ACEA Smart Water Analytics

Abderrahmane ZAYOUDI, Chaimaa EL KABBACH

abderrahmanezayoudi@gmail.com , elkabbach.chaimaa@gmail.com

Hassan 1er University, National School of applied Sciences-Berrechid-.

Supervisor: M. Hamid HRIMECH

Abstract*—*Water is an indispensable resource for human and economical welfare, and modern society depends on complex, interconnected infrastructures to provide safe water to consumers. Given this complexity, efficient numerical techniques are needed to support optimal control and management of water distribution systems. This document is intended to be a position paper on soft computing tools to suitably handle the huge amount of data generated by processes related to smart water applications.

In this paper, we develop an ensemble-learning based predictive-analytics framework for smart water management to predict the amount of water. In the predictive analytics, we first perform, Exploratory Data Analysis: for analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods, then we processed the Data; it used to transform the raw data in a useful and efficient format, feature engineering to select relevant features, and to achieve our goal, we tested several machine learning and deep learning models with different techniques for choosing the best one.

1- INTRODUCTION

The Acea Group is one of the largest Italian multiutility operators in the water services sector supplying 9 million inhabitants. It organized this competition on Kaggle to get help in predicting the water level in various types of water bodies over different seasons of the year. It is crucial for a water supply company to forecast the water level in a waterbody (***water spring, lake, river, or aquifer***) to handle daily consumption.

It can become tedious in scenarios like during fall and winter, water bodies are refilled, but during spring and summer, they start to drain. To help preserve the health of these water bodies it is important to predict the most efficient water availability, in terms of level and water flow for each day of the year.

The Acea Group provides data for 4 types of water bodies, namely, water spring, lake, river, and aquifer. While the primary intention is the same i.e., to predict water availability, the reality is that each waterbody has such unique characteristics that their attributes are not linked to each other. This problem uses datasets that are completely independent of each other. As each waterbody is different from the other, the related features are also different.

To determine how the features influence the water availability of each waterbody poses a challenge. So, we want to design a solution that helps better understand volumes so that we can ensure water availability for each time interval(day/month) of the year. It is of the utmost importance to notice that some features like rainfall and temperature, which are present in each dataset, don't go alongside the date. Indeed, both rainfall and temperature affect water level features. This means, for instance, that rain fell on 1st January doesn't affect the mentioned features right the same day but sometime later. As we don't know how many days/weeks/months later rainfall affects these features, this is another aspect to keep into consideration when analyzing the dataset.

2- Data Collection and Description

we will use nine different datasets, completely independent and not linked to each other. Each dataset can represent a different kind of waterbody. As each waterbody is different from the other, the related features as well are different from each other. So, if for instance we consider a water spring we notice that its features are different from the lake's one. This is correct and reflects the behavior and characteristics of each waterbody. The Acea Group deals with four different types of waterbodies: water spring (for which three datasets

are provided), lake (for which a dataset is provided), river (for which a dataset is provided) and aquifers (for which four datasets are provided).

A brief description of each one:

Auser__Aquifer__ This waterbody consists of two subsystems, called NORTH and SOUTH, where the former partly influences the behavior of the latter. Indeed, the north subsystem is a water table (or unconfined) aquifer while the south subsystem is an artesian (or confined) groundwater.

Petrignano_Aquifer__ The wells field of the alluvial plain between Ospedalicchio di Bastia Umbra and Petrignano is fed by three underground aquifers separated by low permeability septa. The aquifer can be considered a water table groundwater and is also fed by the Chiascio river. The groundwater levels are influenced by the following parameters: rainfall, depth to groundwater, temperatures and drainage volumes, level of the Chiascio river.

Petrignano__Aquifer__ The wells field of the alluvial plain between Ospedalicchio di Bastia Umbra and Petrignano is fed by three underground aquifers separated by low permeability septa. The aquifer can be considered a water table groundwater and is also fed by the Chiascio river. The groundwater levels are influenced by the following parameters: rainfall, depth to groundwater, temperatures and drainage volumes, level of the Chiascio river.

Luco__Aquifer__ The Luco wells field is fed by an underground aquifer. This aquifer not fed by rivers or lakes but by meteoric infiltration at the extremes of the impermeable sedimentary layers. Such aquifer is accessed through wells called Well 1, Well 3 and Well 4 and is influenced by the following parameters: rainfall, depth to groundwater, temperature, and drainage volumes.

Amiata_Water_spring__ The Amiata waterbody is composed of a volcanic aquifer not fed by rivers or lakes but fed by meteoric infiltration. This aquifer is accessed through Ermicciolo, Arbure, Bugnano and Galleria Alta water springs. The levels and volumes of the four sources are influenced by the parameters: rainfall, depth to groundwater, hydrometry, temperatures and drainage volumes.

Madonna di Canneto__Water spring__ The Madonna di Canneto spring is situated at an altitude of 1010m above sea level in the Canneto valley. It does not consist of an aquifer and its source is supplied by the water catchment area of the river Melfa.

