National College of Ireland

Soil Degradation Prediction and Classification using Digital Soil Maps:Boosting Nigerian Food Security

MSc Research Project
Programme Name
Masters in Data Analytic in Data Science

Oyinkansola Shittu
Student ID: X20147406

School of Computing
National College of Ireland

Supervisor: Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Oyinkansola Shittu
Student ID:	X20147406
Programme:	Masters in Data Analytic
Year:	2023
Module:	MSc Research Project
Supervisor:	Catherine Mulwa
Submission Due Date:	14/08/2023
Project Title:	Soil Degradation Prediction and Classification using Digital Soil Maps - Boosting Nigerian Food Security
Word Count:	XXX
Page Count:	46

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Oyinkansola Shittu
Date:	14/08/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Soil Degradation Prediction and Classification using Digital Soil Maps-Boosting Nigerian Food Security

Oyinkansola Shittu
X20147406

14th August 2023

Abstract

The global threat to food security in recent years and the uncertainties around it motivated this research. The experiment is channelled towards assisting the Nigerian government in the plan to improve farmers economic well-being and food security in Nigeria. In this research, eight machine learning models were developed to predict and classify soil pH and soil textures using the Nigerian digital soil map (two for prediction and six for classification). The models are support vector machine for regression, random forest for regression, k-nearest neighbour (2), support vector machine, non-parametric Naive Bayes (2) and Random forest. Soil PH has been rated high as one of the key indicators of soil organic carbon which in turn, scientists have mentioned is one of the main indicators of soil degradation.The developed models successfully predicted and classified both soil pH and soil texture with very high accuracy and negligible errors. Randomforest was found to be the best of the models for both prediction and classification at accuracy of 1 and relative mean square error of 0.006 and all the developed models outperformed the benchmarked existing models on the evaluating metrics.This successful high performance models confirmed the Nigerian soil map dataset can be used to predict and classify soil degradation with the aim of the Nigerian government educating farmers on suitable crops for each soil type to improve the farmers economic power and indirectly resolving the continuous farmers- herders clashes over farm lands.

1 Introduction

The recent awareness of negative impacts of climate change, health pandemic, and global unrest have brought to fore, the crucial importance of food security, soil and Agro- management across the world with its problems of many ecological factors that are known to have increased the natural degradation of soil quality over the years. The year 2020 announcement by the Nigerian minister of environment that 35 percent of Nigerian landmass is heading towards desertification, caused a growing concern amongst the citizens, renewing the government's commitment to land conservation, soil management and land zoning, with the aim of returning the country to its former food and economic glory. Figure 1[1]. As a young child of a Nigerian commercial cocoa and kolanut farming family, excessive farming, over-grazing, inefficient soil management were direct experiences.

[1]Source:The Nigerian Punch Newspaper-18th June 2020

Figure 1: Nigerian land mass threatened by desertification

These and climate change impact rendered hectares of the family farmland unproductive and degraded, forcing its sale to housing developments, like many other farms as a land use change (LUC). The increase in the unproductive family lands and threat of desertification spurred the primary motivation for the need to ensure food security for the citizens of Nigeria and economic improvements of Nigerian farmers by assisting the Nigerian government with its policies on food security via the prediction and classification of soil degradation using machine learning methods.
In addition, the family-head relayed the Nigerian historic arrangements between local farmers and regional cattle headers, where the herdsmen were allowed to graze their animals on the farmers' farmland in exchange for raw milk, cheese, animal skin and manure, however, easy availability of milk and fertilisers, over the years, reduced farmers dependence on herdsmen, distorting this symbiosis. Also, the improved health system that resulted to larger family allocation of fallow farmlands, drastically cut off the grazing routes of the herdsmen who already were at a disadvantage.This is the generator of the farmers-herders conflict in Nigeria. This generational knowledge set the incentive for the need for sustainable conflict resolution that is the sub-research question of this project.

1.1 Research Questions, Objectives and Contributions

For the mentioned problems, this research addressed the following:
Research Question. *How well would the machine learning models (Random Forest and Support Vector Machine for Regression) predict soil pH as key indicator of Nigerian soil degradation, using the Nigerian digital soil map attributes, to support the plans of Nigerian government towards food security and improve the economic power of farmers in Nigeria?*
Sub-Research Question. *To what extent can the classification models(Support Vector Machine, K-Nearest Neighbour, non-parametric Naive Bayes and Random Forest) help with enhancing classification of soil texture and soil pH to help farmers know the kind of crops that suits the type of soil texture and pH level which will lead to reduction of soil degradation and sustainable farm management, making grazing routes and lands available for herdsmen, which will invariably mitigate the farmers- herders conflicts?*

Literature assessment showed regression, machine learning, and geospatial statistics

Table 1: Project Objectives Summary

Digits	Objectives	Explanation
1	Critical review of existing literature	This critically assesses peer-reviewed literature on soil degradation, conflict resolution and soil maps to identify the degradation key causative factor(s) and methods used for prediction and classification of same in Nigeria, Africa and global.
2	Download of the Nigerian digital soil map (NDSM) dataset	Dataset of the Nigerian digital soil map is freely downloaded from the database at Mendeley.com.
3	Data Extraction	The dataset comprises of both shape and database files. Different extraction methods is used to get these data with extraction, transformation and loading (ETL).
3.1	**Extraction of the database file**	The database files contain attributes of multiple records, fields and varied data types stored in a database schema. Database management tool - DBF Viewer is used to view dataset and SQL query commands to extract data.
3.2	Extraction of the digital soil maps file	These shape files have the digital soil maps. R-algorithms and RStudio are used to extract the map data and view the maps.
4	**Data Transformation**	Extracted data are transformed into '.csv' format for ease of merging and upload.
4.1	Database file	The SQL- extracted data from the database schema is exported as '.csv' and saved in a desktop folder.
4.2	Digital soil map file	The use of 'R' algorithms transforms the extracted digital soil maps into '.csv' format and stored in same folder as database file on local directory.
5	Data Loading	The two transformed CSV files are loaded onto Jupyter notebook ('R' kernel) for further operations.
6	Data upload accuracy and integrity check	Same random sample queries from Jupyter notebook (using 'R' algorithms) and DBF Viewer (using SQL query commands) are compared, ensuring the different reports are same to confirm data conversion accuracy and upload.
7	Exploratory data analysis (EDA) and data pre-processing	Dataset is explored for missing data and regression assumptions analysed and pre-processed for outliers' detection and removal, data imbalance correction (SMOTE), normality conversion (Box-cox and log transformations),PCA feature selection and synthetic data generation (Copula).

Table 2: Continuation of Project Objectives Summary

Digits	Objectives	Explanation
8	**Data Modeling**	Applying relevant identified models from objective1, for prediction and classification models for soil degradation using the Nigerian digital soil map (NDSM) dataset.
8.1	**Prediction Models**	Non-linear models with variables interaction capability.
A	Support Vector Machine for Regression (SVR)	Machine learning regression kernel model.
B	Random Forest (RF)	Machine learning regression tree model.

methods fit well with digital soil map attributes, for predicting and classifying soil degradation to which six (6) main models were developed for this projects, outlined in the project objectives summary (tables 1-3). However, both the K-nearest neighbour and Naive Bayes models had one extra embedded model each making a total of eight (8) models- two for prediction and six for classification. Out of these models random forest for regression performed marginally better than the support vector machine for regression

Table 3: Continuation of Project Objectives Summary

Digits	Objectives	Explanation
8.2	**Classification Models**	Non-linear models with variables interaction capability.
C	K-Nearest Neighbour (KNN)	Machine learning distance-based classification model.
D	Support Vector Machine (SVM)	Machine learning hyperplane classification model.
E	Non-parametric Naïve Bayes (NB)	Machine learning non-parametric kernel classification model.
F	Random Forest (RF)	Machine learning classification tree model.
8.3	Variable interaction detection and effect	The dataset is pre-processed and transformed for detection of any interaction amongst the variables and the effect on the soil pH and soil textures with machine learning algorithms.
9	**Evaluation, Results and Models Comparison**	
9.1	Prediction Models	The evaluation methods are coefficient of determination (R2), Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Error (MAE).
9.2	Classification Models	For Classification, they are Accuracy, Precision, Recall and F1 Score.
9.3	Models Comparisons	This section is on three levels- (1) Comparing the evaluated results of all the implemented models to identify that with best result in each category and (2) Comparing results of the models used for both regression and classification (RF, SVR and SVM) to rank method performance. (3) Comparing model results with the existing models results to rank this project's outcome.

with slightly higher evaluation metrics. Amongst the classification models, again, random forest for classification performed marginally best in classifying soil pH with K-nearest neighbour close.

Contributions: The major contributions resulting from this research are the machine learning applied soil degradation prediction and classification models using the Nigerian digital soil map dataset to assist the Nigerian government in farm and conflict management policies to enhance the farmers' economy. Other contributions are the test for possible variables interaction and accuracy measures of models since the creators of the Nigerian digital soil maps dataset neither tested for the interaction between the data attributes nor for result accuracies. The data integrity check is another contribution to ensure accuracy and reliability of data upload from the database schema to the modeling platforms.

Chapter two critically reviewed the soil degradation (2017-2023) at (2.1) starting with the understanding of soil degradation and soil maps(2.2), proffered Nigerian conflict resolutions using soil maps(2.2.1) to the global review of machine learning methods for soil degradation prediction and classification (2.2.2), model comparison (2.3) and concluded with summary of findings and identified gaps (2.4). Chapter three covered the soil degradation scientific methodology and design flow process where the first three steps of followed methodology were described leading to the implementation,evaluation and res-

ults of both the prediction and classification models in chapter 4. Chapter 5 followed with models comparison and discussion where the project's models were compared together and with existing models, and learning outcomes and limitation were highlighted while chapter 6 has the conclusion and future works presented. Acknowledgement and references are found at the end of the project.

2 Critical Review Soil Degradation (2017 – 2023)

2.1 Introduction

A critical review of existing experiments was carried out to identify methods that suit the project's dataset and models to be used as basis of result comparison of this project. The literature covered recent seven years (2017-2023) because of the fast changes in Machine Learning algorithms and the sub-sections are grouped into themes that answer the two research questions. (2.2) laid the foundation of soil degradation and soil map understanding while (2.2.1) reviewed literature where soil maps have been used in suggesting resolution to the Nigerian farmers-herdsmen conflicts (the sub-research question) then to answer the main research question, (2.2.2) explored traditional methods of using soil maps and statistics for predicting and classifying soil degradation, to appreciate the application of machine learning techniques to same, after which the best methods and evaluating metrics were deduced. Section (2.2.2) has a global coverage because of limited relevant peer-reviewed machine learning literature on soil degradation with soil maps in Nigeria and Africa. A comparison of the reviewed models is tabularly presented in (2.3) and summary of findings and identified gaps is in (2.4).

2.2 Understanding Soil Degradation and Soil Map

A viable soil is said to be fertile and sturdy to withstand the impacts of climate change and harsh farm management practices while soil degradation, in its simplicity, can be said to be the decline and/ or loss of this fertility and sturdiness. Soil deteriorate on many factors that are generally grouped into physical, chemical, biological and ecological, Figure 2[2], making it prone to erosion, drought, soil organic carbon (SOC), salinity loss and reduced cation exchange capacity (CEC), which ultimately affect the sustainability and reproducibility of crops and animals- hence the need to monitor and conserve this natural resource. This soil health quality is strongly linked to many features like CEC- an indicator of soil fertility, that measures the soil's ability to hold and supply positively charged cations soil nutrients (Calcium, Magnesium, Sodium, Potassium and Aluminium) to plant, while soil texture (Clay and Silt content) have negatively charged colloids that electrically hold large amounts of cations thereby serving as warehouse of nutrients for plants roots. The department of primary industries, Australia ,explained the stronger the colloids negative charge, the greater its capacity to hold and exchange cations between the soil water and plant roots which is why soil pH is important as the higher the pH (less acidic), the higher the negative charge of the colloids as found in humus and clay soils, (Al-Kaisi & Lowery 2017).
The magnitude of SOC are very important too as SOC provides resistance to physical degradation, erosion , water and nutrient loss (Tadesse et al. 2023) and the carbon content,

[2]Source:Restoring soil quality to mitigate soil degradation- Lal (2015)

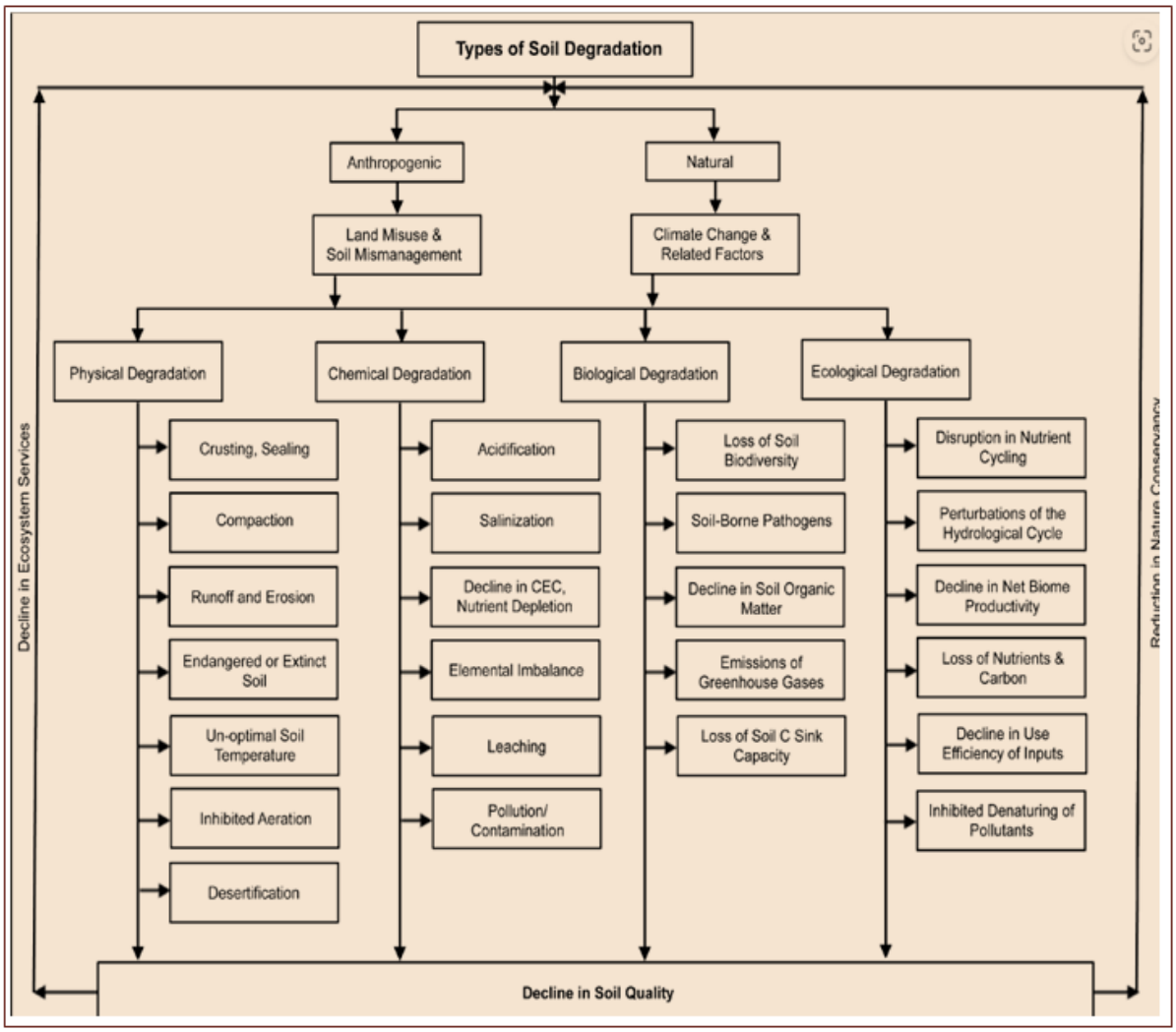


Figure 2: Four-Grouped Types of Soil Degradation

as evidenced in humus soils, can be measured by pH, however,intensive agro-production, land-use mis-management, LUC and excessive rainfall altered these key soil properties in African countries, rendering soil susceptible to degradation causes like erosion (Bennett et al. 2021) making many scientists believe decline in soil health qualities are good indicators of soil degradation (Lu et al. 2022) .
According to the food and agricultural organisation (FAO), Nigerian soils are of medium to high productivity due to the challenges of its soil management practices that result in loss of soil carbon, hence the choice of using PH as the main soil degradation indicator for this project as it measures carbon, cations and colloids.

Soil mapping on the other hand, is a process of finding recurring patterns in soil clusters, marking, grouping and transforming those patterns into map units to reproduce the recognised soil information in spatial map format as points, lines or polygons. Although it is central to the interpretation of soil properties distribution and execution of maintainable practices for soil degradation prevention, the traditional method using soil survey from manual soil testing to produce legacy maps,intrinsically losses spatial variability information for prediction and classification of soil properties, thus the increased

need for high quality soil maps (Grundy et al. 2020) reliable for statistical and machine learning techniques to predict and classify soil properties especially in areas of little information (Shepherd et al. 2022). Recent possibilities of legacy maps being updated with data mining, (Liu et al. 2022) caused the development of quantitative digital soil mapping (DSM), where practitioners cross-reference primary soil observations with secondary data from which a model is used to describe the relationship between the two and predict and/or classify those relationships for other locations, following defined common set of rules (Kidd et al. 2020).

