlorides, sulphides and carbonate compounds. Compounds that dissociate into ions are also known as electrolytes. The more ions present, the higher the conductivity of the water. Similarly, the fewer ions in the water, the less conductive it is. Distilled or deionized water can act as an insulator due to its very low or negligible conductivity value. In contrast, seawater has a very high conductivity. [9]

The units of measurement for conductivity are typically micro- or millisiemens per centimeter (uS/cm or mS/cm). In micromhos or millimhos per centimeter (umhos/cm or mmhos/cm), it can also be stated. [10]

4.2.4 Total Dissolved Solids (TDs)

Total dissolved solids (TDS) concentration represents the amount of organic and inorganic materials, such as metals, minerals, salts and ions, dissolved in a specific volume of water, encompassing the sum of all ionic-sized particles less than 2 microns. TDS is basically a measurement of everything dissolved in water except water molecules. TDS in water can come from a variety of

sources such as natural water resources, chemicals used to treat municipal water and runoff from roads and waterways. [11]

The total dissolved solids concentration is expressed in mg/L. TDS can be measured gravimetrically or calculated by multiplying a conductivity value by an empirical factor. The evaporation method for determining TDS is longer but useful when the composition of the water source is unknown. However, estimating TDS from conductivity is faster, adapted to field measurements and continuous monitoring. [9]

- **Relation between conductivity and total dissolved solids**

Since measuring total dissolved solids (TDS) can be time-intensive, it is common to estimate TDS from electrical conductivity (EC) measurements. This approach assumes the dissolved solids are primarily ionic species present at concentrations low enough to yield a linear relationship between TDS and EC.

$$TDS = ke \times EC \qquad (1.1)$$

Where ke: is a constant of proportionality. [12]

The higher the conductivity the higher the total solids dissolved and vice versa.

4.2.5 Dry residue

The dry residue gives information on the content of non-volatile dissolved substances (the rate of mineral elements). This quantity can range from less than 100 mg/l (water from crystalline massifs) to more than 1000 mg/l, depending on the water's place of origin. [13]

4.2.6 Total hardness TH

Water hardness is defined by its ability to react and produce foam in the presence of soap. Currently, hardness, also known as hydrotimmetrictitre (TH), is the total amount of alkaline-earthy cations present in water. In practice, only cations with concentrations above 1 mg/L are considered, such as calcium and magnesium ions. These ions are found in water as salts such as chloride, sulfate or hydrogen carbonate. [14]

According to REJSEK (2002), the classification of water hardness as a function of the THt value is as follows:

- 00 to 10°F: very fresh water.
- 10 to 20°F: moderately fresh water.
- 20 to 30°F: hard water.
- Over 30°F: very hard water.

4.3 Microbial characteristics

4.3.1 All germs

They live in aerobic environments. Their presence indicates the presence of polluting bacteria. Their counting makes it possible to obtain data on the hygienic quality of the water. [15]

4.3.2 Coliforms

Coliforms are the family Enterobacteriaceae includes coliforms, which are rod-like, non-sporogonous, Gram-negative, oxidase-negative, facultative aero-anaerobic bacteria. These bacteria are capable of fermenting lactose and producing acid and gas within 48 hours at temperatures of 35 and 37°C.

4.3.3 Fecal organisms or thermo-tolerant organisms

Fecal coliforms refer to organisms that have the same characteristics (organism characteristics) after being incubated at a temperature of 44°C. [16]

4.3.4 Fecal streptococci

Fecal streptococci are often considered as witnesses of Urine pollution. They are Gram positive, composed of chains, facultative anaerobes and immobile.

Natural water sources, chemicals used to treat the water supply of municipalities, runoff from roads and yards and even the plumbing system of your house.

Total dissolved solids are reported in mg/L. TDS can be measured by gravimetry or calculated by multiplying a conductivity value by an empirical factor.

Determination of TDS by evaporation takes more time, it is useful when the composition of a water source is not known. The derivation of TDS from conductivity is faster and suitable for both field measurements and continuous monitoring. [15]

4.4 Other characteristic

4.4.1 Clorides (Cl^-)

Chlorides, which are significant inorganic anions, are typically found in natural waters as sodium salts (NaCl) and potassium (KCl) at varying amounts. They are frequently employed as a pollution index. They have an impact on plant growth as well as aquatic flora and animals. [17]

4.4.2 Magnesium (Mg^{2+})

Magnesium is a key contributor to water hardness, discussed in another section. It can also alter the taste of water. According to studies, the detection threshold for magnesium taste is estimated to be about 100 mg/L for sensitive individuals and about 500 mg/L for the average population. [18]

4.4.3 Calcium (Ca^{2+})

In nature, calcium is a very sensitive alkaline earth metal that is found mostly in limestone rocks as carbonates. It is a key element of calcium, which makes up the majority of the water's hardness overall, it is typically found in drinking water. It mostly manifests as hydrogen carbonates. [18]

4.4.4 Sodium (Na^+)

Sodium salts (for example, sodium chloride) are found in virtually all foods (the main source of daily exposure) and in drinking water. Although sodium concentrations in drinking water are generally below 20 mg/litre, they can well exceed this threshold in some countries. Levels of sodium salts in the air are generally low compared to those in food or water. It should be noted that some water softeners can significantly contribute to increasing the sodium content of drinking water. [17]

4.4.5 Potassium (K^+)

Potassium is a naturally occurring element in water, but can also be introduced through human activity, such as from salt mines, the glass industry and the use of fertilizers. Like magnesium, potassium plays an essential role in the proper functioning of the nervous system, among other functions. [19]

4.4.6 Bicarbonates (HCO^-)

The presence of the bicarbonate ion in most running waters is caused by the action of bacteria which produce CO_2 from minerals containing carbonates. [15]

4.4.7 Nitrates (NO^{3-})

Nitrates are present in our daily diet, for example in some fruits and vegetables and are also found in fertilizers used in agriculture. Over-application of fertilizer to agricultural soils can lead to the risk of excessive nitrate concentrations in water. For this reason, the World Health Organization (WHO) has established a maximum limit of 50 mg/l for the concentration of nitrates in water. [19]

4.4.8 Sulfate (SO_4^{2-})

Sulfate (SO_4) is a constituent of nearly all natural water. The majority of sulfate compounds come from industrial wastes, shales, or the oxidation of sulfite ores.

One of the main dissolved substances in rain is sulfate. High sulfate concentrations in drinking water can be laxative when paired with the two most common hardness ingredients, magnesium and calcium. [20]

5 Drinking Water standards

Criteria for water potability are determined based on scientific data establishing maximum allowable doses (MADs). DMA is the amount of a substance that a person can ingest daily, without risk, throughout his life, taking into account all intakes (food, water). This approach determines the maximum amount a water can contain, with the addition of an adequate safety margin. It is important to note that these maximum amounts are always calculated by considering the most vulnerable individuals (babies, pregnant women, immunocompromised people...), thus ensuring enhanced protection for healthy adults. In addition, for various chemicals naturally occurring in water or from human activities (such as antimony, arsenic, cadmium, chromium, cyanide, certain hydrocarbons, mercury, nickel, nitrates, lead, selenium, certain pesticides), drinking standards are established taking into account a characteristic "margin of uncertainty" in toxicology. This means that they set limits below the thresholds considered acceptable. [21]

5.1 International standards: (OMS 2017 4th edition)

The international waste water standards established by the World Health Organization are listed in a table below :

Parameters	Units	Indicative values
Color	Mg/l platine	15
Turbidity	NTU	Lessthan 5
Calcium	Mg/l	Not mentioned
Chloride	Mg/l	No value but a taste can be noted from 250 mg/ l
PH	PH unit	6,5-8,5
Hardness (TH)	ppm of CaCO3	200
Manganese	Mg/l	0.4
Potassium	Mg/l	Not mentioned
Sodium	Mg/l	200
Sulfates	Mg/l	250
Nitrate as No3	Mg/l	50
Nitrite as No2	Mg/l	3
Temperature	°C	Not mentioned

Table I-1 : International release standards according to OMS.

5.2 Algerian standards: (JORADP 2014)

The maximum effluent discharge limit values in accordance with Algerian standards are listed in the table below :

Parameter group	Parameters	Units	Indicative values
Organoleptic parameters	Color	Mg/l platine	15
	Turbidity	NTU	5
	Smell	Dilution rate	4
	Taste	Dilution rate	4
Physico-chemical parameters	Calcium	Mg/l	200
	Chloride	Mg/l	500
	PH	PH unit	≥ 6,5 et ≤ 9
	Conductivity at 20°C	µS/cm	2800
	Hardness (TH)	Mg/l of CaCO3	500
	Manganese	µg/l	50
	Potassium	Mg/l	12
	Sodium	Mg/l	200
	Sulfates	Mg/l	400
	Temperature	°C	25

Table II-2 : Algerian release standards according to JORADP.

6 Methods of analysis of water

Three categories of field test methods for monitoring chemical water quality are commonly used :

- **Test strips:** These small single-use strips change colour to indicate the concentration of a specific chemical. The user activates the strip by dipping it in the water sample and stirring it, or holding it in a stream of water. After a short waiting time, the user compares the color of the strip with a color chart to read the concentration of the chemical. Although these kits are extremely simple, their accuracy is less than other methods, especially if users do not follow the instructions.
- **Colored disc kits:** These kits are designed for a diverse array of chemical tests. In a typical configuration, the user adds a powder bag or a few drops of a liquid reagent to a water sample in a reusable plastic tube. The user then places the sample tube in a small plastic display box. This box contains a plastic disc with a printed color gradient. The user rotates the colored disc to find the part that best matches the color of the sample, then reads the concentration of the chemical on the disc. These colorful disc kits usually have several steps and often include prescribed wait times, which makes them a little more complex and expensive, but usually more accurate.
- **Portable digital instruments:** Lightweight and portable digital devices such as meters, colorimeters and photometers are available for water testing. They offer the most accurate results among these three test methods, but are also more expensive and tricky than the previous options. These instruments require batteries and calibration. [22]

7 Measurement of specific parameters

7.1 Ph meter

The pH meter is a device used to measure the pH of a solution, consisting of two elements : an electronic box displaying the pH and an electrode measuring this value. Its operation is based on the relationship between the $H3O^+$ ion concentration and the difference in electrochemical potential in the glass electrode. In general, this electrode is combined, with one electrode with constant potential and another whose potential varies according to the pH. The potential between these two electrodes is neutral at pH=7, making it possible to determine the pH by correlating the potential variation with the pH. [23]

Figure I7-2 : PH meter. [24]

7.2 Conductivity meter

A conductivity meter, also known as a conductivity meter, is a measuring device that assesses the conductivity of a substance or solution. Its operation is based on the immersion of a probe in the solution to pass an electric current. The propagation speed of this current is recorded by the probe, then transmitted to a housing equipped with a screen that displays a value in "S.m" or "S.cm" (Siemens per meter or per centimeter). The conductivity of a solution depends directly on its concentration in various elements, including minerals, which allows the conductivity meter to provide information on the composition of the liquid. Proper calibration of a conductivity meter is essential to ensure accuracy. [25]

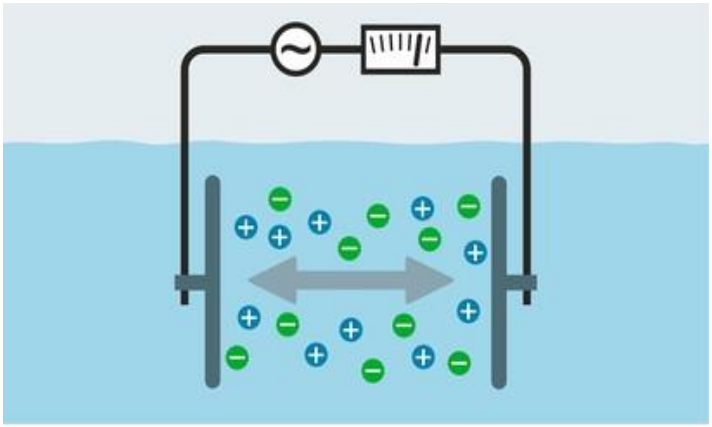


Figure I-3 : conductivity meter. [26]

7.3 Turbidimeter

A turbidimeter is a portable or installation instrument to measure suspended particles in a liquid or a colloidal gas. A turbidimeter measures the suspended particles with a light beam (beam source) and a light detector set at 90 ° from the original beam. The density of the particles is a function of the light reflected by the particles suspended in the detector. The amount of light reflected for a given density of particles depends on the properties of particles such as their shape, color and reflectivity.

The turbidimeter is calibrated with a known particulate material, commonly known as Arizona street dust. Subsequently, environmental factors (K factors) are used to compensate for light or darker dust. K-factors are determined by the user by activating the turbidimeter near an air sampling pump and comparing the results. [27]

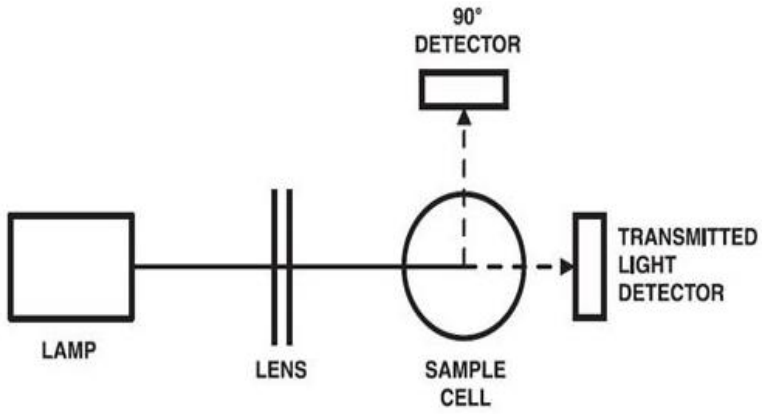


Figure I-4 : Turbidimeter. [28]

7.4 Oximeter

The oximeter is a device used to measure the amount of oxygen dissolved in water, essential for aquatic life. There are two types of probes for this measurement : electrochemical probes that evaluate the electrical voltage from an electrolyte reaction, and optical probes that rely on fluorescence to measure dissolved oxygen. The optical probe is often preferred for its ease of use and maintenance.