Lupa__Water spring__ this water spring is in the Rosciano Valley, on the left side of the Nera River. The waters emerge at an altitude of about 375 meters above sea level through a long draining tunnel that crosses, in its final section, lithotypes and essentially calcareous rocks. It provides drinking water to the city of Terni and the towns around it.

Arno__River__ is the second largest river in peninsular Italy and the main waterway in Tuscany and it has a relatively torrential regime, due to the nature of the surrounding soils (marl and impermeable clays). Arno results to be the main source of water supply of the metropolitan area of Florence-Prato-Pistoia. The availability of water for this waterbody is evaluated by checking the hydrometric level of the river at the section of Nave di Rosano.

Bilancino__Lake is an artificial lake located in the municipality of Barberino di Mugello (about 50 km from Florence). It is used to refill the Arno River during the summer months. Indeed, during the winter months, the lake is filled up and then, during the summer months, the water of the lake is poured into the Arno River.

***NB*:** Each waterbody has its own different features to be predicted.

The most importance to notice that some features like rainfall and temperature, which are present in each dataset, don't go alongside the date. Indeed, both rainfall and temperature affect features like level, flow, depth to groundwater and hydrometry sometime after it fell. This means, for instance, that rain fell on 1st January doesn't affect the mentioned features right the same day but sometime later. As we don't know how many days/weeks/months later rainfall affects these features, this is another aspect to keep into consideration when analyzing the dataset.

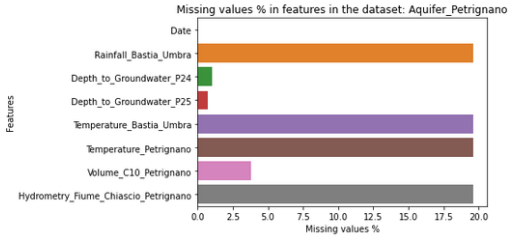
Our goal from those datasets is generate four mathematical models, one for each category of water body (aquifers, water springs, river, lake) that might be applicable to every single waterbody.

3- Exploratory data analytics

is an approach to analyzing data sets to summarize their main characteristics in easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis Exploratory data analysis was promoted by John Tukey to encourage statisticians to examine visually their data sets, to formulate hypotheses that could be tested on new datasets. These visualization capabilities also allowed statisticians to identify outliers, trends and patterns in data that merit further study.

In this paper, we used a few graphs to do the analysis of data, then we explained each graph to explore the data, to understand them well, in this paper, we will display some of them and their exploration.

For Aquifer Petrignago, the maximum number of missing values is in the features *Rainfall_Bastia_Umbra,Temperature_Bastia_Umbra*, *Temperature_Petrignano* and *Hydrometry_Fiume_Chiascio_Petrignano* with about 20% missing values and apart from the non-null features, the minimum number of missing values is in feature *Depth_to_Groundwater_P24*, *Depth_to_Groundwater_P25* with about 1% missing values.

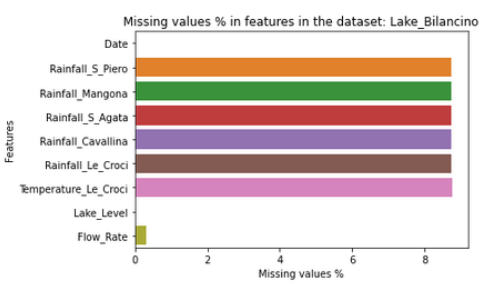


For Aquifer Doganella, the maximum number of missing values is in the Volume features with about 78% missing values and apart from the non-null features, the minimum number of missing values is in Rainfall features with about 10% missing values.

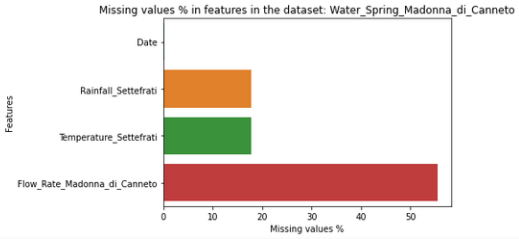
For Aquifer Auvser, the maximum number of missing values is in feature *Depth_to_Groundwater_DIEC* with about 60% missing values and apart from the non-null features, the minimum number of missing values is in feature Hydrometry_*Monte*_S_Quirico with about 11% missing values.

For Aquifer Luco, the maximum number of missing values is in the features of *Depth_to_Groundwater* and *Rainfall_Siena_Poggio_al_Vento* and *Rainfall_Ponte_Orgia* with about 80 to 85% missing values and apart from the non-null features, the minimum number of missing values is in feature *Rainfall_Simignano* with about 6% missing values.

For Lake Bilancino, the maximum number of missing values is in the Rainfall and Temperature features with about 85% missing values and apart from the non-null features, the minimum number of missing values is in feature *Flow_Rate* with about 0.2% missing values.

