Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.9854907

A STUDY ON SUSTAINABLE DEVELOPMENT GOALS IN SOUTH INDIA USING DATA MINING AND MACHINE LEARNING APPROACHES

B. Santhosh Kumar [1] and 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, (Deputed from Department of Computer and Information Science, Annamalai University) Tamilnadu, India
Email: rajeshdatamining@gmail.com

Abstract

Data mining is the best tool for process of discovering interesting result or patterns in large datasets involving methods for using machine learning and statistical methods. It is an interdisciplinary or multidiciplanary subfield of computer science and analytics. Data mining is used to uncover insights such as patterns and trends, and user preferences. In this paper consider the 17 SDG dataset relating to South India, including five states, namely 1. Andhra Pradesh, 2. Karnataka, 3. Kerala, 4. Tamil Nadu, and 5. Telangana. Numerical illustrations were also provided to prove the results and discussions using the Gaussian process, linear regression, random forest, and REP tree with accuracy parameters.

1. Introduction

NITI Aayog is an institution of the Indian government, established with achieve the sustainable development goals with involvement of State Governments of India in the economic policy using a bottom-up approach. It replaced with differernt modification finally name as the Planning Commission of India, which was set up in 1950, it was the major advisory body to the Government of India.

Data Mining is the best data analytical process of extracting and discovered lot of useful knowledge from large amounts of data. It involves the process of discovering patterns and insights from data sets with the help of different major data mining techniques such as clustering, classification, regression, and association rules. These data mining techniques help to identify relationships, trends, and patterns in data that can lead to valuable business insights.

ML is a subset of AI and is concerned with the design of algorithms that can learn and make predictions using pre-existing data. It uses algorithms to analyze data, identify patterns, and make predictions. These algorithms can be trained to detect patterns in data and to make decisions without being explicitly programmed to do so. This can be done through supervised or unsupervised learning methods. Supervised learning methods use labeled data (data with a known outcome) to train the algorithm, while unsupervised learning methods use unlabeled data (data with an unknown outcome) to train the algorithm. Machine learning algorithms can be used for a wide range of tasks, including pattern recognition, image recognition, text analysis, language processing, and more.

The Sustainable Development Goals (SDGs) are a set of 17 global goals set by the United Nations to end poverty, protect the planet and ensure that all people enjoy peace and prosperity. These goals are a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity. The SDGs cover a broad range of issues, from education and gender equality to climate change, clean water and sanitation, and economic growth. The 17 SDGs are organized around five key themes: people, planet, prosperity, peace, and partnership. The 17 SDGs' goals with different parameters as mentioned below.

No Poverty: End poverty in all its forms everywhere. (2) Zero Hunger: End hunger, achieve food security and improved nutrition and promote sustainable agriculture. (3) Good Health and Well-Being: Ensure healthy lives and promote well-being for all at all ages. (4) Quality Education: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. (5) Gender Equality: Achieve gender equality and empower all women and girls. (6) Clean Water and Sanitation: Ensure availability and sustainable management of water and sanitation for all. (7) Affordable and Clean Energy: Ensure access to affordable, reliable, sustainable and modern energy for all. (8) Decent Work and Economic Growth: Promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all. (9) Industry, Innovation, and Infrastructure: Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation. (10) Reduced Inequalities: Reduce inequality within and among countries. (11) Sustainable Cities and Communities: Make cities and human settlements inclusive, safe, resilient and sustainable. (12) Responsible Consumption and Production: Ensure sustainable consumption and production patterns. (13) Climate Action: Take urgent action to combat climate change and its impacts. (14) Life Below Water: Conserve and sustainably use the oceans, seas and marine resources for sustainable development. (15) Life on Land: Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss. (16) Peace and Justice Strong Institutions: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels. (17) Partnerships to Achieve the Goal: Strengthen the means of implementation and revitalize the global partnership for sustainable development.

In India, the SDGs are a key part of the Government's vision for a "New India" by 2022. The Government of India has adopted the SDGs as part of its National Development Agenda. In 2018, the Government of India released the Sustainable Development Goals India Index 2018, which assesses the country's progress on the SDGs. The report found that India is making good progress on some of the goals but is lagging behind on others. The Government of India has also launched several initiatives to promote the SDGs, including the Swachh Bharat Mission and the Pradhan Mantri Ujjawala Yojana. India is also part of the Global Climate Action Agenda and has committed to achieving the goals of the Paris Agreement.

