Paper Nr
371

INTERNATIONAL CONFERENCE

GROUNDWATER, KEY TO THE SUSTAINABLE DEVELOPMENT GOALS

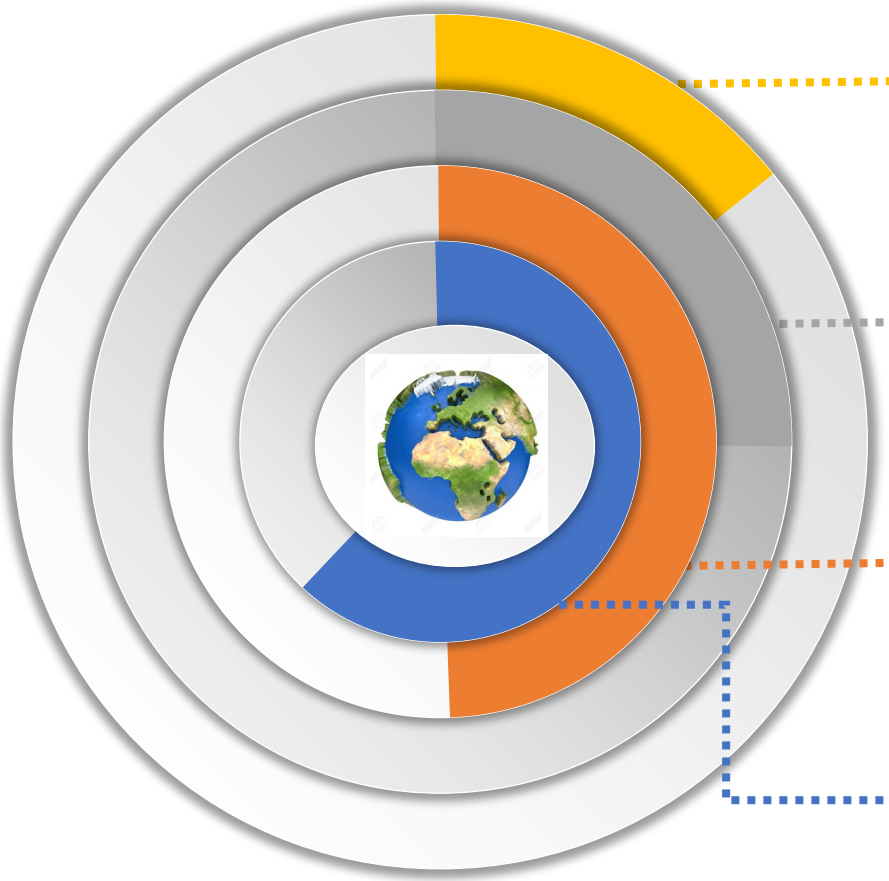
PARIS - May 18 -20, 2022

# Exploring Machine Learning Models in Predicting Irrigation Groundwater Quality Indices for Effective Decision Making in Medjerda River Basin, Tunisia

**Dr. Fatma TRABELSI, Ing. Salsebil BEL HADJ ALI**

**Higher School of Engineers of Medjez El Bab, University of Jendouba**

# OUTLINE

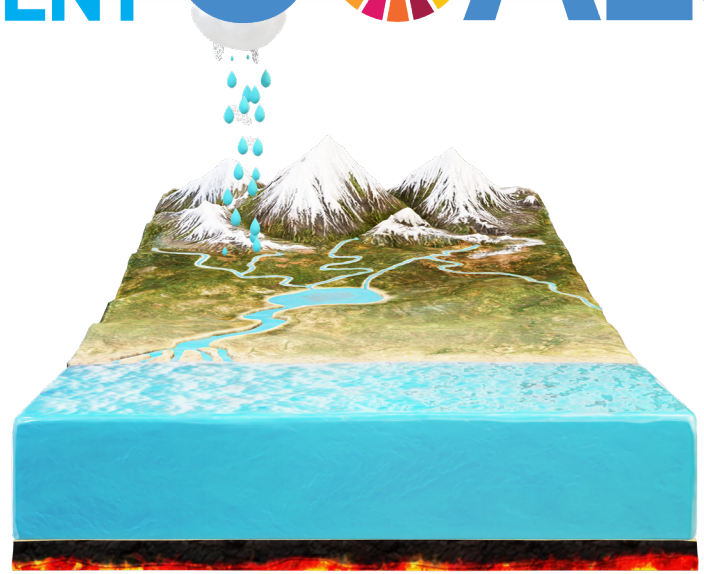


- STUDY CONTEXT
- METHODOLOGY AND DATASETS
- RESULTS
- DISCUSSION & CONCLUSION

# BACKGROUND

## SDGs 6 :Clean Water and Sanitation

***To achieve sustainable development goals (SDGs 6) related to the efficient use of water, it is crucial to improve water management, rationalize the water irrigation uses and improve the tools of groundwater quality assessment.***



- The assessment of groundwater quality require a sampling protocol, laboratory analysis, and at a larger scale, testing and data management which increase the cost and study time of water quality assessment and affects the decision-making on water quality management planning.

- To cope with these issues, it is crucial to develop a powerful and cost-effective approach for quick and accurate assessment of irrigation water quality.

