Journal of Geoscience and Environment Protection, 2024, 12, 1-29
https://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Machine Learning and Statistical Analysis in Groundwater Monitoring for Total Dissolved Solids Assessment in Winkler County, Texas

Azuka I. Udeh[1], Osayamen J. Imarhiagbe[1], Erepamo J. Omietimi[2*], Abdulqudus O. Mohammed[1], Oluwatomilola Andre-Obayanju[3]

[1]Department of Geoscience, University of Texas Permian Basin, Odessa, USA
[2]Department of Geology, University of Pretoria, Pretoria, South Africa
[3]Department of Geology, University of Benin, Benin, Nigeria
Email: *erepamo.omietimi@tuks.co.za

How to cite this paper: Udeh, A. I., Imarhiagbe, O. J., Omietimi, E. J., Mohammed, A. O., & Andre-Obayanju, O. (2024). Machine Learning and Statistical Analysis in Groundwater Monitoring for Total Dissolved Solids Assessment in Winkler County, Texas. *Journal of Geoscience and Environment Protection, 12,* 1-29.
https://doi.org/10.4236/gep.2024.126001

Received: April 25, 2024
Accepted: June 3, 2024
Published: June 6, 2024

Abstract

This research aims to develop reliable models using machine learning algorithms to precisely predict Total Dissolved Solids (TDS) in wells of the Permian basin, Winkler County, Texas. The data for this contribution was obtained from the Texas Water Development Board website (TWDB). Five hundred and ninety-three samples were obtained from two hundred and ninety-eight wells in the study area. The wells were drilled at different county locations into five aquifers, including Pecos Valley, Dockum, Capitan Reef, Edward Trinity, and Rustler aquifers. A total of fourteen different water quality parameters were used, and they include Potential hydrogen (pH), Sodium, Chloride, Magnesium, Fluoride, TDS, Specific Conductance, Nitrate, Total Hardness, Calcium, Temperature, Well Depth, Sulphate, and Bicarbonates. Four machine learning regression algorithms were developed to get a good model to help predict TDS in this area: Decision Tree regression, Linear regression, Support Vector Regression, and K-nearest neighbor. The study showed that the Decision Tree produced the best model with attributes like the coefficient of determination R^2 = 1.00 and 0.96 for the training and testing, respectively. It also produced the lowest score of mean absolute error MAE = 0.00 and 0.04 for training and testing, respectively. This study will reduce the cost of obtaining different water quality parameters in TDS determination by leveraging machine learning to use only the parameters contributing to TDS, thereby helping researchers obtain only the parameters necessary for TDS prediction. It will also help the authorities enact policies that will improve the water quality in areas where drinking water availability is a challenge by providing important information for monitoring and assessing

groundwater quality.

Keywords

Machine Learning, Regression, Aquifers, Winkler County, Sinkholes

1. Introduction

The lack of perennial surface water bodies in Winkler County makes it unavoidable to utilize the aquifers in that area for domestic, irrigation, and industrial purposes. Water contamination is influenced by numerous factors, including anthropogenic (oil and gas mining, agriculture) and natural causes (formation dissolution and subsequent introduction of compounds into the aquifer system) (Kim et al., 2019; Shi et al., 2019; English et al., 2020). With increased development and economic growth came population growth, further straining the available water resources. Two sinkholes formed in the Hendrick Field in June 1980 and May 2002, and the mechanism of formation has been documented by several studies (Kim, Lu, & Degrandpre, 2016; Johnson, 2005; Baryakh and Fedoseev, 2011; Kim et al., 2019; English et al., 2020) along with the cause which had to do with the impact of oil and gas activities and the dissolution of naturally occurring evaporites formations (Frumkin et al., 2011; Kim et al., 2019; English et al., 2020) thereby increasing the inorganic and organic TDS contents of the aquifers. Studies around the Hendrick field surrounding the sinkholes show that the impact of oil and gas activities in Winkler County helped in accelerating the formation of the sinkholes through the introduction of meteoric water into the evaporites (Castile and Salado formation) and other bad mining practices such as the lack of intermediate casing (Johnson, 2005), the use of dynamite to blast hard rock units present in the floor of disposal pits (Heithecker, 1932). Most sinkholes result from natural subsurface drainage, drought, and extreme flooding (Gutiérrez et al., 2016). In the case of Winkler County, the sinkhole formation can be attributed to both natural and anthropogenic causes (Kim et al., 2019; Shi et al., 2019; English et al., 2020). In the present study, Decision tree, linear regression, Support Vector Regression Model and K-Nearest Neighbor Regression machine learning algorithms were used to detect TDS levels. Understanding the TDS levels can provide insights into the water quality and potential environmental risks associated with dissolved substances in the water. This information is important for monitoring the impact of sinkholes and dissolved formations on local water resources. Predictive models can help assess the risk of groundwater contamination or salinization due to dissolved solids. This is particularly important in areas prone to sinkholes like Winkler County, as they can serve as conduits for contaminants to reach aquifers. Predictive models as produced from this study can assist in managing water resources more effectively by providing early warnings of potential water quality issues.

This information can inform decisions regarding water usage, treatment, and allocation. Moreover, understanding the factors influencing TDS levels can aid in the development of mitigation strategies to reduce the impact of dissolved solids on water quality. This could involve implementing measures to prevent further dissolution of soluble formations or implementing water treatment techniques to reduce TDS levels. Studying TDS levels in an area with sinkholes can contribute to scientific knowledge about the interactions between geological features, hydrology, and water quality. This research can also serve as an educational resource for students and professionals interested in environmental science, hydrology, and geology. Overall, using machine learning to predict TDS in Winkler County can help address environmental challenges associated with sinkholes and dissolved formations while facilitating informed decision-making and sustainable management of water resources.

2. Natural Causes of the Wink Sinks

The dissolution of the Salado Formation by natural means in the Delaware Basin has been recorded by many researchers (Kirkland and Evans, 1976; Lambert, 1983; Johnson, 1986; Kim et al., 2019; Shi et al., 2019; English et al., 2020). One of the proofs of this assertion is the abnormal and abrupt intercalation of thin and thick salt units (Johnson, 1986). The movement of groundwater can cause the natural dissolution of salt units; however, the artesian water flow is considered the primary cause of the natural dissolution of salt units. This is due to the existing fractures in the Capitan Reef, Tansill, and Yates Formation caused by differential compaction of overlying sediments, which serves as a passageway for meteoric water under artesian conditions (Anderson and Kirkland, 1980; Baumgardner et al., 1982; Kim et al., 2019; English et al., 2020).

Fractures near the Wink Sinks

Identification of fracture systems in the Wink Sinks area has been well documented (Heithecker, 1932; Adams, 1944; Baumgardner et al., 1982; Johnson, 1986; Kim et al., 2019; Shi et al., 2019; English et al., 2020). During the early days of oil production from the Hendrick fields, very little mining practice was observed. Hence, most brine produced was disposed of using unlined, natural, and artificial earthen pits for evaporation (Heithecker, 1932). The nature of the topmost formation (Cenozoic Alluvium), which consists of loose, unconsolidated sand, gravel, silt, and clay, made it very easy for the brine to percolate through the porous and permeable surface material to reach the groundwater (Johnson, 1986). Furthermore, dynamite was employed to blast hard rock units found in pit floors (Heithecker, 1932), which further increased the permeability of the rock (Triassic Santa Rosa formation) and topsoil (Cenozoic Alluvium). Improper grouting and drilling, removal of casings following final filling of boreholes, and corrosion from saltwater can all increase the permeability for downward groundwater migration between permeable layers below the salt sequence and

shallow aquifers. Inherent fractures in the Salado Formation are also a result of deeper solution and collapse, small faults, or warping of the younger layers over the underlying Capitan Reef (Adams, 1944; Baumgardner et al., 1982; Kim et al., 2019; English et al., 2020).

Many countries have monitoring systems that monitor the water quality through various water quality parameters (Mohd Zebaral Hoque et al., 2022). These parameters include Total Dissolved Solids (TDS), Potential Hydrogen (pH), Dissolved Oxygen (DO), and Biochemical Oxygen Demand (BOD). These parameters have been widely used to assess and categorize surface and groundwater quality (Berhe, 2020; Asadollah et al., 2021; Prabowo et al., 2021; Zakir et al., 2022). In Texas, TWDB is a good example of a water monitoring authority. Artificial intelligence and machine learning have been applied to make meaning of the enormous data continually generated from water quality assessments to solve many environmental engineering problems, including underground water quality prediction modeling (Shah & Joshi, 2017; Alizadeh et al., 2018; Haghiabi et al., 2018; Ahmed et al., 2019).

Regression techniques like Linear regression and Decision Tree regression are used by researchers for the prediction of various problems such as ozone depletion prediction, solar thermal system forecasting, and prediction of water quality parameters (Djarum et al., 2021; Noori et al., 2011; Chen et al., 2020). Because there is no single algorithm that can solve all these problems, all the algorithms must be subjected to further scrutiny to get the algorithm that gives the best results: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE).

In this study, we applied four algorithms to the same dataset to identify the best performer as a guideline for future research in this area. The result revealed that the Decision Tree regression performed the best among the four regression techniques. The Decision Tree got the lowest MAE, the lowest RMSE, the best coefficient of determination, and the lowest MSE. This study aims to 1) Evaluate the groundwater quality of Winkler County using TDS as the target parameter, 2) Use machine learning and statistical analysis for the prediction of TDS in Winkler County, and 3) identify the best machine learning model for groundwater monitoring and assessment. The best model produced from trying all four algorithms would be used for future endeavors in groundwater monitoring and assessment and for predicting TDS or other water quality parameters from data obtained in West Texas or from other locations.

3. Data and Methodology

3.1. Study Area

With an estimated area and population of about 1427.49 square kilometers (about half the area of Yosemite National Park) and 8000 inhabitants respectively, as of 2023, West Texas' Winkler County is located at the southeast border of New Mexico and lies mainly within the Pecos River Valley; the northeastern

half is part of the High Plains. The production and refinement of oil are its primary industries, and ranching is also a significant industry. The weather is semi-arid, with temperatures varying greatly from 30 to 107 degrees Fahrenheit. Little precipitation 355.6 mm (about 1.17 ft) annually on average, and significant rates of evaporation are predominant, with spring and fall having the most precipitation, making the county experience arid or semi-arid weather (Ashworth, 1990; Menne et al., 2019). Grass, bushes, and brush comprise most of the land cover, followed by arid lands, developed areas, and crops (English et al., 2020). Elevated levels of oil and gas drilling, production, and exploration operations are another characteristic of this region. With oil and gas operations, the semi-transient population changes.

