Application of Machine Learning Techniques for Soil Type Classfication of Karanataka

MSc Research Project
Data Analytics

Shozab Raza Ansari
x17106974

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Shozab Raza Ansari
Student ID:	x17106974
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Noel Cosgrave
Submission Due Date:	13/08/2018
Project Title:	Application of Machine Learning Techniques for Soil Type Classfication of Karanataka
Word Count:	6880

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 encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	16th September 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. 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):	

Application of Machine Learning Techniques for Soil Type Classfication of Karanataka

Shozab Raza Ansari
x17106974
MSc Research Project in Data Analytics

16th September 2018

Abstract

Soil science and its integration with machine learning has been into practice since the past few decades. Within the agriculture domain soil classification is an essential work that has to be conducted so as to provide good classifiaction sytmes for the soil types. Karnataka state has registerd the highest suicides in India. With the data about soil health of Karnataka state the different types of soils was analyzed and classified using different machine learning techniques. This research study for classification of soil types was conducted using tree-based model Decision Tree (C5.0), Random Forest (RF). Support Vector Machines (SVM) and eXtreme Gradient Boosting (XGBOOST). Accuracy and Kappa values suggested XGBOOST performed the best whereas the time of execution for these models differed and Random Forest had the most effient compution time with relatively comparable accuracy.

"Conducting soil research without proper classification would be comparable conducting a field experiment with green plants or a laboratory experiment with some extremely small soil animals" (Hartemink; 2015)

1 Introduction

The assessment and analysis of soil quality has been on a rise from past few decades. This in turn has been a result of emphasis on the land use, soil erosion control as well as soil management. The main factors which directly relate to the soil quality are dynamic, inherent biological and chemical properties. Inherent properties that are predominant within most soils are pH, Organic Matter (OM), bulk density etc (Karlen et al.; 2003). Dynamic properties of the soil are likely to change easily with change in the management practices, whereas the inherent properties of the soils are comparatively hard to change SQI [1].

(Karlen and Rice; 2015) stated that the soil degradation problem has now become global and there are various reasons behind this issue such as rotation of crops in an unsuitable manner, felling and cutting of trees on a large scale, increase in the construction(urbanization), high grazing of the crop areas and many more.

[1] https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/health/

(Hartemink; 2015)in his paper about the use of soil classification papers puts the importance of soil taxonomy research papers. (Hartemink; 2015) has acknowledged the fact that very small amount of time has been spent in the past four decades for developing proper soil classification systems. The number of people who work as soil scientists or pedologist have also decreased, though the information about soil classification present on the web is ample but they lack to take advantage of it.

Machine learning has been on the rise in agriculture domain. Multifaceted problem within the agriculture domain can be countered using the techniques of data mining. (Mucherino et al.; 2009) reported different mining techniques both supervised and unsupervised such as K Nearest Neighbors (KNN), Artificial Neural Networks (ANN), k-means as well as Support vector Machines (SVM) are viable to tackle problems within agriculture domain. Application of the data mining techniques in the agriculture domain are diverse ranging from weather forecasting, soil characteristic study, wine fermentation, pig cough sound recognition etc.

For this study the focus was to study and implement Supervised Machine learning algorithms C5.0, Support Vector Machine (SVM), Random Forest (RF) and eXtreme Gradient Boosting (XGBOOST) were implemented to tackle the problem of classification of different soil types in the region of Karnataka, India. The study was conducted using the data acquired from ICRISAT [2] about soil health status across the Karnataka state.

In the following sections the different phases of this study have been presented. Section 1.1 provides the details to the motivation of taking up this study. Section 1.3 describes the problems related to the domain of the conducted study. Section 2 exhaustively presents the related work previously done in the same domain. Section 3 outlines the methodology taken up for the study. Section 4 presents the detailed implementation procedure of the study conducted. Section 5 and 6 provide the evaluation and result finding of the study respectively. Finally, Section 7 provides the scope of future work that can be taken up.

1.1 Motivation and Objective

Herein this section, the motivation gathered to take up this study and propose a classification system is described.

(Kumar and Hashim; 2017) outlined the problem of farmers in the Karnataka state of India, the authors stated that suicide rates of Karnataka farmers (17.8%) were found to be higher than the national average of suicides happening in India (10.6%). While delineating the suicides of Karnataka farmers various characteristics were presented such age the age groups involved, past history, alcohol intoxication etc. as well as the mode that they used to commit the suicide. Within the factors that caused the suicides it was noted that maximum number of suicides had taken place within the farmers belonging to the lower socioeconomic status.

Studying about the Soil Health Mission conducted by the Department of Agriculture, Government of Karnataka provided a basis for this study. The Department of Agriculture did the study about the nutrient of soil present in the soils of Karnataka. The study analyzed that though excessive use of fertilizers and the irrigation have increased the production of food grain, it is in turn resulted in the depletion of soil nutrients such as Zinc (Zn), Iron (Fe), Boron (B), Manganese (Mn), Copper (Cu) etc. Within the soil health mission, the different samples of soils were analyzed for macro, micro and secondary nutrients. Important components of the soil such pH and Electrical Conductivity (EC)

[2] http://www.icrisat.org/

were also analyzed. The mission aimed to provide recommendation about the soil samples to the farmers of the Karnataka state. As of 2015-16 study 78,32,000 farm infrastructures are established and within those 1.30-1.35 samples of the soils are taken under analysis annually. The mission aimed to complete the objectives of issuing soil health card to the farmers of Karnataka state so as to overcome the inaccurate fertilization practices. Provide recommendation of fertilizers to the farmers. Soil Health Mission -Karnataka[3]

For this study the components and properties of different types of soils were used as described by the United States Department of Agriculture (USDA) [4]. Since the chemical indicators and heavy metals present such as Aluminum (Al), Copper (Cu), Zinc (Zn) etc. play an important role in the saturation, citation-exchange capacity, pH, Phosphorus and Electric Conductivity (EC) are also important in soil quality assessment.

Developing a classification model which could classify the different types of soils while taking soil health indicators as the predictors would help in proper agricultural practices such as fertilizer use and land reuse over the different soil types.

1.2 Research Question

Using classifiication techniques, how can the different types of soils in Karnataka region would be classified?
Amongst the different classification techniques applied, how accuractly can the best model be evaluated?