☞ ***Machine Learning models have opted for non-physical tool that successfully predicting groundwater quality***

# STUDY CONTEXT

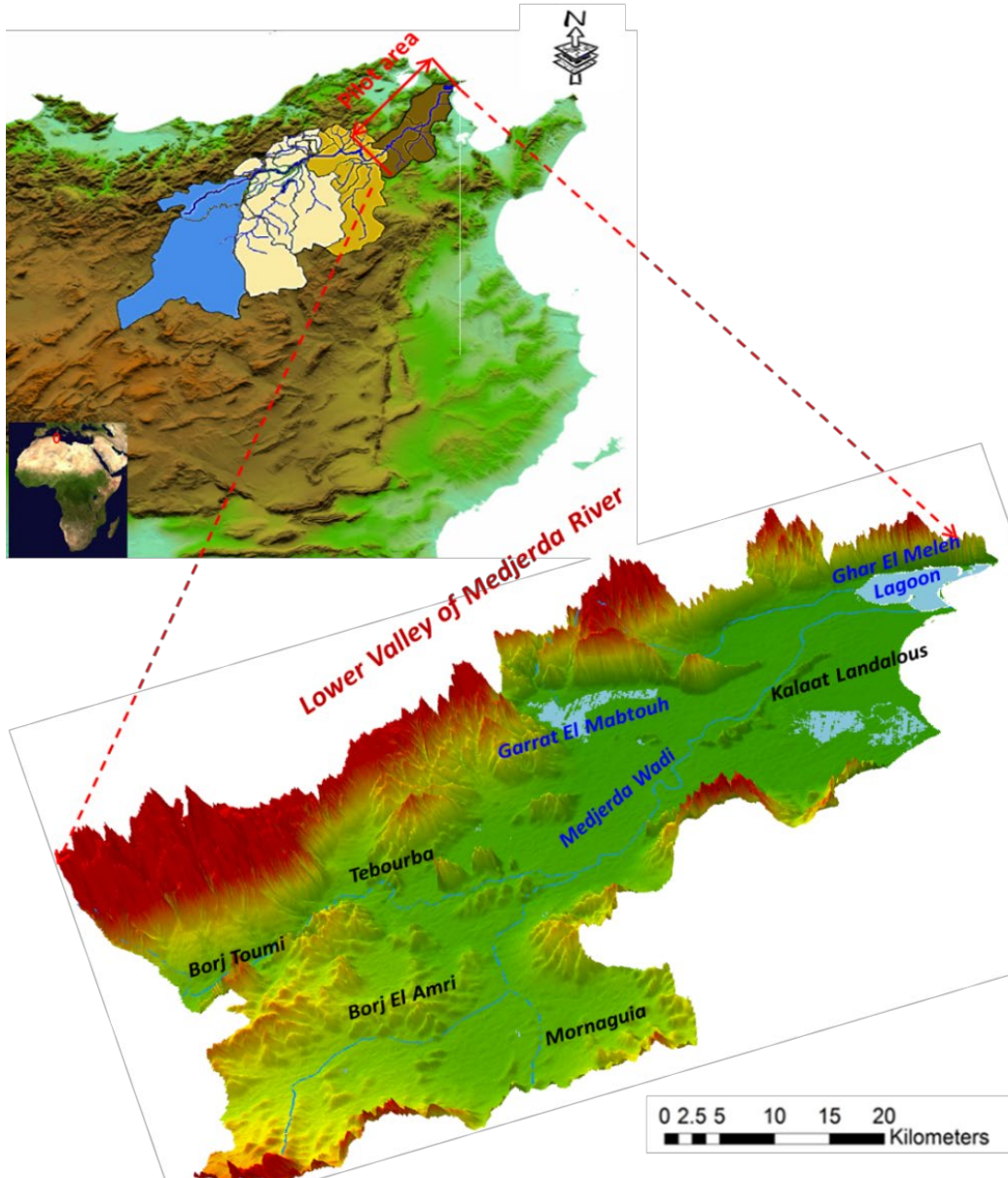
**The Lower valley (LV) of Medjerda basin is part of the largest watershed of Tunisia, where it supplies about half of the country's drinking water.**

It is essentially agricultural, where irrigation water supply depends **on surface water in conjunction with groundwater resources**.

Despite the importance of groundwater in the Medjerda basin, there is currently **a huge lack of data regarding its quality** that undermines the ability of decision makers and users to manage it properly.

Few groundwater sampling campaigns and analyses were conducted, and they are **therefore insufficient to fill the existing data gap and to give a real time information** about suitability of groundwater use.

Thus, improving the water quality evaluation process based on **non-cost data** using an objective tool with flexibility in its decision-making capacity for water management and planning is essential in the LV of Medjerda basin.