2.2.1 Critical Review of Conflict Resolutions using Soil Maps in Nigeria

There are two generally known approaches applied to conflict resolutions in Nigeria: The western approach that focuses on litigation and the ethnical African approach that focuses on mediation. Scholars claim western approach to conflict resolution is rooted in the community's individualistic nature while the African system is based on the holistic or communal nature of African communities. Result of an analysed survey of (Okeke-Ogbuafor et al. 2019) showed although beneficiaries applauded both conflict resolution approaches, some do not feel academically skilled for and/or afford to use the western approach while some perceive the ethnical African systems are more of a conflict generator than conciliation due to the ingrained potential bias. Also, (Mosweu Plefhile 2023) found archival materials like drawings, plans, maps and scientific documents are accepted as significant influential evidences in the international Court of Justice for border conflict resolutions in Nigeria and Africa, although reliance on maps alone is cautioned on as past tribunals have placed reservations on some (Olademo et al. 2021). Since many Nigerian farmers and herdsmen are uneducated and relatively poor, the western approach is financially unattainable and as the use of archived maps influence strong conflict resolution in both the western and African approaches, a review of how soil maps have been used to suggest conflict resolutions is essential for the sub-research question.

In Nigeria, agricultural and zoning policies are based on geopolitical borders but with population increase, tribalism, poor soil nutrient management and lack of monitoring, border crossing for both farming and animal grazing has become an underlying cause of the herdsmen- farmers conflict. (Wonah & Bullem 2019) explained how herdsmen search of grazing lands resulted in cultural and religious affiliations, social life and marriages amongst the settlers overtime, making maps of similar interest easy for conflict policy dissemination through stakeholders education. This encouraged some to advocate for grazing bill to amalgamate these similar interests but opposing the suggestion of legislating a grazing bill, spatial analysis of soil maps by (Amusan Lere 2017) showed the bill will affect the Nigeria's cultural diversity and aggravate the conflict thus orated sedentary cattle ranching is better as a resolution. These soft approaches align with both the food and agricultural organisation education in 2018 for Nigerian farmers, and the conclusion of (Wonah & Bullem 2019) that success of such educational practises is evidenced in the reduced Ghanaian farmers-herdsmen clashes although the effectiveness relies on herdsmen willingness to be established in ranches but the ranch-settlement is being vehemently resisted by the Nigerian fulani herdsmen body- The Miyelti Allah cattle breeders association.

These soil maps use for conflict resolution tilt towards the ethnical African approach and as it can be scientifically evidenced, the probable bias in this approach is greatly reduced ,thus this project followed the educational suggestion based on machine learning

classification soil textures.

2.2.2 Justifying Project Machine Learning Techniques- Global Soil Degradation Prediction and Classification

Predictive soil mapping, a sub-set of applied predictive modeling, is a cross-sectional field of soil science, machine learning and statistics, that aims to produce the most precise and impartial predictions of soil variables for specific requirements. Developing Composite Soil Degradation Index (CSDI) of Nigerian cocoa agroecosystems, (Aderele et al. 2017) used series of multivariate-statistical techniques on various soil samples of south-western cocoa farms at a maximum soil depth of 20cm to predict and classify soil degradation, using factor analysis with principal component analysis and stepwise discriminant analysis, to get the final four (4) key soil degradation indicators- Clay, CEC, Zinc and Organic matter- with the classification validated with combinatorial data analysis and (Falaki et al. 2020) computed quantitative area for Katsina state in Nigeria, using Landsat images and supervised classification and evaluated with accuracy and Kappa metrics of the classified images and IDIRISI Kilimanjaro software used to predict desertification up to year 2030. As spatial soil samplings of legacy data are driving and limiting factors to DSM-model performances, (Lagacherie et al. 2020), in Tunisia, evaluated the impact of these factors using Random Forest (RF) on varied soil sizes, spacing, distribution types and varied legacy data handling, found that performance increased with large spacing, complete and even distribution and local uncertainties were underestimated for sparse samplings and vice-versa.

Leveraging on the benefits of machine learning (ML) scrutiny of complex data to recognise unidentified variables patterns and compatibility with remote sensing data (RSD), (Hengl et al. 2021) re-worked the 2017 African soil information system (AfSIS) project by applying spatially-adjusted 2-scale ensemble ML technique to predict and classify African soil nutrients, incorporating spatial point clusters analysis with sentinel-2 images and extended with soil chemicals (pH and CEC) and physical properties (bulk density, clay, sand and silt). The coefficient of determination (R2) result showed pH, clay and SOC are most correlated for predictability with five-fold spatial cross validation showing pH had the best prediction accuracy at the three depth levels of 0, 20cm and 30 cm similar to (Pahlavan-Rad et al. 2020) in predicting soil water infiltration with RF and multiple linear regression (MLR) where RF predicted better than MLR with root mean square error (RMSE), MAE and 10-fold cross validation results. Likewise, (Yu et al. 2019) predicted and classified indicators of grassland degradation in West Jilin, China using decision tree (DT), partial least squares regression (PLSR) and object-based image analysis to measure spatial distribution of landsat OLI images, RF for variable selection and regression for prediction and (Amuyou et al. 2022) predicted above-ground biomass in Cross-river, Nigeria with sentinel-2 images and recursive wrap with inbuilt RF method on vegetative and climatic variables, both used RMSE and R2 for evaluation to confirm vegetative indices, topography, soil salinity and air temperature are important predictors for soil degradation. To overcome the limitation of RSDs and hyper-spectral experiments in under-estimating accuracies for large areas, (Peng et al. 2019) predicted soil nutrient content in China, with three separate models- PLSR, Back propagated neural network (BPNN) and Genetic algorithm-back propagated neural network (GA-BPNN) to images and discovered GA-BPNN had best prediction accuracy, when GA was used for optimisation with relative root mean square error (RRMSE) as evaluation metric.

For experiments that used digital soil maps (DSM), (Baltensweiler et al. 2021) used ML to produce DSM from legacy (paper) map data in Switzerland, across six (6) models-lasso, robust external-drift kriging, geoadditive modelling, quantile random forest regression (QRF), cubist and support vector machines with each map model's predictability enhanced by weighted model average approach on multi-scale terrain variables, (pH, soil organic carbon, clay, sand, gravel, soil density) and remote sensing vegetative cover data, discovering QRF performed best from R2 result while (Haghighi et al. 2020) predicted soil stability indices and SOC in Iran, using DSM for five (5) ML models (RF, k-nearest neighbours (KNN), support vector machine (SVM), artificial neural network (ANN) and the ensemble of four single models), trained with repeated 10-fold cross-validation method and both found KNN and SVM models were best for SOC, RF best for soil stability index mean weighted diameter (MWD), and the ensemble model increased the prediction accuracies for all and both used evaluation metrics of R2, RMSE, mean absolute percentage error (MAPE), mean absolute error (MAE), while (Haghighi et al. 2020) added normalized RMSE (nRMSE), as extra.

Despite extensive soil studies, accurate mapping of DSM is still difficult, a challenge picked up by (Chen et al. 2019) in predicting SOC in China where six (6) DSM models were compared: geostatic models- ordinary kriging (OK), geographically weighted regression (GWR); ML models- multilayered perceptron neural network with back propagated algorithm (MLPNN), SVM for regression (SVR); hybrid models- ANNkriging (ANNK), GWRkriging (GWRK) with accuracy values found reducing in the order ANNK, SVR, ANN, GWRK, OK and GWR with R2, RMSE, and MAE. Also ML and hybrid models found to be more suitable for regional terrains and desperateness with environmental, soil properties, climate, topography and RSD used as variables as (Guo et al. 2017) stated GWRK better performed than partial least squares regression kriging (PLSRK).

Many of the ML techniques can be used for both prediction and classification as evidenced in the works of (Cho et al. 2023); (Pham et al. 2021); (Padmapriya. & Sasilatha 2023) to classify soil using (C4.5 DT); (Tree Algorithm, ANN and Adaboost); (Naïve Bayes, KNN, SVM and Deep Learning) respectively all in India with evaluation based on accuracy, precision, F1score, recall and confusion matrix. Although one main challenge of DSM-complicated models, is the inability to clearly quantify and evaluate the importance of each covariate of a model, researchers suggested accuracy of any ML-DSM-model can be improved with model-diagnostic interpretation tools like partial dependence plot, SHapley values, and permutation approaches for evaluating feature importance analysis (Taghizadeh-Mehrjardi et al. 2021).

2.3 Comparison of Soil Degradation Reviewed Models

A comparison of the reviewed models that gave better evaluation result in each study and relatable to this project's objectives is presented in table 4 below. These models were carefully selected for the project's execution on the basis of the high result for prediction and classification, the data sizes, capability to handle feature interaction and non-linearity.

2.4 Summary of Findings and Identified Gap

The critical review unveiled small soil sample size does not debar the development of models for machine learning techniques with clay, silt, CEC, SOC, PH, topography and

Table 4: Reviewed Models Comparison

Experiment Summary	Model Types	Evaluation Metrics	Results	Data Size	Author(s)
Organic carbon prediction	Support Vector Regression	RMSE,ME, R-squared	RMSE=8.61, ME=1.63 R-sq=0.30	395	Chen et al. (2019)
Organic carbon prediction	Random Forest Regression	RMSE, MSE, ME, R-squared	RMSE=0.66, ME= 0.02, MSE=0.44, Rsq=0.97	241	Zhang et al. (2017)
Classification of soil types	Tree algorithm	Confusion matrix	Classified Training =328 Testing = 74	440	Pham et al. (2021)
Multi-labeled soil classification	Support Vector Machine (SVM)	Accuracy (A) Precision (P) Recall (R) F1-score (F)	Result ranges A= 0.82 - 0.92 P= 0.84 -0.91 R= 0.83-0.89 F=0.86-0.91	5938	Padmapriya and Sasilatha (2023)
Multi-labelled soil classification	K-Nearest Neighbour	Accuracy (A) Precision (P) Recall (R) F1-score (F)	Result ranges A=0.75 - 0.81 P= 0.71- 0.75 R=0.70 - 0.75 F= 0.69 - 0.78		
Multi-labelled soil classification	Naive Bayes	Accuracy (A) Precision (P) Recall (R) F1-score (F)	Result ranges A= 0.76 - 0.81 P= 0.76 - 0.83 R= 0.76 - 0.80 F= 0.76 - 0.81		

air temperature being the recurring important indicators for soil degradation across the various studies and soil map attributes database can be used successfully to predict and classify soil degradation with high accuracy irrespective of the geological, climatic or vegetative environment of the area of study. It also revealed soil PH is the most important predictor for soil organic carbon (Zhang et al. 2017), (Hengl et al. 2021),- hence the choice of soil PH as the dependent variable for this project.
Identified gaps are the inadequate use of machine learning techniques for farmers- herdsmen conflict resolution in Nigeria, the need to use model-diagnostic tools to enhance the accuracy and interpretebility of model results, and quantifiably evaluate the interaction between the soil components to suggest 'how to replenish lost soil element(s).

3 Scientific Methodology and Design Flow Process

3.1 Introduction

Legacy soil data (paper maps) have been produced for years in majority of countries but sadly such data information and knowledge are still currently fragmented and at risk of getting lost if they remain in a paper format, hence the need to reproduce them

digitally. The dataset used for this project is the Nigerian digital soil map and soil database (NDSM) that was developed from the Nigerian legacy data, digitally processed into consistent, non-spatial quantitative soil information of high resolution (Nkwunonwo & Okeke 2013).

3.2 Soil Degradation Scientific Methodology

The scientific methodology pipeline followed in this project is the knowledge discovery in database (KDD) which was modified into five (5) steps- Data download, Extraction and transformation (ETL), exploration and pre-processing (EDA), modeling and analytics then evaluation and results. These methodology steps are shown in figure 3.

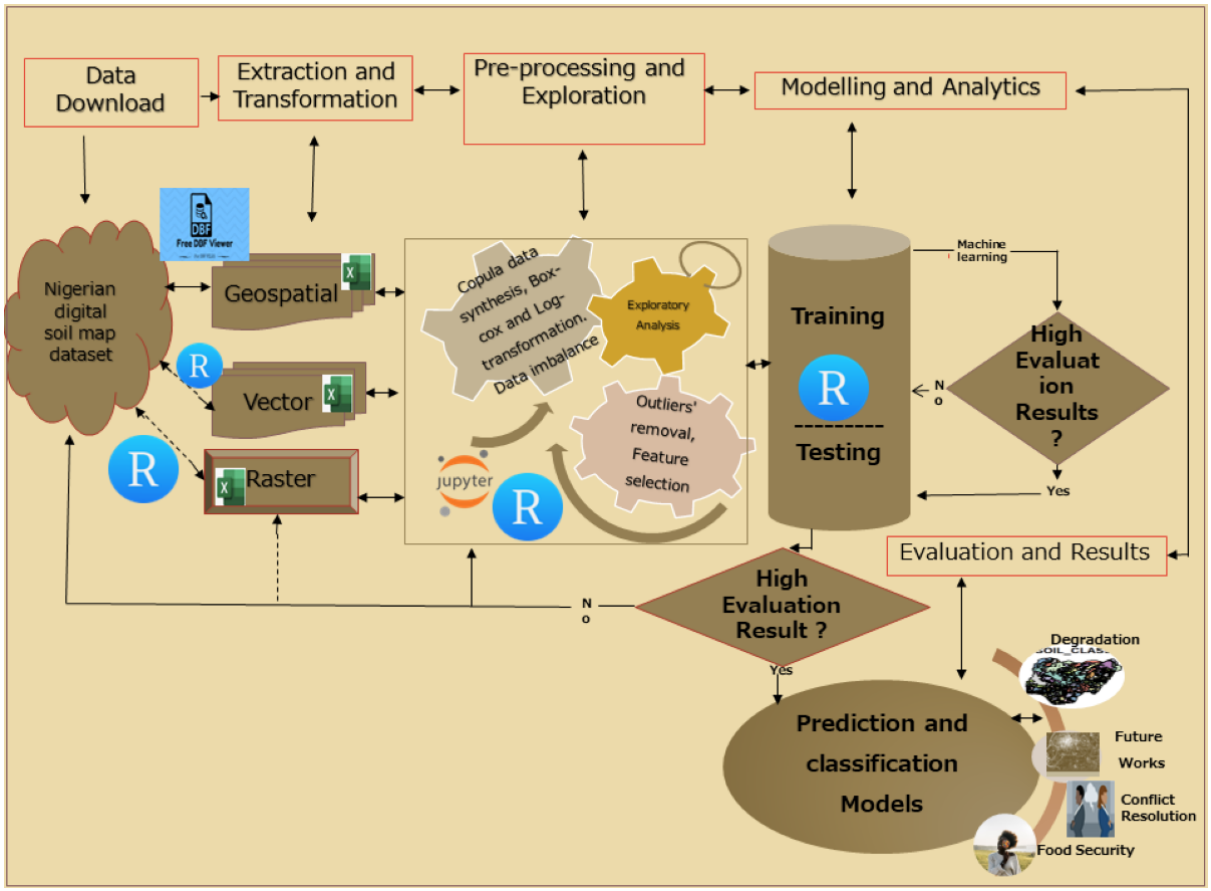


Figure 3: Soil degradation methodology

3.3 Soil Degradation Design Flow Process

The design flow of this project shown in figure 4 below is the details of the implementation (i.e modeling and analytics) step of scientific methodology. Project's objectives 2 – 7 were achieved after the first three (3) methodology steps in figure 3 while objectives 8 and 9 (including project's second contribution) are fully achieved after these implementation steps, although part of the variable interaction effect was tested during the pre-process stage through the careful selection of non-linear models that had the capability to accommodate non-linearity of the data.