1.3 Problem Statement

Various problems are faced in the recent times within the domain of soil science and agriculture. These problems must be considered as challenges to overcome so as to improve the soil quality and fertility and hence, in turn, improve the agriculture practices.
Land Degradation - Soil productivity is highly affected by land degradation. The acknowledgment is taken towards the soil importance in agriculture as well as food security and nutrition. A more extensive association between the land degradation and the productivity of soil cannot be understood clearly. Lenka et al. (2017)
Change of land under use - To what extent the biomass in available to the soil after the land usage is a major factor when assessing the nutrients present in the soil. The erosion of soil and loss of nutrient is highly affected by what type of crop and at what time the cropping is being done. AGBOOLA et al. (2017)
Different council of the national research has discussed the soil properties and then published various reports and papers of strategies regarding the same. Pinning down to one single point all the council has reported that the quality of the soil is decreasing rapidly and the need of improving the soil quality is very much to be taken care of.Karlen and Rice (2015)

2 Related Work

Previously studies have been conducted under the area of and machine learning relating with the soil science.

[3] http://raitamitra.kar.nic.in/ENG/Document/SHM.pdf

[4] https://www.nrcs.usda.gov/wps/PA$_{N}RCSConsumption/download?cid = stelprdb1269818ext = pdf$

2.1 Machine learning work done

Research have been conducted previously within the domain of machine learning and data mining. Reviewed below are some of the relevant work that have been published in the past in the soil since domain integrated with machine learning.

(Bhattacharya and Solomatine; 2006) in the previous decade acknowledged the association of machine learning with the classification of soil and conducted their study using the data about Cone Penetration Testing (CPT) which is a less expensive method. A novel algorithm namely Constraint Clustering and Classification (CONCC) was used for segmenting the data and the classification. Further, for classification the classes of the segmented data were labeled by experts. Ultimately, machine learning algorithms were used Decision Trees (DTs), Artificial Neural Networks (ANNs) and Support Vector Machines (SVM). The outcome classes which were classified were *na*mely Clayey Peat, Clayey sand, Peat, Clay and Sand. The authors under 3 classification problem where they classified soil into sandy or not (binary), sandy, clay or peat (3 class-multi class) and seven class classification. The classifiers were trained using optimum parameters which were exhaustive and un-exhaustive respectively. Accuracy of 83% was achieved by the authors on the test set where DT had outperformed the other algorithms.

Specifically concentrating over the Bayesian classifier (Bhargavi and Jyothi; 2009) applied five different types of BN classifiers such as Naive Bayes (TAN), Bayesian network augmented Naive Bayes (BAN), Bayesian multi-nets etc. to handle the problem of soil classification. Adhering the soil classification system proposed by the Unified Soil Classification System (USCS) the authors classified different types of soils- Clay, Clay loamy, Loam, Sand, Sandy loam etc. the data acquired was from the Chittoor district of Andhra Pradesh. The data was transformed and pre-processed using the Excel tool which was then fed into the WEKA software for applying the data mining algorithms. Judging by the accuracy the authors proposed the 100% classification rate for soil data using Naive Bayes classifier. Unexpectedly, the root mean squared error, relative absolute error metrics were also considered which are not the best metrics to evaluate classification models.

(Kumar and Kannathasan; 2011) presented a detailed overview of types of techniques and data mining algorithms that have come into practice to tackle problems involved with the soil research. Taking into consideration the dataset from World Soil Information International Soil Reference and Information Centre (ISRIC) authors conducted a survey of previously published researches done the soil and data mining domain. The authors did a thorough survey of what all techniques and algorithms ca be used for the classification of soil. Describing the algorithms such as Support Vector Machines, Decision Trees (DTs), k Nearest Neighbors (k-NN), Bayesian Networks (BNs), k Means and Artificial Neural Networks (ANNs). The authors also described the techniques to optimize the parameters for the above-mentioned algorithms using Particle Swarm Optimization (PSO) and Simulated Annealing (SA). Finally, the authors concluded that mixture of various data mining algorithms can be effective in soil classification problem.

(Gambill et al.; 2016) after acquiring the digital soil data of about 131 countries from the USDA (United States Department of Agriculture) did an exhaustive research for classifying the soils by the standards of USCS (Unified Soil Classification System) and USDA. The authors classified soils such as Clayey gravel, Silty sand, Clayey sand etc. by using 15 variables such as the Organic Matter, Drainage class, Available water capacity, Bulk density etc. as the predictor variables. The research was focused upon a single machine learning algorithm namely Random Forest. Various models of the same algorithms were developed amongst which the Model-1 provided the highest accuracy of 99.1%. Furthermore, important variables involved in the prediction were predicted using the same algorithm. Removal of these variable lead to a sudden increase of the error rates. It can wise correct to say that Random Forest which acts as an ensemble technique is good for classification problem. When considering the Model-1 it was found out that Organic matter as a predictor is of low importance for soil classification. The remarkable work done by (Gambill et al.; 2016) has been taken into consideration in this study.

Relating the importance of soil classification to the agricultural as well economic development in India (Hemageetha; 2016) conducted their research on the classification of soil by taking 8 attributes for training the classifiers which were ph, Electric Conductivity (EC), Organic Carbon (OC), Phosphorus (ps) etc. the classifiers built were Naive Bayes, J45(C.5), k- means. The types of soils that were classified were Alluvial, Black, Laterite, Mountain and Red. J48(C5.0) was found to be the most efficient with an accuracy of 91.90% for the classification. Authors used the Apriori algorithm for finding the combination of crops that are grown on different types of soil. By applying the Apriori algorithm it was found out that the Rice, Sugarcane and Wheat crops were most frequent. Though the research was quite informative the dataset only had 108 instances which can be said less in terms of data mining applications.

As discussed in the earlier section grass land degradation is a problem that has long existed and has been on the rise even since. (Li et al.; 2017) in their research predicted the Carbon (C), Nitrogen (N), and Phosphorus (P) so as the study the extent of grassland degradation. The data collected about the Jilin Province of China was used for building models namely Radial Basis Function Neural Networks (RBFNN) and Support Vector machines (SVMs). The analysis was conducted on MATLAB and careful consideration was taken when selecting the parameter for the model (SVM) such penalty parameter and kernels. The distribution of ratio of Carbon, Nitrogen and Phosphorus was divided into 5 classes. The degradation level was then studied on the different levels and it was deduced that the degradation was mostly present in 3rd and 4th levels. It was concluded that the RBFNN and SVM are adequate techniques in the prediction of carbon, nitrogen and phosphorus content.