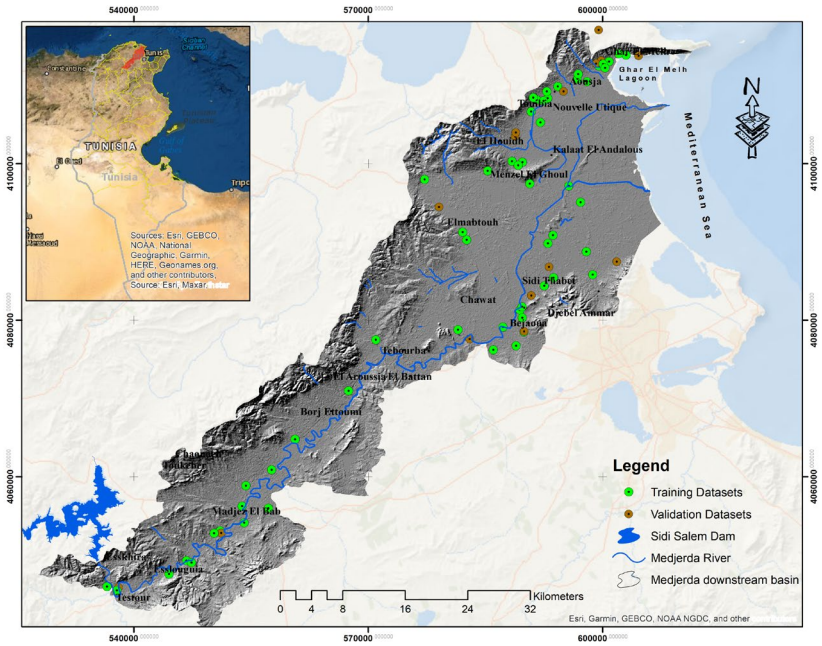
# GOALS

*The focus of this study was to test the performance of the novel approach to predict the suitability of groundwater quality for irrigation purposes useful to support decision-making under uncertainty in water resource management.*

- to evaluate the effectiveness of machine learning (ML) models to predict **the irrigation water quality parameters (IWQ): TDS, PS, SAR, ESP, and MAR** using physico-chemical parameters as input variables

- to evaluate the accuracy of the implemented models;

- to analyse the uncertainty and sensitivity of the tested models



# METHODOLOGY

**INPUT DATA**
**Physico-chemical parameters**
**Irrigation Water Quality Indices (IWQI)**

## Data preparation

Data checking reliability

### Cleaning Datasets

*1008 variables (14 columns & 72 lines)*

- Statistical analysis and imputation
  *Solve for missing values*
  *Delete inherent values ( i.e. pH= 0.5, T°= 88)*
- Reliability check of the data using Ionic balance
  *(>5% rejected)*

### Evaluation of correlation index r

*Scatterplot of sum cations vs sum anions*

Data Exploration

### Irrigation water quality indices

Basic statistical characteristics
Matrix correlation analysis

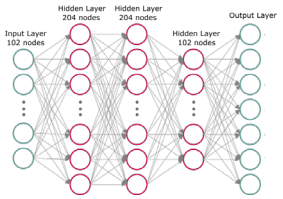
Data Normalization

$x\ normalized = (x - x\ min) / (x\ max - x\ min)$

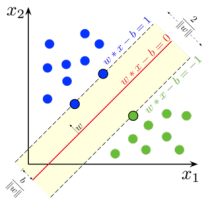
## Machine Learning Models

**Training 80%**

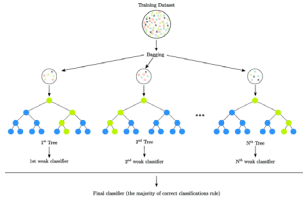
### Artificial Neural Network (ANN)