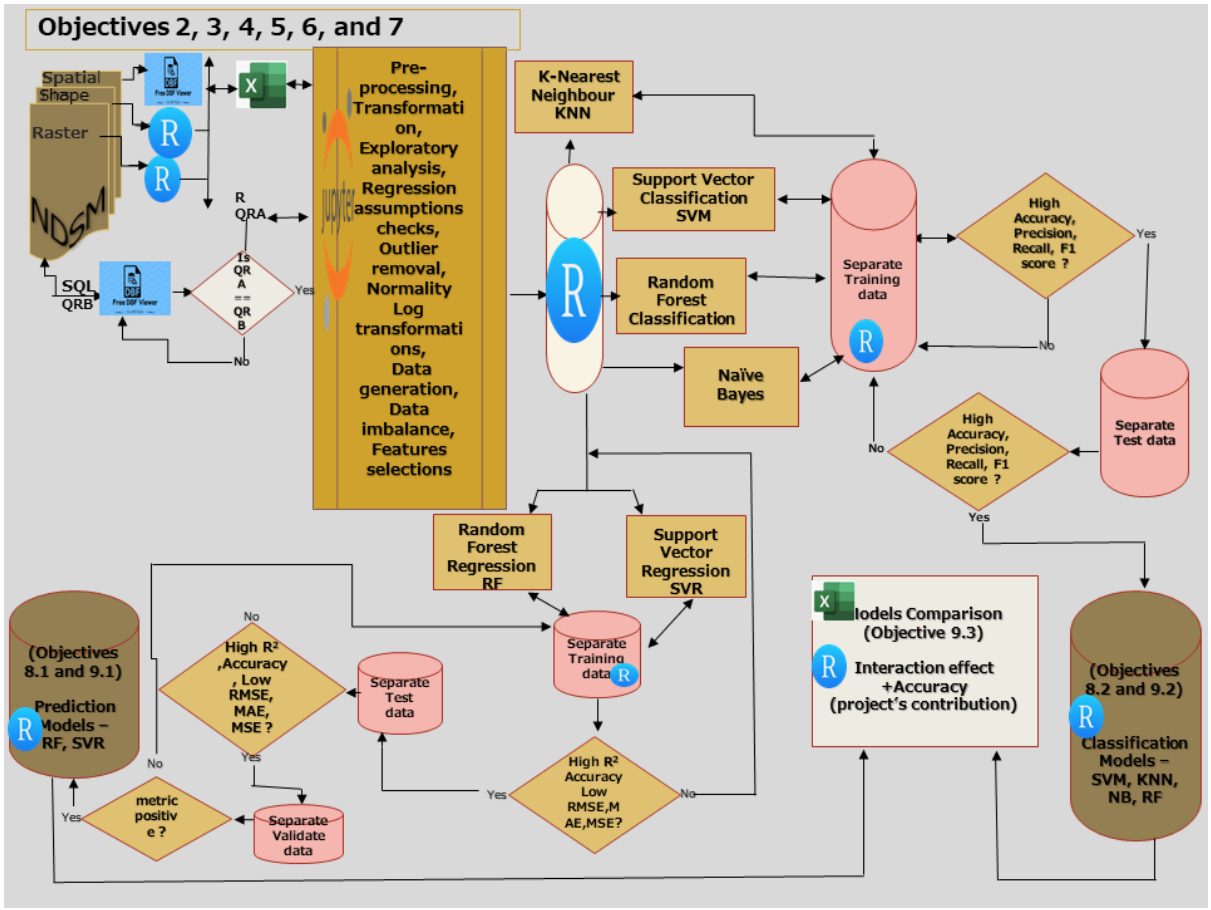


Figure 4: soil degradation design flow process

3.4 Extraction,Loading,Exploration and Data Pre-processing

Data analytic and science is about assessing, exploring and preparing a dataset for model implementation while preserving the data integrity as much as possible. These aspects, mentioned in 3.2, where the dataset was downloaded, how it was extracted, transformed, loaded (ETL) and data integrity checked were discussed in 3.5, and the exploratory data analysis (EDA) and pre-processing in 3.6. The pre-processing part that entailed the checks for regression assumptions, attempts to correct the non-linearity, non-normality and outliers removal, synthetic data generation and imbalance correction are all discussed from 3.3.1 to 3.6.3 with 3.7 concluding this chapter. A visual overview of the first three of the methodology steps are shown in figure 5 for further understanding.

3.5 Dataset Download,Extraction,Transformation and Loading

Data download: This is the first step of the methodology that was effected freely from Mendeley.com repository on to a local desktop folder. The zipped downloaded file had six(6) different ArcGIS files that can be split into the optional files with extensions .prj, .sbn and .sbx and the compulsory files with extensions .shp, .dbf and .shx. The project file (.prj) contains the metadata of the shapefiles without which you cannot get the coordinates and projections of the maps while the spatial index file (.sbn) is used to enhance spatial queries in conjunction with the spatial index file (.sbx) that speeds up

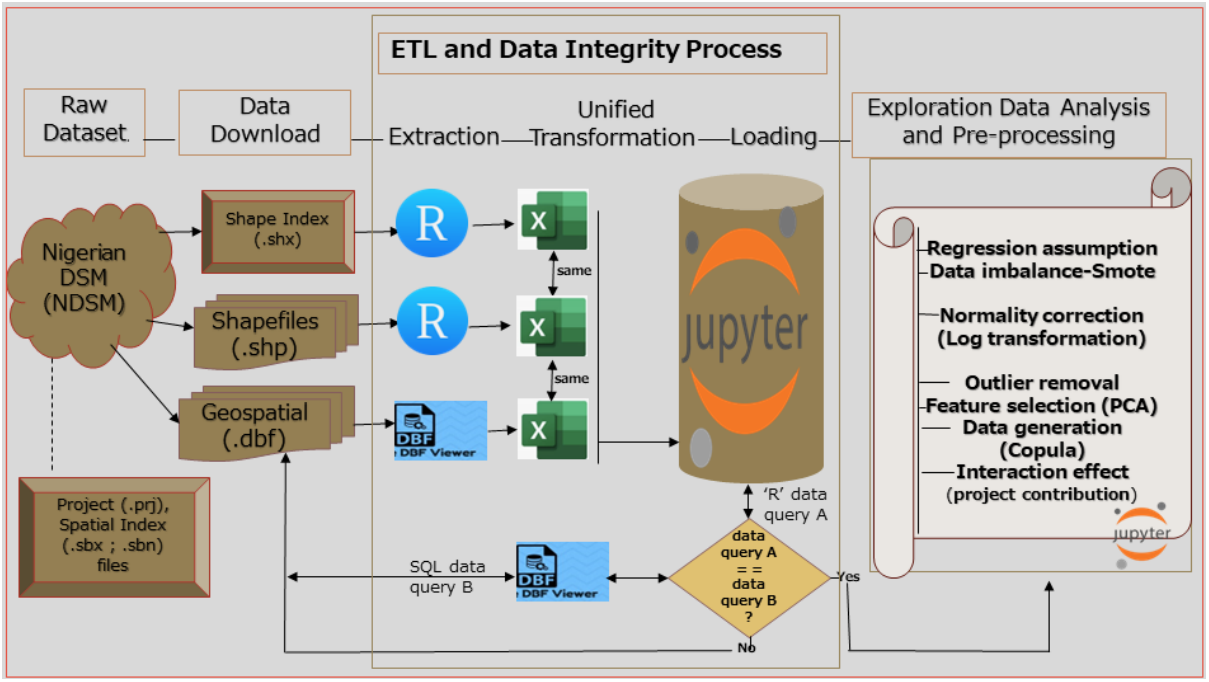


Figure 5: First three (3)methodology steps of soil degradation prediction and classification models

the shape loading. The main file (.shp) has the spatial vector data (polygons, lines or points) that give the features its geometry, index file (.shx) is the search engine for the shape index positions and dBase file (.dbf) is the database file for storing the shapefiles features and shape IDs. The compulsory files were extracted, transformed and loaded for this project.

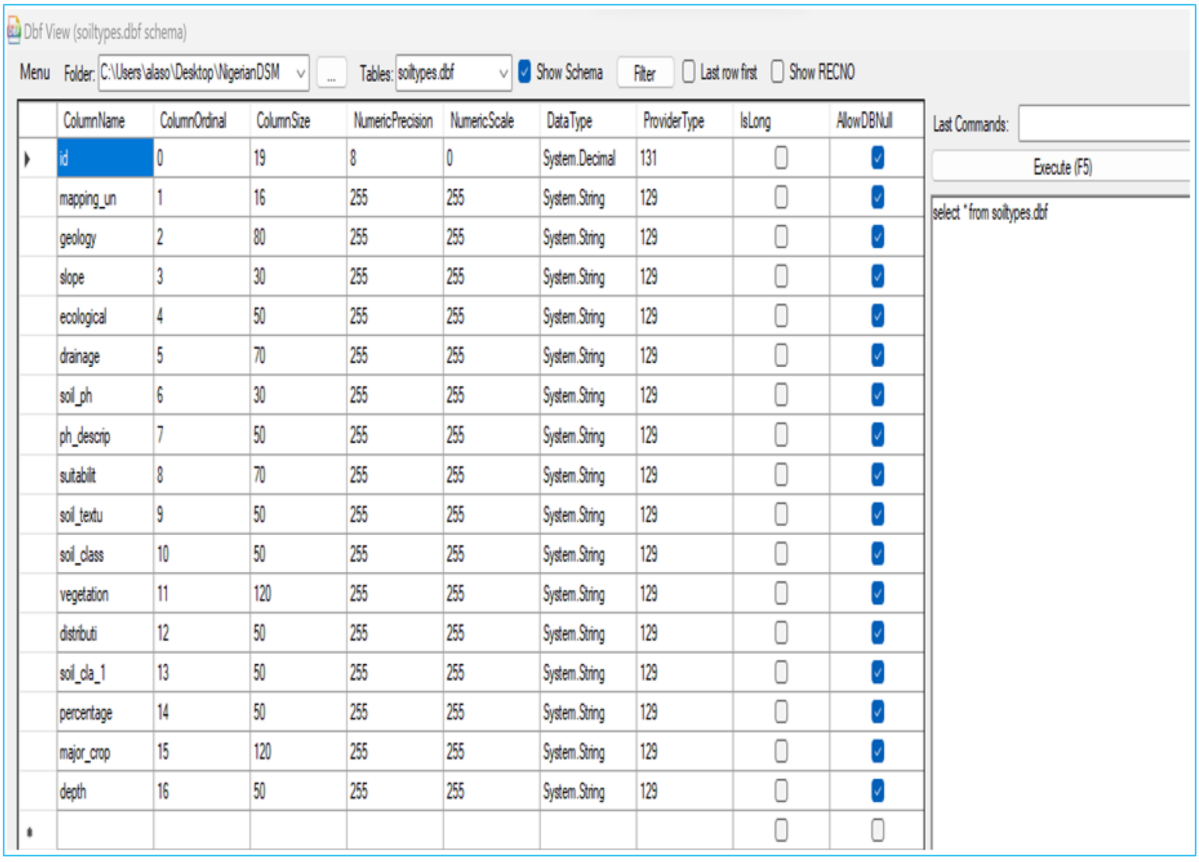
Extraction, transformation and loading: This is the second step that incorporates the data integration process of extracting data from a relational database (NDSM in this case), transforming the various data (here, the .shx, .shp and .dbf files) into a unified format ('.csv' in this project) and loading into a target system (Jupyter notebook for this project). The dataset file was unzipped to get the individual files separately for extraction.

Main shape file (.shp): Several 'R' functions from different 'R' packages, including shapefiles and raster, were used to extract the shapefile data and imported into Rstudio using the st-read function (soilShape and soilIndex coding in the artefact). The file showed 658 multipolygon rows as geometry attribute.This extracted geospatial file was transformed, with R algorithms, into both Microsoft comma separated value files (.csv)- to view the geocoding and microsoft excel worksheet file (.xl)- for headings, saved in a local desktop folder and loaded into Jupyter notebook.

Index file (.shx): The same extraction, transformation and loading for the .shp files was carried out for the .shx files giving same information which confirms data of both is same and .shx just used for searching the geospatial index.

Database file (.dbf): The database tool pack – DBFviewer was used to view the data in this file and as it is stored as a structured query language (SQL) database, SQL commands were used to extract the 658 rows of the soiltypes table using the query command, select*from soiltypes.dbf, from the schema. Figure 6 shows the dbf has 17

columns of which Id is the only decimal and others are strings while figure 7 gives a description of the attributes. This file was also transformed into a comma separated value file (.csv), exported and saved in the same local desktop folder as the .shp and .shx files then loaded into Jupyter notebooks. Both the .shp and .dbf files showed same data, so only .dbf file was used for the modeling.



ColumnName	ColumnOrdinal	ColumnSize	NumericPrecision	NumericScale	DataType	ProviderType	IsLong	AllowDBNull
id	0	19	8	0	System.Decimal	131	☐	☑
mapping_un	1	16	255	255	System.String	129	☐	☑
geology	2	80	255	255	System.String	129	☐	☑
slope	3	30	255	255	System.String	129	☐	☑
ecological	4	50	255	255	System.String	129	☐	☑
drainage	5	70	255	255	System.String	129	☐	☑
soil_ph	6	30	255	255	System.String	129	☐	☑
ph_descrip	7	50	255	255	System.String	129	☐	☑
suitabilt	8	70	255	255	System.String	129	☐	☑
soil_textu	9	50	255	255	System.String	129	☐	☑
soil_class	10	50	255	255	System.String	129	☐	☑
vegetation	11	120	255	255	System.String	129	☐	☑
distributi	12	50	255	255	System.String	129	☐	☑
soil_cla_1	13	50	255	255	System.String	129	☐	☑
percentage	14	50	255	255	System.String	129	☐	☑
major_crop	15	120	255	255	System.String	129	☐	☑
depth	16	50	255	255	System.String	129	☐	☑
							☐	☐

Figure 6: SQL- extracted attributes from DBFviewer schema

Data integrity check: Exploration of the loaded data showed all the 658 rows and 17 columns were uploaded and all features were of the same data type as in the database except for variable 'percentage' that was strings in database but double (decimal) in jupyter notebook. This is a minor difference that was corrected during pre-processing. Several SQL command queries were executed in Dbf Viewer, to subset columns and rows, for example, tested SQL query report B for soil pH range greater or equal to '6.0 -'6.0' with corresponding id, soil texture and major crops was the same 161 rows when compared with R-query A report in jupyter notebook with R codes :
$SELECTid, soil - pH, soil - textu, major - cropFROMsoiltypes.dbfWHEREsoil - pH >= 6.0 - 6.0$
and
$soilSubset1 < -subset(soilDbf)(as.character(soilDbf[SOIL - PH])) >= 6.0 - 6.0$
$soilSubset1[c(ID, SOIL - TEXTU, MAJOR - CROP)]$
This confirms accuracy of the data conversion and upload.
Objective 6 of chapter1, sub-section 1.1 has been achieved.

3.6 Exploratory Data Analysis and Data Pre-processing

3.6.1 Exploratory Data Analysis

S/No	Dataset Features	Description
1	Mapping unit	58 units from 1a – 24b.
2	Geology	The bedrock cover like migmatite and sandstone.
3	Slope	Different gradient in percentage range 0%- 55%.
4	Ecological zone	Rainforest, Savannah and Wetlands
5	Drainage	Five types: Imperfect, Poorly, Shallow, Moderate and highly drained.
6	Soil PH	Measurement range from 3.6– 9.1
7	PH description	Acidic, Basic, Neutral or variety of either Basic or Acidic.
8	Suitability to mechanised farming	Moderate, Fair, Marginal or Unsuitable
9	Soil texture	Eight types: Sandy, Sandy clay, Sandy loam, Loamy fine sand, Clay loam, Concretionary clay, Silty clay, Silty loam.
10 and 11	Soil class and soil_cla_1	These are United States department of Agriculture (USDA) or Food and agriculture organisation (FAO) classification categories
12	Vegetation	Several types like fallow, grass, forest, crops , shrubs and many more.
13	Distribution	Spread of the mapping units across Nigeria
14	Major Crops	Major crops grown in each mapped unit like
15	Depth	Four types: Deep / mostly deep, Moderate / generally deep, Shallow, Very deep
16	Soil class percentage	Ratio of the soil class in percentage

Figure 7: Dataset attributes description

The exploratory analysis of the dataset identified,Soil pH variable (SOIL-PH), the target is a continuous data, while variables'Id', 'Mapping-un'and'distribution' are numeric and discrete and other predictors are categorical (stored as characters).These attributes and description are shown in figures 6 and 7. Since'id' and 'mapping-un' are discrete, they are both irrelevant to the modeling, and so are'pH-description', 'soil-class' and 'percentage' since 'pH-description' described the pH values and soil-class is a replication of Soil-CLA-1. All these were removed during pre-process.

The explored bar charts showed 'ECOLOGICAL' variable had varied spacing width for the 'Rainforest' labels and while checks for data imbalance showed all the variables had data imbalance and there were no missing values (figure 8).

3.6.2 Data Pre-processing

The identified 'id', 'mapping-un', 'pH-description', 'soil-class' and 'percentage' were removed. The discovered varied spacing width for the 'Rainforest' labels were corrected by first converting the column to character then applied str-squish() and reconverted into factors as seen in figure 9 .This pre-process eliminated prediction or classification bias.

