IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 8 Issue 7, July 2021
ISSN (Online) 2348 – 7968 | Impact Factor (2020) – 6.72
www.ijiset.com

Machine Learning based Soil Fertility Prediction

NIYITEGEKA JANVIER [1], NSHIMIYIMANA Arcade [2], NGABOYERA ERIC[2], NSENGUMUREMYI Jean [2]
[1],[2]Department of Electronics and Telecommunication Technology, IPRC TUMBA
Rulindo, Northern province, Rwanda
[1]nijas2012@yahoo.com
Kigali, Rwanda

***Abstract*—Soil fertilization activities contribute a lot in crops production volume. However, if the quantity of soil composition (fertilizer) is not controlled and maintained consistent, this may lead to less crop production volume. Regulating the quantity of soil fertilizer is one of the measures to be taken prior for preventing the inferior quality and less quantity of the crop production. Thus, the measurement of soil nutrients is greatly required for better plant growth and fertilization. Therefore Calcium Ca, Phosphorous P and pH level are among the parameters commonly measured to monitor the soil fertility as they are the ones mostly important and informative soil parameters to determine the soil fertility.**

In our current paper we used a Machine Learning ML algorithm to build a model which may help the farmers to predict the quantity of soil properties which include the Calcium Ca, Phosphorous P and pH values. These parameters may help the farmers to know the quantity of fertilizers to be added to the soil sample according to the values of Ca, P and pH measured for keeping the soil fertility consistent. Through this paper, different machine learning algorithms such as Linear Regression, Random Forest, Gradient Boosting, Random Forest,K-Nearest Neighbors ,Decision Tree and Artificial Neural network were tested for optimizing the prediction accuracy using python programming packages.

***Index Terms*—Machine learning, soil fertility, Linear Regression, Random Forest, Gradient Boosting, Ridge,K-Nearest Neighbors, Decision Tree and Artificial Neural network.**

I. Introduction

Nowadays, the usage of ICT in agriculture sector is increasing tremendously in the different countries of Africa Continent country. An agriculture sector is one among the factors that play a vital role in the development of the African continent. These are justified by a high percentage of citizens who are participating in this sector. The rate of increasing of the population in Africa is high and the demand for increasing the food production must be proportional. So, it is very important for taking measures on how to increase the crop production.
The Crop production is mainly depending on the soil properties of plant interaction. It is very important for the farmers to determine the soil fertility requirement for better and economical crop production. Soil pH is the most important soil parameter because it gives more information about many aspects of the soil fertility [1]. Major soil nutrients present in soil and that contribute in yield production include Phosphorus ,Potassium, Nitrogen,Calcium and pH [1]–[4]. Due to the insufficient rate of nutrients or excess fertilization may lead to the lower yield in the crop production [4]. However, the appropriate quantity of fertilizer is required for better plant growth.
In Rwanda and even in the other countries of African continent, most farmers use to imagine the quantity fertilizer to be used during of soil fertilization where the problem of using much or less fertilizers is possible. Measuring the nutrients concentration present in the soil can help to get the soil nutrients to be provided and select the suitable crop for the specific soil sample identified. Thus, the use of ICT in agriculture activities could play a vital role in sustainability of optimum agriculture and reducing the environmental impacts and economic losses.
It is in line the researcher in this current research, used the Machine Learning Algorithms concept to generate a predictive models that might help the farmers to know the quantity of soil composition and identification of the crops to be planted in the environment based on the soil composition predicted or identification of the quantity of nutrients to be added in the soil for keeping the soil fertility consistent. The dataset used by the Machine Learning Algorithm includes different features such as soil mid-infrared absorbance capacity and spatial spectral features provided by remote sensing data source like NASA [5] because determining the soil nutrients the spatial data can also play the great importance [6].
Therefore, there a need of an adequate technology to provide the adequate spatial data. The targeted values were soil Calcium ca, Phosphorous P and pH. For selecting the best predictive model different Machine Learning algorithms such Linear Regression, K-Nearest Neighbors (KNN) , Decision Tree, Random Forest, Gradient Boosting, and Artificial Neural network were implemented and investigated using python programming language packages to confirm the best algorithm for this particularly dataset. Python has seem to be a stable, flexible and popular language and makes many tools available for the researchers from development to deployment and maintenance of an AI project [7]–[9].

