Analysis of Age Sage Classification for Students' Social Engagement Using REPTree and Random Forest

Jigna B. Prajapati

Analysis of Age Sage Classification for Students' Social Engagement using REPTree & Random Forest

Jigna B Prajapati,

Acharya Motibhai Patel Institute of Computer Studies,

Ganpat University, Gujarat, India

jignap15@gmail.com

Abstract: Study and analysis of train dataset along with various ML algorithms is used widely in different sectors. The accuracy parameters can be clarified to have prediction of different score levels. This study covers the extension work of Students' social engagement during covid-19 pandemic. The study was initiated with students' social connection during the pandemic. We had compared various machine learning algorithms with its performance about the engagement of students in various social network. After studied, analyzed & compared, we derived that the most of students' social engagement found in WhatsApp, YouTube & Instagram. The current study is foreseeing age wise social media connection. It correlates between student & their social engagement during the pandemic phase. In which age group, which social media is one of the most popular one. This study focuses on age wise classification using Machine Learning. In this paper, the decision-making classification is compared. The Reduced Error Pruning Tree (REPTree) and Random Forest algorithm is implemented on train dataset with diverse nodes. The attributes are focused as age & time spent on social media as per necessity of study. This paper includes the study and analysis of RAE & RMSE along with ML tree approach. The discoveries of this study can lead better classification in regards of students' age and duration which they have spent on social media for derived social platform.

Keywords: Social Media Platform, REPTtree, Random forest, Classification, Decision Tree

1. Introduction

The virtual social engagement is drastically increasing day by day. The people are using n number of social platforms to make them engage in various community. This is also adopted by the student fraternity for social connection & entertainment to be in various social media platform.

In our previous work entitled "Performance Comparison of Machine Learning Algorithms for Prediction of Students' Social Engagement", we have studied Students engagement in different social media platform. Such platform as LinkedIn, Facebook, Instagram, Reedit, Snapchat, Talklife, Telegram, Twitter, WhatsApp, YouTube & etc along with various attributes. The attributes covers the broad spectrum of student's routine activities during the pandemic. The data set collected with residence area, age, duration hours for online class, rating of online class experience, medium for online class, duration hours for self-study, duration hours for fitness, duration hours for sleep, duration hours for social media, social

media platform, duration hours for TV, Number of meals per day, Change in weight, Health issue during pandemic, Stress busters, Time utilized for more than 1200 instances. After analyzed suitably we derived that the most of students were majorly engaged in WhatsApp, YouTube & Instagram [1].

Our dataset consists of student from all age group. Social engagement has been found to be associated with many factors for the student crowd as social connection, entertainment, relaxing phase, news-updates of others and many more. The dataset collected for this study includes residence region, age, online class hours, online education rating, online education medium, self-study hours, fitness hours, Sleep hours, social media platform used, time spent on specific social media, TV hours, meals plan, weight increase or decrease, health issues recorded, noticeable stress factor. The different age group many have their own reasons for joining & remaining on various social platforms. The WhatsApp, YouTube & Instagram have been found the most popular connection network during the last study [1]. Here the focus on age wise connection popularity for well said social media platform. The collected data is processed by feature abstraction, feature alignment very initially. After this noise removing, splitting & labeling of data is being done well during the study about the most popular social media platform [1]. To study accurately about the popular social medial in particular age wise group, the data are again pre-processed as per necessity of REPTree and Random Forest algorithm.

REPTree and Random Forest algorithm are effective to support decision making in real time environment[2]. These are supervised chaffier and popular to derive various classification results [3,4].

2. Related Work

To analyze & process the data in well structural manner, studied the various researchers work in the same domain. The ML algorithm have been used find appropriate results from heterogeneous sectors [5].