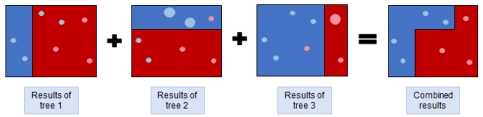

### Support Vector Machine (SVM)




### Random Forest (RF)

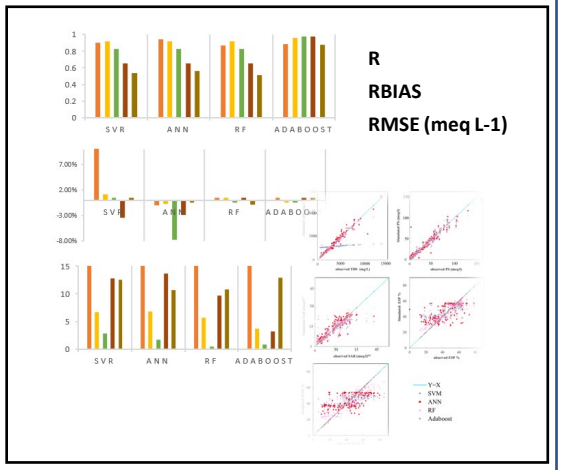


### Adaptive Boosting (Adaboost)

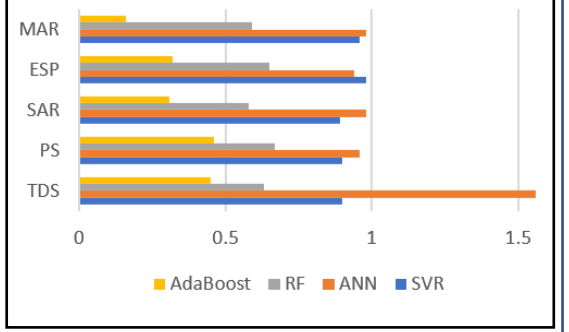



## Metric Validation

**Validation 20%**




## Generalization Ability

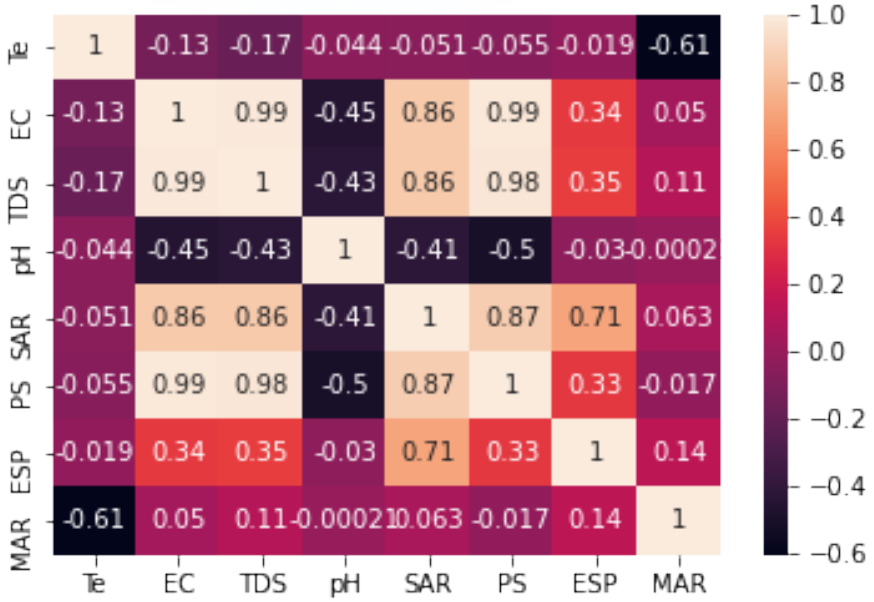



## Uncertainty and sensitivity Analysis

| Parameter | Error | SVR | ANN | RF | AdaBoost |
|---|---|---|---|---|---|
| TDS (mg $L^{-1}$) | E | 412.4 | -27.01 | 4.79 | 11.57 |
| | CB (95%) | 142.5 | 55.07 | 50.65 | 27.55 |
| PS (meq $L^{-1}$) | E | 0.45 | -0.27 | 0.21 | -0.09 |
| | CB (95%) | 0.96 | 1 | 0.97 | 0.91 |
| SAR ($meq^{0.5} L^{-0.5}$) | E | 0.04 | -0.36 | -0.01 | -0.02 |
| | CB (95%) | 0.37 | 0.47 | 0.09 | 0.04 |
| ESP (%) | E | -1.45 | -1.31 | 0.13 | 0.56 |
| | CB (95%) | 1.89 | 1.69 | 1.14 | 0.74 |
| MAR (%) | E | 0.27 | -0.05 | -0.02 | 0.19 |
| | CB (96%) | 2.47 | 2.01 | 1.47 | 0.69 |

# RESULTS

## 1. Statistical Analysis

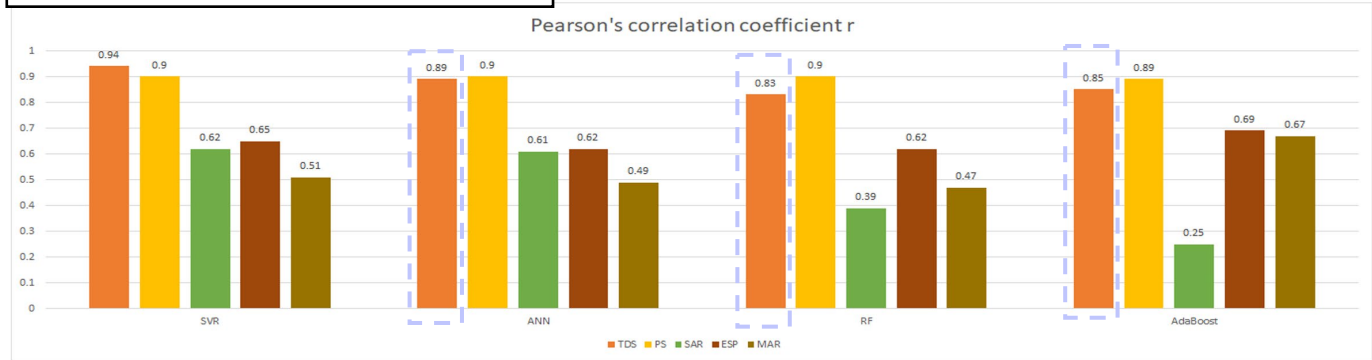


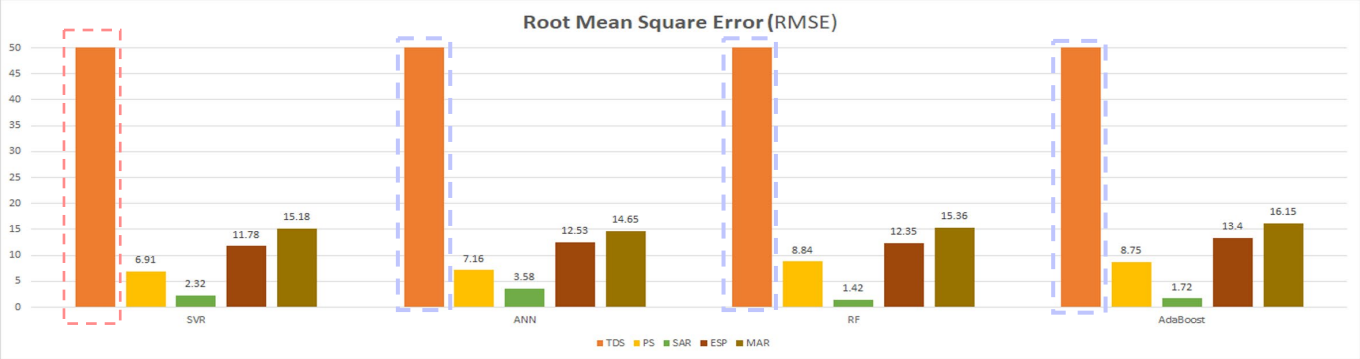

- Electrical conductivity (EC) is a **more correlated input variable** with the predicted parameters than pH and temperature.
- The lowest correlations between T, pH, and EC prove that these parameters are **separable and non-redundant** and, therefore, are useful for improving the predictive accuracy of machine learning

