EVALUATION OF MULTI-TARGET REGRESSION MODELS ON AFRICA SOIL PROPERTIES

Folorunso, S. O[1]., Ayo, F. E.[2], Adigun, A. A.[3], Olaniyan, O. E.[4]
[1,4]Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria
[2]Department of Physical and Computer Sciences, McPherson University, Seriki Sotayo, Nigeria
[3]Department of Information and Communication Technology, Osun State University, Osogbo, Nigeria
sakinat.folorunso@oouagoiwoye.edu.ng[1]

Keywords: Multi-output, Multi-target Regression, Ensemble of Regressors, Machine Learning Models, Regressor Chain

Abstract: *Intensifying sustainable agriculture and management of natural resources can be achieved with digital mapping of soil functional properties in data sparse regions of Africa. This research is aimed at evaluating different multi-output Regression models (Random Forest Regressor (RFR), Linear Regressor (LR), Extremely Randomized Trees Regressor (ET) and Bagging Regressor using LR as base classifier (BLR))to predict five soil properties (Calcium (Ca), Phosphorus (P), Potential of Hydrogen (pH), Soil Organic Carbon (SOC) and Sand of different soil sample) simultaneously on African Soil sample dataset. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Correlation (R^2) and Explained Variance metrics were used to evaluate the performances of these models on the dataset. The result obtained revealed that RFR performed best based on RMSE, MAE and Explained Variance that other models. LR performed inferior to other models but its ensembles with BLR improve its predictive performance.*

1. INTRODUCTION

Soil is a loose material formed by weathering the physical and chemical breakdown of rocks and are also key elements for agriculture and natural resources[1]. Its quality is a determinant to the level of agricultural productivity of a particular region. While the content of soil to support life are minerals, gases, organic matter, liquids and organisms, its quality can be determined by taking the measurement of its properties[2].The physical and compound properties of soil in different zones gives vital information for land the board, supportability and water yield of the particular zone [3]. Some of soil's assets related to organic matter are Nitrogen (N), Carbon (C), potential of hydrogen (pH), Magnesium (Mg), Calcium (Ca) and Potassium (K)[4]. These properties are helpful to farmers in the evaluation of its fertility, designing of cultivation plan and predicting crop productivity [5]. Intensifying sustainable agriculture and management of natural resources can be achieved using automation of soil functional properties in data sparse regions of Africa. Building a multi-target regression model to predict these properties at un-sampled locations is key to the success of the planning. This automation is advantageous over traditional expert-based soil mapping. Some key advantages are listed here [6]. Multi-target/output regression model predicts many continuous dependent variables built on a set of input/independent variables concurrently. Multi-target regression tasks are popular in various domains like ecology [7], soil samples [6], [8] and Cancer [9]. This research is focused on the evaluation of multi-output models to predict five multiple target variables of soil functional properties from diffuse reflectance infrared spectroscopy measurements (Ca, P, pH, SOC and Sand) of different soil sample concurrently. The ensemble modelling approach was adopted and compared with a base learner. Linear regression (LR), Extremely Randomized Trees Regression (ET) [10], Random Forest Regression (RFR) [11], and Bagging Regression(BLR) [12] were the models considered.The remainder of this paper is organized as follows: In Section 2 describes the related work to multi-target regression in Section 3 the methodology adopted and the experimental setup for data analysis. In Section 4, the result obtained is presented, discussed and the models were compared. Finally, the conclusion to the study is outlined in Section 5.

2. RELATED WORK

2.1 Multiple-Output Regression

A multiple-output regression learns the association between the input and output pair when presented with a set of train data. This input is a vector with many target variables [13]. Formally, given an$N \times D$input matrix $X = [x_1, x_2, \cdots x_N]^T$ and $N \times K$ output matrix $Y = [y_1, y_2, \cdots y_N]^T$, the multiple-output regression learns the functional association between the inputs $x_n \in \mathbb{R}^D$and the outputs $y_n \in \mathbb{R}^K$. LR is displayed in (1)

$$y_n = W^T x_n + b + \epsilon_n \forall_n 1, \cdots, N \qquad (1)$$

Here, $W = [w_1, \cdots w_K]$denotes the $D \times K$ matrix where $w_k$denotes the regression coefficientof the k-th output,$b = [b_1, \cdots b_k]^T \in \mathbb{R}^K$is a vector of bias terms for the K outputs, and$\epsilon_n = [\epsilon_{n1}, \cdots \epsilon_{nK}]^T \in \mathbb{R}^K$is a vector consisting of the noise for each of the *K*outputs. The noise, though uncorrelated across the *K* outputs is presumed to be Gaussian with a zero mean.

