



Utilizing Data Mining Algorithms to Predict Groundwater Levels: An Applied Study on the Rising Groundwater Issue in Zliten, Libya

Abdalkarim .A.Erhab⁽¹⁾ , Abd El-Salam M.Gnedela⁽²⁾ , Rabieaa .A. Jaballa⁽³⁾

⁽¹⁾computer Department, Higher Institute for Science and Technology Algarabelli ,Libya

⁽²⁾ Statistics Department, College of Science, Al-asmarya University ,Libya

⁽³⁾computer Department, Higher Institute for Science and Technology Algarabelli ,Libya

Received: 15-12-2025; Revised: 22-12-2025; Accepted: 2-1-2026; Published: 11-1-2026

1. Abstract:

Among the most valuable environmental challenges facing the city of Zliten, Libya, which is depending totally on groundwater resources, is the abnormal rise in groundwater level. The study will try to use data mining techniques for extracting an accurate predictive model for groundwater levels based on monthly observations from 31 monitored wells collected over a continuous period of ten months. In this context, two machine learning-based algorithms for prediction were applied: Linear Regression and Decision Tree Regression, using the specialized analytic environment RapidMiner.

These results show that, based on the key statistical performance measures of RMSE, MAE, and R², the Decision Tree algorithm obviously outperformed the Linear Regression model and thus can be effective in forecasting future groundwater levels. Such findings bring into focus the need to adopt artificial intelligence tools to support the management of groundwater resources and environmental sustainability in the face of fluctuating and unstable water supplies.

Keywords:

Data Mining; Groundwater Level; Decision Tree; Linear Regression; RapidMiner; Prediction.

2. Introduction

Numerous cities in coastal and arid areas like Libya Zliten are facing complicated environmental dilemmas concerning water resources management. These dilemmas are not confined to water shortage. In certain regions, an abnormal rise in groundwater levels has emerged, ultimately creating serious challenges for

urban systems, stagnant water bodies, and pollution of potable drinking water. Unlike drought, which easily attracts attention, rising groundwater levels are inconspicuous and as such need accurate prediction because it is an incredibly useful planning mechanism and a way to mitigate chronic environmental risks.[1] In response, recent research has been emphasizing more the utilization of artificial intelligence and data mining techniques to construct predictive models that can predict future changes in groundwater levels from empirical observation data aggregated over large stretches of continuous time. Data mining offers a salient alternative to traditional statistical methodologies, especially when faced with complex, high-dimensional, and nonlinear data sets.[2]

Linear Regression, a classical and easy predictive algorithm, finds a linear relationship between independent feature variables and a dependent feature variable.[3] Decision Tree Regression, a powerful algorithm, splits data based on conditions on attributes. The ability to partition data provides a transparent structure that allows for interpretation of results, which is well suited for environmental applications.[4]

As outlined, this study is aimed at predicting future groundwater levels in 31 wells distributed throughout the city of Zliten by means of both Linear Regression and Decision Tree Regression algorithms. The data used in this study encompasses ten consecutive months of measurement data. Modeling and analysis were undertaken in RapidMiner due to its easy-to-use and visually interactive interface, which provided support through all stages of data analysis from preprocessing to modeling to performance measurement.

The research gap lies in the scarcity of local studies that employ data mining algorithms to interpret and predict groundwater behavior in the city of Zliten, which renders this study a significant scientific contribution.

3.Research Problem :

The research problem is represented by the absence of predictive models based on real-world data to explain the abnormal rise in groundwater levels in the city

of Zliten. Accordingly, this study seeks to address the following main research question:

To what extent can Linear Regression and Decision Tree algorithms provide an accurate model for predicting groundwater levels in the city of Zliten?

4. Research Objectives :

The objective of this study is to predict the groundwater levels in Zliten City in Libya, using data mining techniques. The objectives of the research are:

4.1 To determine and analyze the time trend of groundwater levels in thirty-one wells in the city of Zliten over ten successive months.

4.2 To create a predictive model using Linear Regression to forecast the groundwater levels in Zliten City.

4.3 To create a second alternative predictive model using Decision Tree Regression with the aim of predicting groundwater levels in the same area as the first model, and comparing both models.

4.4 To employ expected statistical measures to evaluate the forecasted groundwater levels; namely Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).

4.5 To provide technological and procedural recommendations to control the phenomenon and aid environmental decision-making with the predictive character of the data collected.