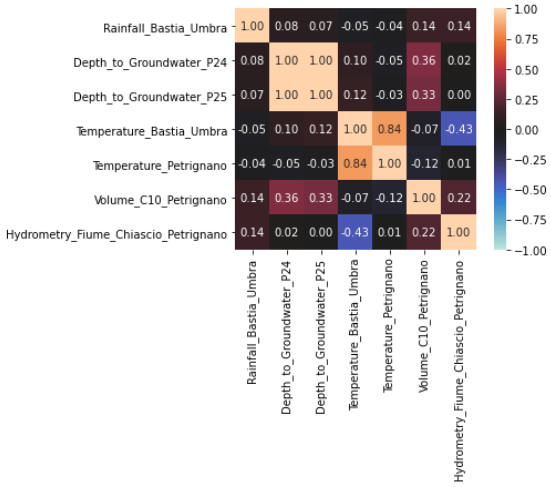


For Water Spring Madonna Di Canneto, the maximum number of missing values is in feature *Flow_Rate_Madonna_di_Canneto* with about 55% missing values and apart from the non-null features, the minimum number of missing values is in features *Rainfall_Settefrati* and *Temperature_Settefrati* with about 18% missing values.



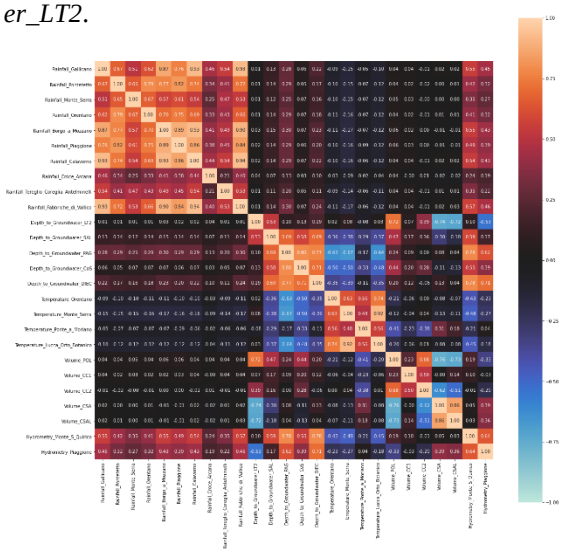
i) Correlation in features

For Aquifer Petrignago, positively strongly correlated features are *Depth_to_Groundwater_P24,Depth_to_Groundwater_P25*, *Temperature_Bastia_Umbra*, *Temperature_Petrignano* and negatively correlated features are Temperature_Bastia_Umbra, *Hydrometry_Fiume_Chiascio_Petrignano*.



For Aquifer Auser, positively strongly correlated features are *Rainfall_Gallicano*, *Rainfall_Borgo_a_Mozzano,Rainfall_Calavoro*, *Rainfall_Fabbriche_di_Vallico,Rainfall_Piaggione* and negatively strongly correlated features are *Volume_CSA*, *Volume_CSAL*, *Volume_POL*,

Depth_to_Groundwater_LT2,Depth_to_Groundwater_LT2.



ii) Target feature analysis

The Depth to groundwater of LT2 was the deepest in 2006 (earlier data is missing) at -15 m, then it is seen to be gradually decreasing over the years and around initial 2020 it hit the lowest for a brief period and then closed at around -12 meters.

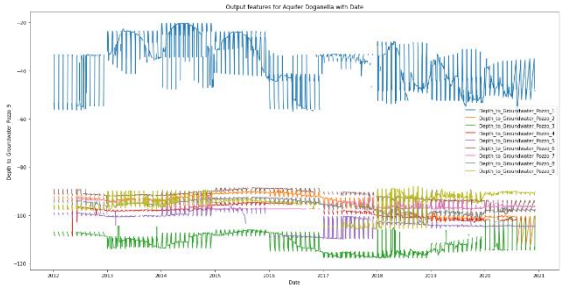
The Depth to groundwater of CoS lie somewhere in the range of -4 to -8 m throughout the years data and closes at -4 m.

The Depth to groundwater of SAL was the deepest in 2006 (earlier data is missing) at -7 m, subsequently, the range varies between -4 to -8 over the years and in 2020, it is recorded approximately -7 m.

The Depth to Groundwater Pozzo1 was observed on average around -40 m in 2012, after that it started decreasing in later years then improved back at closed at -40 m before 2021.

The Depth to Groundwater Pozzo3, Pozzo5 and Pozzo8 show remarkable improvement since the initial recordings in 2012. They close upto -110 m, -100 m and -95 m respectively before 2021.

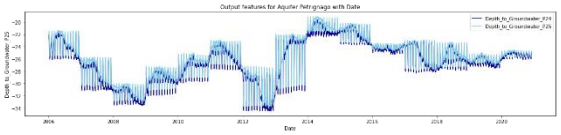
All other Depth to Groundwater features show similar values over the years.



The Depth to Groundwater in 2006 was at -24 m in 2006 and recorded around at the same after 2020.

