Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

Agricultural Data Analysis with Weather and Soil Using Machine Learning Models

S. Dhanavel[1], A. Murugan[2]

[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Science, Periyar Arts College, Cuddalore, (Deputed from Annamalai University, Annamalainagar) Tamil Nadu, India

(Received: 02 September 2023 Revised: 14 October Accepted: 07 November)

KEYWORDS

Machine Learning, Data Mining, Agricultural Data, Correlation, Test Statistics.

ABSTRACT:

Data mining involves extracting valuable insights, patterns, correlations, and trends from large datasets stored in databases or data repositories. It employs statistical, mathematical, and computational techniques to unveil information not easily visible to humans. The main objective is to convert raw data into actionable knowledge. Creating a decision tree involves recursively dividing the data into subsets based on different attributes, aiming to achieve homogeneity (for classification) or minimize variance (for regression) within each subset. This paper considers agriculture and its soil chemical-related dataset for applying data mining techniques to find suitable variables for future predictions. The five decision tree approaches are decision stump, M5P, random forest, random tree, and REP tree**.** Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

1. Introduction and Literature Review

Data mining encompasses diverse methods like clustering, classification, regression analysis, association rule mining, anomaly detection, text mining, and time series analysis. Its significance spans various fields such as business, marketing, finance, healthcare, and scientific research. Data mining provides valuable insights by analyzing structured and unstructured data (e.g., text, images, images, and videos). However, ethical considerations, privacy, and security should be prioritized, especially when dealing with sensitive or personal information [1] and [2].

The problem of knowledge acquisition and efficient knowledge exploitation is also prevalent in agriculture. One of the methods for knowledge acquisition from the existing agricultural databases is classification. In agricultural decision-making, weather and soil characteristics play an essential role. This research aimed to assess the various classification techniques of data mining and apply them to a soil science database to establish if meaningful relationships can be found. A large data set of soil databases is extracted from the Soil Science & Agricultural Department, Kanchipuram, and National Informatics Centre, Tamil Nadu. Data mining techniques have never been applied to Tamil Nadu soil data sets. The research compares the different classifiers, and the outcome of this research could improve the management and systems of soil uses throughout many fields, including agriculture, horticulture, and environmental and land use management [3].

New data mining techniques and apply them to a soil science database to establish if meaningful relationships can be found. A large data set of Soil database is extracted from the Department of Soil Sciences and Agricultural Chemistry, S V Agricultural College, Tirupati; The database contains measurements of soil profile data from various locations of Chandragiri Mandal, Chittoor District. The research establishes whether Soils are Classified Using multiple data mining techniques. In addition, a comparison was made between the Naive Bayes classification and analyze the most effective method. The research outcome may benefit agriculture, soil management, and the environment [4].

The page offers an overview of soil chemical properties, encompassing concentrations of specific chemicals (such as phosphorus, nitrogen,

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

carbon, major cations like calcium, magnesium, sodium, potassium, sulfur, trace metals, and elements), pH, cation exchange capacity, base saturation, salinity, sodium adsorption ratio, enzymes, and electrical conductivity. These properties influence nutrient cycling, biological activity, soil formation, pollutant fate, and erosion. Additionally, the page addresses the effects of human activities, discusses stormwater applications, and provides links to related topics, including information on sampling, testing, and soil health assessments. [5].

Data mining is a valuable tool for the practice of examining large pre-existing databases to generate previously unknown helpful information; in this paper, the input for the weather data set denotes specific days as a row, attributes denote weather conditions on the given day, and the class indicates whether the conditions are conducive to playing golf. Attributes include Outlook, Temperature, Humidity, Windy, and Boolean Play Golf class variables. All the data are considered for training purpose, and it is used in the seven-classification algorithm likes J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP) and Random Forest (RF) are used to measure the accuracy. Out of seven classification algorithms, the Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [6].

A simple cost analysis showed that soil chemical analytical costs are too significant for economical use in precision agriculture, with higher prices in Australia than in the United States. This paper concludes that developing field-deployed, 'on-the-go' proximal soil sensing systems and scanners is essential for large-scale implementation of precision agriculture. These sensing systems or scanners aim to overcome current problems of high cost, labor, time, and, to some extent, imprecision of soil sampling and analysis to represent the spatial variability of the measured properties more efficiently and accurately [7].

