Soil-Crops Suitability Analysis Modeling (SCSAM) Using Machine Learning Approach

Dan Ayana
HiLCoE, Computer Science Programme, Ethiopia
Engineering Corporation of Oromia
danpro.it@gmail.com

Tibebe Beshah
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

For a country whose economy is strongly dependent on agriculture, conducting land and environmental study and assessing the land's physical nature, chemical characteristics, and topographic formation are critical for improvement. While we are in the era of AI (Artificial Intelligence) where most of the industries are actively using ML (Machine Learning)/Data Mining (DM), Soil Suitability Classification (SSC) remains one of the tedious manually crafted and resource-intensive activities in the agricultural sector.

Thus, to enhance these processes, following Design Science Research Methodology and Cross-Industry Standard Process for Data Mining (CRISP-DM) design and development frameworks, a study was made experimenting with classification algorithms of Rapidminer, WEKA, and Python DM/ML libraries.

In the Rapidminer tool, experiments were done on various ML algorithms using overall crops, Six Crops, and Crop-Based datasets. Prediction accuracies of 94.7%, 96%, and 99% were achieved using Naïve Bayes, Generalized Linear Model, and Gradient Boosted Tree algorithms, respectively. Similarly, in the WEKA tool, prediction accuracy of 96% was found on the J48 tree algorithm. In Python, using a crop-based dataset, different experiments were done with the default and applying the Oversampling technique to minority classes. The algorithms' generalization capability was also evaluated with 10-fold cross-validation and other evaluation matrices.

From the experiments, selecting the top-scored algorithms (decision tree, random forest, extra tree, Multilayer Perception (MLP), and K-Nearest Neighbor (KNN)) and settings, the Soil-Crop Suitability Prediction Prototype was created for six crops and evaluated with user data. In the evaluation, professionals confirmed that the predicted result by the prototype was 88%-97% matching with manually classified crop suitability data. Likewise, in their feedbacks, professionals indicated that the SCSAM (Soil-Crops Suitability Analysis Modeling) artifact is an opportunity for soil suitability analysis to minimize experts' efforts and resources.

Keywords: Design Science Research; CRISP/DM; Soil Suitability; Classification

1. Introduction

For a country whose economy strongly depends on agriculture, land and water resource utilization is very crucial for the benefits of its citizens [1].

Over the past two to three decades, the rapid population growth, frequent technological advances, industrial development, and agricultural expansion have led to inequality and imbalances in the supply of adequate water, energy, and other basic resources to inhabitants.

From environmental resources, land is used as the basis for agriculture, industry, settlement, and other socio-economic development activities.

Knowing the soil's physical nature, chemical characteristics, and topographic formation helps to understand its potential and improve productivity. Land and environmental study helps to ensure long-term productivity and sustainable development through the appropriate use of land resources [2]. Thus, in this study, Land Suitability Assessment (LSA) is used to find where suitable land is and to identify its potential as to what kind of uses it is best suited [3].

Soil suitability analysis and assessment are usually done through the determination of the most important factors such as availability of nutrient elements, pH, cation exchange capacity, oxidation-

reduction conditions, water holding capacity, and the like [4] that have a direct impact on soil properties. In soil suitability assessment, professionals are expected to apply Soil Science knowledge in the collection, mapping, analysis, testing, and interpretation process.

Although results often depend on several factors, the evaluation and interpretation processes are not an effortless task [5]. It requires multidisciplinary and cross-sectoral approach to contemplate the biophysical, socio-economic, and environmental characteristics. As a result, it consumes too much time and resources and it is highly vulnerable to personal opinions.

Yet, there are Artificial intelligence (AI) and Machine Learning (ML) technology-based applications on the market to automatically learn from data and improve from experience without being explicitly coded [6]. Banks, government offices, service-giving sectors, insurances, etc. apply those technologies to their problems and have begun to reap the benefits. However, land-use study and soil analysis continued as the most time taking and expensive activities in the agricultural sector.

In Ethiopia, especially in Oromia and surrounding regions, the Regional State Land Administration Bureaus collaborating with different NGOs funded many Integrated Land-Use Plan (ILUP) studies for major rivers and sub-basin lands [7]. However, project owners and professionals have not still started recognizing ML to estimate and forecast similar ILUP projects. Instead, experts have followed common steps and procedures for similar land-use study projects that often lead to delays and extra budgets.