The SDGs offer a comprehensive framework for governments, businesses, civil society, and individuals to work together to create a better world. By setting out a clear and ambitious agenda, the SDGs help to focus efforts and resources on the most pressing global challenges. Furthermore, they provide a common language for governments, businesses, and other stakeholders to communicate their progress and identify areas where more action is needed. The SDGs also provide an opportunity for businesses to demonstrate their commitment to sustainability and to showcase their positive impact on society and the environment. Finally, the SDGs offer a platform for individuals to act and to advocate for global change.

2. Review of Literature

The authors explain various methodologies related to SDGs and review 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 [1].

The authors discuss various works by different researchers on linear regression and polynomial regression and compare their performance using the best approach to optimize prediction and precision. Almost all of the articles analyzed in this review are focused on datasets; in order to determine a model's efficiency, it must be correlated with the actual values obtained for the explanatory variables [2].

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 [3]. The authors explain clearly to identify synergetic SDGs using Boosted Regression Trees model, which is a machine learning and data mining technique. In this study, the 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 [4].

The analysis is organized into six categories or perspectives of human needs: life, economic and technological development, social development, equality, resources and natural environment. Finally, a closing discussion is provided about the prospects, key guidelines, and lessons learned that should be adopted for guaranteeing a positive shift of artificial intelligence developments and applications towards fully supporting the SDGs attainment by 2030 [5]. The authors explain various data mining and machine learning algorithms with accuracy for various decision tree approaches using the WEKA tool to stumble on important parameters of the tree structure. Seven classification algorithms such as 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. The data mining tool WEKA (Waikato Environment for Knowledge Analysis) has been used for finding experimental results of weather data sets. Out of seven classification algorithms, Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [6].

The main objective of this paper is to analyze the SDGs by various independent metrics in the states of Tamil Nadu, Kerala, and Karnataka in India and by taking into consideration three different state SDGs index using data mining and statistical approaches for retrieving various hidden information. Numerical illustrations are also used to prove the proposed results [7]. This stocktaking report attempts to provide an overview of big data, its use in the policymaking context, the stakeholders and their roles, and provides some suggested actionable steps as a discussion stimulus for the "Big Data and the 2030 Agenda for Sustainable Development: Achieving the Development Goals in the Asia and the Pacific Region" meeting in Bangkok on 14 - 15 December 2015. Critical data for global, regional, and national development policymaking are still lacking. Many governments still do not have access to adequate data on their entire populations. This is particularly true for the poorest and most marginalized, the very people that leaders will need to focus on if they are to achieve zero extreme poverty and zero emissions and to 'leave no one behind' in the next 15 years [8].

The Sustainable Development Goals (SDGs) and the Paris Agreement on Climate Change call for deep transformations in every country that will require complementary actions by governments, civil society, science and business. Yet stakeholders lack a shared understanding of how the 17 SDGs can be operationalized. Drawing on earlier work by The World in 2050 initiative, we introduce six SDG Transformations as modular building blocks of SDG achievement: (1) education, gender, and inequality; (2) health, well-being, and demography; (3) energy decarbonization and sustainable industry; (4) sustainable food, land, water, and oceans; (5) sustainable cities and communities; and (6) digital revolution for sustainable development. Each Transformation identifies priority investments and regulatory challenges, calling for actions by well-defined parts of government working with business and civil society. Transformations may therefore be operationalized within the structures of government while respecting the strong interdependencies across the 17 SDGs. We also outline an action agenda for science to provide the knowledge required for designing, implementing, and monitoring the SDG Transformations [9].

3. Material and Methods

3.1 Gaussian Processes

Gaussian processes are a type of machine learning algorithm used for regression and classification. They are based on a probabilistic model in which observations are assumed to be generated by a Gaussian process. Gaussian processes have been used in a variety of applications, including time-series prediction, robotic control, and image recognition. The main advantage of using Gaussian processes is its ability to capture nonlinear relationships and incorporate prior knowledge into the model. Furthermore, Gaussian processes can be used to make predictions with uncertainty, which is useful in cases where we are uncertain about the data or need to make decisions with a high degree of confidence. The formula for Gaussian processes is:

$$GP(x) = \mu + \sigma * N\big(0, K(x,x)\big) \qquad \dots (1)$$

where:

- GP(x) is the output of the Gaussian process model
- μ is the mean of the process
- σ is the standard deviation
- N(0, K(x, x)) is a multivariate normal distribution with mean 0 and covariance matrix K(x, x)
- K(x, x) is a kernel function that measures the similarity between two inputs x and x.

