A STUDY ON VARIABLE SELECTIONS AND PREDICTION FOR SUSTAINABLE DEVELOPMENT GOALS USING DATA MINING WITH MACHINE LEARNING APPROACHES

B. Santhosh Kumar[1], Dr. P. Rajesh[2*]

Article History: Received: 17.09.2022 **Revised:** 02.11.2022 **Accepted:** 18.12.2022

Abstract

Machine learning, a subset of artificial intelligence (AI), centers on creating algorithms and statistical models. ML tools empower computer systems to acquire knowledge and autonomously make predictions without explicit programming. Data mining involves extracting valuable insights, patterns, correlations, and trends from large datasets stored in data repositories. The main objective is to convert raw data into actionable knowledge. This paper considers sustainable development goals-related datasets for applying ML techniques to find suitable variables for future predictions. The ten familiar machine learning approaches are Gaussian processes, linear regression, random forest, and REP tree**.** Numerical illustrations are provided to prove the proposed results with test statistics.

Keywords: Machine learning, data mining, decision tree, accuracy, and SDG.

[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar, Tamil Nadu, India
Email: santhoshcdm@gmail.com

[2*]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram – 608 102, (Deputed from Dept. of Computer and Information Science, Annamalai University, Annamalainagar-608 002) Tamil Nadu, India.
[*]Email: rajeshdatamining@gmail.com

DOI: 10.48047/ecb/2022.11.12.164

1. Introduction and Literature Review

The analysis shows that IAMs cover the SDGs related to climate because of their design. It also shows that most IAMs cover several other areas that are related to resource use and the Earth system as well. Some other dimensions of the 2030 Agenda are also covered, but socio-political and equality goals, and others related to human development and governance, are not well represented. Some of these are difficult to capture in models. Therefore, it is necessary to facilitate a better representation of heterogeneity (greater geographical and sectoral detail) by using different types of models (e.g. national and global) and linking different disciplines (especially social sciences) together. Planned developments include increased coverage of human development goals and contribute to policy coherence [1].

Robustness of Earth Observation data for continuous planning, monitoring, and evaluation of SDGs. The scientific world has made commendable progress by providing geospatial data at various spatial, spectral, radiometric, and temporal resolutions enabling usage of the data for various applications. This paper also reviews the application of big data from earth observation and citizen science data to implement SDGs with a multi-disciplinary approach. It covers literature from various academic landscapes utilizing geospatial data for mapping, monitoring, and evaluating the earth's features and phenomena as it establishes the basis of its utilization for the achievement of the SDGs [2].

Various works by researchers on linear and polynomial regression and compares their performance using the best approach to optimize prediction and precision. Almost all of the articles analyzed in this review is focused on datasets; in order to determine a model's efficiency, it must be correlated with the actual values obtained for the explanatory variables [3].

Explores the spatial relationship between mining and agricultural activities towards meeting the United Nations (UN) Agenda 2030 Sustainable Development Goals (SDGs) in Northwest Ghana. Agenda 2030 SDGs highlight the importance of poverty reduction, livelihood enhancement, and food security. A state's natural resources include both nonagricultural and agricultural resources. There is a renewed interest in large-scale mining in Ghana, entering into previously underexplored areas in the Northwest, an area dominated by agriculture. With the emergence of mining in this region, this study combines both satellite imagery, covering years 2000, 2010 and 2018, and ground truthing data to conduct baseline studies and assess changes in land use over time. We compared known data sets and field knowledge with satellite data to objectively measure changes in the distribution of surface water, farmlands and grasscover over time. The study finds increasing areas of surface water, abundant grasscover and farmlands within leases in the area. These growing abundance of land use and land cover types provide opportunities for commercial livestock keeping, extensive and intensive crop farming. The classified satellite images revealed the existence of more farmlands and potential cultivable areas than reported by agriculture extension offices. Most of these areas overlap with mining concessions and could be modelled for commercial food production and local job creation. The occurrence of mining and agricultural activities in rural subsistence farming communities often indicate conflict. However, a co-exitence of both sectors has a strong opportunity to drive inclusive growth for smallholder farmers; reduce poverty, generate income and uphold sustainable development [4].