The geology of Winkler County is as complex as it is varied. Winkler County is on the western side of the Monument Draw trough, on the shelf margin between the Central Basin Platform and the Delaware Basin on the western side of the Permian Basin (Meyer et al., 2012). The oldest investigated rocks in Winkler County are igneous rocks located 10,000 feet below the surface in the Keystone oil field, approximately four miles northeast of Kermit (Jones et al., 1949). This is overlain by sedimentary rocks of the Ordovician, Silurian, Permian, Triassic etc. The Central Basin Platform separates the Delaware and Midland sub-basins of the Permian Basin in West Texas.

Three distinct rock sequences were deposited simultaneously along the margins of the Delaware Basin during the late Guadalupe period due to a large reef known as the Capitan. These sequences include the deep-water marine facies in the Delaware Basin, which are represented by sandstone, shale, and limestone; a reef zone, which is represented by massive crystalline dolomite or limestone; and shelf or lagoonal deposits, which are represented by fossiliferous limestone and shale, dolomitic limestone, saline evaporites, and onshore clastics (Garza and Wesselman, 1962). The Delaware Basin is a structural depression in western Texas and southern New Mexico. It is a depression filled and overlain by chemical deposits of the Late Permian age, which are now either partially eroded west of the Pecos River or east of it (Lang, 1939). The shelf deposits near the reef are usually thin-bedded limestone or dolomite that grades into clastics and evaporites. The Capitan Limestone is a representation of the reef deposits. Back-reef deposits include the Tansill formation of the Whitehorse group, the Yates Sandstone, and the Queen, Grayburg, and Seven Rivers formations. In the Delaware Basin, the sandstones of the Guadalupe age are replaced by the anhydrite of the Castile formation of the Ochoa age.

The Triassic period represents a basin-wide continental depositional event with the formation of the Dockum group. Cretaceous marine transgression was followed by an elevation and Cretaceous deposit erosion during the Laramide orogeny. Rustler, Dewy Lake, and Dockum formations collapsed throughout the Cenozoic due to the dissolution of the Salado and Castille evaporites, and the Pecos Valley Alluvium was later filled in (**Figure 1**).

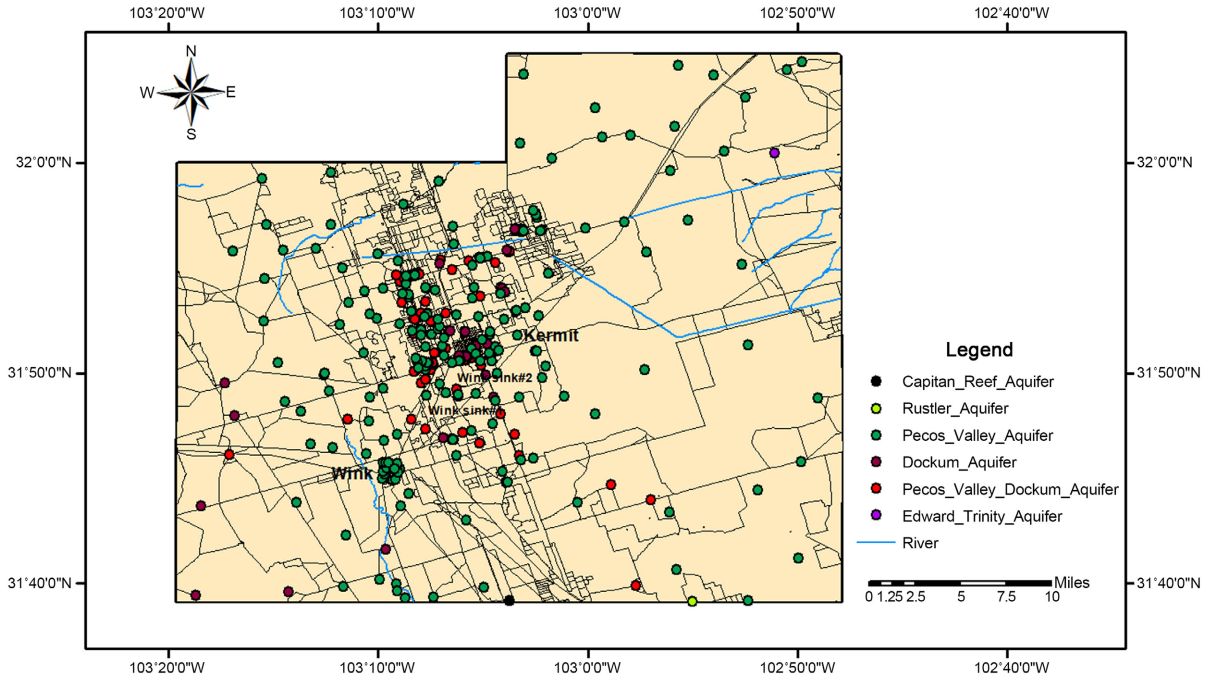


Figure 1. Map of the study area showing 298 wells sunk in different aquifers.

3.2. Data Description

The data for this study was obtained from the Texas Water Development Board (TWDB) website for Winkler County. The data was collected over seventy years (between 1940-2022). The methodology can be broadly categorized into four categories: Data collection, Data Preprocessing, Model Training, Model Evaluation, and Data plotting. Water samples 593 were taken from 298 wells of Winkler County and five aquifers (Edward Trinity, Rustler, Pecos Valley, Dockum, and Capitan Reef). Sixteen parameters were used for this study: Chloride, Total Dissolved Solids (TDS), bicarbonates, pH, Sodium, Magnesium, Sulphate, Fluoride, Calcium, Nitrate, Temperature, Well Depth, Total Alkalinity, Total Hardness, and Specific conductance.

Table 1 below gives a statistical summary all the parameters used for this study. TDS was chosen as the target parameter for all modeling in this study because it is a widely used water quality parameter for estimating dissolved solids in water. TDS is the amount of substance left (organic + inorganic) after a liter of water is evaporated from a container. A low amount of TDS reflects high-quality water and indicates fresh water; a high TDS level suggests brine or low water quality. It is pertinent to state that not all the parameters impact TDS concentration in this area. Heavy metals tend to dissolve faster in acidic water and form toxic compounds with available anions in the water. Specific conductivity is an indicator of ionic salt contamination used to determine the concentration of harmful ionic salts in water. High water-specific conductance is destructive to piping infrastructure. The correlation coefficient reveals that Sodium, Total Hardness, Specific conductance, Magnesium, Chloride, and Sulphate had the most impact on TDS.

Table 1. Statistical Analysis of all fourteen Parameters (593) samples from Winkler County.

Parameters	Statistical Analysis of all fourteen Parameters (593) samples from Winkler County				
	Unit	Min	Max	Mean	S/D
Sodium (Na)	mg/L	6.000	25100.000	40.360	1357.240
Calcium (Ca)	mg/L	5.600	1880.000	159.840	219.240
Magnesium (Mg)	mg/L	0.120	2920.000	40.360	134.440
Bicarbonate ($CaCO_3$)	mg/L	0.000	508.880	160.400	67.710
Sulfate (SO_4^{2-})	mg/L	0.000	5040.000	271.990	491.830
Temperature	˚C	10.000	31.600	21.480	1.890
Chloride (Cl)	mg/L	3.200	41000.000	467.560	2361.560
Fluoride (F)	mg/L	0.000	5.700	1.490	0.770
Nitrate (NO_3)	mg/L	0.000	210.000	7.630	15.430
PH		4.400	9.100	7.570	0.370
TDS	mg/L	105.000	71121.000	1303.000	4165.110
Total Hardness	mg/L	27.000	14334.000	563.390	931.510
Well depth	feet	10.000	4400.000	265.430	331.430
Specific Conductance	µS/cm	175.000	108416.000	2348.300	6933.310

Winkler County's elevated levels of water contamination in the past resulted from the activities associated with crude oil mining and exploration (anthropogenic) and natural causes.

3.3. Data Preprocessing

Data preprocessing is crucial in data analysis to improve the performance and quality of the data. The raw data obtained from TWDB was riddled with missing data and may not perform optimally when used in that state. Hence, it is crucial to perform some preprocessing on the data. The missing cells were filled with the mean of each parameter. This is done to preserve central tendency and maintain the sample size. Some studies will exclude some features when attempting to create a model; in this study, we used all fifteen features as input parameters.

The concentration of TDS in this study for the wells of Winkler County was predicted using four different machine learning regressor algorithms: Decision Tree, Support Vector Machine, Linear Regression, and K-Nearest Neighbor. To get the best out of our modeling, 80% of our dataset was used for training, and 20% was used for testing. A developed model can only be deemed sustainable if, under an increment in the amount of dataset initially used or when a different data is introduced, it can yield particularly reliable results. A model that crumbles upon introducing a new dataset is not necessarily good. For this, cross-validation was performed using a k-fold value of 10, and performance evaluation of the models was done using statistical parameters like MSE, MAE, RMSE, and EVS. With this approach, a fair and accurate assessment of the algorithms and

modeling precision can be obtained by applying consistent training and testing inputs in trials. This also helps in unveiling the biases and shortcomings of the different models. All modeling and analysis were performed using Python v3.8 in Jupyter Notebook in the Department of Geoscience UTPB computer lab.

4. Machine Learning (ML) Algorithms

4.1. Machine Learning Algorithms Used

4.1.1. Regression

Regression analysis uses a set of records containing X and Y values to learn a function. This function can then be applied to predict Y from an unknown X. To obtain the value of Y in a regression given X as independent characteristics, a function that predicts continuous Y is needed. Here, X is referred to as an independent variable, also known as Y's predictor, while Y is referred to as the dependent or target variable. Regression can make use of a wide variety of functions or modules. The most basic kind of function is a linear function. Every kind of regression machine learning model starts with the standard regression equation, which may be computed using the formula below (Kayanan & Wijekoon, 2020)

$$Y = X\beta + e \quad (1)$$

The variables in this equation are the TDS as the dependent variable, X denotes the independent variables (i.e., water quality indicators), β denotes the estimated regression coefficients, and e denotes the errors and residuals.

4.1.2. Linear Regression

Linear regression is one of the most fundamental and popular machine learning techniques. It is a technique that uses mathematics for predictive analysis. One can project continuous, real, or mathematical variables with linear regression (Kanade, 2023). The link between variables under examination is evaluated and quantified using linear regression. Regression models state that the independent factors can predict the dependent variables. Because the independent variable "x" has a range of values, regression analysis estimates the dependent variable "y" value. A case model with a single independent variable is called a simple linear regression. The variable's dependence is defined by simple linear regression. The influence of independent variables is separated from the interaction of dependent variables using simple regression.

