Penerbit UTHM

© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm

Journal of Soft Computing and Data Mining

Soil Classification Based on Machine Learning for Crop Suggestion

Siti Nor Fatin Liyana Mohd Azmin[1], Nureize Arbaiy[1]*

[1]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2022.03.02.009
Received 12 July 2022; Accepted 20 October 2022; Available online 01 November 2022

Abstract: A system for classifying and arranging information about soil is known as soil classification. This category of soil was formed in response to a need for a simple, consistent, and easy-to-understand way to classify lands, which is especially important for plantation and agricultural decision-making. However, the current method of assessing soil type is time consuming and heavily relied on agricultural experts. The implementation of machine learning is expected for better soil classification to suggest the crop. The three algorithms are tested, which is Random Forest, Naïve Bayes, and k-Nearest Neighbor (k-NN). Classification techniques are being chosen as a data mining task to produce a classify model. Random Forest has the best accuracy (97.23 percent), Naïve Bayes has the second highest accuracy (96.82 percent), and k-Nearest Neighbor (k-NN) has the lowest accuracy (92.92 percent).

Keywords: Soil classification, machine learning, Random Forest, Naïve Bayes, k-Nearest Neighbor (k-NN)

1. Introduction

Soil is a critical factor for successful agriculture. Soil and crop health is supported by a good soil structure, which allows water and air to flow freely to and through the soil profile. Soil holds water for agricultural growth and facilitates the movement of machines and animals. There are several components that are required for plant growth, most of which are obtained by root uptake from the soil [1]. Agriculture is considered as the main and the foremost cultural practices since the ancient period. Because of soil is such a vital facet of agriculture, it's crucial to grasp what kind of soil to use for agricultural cultivation to possess the most effective crop yield. A range of methods, including traditional methods, expertise, and technology can be used to determine the soil type [2]. Understanding soil categorization aids in predicting soil behavior. Soil behavior can be used to forecast how well a soil would perform when growing agricultural crops. Soil classification is a technique used for determining the kind of soil [3][4].

In Peninsular Malaysia, the soil classification system is based on the USDA Soil Taxonomy, which has been updated to account for local characteristics. Soils expressed in lowlands, for example, are not the same as highlands. Temperature, humidity, pH, rainfall, and other factors often differ from each other. The type of trees available and the way they are planted are often determined by the suitability of the soil. Currently, there are no specific data simulations that can make a classification of available soil types. It should be emphasized that the current method of determining soil type takes time and is highly dependent on agricultural experts in the Malaysian Agriculture Office. For example, Institut Agro Usahawan, or iGROW, is a private skills training center that offers short-term and long-term farming workshops. For soil knowledge on which soils are suitable for which plantations and crops, the iGROW organization relies on Malaysian Department of Agriculture.

*Corresponding author: nureize@uthm.edu.my

Determination of soil type is highly dependent on this specialist. Sometimes, experts and iGROW must physically visit the land area to inspect and determine the type of soil. Then the appropriate plant suggestions are given. The key issue here is the reliance on human specialists for soil type knowledge information. While the different soil types found in each area make it difficult for specialists like iGROW to remember them. Soils expressed in lowlands, for example, are not the same as highlands. Temperature, humidity, pH, rainfall, and other factors often differ from each other. The type of trees available and the way they are planted are often determined by the suitability of the soil. Currently, there are no specific data simulations that can classify the available soil types although these data are very important for the agricultural sector.

The results of this soil type classification are very important to be used as a reference for deciding on the appropriate crop type. Therefore, soil classification using machine learning is proposed in research because of its ability to review large volumes of data and identify patterns and trends that may not be clear or take a long time to be identified by humans. For this study, three algorithms were employed to classify the soil for crop suggestion: Naïve Bayes, Random Forest, and k-Nearest-Neighbours (k-NN). The acquired soil data will be used to train and test these three algorithms for soil classification. The results of these three algorithms will be compared to determine which machine learning model is better for crop prediction. Finally, the suggested research will assist the iGROW community in classifying soil to make the best crop recommendations for optimum yields.