Thus, the objective of this paper is to investigate soil suitability analysis and build an automated knowledge-based Soil-Crop Suitability Classification Prototype that can enhance suitability analysis and the interpretation processes of land-use study.

2. Related Work

Land evaluation for crop production has provided useful information to stakeholders to improve crop yield and soil management.

The work in [9] used a Parallel Random Forest (PRF) classifier for an ML prediction model for the Sorghum crop. The authors set-up experiments using PRF, Linear Regression (LR), Linear Discriminate Analysis (LDA), KNN, Gaussian Naïve Bayesian (GNB), and Support Vector Machine (SVM). 10-fold cross-validation was used to evaluate performance. According to their findings, PRF has a better accuracy result of 96%.

The work in [10] used ML techniques to predict yield by analyzing soil datasets. The crop yield prediction problem was formalized as a classification rule where Naïve Bayes and KNN methods were used.

Similar research was conducted on crop suitability classification to improve the LIMEX system [11]. The study used a hybrid approach (c4.5 tree) considering 11 attributes.

A newly proposed ensemble classifier generation technique called RotBoost, which is a combination of Rotation Forest and AdaBoost, is used for land suitability classification in [8]. The researchers found that RotBoost generates ensemble classifiers with significantly higher prediction accuracy.

So far, this research was focused on the prediction of the suitability of the soil to the required crops that are grown in most areas of the country and has complete features. The main difference or feature of this study is that, unlike [9] and [10], we have included most of the determinant chemical features and used more than 22,000 soil sample data collected from the different agro-ecological zones of Ethiopia.

3. The Proposed Solution

In order to automate and enhance the soil suitability analysis work for medium to large scale land use planning, we have made a systematic investigation using the design science research methodology [12] and Cross-industry standard process for data mining (CRISP-DM) framework. Afterward, an attempt was made to develop ML-based soil suitability prediction model.

Our proposed method to predict Soil-Crop suitability is through the implementation of supervised classification machine learning algorithms using Decision tree, Random forest, Extra tree, SVM, and KNN classifiers.

After understanding the business process, essentials, and challenges, important data were collected from Oromia Water Works Design &

Supervision Enterprise (OWWDSE), Land-Use Planning Study and Environmental Protection Subprocess Division. The collected raw data are the soil survey data based on a 1:50,000 mapping scale for over 41 million hectares of land representing more than 50,000 soil mapping units from Oromia, Somali, Gambela, and Afar regions.

For a machine to learn the problem and make predictions effectively, data preprocessing and feature engineering are the fundamental processes. Therefore, to transform raw data into important features, a hybrid feature selection technique (Knowledge Expert Feature Selection and Traditional Attribute Selection Method) was implemented. First, to reduce and keep only very determinant attributes, Knowledge Expert Feature Selection method [13] based on expert consultation was used. Then those identified features are further preprocessed using traditional attribute selection methods found in each ML tool.

Discussing with experts, the possible soil-crop suitability attributes from physical, environmental, and chemical features are identified. Table 1 shows suitability determinant chemical features that are considered in the soil suitability analysis for the required crop.

Table 1: Soil Suitability Determinant Chemical Features

No.	*Chemical properties*	*Code*
1.	Cation Exchange Capacity	CEC
2.	Base saturation percentage	BS
3.	Electrical conductivity	EC
4.	Exchangeable Calcium	Ca
5.	Exchangeable Magnesium	Mg
6.	Exchangeable Potassium	K
7.	Exchangeable Sodium	Na
8.	Exchangeable sodium percentage	ESP
9.	Total Nitrogen	TN
10.	Organic Matter	OM
11.	Organic Carbon	OC
12.	Soil Reaction	PH
13.	Available Phosphorus	AvP
14.	Carbon to Nitrogen Ratio	C: N
15.	Exchangeable Calcium to Exchangeable magnesium ratio	Ca: Mg
16.	Calcium carbonate	Caco3
17.	Potassium Chloride	KCL

Along with that, non-relevant attributes and a large number of missing values were also discarded from the data through the Listwise or case deletion method [14]. In the meantime, the attribute values measured in different parameters were uniformly assigned with similar values. Inconsistently labeled attribute values were replaced with consistent labeling. Besides, the data details that differ significantly from other observations are handled in the standard ways [15].