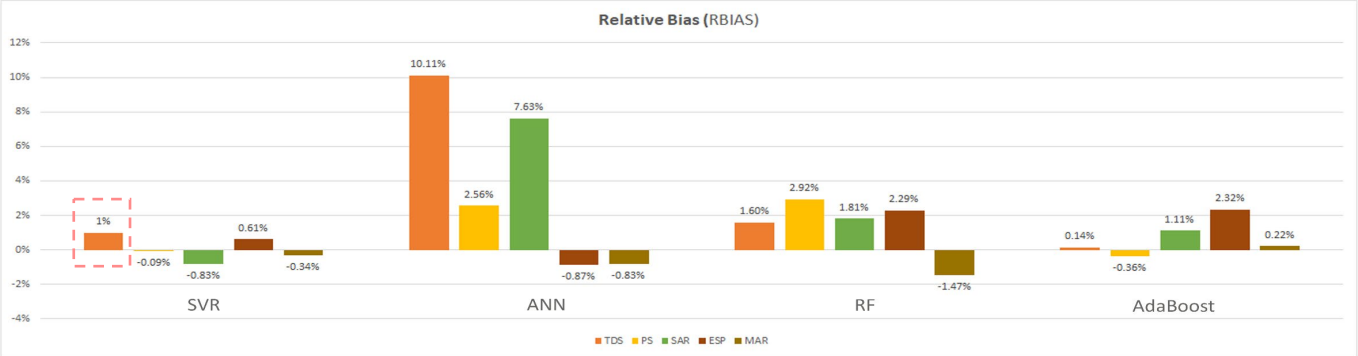
# 2. Implementation and Evaluation of Model

## RESULTS

**Training Model *Performance***



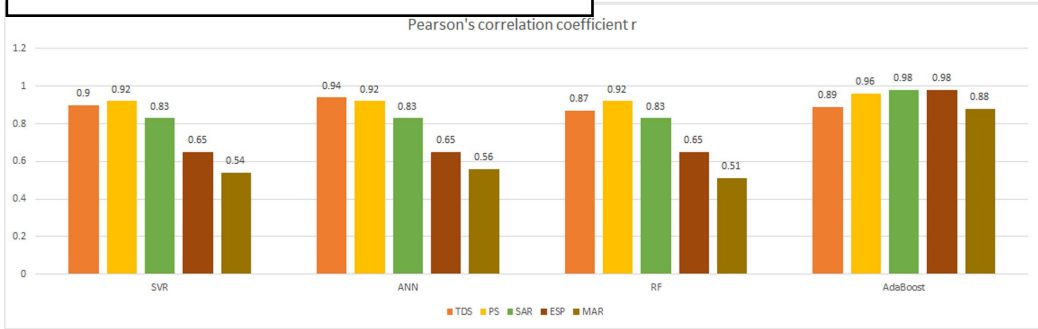



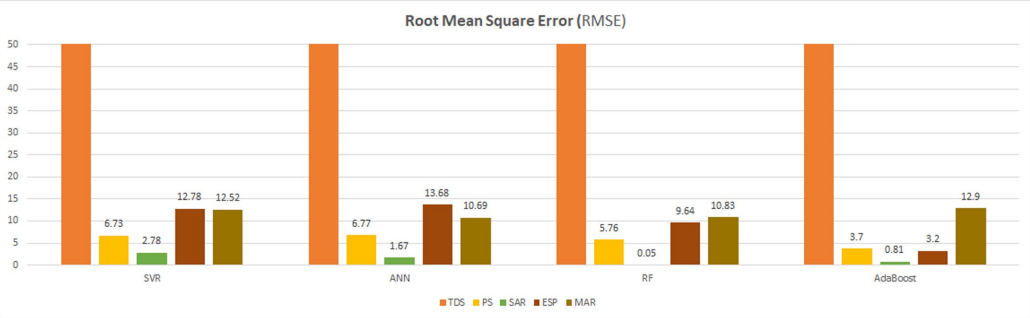

- SVR model has significant values of RBIAS and RMSE compared with the other models for predicting the TDS parameter
- The ANN, RF, and AdaBoost models revealed high accuracy in predicting the TDS parameter during the learning process
- The AdaBoost model had a good performance in predicting all the IWQs parameters.
- The four models perform satisfactory for the prediction of the sodium absorption ratio (SAR) and the percent exchangeable sodium (ESP)
- Random Forest and artificial neural network models were unable to predict the MAR parameter

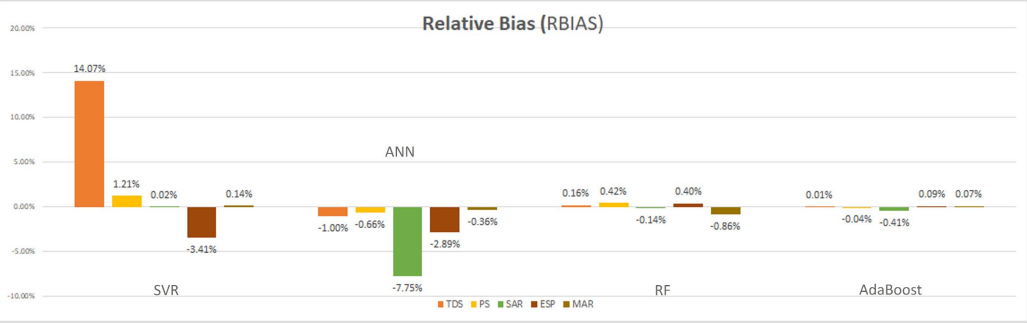
# 2. Implementation and Evaluation of Model

