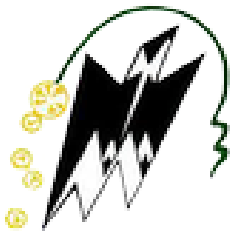
République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mouloud Mammeri de Tizi-Ouzou



Faculté De Génie Électrique et d'Informatique
Département de Télécommunications

Mémoire de Fin d'Etudes

de MASTER ACADEMIQUE
Filière :
Télécommunication
Spécialité :
Réseaux et Télécommunications

Par

CHETTOUM SELMA
AMAZOUZ MELLISSA

Thème

Optimisation de l'utilisation des ressources d'eaux en Algérie à l'aide des méthodes d'apprentissage automatique

Soutenu le : 27/06/2024

Devant le jury :

Président :	Mme. HAMMAR KARIMA	Univ. UMMTO
Promoteur :	Mme. KHIALI LYNDA	Univ. UMMTO
Examinateurs :	Mme. BECHA TASSADIT	Univ. UMMTO

Abstract

In the field of water resource management, local authorities and distribution network managers are particularly interested in predicting water use to make informed decisions, minimize the risk of shortages, develop optimized distribution strategies, and improve operational efficiency. As part of our final project, we focused on modeling and predicting the amount of water consumed in the region of Tizi Ouzou town. We achieved this by using time series combined with machine learning algorithms and deep learning. The data used for this study includes water consumption time series for each user going back 21 years. We tested a variety of forecasting methods to determine the most accurate ones, including linear regression, KNN, SVR, decision trees, random forests, and XGBoost for machine learning methods; as well as artificial neural networks (ANN), convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) models for deep learning methods. The performance of these models was evaluated using metrics such as Mean Squared Error (MSE) and the coefficient of determination (R^2). The results show that deep learning models, particularly RNN, offer superior accuracy compared to traditional machine learning methods.

Keywords: Machine learning, time series forecasting, water use prediction, deep learning, forecasting.

Résumé

Dans le domaine de la gestion des ressources en eau, les autorités locales et les gestionnaires des réseaux de distribution sont particulièrement intéressés par la prédiction de la consommation d'eau afin de prendre des décisions éclairées, minimiser les risques de pénurie, élaborer des stratégies de distribution optimisées, et améliorer l'efficacité opérationnelle. Dans le cadre de notre projet de fin d'étude, nous nous sommes concentrés sur la modélisation et la prédiction de la quantité d'eau consommée dans la région de Tizi Ouzou Ville. Nous avons réalisé cela en utilisant des séries temporelles combinées avec des algorithmes d'apprentissage automatique et d'apprentissage profond. Les données utilisées pour cette étude comprennent les historiques de consommation d'eau pour chaque utilisateur sur une période de 21 ans. Nous avons testé une variété de méthodes de prévision pour déterminer les plus précises, notamment : la régression linéaire, KNN, SVR, les arbres de décision, les forêts aléatoires, et XGBoost pour les méthodes d'apprentissage automatique ; ainsi que les réseaux de neurones artificiels (ANN), les réseaux de neurones convolutifs (CNN), les réseaux de neurones récurrents (RNN), et les modèles de mémoire à long terme (LSTM) pour les méthodes d'apprentissage profond. La performance de ces modèles a été évaluée à l'aide de métriques telles que l'erreur quadratique moyenne (MSE) et le coefficient de détermination (R^2). Les résultats montrent que les modèles d'apprentissage profond, en particulier RNN, offrent une précision supérieure par rapport aux méthodes traditionnelles d'apprentissage automatique.

Mots-clés : Apprentissage automatique, prévision de séries temporelles, prédiction de la consommation d'eau, apprentissage profond.

Table of Contents

List of Figures

List of Tables

Introduction Générale

Sur Terre, il existe plusieurs ressources naturelles indispensables à la vie humaine, jouant un rôle crucial dans le fonctionnement des écosystèmes, ainsi que dans la santé économique, sociale et environnementale du monde[1]. La ressource la plus vitale est l'eau; aucun organisme ne peut vivre sans elle[2], en particulier sans eau douce.

Cependant, nous n'avons pas accès à toute l'eau douce, car elle est soit confinée sous forme de glace, soit se trouve profondément dans des milieux souterrain, ce qui fait qu'environ 1% seulement est facilement accessible. Cette eau douce est utilisée dans plusieurs activités, notamment l'agriculture en consommant une proportion significative par rapport à d'autres usages.

L'Algérie compte parmi les pays qui disposent de ressources en eau limitées, provenant à la fois des milieux souterrains et de surface. Cependant, ces ressources sont constamment menacées par des pénuries, particulièrement dans le nord du pays, en raison de problèmes liés au développement économique et social. Ces problèmes comprennent des pertes et gaspillages d'eau, des traitements inefficaces, la dégradation des infrastructures et une protection insuffisante des ressources en eau.[3]. Pour cela, il est important de concevoir des stratégies et de trouver des solutions.

Afin de remedier a cette problématique, nous proposons une étude qui permet de prédire la consommation future d'eau dans la Wilaya de Tizi-Ouzou. Nous utilisons des séries temporelles (données historiques) allant de 2003 à 2023 pour chaque utilisateur de chaque région de la province, ainsi que des méthodes d'apprentissage automatique et d'apprentissage profond pour analyser ces données.

Notre étude se divise en quatre parties essentielles que nous allons décrire pour une meilleure compréhension de notre travail :

Dans la première partie (Chapitre I), nous introduison le contexte général de notre étude.

La deuxième partie (Chapitre II) nous avons résumé des articles et études précédentes portant sur la prédiction de la consommation d'eau à travers ces méthodes. Nous récapitulons les conclusions et les approches méthodologiques employées dans ces recherches afin de situer notre propre démarche dans ce contexte.

La troisième partie (Chapitre III) est dédiée à la méthodologie de notre étude. Nous décrivons en détail le schéma général de notre étude, ainsi que chaque étape de notre processus de prédiction. De plus, nous présentons et définissons les différentes méthodes utilisées, tant en appprentissage automatique qu'en apprentissage profond. Nous expliquons également les différentes métriques utilisées telles que : R-squared et MSE pour l'evaluation des performances de chaque méthode.

La quatrième partie (Chapitre IV) traite de la description des données utilisées dans notre étude. Nous définissons l'ADE (Algérienne des eaux) et ses fonctionnalités. Nous décrivons aussi le prétraitement des données et présentons une analyse détaillée comprenant des données relatives telles que les types d'abonnés, les quantités d'eau consommées, etc.

La cinquième partie (Chapitre V) présente les résultats expérimentaux de chaque méthode utilisée dans notre étude. Nous analysons les performances de chaque méthode à l'aide des métriques quantitatives.

References

[1] Importance of Natural Resources In Our Lives - Earth Reminder. (s. d.). Earth Reminder.

https://www.earthreminder.com/importance-of-natural-resources/: :text=Natural%20resources%20play%20an%20important%20role%20in%20our,role%2

[2] Why Are Natural Resources Important ? (25 Reasons). (s. d.). Enlightio.

https://enlightio.com/why-are-natural-resources-important: :text=Natural%20resources%20play%2

[3] Freshwater Resources. (s. d.-a). Education | national geographic society. https://education.nationalgeographic.org/resource/freshwater-resources/

Chapter I

Introduction

Natural resources are substances found on Earth that sustain life and fulfill human needs. These naturally occurring products include stone, sand, metals, oil, coal, natural gas, sunlight, air, water, soil, plants, animals, and birds. Natural resources are vital to our lives and are crucial for the world's economic, social, and environmental health [1]. They establish the foundation that sustains all living organisms. From the air, we inhale, courtesy of oxygen generated by planet life, to the nourishment we obtain from both plants and animals, our basic survival is deeply intertwined with these resources.

Water, perhaps the most vital natural resource, is fundamental to our existence. Every organism, from the smallest bacteria to the largest blue whale, relies on water to survive [2]. Potable water is a valuable asset on Earth's surface. It also serves as the habitat for numerous diverse fish, plant, and crustacean species. The environment the freshwater ecosystem offers encompasses lakes, rivers, ponds, wetlands, streams, and springs [3].

The majority of the world's freshwater is not readily reachable by humans. Roughly 69% of Earth's freshwater is confined in the shape of ice in glaciers and polar ice caps, and the additional 30% of Earth's freshwater is situated beneath the surface as groundwater. This results in only about one percent of Earth's freshwater being easily accessible for human utilization.

Accessibility to freshwater is also vital for economic progress. For instance, freshwater reservoirs facilitate the establishment of fisheries. Individuals across the globe extract fish from these habitats, supplying enough animal protein to sustain 158 million worldwide.

Apart from serving as a habitat for freshwater organisms, freshwater also plays a crucial role in other economic activities, such as agriculture. As per one estimation, approximately 70% of the world's freshwater is utilized for agriculture. Farmers globally employ

irrigation to convey water from the surface and groundwater sources to their fields. These agricultural activities engage over 1 billion people worldwide and produce over $2.4 trillion in economic value every year. In the future, demand for agricultural freshwater is expected to rise as global populations grow. According to one estimate, demand is projected to rise by 50% by 2050. This escalation in water usage will further strain Earth's limited freshwater supplies and emphasize the importance of accessing fresh water.

As in all other countries, in Algeria, fresh water is essential to meet the needs of the population, agriculture, and industry. Unfortunately, Algeria is among the countries with limited water resources. The available water mainly comes from surface water sources, such as rivers and dams, as well as groundwater, extracted through drilling and wells.
The availability of this vital resource is constantly threatened by shortages, particularly in the northern part of the country. These problems are mainly due to chaotic economic and social development, such as significant losses, waste, inefficient treatments, infrastructure degradation, and lack of protection for water resources. These difficulties are compounded by already unfavorable natural conditions[4].

Among the most important dams in Algeria is the Taksebt Dam. Located in the Tizi-Ouzou province, it plays a crucial role in supplying water to several provinces, including Boumerdès, Algiers, and Tizi-Ouzou. In addition to meeting the domestic needs of these regions, it also contributes to agricultural irrigation and supports industrial activities. In 2020, according to information from the Water Resources Directorate, the water level of the Taksebt Dam is only 18.8%, representing around 34 million cubic meters, of which only 20 million cubic meters can be mobilized. Consequently, the Directorate has decided to reduce the volume of available water[5]. The supply of drinking water in the Tizi-Ouzou province, as in all other provinces in the country, is severely disrupted. This distribution is becoming increasingly problematic over time due to the scarcity of water resources[6].

To address these management water issues, several measures need to be implemented. First, it is essential to modernize existing infrastructure to reduce losses and waste. Next, investing in more efficient and sustainable water treatment technologies is crucial. Additionally, implementing rigorous water management policies, including the conservation and protection of water resources, is indispensable. Finally, raising public awareness about the importance of water conservation and encouraging sustainable agricultural and industrial practices will also help mitigate these challenges.
We must also determine how to reduce water consumption to conserve rivers for all users. As populations increase and climate change modifies precipitation patterns globally, these conflicts over water will persist and become more frequent in the future. By taking these

comprehensive steps, we can address current drinking water issues while also preparing for and preventing future conflicts over water resources[3].

Our study proposes a promising solution to help manage water resources more effectively: the implementation of a water use prediction framework.

Such a prediction framework can help optimize water usage by forecasting demand based on various factors such as population growth, agricultural needs, and climate change. By analyzing time series data (historical data) while predicting consumption patterns, this framework can contribute to better allocation and conservation of water resources.

In this article, our objective is to forecast the future water use of the wilaya of Tizi-Ouzou using time series data collected from 2003 to 2023 as input through machine learning and deep learning methods and to compare their effectiveness.

References

[1] Importance of Natural Resources In Our Lives. Earth Reminder. https://www.earthreminder.com/importance-of-natural-resources

[2] Why Are Natural Resources Important ? (25 Reasons). (s. d.). Enlightio. https://enlightio.com/why-are-natural-resources-important

[3] Freshwater Resources. (s. d.-a). Education | national geographic society. https://education.nationalgeographic.org/resource/freshwater-resources/

[4] Boudjadja, A., Messahel, M. Pauc, H. (2003). Ressources hydriques en Algérie du Nord. Revue des sciences de l'eau / Journal of Water Science, 16(3), 285–304. https://doi.org/10.7202/705508ar.

[5] Barrage de Taksebt de Tizi Ouzou : Un taux de remplissage de 18.8% - Régions : EL Moudjahid.

https://www.elmoudjahid.dz/fr/regions/barrage-de-taksebt-de-tizi-ouzou-un-taux-de-remplissage-de-18-8-1679.

[6] Alimentation en eau potable à Tizi-Ouzou : Une nouvelle saison sèche en perspective. (s. d.). La Sentinelle. https://lasentinelle.dz/index.php/2022/04/07/alimentation-en-eau-potable-a-tizi-ouzou-une-nouvelle-saison-seche-en-perspective/.

Chapter II

Related Works

After examining several studies on water demand forecasting, we have identified different approaches and methodologies. These studies cover various aspects of water demand prediction, ranging from irrigation to urban consumption, as well as forecasting water levels and floods.
To better understand the challenges in this field, we have grouped the articles according to their main themes, which allow us to compare different modeling approaches, evaluate their effectiveness, and assess their applicability in different contexts. This synthesis of existing knowledge will enable us to better guide our own research and contribute to improving water resource management in the face of current and future challenges.

- **Short-term water demand forecasting:** In [1], the researchers introduce a methodology to improve daily irrigation district management. The study successfully created an ANN model for short-term forecasting of daily irrigation water demand. The created model combine ANN architecture, the Bayesian framework, and genetic Algorithms (GA), the methodology outperformed previous models, offering more accurate predictions. This methodology was implemented in the Bembézar MD Irrigation District in Southern Spain. Water volume records were compiled at the daily level for the 2010, 2012, and 2013 irrigation season. The data was split into two subsets, a training set, and a testing set, and for the features they used: water demand in the previous day (Demand_1) in m^3, and the water demand in two previous days(Demand_2) in m^3 day. The prediction accuracy of the models was assessed using Standard Error prediction(SEP) and determination coefficient(R^2) values, the top-performing models achieved R^2 values of 91% and 96%, with SEP values of 13.50% and 8.66% respectively.

- **Comparison of models for urban water consumption forecasting:** The objective of the study presented in [2], is to compare the effectiveness of Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) in predicting daily water

consumption. The aim is to determine which method, ANN or MLR, provides more accurate predictions for future daily water consumption values based on antecedent records of water consumption and meteorological variables, such as humidity. The study focuses on a relatively small urban area, exemplified by the city of Torun in Poland.

The data used in the study was obtained from Toruńskie Wodociągi LLC [3] and covered the years 2011 to 2013, comprising 1096 records. The input data for predicting water consumption included three antecedent values of water consumption, humidity as a meteorological variable, and dummy variables indicating the type of day(working days, Saturdays, Sundays, public holidays)and month. This processed data was split into three subsets: 767 records for training, 165 records for validation, and 164 records for testing.

The ANN model that included humidity(ANN_H) achieved the lowest mean absolute percentage error(MAPE) of 2.28%, outperforming other models. The analysis showed that including humidity improved the performance of ANN models for working days and Saturdays but not for Sundays and public holidays. The results indicated that ANN models were superior to MLR models. The study confirmed that ANN models are suitable for water demand forecasting.

In [4], the researches aims to predict the water demand of Shenyang City in China using a combination of Principal Component Analysis (PCA) and Back Propagation (BP) neural networks. The study initially considered nine socio-economic factors. After applying PCA, which was used to reduce these factors, five key features were retained: agricultural population, non-agricultural population, effectively irrigated area, industrial added value, and precipitation. These five principal components explain 97.6% of the total variance. The BP neural network model with these five retained features employed five input layers, ten hidden layer units, and one output layer, using training data from 2000 to 2011, and test data from 2012 and 2013 to predict 2014 and 2015. The prediction results were compared with those obtained using the index quota method, which is the most widely used method for water demand prediction in China. The model was evaluated by calculating the relative error between the actual and predicted values, which yielded a maximum error of 3.71% and a minimum error of 0.19%. The conclusion was that the BP model predictions are more accurate than those obtained by the index quota method in predicting water demand.

The purpose of the research conducted in [5], is to improve urban water demand forecasting in Telford, a small community in southwestern Pennsylvania, USA. By

incorporating probabilistic forecasts and considering the inherent uncertainty in the data, they used hybrid machine learning models based on conformal prediction, combining Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) methods. The features employed in this study include meteorological data (temperature, precipitation, humidity), historical water demand data, demographic data, economic data, days of the week, and seasons. The CNN method was used to extract these features from the water demand and mentioned time series data, which were then input into BiLSTM to capture complex temporal relationships and long-term dependencies in the data. The data for this study is divided into training, validation, and test sets. The final dataset consisted of 4104 hourly time series records of demand data collected between January 1, 2022, and June 21, 2022. This dataset is divided as follows: the first 2626 hourly records were allocated for training prediction models. Next, the subsequent 657 hourly demand records were used to validate the model, while the remaining 821 data points were used to test and evaluate the model's generalization performance. To compare the precision and reliability performance of the hybrid model, machine learning methods are used as benchmarks such as Quantile Regression (QR), Quantile Multilayer Perceptron (QMLP), and Quantile Gradient Boosting Regression Tree (QGBRT)... In summary, the CQRCTN (Convolutional-BiLSTM Quantile Regression model) showed superior performance in terms of both deterministic and probabilistic demand prediction accuracy compared to other comparison models, with a coefficient of determination (R2) of 0.94 for 1-hour ahead water demand forecasting, indicating very high precision.

The objective of the study described in [6]is to predict water demand in the Beijing–Tianjin–Hebei region of China using eleven statistical and machine learning methods; Statistical methods like linear regression (LR), ridge regression, lasso regression (least absolute shrinkage and selection operator), kernel ridge regression (KRR), and Bayesian ridge regression (BRR). For machine learning models: Single predictors comprise backpropagation neural network (BPNN), decision tree (DT), and support vector machine (SVM); Ensemble methods include random forest (RF), adaptive boosting (AdaBoost), and gradient boosting decision tree (GBDT). RF is a parallel integration algorithm, while AdaBoost and GBDT are serial integration algorithms. These models were implemented using the Python Scikit-Learn library. To do this, explanatory variables related to the economy, community, water usage, and resource availability were identified. The data used covers the period from 2004 to 2019 and was preprocessed using a normalization method to scale the features. Two prediction scenarios were considered:

1. **Interpolation Prediction Scenario (IPS):** For each model, 10-fold cross-validation was applied to a training set representing 80% of the data, randomly selected and covering the period from 2004 to 2019. The adjusted models were then tested on the remaining 20% of the data to verify their prediction performance.
2. **Extrapolation Prediction Scenario (EPS):** For each model, 10-fold cross-validation was applied to the training data covering the period from 2004 to 2018. The adjusted models were then tested on the 2019 data.

Additionally, water demand for the next two years (2020 and 2021) was predicted using a model similar to that used for the EPS scenario.
To evaluate the performance of each model, three metrics were used: mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2). The GBDT (gradient boosting decision tree) model demonstrated the best performance in both scenarios, achieving the lowest MSE and MAE values, as well as R^2 scores of 99.9999% and 99.9578% in the IPS and EPS scenarios, respectively[7].
The article describes a project for developing an intelligent system for predicting future hourly water consumption, called SWAP, based on historical data collected from 14 university campus buildings during the autumn semester of 2018. The goal is to design an effective model for predicting future water consumption using models such as probabilistic discriminative graphs and deep learning models, specifically conditional Gaussian random fields (GCRFs) and long short-term memory (LSTM) neural networks. The models are compared with linear regression and ARIMA models using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The data are divided into two parts: 75% for training and 25% for testing. For GCRF and LSTM models, 24-hour data sequences are used to predict the next 12 hours. This division allows for a direct comparison of the performance of the two types of models. Additionally, the dataset contains approximately 0.3% missing values, which are filled using linear regression. The results show that the GCRF model outperforms the LSTM-based deep learning model, offering improvements of 50% and 44% on average respectively.