5. Significance of the Study:

The study is relevant because it is addressing one of the important environmental issues affecting urban development in the city of Zliten, Libya, namely, the abnormal rise of the groundwater levels. The significance of this study can be articulated as follows:

5.1 The study offers a practical model based on artificial intelligence and data mining to monitor and predict groundwater levels' change in Zliten, which is more advanced than traditional techniques that relied solely on observation or direct measurements.

5.2 The model provides nearly accurate predictions of total groundwater rise levels that can be usefully designed to act as a potential support mechanism or tool for decision makers in Zliten and Libya environmental authorities to anticipate useful future developments in relation to the groundwater rising in the city.

5.3 The study highlights the significance of using specific software tools such as RapidMiner to mine and develop environmental data, which has opened wider options for the use of artificial intelligence in evaluating developing urban prospects and natural resource management.

6. Research Questions:

Following the afore-stated clearly defined problem and research objectives, this paper aims at giving solutions to the following research questions:

6.1 To what extent, in terms of time horizon, will the monitoring period of groundwater levels in the wells of the Zliten city stretch?

6.2 Is there a possibility that the Linear Regression algorithm can make accurate forecasts of upcoming groundwater levels in Zliten?

6.3 Can the Decision Tree Regression algorithm overcome the capability of Linear Regression in predicting groundwater levels?

6.4 What is the predictive or accurate value of the models used in the study, and how would this be measured through statistical parameters like RMSE, MAE, and R^2 ?

6.5 What do the implications of the results drawn from this research signify for the application of proactive environmental decision-making by the Libyan city of Zliten?

7. Scope of the Study:

This study seeks to investigate the current state and rising groundwater levels in the city of Zliten, Libya, using primary data from 31 wells located in different districts in the city. The groundwater levels were recorded monthly for a period of ten continuous months. The study is limited to the use of two predictive

algorithms: Linear Regression and Decision Tree Regression, using RapidMiner as the software-based analytical tool. The analysis focuses on assessing the performance of the two models in terms of accuracy using three standard statistical performance measures RMSE, MAE, and R^2 with the goal to find the best model in forecasting groundwater levels in Zliten.

8. Research Methodology:

The research utilizes an applied analytical research approach to build authentic predictive models that estimate and predict Zliten groundwater levels. The research is based on quantitative data collected during the fieldwork. It is the most appropriate approach for this research owing to its applicability in the nature of the research in using actual observation data, examining temporal changes, and predicting values in the future by using artificial intelligence techniques.

8.1 Data and Source:

A total of 31 wells in various neighborhoods in Zliten, Libya, were used in the data collection. Groundwater level measurements were conducted on a monthly basis for a total of ten months between the months of February and November 2024.

8.2 Analytical Tool

The main tool used in conducting analysis and modeling was RapidMiner. The tool provides a visual and interactive interface that supports more than one stage of analysis, and these include:

- Feature Selection
- Performance Evaluation

8.3 Data Preprocessing

There were complete preprocessing steps done before constructing the models, including:

- Handling and imputing missing values
- Standardization of units and measurement
- Extraction of the research material from the dataset

8.4 Models Used:

There were two predictive algorithms employed:

- Linear Regression: Used for finding a relationship between a dependent variable (groundwater level) and one or multiple independent variables (month, well ID).

8.5 Performance Evaluation:

These two models were compared using the criteria of three main statistical metrics:

- Root Mean Square Error (RMSE)
- Mean Absolute Error
- Coefficient of Determination (R^2)

All three of these methods give a good insight into the fitness of the forecast values to the actual values and the overall best performing model [5].

8.6 Comparison of Models:

After the development of both models, it was possible to conduct a comparison of the two models quantitatively to find which one is capable of representing our area of study more accurately so that accurate predictions of the levels of the groundwater for the future can be ensured.

9. Literature Review:

In recent years, there has been a growing recognition of the potential of artificial intelligence approaches in predicting groundwater levels. Below are a few relevant studies of recent interest.

9.1 Groundwater Level Prediction in Landslide Area [6]:

From fieldwork at a landslide area in Shuping region of the Three Gorges Dam in China, this study aimed to understand the relationship between influencing factors (such as rainfall and levels in the reservoir) and fluctuations in groundwater level. A Classification and Regression Tree (CART) model was applied to predict the groundwater levels and this was compared to the performance of Support Vector Machine (SVM) with the same set of variables.

The findings showed that the CART model performed better than SVM. The CART model predicted mean absolute steps of only 0.28 meters (compared to SVM of 1.53 meters), while relative step errors were very similar at just 1.15% (CART) and 6.11% (SVM).

This study concluded CART has excellent generalization and interpretability capabilities making it a very good tool for groundwater dynamics especially in a complex geological setting or landslide area.