II. PROBLEM STATEMENT

The crop production mainly depends on the rate of soil nutrients. It is important for the farmers to determine the soil fertility requirement for increasing the crop production yield. Due to having insufficient information needed about soil composition, this might cause a serious problem of planting without nutrients present in the soil knowledge. However, this may lead to low agriculture production quality and less

crop production yield because using inappropriate rate of soil nutrients which in return lead to the crops degradation. Improper usage of soil fertilizers may results into the poor quality of production [1]. For instance in China inappropriate usage of fertilizer caused the low product quality and even critical environmental problems [10]. Thus this current research evaluated various Machine Learning algorithms [11], [12] to confirm a good predictive model which allow the farmer to predict the quantity of soil nutrients present in the soil.

III. OBJECTIVE OF THE STUDY

The main objective of this current research study project is to implement and compare different Machine Learning Algorithms by using python libraries for generating a best predictive model to be used for helping the famers to know the quantity of the soil composition specifically Calcium Ca, Phosphorus P and pH based on the type of the soil samples provided such as mid-infrared absorbance capacity and soil spatial spectral features generated by a remote sensing data source (Satellite).

IV. PREVIOUS WORKS

This section gives the brief description and analysis of existing related research projects. These include the problem investigated by the previous researchers, proposed technical solution, methodology used, and results found. Finally, the gaps identified through those existing project have been explained as the motivation of the current proposed research. Through [1], Shylaja S.N. and Dr.Veena M.B.have developed a Wireless Sensor Network architecture for collecting the soil nutrients information. Through this project the major soil nutrients collected were Potasium K, Phosphorous P and Nitrogen N. These data were collected by using the sensor technology and in return the collected information were sent to IBM cloud platform database to be stored through IoT backbone architecture. This system provided the people interface for accessing to the data through their mobile phone.

In [13], R. Ajith kumar et. al, have done a deep research on statistical soil nutrients analysis on three thousand and eight hundred soil samples of Thrissur district by using R software. The main nutrients analyzed were soil pH value, oganic carbon, and electrical conductivity, phosphorus, potassium, calcium, Magnesium, Sulfur, Zinc, Boron, Iron, copper and Manganese. Their results of the study shows that there is a strong correlation between those soil nutrients in Thrissur district .

Amrutha A et.al [3] developed a system for collecting the soil nutrients information from the soil by using sensors technology. The interested soil properties were Phosphorous, Potassium and Calcium. At the end they have developed an automated system for identifying the amount of nutrients to be added based on the measured nutrients from the soil for avoiding using excess or insufficient fertilizers which may lead to plant degradation.

The researcher in [4], identified a soil fertility analyzer system with a Soil Test Kit. His study aimed to predict the soil nutrients present in the soil using image processing and artificial neural network implemented by using Matlab. Those nutrients include soil pH ,Zinc, Phosphorus, Potassium, Nitrogen and Calcium.

Viraj. A. Gulhane et.al [14] preliminary have investigated the soil properties which include electrical conductivity, carbon, pH level, phosphorous and potassium contents of soil samples. Secondary they have collected soil spatial data by using the remote sensing data source which is satellite from improving the predictive model . Then the data have been analyzed via Matlab and the results showed the strong correlation between soil nutrients and wavelet transformation.

V. METHODOLOGY

A. Data Collection

The dataset used in this research is accessible online and has been prepared by the Africa Soil Information Service (AfSIS) in their project dated from 2009 up two 2012 [5]. The dataset contains 3594 inputs (features) which include various soil mid-infrared absorbance capacity measured by using spectroscopy and spatial data collected by remote sensing data source (Satellite). The response or targeted values involve Calcium Ca, Phosphorous P and pH level.

B. TRAINING A MODEL

In this section the training inputs data are soil mid-infrared absorbance capacity information and the spatial data measured by remote sensing data source like NASA. Each dataset used contains 3597 elements that include 3594 features, three targets and 1,157 samples. The training labels (targets) were Calcium Ca, Phosphorous P and pH level [5]. These labels are a column matrix of three elements and 1,157 samples matching with features samples. During of model training the dataset was first split into three small datasets for predicting pH, Calcium and Phosphorus respectively. Each dataset among those small dataset generated contains 3594 features, one target and 1157 samples. After dividing the original dataset into three small datasets, each dataset was split into training data and test data. The training process has been carried out via jupyter notebook software package provided by anaconda distribution software.

During the training process of Artificial Neural Network, the back-propagation neural network was used for improving the accuracy of the algorithm by capturing different interaction between different parameters [15]. For analyzing the target parameter, each parameter has been analyzed alone using the same inputs elements and the same number of samples because the dataset contains more target parameters. During of model training, the given dataset was split into two sub-datasets which include 85% and 15% for training and testing the model respectively. Each Neural Network Diagram used for each specific target is described in the following section.