Li Yang has discussed about cardiovascular disease prediction for the area of eastern China. They have used random forest mechanism. His results shown that random forest produced more sound for significant improvement for CVD prediction [6]. Joske Ubels has discussed about prediction of treatment benefit using random forest . They have used failed clinical drug trials. [7]. Fabián Santos has discussed about the evaluation of forest change drivers using random forest for Northern Ecuadorian Amazon. This approach demonstrated the advantages for integrating remote data & remote sensing-derived products [8]. Martin Hanko has discussed about traumatic brain Injury using random forest to predict mortality in patients. They have constructed data for 6-month mortality and derived enhanced prediction results [9]. Toby G Pavey has discussed about wrist-worn accelerometer data in concern random forest classifier. He claims accurate group level prediction for controlled conditions using random forest classifier for wrist accelerometer [10]. Eric S Walsh has discussed about estuarine system aligning with the spatial distribution of sediment pollution using Random Forest [11]. Ishwaran H has discussed regression, classification, survival Statistics in medicine by SE (standard errors) & CE (confidence intervals) for variable with random forest [12]. Eric Ariel L Salas has discussed about the forest image classification of agricultural systems using random forest. They have used airborne hyperspectral datasets [13]. Samad Jahandideh has used a random forest classifier for improvement of chances of successful protein structure determination [14]. F Chris Jones has discussed about the Random forests as cumulative effects models for a case study of lakes and rivers in Muskoka, Canada [15]. Shiyang Li has discussed about the nitrate concentration and load estimation using data mining techniques. This study was focused on

different type of watersheds. They have predicted nitrate levels and derived that REPTee has given better performance in concentration and load results [16]. Sankaralingam Mohan has discussed about summertime ground-level ozone concentration to forecast O3 concentrationfor the surface level using REPtree[17]. Mahfuzur Rahman has discussed about delineating multi-type flooding in Bangladesh using stacking hybrid machine learning algorithms[18]. Elizabeth Goya-Jorge has discussed about the chemical-induced estrogencity in silico and in vitro methods[19], Sunil Saha has discussed about forecast of probability of deforestation about Gumani River Basin, India using random forest, REPTree & binary logistic regression [20]

3. Proposed Work

In the previous study, the process of data acquisition, feature abstraction, feature alignment, noise removal, splitting and labeling data has been carried out [1] but Some data structuring is necessary to apply REPTree and Random Forest algorithm. As mentioned in Fig.1, feature selection is done in the reference of age wise, social media usage with time spent. The train dataset is prepared to be applicable in REPTree and Random Forest. This train dataset is using age of student with Time spent on social media. It analyzes & compares the derived result using Weka tool.

Classification algorithm implements classifier to maps input data to a particular category. It is an instance of supervised learning. Here, the train dataset is used to identify classification observations for age wise (social media platform & duration they spent on social media). Classification can help us to have step-by-step observation. This study uses the 10-fold cross validation to implements various models.

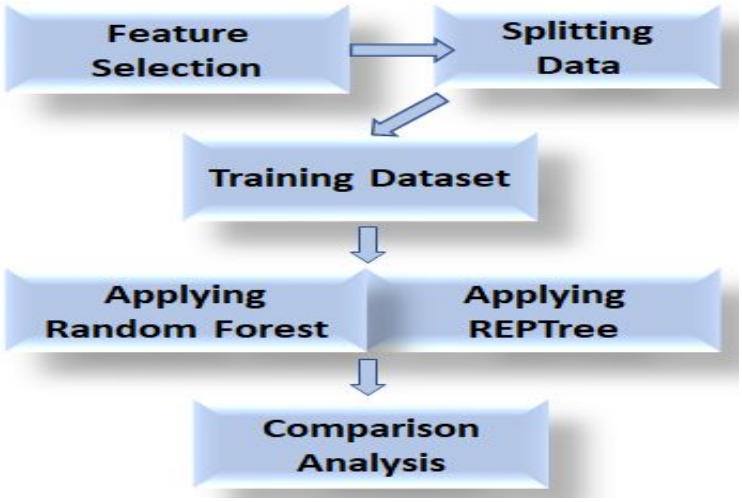

Fig. 1 Work Flow Diagram

3.1 RepTree

Reptree focus on each node which represents a decision based on input, and move to the next node & next until the predicted output. With the use of regression tree mechanism, it creates multiple node for specified tree in multiple iterations. Once

the iteration done, it selects best one as representative. It is one of the fast decisions.