There are five sections in this article. The first section describes the research context, namely soil classification with machine learning algorithms. Related work is described in the second section. Research methodology and solution methods are discussed in the third section. The fourth section discusses the results and its analysis. Conclusions are presented in the final section.

2. Related Work

2.1 Soil Classification

Soil is the loose surface material that covers most of the earths [5]. It consists of organic matter and inorganic particles that together support life. The soil structurally supports the agriculturally plants because it is one of the important functions as soil which is as a medium for plant growth [6]. It also is their source of water and nutrients. Soils differ greatly their physical and chemical properties. Processes such as weathering, leaching and microbial activity combine to form a whole range of different soil types, each with distinct characteristics that provide growing benefits and limitations for agricultural production [7]. Identifying the kind of soil needed for a project is paramount to support the healthy growth of plant life.

Soil classification is the separation of soil groups or classes, each of which potentially similar behavior and having similar characteristic which can be geo-referenced and mapped [8]. Soil classification is based on the measurement and description of various characteristics of representative soil profiles that are indicative soil formation processes [9]. However, these characteristics can also be used singly or in combination to create soil datasets or maps the classify soils according to the specific needs of an end user.

It is important to note that the existing process of assessing soil type is time-consuming and heavily reliant on agricultural experts at the Malaysian Agriculture Office. This specialist's ability to determine soil type is crucial. Because iGROW relies primarily on the expertise of experts at the Malaysian Department of Agriculture to acquire soil and type information, decisions have been sluggish. This is because human expert classification can be delayed and unreliable at times. The outcomes of human specialists also can differ from one to the next. As a result, the information's accuracy is inconsistent.

iGROW also identifies soil types based on their experienced trainees, in addition to experts from the Department of Agriculture. Trainer must visit trainee land to assess the soil's condition. Then, based on their previous experience, they determine which crops can be planted. Similar problems occur, with the accuracy and outcomes of these expert trainees being inconsistent and perhaps conflicting at times. To overcome the issues, the application of data mining algorithms to the classification of soil type data is critical. This study should be able to provide precise soil categorization information as well as accuracy in crop forecasting. At the same time, this research could help iGROW trainees apply smart farming in a more efficient manner.

2.2 Classification Method

Machine learning is an Artificial Intelligence (AI) technique and computer science which focuses on the employment of data and algorithms to imitate the way that human learns and gradually improving its accuracy [10]. Machine learning algorithm is an evaluation of the regular algorithm and generally defined as supervised, unsupervised, and reinforcement [11]. There are a few types of the machine learning algorithms, for instance linear regression, decision tree, random forest, naïve bayes, artificial neural network, and more.

Classification is a supervised machine learning approach, within which the algorithm learns from the data input provided thereto, then uses this learning to classify new observations [12]. This method used to determine what data should be recognized to provide a set of sample data that has its classes. It consists of two phases when constructing a classifier. In the training phase, the training set needs to decide how the parameter should be a focus on and combine

the separate type of data into one type of data. As for the testing, the set will be tested by applying to the test data with a known target and comparing the training set with selected data. For additional information in the testing set the result that will be produced on how long it takes to shows the result on each data with the accuracy the data interpret either it has high accuracy or not. Three algorithms in classification method in discussion are Random Forest, Naïve Bayes, and k-Nearest-Neighbor (k-NN).

2.2.1 Random Forest

Random forest is a type of ensemble method used to predict the average of several independent base models was introduced for the purpose in classification and regression method for random forest framework [13]. Ensemble methods stand where it is using multiple learning algorithms to obtain better predictive performance in classification and regression. One of the ensemble methods use in Random Forest is Bagging (Bootstrap Aggregation). Bagging is one of the techniques that perform in a decision tree that used to reduce the variance of a decision tree.
In terms of classification, the Random Forest model is structured similarly to a decision tree model. Because each decision tree is generated using a random subset of the training data, this method was introduced. The predicted Information, Entropy, and Information Gain decision tree can be seen in Equations (1), (2), and (3).