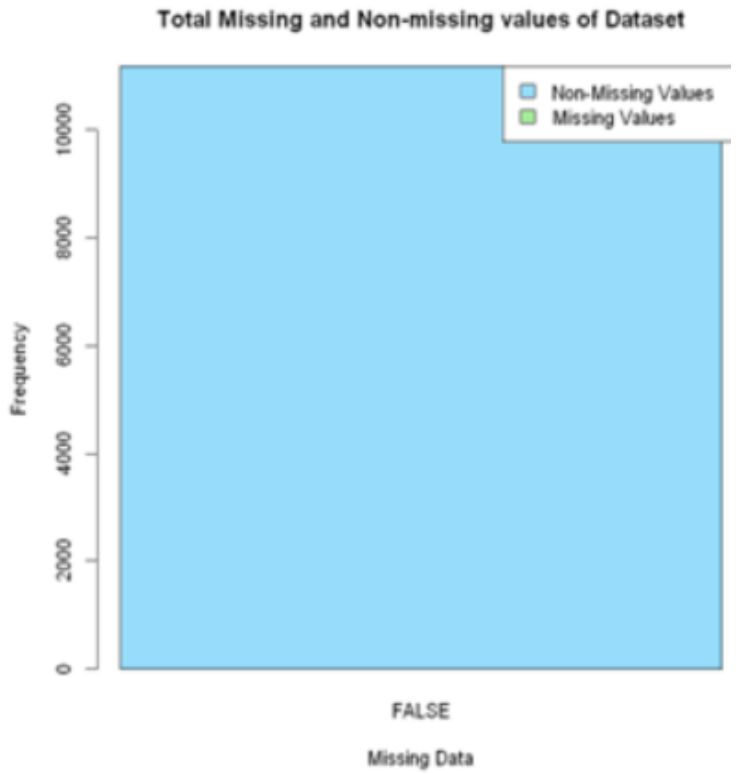


Figure 8: Dataset Non-missing values

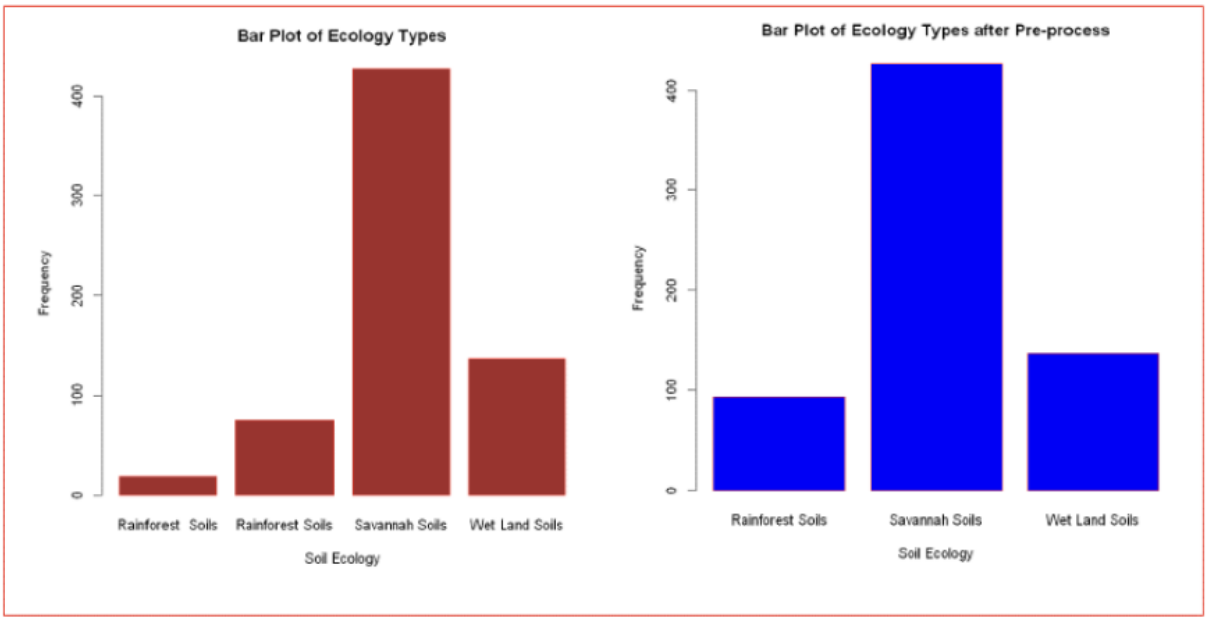


Figure 9: Pre-processed 'Ecological' variable- before and after

Regression Models Assumption Checks:
Although soil data is generally non-linear by nature, the dataset was tested for regression assumptions to ascertain and justify the choice of regression models used that do not require the assumption of linearity. In spite the reviewed studies had small sample

sizes that are akin to the data size of 658 except the classification models (table 4), this project generated synthetic data, using copula, to augment the sample size to 5658 to be similar to the benchmarked classification models and conducted in Jupyter notebook.
Independence: As the dataset was generated from the Nigerian legacy data that was derived from varied soil sampling areas over 5 years, the assumption of independence was assumed correct. However, statistically checking with chi square test, the p-value was lower than the set alpha of 0.5 at 2.2 exp-16, meaning there is insufficient evidence not to reject the null hypothesis that no existing relation between dependent variable, SOIL-PH, and other variables, so relationship exists between SOIL-PH and at least one of the independent variables. This violates the independence assumption.
Multicollinearity: This is to ascertain there is no correlation between the predictors to avoid multiple influence on predictions and if any, to select the best predictor representing the correlated variables. Here, xtabs() function was applied to the dataset, as they are categorical, and contingency table with all the predictor values as zero (0) was the result, meaning there is no multicollinearity between the predictors, therefore all can be used for modeling. The data was converted to numeric values to asses via correlation matrix and it showed the predictors are negligibly to marginally correlated, confirming result above, with highest being absolute 0.418 between Suitability and Slope while Ecological had highest absolute correlation of 0.189 with dependent soil PH, which is also negligible. Thus no correlation for both within predictors and between predictors and dependent variable. However, since correlation measures linearity, it is of no surprise that it is low for this non-linear data, thus the two predictors (Suitability and slope) were not removed yet for further tests. (correlation matrix is in the configuration manual).
Linearity: Since the remaining predictors are all categorical, linearity assumption is usually ignored, if other assumptions pass since categorical variables do not have measures in the Euclidean space, so will always be non-linear. As independence check failed, linearity relationship was checked with pair-wise test (configuration manual) and residual-fitted plots (figure 10) which all showed non-linearity since the residual plots line did not correctly capture the data distribution violating linearity assumption.
Normality check: Shapiro-Wilk test was carried out on the original data because the size was less than 5000 and many researchers adduced it is best normality test for such sample size. The result showed that the tested dependent variable, soil pH with ecological and texture predictors had p-values of 2.649 exp-13, 2.2 exp-16 and 2.2 exp-16 respectively, lower than alpha value set at 0.05 which was rechecked with Pearson chi-square test (since it involves categorical variables and takes skewness into consideration) with the three having 2.2 exp-16 each. This non-normality was visually confirmed with residual- plots (figure 10) and histogram (figure 11) and thus the conclusion that there is sufficient evidence the data sample is not from a population of normal distribution.

Outlier detection: Respective boxplots(figure 12) shows 'Ecological' and 'Soil Texture' have outliers presence like the qqplots in and cooks distance of residuals vs leverage plots in figure 10.

3.6.3 Treatment of Outliers, Normality and Linearity

To transform the data to conform to the linearity and normality conditions of regression models, the outliers seen in both the residual plots and box plots were statistically confirmed and removed before series of transformation algorithms were applied as discussed

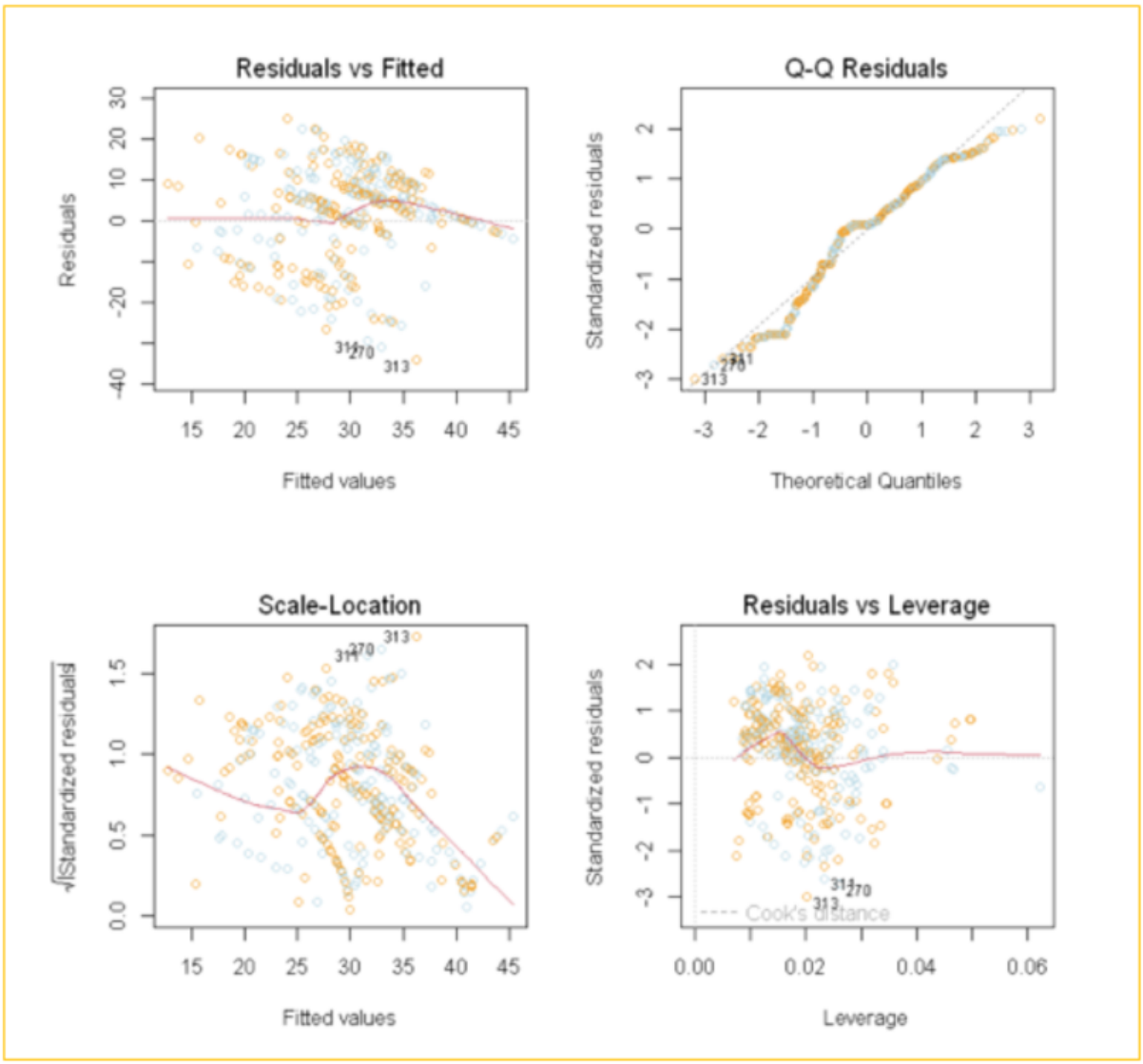


Figure 10: Soil degradation Residual plot

below.

Outliers Detection and Removal: Three sets of functions were applied to the dataset as a whole; the first set was a function to transform the integer columns to numeric to enable the other two functions work on the data. The second is an outlier detection function based on the formula- ‘ inter-quantile range *1.5 ‘ where quantile 1 and quantile 3 were set as 0.25 and 0.75 respectively with inter-quantile range being the difference between the two which was used to set the upper and lower quantile limits. This function detected outliers as those data falling outside the quantile upper and lower limits based on the quantile formula above. The third algorithm is an outlier removal function in the form of a ’for-loop’ that called the outlier function above, and created another dataset of ’no-outliers’ as it looped through each column and row of the numeric dataset which effectively removed 112 rows of outliers. The loop-created new data was algorithmically written to a .csv file that was saved on local desktop, loaded into Jupyter notebook and rechecked for normality which was still not complied with. It means outliers were not the root-cause of the non-normality.

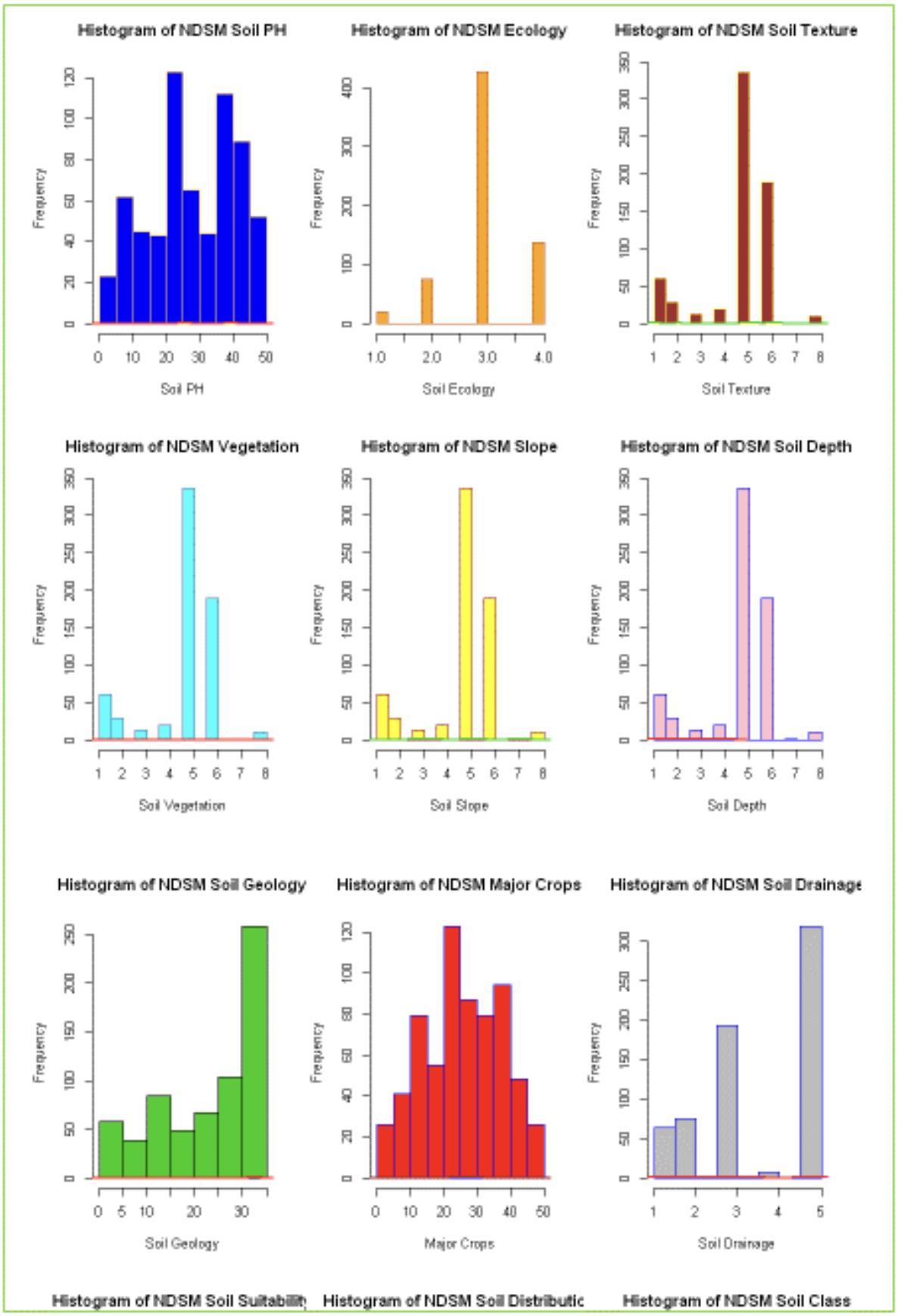


Figure 11: Soil degradation histograms

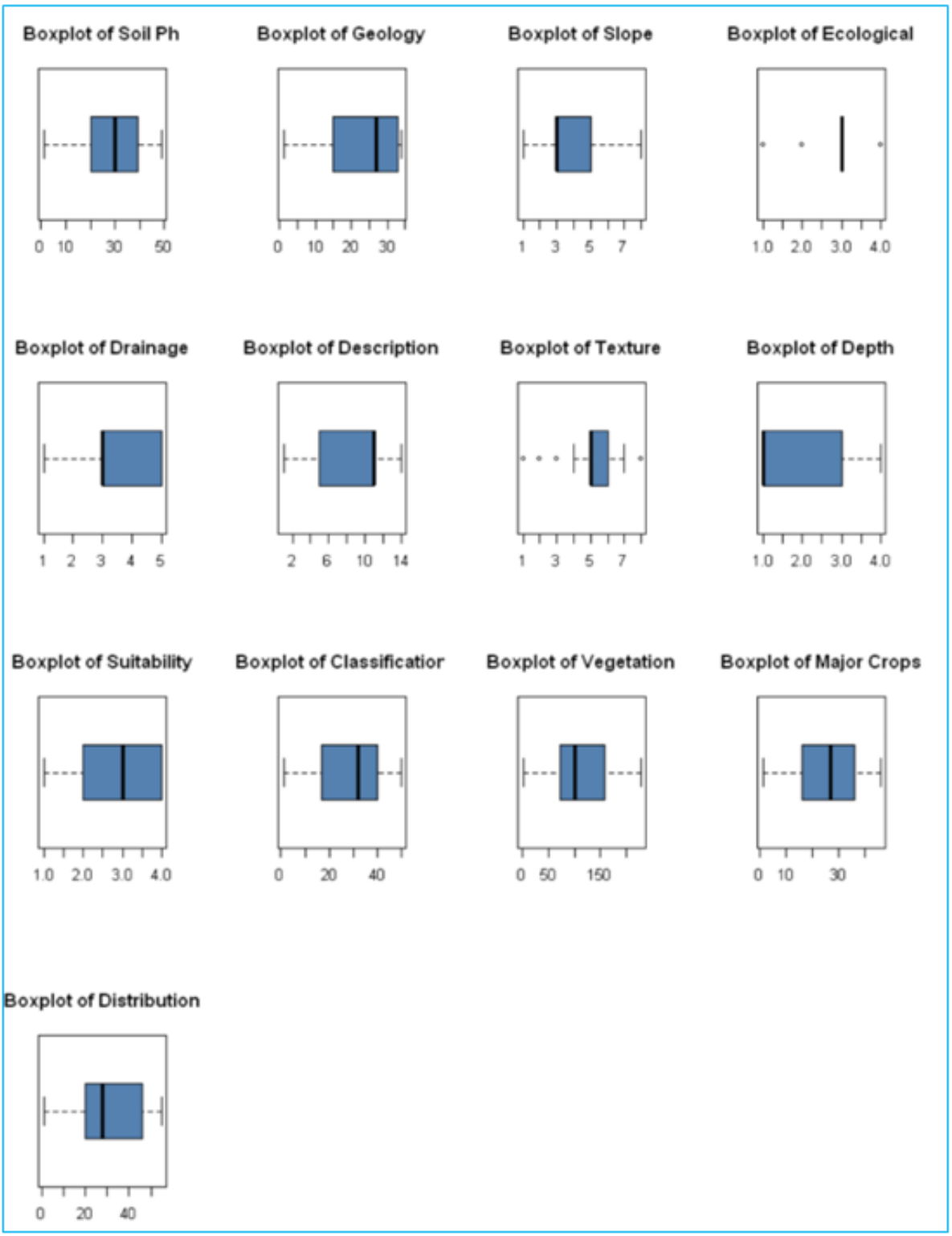


Figure 12: Soil degradation box-plots

Normality and Linearity correction: To transform the skewed variables and correct the evidenced non-linearity, several log transformations were applied to the 'no-outlier' dataset starting with Box-cox transformation that is believed to transform most distribution types, unchanged the predictors' distribution and linearity and worsened soil-ph and major-crop, both formerly of near-normal distribution. Similarly to this outcome are those of log base 10, square root, cube root and mix of squared and cubed interactive variable transformations (figure 13). Since there was no success in the normality and non-

