A Study on World Sustainable Development Goals Using Machine Learning Approaches

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

[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

Abstract:
Sustainable Development Data encompasses the gathering, examination, and distribution of information pertaining to Sustainable Development Goals (SDGs) and the broader notion of sustainability. Sustainable development constitutes a worldwide effort directed at advancing economic, social, and environmental well-being while safeguarding the welfare of both present and future generations. Data plays an indispensable role in monitoring and attaining these sustainable development objectives, offering essential insights for evaluating advancements, pinpointing obstacles, and guiding well-informed choices.This paper considers different countries sustainable development goalsrelated dataset like country, year, SDG 1 to SDG 17 and overall score. The machine learning approaches which is used to analysis and predict the dataset using linear regression, multilayer perceptron, SMOreg, random forest, random tree, and REP tree**.** Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

Keywords: Machine learning, sustainable development goals, decision tree, correlation coefficient, and test statistics.

DOINUMBER:10.48047/NQ.2022.20.19.NQ99451 NEUROQUANTOLOGY2022;20(19):4896-4904

1. Introduction and Literature Review

Leveraging machine learning and data mining for Sustainable Development Goals (SDGs) research represents a cutting-edge and potent strategy for tackling the intricate and interlinked obstacles connected to the pursuit of sustainability.

Machine learning finds extensive applications in diverse fields such as image and speech recognition, natural language processing, recommendation systems, predictive analytics, autonomous vehicles, fraud detection, and numerous other areas. It has evolved into a pivotal instrument across various industries, facilitating automation, optimization, and data-driven decision-making.Data mining plays a crucial role in decision-making by enabling organizations to make informed decisions, unearth market trends, streamline operations, and enhance overall efficiency. Additionally, it has the capacity to reveal concealed insights within extensive datasets that might remain concealed using conventional analytical approaches. Furthermore, data mining methods frequently intersect with machine learning, as they leverage various machine learning algorithms

for predictive modeling and the identification of patterns in data mining tasks.

A Boosted Regression Trees model, a machine learning and data mining technique, to identify synergistic Sustainable Development Goals (SDGs). It assesses the contributions of all SDGs to the SDG index and conducts a "what-if" analysis to understand the significance of goal scores. The findings highlight that SDG3, "Good health and well-being," SDG4, "Quality education," and SDG7, "Affordable and clean energy," exhibit the most synergy when their scores exceed 60%. These findings offer valuable insights for decision-makers to prioritize synergistic goals and allocate resources effectively [1].

Author introduces a hybrid framework that combines Shannon entropy for data-driven weights, a technique for order preference by similarity to ideal solution for relative ranking, and data envelopment analysis for generating performance efficiency and effectiveness indices. This approach offers benchmarks for setting targets and provides analytical insights for decision-making. The study validates the hybrid methods using the 2014, 2016, and 2018 EGDI datasets covering 193 countries. The results reveal significant changes in performance indices and a decrease in the SDG performance trend due to less efficient resource utilization, despite an increase in outcome effectiveness [2].

The first empirical study of the relationship between the SDGs and populism, introducing a "Sustainability-Populism Framework." It classifies 39 countries into four categories based on their performance on the 17 SDGs and electoral support for populist parties. Regression analysis suggests that a 1-point increase in the aggregate SDG Index corresponds to a 2 percentage point drop in the vote share of populist parties. The findings imply that a strong commitment to the SDGs, especially SDGs 1, 2, 11, and 15, could mitigate populism. This study aims to stimulate a debate on the complex relationship between populism and sustainable development [3].

A comprehensive overview of the intersection between artificial intelligence technologies and the SDGs. It reviews existing literature and conducts SWOT analyses to identify the strengths, weaknesses, opportunities, and threats associated with AI-driven technologies in relation to each SDG. The analysis highlights efforts, opportunities, challenges, and target areas for progress, organized into six perspectives of human needs. The study concludes with a discussion on prospects, guidelines, and lessons for aligning AI developments with SDG attainment by 2030 [4].