Expected Information

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$
$$Info(D) = I(p_i, n_i) = \frac{p_i}{p_i+n_i}\log_2\left(\frac{p_i}{p_i+n_i}\right) - \frac{n_i}{p_i+n_i}\log_2\left(\frac{n_i}{p_i+n_i}\right) \quad (1)$$

where:
i. $\mathrm{p_i}$ is a true and positive value of data.
ii. $\mathrm{n_i}$ is a false and negative value of data.
iii. $\mathrm{I(p_i, n_i)}$ is a value in data.
iv. $\mathrm{Info(D)}$ is an expected information value from $\mathrm{p_i}$ and $\mathrm{n_i}$ of the data.

Entropy

$$E(A) = \frac{A1(p_i+n_i)}{p_i+n_i} I\big(A1(p_i+n_i)\big) + \cdots + \frac{Ax(p_i+n_i)}{p_i+n_i} I\big(x(p_i+n_i)\big) \quad (2)$$

where:
i. $E(A)$ is a value of Entropy that act as a root of the decision tree.
ii. $A1(p_i + n_i)$ is a value of data from feature or attribute of the data.

Information gain.

$$\mathrm{Gain(A) = Info(D) - Info_A(D)} \quad (3)$$

In random forest, the data sampling utilized in data training and subset is random. The randomization of the random forest approach cannot be applied without the independent decision tree. To use a random forest strategy, first create a decision tree using the expected information, entropy, and information gain equation, known as decision tree (CART). Data can have both true and false random values. Gini Impurity is one of the random forest's measurements. The Gini Impurity equations for the first and second layers of the random forest model are shown in Equations (4) and (5). The first layer of Gini Impurity:

$$\mathrm{I_G(n) = 1 - \sum_{i=1}^{J} (P_i)^2} \quad (4)$$

The second layer of Gini Impurity:

$$\mathrm{I_{second\ layer} = \frac{n_{left}}{n_{parent}} * I_{left\ node} + \frac{n_{right}}{n_{parent}} * I_{right\ node}} \quad (5)$$

Random forest has the advantage of being able to manage missing data values while maintaining the accuracy of the missing data. The random forest uses a combination of random and numerous learning approaches to improve the performance and accuracy of the decision tree method.

2.2.2 Naïve Bayes

Based on Thomas Bayes Theorem, which was introduced in 1702, Naïve Bayes is a probabilistic machine learning technique that may be used for a range of classification applications [14]. The Bayes theorem is based on the premise that a feature or predictor is independent of others in the dataset. Equation (6) gives the Bayes Theorem:

$$P(A|B) = \frac{P(B|A)(P|A)}{P(B)} \tag{6}$$

Multinomial Naïve, Bernoulli Naïve, and Gaussian Naïve Bayes are the three forms of Naïve Bayes. For document classification problems, the multinomial Naïve was the most commonly employed method. Meanwhile, Boolean values such as true and false, as well as 0 and 1, are commonly used by Bernoulli Naïve in prediction. The Gaussian Naïve also used prediction by using continuous value data sets. Using Naïve Bayes has both advantages and disadvantages. If the prediction is correct, for example, the performance improves when compared to other methods, such as logistic regression with less training data. It's also straightforward to set up, but there's a good probability it'll reduce data accuracy. Naïve Bayes is a supervised machine learning approach to classification that can infer the presence of a specific feature in a class without regard to the presence of other characteristics.

2.2.3 k-Nearest Neighbor (k-NN)

The supervised classification method k-Nearest-Neighbor (k-NN) can classify non-attributes by assigning them to a similar attribute in the class [15]. Based on the articles and journal that related to (k-NN), The Bayes minimum probability of error is better for making the most impact article in pattern recognition because the probability of error of simple classification rule is bounded with the Bayes minimum probability of error [16].