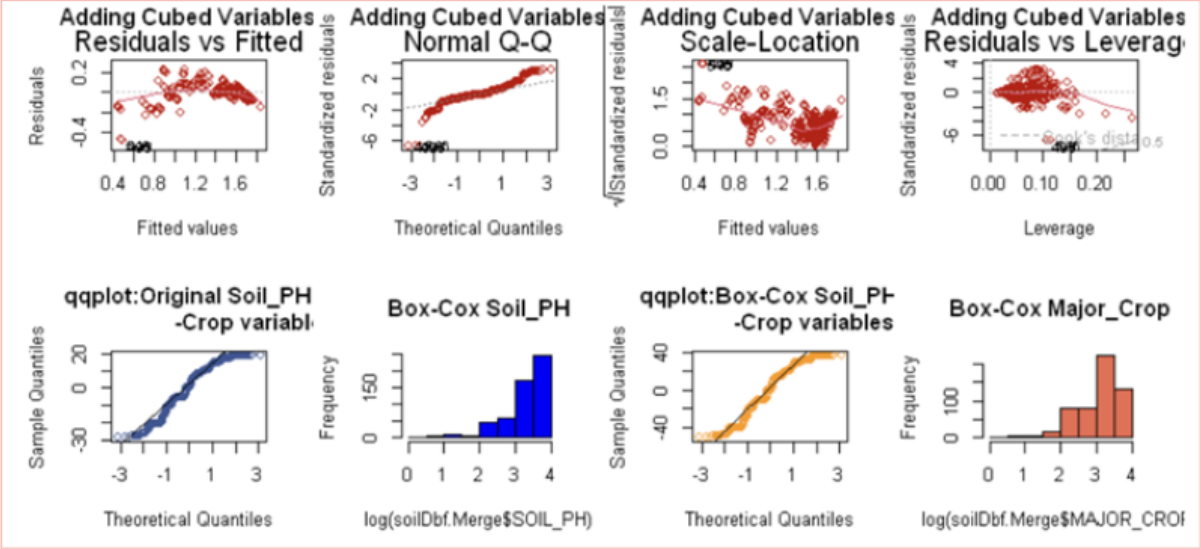


Figure 13: Box-Cox and cubed transformation plots

linearity transformations attempts, it is safer to claim Nigerian digital Soil Map dataset is of non-linear and non-normal nature fitting the non-parametric models developed in chapter 4.

3.6.4 Data Generation, Features Selection and Data Imbalance

For non-parametric models, large datasets are required to ensure smoothening and although project's data size is about those of the reviewed literature, this project synthesised more data to fulfil the smoothening requirement.

Data generation was achieved with the use of copula algorithm to generate five thousand (5000) more data giving a total of 5,658 observations. This algorithm was chosen because it synthesises new data using correlation coefficients to emulate original dataset's non-linear dependences of the variables' different distribution, correlation and data types. The original dataset was pre-processed with VineCopula's pobs() and BiCopSelect() which gave bivariate copula family that suited each variable based on Bayesian information criterion (BIC) and Akaike information criterion (AIC) and these family numbers (0,1) were then used to generate the 5,000 new data with each par1 and par2 values then merged with the original data of 568 rows using rbind(), however, the original data categorical predictors were earlier transformed to estimated numeric values based on impact of each factor level on the dependent variable using SQL commands in tidymodels package. Video explains further.

Features selection: Principal component analysis (PCA): Ascertaining variables to discard, PCA was applied to know those features having major contribution to the dataset

through scree plot, Eigen values, and significance of PCAs and result showed ecological, suitability, soil class and soil texture (marginally) are insignificant. However, superior regress of the predictors' PCA on dependant, showed all 12 PCs are statistically significant. Figure 14 shows the scree plot showing only PC1 is relevant, having 95 percent of the variance.

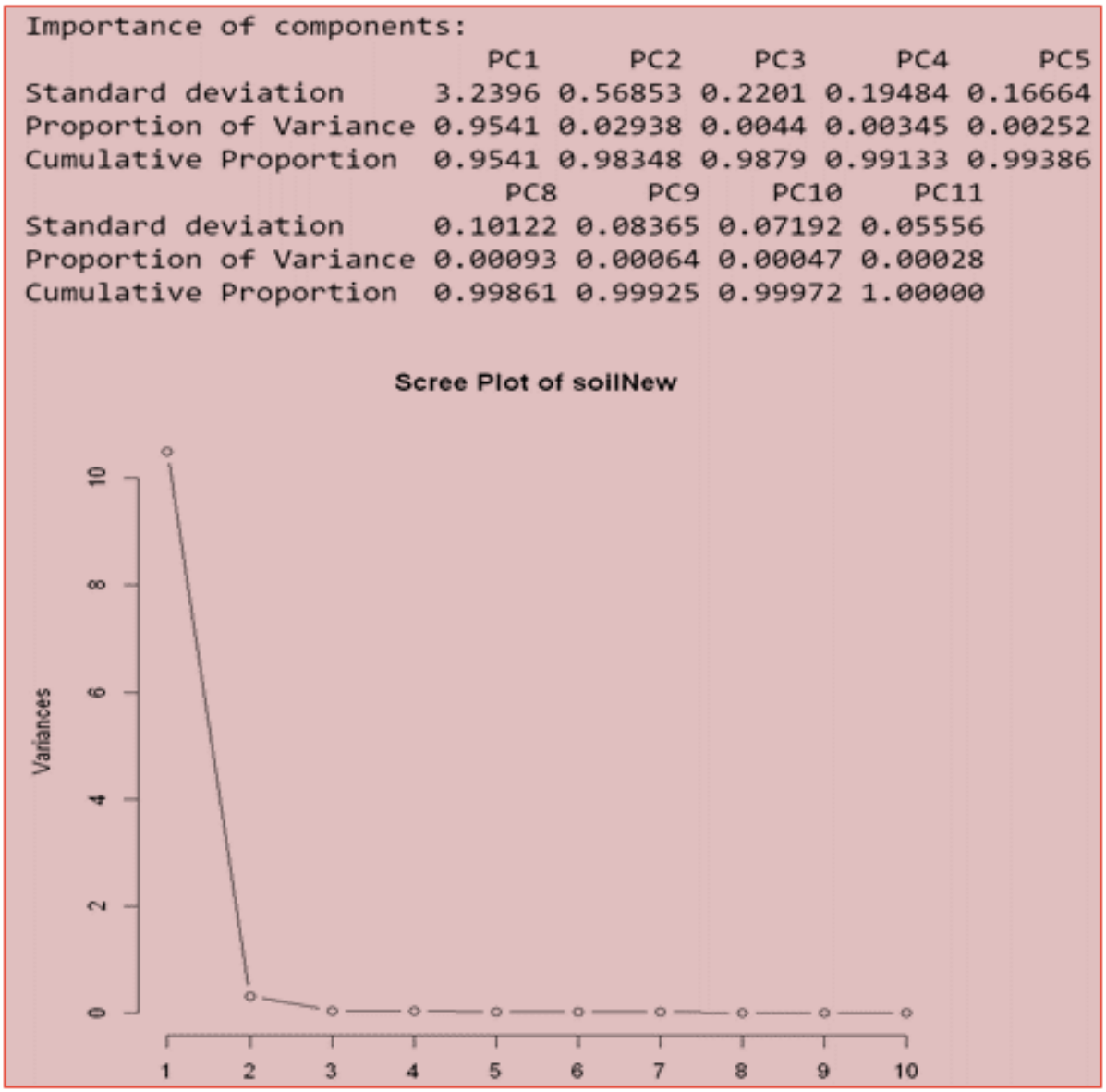


Figure 14: Soil degradation Screeplot

Outliers recheck and removal: Outliers were rechecked for and a total of 653 observations were removed leaving an NDSM dataset of 5,005 observations and 12 features.
Data imbalance: All the variables have data imbalance of various dimensions, especially the target- 'Soil-PH'. To mitigate prediction and classification bias, these were treated using SMOTE() function from smotefamily package in R on randomly subsetted 30 percent that yielded 2,997 observations and 13 features data used throughout the project. This is to avoid clog caused by the high volume factors and several split ratios were tried between the training and test data for the 70:30 to be the best fit for the SMOTE

algorithm.

3.7 Conclusion

The soil degradation methodology for this project (figure 3) captured all the scientific steps which is channelled to ascertain the two research questions were addressed with which the objectives (tables 1-3) were tailored towards. The design flow process (figure 4) detailed the implementation step and highlighted stages of achievement of the objectives. A combination of different tools were used- DBF Viewer (A Structured Query Language (SQL), R 4.3.1, Jupyter notebook (IR-Kennel), Tidymodels and Excel- to achieve the download, data exploration analysis and data pre-processes. The modeling, evaluation and results, model comparisons and variable interaction are discussed in the next chapter 4.

Objectives 1 -7 in chapter 1, sub-set 1.1 have been achieved.

4 Implementation, Evaluation and Results of Soil Degradation Prediction and Classification Models

4.1 Introduction

As mentioned in chapter 3, this chapter discusses the implementation step where the prediction and classification models were developed and respective analytics carried out. It also describes the metrics used to evaluate the models' performances with the results of the models.

Implementations:

For this project, eight (8) models were developed- two prediction machine learning (Random forest, for regression and Support vector machine for regression) and six machine learning (2 x K-nearest neighbour models ,Support vector machine, 2 x Naïve Bayes and random forest), for classification. The chapter is split such that the implementation, evaluation and result of a model are discussed for each model separately.

Evaluation Metrics Description:

For each data point of a dataset, a prediction is made by a model and so to check how accurate the model is, the prediction error of the model needs to be calculated. From the various reviewed literature the common evaluation metrics for prediction models are MAE, MSE, RMSE, and R2 while those for classification models are accuracy, precision, recall and F1 score. These metrics are explained below.

Mean absolute error (MAE): This is the actual difference between the recorded data value (y-i) and the predicted data value (x-i) and MAE is arrived at when computed for each data point and all the errors summed up and the mean of this sum is calculated, expressed as:

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \tag{1}$$

where D is data sample size

Mean square error (MSE): This measures how good a model fits the data. The smaller the MSE value, the closer the regression line is to the data meaning the lower the

prediction error and calculated by squaring the MAE, as shown:

$$MSE = \sum_{i=1}^{D}(x_i - y_i)^2 \tag{2}$$

Relative mean square error (RMSE): This is the most common metric used in ML and the lower its computed value, the better. It is the Square root of MSE, giving the prediction value deviation.
Accuracy: This measures how 'correct on average' a measurement of central tendency is but not the absolute dispersion from actual data. The higher it is, the better as it shows low error. As this project is a multiclass classification, accuracy in the confusion matrix is the fraction of all classifications in percentage:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Accuracy = Correct classifications / Total classifications

Where TP = True positive; FP = False positives; TN = True negative ; FN= False Negatives
Precision: This measures how close several measurement of same data point is making it a good measure of model reproducibility, thus high precision is better as it means low error in reproducing the model even though it also does not tell how close the measure is to actual data. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall: This is the sensitivity result in confusion matrix and it is the measure of how the model can correctly identify and retrieve the targeted data from a dataset and is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F1-Score: Technically, this is the harmonic mean of precision and recall which simply means how accurate a model classifies a dataset. It is calculated as:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{6}$$

4.2 Implementation, Evaluation and Results of Soil Degradation Prediction Models

4.2.1 Implementation, Evaluation and Results of Support Vector Machine for Regression Model

Implementation:

Support vector machine for regression, SVR, is a non-linear machine learning technique that uses kernels to convert the data to as near best fit to linearity as possible except it uses values to predict unlike classification

by SVM. It does not depend on data distribution and is used when the target variable is numerical with its performance usually improved by fine-tuning the cost (to avoid over-fitting) and value of epsilon (error) through a grid search which trains the data via large number of models for selection of the best.

For this project, SVR modeling was performed in Rstudio with packages e17071, caret, mass, kernlab, ggplot2, and lattice.
Two models were initially developed, one using the Polynomial kernel (svmPoly) and the second using the radial kernel (svmRadial), to ascertain the kernel to use for the modeling based on lower RMSE. The Polynomial kernel (svmPoly) had a lower RMSE of 0.01 to radial of 0.50 and was selected for the modeling, although it would have been selected if otherwise because of the advantage of handling variable interactions which is one of the contributions of this project.

The pre-processed and balanced data was loaded into Rstudio with 'read.csv'. The dataset Index was split with the target column at 70:20:10 ratio of training:testing:validating respectively to get the associated data samples used for the model.
A set of hyper-parameters were created from 10-fold cross validation (the control value to mitigate model over-fitting) and tuning parameters (degree, scale and center),from tunegrid function, to optimise the model's performance. The selected best parameters from different dataset subsets, improved prediction errors of both test and validation data.

Evaluation and Results:

In line with the benchmarked work of (Chen et al. 2019), this project's SVR model was evaluated with RMSE, MAE and R2 as seen in figure 15. The polynomial kernel was selected with the lower RMSE result (0.01) over radial kernel RMSE of 0.50 and the advantage of feature interaction capability, if any. Prediction results of both testing and validating data are very similarly good with near-zero prediction errors,(test-RMSE=0.011, validate-RMSE=0.010) with similar pattern in MAE results of both. Coefficient of determination,R2 of almost 1 means the model is a perfect fit for the dataset and the independent variables statistically perfectly predicted the dependent variable, soil pH.

There is the issue of over-fitting even though the control parameter was to mitigate against this.

For feature interaction:the fact that the 10-fold cross validation resulted in degree 1 for a non-linear data (yellow highlight in figure 15) indicates a linear model meaning there is likely no features interaction and effect on

this model or could be a way cross validation is minimising the error or over or under fitting too. For this model, the choice of degree 1 shows no feature interaction occurred.

The objective 8.1 (A) in chapter 1,sub-section 1.1 has been achieved.

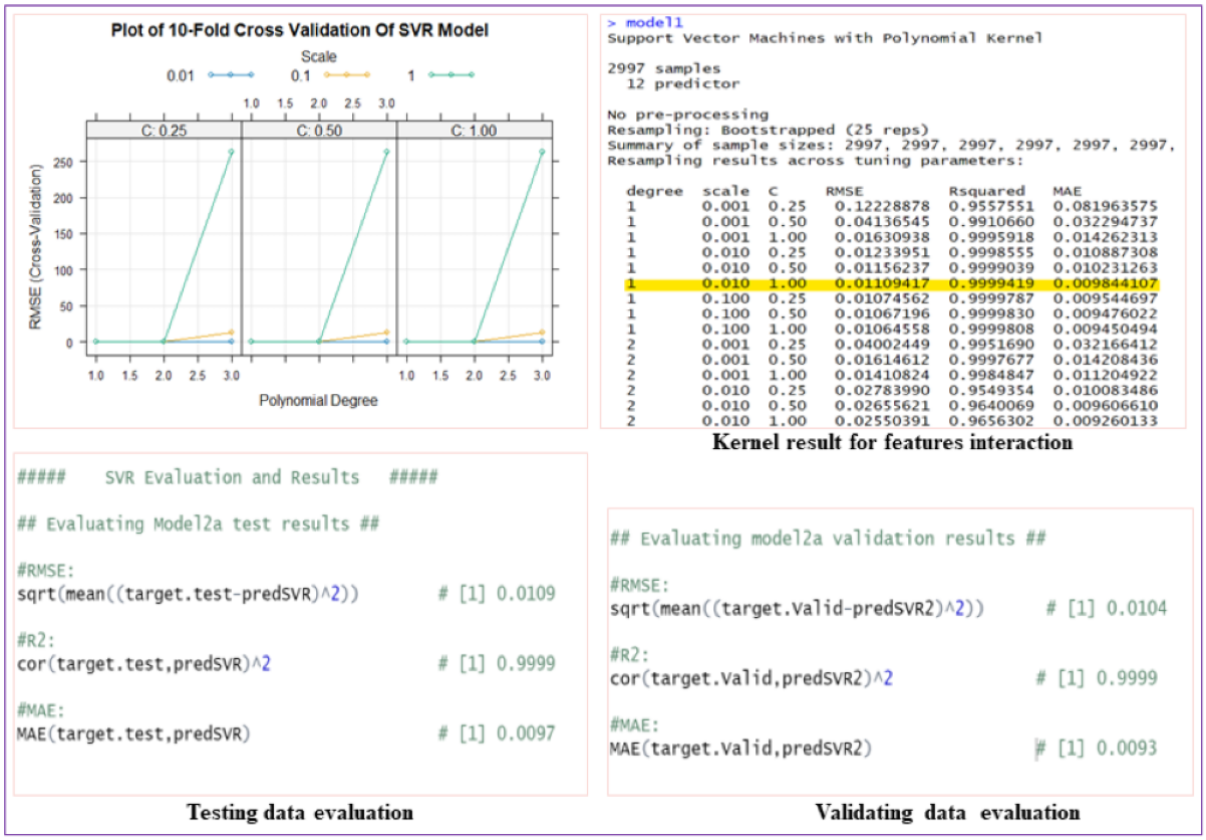


Figure 15: Soil degradation support vector for regression model results

4.2.2 Implementation, Evaluation and Results of Random Forest for Regression Model

Implementation:

Random forest is a suit of decision trees that take random training data samples, selecting random predictors as random initial variables to build a model of combined multiple decision trees that gives more accurate prediction decisions. It is a non-linear prediction and classification algorithm whose accuracy is evaluated by the out-of-bag error (OOBE) percentage.