The growing interest in statistical analysis of remote sensing data for environmental, agricultural, and sustainable development measurements has led to increased collaboration between earth science and statistics. This review focuses on statistical machine learning methods applied to remote sensing data in the context of the United Nations World Bank Sustainable Development Goals. It provides insights into the methods and their applications, particularly in agriculture, forests, and water quality. The review offers guidance, examples, and case studies for remote sensing practitioners and applied statisticians, covering pre- and post-analysis steps for remote sensing data [5].

Data mining is a valuable tool for extracting valuable information from large databases. In this paper, data mining techniques are applied to a weather dataset that includes attributes related to weather conditions and a class variable indicating whether conditions are conducive to playing golf. Seven classification algorithms, including J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP), and Random Forest (RF), are employed to measure accuracy. The Random Tree algorithm outperformed other algorithms, achieving an accuracy of 85.714% [6].

The increasing influence of artificial intelligence (AI) across various sectors, it becomes essential to evaluate its impact on the achievement of the Sustainable Development Goals. Through a consensus-driven expert elicitation process, our findings reveal that AI has the potential to facilitate the attainment of 134 targets spanning all goals, but it may also hinder progress toward 59 targets. Nevertheless, existing research priorities may overlook critical aspects. To ensure the sustainable development enabled by AI, it is imperative to complement its rapid advancement with appropriate regulatory

oversight and insight. Neglecting this imperative could lead to deficiencies in transparency, safety, and ethical standards [7].

The question of establishing relationships between night satellite monitoring data and sustainable development indicators using the case of Ukraine. It outlines the method for obtaining zonal statistics data for administrative units of varying levels based on long-term nighttime illumination estimates. The study justifies the use of local correlation and regression metrics to identify spatial variations in the strength of the association between nighttime light intensity and sustainable development indicators [8].

The author explain to identify development indicators by examining the development metrics of high-ranking and high-income countries with strong innovation index scores. It attempts to elucidate the correlation between Sustainable Development Goal SDG9 (Industry, Infrastructure, and Innovation) and Global Innovation Index scores in the top ten GII-ranked countries and BRICS. This analysis is based on cross-sectional data encompassing various economic, social, and technological development indicators from UN and World Bank reports for the year 2018. The trends in the variables suggest a need for increased strategic investments in ICT and R&D to fortify ICT infrastructure for digital readiness. Additionally, the analysis underscores a high correlation (R2 = 0.70 to 0.90) between ICT development indicators and GDP per capita (PPP), indicating that ICT serves as a leading indicator for sustainable development, innovation, and infrastructure [9].

This paper utilizes stochastic modeling and data mining approaches to assess groundwater levels, rainfall, population, food grains, and enterprises. The research employs data assimilation analysis to predict groundwater levels effectively. The findings reveal the potential for efficient groundwater level estimation [10] and [11].

A dataset related to chronic diseases is analyzed using five classification algorithms. The study compares the accuracy of these algorithms, and the M5P decision tree approach is found to be the most effective for building the model [12].

Data mining decision trees are widely used for classification and regression tasks. They visually depict sequences of decisions and potential outcomes in a tree-like structure. The CART (Classification and Regression Trees) algorithm is commonly used to build decision trees [13].

2. Backgrounds and Methodologies

2.1Linear 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:

$$y = m_x + b \quad \text{… (2)}$$

Wherey 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.

2.2Multilayer Perception

A Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of interconnected nodes or neurons. It's a fundamental architecture in deep learning and is used for various tasks, including classification, regression, and more complex tasks like image recognition and natural language processing. The architecture of an MLP typically includes three types of layers:

i. **Input Layer:** This layer consists of neurons receiving input data. Each neuron corresponds to a feature in the input data, and the values of these neurons pass through the network.

ii. **Hidden Layers:** These layers come after the input layer and precede the output layer. They are called "hidden" because their activations are not directly observed in the final output.

iii. **Output Layer:** This layer produces the network's final output. The number of neurons in the output layer depends on the problem type.