Standard parameter estimation for (1) involves maximizing the (penalized) log-likelihood of the model, or equivalently minimizing the (regularized) loss function over the training data:

$$\arg\min_{W\,b} \operatorname{tr}\left(\left(Y - XW - 1b^T\right)\left(Y - XW - 1b^T\right)^T\right) + \lambda R(W) \qquad (2)$$

where $\operatorname{tr}(.)$denotes matrix trace, 1 an $N \times 1$vector of all 1s and $R(W)$the regularizer on the weightmatrix$W$consisting of the regression weight vectors of all the outputs. For a choice of $R(W) = \operatorname{tr}(W^T W)$ (the ℓ_2-squared norm, equivalent to assuming independent, zero-mean Gaussian priorson the weight vectors), solving (2) amounts to solving K independent regression problems and this solution ignores any correlations among the outputs or among the weight vectors[13].

2.2 Related Work On Multi-Target Regression (MTR) Models

MTR models are effective for the prediction of multi-output tasks. It is a well-researched area due to its applicability in a wide range of domains. Two common MTR methods reported in the literature are:

a) Problem Transformation (PR) methods: This approach is local to the model. It transforms the task into independent single-output problems to be by solved by a single-output regression model, and

b) Algorithm Adaptation (AA) methods: This is global to the model. This approach adapts a particular single-output method to predict multi-output datasets simultaneously [14].

The two common approaches adopted in the prediction of a multi-target dataset are to build a model for each target separately (Single-Target) or to build one model to predict all targets simultaneously (multi-target).A Single-Target (ST) approach applies a one-versus-all (binary relevance) baseline, stacked generalization and regressor chains to learn a model for each target separately as there are no dependencies amongst the targets. A Multi-Target (MT) approach applies algorithm adaptation methods to learn one model for all targets simultaneously. This approach learns faster, builds smaller models and explains dependencies between different target attributes [7]. Table 1 gives the summary of related research in MTR. The third column consists of the type of dataset used and its proportion. For example, in a: b:c, a represents the number of examples contained in the soil data, b is the number of features while c is the number of dependent variables.

3. METHODOLOGY

This section discusses the methods and materials used for this study. Figure 1 depicts the flow of process adopted for this study.

Table 1 Summary of related work

Author	Models	Dataset	Method	Remark
[7]	Regression Trees (RT), Ensembles of RT, MT-RT and ensembles of MT-RT	16967:40:7ecological/ remote sensed data	Ensemble + Base Learner	RF with RT as base learner
[9]	Bayesian Classifier Chains (BCC), Classifier Chain (CC) and Class Relevance (CR) with Decision Tree (DT) and Decision Stump as the base learner and RF	858:33:3 Cancer data from UCI	Ensemble + base learner	BCC + Decision Stump performed best
[15]	MTR Stacking (MTRS), Ensemble of Regressor Chains (ERC) and Ensemble of Regressor Chains Corrected (ERCC). Ridge, Support Vector Machine (SVM), Stochastic Gradient Boosting and Bagging regressors were used.	6: ≥ 24 4 new + 2 publicly available	Ensemble	ERCC performed best
[16]	A new MTAugmented Stacking (MTAS), compared with ST and other three MT methods (Stacked ST, ERC, and Deep Regressor Stacking (DRS) Using different base learners SVM, RF and Classification and Regression Tree (CART).	40:12	Ensemble	MT outperform ST.
[17]	Extend ET based to Predictive Clustering Trees (PCTs) framework and compared with Forest-PCTs	13: ≥ 2ecological/ remote sensed data	Ensemble	Extra-PCTs
[18]	Pre-processing: Moving Average (MA), Median Filtering (MF), Gaussian Smoothing (GS) and Savitzky Golay Smoothing (SGS) Model: Artificial Neural Network (ANN), Principal Component Regression (PCR), Principal Least Square Regression (PLSR) models on the resultant dataset.		Pre-Processing + Base Learners	SGS + PLSR model outperforms all other models
[19]	Ensemble of MTR that constructs new target variables via Random Linear Combinations (RLC) of existing targets. Compared with ST with gradient boosting as the baseline and Multi-Objective RF	12: ≥ 2 from UCI	Ensemble	RLC
[20]	LR, SVR, RR, Logistic Regression (LGR), PR, PLR	1,157: 3594:5 soil sample	Base Learners	LGR performed best
[8]	PCR, PLSR, LS-SVM, Cubist	140:3 soil samples 40 features	Base Learners	LS-SVM
[6]	LR and RF Kriging	28,000:14 soil samples	Ensemble	RFkriging