9.2 Predicting Groundwater Levels in Texas Using Linear Regression[7]:

This project was concerned with the ongoing declines in groundwater levels that are becoming a significant constraint on economic and social development, especially in agricultural and industrial contexts. A predictive model was developed by the researchers using Partial Least Squares (PLS) Regression, which is a generalization of linear regression, and tested with historical data from the state of Texas. This model was able to produce very accurate predictions and produced well defined, generalizable forecasting formulae. When examined in comparison with other algorithms, the PLS-based model struck a positive balance of simplicity and accuracy, and it has potential value in both long-term groundwater monitoring and forecasting. The project addressed the strategic function of groundwater as a direct input for agricultural, domestic, and industrial purposes, and established the value of predictive models to contribute to strategic environmental decisions.

9.3 Forecasting Groundwater Levels in Ardabil Plain Using SVR and M5 Decision Tree

[8]:

The Ardabil Plain in northwestern Iran is facing severe groundwater level reduction due to increased rainfall variability, prolonged drought, and overexploitation of groundwater for agricultural uses. The aim of the current study was to develop a groundwater prediction model using two algorithms, Support Vector Regression (SVR) and M5 Decision Tree.

The models were built on a monthly reading of 24 observation wells (piezometers) and this led to training as well as testing the models for predicting groundwater levels. The independent variables were previous groundwater levels, the amount of rainfall, and discharge of the well whereas the target variable was the this month's groundwater level.

Both models had a good ability to predict groundwater levels, but based on simplicity and interpretability compared to SVR, the M5 Decision Tree was used for the analysis. The SVR model was more complicated but had the best prediction performance in predicting groundwater levels of certain wells with a Correlation Coefficient of $R = 0.996$.

10. RapidMiner Platform:

RapidMiner is a powerful and popular platform for data mining and machine learning in both academic research and practitioner audiences. Its visual interface is easy to use, flexible, and allows the complex set of analytical processes to be run with little to no explicit programming or coding aspects required by the user[9]. The platform has been used successfully in consumer behavior, health crisis prediction, and environmentally based resource monitoring. In addition, it provides solid support for performance measures related to RMSE, MAE, R^2 , etc. It also has an impressive array of built-in algorithms for predictive modeling and data analysis.

11. Linear Regression Algorithm:

Linear regression is among the most common or typical predictive modeling techniques. In it, a linear relationship is estimated between a dependent variable and one or more independent (or explanatory) variables through a mathematical equation.

Linear regression can be relatively simple to use, it is still popularly used in the water sciences and hydrology when forecasting variables like groundwater levels, if an approximate linear relationship can be assumed between these variables.

As of late, research shows that linear regression performs well, when the data set is clean and likely void of bias (noise/variation), but amplitude fluctuates, with likelihood rankings dropping the with increasing variance and nonlinearity in the data increasingly heterogeneous datasets [10].

12. Decision Tree Regression Algorithm:

Decision Tree Regression, considered a nonlinear predictive algorithm, is frequently used because it performs well when handling complex data and interactions among many variables. The mechanism for Decision Tree regression is to split the dataset recursively into smaller and smaller subsets that are more homogeneous. This happens through decision nodes that reflect the variable, or variables, that has the greatest influence on the outcome.[11]

Research has shown that Decision Tree models will have an appropriate balance between accuracy and explain ability. In applied fields such as groundwater management, where understanding the rationale of model predictions is as critical as the exact results.

13. Implementation and Results:

The dataset used in this study was obtained from an official report issued by the **Libyan Ministry of Water Resources**, entitled “*Study on the Phenomenon of Rising Groundwater Levels in the City of Zliten.*” [12]. The report includes groundwater level readings collected from **31 wells** located throughout the city of Zliten .

Measurements were recorded over several time intervals spanning a continuous period of **ten months**, from **February to November 2024**. The data reflects actual field observations of groundwater levels across various locations in the city, **Table (1)** presents a sample of the groundwater level readings that were used in the analysis.