Maximum depth to groundwater was seen in 2013 upto -34 m.

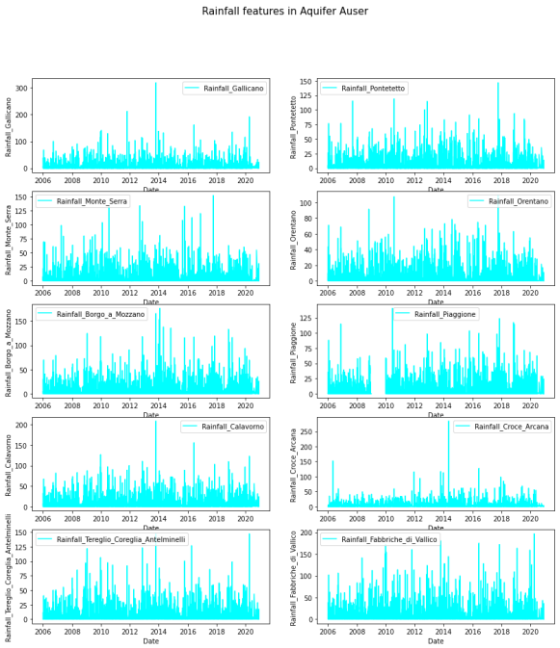
Minimum depth to groundwater was seen in 2015 as low as -20 m.



iii) Feature analysis with date

The most rainfall is received in the years 2008, 2010, 2014, 2018 and 2020.

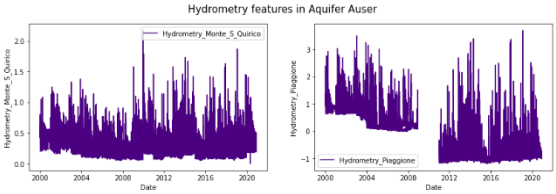
Maximum rainfall is received in Pontetetto, Tereglio Coreglia Antelminelli and Fabriche di Vallico area and the least in Orentano



An all-time high was recorded for Monte S Quirico hydrometric station in 2010 at 2 m.

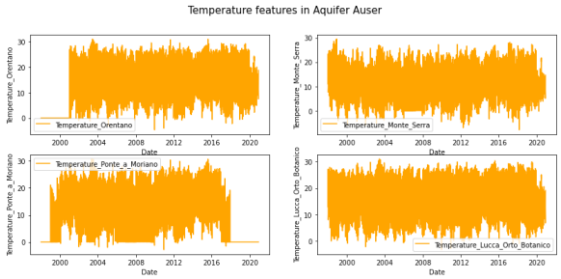
An all-time high was recorded for Piaggione hydrometric station in 2020 above 3 m.

Negative ground water level is also observed for Piaggione for 2012 to 2020 which is not so in case of Monte S Quirico.



Temperature varies between 0 to 30 degrees for all years.

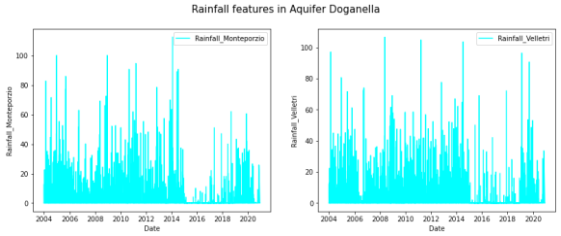
Monte Serra has more variations in temperature than other areas.



Low rainfall was recorded for initial 2020.

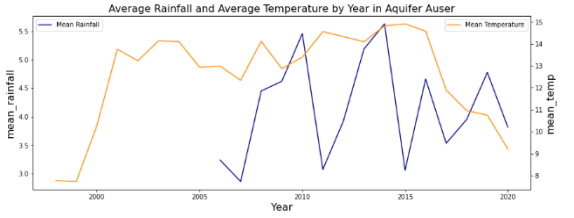
Maximum rainfall was above 100 mm in both areas in the years 2014 and 2008 respectively.

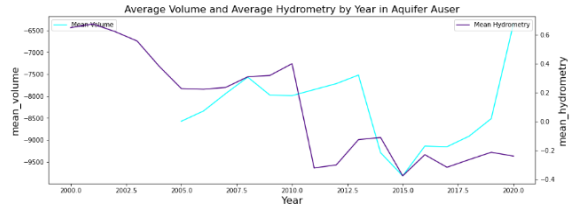
2016-17 was comparatively scanty for both areas.