For a machine-learning algorithm to improve its prediction performance and to reduce overfitting, selecting the best features among all the data is very essential [16]. Therefore, from available attribute selection methods, in the RapidMiner, the Remove Correlated method was used to clean above 0.9 correlated attributes and found a total of 9 nominal and 19 numeric attributes. Similarly, in WEKA, "Info Gain attribute Eval" filter method was used to rank the attributes and to remove those with score below the threshold. In the process, 9 nominal and 20 numeric attributes were found. In Python using Sklearn.feature_selection, the attribute Variance Thresholds are evaluated. The possible determinant number of attributes "K" was identified and using the SelectKBest method, only very important features (K-number of features) were selected.

After applying data preprocessing and preparation techniques for the collected ILUP data, we got three dataset options with 22,532 rows of soil-related processed datasets. Then for further processing and experiments, the generated datasets were transformed into the model experimentation phase to train and test the ML algorithms.

4. Experiments and Results

Since the right model can only be identified through experimentation, different classification algorithms and settings in the RapidMiner, WEKA, and Python were tested.

4.1 Experiment using RapidMiner DM/ML

Cleaning and preprocessing the raw data using data cleansing feature and removing 0.9 correlated attributes, experiments were made using preprocessed dataset. In the experiment, different classification algorithms, prediction accuracy, and the performance of a learning model were tested

using ILUP soil datasets. Among the experimented five algorithms using OverAllCrops, the best accuracy score of 94.7% was obtained in the Deep learning model. In the second dataset option using the Six Crops dataset, the best score was found on Generalized Linear Model and Deep Learning algorithms.

For crop-based experiments, applying auto and dummy encoding to the categorical attributes, two types of experiments were made for Maize and Onion crops. From the tested algorithm and settings, Naïve Bayes, Generalized Linear, and Gradient Boosted Trees algorithms show a better score in the dummy encoded method. Table 2 shows the experimented algorithms result in terms of accuracy.

Table 2: RapidMiner Experiment Summary

Algorithm	*OverAll Crop*	*SixCrop*	*Maize Auto*	*Onion Auto*	*Maize Dummy*	*Onion Dummy*
Naive Bayes	93.9%	92%	84%	85.48 %	90.2%	90.2%
Generalized Linear Model	-	96%	97%	97.2%	98.2%	99.02%
Logistic Regression	-	94%	98%	98.90 %	97.80%	98.90%
Fast Large Margin	-	Error	72%	85.1%	95%	95.59%
Deep Learning	94.7%	95%	98%	99.05 %	98.2%	98.99%
Decision Tree	11.4%	40%	87.00 %	93.72 %	79.3%	68.15%
Random Forest	18.5%	36.00%	72.00 %	85.40 %	72%	65.56%
Gradient Boosted Trees	-	Error	99%	99.1%	97%	99.18%

4.2 Experiment using WEKA

The raw data were preprocessed using the “Info Gain attribute Eval” filtering method and 9 nominal and 20 numeric attributes were prepared. Then, using available classification algorithms or learning schemes, different soil suitability experiments were made with the preprocessed datasets. To find out the most suitable algorithm, multiclass supporting classification algorithms were tested. Meanwhile, to rank their performance, Cross-validation, K-fold, and 70:30 training test splitting techniques were used.

Table 3: Weka based Classification Accuracy Result

Algorithm	*Validation*	*Overall Crop*	*Six-crops*	*Crop based*					
				Maize	*Onion*	*Rice*	*Sesame*	*Sugarcane*	*Tomato*
Naive Bayes Updatable Auto encoding	10-fold	73.89%		71.44%		-	-	-	-
	70:30:00	73.22%	-	71.88%	-	-	-	-	-
Naive Bayes Updatable Nominal to Binary encoding	10-fold	-	-	-	-	-	-	-	-
	70:30:00	-	-	-	-	-	-	-	-
Decision Tree	10-fold	96.27%	97.00%	98.14%	98.72%	95.41%	97.31%	99.81%	99.65%
	70:30:00	95.86%	96.64%	97.97%	98.73%	98.73%	97.09%	99.69%	99.69%
J48	10-fold	97.02%	97.46%	98.69%	98.86%	98.86%	98.21%	99.47%	99.49%
	70:30:00	96.60%	97.10%	98.76%	98.77%	98.77%	98.00%	99.33%	99.36%
Random Forest Classification	10-fold	96.36%	97.04%	98.44%	98.75%	98.75%	97.29%	99.87%	99.69%
	70:30:00	96.01%	96.49%	98.17%	98.55%	98.55%	96.98%	99.81%	99.67%
GaussianNB	10-fold	90.28%	-	-	-	-	-	-	-
	70:30:00	86.57%	-	-	-	-	-	-	-
K- Nearest Neighbors Algorithm	10-fold	69.45%	-	84.06%	-	-	-	-	-
	70:30:00	-	-	-	-	-	-	-	-