Dissolved oxygen measurement is crucial to assess water quality, as low concentrations can negatively impact aquatic organisms and ecosystems. This measure is widely used in aquaculture, drinking water treatment and water quality monitoring. [29]

8 Difficulty with water analysis

To find an affordable device to measure water potability parameters, it is essential to consider the specificities of each region and to favor a reliable method to assess water quality. This method must be simple, reproducible and adapted to random sampling, with intensification in times of crisis. When developing chemical verification, it is crucial to take into account the availability of suitable analytical tools, the cost of analysis, the possible degradation of samples, the stability of pollutants, their potential presence in various supplies, as well as the optimal location for monitoring and frequency of sampling. [30]

9 Conclusion

In this chapter, after an extensive examination of water, we have determined that humans can not consume water directly. Therefore, water must meet specific drinking water standards.

This is accomplished by testing various physical and chemical properties of the water as PH, Temperature, Turbidity, Conductivity, …etc, to ensure they align with the established criteria. Water must be treated with a high degree of precision.

As a result, we have found that humanity has developed sophisticated technology to predict and monitor water quality, which we will explore further in the next chapter.

Chapter 2 : Artificial Neural Networks (ANNs)

Chapter 2 :

Artificial Neural Networks (ANNs)

1 Introduction

In the present era, artificial intelligence (AI), machine learning (ML) and deep learning (DL) have become widely recognized terms that are occasionally used synonymously to depict systems or software capable of exhibiting intelligent behavior.

In this chapter we made contact to the definition of machine learning and the different types of learning approaches. We placed particular emphasis on examining several algorithms commonly used in machine learning. Following that, we redirected our attention towards the field of deep learning. We began by discussing deep neural networks RNN (Recurrent Neural Networks), CNN (Convolutional Neural Network) and ANN (Artificial Neural Networks) and subsequently shifted our focus to artificial neural networks, aiming to delve deeper into their composition and internal performance.

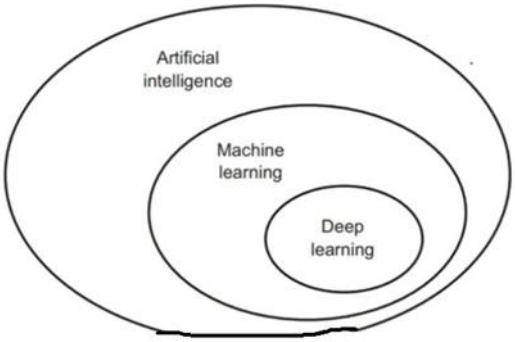


Figure II-1 : Relation between AI and ML and DL.

2 Artificial intelligence (AI)

Artificial intelligence (AI) involves the application of scientific and engineering principles to create smart machines, particularly intelligent computer programs. It is closely connected to the field that explores the use of computers to comprehend human intelligence. [31]

It branches into deep learning and machine learning as described in the following chart :

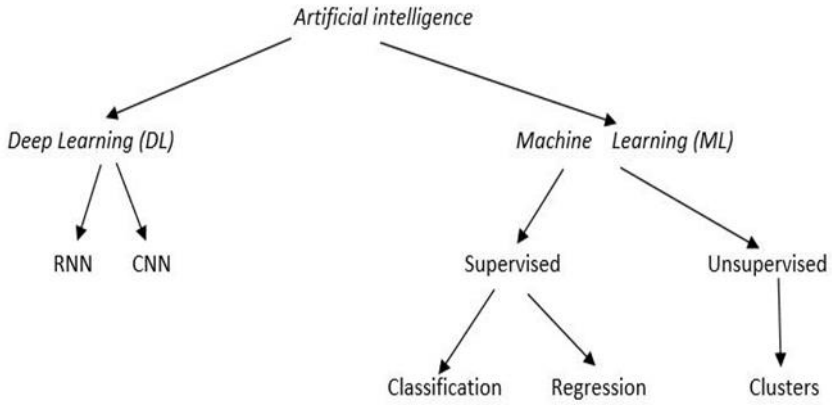


Figure II-2 : Chart of artificial intelligence.

3 Maching learning (ML)

Machine learning is a field of research that focuses on computer algorithms designed to enable computers to learn autonomously from data and past experiences. Its objective is to identify patterns within the data and make predictions without requiring human intervention. Machine learning, along with its applications in various domains, is widely recognized as a fundamental aspect of artificial intelligence. [32]

3.1 Types of learning

Machine Learning can be roughly classified into two primary categories, Supervised Learning and Unsupervised Learning, depending on whether labeled information is available for the examples in the dataset.

3.1.1 Supervised learning

Supervised learning involves deriving behavioral rules from a database that consists of pre-labeled examples. Specifically, this database comprises a collection of random input-output pairs $\{(X_i, Y_i)\}1 \leq i \leq n$. The main goal is to develop the ability to predict the output Y for any new input X. Classification and regression algorithms are two common examples of supervised algorithms. [33]

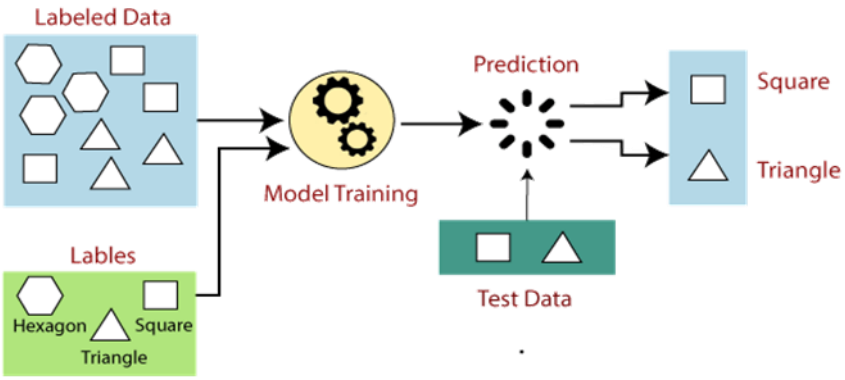


Figure II-3 : Supervised learning. [34]

3.1.1.1 Classification

Classification is a technique in data analysis that aims to identify and define significant categories within the data. These categories are represented by models known as classifiers, which make predictions about discrete and unordered class labels. For instance, a classification model can be created to categorize bank loan applications as either safe or risky. This analytical approach allows for a deeper comprehension of the overall data.

Classification finds wide-ranging applications in various fields such as fraud detection, targeted marketing, performance prediction, manufacturing, and medical diagnosis. In the following section, we provide a summary of the typical problems that arise in classification tasks. [35]

- **Binary Classification** : Binary classification involves the assignment of two class labels, such as "true and false" or "yes and no", to data instances. One class represents the normal or expected state, while the other class represents the abnormal or undesired state. For instance, in medical testing, the normal state could be "cancer not detected", while the abnormal state could be "cancer detected". Likewise, in email filtering, the binary classification labels could be "spam" and "not spam".
- **Multiclass Classification :** Multiclass classification refers to tasks where there are more than two class labels involved. In contrast to binary classification, there is no notion of normal and abnormal outcomes. Instead, examples are categorized into one of several predefined classes. [36]

3.1.1.2 Regression

Regression analysis encompasses various machine learning methods that enable the prediction of a continuous outcome variable (y) based on the values of one or more predictor variables (x). The key difference between classification and regression lies in their respective prediction targets. While classification predicts discrete class labels, regression focuses on predicting a continuous quantity. The figure below illustrates this distinction between classification and regression models. Regression models find extensive application in diverse fields such as financial forecasting, cost estimation, trend analysis, marketing, time series estimation, drug response modeling and many others. [36]

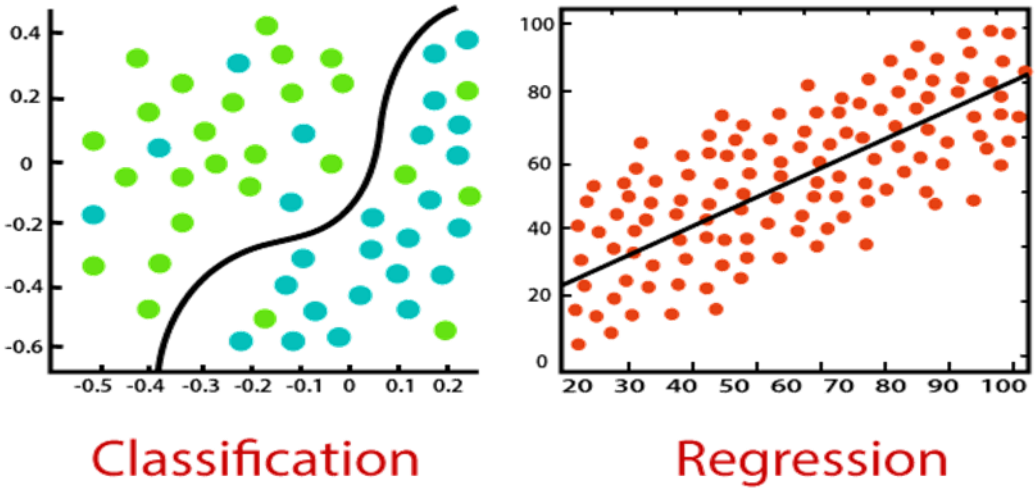


Figure II-4 : Classification vs regression. [37]

- **Types of regression**

✓ **Linear regression**

Linear regression is a statistical approach that models the linear relationship between a dependent variable (Y) and one or more independent variables (X).

✓ **Logistic regression**

Logistic regression is a statistical approach employed to estimate the probability of a binary event occurring (for example : yes/no, success/failure).

✓ **Polynomial regression**

Polynomial regression is a statistical approach that allows for modeling and understanding the relationship between a dependent variable (Y) and one or more independent variables (X), in cases where the relationship is not a linear one. [38]

3.1.2 Unsupervised learning

In unsupervised learning, the focus is on situations where only the inputs $\{X_i\}1 \leq i \leq n$ are available, without any corresponding outputs. The primary objective in unsupervised learning is to perform data partitioning, often referred to as clustering. This involves grouping observations into distinct and homogeneous clusters, ensuring that the data within each cluster share common characteristics. [39]

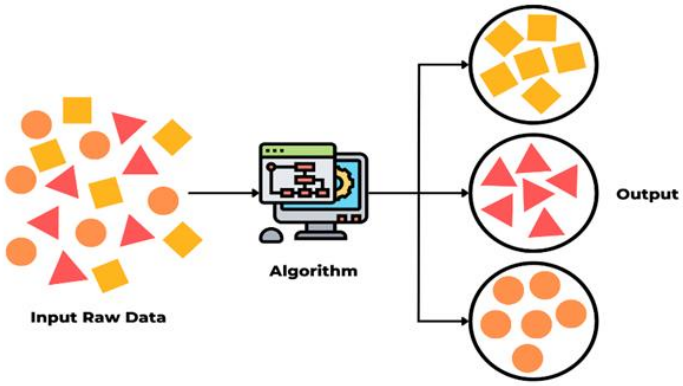


Figure II-5 : Unsupervised learning. [39]

3.1.2.1 Clustering

Cluster analysis, also known as clustering, is the process of dividing a set of data objects or observations into subsets, where each subset represents a cluster. Clusters are defined such that the objects within a cluster exhibit similarity to one another but exhibit dissimilarity to objects in other clusters. The result of a cluster analysis is a set of clusters, collectively known as a clustering. It is important to note that different clustering methods can yield different clusterings when applied to the same dataset. The partitioning of data into clusters is performed by clustering algorithms rather than by humans, making clustering a valuable tool for uncovering previously unknown groups within the data.

Cluster analysis has found extensive application in various fields, including business intelligence, image pattern recognition, web search, biology and security. In the context of business intelligence, clustering is particularly useful for organizing a vast customer base into distinct groups based on shared characteristics. This enables businesses to develop targeted strategies for improved customer relationship management. [40]

- **Clustering methods**

Below are some of the most commonly recognized clustering methods :

✓ **K-Mean Clustering or partitioning**

The process of k-means clustering is relatively straightforward. Initially, we determine the number of clusters, denoted as K and assume the initial centers for these clusters. The initial centers can be randomly chosen objects or the first K objects in the sequence. The k-means algorithm then follows these three steps iteratively until convergence or stability is achieved :

1. Determine the coordinates of the cluster centers.

2. Calculate the distance between each object and the cluster centers.

3. Group the objects based on the minimum distance, assigning them to the nearest cluster.

This process continues until no objects change their assigned groups, indicating that the algorithm has reached stability.

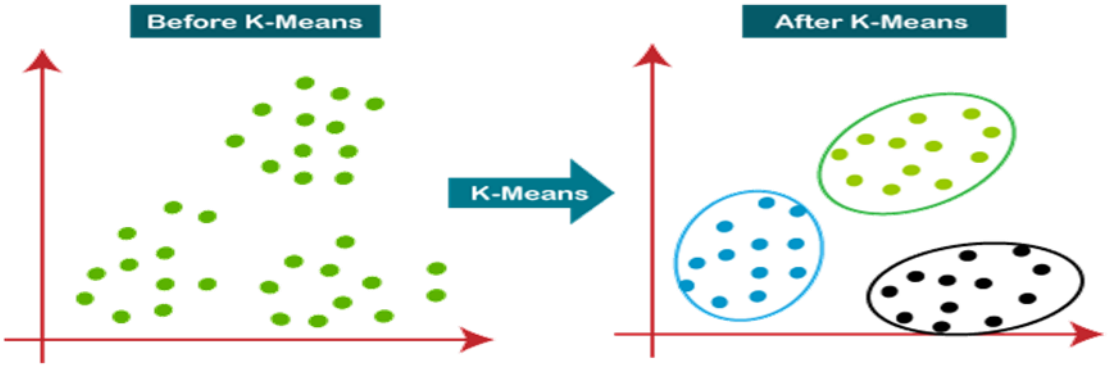


Figure II-6 : Before K-means and after. [41]

✓ **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

DBSCAN begins with a randomly selected point and determines its neighbors based on a specified distance threshold called epsilon. It can identify noise points as well, which do not belong to any cluster.