The REPTree is applied on dataset with 10 fold cross validation methods on different attributes wise Social media platform usage on Time spent and age. The REPTree classifies with predictive social media as mentioned in a. Predictive social media with size of 31 tree nodes.

3.1.1 Predictive social media Test mode: 10-fold cross-validation

```
Age < 16.5
|  Time spent on social media < 1.75
|  |  Age < 14.5
|  |  |  Time spent on social media < 0.25 : Youtube (8/4) [6/4]
|  |  |  Time spent on social media >= 0.25
|  |  |  |  Age < 11.5 : Whatsapp (8/0) [3/2]
|  |  |  |  Age >= 11.5
|  |  |  |  |  Age < 12.5 : Youtube (7/4) [4/1]
|  |  |  |  |  Age >= 12.5 : Whatsapp (39/18) [13/5]
|  |  Age >= 14.5 : Whatsapp (42/27) [20/12]
|  Time spent on social media >= 1.75
|  |  Time spent on social media < 5.5
|  |  |  Time spent on social media < 2.5 : Youtube (37/18) [14/9]
|  |  |  Time spent on social media >= 2.5
|  |  |  |  Age < 13.5 : Youtube (11/0) [3/1]
|  |  |  |  Age >= 13.5
|  |  |  |  |  Time spent on social media < 3.5
|  |  |  |  |  |  Age < 14.5 : Youtube (3/2) [2/1]
|  |  |  |  |  |  Age >= 14.5
|  |  |  |  |  |  |  Age < 15.5 : Instagram (2/0) [3/1]
|  |  |  |  |  |  |  Age >= 15.5 : Youtube (3/1) [0/0]
|  |  |  |  |  Time spent on social media >= 3.5 : Youtube (7/2) [4/2]
|  |  Time spent on social media >= 5.5 : Instagram (4/2) [1/1]
Age >= 16.5
|  Age < 29.5
|  |  Time spent on social media < 0.45 : Youtube (11/7) [10/7]
|  |  Time spent on social media >= 0.45
|  |  |  Time spent on social media < 1.25 : Whatsapp (153/105) [70/48]
|  |  |  Time spent on social media >= 1.25 : Instagram (403/232) [215/128]
|  Age >= 29.5 : Whatsapp (50/24) [26/13]

Size of the tree : 31
```

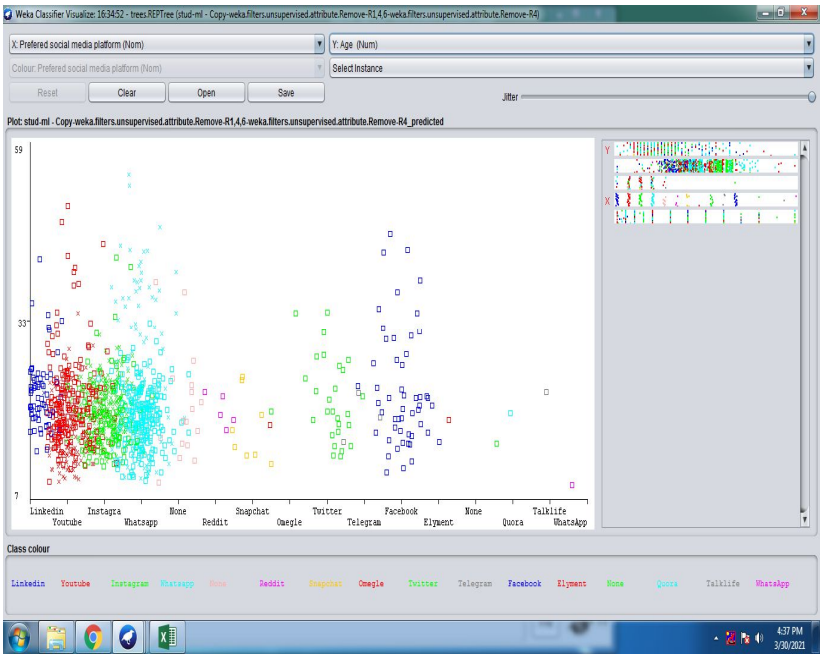

Fig. 2 Reptree visualize(predictive social media))