3.2 Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

$$y = c + b * x \qquad \ldots (2)$$

where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable [10].

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors. Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

3.3 Random Forest

Random forest is a type of supervised machine-learning algorithm that is used for both classification and regression. It is a type of ensemble learning method, which means it uses multiple decision trees to make predictions. The random forest algorithm creates a collection of decision trees from a randomly selected subset of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object. It is one of the most accurate and powerful algorithms and is used widely in machine learning and data mining applications.

Random forest algorithm = Ensemble learning method + multiple decision trees + randomly selected subset of training set + aggregated votes from decision trees …(3)

Step 1: Select random samples from the dataset with replacements.
Step 2: Construct a decision tree for each sample and get a prediction result from each tree.
Step 3: Perform a vote for each predicted result.
Step 4: Select the prediction result with the most votes as the final prediction.

3.4 REP Tree

The Reduced Error Pruning (REP) tree is a type of decision tree used in supervised machine learning algorithms. It is an iterative process of pruning the decision tree to reduce the error associated with the decision tree. The process begins with a fully grown tree and then removes branches that do not contribute to the overall accuracy of the decision tree. REP trees are often used in classification problems, and the goal is to create a decision tree that has the lowest possible error rate. REP trees are typically constructed using a top-down approach, where each node is split into two branches based on a certain criterion. The process of splitting is repeated until the tree is fully grown. Once the tree is fully grown, the process of pruning is initiated, where the branches that do not contribute to the accuracy of the decision tree are removed.

1. Begin with a dataset containing a set of relevant attributes and their associated records.

2. Calculate the entropy of the dataset.
3. Select the attribute with the highest information gain and use it as the root node of the decision tree.
4. Split the dataset into subsets, each containing records with the same value for the attribute selected in Step 3.
5. For each subset, calculate the entropy.
6. If the entropy of the subset is 0, then the attribute associated with that subset is the leaf node of the decision tree.
7. If the entropy of the subset is greater than 0, then repeat Steps 3-6 for the subset.
8. Repeat Steps 2-7 until all subsets have an entropy of 0.

3.5 R2 Score

The R2 score, also known as the coefficient of determination, is a metric used to measure the performance of a regression model. It is a statistical measure that indicates how well a model explains and predicts the variance of a dependent variable given an independent variable. The R2 score ranges from 0 to 1, with 0 indicating that the model does not explain the variance in the dependent variable and 1 indicating that the model perfectly explains the variance in the dependent variable [11]. The formula for the R2 score is:

R2 = 1 - (Sum of Squared Errors / Total Sum of Squares) …(4)

where Sum of Squared Errors = (Observed Value - Predicted Value)^2 and Total Sum of Squares = (Observed Value - Mean Value)^2

3.6 Mean Absolute Error (MAE)

Mean absolute error (MAE) is a popular metric because, as with Root mean squared error (RMSE), see next subsection, the error value units match the predicted target value units. Unlike RMSE, the changes in MAE are linear and therefore intuitive. MSE and RMSE penalize larger errors more, inflating or increasing the mean error value due to the square of the error value. In MAE, different errors are not weighted more or less, but the scores increase linearly with the increase in errors. The MAE score is measured as the average of the absolute error values. The Absolute is a mathematical function that makes a number positive. Therefore, the difference between an expected value and a predicted value can be positive or negative and will necessarily be positive when calculating the MAE [12].

The MAE value can be calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_1 - \hat{y}_i|^2 \qquad \dots (5)$$

3.7 Root Mean Square Error (RMSE)

RMSE is the measure of the root of the mean of the squared errors between the predicted and observed/actual values [13]. The chapter uses W/m2 as the unit of measurement for RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{p,i} - Y_{a,i})^2} \qquad \dots (6)$$

where $Y_{p,i}$ is the predicted output and $Y_{a,i}$ is the actual output.

Both errors are expected to be minimum for a good forecasting model. The values near zero represent the model stably tracing the observed value of GHI. On the other hand, larger MAPE and RMSE values represent poor model performance.

4. Numerical Illustrations

NITI Aayog is an institution of the Government of India, established with the aim to achieve sustainable development goals with cooperative federalism by fostering the involvement of State Governments of India in the economic policy-making process using a bottom-up approach. The following dataset (table 1) was collected from SDG India Index & Dashboard 2020-21 [14].