- ✓ **EM (Expectation Maximization) Clustering using Gaussian Mixture Models (GMM)**

In this method, data points are assumed to be distributed in a Gaussian field. The clusters are determined by estimating the mean and standard deviation of the Gaussian distributions. EM clustering aims to find the maximum likelihood estimation of the parameters.

- ✓ **Agglomerative Hierarchical Clustering**

Unlike other clustering methods, Agglomerative Hierarchical Clustering does not require the specification of the number of clusters in advance. It starts by considering each data point as an individual cluster and then iteratively merges the closest clusters based on a chosen distance metric. This algorithm is also insensitive to the choice of distance metric used during clustering. [42]

3.2 Learning algorithms

3.2.1 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are supervised machine learning algorithms designed to tackle classification and regression problems. They were conceptualized in the 1990s, stemming from the statistical learning theory developed by Russian computer scientists Vladimir Vapnik and Alexey Chervonenkis, known as the Vapnik-Chervonenkis theory. SVMs were quickly adopted due to their ability to handle large-scale data, strong theoretical foundations and impressive practical performance.

A key principle behind SVMs is to transform a classification or discrimination problem into a hyperplane (feature space) where the data is separated into distinct classes with the maximum possible margin between them. This is why SVMs are also referred to as "wide-margin separators". Additionally, SVMs require relatively few parameters, making them relatively straightforward to use. [43]

3.2.2 K nearest neighborsalgorithm (K-NN)

KNN is a supervised machine learning method that can be used for both regression and classification tasks. Unlike other predictive models like logistic regression or linear regression, the KNN algorithm does not generate a predictive model from a training set.

Instead, to make a prediction for a new observation, the KNN algorithm examines the entire dataset. It identifies the K instances that are closest to the new observation and then uses the output variables (y) of those K nearest neighbors to determine the predicted output.

Specifically :

For regression tasks, the predicted output is calculated as the mean or median of the y values of the K nearest neighbors.

For classification tasks, the predicted output is the mode (most common class) among the y values of the K nearest neighbors.

The KNN method relies on the entire dataset to make predictions, rather than fitting a model to a training set. This approach, which does not involve building a predictive model, is a key distinguishing feature of the KNN algorithm. [44]

3.2.3 Random Forest algorithm

Random Forest is a widely-used machine learning algorithm that was trademarked by Leo Breiman and Adele Cutler. It combines the outputs of multiple decision trees to arrive at a single result, which contributes to its ease of use and flexibility, making it a popular choice.

The Random Forest algorithm has three main hyperparameters that must be set before training: node size, the number of trees and the number of features sampled. Once these are configured, the Random Forest classifier can be used for both regression and classification problems.

The Random Forest algorithm is composed of a collection of decision trees. Each individual tree is trained on a bootstrap sample - a data sample drawn from the training set with replacement. From this training sample, one-third is set aside as an out-of-bag (OOB) sample, which is used later for cross-validation.

An additional layer of randomness is introduced through feature bagging, which further diversifies the dataset and reduces correlation between the individual decision trees.

The final prediction depends on the type of problem :

- For regression tasks, the individual decision tree outputs are averaged.
- For classification tasks, the most common predicted class among the decision trees is selected.

Finally, the OOB sample is used to perform cross-validation and finalize the prediction. [45]

3.2.4 Decision Trees

Decision Trees are a non-parametric supervised learning approach used for both classification and regression problems. The aim is to build a model that can predict the value of a target variable by learning simple decision rules derived from the data features. A Decision Tree can be interpreted as a piecewise constant approximation. [46]

4 Deep learning

Deep learning is a field within the realms of machine learning and artificial intelligence that aims to represent intricate data abstractions by utilizing multiple layers of neurons, which are composed of intricate structures or nonlinear transformations. As the volume of data and computational capabilities have grown, neural networks with more intricate architectures have gained significant popularity and have been employed across diverse domains. [47]

4.1 Biological neuron

A cell consists of a cell body and a nucleus. The cell body branches out to form extensions called dendrites. Dendrites are used to convey information from the outside to the neuron body. Once the information is processed by the neuron, it propagates along the axon to be transmitted to other neurons. [48]

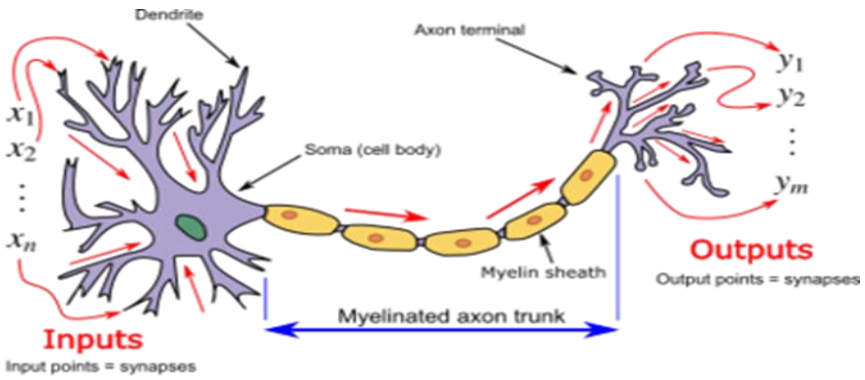


Figure II-7 : Biological neuron. [49]

4.2 Artificial neuron

Any artificial neuron is considered an elementary processor that processes information. It receives a variable flow of inputs from other neurons or sensors present in the machine to which it belongs. Each input is weighted according to the strength of the associated connection. The

elementary processor then has a single output that then splits to provide information to a number of neurons located downstream.[50]

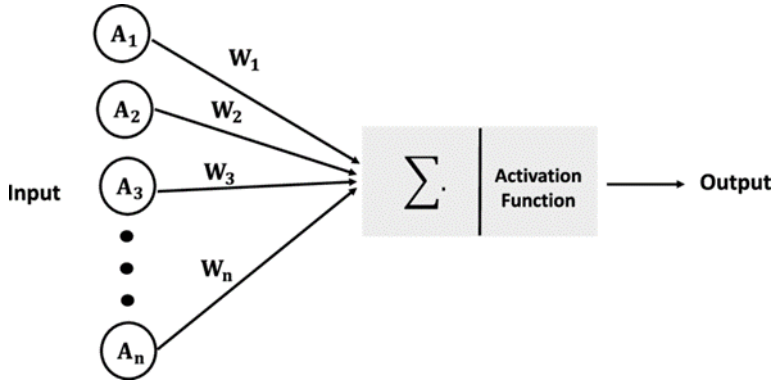


Figure II-8 : Artificial neuron. [51]

4.3 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are computational systems that are heavily influenced by the way biological nervous systems, like the human brain, function. ANNs consist of numerous interconnected computational nodes, known as neurons, which collaborate in a distributed manner to acquire knowledge from input data and optimize the final output. The input, typically presented as a multidimensional vector, is fed into the input layer and propagated through the hidden layers. In the hidden layers, decisions are made based on the information from the preceding layer and the network evaluates how internal adjustments impact the overall output. This iterative process of adjusting and improving the network's performance is known as learning. When multiple hidden layers are stacked together, it is commonly referred to as deep learning. [51]

4.4 Activation function

Activation functions are utilized in neural networks to determine the network's output. They are predefined mathematical expressions that apply either linear or nonlinear operations within the network. For optimization purposes, activation functions need to satisfy certain criteria such as being bounded, continuous, monotonic and continuously differentiable with respect to the weights.

The sigmoid function is the most widely used activation function, but there are other options available, including the arc-tangent function, hyperbolic-tangent function, ReLU and Leaky ReLU. The activation function f(v) serves as the decision-making component that establishes the decision boundary in the input space by setting a threshold in the induced local field. In the absence of an activation function, the output signal becomes a simple linear function. It is important to note that

linear functions are only capable of producing single-grade polynomials, effectively transforming the neuron into a linear regression model. Regardless of the number of linear functions stacked, the resulting output will always be linear. However, by activating the linear combiner using an appropriate activation function, it becomes possible to create complex decision boundaries by employing a combination of multiple neurons.

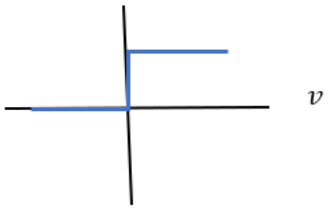
The simplest form of an activation function can be defined as a step function, where the induced local field, denoted as 'v', determines the output values. By implementing the signum function, the step function can be adjusted to produce values within the range of 1, 0, and -1.[52]

The neuron's output, denoted as y, is determined by evaluating the potential v using the activation function f(v). [53]

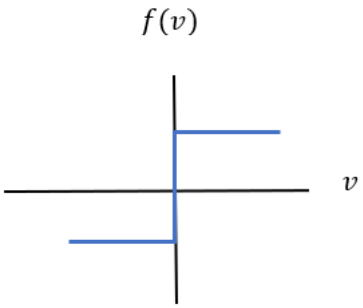
It is therefore :

$$f(v) = y \rightarrow y = f\left(\sum_{j=0}^{n} w_j x_j\right) \quad (2.1)$$

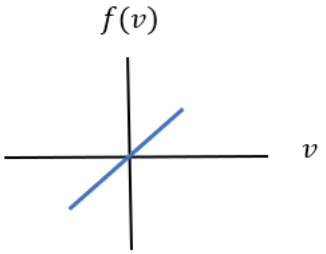
- Step function : $f(v) = \begin{cases} 0 & if\ v < 0 \\ 1 & if\ v \geq 0 \end{cases}$ (2.2)



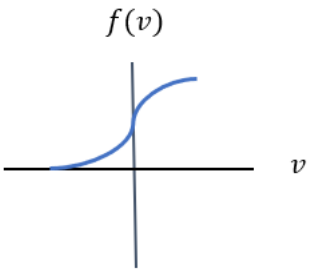
- Sign function : $f(v) = \begin{cases} -1 & if \quad v < 0 \\ +1 & if \quad v \geq 0 \end{cases}$ (2.3)



- Linear function : f(v) = v (2.4)



- Sigmoid function : $f(v) = \frac{1}{1-e^{-v}}$ (2.5)



- ReLU

ReLU (Rectified Linear Unit) is a widely used activation function in artificial neural networks, especially in deep learning models. It is a non-linear function that introduces non-linearity into the network. This non-linearity helps address the vanishing gradient problem, which can occur in neural networks. By incorporating ReLU, neural networks are better able to learn and model more complex relationships within the data. [54]

4.5 The layers of a neural network

Artificial neural networks consist of a limited number of neurons organized in layers. The neurons of the adjacent layers are connected by weights. Information spreads from layer to layer in the network, which characterizes their "feed-forward" nature. There are three types of layers in these networks :

- **Input layer :** Features within this layer obtain input values from the network and convey them to the hidden neurons.
- **Hidden layers :** consist of neurons that gather input from multiple preceding layers. These neurons then calculate a weighted sum of the inputs and apply an activation function, commonly a sigmoid function, to modify the result. The modified response is then transmitted to the neurons in the subsequent layer.
- **Output layer :** The output layer performs a similar function to the hidden layers, but with one key distinction : the neurons in the output layer do not have any connection to other neurons. [49]

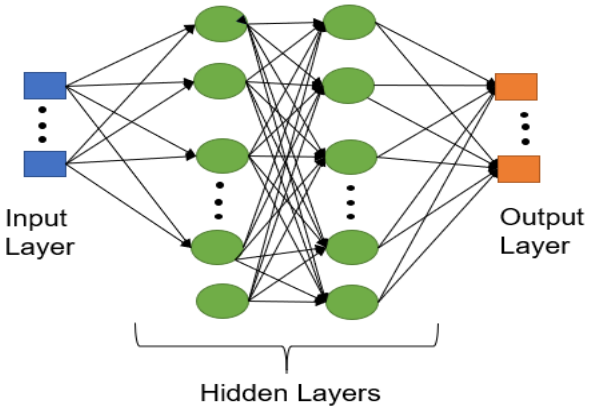


Figure II-9 : The layers of a neural network.

4.6 Feedforward and Feed backward

4.6.1 Feedforward

Feedforward neural networks were the initial and perhaps the most straightforward form of artificial neural network created. In these networks, information moves unidirectionally, progressing from the input nodes through any intermediate nodes (if they exist) and ultimately reaching the output nodes. The network does not contain any loops or cycles. Feedforward networks, also referred to as associative networks, can be built using various types of units. Typically, continuous neurons with sigmoid activation functions are employed, particularly in the context of error backpropagation. [56]

4.6.2 Feed backward

To put it in different words : The backpropagation algorithm is a technique that builds upon the Widrow-Hoff learning rule. It operates through supervised learning, where the algorithm is provided with input-output examples that the network should be able to produce. By calculating the error, the algorithm adjusts the randomly assigned weights in order to minimize this error. The standard backpropagation algorithm follows a gradient descent approach, where the network weights are adjusted in the opposite direction of the performance function's gradient. The optimal combination of weights that minimizes the error is considered a solution to the learning problem. The backpropagation algorithm requires an activation function that is differentiable, with the most commonly used ones being tan-sigmoid, log-sigmoid and occasionally linear functions. Feed-forward networks often consist of hidden layers of sigmoid neurons followed by an output layer of linear neurons. This structure allows the network to learn both linear and nonlinear relationships between the input and output data. The presence of a linear output layer enables the network to generate values outside the range of -1 to +1.