2.3SMO

SMO stands for "Sequential Minimal Optimization," an algorithm used for training support vector machines (SVMs), machine learning models commonly used for classification and regression tasks. The SMO algorithm is particularly well-suited for solving the quadratic programming optimization problem that arises during the training of SVMs.

Step 1. Initialization
Step 2. Selection of Two Lagrange Multipliers
Step 3. Optimize the Pair of Lagrange Multipliers
Step 4. Update the Model
Step 5. Convergence Checking
Step 6. Repeat: If convergence hasn't been reached, repeat steps 2 to 5 until it is.

2.4Random 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 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)

2.5Random 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 improve the model's generalization performance. Random Trees are commonly used as building blocks in ensemble methods like Random Forests.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):

2.6REP 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.Various steps involved in REP Tree

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

2.7Accuracy 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.

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

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

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

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

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

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.

$$RAE = \frac{\sum|y_i - \hat{y}_i|}{\sum|y_i - \bar{y}|} \quad \ldots (5)$$

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.

$$RRSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}} \quad \ldots (6)$$

Equation 3 to 7are 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.

Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository. The sustainable development goalsdataset includes21 parameters which have different categories of data like country, year, SDG 1 to SDG 17 and overall score[17]. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. Sustainable development goalssampledataset

Country Code	country	year	Goal 1 score	Goal 2 score	Goal 3 score	Goal 4 score	…	Goal 17 score	SDG INDEX SCORE
AFG	Afghanistan	2000	28.8	27.3	19.2	1.6	…	34.2	36
AFG	Afghanistan	2001	28.8	30.6	19.4	1.6	…	34.2	36.3
AFG	Afghanistan	2002	28.8	30.7	19.7	1.6	…	34.2	36.3
AFG	Afghanistan	2003	28.8	32.5	19.9	1.6	…	34.2	36.7
AFG	Afghanistan	2004	28.8	32.1	21.1	1.6	…	34.2	37.1
AFG	Afghanistan	2005	28.8	35.9	22.6	1.6	…	34.2	37.5
AFG	Afghanistan	2006	28.8	36.5	22.7	1.6	…	34.7	37.6
AFG	Afghanistan	2007	28.8	39.5	24.4	1.6	…	34.8	38
AFG	Afghanistan	2008	28.8	37.8	25.9	1.6	…	36	37.3
AFG	Afghanistan	2009	28.8	43	28.1	1.6	…	36.7	38.3

Table 2: Machine Learning Models with R2 Score

ML Approaches	Correlation coefficient
Linear Regression	0.9948
Multilayer Perceptron	0.9990
SMOreg	0.9938
Random Forest	0.9990
Random Tree	0.9947
REP Tree	0.9923

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

ML Approaches	MAE	RMSE
Linear Regression	0.8515	1.1020
Multilayer Perceptron	0.3701	0.5032
SMOreg	0.7973	1.2010
Random Forest	0.3610	0.4930
Random Tree	0.6998	1.1073
REP Tree	0.9128	1.3335

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

ML Approaches	RAE (%)	RRSE (%)
Linear Regression	9.5025	10.2239
Multilayer Perceptron	4.1303	4.6685
SMOreg	8.8978	11.1431
Random Forest	4.0281	4.5741
Random Tree	7.8096	10.2738
REP Tree	10.1860	12.3723

Table5: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Timetaken (seconds)
Linear Regression	0.3800
Multilayer Perceptron	10.0900
SMOreg	55.6100
Random Forest	4.5500
Random Tree	0.0600
REP Tree	0.6600