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

4.1.3. Decision Regression Tree

While the decision tree-based supervised learning approach is technically described as a rule-based binary tree-building technique, a more straightforward interpretation is as a hierarchical domain division strategy. The data is subjected to the categorization standards for attaining high data homogeneity up until the point at which the nodes can no longer be further subdivided. Decision Trees are

known for their interpretability and transparency, making them valuable for understanding the factors that influence predictions (Czajkowski et al., 2023). By observing an object's attributes and training a model within a tree's structure, decision tree regression generates meaningful continuous output by predicting data in the future. Continuous output denotes a result or output that is not discrete, not solely represented by a known, discrete collection of numbers or values. They are, nevertheless, susceptible to overfitting, particularly if the tree is permitted to grow excessively deeply. This problem can be lessened by employing pruning, establishing a maximum depth, or ensemble techniques like Random Forest or Gradient Boosting (Breiman & Ihaka, 1984).

4.1.4. Support Vector Regression (SVR)

Regression analysis uses a machine learning method called Support Vector Regression (SVR). SVR is a supervised machine learning technique that (Vapnik, 1995) provided. It can be used to address issues with pattern recognition, regression, and classification.

Finding a function that minimizes the prediction error and represents the relationship between the input variables and a continuous target variable is the aim of support vector machines (SVR). SVR's goal is to minimize the coefficients rather than the squared error, or more precisely, the l2-norm of the coefficient vector. Support Vector Regression (SVR) looks for a hyperplane in a continuous space that best matches the data points. The process involves projecting the input variables onto a high-dimensional feature space and then identifying the hyperplane that minimizes the prediction error and maximizes the margin, or distance, between the hyperplane and the nearest data points.

The unweighted average throughout the collection is the random forest forecast for regression. Because of this, it's an effective tool for regression problems where the goal and input variables may have intricate interactions. The challenge of regression analysis is finding a function that maps an input domain to real numbers using a training sample. The SVR seeks to match the best line within a threshold value, unlike other regression models that aim to minimize the error between the real and projected value. The distance between the boundary line and the hyperplane is known as the threshold value. Scaling SVR to datasets with more than ten thousand samples is challenging since the fit time complexity increases more than quadratically with sample count.

4.1.5. K-Nearest Neighbor for Regression (KNN)

The K-nearest neighbor method can be applied to regression analysis in the same way that it can be applied to classification. It is typically used as a classification technique, based on the notion that equivalent points can be located adjacent to each other, even if it can be applied to regression or classification problems.

Just as it may be used for classification, regression analysis can also be carried out using the K-nearest neighbor method. Based on the notion that comparable

points can be found adjacent to each other, it is typically used as a classification procedure, though it can also be applied to regression or classification problems. Because it does not make any assumptions about the underlying data distribution, it is regarded as a non-parametric method. To put it simply, KNN looks at the data points nearby to decide what group a given data point belongs to. This algorithm, also known as a lazy learner or lazy learning algorithm, does not do any training when you provide the training data. Rather, it makes no computations during the training phase; it simply saves the data. It waits to develop a model until the dataset is queried. KNN is hence perfect for data mining. Similar concepts are used in classification and regression problems; however, in regression problems, a prediction about classification is made by taking the average of the k nearest neighbors. Regression is used with continuous values, while classification is utilized with discrete ones in this case. However, the distance must be established before a categorization can be made.

4.2. Performance Measurement of the Different Regression Algorithms

The methods discussed below were used to evaluate and compare the performance of all the models.

4.2.1. Linear Correlation Coefficient (R)

Measured by the linear correlation coefficient (R), a model's ability to correctly predict the observed (real) data. Normally, R values fall between −1.0 and 1.0 A complete positive correlation (a value of 1.0) exists when there is no difference between the observed and the anticipated, and vice versa. A value that expresses the direction and strength of the linear relationship between two variables, x, and y, is another name for it. The computation of this value involves determining the covariance ratio between the two variables and multiplying their standard deviations by one another.

$$R = \frac{n\sum y \cdot y' - \left(\sum y\right)\left(\sum y'\right)}{\sqrt{\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]\left[\left(n\sum y'^2\right) - \left(\sum y'\right)^2\right]}} \tag{2}$$

4.2.2. Coefficient of Determination (R^2)

The R^2 calculates the extent to which the model prediction accounts for the variation in the observed values. A higher R^2 value indicates a better prediction accuracy for the model. The degree to which a statistical model accurately predicts a result is indicated by a value between 0 and 1 called the coefficient of determination. The coefficient of determination is always positive, even when the correlation is negative (Figure 2)

$$R^2 = \left(\frac{\sum_{i=1}^{n}\left(t_i - \underline{t}\right)\left(y_i - \underline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(t_i - \underline{t}\right)^2 \sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2}} \right)^2 \tag{3}$$

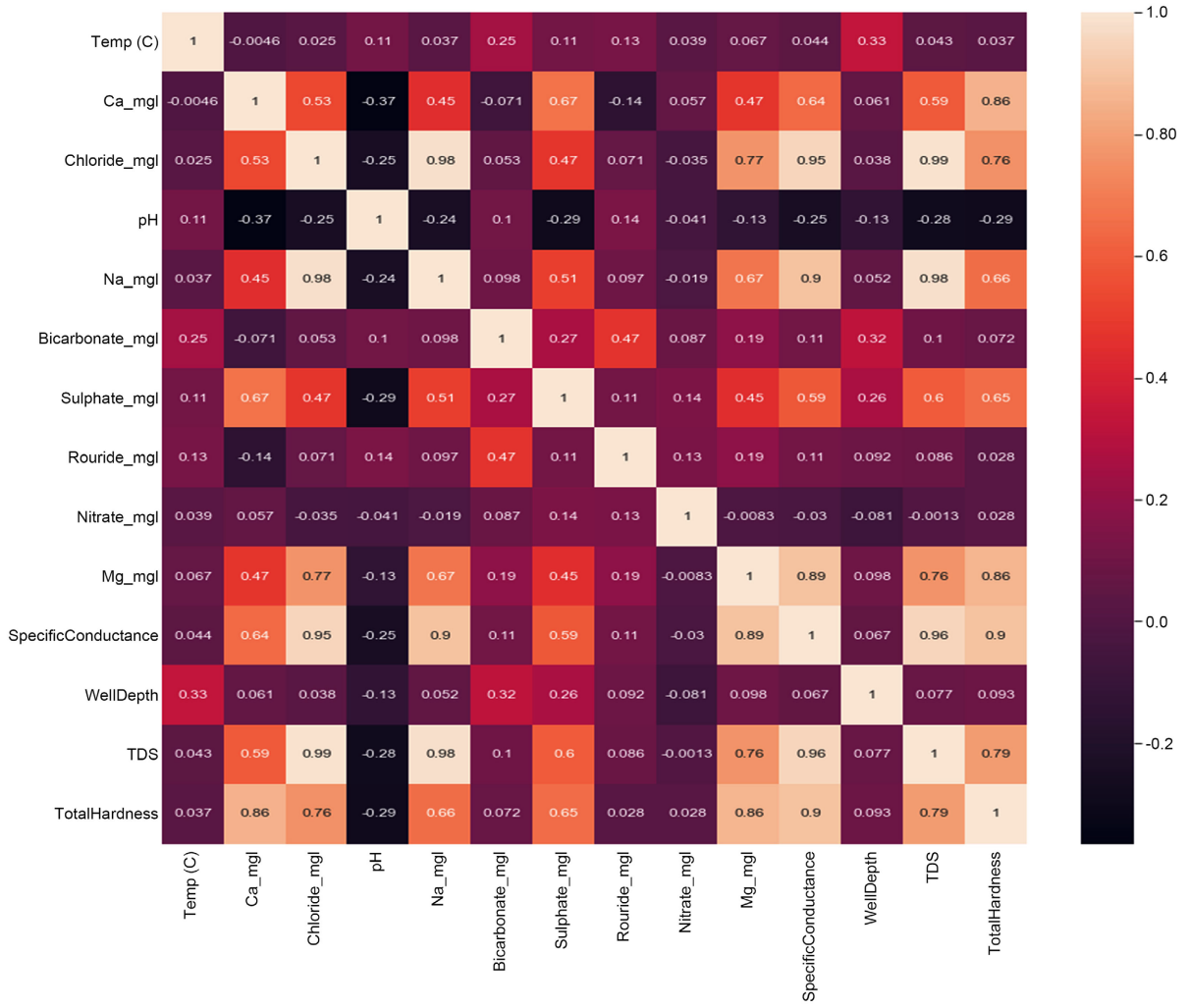


Figure 2. Correlation coefficients of statistically input parameters on TDS.

4.2.3. Root-Mean-Squared Error (RMSE)

The square root of the mean square error is called root-mean-squared error, or RMSE for short. The root mean square error (RMSE) can be defined as the standard deviation of the prediction errors or the average distance of an observed data point from the measured model line. The equation that follows provides the RMSE. The RMSE measures the degree to which these residuals are scattered, providing insight into how well the observed data adheres to the expected values. The RMSE decreases as the data points approach the regression line because the model has fewer errors. Predictions made by a model with a lower error are more accurate.

$$\mathrm{RMSE} = \sqrt{\frac{\sum (y' - y)^2}{n}} \tag{4}$$

4.2.4. Mean Absolute Error (MAE)

The statistical measure of a model's prediction ability is the mean absolute error (MAE), which is the arithmetic of the absolute errors. Since the MAE shows the

relative overall fit or goodness of fit, it is frequently utilized in quantitative predictive models. MAE, one of the most often used loss functions for regression issues, aids users in turning learning challenges into optimization issues. Additionally, it provides regression problems with an easily comprehensible, quantitative measurement of errors.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y - y'|}{n} \tag{5}$$

4.2.5. Prediction Error

A measurement of the difference between expectation and reality is called prediction error. It is frequently used to evaluate prediction accuracy in the context of statistical and machine learning models. Usually, the prediction error is computed as follows:

Actual Value − Predicted Value = Prediction Error.

4.2.6. Cross Validation (K-Fold Method)

Validation is the process of determining whether the numerical results quantifying proposed relationships between variables are appropriate for characterizing the data. Therefore, we need a procedure that leaves enough data for both the validation and training of the model. That is precisely what K-Fold cross-validation achieves. The data in K-Fold cross-validation is separated into k subgroups. The holdout approach is now performed k times, with each repetition using one of the k subsets as the test or validation set and the remaining k-1 subsets combined to create a training set. We use the average error estimation over the k trials to determine our model's overall effectiveness. Each data point appears exactly once in a validation set and k times in a training set. Because the majority of the data is also utilized in the validation set, this considerably minimizes variance as well as bias because the majority of the data is used for fitting. Because the majority of the data is also utilized in the validation set, this considerably minimizes variance as well as bias because the majority of the data is used for fitting. This technique is made more effective by switching up the training and test sets. Based on empirical evidence and general guidelines, k = 5 or 10 is typically favored. However, it can take any value. In this study, we used k = 10 (Table 2).

Table 2. Training and testing results for Cross Validation (Coefficient of determination).