During the learning process, the data is divided into two sets : the training data set, which is used to calculate the error gradients and update the weights and the validation data set, which helps determine the optimal number of iterations to avoid overfitting. As the number of iterations increases, the training error decreases while the validation error initially decreases, reaches a minimum and then starts to increase. Continuing the learning process after reaching the minimum validation error leads to overfitting. Once the learning process is complete, another data set, known as the test set, is used to assess and confirm the accuracy of predictions. Well-trained backpropagation networks generally provide reasonable outputs when presented with new inputs.

Typically, in artificial neural network (ANN) approaches, data normalization is necessary before training to ensure that the influence of input variables is not biased by their magnitude or range of values. The normalization technique involves linearly transforming the input/output variables to a specific range. [55]

4.7 Perceptron

The perceptron is a member of the group of neural networks that propagate information in a forward direction, going from input to output. It can be classified into different types, such as single-layer and multilayer perceptrons, but the emphasis is placed on the multilayer perceptron (MLP).

4.7.1 Multilayer Perceptron (MLP)

This is an extension of the previous structure, which includes one or more hidden layers between the input and the output. Each layer contains neurons that are connected to all neurons of the previous and next layer (except input and output layers) and there are no connections between neurons of the same layer. The activation functions commonly used in this type of network are threshold or sigmoid functions. This type of network is capable of solving non-linear separable problems as well as more complex logical problems, including the famous XOR problem. It also uses supervised learning based on the error correction rule. [56]

4.8 Deep neural networks (DNN)

Deep neural networks are characterized by their extensive use of hidden layers. There are two primary categories of deep neural networks :

- Convolutional neural networks (CNNs) are one-way, where specific layers carry out pre-processing tasks.

- Recurrent neural networks (RNNs) allow information to flow in both directions. [57]

4.8.1 Convolutional Neural Network (CNN)

A widely used deep learning architecture, known as CNN, is gaining popularity due to its ability to learn directly from input data without the need for human-defined features. This feature distinguishes CNN from traditional artificial neural networks (ANNs) like regularized MLP networks and leads to improved design. Each layer of a CNN considers optimal parameters to produce meaningful outputs and simplifies the complexity of the model.

CNNs are specifically designed to handle diverse two-dimensional shapes, making them widely applicable in tasks such as visual recognition, medical image analysis, image segmentation, natural language processing and more. The automatic discovery of important input features without human intervention gives CNNs a greater advantage over traditional networks in terms of power and effectiveness. [58]

4.8.2 Recurrent Neural Networks (RNN)

The RNN consists of interconnected layers that create cyclical structures, meaning that each layer's output is fed into the next layer as input. Recurrent neural networks are commonly employed for tasks such as image captioning, automatic translation and natural language processing, as they effectively handle sequential or temporal information. They have been utilized in various applications, including the detection of sleep apnea from overnight ECG signals and automatic speech processing. [59]

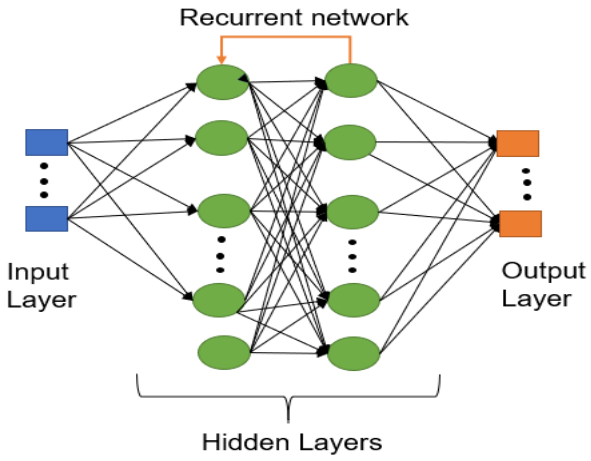


Figure II-10 : General form of RNNs.

4.9 Evaluation

In the field of forecasting, accuracy is a critical factor that determines the success of any forecasting model. To assess the accuracy of a forecast, precision measurements serve as anessential tool.

4.9.1 Mean Squared Error (MSE)

Mean Squared Error (MSE) is a widely used metric for assessing the performance of regression models. It calculates the average of the squared differences between the predicted values and the actual values in a dataset. A lower MSE value suggests a better fit for the model, as it indicates the predicted values are closer to the true values. [60]

It is calculated as :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2.6)$$

Where :

- y_i is the real value.
- $\hat{y}_i$ is the corresponding predicted value.
- n is the number of observations.

4.9.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a straightforward yet robust metric used to evaluate the accuracy of regression models. It calculates the average of the absolute differences between the predicted values and the actual target values. Unlike other metrics, MAE does not square the errors, which means it assigns equal importance to all errors, regardless of their magnitude or direction. [61]

The formula for calculating MAE is as follows :

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \qquad (2.7)$$

4.9.3 Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is one of the most widely used metrics for evaluating the quality of predictions. It provides a measure of the average magnitude of the errors, using Euclidean distance to quantify how far the predictions fall from the true observed values. [62]

Root mean square error can be expressed as :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \|y(i) - \hat{y}(i)\|^2}{n}} \qquad (2.8)$$

To assess the results, we will utilize several parameters, including Accuracy, Precision and Recall, which will be explained in further detail.

4.9.4 Accuracy (ACC)

This metric represents the percentage of predictions that correctly match the actual or real values.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.9)$$

4.9.5 Precision (P)

This metric indicates the percentage of accuracy for the positive situations that have been identified or determined previously.

$$P = \frac{TP}{TP+FP} \qquad (2.10)$$

4.9.6 Recall (R)

Reacallalso known as Sensitivity, is the percentage of the real or actual situations that were accurately identified or determined.

$$R = \frac{TP}{TP+TN} \qquad (2.11)$$

4.9.7 F1 Score

The F1 Score is a metric that aims to balance precision and recall. It is calculated as the harmonic mean of precision and recall. The F1 Score is useful when trying to achieve a balance between high precision and high recall, as it penalizes extremely low values for either of those components. [63]

$$F1 = \frac{2(P \times R)}{(P+R)} \qquad (2.12)$$

5 Conclusion

In conclusion, machine learning and deep learning are two powerful branches of artificial intelligence that have revolutionized various industries and fields.

Machine learning focuses on developing algorithms that allow systems to learn from data and make predictions or decisions without explicit programming.

Deep learning, on the other hand, is a subset of machine learning that utilizes artificial neural networks with multiple layers to extract complex patterns and representations from data to reduce and improve error thanks to back propagation algorithm.

Chapter 3 : Application of artificial neural network in water quality

Chapter 3 :
Application of artificial neural network in water quality

1 Introduction

Following the usage of the key ideas and concepts from the preceding chapters (1&2), this chapter will address the application of our research topic, which is the prediction of a characteristic (in this case, the pH) in the field of water potability via the use of artificial intelligence. Before beginning the actual model creation, it is necessary to have a thorough understanding of our working environment and to do data analysis.

2 Workplace conditions

The used software is Python which gives whith its extended librairies a free and environment of execution. The attributes of the equipment (for computing) that we deployed in our work are shown in the following table:

	Desktop
PC	DELL
Windows type	Windows 10 Professionnel
Processor	Intel(R) Core (TM) i5-8350U
RAM	8.00 GB
System	64-bit

Table III-2 : The attributes of the equipment for computing

3 Prediction of pH of water

3.1 Data Description

The dataset used in this study comes from a variety of water quality monitoring sources. It includes a total of 3276 observations distributed on 9 variables. These variables represent various physicochemical parameters of water quality: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, carbon, Trihalomethanes and Turbidity.

This data is taken from a universal source (https://www.kaggle.com/datasets/nayanack/water-probability/data).

In addition “figure III-1” show the variable distribution.

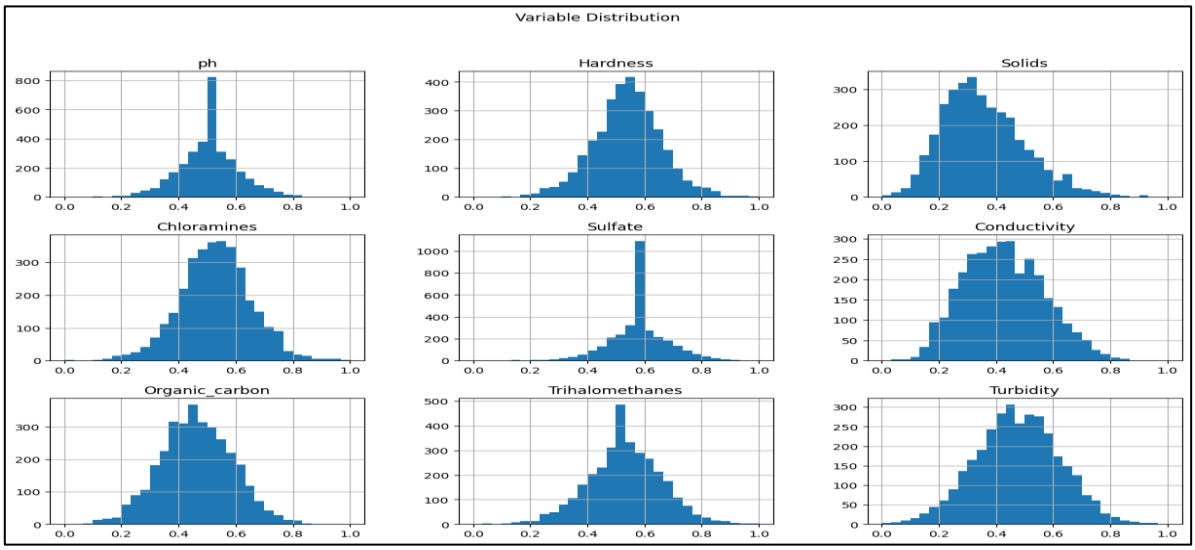


Figure III-1 : Variable distribution

3.1.1 Correlation matrix

Correlation matrix heatmap is a graphical representation of the correlation coefficients between different variables in a dataset.

Correlation values range from -1 to 1, where 1 indicates a perfect positive correlation (variables tend to increase/decrease together), -1 a perfect negative correlation (variables move in opposite directions) and 0 no correlation. The colors indicate the strength and direction of the correlation.

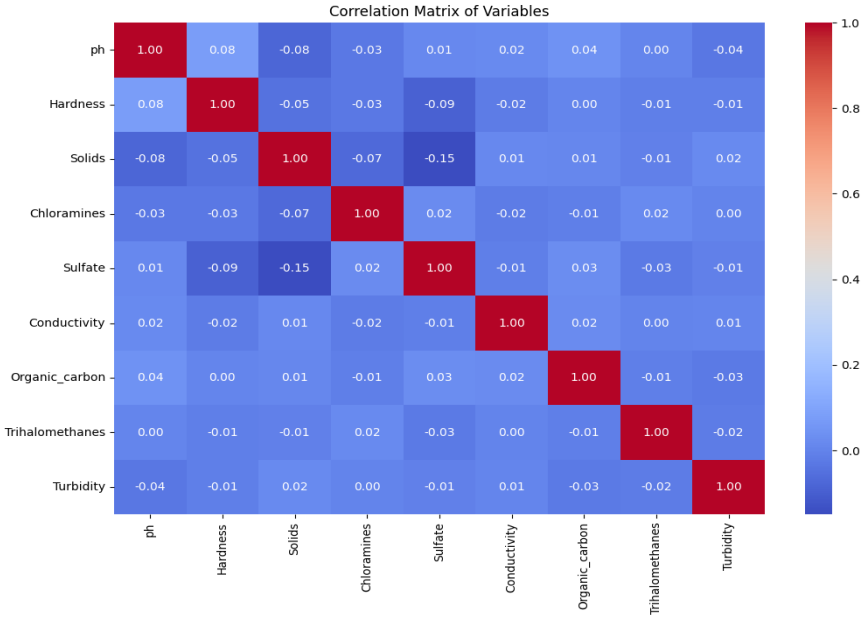


Figure III-2 : Correlation matrix of variables

The matrix correlation in “figure III-2” shows that pH has relatively weak relationships, both positive and negative, with the various water quality parameters. The strongest relationship is the weak positive correlation with Hardness and Organic_carbon, while the weak negative correlations with Solids and Turbidity are also noteworthy.

3.2 Creation of models and results

In the following, we will create 5 models. Each of them will be trained and test, then evaluated and at the end compared.

3.2.1 Intelligent model 1

This model is a sequential neural network model for regression using Keras. The model is designed to predict a continuous target variable from a set of 8 input features and involves multiple dense layers with ReLU activation. It employs L2 regularization to prevent overfitting and utilizes dropout layers to improve generalization, using Adam optimization algorithm, efficiently updates the model's weights durant training. In the following, we can see a part of the programme that we used to create the model.

```
# Définition de l'architecture du modèle avec régularisation
model = Sequential()
model.add(Dense(64, input_dim=8, activation='relu', kernel_regularizer=l2(0.01)))
model.add(Dropout(0.2))
model.add(Dense(32, activation='relu', kernel_regularizer=l2(0.01)))
model.add(Dropout(0.2))
model.add(Dense(16, activation='relu', kernel_regularizer=l2(0.01)))
model.add(Dense(1))

# Model compilation
model.compile(optimizer='adam', loss='mse', metrics=['mae'])
```

Figure III-3 : Model 1 configuration

Once the model is trained, we can generate a graph that shows how the loss function value decreased during training for both the training and validation datasets.