Fig. 1. R2 Score for Machine Learning Approaches

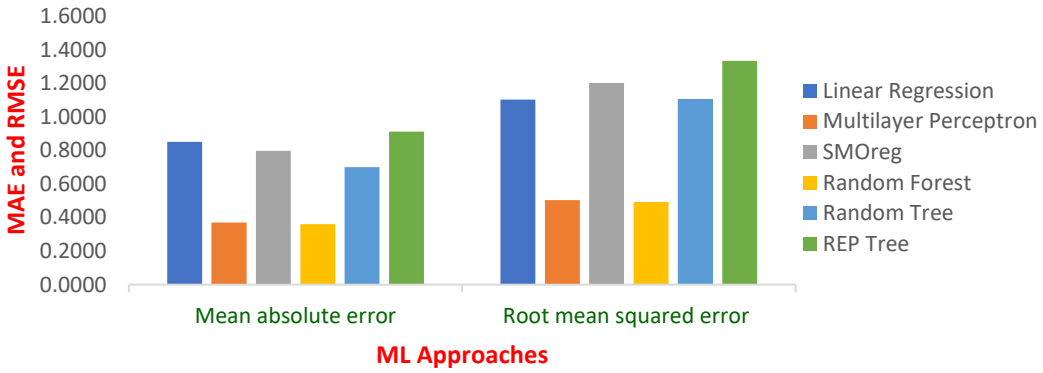

Fig. 2.Machine Learning Models with MAE and RMSE

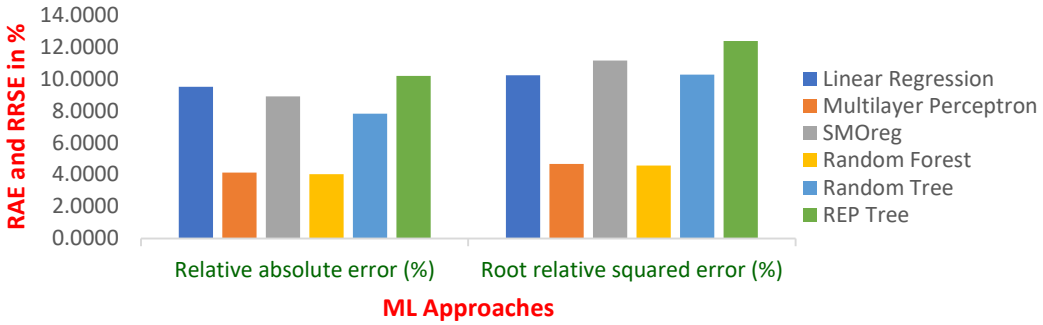

Fig.3. Machine Learning Models with RAE (%) and RRSE (%)

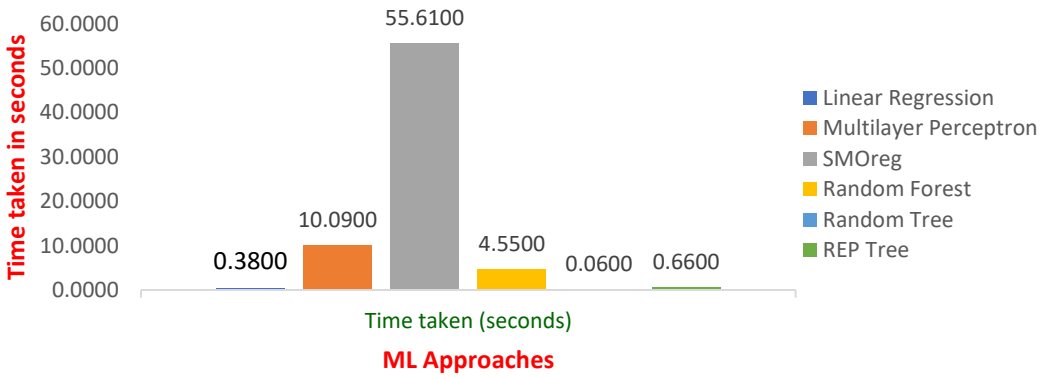

Fig. 4. Machine Learning Models and its Time Taken to Build the Model (Seconds)

3. Results and Discussion

In Table 1, we delineate 21 parameters spanning categories like country, year, SDG 1 to SDG 17, and overall score. These parameters serve as the foundation for employing six distinct machine learning decision tree techniques aimed at unveiling hidden patterns and determining the most influential factors for future predictions. The outcomes and numerical representations are conveyed through Table 1 to Table 5 and Figure 1 to Figure 4.