Research aims to provide a comprehensive overview of how these domains are currently interacting, by illustrating the impact of Big Data on sustainable development in the context of each of the 17 UN SDGs [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].

The results demonstrate that citizen science is “already contributing” to the monitoring of 5 SDG indicators, and that citizen science “could contribute” to 76 indicators, which, together, equates to around 33%. Our analysis also shows that the greatest inputs from citizen science to the SDG framework relate to SDG 15 Life on Land, SDG 11 Sustainable Cities and Communities, SDG 3 Good Health and Wellbeing, and SDG 6 Clean Water and Sanitation. Realizing the full potential of citizen science requires demonstrating its value in the global data ecosystem, building partnerships around citizen science data to accelerate SDG progress, and leveraging investments to enhance its use and impact [7].

Though urbanization is often linked to development gains, some regions in Asia, Latin America, and Sub-Saharan Africa have grown in urban population, while remaining bereft of basic services like reliable electricity. Daytime optical remote sensing has tracked urban land cover change for decades, but there have been few studies that have monitored whether infrastructure is keeping pace with demographic and land transitions. Here, we explore how fusing multi-temporal population and land data with nighttime lights data, derived from the Suomi-NPP VIIRS Day Night Band, can add to our understanding of urban infrastructural transitions. We classify urban changes in India and the US, using these three measures in tandem to create a typology of urban development processes. When compared against survey data, our results indicate the classification can track rural electrification and identify growing informal settlements with inadequate infrastructure, and is therefore useful for monitoring progress towards two Sustainable Development Goals: Goal 7.1 (ensure universal access to affordable, reliable and modern energy services) and Goal 11.1 (ensure access for all to adequate, safe and affordable housing and basic services and upgrade slums). The classification results also illustrate the diversity of urban development processes, and how uni-dimensional measures of urbanization, greatly under-represent urban change, particularly in high-income countries [8].

The objective of this paper is to identify synergetic SDGs using Boosted Regression Trees model which is a machine learning and data mining technique. In this study, contributions of all SDGs to form the SDG index are identified and a “what-if” analysis is conducted to understand the significance of goal scores. Findings show that SDG3, “Good health and well-being”, SDG4, “Quality education”, and SDG7, “Affordable and clean energy”, are the most synergetic goals, when their scores are >60%. The findings of this research will help decision-makers implement effective strategies and allocate resources by prioritizing synergetic goals [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 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 Gaussian Processes

A Gaussian process is a stochastic process, such that every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed. Gaussian Process (GP) is a powerful supervised machine learning method that is largely used in regression settings. This method is desirable in practice regime.it performs pretty well in small data regime, highly interpretable, estimates the prediction uncertainty.

This last point is what sets GP apart from many other machine learning techniques: for a GP model, its prediction f(x) at a location x is not a deterministic value, but rather a random variable following a normal distribution,

$$\mathbf{f(x) \sim N(\mu(x), \sigma^2(x)) \qquad \ldots (1)}$$

Here, $\mu(x)$ denotes the prediction mean, and $\sigma^2(x)$ is the prediction variance, which indicates the prediction uncertainty.

2.2 Linear Regression

Linear regression is a statistical technique employed to comprehend and forecast the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition. The core idea of linear regression is to find the best-fitting straight line (also called the "regression line") through a scatterplot of data points. This line represents a linear equation of the form:

$$\mathbf{y = m_x + b \qquad \ldots (1)}$$

Where y is the dependent variable, x is the independent variable, m is the slope of the line, representing how much, y changes for a unit change in x and b is the y-intercept, indicating the value of y when x is 0.

2.4 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 steps involved in random forest.

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)

2.6 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 incremental pruning and error-reduction techniques. the key steps involved in building a rep tree are as follows recursive binary splitting, pruning and repeated pruning and error reduction. 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

2.7 Accuracy and Performance Metrics

The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is comparing the original target with the predicted one and using metrics like R-squared, MAE, MSE, and RMSE to explain the errors and predictive ability of the model [14]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [15] and [16].