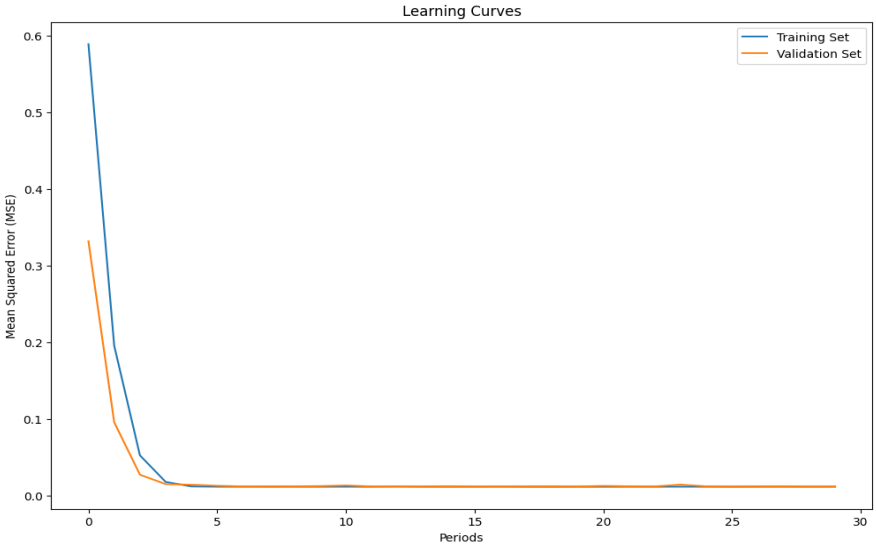


Figure III-4 : Learning curves

After training the model, we can evaluate the model by calculating performance measures. Three metrics (tools) were implemented for the evaluation: Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), which is a commonly used tool. The following are the obtained results :

```
ANN- MSE : 0.010499338619410992
ANN- MAE : 0.07279794663190842
ANN- RMSE : 0.10246628040194976
```

Figure III-5 : Calculate model 1 performance

The graph figure III-6 visualizes predicted values alongside the actual real values when the red points represent the actual values, while the blue points show the model's predictions for those values.

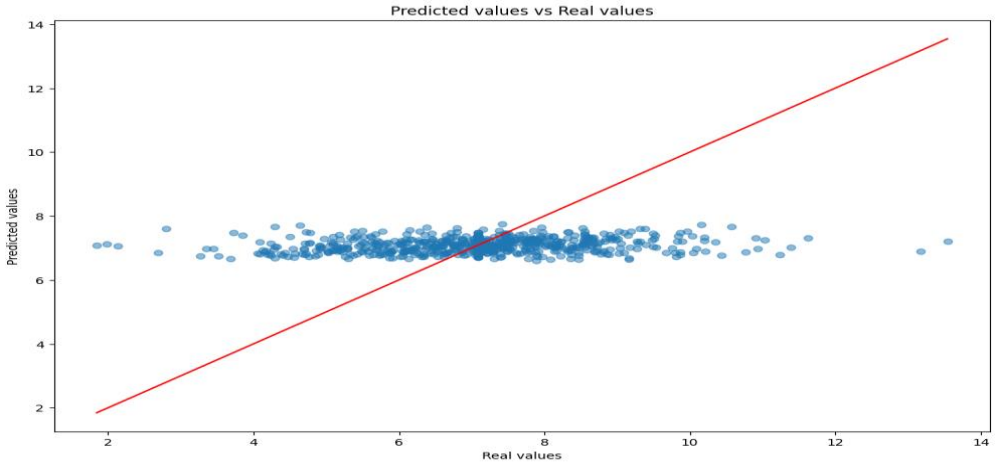


Figure III-6 : Predicted values vs real values of model 1

3.2.2 Intelligent model 2

Model 2 leverages linear regression to minimize the gap between its predicted values and the true target values. The accompanying figure illustrates how to train and utilize such a model with scikit-learn in Python.

```
# Linear regression model training
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
```

Figure III-7 : Model 2 configuration

The following figure presents the results of the performance evaluation metrics calculated on the testing data after the model training process.

```
Linear regression- MSE : 0.01122561550770262
Linear regression- MAE : 0.07741412185654167
Linear regression- RMSE : 0.15015009155809972
```

Figure III-8 : Calculate model 2 performance

The graph figure III-9 visualizes predicted values alongside the actual real values when the red points represent the actual values, while the blue points show the model's predictions for those values.

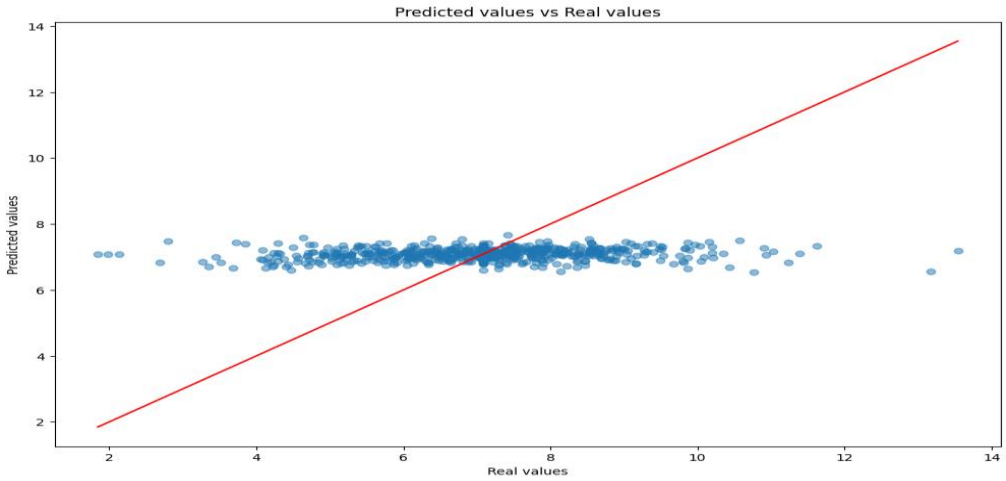


Figure III-9 : Predicted values vs real values of model 2

3.2.3 Intelligent model 3

SVR a method from scikit-learn for regression, creates a high-dimensional plane (hyperplane) during training. This plane minimizes the difference between its predictions and the actual values.

Figure III-10 details the SVR parameters used for model creation.

The following figure presents the results of the performance evaluation metrics calculated on the testing data after the model training process.

```
SVR (MSE): 0.02
SVR (MAE): 0.11
SVR (RMSE): 0.15
```

Figure III-10 : Calculate model 3 performance

This graph figure visualizes predicted values alongside the actual real values when the red points represent the actual values, while the blue points show the model's predictions for those values.

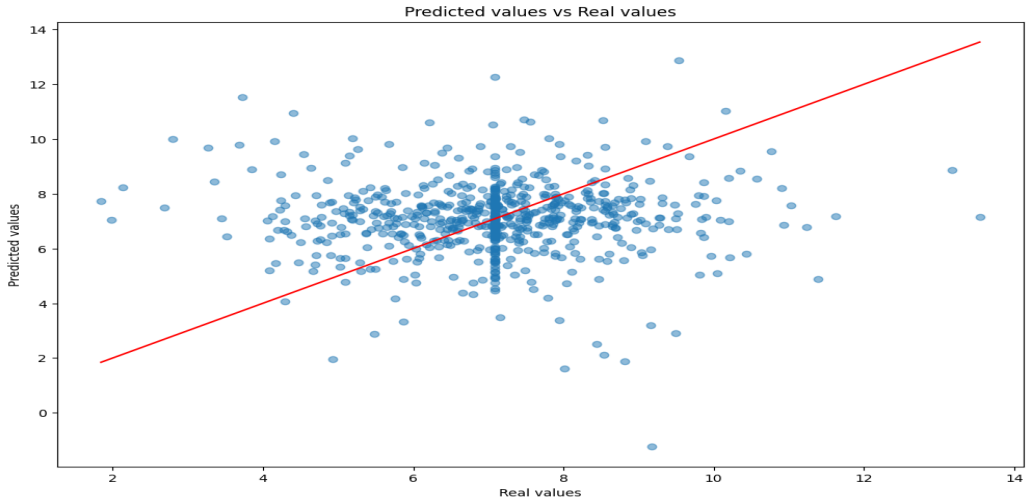


Figure III-11 : Predicted values vs real values of model 3

3.2.4 Intelligent model 4

The KNeighborsRegressor class implements k-Nearest Neighbors regression in scikit-learn. We can control its behavior through parameters like the number of neighbors (set to 5 here) and the weighting scheme (uniform weights).

The following figure presents the results of the performance evaluation metrics calculated on the testing data after the model training process.

```
KNN - MSE: 0.011725323280340788
KNN - MAE: 0.08187729781804323
KNN - rmse : 0.1082835318981644
```

Figure III-12 : Calculate model 4 performance

The graph figure III-13 visualizes predicted values alongside the actual real values when the red points represent the actual values, while the blue points show the model's predictions for those values.

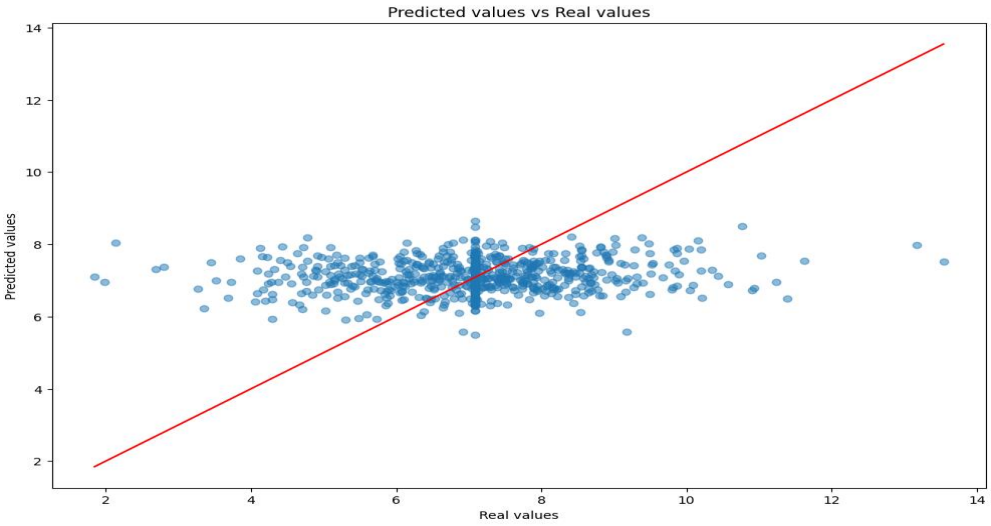


Figure III-13 : Predicted values vs real values of model 4

3.2.5 Intelligent model 5

Decision trees use a recursive approach to split data, aiming to reduce variation at each step. This leads to a tree structure that predicts the target variable based on learned decision rules from the data's features. Figure III-14 details decision tree parameters used for model creation.

```
(class) DecisionTreeRegressor(*, criterion: str =
"squared_error", splitter: str = "best", max_depth: Any |None
= None, min_samples_split: int = 2, min_samples_leaf: int =
1, min_weight_fraction_leaf: float = 0, max_features: Any |
None =None, random_state: Any | None = None, max_leaf_nodes:
Any | None = None, min_impurity_decrease: float = 0,
ccp_alpha: float = 0)
```

Figure III-14 : Model 5 configuration

The following figure presents the results of the performance evaluation metrics calculated on the testing data after the model training process.

```
Decision Tree - MSE: 0.010493189643299018
Decision Tree - MAE: 0.07502988549248447
Decision Tree - RMSE : 0.10243627113136743
```

Figure III-15 : Calculate model 5 performance

This graph figure visualizes predicted values alongside the actual real values when the red points represent the actual values, while the blue points show the model's predictions for those values.

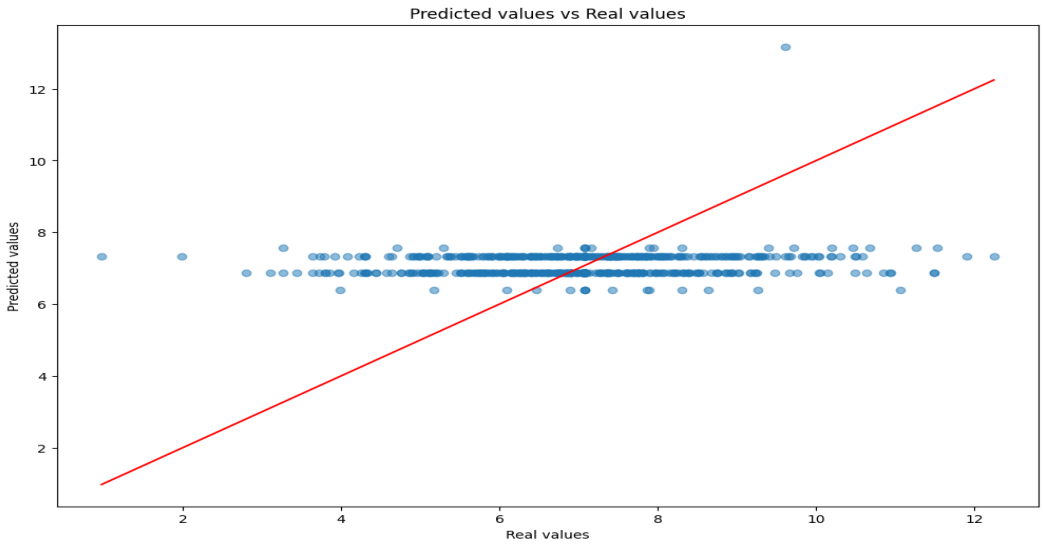


Figure III-16 : Predicted values vs real values of model 5

3.2.6 Comparison of results

Below, we present two methods for comparing performance metrics (MSE, MAE, RMSE) of the 5 models using both column charts and a comparative table :