RH and RHC amendment increased the saturated hydraulic conductivity, saturated water content, plant available water, and field capacity but decreased the bulk density of soil. Crop growth components at harvest revealed that the highest plant height was recorded in RH4%. However, for the panicle length, weight, and number of tillers, the highest value was found in RHC2%, 14.2 cm, 4.0 g, and 28.8 cm, respectively. Furthermore, the number of panicles, 1000-grain weight, and grain yield were also highest in RHC2%, 22.4 g, and 4.41 t/ha, respectively. However, the highest value was found in RH4%, 79.0, and 88.5 for the number of grains per panicle and percentage of filled grains. The grain yield increased by 38%, 28%, 18%, and 22%, and the biological yield increased by 27%, 18%, 14%, and 16% for RHC2%, RHC4%, RH2%, and RH4%, respectively, compared to that of the control; however, the significant difference was found only for RHC2% for both. The harvest index increased under all RH and RHC application rates compared to the control [8].

Investigate the spatial variability of selected soil properties, such as soil pH, electrical con-ductility (EC1:5), carbonate content (CaCO3) and total organic matter (TOM), and soil texture using different conventional analytical methods. The spatial distribution of these soil properties was elaborated using the Kriging method. The results showed that the soil is alkaline, with a pH ranging from 7.20 to 8.78. TOM varies from 0.11 to 1.05 dS/m, and its texture varies from sand to loam content. Nonetheless, EC values fall within the slight, moderate, and vigorous degrees of salinity. According to these results, it can be concluded that the soil is saline. Interrelationships between the parameters analyzed and the different samples were investigated by multivariate analyses, principal component analysis (PCA), and hierarchical cluster analyses (HCA). HCA classified the five groups, which were compared to PCA visualizations. The Boxplots show that the values fractions (sand, silt, clay) were very variable. Pearson correlation coefficient indicates a high correlation between carbonate content and pH, organic matter, and silt fraction [9].

Data mining is discovering hiding information that efficiently utilizes the prediction by stochastic sensing concept. This paper proposes an efficient assessment of groundwater level, rainfall, population, food grains, and enterprises dataset by adopting stochastic modeling and data mining approaches. Firstly, the novel data assimilation analysis is proposed to predict the groundwater level effectively. Experimental results are done, and the various expected groundwater level estimations indicate the sternness of the approach [10] and [11].

The input for the chronic disease data denotes a specific location as a row; attributes denote topics, questions, data values, low confidence limit, and high confidence limit. All the data are considered for training and testing using five classification algorithms. In this paper, the authors present the various analysis and accuracy of five different decision tree algorithms; the M5P

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

decision tree approach is the best algorithm to build the model compared with other decision tree approaches [12].

2. Backgrounds and Methodologies

A data mining decision tree is a widely used machine learning technique for classification and regression tasks. It visually depicts a sequence of decisions and their possible outcomes in a tree-like structure. Each internal node represents a decision based on a specific feature, and each branch corresponds to the potential result of that decision. The tree's leaf nodes represent the final decision or the predicted outcome. The "CART" (Classification and Regression Trees) algorithm is the most used algorithm for building decision trees [13].

2.1 Decision Stump

A decision stump is a straightforward machine-learning model for binary classification tasks. It is the most basic form of a decision tree with a single level or depth of one. In a decision stump, only one feature (attribute) of the data is used to decide, and it splits the data into two subsets based on a threshold value for that feature. The decision stump can be understood as choosing one part, choosing a threshold, assigning classes, and predicting.

Steps involved in decision stump

Step 1. Selecting the feature
Step 2. Choosing the threshold
Step 3. Assigning class labels
Step 4. Making predictions
Step 5. Training and evaluation

2.2 M5P

M5P is a machine learning algorithm used for regression tasks. It is an extension of the decision tree-based model called M5, which Ross Quinlan developed. The M5 algorithm combines decision trees and linear regression to create more accurate and flexible regression models. M5P, specifically, stands for M5 Prime. It enhances the original M5 algorithm to improve its predictive performance. M5P uses a tree-based model to divide the data into subsets based on feature values recursively and then fits linear regression models to each of these subsets. The result is a piecewise linear regression model, where different linear regressions are used for other regions of the input feature space.

Steps involved in the M5P

Step 1. Building the initial decision tree (M5 model): Recursive Binary Splitting and Pruning (optional)
Step 2. Linear Regression Model: Leaf Regression Models and Model Parameters
Step 3. Piecewise Linear Regression: Piecewise Prediction
Step 4. Model Evaluation: Training and Testing.