R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The values from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} \qquad \ldots (2)$$

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \qquad ... (3)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \qquad ... (4)$$

Relative Absolute Error (RAE) is a metric used in statistics and data analysis to measure the accuracy of a forecasting or predictive model's predictions. It is particularly useful when dealing with numerical data, such as in regression analysis or time series forecasting.

$$\mathrm{RAE} = \frac{\Sigma|y_i-\hat{y}_i|}{\Sigma|y_i-\bar{y}|} \qquad \dots (5)$$

Root Relative Squared Error (RRSE) is another metric used in statistics and data analysis to evaluate the accuracy of predictive models, especially in the context of regression analysis or time series forecasting.

$$\mathrm{RRSE} = \sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y})^2}} \qquad \dots (6)$$

Equations 2 to 6 used to find the model accuracy which is used to find the model performance and error. Where Y_i represents the individual observed (actual) values, $\hat{Y}_i$ represents the corresponding individual predicted values, $\bar{Y}$ represents the mean (average) of the observed values and Σ represents the summation symbol, indicating that you should sum the absolute differences for all data points.

3. Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository. The sustainable development goals dataset includes 17 parameters which have different categories of data like SDG 1, SDG 2, SDG 3, SDG 4, SDG 5, SDG 6, SDG 7, SDG 8, SDG 9, SDG 10, SDG 11, SDG 12, SDG 13, SDG 14, SDG 15, SDG 16, SDG 17 [17]. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. a) Sample SDG dataset (SDG 1 to 9)

SGD 1	SGD 2	SGD 3	SGD 4	SGD 5	SGD 6	SGD 7	SGD 8	SGD 9
63.8	40.3	37.9	6.9	24.6	49.3	52	31.2	3.8
99.7	48.8	81	84.8	73.5	93.1	82.3	55.6	27.3
99.5	50.7	76.6	77.8	49.3	61.9	79.8	53.7	28.8
60.4	44.1	31.4	44.8	61.8	57.5	35.6	48.9	7.7
99.8	69.1	80.8	88.6	78.5	100	85.6	61.5	38.3
98.9	54.1	76.9	88.2	69.9	73.4	93.6	55.5	29.2
99.9	59.6	95.3	96.2	79.9	86.7	84	86.8	81.5
99.6	80.2	93.7	82.2	77.1	94.4	89.1	87.6	79.5
100	57.9	74.3	90.1	68.2	76.4	85.7	61.2	38.7
99.9	66.9	88.9	84.4	58.1	32.9	88.7	86	50.2
97	46.8	60.3	63.9	57	77.3	42.9	63.1	18
99.9	59.5	81.4	93.3	82.1	95.4	83.4	79.1	31.5
99.7	80.8	93.1	89.4	84	77	85.9	85.5	73.4
84.7	58.5	74.5	83.1	60.7	90.4	89.3	56.3	18.7
45.8	49.3	46.9	38.7	43.1	58.3	10.7	64.5	7.6

Table 1. b) Sample SDG dataset (SDG 10 to 17)

SGD 10	SGD 11	SGD 12	SGD 13	SGD 14	SGD 15	SGD 16	SGD 17
65.4	39.4	82.1	87.7	48.4	50.8	47.6	55
59.9	75.4	74.4	75.4	38.8	75.1	65.8	60.6
88.9	68.7	81.6	90.8	44.4	62.2	65.2	74
45.2	44.2	80.6	87.4	43.1	63.8	38.8	47.3
39.8	83.6	69.9	89.1	44.5	50.5	58.8	56.6
50.2	75.7	80.9	92.2	48.4	59.9	71.1	59.5
77.1	84.5	50.7	23.3	55	37.7	81.9	59
87.5	83.9	51.5	83.2	52.4	64.6	86.7	66
68.4	79	77.2	82.3	48.4	66.3	72.3	57.2
62.5	61.8	73.9	55.9	40.6	53.4	68.7	47.5
76.5	39.4	77.5	81.1	51	54.7	60.7	41.7
85.9	81.1	82.1	91.1	48.4	70.2	67.9	59.5
93.6	84.5	51.9	79	47.2	75.2	83.9	58.9
36.5	70.9	63.8	85	24.6	41.6	47.1	73.8
36.3	50.2	81.3	84.3	41.5	62.2	54.5	57.5