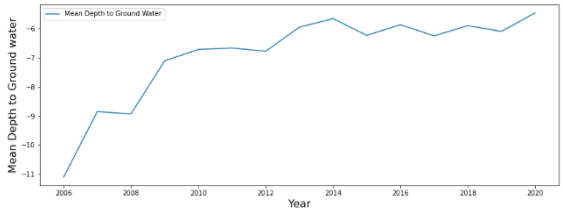


iv) Average yearly and monthly comparison of all features with target features

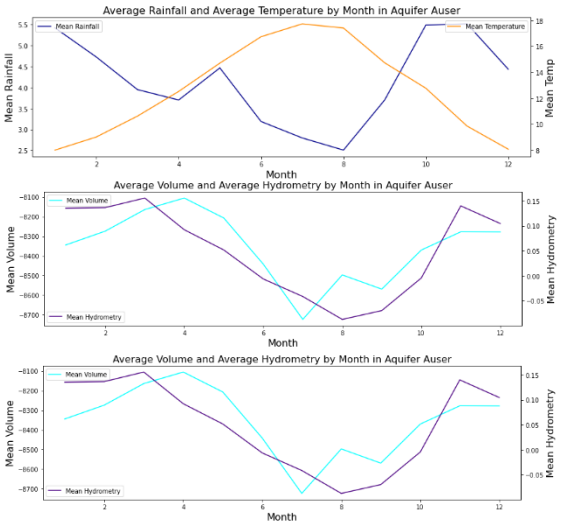
The mean depth to ground water per year keeps on dipping at drops to an average of -6 mm into 2020. This happens when mean rainfall and temp are 4 mm and 9 degrees respectively, which are moderate as compared to other years means. Also mean volume and mean hydrometry is -6500 cm and -0.2 m in 2020 which is very low.



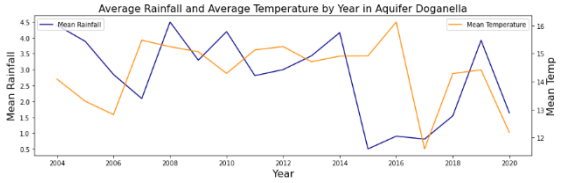


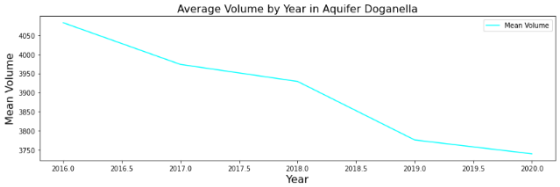


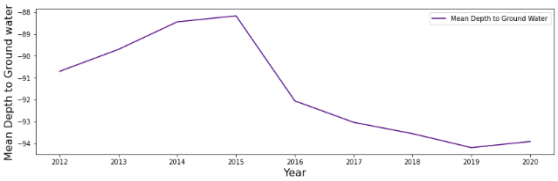
The maximum average rainfall is recorded in the month of October or November on an average. The maximum temperature is found in July or August. The mean volume is also highest at -8700 cm and mean hydrometry at -0.10 m in July and august. This is around the same time the average depth to groundwater is found at maximum at -6.7 m.



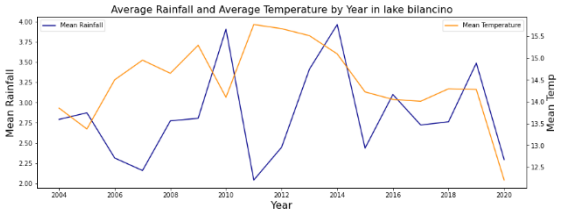
Mean Temperature and Rainfall are quite high for 2019. Mean Volume is about the lowest in 2019. So, the average depth to ground water is quite deep at -94 m in 2019 to 2020.

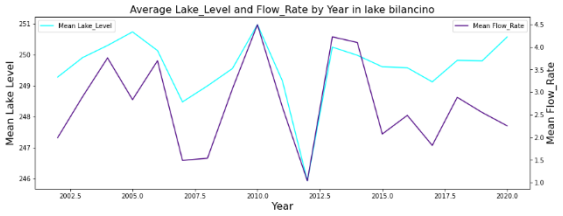




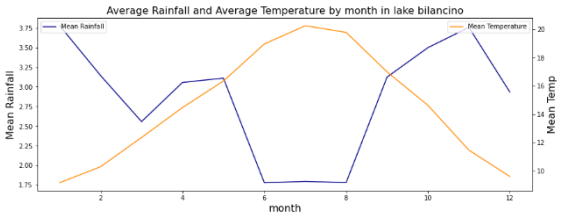


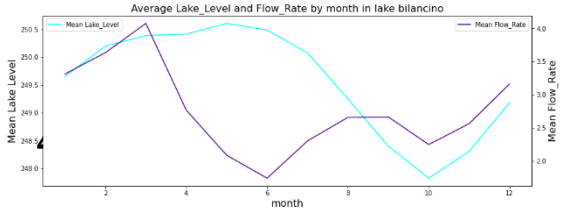
The patterns in mean lake level and mean flow rate are almost overlapping. Both seem to increase or decrease proportionally. Lake level, hydrometry is low when mean temperature is high, and rainfall is low like in 2012.





Average lake level and flow rate appear to be contradicting each other in the months of March to September and linearly similarly for rest of the months. This happens when there is very low rainfall and high temp from March to September resulting in high lake level and low flow rate.





5- Data preprocessing

in Machine Learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In simple words, data preprocessing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format.