2.3 Random Forest

Random Forest is a popular machine learning ensemble method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy, robustness, and ability to handle complex datasets. Random Forest is widely used in various domains, including data science, machine learning, and pattern recognition. The main idea behind Random Forest is to create an ensemble (a collection) of decision trees and combine their predictions to make more accurate and stable predictions. The following steps describe what Random Forest works like.

- Bootstrap Aggregating (Bagging)
- Decision Tree Construction
- Voting for Classification, Averaging for Regression

The key advantages of Random Forest are:

- Reduced overfitting
- Robustness
- Feature Importance

Steps involved in Random Forest

Random Forest is an ensemble learning method combining multiple decision trees to make more accurate and robust predictions for classification and regression tasks. The steps involved in building a Random Forest are as follows:

Step 1. Data Bootstrapping
Step 2. Random Feature Subset Selection
Step 3. Decision Tree Construction
Step 4. Ensemble of Decision Trees
Step 5. Out-of-Bag (OOB) Evaluation
Step 6. Hyperparameter Tuning (optional)

2.4 Random Tree

In machine learning, a Random Tree is a specific type of decision tree variant that introduces randomness during construction. Random Trees are similar to traditional decision trees but differ in how they select the splitting features and thresholds at each node. The primary goal of introducing randomness is to create a more diverse set of decision trees, which can help reduce overfitting and

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

improve the model's generalization performance. Random Trees are commonly used as building blocks in ensemble methods like Random Forests. The critical characteristics of Random Trees are as follows:

- ❖ Random Feature Subset
- ❖ Random Threshold Selection
- ❖ No Pruning
- ❖ Ensemble Methods

Steps involved in Random Tree

Step 1. Data Bootstrapping:
Step 2. Random Subset Selection for Features:
Step 3. Decision Tree Construction:
Step 4. Voting (Classification) or Averaging (Regression):

2.5 REP Tree

REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm constructs a decision tree using a combination of incremental pruning and error-reduction techniques. The key steps involved in building a REP Tree are as follows:

- ❖ Recursive Binary Splitting
- ❖ Pruning
- ❖ Repeated Pruning and Error Reduction

Steps involved in REP Tree

REP Tree (Repeated Incremental Pruning to Produce an Error Reduction Tree) is a machine learning algorithm for classification and regression tasks. It is an extension of decision trees that incorporates pruning to reduce overfitting and improve the model's generalization performance. Below are the steps involved in building a REP Tree.

Step 1. Recursive Binary Splitting
Step 2. Pruning
Step 3. Repeated Pruning and Error Reduction
Step 4. Model Evaluation

2.6 Correlation coefficient

The correlation coefficient, often denoted by the symbol "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is commonly used to assess the degree to which changes in one variable are associated with changes in another. The correlation coefficient takes values between -1 and 1:

- A correlation coefficient of +1 indicates a perfect positive linear relationship where the two variables increase together.
- A correlation coefficient -1 indicates a perfect negative linear relationship, where one variable increases as the other decreases.
- A correlation coefficient close to 0 indicates a weak or no linear relationship between the variables.

Steps involved in the Correlation coefficient.

Calculating the correlation coefficient between two variables involves several steps. Let's assume we have two datasets: X and Y, each containing n data points. Here are the steps to calculate the correlation coefficient:

Step 1. Compute the means of X and Y
Step 2. Calculate the deviations from the norm for both X and
Step 3. Compute the product of the variations for each data point
Step 4. Sum up the effects of deviations
Step 5. Calculate the square of the variations for both X and Y
Step 6. Sum up the squared deviations for X and Y
Step 7. Compute the correlation coefficient (r). The correlation coefficient (r) is calculated using the formula:

$$r = \Sigma ((X - \bar{x})(Y - \bar{y})) / \sqrt{(\Sigma (X - \bar{x})^2 * \Sigma (Y - \bar{y})^2)} \quad \text{... (1)}$$

Step 8. Interpret the correlation coefficient. The resulting value of "r" will be between -1 and 1, indicating the strength and direction of the linear relationship between X and Y:

- ❖ r ≈ +1: A strong positive correlation (as X increases, Y increases).
- ❖ r ≈ -1: A strong negative correlation (as X increases, Y decreases).
- ❖ r ≈ 0: Little to no linear correlation (no consistent relationship between X and Y).