With the aim to compare different machine learning algorithms over the data gathered from distinct places namely United States of America and New Mexico (Brungard et al.; 2015) classified the soil taxonomic classes. Digital soil mapping was conducted previously to predict the soil classes. For the different area classification algorithms were implemented such as k-nearest neighbors (KNN), Linear discriminant analysis (LDA), Linear Support Vector Machines (SVML), Radial-basis Support Vector Machines (SVMR), Multi-layer perceptron neural networks (MLP) etc. Not surprisingly Random Forest (RF) using the covariates set 3 (predictors) had outperformed the other classifier as it had the highest kappa value as well. Within the Random Forest (RF) the recursive elimination method was done so as to identify the best set of covariates for classification. Radial-basis Support Vector Machines (SVMR) and single-hidden-layer neural network (NNET) also performed good.

(Kovačević et al.; 2010) while working with the chemical and physical properties of soil applied Support Vector Machines (SVMs) to classify the soil types. The chemical and physical properties used as the predictors for the classification were Soil Organic Matter, Soil pH, Nitrogen (N), Potassium (K), Phosphorus (P) etc. After giving a brief description of how support vector machines work and the concept of Hyperplane within it, the authors explained the model that had built for the classification. The dataset was gathered from the Institute of Soil and Melioration at the faculty of Agriculture, University of Belgrade. The model implemented was by using the Gaussian and linear kernel separately. The analysis was conducted on the WEKA tool and metrics chosen for classification evaluation were F1 measure and kappa statistics. The classification model was built to predict the 8 soil types using 7 chemical and 3 physical properties. Apart from the (SVM), logistic regression was also implemented. It was concluded that both the techniques are suitable for classification of soil types. I was also noted that linear (SVM) can be good when the observations per class is small.

In the practice of agriculture, the wetting and drying of soil is an important process the moistness of soil can cause delays in the practice and can be costly for agriculture (Coopersmith et al.; 2014) in their study the data gathered from NEXRAD about the precipitation over a period of time was used. Classification algorithms namely K- Nearest Neighbors (KNN), Decision Trees (DTs) and boosted perceptron were used. The problem was basically binary in nature where the prediction was used to tell whether the soil will be ready or not. (KNN) was noted to outperform the other two algorithms with an accuracy of 93% followed by boosted perceptron algorithm.

(Harlianto et al.; 2017) in their study for soil type classification worked with 10 attributes to classify the 12 classes. The authors narrowed down to Neural Networks, Naive Bayes, Decision Trees (DTs) and Support Vector Machines (SVMs) as the techniques for the research. For the (SVM) linear kernel was used while building the model. In their experiment (SVM) outperformed the other algorithms with accuracy of 82.3% followed by Neural Nets. Though reduction of attributes increased the accuracy of Neural Nets and Naive Bayes, it failed to do so with SVM. Their research was conducted on the Rapid Miner tool. The authors failed to provide any emphasis on parameter optimization for the models.

(Sirsat et al.; 2017) did classification of agricultural soils in India using various parameters and different classifiers. The authors also acknowledged the importance of providing a nutrient management system for soil. Various classification problems were undertaken by the authors such as classification of soil nutrients, classification of soil pH (salinity), classification of soil types as well as classification of crops. Specifically, within the soil types classification, loamy, mixed etc. soils were classified using different machine learning techniques namely Logistic regression (LR), Support Vector Machines (SVMs), K Nearest Neighbors (KNN), Artificial Neural Networks and Random Forests (RF). The implementation was done using JAVA and MATLAB tool. For the soil type classification, the best classifier was found out to be a hybrid of decision table and Nave Bayes with a kappa value of 97.82% and Random Forest classifier (non-pruned) with kappa value of 96.80%. This extensive research is also taken into consideration while building the Random Forest model for this proposed study. The research conducted by the authors was for the state of Maharashtra in India, which can be related to this study which focuses on the study about soil classification and soil parameters in Karnataka state of India. Their research was focused on providing recommendation to the government for the proper utilization of fertilizers and improving the soil quality around the state.

(Stevenson et al.; 2015) Conducted their research for the data acquired from about 700 locations in New Zealand. For their research the indicators which were listed as the predictors were pH, Carbon (C), Nitrogen, (C)/(N) ratio Bulk density etc. which are key components of the soil play an important part in the soil quality assessment. Different hypotheses were proposed by the authors which were regarding the distinctions of native and managed sites, clustering of managed sites and relation amongst the soil quality and land use pattern. After successful soil sampling the data analysis was conducted on the R software. To find out the correlation between the indicators of the soil authors conducted Principal Component Analysis. (PCA) results indicated that the 90% of variance was showed by 4 principal components. Authors applied the fuzzy c-means to find the clusters in the data. It was also concluded that clusters occurred within the managed sites.

(Shastry et al.; 2014) developed various soil types classifiers using the data from Nation Bureau of Soil Sciences (NBSS), India. The authors analyzed the 2593 samples of soils, and implemented the tree based C4.5 and CART algorithms. The attributes used as the predictor variables were pH, Electric Conductivity (EC) and Exchangeable Sodium Percentage (ESP). This data was from the Karnataka state of India. The authors achieved an accuracy of 91% for CART, 92.5% and 97% with the proposed novel method. For their novel method the authors considered the limitation of CART and proposed to calculate the Gini index of the attributes beforehand.

3 Methodology

This section provides an exhaustive explanation of the methodology followed for this study. Knowledge Discovery in Databases (KDD) methodology was followed for this study as the problem at hand is of environmental concern and not business oriented.

(Fayyad et al.; 1996) gave a rationale of manual data analysis in various domains such as health-care industry where the trend analysis about the changes in the domain are analyzed on quarterly basis, remote-sensed images practiced by geologists etc. In these scenarios the data is costly, big as well as subjective in nature. One may say that manual data analysis fails in these scenarios. Knowledge Discovery in Databases (KDD) is basically the extracting knowledge completely from the data at hand, data mining is crucial step within the KDD process. Some additional step that KDD involves are data selection, cleaning preprocessing etc. The idea behind KDD is to automate the process of data inference with the help of computers, this is where the data mining steps comes into play. (Fayyad et al.; 1996) defined KDD as "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

The entire KDD process followed is described in a figurative manner in is shown in [Fig. 1] Adhering to the steps involved in the KDD, the study was conducted, and the various steps were covered in a sequential manner.

1. Getting to know about the domain of this study was crucial, so as to understand the soil science. While studying about different types of soils application of machine learning to soil research was also understood. On a broader level machine learning plays important role in agricultural domain as described by (Mucherino et al.; 2009) and narrowing down specifically, soil classification and analysis related to soil science has also been in existence since a long time (Hartemink; 2015). Socio-economic status of farmers of Karnataka, India and the Soil Health Mission conducted by the Government of Karnataka, Department of Agriculture also inspired to take up this study.