Table 2. Gaussian Processes

SDG	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root Relative Squared error (%)	Time taken
SDG1	0.7570	10.7891	15.5142	60.5032	64.9212	0.0900
SDG2	0.8087	6.3556	7.9684	56.4148	58.4708	0.0200
SDG3	0.9079	6.2709	7.9939	38.4758	41.7510	0.0300
SDG4	0.8441	9.9433	12.6395	52.1977	53.3965	0.0500
SDG5	0.7809	7.8040	9.8072	60.5298	62.2532	0.0300
SDG6	0.6273	10.0442	13.1917	70.8344	77.1452	0.0600
SDG7	0.8727	11.1496	13.7127	46.9840	48.5559	0.0700
SDG8	0.8114	7.2905	9.5593	55.1861	58.0641	0.0200
SDG9	0.8902	8.2738	10.9634	41.4499	45.3846	0.0300
SDG10	0.5094	16.1885	19.8621	85.5955	85.5929	0.0300
SDG11	0.7511	7.8653	10.4544	59.4478	65.6435	0.0500
SDG12	0.7796	5.5673	7.7992	55.1258	62.0414	0.0200
SDG13	0.4094	7.7115	10.6479	94.8512	91.4682	0.0600
SDG14	0.0610	7.3074	10.3778	103.7681	101.7286	0.0200
SDG15	0.2620	10.2820	12.6388	98.7763	96.5176	0.0200
SDG16	0.7917	6.6322	8.3026	60.2161	60.8872	0.2600
SDG17	0.0350	11.4996	15.4311	100.7687	101.8449	0.0600

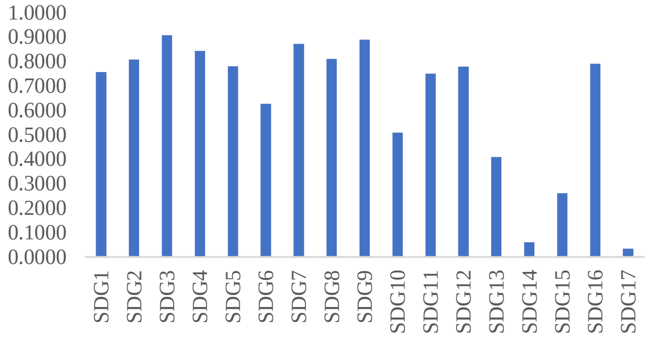

Fig. 1: Gaussian process with R2 Score

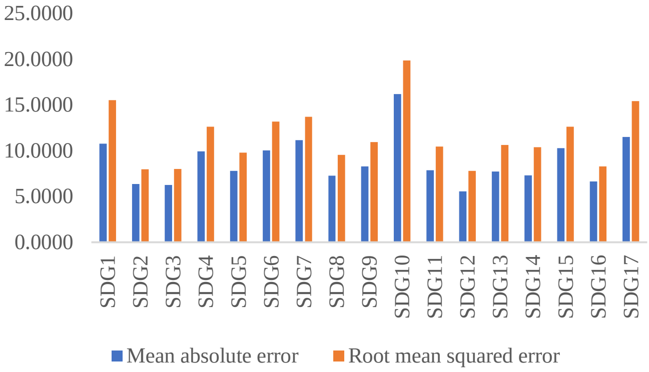

Fig. 2. Gaussian process with MAE and RMSE

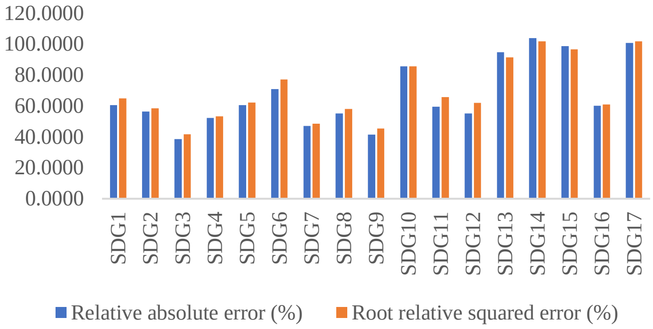

Fig. 3. Gaussian process with RAE (%) and RRSE (%)