Using Auto and nominal to binary encoding methods, two types of experiments were made with the OverAllCrop dataset. The experiment result indicated that auto encoder has a better score than Nominal to binary encoding. From the tested algorithms, an accuracy score of >95% was found on Trees.J48, Decision Tree Classifier, Naïve Bayes, and Random Forest classifier in both 10-fold cross-validation and 70:30 splitting methods. Similarly, an experiment was made with SixCrop and the crop-based dataset and a better score was found with Trees.J48, Decision Tree Classifier, and Random Forest algorithms in the K-folding cross-validation and training test split techniques. The summary results are listed in Table 3.

4.3 Experiment using Python with its ML/DM Library

In contrast to the two data mining tools, to preprocess raw data, as a programming language, Python's ML libraries need to be coded. Therefore, further data processing and suitability experiment were made using a crop-based dataset. In the preprocessing, outlier values were handled using interquartile and replace methods. Using the round() function, numerical values were uniformly set to have only 3 digits. Empty and inconsistent value fixing code was written.

Since all ML models are mathematical models, the categorical values in the dataset need to be converted into numeric values. Alhough Python provides categorical to numeric encoding techniques [17], OneHotEncoding, oneLabelEncoder and Binary Encoding were not able to assign similar values in training test and user data provision cases. Therefore, to handle this issue, we wrote a categorical data handling code and were able to replace an identical numeric value with similar objects in all cases.

To select the best determinant features from the dataset, using the Sklearn Variance Threshold class parameter (Threshold = .51), the maximum number of determinant attributes K = 27 were identified. To identify the K number of attributes in each crop, SelectBest 'K' was implemented [18]. In each crop, the number of best determinant features were identified as follows:

- Maize [SMU, Major_Soil, Texture, Drainage, Depth, Slope, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Onion [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Rice [SMU, Major_Soil, Texture, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Sesame [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM.3, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Sugarcane [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]
- Tomato [SMU, Major_Soil, Drainage, Depth, Slope, Temperature, Elevation, PH_Water, PH_Kcl, EC, Na, K, Ca, Mg, CEC, BS, TN, OC, OM, C_N, P2O5, CaCO3, ESP, Ca_Mg, K_Mg, Ca_Mg_k, SUM]

Along with that, to deal with imbalanced data issues of the target attribute, from oversampling techniques Synthetic Minority Oversampling Technique (SMOTE) and Random Over Sampler objects were used to fit the training set.

First using the Maize crop dataset, a prediction experiment was made for ten ML algorithms (Decision Tree, Random Forest, Extra Trees, MLP Classifier, K-Neighbors, Logistic Regression, Gaussian NB, SVM, Xgboost, and AdaBoost). Then among these, the algorithms which return the highest evaluation scores were used to test for the remaining crops. The models' performance was evaluated using the K-fold Cross-validation and Training test 70:30 splitting technique.

From the experimented ML algorithms and settings on Python, the best-suited algorithms for the suitability classification model were identified through the implementation of the Oversampling technique to minority classes.

Table 4: Python Crop-based Algorithms' Prediction Accuracy and Performance Summary