- **Water level and flood forecasting:** The study aimed to investigate and compare different forecasting methods for accurately predicting water levels, particularly for early flood warnings. The study focused on three methods: SARIMA, RF, and LSTM.

  The experiment was conducted using data collected from Red River flow gauging stations, including Pembina, Drayton, and Grand Forks. The data included water-level time series collected at different frequencies from 2007 to 2017 for each sta-

tion. The researchers divided the collected data into training (70%), validation (15%), and testing sets (15%). The features comprised water-level measurement at various time intervals, including six hours, twelve hours, one day, three days, and one week.

The LSTM method consistently outperformed SARIMA and RF across all stations and prediction times, with significant reductions in RMSE values, for one-week-ahead predictions, LSTM achieved RMSE values of 0.190, 0.151, and 0.107 for Pembina, Drayton, and Grand Forks stations, respectively. These findings suggest that LSTM is a highly precise and reliable method for water-level forecasting, particularly for flood warning systems[8].
The objective of the analysis described in [9], is to study and apply machine learning models to forecast water levels 5 days ahead for the flood warning system of the Mekong River, focusing on the Thakhek station. The methods used are LASSO, Random Forests, and Support Vector Regression (SVR) to provide accurate forecasts of water levels, with the article detailing the principles and techniques of these models. Continuous data from 1994 to 2003 were used in this study, with data from 1994 to 2000 for initial training and data from 2001 to 2003 for testing. At each iteration, a sample of test data is added to the training data. The first method employed is LASSO, which helped identify the most significant variables for forecasting. The second method is Random Forest, which models the complex nonlinear relationships between the explanatory variables and water levels. Finally, the last method, SVR, was applied to capture the relationships between the input variables and water levels. Model performance was evaluated using criteria such as MAE and RMSE, as well as cross-validation (two 10-fold cross-validation repetitions) with values meeting the Mekong River Commission's requirements for flood forecasting. The best method meeting the requirements of the Mekong River Commission appears to be Support Vector Regression (SVR) with the lowest values for error measures in RMSE and an MAE of 0.486 m.

- **Impact of human mobility on short-term water demand forecasting:** The study [10], aims to study the influence of independent variables related to human mobility on short-term water demand forecasting in five DMAs (District Metered Areas) in Wroclaw, Poland. By employing features that include water consumption data, mobile phone data, spatiotemporal records, characteristics of the District Metered Areas (DMAs), population characteristics, and weekly cyclicity. This study uses classical and machine learning methods, including ARIMA (Auto-regressive Integrated Moving Average), Support Vector Machines (SVMs), tree-based ensemble methods (including random forests and extremely randomized trees), as well as the blind approach. These models were enriched with human mobility data cor-

related with water consumption data. The study uses precisely 51 days of water consumption data from January to March 2018, along with over 7 million geolocated mobility records in urban areas. These data are merged and divided into two sets: 14 days for training and testing, which are combined and shuffled during the 3-fold cross-validation process for model parameter tuning, and 7 days for validation (66.7% and 33.3%, respectively). In the results and experimental section of the article, several types of water demand predictions were made, including predictions for each week, each day, several months, etc., each type using different methods and data mentioned at the beginning of the article. These predictions allowed for a comparison of the performance of different models in terms of accuracy and robustness. The results show that mobility data can improve the accuracy of water demand forecasts. The models based on random forests showed the best performance with an accuracy of 90.4% for one-week forecasts.

- **Water demand forecasting using time series models:** The objective of the study presented in [11] is to propose a short-term forecasting model, termed Multivariate Long Short-Term Memory (M-LSTM), for predicting hourly water demand using weather data. The aim is to develop a robust model capable of accurately forecasting water demand and compare its performance with MLP and SARIMA models to evaluate its effectiveness in improving forecasting accuracy. This comparison will help identify whether the M-LSTM model offers superior predictive capabilities over MLP and SARIMA models.

  The study utilizes data from a small Water distribution system (WDS) located in northern Italy, serving approximately 5000 users. The dataset includes hourly time series data for water demand, temperature, humidity, solar radiation, and rainfall, spanning nearly 7 years from January 2013 to September 2019.

  The M-LSTM model demonstrated superior performance over MLP and SARIMA models. After tuning, the M-LSTM model's configuration with specific layers and neuron combinations yielded optimal results. The M-LSTM achieved a lower Mean absolute Percentage Error (MAPE) of 8.82%, indicating higher predictive accuracy compared to the other models. Similarly, the M-LSTM model exhibited a high R^2 of 0.905, signifying better predictive capability. These results underscored the effectiveness of the M-LSTM model in accurately forecasting water demand .

References

[1] Perea, R. G., Poyato, E. C., Montesinos, P., & Díaz, J. A. R. (2019). Optimization of water demand forecasting by artificial intelligence with short data sets. Biosystems engineering, 177, 59-66.

[2] Piasecki, A., Jurasz, J., & Kaźmierczak, B. (2018). Forecasting daily water consumption: a case study in Torun, Poland. Periodica Polytechnica Civil Engineering, 62(3), 818-824.

[3] https://www.wodociagi.torun.com.pl/

[4] Zhou, X., Zhang, S., Xie, X., Yang, M., Bi, Y., & Li, L. (2014, October). Application of BP Neutral Networks to Water Demand Prediction of Shenyang City Based on Principle Component Analysis. In 2014 7th International Conference on Intelligent Computation Technology and Automation (pp. 912-915). IEEE.

[5] Iwakin, O., Moazeni, F. (2024). Improving urban water demand forecast using conformal prediction-based hybrid machine learning models. Journal of Water Process Engineering, 58, 104721.

[6] Bejarano, G., Kulkarni, A., Raushan, R., Seetharam, A., Ramesh, A. (2019, November). Swap: Probabilistic graphical and deep learning models for water consumption prediction. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (pp. 233-242).

[7] Shuang, Q., & Zhao, R. T. (2021). Water demand prediction using machine learning methods: A case study of the Beijing–Tianjin–Hebei region in China. Water, 13(3), 310.

[8] Atashi, V., Gorji, H. T., Shahabi, S. M., Kardan, R., & Lim, Y. H. (2022). Water level forecasting using deep learning time-series analysis: A case study of red river of the north. Water, 14(12), 1971.

[9] Nguyen, T. T., Huu, Q. N., Li, M. J. (2015, October). Forecasting time series water levels on Mekong river using machine learning models. In 2015 Seventh Interna-

tional Conference on Knowledge and Systems Engineering (KSE) (pp. 292-297). IEEE.

[10] Smolak, K., Kasieczka, B., Fialkiewicz, W., Rohm, W., Siła-Nowicka, K., Kopańczyk, K. (2020). Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models. Urban Water Journal, 17(1), 32-42.

[11] Zanfei, A., Brentan, B. M., Menapace, A., & Righetti, M. (2022). A short-term water demand forecasting model using multivariate long short-term memory with meteorological data. Journal of Hydroinformatics, 24(5), 1053-1065.

Chapter III

Methods

Our study aims to forecast the water use in Algeria using both machine learning and deep learning algorithms. This analysis is based on a rigorous and comprehensive process, as depicted in the diagram below (figure III.1). Each step is carefully designed to ensure the relevance and accuracy of our predictions.

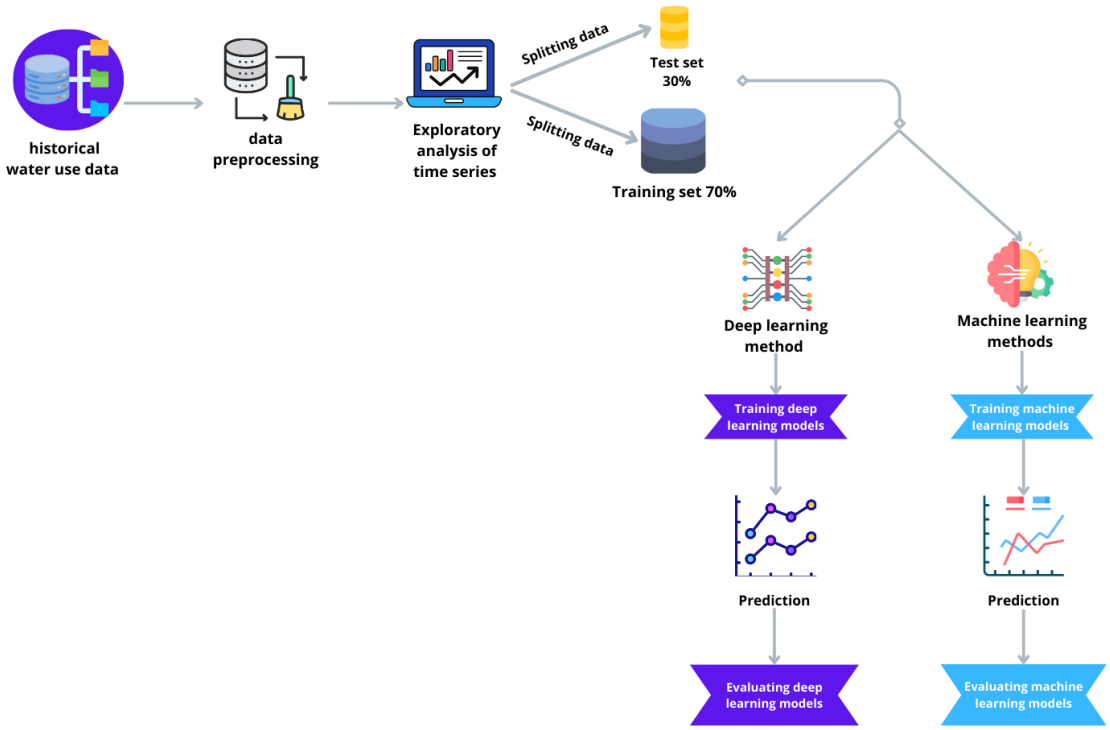


Figure III.1: Diagram of the water consumption prediction process.

Our framework includes six steps :

- **Data Collection:**
  In our study, data collection is carried out in collaboration with ADE (Algérienne des Eaux). This organization is responsible for water resource management in Al-

geria. The historical data we use comes from the Tizi-Ouzou Province, containing detailed information on water use for each subscriber in 21 regions of the province. The provided dataset includes observations covering a period ranging from March 31, 2003, to December 31, 2023. The data description is available in chapter IV (Data)

- **Data Preprocessing:**
  This process involves cleaning and preparing the data for analysis. In the cleaning phase, we filtered the data to keep only the consistent subset. Thus, the missing data are ignored in our study. Once the data are cleaned, they are divided into monthly and quarterly data. The monthly data was converted into quarterly data and combined with the latter.
  Next, we added lag features to our dataset ranging from 1 to 10 to train the various machine learning and deep learning models. This means that for each observation, we included the values from the previous 10 observations as additional features.

- **Exploratory Data Analysis (EDA):**
  Using the most popular visualization library, Matplotlib, we successfully plotted the time series graphs and decomposed them into three components: trend, seasonality, and noise. The charts representing these three components are included in the chapter IV.

- **Data splitting:**
  The historical data covers a period from March 31, 2003, to December 31, 2023. To perform the split, we applied the most popular method of dividing into training and test sets to achieve a good balance between providing enough data for the model to learn and reserving a significant amount of data for reliable evaluation. The training set comprises 70% of the total data. While the test set consists of the remaining 30% of the data.

- **Model training:**
  We trained various machine learning and deep learning algorithms using the training data. For each algorithm, we adjusted their parameters to optimize the prediction performance of the resulting model.

- **Prediction:**
  After training the models, We use them to make predictions on the test data. Then, we plotted graphs to compare the model's predictions with the actual data for each method used.

- **Model evaluation:**
  We evaluated the performance of the various resulting models, by calculating met-

rics such as MSE and R^2 for each data configuration. These metrics estimate the difference between the predicted and actual values. Finally, we compare the different results to identify the most effective combination: data configuration and model. The results of the evaluation step are given in chapter V (Experiments and Results).

III.1 Machine Learning methods

III.1.1 Linear Regression:

Linear regression is a statistical technique used in data analysis and supervised machine learning to create a model that describes the relationship between a dependent variable and one or more independent variables. It mathematically models the relationship between these variables in the form of an equation.
Linear regression models are relatively simple and provide an easily interpretable mathematical formula for generating predictions.

The fundamental operation of linear regression involves plotting a linear graph between two data variables, x and y. The independent (explanatory or predictive) variable x is plotted along the horizontal axis, and the dependent (predicted) variable y is plotted along the vertical axis. As shown in the figure III.2

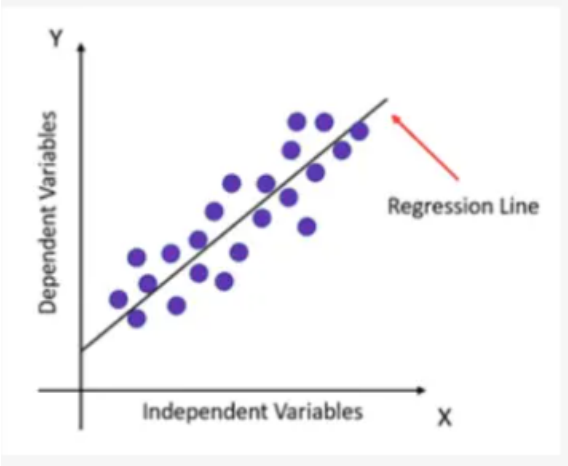


Figure III.2: Example of linear regression.

There are several types of linear regression, among them: simple linear regression and multiple Linear Regression

1. **Simple linear regression:**
   The objective of simple linear regression is to predict the value of a dependent variable based on an independent variable. The stronger their linear relationship, the more accurate the prediction, as shown in the following figure:

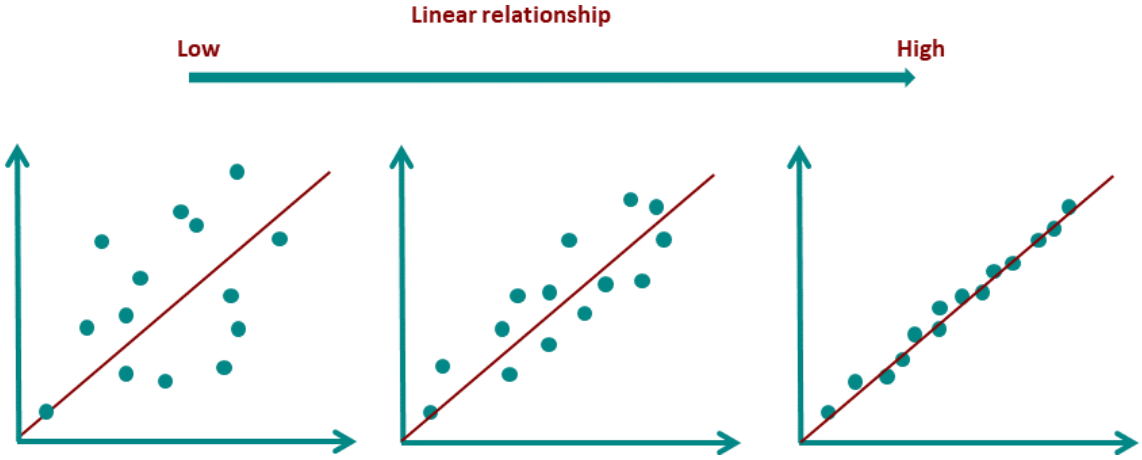


Figure III.3: Simple linear regression [1].

To determine the straight line that is drawn on these scatter plots, linear regression uses the method of least squares, and it can be described by the equationIII.1:

$$Y = \beta_0 \cdot X + \beta_1 + \epsilon \tag{III.1}$$

Where β_0 and β_1 are two unknown constants representing the regression slope, while ϵ is the error term. It is used to model the relationship between two variables.

2. **Multiple Linear Regression:**
   Compared to simple linear regression, multiple linear regression stands out taking into account more than two independent variables intending to estimate one variable based on several other variables [1]. In this case, the function of the linear regression line changes to include more factors as presented in the equation III.2 [2]:

$$Y = \beta_0 \cdot X_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n + \epsilon \tag{III.2}$$

   Each time the number of predictor variables increases, the β constants also increase accordingly[3]. This method is commonly used in empirical social research as well as in market studies. In these two fields, it is interesting to determine the influence of different factors on a variable [1].

III.1.2 Support Vector Regression algorithm(SVR):

The objective of SVR is to find an approximate function that defines the relationship between the input domain and the real values, based on a training sample. As represented

in the graph III.4, the idea is to see these two red lines as the decision boundary, and the green line as the hyperplane. Our goal is to consider the points located within this decision boundary. The optimal hyperplane is the one that maximizes the number of points that best adhere to it.

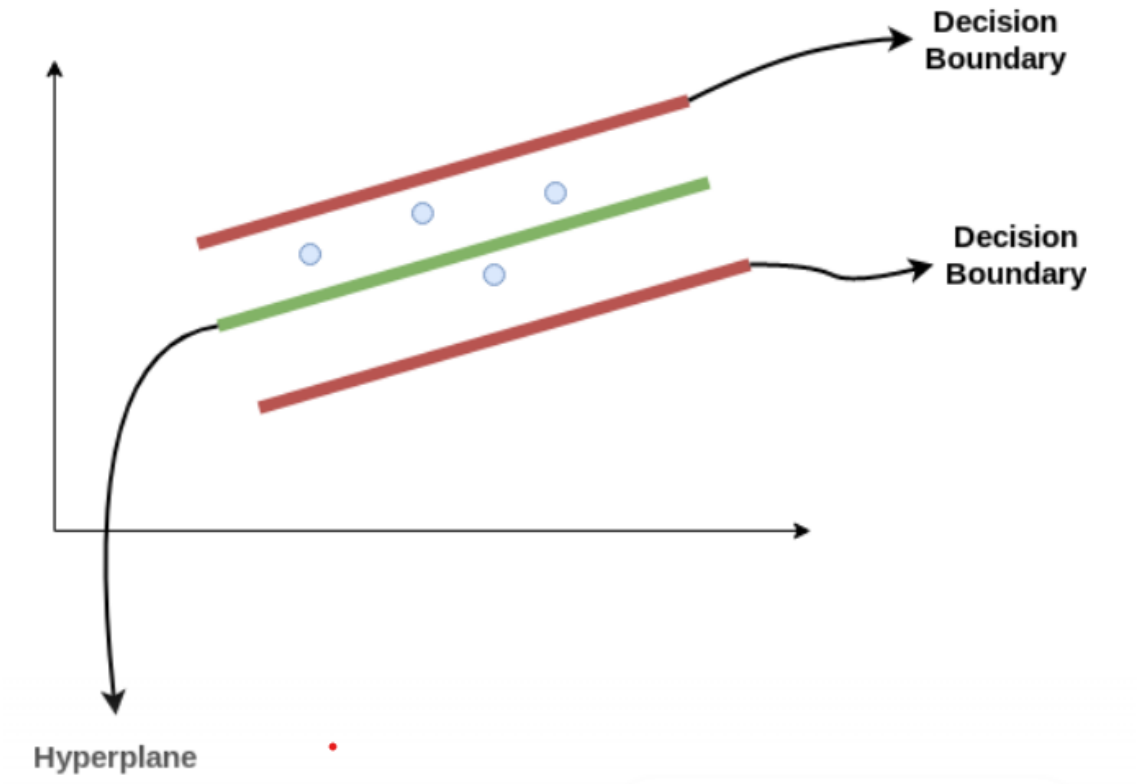


Figure III.4: Hyper plane of SVR.