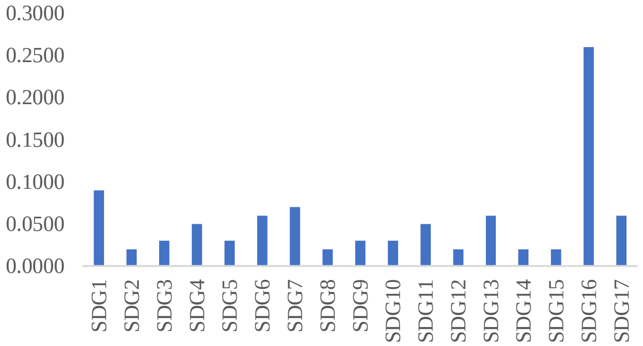

Fig. 4. Gaussian process and the time taken to build the model (seconds)

Table 3: Linear Regression

SDG	**Correlation coefficient**	**Mean absolute error**	**Root Mean Squared error**	**Relative Absolute error (%)**	**Root relative squared error (%)**	**Time taken**
SDG1	0.8029	10.3127	14.1679	57.8313	59.2874	0.1200
SDG2	0.8412	6.0565	7.3427	53.7593	53.8797	0.0200
SDG3	0.9419	4.8936	6.3904	30.0251	33.3759	0.0000
SDG4	0.8814	8.5357	11.1279	44.8082	47.0105	0.0600
SDG5	0.8269	7.1305	8.8329	55.3064	56.0683	0.0100
SDG6	0.7011	9.2460	12.1520	65.2054	71.0653	0.0000
SDG7	0.9079	9.2292	11.7119	38.8915	41.4713	0.0200
SDG8	0.8367	6.7496	8.9499	51.0915	54.3628	0.0000
SDG9	0.8820	8.6695	11.3613	43.4322	47.0316	0.0000
SDG10	0.5112	16.2811	20.0386	86.0849	86.3536	0.0000
SDG11	0.7731	7.6756	10.0890	58.0136	63.3489	0.0100
SDG12	0.7935	5.4595	7.5891	54.0583	60.3694	0.0600
SDG13	0.4024	7.8792	11.0095	96.9139	94.5741	0.0100
SDG14	0.1086	7.4543	10.4305	105.8533	102.2453	0.0600
SDG15	0.2939	10.2463	12.6563	98.4328	96.6512	0.0000
SDG16	0.7931	6.6290	8.2896	60.1870	60.7916	0.0100
SDG17	0.0066	12.0106	15.8987	105.2458	104.9309	0.0000