Algorithm	*Crop*	*Sampling Strategy*	*Accuracy*	*Cross-validation*
Decision Tree	Maize	SMOTE "auto"	97.38%	94.31%
	Onion	SMOTE "auto"	99.14%	96.77%
	Rice	SMOTE "auto"	97.06%	95.62%
	Sesame	SMOTE "auto"	98.18%	94.88%
	Sugarcane	SMOTE "auto"	100.00%	98.03%
	Tomato	Random Oversampling "all"	99.59%	98.60%
Random Forest	Maize	Random Oversampling "all"	97.28%	94.81%
	Onion	SMOTE "auto"	98.71%	97.07%
	Rice	SMOTE "auto"	96.88%	95.26%
	Sesame	SMOTE "auto"	97.87%	94.38%
	Sugarcane	Random Oversampling "all"	100.00%	98.28%
	Tomato	SMOTE "auto"	99.82%	98.31%
Extra Tree	Maize	Random Oversampling "all"	97.26%	94.31%
	Onion	Random Oversampling "all"	98.46%	96.59%
	Rice	Random Oversampling "all"	96.89%	94.91%
	Sesame	SMOTE "auto"	98.18%	94.62%
	Sugarcane	Random Oversampling "all"	100.00%	97.18%
	Tomato	Random Oversampling "all"	99.82%	97.72%
Multi-layer Perceptron	Maize	SMOTE 'auto'	96.72%	93.05%
	Onion	SMOTE 'auto'	99.23%	96.63%
	Rice	SMOTE 'auto'	96.48%	94.86%
	Sesame	Random Oversampling "all"	97.46%	94.26%
	Sugarcane	Random Oversampling "all"	99.75%	97.00%
	Tomato	Random Oversampling "all"	99.72%	97.02%
K-Nearest Neighbor	Maize	SMOTE 'auto'	96.45%	92.88%
	Onion	Random Oversampling "all"	96.46%	96.33%
	Rice	Random Oversampling "all"	95.27%	94.37%
	Sesame	SMOTE "auto"	96.12%	95.14%
	Sugarcane	Random Oversampling "all"	99.78%	91.80%
	Tomato	SMOTE "auto"	99.89%	97.54%

5. The Prototype

Initially, the SCSAM deployment plan was to deploy on the Flask web framework. Accordingly, using this popular Python web framework, the Soil-Crop Suitability Prediction Model for Maize crop was developed and tested on the flask framework. However, the bulk sample prediction for the entire soil sample could not be achieved through the framework. It did not allow the user to upload multiple row datasets to the classification model and get the prediction results.

Hence, instead of using flask as the main SCSAM deployment platform, we used a Python GUI library called Tkinter. The deployment process and the Soil suitability Prediction prototype structure is depicted in Figure 1.

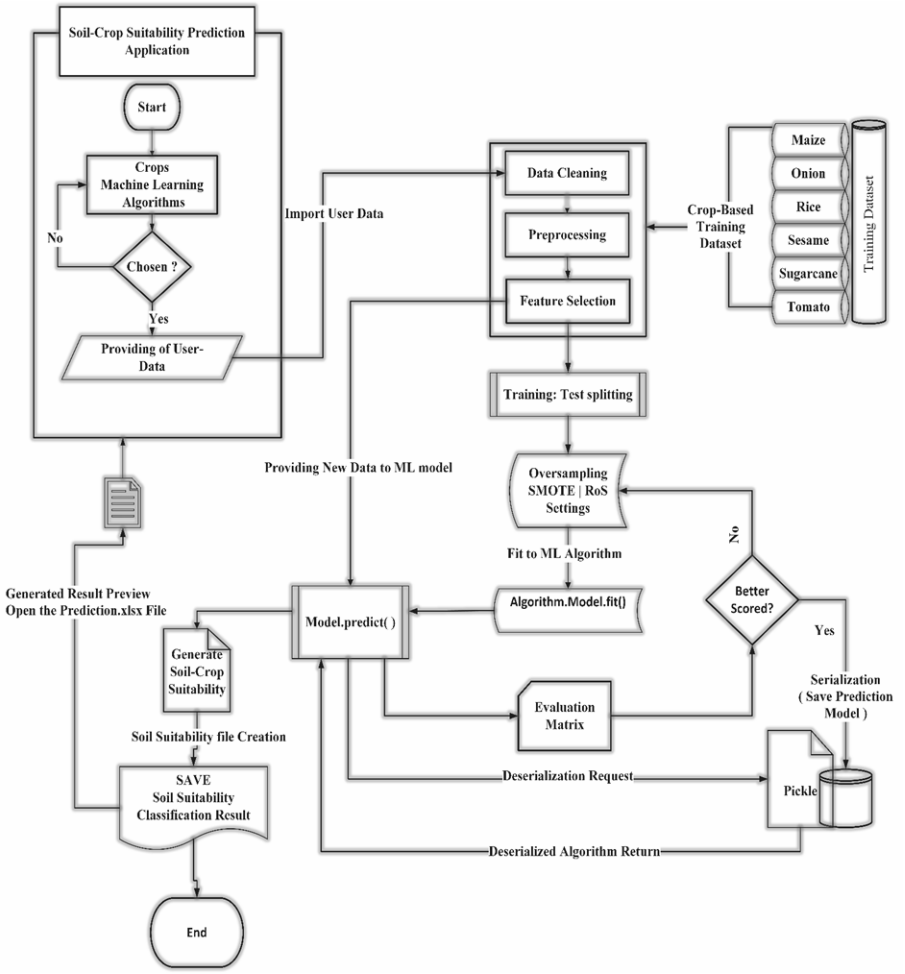

Figure 1: Flowchart of GUI Based Soil-Crop Suitability Analysis Model for Selected Crops

The SCSAM application interface is created with Tk() top-level window (root) object. Figure 2 depicts the soil-crop suitability prediction model prototype user interface.