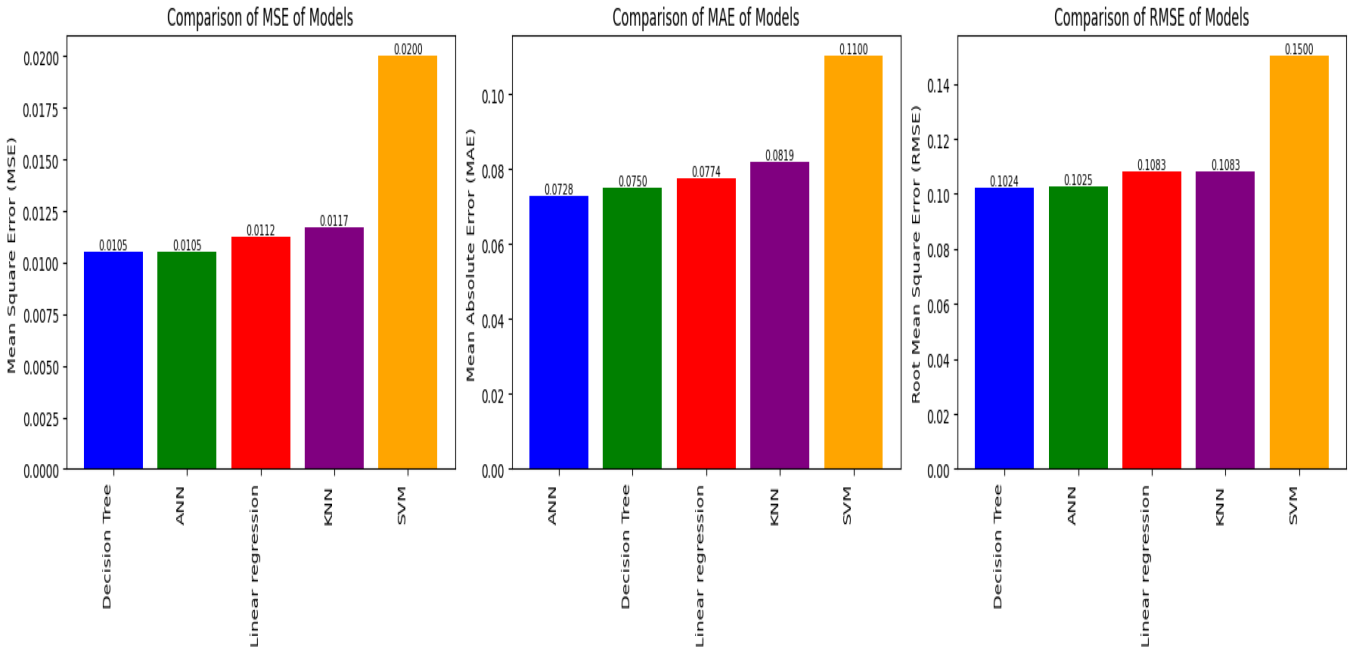


Figure III-17 : Graphic columns for comparison performance of models

Models	MSE	MAE	RMSE
Artificial neural network	0.010499338619410992	0.07279794663190842	0.10246628040194976
Linear regression	0.01122561550770262	0.07741412185654167	0.1082835318981644
SupportVector Regression	0.02	0.11	0.15
k-Nearest Neighbors	0.011725323280340788	0.08187729781804323	0.1082835318981644
Decision tree	0.010493189643299018	0.07502988549248447	0.10243627113136743

Table III-2 : Comparison of model results

According to the graph figure and the comparative table, we compare the performance metrics :

- MSE (Mean Squared Error) : The Artificial Neural Network (ANN) and Decision Tree models achieve the lowest MSE of 0.0105, indicating they generally provide more accurate predictions compared to Linear Regression (0.0112), Support Vector Regression (0.0200) and k-Nearest Neighbors (0.0117).
- MAE (Mean Absolute Error) : The ANN model has the lowest MAE of 0.0728, followed closely by the Decision Tree model with 0.0750. Linear Regression and k-Nearest Neighbors have higher MAE values, indicating slightly less accuracy in predicting individual data points.
- RMSE (Root Mean Squared Error) : Both the ANN and Decision Tree models show the lowest RMSE of 0.1025 and 0.1024, respectively, suggesting they are more precise in predicting deviations from the actual values compared to other models.

In summary, the Artificial Neural Network (ANN) and Decision Tree models exhibit superior performance across all three metrics (MSE, MAE, RMSE) compared to Linear Regression, Support Vector Regression and k-Nearest Neighbors. This suggests that the ANN model, in particular, is well-suited for the task of predicting the target variable based on the given data, offering lower errors and better predictive capabilities.

4 Water potability classification

4.1 Data Description

We use the same previous data just add a column of potability (0,1), It includes a total of 3276 observations distributed on 10 variable and figure show the variable distribution:

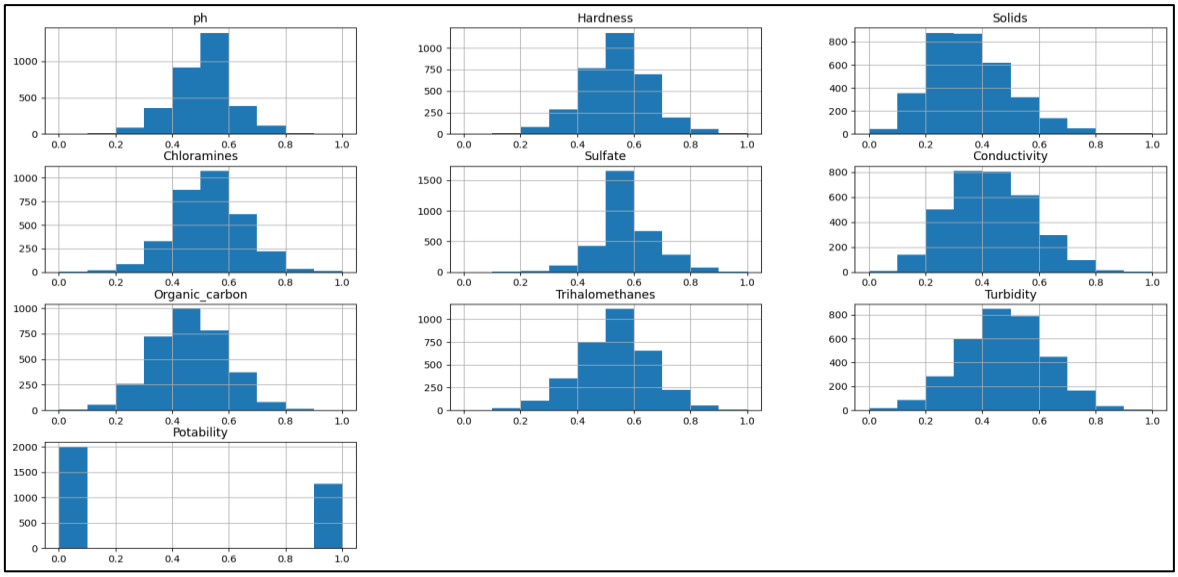


Figure III-18 : Variable distribution 2

We have 1998 variables not potable (0) and 1278 variables potable (1) like the figure III-19 show :

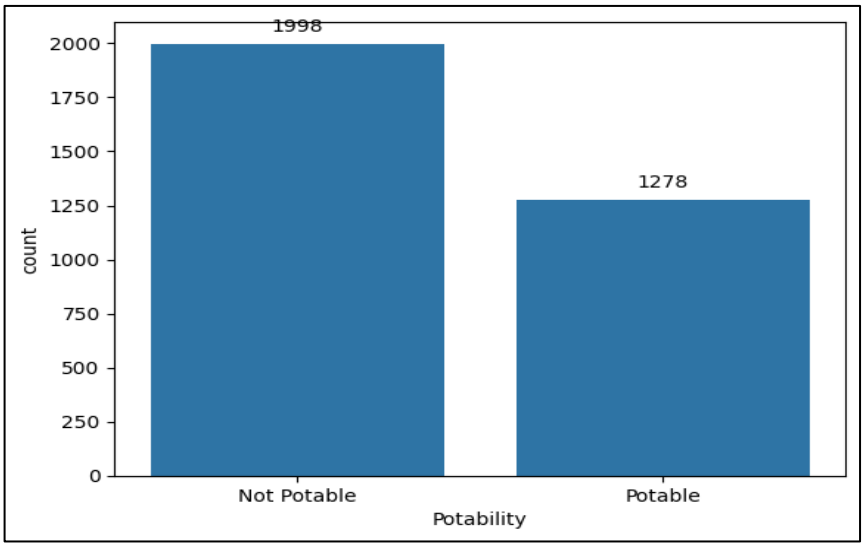


Figure III-19 : Potability

4.1.1 Correlation matrix

From the correlation matrix provided, the key points regarding the relationship between Potability and the other variables are :

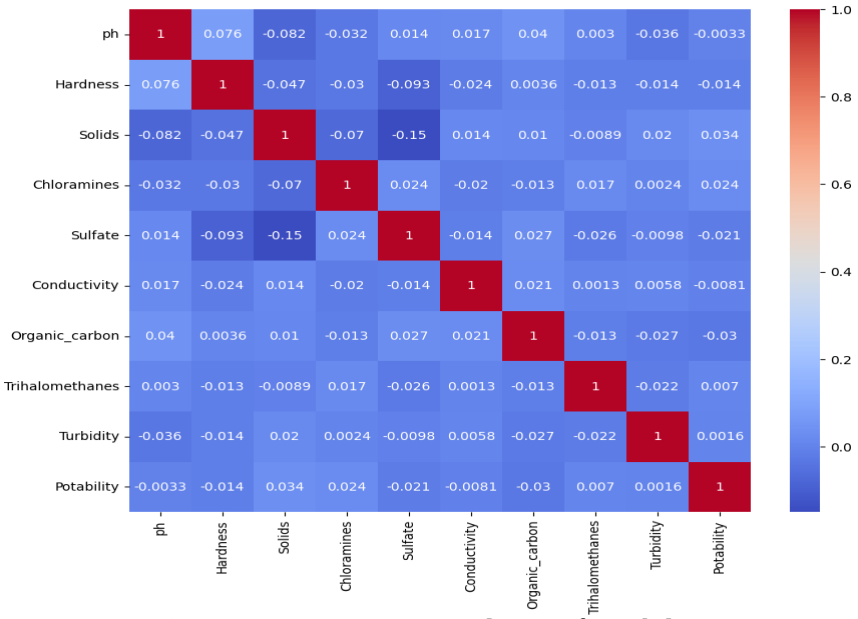


Figure III-20 : Matrix correlation of potability

The matrix correlation in “figure III-20” shows that potability has relatively weak relationships, both positive and negative, with the various water quality parameters. The strongest relationship is the weak positive correlation with Hardness, while the weak negative correlations with Sulfate and Organic_carbon are also noteworthy.

4.2 Creation of models and results

4.2.1 Intelligent model 1

The provided code defines a sequential neural network architecture with several fully connected layers, Dropout layers for regularization and a final sigmoid activation layer for binary classification. The model is compiled with RMSProp optimization, binary cross-entropy loss and accuracy as the evaluation metric.

```
# Define neural network architecture
model = Sequential([
    Dense(128, activation='relu', input_shape=(X_train.shape[1],)),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dropout(0.2),
    Dense(128, activation='relu'),
    Dense(288, activation='relu'),
    Dense(1, activation='sigmoid')
])

# Compile the model
model.compile(optimizer=RMSprop(), loss='binary_crossentropy', metrics=['accuracy'])
```

Figure III-21: Configuration of model 1 of potability1

After training the model, we assess its performance by computing standard metrics such as accuracy, precision, recall and F1 score, which are commonly used evaluation criteria. Below are the obtained results :

Artificial neural network accuracy : 0.6814.

	precision	recall	f1-score	support
0.0	0.71	0.83	0.77	412
1.0	0.60	0.43	0.50	244
accuracy			0.68	656
macro avg	0.66	0.63	0.63	656
weighted avg	0.67	0.68	0.67	656

Figure III-22 : Performance of model 1 of potability

The two following plots provided show the performance of the model in terms of Accuracy and Loss during the training and validation phases :

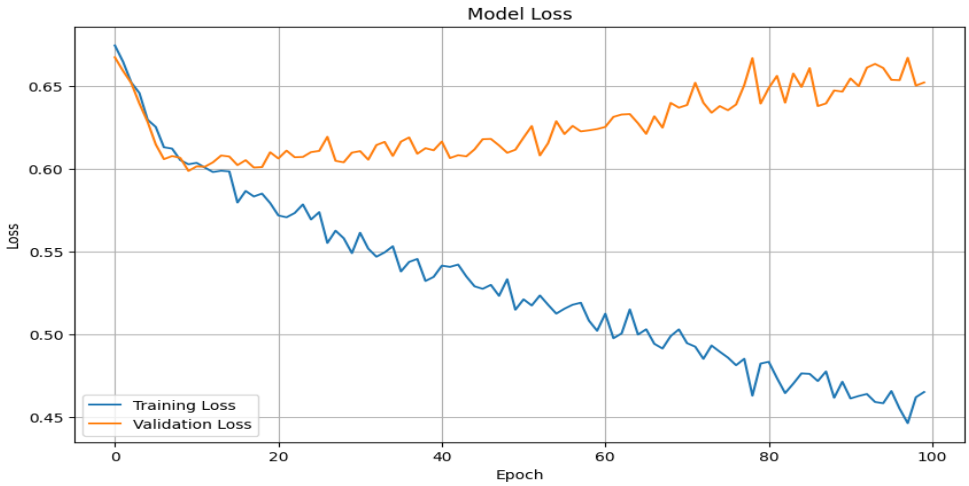


Figure III-23 : Loss during the training and validation phases

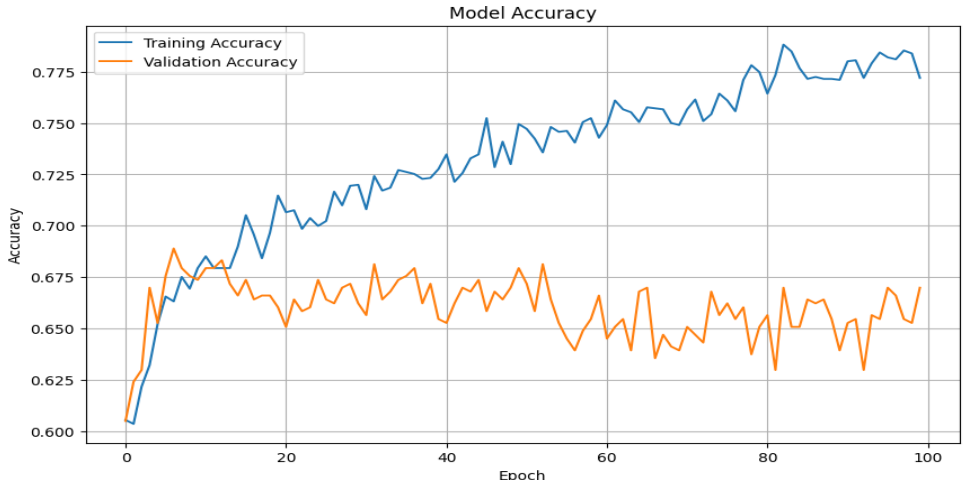


Figure III-24 : Accuracy during the training and validation phases

The validation loss is generally higher than the training loss, where we note that the training Loss converge towards a low level of loss, indicating the model is effectively minimizing the error during training.