2. To acquire the data about the soil health, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) database was used. ICRISAT is an organisation

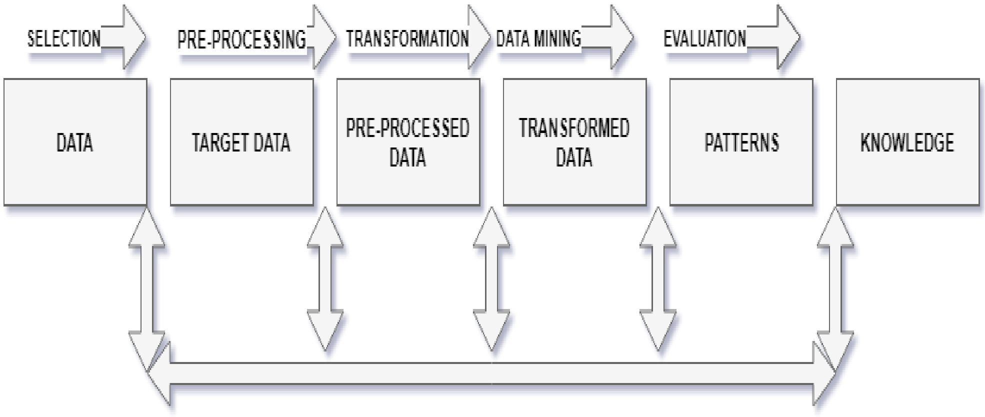

Figure 1: KDD Process
(Fayyad et al.; 1996)

which serves on an international level. It profoundly conducts researches in the rural areas. Soil health data gathered contained observations of different soil types present in the region and their specific compositions. The composition contained chemical indicators as well as some inherent features present in the soil, previously discussed in Section 1.

3. As depicted in the KDD data cleaning and pre-processing, outlier removal etc. are the next step. For the study Microsoft Excel and R Studio(RStudio Team; 2015) were used for cleaning and pre-processing purposes. The dataset had spelling mistakes within the names of soil types (Exp Latirite to Laterite), manual correction of the names of soils was done in Microsoft Excel. Though the data observations of various types of soils, the dominant types of Soil were chosen after referring to the Geography of Karnataka [5]. Soil heath data also contained missing values in some the features which were tackled using the MICE (Multivariate Imputation by Chained Equations) function the R Studio. Some algorithms such as Support Vector Machine (SVM) and Principal Component Analysis (PCA) cannot handle categorical data but work only on the numeric data, hence normalization of the data was also done as a part of pre-processing. An important aspect when dealing with classification problem is the class imbalance. Amongst the 5 classes under study class imbalance was found which was later on fixed by the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al.; 2002).

4. Fourth step in the KDD involves data reduction and projection. The problem at hand for the for the classification of soil types was a multiclass classification problem. Since the soil types (outcome variable) that were to be classified had 5 classes namely Black, Laterite, Mixed, Red and Sandy soil. It was essential to check whether the data was linearly separable or not. Implementing Principal Component Analysis (PCA) so as to find clusters within the data provided insights about the features. PCA was also implemented to for the dimensionality reduction and important feature selection as well as the variance showed by different features acting as the predictor variables. To check the relationship amongst the features in the data correlation check was also performed. Though Random Forest was

[5] https://en.wikipedia.org/wiki/Geography$_{o}f_{K}arnatakaSoil_{t}ypes$

implemented as a model for classification, its application for selecting the important features acting as the predictor was taken into consideration(Breiman; 2001).

5. With the objective of the study already in place (classification of soil types) for this step of the methodology it was crucial to select the data mining techniques to that were apt for tackling the multiclass classification problem. As discussed in the Section 1 machine learning algorithms were used in soil types classification.

6. The data mining techniques chosen for this study were shortlisted after carefully doing the extensive literature review in the soil classification field.

   Previously researcher have implemented algorithms which have performed exceptionally for classification of soil. Gambill et al. (2016) while classifying the soil types by using Random Forest models achieved accuracy of 99.1% similarly (Sirsat et al.; 2017) achieved an accuracy of 96.8% for the Random Forest as discussed in the Related work section. Likewise, The performance of Support Vector Machines Radial Basis (SVMR) in the study by (Brungard et al.; 2015) was commendable. The repeated use of Decision Trees, C5.0 was taken into account and was also implemented for this study

7. Implementation of the different models that had been shortlisted were done using the R studio. The 4 specific algorithms chosen namely, C5.0, Random Forest (RF), Support Vector Machines (SVM) and eXtreme Gradient Boosting (XGBOOST) were implemented on the soil health data. While conducting the study the tree-based models C5.0, Random forest had shown good results and comparatively less computation time, analysing this, XGBOOST algorithm was implemented to achieve even a higher classification accuracy. (Chen and Guestrin; 2016) proposed that XGBOOST is system which boosts the trees in the model.

8. For the evaluation and interpretation the prediction/performance of the model is by analysing the confusion matrix produced by the model. Unlike the binary classification problem, the confusion matrix that was made by the different models was 5x5 as there were 5 classes to be classified. The matrix provides the correctly classified and misclassified instances of the model. Within the confusion matrix, accuracy as well as the kappa values were taken to judge the models performance.

9. Use of the discovered findings and the knowledge inferred by applying the data mining techniques and algorithms were for academic purposes. Usage of this study and its proposed system may help in better soil classification when used in more improvised manner.

3.1 Models And Techniques

3.1.1 C5.0

C5.0 is a logical model which resemble the structure of a tree. Following its tree-based approach the decisions are made at each node. After analyszing the best decision at the node the splitting happens finally the outcome is predicted at the end node also known as the terminating node. C5.0 is an advanced version of C4.5 algorithm which had previously been discovered for classification as well as regression problems (Simon; 2018)

3.1.2 Random Forest

Random forest is basically mixture of trees that are used for prediction. Random vectors are generated and further sampled for all the trees in the model. Random forest occurs when a large number of trees are made, and the most popular class is voted for. Random forest help in the variance error reduction by constructing various decision trees. In definition the Random Forests classifier is combination of multiple classifiers which are in a tree structure. It can be further represented as:

(h(x,θ_k)$, k = 1......$)

Herein, $(\theta_k) i.e.$

Random vectors which are independent and identically distributed.

x = input for the classes The basic idea is that a vote is put in by all the trees so as to get the most popular class by taking in the x. (Breiman; 2001)

3.1.3 Support Vector Machines(SVM)

In its binary form the (SVM) acts as a classifier that divides the classes by introducing a hypeplane between the classes. The classes are divided by maximizing the margin distance between the hyperplane and the points. The nearest point to the hyperplane ac as the support vectors. Strictly referring to our study (SVM) also has the ability to handle multi class classification problem. For multi class problem usually One versus One and One versus All concept comes into play. For the One versus One technique a single classifier is built for every class pair present in the data. However, in the case of One versus All technique various classifiers are built which in turn act in splitting a particular class from all the other classes. Whenever a new observation is accounted for the classifier which has the best function for decision is chosen. More hyperplanes are constructed in multiclass (SVM) (Sun et al.; 2018)

3.1.4 XGBOOST (eXtreme Gradient Boosting)

XGBoost has the ability to implement machine learning algorithms. The frameworks used by XGBoost are gradient boosting framework and parallel tree boosting. The various trees generated by the tree-based models are combined which have low accuracy so as to generate a model that is more efficient.