This project applied 'caTools', 'caret' and 'randomForest' packages in Rstudio to the loaded previously pre-processed and balanced data. The random forest for regression modeling process was split into two- the first part involved factorising the target, SOIL-PH, in to binary levels (ph

values greater or equal to 0.5 = 2 and lower than 0.5 =1) to allow for binary classification of the data which was necessary to select the optimal hyper-parameters (number of tree nodes (ntrees), tuning (tuneGrid, mtry), times to retrain the model(trControl) and 10-fold cross-validation (for resampling) that were all modeled under the 'RMSE' metric. These parameters were used for the RF regression (2nd part) to fine-tune the final model with split data of train (0.7), test (0.2) and validate (0.1) and SOIL-PH not factorised this time. The trained model was used to predict the test and validate data that both resulted in high evaluation metrics (RMSE,MAE,MSE and R2).

Evaluation and Results

The evaluated error metrics of the Random Forest for regression model all gave approximately zero values meaning the prediction accuracy is very high (accuracy not evaluated here as Rf was used for regression) and the error rate by tree plot backed this up with the tuning cost of 0.5 being that with lowest RMSE. R2 at 0.99 shows the predictors are statistically

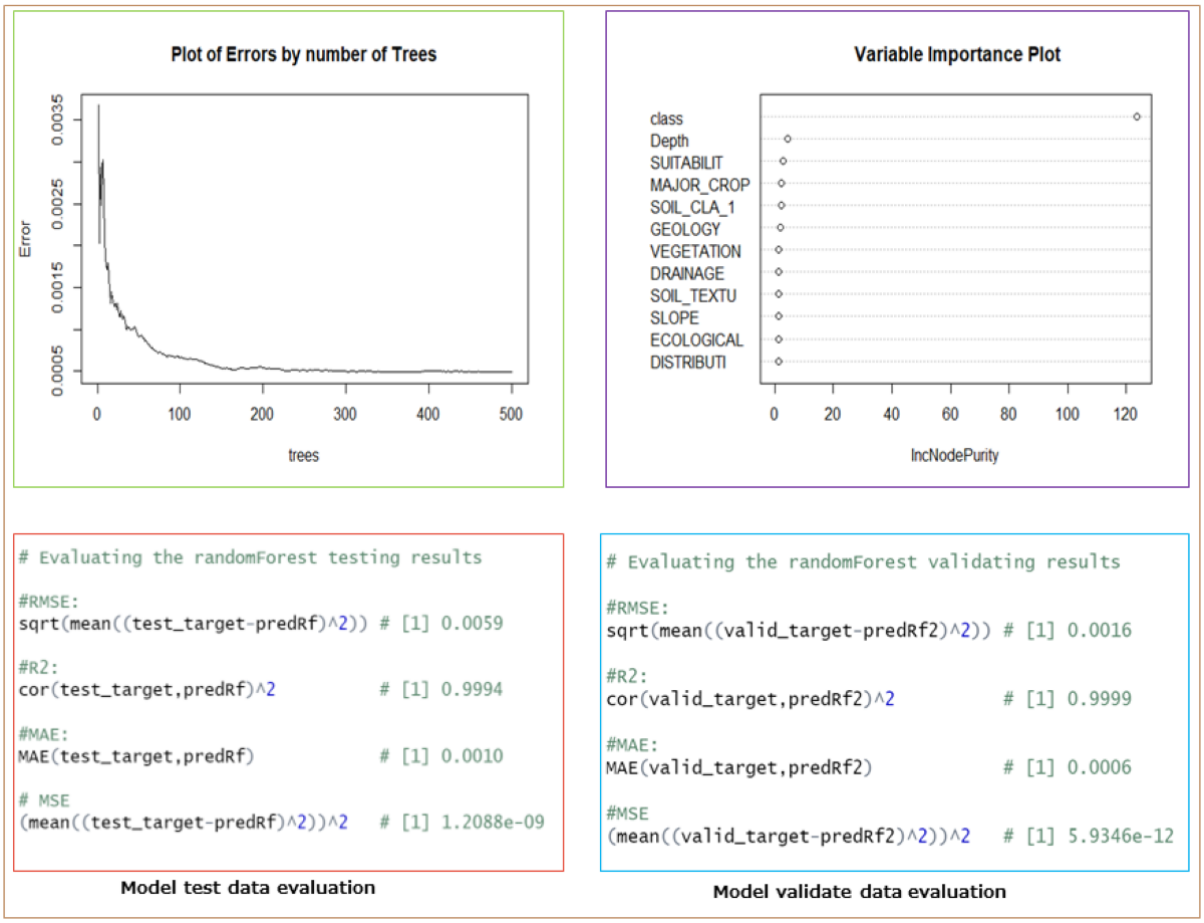


Figure 16: Soil degradation random forest for regression model result

significant in predicting soil ph at a level seen from the variable importance plot where soil depth, suitability and major crops of the area are the topmost three, (in descending order) and ecology is the least relevant to soil ph (figure 16)

The objective 8.1(B) in chapter 1, sub-section 1.1 has been achieved.
Objective 9.1 in chapter1, sub-section 1.1 has also been achieved.

4.3 Implementation, Evaluation and Results of Classification Models

4.3.1 Implementation, Evaluation and Results of K-Nearest Neighbor Models

Implementation:

K-Nearest Neighbour (KNN) is a supervised non-parametric ML technique that uses the ideology of grouping together data that are nearest to each other to predict and classify a dataset. It is referred as non-parametric because it does not rely on the assumption of any distribution and as such, can be used on a wide variety of dataset type with the value of the 'k' determining the number of data points to be considered nearest to each other based on preset distance metric which could be Euclidean or Minkowski or Manhattan.

For this project, KNN was effected in Rstudio with R-packages 'e1071', 'caTools' and 'class. The loaded pre-processed and balanced data was subsetted to remove the target column and normalized. This normalized data was split into training-0.8 and testing data -0.2 (Train.knn and Test.knn) while the target variable (SOIL-PH) was first transformed into factors of binary levels, with ph value of 0.5 being the partition, then used as the classifier labels after being split like the predictors.

The rule of thumb that value of k should be square-root of the training rows, prompted the building of two sub-models of k=48 and k=49, comparison of which led to random selection of k-value=48 used for modeling with knn() since they both had same accuracy of 0.99and same mis-classified total of 4. To avoid a clog on the system, a function was created to calculate the evaluation metrics of accuracy, precision, recall and f1-score based on the statistical formulae.

Two KNN models were developed, one to classify the soil ph to compare with project's models and the second to classify soil textures to compare the project's result with existing benchmarked models. Both models gave high evaluation metrics with the soil texture model,having soil

texture variable grouped into four (ClayLoam, Loamyfinesand, Sandyclay and Siltyclay) to be as near the benchmarked model's classified groups metrics while soil PH model has soil PH variable factored into binary level.

Evaluation and Results:

Results seen in figures 17 and 18, show that the classified soil pH has high accuracy, precision, recall and F1 score and the confusion matrix confirms the soilPh model is a good fit as only 4 points were mis-classified.

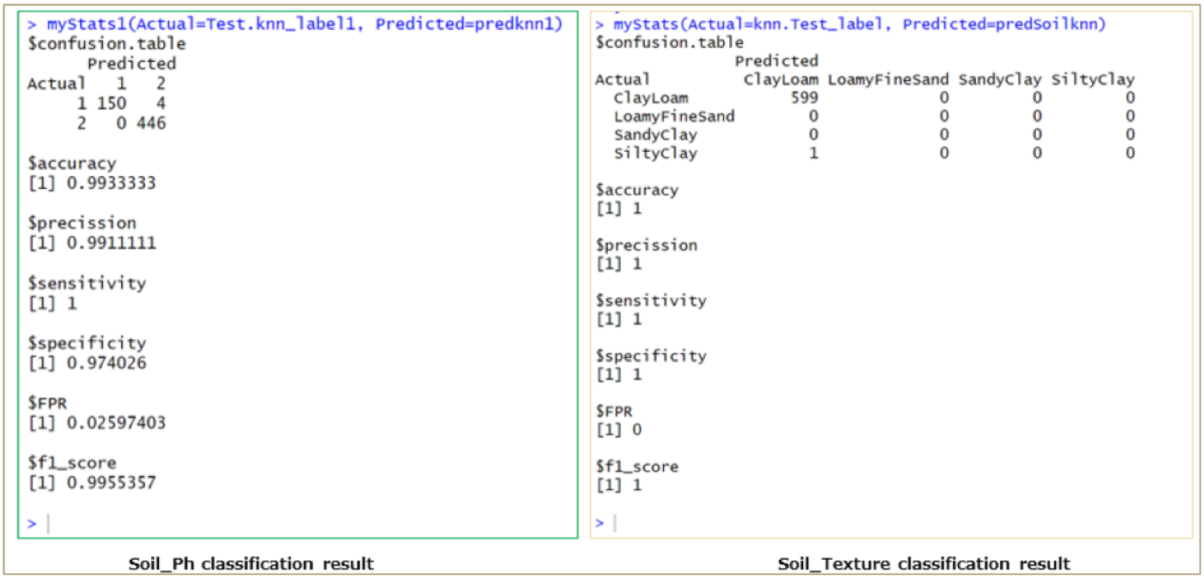


Figure 17: Soil degradation K-nearest neighbour models results

The objective 8.2(C) in chapter 1, sub-section 1.1 has been achieved.

4.3.2 Implementation, Evaluation and Results of Support Vector Machine Model

Implementation:

Support vector machine (SVM) is also a supervised ML method that is used for both classification (SVM) and regression (SVR) ideas but more for classification problems. The principle of SVMs are based on discovering best hyperplane that best separate data to classes, for example, two linear data classes- straight line and hyperplanes using kernel functions, for non-linear data, and the further away the data points are from the hyperplanes, the clearer the classification. Though it gives high accuracy, works better

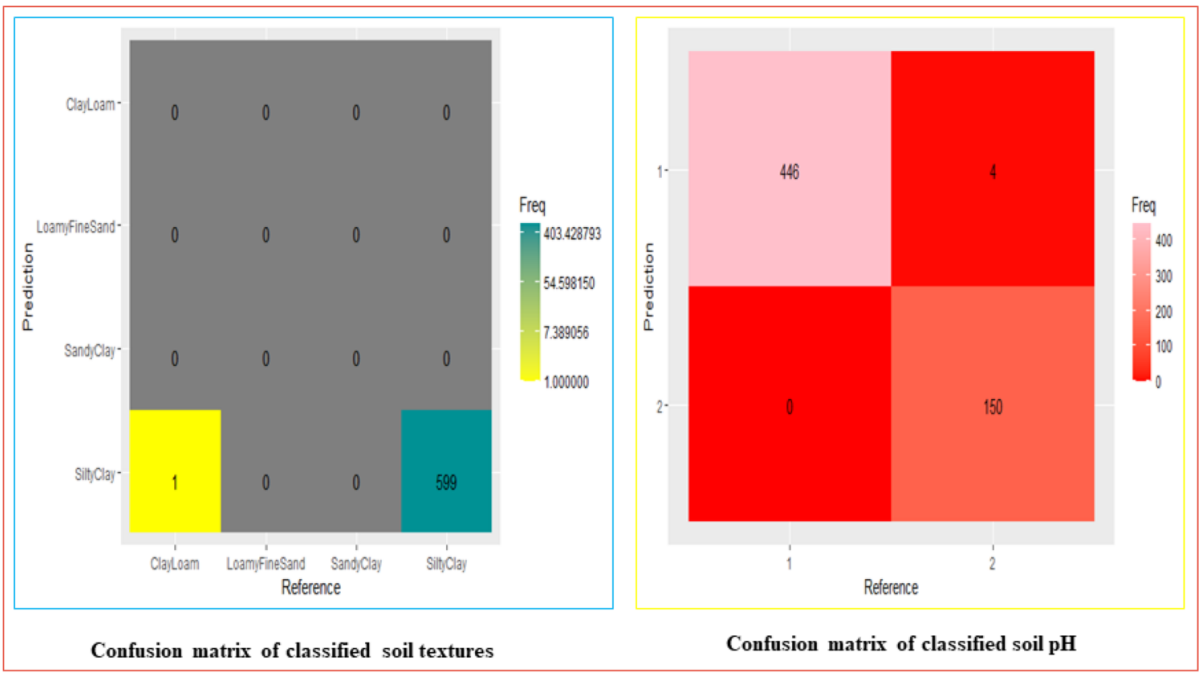


Figure 18: Soil degradation K-nearest neighbour models confusion matrices

with small size dataset and use of training subsets makes it efficient, it is less suitable for noisy datasets and kernel choice will determine the classification effectiveness of the non-linear data. Equations (7) and (8) are respectively for linear SVM and RBF kernel non-linearity function

$$f(x) = w^T x + b \tag{7}$$

$$k(w, x) = e(-\frac{||x_i - x_j||^n}{2(\sigma^2)}) \tag{8}$$

Where wT is the weight vector, x, the data and b, the coefficient from training data, n is sample size and sigma the kernel variance. SVM algorithms suit this project for the above pros and cons with original size. For this project SVM was carried out in Rstudio with package e17071,caret and ggplot2. The pre-processed balanced data was normalised, split into train and test data (80:20), then model fitted and predicted using ksvm() to enable choices of kernels. Algorithm chosen kernel, radial basis function (RBF) (for data non-linearity), and polynomial kernels (for both non-linearity and feature interactions were tested resulting in **evaluation of three (3) SVM models** with the polynomial model chosen due to the project's contribution, even though it is not the optimal model.

Evaluation and Results:

Results of the Support Vector Machine (SVM) model is presented in figure 19. The polynomial kernel was chosen because of the feature inter-

action advantage out of the three models although the balanced accuracy values are almost same.
This model has near to 1 values for the evaluation metrics with very low mis-classified points 6 in total which is backed with the low false positive rate of 0.03 testifying to the classification accuracy and high recall (sensitivity of 0.997, making this model a good fit to the dataset

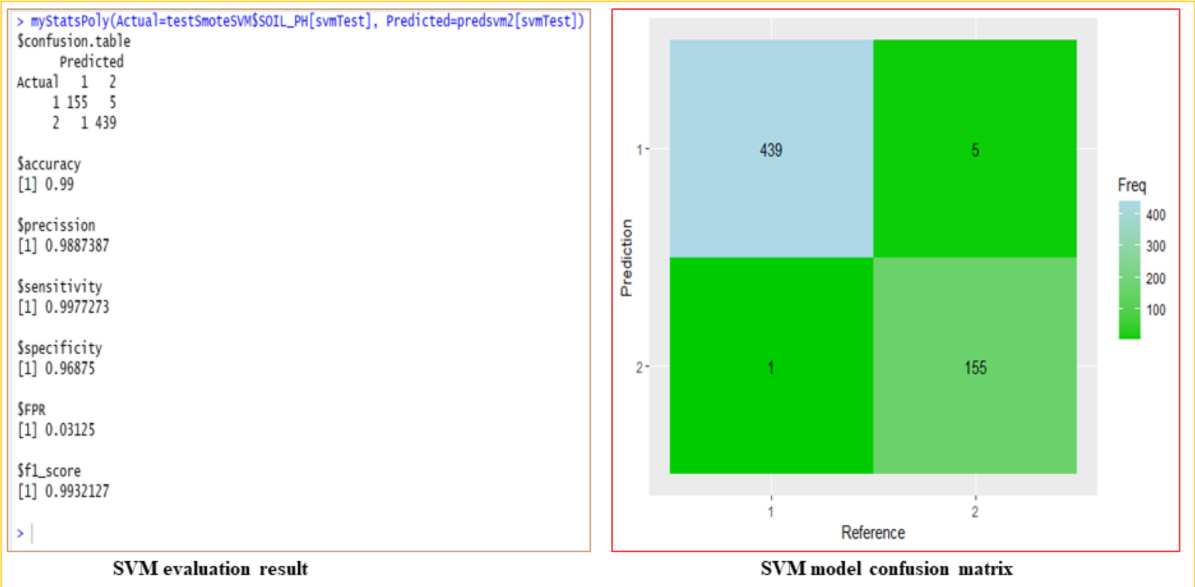


Figure 19: Soil degradation support vector machine model results and confusion matrix

The objective 8.2(D) in chapter 1, sub-section 1.1 has been achieved.