This model consists of a few of the equations that estimated the distance between the variable of the data set. In addition, there is another distance measure where it uses to Normalize distance. When a training set has a combination of variables, such as numerical and categorical, this standardized distance occurs. Equation (7) shows the equation of standardized distance.

$$\text{MinMax} = \frac{(v - \text{Min } x)}{\text{Max } x - \text{Min } x}(\text{newMax} - \text{newMin}) + \text{newMin} \tag{7}$$

where:

i. Min is referring to the minimum value of the attribute soil data set.
ii .Max is referring to the maximum value of the attribute soil data set.
iii. v is the pick value of the row on each attribute of the soil data set.
iv. newMax is set the maximum value as 1.
v. $newMin$ is the minimum value as 0.

Because the k-NN model is simple to construct and understand, it has both advantages and disadvantages. The k-NN model does not require any training, and this makes it known as the Lazy Leaner. This model algorithm outperforms other models like linear regression and Support Vector Machine in terms of training speed (SVM). However, when dealing with large data sets, this paradigm has a significant drawback, namely slow performance. Another issue with k-NN is that it is prone to noisy data and missing values, which must be filled in manually.

In this project, these three algorithms were included in the experiments namely Random Forest, Naïve Bayes, and k-Nearest Neighbor (k-NN) to identify the better performing models.

2.3 Comparative Study

Three relevant research papers were reviewed and summarized as follows. The Machine learning is used in the study of [17] to forecast mustard crop output based on soil analysis. k-Nearest Neighbor (k-NN), Nave Bayes, Multinomial Logistic Regression, Artificial Neural Network (ANN), and Random Forest are five supervised machine learning methods used in this study to estimate Mustard Crop yield from soil data. Based on the results of the experiment, they concluded that machine learning techniques may be successfully applied for yield prediction. k-NN and Random Forest predicted the highest accuracy (88.67 percent and 94.13 percent, respectively), while Nave Bayes predicted the lowest accuracy (72.33 percent), ANN predicted 76.86 percent, and Multinomial Logistic Regression predicted 80.24 percent, according to the results of this study's experiment.

To assess soil type, [18] used k-Nearest Neighbor (k-NN), Random Forest, Decision Tree, and Nave Bayes [18]. These algorithms collect information from soil data using classifier techniques. The main purpose of these four classifiers is to discover the best machine learning for soil categorization. When compared to Naïve Bayes (69.23 %), Decision Tree, and Random Forest, k-NN has the highest accuracy of 84 %, according to the testing data (53.85 %). Therefore, it outperforms other classification algorithms. The data suggest that k-NN could be effective in agricultural soil type classification.

Using a publicly available agricultural soil dataset, [19] examined the performance of three well-known classification models: k-Nearest-Neighbor (k-NN), Nave Bayes, and Decision Tree. The investigation helps to get the domain knowledge of soil science and machine learning techniques to solve the various soil research problem. Each classifier model is implemented and evaluated with the same dataset for the performance evaluation and calculation of accuracy. From the experiments, the Naïve Bayes algorithm had the lowest accuracy of 72.90% and for the k-NN the accuracy result is 73.56%. Meanwhile, Decision Tree had the most outstanding accuracy of 80.84%. When the fused or suggested technique is applied to the same dataset, it achieves a greater accuracy of 84.14% than the remaining three classifiers. According to the findings of the study, it concluded that the suggested ensemble classifier outperformed the popular three classifiers in terms of accuracy. The accuracy of the algorithms in three research papers that used the soil data set to identify the soil for crop recommendations are compared in Table 1. This accuracy will be compared to the algorithm utilized in this suggested research at the end of the research undertaking.