Table (1): Groundwater Level Measurements at Selected Wells from the Study Sample during 2024

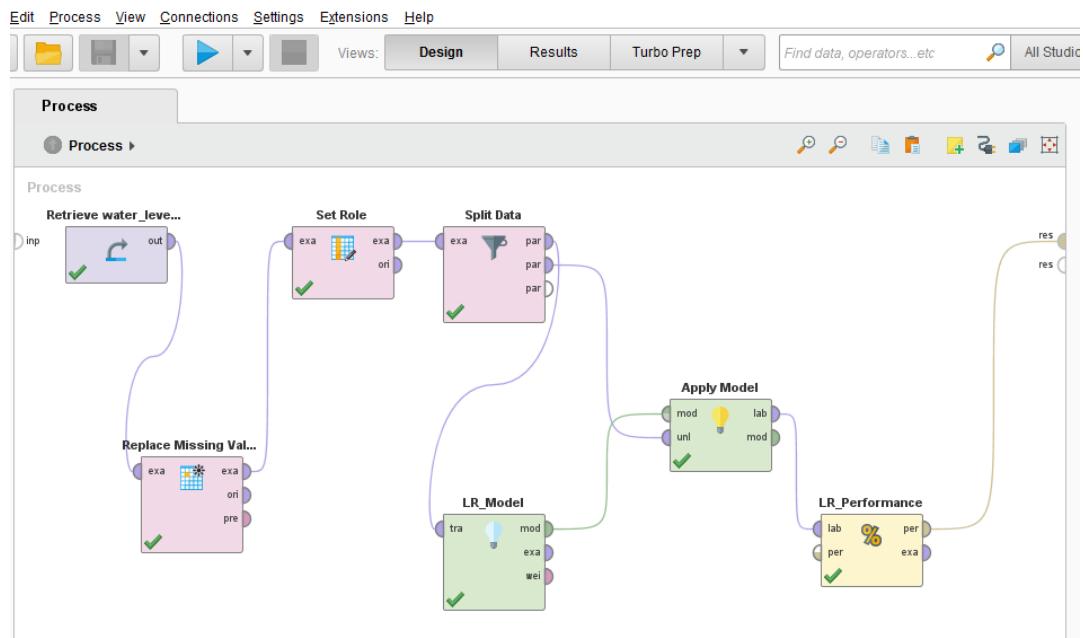
Well No.	Measurement date 2024 / Depth to water level from ground surface (meters)									
	2	3	4	5	6	7	8	9	10	11
W1	2.93	3.1	3.09	3.34	3.5	3.6	3.66	3.86	3.81	3.88
W2	5.31	5.15	5.15	5.33	5.43	5.58	5.65	5.83	5.82	5.85
W3	0.4	0.58	0.72	0.87	1.08	1.15	1.25	1.2	1.26	1.34
W4	7.03	7.11	7.14	7.27	7.39	7.52	7.6	7.66	7.72	7.77
W5	4.88	4.95	4.68	5.1	5.23	5.36	5.44	5.49	5.59	5.67
W6	1.47	2.02	2.06	2.12	2.25	2.39	2.49	2.5	2.54	2.62
W7	12.92	12.91	12.92	13.02	13.12	13.25	13.32	13.34	13.37	13.41
W8	12.56	12.6	12.91	12.67	12.76	12.89	12.97	12.98	13.04	13.06
W9	5.6	5.6	5.65	5.86	5.92	6.11	6.17	6.07	6.21	6.36
W10	14.08	14.08	14.1	14.21	14.31	14.45	14.5	14.51	14.56	14.58

However, prior to being exported for analysis and modelling in RapidMiner (version 9.9), the dataset was first entered and preprocessed in Microsoft Excel. In an attempt to forecast future values for the groundwater level in Zliten, two modelling approaches had to be applied: one for Linear Regression and the other for Decision Tree Regression.

The treatment of the missing values in the dataset required the application of data transformation methods, where the missing values were replaced using the arithmetic mean in order to make all the statistics consistent. The groundwater level was then selected as the aim to be achieved by being the dependent variable or the ‘label’ in the modelling context.