Model	Cross Validation (K-Fold Method)	
	Training	Testing
Decision Tree Regression	0.963	0.911
Support Vector Regression	0.831	0.756
Linear Regression	0.852	0.893
K-Nearest Neighbor	0.974	0.910

5. Results and Discussion

The mean concentrations of Chloride, Sodium, and Sulphate are 467 mg/L, 239 mg/L, and 271 mg/L, respectively, in the Winkler County wells. This is above the EPA (Environmental Protection Agency) recommended number for safe drinking water. This is probably due to the oil and gas activities in the area. The presence of evaporites in the Salado Formation could be the reason there is elevated sodium concentration from the data used. The use of saline water for the secondary recovery of oil in the Hendrick oil field and the subsequent disposal of the water in earthen pits close to the wellhead could be one of the reasons sodium concentrations are quite high in analyzed samples.

Table 3 below summarises all the performance indices in training and testing for all the regression models used for this study.

The correlation between TDS and other input parameters was evaluated. The results show that there is a high correlation between Sodium and TDS (R = 0.98), Chloride and TDS (R = 0.99), Specific conductance and TDS (R = 0.95), and Magnesium and TDS (R = 0.76). These parameters have the greatest influence on the predictive power of the models generated.

5.1. Evaluation of Model Performances

This study has focused on identifying the best-performing regression model for predicting TDS from the four regression models utilized. As stated earlier, the entire dataset was divided into two; 80% was dedicated to training while the remaining 20% was for testing. It should be noted that a couple of other split ratios were tried, such as 70% and 30%, 60% and 40%. The 80%/20% ratio gave the best results (Figure 3).

5.2. Decision Tree Regressor Model

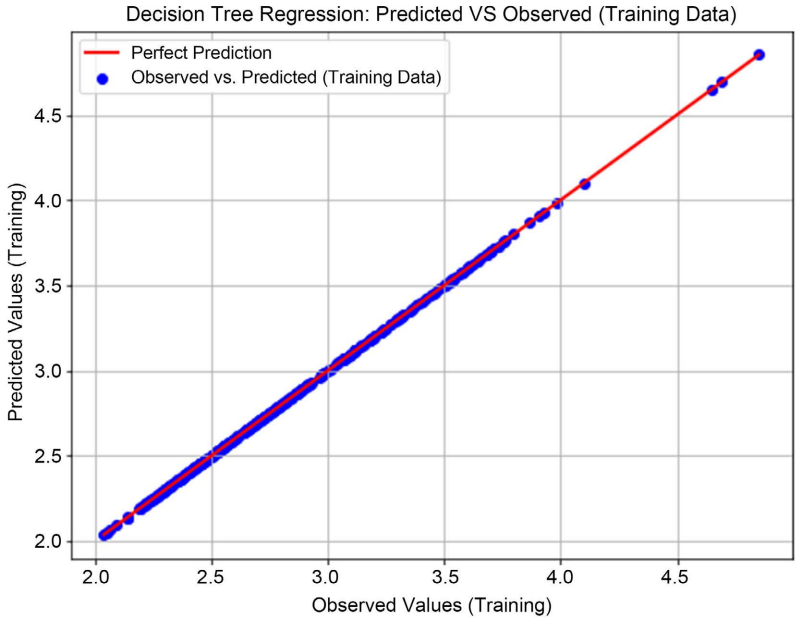
The decision tree regression algorithm yielded impressive results for the training and testing R^2 = 1.0 and R^2 = 0.96, respectively (see Figures 4-6). It showed one of the least MAE, RMSE, and MSE; 0.0, 0.0, and 0.0 for training and 0.05, 0.09,

Table 3. The Performance of different machine Learning algorithms for estimation of TDS.

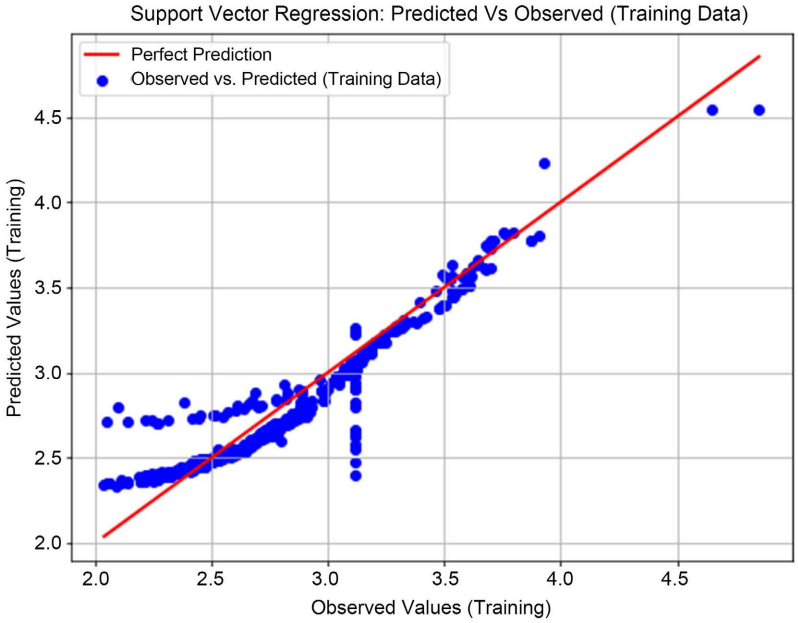
Model	Performance evaluation of the algorithms			
	R^2	MAE	RMSE	MSE
DT: Training	1.000	0.000	0.000	0.000
DT: Testing	0.954	0.050	0.090	0.010
SVR: Training	0.928	0.090	0.110	0.010
SVR: Testing	0.897	0.100	0.130	0.010
LR: Training	0.902	0.080	0.130	0.010
LR: Testing	0.848	0.120	0.240	0.060
KNN: Training	0.974	0.030	0.060	0.000
KNN: Testing	0.977	0.040	0.060	0.000

LR = Linear Regression, **DT** = Decision Tree, **SVR** = Support Vector Regression, **KNN** = K-Nearest Neighbor.

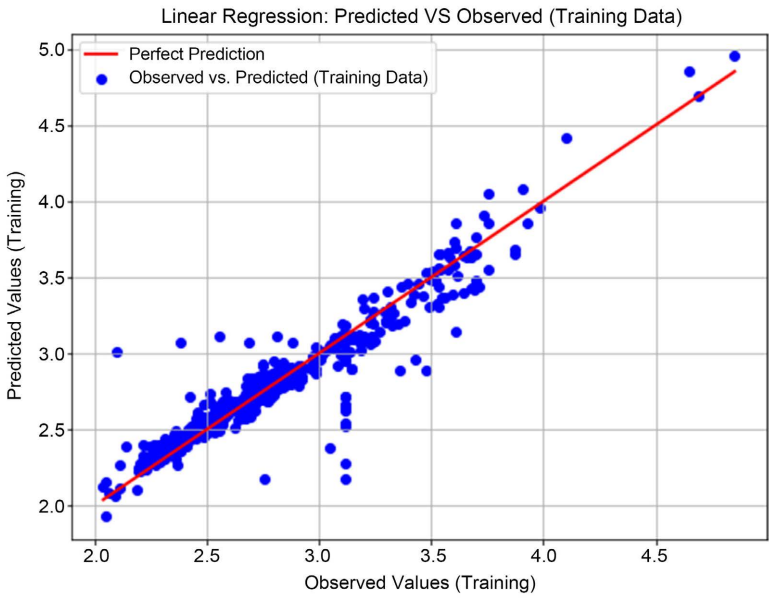
and 0.01 for the testing datasets, as shown in the table above (Table 3). The DTR algorithm was used to predict the concentration of TDS in this study, and it yielded excellent results. Feature importance from DTR reveals that not all the parameters impacted TDS in the samples. Sodium, Total Hardness, Magnesium, Chloride, Specific Conductance, and Calcium were solely responsible, with Sodium contributing more than sixty percent. Cross-validation (k-fold) reveals a mean squared error of 0.013 for training and 0.012 for testing. This shows that our model will give accurate predictions when tested with data from other samples. The benefit of this approach is that parameters that had a major impact on the TDS could be



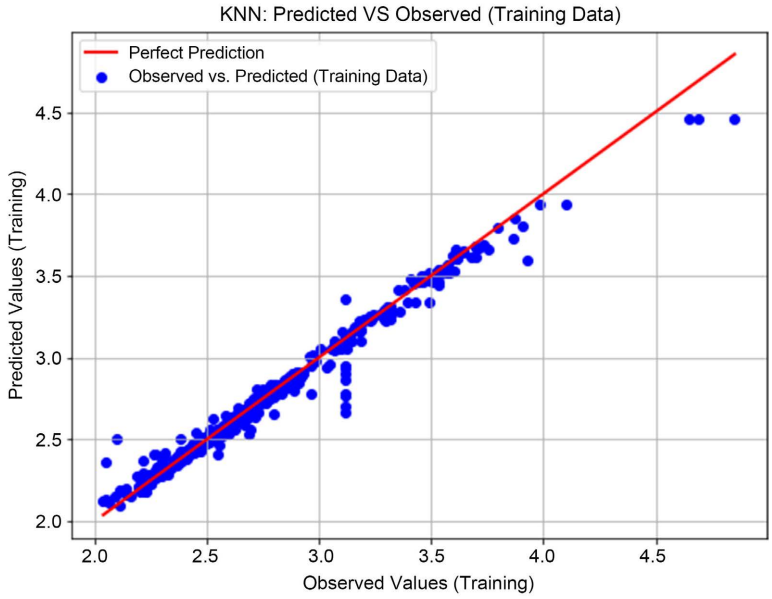
(a)



(b)



(c)



(d)

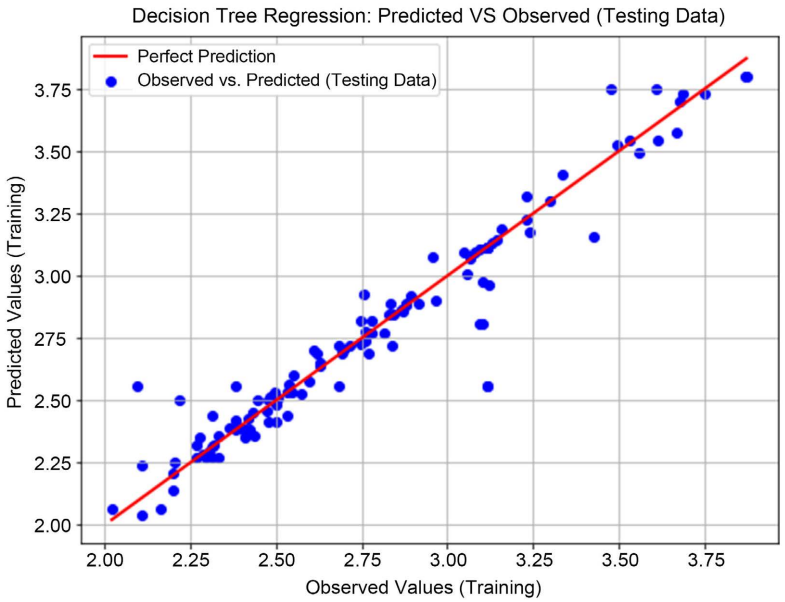
Figure 3. Prediction accuracy for the training dataset (a) Decision Tree Regression (b) Support Vector Regression (c) Linear Regression (d) K-Nearest Neighbor.

directly known, including their dosages and any positive or negative impacts. By conducting parameter evaluations one at a time, according to the decision tree diagram, and modifying the course based on analysis results, it is feasible to use the decision tree diagram to minimize future physicochemical analyses required for TDS doses (Hichem et al., 2022).