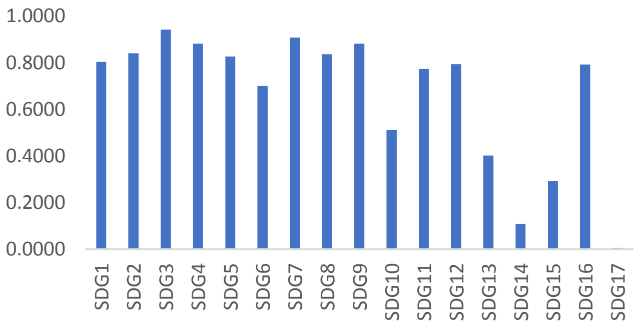

Fig. 5. Linear Regression with R2 Score

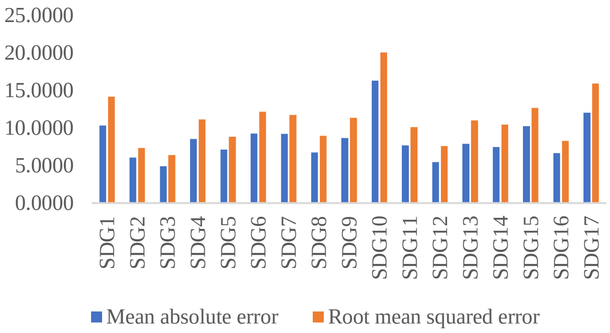

Fig. 6. Linear Regression with MAE and RMSE

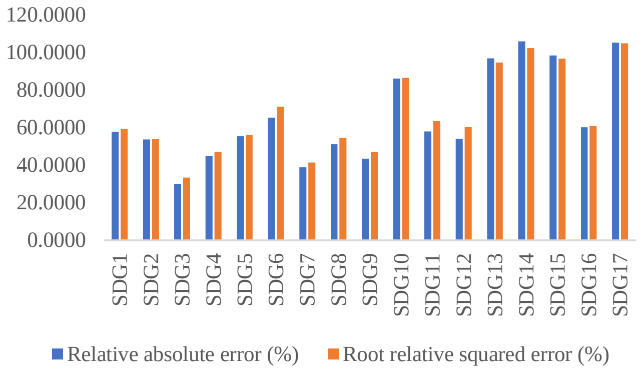

Fig. 7. Linear Regression with RAE (%) and RRSE (%)

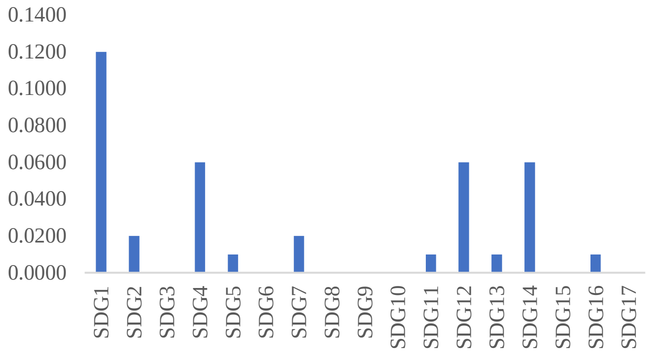

Fig. 8: Machine Learning Models and its Time Taken to Build the Model (Seconds)

Table 4: Random Forest

SDG	Correlation coefficient	Mean absolute error	Root Mean Squared error	Relative Absolute error (%)	Root relative squared error (%)	Time taken
SDG1	0.8171	7.8723	13.7247	44.1461	57.4329	0.6300
SDG2	0.8553	5.8279	7.0215	51.7302	51.5227	0.3300
SDG3	0.9529	4.2773	5.8275	26.2436	30.4361	0.3700
SDG4	0.9128	6.8500	9.7328	35.9591	41.1170	0.2000
SDG5	0.8068	7.2482	9.2981	56.2188	59.0214	0.3400
SDG6	0.7428	8.4910	11.3666	59.8806	66.4719	0.1300
SDG7	0.9150	8.3872	11.3237	35.3432	40.0964	0.1700
SDG8	0.8059	7.2486	9.6855	54.8693	58.8308	0.3900
SDG9	0.9327	6.5300	8.7754	32.7138	36.3270	5.7000
SDG10	0.6827	13.8755	17.0133	73.3656	73.3163	0.3300
SDG11	0.7859	7.5204	9.7896	56.8407	61.4689	0.3400
SDG12	0.7871	5.1486	7.6840	50.9795	61.1245	0.1900
SDG13	0.4362	7.4693	10.5490	91.8726	90.6190	0.9400
SDG14	0.1527	7.2416	10.2426	102.8339	100.4041	0.6700
SDG15	0.3422	9.7399	12.2411	93.5678	93.4804	0.1500
SDG16	0.7947	6.5364	8.2501	59.3456	60.5020	0.2400
SDG17	0.1868	11.1949	14.9569	98.0981	98.7151	0.3900