There are various age grouping whose engagement in different social media. The node 1 to 20 shown in *fig.3* present the age slots. If age group less than 16 and greater than 14, majorly used YouTube & WhatsApp. The greater than 16 and less than 29 age group is using Instagram also. Fig. 4 display REPTree predictive age from weka 3.9.

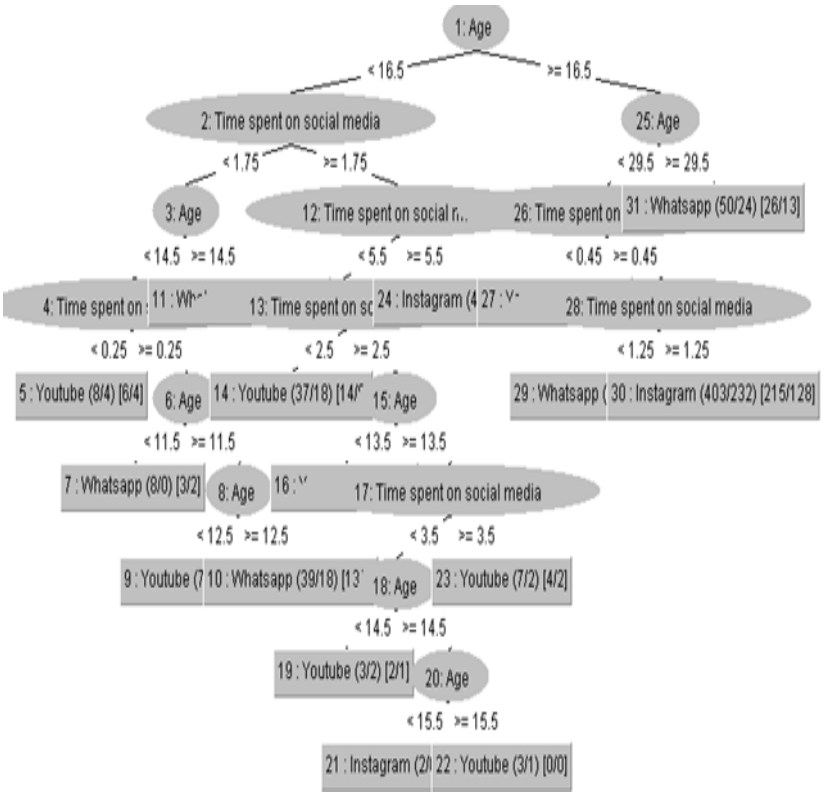

Fig. 2 Tree visualiser Reptree(age)

3.1.2 **Predictive rules for Social Media** Test mode: **10-fold cross-validation**

Prefered social media platform = Linkedin : 22.26 (41/8.51) [20/26.91]
Prefered social media platform = Youtube
| Time spent on social media < 0.15 : 16 (8/34.69) [4/10.81]
| Time spent on social media >= 0.15
| | Time spent on social media < 5.5
| | | Time spent on social media < 2.5
| | | | Time spent on social media < 1.5 : 18.59 (77/27.61) [33/25.99]
| | | | Time spent on social media >= 1.5 : 20.32 (55/55.4) [38/39.72]
| | | Time spent on social media >= 2.5
| | | | Time spent on social media < 3.25: 17.87 (24/13.71) [15/13.5]
| | | | Time spent on social media >= 3.25: 19.41 (33/18) [16/20.7]
| | Time spent on social media >= 5.5: 21.09 (6/6.33) [5/8]
Preferred social media platform = Instagram : 19.83 (246/9.19) [106/5.28]
Preferred social media platform = Whatsapp : 20.67 (211/46.49) [125/52.82]
Preferred social media platform = None : 18.41 (13/33.3) [4/74.81]
Preferred social media platform = Reddit : 18.8 (5/2.16) [0/0]
Preferred social media platform = Snapchat : 18.25 (7/15.14) [1/4]
Preferred social media platform = Omegle : 21 (0/0) [1/0.72]
Preferred social media platform = Twitter
| Time spent on social media < 2.5
| | Time spent on social media < 1.5 : 19.5 (4/2.75) [4/9.25]