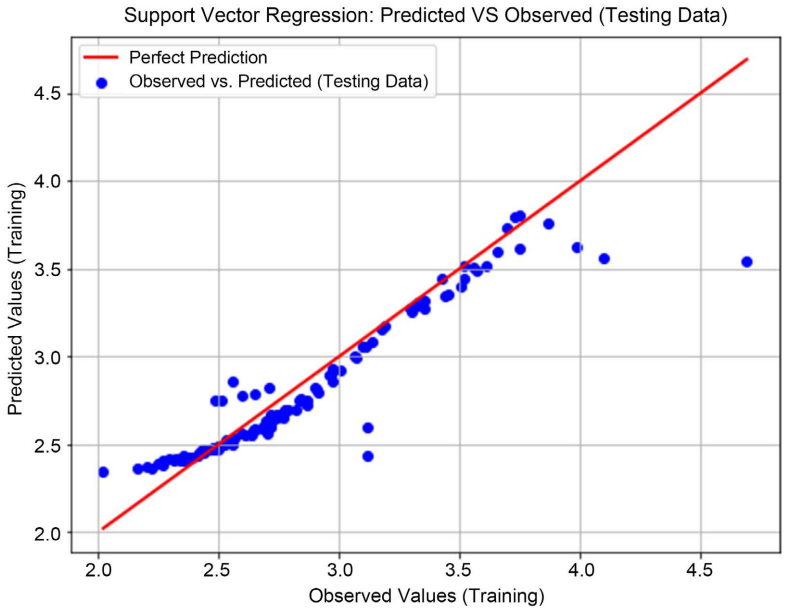
5.3. Linear Regression Model

Many authors believe that simple linear regression analysis cannot accurately

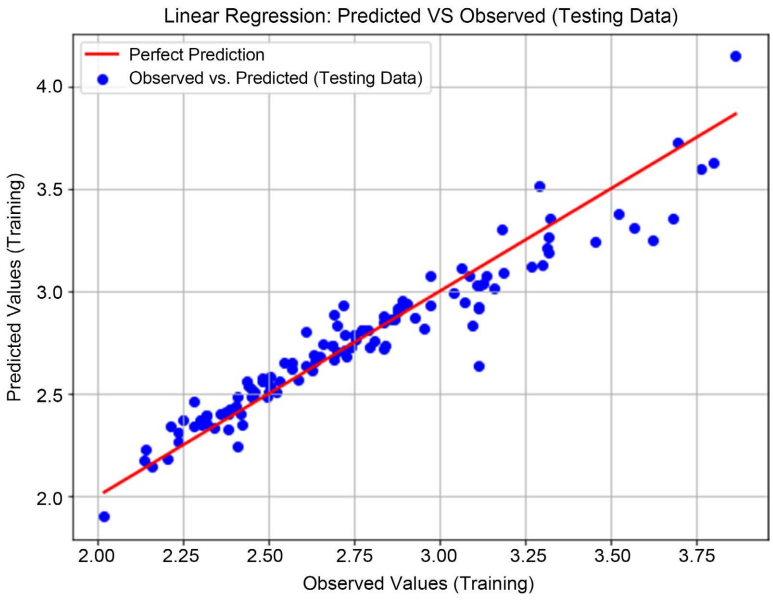
forecast water quality because of complicated linear and nonlinear relationships in the water quality dataset. However, this study aims to compare different machine learning models to get the best model that can accurately predict TDS. When determining the linear relationship between a goal and one or more predictors, linear regression is utilized. Finding the line that best fits the data is the main concept. The line with the lowest total prediction error (across all data points) is the best match. The gap between the point and the regression line is called the error. Linear regression generated one of the lowest coefficients of determination (R^2), 0.91 and 0.73, for training and testing in this study. The MAE, RMSE, and MSE values are 0.08, 0.13, and 0.01 for the training (Table 3).



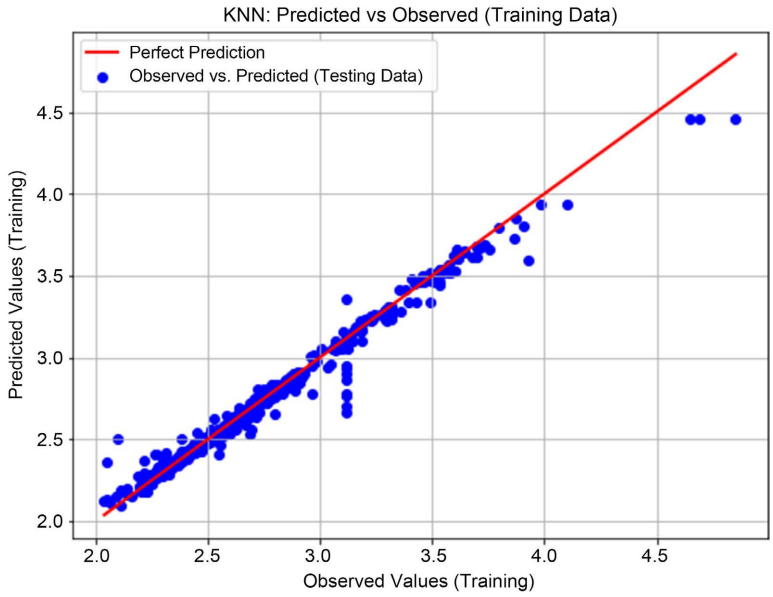
(a)



(b)



(c)

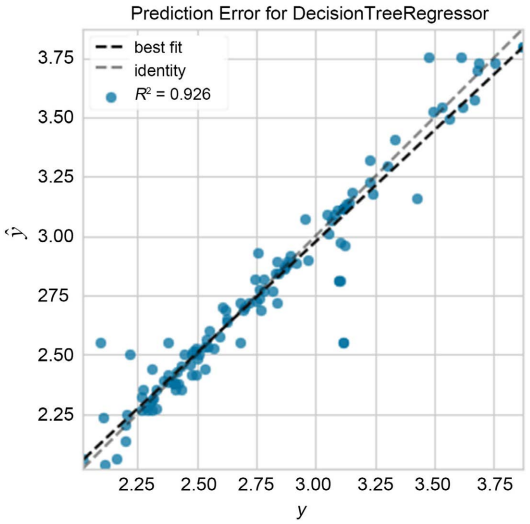


(d)

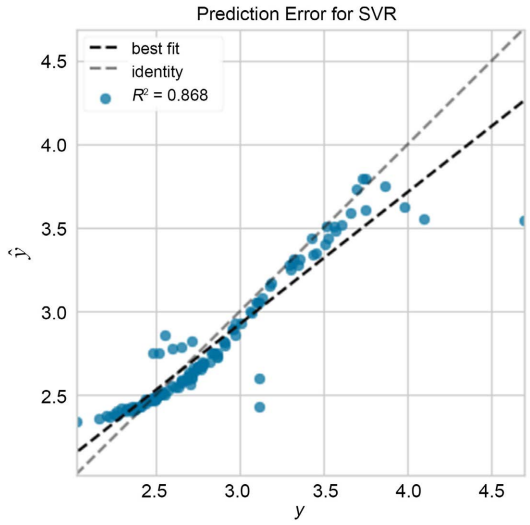
Figure 4. Prediction accuracy for the testing dataset (a) Decision Tree Regression (b) Support Vector Regression (c) Linear Regression (d) K-Nearest Neighbor.

5.4. Support Vector Regression Model

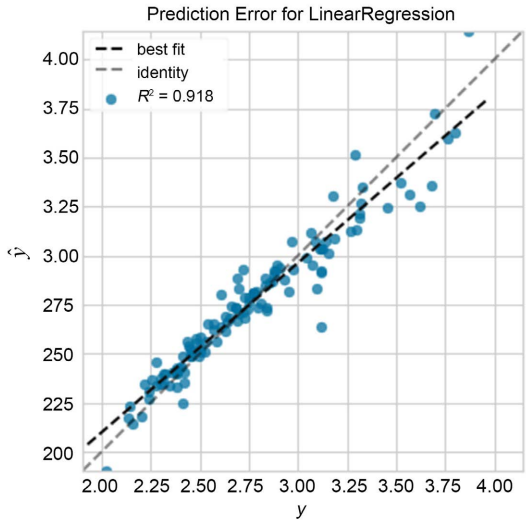
The SVR method's fundamental goal is to minimize structural risk, which is accomplished by comparing the high-limit error to the usual local training error in other machine learning techniques. Using an appropriate kernel function, the original data sets from the input space are mapped into a high-dimensional or even infinite-dimensional feature space, where a maximal separation plane (SP) is created using the SVM (Support Vector Machine) technique. In this study, the SVR gave the highest (MAE = 0.09 and RMSE = 0.11 in training and MAE = 0.10 and RMSE = 0.13 for


Prediction Error for DecisionTreeRegressor
best fit
identity
R^2 = 0.926
3.75
3.50
3.25
3.00
275
2.50
2.25
$\hat{y}$
2.25 2.50 2.75 3.00 3.25 3.50 3.75
y

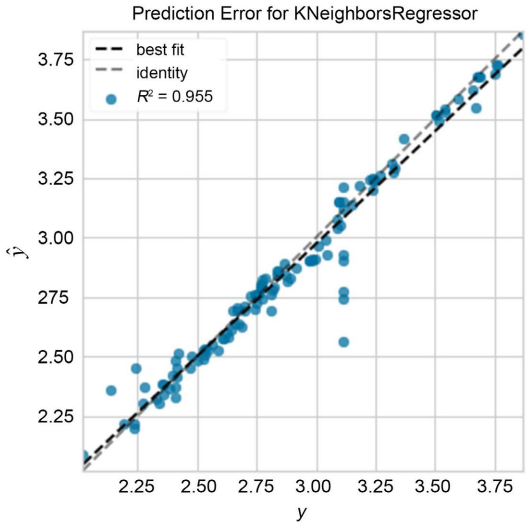
(a)


Prediction Error for SVR
best fit
identity
R^2 = 0.868
4.5
4.0
3.5
3.0
2.5
$\hat{y}$
2.5 3.0 3.5 4.0 4.5
y