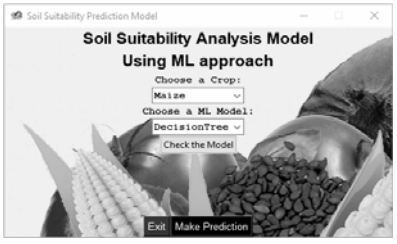

Figure 2: Tkinter Base Soil-Suitability Analysis Model for Selected Crops

- When the SCSAM application starts, it provides an interface to choose the type of crop and prediction algorithm.
- If the user wants to be sure which crops and models are chosen, the "Check the model" button provides that information.
- The "Make prediction" button is used to start the prediction process based on the selected crop and ML algorithm.
- Then it provides a mechanism for users to feed the soil-related dataset into the system.
- To make the user data suitable for ML algorithms, data cleaning and preprocessing works will be done and transform the user data into the prediction process.
- Then in the ultimate stage, the selected ML model makes the suitability prediction and returns the predicted result.

6. Discussion

Model building and prototyping is an approach used to evaluate the solution's functionality, interaction, feasibility, and usability of the product. Once the prototype is developed, it needs to be tested by users.

To test the system's functionality with the user dataset, the prototype was given to senior soil experts, agronomists, and environmentalists in OWWDSE. Twelve user-test evaluation questions were prepared according to [19, 20] and given to them.

In the prototype functionality test, experts tested SCSAM's prediction accuracy with the soil-related data for Maize, Sugarcane, and Tomato crops.

Since the model learning process is initially split from the prediction process, the soil experts highly admired the quick output generation capability for the entire sample. Experts evaluated the prediction performance providing 347 rows of soil dataset. When the prototype returns the prediction result, they compared the output with manually classified crop suitability data. Table 5 shows the total number of correctly classified instances among 347 user-provided data in each crop and the accuracy.

Table 5: The Matching Result of the 347 Rows Containing User Data

Algorithm	*Crop*	*Correctly classified instances*	*Accuracy*
Decision Tree	Maize	320	92.22%
	Onion	334	96.25%
	Rice	332	95.68%
	Sesame	331	95.39%
	Sugarcane	337	97.12%
	Tomato	338	97.41%
Random Forest	Maize	327	94.24%
	Onion	336	96.83%
	Rice	331	95.39%
	Sesame	325	93.66%
	Sugarcane	338	97.41%
	Tomato	336	96.83%
Extra Tree	Maize	324	93.37%
	Onion	333	95.97%
	Rice	328	94.52%
	Sesame	320	92.22%
	Sugarcane	335	96.54%
	Tomato	333	95.97%
Multi-layer Perceptron	Maize	310	89.34%
	Onion	329	94.81%
	Rice	320	92.22%
	Sesame	314	90.49%
	Sugarcane	332	95.68%
	Tomato	332	95.68%
K-nearest neighbor	Maize	309	89.05%
	Onion	331	95.39%
	Rice	315	90.78%
	Sesame	324	93.37%
	Sugarcane	336	96.83%
	Tomato	334	96.25%

Besides, the prototype's performance evaluation questionaries were also supplied to the four senior soil experts and six agronomists found in the enterprise. The questions prepared were designed to obtain the user opinion evaluating the prototype's usability, ease of learning, usefulness, capacity, performance, and output correctness.

The experts ranked that the prototype is lightweight, user-friendly, easy to use, fast and interactive. After evaluating the prototype when we communicate with those professionals, they mentioned that most of the incorrectly classified values in the dataset are tolerable. Even in reality, sometimes, the physical and environmental characteristics of the soil units might have multiple suitability class features and in this kind of situation, experts may decide following their judgments.