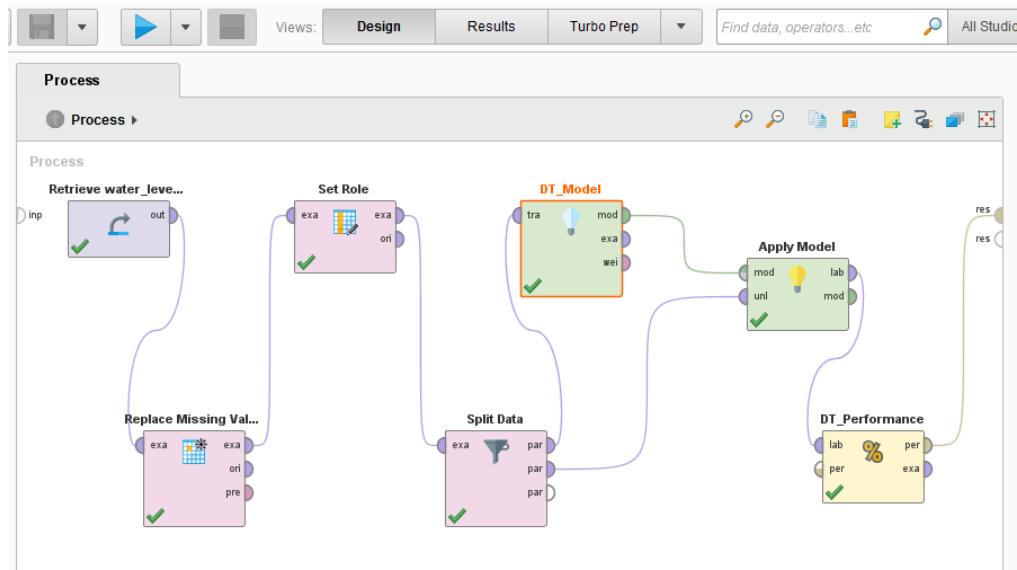
In order to prevent any subjective assessment on the efficiency of the predictive models, a Split Data approach was applied to split the dataset into subsets, where 70% was utilized for training, followed by 30% for the purpose of testing. In order for there to be a proper assessment on the accuracy as well as the efficiency of the models, a performance (regression) operator was applied in the final process in order to measure the efficiency of models developed using statistical regression test standards.

Figure (1) illustrates the steps that are employed during the development of a predictive model using a linear regression algorithm.



Figure(1) illustrates the implementation of the Linear Regression algorithms.

Figure(2) illustrates the steps followed in the development of the predictive model using the decision tree regression technique.



Figure(2) illustrates the implementation of the Decision Tree algorithms.

Table (2) shows the sample running of the predictive algorithms. This table presents 10 wells that are chosen from the study sample, and it shows the average

actual values of groundwater levels with the average predicted values based on Linear Regression and Decision Tree Regression algorithms.

Table (2) presents a selection of the results obtained from the predictive modeling process.

well-id	Actual water level value	prediction water level(Decision Tree)	prediction water level(Linear Regression)
W1	3.477	3.324	7.143
W2	5.510	3.324	7.003
W3	0.985	3.324	6.863
W4	7.421	6.33	6.723
W5	5.239	6.33	6.583
W6	2.246	2.246	6.443
W7	13.158	10.652	6.304
W8	12.844	10.652	6.164
W9	5.955	10.652	6.024
W10	14.338	15.002	5.884

Figure (3) illustrates the comparison curve between the actual and predicted groundwater levels in the city of Zliten, using the Linear Regression algorithms.

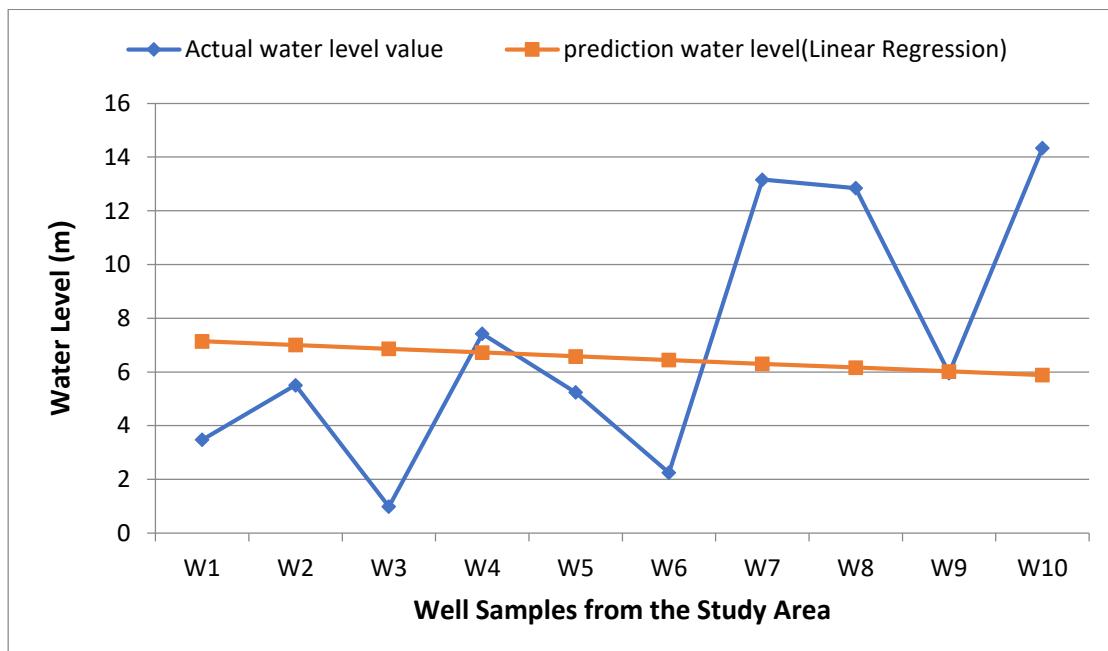


Figure (3) presents a comparison between the actual and predicted groundwater levels using the Linear Regression algorithms.