Within XGBoost there exits parameters that can be optimized the so as to get a good accuracy by the model built. Parameter like learning rate (for preventing the overfitting), maximum depth of a tree so as to increase the complexity of the model, gamma value so as to minimize the loss while creating a new tree etc. Increasing the number of trees produced by the model and lowering the learning rate can make the model run on high computation cost (Zhang and Zhan; 2017). XGBoost has the ability to run fast that too with low memory provision. This feature of XGBoost makes it scalable to a large extent. The scalability of XGBoost is highly dependent on the machine type as well as the optimization of algorithm. (Chen and Guestrin; 2016).

3.1.5 SMOTE

When the proportion of the classes varies a lot, it is considered as imbalanced. Previously while building the models under sampling of the majority class was done usually. With the help of SMOTE the oversampling of the minority class is done. This is done using the minority class samples and further putting in the synthetic examples to it. This technique even helps in better generalization of the decision trees (Chawla et al.; 2002).

4 Implementation

4.1 Flowchart of the process/study

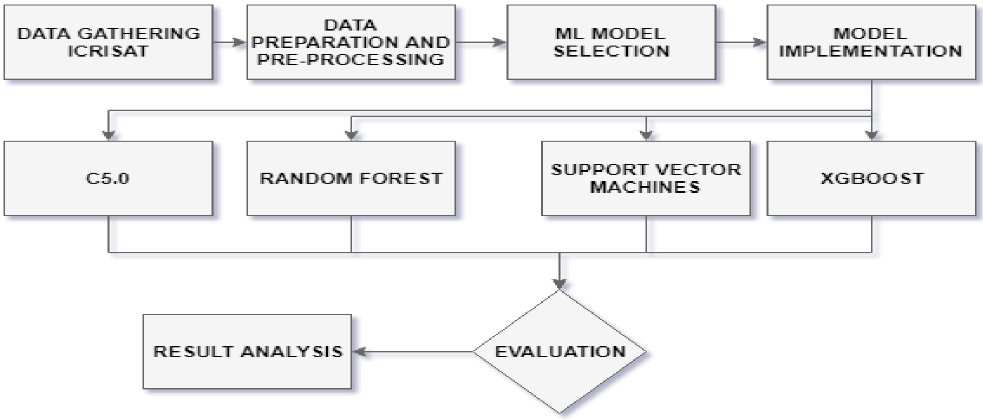

Figure 2: Flow Diagram for the study

The implementation of this study was conducted using the R programming language. R Studio software, as well as Microsoft Excel, was used. R Studio is a powerful tool for statistical analysis.

The data downloaded from the ICRISAT was very untidy data, it not only had missing values but also the levels in the outcome feature (Types_of_soil) were miss-spelled. Therefore, the spellings of the levels before loading in the R studio were manually corrected in Microsoft Excel. After fixing the miss-spelled data, the data was loaded into the R studio for further analysis. The analysis began with installation and loading of the necessary packages required for the analysis. Packages included were (dplyr) for data manipulation, (caret) classification and regression training package, (xgboost) Extreme Gradient Boosting for tree bosting, (corrplot) to plot the correlation matrix of the features, (mice) to impute the missing values in the data, (VIM) for visualizations of the missing values, (DMwR) data mining with R which provides function used while performing data mining, (factoextra), (FactoMiner) for the analysis of statistical findings and (prroc) to plot the precision-recall curves.

Initially, the data that was read into R studio was checked for missing values, the missing values were check using the built-in function of R anyNA and is.na. With the use of mice library functions md.pair and md.patters the missing values were further explored.

It was analyzed that the data were missing at random and had to be imputed. (Aggr) function from the VIM library provided the understanding of the proportion of missing data. Mice function was used from the mice library to impute the missing values. Once the missing values were imputed the (complete) function was used to create the complete imputed data. As the mice took a long time for computation the data was written into a new CSV file and imported again. Again, the missing values were checked and verified. The next step was to give factors to the outcome feature the soil type. Some

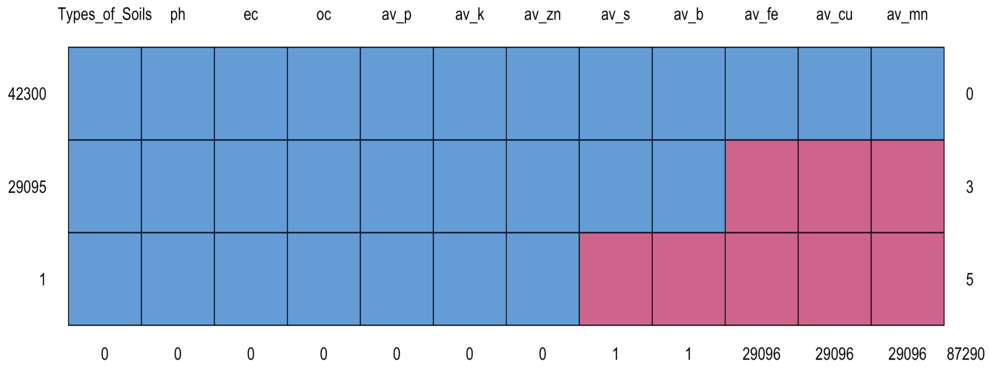

Figure 3: Pattern of missing values

of the features had a different datatype from the others so all the features were given numeric datatype expect the categorical feature. With the data almost prepared and preprocessed, a subset of numerical features of data prepared so as to normalize the data. Since the problem at hand was classification problem (multiclass), class imbalance [Fig.4] was checked for the different soil types.