(b)

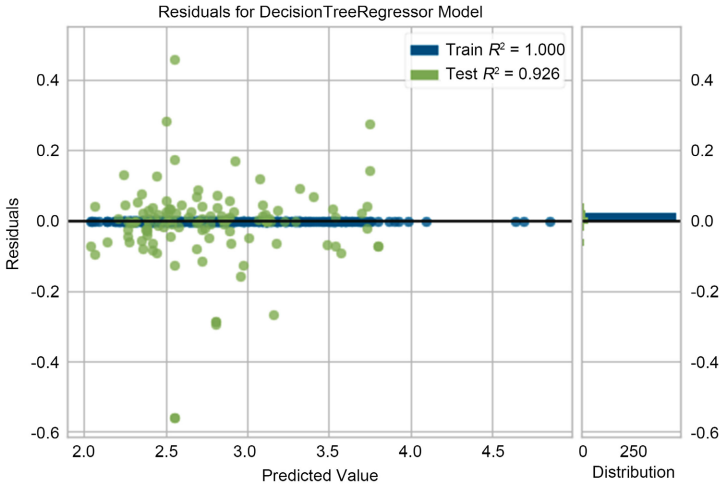

Prediction Error for LinearRegression
best fit
identity
R^2 = 0.918
4.00
3.75
3.50
3.25
3.00
275
250
225
200
$\hat{y}$
2.00 2.25 2.50 2.75 3.00 3.25 3.50 3.75 4.00
y

(c)

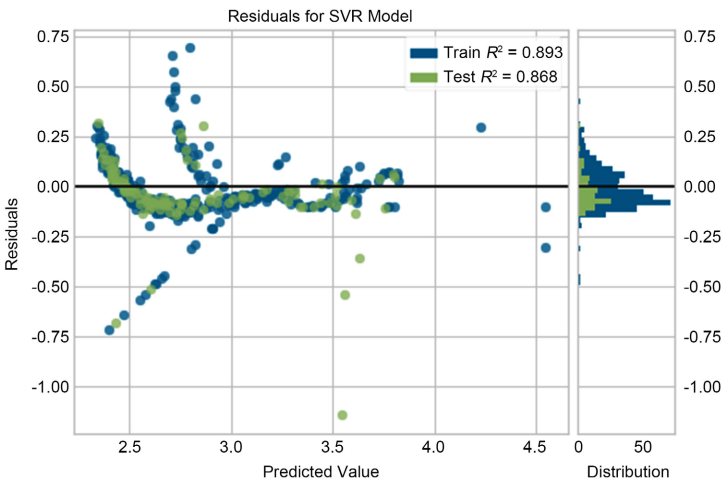


(d)

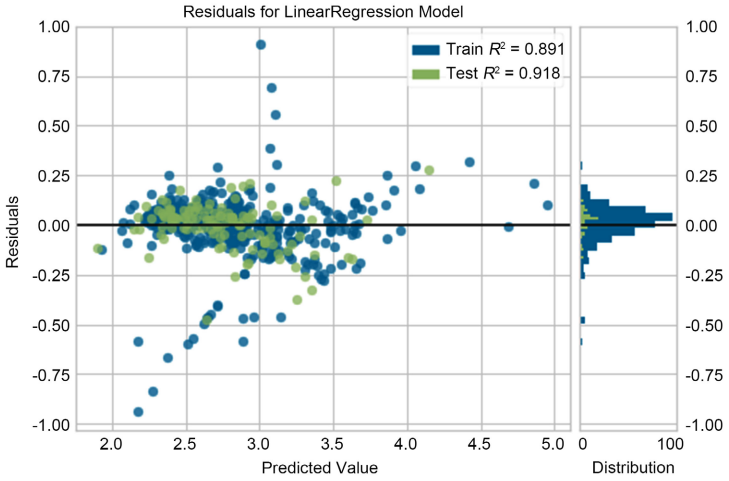
Figure 5. Prediction error for the models (a) Decision Tree Regression (b) Support Vector Regression (c) Linear Regression (d) K-Nearest Neighbor.



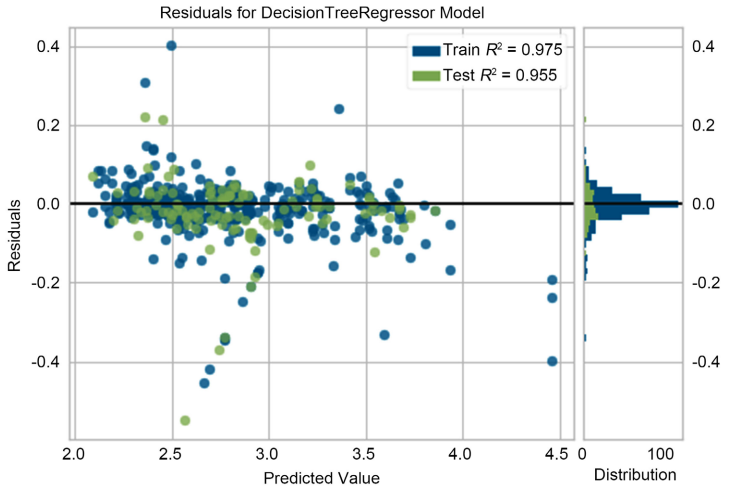
(a)



(b)



(c)



(d)

Figure 6. Prediction error for the models (a) Decision Tree Regression (b) Support Vector Regression (c) Linear Regression (d) K-Nearest Neighbor.

testing) (see **Figures 4-6**) compared to other algorithms. This algorithm gives us the flexibility to be able to define how many errors are acceptable in our models and find an appropriate line (hyperplane) to fit the data.

5.5. K-Nearest Neighbor Regression

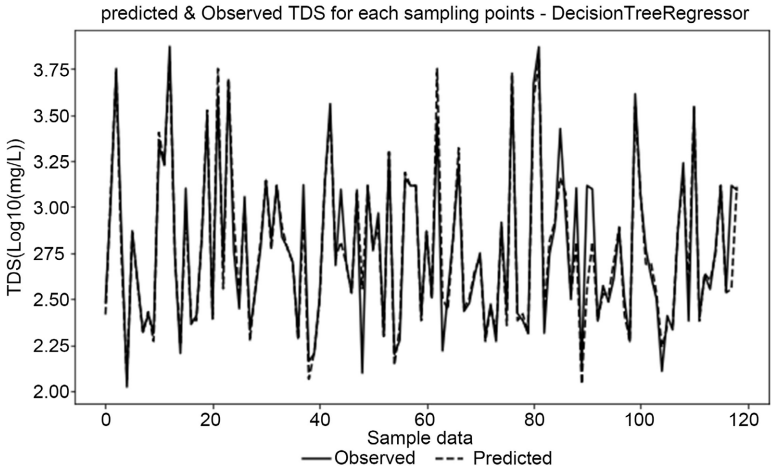
Through the simple process of averaging observations within the same neighborhood, KNN regression is a non-parametric technique that estimates the relationship between independent variables and the continuous result (Youssef et al., 2022). To minimize the mean-squared error, the analyst must determine the size of the neighborhood, or it can be selected via cross-validation. In the KNN regression model from this study, one can deduce that the KNN algorithm performed well (see **Figures 4-6**), better than the SVR algorithm. The coefficient of determination is 0.98 and 0.98, for both the training and testing datasets. 0.03 and 0.04 were obtained for MAE for both training and testing,

respectively. 0.06 and 0.06 were obtained for the RMSE for the training and testing datasets. In summary, the KNN models appear to perform well based on the metrics above.

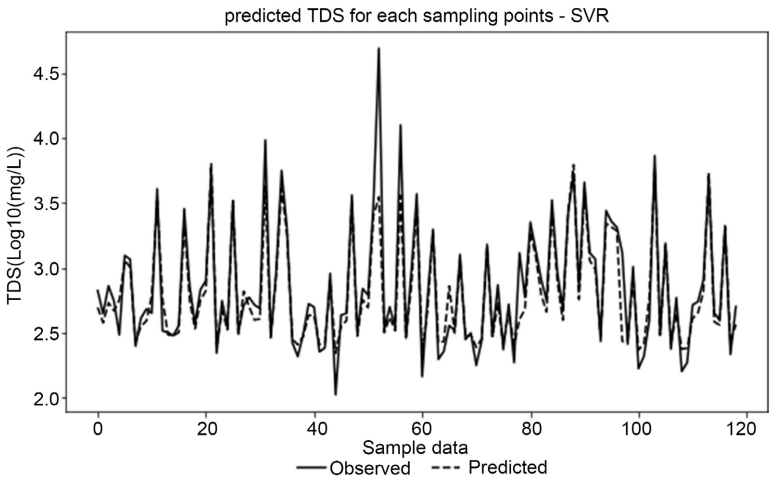
6. Comparing the Different Machine Learning Models

A comparative assessment of the four machine learning models reveals that they are generally aligned with the observed and estimated input parameters of the study area. The DTR gave the best prediction model among the four algorithms (Figure 3). DT showed the best model with zero values for MAE, RMSE, and MSE and a coefficient of determination of 1 for training and 0.050, 0.090, and 0.010 for MAE, RMSE, and MSE respectively for the testing with an R^2 of 0.954 (Figure 7).

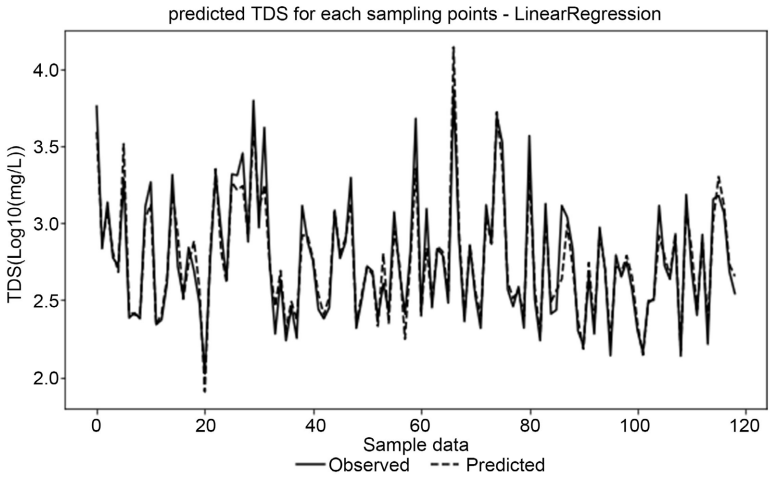
Linear regression performance, though impressive, performed low when compared to the other three models. It has the lowest coefficient of determination of 0.902 and 0.848 for training and testing. All the models agree with their observed and estimated TDS parameters in the study area. Linear regression has the most error in predicting TDS in the dataset. The results from prediction error



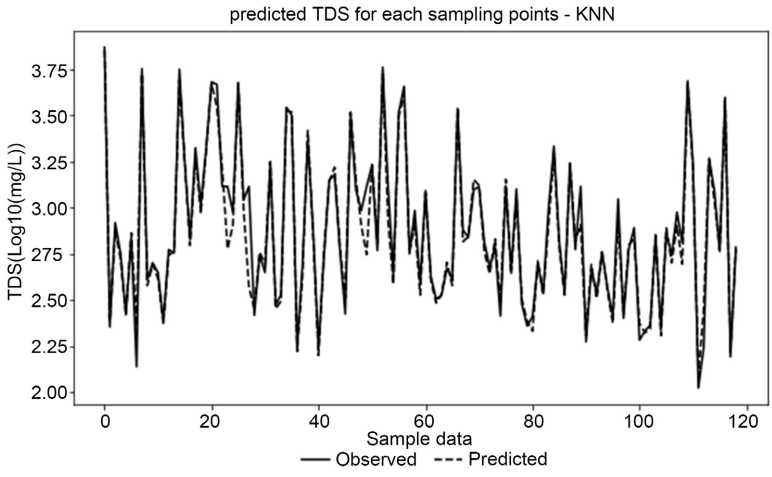
(a)



(b)



(c)



(d)

Figure 7. Performance of the models (a) Decision Tree Regression (b) Support Vector Regression (c) Linear Regression (d) K-Nearest Neighbor.

can obtain the overall performance of the prediction for each of the regression algorithms. If the difference between the model and the real data becomes smaller, the prediction will be more accurate (Nouraki et al., 2021).