| | Time spent on social media >= 1.5 : 24.33 (7/34.12) [2/8.59]
| Time spent on social media >= 2.5 : 20.82 (9/2.47) [2/3.09]
Preferred social media platform = Telegram : 19 (2/9) [1/0]
Preferred social media platform = Facebook : 23.48 (36/77.47) [16/29.36]
Preferred social media platform = Elyment : 22 (1/0) [0/0]
Preferred social media platform = None : 14 (1/0) [0/0]
Preferred social media platform = Quora : 20 (1/0) [0/0]
Preferred social media platform = Talklife : 20 (0/0) [1/0.02]
Preferred social media platform = WhatsApp : 12 (1/0) [0/0]

Size of the tree : 45

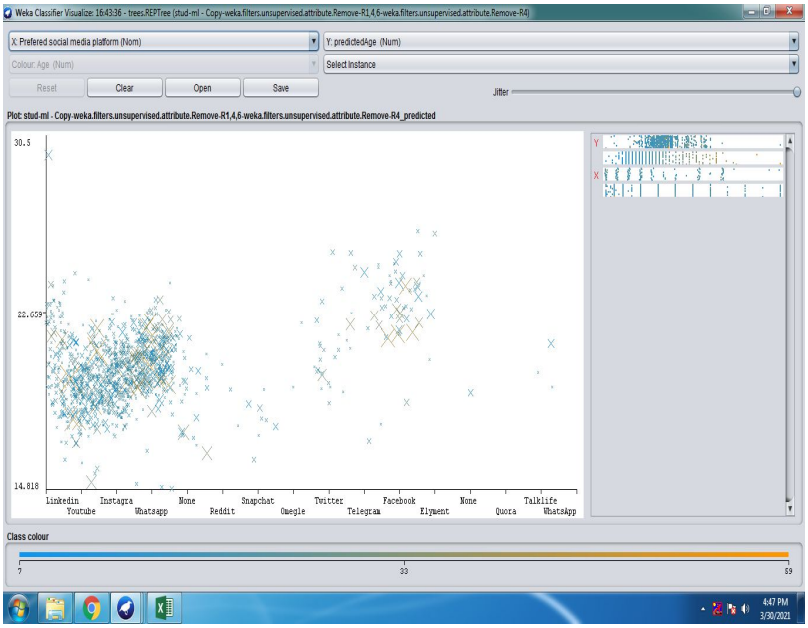

Fig. 3 Reptree(predictive age)

3.2. Random forest

Random forest Tree is a mechanism of supervised Classifier. Here, nodes are divided in suitable subset randomly. Randomly chosen node are related with super node & same tree structure. Radom forest can help to derive results in classification and regression problems. The random trees classifier classifies each set of nodes with each tree in the forest and generates outputs as majorly voted. Suppose we discusses regression, the average of the responses for all the trees in the forest is concentrated to focus more appropriate answer [17]. Fig.5 shows Random Forest algorithm with predictive social media factor. It focusses on the age on y-axis and social media on x-axis. The particular age group is using particular type of social media which is represented in broadly blue, green and red colors. Whatapp & youtube is being use mostly in age around 20.