Table 1 - Comparison of the accuracy among the algorithm in three research paper

Paper	Algorithm (s)	Accuracy (%)
[17]	Naïve Bayes	72.33
	k-Nearest-Neighbor (k-NN)	88.67
	Multinomial Logistic Regression	80.24
	Random Forest	94.13
	Artificial Neural Network (ANN)	76.86
[18]	Naïve Bayes	69.23
	Decision Tree	53.84
	Random Forest	53.84
	k-Nearest-Neighbor (k-NN)	84.61
[19]	Decision Tree	80.84
	k-Nearest-Neighbor (k-NN)	73.56
	Naïve Bayes	72.90
	Ensemble Classifier	84.14

3. Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) [20] was employed in this study. There are five steps in this procedure. Figure 1 shows the CRISP-DM process model.

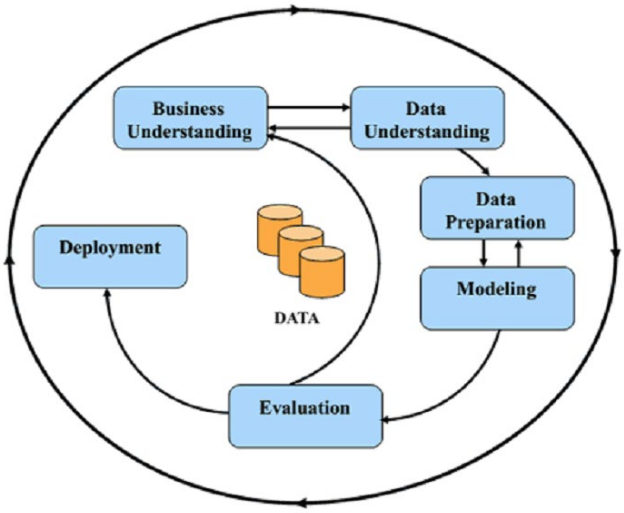

Fig. 1 - CRISP-DM process model [20]

3.1 Research Framework

The research framework is described in Figure 2, which begins with data selection, data pre-processing, data transformation, and feature selection, after which it will be trained and tested using classification algorithms.