4.3.3 Implementation, Evaluation and Results of Non-Parametric Naive Bayes Models

Implementation:

This is another supervised kernel classifier that relies on the probability and Bayes theorem with a strong naïve assumption of independence amongst the variables. As a kernel-based algorithm, it affords the option to be used as classifier for different distributions, making it suitable for non-linear datasets. Naïve Bayes gives the conditional probability of an event A based on the fact that occurrence of another event B has happened and mathematically represented in equation (9) as:

$$P(B|A)P(A)P(A|B) = P(B) \tag{9}$$

Where: P(A—B) = conditional probability of A given B occurred, P(B—A) = conditional probability of B given A occurred, P(A) = probability of event A occurring, P(B) = probability of event B occurring.

For this project, Naïve Bayes' kernel density estimate (KDE) was the chosen kernel with usekernel set to 'True' and Poisson set to 'False'. The KDE ensures non-parametric measures are used to get the probability estimates while Bernoulli equals 'True' ensured the factored target is treated as discrete. The pre-processed, balanced data was loaded in to Rstudio and with e1071, caTools, ggplot2 and caret packages was normalised and split into train (0.8) and test (0.2) data with the target variable factorised and transformed to binary level with 0.5 being the partitioning point. **Two models were built**- one for classifying Soil ph (to compare with other models of the project) while the second classified the soil textures (to compare with the benchmarked work of (Padmapriya. & Sasilatha 2023). The two models performed highly as detailed in evaluation section below.

Features interaction effect could not be tested for these models as Naive Bayes algorithm could not handle interaction terms.

Evaluation and Results:

Evaluated result of the two models are seen in figures 20 and 21 and both appear to be over-fitted with accuracy being 0.98 (for the model classifying soil ph) and 1 (model classifying the soil textures grouped to be similar to the benchmarked model) like the other metrics. The confusion table of the soil pH model is very good at few mis-classifications unlike the soil texture model that had for only two of the soil texture groups, albeit low and nil value for others. This could be as a result of the probabilistic numeric conversion of the categorical variable, by tidymodels package, which gave near zero values as such it is possible those soil texture types were accurately predicted too.

The objective 8.2(E) in chapter 1, sub-section 1.1 has been achieved.

4.3.4 Implementation, Evaluation and Results of Random Forest for Classification Model

Implementation:

Random forest is a suit of decision trees that takes random training data samples, selecting random predictors as random initial variables to build a model of combined multiple decision trees that gives more accurate prediction decisions. It is a non-linear prediction and classification algorithm whose accuracy is evaluated by the number of trees per error.

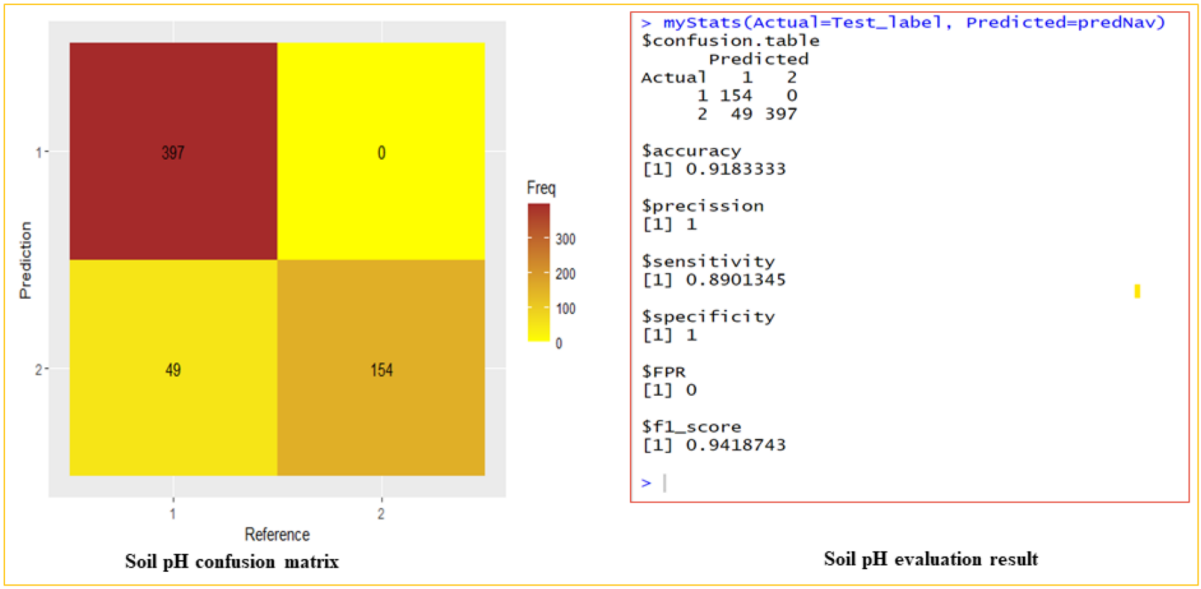


Figure 20: Soil degradation Naive Bayes Soil PH model results and confusion matrix

This project applied 'caTools'and 'randomForest' packages in Rstudio to the loaded previously pre-processed and balanced data. The start of the random forest for classification modeling was with the split of the data in to train (0.8) and test (0.2) and target, SOIL-PH factorised and converted into binary levels at which 0.5 was the partition.
The randomForest algorithm was passed to the train set to create the model that was used to classify the test data. Evaluated result and confusion matrix are discussed in result evaluation.

Evaluation and Result:

Result of the classifier model is seen in figure 22 below which showed perfect 1 for all the evaluation metrics as evidenced with the confusion matrix which showed zero mis-classified class. This is buttressed by the number of trees error plot which confirmed the result as the error rate stabilised at almost 0.005 at about 50 number of trees.

The objective 8.2(F) in chapter 1, sub-section 1.1 has been achieved. Objective 9.2 in chapter 1 subsection 1.1 has also been achieved.

4.4 Implementation, Evaluation and Results of Checking for Features Interaction

4.4.1 Implementation

In the review of the production of the Nigerian digital soil map dataset, the authors declared features interaction was neither checked nor accoun-

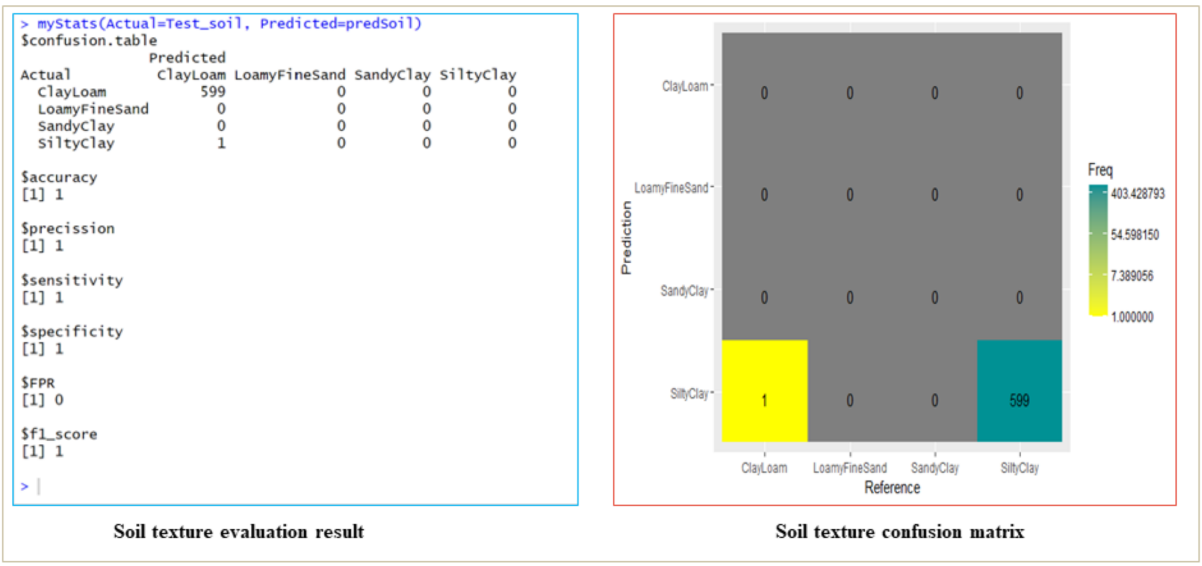


Figure 21: Soil degradation Naive Bayes soil texture model results and confusion matrix

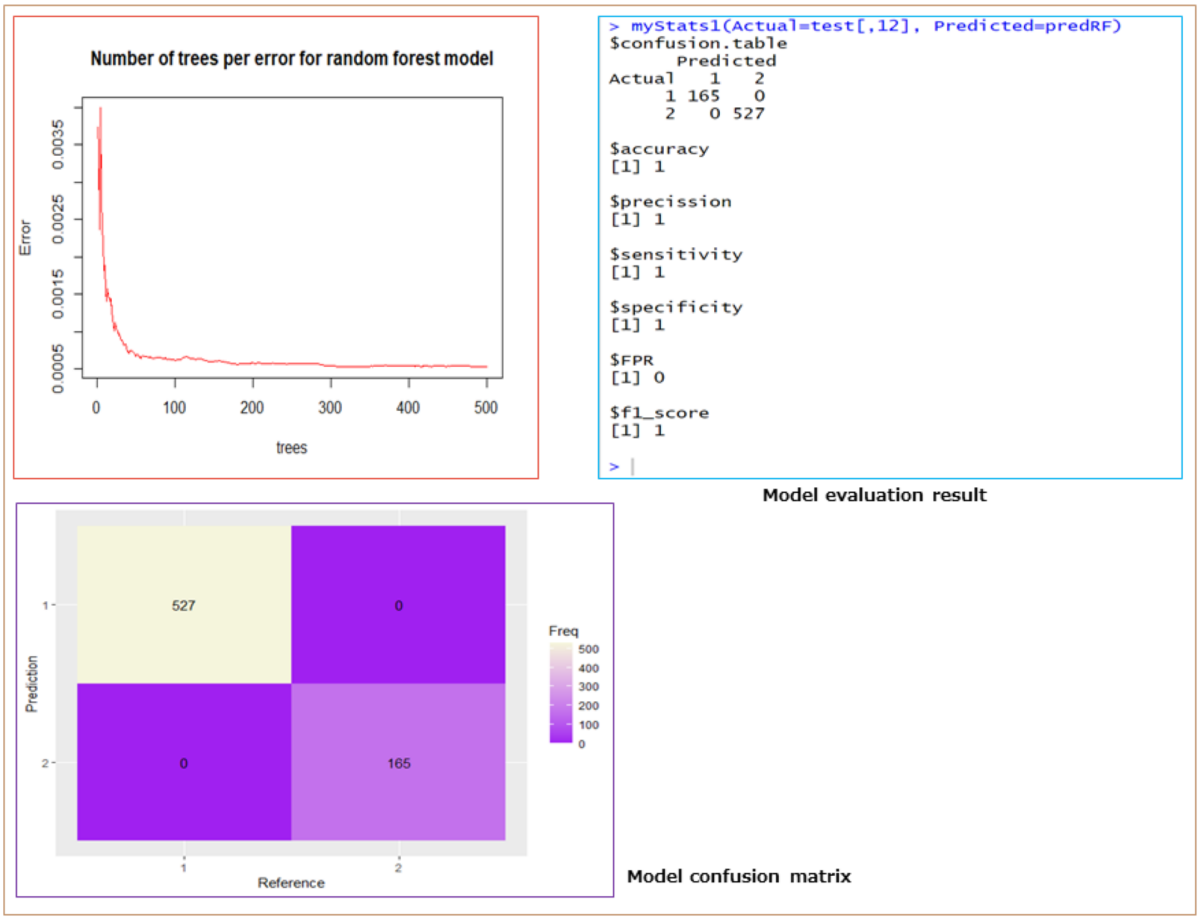


Figure 22: Soil degradation random forest for classification model results and confusion matrix

ted for, an objective this project achieved. The machine learning models used for this project automatically accounted for features interaction through the selected polynomial kernels (support vector machine, support vector machine for regression), recursive conditional splitting (random forest),'kernel density estimate (non-parametric Naïve Bayes) and although K-nearest neighbour does not directly account for feature interactions, the distance-based approach using similarity in observations of all variables imply consideration of feature interaction. This project went further to check the interaction for each model using a modified linear model (equation 10) where polynomials were added for the data non-linearity and degree2 introduced to account for all possible interactions in the data.

$$lm(Y\ X + I(X^2)) \tag{10}$$

Where: Y= Dependent variable X= Predictor I = Polynomial interaction term

4.4.2 Evaluation and Results

The result of the interaction algorithm pvalues showed geology,vegetation and major crop had meaningful interaction in the data while drainage slope and soil texture have marginal interaction, and further analysis using ANOVA showed those variables have pvalues less than 0.001 which confirms the high presence of interaction amongst them. figure 23 is an example of the ANOVA result for interaction check between geology and vegetation variables. It is clearly seen that the effect of the interaction of the two variables on the variation of the model outcome is fairly okay at a partial sum of squares value of 7.02. However it is worthy to note that the inability to correctly quantify these interactions have been a challenge in machine learning based digital soil map predictions (Taghizadeh-Mehrjardi et al. 2021).

The objective 8.3 in chapter 1, sub-section 1.1 has been achieved.

4.5 Conclusion

The developed prediction and classification soil degradation models will enable Nigerian government to educate farmers on how to replenish lost soil-elements, invest in the right type of vegetation suitable for the soil types and assist policy designs in formalising suitable grazing route for the herders for both dry and wet seasons which would be government-monitored not to encroach on the farmers land and vice-versa , thereby

```
> anova(Geo)
                Analysis of Variance          Response: SOIL_PH

 Factor                                                d.f. Partial SS MS        F     P
 GEOLOGY  (Factor+Higher Order Factors)                   6  40.734117 6.7890195 38.35 <.0001
  All Interactions                                        3   7.478721 2.4929071 14.08 <.0001
  Nonlinear (Factor+Higher Order Factors)                 4  23.054368 5.7635920 32.56 <.0001
 VEGETATION  (Factor+Higher Order Factors)                4   7.489798 1.8724496 10.58 <.0001
  All Interactions                                        3   7.478721 2.4929071 14.08 <.0001
 GEOLOGY * VEGETATION  (Factor+Higher Order Factors)      3   7.478721 2.4929071 14.08 <.0001
  Nonlinear                                               2   7.021799 3.5108997 19.83 <.0001
  Nonlinear Interaction : f(A,B) vs. AB                   2   7.021799 3.5108997 19.83 <.0001
 TOTAL NONLINEAR                                          4  23.054368 5.7635920 32.56 <.0001
 TOTAL NONLINEAR + INTERACTION                            5  23.230805 4.6461611 26.25 <.0001
 REGRESSION                                               7  41.078253 5.8683219 33.15 <.0001
 ERROR                                                 2989 529.140632 0.1770293
> # tests for interaction (shape differences across T, 3 d.f.)
> # anova includes a test for nonlinear interaction
```

Figure 23: Result sample showing interaction effect of geology and vegetation variables

minimising conflicts as both parties are likely to trust a data-based decision over human due to historical bias.

All the objectives have been achieved at this stage except the objectives on model comparisons which are going to be done in the next chapter 5.

5 Comparison of the Developed Prediction and Classification Models and Discussion

This chapter presents a comparison of the developed models, existing models and discussions.

5.1 Comparison of the Developed Prediction Models and Discussion

It can be seen from table 5 below that the evaluation metrics of random forest algorithm used for regression purposes are better than that of the support vector algorithm used for regression purposes, albeit marginally, except for the coefficient of determination which support vector regression had a higher value. This means with the lower error values random forest for regression ranks higher in performance to support vector machine for regression.

Table 5: Comparison of the Developed Prediction Models

Project Prediction Models			
	Model Types		
Model Metrics	Support Vector for Regression SVR	Random Forest Regression RFR	Ranking: if RFR=1 SVR=2
RMSE	0.01	0.0059	1
ME	0.0097	0.001	1
R-squared	0.9999	0.9994	2

5.2 Comparison of the Developed Classification Models and Discussion

When the classification models were ranked, random forest method used for classification also topped the rank amongst the classifications of soil pH (table 6) in all the metrics.
For the other four models; support vector machine is the best in accuracy, k-nearest neighbour best in recall and naive bayes best in precision. Since the models' algorithms are based differently, assessing on unified F1 score still made random forest the best closely by k-nearest neighbour, then support vector machine and naive bayes.