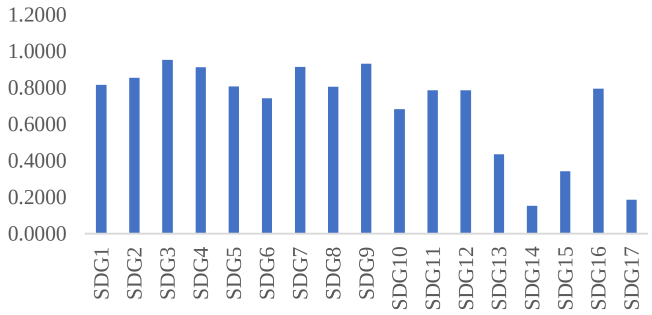

Fig. 9. Random Forest with R2 Score

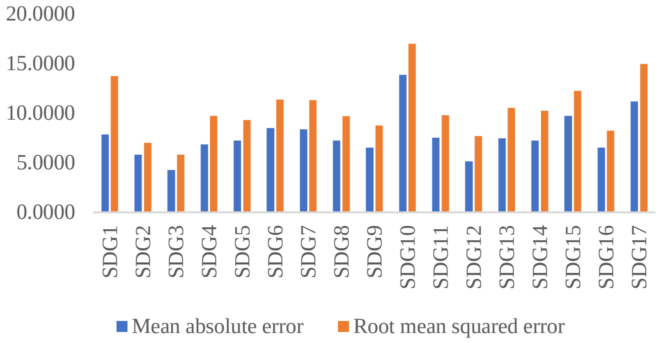

Fig. 10. Random Forest with MAE and RMSE

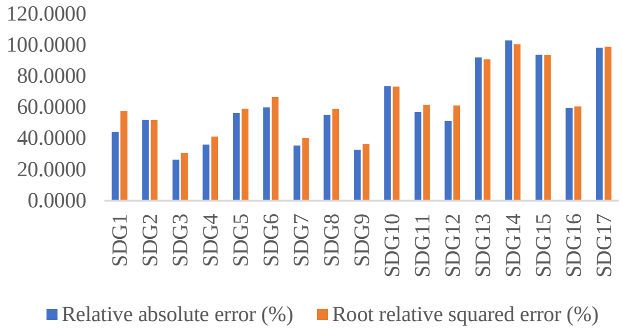

Fig. 11. Random Forest with RAE (%) and RRSE (%)

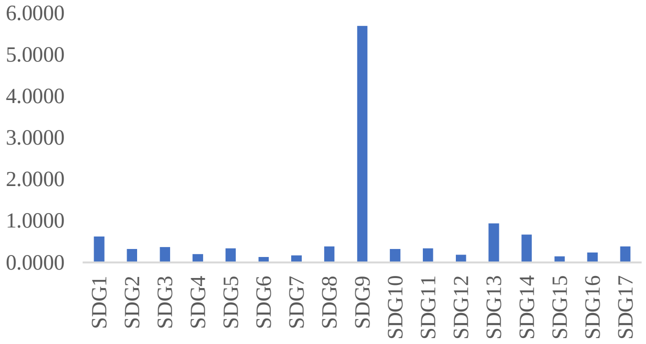

Fig. 12. Random Forest and its Time Taken to Build the Model (Seconds)

4. Result and Discussion

Table 1(a, b) indicates 17 parameters, which include SDG 1 to SDG 17. Based on the dataset, it is evident that different machine learning functions and decision tree approaches are used to find the influencing parameter to decide future predictions. Related results and numerical illustrations are shown between Table 1 to Table 4 and Fig. 1 to Fig. 12.

Based on Table 2, Fig. 1, Fig. 2, Fig. 3, and Fig. 4, the analysis and prediction using the Gaussian process. In this case, to find the R2 score by comparing 17 parameters. Numerical illustrations suggest that there may be a significant difference from one parameter to another. In this case, selecting one parameter compared to the other 16 parameters using the Gaussian process returns a robust, strong positive correlation of nearly 0.9 for using SDG 3 with minimum error and time-consuming to build the model.