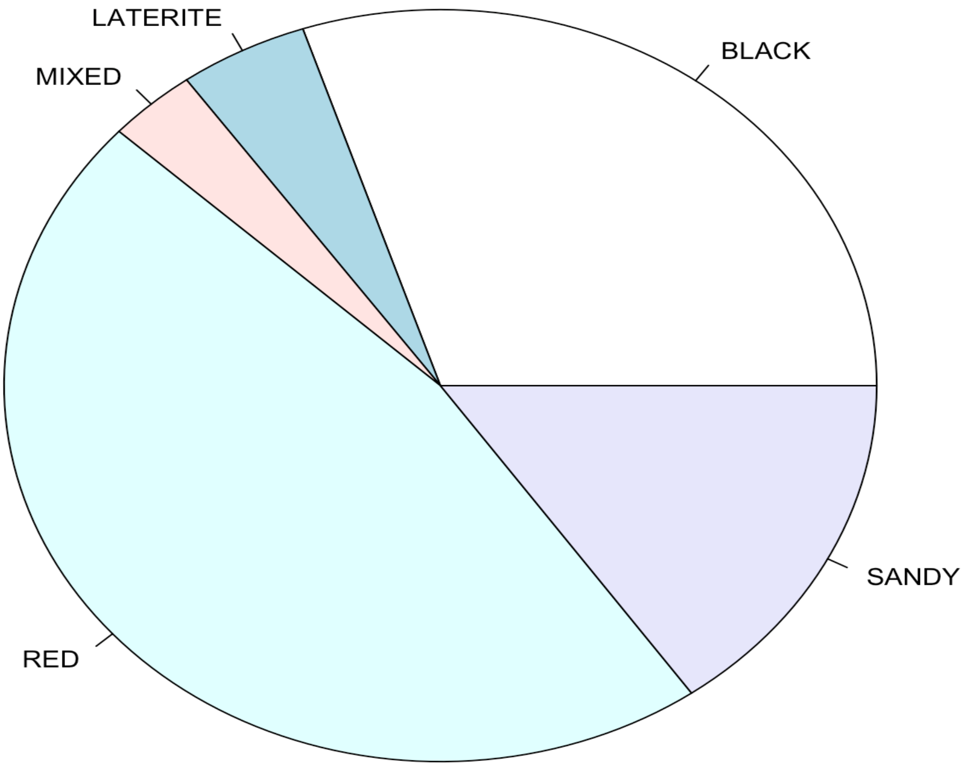

Figure 4: Class Imbalance in Soil Types

Correlation test [Fig.5] to find out the relationship between the features was also done. Test for homogenity of variance was done, Bartlett's test using the bartlett.test function in R. The value for p in the Bartlett's test suggested that the variance of the components was not equal. The correlation plot which was made using the corrplot package. Principal Component Analysis (PCA) was conducted on the normalized data. The PCA plot [Fig.6] suggested that the data under study was not linearly separable.

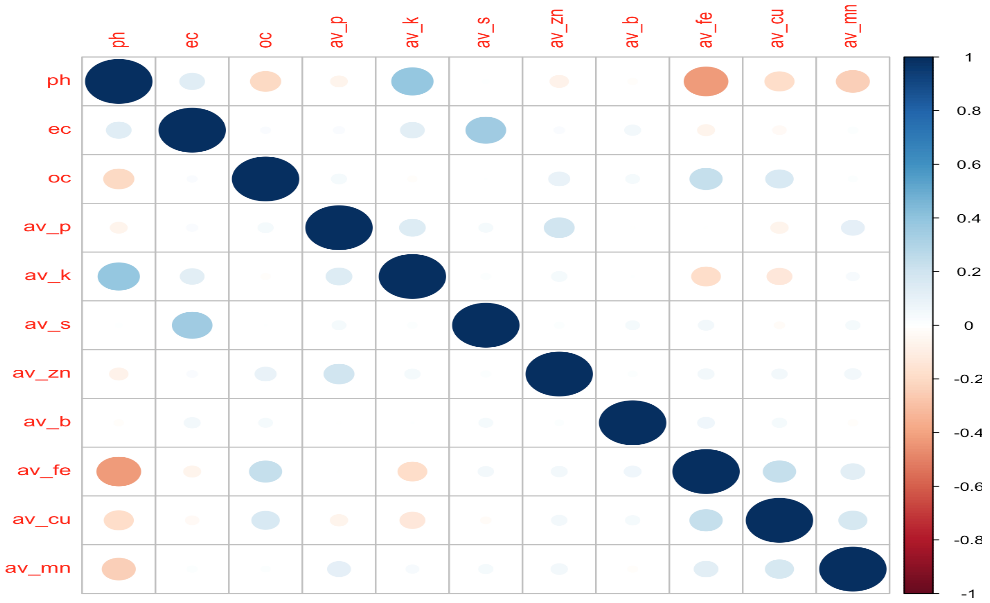

Figure 5: Correlation amongst indicators

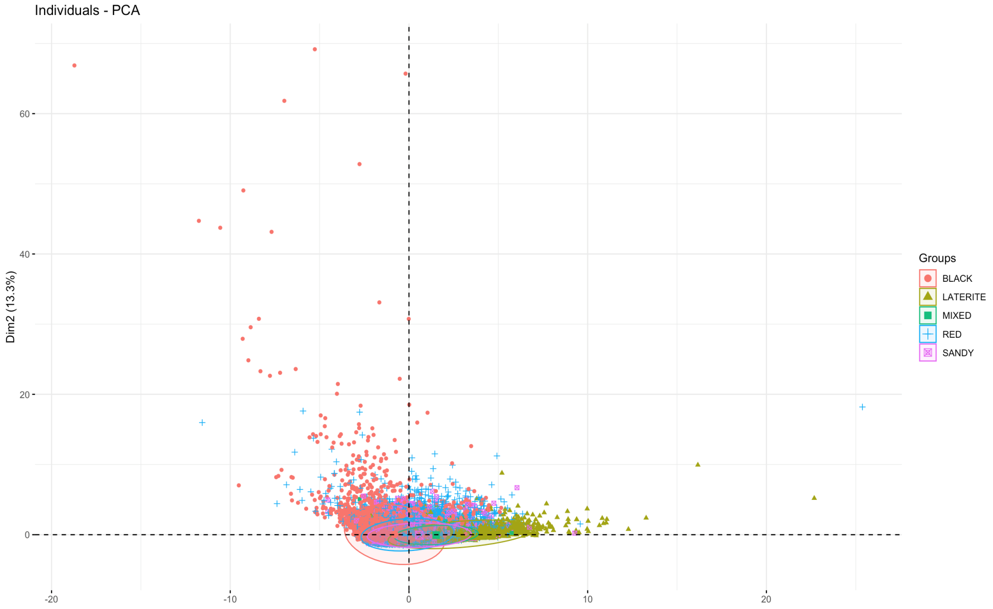

Figure 6: PCA Plot

There were no clusters found in the data. Though the Scree plot [Fig.7] depicted a slightly higher variance within the first principal component, 90% of the variance in total was explained by 9 out of 10 components.