Table 1. SDG Goals in South India

State SDG's	Andhra Pradesh	Karnataka	Kerala	Tamil Nadu	Telangana
SDG 1	81	68	83	86	68
SDG 2	52	53	80	66	50
SDG 3	77	78	72	81	67
SDG 4	50	64	80	69	63
SDG 5	58	57	63	59	41
SDG 6	92	85	89	87	96
SDG 7	100	100	100	100	100
SDG 8	67	66	62	71	73
SDG 9	52	64	60	71	59
SDG 10	74	67	69	74	67
SDG 11	78	78	75	79	76
SDG 12	84	89	65	78	73
SDG 13	63	62	69	61	43
SDG 14	79	60	53	11	79
SDG 15	69	67	77	63	81
SDG 16	77	76	80	71	71
SDG 17	72	72	75	74	69

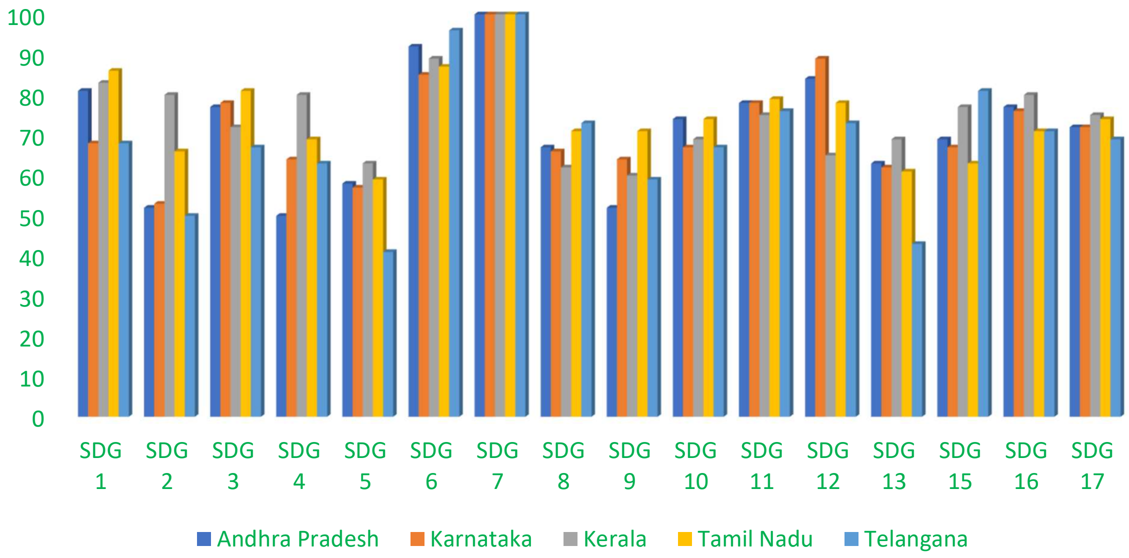

Fig. 1. Comparison between SDG Goals and South India

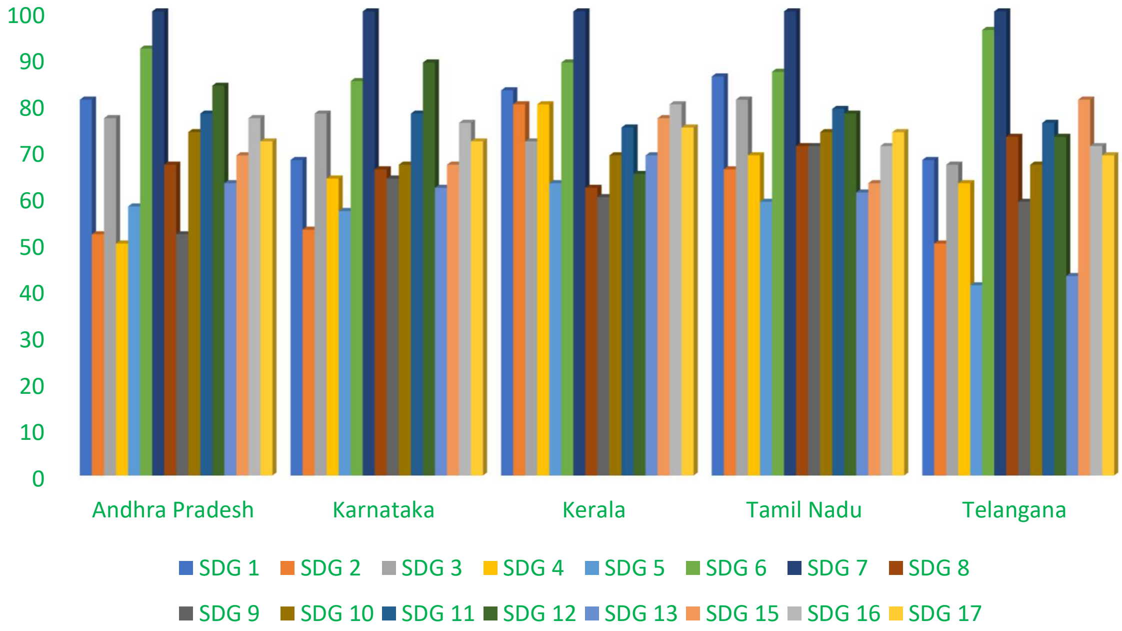

Fig. 2. Comparison between South India and their Goals

Table 2. Descriptive Statistics

Attributes	Mean	StdDev
Andhra Pradesh	72.0590	14.0150
Karnataka	70.9410	12.1730
Kerala	73.6470	11.5540
Tamil Nadu	70.6470	18.5200
Telangana	69.1760	15.7810

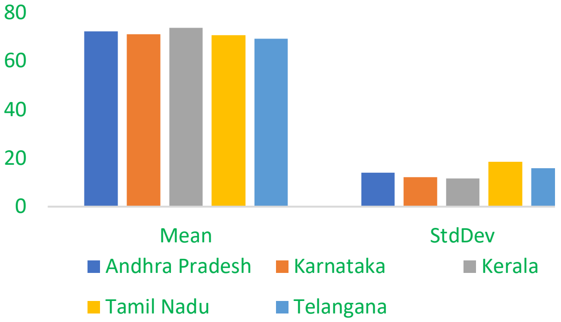

Fig. 3. Mean and Standard Deviation in SDG's Goals

Table 3. ML Algorithms with R2 Score