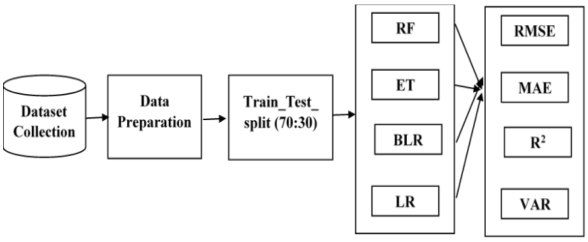


Fig 1: Design methodology of the Multi-Output model

3.1 Dataset

The publicly available African Soil data was used for this research [21]. The train and test data were split along Sentinel Landscape levels. This dataset contained 1158 instances with 3,600 attributes. The five target variables to be predicted simultaneously are Soil Organic Carbon (SOC), Potential of Hydrogen (pH), Mehlich-3 extractable Calcium (Ca), Mehlich-3 extractable Phosphorus (P), and Sand content. Table 2 displays the dataset attributes and Sand content. Table 2 displays the dataset attributes.

Table 2: The African Soil Dataset

Dataset	Instances	Features	Targets
Africa Soil [21]	1158	3595	5

3.2 Multi-Output Models Used

Algorithm adaptation method was adopted for the multi-output regression model for this study. These methods can predict all target variables simultaneously with a lone model capable of capturing all dependencies and internal relationships between these variables. The advantages of this method over problem transformation methods is that they build smaller and faster models [7] and are better interpreted than many single-target models. Also, when these target variables are correlated, there is improved predictive performance [22].

3.2.1 Linear Regression

Linear regression learns the linear association between a dependent variable Y and one or more independent variables X. Y must be continuous values, while X may be either binary, continuous, or categorical [23].

3.2.2 Random Forest

Random Forest (RF) [11] is an ensemble of classification and regression algorithm based on the bagging [12] and random subspace methods [24]. For the purpose of this study, regression was used. Random Forest Regressor (RFR) ensemble is a variant of Bagging where the tree in the ensemble is grown from a sample drawn with replacement from the train data. When the tree is being grown, the best split node is selected among an arbitrary subset of the features. Consequently, this randomness slightly increases the bias of the forest and decreases the variance due to averaging to building an overall better model.

3.2.3 Bagging Regressor

Bootstrap AGGregatING (Bagging) [12] is an ensemble of base learners for classification or regression based on a majority vote on trained bootstrapped samples of the training dataset from base learners. It generates A bootstrapped examples of the train data using arbitrary sampling with replacement. It trains the classifier/ regression function using each bootstrap sample. For the regression task, it averages on the predicted values thereby reducing variation.

3.2.4 Extra Trees Regressor

The Extra-Trees (ET) [10] model is an ensemble of trees built on extraordinary randomization of the tree construction algorithm. For each node of the tree, k attributes are arbitrarily selected where the best split is chosen at the node upon evaluation.