The validation accuracy is generally lower than the training accuracy, but they both converge towards a high level of accuracy, indicating the model is learning and generalizing well.

In summary, the model appears to be learning effectively, as evidenced by the increasing accuracy and decreasing loss over the training epochs.

- Predicted vs Actual values :

  This graph figure visualizes the predicted values alongside the actual values, where the red points represent the actual values and the blue points denote the model's predictions corresponding to those values.

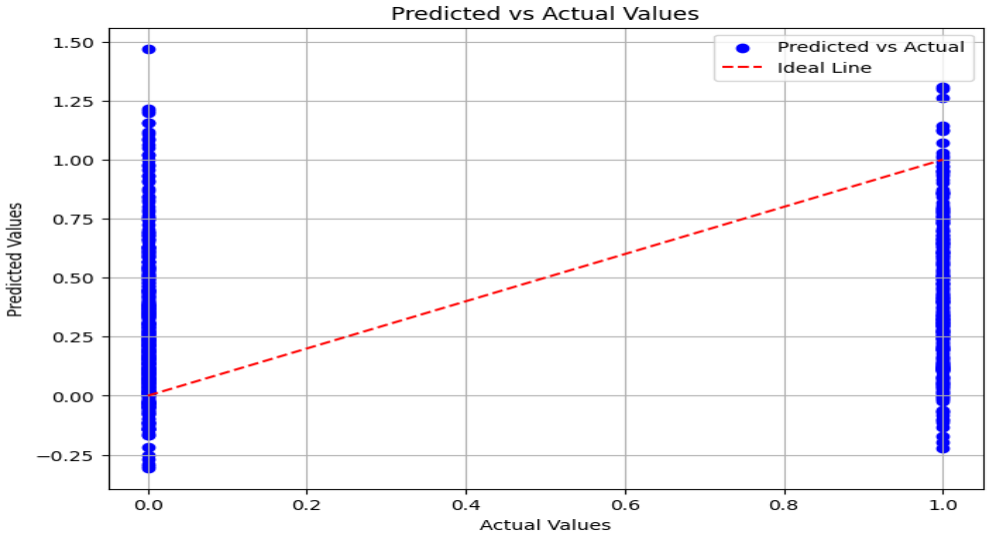


Figure III-25 : Predicted vs actual values of model 1 of potability

The blue scatter points represent the predicted values plotted against the actual values.

The scatter points are generally distributed along the dotted red "Ideal Line", which represents a perfect 1:1 correspondence between predicted and actual values.

This figure shows that the model is performing well, with a strong linear relationship between the predicted and actual values, minimal outliers or bias, and a good coverage of the full range of the data. This visually confirms the model's ability to accurately predict the target values based on the input features.

4.2.2 Intelligent model 2

The LogisticRegression class in scikit-learn implements logistic regression, a popular algorithm used for both binary and multi-class classification tasks. The figure below illustrates its configuration.

```
(class) LogisticRegression(penalty: str = "l2", *, dual: bool =
False, tol: float = 0.0001, C: float = 1, fit_intercept: bool = True,
intercept_scaling: int = 1, class_weight: Any | None = None,
random_state: Any | None = None, solver: str = "lbfgs", max_iter:
int = 100, multi_class: str = "auto", verbose: int = 0, warm_start:
bool = False, n_jobs: Any | None = None, l1_ratio: Any | None = None)
```

Figure III-26 : Configuration of model 2 of potability

The following figure illustrates the results of the performance evaluation metrics calculated on the testing data after the model training process.

Logistic Regression accuracy : 0.6284658040665434.

	precision	recall	f1-score	support
0.0	0.63	1.00	0.77	680
1.0	0.00	0.00	0.00	402
accuracy			0.63	1082
macro avg	0.31	0.50	0.39	1082
weighted avg	0.39	0.63	0.49	1082

Figure III-27 : Performance of model 2 of potability

4.2.3 Intelligent model 3

The DecisionTreeClassifier is a class in the scikit-learn library that implements the decision tree algorithm for classification tasks. The class has several configuration parameters that can be adjusted to customize the behavior of the decision tree model and below the configuration of our model :

```
(class) DecisionTreeClassifier(*, criterion: str = "gini",
splitter: str = "best", max_depth: Any | None = None,
min_samples_split: int = 2, min_samples_leaf: int = 1,
min_weight_fraction_leaf: float = 0, max_features: Any | None
= None, random_state: Any | None = None, max_leaf_nodes: Any
| None = None, min_impurity_decrease: float = 0,
class_weight: Any | None = None, ccp_alpha: float = 0)
```

Figure III-28 : Configuration of model 3 of potability

The following figure illustrates the results of the performance evaluation metrics calculated on the testing data after the model training process.

Decision Tree accuracy : 0.6128048780487805.

	precision	recall	f1-score	support
0.0	0.65	0.81	0.72	412
1.0	0.47	0.28	0.35	244
accuracy			0.61	656
macro avg	0.56	0.54	0.54	656
weighted avg	0.58	0.61	0.58	656

Figure III-29 : Performance of model 3 of potability

4.2.4 Intelligent model 4

The KNeighborsClassifier is a class in the scikit-learn library that implements the k-nearest neighbors (KNN) algorithm for classification tasks. The class has several configuration parameters that can be adjusted to customize the behavior of the KNN model like the code below show :

```
(class) KNeighborsClassifier(n_neighbors: int = 5, *, weights: str =
"uniform", algorithm: str = "auto", leaf_size: int = 30, p: int = 2, metric:
str = "minkowski", metric_params: Any | None = None, n_jobs: Any | None = None)
```

Figure III-30 : Configuration of model 4 of potability

The following figure illustrates the results of the performance evaluation metrics calculated on the testing data after the model training process.

K-nearest neighbors accuracy : 0.6585365853658537.

	precision	recall	f1-score	support
0.0	0.67	0.91	0.77	412
1.0	0.61	0.23	0.34	244
accuracy			0.66	656
macro avg	0.64	0.57	0.55	656
weighted avg	0.64	0.66	0.61	656

Figure III-31 : Performance of model 4 of potability

4.2.5 Intelligent model 5

The SVC (Support Vector Classifier) is a class in the scikit-learn library that implements the support vector machine (SVM) algorithm for classification tasks. The code below show our configuration with this class :

```
(class) SVC(*, C: float = 1, kernel: str = "rbf", degree: int = 3, gamma: str = "scale",
coef0: float = 0, shrinking: bool = True, probability: bool = False, tol: float = 0.001,
cache_size: int = 200, class_weight: Any | None = None, verbose: bool = False,
max_iter: int = -1, decision_function_shape: str = "ovr", break_ties: bool =
```

Figure III-32 : Configuration of model 5 of potability

The following figure illustrates the results of the performance evaluation metrics calculated on the testing data after the model training process.

Support Vector Classifier accuracy : 0.6692073170731707.

	precision	recall	f1-score	support
0.0	0.69	0.86	0.77	412
1.0	0.60	0.34	0.43	244
accuracy			0.67	656
macro avg	0.64	0.60	0.60	656
weighted avg	0.65	0.67	0.64	656

Figure III-33 : Performance of model 5 of potability

4-2-6 Comparison of results

Below, we present two methods for comparing performance metrics (accuracy, precision, recall and f1 score) of the 5 models using both column charts and a comparative table :

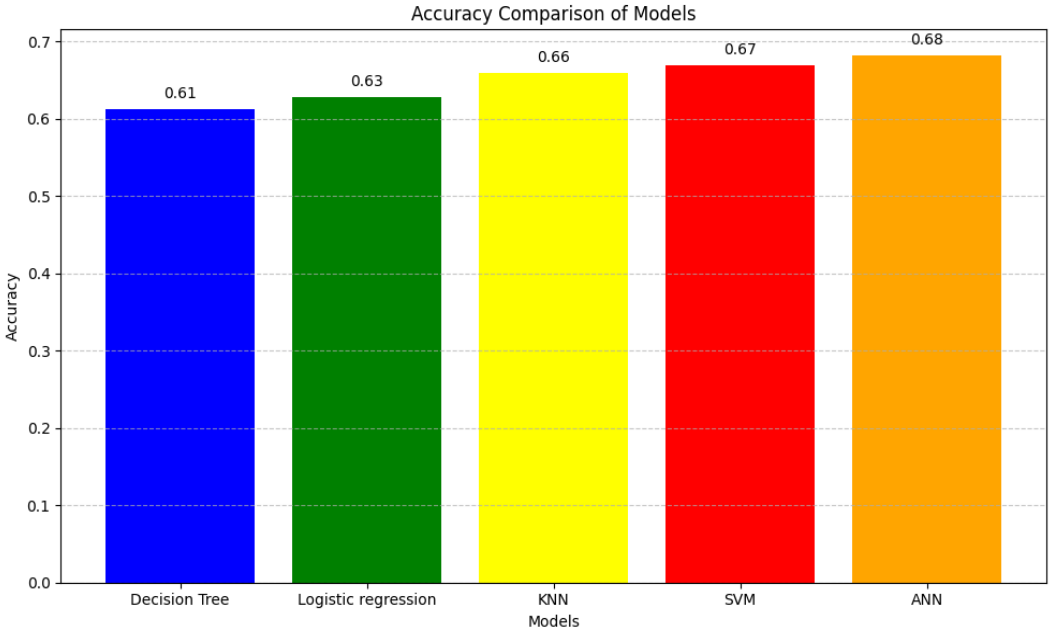


Figure III-34 : Accuracy comparison of models

Models	Accuracy	Precision (class 0)	Precision (class 1)	Recall (class 0)	Recall (class 1)	F1-score (class0)	F1-score (class1)
Artificial neural network	0.68	0.70	0.57	0.82	0.41	0.75	0.48
Logistic Regression	0.63	0.63	0	1	0	0.77	0
Decision Tree	0.61	0.65	0.47	0.81	0.28	0.72	0.35
Support Vector Regression	0.66	0.67	0.61	0.91	0.23	0.77	0.34
k-Nearest Neighbors	0.67	0.69	0.60	0.86	0.34	0.77	0.43

Table III-3: Comparison of model results of potability

According to the graph figure and the comparative table, we conclude that the ANN model outperforms the other models in terms of overall accuracy, precision class 0 and recall and F1-score for class 1. The only areas where the ANN is not the top performer are recall for class 0 (where Logistic Regression is better) and F1-score for class 0 (where Logistic Regression, Support Vector Regression and kNN are slightly better).

The key strengths of the ANN model are its ability to maintain a good balance between precision and recall. This makes it a strong contender overall, particularly if the goal is to maximize performance across both classes.

5 Conclusion:

The chapter's findings demonstrate that ANNs excel in both predicting and classifying data. Furthermore, their built-in fault tolerance allows them to maintain reasonable output even with partial corruption, making them less susceptible to errors compared to conventional prediction techniques.

General Conclusion

General Conclusion

After the studies we conducted in the first, second and third chapters, we conclude that:

Artificial neural networks (ANNs) offer a powerful approach to prediction because they can effectively handle incomplete, noisy and complex data. This makes them well-suited for situations where data quality may not be ideal. ANNs also achieve high accuracy and continuously learn and improve with exposure to new information through training on additional data. And the more data they are fed, the better they become.

These capabilities make ANNs a valuable asset in various fields that require precise predictions to inform decision-making, such as water quality assessment. In our case, that makes (ANNs) a perfect fit for prediction.

And also compared to conventional analytical methods, ANNs can provide more affordable water quality evaluations. They can assist water managers in making informed decisions to protect water resources and quickly respond to potential pollution or contamination issues.

The increasing use of neural network-based approaches in water quality monitoring represents a significant step forward in ensuring access to safe and sustainable water, which is essential for human and environmental well-being.

In the end, the hope is that this work can serve as a foundation for more in-depth studies and further development in this area. We hope that this contribution can pave the way for new advancements in the field of water quality assessment and monitoring. Ultimately, the goal is to stimulate deeper exploration and research into this crucial issue, driving continued progress and innovation

Bibliographic References

Bibliographic References

[1]https://phys.org/news/2024-07-efficient-quality-future-scarcity.html; Consulted on 2/6/2024.

[2]culligan.fr/conseilsqueeau/?fbclid=IwAR0USBIgDRCnjGB7RGXvWx2v6C5Lgm42aGrL36Ryxd5emt7Cv2LNQdlwuNY;Consulted on 3/3/2024.

[3]https://www.cieau.com/espace-enseignants-et-jeunes/les-enfants-et-si-on-en-apprenait-plus-sur-leau-du-robinet/la-definition-de-leau-potable); Consulted on 3/3/2024.

[4] J. Rodier, " L'Analyse de l'Eau : Eaux Naturelles, Eaux Résiduaires, Eau de Mer ", 9 th edition, Paris, p-100, 2009.