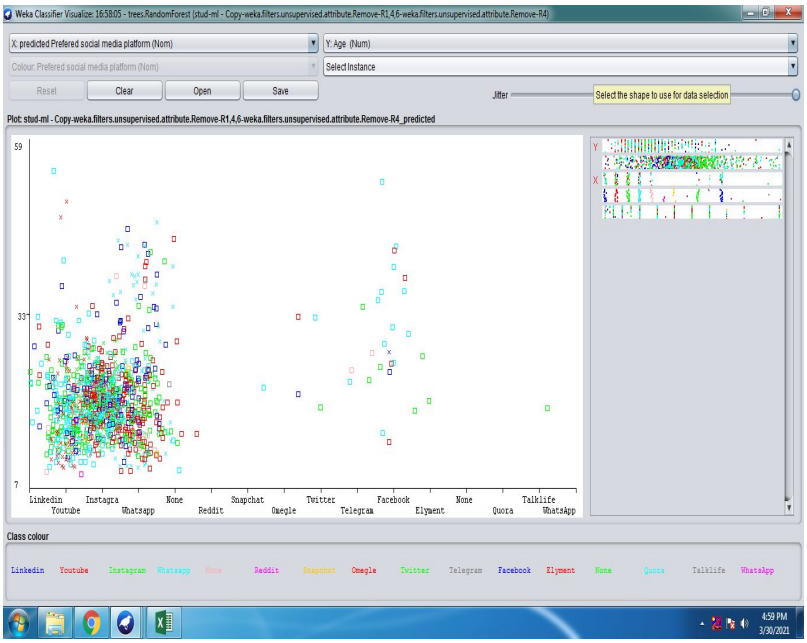

Fig. 5 Ramdom Forest (predctive Social media for age)

4. RESULTS AND DISCUSSION

Decision tree is one of the most popular techniques. Sometimes due to large structure of data size, the results are complex to analyze. REPTree can produce a simple tree structure. Such tree structure will work on accurate classification with Pruning Methods. Random Forest is applied to add on the age wise student engagement in social media using 10 folds validation method. The 66 percentage of slit is used on dataset for further decision process. The REPTree is implemented with different attributes. The REPTree is structed with 31 node to classify the train dataset for age and 45 node tree structure for time spent on social media. The Table-1 shows the results for REPTree. Table-2 shows the results from Random Forest. The Random Forest suggested 165 node structure for same train data set for age attribute while 269 for time spent on social media attributes. The performance measures and errors plots are analyzed for classify the students' age wise engagement.

TABLE 1: Weka measurements: REPTree

Error	Social Media	Age	Time Spent
MAE	0.09	3.5	1.23
RMSE	0.22	5.46	1.7
RAE	94.58	99.98	93.64
RRSE	99.17	98.96	96.34

TABLE 2: Weka measurements**:**.RandomForest

Error	Social Media	Age	Time Spent
MAE	0.09	3.58	1.29
RMSE	0.22	5.54	1.79
RAE	94.1	99.98	93.05
RRSE	99.98	100	99.98

The MAE is popular for continuous variable data. The Lower value of Mean Absolute error is shown as better performance of predicted model. The RMSE is popular for high or low values for large errors. Again, the lower RMSE direct towards the more appropriate model results. The REPTree MAE & RMSE is low compare to Randon Forest in all factors as age wise social media platform time spent on social media.

5. Conclusion

The REPTree & Random Forest algorithms are popular for decision making in quick mode. These are is implemented on preprocessed dataset using weka 3.9 . The REPTree is implemented with 31 node size to determine the age wise, social media wise and time wise usage as shown in tree visualizer Fig.3 . The different decision rules confirm that age greater than 16 and less than 29 majorly engaged in WhatsApp and YouTube. The decision rules defines also that age less than 16 are majorly engaged in YouTube. The same data is implemented with Random Forest and shown similar outcomes about age wise social engagement. *Fig.*5 is presentation of age and social media connection using Weka tool. The cluster created on greater than 5 and less than 29 which is concentrated around in between age number. Such age number supports the major engagement in WhatsApp & you tube as shown in *Fig.5.* The study accomplishes the cluster on age nearby 16 is engaged in WhatsApp and you tube. The REPTree is more appropriate algorithm for implemented train dataset with minimum & structure node tree.