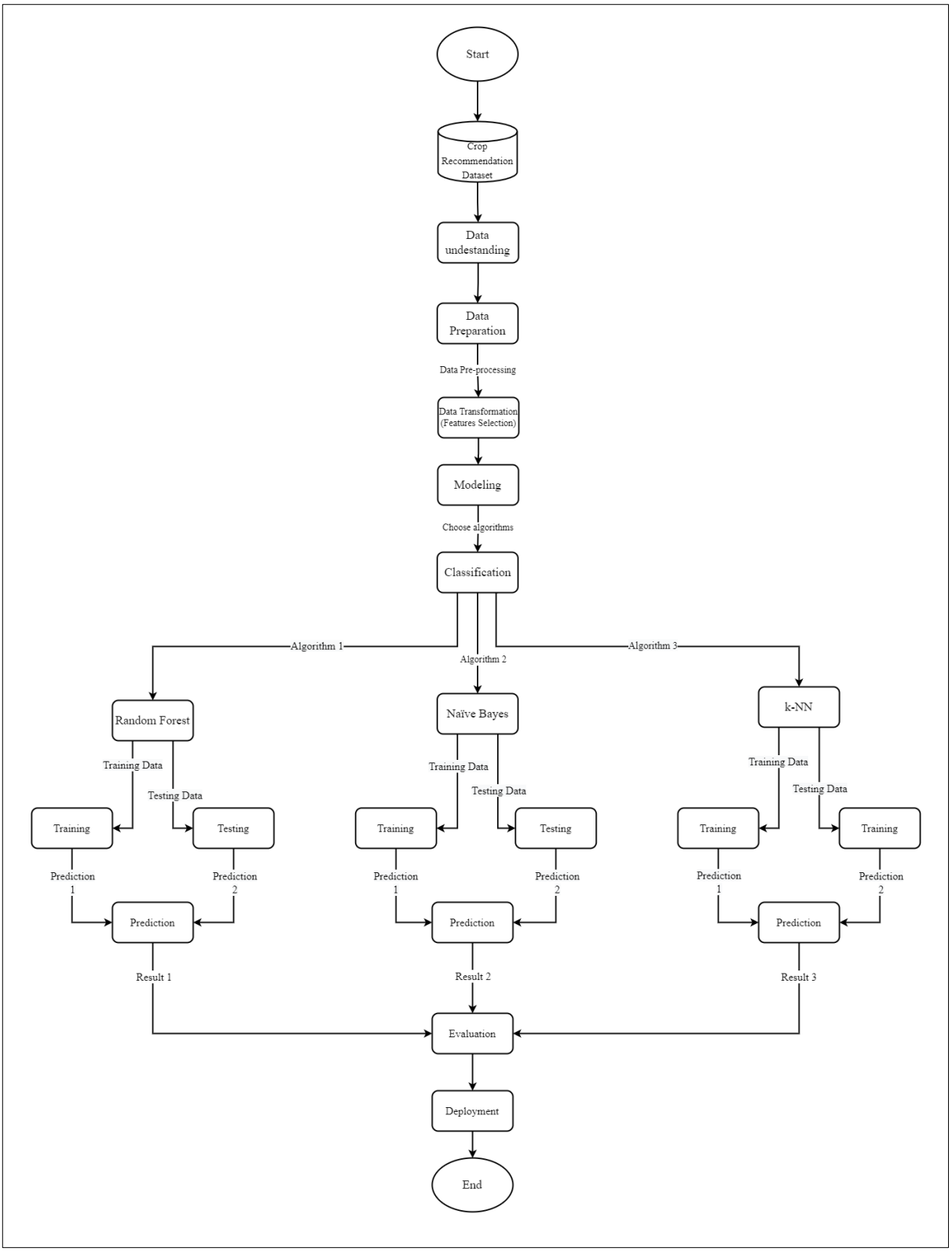

Fig. 2 - Research framework

3.2 Data Selection

This study will use two types of data that will be combined both. First one, the Crop Recommendation dataset from Kaggle (https://www.kaggle.com/atharvaingle/crop-recommendation-dataset). This dataset was built by augmenting datasets of rainfall, climate and fertilizer data that classify the soil for a total of 22 different crops. With 22 crops, this dataset has a total of 2200 data with 7 data fields and one class label. Second, the real-world data or the raw data from

iGROW with 6 data 6 fields and one class label. The context of the data fields is N, P, K which means the ratio of Nitrogen, Phosphorous and Potassium content in soil. Next is temperature in degree Celsius, humidity in percent, pH value of the soil and the last one is the rainfall in mm, but the raw data doesn't have the rainfall attributes. This data was in numerical form in continuous value that was suitable by using a classification in the classifying the accuracy of each method.

3.3 Data Pre-processing

Some of the approaches will be applied on the crop recommendation dataset in the data pre-processing phase before data simulation using the selected algorithm, which includes Random Forest, Naïve Bayes, and k-NN. Data cleaning, normalization, and reduction are the most common data processing methods. It will be used to clean data in the crop recommendation dataset if there are any missing values that need to be fixed. However, in this research, the crop suggestion dataset had no missing values. As a result, there will be no data cleanup. However, the real-world data from iGROW necessitates the pre-processing phase in order to obtain better data with fewer missing values.

Data reduction, on the other hand, is a procedure in which a large dataset is reduced in order to speed up the process of obtaining results. After all, the fewer datasets used, the more precise the data produced.

3.4 Parameter Selection

The process of determining the selected parameter for each technique algorithm that will be employed during the training of each of the data is known as parameter selection. Each strategy will be chosen based on how manipulating a particular parameter affects that parameter. For Random Forest techniques the parameter used was ntree. For the k-NN, k will use as a parameter, and Naïve Bayes will use the feature as a chosen parameter for tuning in every technique to increase the performance of the accuracy. Tables 2 and 3 show the selected parameter of each technique with the description of each parameter.