3.3 EXPERIMENTAL SETUP

All of the multi-output regression models were implemented on the open-source Python Language Scikit-Learn [25] in Jupyter computing environment with the following packages: RandomForestRegressor [11], BaggingRegressor [12], ExtraTreesRegressor [10] and linear regressor. The models were adapted to multi-output form using MultiOutputRegressor for regression model building.The ensemble of unpruned multi-target regression trees was built with 100 unpruned trees as suggested in[12], [11].The predictions of the trees were gotten by averaging the predictions from each tree and combine them together. The maximum depth at which nodes were expanded is 10, the number of trees grown in the forest was 50 and the number of jobs to run in parallel for both fit and predict state using all processors for RFR.For the ET model, 10 trees were grown in the forest, 32 features were for the best split while the random state is zero.BLR and LR were with their default values. Model performance on unseen data was estimated by dividing the whole dataset into train and test data to the ratio of 70%-30% respectively. The predictive performance of these models was compared based on MAE, RMSE, R^2 and Explained Variance [25] as displayed in Fig. 2

$$\text{MAE} = \frac{1}{N}\sum_{j=1}^{N} |y_j\hat{y}_j|$$

Where N is the number of instances, y_j is actual value and $\hat{y}_j$ is predicted value.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N} (y_j - \hat{y}_j)^2}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

Where SSR is the regression sum of squares
SST is the measure of total variation in the Y variable

$$\text{Explained\_Variance}\,(y, \hat{y}) = 1 - \frac{\text{Var}\{y - \hat{y}\}}{\text{Var}\{y\}}$$

Where $\hat{y}$ is the estimated target output, y is the corresponding (correct) target output
Var is variance is the square of the standard deviation

Fig. 2: The predictive performance measure

4. RESULTS AND DISCUSSION

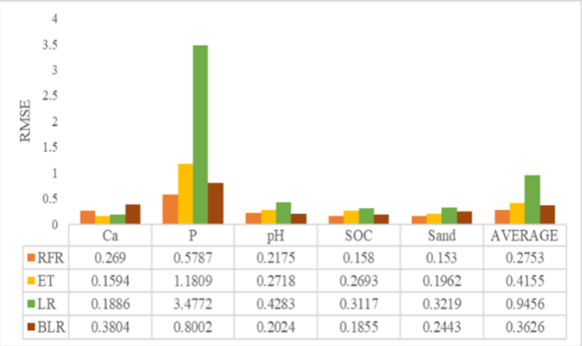
Results obtained were presented and discussed in this section. The superiority of the performances of each model (LR, RFR, ET and BLR) based on RMSE, MAE, R^2 and explained variance is reported in this section.

In the following figures (Fig 3-Fig 6), the performance of the multi-output models was presented. The figures show the performance of the specific models performed against one another based on the discussed metrics. For RMSE and MAE metrics, the closer their values is to zero, the better the model. For R^2 and Explained variance, the closer their values are to 1, the better the model.

From the results obtained, Phosphorus (P) was the most difficult soil property to be predicted by all the models. It was particularly difficult for LR to predict P for all metrics. But, when cast as base learner for the ensemble

Bagging, there is an improvement in the performance.

Fig.3 presents the result of the comparison of the RMSE of the four models with respect to the prediction of the five target variables. As observed in Fig.3, RFR outperforms all other models for the prediction of all the properties on the average. In the prediction set, ET provided the best prediction for Ca (RMSE = 0.1594). P which was most difficult to be predicted by all models and was best predicted by RFR (RMSE = 0.5787). pH was best predicted by BLR with (RMSE = 0.2024). Sand was also best predicted by RF with RMSE = 0.1530. LR performed worst on all predictions of individual soil properties and on the average. The average result obtained from the three homogenous ensembles is close to each other.