These results, tied to Equation 2, Table 2, and Figure 1, facilitate the calculation of the R2 score or correlation coefficient across the 21 parameters. The numerical depictions underscore the substantial variance that can exist among different parameters. Notably, our data analysis centers on the sdg_index_score, a metric signifying a country's economic output per capita. When assessed using six different machine learning approaches, this analysis reveals a strong positive correlation of nearly 0.9.

Furthermore, our data analysis highlights a gradual enhancement in test scores over time. To assess model errors, we employ the Mean Absolute Error (MAE) as described in Equations 3. Six machine-learning algorithms are employed, with the random forest exhibiting the lowest error rate based on MAE test statistics.

We also apply the Root Mean Square Error (RMSE), which gauges the disparities between predicted and actual values using Equation 4. In this context, the random forest approach yields the least error as determined by RMSE test statistics, while the REP Tree incurs the maximum error. The relevant numerical representation is presented in Table 3 and Figure 2.

Additionally, the study introduces the Relative Absolute Error (RAE) to gauge accuracy by comparing discrepancies between predicted and actual values in percentage terms using Equation 5. Notably, the Random Forest approach results in the lowest error. A similar pattern emerges in the Relative Root Square Error (RRSE), as seen in Table 4 and Figure 3.

The duration of time required represents a pivotal aspect in machine learning processes. As evidenced in Table 5 and Figure 4, linear regression, REP tree, and random tree demonstrate the quickest model-building times, while multilayer perception, SMOreg, and random forest are characterized by lengthier time requirements.

4. Conclusion and Further Research

In concluding this study, it is imperative to acknowledge certain limitations. The relatively modest sample size within each group could potentially impact the generalizability of our results, and there might be other variables influencing SDG energy performance that warrant consideration. Nonetheless, our findings contribute to the understanding that all parameters exhibit robust positive correlations.In the realm of machine learning, this research showcases that the majority of approaches result in minimal errors and streamlined processing times. Future investigations may extend these findings, exploring SDG-appropriate variables for enhanced predictive accuracy through diverse machine learning and decision tree methodologies.

5. Reference

[1]. Asadikia, A., Rajabifard, A. and Kalantari, M., 2021. Systematic prioritisation of SDGs: Machine learning approach. World Development, 140, p.105269.

[2]. Osman, I.H. and Zablith, F., 2021. Re-evaluating electronic government development index to monitor the transformation toward achieving sustainable development goals. Journal of Business Research, 131, pp.426-440.

[3]. Kroll, C. and Zipperer, V., 2020. Sustainable development and populism. Ecological economics, 176, p.106723.

[4]. Palomares, I., Martínez-Cámara, E., Montes, R., García-Moral, P., Chiachio, M., Chiachio, J., Alonso, S., Melero, F.J., Molina, D., Fernández, B. and Moral, C., 2021. A panoramic view and swot analysis of artificial intelligence for achieving the sustainable development goals by 2030: Progress and prospects. Applied Intelligence, 51(9), pp.6497-6527.

[5]. Holloway, J. and Mengersen, K., 2018. Statistical machine learning methods and remote sensing for sustainable development goals: a review. Remote Sensing, 10(9), p.1365.

[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]. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M. and Fuso Nerini, F., 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. Nature communications, 11(1), pp.1-10.

[8]. Zgurovsky, M., Putrenko, V., Dzhygyrey, I., Boldak, A., Yefremov, K., Pashynska, N., Pyshnograiev, I. and Nazarenko, S., 2018, October.

Parameterization of Sustainable Development Components Using Nightlight Indicators in Ukraine. In 2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC) (pp. 1-5). IEEE.

[9]. Mavuri, S., Chavali, K. and Kumar, A., 2019, November. A study on imperative innovation eco system linkages to map Sustainable Development Goal 9. In 2019 International Conference on Digitization (ICD) (pp. 142-147). IEEE.

[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]. 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(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=sdg_index_2000-2022.csv