In this section we will process the data in order to remove the anomalies found in EDA.

Our major issue is the missing values in all the 9 datasets. We will preprocess in three steps:

1- First step: choice of the metric.

To choose the best metrics that we will work with in the following parts, we calculate Mean Absolute Error scores, Median Absolute Error scores, MSE scores, RMSE scores, RMSLE scores and R2 scores for different k value in the KNN Imputer in one of our datasets. Than we compare them:

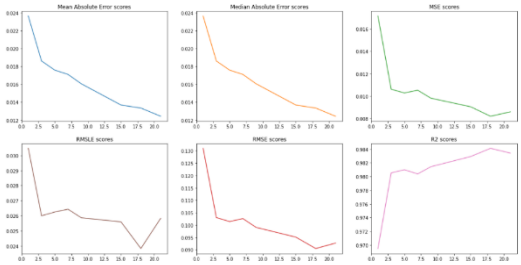


Figure: The performance of various metrics for different k values

We chose our metrics for the problem as Median Absolute Error (MAE), Root Mean Square Log Error (RMSLE) and R Squared (R^2) because of the following reasons:

- MAE is robust to outliers whereas RMSE is not. Using median is an extreme way of trimming extreme values. Hence median absolute error reduces the bias in favor of low forecasts. Also, MAE is really suited from an interpretation standpoint.
- RMSLE is used because the underestimation of the target variable is not acceptable, but overestimation can be tolerated. The RMLSE incurs a larger penalty for the underestimation of the actual value. Also, we don't want to penalize huge differences in the predicted and the actual values when both predicted and actual values are huge numbers. RMSLE metric (unlike RMSE) only considers the relative error between and the Predicted and the actual value and the scale of the error is not significant.
- High R^2 means that the correlation between observed and predicted values is high. It tells how good our regression model is as

compared to a very simple model that just predicts the mean value of target from the train set as predictions.

2- Second step:

We take only the rows where target features are not null. Impute the missing values in the independent features using KNN imputation with the selected least error k value (this is done by plotting k values with our chosen metrics MAE, RMSLE, and R2 score).

3- Third step:

We use the complete dataset (with filled features) to predict missing values in the respective target feature with Random Forest, Linear Regression, Decision Tree and KNN, with different values of k.

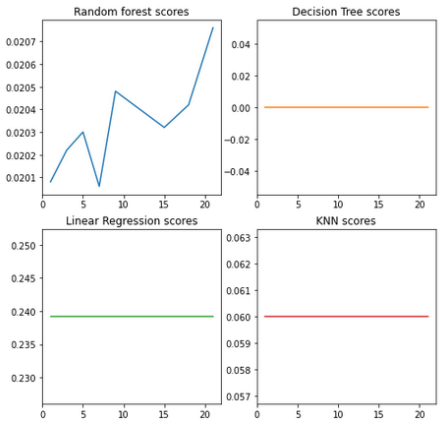


Figure: Various models MAE performance for different k values

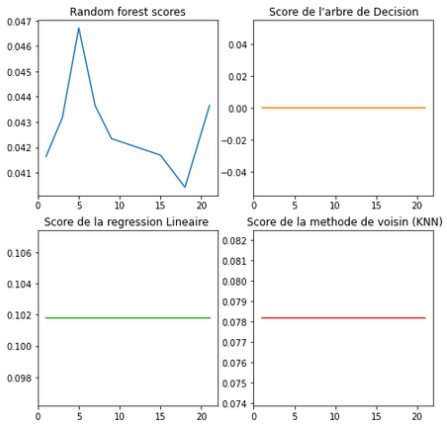


Figure: Various models RMSE performance for different k values

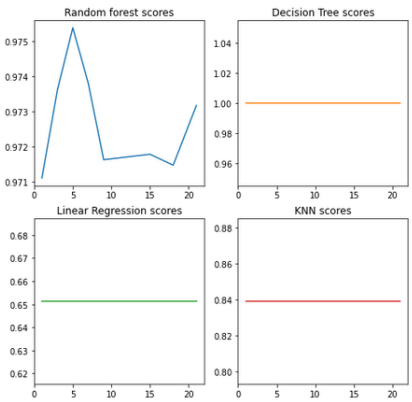


Figure: Various models R2 performance for different k values

Like we see in these figures, different plots for comparing these models error metrics, and the best one is Random Forest model.

After that we test the performance for different k values like we see in the figure below, and for that example the best one is the k = 18.

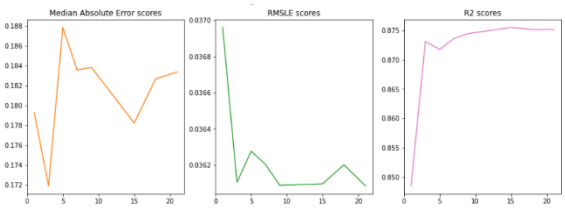


Figure: Various metrics performance for different k values.