2.7 Mean Absolute Error

Mean Absolute Error (MAE) is a metric used to measure the average absolute difference between predicted and actual (true) values in a regression problem. It is commonly used to assess the accuracy of a regression model's predictions [14]. The formula to calculate Mean Absolute Error (MAE) is as follows:

$$MAE = \Sigma |(\text{Actual Value} - \text{Predicted Value})| / n \quad \text{... (2)}$$

Where:

- ❖ Σ represents the summation symbol, which sums up the values for all data points.
- ❖ | | denotes the absolute value, ensuring the differences are positive.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

In this formula:

- Actual Value: Refers to the true value of the target variable (ground truth) for a specific data point.
- Predicted Value: Refers to the value predicted by the regression model for the same data point.
- n: Represents the total number of data points in the dataset.

Calculating the Mean Absolute Error (MAE) involves the following steps

Step 1. Collect Data

Step 2. Compute the Absolute Errors: This is done by taking the absolute value of the difference: |Actual Value - Predicted Value|.

Step 3. Sum the Absolute Errors: Add up all the absolute errors obtained from Step 2 to get the total sum of fundamental mistakes.

Step 4. Calculate the Mean: Divide the total sum of absolute errors obtained from Step 3 by the total number of data points (n) to find the Mean Absolute Error (MAE).

MAE = Sum of Absolute Errors / n

Step 5. Interpret the MAE: The resulting MAE value represents the average absolute difference between the model's predictions and the actual values. It indicates how far, on average, the model's predictions deviate from the actual values.

Step 6. Compare the MAE (Optional): compare the MAE of different regression models or iterations of the same model to assess their performance. A model with a lower MAE is considered to have better predictive accuracy.

2.8 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a commonly used metric to assess the accuracy of a regression model's predictions. It measures the average magnitude of the errors between the predicted and actual (true) values, considering both the direction and magnitude of the errors. The formula to calculate Root Mean Squared Error (RMSE) is as follows [15]:

$$RMSE = \sqrt{\frac{\Sigma (\text{Actual Value} - \text{Predicted Value})^2}{n}} \quad \text{... (3)}$$

Where:

- Σ represents the summation symbol, which sums up the values for all data points.
- (Actual Value - Predicted Value) ² denotes the squared difference between the actual and predicted values for each data point.
- n is the total number of data points in the dataset.

Steps involved in Root mean squared error

Step 1. Collect Data - Gather the actual (true) values and the corresponding predicted values from a regression model. Each data point should have the real deal (ground truth) and the model's expected value.

Step 2. Compute the Squared Errors - For each data point, calculate the squared difference between the actual value and the predicted value:

(Actual Value - Predicted Value) ²

Step 3. Sum the Squared Errors - Add up all the squared errors obtained from Step 2 to get the total sum of squared errors.

Step 4. Calculate the Mean - Divide the total sum of squared errors obtained from Step 3 by the total number of data points (n) to find the Mean Squared Error (MSE).

MSE = Sum of Squared Errors / n

Step 5. Calculate the Square Root - Take the square root of the Mean Squared Error (MSE) obtained from Step 4 to get the Root Mean Squared Error (RMSE).

RMSE = √MSE

Step 6. Interpret the RMSE - The resulting RMSE value represents the square root of the average of the squared errors between the model's predictions and the actual values. It measures the average magnitude of the errors, giving more weight to more significant errors due to the squaring. A lower RMSE value indicates that the model's predictions are, on average, closer to the actual values, while a higher RMSE value indicates more significant prediction errors.

Step 7. Compare the RMSE (Optional) - You can compare the RMSE of different regression models or different iterations of the same model to assess their performance. A model with a lower RMSE is considered to have better predictive accuracy.

2.9 Relative Absolute Error (RAE)

Relative Absolute Error (RAE), also known as Mean Absolute Percentage Error (MAPE), is a metric used to evaluate the accuracy of predictions in regression tasks. It measures the

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

average percentage difference between the absolute and actual (valid) values, providing a relative measure of the prediction errors [16]. The formula to calculate Relative Absolute Error (RAE) is as follows:

RAE = (Σ |Actual Value - Predicted Value| / Σ |Actual Value|) * (100 / n)... (4)

Where:

- Σ represents the summation symbol, which sums up the values for all data points.
- | | denotes the absolute value, ensuring the differences are positive.
- n is the total number of data points in the dataset.