Figure (4) illustrates the comparison curve between the actual and predicted groundwater levels in the city of Zliten, using the Decision Tree Regression algorithms.

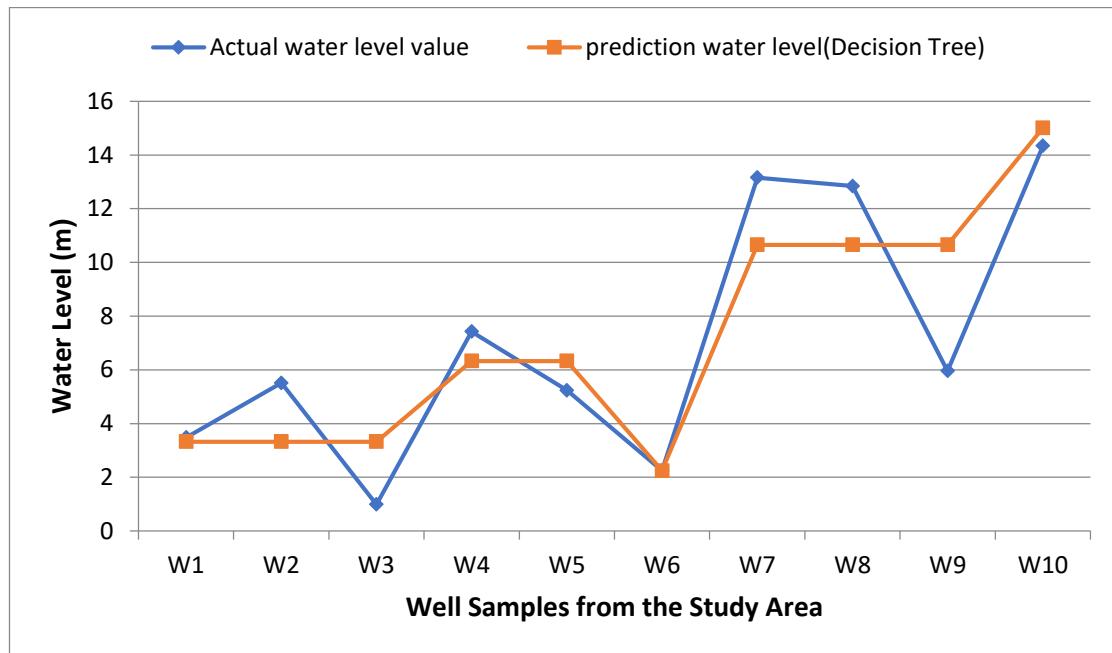


Figure (4) presents a comparison between the actual and predicted groundwater levels using the Decision Tree Regression algorithms.

The previously reviewed figure illustrates the material differences between average actual groundwater levels from the selected wells that show significant variability and nonlinearities. Groundwater level data are non-stationary and change over time in an irregular fashion.

On the other hand, the predicted values from the Decision Tree algorithm showed a moving average curve that smoothly transitioned up and down, moving up and showing a degree of similarity to the movement of the actual values indicating the model was capturing linear and nonlinear trends.

In contrast, the values predicted by the Linear Regression model look almost like a straight line sloping from left to right, indicating that the model assumes a linear model which did not closely approximate the actual fluctuating data. Thus, the linear model did not adequately capture true variation.

Both algorithms were evaluated based on three selected statistical performance metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). In Table(3), we provide a summary of results calculated for the performance metrics.

Table (3) presents the values of statistical performance indicators for both the Linear Regression and Decision Tree Regression models.