References

1. Prajapati JB, Patel SK. Performance Comparison of Machine Learning Algorithms for Prediction of Students' Social Engagement. In2021 5th International Conference on Computing Methodologies and Communication (ICCMC) 2021 Apr 8 (pp. 947-951). IEEE.
2. Y. J. Sheela and S. H. Krishnaveni, "A comparative analysis of various classification trees," 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), 2017, pp. 1-8, doi: 10.1109/ICCPCT.2017.8074403.
3. S. A. Shubho, M. R. H. Razib, N. K. Rudro, A. K. Saha, M. S. U. Khan and S. Ahmed, "Performance Analysis of NB Tree, REP Tree and Random Tree Classifiers for Credit Card Fraud Data," 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038578Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, vol. 2, issue 10,2014 pp.155-160.Classification Using REPTree", International Journal of Advance
4. Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, vol. 2, issue 10,2014 pp.155-160.
5. D. Anguita, A. Ghio, N. Greco, L. Oneto and S. Ridella, "Model selection for support vector machines: Advantages and disadvantages of the Machine Learning Theory," The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1-8, doi: 10.1109/IJCNN.2010.5596450.
6. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific reports. 2020;10(1):5245.
7. Ubels J, Schaefers T, Punt C, Guchelaar HJ, de Ridder J. RAINFOREST: a random forest approach to predict treatment benefit in data from (failed) clinical drug trials. Bioinformatics (Oxford, England). 2020;36(Suppl_2):i601-i9.
8. Santos F, Graw V, Bonilla S. A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. PloS one.

2019;14(12):e0226224.
9. Hanko M, Grendár M, Snopko P, Opšenák R, Šutovský J, Benčo M, et al. Random Forest-Based Prediction of Outcome and Mortality in Patients with Traumatic Brain Injury Undergoing Primary Decompressive Craniectomy. World neurosurgery. 2021;148:e450-e8.
10. Pavey TG, Gilson ND, Gomersall SR, Clark B, Trost SG. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. Journal of science and medicine in sport. 2017;20(1):75-80.
11. Walsh ES, Kreakie BJ, Cantwell MG, Nacci D. A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system. PloS one. 2017;12(7):e0179473.
12. Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Statistics in medicine. 2019;38(4):558-82.
13. Salas EAL, Subburayalu SK. Modified shape index for object-based random forest image classification of agricultural systems using airborne hyperspectral datasets. PloS one. 2019;14(3):e0213356.
14. Jahandideh S, Jaroszewski L, Godzik A. Improving the chances of successful protein structure determination with a random forest classifier. Acta crystallographica Section D, Biological crystallography. 2014;70(Pt 3):627-35.
15. Jones FC, Plewes R, Murison L, MacDougall MJ, Sinclair S, Davies C, et al. Random forests as cumulative effects models: A case study of lakes and rivers in Muskoka, Canada. Journal of environmental management. 2017;201:407-24.
16. Li S, Bhattarai R, Cooke RA, Verma S, Huang X, Markus M, et al. Relative performance of different data mining techniques for nitrate concentration and load estimation in different type of watersheds. Environmental pollution (Barking, Essex : 1987). 2020;263(Pt A):114618.
17. Mohan S, Saranya P. A novel bagging ensemble approach for predicting summertime ground-level ozone concentration. Journal of the Air & Waste Management Association (1995). 2019;69(2):220-33.
18. Rahman M, Chen N, Elbeltagi A, Islam MM, Alam M, Pourghasemi HR, et al. Application of stacking hybrid machine learning algorithms in delineating multi-type flooding in Bangladesh. Journal of environmental management. 2021;295:113086.
19. Goya-Jorge E, Amber M, Gozalbes R, Connolly L, Barigye SJ. Assessing the chemical-induced estrogenicity using in silico and in vitro methods. Environmental toxicology and pharmacology. 2021;87:103688.
20. Saha S, Saha M, Mukherjee K, Arabameri A, Ngo PTT, Paul GC. Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, REPTree: A case study at the Gumani River Basin, India. The Science of the total environment. 2020;730:139197.