Attributes	Gaussian Process	Linear Regression	Random Forest	REP Tree
Andhra Pradesh	0.6204	0.7698	0.8137	0.4681
Karnataka	0.6379	0.8028	0.6368	0.4120
Kerala	0.0124	0.5197	0.1229	-0.1229
Tamil Nadu	0.1917	0.4028	0.3190	-0.4628
Telangana	0.6388	0.5966	0.7641	0.2821

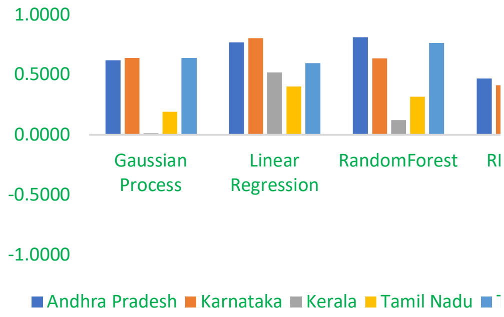

Fig. 4. R2 Score Comparison in South India

Table 4. ML Algorithms with MAE

Attributes	Gaussian Process	Linear Regression	Random Forest	REP Tree
Andhra Pradesh	9.1505	7.3157	6.8521	10.2249
Karnataka	8.7024	5.6045	6.9674	9.4346
Kerala	9.6308	9.296	10.4223	10.5574
Tamil Nadu	10.8217	13.5737	9.8079	12.1993
Telangana	10.0267	10.9147	7.9361	10.5774

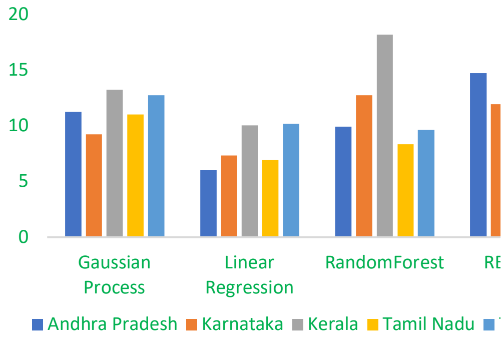

Fig. 5. Comparison between SDG's using MAE

Table 5. ML algorithms with RMSE

Attributes	Gaussian Process	Linear Regression	Random Forest	REP Tree
Andhra Pradesh	11.4207	9.1949	8.4827	12.5814
Karnataka	10.6311	7.5143	9.1514	10.9402
Kerala	11.6159	10.7999	12.0523	12.6574
Tamil Nadu	18.2224	19.0645	17.3863	20.2923
Telangana	12.8132	12.8077	10.0763	14.8632