Model	RMSE	MAE	R^2
Linear Regression	3.911	3.181	0.098
Decision Tree	1.579	1.113	0.853

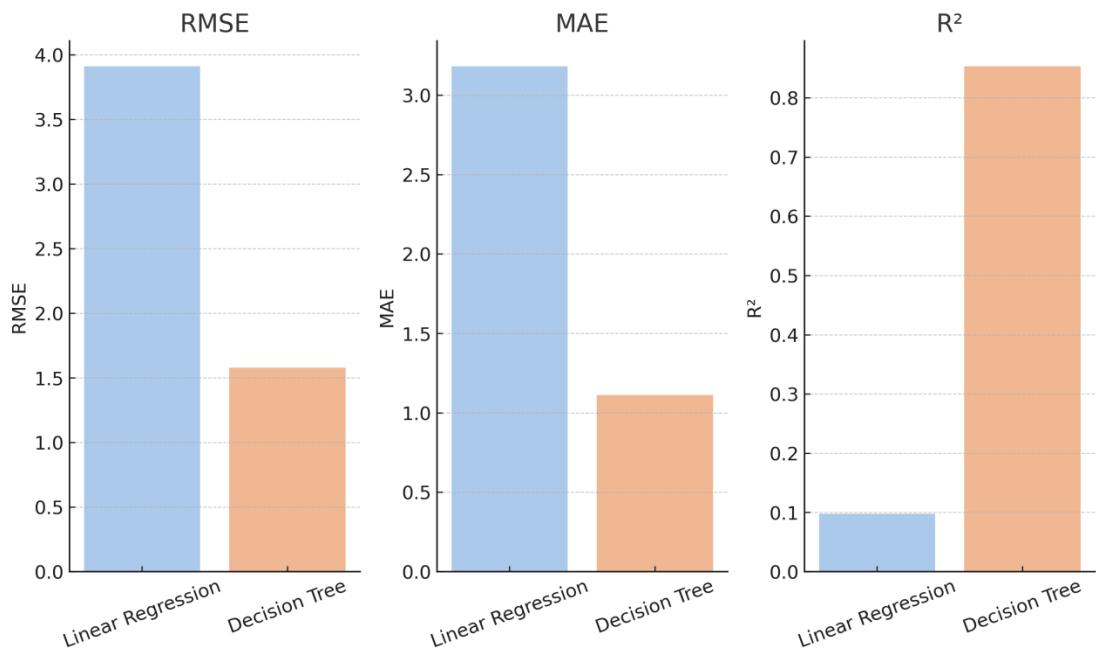
As summarized in Table (3), the Root Mean Square Error (RMSE) of the Linear Regression model is 3.911, whereas the RMSE for the Decision Tree model is considerably lower at 1.579. Thus, since RMSE is largely impacted by larger errors (in this case, the Decision Tree model's predicted mean value was a closer estimate of groundwater levels than the Linear Regression model), one can assume that the Decision Tree model will provide more accurate predictions of groundwater levels than the linear model.

Next, we consider Mean Absolute Error (MAE) and find the Linear Regression model had an MAE of 3.181, whereas the Decision Tree model had an MAE of only 1.113. MAE is also similar to RMSE in that it establishes an average magnitude of errors, regardless of their direction. Just as with RMSE, MAE reveals a distance between estimates and observed values for all predictions, indicating the Decision Tree produced better estimates for groundwater levels than the linear model.

Finally, we looked at the Coefficient of Determination (R^2) to assess how much variation in groundwater levels could be attributed to the input dataset. The Linear Regression model only produced a very low R^2 value of 0.098, indicating that the Linear Regression model has weak explanatory power in relation to the data because only 9.8% of total variation can be attributed to this model. By contrast, we reported an R^2 of 0.853 for the Decision Tree model, meaning this model can

explain 85.3% of the variation in the data. This puts forward evidence supporting the Decision Tree as the most suitable model for capturing the relationship that the input dataset has with groundwater levels.

Figure(5) illustrates the statistical performance metrics for both the Linear Regression and Decision Tree Regression models.



Figure(5) illustrates the comparative chart of statistical performance indicators between the Linear Regression and Decision Tree models.

Previous results show the type of relationship between influencing variables and groundwater levels within our study area. The relationship is mainly nonlinear; hence, the Decision Tree model will better capture and portray such a complex pattern.

Whereas in the Linear Regression model, there is an implicit assumption that the relationship between the input and outputs is linear, the Decision Tree sequentially splits on thresholds that provide the best prediction. The tree would, therefore, be more capable of functioning under nonlinear assumptions or to allow conditional or interaction effects that involve some intervening variable.

These noted characteristics strengthen the rationale for using flexible algorithms which can model nonlinear relationships when conducting research for predicting groundwater levels.

14. Conclusion

This study aimed to forecast groundwater levels in the city of Zliten, which in recent years has experienced a high rise in groundwater tables with rising urban, health, and environmental issues.

In an attempt to eliminate this, two data mining methods Linear Regression and Decision Tree Regression were used via the RapidMiner software. The models were also trained with 31 well monthly data spanning ten months.