Table 2 - Parameter setting for each technique in crop dataset

No.	Techniques	Default Parameter	Parameter selected (testing)
1	Random Forest	ntree: 500	ntree: range (200-2570)
2	Naïve Bayes	feature:	feature: All feature
3	k-NN	k=5	k=1-50

Table 3 - The description of each selected parameter of each technique

No.	Techniques	Parameter	Parameter Description
1	Random Forest	ntree	ntree: The number tree to grow.
2	Naïve Bayes	feature	feature: The attribute of the data set.
3	k-NN	k	k: the number of nearest neighbors of the model will be consider based on length of the data.

3.5 Experimental Setting

The experimental setup that will be employed in this research project is described in this section. The Random Forest, Naïve Bayes, and k-NN as a training and testing model were utilized. The training and testing model are required in this experiment to generate these two models for each method. The reason for using these two models is that in the training model, each algorithm will be used to ensure that the algorithm can be trained using the crop data set because each algorithm cannot be applied directly due to data and algorithm incompatibility, so the algorithm must be trained first before being put to the real test.

The crop recommendation dataset will be divided into two parts: data training and data testing. For each categorization technique, the training and testing ratios will be used in the same context. So, to carry out this experiment, testbed will be needed, often known as a platform. Microsoft Azure, MATLAB, RapidMiner, and RStudio are just few of the platforms available. RapidMiner Studio will be used as a platform for this test. The expected outcome of the result is to meet or not meet what was proposed in this research project at the end of each algorithm's result validation and evaluation.

3.6 Research Activities

There are total of five phases will be used from CRISP-DM model. Table 4 shows the research activities or milestone in each phase to be conduct and followed during the entire research.

Table 4 - Research activities

Phase	Task	Output
Data understanding	Collect initial data (acquire the data listed in the project resources for soil classification such as humidity, soil pH, soil nutrient)	Initial data collection report consists of list of the dataset acquired, the method used to acquire it and any problems encountered.
	Describe data (examine the properties of the acquired data and report on the results)	Data description report evaluate whether the data acquired satisfied the relevant requirements.
	Explore data by querying, visualization, and reporting techniques.	Data exploration report (including first findings or initial hypothesis and impact on the remainder of the project)
	Verify data quality (examine the data quality, correct, or contain any error)	Data quality report (list of the results of the quality verification)
Data preparation	Select data (deciding the dataset to be used)	Dataset description
	Clean data	Data cleaning report
	Construct data	Derived attributes and generate records.
	Integrate data	Merged data
	Format data	Reformatted data
Modeling	Select the modeling techniques (machine learning algorithms)	Modeling techniques (Random Forest, Naïve Bayes, k-Nearest-Neighbor (k-NN)).
	Generate test design	Test design.
	Build model by running the modeling tool on the prepared dataset to create models	Parameter settings, model and model description.
	Assess the model	Model assessment, revised parameter settings. All the revisions and assessment were documented.
Evaluation	Evaluate the results.	Assessment of data mining results and approved model.
	Review process also covers quality assurance issues.	Summarize the process review.
	Determine the next step, decide how to proceed.	List of possible further actions and decision making.
Deployment	Plan deployment and strategy for deployment.	Summarize the deployment strategy.
	Plan monitoring and maintenance.	Summarize the monitoring and maintenance strategy.
	Produce the final report.	Final report and final presentation.
	Review the report	Experience documentation

3.7 Testbed (RapidMiner Studio)

This research project necessitates the use of a testbed, also known as a platform, to conduct experiments. Microsoft Azure, MATLAB, RapidMiner Studio, and RStudio are a few examples of platforms. RapidMiner Studio is used to complete this project and Figure 3 shows the tool's user interface.