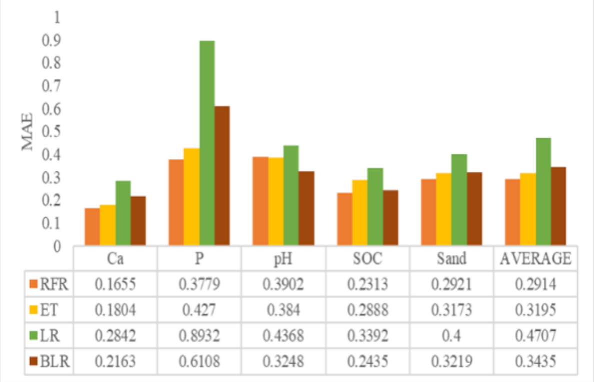
	Ca	P	pH	SOC	Sand	AVERAGE
🟧RFR	0.269	0.5787	0.2175	0.158	0.153	0.2753
🟨ET	0.1594	1.1809	0.2718	0.2693	0.1962	0.4155
🟩LR	0.1886	3.4772	0.4283	0.3117	0.3219	0.9456
🟫BLR	0.3804	0.8002	0.2024	0.1855	0.2443	0.3626

Fig.3: Model comparison using Root Mean Square Error (RMSE) on soil properties.

Fig.4 presents the result of the comparison of the MAE of the four models with respect to the five target variables. As observed in Fig.4 also, RF outperforms all other models for the prediction of all the properties on the average.

In the prediction set, RF provided the best prediction for Ca (MAE = 0.1655). P which was most difficult to be predicted by all models was best predicted by RFR (MAE = 0.3779).

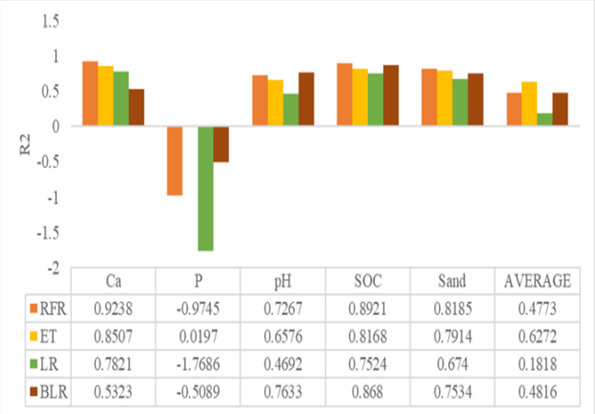
pH was best predicted by BLR with (MAE = 0.3248). SOC was also best predicted by RF with MAE = 0.2313. Sand was also best predicted by RF with MAE = 0.2921. LR performed worst on all predictions of individual soil properties and on the average.



	Ca	P	pH	SOC	Sand	AVERAGE
🟧RFR	0.1655	0.3779	0.3902	0.2313	0.2921	0.2914
🟨ET	0.1804	0.427	0.384	0.2888	0.3173	0.3195
🟩LR	0.2842	0.8932	0.4368	0.3392	0.4	0.4707
🟫BLR	0.2163	0.6108	0.3248	0.2435	0.3219	0.3435

Fig. 4: Model comparison using Mean Absolute Error (MAE) on soil properties.

Fig.5 presents the result of the comparison of the R^2 of the four models with respect to the five target variables. As observed in Figure 6 also, ET outperforms all other models for the prediction of all the properties on the average. In the prediction set, RF provided the best prediction for Ca (0.9238). P which was most difficult to be predicted by all models was best predicted by ET (0.0197). pH was best predicted by BLR with (0.7633). SOC was also best predicted by RFR with value 0.8921. Sand was also best predicted by RF with value 0.8184. LR performed worst on all predictions of individual soil properties and on the average.

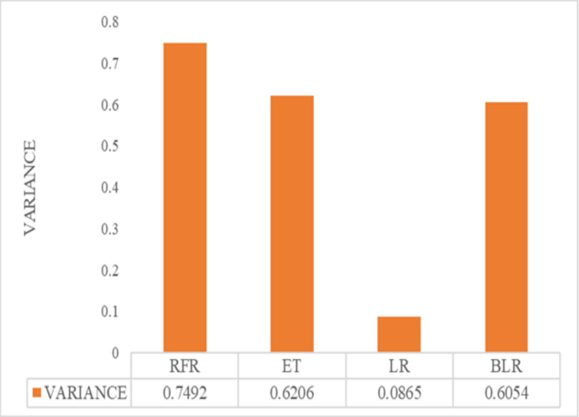


	Ca	P	pH	SOC	Sand	AVERAGE
🟧RFR	0.9238	-0.9745	0.7267	0.8921	0.8185	0.4773
🟨ET	0.8507	0.0197	0.6576	0.8168	0.7914	0.6272
🟩LR	0.7821	-1.7686	0.4692	0.7524	0.674	0.1818
🟫BLR	0.5323	-0.5089	0.7633	0.868	0.7534	0.4816

Fig. 5: Model comparison using R Square Error (R^2) on soil properties.