RMSE and R^2 values from the Outcome of evaluation showed that the Decision Tree model performed significantly better than the Linear Regression model, as it had a lower error value (RMSE=1.579) and a higher R^2 value ($R^2=0.853$), which were indicators of a high level of prediction accuracy. On the other hand, the Linear Regression model performed poorly, which indicated that it was not suited for modeling a nonlinear or highly varying natural dataset.

Such findings emphasize the need to integrate data mining methods in water resource management, especially in areas where changes are happening quite rapidly in groundwater system dynamics.

15. Recommendations

Based on the findings of the study, it is suggested that:

15.1 Advanced data analysis algorithms should be used for ongoing monitoring of groundwater levels because of their good forecasting potential and future trend analyses.

15.2 The Decision Tree model should be used as a budgeting model for groundwater applications in Zliten, partly due to its reporting of nonlinear relationships and ease of interpretation.

15.3 The study should look into additional influencing variables, such as rainfall levels, temperature variations, and groundwater pumping, to improve the predictive value.

15.4 The study should be developed with artificial intelligence technologies integrated with decision support systems to help municipal and environmental agencies perform forward-thinking infrastructure and public health planning.

15.5 Predictive models integrated with Geographic Information System (GIS) technology can create mapping to identify vulnerable or high-risk areas, assisting the agencies with timely or targeted remedial actions.

15.6 Governmental departments should be encouraged to create and maintain databases for consistent recording of groundwater levels, and to make these databases publicly accessible to affected parties and researchers to develop sustainable water knowledge and planning.

16. References:

[1] Ashmil.A , M., Naji. A, Abu Zandah.M , Shakhtur.O & Kahil.H (2024). Chemical and physical properties of shallow aquifer water in Zliten, Libya [in arabic], 18(8),PP.49, <http://jsba.misuratau.edu.ly/ojs/index.php/jsba/article/view/219>

[2] Sebt. M. V., Sadati-Keneti.Y, Rahbari. M., Gholipour. Z., & Mehri. H (2024). Regression Method in Data Mining: A Systematic Literature Review. *Archives of Computational in Engineering*, 31(6),PP.3515-3534, <https://doi.org/10.1007/s11831-024-10088-5>

[3] Kecheng Qu * (2024). Research on linear regression algorithm In MATEC Web of Conferences (pp. 01046 ,Vol. 395, <https://doi.org/10.1051/matecconf/202439501046>

[4] Mienye, I. D., & Jere, N. (2016). A survey of decision trees: Concepts, algorithms, and applications. *IEEE access*,v(4) ,pp.86716-86727 , <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10562290>

[5] Hodson, T. O. (2022). *Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not.* Geoscientific Model Development, V(15),PP.5481–5487, <https://doi.org/10.5194/gmd-15-5481-2022>

[6] Zhao, Y., Li, Y., Zhang, L., & Wang, Q. (2016). Groundwater level prediction of landslide based on classification and regression tree. *Geodesy and Geodynamics*, 7(5), 348-355, <https://doi.org/10.1016/j.geog.2016.07.005>

[7] Kommineni, M., Reddy, K. V., Jagathi, K., Reddy, B. D., Roshini, A., & Bhavani, V. (2020, March). Groundwater level prediction using modified linear regression. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1164-1168). IEEE , <https://doi.org/10.1109/ICACCS48705.2020.9074313>

[8] Sattari, M. T., Mirabbasi, R., Sushab, R. S., & Abraham, J. (2017). Prediction of groundwater level in Ardebil plain using support vector regression and M5 tree model. *Groundwater*, 56(4), 636-646, <https://doi.org/10.1111/gwat.12620>

[9] Madyatmadja, E. D., Jordan, S. I., & Andry, J. F. (2021). Big data analysis using rapidminer studio to predict suicide rate in several countries. *ICIC Express Letters, Part B: Applications*, 12(8),pp. 757-764, <http://www.icicelb.org/ellb/contents/2021/8/ellb-12-08-12.pdf>

[10] Qu, K. (2024). Research on linear regression algorithm. In *MATEC Web of Conferences* (Vol. 395, p. 01046). EDP Sciences, <https://doi.org/10.1051/matecconf/202439501046>

[11] Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163(8),pp.15-19, <https://www.academia.edu/download/68473530/ijca2017913660.pdf>

[12] Emergency Committee. (2025). Report on the phenomenon of rising groundwater levels in the city of Zliten (Version No. 2). Ministry of Water Resources, Government of National Unity, State of Libya. (Unpublished report)