As we can see in figure III.5, the decision boundary represents a certain distance called "ϵ" from the hyperplane, whether on the positive side ($+\epsilon$) or the negative side ($-\epsilon$). Therefore, ϵ represents the margin of error or the margin of tolerance around the hyperplane [4].
To better understand, here is a second diagram that explains this more precisely.

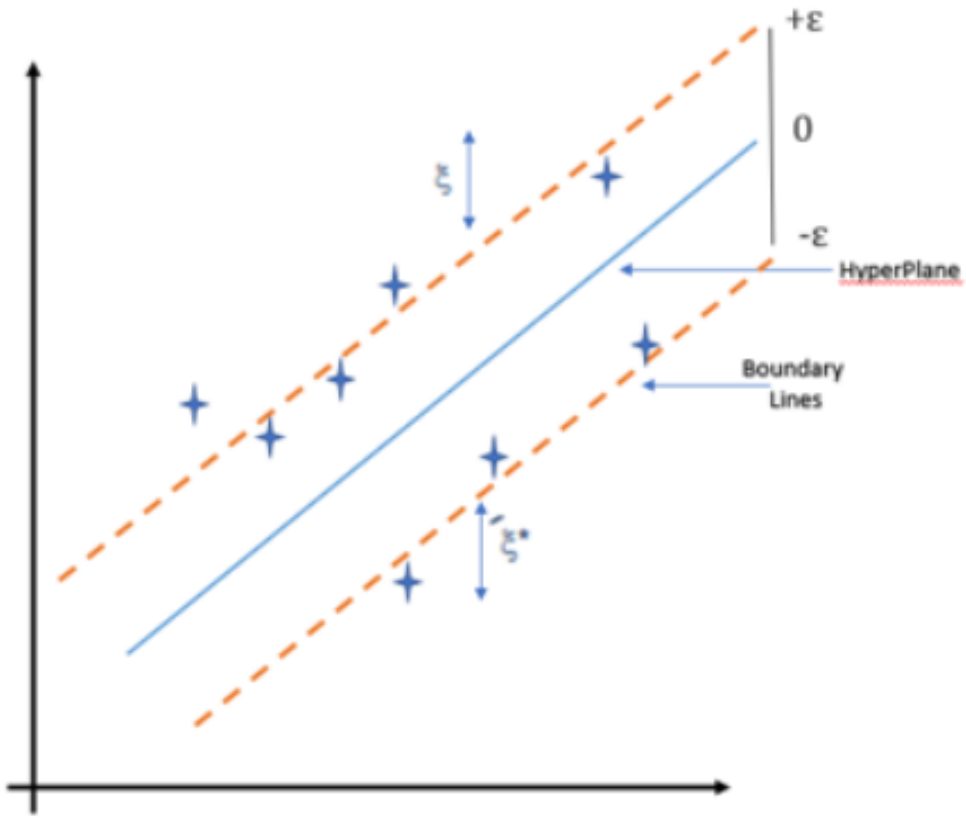


Figure III.5: Support Vector Regression (SVR) with Margin and Boundary Lines [5].

Support vector regression has some advantages such as it is resistant to outliers, decision models can be readily updated, and finally, it is easy to implement. However, for the disadvantages: large datasets are inappropriate for them, the decision model does not perform well when there is a lot of noise in the dataset, and when the number of features per data point is greater than the number of training data samples, its performance becomes insufficient [6].

III.1.3 The regression trees algorithm:

Regression trees are used for explaining and predicting the values taken by a quantitative dependent variable, based on both quantitative and qualitative explanatory variables [7].

Figure III.6 shows the structure of a decision tree.

The construction of a decision tree stars, with the complete dataset at the root node. To divide the data into subsets, the algorithm selects a feature and a threshold. To achieve this, it evaluates all possible splits and selects the one that reduces the variance of the target variable. This is accomplished using criteria such as Mean Squared Error (MSE) or

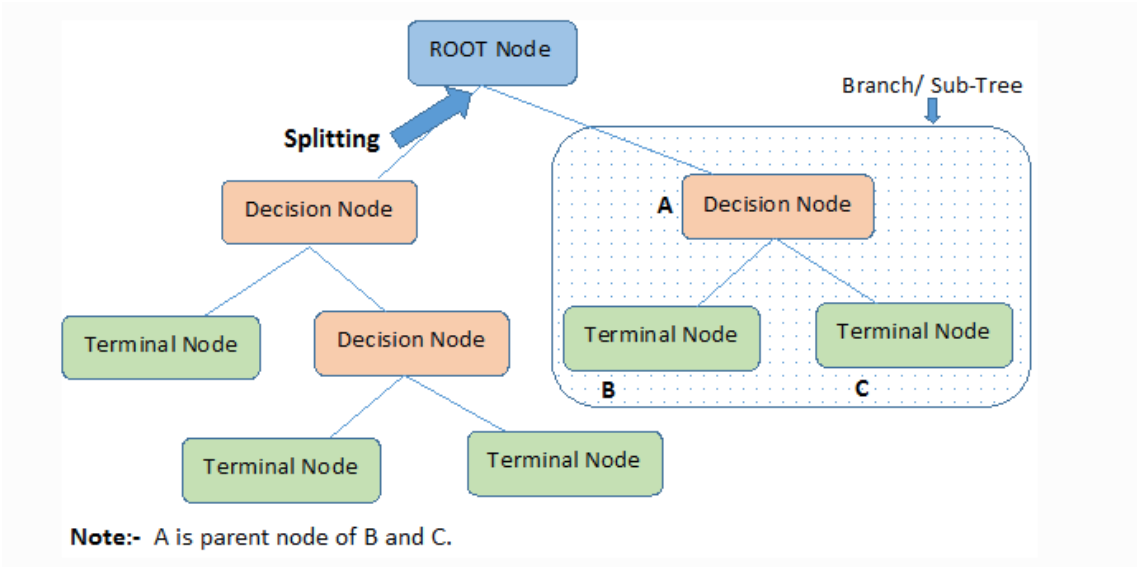


Figure III.6: Structure of a decision Tree.

Mean Absolute Error (MAE). The selected feature and threshold establish the condition of that node.

After the first split, the dataset is split into two subsets based on the selected feature and threshold. The same splitting process is then applied individually to each of these subsets. This recursive process continues until a stopping condition is satisfied, which could be a maximum depth limit, a minimum number of samples in a node, or other criteria.

Figure shows III.6 the key concepts related to decision trees:

- **Root node:** It illustrates the entire population or sample, which is subsequently split into two or more homogeneous sets.
- **Decision node :** When a sub-node divides into further sub-nodes, then it is referred to as a decision node.
- **Leaf :** Nodes that do not split are referred to as leaf or terminal nodes.
- **Branch/Sub-Tree :** A subsection of the entire tree is referred to as a branch or sub-tree.
- **Parent and child node:** A node, which is split into sub-nodes, is known as the parent node of sub-nodes, whereas sub-nodes refer to the child of a parent node [8].

Decision tree regression can capture both linear and non-linear relationships in the data, is simple to understand, and can manage both numerical and categorical data. As for

its challenges, it can be influenced by small variations in the data, and trees can become intricate and deep, making them challenging to interpret in some cases [9].

III.1.4 K-Nearest Neighbours:

K-Nearest Neighbours is a supervised machine learning technique used to solve regression and classification issues, KNN is a flexible and popular machine learning method that is mostly employed due to its simplicity and implementation ease. It doesn't call for any presumptions on the distribution of the underlying data.

The operating principle of the KNN can be summarized by writing it in pseudo-code as follows:
Start algorithm:

- Prepare the data.
- Compute the distance: The Euclidean distance is mainly used to calculate how similar the target and training data points are to each other.
- The target point and every data point in the dataset are separated by a computed distance.
- Identifying the Nearest Neighbours: The closest neighbors are the k data points that have the smallest distances to the target point.

End algorithm

The majority vote is used to determine the class labels in the classification problem. The predicted class for the target data point is the one with the highest frequency among the neighbors. In the regression problem the target data point's expected output is determined by the computed average value [10].

KNN provides many key advantages, due to its simplicity and ease of comprehension, the KNN algorithm is an effective choice for beginners in machine learning. Additionally, There is no training step necessary for the algorithm, and it can manage big datasets without experiencing the dimensionality curse, which is a typical issue with other machine learning techniques, likewise, it is known for being accurate as well as efficient, in addition KNN algorithm has also some disadvantages The computational cost of the KNN technique can be high, especially for huge datasets. This is due to the algorithm's need to calculate the distance between each test and training data point, which can be time-consuming. Additionally, The performance of the KNN method can be greatly impacted by data outliers because it can be sensitive to them. Outliers are data points that are significantly different from the rest of the data [11].

III.1.5 Random forest algorithm:

Random forest is a machine learning algorithm that constructs an ensemble of multiple decision trees to achieve a singular, more accurate prediction or result [12]. It can be used for both classification and regression tasks. In prediction, the algorithm combines the result of all the trees, either by voting(for classification tasks) or by averaging (for regression tasks). For making predictions Random Forest uses bagging also known as bootstrap aggregation, which is an ensemble learning, where multiple weak models are developed on different subsets of the training data.

A basic structure of a Random forest is shown in figure III.7:

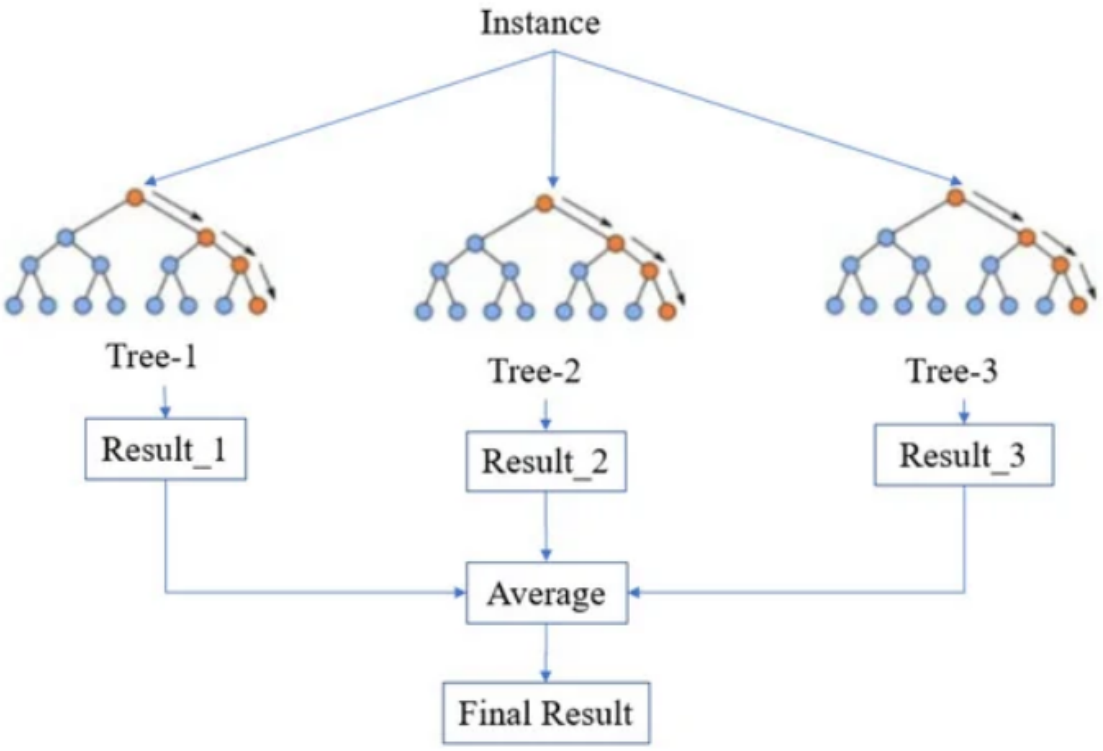


Figure III.7: [13]Structure of a Random forest Algorithm

The random forest has some key features like Resistive to overfitting and can handle large datasets and missing values [14]. The main limitation of random forest is that a large number of trees may render the algorithm too slow and ineffective for real-time predictions. In general, these algorithms are quick to learn but take longer to make predictions once they're trained. A more accurate prediction necessitates more trees, leading to a slower model [12].

III.1.6 XGBoost (Extreme Gradient Boosting) Algorithm:

XGBoost stands for "Extreme Gradient Boosting" and has emerged as one of the most popular and widely adopted machine learning algorithms, due to its capability to achieve

state-of-the-art performance in many machine learning tasks such as classification and regression [15]. It is also an ensemble learning method that merges the predictions of multiple weak models to create a stronger prediction. A simplified structure of xgboost is shown in the figure III.8:

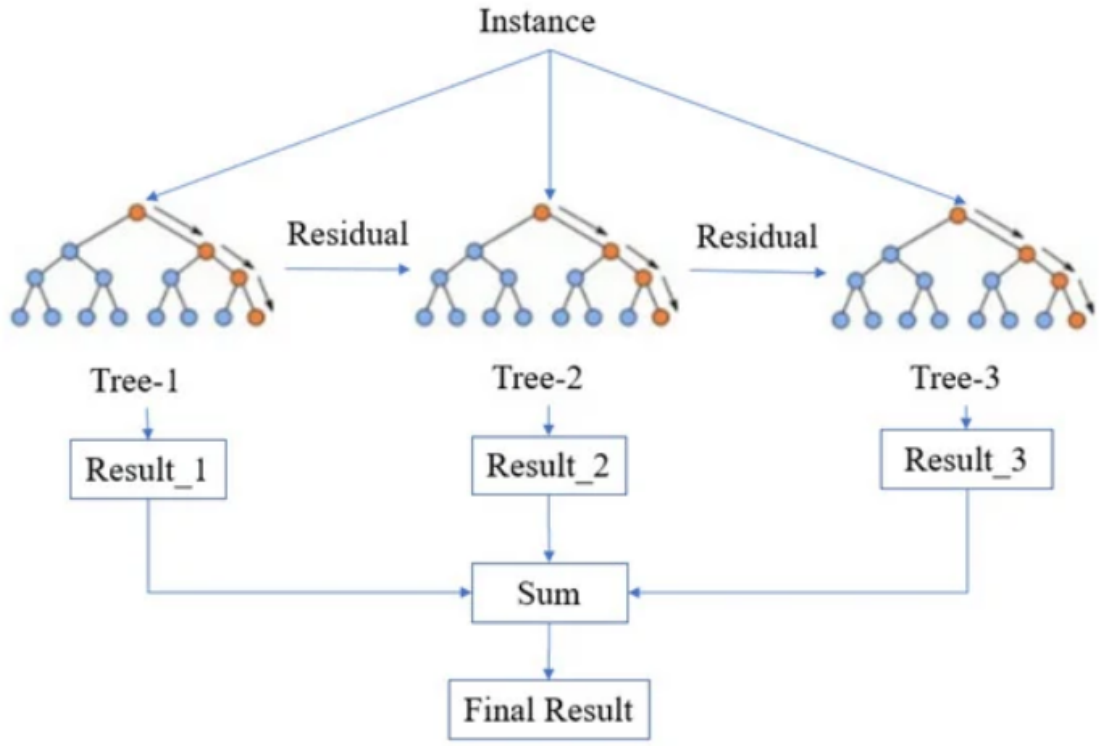


Figure III.8: [16] Simplified structure of xgboost.

The operational principle of Xgboost can be summarized by the following steps:
Start algorithm:

- Create a predictive model S1 using the full training data.
- Calculate the difference between the original target variable “y” and the predicted one”y1”.
- Construct another predictive model S2, to correct mistakes made by S1.
- Repeat step 2.
- Create another predictive model S3, which corrects the mistakes of S2.
- Continue repeating these steps until the error becomes zero, and the model becomes accurate [17].

End algorithm

Xgboost can produce high-quality results, and handle missing values and it is suitable for large datasets. However, xgboost can be susceptible to overfitting, particularly when trained on small datasets or when there are too many trees, and may consume a lot of memory when dealing with large datasets [15].

III.2 Deep learning methods

Deep learning is one of the main approaches used in machine learning[18]. It is a category of artificial intelligence that operates artificial neural networks (ANN)[19]. ANN are inspired by the biological neural networks of human brains[20]. They are based on a set of connected units called neurons or nodes, which are organized into interconnected layers. These nodes are the most important components of this method; they are calculation units that enable the neural network to function.[21] All the neurons in a single layer are grouped and perform a similar type of function[22].

Layers can contain millions of neurons, depending on the complexity of the neural network. The greater the number of layers, the deeper and more powerful the network[18]. ANN typically consists of three main layers: an input layer, a hidden layer, and an output layer. The input data passes through the input layer to the hidden layers and finally produces an output on the output layer[23]. Figure III.9 illustrates the architecture of the ANN:

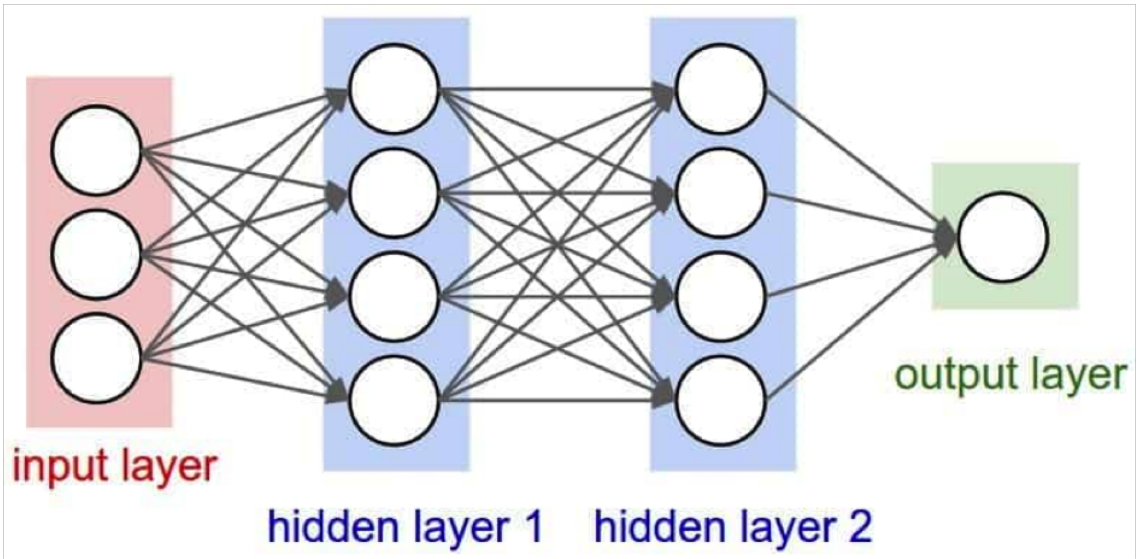


Figure III.9: Structure of artificial neural networks.

- Input Layer: The first layer receives raw information inputs, a bit like the human

optic nerves when processing visual signals. Neurons process this information and transmit it to neurons in the next layer [22]. The number of neurons in the input layer is determined by the number of input features[24].

- Hidden layer: Hidden neurons are located between the input layer and the output layer. They take output data from previous input neurons, calculate new output data, and pass it on to successive layers[22]. The number of hidden layers and neurons in each layer can vary depending on the complexity of the problem[24].
- Output Layer: The last layer produces the results of the system, which can be predictive or probabilistic values[22].

III.2.1 Weights and Biases

Information processing in neural networks often follows the same sequence; each neuron is assigned a unique weight that enables the definition of what information can enter the system[25]. Weights are essential components that are crucial to the network's capacity for learning and prediction. They resemble the synapses in biological neural networks, enabling connections between neurons across different layers[26].
A weight is applied to the input of each neuron. Neural networks update these weights continuously. There is thus a feedback loop implemented in most neural networks[22].