Fig.6 presents the result of the comparison of Explained Variance of the four models with respect to the five target variables. As observed in Figure 7 also, RF outperforms all other models for the prediction of all the properties on the average with the value of 0.7492). And again, LR performed worst on all predictions of soil properties with the value of

0.0865. By this metric, this is an extremely poor value showing that the model performed poorly predicting these values.



	RFR	ET	LR	BLR
■ VARIANCE	0.7492	0.6206	0.0865	0.6054

Fig. 6: Model comparison using Explained Variance on soil properties.

By all metrics, RFR performed well often more than the other models. One of the reasons could be that, from every example drawn with substitution from the train data, the best split node is selected amongst an arbitrary subset of the features.

Consequently, this randomness slightly increases the bias of the forest and decreases the variance due to averaging to building an overall better model. And also that the base learner is a randomize tree. ET ensemble model which is also a variant of Bagging performed well on the predictions. One of its strength is that its base learner is an extremely randomized tree. Hence, better performance.

LR as a base learner performed worse in all predictions. This could be due to the fact that it learns a linear relationship better. But, as a base learner for Bagging ensemble model, there is a great improvement in the performance.

LR took advantage of bootstrapped samples of the training data using random sampling with replacement property of Bagging. Bagging averages on the predicted values of LR thereby reduce variation.

5. CONCLUSION

This research compared three (3) multi-output regression models: Random Forest Regression, Extra-Trees Regression, Bagging with Linear Regression as the base learner and Linear Regression model for the prediction of soil properties to predict 5 target outputs of different soil properties simultaneously.

Based on Agriculture, the use of machine learning models can be used to ease Agriculturist and Soil Scientist work by predicting the properties of soil faster and effectively.

Though ensembles give precise models, such a large number of precise models may restrict their useful application. To be attainable and focused, it is vital that the regression models keep running in sensible time.

Based on the regression accuracy measure with RMSE, MAE, R^2 and variance, the efficiency of the models was measured on the target variables.

It was observed that the Random Forest outperforms all other models based on RMSE, MAE and the variance while Extra-Trees performed best for the R^2 metric and that Phosphorus prediction was worse for the models under consideration.

In future research work, several versions of Bagging from the literature such as Pasting and Random Patches will be considered and improvement on the prediction of Phosphorus P.

6. REFERENCES

[1] K. Azadeh and M. G. Mohammad, "Effects of land use change on the soil physical and chemical properties and fertility of the soil in Sajadrood catchment," Agric Eng Int: CIGR Journal, p. 16, 2014.

[2] M. Casanova, E. Tapia, O. Seguel and O. Salazar, "Direct measurement and prediction of bulk density on alluvial soils of central Chile," CHILEAN JOURNAL OF AGRICULTURAL RESEARCH, vol. 76, no. 1, pp. 105-113, January-March 2016.

[3] D. M. Michele, H. G. Sérgio , R. M. Carlos , R. O. Phillip and C. Nilton , "Knowledge-based digital soil mapping for predicting soil properties in two representative watersheds," Sci. Agric., vol. 75, pp. 144-153, 2018.

[4] A. S. Melissa and . R. W. Ray, "The Relationship between soil quality and crop productivity across three tillge systems in south central Hondursa," American Journal of

Alternative Agriculture, vol. 17, pp. 292-297, 2002.

[5] S. S. Manisha and F. D. Manuel, "Machine Learning for the Management of Agricultural Soil Data," Centro Singular de Investigación en Tecnoloxías da Información, pp. 1-17, 2016.