In this formula, actual value refers to the true value of the target variable (ground truth) for a specific data point. Predicted value refers to the value predicted by the regression model for the same data point.

Steps involved in Relative absolute error

Step 1. Collect Data - Gather the actual (true) values and the corresponding predicted values from a regression model. Each data point should have the real deal (ground truth) and the model's expected value.

Step 2. Compute the Absolute Errors - For each data point, calculate the absolute difference between the actual and predicted values: |Actual Value - Predicted Value|.

Step 3. Compute the Absolute Percentage Errors - For each data point, calculate the absolute percentage difference between the absolute error and the actual value:

|Actual Value - Predicted Value| / |Actual Value|.

Step 4. Sum the Absolute Percentage Errors - Add up all the absolute percentage errors obtained from Step 3 to get the total sum of fundamental percentage errors.

Step 5. Calculate the Mean Absolute Percentage Error (MAPE) - Divide the total sum of absolute percentage errors obtained from Step 4 by the total number of data points (n) to find the Mean Absolute Percentage Error (MAPE).

MAPE = (Sum of Absolute Percentage Errors) / n

Step 6. Convert MAPE to Relative Absolute Error (RAE) - Multiply the MAPE obtained from Step 5 by 100 to get the Relative Absolute Error (RAE).

RAE = MAPE * 100

Step 7. Interpret the RAE - The resulting RAE value represents the average percentage difference between the absolute errors and the actual values. It provides a relative measure of the prediction errors, allowing for easier comparison across datasets and target variable scales.

Step 8. Compare the RAE (Optional) - Compare the RAE of different regression models or different iterations of the same model to assess their performance.

2.10 Root Relative Squared Error (RRSE)

"Root Relative Squared Error" is not a standard or widely recognized metric in statistics or machine learning. It appears to be a combination of the terms "Root Mean Squared Error (RMSE)" and "Relative Absolute Error (RAE)." It's possible that the time was created or used in a specific context or literature, but it is not a commonly used or established metric. For clarity, let's briefly define the two individual metrics mentioned:

- Root Mean Squared Error (RMSE): As explained earlier, RMSE is a commonly used metric to evaluate the accuracy of regression models. It measures the average magnitude of the errors between the predicted and actual values, considering both the direction and extent of the errors. The formula to calculate RMSE is:

RMSE = √(Σ (Actual Value - Predicted Value)² / n)

- Relative Absolute Error (RAE): Also known as Mean Absolute Percentage Error (MAPE), RAE measures the average percentage difference between the absolute errors and the actual (true) values, providing a relative measure of the prediction errors. The formula to calculate RAE is:

RAE = (Σ |Actual Value - Predicted Value| / Σ |Actual Value|) * (100 / n)

As there is no established metric called "Root Relative Squared Error," it's crucial to use standard evaluation metrics such as RMSE, RAE (MAPE), or others that are well-known and have clear interpretations in the context of your specific problem.

Steps involved in Root relative squared error

Root Relative Squared Error is not a standard or commonly used metric in statistics or machine learning. Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE) are widely used to evaluate the accuracy of regression models. These are well-known metrics with clear interpretations:

Step 1. Root Mean Squared Error (RMSE):

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

- Collect the actual (true) values and the corresponding predicted values from a regression model.
- For each data point, calculate the squared difference between the actual and predicted values: (Actual Value - Predicted Value) ².
- Sum up all the squared errors to get the total sum of squared errors.
- Divide the total sum of squared errors by the total number of data points (n) to find the Mean Squared Error (MSE).
- Take the square root of the Mean Squared Error (MSE) to get the Root Mean Squared Error (RMSE).

Step 2. Relative Absolute Error (RAE) or Mean Absolute Percentage Error (MAPE):

- Collect the actual (true) values and the corresponding predicted values from a regression model.
- For each data point, calculate the absolute difference between the actual and predicted values: |Actual Value - Predicted Value|.
- For each data point, calculate the absolute percentage difference between the absolute error and the actual value: |Actual Value - Predicted Value| / |Actual Value|.
- Sum up all the absolute percentage errors to get the total sum of absolute percentage errors.
- Divide the total sum of absolute percentage errors by the total number of data points (n) to find the Mean Absolute Percentage Error (MAPE) or Relative Absolute Error (RAE).

Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository [17]. The agriculture and soil types of data include 45 parameters which have different categories of data like Meteorology, Types of Land, Chemical fertilizers, Soil types, and Soil Information [18]. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. Soil Chemical dataset

District	tsp	Noncalcareous Grey Floodplain Soil
Year	mp	Noncalcareous Dark Grey
Area	DAP	Floodplain
avg_rainfall	inundationland_Highland	Soil
max_temperature	inundationland_mediumhighland	Peat
min_temperature	inundationland_lowland	Made-Land
aus	inundationland_mediumlowland	Noncalcareous Brown Floodplain
Oman	inundationland_verylowland	Soil
paddy	Miscellaneous Land	Shallow Red-Brown Terrace Soil
wheat	Calcareous Alluvium	Deep Red-Brown Terrace Soil
potato	Noncalcareous Alluvium	Brown Mottled Terrace Soil
jute	Acid Basin Clay	Shallow Grey Terrace Soil
humidity	Calcareous Brown Floodplain Soil	Deep Grey Terrace Soil
storm	Calcareous Grey Floodplain Soil	Grey Valley Soil
urea	Calcareous Dark Grey Floodplain Soil	Brown Hill Soil
		Grey Piedmont Soil
		Soil moisture

Table 2: Machine Learning Models with Correlation coefficient

Decision Tree	R2 Score
Decision Stump	0.8642
M5P	0.9797
Random Forest	0.9989
Random Tree	0.9986
REP Tree	0.9924

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

Table 3: Machine Learning Models with Mean Absolute Error and Root Mean Squared Error

Decision Tree	MAE	RMSE
Decision Stump	46591.7588	56105.9985
M5P	14187.7693	22370.4088
Random Forest	3577.6776	5698.9189
Random Tree	1441.1155	6083.9728
REP Tree	6120.3275	14219.5551

Table 4: Machine Learning Models with Relative Absolute Error (%) and Root Relative Squared Error (%)

Decision Tree	RAE (%)	RRSE (%)
Decision Stump	47.0565	49.7113
M5P	14.3293	19.8207
Random Forest	3.6134	5.0494
Random Tree	1.4555	5.3905
REP Tree	6.1814	12.5989

Table 5: Machine Learning Models with Time Taken to Build Model (Seconds)

Decision Tree	Time (Seconds)
Decision Stump	0.0200
M5P	0.3600
Random Forest	0.1600
Random Tree	0.0100
REP Tree	0.0400

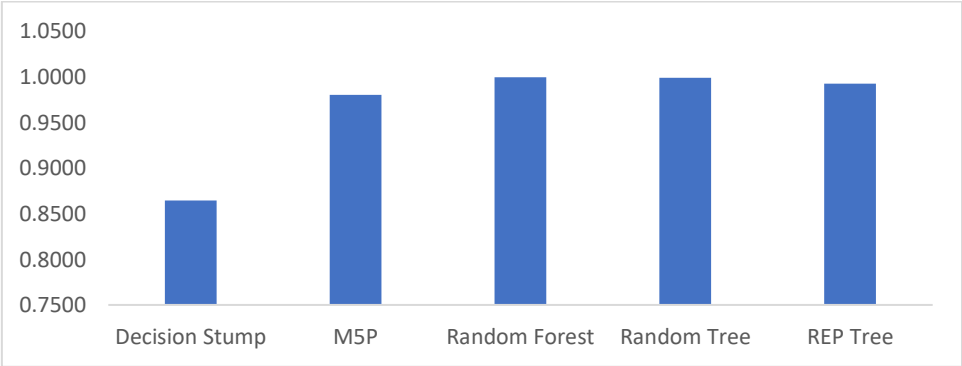

Fig. 1. R2 Score for Machine Learning Approaches

Journal of Chemical Health Risks
www.jchr.org
JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