These numerically based parameters, are iteratively optimized to minimize the discrepancy between the network prediction and the actual values throughout the training phase of the neural network[27]. As the input signal passes through the various layers, it undergoes multiplication by these weights, collectively influencing the network's final output[26].

Biases are the self-learning values of our neural network, like the weights[22], they are defined as a constant added to the product of features and weights at the output layer for calculation of the results. This parameter helps the model shift the activation function to the positive or negative side[28].

As shown in the figure III.10, the neuron initially calculates the weighted sum of its inputs and subsequently incorporates a bias into the computation.

$$y = \sum(\text{weights} \times \text{inputs}) + \text{bias}$$

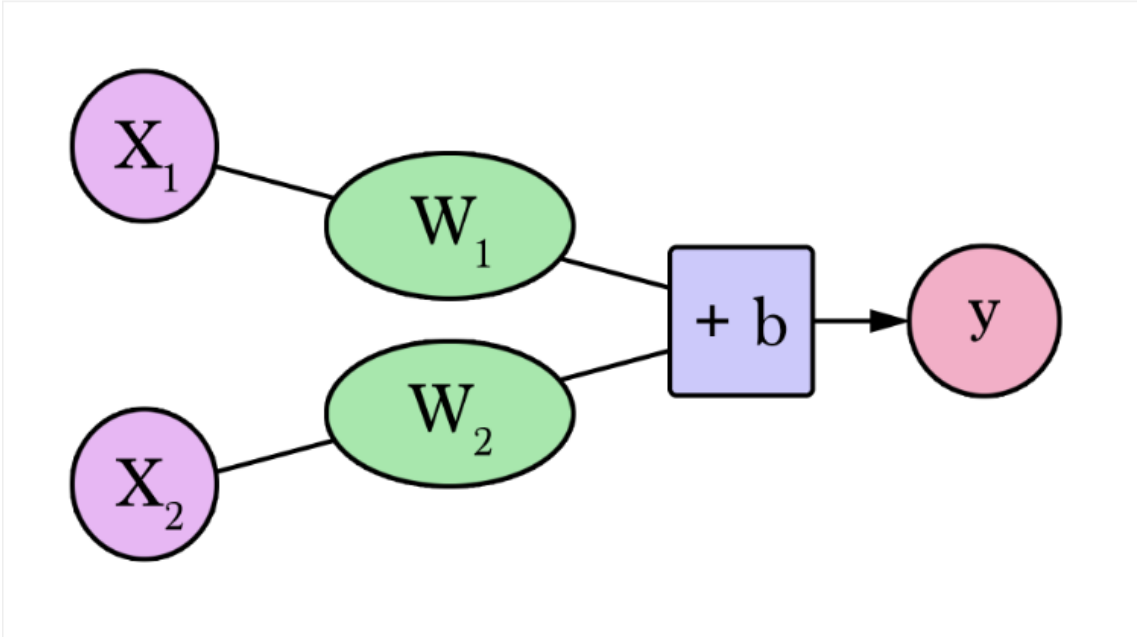


Figure III.10: The processing done by a neuron [29].

III.2.2 Activation function

In the biological context, when comparing the activation function to neurons in the human brain, it plays a role in deciding what information to send to the next neuron, in the artificial neural network, this is precisely its function; it takes the output signal from the previous layer and converts it into a format that can be inputted into the next layer[30] or the previous one depending on the network's complexity. The activation function determines whether the neuron should be activated by computing the weighted sum and adding bias to it[31]. As shown in Figure III.11 the neuron first computes the weighted sum of the inputs, then a bias will be added to the operation. Next the calculated value is included to the activation function.

The main feature of the activation function is its ability to introduce non-linearity into a neural network[30] helping to identify underlying patterns[19], without an activation function a neural network is essentially just a linear regression model.

Neural networks have different types of activation functions. Each of these functions has specific properties and corresponds to certain uses [31]:

1. **Binary step function:** binary step function relies on the threshold values of whether a neuron should be activated or not, the input passed to the activation function is evaluated to a certain threshold, if the input exceeds it the neuron activates otherwise it deactivates, meaning that its output is not transmitted to the next hidden layer[32]. mathematically it can be represented by the equation III.3:

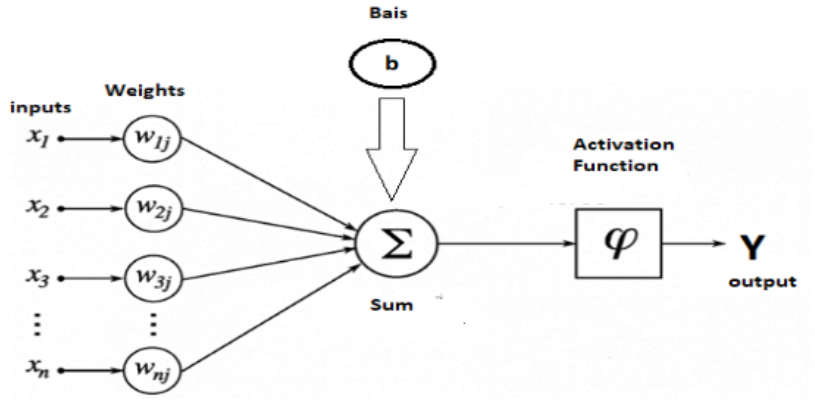


Figure III.11: Activation function [22].

$$x = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases} \tag{III.3}$$

It's recommended better use a binary step-like activation function in the output layer for the binary classification problems in neural networks.

2. **Linear Activation Function:** It is a simple function and it is used when predicting numerical values, which is common in regression problems. However, the linear activation function is rarely used in hidden layers of neural networks because it does not introduce any non-linearity [33].It is defined by the equation III.4:

$$f(x) = x \tag{III.4}$$

Where :

$$x = \sum(\text{input} \times \text{weight}) + \text{bias} \tag{III.5}$$

3. **Non-Linear Activation Functions:** Non-linear activation functions play an essential role in ANN by introducing non-linearity into the network and allowing the network to learn from the pattern [34]. Among these nonlinear functions, there are several commonly used options:

- **Sigmoid function:** Also known as a logistic function, it's a function that's very familiar from the early days of neural networks. It is represented by $\sigma(x)$, which has an "S"- shaped curve approaching 0 if the values are negative and 1 if the values become positive. The interpretation of the outputs can be as probabilities, which is good for binary classification problems[32]. The sigmoid function is defined by the equation III.6:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{III.6}$$

The figure III.12 shows the graph of the sigmoid activation function:

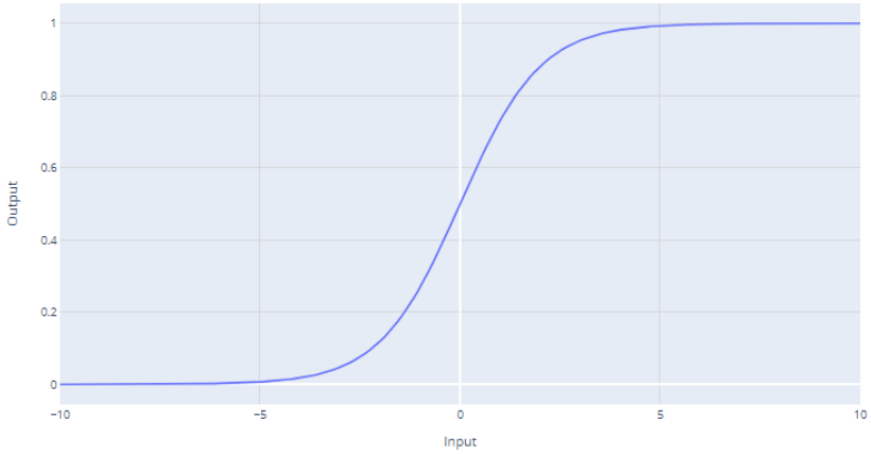


Figure III.12: Sigmoid activation function [31].

- **ReLU (Rectified Linear Unit) function:** One of the best-known functions used in deep learning today is the ReLu function[35], which sets the output to zero for negative input, and the input itself for positive values[32]. The ReLU function is defined by the equation III.7:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases} \tag{III.7}$$

Figure III.13 shows the graph of the Relu activation function:

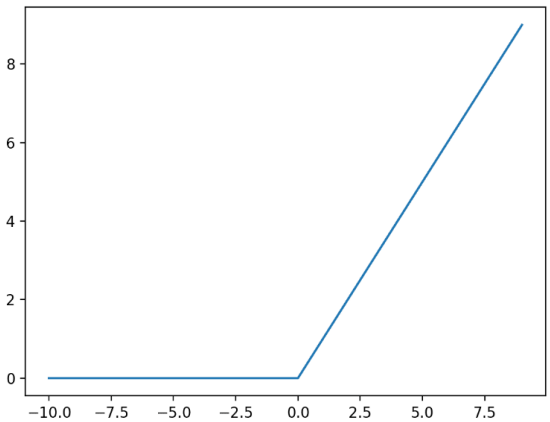


Figure III.13: [36]plot of inputs vs outputs for the ReLu activation function

This means that if the input value (x) is negative, the value 0 is returned, otherwise the value is returned.

- **Leaky ReLU function:** Leaky ReLU function is an enhanced version of the ReLU function. Instead of setting the output to zero for negative values, it defines a slight linear component of x.it can be defined as: III.8:

$$f(x) = \max(0.01x, x) \tag{III.8}$$

Figure III.14 displays the graph depicting the Leaky Leaky ReLU function

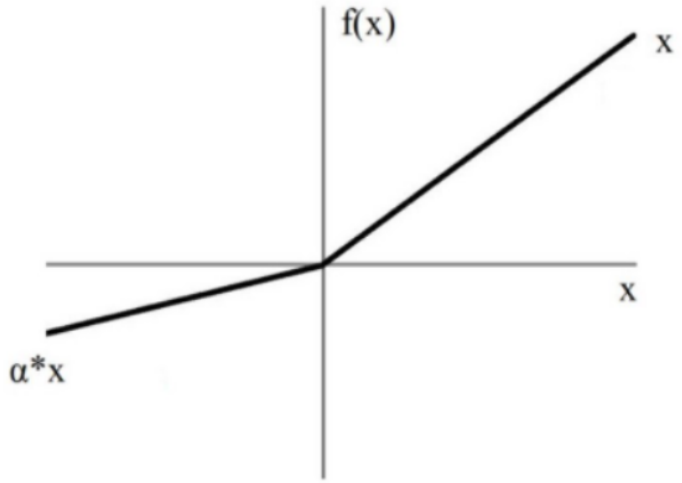


Figure III.14: Plot of inputs vs outputs for the Leaky ReLu activation function [34].

The difference is to exchange the horizontal line for values below zero with a non-horizontal straight line. The slope of this straight line is represented by the parameter α, which is multiplied by the input x [35].

- **Tanh (hyperbolic tangent) function:** The tanh function can be employed as a non-linear activation function in the hidden layers, producing values between-1 and 1. The tanh function is defined by the equation III.9:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{III.9}$$

Figure III.15 shows the S-shape of the Tanh activation function.

The hyperbolic tangent activation function looks just like the sigmoid activation function, but there's a difference in the output: instead of the derivative of the function approaching zero, it's centered on zero [34]. The larger (positive) the input, the closer the output value will be to 1.0, while the smaller (negative) the input, the closer the output will be to -1.0 [36].

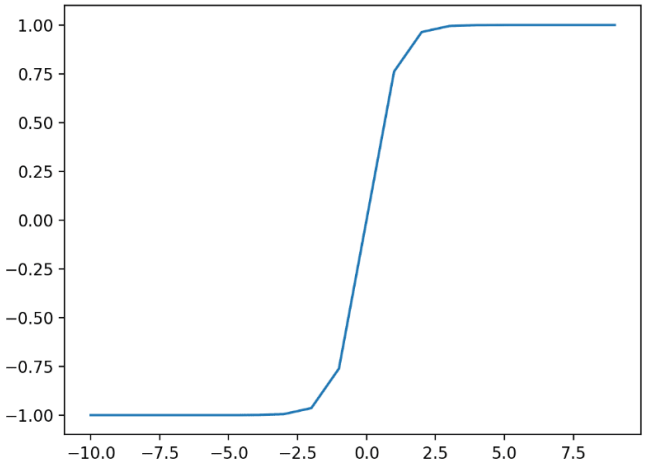


Figure III.15: [36]plot of inputs vs outputs for the Tanh activation function

- **Softmax function:** The Softmax function is applied to the prediction of probability values on the output layers of the neural network to solve classification problems[34].
  This function computes the probability distribution for " n " different events. In general, this function calculates the probabilities of each target class among all possible target classes. The calculated probabilities will help determine the target class for the given inputs [37]. The equation III.10 defines the softmax function:

$$f(x_i) = \frac{e^{x_i}}{\sum e^{x_j}} \tag{III.10}$$

III.2.3 Types of neural networks

There are numerous types of neural networks accessible. They can be categorized based on their: Structure, Data flow, Neurons used and their density, Layers and their depth, and activation filters. Here are some of the most common types:

1. **Perceptron:** It is one of the most basic and oldest models of a Neuron [38], also known as the monolayer neural network [39]. It is the smallest unit of a neural network, containing just two layers: an input layer and an output layer [38], as shown figure III.16:

   The perceptron takes the input values and calculates their weighted sum for each input node. This weighted sum is then sent through an activation function to produce the output [39]. The perceptrons categorize data into two classes [38] for classification in linear or binary models [39]. However, it's important to note that perceptrons can only categorize sets of vectors that can be linearly separated [40].

2. **Feed-forward neural network (FFNN):** It is the simplest form of neural network

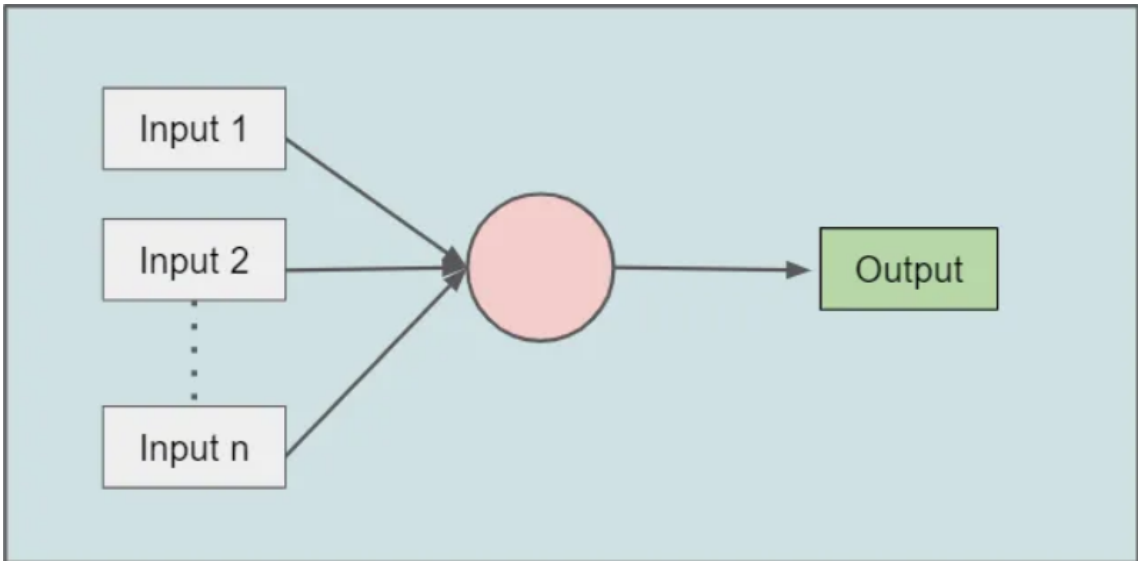


Figure III.16: Single neuron neural network [39].

where input data moves in one direction only, from the input nodes, through the hidden nodes (if any), and finally, to the output nodes[38], There are no loops or cycles in this network.

A feedforward neural network operates in two distinct stages: the feedforward phase and the backpropagation phase: for the first phase (feedforward), the process followed is simple, and is completed when the prediction is achieved in the output layer. Once the prediction has been achieved, we move on to the backpropagation phase, which consists of calculating the error between the actual and predicted values, which is then transmitted to the network by adjusting the weights to reduce this error. This process is carried out using the gradient descent optimizer[40]. This type of neuron network is used in a wide range of tasks like pattern recognition, classification, regression analysis, Image recognition, etc [41].

3. **Multilayer Perceptron (MLP):**

   The Multilayer Perceptron is a type of feedforward neural network because inputs are weighted and processed with an activation function, similar to the perceptron. However, unlike the Perceptron, the Multilayer Perceptron passes each weighted sum to the next layers. Each layer transmits its computed result to the next layer. This extends throughout the hidden layers to the output layers [42].

4. **Recurrent Neural Network (RNN):** Recurrent Neural Network is designed to store the output of a layer. RNN is looped back to the input to help predict the output of the layer [38]. The hidden state of RNN is crucial because it preserves certain information about a sequence. This state is also called the memory state because it retains information from the previous input to the network. Figure III.17

demonstrates the architecture of RNN.

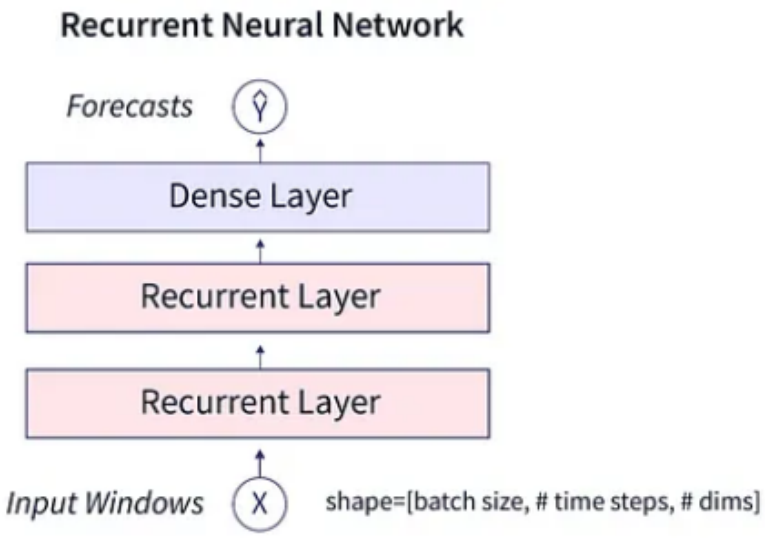


Figure III.17: The architecture of Recurrent Neural Network RNN [45].

When the RNN network produces a forecast that doesn't match the true label, it computes the loss, which measures the discrepancy between the prediction and the actual target. This loss is then transmitted through the network, This process is called backpropagation. During the backpropagation, the network determines how each weight and bias influenced the error, and it modifies these parameters to minimize the loss. The aim is to adjust the weights and biases associated with the hidden layers and input to minimize the loss function [43].

The advantages of RNNs include the capability to process inputs of any length, retain the memory of previous inputs, and the model's size remains constant without regard to the length of the input sequence.

5. **Long Short-Term Memory (LSTM)**: In a traditional RNN, a single hidden state passes through time. This can make it challenging for the network to acquire long-term dependencies. To address this problem, LSTMs introduce a memory cell, a container that can store information for an extended period. Learning long-term dependencies in sequential data is possible for LSTM networks, This makes them suitable for tasks like language translation, speech recognition, and time series forecasting.