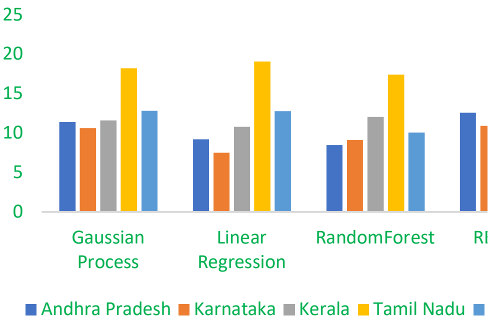

Fig. 6. Comparison between SDG's using RMSE

5. Result and Conclusion

The numerical illustrations are based on different tables and graphs shown in Table 1 to Table 5 and Figure 1 to Figure 6. From Table 1, we can see that overall rank of 1 to 17 sustainable development goals relating to five different states, namely Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, and Telangana in south India. This table indicates SDG 7, namely affordable and clean energy, received a maximum score (100) in South India. The results and discussion are shown in Table 1, figure 1, and Figure 2.

Based on comparative analysis, Kerala received a maximum mean score of 73.6470 and a minimum standard deviation of 11.5540. Andhra Pradesh received second place and had a mean score of 72.0590 and a standard deviation is 14.0150. Karnataka received a mean of 70.9410 and an SD of 12.1730. Tamil Nadu having a fourth place, received a mean score is 70.6470 and an SD of 18.5200. Finally, Telangana received a mean score of 69.1760 and the SD score is 15.7810, respectively. The results and discussion are shown in Table 2 and Figure 3.

The results indicate that there is a significant correlation between the independent and dependent variables. These are presented in Table 3, and Figure 4 shows that Andhra Pradesh returned the first highest R2 score of 0.8137 for using random forest. Karnataka received the second-highest R2 score of 0.8028 for using linear regression. Andhra Pradesh received the third highest R2 score, 0.7698, for using linear regression. Kerala and Tamil Nadu return a negative correlation score of 0.9160 for using the REP tree. The related results are shown in Table 3 and Figure 4.

The results of this study suggest that the test statistics, namely mean absolute error (MAE), indicate the Karnataka return minimum error of 5.6045 for using linear regression. Tamil Nadu returns the maximum error rate for using linear regression and REP tree. The related results are shown in Table 4 and Figure 5. Karnataka returned a minimum RMSE (7.5143) for using linear regression, and the REP tree returned a maximum error of 20.2923 for Tamil Nadu. The related results and discussion are shown in Table 5 and Figure 6.

6. Further Research

The results are consistent with previous studies that have shown similar correlations between independent and dependent variables. Additionally, the results may be affected by other factors that were not considered in this study. Therefore, further research is needed to include all the states in India using different machine learning algorithms, which are used to increase accuracy and also minimize errors.

Reference

1. Avtar, R., Aggarwal, R., Kharrazi, A., Kumar, P. and Kurniawan, T.A., 2020. Utilizing geospatial information to implement SDGs and monitor their Progress. *Environmental monitoring and assessment*, *192*(1), pp.1-21.
2. Maulud, D. and Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), pp.140-147.
3. Moomen, A.W., Bertolotto, M., Lacroix, P. and Jensen, D., 2019, July. 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). IEEE.
4. Asadikia, A., Rajabifard, A. and Kalantari, M., 2021. Systematic prioritization of SDGs: Machine learning approach. World Development, 140, p.105269.
5. 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.
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. Rajesh, P. and Santhosh Kumar, B., 2020. Comparative studies on Sustainable Development Goals (SDG) in India using Data Mining approach. J. Sci, 14(2), pp.91-93.
8. Maaroof, A., 2015. Big data and the 2030 agenda for sustainable development. Report for UN-ESCAP.
9. Sachs, J.D., Schmidt-Traub, G., Mazzucato, M., Messner, D., Nakicenovic, N. and Rockström, J., 2019. Six transformations to achieve the sustainable development goals. Nature Sustainability, 2(9), pp.805-814.
10. Linear Regression, https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/
11. R2 Score, https://www.investopedia.com/terms/r/coefficientofdetermination.asp
12. MAE, https://www.sciencedirect.com/topics/computer-science/mean-absolute-error
13. RMSE, https://www.sciencedirect.com/topics/engineering/root-mean-square-error

14. SDG India & Dashboard 2020-21, NITI Aayog, Government of India, Sansad Marg, New Delhi - 110001, Source: sdgindiaindex.niti.gov.in.