[5] E. Popek, " Sampling and Analysis of Environmental Chemical Pollutants: Practical Approach to Sampling ", second edition, USA, P- 195, 2018.

[6] J. Rodier, " L'Analyse de l'Eau : Eaux Naturelles, Eaux Résiduaires, Eau de Mer ", 9 th edition, Paris, p-88, 1984.

[7] OMS.,1986 : Directives de qualité pour l'eau de boisson. Volume 2: 1ere Edition, Genève, P 134.

[8] https://www.reagent.co.uk/blog/what-is-ph-scale/; Consulted on 5/3/2024.

[9]https://www.Fondriest.Com/Environmental/Measurements/Parameters/WaterQuality/Conductivity-Salinity-Tds/; Consulted on 5/3/2024.

[10]https://www.ElectronicsNotes.Com/Articles/Basic_Concepts/Resistance/ElectricalConductivity-Conductance.Php; Consulted on 5/3/2024.

[11]https://Www.Freshwatersystems.Com/Blogs/Blog/What-Is-Tds-In-Water-Why-Should-YouMeasure-It; Consulted on 6/5/2024.

[12]https://pubmed.ncbi.nlm.nih.gov/29722685/#:~:text=Because%20TDS%20measurement%20is%20time,is%20a%20constant%20of%20proportionality; Consulted on 7/3/2024.

[13] Berne. F, Jean. C, 1991 : Traitement des eaux, Édition TECHNIP, P 306.

[14] Si Abderahmane O., 2016 : Contribution à l'évaluation du système management et qualité des paramètres physico-chimiques, bactériologiques et organoleptiques des eaux des stations de

traitement Taksebt et Boudouaou. Mémoire de master : Université Mouloud Mammeri de Tizi-Ouzou, Algérie.

[15] J. Rodier, "L'Analyse de l'Eau : Eaux Naturelles, Eaux Résiduaires, Eau de Mer : chimie, physico-chimie, microbiologie, biologie, interprétation des résultats", 8 th edition, Paris, p- 200 ,2005.

[16] S.C. Edberg, E.W. Rice, R.J. Karlin, M.J. Allen, " *Journal of Applied Microbiology* ", Volume 88, Issue S1, Pages 106S–116S, 2000.

[17] OMS. (2006). P436. World Health Organization.

[18] National Academy of Sciences. (1977). Drinking water and health. Part 1. Ch. 1–5. A report of the Safe Drinking Water Committee Advisory Center on Toxicology Assembly of Life Sciences. Washington, DC: U.S. National Research Council.

[19] https://www.inbw.be/parametres-et-normes; Consulted on 8/3/2024.

[20] https://www.lenntech.fr/sulfates.htm; Consulted on 8/3/2024.

[21]https://www.cieau.com/leau-et-votre-sante/qualite-de-leau/quelles-normes-de-qualite-pour-leau-potable/; Consulted on 10/3/2024.

[22]https://blogs.worldbank.org/water/how-test-water-quality-chemical-tests limitedbudgets#_ftnref1; Consulted on 12/3/2024.

[23]https://www.ipgp.fr/~losno/Manips/pH/appareilsdemesure.html#:~:text=Le%20pHm%C3%A8tre%20est%20un,%C3%A9lectrode%20qui%20mesure%20cette%20valeur;Consulted on 12/3/2024.

[24]https://www.mrs-scientific.com/equipment/instruments/ph-meters/benchtop-ph-meters/hanna-white-edge-ph-meter-with-ph-probe-hi-11310/; Consulted 14/3/2024.

[25]https://www.francebiotechnologies.fr/sous-rubrique/mesure-conductivite-et-turbidite; Consulted 14/3/2024.

[26]https://www.fr.endress.com/fr/instrumentation-terrain-sur-mesure/analyse-liquides produits/capteurs-transmetteurs-conductivite; Consulted 20/3/2024.

[27] https://kalstein.eu/how-does-a-turbidimeter-work/?lang=en; Consulted 20/3/2024.

[28]https://www.google.com/amp/www.water-chemistry.in/2010/11/working-principle-of nephelometric-turbidity-meter/amp/; Consulted 20/3/2024.

[29]https://www.shop.cifec.fr/11oxymetre#:~:text=Cet%20outil%20mesure%20la%20quantit%C3%A9,eau%20est%20l'oxym%C3%A9trie%20optique; Consulted 22/3/2024.

[30] https://wikiwater.fr/e27-methodes-et-moyens-disponibles; Consulted 22/3/2024.

[31]GASPAR D.,2010 : Applications de l'apprentissage artificiel à la modélisation systémique de la chaîne hydrometeorologique pour la prévision des crues éclair.

[32] Soori,B.Arezoo,R.Dastres, "*Machine learning and artificial intelligence in CNCmachine tools, A review*",Sustainable Manufacturing and Service Économies,1(03950000),pp.3.2023

[33] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[34]https://www.linkedin.com/pulse/applications-supervised-learning-priya-ramesh-zgfcc;

Consulted on 2/4/2024.

[35] Santos, I. R., & Burnett, W. C. (2011). Tracing of submarine groundwater discharge. In W. C. Burnett, H. Dulaiova, M. R. Haese, & P. K. Swarzenski (Eds.), *Measurement Techniques in Space Plasmas: Fields* (pp. 978-080). Academic Press.

[36] Ben-Akiva, M., &Bierlaire, M. (2021). Discrete choice analysis: Theory and application to travel demand. *European Transport Research Review*, 13(3), Article 76.

[37] https://www.javatpoint.com/regression-vs-classification-in-machine-learning; Consulted on 4/4/2024.

[38] https://statorials.org/types-de-regression/; Consulted on 6/4/2024.

[39] https://eastgate-software.com/what-is-unsupervised-learning/; Consulted on 6/4/2024.

[40] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. In W. C. Burnett, H. Dulaiova, M. R. Haese, & P. K. Swarzenski (Eds.), *Measurement Techniques in Space Plasmas: Fields* (pp. 978-080). Academic Press.

[41] https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning; Consulted on 6/4/2024.

[42] Han, J., &Kamber, M. (2000). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann Publishers.

[43]https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligenceartificielle/1501879-machine-a-vecteurs-de-support-svm-definition-et-cas-d-usage/; Consulted 10/4/2024.

[44] https://mrmint.fr/introduction-k-nearest-neighbors; Consulted on 10/4/2024.

[45]https://www.ibm.com/topics/randomforest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems; Consulted on 12/4/2024.

[46] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.

[47] World Scientific. (2016). Understanding deep learning: A review. *International Journal of Neural Systems*, 26(7), Article 16500045.

[48]C.Touzet,"Les reseaux de neurones artificiels,introduction au connexionisme", V(01338010), pp. 30. 2016.

[49] https://en.wikipedia.org/wiki/Biological_neuron_model; Consulted 15/4/2024.

[50] A.Mohamed Yessin, "Miseen œuvredereseauxde neuronespour la modelisationde cinetiques reactionnelles en vue de la transposition BATCH/CONTINU". thèse de doctorat, DEA Procédés et Matériaux, École Nationale d'ingénieurs de Sfax, Tunisie, 2007.

[51] https://www.researchgate.net/figure/Sketch-of-an-artificial-neuron_fig2_355514423; Consulted 15/4/2024.

[51] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1511.08458*.

[52] Feng, J., Ma, S., & Zheng, J. (2013). An extended linear combination method for complete ranking of fuzzy numbers. *Fuzzy Sets and Systems*, 231, 89-99.

[53] GASPAR D.,2010 : Applications de l'apprentissage artificiel à la modélisation systémique de la chaîne hydrometeorologique pour la prévision des crues éclair.

[54] https://builtin.com/machine-learning/relu-activation-function; Consulted on 25/3/2024.

[55] Wang, Z., & Xu, W. (2021). Mobile Edge Computing: Fundamentals and Challenges. *Handbook of Big Data Technologies*, 11, 229-245. https://doi.org/10.1016/B978-0-12-820601-0.00011-2

[56] Zhang, Q., Wu, W., Liu, Y., Wang, Z., & Zou, H. (2017). Machine learning and its applications in geotechnical engineering. *Handbook of Statistics*, 35, 79-121.

[57] HOUAMED I.,201: Détection de l'onde P dans un Signal ECG.Magister en électronique.

[58] D.Etiemble,F.Auzanneau,"*Réseaux de neurones profonds(DNN)*".V1(3730),pp.5.2023.

[59] Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* 2, 420 (2021).

[60]https://fastercapital.com/fr/sujet/exploration-de-l'erreur-quadratique-moyenne-(mse).html; Consulted on 22/3/2024.

[61]https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df; Consulted on 22/3/2024.

[62] https://c3.ai/glossary/data-science/root-mean-square-error-rmse/; Consulted on 24/3/2024.

[63] https://klu.ai/glossary/accuracy-precision-recall-f1; Consulted on 24/3/2024.

Annexes

Libraries used

Annexe 1: Numpy

NumPy is the abbreviation for «Numerical Python» and it is a fundamental set for scientific computation in Python. NumPy provides Python with an extensive mathematical library that can perform numerical calculations effectively and efficiently in order to work with multidimensional arrays and matrix data structures, very common in data science and machine learning.

Annexe 2 : Pandas

Pandas is a package for data manipulation and analysis in Python. The name Pandas is derived from the term «Panel Data». Pandas integrates two additional data structures into Python, namely Pandas Series and Pandas DataFrame. These data structures allow us to work with labeled and relational data in a simple and intuitive way.

The recent success of machine learning algorithms is in part due to the huge amounts of data we have to train our algorithms. However, when it comes to data, quantity is not the only thing that matters, data quality is just as important. Often large data sets are not ready to be integrated into learning algorithms. Most often, large datasets will often have missing values, outliers, incorrect values, etc... Having data with many missing or bad values, for example, will not allow algorithms machine learning to work well. Therefore, a very important step in machine learning is to first examine the data and ensure that it is well suited to your training algorithm by performing a database analysis. This is where pandas come in. Pandas Series and DataFrames are designed for fast data analysis and manipulation, while being flexible and easy to use. Here are some features that make Pandas an excellent set for data analysis:

- Allows the use of labels for rows and columns.
- Can calculate continuous statistics on time series data.
- Easy handling of NaN values.
- Is able to load data of different formats into DataFrames.
- Can combine and merge different data sets.
- It integrates with NumPy and Matplotlib.

For these reasons, among others, Pandas DataFrames have become one of the most used Pandas objects for data analysis in Python.

Annexe 3: Matpolib

Matplotlib is currently the most popular Python library for producing plots and other 2D data visualizations. Since data analysis requires visualization tools, Matplotlib is the most suitable library for this task.

With Matplotlib, one can easily generate a wide variety of charts, such as charts, histograms, bar charts, scatter charts, etc., using only a few lines of Python code. This allows you to focus on generating diagrams for faster data analysis and exploration, rather than wasting time looking for solutions.

Key features that have made Matplotlib the most widely used tool for graphical representation of data include:

- Progressive development and interactive data visualization
- Finer control of graphic elements
- The ability to export visualizations in many formats, such as PNG, PDF, SVG and EPS

In addition, the Seaborn and Pandas libraries' built-in layout features are built on the basis of Matplotlib.

Annexe 4: Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of Matplotlib and integrates closely with Pandas data structures.

Seaborn helps explore and understand data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of plots mean, rather than on the details of how to draw them.

Annexe 5: Scikit-Learn

Scikit-Learn is a powerful open-source machine learning library for Python. It relies on the NumPy, SciPy and Matplotlib libraries to provide simple and effective tools for common data analysis tasks.

Key features of Scikit-Learn include:

- Comprehensive algorithm set: Scikit-Learn implements a wide variety of supervised and unsupervised learning algorithms, such as linear models, decision trees, random forests, support vector machines, nearest k neighbors, etc.

- Consistent and intuitive API: Scikit-Learn follows a consistent API design, making it easy to learn and use. All models share common methods like fit(), predict() and score(), simplifying the experimentation of different algorithms.
- Data preprocessing: Scikit-Learn offers many tools for data preprocessing, such as missing value management, categorical variable encoding, feature scaling, and dimensionality reduction.
- Model evaluation and selection: Scikit-Learn includes various metrics and cross-validation techniques to evaluate model performance. It also provides tools for model selection, such as grid search and random search, to find the best hyperparameters.
- Efficiency and scalability: Scikit-Learn is designed to be efficient and scalable, with many algorithms implemented in Cython or using optimized libraries NumPy and SciPy.

Annexe 6: TensorFlow

TensorFlow is an open-source digital computing and machine learning library developed by Google. It is mainly used for building and deploying deep learning (deep learning) models and other machine learning algorithms.

To improve the performance of machine learning models, TensorFlow allows execution on both processors (CPUs) and graphics cards (GPUs). However, the higher performance capabilities of TensorFlow can be discovered when using GPUs.

TensorFlow currently supports frontend interfaces for a number of programming languages. Although TensorFlow has official APIs in several languages, the Python API is currently the most comprehensive and widely adopted by machine learning and deep learning practitioners. In addition to the Python API, TensorFlow also has an official API in C++.

TensorFlow leverages a broad ecosystem of related components, including libraries like TensorBoard for visualization, as well as deployment and production APIs to help scale models in real-world environments.

Annexe 7: Keras

Keras is a high-level deep learning API that greatly simplifies the design, training and use of neural networks. It can work using different machine learning frameworks like TensorFlow, Theano or Microsoft Cognitive Toolkit (formerly CNTK).

TensorFlow provides its own implementation of the Keras API, called tf.keras, which benefits from some advanced features of TensorFlow, such as efficient data loading.

Although Keras remains a separate library from TensorFlow, it can now be officially imported and used directly within TensorFlow, without requiring a separate installation.

The Keras API is particularly easy to use, allowing developers to build models by simply adding layers one after the other by simple function calls. This highly abstract and intuitive approach greatly facilitates the prototyping and development of deep learning solutions.