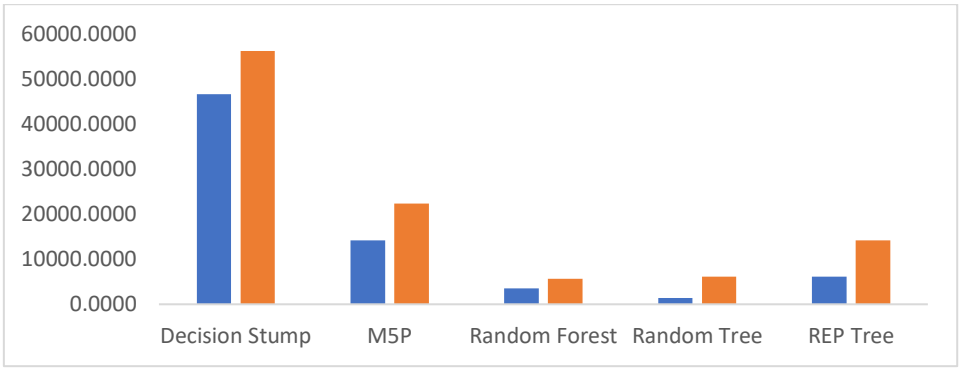

Fig. 2. Machine Learning Models with MAE and RMSE

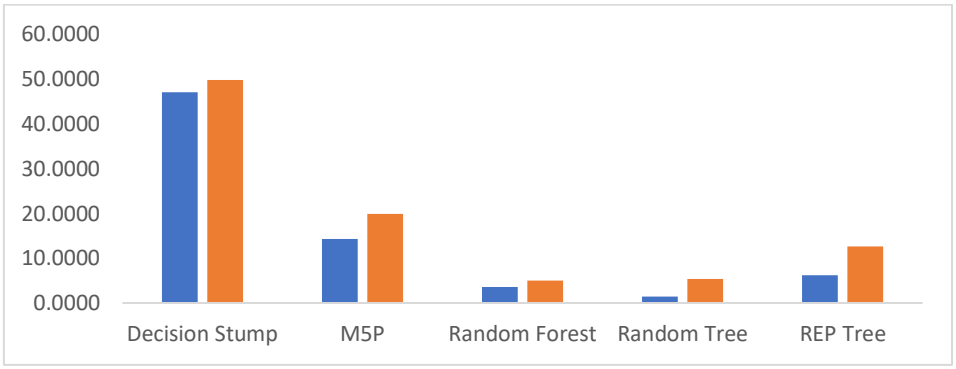

Fig. 3. Machine Learning Models with RAE (%) and RRSE (%)

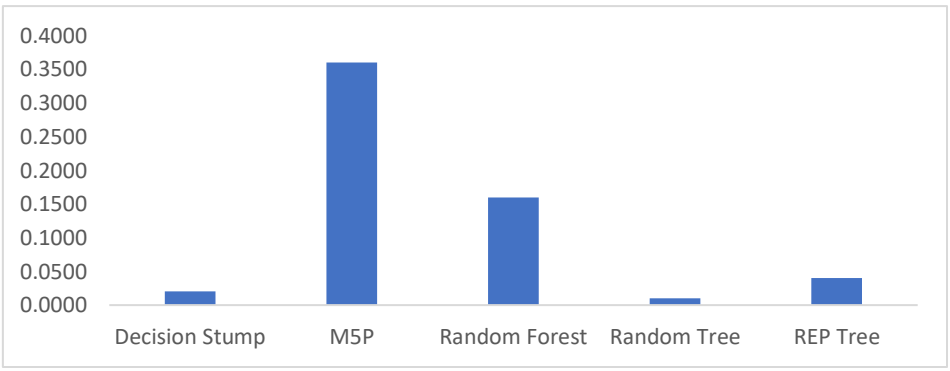

Fig. 4. Machine Learning Models and its Time Taken to Build the Model (Seconds)

3. Result and Discussion

Table 1 explains 45 parameters which include different categories of data like Meteorology, Types of Land, Chemical fertilizers, Soil Type, and Soil Information. Based on the dataset, it is evident that five different machine learning decision tree approaches are used to find the hidden patterns and which is the best or influencing parameter to decide future predictions. Related results and numerical illustrations are shown between Table 1 to Table 5 and Figure 1 to Figure 4.

They are based on Equation 1, Table 2, and Figure 1, which is used to find the R2 score or correlation coefficient by comparing 44 parameters. Numerical illustrations suggest that there may be a significant difference from one parameter to another. In this case, using five different decision tree approaches among these results, random forest, random tree, REP tree, and M5P, return a robust, strong positive correlation of nearly 0.9 when using different soil properties. Decision stump machine learning approaches also produce positive correlations of 0.8642.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

Further data analysis revealed a gradual improvement in test scores over time. The MAE is used to find model errors using Equations 2. Five machine-learning algorithms will be used in this case. The random forest, random tree, and REP tree return a minimum error for using MAE test statistics. Decision stump and M5P approaches produce the maximum error. The RMSE (root mean square error) measures the difference between predicted and actual values using Equation 3. In this case, the random forest, random tree, and REP tree return a minimum error for using MAE test statistics. Decision stump and M5P approaches produce the maximum error. The related numerical illustration is shown in Table 3 and Figure 2.