It's crucial to remember that the qualities of the features, the characteristics of the dataset, and the problem you are attempting to solve can all have an impact on how well a machine learning algorithm performs. Depending on the task at hand and the characteristics of the data, some algorithms may perform better than others in specific situations.

6.1. Sinkhole Formation

The two main factors behind the formation of the two sinkholes (winksink #1 and winksink #2) are natural and anthropogenic. Anthropogenic factors deal with human-induced causes, including oil and gas development and agricultural influences on the land surfaces. The natural factors are comprised of the geologic

processes (tectonism, chemical weathering) of the formation from the time was deposited.

6.2. Groundwater of Winkler County

The Santa Rosa sandstone-saturated zones and the Cenozoic alluvium serve as Winkler County's main freshwater aquifers. They are a component of vast aquifers that support a sizable portion of West Texas and eastern New Mexico. Precipitation is the only source of fresh water for the aquifers; some of the water comes from precipitation inside the county and some from regions to the north and northeast. The foot of Santa Rosa is the lowest point of fresh water in the county, and Santa Rosa gets its water primarily from water seeping through the Cenozoic deposits where the two formations meet (Garza and Wesselman, 1962).

With increased urbanization comes a reduction in natural land surfaces through which water percolation can occur during precipitation. Because anthropogenic changes to ground cover have reduced the surface area where infiltration can occur, there is a greater chance that new catchment areas and runoff paths will emerge in certain areas. Groundwater flow patterns and recharge zones may be affected by the concentrated intake of groundwater in certain areas. Increased inter-aquifer flow and active dissolution are caused by the increased water burden in the strata above, higher withdrawal, and lowered hydraulic head in the Rustler and other underlying aquifers (English et al., 2020). The process is sped up by changing the land's topography by establishing additional catchment regions and reducing recharge areas. Dissolution started far into the tertiary and is still going on today. Water percolates through the overburden of the atmosphere as it flows through, dissolving CO_2 to create weak carbonic acid (H_2CO_3). The underlying carbonate and evaporite formation dissolves as the H_2CO_3 permeates it. TDS concentrations can be used to assess the amount of dissolution. Dissolution rates within these evaporite layers can be up to three times faster than those seen in limestone lithology, especially with the addition of water that is undersaturated relative to the dominant salt within the evaporite. This dissolution occurs especially quickly around the areas of focused infiltration. The growth of sinkholes might significantly speed up as additional waters seep through the underlying Salado and Castille, increasing their number and pace of growth.

6.3. Hydrocarbon Exploration and Development in Winkler County

The removal of early land cover types, the construction of substantial concrete building surfaces and the creation of new transportation networks during oil and gas production alter the land cover map. The oil and gas installations frequently encroach on grasslands, agricultural fields, and forests. Workers at oil and gas sites often relocate to and congregate in a nearby city or town. Additionally, deforestation brought on by the extraction of oil or the conversion of grassland into developed land has significant negative consequences on the ecosystem, in-

cluding the loss of habitat for several animal and plant species, and may play a substantial role in localized climate change. This regional disruption can result in changes in soil moisture and severe temperatures, which can be harmful to both humans and plants.

Due to brine and crude oil leaks, oil and gas production significantly pollutes the soil (Meng, 2017). In oil-producing regions, spills of crude oil from well sites and pipelines are additional significant causes of soil pollution. Such spills may travel upward through the soil and contaminate the atmosphere or downward through the soil and damage the groundwater. Brines can contain naturally occurring radioactive materials, hazardous trace metals, and elevated salinity levels. According to Carls, Fenn, & Chaffey (1995), the purposeful, unintentional, and incidental discharge of drilling fluids, crude petroleum, and refined petroleum products is the main cause of soil pollution at oil and gas production wells. In contrast to modern, highly engineered, and scientific approaches, the shattering of rocks with explosives generated uncontrolled fractures throughout the zone. These fractures improved oil production permeability while also providing the necessary channel for unsaturated waters to have ongoing access to the weakening-prone underlying formations (Lambert, 1983).

Regulations were far different from what they are now, and there are few records of wells being plugged and abandoned. There is not enough information to know where or how many of these poorly filled wells are, which have opened new channels for water to move through the overburden. The well itself may turn into a direct route to the delicate underlying layers as a pipe corrodes with age. Potential paths also include wells that did not use secondary and tertiary casing and cementing. Meteoric waters will use these paths to flow if the wellbore is deficient in cement along the whole borehole above the targeted zone:

In West Texas, population increase coupled with an arid climate has led to the shrinkage of the aquifer system in this area, which serves as the primary source of drinking water, agricultural production, and oil exploration. The importance of accurate prediction of water quality parameters in monitoring the pollution caused by indiscriminate use of water in agriculture, oil mining, etc., in West Texas cannot be overemphasized. This study employed four machine learning methods (LR, DT, SVR, and KNN) to predict the TDS concentration of 593 samples from wells in Winkler County. Fifteen input parameters were used to estimate TDS for all the machine learning methods. All the methods showed impressive figures for coefficients of determination, MAE, MSE, RMSE, and EVS, with DTR performing exceptionally well in both the training and testing datasets. Machine learning models are, therefore, an appropriate alternative to physically based modeling in predictive conditions (Nouraki et al., 2021). This statement shows that machine learning models can reduce the cost of water quality monitoring and assessment and save time. In addition, machine learning models can provide the foundation for managers, engineers, and the government to design, manage, and make significant decisions in different aquifers (Nouraki et al., 2021).

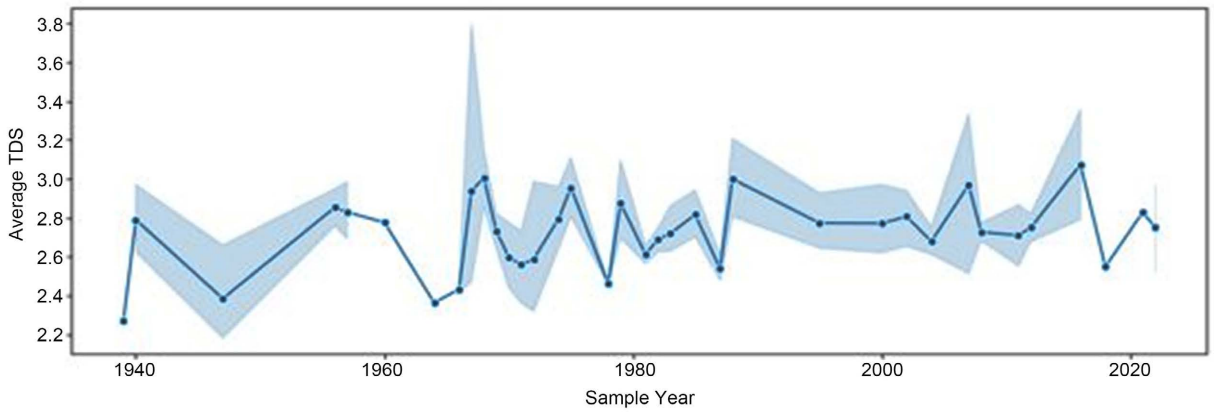


Figure 8. Innovative trend Analysis for average TDS concentration in Winkler County over seventy years (1940-2022).

Figure 8 shows the trend of TDS over seventy years. It can be deduced from the figure that TDS concentration peaked between 1950 and 1980. This is a result of increased oil and gas activities in combination with natural processes like erosion of underground formation (Salado Formation), which comprises halite and evaporites, going on in the area. From 1980 to today, the TDS concentration went steadily down. This can be attributed to government policies and increased awareness of the danger of destructive mining practices like uncased wells.

Furthering this study will include using hybrid machine learning models to estimate TDS, Specific conductance, and Total hardness of different aquifer systems in West Texas to determine their water quality.