## RESULTS

**Validation Model Performance**



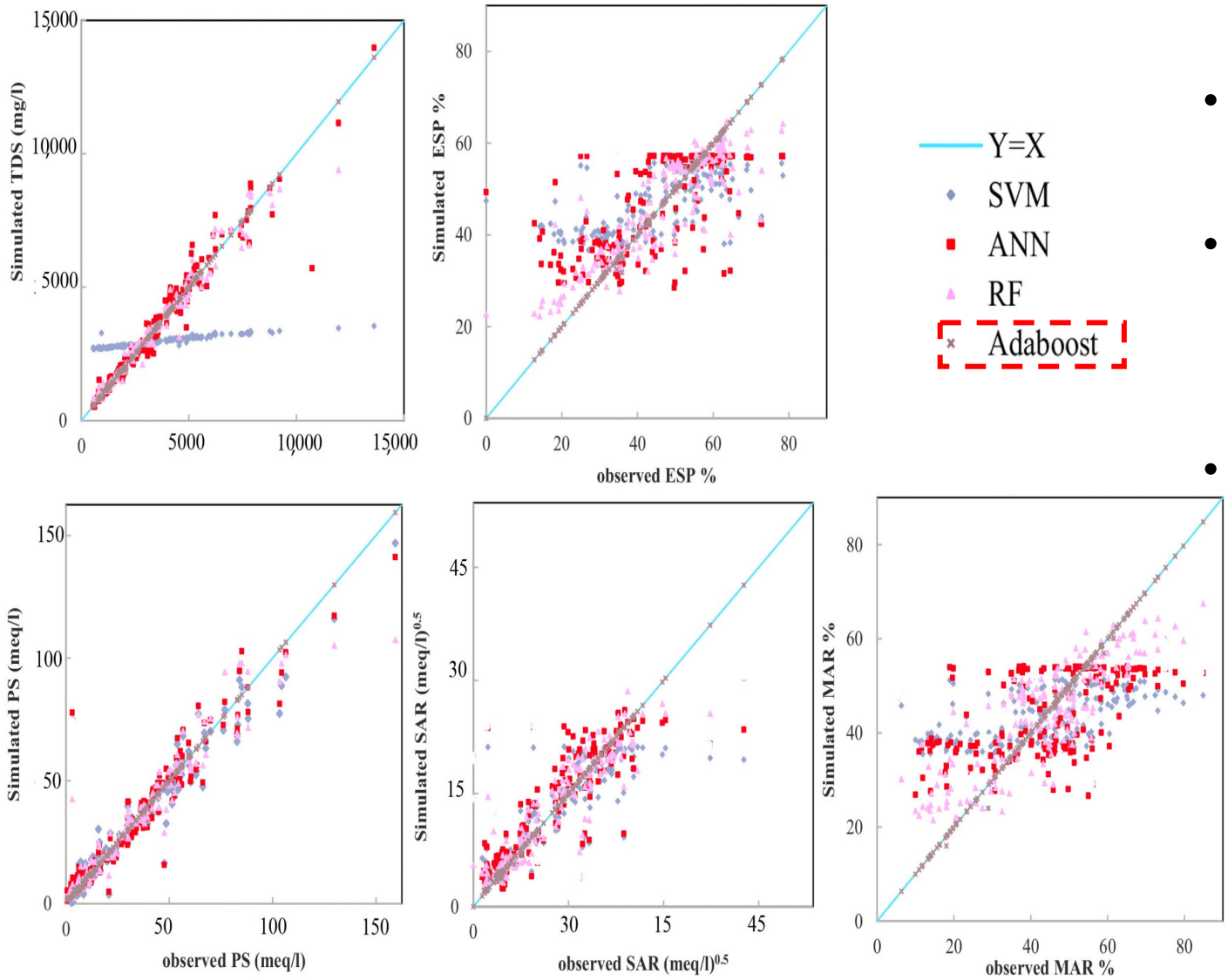




- The Pearson's coefficient values range from 0.65 to 0.94 for TDS, PS, SAR, and ESP over ANN and SVR models.
- RMSE showed an unacceptable performance for all models for the simulation of the TDS and MAR parameters.
- RBIAS showed a lowest performance for the SVR model for the simulation of the TDS and MAR parameters

ANN and SVR models presented very close results during the two processes for the prediction of all IWQs parameters