Relative Absolute Error (RAE) measures accuracy using equation 4 to compare the difference between predicted and actual values in percentage. In this research, taking into consideration five ML classification algorithms, except decision stump remaining four algorithms return a minimum error. Similar error approaches are reflected in RRSE. Similar numerical illustrations are shown in Table 4 and Figure 3.

Time taken is one of the significant tasks in machine-learning approaches. Based on Table 5 and Figure 4, random tree take minimum time to build the model. Subsequently, the Decision stump, REP tree, and random forest take the time to make the models. Finally, M5P takes the maximum time to build the model. Similar approaches are reflected in the mentioned visualization.

4. Conclusion and Further Research

It is essential to consider the limitations of this study. The sample size of each group was relatively small, which could impact the generalizability of the results. Additionally, other variables could influence agriculture with soil chemical performance. The findings presented in this study contribute to our understanding that all the parameters return robust positive correlations. In this research, the maximum of the machine learning approaches returns a minimum error with less processing time. Future studies can build upon these, finding the suitable variable for future prediction with increased accuracy using different machine learning and decision tree approaches.

Reference

1. United Nations Framework Convention on Climate Change. (2021). What is climate change? Retrieved June 8, 2021, from https://unfccc.int/topics/what-is-climate-change
2. McMichael, A.J., et al., Climate Change and Human Health: Risks and Responses. Lancet, 2006. 367(9513): p. 859-869.
3. Vamanan, R. and Ramar, K., 2011. Classification of agricultural land soils a data mining approach. International Journal on Computer Science and Engineering, ISSN, pp.0975-3397.
4. Bhargavi, P. and Jyothi, S., 2009. Applying Naive Bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security, 9(8), pp.117-122.
5. https://stormwater.pca.state.mn.us/index.php?title=Soil_chemical_properties_and_processes
6. Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the Weka tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.
7. Rossel, R.V. and McBratney, A.B., 1998. Soil chemical analytical accuracy and costs: implications from precision agriculture. Australian Journal of Experimental Agriculture, 38(7), pp.765-775.
8. Mishra, A., Taing, K., Hall, M.W. and Shinogi, Y., 2017. Effects of rice husk and rice husk charcoal on soil physicochemical properties, rice growth, and yield. Agricultural Sciences, 8(9), pp.1014-1032.
9. Oumenskou, H., El Baghdadi, M., Barakat, A., Aquit, M., Ennaji, W., Karroum, L.A. and Aadraoui, M., 2019. Multivariate statistical analysis for spatial evaluation of physicochemical properties of agricultural soils from Beni-Amir irrigated perimeter, Tadla plain, Morocco. Geology, Ecology, and Landscapes, 3(2), pp.83-94.
10. Rajesh, P., Karthikeyan, M. and Arulpavai, R., 2019, December. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model. In AIP Conference Proceedings (Vol. 2177, No. 1). AIP Publishing.
11. Rajesh, P. and Karthikeyan, M., 2019. Data mining approaches to predict the factors that affect agriculture growth using stochastic models. International Journal of Computer Sciences and Engineering, 7(4), pp.18-23.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2008-2018 | ISSN:2251-6727

12. Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., 2019. Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), pp.1472-1477.
13. Kohavi, R., & Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).
14. Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/
15. S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, "Root mean square error (RMSE): A comprehensive review," International Journal of Applied Mathematics and Statistics, vol. 59, no. 1, pp. 42–49, 2019.
16. Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566
17. https://www.kaggle.com/datasets/tanhim/agricultural-dataset-bangladesh-44-parameters
18. Gaihre, Y.K., Singh, U., Islam, S.M., Huda, A., Islam, M.R., Satter, M.A., Sanabria, J., Islam, M.R. and Shah, A.L., 2015. Impacts of urea deep placement on nitrous oxide and nitric oxide emissions from rice fields in Bangladesh. Geoderma, 259, pp.370-379.