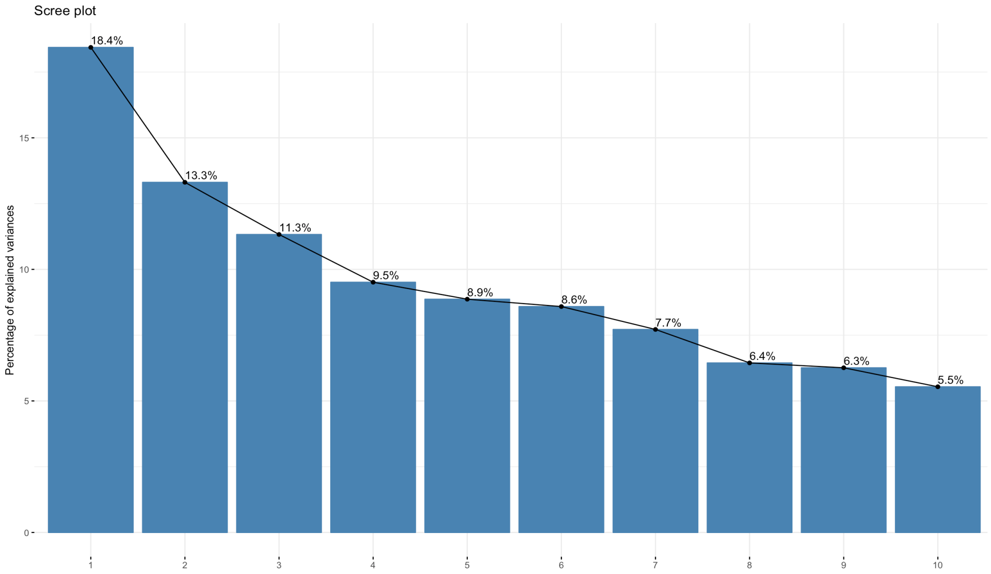

Figure 7: Scree Plot

As discussed earlier, to fix the class imbalance problem the SMOTE technique was used within the caret package (Chawla et al.; 2002). For the application of SMOTE technique as well as repeated cross-validation of the training data, a specific control parameter was defined again using the (caret) package. After the exploration of soil health data, machine learning techniques shortlisted were to be built. The building of the different models implemented was carried out by the using the train function within the (caret) pack The approach while implementation of the models for the study was to apply simpler models initially and then take up more complex model. C5.0, Random Forest (RF), Support Vector Machines (SVM) and XGBOOST were applied one after the other.

To start with training the models, the data was split into training and testing datasets (hold out simple). The data was divided into 80% and 20% ratio after the preprocessing. The same holdout was used for the Principal Component Analysis (PCA). It is to be noted that numeric data works within the PCA.

Once the holdout was done the data was normalized so that it could be fed into the PCA and SVM (SVM). A control parameter was set up using the traincontrol function from the caret package.

Initially, the C5.0 algorithm was applied using the (C50) package in the R studio. The C5.0 model was trained using the controlled parameters (Cross-validation and SMOTE). Similarly, the next tree-based model, Random forest was trained. After Random Forest (RF) model the Support Vector Machine (SVM) was trained using different kernels Radial Basis and Polynomial. Within the (SVM) the Linear kernel was not implemented as the data observation was not linearly separable. For efficient results in the (SVM) model, the parameter in the (SVM) model was optimized using the tunelength function from the caret package. Tune function gives the most optimal features with which the model had been trained. Finally, the eXtreme Gradient Boosting (XGBOOST) model was implemented using the (xgboost) package. For XGBOOST model a new control parameter was set up. For the parameter tuning of XGBOOST, the grid search technique was used so as to get the optimal parameters for the model when under training. A tuning grid was defined before running the model using the expand. grid function. This grid search helps in parameter optimization by setting up values for gamma, the depth of tree etc.

Ultimately, the prediction of the train data was tested against the test dataset using the predict function from (caret) package. Confusion matrix was obtained using the prediction done by the model.

Accuracy and kappa values were observed for each model. Execution times of the different models were also noted.

5 Evaluation

Depending upon the problem at hand, the prediction of model and how the performance is interpreted of the model usually varies. While classifying the distinct soil types confusion/classification matrix was considered as it provides with the predicted classification against the original classification. With the help of confusion matrix, the classification rate of the models and kappa values were noted. Individual classification of the different classes was also carefully analyzed. For the evaluation of models that were implemented for this study, classification accuracy and Cohen's kappa were focused upon.

- Accuracy – Accuracy is the correctly classification done by the model of the data. The accuracy can be calculated from the confusion matrix. For multiclass problems the accuracy is calculated using the true classification rate.

  True classification rate is calculating the true classification for a particular class and dividing it by the summation of false classification for the number of classes. It can also be described as:

  True Classification rate$_i$ = $\frac{TrueClassification_i}{\sum_n^{i=1} FalseClassification}$

where n = number of the classes.

The overall accuracy is the summation of the True classifications in the confusion matrix upon the all the cases in the confusion matrix. i.e.

Overall Accuracy$_i$ = $\frac{\sum_{i=1}^{n} TrueClassification_i}{TotalCases}$

where n = number of classes

i is the class number. Simon (2018)

- Kappa – The range of kappa value varies between 0 to 1. The higher value of kappa the suggest that how good the classification for true values predicted by the model. In the case of multiclass classification problems, the accuracy described the models is often not a complete metric to judge the model. The Cohen's Kappa is an efficient metric to rely on when considering the imbalance class problem within the multiclass classification. (Simon; 2018)(; kappa)

- Time of execution – Since machine learning models had to be trained for the for the specific algorithms, the time taken to train the model was also noted. As the training of the models is a tedious task, over the more, application of cross validation of where the data is trained multiple times, some algorithms took more time than the others. Registering the system time just before training the model and using the system time function gave the total time took by the model to train.