7. Conclusion and Future Work

The raw data for this research work was collected from OWWDSE and using diverse machine learning and data mining tools, the collected data were preprocessed. Subsequently, the processed data were used to train the ML algorithms. In the process, various experiments were done on RapidMiner, WEKA, and Python ML tools. From the experiments, the best-suited algorithms and settings were identified in each ML tool. Python and its ML libraries were found to perform better than the others.

Thus, using the best-scored algorithms and settings, SCSAM was developed on the Tkinter platform.

To test prediction capability, the prototype was given to professionals along with 12 evaluation questions. The professionals' evaluation summary and feedback show that the prediction result of the prototype was 88%-97%.

Because of data quality issues, this work could not be able to include all-climate zone data in the ML model. Hence, to make the prototype effective for any large-scale project, future work is required to include the soil-related data of all climate zones. Also, instead of combining both irrigation and rain-fed agriculture data into a single dataset, preparing a separate dataset based on essentials is a future work.

References

[1] OWWDSE, "Hidha-Sombo Small Scale Irrigation Project, Soil and Land Evaluation," Oromia Water Works Design and Supervision Enterprise, Addis Ababa/Finfinne, March 2019.

[2] Hellkamp, A. S., Bay, J. M., Campbell, C. L., Easterling, K. N., Fiscus, D. A., Hess, G. R., McQuaid, B. F., Munster, M. J., Olson, G. L., Peck, S. L., Shaver, S. R., Sidik, and Tooley, M. B., "Assessment of the Condition of Agricultural Lands in Six Mid-Atlantic States," Journal of Environmental Quality, pp. 795-804, 2000.

[3] FAO, "A Framework for Land Evaluation", FAO Soils Bulletin 32, 2nd ed., Vol. 32, Rome, 1976.

[4] FAO, "Soil and Plant Testing and Analysis," in FAO BullETIN 38/1, Rome 13-17, June 1977.

[5] Thom, William O., and Sikora, Frank J, "Soil Testing for Agronomic and Environmental Uses," Vol. 21, No. 4, 2000.

[6] S. Stanford, "The Best Public Datasets for Machine Learning and Data Science," Machine Learning Memoirs Inc.

[7] Oromia Water Works Design and Supervision Enterprise, "Integrated Land Use Plan Study projects," 2001-2019.

[8] M. Mokarram, S. Hamzeh, F. Aminzadeh, and A. Rassoul Zarei, "Using Machine Learning for Land Suitability Classification," West African Journal of Applied Ecology, Vol. 23, No. 1, pp. 63–73, 2015.

[9] Kennedy Senagi, Nicolas Jouandeau, and Peter Kamoni, "Using parallel random forest classifier in predicting land suitability for the crop," Journal of Agricultural Informatics, Vol. 8, 2017.

[10] M. Paul, S. K. Vishwakarma, and A. Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," International Conference on Computational Intelligence and Communication Networks (CICN), pp. 766-771, Jabalpur, 2015.

[11] Ratnmala Bhimanpallewar and M. R. Narasinagrao, "A Machine Learning Approach to Assess Crop Specific Suitability for Small/Marginal Scale Croplands," International Journal of Applied Engineering Research, Vol. 12, pp. 13966-13973, 2017.

[12] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee, "A Design Science Research Methodology for Information Systems Research," Journal of Management Information Systems, Vol. 24, No. 3, pp. 45-77, December 2007.

[13] Vorhies, William, "CRISP-DM – a Standard Methodology to Ensure a Good Outcome," July 2016.

[14] David Camilo Corrales, Emmanuel Lasso, Agapito Ledezma, and Juan Carlos Corrales, "Feature selection for classification tasks: Expert knowledge or traditional methods?", Journal of Intelligent & Fuzzy Systems, Vol. 34, pp. 2825–2835, 2018.

[15] K. J. Anesthesiol, "The prevention and handling of the missing data", Vol. 64(5), May 2013.

[16] Yeshanew Ayalew, "Design and Development of Predictive Model for Potential ATM Card User: Case of Dashen Bank S.C", Unpublished Masters Thesis, January 2019.

[17] Saurav Kaushik, "Introduction to Feature Selection Methods with an Example," Analytics Vidhya, December, 2016.

[18] M. Pathak, "Handling Categorical Data in Python", Datacamp, January, 2020.

[19] Dhandre, Pravin, "Selecting Statistical-based Features in Machine Learning application," Packt hub, https://hub.packtpub.com/selecting-statistical-based-features-in-machine-learning-application/.

[20] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," Vol. 13, pp. 319-340, September 1989.