The architecture of LSTM in shown in Figure III.18:

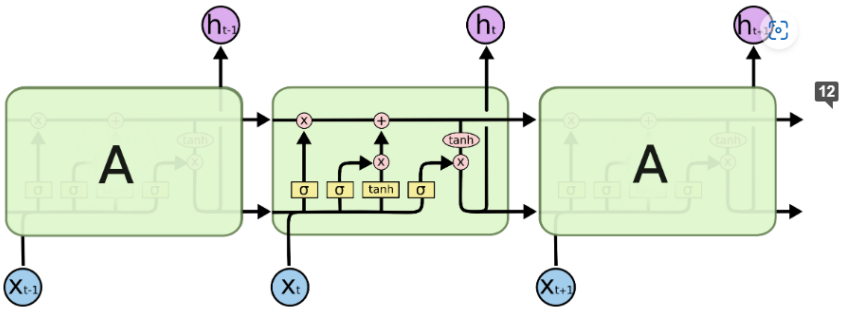


Figure III.18: LSTM architecture [48].

Controlling the memory cell involves three gates: the input gate, the forget gate, and the output gate. The gates are responsible for dealing with which information to add, remove, and output from the memory cell [46].

- **input gate:** The input gate is responsible for determining which information will be stored in long-term memory. The current input and short-term memory from the previous step are the only sources of information it uses. This gate removes information from variables that are not useful.
- **Forget gate:** The forgotten gate chooses whether to keep or discard old information.
- **Output gate:** The output gate will generate new short-term memory by combining the current input, previous short-term memory, and newly computed long-term memory, and passing it on to the cell in the next step. The output of the current time step can also be derived from this hidden state.
- **Update date:** The update gate determines the amount of previous information that needs to be transmitted to the next state.
- **Reset gate:** The model's reset gate decides how much information must be forgotten. The reset gate is responsible for storing relevant information from the past time step into new memory content [47].

6. **Convolutional neural network (CNN):**

Convolutional neural networks (CNNs), also known as ConvNet, are an enhanced version of feed-forward Artificial Neural Networks (ANNs), frequently used for computer vision tasks, sorting visuals for object recognition, and image classification. They use convolution and clustering of neural networks to extract features from videos or images (lines, curves, etc.), and then exploit these features to classify or detect objects or the environment [48].
Convolutional neural networks are built up of several layers, such as the convolution layer,

the pooling layer, and the fully connected layer, for each layer, the CNN amplifies its complexity [49], thus increasing its ability to identify and differentiate various aspects or areas of an image[50]. The figure III.19 demonstrates the CNN architecture

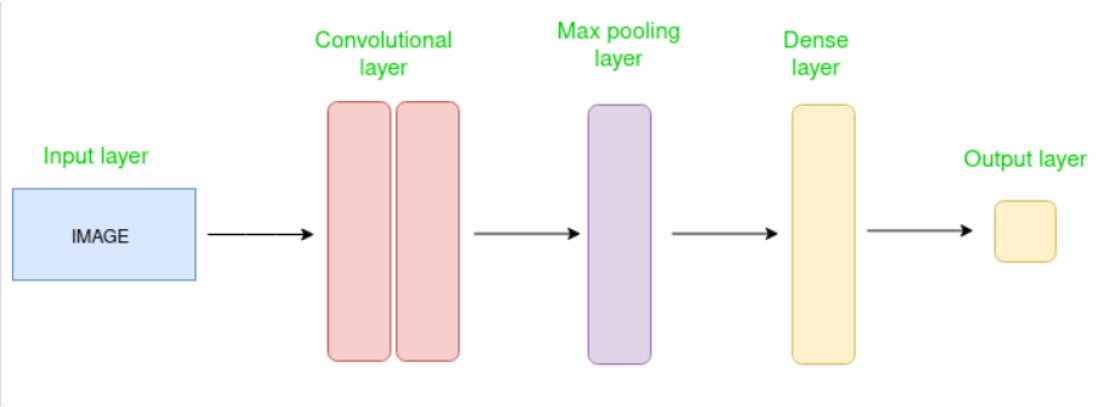


Figure III.19: CNN architecture [49].

Here are the different layers of the CNN with explanations of each one :

- **The convolutional layer:** The convolutional layer is the basis of CNN construction, where the majority of the calculations take place, not needing many components; just the input data, a feature map, and a filter.Convolutional layers enable the network to understand and extract valuable patterns.

- **The pooling layer:** Pooling layers, also known as subsampling, are used in the neural network to minimize the dimensions of the input data. Unlike convolutional layers, pooling layers simply move a filter over the entire input without using weights. This filter consists of applying a function (a maximum or minimum function) to the numerical values of the receiver field to generate the output.
  Although some information may be lost in the pooling operation, it offers many advantages to CNNs. It helps reduce the risk of overfitting by lowering the complexity of the model while improving its efficiency by reducing the dimensionality of the data. The objective of the pooling layer is to draw out the significant features from the convoluted matrix.

- **Fully connected layers (FC):** Fully connected (FC) layers As the name suggests, each node in an FC layer is connected to all the features in the preceding layers through their filters. Unlike pooling and convolution layers, which often use ReLU activation functions, FC layers generally use a softmax activation function, which transforms the outputs into probabilities of 0 or 1. These are responsible for learning and identifying various features of a given data [51].

The advantages and disadvantages of a convolutional neural network are:

- Efficient detection of features and patterns.
- They can process huge amounts of data while maintaining high accuracy.
- Training requires high computational intensity and memory capacity.
- Need large quantities of labeled data[48].

III.3 Metrics

When we develop an artificial intelligence model, it is essential to measure its performance to know whether the model is good at accomplishing the requested task. In the case of regression methods, we use metrics that assess the closeness between predicted and actual values.

Among the most common metrics, we find the mean square error (MSE), the coefficient of determination (R-squared).

To be able to compare our resulting models to the existing models, we use the MSE and the R-squared:

- **R-squared (R^2):** It is a statistical measure used to evaluate the performance of machine learning and deep learning models [52]. The R² is the proportion of the variation in the dependent variable that is explained by the independent variables in a regression model[53], In other words, the R² indicates how well the data fit the regression model, i.e. the quality of the fit [54]. Unlike other measures, such as RMSE, it does not indicate the degree of forecast accuracy.

  This metric is a derived measure, calculated from other measures, it is calculated mathematically by comparing the sum of squared errors (SSE) or the sum of squared residuals (SSR) to the total of squares (SST). the formula is represented as follows III.11:

  $$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{III.11}$$

  The formula of the Sum of Squared Errors SSE is :

  $$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{III.12}$$

  The formula of the Total Sum of Squares SST is :

  $$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{III.13}$$

where:
- SSE: The sum of squared differences between the actual values of the dependent variables and the predicted values of the regression model.
- SST: The SST is the total variation of the dependent variable.

Once the SSE and SST are calculated, the R^2 is determined by dividing the SES by the SST, then subtracting this result from 1 [52]. The interpretation of R-squared is reasonably easy. It is a decimal number in the range of 0 and 1, and it is frequently expressed as a percentage when referring to model fit. An R-squared of 100% illustrates that all changes in the dependent variable are entirely explained by changes in the independent variables. In contrast, an R^2 of 0% illustrates that the dependent variable cannot be predicted from the independent variable [55].

The R^2 is particularly sensitive to unnecessary features. Indeed, the R^2 value will either go up or remain the same when more variables are added, even if those variables are feebly related to the response. This may result in overfitting, particularly when dealing with numerous features.

- **Mean squared error(MSE):** MSE is a metric used to measure the amount of errors [56] and evaluate the effectiveness of regression models [55]. This metric is also known as the Mean Squared Deviation (MSD) [56].

MSE quantifies prediction error by calculating the average squared difference between actual values and predicted values [55]. The calculations for this metric are similar to those of variance [56], and its mathematical formulation is defined by the equation III.14:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{III.14}$$

Where:
n: represents the sample size.
y_i: represents the observed value of the dependent variable for the i-th observation.
$\hat{y}_i$: represents the predicted value of the dependent variable for the i-th observation.

The mean squared error imposes a stronger penalty on large errors by squaring the residuals. This means that the greater MSE will result from models that produce larger errors, therefore, our objective while employing MSE as a metric is to reduce its value. MSE indicates how closely the observed data and the model are related

to each other. A closer fit to the data is shown by a lower MSE.

As an advantage, MSE emphasizes larger errors, which means it gives bigger errors greater weight by squaring the residuals. This can be advantageous when larger errors are especially unsuitable. One disadvantage of MSE is that it is sensitive to outliers, because MSE squares the residuals, and a single outlier may have a significant impact on the MSE[55].

References

[1] Calculatrice de statistiques en ligne : test t, khi-deux, rÃ©gression, corrÃ©lation, analyse de variance.

https://datatab.fr/tutorial/linear-regression.

[2] Kavitha, S., Varuna, S., Ramya, R. (2016, November). A comparative analysis on linear regression and support vector regression. In 2016 online international conference on green engineering and technologies (IC-GET) (pp. 1-5). IEEE.

[3] Qu'est-ce que la régression linéaire ? – La régression linéaire expliquée – AWS. (s. d.). Amazon Web Services, Inc. https://aws.amazon.com/fr/what-is/linear-regression/

[4] Support Vector Regression In Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/

[5] Support Vector Regression | Learn the Working and Advantages of SVR. (s. d.). EDUCBA. https://www.educba.com/support-vector-regression/.

[6] Unlocking the True Power of Support Vector Regression | by Ashwin Raj | Towards Data Science. https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0

[7] Arbres de classification et de régression. (s. d.). XLSTAT, Your data analysis solution.https://www.xlstat.com/fr/solutions/fonctionnalites/arbres-de-classification-et-de-regression

[8] Decision Tree Algorithm, Explained - KDnuggets. (s. d.). KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

[9] How to tune a Decision Tree in Hyperparameter tuning - GeeksforGeeks. (s. d.). GeeksforGeeks. https://www.geeksforgeeks.org/how-to-tune-a-decision-tree-in-hyperparameter-tuning/

[10] K-Nearest Neighbor(KNN) Algorithm - GeeksforGeeks. (s. d.). GeeksforGeeks. https://www.geeksforgeeks.org/k-nearest-neighbours/.

[11] Advantages and Disadvantages of KNN Algorithm. (s. d.). AspiringYouths. https://aspiringyouths.com/advantages-disadvantages/knn-algorithm

[12] Random Forest : A Complete Guide for Machine Learning | Built In. (s. d.). Built In. https://builtin.com/data-science/random-forest-algorithm

[13] Wang, W., Chakraborty, G., Chakraborty, B.: Predicting the risk of chronic kid ney disease (ckd) using machine learning algorithm. Applied Sciences 11(1), 202 (2020

[14] Random Forest Regression in Python - GeeksforGeeks. (s. d.). GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-regression-in-python/.

[15] XGBoost - GeeksforGeeks. https://www.geeksforgeeks.org/xgboost/.

[16] Wang, W., Chakraborty, G., Chakraborty, B.: Predicting the risk of chronic kid ney disease (ckd) using machine learning algorithm. Applied Sciences 11(1), 202 (2020)

[17] Data Science Interview Questions for IT Industry Part-3 : Supervised ML - Thinking Neuron. (s. d.). Thinking Neuron - Data Science application to real world problems !

https://thinkingneuron.com/data-science-interview-questions-for-it-industry-part-3-supervised-ml/XGBoost.

[18] Artificial Neural Network (ANN) with Practical Implementation | by Amir Ali | The Art of Data Scicne | Medium. https://medium.com/machine-learning-researcher/artificial-neural-network-ann-4481fa33d85a.

[19] Qu'est-ce-que le Deep Learning ? - Sage Advice France. Deep Learning : définition. (s. d.). Sage Advice France. https://www.sage.com/fr-fr/blog/glossaire/deep-learning-definition/.

[20] An Introduction to Artificial Neural Networks | by Srivignesh Rajan | Towards Data Science. https://towardsdatascience.com/an-introduction-to-artificial-neural-networks-5d2e108ff2c3.

[21] Tutoriel de réseau de neurones artificiels avec des exemples TensorFlow ANN. (s. d.). Guru99. https://www.guru99.com/fr/artificial-neural-network-tutorial.html.

[22] Comprendre les réseaux de neurones. (s. d.). MonCoachData. https://moncoachdata.com/blog/comprendre-les-reseaux-de-neurones/.

[23] Artificial Neural Networks and its Applications - GeeksforGeeks. (s. d.). GeeksforGeeks. https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

[24] L'équipe éditoriale IONOS. (2020, 10 mars). Réseau de neurones artificiels : quelles sont leurs capacités ? IONOS Digital Guide. https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/quest-ce-quun-reseau-neuronal-artificiel/.

[25] DeepAI. (2019, 17 mai). Weight (Artificial Neural Network). https://deepai.org/machine-learning-glossary-and-terms/weight-artificial-neural-network.

[26] Weights and Bias in Neural Networks - GeeksforGeeks. (s. d.). GeeksforGeeks. https://www.geeksforgeeks.org/the-role-of-weights-and-bias-in-neural-networks/

[27] Turing. (2022, 29 septembre). Importance of Neural Network Bias and How to Add It. AI-Powered Engineering Services, LLM Training, Teams | Turing. https://www.turing.com/kb/necessity-of-bias-in-neural-networks

[28] Everything you need to know about "Activation Functions" in Deep learning models | by Vandit Jain | Towards Data Science. https://towardsdatascience.com/everything-you-need-to-know-about-activation-functions-in-deep-learning-models-84ba9f82c253.

[29] A Visual and Interactive Guide to the Basics of Neural Networks – Jay Alammar – Visualizing machine. https://jalammar.github.io/visual-interactive-guide-basics-neural-networks/.

[30] Activation functions in Neural Networks - GeeksforGeeks. (s. d.-b). GeeksforGeeks. https://www.geeksforgeeks.org/activation-functions-neural-networks/.

[31] Ali, M. (2023, 9 novembre). Introduction to Activation Functions in Neural Networks. Learn Data Science and AI Online | DataCamp. https://www.datacamp.com/tutorial/introduction-to-activation-functions-in-neural-networks.

[32] Understanding Linear and Non-linear Activation Functions in Deep Learning. (s. d.). Machine Mindscape. https://machinemindscape.com/understanding-linear-and-non-linear-activation-functions-in-deep-learning/.

[33] AI | Neural Networks | Binary Step Activation Function | Codecademy

[34] Activation Functions in Deep Learning : Sigmoid, tanh, ReLU - KI Tutorials. (s. d.). KI Tutorials. https://artemoppermann.com/activation-functions-in-deep-learning-sigmoid-tanh-relu/.

[35] Understanding ELU Activation Function: A Comprehensive Guide with Code Example | by DataScience-ProF | Medium. https://medium.com/@TheDataScience-ProF/understanding-elu-activation-function-a-comprehensive-guide-with-code-example-de7b152886a1.

[36] How to Choose an Activation Function for Deep Learning - MachineLearningMastery.com. (s. d.). MachineLearningMastery.com. https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/.

[37] Activation Functions Neural Networks : A Quick ; Complete Guide. (s. d.). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/activation-functions-and-their-derivatives-a-quick-complete-guide/.

[38] Types of Neural Networks and Definition of Neural Network. (s. d.-a). Great Learning Blog : Free Resources what Matters to shape your Career ! https://www.mygreatlearning.com/blog/types-of-neural-networks/.

[39] 6 Types of Neural Networks Every Data Scientist Must Know https://towardsdatascience.com/6-types-of-neural-networks-every-data-scientist-must-know-9c0d920e7fce

[40] DeepAI. (2019, May 17). Feed Forward Neural Network. https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network

[41] Biswal, A. (2020, April 24). Power of Recurrent Neural Networks (RNN): Revolutionizing AI. Simplilearn.com. https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn

[42] Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis | by Carolina Bento | Towards Data Science https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141

[43] Step 6: Understanding Recurrent Neural Networks | by Gourav Didwania | . | Medium.https://medium.com/aimonks/step-6-understanding-recurrent-neural-networks-1618a4e982b2.

[44] Types of Recurrent Neural Networks (RNN) in Tensorflow - GeeksforGeeks. (n.d.). GeeksforGeeks. https://www.geeksforgeeks.org/types-of-recurrent-neural-networks-rnn-in-tensorflow/.

[45] Introduction to Recurrent Neural Network - GeeksforGeeks. (n.d.). GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/.

[46] What is LSTM - Long Short Term Memory? - GeeksforGeeks. (n.d.). GeeksforGeeks. https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/

[47] LSTM Vs GRU in Recurrent Neural Network : A Comparative Study – AIM. (s. d.). AIM – Artificial Intelligence, And Its Commercial, Social And Political Impact. https://analyticsindiamag.com/lstm-vs-gru-in-recurrent-neural-network-a-comparative-study/.

[48] Introduction to Convolution Neural Network - GeeksforGeeks. (n.d.). GeeksforGeeks. https://www.geeksforgeeks.org/introduction-convolution-neural-network/

[49] Qu'est-ce qu'un réseau de neurones convolutifs ? | IBM. (s. d.). IBM - United States. https://www.ibm.com/fr-fr/topics/convolutional-neural-networks.

[50] Réseaux neuronaux convolutifs (CNN) : Introduction | Geekflare. (s. d.). Geekflare France. https://geekflare.com/fr/convolutional-neural-networks/.

[51] Keita, Z. (2023, 14 novembre). An Introduction to Convolutional Neural Networks : A Comprehensive Guide to CNNs in Deep Learning. Learn Data Science and AI Online | DataCamp. https://www.datacamp.com/tutorial/introduction-to-convolutional-neural-networks-cnns.

[52] R Squared : Understanding the Coefficient of Determination. (s. d.). Arize AI. https://arize.com/blog-course/r-squared-understanding-the-coefficient-of-determination/

[53] (19)Understanding Evaluation Metrics in Machine Learning: R-squared, Adjusted R-squared, MSE, MAE, and RMSE. https://medium.com/@brandon93.w/regression-model-evaluation-metrics-r-squared-adjusted-r-squared-mse-rmse-and-mae-24dcc0e4cbd3

[54] R-Squared. (n.d.). Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/r-squared

[55] Regression Model Evaluation Metrics: R-Squared, Adjusted R-Squared, MSE, RMSE, and MAE | by Brandon Wohlwend | Medium. https://medium.com/@brandon93.w/regression-model-evaluation-metrics-r-squared-adjusted-r-squared-mse-rmse-and-mae-24dcc0e4cbd3.

[56] Mean Squared Error (MSE). (s. d.-b). Statistics By Jim. https://statisticsbyjim.com/regression/mean-squared-error-mse

Chapter IV

Data

The responsible management of potable water, an irreplaceable resource, is essential to meet the current and future needs of humanity and the planet. In this study, we analysed the water use in Algeria using a dataset collected over 21 years provided from the Algerian Water Company (ADE).

The ADE is responsible for supervising water resources in Algeria and plays a crucial role in providing potable water to populations. It is a public water company created in 2001 governed by the Ministry of Water Resources.

This company carried out the management and supply of potable water, wastewater treatment, supervision of the infrastructure (such as pumping stations, reservoirs, distribution networks, and water treatment plants), as well as achieving national objectives in potable water, sanitation, and water resource preservation.

It ensures the daily production and distribution of 5.5 million m^3 of potable water, serving over 28.5 million inhabitants nationwide. To transport the produced water to citizens and the most remote areas of the national territory, the ADE uses a network of potable water pipelines called AEP, which spans 83,407 kilometers, including 29,863 kilometers of conveyance pipelines. This complex network is utilized for the production, transfer, storage, and distribution of potable water. It comprises 4,065 boreholes, 351 sources, and 1,792 pumping stations, enabling the collection of this resource and its delivery to 91 treatment stations tasked with making it potable before distribution through the 7,543 storage reservoirs managed by the ADE. Additionally, the ADE has two iron removal stations, 19 demineralization stations, and seven monobloc desalination stations, which are used to reduce the mineral and salt content in the treated waters.

The water sources it uses are distributed as follows:

- Approximately 45% come from groundwater sources such as rivers, lakes, and reservoirs, which are often vital sources of potable water.

- Approximately 39% are surface waters stored in aquifers and groundwater, typically less exposed to pollution and benefiting from natural protection through layers of soil and rock.