6 Result

The various algorithms implemented to classify the soil types preformed differently and took different computation time. The table below compares the values for accuracy, kappa and computation time took by each model. Judging by the table shown below the best accuracy and kappa values were for the XGBOOST model (66.1%), followed by the Random Forest (RF) (65.1%), C5.0 (63.1%) and the Support Vector Machines (SVM) respectively.

SOIL TYPE CLASSIFICATION			
MODEL	**ACCURACY**	**KAPPA**	**TIME EXE**
C5.0	**63.1%**	**0.4504**	**6.43**
RANDOM FOREST (RF)	**65.1%**	**0.4667**	**3.89**
SVM (RADIAL)	**60.3%**	**0.4112**	**7.85**
SVM (POLY)	**56.8%**	**0.3625**	**41.9**
XGBOOST	**66.1%**	**0.4619**	**182.66**

Although XGBOOST outperformed the other algorithms implemented, the computation time for the XGBOOST was the highest. Radial Basis and Polynomial the (SVM) were kernels implemented which took relatively less time. The computation time (in minutes) of the (C5.0) was similar to the (SVM) models whereas, the Random Forest (RF) had the lowest computation time.

7 Conclusion and Future work

It was concluded after analyzing the results that for the classification of soils of Karnataka, machine learning can be an efficient approach. Amongst the model applied Random Forest (RF) was analyzed to be the best. Though, the selection of model depends upon the business objective. Inclusion of better systems with higher computation power may improve the time of XGBOOST model, although it can be expensive and resource consuming. For academic purposes like this study Random forest (RF) is quite competent.

In the future the including the data about biological indicators as described by (USDA) should be included. More classes soils for the Karnataka region should be considered for future analysis. It was understood that since computation time of complex algorithm XGBOOST is higher, simpler algorithm with better parameters and optimal settings may provide even better results. Since pH feature in the dataset showed importance using the Random Forest model, it can be considered as a dependant parameter for the classification of the soil based on salinity levels of the soils in the future.

8 Acknowledgement

I would like to thank my supervisor Mr. Noel Cosgrave who's feedback and guidance the throughout this research study helped me immensely.I thank the National College of Ireland library for providing support and resources needed for this research study. I would also like to express my gratitude towards the School of computing, National College of Ireland.

References

(kappa).
URL: *http://thedatascientist.com/performance-measures-cohens-kappa-statistic/*

AGBOOLA, O. O., AKINSOJI, A. and OYEDEJI, S. (2017). Changes in nutrient contents of soil across different land uses in a forest reserve, *Notulae Scientia Biologicae* **9**(3): 414–421.

Bhargavi, P. and Jyothi, S. (2009). Applying naive bayes data mining technique for classification of agricultural land soils, *International journal of computer science and network security* **9**(8): 117–122.

Bhattacharya, B. and Solomatine, D. P. (2006). Machine learning in soil classification, *Neural networks* **19**(2): 186–195.

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A. and Edwards Jr, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes, *Geoderma* **239**: 68–83.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, pp. 785–794.

Coopersmith, E. J., Minsker, B. S., Wenzel, C. E. and Gilmore, B. J. (2014). Machine learning assessments of soil drying for agricultural planning, *Computers and electronics in agriculture* **104**: 93–104.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data, *Communications of the ACM* **39**(11): 27–34.

Gambill, D. R., Wall, W. A., Fulton, A. J. and Howard, H. R. (2016). Predicting uscs soil classification from soil property variables using random forest, *Journal of Terramechanics* **65**: 85–92.

Harlianto, P. A., Adji, T. B. and Setiawan, N. A. (2017). Comparison of machine learning algorithms for soil type classification, *Science and Technology-Computer (ICST), 2017 3rd International Conference on*, IEEE, pp. 7–10.

Hartemink, A. E. (2015). The use of soil classification in journal papers between 1975 and 2014, *Geoderma Regional* **5**: 127–139.

Hemageetha, N. (2016). A survey on application of data mining techniques to analyze the soil for agricultural purpose, *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, IEEE, pp. 3112–3117.

Karlen, D. L., Ditzler, C. A. and Andrews, S. S. (2003). Soil quality: why and how?, *Geoderma* **114**(3-4): 145–156.

Karlen, D. L. and Rice, C. W. (2015). Soil degradation: Will humankind ever learn?

Kovačević, M., Bajat, B. and Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines, *Geoderma* **154**(3-4): 340–347.

Kumar, A. and Kannathasan, N. (2011). A survey on data mining and pattern recognition techniques for soil data mining, *IJCSI International Journal of Computer Science Issues* **8**(3).

Kumar, R. S. and Hashim, U. (2017). Characteristics of suicidal attempts among farmers in rural south india, *Industrial psychiatry journal* **26**(1): 28.

Lenka, N. K., Jaiswal, S., Thakur, J., Lenka, S., Mandal, A., Dwivedi, A., Lakaria, B., Biswas, A., Shukla, A. and Yashona, D. (2017). Soil degradation effect on soil productivity, carbon pools and soil enzyme activity, *CURRENT SCIENCE* **112**(12): 2434.

Li, Y., Liang, S., Zhao, Y., Li, W. and Wang, Y. (2017). Machine learning for the prediction of l. chinensis carbon, nitrogen and phosphorus contents and understanding of mechanisms underlying grassland degradation, *Journal of environmental management* **192**: 116–123.

Mucherino, A., Papajorgji, P. and Pardalos, P. M. (2009). A survey of data mining techniques applied to agriculture, *Operational Research* **9**(2): 121–140.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA.
URL: *http://www.rstudio.com/*

Shastry, K. A., Sanjay, H. and Kavya, H. (2014). A novel data mining approach for soil classification, *Computer Science & Education (ICCSE), 2014 9th International Conference on*, IEEE, pp. 93–98.

Simon, C. R. (2018). Data mining, nci, National College of Ireland.

Sirsat, M., Cernadas, E., Fernández-Delgado, M. and Khan, R. (2017). Classification of agricultural soil parameters in india, *Computers and electronics in agriculture* **135**: 269–279.

Stevenson, B., McNeill, S. and Hewitt, A. (2015). Characterising soil quality clusters in relation to land use and soil order in new zealand: An application of the phenoform concept, *Geoderma* **239**: 135–142.

Sun, Y., Feng, X. and Yang, L. (2018). Predicting tunnel squeezing using multiclass support vector machines, *Advances in Civil Engineering* **2018**.

Zhang, L. and Zhan, C. (2017). Machine learning in rock facies classification: An application of xgboost, *International Geophysical Conference, Qingdao, China, 17-20 April 2017*, Society of Exploration Geophysicists and Chinese Petroleum Society, pp. 1371–1374.