# 3. Relationship between observed & simulated variables over IWQs parameters

RESULTS



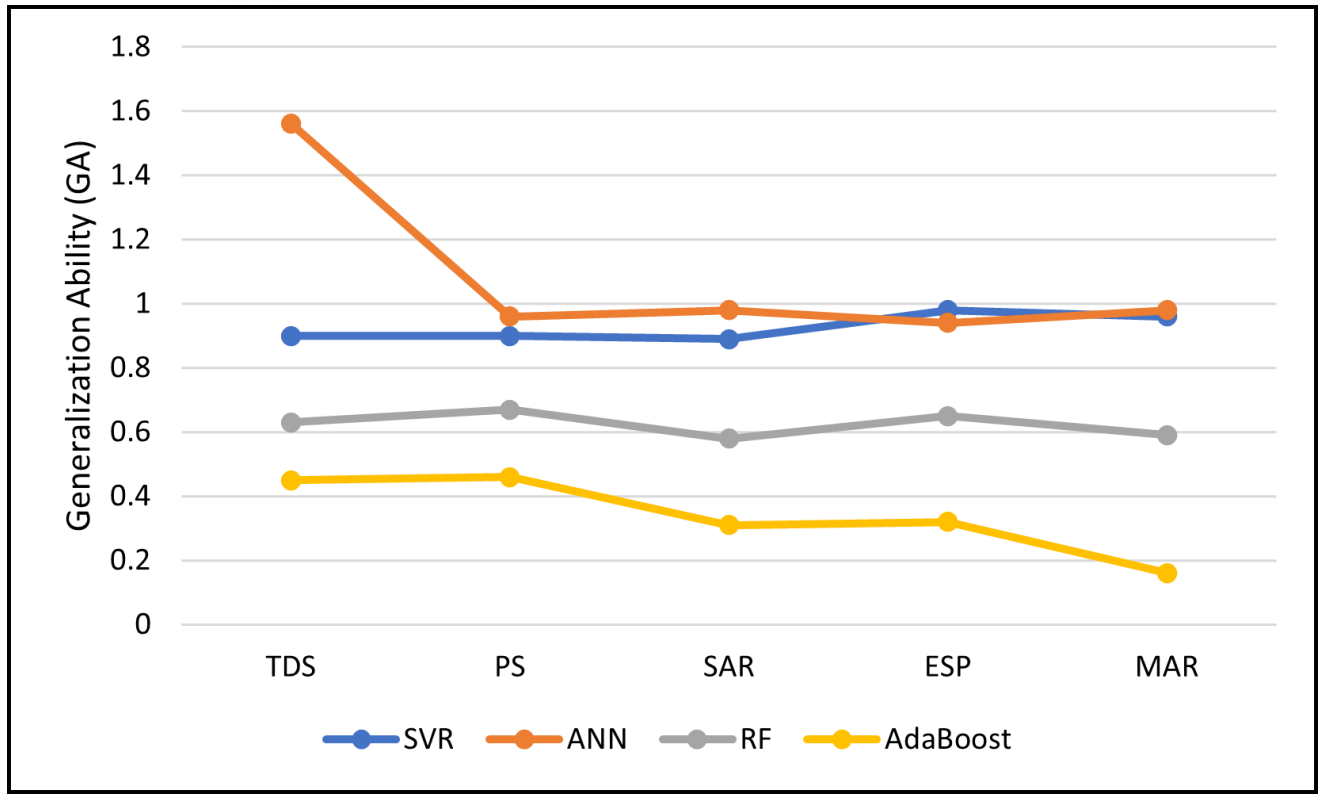

- It identifies a better distribution on the X = Y line for the random forest for all models
- SVR model has the weakest performance in predicting PS and SAR parameter followed by ANN and the RF in predicting TDS, PS, and SAR
- The predicted values are very close to the observed values for the AdaBoost model except for the MAR parameter

AdaBoost model has the best performance in predicting all IWQ parameters

# 4.Generalization ability (GA) indices of the models



To have useful models to predict new data sets, while avoiding errors, it is necessary to test its generalization capability.




- The ANN model for TDS is overfitted while for all other models are underfitted.
- The GA indices of the random forest and AdaBoost models are weaker than the ANN and SVR models.

## 5. Model Uncertainty Analysis

| Parameter | Error | SVR | ANN | RF | AdaBoost |
|---|---|---|---|---|---|
| TDS (mg $L^{-1}$) | E | 412.48 | -27.01 | 4.79 | 11.57 |
| | CB (95%) | 142.56 | 55.07 | 50.65 | 27.55 |
| PS (meq $L^{-1}$) | E | 0.45 | -0.27 | 0.21 | -0.09 |
| | CB (95%) | 0.96 | 1.00 | 0.97 | 0.91 |
| SAR ($meq^{0.5}$ $L^{-0.5}$) | E | 0.04 | -0.36 | -0.01 | -0.02 |
| | CB (95%) | 0.37 | 0.47 | 0.09 | 0.04 |
| ESP (%) | E | -1.45 | -1.31 | 0.13 | 0.56 |
| | CB (95%) | 1.89 | 1.69 | 1.14 | 0.74 |
| MAR (%) | E | 0.27 | -0.05 | -0.02 | 0.19 |
| | CB (96%) | 2.47 | 2.01 | 1.47 | 0.69 |