- Approximately 16% are desalinated waters, primarily from seawater, treated to remove salts and other impurities, making the water suitable for human consumption, agricultural irrigation, and other uses[1].

To carry out our study, ADE provided us with numerical data on user consumption throughout the Tizi-Ouzou province. The Tizi-Ouzou province serves as the capital of the Kabyle region and ranks as the second most populous city in Kabylie after Bejaia. It is situated around 30 kilometers south of the Mediterranean coast and 100 kilometers east of the capital Algiers. This province includes 308634205 subscribers located in 21 regions.

In the provided data, each region is identified by a number: (1) Tizi-Ouzou town, (2) Aïn El Hammam, (3) Azazga, (4) Draâ El Mizan, (5) Larbaâ Nath Irathen, (6) Tigzirt, (7) Ouacifs, (8) Azeffoun, (9) Draâ Ben Khedda, (10) Boghni, (11) Ouadhias, (12) Tizi Rached, (13) Beni Douala, (14) Bouzguene, (15) Ouaguenoun, (16) iferhounen, (17) Maatkas, (19) Mekla, (20) Makouda, (33) freha. The geographic map presented in figure IV.1 shows the localization of all regions of the Wilaya of Tizi-Ouzou.:

Figure IV.1: Geographic map depicting all regions of the Wilaya of Tizi-Ouzou.

Our Dataset consists of 21 data tables, stored in separate files. Each table stands for one region. The data table contains time series related to the subscriber's consumption collected from 2003 up to 2023. Figure IV.2 shows an example of a data table.

	NUM	CENT	TRIM	MONT	QTE	CAT_ACT	TYPABON	ETATCPT	ADRESSE
2	110001	02	20030331	256.80	0	10	10	10	TIZI OUMALOU N°1
3	110001	02	20030630	256.80	0	10	10	10	TIZI OUMALOU N°1
4	110001	02	20030930	265.60	1	10	10	10	TIZI OUMALOU N°1
5	110001	02	20031231	586.22	30	10	10	20	TIZI OUMALOU N°1
6	110001	02	20040331	586.22	30	10	10	20	TIZI OUMALOU N°1
7	110001	02	20040630	256.80	0	10	10	10	TIZI OUMALOU N°1
8	110001	02	20040930	586.22	30	10	10	20	TIZI OUMALOU N°1
9	110001	02	20041231	309.63	6	10	10	10	TIZI OUMALOU N°1
10	110001	02	20050331	2078.58	62	10	10	10	TIZI OUMALOU N°1
11	110001	02	20050630	778.90	29	10	10	10	TIZI OUMALOU N°1

Figure IV.2: Part of a data table from one of the regions of Tizi-Ouzou called Ain El HEMMAM.

As shown in figure IV.2, the data table contains 11 columns:

- The 'NUM' column, designates a reference number assigned to each subscriber in the region used.
- The 'CENT' column, represents the number of each region as depicted on the geographic map in Figure IV.1.
- The 'TRIM' column is in date format. It stands for the timestamps determined to check the meter and report the water index. The water index allows us to figure out the water quantity used by the subscriber. It is reported periodically depending on the subscription type. There are two types of subscriptions: quarterly subscription and monthly subscription. The timestamps defined for the quarterly subscribers are 31/03, 30/06, 30/09, and 31/12, while the timestamps defined for the monthly subscribers start on 31/01 and end on 31/12.
- The 'QTE' column represents the quantity consumed in cubic meters per subscriber for each timestamp. It may contain zeros, indicating zero consumption.
- The amount to be paid is listed in the previous column with the name 'MONT' assigned to each consumed quantity, even for quantity 0 (they have set a symbolic price that has increased over the years).
- The 'TYPABON' column indicates the number of the subscriber type. There are 33 types of subscribers. A different number is assigned for each subscriber type. The different subscribers types are listed in table IV.1:

Numéro	Type d'abonnees
10	Ménage Individuel
11	Ménage divisionnaire
12	Ménage Coll. EPEOR
13	Ménage Coll. OPGI
14	Autre Ménage Coll
15	Vente en gros
20	Établissement/APC
21	Établissement Militaire
22	Établissement /Wilaya
23	Sante
24	Enseignement
25	Administrations
26	P.T.T
27	D.G.S.N
28	GENDARMERIE
29	Enseignement Super
30	Commerce
31	Bain/Douche
32	Agence Commerciale
33	Hôtellerie N/Classée
34	Brt cl Provisoire
40	Activité industriel
41	Entrepôt
42	Élevage
43	Agriculture
44	Hôtellerie Classée
45	Complexe Touristique
46	Camp de Vacances
47	Chantier
48	Gros Consommateur
49	Bains & Douches
60	Eau Agricole
80	Structure ADE

Table IV.1: The list of the subscriber's types and their reference numbers.

- The 'ETATCPT' column, stands for the meter status. It shows numerical values. Each value indicates a different state, as listed in the table IV.2:

Numéro Du compteur	**État du compteur**
10	En marche
20	A l'arrêt
21	Horlogerie Cassée
20	Sans compteur
40	Résilier
41	Non Branché

Table IV.2: The Meter status and their reference number.

- The last column displays the address of each user under the name 'ADRESSE'.

After Analysing our dataset, we noticed the presence of missing data for certain users, which means that there are missing timestamps for both monthly subscribers as well as quarterly subscribers. To elaborate further, we counted the number of occurrences for each subscriber. We found that subscribers with 84 occurrences stand for complete quarterly data (4 quarters*21 years=84 occurrences), and those with 252 occurrences came from complete monthly data (12 months*21). If the occurrence is less than 84 and 252, the data are considered missing data.

Based on this analysis, first, we divided our data into two parts: Monthly data and quarterly data. Then, we converted the monthly data into quarterly data to be able to conduct our study on both of them simultaneously. Next, we excluded the missing data and kept only the complete data by filtering the subscribers with 84 occurrences from the converted data and the quarterly data. Finally, we combined them, resulting in a dataset containing only 84 occurrences, and we applied this to each region.

The table IV.3 presents the distribution of subscribers by region, highlighting the number of subscribers who have complete quarterly data:

Region	Number of subscribers	84 occurrences subscribers
DBK	11,292 subscribers	3157 subscribers
AZEFFOUN	17,715 subscribers	3487 subscribers
BOGHNI	17,463 subscribers	2561 subscribers
FRAHA	8,964 subscribers	2848 subscribers
IFERHOUNENE	5,575 subscribers	1296 subscribers
MAATKAS	14,026 subscribers	6129 subscribers
DRAA EL MIZAN	20,733 subscribers	415 subscribers
LNI	19,755 subscribers	8709 subscribers
OUADHIAS	13,138 subscribers	4914 subscribers
BENI DOUALA	19,726 subscribers	7,147 subscribers
OUAGUENOUN	22,884 subscribers	6692 subscribers
BOUZEGUENE	11 ,197 subscribers	3646 subscribers
Ain-EL-HAMMAM	16,220 subscribers	5134 subscribers
TIZI RACHED	10,260 subscribers	3744 subscribers
MEKLA	11,625 subscribers	3334 subscribers
TIGZIRT	20,478 subscribers	1488 subscribers
MAKOUDA	11,741 subscribers	1815 subscribers
OUACIFS	15,333 subscribers	7298 subscribers
AZAZGA	22,198 subscribers	7,232 subscribers
TIZI OUZOU TOWN	42,000 subscribers	10,403 subscribers
TIZI OUZOU NEW TOWN[2]	38,454 subscribers	10,561 subscribers

Table IV.3: The distribution of the number of subscribers by region and the number of subscribers with complete quarterly data

As shown in table IV.3, Tizi-Ouzou town has the largest number of subscribers. Based on this observation, this region is selected for conducting our analysis. Figure IV.3 shows an overview of the data table of the Tizi-Ouzou town:

	NUM	CENT	TRIM	MONT	QTE	CAT_ACT	TYPABON	ETATCPT	ADRESSE	AGENT
2	350292	01	20060930	952.53	34	10	10	10	RUE GUERRABA AHCENE	68
3	350292	01	20061231	883.08	32	10	10	10	RUE GUERRABA AHCENE	68
4	350292	01	20070331	639.99	25	10	10	10	RUE GUERRABA AHCENE	68
5	350292	01	20070630	639.99	25	10	10	10	RUE GUERRABA AHCENE	68
6	350292	01	20070930	987.26	35	10	10	10	RUE GUERRABA AHCENE	68
7	350292	01	20071231	1160.89	40	10	10	10	RUE GUERRABA AHCENE	68
8	350292	01	20080331	1647.07	54	10	10	10	RUE GUERRABA AHCENE	68
9	350292	01	20080630	1473.44	49	10	10	10	RUE GUERRABA AHCENE	68
10	350292	01	20080930	5537.71	117	10	10	10	RUE GUERRABA AHCENE	68

Figure IV.3: A section of the Excel file for the TIZI-OUZOU Town.

Once the data are preprocessed (cleaned), we obtained the results described in table V.1:

Data preprocessing	Results
ADE's number of subscribers	42000 subscribers
Quarterly subscribers with 84 occurrences	10403 subscribers
Monthly subscribers with 252 occurrences	0 subscribers
Number of subscribers between 0 and 84 occurrences	31330 subscribers
Number of subscribers between 85 and 252 occurrences	181 subscribers
Number of subscribers after converting the monthly subscribers to quarterly subscribers (with 84 occurrences)	157 subscribers
Total number of NAN	15 NAN
Number of rows where QTE=0	285600 rows
Number of rows where $QTE \neq 0$	1603078 rows
Number of NAN after removing QTE=0	14 NAN
Number of subscribers after removing the rows where QTE=0	40698 subscribers
The total quantity consumed by all subscribers	$93110650m^3$
Number of rows in the dataframe	1888678 rows

Table IV.4: The results of the data preprocessing of Tizi-Ouzou town.

The following table shows the results of our data preprocessing for the city of Tizi Ouzou, on which our study is based. We chose this region because it has the largest number of subscribers among the other 21 regions, with exactly 42,000 subscribers. As previously mentioned, there are three types of data: complete data, with no missing values, totaling 10,403. Complete monthly data is not available, and finally, missing data, including quarterly and monthly data, totals 31,511.
After converting our monthly data to quarterly data, we obtained 157 new data points. Adding this number to the complete quarterly data gives a total of 10,560 quarterly subscribers.

The graph below IV.4, which represents the amount of water consumed by all users in the Tizi Ouzou city region, is derived from a table with two columns: the quarterly dates from 03/31/2003 to 12/31/2023, and the amount of water consumed by all users for each of these dates. The table contains exactly 84 rows.

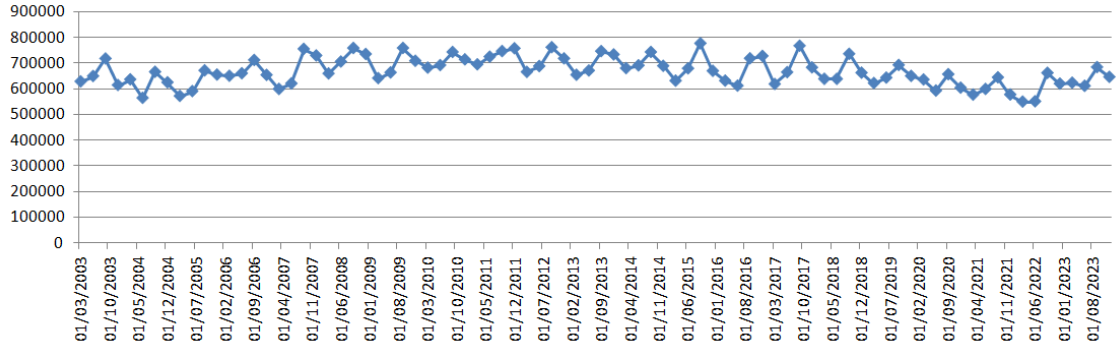


Figure IV.4: The graph of consumed quantity over quarterly dates in Tizi-Ouzou town.

The graph IV.4 represents a time series where each point on the graph illustrates the quarterly consumption at a specific date. It shows a certain stability with periodic fluctuations oscillating between approximately 500,000 and 800,000 m^3. There is no evident upward or downward trend over the entire period. It shows distinct seasonal variations. It can be observed that water consumption generally increases in summer, reaching high peaks. In contrast, water consumption decreases in winter, reaching lows.

References

[1] Présentation - Algérienne des Eaux. (s. d.). Accueil - Algérienne des Eaux. https://www.ade.dz/presentation

Chapter V

Experiments and results

In our study, we conducted a comparative experimentation involving a variety of regression algorithms. These algorithms included machine learning methods as well as deep learning methods. Machine learning methods involve decision trees, random forests, SVR, KNN, XGBoost, and linear regression, while deep learning methods encompass Artificial Neural networks (ANN), Long-term Term Memory (LSTM), Recurrent Neural networks (RNN) and Convolutional Neural Network (CNN).

To evaluate the performance of each technique, we used two metrics: Mean Squared Error (MSE) and R-squared (R^2) to assess the accuracy of our models.

The approaches are implemented in Python. Experiments are carried out on a MacBook Pro, 2.9GHz Intel Core i7 with 8Gb of RAM, HP 2.5GHz intel core i5 with 4Gb, and HP Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71 GHz with 8Gb of RAM.

To perform our experiments, we first divided the data into training and test sets:

- Training set: It helps the model to understand the patterns in the data,
- Test set: This set evaluates the accuracy of the resulting model[1].

Therefore, 70% was used for training, while 30% of the dataset was utilized for testing. The training subset contained data from 31/03/2003 to 30/09/2017, while the tested subset consisted of data from 30/12/2017 to 31/12/2023. Our goal is to forecast water use for 2020 to 2023 using observed data from 2003 to 2017.

Figure V.1 represents the data split, where the part in blue is the training subset and the one in orange is the test subset. This graph is a visual representation of the water use of all users in the region of Tizi-Ouzou town for each quarter from the year 2003 to 2023.

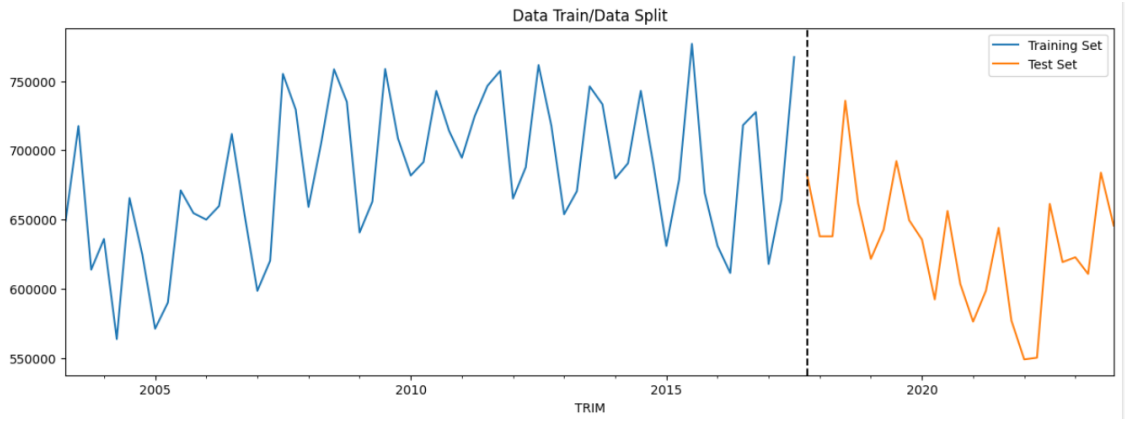


Figure V.1: The data split.

In Figure V.2, we have explored our time series in detail, dividing it to analyze the various components that characterize it (trend, seasonality, noise...), and have obtained the following results:

- The first graph shows our time series itself, which represents the evolution of quantities consumed over the years.
- The second graph shows the trend component of the time series. The trend line shows a general upward and downward movement.
- The third graph shows the seasonality component of the time series. It captures the repeating patterns or cycles at regular intervals.

For our experiment, we incorporated lag features (values from previous timestamps) from t-1 to t-10 to capture temporal dependencies and enhance the predictive performance of both machine learning and deep learning models.

To determine the best performance of the resulting models, we used two metrics: MSE and R-squared. It is essential to look for the lowest value of mean square error (MSE) and the highest value of the coefficient of determination (R-squared).

V.1 Machine learning methods:

- **K-nearest neighbors:**
  Considering the results presented in the table V.1 using lag features from t-1 to t-10, we observed varying impacts on the model performance. Notably, 4 and 5 lag

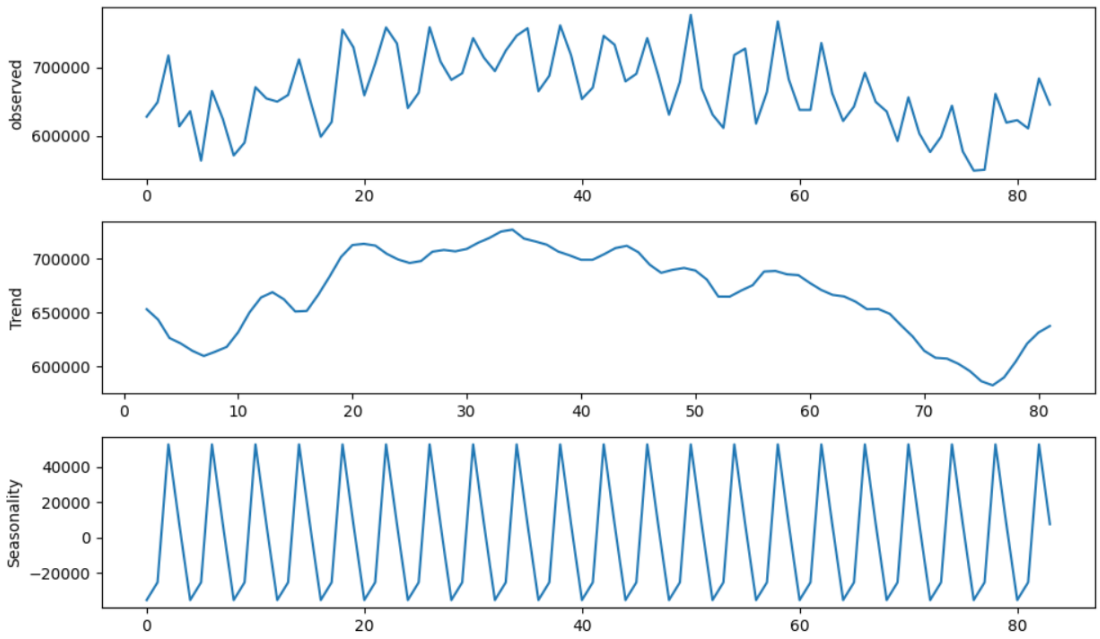


Figure V.2: The decomposition of the time series data into trend and seasonality.

features had the best results, with R^2 values of 0.1468 and 0.2206, respectively, and lower values of MSE of 1567900387.616666 and 1567900387.616666. However other lag values resulted in poorer model performance as indicated by negative R^2 and higher MSE.

lag features	R^2	MSE
1 lag	-1.0067	3758959161.8928
2 lags	-1.0094	3764106507.5664
3 lags	-0.5287	2863644840.0304003
4 lags	0.1468	1567900387.616666
5 lags	0.2206	1432401220.1316664
6 lags	-0.1788	2166280966.491667
7 lags	-0.4255	2619653376.235001
8 lags	-0.5075	2886032564.683479
9 lags	-0.5010	2873506948.010434
10 lags	-0.7941	3434655338.3200006

Table V.1: The evaluation of the performance of the K-nearest neighbors model using different lag features.

- **Support vector regression:**
  The results presented in the table V.2 show that the model is not effectively capturing the temporal dependencies in the data. This is evidenced by the consistently negative R^2 and high Mean squared error (MSE).