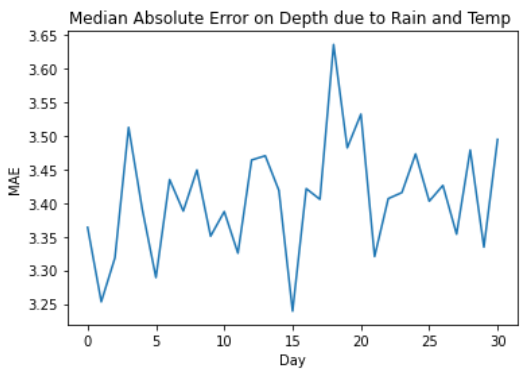
6- Feature Engineering

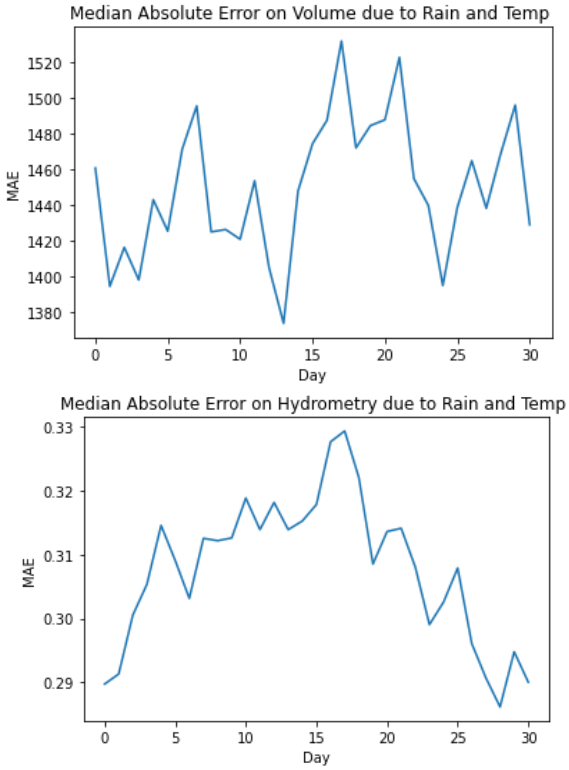
In this step, we have used different water bodies combined as one of each type - aquifer (4) - Lake (1) - Rivers (1) - Springs (3), we have taken average based on different time durations, i.e., Daily, Weekly, Monthly and Yearly.

	Date	Mean_Rainfall	Mean_Temp	Mean_Volume	Mean_Hydrometry	Mean_Depth	Source
0	05/03/1998	0.791111	0.0000	-8368.941841	0.058611	-5.87372	Auser
1	06/03/1998	2.957778	2.5125	-8267.066214	0.056944	-5.85732	Auser
2	07/03/1998	5.171111	4.8000	-7812.676449	-0.068611	-6.06028	Auser
3	08/03/1998	2.051111	6.3125	-7770.275060	-0.035556	-6.15872	Auser
4	09/03/1998	2.053333	6.0625	-7725.811777	-0.035556	-6.17936	Auser
...	...	...	...	...	...	...	...
8149	26/06/2020	0.000000	16.9250	-6698.827973	-0.415000	-4.61754	Auser
8150	27/06/2020	0.000000	17.3125	-6698.827973	-0.410000	-5.58138	Auser
8151	28/06/2020	0.000000	17.5750	-6363.886575	-0.400000	-5.88770	Auser
8152	29/06/2020	0.000000	16.2250	-6698.827973	-0.395000	-5.88498	Auser
8153	30/06/2020	0.000000	17.2250	-6698.827973	-0.420000	-5.78870	Auser

After this step, we taken the average, then we tried to out the median absolute trends in each of these duration models when the variables are shifted by durations (Week, Month, Year, Days).

We started by daily average, then we combined the 4 aquifer datasets to apply the shifts to the combined data, for example, in the case that we have daily, we shifted the data by combination of 1 to 31 days, our goal is the least error for the model tha we used by the random forest, i.e., that number of days shows the real effect of rainfall and temperature.





We observed that the least error i.e., the actual effect of rainfall and temperature is observed when the values are recorded after 26 days. So, we will shift the values by 26 days (the same method to the other durations).

As a conclusion of observing best results at 26 days, 8 weeks, and 2 months, let's have the final aquifers dataset by shifting the values by 56 days (8 weeks or 2 months) backward.