**SVR model has the highest confidence bound values (95%) , followed by the ANN, RF, and AdaBoost models**

# 6. Sensitivity analysis results

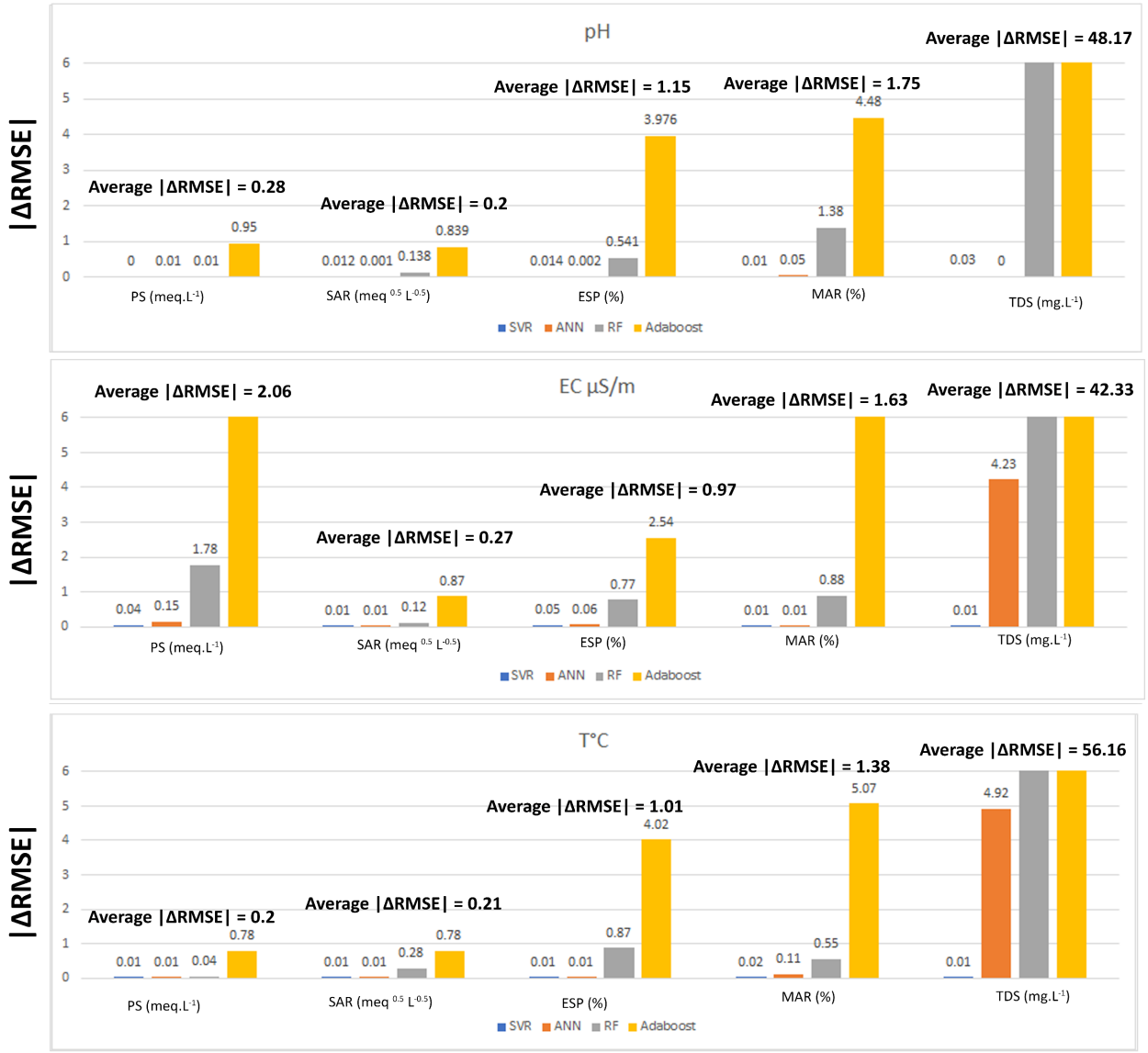
RESULTS

**The sensitivity of the model provides an overview of the impact of input variables on the output. This analysis is necessary to assess how the model acts according to shifts in input values.**

The models are more sensitive to:

- EC followed by temperature and pH, respectively for predicting TDS and MAR;
- pH for predicting ESP parameter;
- EC followed by the pH and the temperature, respectively for predicting PS and SAR.

**AdaBoost model is the most sensitive model**



# CONCLUSION

- This research presents an effective use of machine learning models in forecasting the irrigation groundwater quality indices through low-cost data; using only physicochemical parameters as input variables without decreasing the efficiency of the models.

- These findings can be used as decision support systems (DSS) tool for sustainable water management in LV of Medjerda basin.

- The traditional simulation modelling approaches are dependent on datasets that involve a large amount of unknown or unspecified input data and generally consist of high-cost time-consuming processes.

- Therefore, setting up a DSS based on machine learning models will boost the efficient use of water and rationalize its use by all water stakeholders at watershed level.

*sustainability*

MDPI

*Article*

# Exploring Machine Learning Models in Predicting Irrigation Groundwater Quality Indices for Effective Decision Making in Medjerda River Basin, Tunisia

**Fatma Trabelsi * and Salsebil Bel Hadj Ali**

Research Unit Sustainable Management of Water and Soil Resources, Higher School of Engineers of Medjez El Bab (ESIM), University of Jendouba, Jendouba 8189, Tunisia; belhadjali.salsebil@esim.u-jendouba.tn

* Correspondence: fatma.trabelsi@esim.u-jendouba.tn

Sponsors : SMART IWRM Medjerda "Improving Sustainable Groundwater Management of the Lower valley of Medjerda basin" is the Research & Development project funded by the PEER cycle 7 program (NAS_USAID) and led by the Higher School of Engineers of Medjez El Bab (ESIM) and the United States Geological Survey (USGS).

Thanks for your attention

Dr. Fatma TRABELSI : Associate Professor of Higher Education ,UR-Sustainable Management of Water and Soil Resources, Higher School of Engineers of Medjez El Bab (ESIM)
trabelsifatma@gmail.com