7. Conclusion

Groundwater is a vital supply of potable water worldwide. The current study examined and contrasted four machine learning algorithms to predict groundwater quality in Winkler County, revealing concerning levels of TDS recommendations for safe drinking water. The results of our study indicate that among the four algorithms tested, the decision tree (DT) algorithm demonstrates a very accurate prediction model for estimating TDS levels in the study area. This underscores the use of decision tree as a valuable tool in groundwater quality prediction, offering insights that can inform resource management and environmental conservation efforts in the region. By employing decision trees and other machine learning algorithms in this work, which have shown high coefficient of determination, low MAE, and RSME, we can enhance our ability to identify patterns and factors contributing to groundwater quality degradation, thereby aiding in the prevention of future sinkhole occurrences. Moving forward, continued research into refining and optimizing machine learning techniques for groundwater analysis will be crucial in addressing environmental challenges and ensuring the sustainable management of water resources. The formation of sinkholes in Winkler County poses significant challenges, with potential causes including the dissolution of underground Castile and Salado formations by groundwater and the influence of oil and gas activities. In terms of remediation techniques,

the insights provided by machine learning models can inform the implementation of targeted measures to mitigate sinkhole formation. This may include continuous monitoring of groundwater quality parameters that can give early warnings of potential sinkhole formation by detecting changes in hydrological conditions, proper land use planning, which can inform land use planning decisions, helping to avoid development in areas susceptible to sinkhole formation based on predictive analysis of groundwater quality, proper management of injection wells used in oil and gas activities can help mitigate the risk of sinkhole formation by minimizing the introduction of contaminants into groundwater systems and promoting vegetation growth in vulnerable areas can help stabilize soil and reduce erosion, thereby mitigating the risk of sinkhole formation.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adams, J. E. (1944). Upper Permian Ochoan Series of Delaware Basin, West Texas, and Southeastern New Mexico. *AAPG Bulletin, 28,* 1596-1625. https://doi.org/10.1306/3D9336D8-16B1-11D7-8645000102C1865D

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, S., Ehteram, M., & Elshafie, A. (2019). Machine Learning Methods for Better Water Quality Prediction. *Journal of Hydrology, 578,* Article ID: 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

Alizadeh, M. J., Kavianpour, M. R., Danesh, M., Adolf, J., Shamshirband, S., & Chau, K. (2018). Mechanics Effect of River Flow on the Quality of Estuarine and Coastal Waters Using Machine Learning Models. *Engineering Applications of Computational Fluid Mechanics, 12,* 810-823. https://doi.org/10.1080/19942060.2018.1528480

Anderson, R. Y., & Kirkland, D. W. (1980). Dissolution of Salt Deposits by Brine Density Flow. *Geology, 8,* 66-69. https://doi.org/10.1130/0091-7613(1980)8<66:DOSDBB>2.0.CO;2

Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River Water Quality Index Prediction and Uncertainty Analysis: A Comparative Study of Machine Learning Models. *Journal of Environmental Chemical Engineering, 9,* Article ID: 104599. https://doi.org/10.1016/j.jece.2020.104599

Ashworth, J. B. (1990). *Evaluation of Ground-Water Resources in Parts of Loving, Pecos, Reeves, Ward, and Winkler Counties, Texas.* TWDB (Texas Water Development Board) Report 317.

Baryakh, A., & Fedoseev, A. K. (2011). Sinkhole Formation Mechanism. *Journal of Mining Science, 47,* 404-412. https://doi.org/10.1134/S1062739147040022

Baumgardner, R. W., Hoadley, A. D., & Goldstein, A. G. (1982). *The Wink Sinks; a Case History of Evaporite Dissolution and Catastrophic Subsidence: Formation of the Wink Sink, a Salt Dissolution and Collapse Feature.* Winkler County, Texas, Bureau of Economic Geology, The University of Texas at Austin, Report of Investigations No. 114. https://doi.org/10.23867/RI0114D

Berhe, B. A. (2020). Evaluation of Groundwater and Surface Water Quality Suitability for Drinking and Agricultural Purposes in Kombolcha Town Area, Eastern Amhara Region, Ethiopia. *Applied Water Science, 10,* Article No. 127. https://doi.org/10.1007/s13201-020-01210-6

Breiman, L., & Ihaka, R. (1984). *Nonlinear Discriminant Analysis via Scaling and ACE.* Department of Statistics, University of California.

Carls, E., Fenn, D., & Chaffey, S. (1995). Soil Contamination by Oil and Gas Drilling and Production Operations in Padre Island National Seashore, Texas, USA. *Journal of Environmental Management, 45,* 273-286. https://doi.org/10.1006/jema.1995.0075

Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020). A Review of the Artificial Neural Network Models for Water Quality Prediction. *Applied Sciences, 10,* Article 5776. https://doi.org/10.3390/app10175776

Czajkowski, M., Krzysztof, J., & Kretowski, M. (2023). Steering the Interpretability of Decision Trees Using Lasso Regression—An Evolutionary Perspective. *Information Sciences, 638,* Article ID: 118944. https://doi.org/10.1016/j.ins.2023.118944

Djarum, D. H., Ahmad, Z., & Zhang, J. (2021). River Water Quality Prediction in Malaysia Based on Extra Tree Regression Model Coupled with Linear Discriminant Analysis (LDA). *Computer Aided Chemical Engineering, 50,* 1491-1496. https://doi.org/10.1016/B978-0-323-88506-5.50230-8

English, S., Heo, J., & Won, J. (2020). Investigation of Sinkhole Formation with Human Influence: A Case Study from Wink Sink in Winkler County, Texas. *Sustainability, 12,* Article 3537. https://doi.org/10.3390/su12093537

Frumkin, A., Ezersky, M., Al-Zoubi, A., Akkawi, E., & Abueladas, A. R. (2011). The Dead Sea Sinkhole Hazard: Geophysical Assessment of Salt Dissolution and Collapse. *Geomorphology, 134,* 102-117. https://doi.org/10.1016/j.geomorph.2011.04.023

Garza, S., & Wesselman, J. B. (1962). *Geology and Ground-Water Resources of Winkler County, Texas.* Water Supply Paper 1582.

Gutiérrez, F., Fabregat, I., Roqué, C., Carbonel, D., Guerrero, J., García-Hermoso, F., Zarroca, M., & Linares, R. (2016). Sinkholes and Caves Related to Evaporite Dissolution in a Stratigraphically and Structurally Complex Setting, Fluvia Valley, Eastern Spanish Pyrenees. Geological, Geomorphological and Environmental Implications. *Geomorphology, 267,* 76-97. https://doi.org/10.1016/j.geomorph.2016.05.018

Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water Quality Prediction Using Machine Learning Methods. *Water Quality Research Journal, 53,* 3-13. https://doi.org/10.2166/wqrj.2018.025

Heithecker, R. E. (1932). *Some Methods of Separating Oil and Water in West Texas Fields, and the Disposal of Oil-Field Brines in the Hendrick Oil Field, Texas.* University of Michigan Library.

Hichem, T., Abdeltif, A., Belhadj, A. E., & Zhang, J. (2022). Modeling the Organic Matter of Water Using the Decision Tree Coupled with Bootstrap Aggregated and Least-Squares Boosting. *Environmental Technology & Innovation, 27,* Article ID: 102419. https://doi.org/10.1016/j.eti.2022.102419

Johnson, K. (1986). *Salt Dissolution and Collapse at the Wink Sink in West Texas.* Office of Nuclear Waste Isolation, Battelle Memorial Institute.

Johnson, K. (2005). *Salt Dissolution and Subsidence or Collapse Caused by Human Activities.* Humans as Geologic Agents. https://doi.org/10.1130/2005.4016(09)

Jones, T. S. et al. (1949). *East-West Cross Section through Southern Permian Basin of West Texas*. West Texas Geological Society Publication, 49-17.

Kanade, V. (2023). *What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022.* https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression

Kayanan, M., & Wijekoon, P. (2020). Stochastic Restricted LASSO-Type Estimator in the Linear Regression Model. *Journal of Probability and Statistics, 2020,* Article ID: 7352097. https://doi.org/10.1155/2020/7352097

Kim, J. W., Lu, Z., & Degrandpre, K. (2016). Ongoing Deformation of Sinkholes in Wink, Texas, Observed by Time-Series Sentinel-1A SAR Interferometry (Preliminary Results). *Remote Sensing, 8,* Article 313. https://doi.org/10.3390/rs8040313

Kim, J.W., Lu, Z., & Kaufmann, J. (2019). Evolution of Sinkholes over Wink, Texas, Observed by High-Resolution Optical and SAR Imagery. *Remote Sensing of Environment, 222,* 119-132. https://doi.org/10.1016/j.rse.2018.12.028

Kirkland, D. W., & Evans, R. (1976). Origin of Limestone Buttes, Gypsum Plain, Culberson County, Texas. *The American Association of Petroleum Geologists Bulletin, 60,* 2005-2018. https://doi.org/10.1306/C1EA3A1E-16C9-11D7-8645000102C1865D

Lambert, S. J. (1983). *Dissolution of Evaporites in and around the Delaware Basin, Southeastern New Mexico and West Texas.* Sandia National Labs.

Lang, W. B. (1939). Salado Formation of the Permian Basin: GEOLOGICAL NOTES. *AAPG Bulletin, 23,* 1569-1572. https://doi.org/10.1306/3D93312A-16B1-11D7-8645000102C1865D

Meng, Q. M. (2017). The Impact of Fracking on the Environment: A Total Environmental Study Paradigm. *Science of the Total Environment, 580,* 953-957. https://doi.org/10.1016/j.scitotenv.2016.12.045

Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R. S., Gleason, B. W. et al. (2019). *Global Historical Climatology Network—Daily (GHCN-Daily), Version 3 Daily Summaries.* https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00861#

Meyer, J. E., Wise, M. R., & Kalaswad, S. (2012). *Pecos Valley Aquifer, West Texas: Structure and Brackish Groundwater.* TWDB (Texas Water Development Board) Report 382.

Mohd Zebaral Hoque, J., Ab Aziz, N. A., Alelyani, S., Mohana, M., & Hosain, M. (2022). Improving Water Quality Index Prediction Using Regression Learning Models. *International Journal of Environmental Research and Public Health, 19,* Article 13702. https://doi.org/10.3390/ijerph192013702

Noori, R., Karbassi, A. R., Mehdizadeh, H., Vesali-Naseh, M., & Sabahi, M. S. (2011). A Framework Development for Predicting the Longitudinal Dispersion Coefficient in Natural Streams Using an Artificial Neural Network. *Environmental Progress & Sustainable Energy, 30,* 439-449. https://doi.org/10.1002/ep.10478

Nouraki, A., Alavi, M., & Golabi, M. (2021). Prediction of Water Quality Parameters Using Machine Learning Models: A Case Study of the Karun River, Iran. *Environmental Science and Pollution Research, 28,* 57060-57072. https://doi.org/10.1007/s11356-021-14560-8

Prabowo, R., Bambang, A. N., & Sudarno, S. (2021). Water Quality Index of Well Water in the Converted Agricultural Land. *Jurnal Pendidikan IPA Indonesia, 10,* 560-570. https://doi.org/10.15294/jpii.v10i4.31790

Shah, K. A., & Joshi, G. S. (2017). Evaluation of Water Quality Index for River Sabarmati, Gujarat, India. *Applied Water Science, 7,* 1349-1358. https://doi.org/10.1007/s13201-015-0318-7

Shi, Y., Tang, Y., Lu, Z., Kim, J. W., & Peng, J. (2019). Subsidence of Sinkholes in Wink, Texas from 2007 to 2011 Detected by Time-Series InSAR Analysis. *Geomatics, Natural Hazards and Risk, 10,* 1125-1138. https://doi.org/10.1080/19475705.2019.1566786

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer, New York. https://doi.org/10.1007/978-1-4757-2440-0_1

Youssef, N., Leon, S., & Tim, L. (2022). *DNNR: Differential Nearest Neighbors Regression.* Cornell University.

Zakir, H. M., Sharmin, S., Akter, A., & Rahman, M. S. (2022). Assessment of Health Risk of Heavy Metals and Water Quality Indices for Irrigation and Drinking Suitability of Waters: A Case Study of Jamalpur Sadar Area, Bangladesh. *Environmental Advances, 2,* Article ID: 100005. https://doi.org/10.1016/j.envadv.2020.100005