- **Linear regression:**

lag features	R^2	MSE
1 lag	-0.7828	3339524976.4225984
2 lags	-1.4143	4522524603.673753
3 lags	-0.7130	3208798960.505785
4 lags	-0.4758	2712177420.270154
5 lags	-0.4659	2693972690.4417214
6 lags	-0.5558	2859288488.058423
7 lags	-0.5659	2877731983.677871
8 lags	-0.5171	2904282558.1092963
9 lags	-0.7082	3270241725.5186462
10 lags	-0.8915	3621187477.660597

Table V.2: The evaluation of the performance of the support vector regression model using different lag features.

The best result in Table V.3 is 0.3834 with lags 5 which is quite low. This indicates that the linear regression model explains only 38.34% of the variance in the target variable with 5 lag features.

lag features	R^2	MSE
1 lag	-0.5827	2964787060.1963806
2 lags	-1.0014	3749187386.9870152
3 lags	-0.2723	2383346206.952019
4 lags	0.3082	1271357651.7156637
5 lags	0.3834	1133230037.9226215
6 lags	0.2046	1461728533.2492008
7 lags	0.1828	1501838080.6280174
8 lags	0.0704	1779561410.2068725
9 lags	0.1663	1596085192.9339826
10 lags	-0.1631	2226727931.7381544

Table V.3: The evaluation of the performance of Linear regression model using different lag features

- **Decision tree:**
  Based on the provided results in Table V.4, we observe that the model with 4 and 5 lags performs the best, as indicated by positive R^2 values (0.3259 and 0.3259, respectively) and lower MSE values (1238816184.5746622 and 1214475838.6011693, respectively). This suggests that these lag features are capturing some of the underlying temporal patterns in the data.

lag features	R^2	MSE
1 lag	-1.5716	4817123117.0320015
2 lags	-1.0300	3802751209.9466662
3 lags	-0.4193	2658728784.1712008
4 lags	0.1891	1490329430.3148148
5 lags	0.3810	1137567772.5154316
6 lags	-0.6267	3249761170.6672716
7 lags	0.4244	1057801138.7549998
8 lags	-0.1442	2190428786.9809785
9 lags	-0.2766	2444039770.728695
10 lags	-0.3747	2631732933.860871

Table V.5: The Evaluation of the performance of random forest model using different lag features

lag features	R^2	MSE
1 lag	-1.7725	5193610941.521506
2 lags	-0.9371	3628721852.734415
3 lags	-0.6140	3023355108.202293
4 lags	0.3259	1238816184.5746622
5 lags	0.3392	1214475838.6011693
6 lags	0.1160	1624570389.8164062
7 lags	0.1642	1535965713.8828125
8 lags	0.1206	1683538548.7900484
9 lags	-0.1243	2152474078.2892523
10 lags	-0.2345	2363396734.3404884

Table V.4: The evaluation of the performance of decision tree model using different lag features

- **Random forest:**
  The results in Table V.5 show that the performance improves from 1 lag to lag 5 with an increase in R^2 and a decrease in MSE. The models with 4,5 and 7 lags perform the best. This configuration seems to capture the underlying temporal patterns.

- **Xgboost:**
  The results of Table V.6 describe the performance of the XGboost model with various numbers of lag features. For 4,5, 6, 7, 8, and 9 lags, the R^2 values are positive. This indicates that the model can capture some of the variance in the data. However, The R^2 is generally low, suggesting that the model explains only a small portion of

the variance.

The model performs best with lag 5 with the highest value of R^2(0.5019) and lowest MSE(915482481.6183268), indicating it captures more variance compared to other lag features.

lag features	R^2	MSE
1 lag	-0.5848	2968745961.921875
2 lags	-0.9948	3736657149.052656
3 lags	-0.2723	2383338648.235156
4 lags	0.4230	1060351424.3836263
5 lags	0.5019	915482481.6183268
6 lags	0.4529	1005438327.9824219
7 lags	0.1828	1501831373.7371418
8 lags	0.0704	1779566366.2693615
9 lags	0.1703	1588370859.195822
10 lags	-0.3847	2650835277.278193

Table V.6: The evaluation of the performance of XGboost model using different lag features

As shown in the results provided in Table V.7, the best results are obtained from lag 5, compared to the other lags. XGboost is the best-performing model with the highest R^2 and lowest MSE, indicating it can explain more variance in the data and make more accurate predictions. Random Forest is the second-best performer, making it an alternative to XGboost. Decision tree, KNN, and linear regression are less effective compared to XGboost and random forest. Among them, Linear regression shows better performance than Decision tree and KNN.

Model	Lag	R^2	MSE
XGBoost	5	0.5019	915482481.6183268
Random forest	7	0.4244	1057801138.7549998
Decision tree	5	0.3392	1214475838.6011693
KNN	5	0.2206	1432401220.1316664
Linear regression	5	0.3834	1133230037.9226215
Support vector regression	5	-0.4659	2693972690.4417214

Table V.7: The comparative evaluation of the results for the different models with various lag configurations

The graphs V.3 V.4 V.5 V.6 V.7 show the alignment between the two lines provides insights into how well the model tracks the real data. The blue line represents the actual data, and the orange line represents the predicted values, The x-axis(TRIM) represents the time, divided into quarters from 2018 to 2023. The y-axis(QTE) represents the quantity

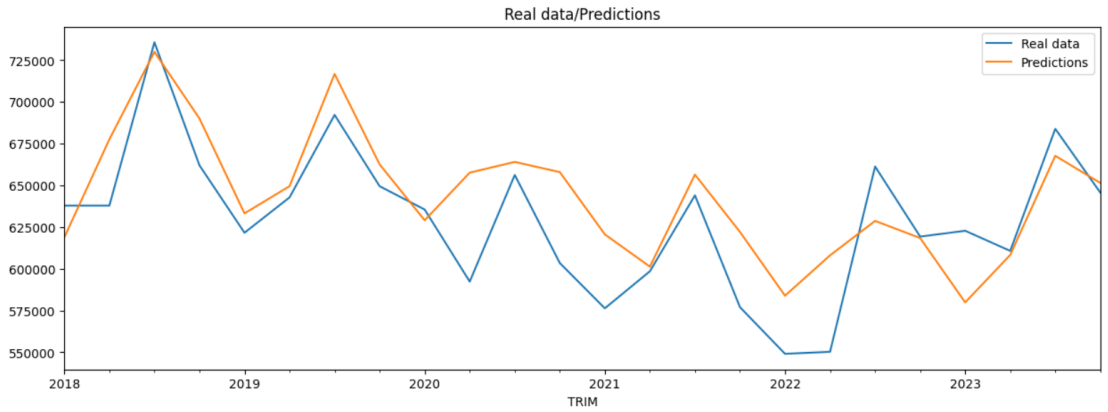


Figure V.3: Comparison between predicted values and real values for XGboost algorithm.

of water consumed.

In graph V.3, the model seems to closely follow the actual data from 2018 to 2019. starting from 2019, the predictions diverge more from the actual values, indicating that the model struggles with capturing more recent patterns. Indeed, the patterns are more complex.

In graph V.4 the model (orange line) follows the overall direction of the actual data (blue line) but not very closely. When the actual data increases (goes up), the model's predictions also show an increase. When the actual data decreases (goes down), the model's prediction also shows a decrease. This means the model recognizes that the value should be falling.

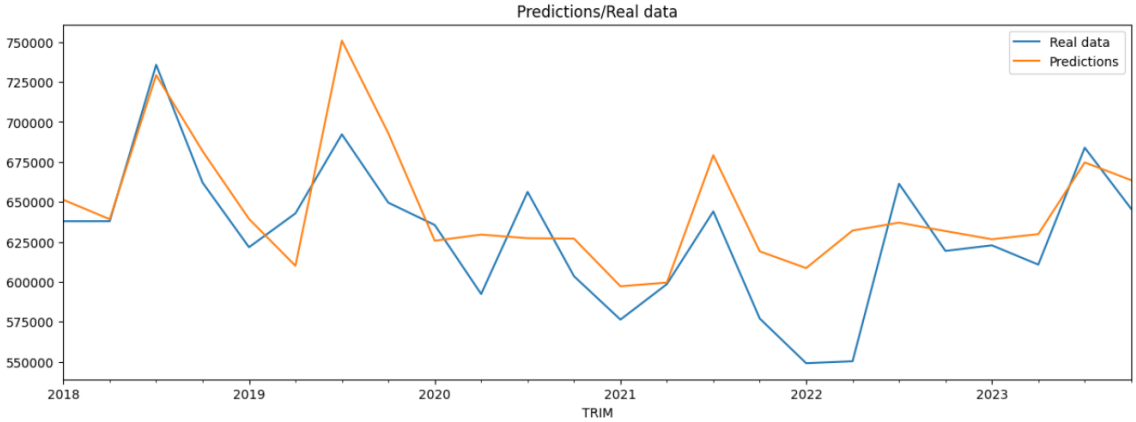


Figure V.4: Comparison between predicted values and real values for Random forest algorithm

For the graph V.5, the decision tree model produces several flats where the predictions do not vary, especially around 2021 and 2022. This suggests the model struggles to capture the variability in the actual data during these periods. Around 2019, the model

predicted higher values than the actual data, this is called overestimation. Around 2022-2023, the model predicts lower values than the actual data, it is known as underestimation.

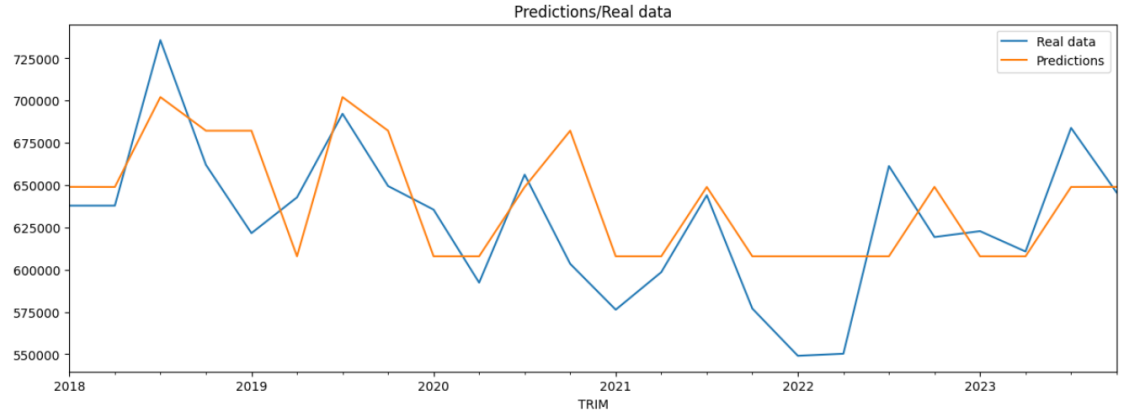


Figure V.5: Comparison between predicted values and real values for Decision tree algorithm.

Graph V.6 shows that the predicted values (orange line) capture some trends and patterns in the real data (blue line). There are periods where the predictions significantly overestimate the real data. For example, in early 2019 and early 2020, the predictions are much higher than the actual values.

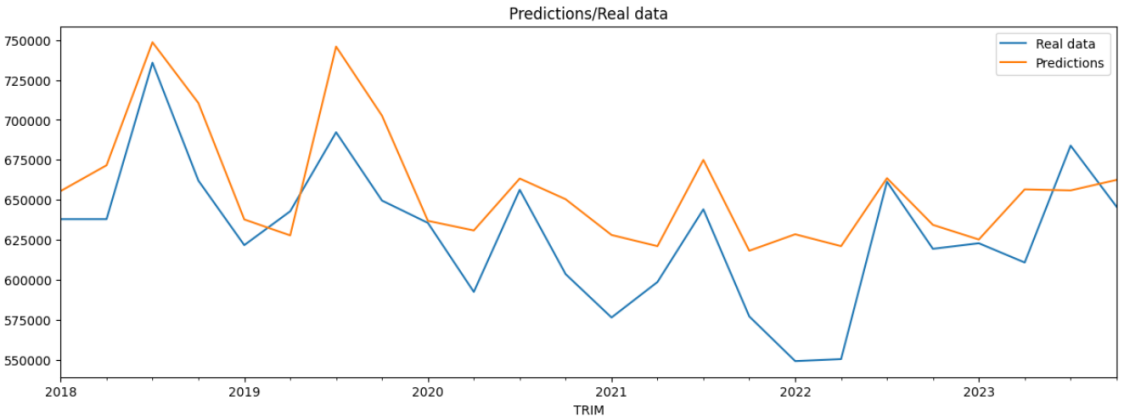


Figure V.6: Comparison between predicted values and real values for KNN algorithm.

For graph V.7 the linear regression predictions (orange line) generally follow the trends in the real data (blue line) better than KNN model did.

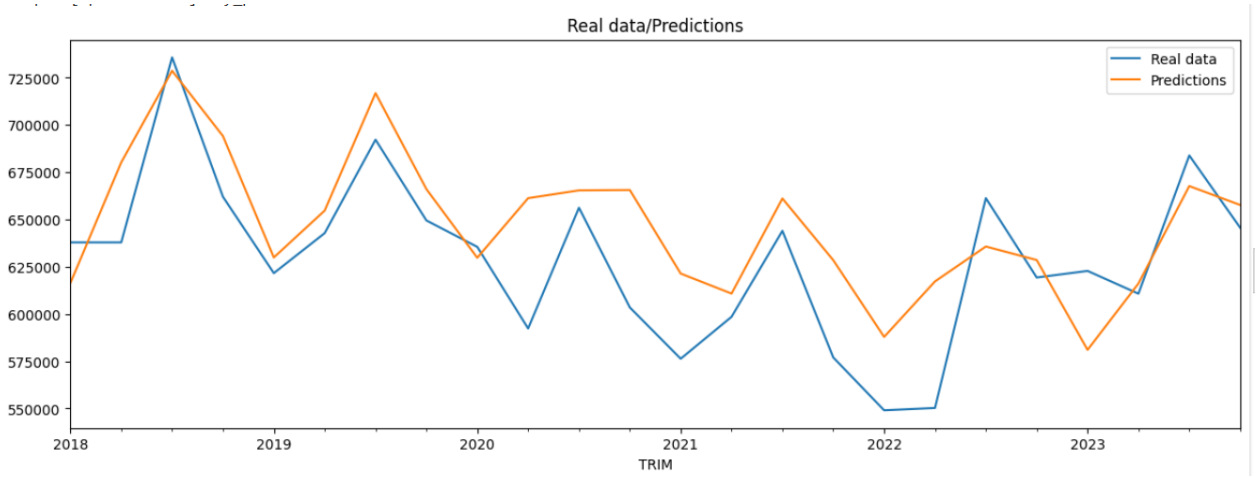


Figure V.7: Comparison between predicted values and real values for Linear Regression algorithm

Figure V.8 represents the graph of the worst model which is the Support vector regression model, The prediction line does not closely follow the real data, indicating a poor fit. There are some periods in mid-2020 where the predictions exceed the actual values. undershoots can be seen in late-2023.

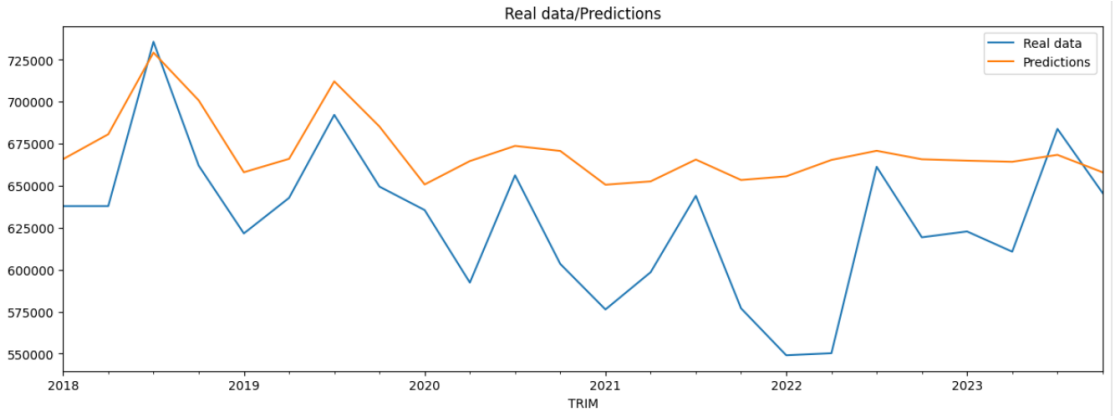


Figure V.8: Comparison between predicted values and real values for SVR algorithm.

The graph V.9 compares the predictions of water consumption over time made by different machine learning algorithms against the actual observed data. XGBoost tends to follow the real data quite closely but with some deviations. However, Random Forest's prediction line is relatively frequently aligned with the actual data, indicating it effectively captures the general trend.

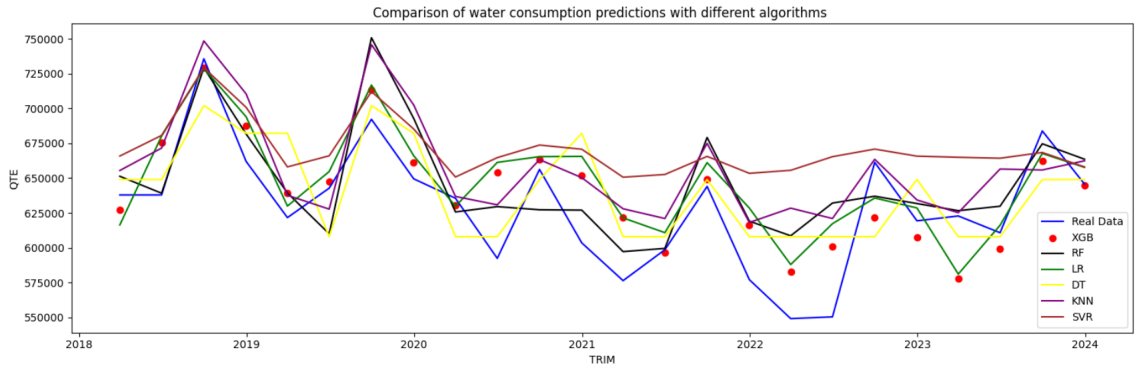


Figure V.9: Comparison of Water Consumption Predictions Using Various Machine Learning Algorithms.

V.2 Deep learning methods:

- **ANN:**
  As shown in table V.8, the ANN model performance best with 5 lags, achieving the highest R^2 (0.75) and the lowest MSE (451,571,170.0). Using too few (1-3) or too many (6-10) lags generally results in poorer performance, as indicated by lower R^2 values and higher MSEs. There's a notable improvement in the model performance when moving from 3 to 4 lags, indicating that more lag features provide significantly better predictive power up to a point (5lags).

lag features	R^2	MSE
1 lag	-1.48	4652198000.0
2 lags	-0.32	2464130000.0
3 lags	-0.50	2803090700.0
4 lags	0.52	876681300.0
5 lags	0.75	451571170.0
6 lags	0.18	1513313300.0
7 lags	0.44	1021421250.0
8 lags	0.51	946757100.0
9 lags	0.39	1164760300.0
10 lags	0.55	870921000.0

Table V.8: The evaluation of the performance of ANN model using different lag features

- **CNN:**
  The results provided in Table V.9 suggest that the model performs best with 3 or 4 lag features, as indicated by the highest R^2 values (0.54 and 0.56) and the lowest MSE values (887779614.11 and 886326596.87). The performance deteriorates significantly when using more than 5 lag features.

lag features	R^2	MSE
1 lag	-0.52	2911697014.4442935
2 lags	0.43	1090043985.300102
3 lags	0.54	887779614.1078465
4 lags	0.56	886326596.8652344
5 lags	0.53	931084488.3556463
6 lags	0.07	1856703316.9582741
7 lags	0.16	1682938511.2070312
8 lags	-0.11	1683409409.8225446
9 lags	-0.41	2139407763.6947544
10 lags	-0.16	1760495393.2066593

Table V.9: The evaluation of the performance of CNN model using different lag features

- **RNN:**
  The results in Table V.10 indicate how the performance of our RNN model varies with different numbers of lag features:

  – The model with just one lag feature performs poorly, as indicated by the very low R^2 (0.05) and high MSE (662851126.3210938). The model explains only 5% of the variance in the target variable.