[6] T. Hengl, G. Heuvelink, B. Kempen, J. Leenaars, M. G. Walsh, K. D. Shepherd, A. Sila, R. A. MacMillan, J. M. Jesus, L. Tamene and J. E. Tondoh, "Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions," PLoS ONE 10(6), pp. 1-26, 2015.

[7] D. Kocev, S. Džeroski, M. D. White, G. R. Newell and P. Griffioen, "Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition," Ecological Modelling Journal, vol. 220, no. 8, pp. 1159-1168, 2009.

[8] A. Morellos, X.-E. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, G. Tziotzios, J. Wiebesohn, R. Bill and A. Mouazen, "Machine Learning based Prediction of Soil Total Nitrogen, Organic Carbon and Moisture Content by Using VIS-NIR Spectroscopy," Biosystems Engineering, pp. 104-116, 2016.

[9] S. G. Fashooto, A. S. Metfula, B. B. Metsebula and B. Y. Fashooto, "Multi-Target Prediction on Cervical Cancer evaluation of Predictive Performance measure," Asian Journal of Information Technology, vol. 17, no. 2, pp. 160-166, 2018.

[10] P. Geurts, D. Ernst and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3-42, 2006.

[11] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[12] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, p. 123–140, 1996.

[13] P. Rai, A. Kumar and H. Daumé III, "Simultaneously leveraging output and task structures for multiple-output regression," in 25th International Conference on Neural Information Processing Systems, 2012.

[14] H. Borchani, G. Varando, C. Bielza and P. Larranaga, "A survey on multi-output regression," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 5, pp. 216-233, September 2015.

[15] E. Spyromitros-Xioufis, W. Groves, G. Tsoumakas and I. P. Vlahavas, "Multi-Label Classification Methods for Multi-Target Regression," CoRR, vol. abs/1211.6581, 2012.

[16] E. J. Santana, B. C. Geronimo, S. M. Mastelini, R. H. Carvalho, D. F. Barbin, E. I. Ida and S. Barbon Jr., "Predicting poultry meat characteristics using an enhanced multi-target regression method," Biosystems Engineering , vol. 171, pp. 193-204, 2018.

[17] D. Kocev and M. Ceci, "Ensembles of Extremely Randomized Trees for Multi-target Regression," in LNAI, vol. 9356, N. Japkowicz and S. Matwin, Eds., Springer International Publishing Switzerland, 2015, pp. 86-100.

[18] S. Minu and A. Shetty, "Prediction Accuracy of Soil Organic Carbon from Ground Based Visible Near-Infrared Reflectance Spectroscopy," Journal of the Indian Society of Remote Sensing, vol. 46, no. 5, p. 697–703, May 2018.

[19] G. Tsoumakas, E. Spyromitros-Xioufis, A. Vrekou and I. Vlahavas, "Multi-target Regression via Random Linear Target Combinations," in Lecture Notes in Computer Science, vol. 8726, T. Calders, F. Esposito, E. Hüllermeier and R. Meo, Eds., Springer, Berlin, Heidelberg, 2014.

[20] I. Akinola and T. Dowd, "Predicting Africa Soil Properties Using Machine Learning Techniques," Electrical Engineering Stanford University, Stanford, CA 94305, pp. 50-62, 2016.

[21] "African Soil Dataset," 2012. [Online]. Available: https://www.kaggle.com/c/afsis-soil-properties/data.

[22] H. Borchani, G. Varando, C. Bielza and P. Larrañaga, "A survey on multi-output regression," WIREs Data Mining Knowl Discov, vol. 5, pp. 216-233, 2015.

[23] A. Schneider, G. Hommel and M. Blettner, "Linear Regression Analysis," Dtsch Arztebl Int, vol. 107, no. 44, pp. 776-782, 2010.

[24] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 20 , Issue: 8 , Aug 1998) , vol. 20, no. 8, pp. 832 - 844, August 1998.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P.

Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in {P}ython," Journal of Machine Learning Research, vol. 12, pp. 2825--2830, 2011.