Date	Mean_Rainfall	Mean_Temp	Mean_Volume	Mean_Hydrometry	Mean_Depth	Actual_Depth	Actual_Volume	Actual_Hydrometry
1998-01-04	0.415556	6.625000	-7850.390763	-0.114444	-6.113340	-6.080260	-8019.271158	-0.083056
1998-01-05	2.054444	6.075000	-7730.606098	-0.032778	-6.178100	-6.064520	-7956.571285	-0.104167
1998-01-06	0.921111	9.087500	-7509.802601	-0.224167	-6.083020	-6.157060	-7715.808854	0.011944
1998-01-07	0.878889	12.325000	-8127.271730	-0.214722	-6.098480	-6.107740	-7731.378766	-0.008611
1998-01-08	0.908889	12.650000	-8243.315107	-0.256389	-6.115560	-6.053100	-7812.676449	-0.072222
...	...	...	...	...	...	...	...	...
2020-12-02	0.005000	9.333125	-7857.712792	1.065000	-33.599629	-33.875983	-8248.071150	1.185000
2020-12-03	0.200000	9.357500	-7238.044134	1.297500	-33.864259	-33.875983	-8248.071150	1.185000
2020-12-04	0.000000	13.914375	-7812.040004	1.057500	-33.798613	-33.875983	-8248.071150	1.185000
2020-12-05	5.740000	16.316875	-6006.015311	1.080000	-33.837573	-33.875983	-8248.071150	1.185000
2020-12-06	0.170000	16.560000	-8248.071150	1.185000	-33.875983	-33.875983	-8248.071150	1.185000

7- Models

we used several algorithms to find the best model, the algorithms used are as follows:

KNN: K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

Multiple Linear Regression: we used this algorithm to estimate the relationship between two or more independent variables and one dependent variable.

Random Forest: The random forest (RF) is a representative supervised learning technique aimed at classification and regression, and it is an ensemble learning method for optimal decision making based on the results of multiple decision trees [28]. The advantage of RF is that the calculation speed is fast, and the accuracy of the predicted results is high. The RF presents the final decision based on a combination of the results estimated from multiple decision trees, leading to high reliability of the obtained results and high model stability

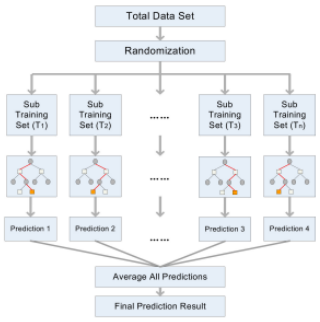


Figure 8. Conceptual diagram of random forest (RF) model.

SGDRegression: is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions.

SGD stands for Stochastic Gradient Descent: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate).

Decision Tree Regressor: Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

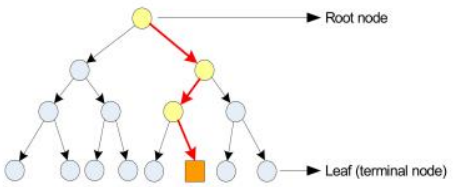


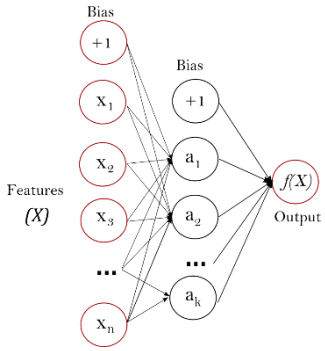
Figure 7. Conceptual diagram of decision tree (DT) model.

XGBoost Regressor: stands for "Extreme Gradient Boosting" and it is an implementation of gradient boosting trees algorithm. The XGBoost is a popular supervised machine learning model with

characteristics like computation speed, parallelization, and performance.

AdaBoost Regressor: is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

MLP Regressor: MLP (Multi-layer Perceptron) is a type of artificial neural network (ANN). Simplest MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer.



here is the comparison table between the performance of the models that are created by different algorithms.

Model	Aquifers	Lakes	Rivers	Springs
knn	0.96	-17.851,-1.254	-5.839	-0.147
Linear Regression	-0.361	-156.95,-0.821	-11.11	-2.141
Random Forest	0.966	-25.89,-0.684	-9.331	0.355
SGRegression	-0.171	-0.781	-11.21	-1.11
Decision Tree	0.956	-0.091	-7.012	0.211
XGBoost	0.958	-0.325	-11.698	0.348
AdaBoost	0.964	-0.053	-4.11	0.309
MLP	0.958	0.076	0.106	0.181

Figure: R2 score with different models

8- Conclusion

In this paper, we trained prediction models for smart water to predict the water level in different waterbodies. After performing data preprocessing and leveling the accuracy between different features of dataset, we tested different models for each waterbody, and chose the best one.

Various types of Machine Learning algorithms like Random Forest, Linear Regression, Decision tree and MLP have been explored for solving the problem. Maybe the next approach would be exploring more neural networks like CNNs, or LSTMs can yield good results. Also, linear statistical models like auto-regressive moving average (ARMA), and auto-regressive exogenous (ARX) can also be tested to capture the seasonal trend in the datasets.

9- References

1. https://www.researchgate.net/publication/223259105_Short-term_water_level_prediction_using_neural_networks_and_neuro-fuzzy_approach
2. https://www.researchgate.net/publication/328086006_Short-term_water_demand_forecasting_using_machine_learning_techniques
3. https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/
4. https://www.geeksforgeeks.org/python-pandas-dataframe-ffill/
5. https://www.kaggle.com/c/acea-water-prediction