Table 6: Comparison of the Developed Classification Models

Project Classification Models							
Model Types							
Model Metrics	Support Vector Machine SVM	Random Forest Classification RFC	K- Nearest Neighbour (Two Models) KNN		Non-parametric Naive Bayes (Two Models) NB		Ranking: if SVM=1 RFC=2 KNN=3 NB = 4
	Soil pH	Soil pH	Soil pH	Soil Texture	Soil pH	Soil Texture	Soil pH
Accuracy	0.9999	1	0.9933	1	0.9183	1	2
Precision	0.9887	1	0.9912	1	1	1	2
Recall	0.9977	1	1	1	0.8901	1	2
FI-Score	0.9932	1	0.9964	1	0.9419	1	2

5.3 Comparison of the Developed Prediction Versus Classification Models and Discussion

Comparing the models used for both prediction and classification types, (SVR,SVM,RFR,RFC) the metrics showed they are all comparatively good but with Random forest classification having 1 all through only means random forest is better for classification than for prediction. The models have very high scores in respective metrics and are comparable. The prediction type models have negligible error values which indirectly mean they are of high accuracy, even though accuracy was not evaluated for prediction type and vice-versa for the classification models. Therefore, the developed models are of high performance and good fit to the dataset.

5.4 Comparison of the Developed Models Versus Existing Models and Discussion

Classification models:
The developed classification models benchmarked against multi-labelled soil types,the work of (Padmapriya. & Sasilatha 2023), were classified with different classifiers of which support vector machine, k-nearest neighbour and naive bayes were relevant to this project in terms of similarity of the classified soil types (clay, silt,loam,humus).For machine learning and comparison basis, this project synthesised data to be within same data size region of the models,(existing model=5938, project-5,568).

Table 7: Comparison of the Developed Classification Models Versus Existing Models

Developed Classification Models Versus Existing Classification Models								
	Support Vector Machine		K- Nearest Neighbour		Naive Bayes		Tree Algorithm	
Model Metrics	SVM	Labelled Soil Classifier	KNN	Labelled Soil Classifier	NB	Labelled Soil Classifier	RF	Soil Types Classifier
	Soil pH	Soil Types	Soil Types	Soil Types	Soil Types	Soil Types	Soil pH	Soil Types
Accuracy	0.99	0.92	0.99	0.81	0.91	0.81		
Precision	0.98	0.91	0.99	0.75	1	0.83		
Recall	0.99	0.89	1	0.75	0.89	0.80		
F1-Score	0.99	0.91	0.99	0.78	0.94	0.81		
Confusion Matrix							Test = 1	Test=0.84

The existing model classified the individual soil types and evaluated separately for all the models, unlike this project, so the existing model's highest score for each metric is used to compare.

The results of this project are substantially higher than those of the existing models across the metrics, as seen in table 7. This difference is more significant between existing models and developed K-nearest neighbour, Naive bayes and random forest models that classified soil types while support vector machine that classified soil pH has little differences in the metrics.This could be as a result of the data source difference; existing models used fresh soil samples, a method exposed to many live counterfactors and this project used data from a paper primary source that is subject to clarity and transformation risk.
Another plausible reason for this could be that the existing models experiment used soils high in pH size (being recent with climatic effects and different countries). Overall, both the existing and developed models had very high metrics evaluation, making them a good fit.
Prediction models: Assessing these two models types also showed steep differences in the evaluation metrics with the developed models having much higher values (table 8 refers). While the existing models values appear to be relatively good considering the fact that soil by nature is spatial, the developed models values appear to show 'over-fitting' which is expected with the source being a static one and the production of paper maps is known to be prone to loss of spatial variability (Grundy et al. 2020).

Table 8: Comparison of Developed Prediction Models Versus Existing Models

Developed Prediction Models Versus Existing Models				
Model Types	Support Vector for Regression		Random Forest for Regression	
Model Metrics	Support Vector Regression	Organic carbon prediction	Random Forest Regression	Organic carbon prediction
	Soil pH	Soil Carbon	Soil pH	Soil Carbon
RMSE	0.99	8.61	0.99	0.66
R-Squared	0.98	0.30	0.99	0.97
ME	0.99	1.63	1.00	0.02
MSE	0.99	N/A	0.99	0.44

Again with both the developed and existing models having high metrics, they are both fit models for the classification and prediction as confirmed by the high R-squared of both developed and existing random forest models indicating good fit although that cannot be said about the existing support vector model.

The objective 9.3 in chapter 1, sub-section 1.1 has been achieved.

5.5 Learning Outcomes and Limitations

The major learning outcome from this research is the honing of data analytical skill sets for prediction and classification modeling field of data science especially for the aspect of non-linear, non-normal data of spatial, irregular nature. It also improved research and project management skills giving way to expertise and confidence in data analytics in data science.

The limitation is on the restricted choices of machine learning methods that could be used with this dataset, for example, the geostatic methods could not be modeled because the dataset did not have the geographic data required.

6 Conclusion and Future Work

The advent of events over the past decades on climate changes, negative natural causes and human-induced actions all have had impact on global food security resulting from soil degradation, which spurred the interest in this project in the quest to finding solutions to Nigerian food insecurity, consistent farmers- herders conflict and assist Nigerian government in the pledge of making Nigerian farmers economically viable again.

To achieve these quests, this project's experiments were set to answer both the main research question of - How well would the machine learning models (Random Forest for Regression, Support Vector Machine for Regression) predict soil pH as key indicator of Nigerian soil degradation, using the Nigerian digital soil map attributes, to support the plans of Nigerian government towards food security and improve the economic power of farmers in Nigeria?
Sub-Research question of - To what extent can the classification models (K-Nearest Neighbour, Support Vector Machine, non-parametric Naive Bayes and Random Forest) help with enhancing classification of soil texture and soil PH level which will lead to reduction of soil degradation and sustainable farm management, making grazing routes and lands available for herdsmen, which will invariably mitigate the farmers-herders conflicts?

The set objectives to answer these research questions were all achieved following a modified knowledge discovery in database methodology. It started from the critical review of soil degradation literature between 2017 and 2023 where relevant and suitable existing models were selected (objective 1) and from which it was discovered that soil pH is a very good

indicator of soil organic carbon which is key causative factor to soil degradation. The Nigerian digital soil map dataset was acquired, extracted, transformed and pre-processed to be compatible with machine learning algorithms while data accuracy and integrity were ensured (objectives 2-6) then exploratory analysis and further pre-processes of the data (objective 7) unveiled the data was non-linear and did not follow the Gaussian distribution but all attempts to get it to conform were conteractive, hence the choice of developing non-linear models.

The developed prediction and classification models covered objective 8 ensuring those models had the capability to handle feature interactions, being one of the contributions of this project. This interaction identification was carried out with ANOVA which revealed the variables with interactivity, level of interactivity with parameters like partial sum of squares, mean square, F- statistics all which could be used to quantify the effect of such variable interaction on the variability of the outcome. This achieved objective 8.3.

The eight (8) developed models successfully predicted soil pH and classified soil textures with very high evaluating metrics (near-zero values for the error metrics and approximate 1 for coefficient of determination, accuracy, precision, recall and f1-score). It was discovered the existing models, from objective 1, had steeped differences for both prediction and classification models, when metrics were compared with those of the developed models, especially for the soil types classifier models. This could be down to many reasons some of which are discussed in model comparison section, achieving objective 9.

With the success of this project, the research questions can be answered thus:

Research question- The machine learning models have accurately predicted soil pH with extremely low error rate, using the Nigerian digital soil map dataset,so the plans of the Nigerian government towards food security and improving farmers economic power can be supported with the models.

Sub-Research answer: Since the classification models accurately classified the soil textures and soilPh with approximately perfect accuracy, the models can be used for Nigerian soil classification to educate the farmers on the type of crops that suits the soil type thereby reducing the need of farmers fallow land which herdsmen can use for grazing, thereby mitigating the conflicts while the farmers economic power increases.

Future Work: As successful as the project experiments are, some improvements would be useful, for example the use of deep learning methods

like Genetic algorithm- back propagated neural network (GA-BPNN). It was mentioned in many of the literature reviewed as exceptional in performance but could not be used as it was out of the project's scope.
It will also be helpful if model-diagnostics can be used to algorithmically and correctly quantify the effect of variable interaction which could be a solution to the possible over-fitting results of the models.

Acknowledgements

My sincere gratitude goes to my supervisor, Dr Catherine Mulwa who challenged and encouraged me in the course of this programme, I say a big thank you for everything..., you will always be remembered.
Worthy of mention are my children and siblings who reduced the physical burden on me but most essential is my God– ALLAH.

There are no words to express my appreciation to Allah for all and everything to me, before and during this course, especially through the many personal challenges at the time of the postgraduate stage of this course.
Although you are more worthy than everything, I still dedicate this thesis to you my GOD.., ALLAH. Alhamdulilah !!

References

Aderele, A. S., de Clercq, W. P. & van Niekerk, A. (2017), 'Development of a composite soil degradation assessment index for cocoa agroecosystems in southwestern nigeria', *Solid Earth* **8**, 827–843.

Al-Kaisi, M. & Lowery, B. (2017), *Soil Health and Intensification of Agroecosystems*, Elsevier Science.
URL: *https://books.google.ie/books?id=2pqpDQAAQBAJ*

Amusan Lere, Abegunde Ola, A. E. T. (2017), 'Climate change, pastoral migration, resource governance and security,: the grazing bill solution to farmer-herdrer conflict in nigeria', *Environmental Economics* **8**, 35–45.

Amuyou, U. A., Wang, Y., Ebuta, B. F., Iheaturu, C. J. & Antonarakis, A. S. (2022), 'Quantification of above-ground biomass over the cross-river state, nigeria, using sentinel-2 data', *Remote Sensing* **14**(22).
URL: *https://www.mdpi.com/2072-4292/14/22/5741*

Baltensweiler, A., Walthert, L., Hanewinkel, M., Zimmermann, S. & Nussbaum, M. (2021), 'Machine learning based soil maps for a wide

range of soil properties for the forested area of switzerland', *Geoderma Regional* **27**, e00437.
URL: *https://www.sciencedirect.com/science/article/pii/S2352009421000821*

Bennett, J., Roberton, S., Ghahramani, A. & McKenzie, D. (2021), 'Operationalising soil security by making soil data useful: Digital soil mapping, assessment and return-on-investment', *Soil Security* **4**, 100010.
URL: *https://www.sciencedirect.com/science/article/pii/S2667006221000071*

Chen, L., Ren, C., Li, L., Wang, Y., Zhang, B., Wang, Z. & Li, L. (2019), 'A comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content', *ISPRS International Journal of Geo-Information* **80**, 40174.

Cho, S., Kim, H.-S. & Kim, H. (2023), 'Locally specified cpt soil classification based on machine learning techniques', *Sustainability* **15**(4).
URL: *https://www.mdpi.com/2071-1050/15/4/2914*

Falaki, M. A., Ahmed, H. T. & Akpu, B. (2020), 'Predictive modeling of desertification in jibia local government area of katsina state, nigeria', *The Egyptian Journal of Remote Sensing and Space Sciences* **23**, 363–370.

Grundy, M. J., Searle, R., Meier, E. A., Ringrose-Voase, A. J., Kidd, D., Orton, T. G., Triantafilis, J., Philip, S., Liddicoat, C., Malone, B., Thomas, M., Gray, J. & Bennett, J. M. (2020), 'Digital soil assessment delivers impact across scales in australia and the philippines', *Geoderma Regional* **22**, e00314.
URL: *https://www.sciencedirect.com/science/article/pii/S2352009420300638*

Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. (2017), 'On calibration of modern neural networks'.

Haghighi, A. T., Darabi, H., Karimidastenaei, Z., Davudirad, A. A., Rouzbeh, S., Rahmati, O., Sajedi-Hosseini, F. & Klöve, B. (2020), 'Land degradation risk mapping using topographic, human-induced, and geo-environmental variables and machine learning algorithms, for the poldoab watershed, iran', *Environmental Earth Sciences* **80**, 1866–6299.

Hengl, T., E, M. A. & Miller (2021), 'African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning', *Scientific Reports* **11**, 2045–2322.

Kidd, D., Searle, R., Grundy, M., McBratney, A., Robinson, N., O'Brien, L., Zund, P., Arrouays, D., Thomas, M., Padarian, J., Jones, E., Bennett, J. M., Minasny, B., Holmes, K., Malone, B. P., Liddicoat, C., Meier, E. A., Stockmann, U., Wilson, P., Wilford, J., Payne, J., Ringrose-Voase, A., Slater, B., Odgers, N., Gray, J., van Gool, D., Andrews, K., Harms, B., Stower, L. & Triantafilis, J. (2020), 'Operationalising digital soil mapping – lessons from australia', *Geoderma Regional* **23**, e00335.
URL: *https://www.sciencedirect.com/science/article/pii/S2352009420300845*

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C. & Nkuba-Kasanda, L. (2020), 'Analysing the impact of soil spatial sampling on the performances of digital soil mapping models and their evaluation: A numerical experiment on quantile random forest using clay contents obtained from vis-nir-swir hyperspectral imagery', *Geoderma* **375**, 114503.
URL: *https://www.sciencedirect.com/science/article/pii/S0016706119322736*

Liu, X., Zhu, A.-X., Yang, L., Pei, T., Qi, F., Liu, J., Wang, D., Zeng, C. & Ma, T. (2022), 'Influence of legacy soil map accuracy on soil map updating with data mining methods', *Geoderma* **416**, 115802.
URL: *https://www.sciencedirect.com/science/article/pii/S0016706122001094*

Lu, W., Zhang, Z., Zhang, S., Zhang, T., Wan, Y. & Li, Y. (2022), 'Screening of six cation exchange resins for high binding capacity, monomer purity and step yield: A case study', *Protein Expression and Purification* **199**, 106155.
URL: *https://www.sciencedirect.com/science/article/pii/S1046592822001127*

Mosweu Plefhile, M. T. (2023), 'The influence of archives in conflict resolution: A case study of botswana and namibia', *African Journal of library, archives and information science* **33**, 23–26.

Nkwunonwo, U. & Okeke, F. (2013), 'Gis based production of didital soil map for nigerai', *Ethiopian Journal of Environmental Studies and Management* **6**, 5–7.

Okeke-Ogbuafor, N., K, A. & T, G. (2019), 'Two approaches to conflict resolution and their applicability to ogoniland, nigeria', *JSTOR* .

Olademo, O., Omotoye, R., Ikibe, S., Ibraheem, L., Tijani, Y., Abubakre, S., Adebisi, A., Aboyeji, A., Fahm, A. & Adimula, R. (2021), 'Internal

mechanisms as tools for conflict resolution: A case study on share-tsaragi, nigeria', *Heliyon* **7**(1), e05974.
URL: *https://www.sciencedirect.com/science/article/pii/S2405844021000797*

Padmapriya., J. & Sasilatha, T. (2023), 'Deep learning based multi-labelled soil classification and empirical estimation toward sustainable agriculture', *Engineering Applications of Artificial Intelligence* **119**, 105690.
URL: *https://www.sciencedirect.com/science/article/pii/S0952197622006807*

Pahlavan-Rad, M. R., Dahmardeh, K., Hadizadeh, M., Keykha, G., Mohammadnia, N., Gangali, M., Keikha, M., Davatgar, N. & Brungard, C. (2020), 'Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern iran', *CATENA* **194**, 104715.
URL: *https://www.sciencedirect.com/science/article/pii/S0341816220302654*

Peng, Y., Zhao, L., Hu, Y., Wang, G., Wang, L. & Liu, Z. (2019), 'Prediction of soil nutrient contents using visible and near-infrared reflectance spectroscopy', *ISPRS International Journal of Geo-Information* **8**(10).
URL: *https://www.mdpi.com/2220-9964/8/10/437*

Pham, B. T., Nguyen, M. D., Nguyen-Thoi, T., Ho, L. S., Koopialipoor, M., Kim Quoc, N., Armaghani, D. J. & Le, H. V. (2021), 'A novel approach for classification of soils based on laboratory tests using adaboost, tree and ann modeling', *Transportation Geotechnics* **27**, 100508.
URL: *https://www.sciencedirect.com/science/article/pii/S2214391220303962*

Shepherd, K. D., Ferguson, R., Hoover, D., Fenny, Sanderman, J. & Ge, Y. (2022), 'A global soil spectral calibration library and estimation service', *Soil Security* **7**, 100061.
URL: *https://www.sciencedirect.com/science/article/pii/S2667006222000284*

Tadesse, Z., Abere, M., Azene, B., Kaiwen, P., Mulatu, Y. & Francis, M. (2023), 'Lowland bamboo (oxytenanthera abyssinica) deforestation and subsequent cultivation effects on soil physico-chemical properties in northwestern ethiopia', *Advances in Bamboo Science* **4**, 100038.
URL: *https://www.sciencedirect.com/science/article/pii/S2773139123000241*

Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Ghebleh Goydaragh, M., Schmidt, K. & Scholten, T. (2021), 'Enhancing the accuracy of machine learning models using the super

learner technique in digital soil mapping', *Geoderma* **399**, 115108.
URL: *https://www.sciencedirect.com/science/article/pii/S0016706121001889*

Wonah, D. E. & Bullem, A. G. (2019), 'Normadic education for national integration in nigeria', *E-Petagogium* **19**, 55–57.

Yu, H., Wang, L., Wang, Z., Ren, C. & Zhang, B. (2019), 'Using landsat oli and random forest to assess grassland degradation with aboveground net primary production and electrical conductivity data', *ISPRS international journal of geo-information* **8**, 0511.

Zhang, H., Wu, P., Yin, A., Yang, X., Zhang, M. & Gao, C. (2017), 'Prediction of soil organic carbon in an intensively managed reclamation zone of eastern china: A comparison of multiple linear regressions and the random forest model', *Science of The Total Environment* **592**, 704–713.
URL: *https://www.sciencedirect.com/science/article/pii/S0048969717303959*