  – Introducing a second lag feature significantly improves the model's performance. The R^2 value increases to 0.67, indicating that the model explains 67% of the variance in the target variable.

  – Adding a third lag feature seems to reduce the model's performance compared to using two lags. The R^2 drops to 0.50, and MSE increases. This could be due to overfitting.

  – With four lag features, the model performs much better, explaining 80% of the variance in the target variable. The MSE is also relatively low, indicating a better fit.

  – Adding a fifth lag feature further improves the model's performance. The R^2 value increases up to 0.83, and the MSE decreases, suggesting that the model is explaining more variance and providing better predictions.

  – With six lag features, the model performs the best, with an R^2 of 0.94 and the lowest MSE. This suggests that the model can explain 94% of the variance in the target variable and provides highly accurate predictions.

  – Adding a seventh lag feature reduces the model's performance compared to six lags. The R^2 drops to 0.71, and the MSE increases. This may indicate overfitting.

- The model's performance deteriorates with eight lags. The R^2 value decreases to 0.60, and the MSE increases, suggesting that the additional lags are not beneficial and may be introducing noise.
- Introducing nine lag features improves the model's performance again, with an R^2 of 0.81 and a relatively low MSE. This indicates that nine lags might capture relevant information better than eight lags.
- The model's performance drops slightly with ten lags, with an R^2 of 0.72 and an increased MSE compared to nine lags.

We conclude that the best performance is achieved by **six lags features** , with the highest R^2 (0.94) and the lowest MSE (45,783,848.89).

lag features	**R^2**	**MSE**
1 lag	0.05	662851126.3210938
2 lags	0.67	230249931.99921876
3 lags	0.50	345527193.140625
4 lags	0.80	157792500.16210938
5 lags	0.83	130172487.30566406
6 lags	0.94	45783848.888671875
7 lags	0.71	221677660.29101562
8 lags	0.60	355637123.22265625
9 lags	0.81	166968894.40104166
10 lags	0.72	245577177.56640625

Table V.10: The evaluation of the performance of RNN model using different lag features.

- **LSTM:**
  The results provided in table V.11, All R^2 values are negative, indicating that the models perform worse. This suggests that the model is not capturing the underlying patterns in the data, and the MSE values are quite large. The negative R^2 values and high MSE values across different lags suggest that the LSTM model is ineffective at predicting the target variable 'QTE' from its lagged values, maybe the model is too simple to capture the dependencies in the data.

lag features	R^2	MSE
1 lag	-0.30	2873992028.739844
2 lags	-8.80	4833471268.4375
3 lags	-0.21	4491686657.496773
4 lags	-3.03	3780667629.057617
5 lags	-0.23	2429592905.7802734
6 lags	-0.46	1947592345.022461
7 lags	-1.66	1726172624.7539062
8 lags	-0.41	4010327929.122396
9 lags	-0.17	3115098413.4453125
10 lags	-0.33	4093066222.7265625

Table V.11: The evaluation of the performance of LSTM model using different lag features

The Table V.12, shows the best results we obtained for different deep learning models using various numbers of lag features:

- The RNN with 6 lag features has the highest R^2 value of 0.94 indicating that the resulting model explains 94% of the variance in the target variable. The MSE is the lowest among the three models, indicating the least prediction error. This model appears to be the best performer among the three, with both high accuracy (high R^2) and low error (low MSE).
- The ANN with 5 lag features has an R^2 value of 0.75, indicating that the mode explains 75% of the variance in the target variable. The MSE is higher compared to the RNN, indicating more prediction error. This model performs reasonably well but it is outperformed by the RNN.
- The CNN with 4 lag features has the lowest R^2 value of 0.56, indicating that the model explains 56% of the variance in the target variable. The MSE is the highest among the three models, indicating the highest prediction error. This model performs the worst among the three, with lower accuracy and higher error.

Model	lag	R^2	MSE
RNN	6 lags	0.94	45783848.888671875
ANN	5 lags	0.75	451571170.0
CNN	4 lags	0.56	886326596.8652344

Table V.12: The comparative evaluation of the results for deep learning models with various lag configurations

The graphs V.10, V.11, V.13 and V.12 present the qualitative evaluation of the 4 deep learning models. Graph V.10 shows predictions from a Recurrent Neural network (RNN)

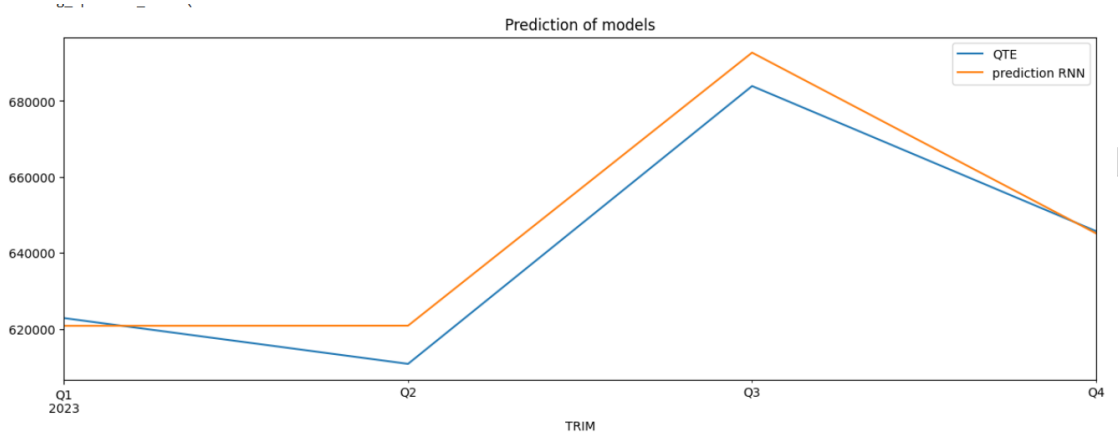


Figure V.10: Comparison between predicted values and real values for RNN model.

model compared to real data for the year 2023. The blue line represents the real data (QTE), and the orange line represents the predictions made by the RNN model.

The RNN predictions closely follow the real data. The model captures the overall trend and changes in the data very accurately.

The visual representation in figure V.11 shows that the ANN model predictions closely follow the actual data, capturing most of the important changes in the data over time. The predicted values are higher than the actual values in some periods. For instance, around the middle of 2018, the orange line (predictions) is above the blue line (actual values). The predicted values are lower than the actual values in other periods. For example, around late 2020, the orange line is below the blue line.

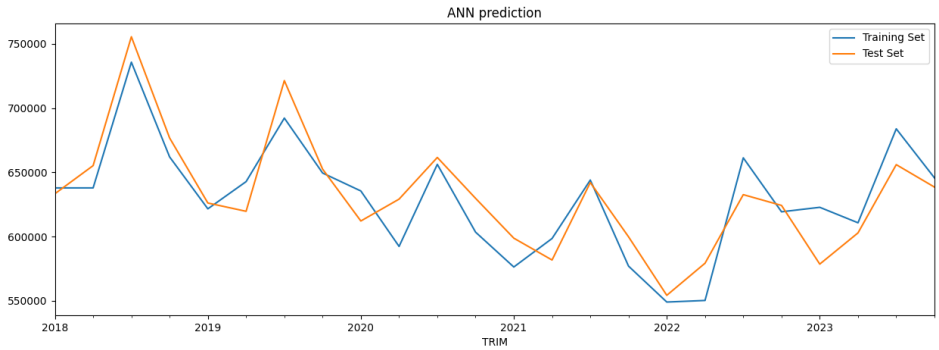


Figure V.11: Comparison between predicted values and real values for ANN model.

The graph V.12 shows that the orange line (CNN predictions) follows the general trend of the blue line (actual data), but there are noticeable differences. Indeed, The model captures some of the peaks and troughs but misses others. For instance: The peak around 2019 is well captured, the troughs in 2020 and 2022 are not accurately predicted, the model predicts higher values than the actual ones in early 2019 and mid-2022, and The

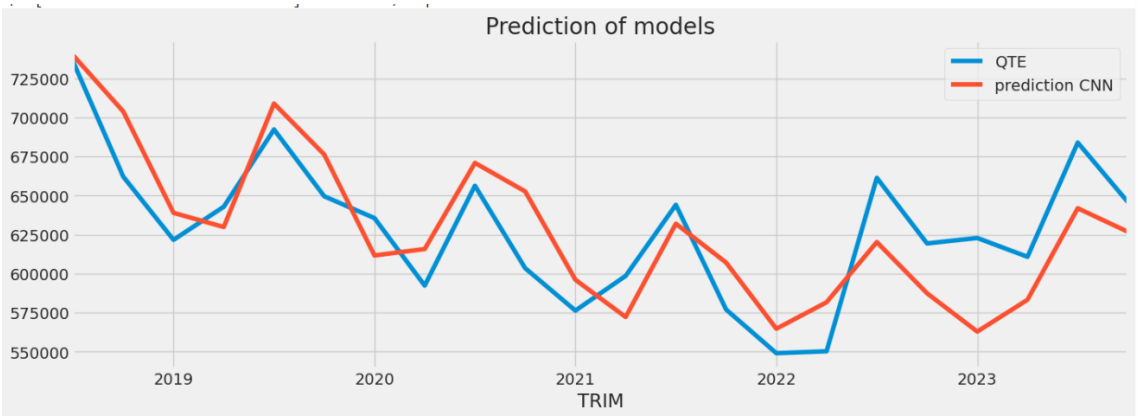


Figure V.12: Comparison between predicted values and real values for CNN model.

model predicts lower values than the actual ones in late 2019 and early 2021.

Finally, Figure V.13 represents the worst model "LSTM". The LSTM model performed poorly in predicting the given dataset, as demonstrated by the negative R^2 value and high MSE. The model did not effectively capture the peaks or troughs.

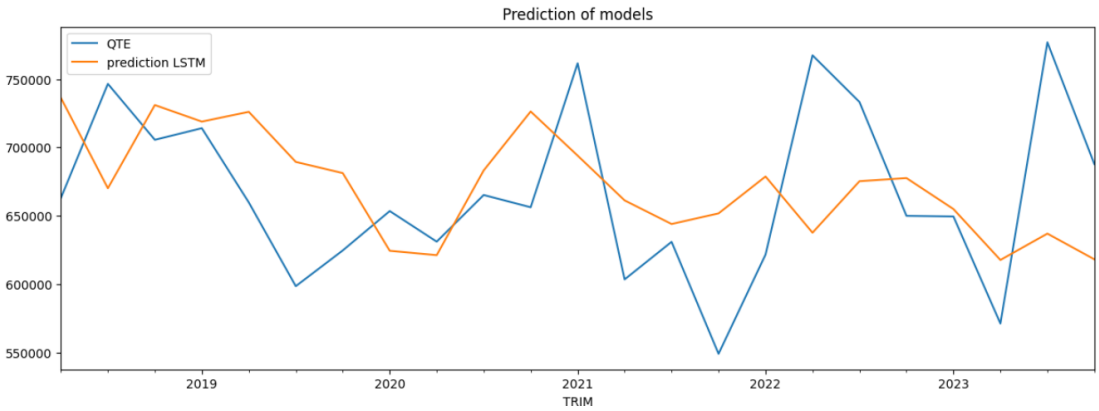


Figure V.13: Comparison between predicted values and real values for LSTM model.

References

[1] A Guide to Data Splitting in Machine Learning | by Data Science Wizards | Medium. https://medium.com/@datasciencewizards/a-guide-to-data-splitting-in-machine-learning-49a959c95fa1.

Conclusion

This research has highlighted the potential of machine learning and deep learning methods to predict water use in Algerie using detailed analysis. We used six major machine learning algorithms: linear regression, KNN, SVR, Decision trees, Random forests, and XGBoost. We investigate also four main deep learning methods: ANN, RNN, LSTM and CNN. The performance of these several algorithms was tested using two essential metrics: the R-squared and the MSE.

We conducted several experiments using ten different lag features from t-1 to t-10. We found that some features harm the prediction, so we had to test for each model the best features that give the best performance.

XGBoost proved to be the best machine learning model with an R^2 of 0.5019 and an MSE of 915,482,481.618. For deep learning, the Recurrent Neural Network (RNN) achieved the highest accuracy with an R^2 of 0.94 and an MSE of 45,783,848.889. The Artificial Neural Network (ANN) also performed well, with an R^2 of 0.75 and an MSE of 451,571,170.0. This study highlights the superior performance of deep learning models, particularly the RNN, compared to traditional machine learning models for predicting water use.

The results of our study have an important benchmark value for water management as a sustainable resource. Sure the establishment of the algorithms is not perfect and there is still much room for improvement but they provide an opportunity for understanding their limitations, and will certainly be valuable to future studies to develop more accurate methods.

In the future, it is essential to continue improving our models and exploring hybrid methods that combine the strengths of different algorithms to increase the accuracy of water consumption predictions. Future research should focus on integrating more diverse data, including real-time data, weather data, demographic data, and economic data.

By optimizing current models and exploring new techniques, we can not only improve the accuracy of predictions but also provide more robust tools for water resource manage-

ment. These improvements will contribute to better planning and management of water resources, crucial for meeting the growing needs of the population of Tizi Ouzou.

Conclusion Générale

Dans ce projet, nous avons exploité divers modèles : modèles d'apprentissage automatique et modèles d'apprentissage profond, afin de prédire la consommation d'eau des abbonées de l'ADE en utilisant des séries temporelles (données historiques). Nous avons ensuite évalué ces modèle afin d'identifier le modèle permettant de réaliser les prédictions les plus précises.

XGBoost s'est révélé être le modèle d'apprentissage automatique le plus performant avec un R^2 de 0,5019 et une MSE de 915,482,481.618. Ce modèle a montré une capacité modérée à capturer la variabilité des données et a fourni les prédictions les plus fiables parmi les approches traditionnelles d'apprentissage automatique.

En ce qui concerne l'apprentissage profond, le Réseau de Neurones Récurrent (RNN) s'est avéré être le modèle le plus efficace dans l'ensemble, avec un R^2 de 0,94 et une MSE remarquablement basse de 45,783,848.889. La capacité de ce modèle à capturer les dépendances temporelles dans les données a abouti à des prédictions très précises. Le Réseau de Neurones Artificiels (ANN) a également très bien performé, atteignant un R^2 de 0,75 et une MSE de 451,571,170.0. Ce modèle a surpassé tous les modèles d'apprentissage automatique, démontrant de fortes capacités prédictives.

Les résultats de cette étude mettent en évidence la performance supérieure des modèles d'apprentissage profond par rapport aux modèles d'apprentissage automatique traditionnels pour la prédiction de notre jeu de données. Le RNN, en particulier, a montré une précision prédictive exceptionnelle. Par conséquent, il est le modèle le plus recommandé pour ce type de données. L'ANN a également montré de bonnes performances. Ainsi, il est une alternative Fiable lorsque la complexité des RNN n'est pas nécessaire. Parmi les modèles d'apprentissage automatique, XGBoost était le plus fiable, suivi par les forêts d'arbres décisionnels.

À l'avenir, il est essentiel de continuer à améliorer nos modèles et d'explorer des méthodes hybrides qui tirent parti des forces de différents algorithmes pour augmenter la précision des prédictions de la consommation d'eau. Les recherches futures devraient

se concentrer sur l'intégration de données plus diversifiées, y compris des données en temps réel, les données météorologiques, les données démographiques et les données économiques. en optimisant les modèles actuels et en explorant de nouvelles techniques, nous pouvons non seulement améliorer la précision des prédictions mais aussi fournir des outils plus robustes pour la gestion des ressources en eau. Ces améliorations contribueront à une meilleure planification et gestion des ressources en eau, cruciales pour répondre aux besoins croissants de la population de Tizi Ouzou.

LIST OF ABBREVIATIONS

AEP	**A**dduction d'**E**au **P**otable
AdaBoost	**A**daptive **B**oosting
ADE	**A**lgérienne **d**es **E**aux
ANN	**A**rtificial **N**eural **N**etwork
ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **A**verage
BiLSTM	**B**idirectional **L**ong **S**hort-**T**erm **M**emory
BP	**B**ack **P**ropagation
BPNN	**B**ackpropagation **N**eural **N**etwork
BRR	**B**ayesian **R**idge **R**egression
CENT	**C**entral
CNN	**C**onvolutional **N**eural **N**etwork
CQRCTN	**C**onvolutional-**BiLSTM FFNN F**eed-**f**orward **n**eural **n**etwork **Q**uantile **R**egress
D.G.S.N	**D**irection **G**énérale de la **S**ûreté **N**ationale
DL	**D**eep **L**earning
DMA	**D**istrict **M**etered **A**reas
DT	**D**ecision **T**ree
EDA	**E**xploratory **D**ata **A**nalysis
ELUs	**E**xponential **L**inear **U**nits
EPS	**E**xtrapolation **P**rediction **S**cenario
ETATCPT	**É**tat du compteur
Établissement/APC	**É**tablissement d'**A**limentation en Eau **P**otable et Assainissement des Eaux Usée
GA	**G**enetic **A**lgorithm

LIST OF ABBREVIATIONS

GBDT	**G**radient **B**oosting **D**ecision **T**ree
GCRF	**G**aussian **C**onditional **R**andom **F**ields
IPS	**I**nterpolation **P**rediction **S**cenario
KNN	**K**-**N**earest **N**eighbors
KRR	**K**ernel **R**idge **R**egression
LAG	**L**ag **F**eatures
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
LLC	**L**imited **L**iability **C**ompany
LR	**L**inear **R**egression
LSTM	**L**ong **S**hort-**T**erm **M**emory
MAE	**M**ean **A**bsolute **E**rror
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
MD	**M**anagement **D**istrict
ML	**M**achine **L**earning
MLP	**M**ultilayer **P**erceptron
MLR	**M**ultiple **L**inear **R**egression
M-LSTM	**M**ultivariate **L**ong **S**hort-**T**erm **M**emory
MRC	**M**ekong **R**iver **C**ommission
MSE	**M**ean **S**quared **E**rror
NAN	**N**ot **A** **N**umber
NUM	**N**umber
P.T.T	**P**ostes **T**élégraphes et **T**éléphones
PCA	**P**rincipal **C**omponent **A**nalysis
QR	**Q**uantile **R**egression
QGBRT	**Q**uantile **G**radient

LIST OF ABBREVIATIONS

Boosting Regression tree	**B**oosting **R**egression **T**ree
QMLP	**Q**uantile **M**ultilayer **P**erceptron
R^2	**C**oefficient of **D**etermination
ReLU	**R**ectified **L**inear **U**nit
RF	**R**andom **F**orest
RNN	**R**ecurrent **N**eural **N**etwork
RMSE	**R**oot **M**ean **S**quared **E**rror
SARIMA	**S**easonal **A**uto**R**egressive **I**ntegrated **M**oving **A**verage
SEP	**S**tandard **E**rror **P**rediction
SVR	**S**upport **V**ector **R**egression
SVM	**S**upport **V**ector **M**achine
SWAP	**S**mart **W**ater **A**nalytics **P**roject
Tanh	**h**yperbolic **t**angent
TYPABON	**T**ype d'abonnees
WDS	**W**ater **D**istribution **S**ystem
XGBoost	**E**xtreme **G**radient **B**oosting