Further data analysis revealed a gradual improvement in test scores over time. Based on Table 3, Fig. 5, Fig. 6, Fig. 7, and Fig. 8, the analysis and prediction used linear regression to find the R2 score by comparing different parameter combinations. In this case, selecting one parameter compared to the other 16 parameters using the linear regression returns a robust, strong positive correlation of nearly 0.9 for using SDG 3 with minimum error and minimum time-consuming to build the model.

The analysis and prediction using one of the familiar decision tree approaches random forest. In this case, selecting one parameter compared to the other parameters using the Random Forest. In this case, the RF returns a robust, strong positive correlation of nearly 0.9529 for using SDG 3 with minimum error. They are based on Table 4, Fig. 9, Fig. 10, Fig. 11 and Fig. 12.

5. 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 of all the parameters. In this research, different ML approaches indicate SDG 3 is the best parameter for predicting the future. Future studies can build upon these, finding the suitable variable for future prediction with increased accuracy using different machine learning and decision tree approaches.

6. Reference

1. H. L.Van Soest, D. P. Van Vuuren, J. Hilaire, J. C. Minx, M. J. Harmsen and G. Luderer, “Analysing interactions among sustainable development goals with integrated assessment models”, Global Transitions, Vol. 1, pp.210-225, 2019.
2. R. Avtar, R. Aggarwal, A.Kharrazi, P. Kumar, and T. A. Kurniawan, “Utilizing geospatial information to implement SDGs and monitor their Progress”, Environmental monitoring and assessment, 192(1), pp.1-21. 2020.
3. D. Maulud and A. M. Abdulazeez. “A review on linear regression comprehensive in machine learning”, Journal of Applied Science and Technology Trends, vol. 1(4), pp.140-147, 2020.
4. A. W. Moomen, M. Bertolotto, P. Lacroix, and D. Jensen, “Exploring spatial symbiosis of agriculture and mining for sustainable development in northwest Ghana”. In 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics) (pp. 1-6), 2019.
5. H. Hassani, X. Huang, S. MacFeely and M. R. Entezarian, “Big data and

the united nations sustainable development goals (UN SDGs) at a glance", Big Data and Cognitive Computing, vol. 5(3), pp.28, 2021

6. P. Rajesh, and M. Karthikeyan, "A comparative study of data mining algorithms for decision tree approaches using the Weka tool". Advances in Natural and Applied Sciences, vol. 11(9), pp.230-243, 2017.
7. D. Fraisl, J. Campbell, U. Wehn, J. Wardlaw, and M. Gold, M. 2020. Mapping citizen science contributions to the UN sustainable development goals. Sustainability Science, vol. 15(6), pp.1735-1751, 2020.
8. E. C. Stokes, and K. C. Seto. Characterizing urban infrastructural transitions for the Sustainable Development Goals using multi-temporal land, population, and nighttime light data. Remote sensing of environment, Vol. 234, p.111430, 2019.
9. A. Asadikia, A. Rajabifard, and M. Kalantari, M., Systematic prioritisation of SDGs: Machine learning approach. World Development, vol. 140, p.105269, 2012.
10. P. Rajesh, M. Karthikeyan, and R. Arulpavai, "December. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model". In AIP Conference Proceedings, vol. 2177(1), 2019.
11. P. Rajesh, and M. Karthikeyan, "Data mining approaches to predict the factors that affect agriculture growth using stochastic models". International Journal of Computer Sciences and Engineering, vol. 7(4), pp.18-23, 2019.
12. P. Rajesh, and M. Karthikeyan, B. Santhosh Kumar, and M. Y. Mohamed Parvees, "Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data". Journal of Computational and Theoretical Nanoscience, vol. 16(4), pp.1472-1477, 2019..
13. R. Kohavi, and M. Sahami. Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286), 1996.
14. A. Akusok, (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. W. Chi, (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/sazidthe1/sustainable-development-report?select=sustainable_development_report_2023.csv