The model architecture used for training an Artificial Neural Network to predict the pH include a hidden layer

of 10 unit of neurons and an output layer with one neuron. The layers used include input layer (3594 features), middle layer (hidden layer) and output layer. All of these layers are interconnected by using the weighted links. During of training, MLPRegressor python library was used for implementing the neural network algorithm. The specific parameters of MLPRegressor considered include $solver = lbfgs, max_iter = 10000, alpha = 0.1, random_state = 1, hidden_layer_sizes = 10, activation function = relu.$

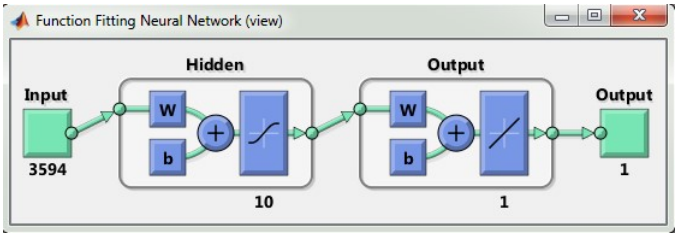

Fig.1: Neural Network diagram for pH predictive model

While training an Artificial Neural Network to predict the Calcium level,a hidden layer of 10 unit of neurons and an output layer with one neuron have been used. The layers used include normalized input layer of 3594 features, middle layer (hidden layer) and output layer. All of these layers are interconnected by using the weighted links. During of training, MLPRegressor python library was used for implementing the neural network algorithm. The specific parameters of MLPRegressor considered include $solver = lbfgs, max_iter = 10000, alpha = 0.1, random_state = 1, hidden_layer_sizes = 10, activation function = relu.$

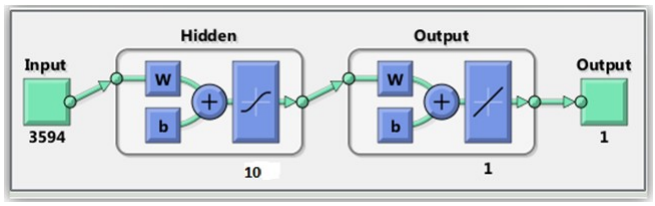

Fig.2:Neural Network diagram for predicting Ca

During of phosphorus predictive model development an Artificial Neural Network of a one hidden layer of 50 unit of neurons and an output layer with one neuron have been used. The layers used include normalized input layer of 3594 features, middle layer (hidden layer) and output layer. All of these layers are interconnected by using the weighted links. During of training, MLPRegressor python library was used for implementing the neural network algorithm. The specific parameters of MLPRegressor considered include $solver = lbfgs, , random_state = 42, hidden_layer_sizes = 50, activation function = relu.$

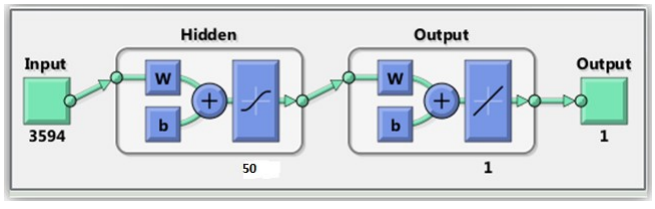

Fig.3:Neural Network diagram for generating Ca and P.

For implementing the remaining Machine Learning algorithms like Linear Regression, Ridge ,Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors the python programming and python library packages were used to prepare the training and testing data, and even to train and evaluating the models. While training and evaluating the models, it was required to normalize the input features based on the type of the model and the target variable. The following tables include the parameters used based on the model target and python library packages used for implementing such machine learning algorithms.

1) pH level training parameters:

Target	Name of Algorithm used	Python Implementation library Package	Features normalization	Parameters used
pH	Linear Regression	Linear Regression	No	defaults
pH	Ridge	Ridge	No	alpha=1
pH	Decision Tree	DecisionTreeRegressor	No	random_state 50, max_depth = 5
pH	Random Forest	RandomForestRegressor	No	Random_state = 42, n_estimators = 9
pH	Gradient Boosting	GradientBoostingRegressor	No	defaults
pH	K-Nearest Neighbors	KNeighborsRegressor	No	n_neighbors = 5

Table1:Summary of pH level training parameters

2) Calcium(Ca) training parameters:

Target	Name of Algorithm used	Python Implementation library Package	Features normalization	Parameters used
Ca	Linear Regression	Linear Regression	No	defaults
Ca	Ridge	Ridge	No	alpha=0.1
Ca	Decision Tree	DecisionTreeRegressor	Yes	random_state=1, mxa_depth=5
Ca	Random Forest	RandomForestRegressor	No	Random_state = 42, n_estimators = 9
Ca	Gradient Boosting	GradientBoostingRegressor	No	defaults
Ca	K-Nearest Neighbors	KNeighborsRegressor	No	n_neighbors = 5

Table2:Summary of Calcium training parameters

3) Phosphorus(P) training parameters:

Target	Name of Algorithm used	Python Implementation library Package	Features normalization	Parameters used
P	Ridge	Ridge	No	alpha=0.1
P	Decision Tree	DecisionTreeRegressor	Yes	max_depth=3
P	Random Forest	RandomForestRegressor	Yes	Random_state=1, n_estimators=6
P	K-Nearest Neighbors	KNeighborsRegressor	No	n_neighbors = 5

Table3:Summary of Phosphorus training parameters

C. MODEL TESTING AND VALIDATION

During of model training and evaluation, each model used 85% of training dataset to train the model. However, 15% of the dataset were used during of model evaluation. During of the model evaluation the prediction accuracy score was used as the performance metric.

VI. RESULTS DISCUSSION

The results show that different machine learning algorithms provided different prediction accuracy for the same dataset applied. This means that the choice of the model is very important for the given specific data because some models may tend to overfitting or underfitting due to the input data. Secondly by applying the same model to the same dataset gives different prediction accuracy when different specific parameters are selected. Thus, the choice of the model parameters is very important for preventing the learning algorithm from overfitting or underfitting and even to increase model performance. The prediction accuracy was increased also by increasing the number of training samples. So, the size of the dataset is an important factor for improving a predictive model. The prediction accuracies generated during of model training and evaluation processes are summarized in the following tables based on the target variable.

A. Calcium Prediction

Target	Name of Algorithm used	Training Accuracy	Testing Accuracy	Observation
Calcium	Neural Network	99.1%	93.5.%	Excellent
Calcium	Linear Regression	100%	65.0%	Model overfitting
Calcium	Ridge	94.6%	90.3%	Very Good
Calcium	Decision Tree	92.7%	86.2%	Very Good
Calcium	Random Forest	96.9%	93.1%	Excellent
Calcium	Gradient Boosting	95.8%	87.9%	Very Good
Calcium	K-Nearest Neighbors	91.3%	84.8%	Very Good

Table4:Calcium Prediction accuracy

B. pH level Prediction

Target	Name of Algorithm used	Training Accuracy	Testing Accuracy	Observation
pH	Neural Network	93.5%	86.5%	Very good
pH	Linear Regression	100%	65.3%	Model overfitting
pH	Ridge	84.1%	81.5%	Very good
pH	Decision Tree	70.4%	69.2%	Good
pH	Random Forest	94.3%	78.3%	Good
pH	Gradient Boosting	94.3%	78.3%	Good
pH	K-Nearest Neighbors	80.1%	68.3	Model overfitting

Table5:pH Prediction accuracy

C. Phosphorus level Prediction

Target	Name of Algorithm used	Training Accuracy	Testing Accuracy	Observation
Phosphorus	Neural Network	75.4%	18.7%	Model overfitting
Phosphorus	Ridge	38.0%	30.6%	Model underfitting
Phosphorus	Decision Tree	33.2%	14.2%	Model underfitting
Phosphorus	Random Forest	83.1%	26.9%	Model overfitting
Phosphorus	K-Nearest Neighbors	40.7%	36.2%	Model underfitting

Table6:Phosphorus Prediction accuracy

This tables show that the predictive models generated could be used for Calcium and pH level prediction perfectly because the predictive accuracy is high in both training and testing process. For predicting the Calcium level contained in the soil, the results show that Artificial Neural Network and Random Forest Machine Learning algorithms are the best algorithms to be used because their training and testing accuracy are very high. The same the results show that the Artificial Neural Network and Ridge algorithms could be used to predict the pH level contained in the soil as the training and even testing accuracy are both high. However, the predictive models generated could not be used for predicting the Phosphorous quantity contained in the soil for the given dataset because the accuracy is very low in both training and testing and model tends always to overfit or underfit. During of model training, the linear regression machine learning algorithm tends always to overfit in all predictors made and this was corrected by using Ridge regression to prevent the model from overfitting.

VII. CONLUSION AND RECOMMENDATION

Finally, the results show that the generated predictive models are able to predict the quantity of pH level, and the Calcium (Ca) by applying unseen data. For predicting the Calcium level contained in the soil, the result shows that the Artificial Neural Network and Random Forest Machine Learning algorithms are the best algorithms to be used because their training and testing accuracy are very high. The results show also that the Artificial Neural Network and Ridge Machine Learning algorithms could be used to predict the pH level contained in the soil as the training and even testing accuracy are high. Furthermore, the results show that increasing the number of hidden layers for Artificial Neural Networks can affect the predictive accuracy. Thus the choice of number of hidden layers to be considered during of model training is very important.

However, the results show also that the predictive model generated could not be used to predict the Phosphorous quantity with the given specific dataset because the predictive accuracy of the model was very low during the training and testing process. These indicates that the correlation between the dataset features and the target variable was very low. For being able to predict the Phosphorous with the given dataset I suggest to add additional features in order to see if the model could be improved.

REFERENCES

[1] D. Vadalia, M. Vaity, K. Tawate, D. Kapse, S. V. Sem, and C. Engg, "Real Time soil fertility analyzer and crop prediction," Int. Res. J. Eng. Technol., vol. 4, no. 3, pp. 3–5, 2017.

[2] S. N. Shylaja, "Real- T ime Monitoring of Soil Nutrient Analysis u sing WSN," 2017 Int. Conf. Energy, Commun. Data Anal. Soft Comput., pp. 3059–3062, 2017.

[3] A. Amrutha, R. Lekha, and A. Sreedevi, "Automatic soil nutrient detection and fertilizer dispensary system," Proc. 2016 Int. Conf. Robot. Curr. Trends Futur. Challenges, 2017, doi: 10.1109/RCTFC.2016.7893418.

[4] J. C. Puno, E. Sybingco, E. Dadios, I. Valenzuela, and J. Cuello, "Determination of soil nutrients and pH level using image processing and artificial neural network," HNICEM 2017 - 9th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag., vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/HNICEM.2017.8269472.

[5] "KAGGLE," https://www.kaggle.com/c/afsis-soil-properties/data, 2019.

[6] H. Zheng, J. Wu, and S. Zhang, "Study on the spatial variability of farmland soil nutrient based on the kriging interpolation," 2009 Int. Conf. Artif. Intell. Comput. Intell. AICI 2009, vol. 4, pp. 550–555, 2009, doi: 10.1109/AICI.2009.137.

[7] Prince Patel, "Why Python is the most popular language used for Machine Learning," 2018. [Online]. Available: https://medium.com/@UdacityINDIA/why-use-python-for-machine-learning-e4b0b4457a77. [Accessed: 20-Mar-2020].

[8] N. Gupta, "Why is Python Used for Machine Learning?," 2019. [Online]. Available: https://hackernoon.com/why-python-used-for-machine-learning-u13f922ug. [Accessed: 20-Mar-2020].

[9] A. Beklemysheva, "Why Use Python for AI and Machine Learning." [Online]. Available: https://steelkiwi.com/blog/python-for-ai-and-machine-learning/. [Accessed: 20-Mar-2020].

[10] D. V Ramane, S. S. Patil, and A. D. Shaligram, "Detection of NPK nutrients of soil using Fiber Optic Sensor," Int. J. Res. Advent Technol. ACGT, no. February, pp. 13–14, 2015.

[11] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," Int. Conf. ICT Knowl. Eng., pp. 1–6, 2018, doi: 10.1109/ICTKE.2017.8259629.

[12] A. Angra, Sheena; Sachin, ",, /,7(5$785(6859(¡," pp. 57–60, 2017.

[13] R. A. Kumar, M. K. M. Aslam, V. P. J. Raj, T. Radhakrishnan, K. S. Kumar, and T. K. Manojkumar, "A statistical analysis of soil fertility of Thrissur district, Kerala," Proc. 2016 Int. Conf. Data Sci. Eng. ICDSE 2016, pp. 7–11, 2017, doi: 10.1109/ICDSE.2016.7823953.

[14] V. A. Gulhane and S. V. Rode, "Correlation analysis on soil nutrients and wavelet decompositions of satellite imagery," 2015 Int. Conf. Ind. Instrum. Control. ICIC 2015, no. Icic, pp. 1225–1230, 2015, doi: 10.1109/IIC.2015.7150934.

[15] P. Padmasree and R. Maheswari, "A Novel Technique for Image Compression in Hand Written Recognition using Back Propagation in Neural Network," vol. 4, no. 06, pp. 763–768, 2013.