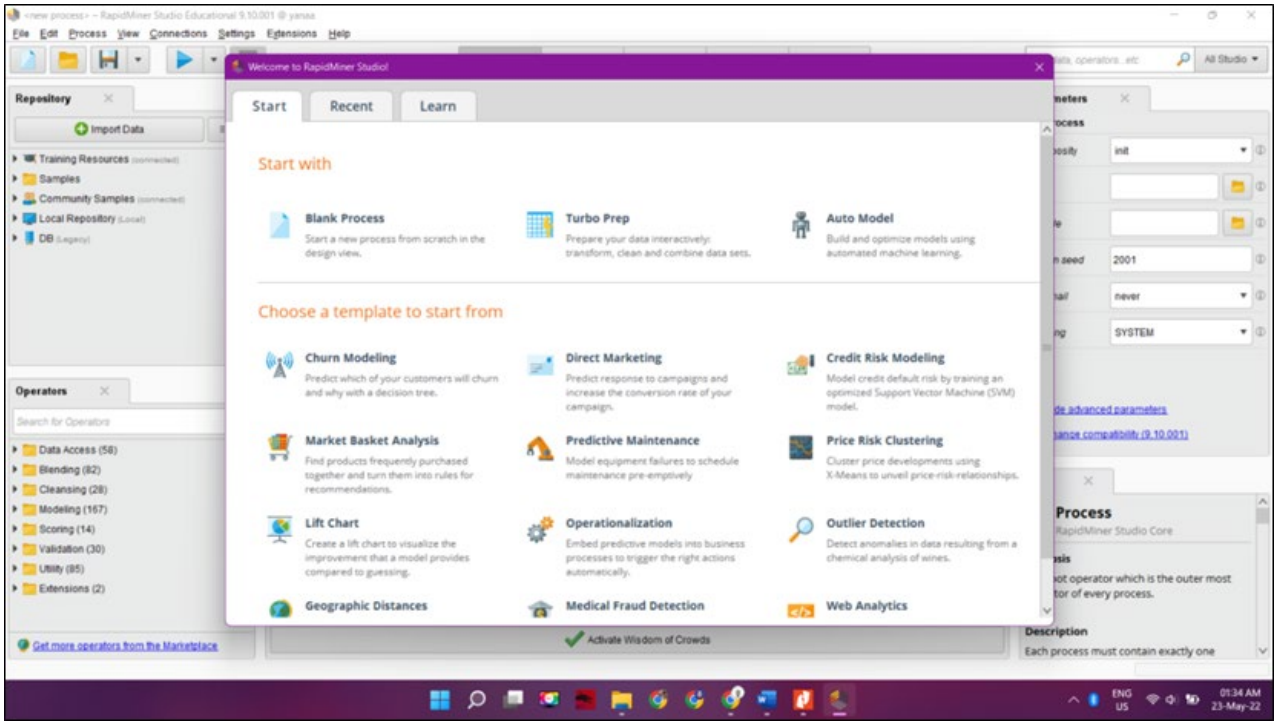

Fig. 3 - Interface of RapidMiner Studio

3.8 Simulation Setup

The dataset needs to be imported to the RapidMiner Studio before starting the experiment. Table 5 shows the data information of the dataset for the experiment. Meanwhile, the splitting ratio of the processed dataset for training data and testing data is shown in table 6.

Table 5 - Data information

No.	Data	Number of Data	Number of the Attribute	Class Label	Column	Row
1.	Crop dataset	2570	6	1	7	2571

Table 6 - Data split for training and testing

No.	Data Splitting		Crop Recommendation Dataset		
			Random Forest	Naïve Bayes	k-NN
1.	50-50	Training (0.5)	1-1285 (1285)	1-1285 (1285)	1-1285 (1285)
		Testing (0.5)	1285-2570 (1285)	1-1285 (1285)	1-1285 (1285)
2.	60-40	Training (0.6)	1-1542 (1542)	1-1542 (1542)	1-1542 (1542)
		Testing (0.4)	1543-2570 (1028)	1543-2570 (1028)	1543-2570 (1028)
3.	70-30	Training (0.7)	1-1799 (1799)	1-1799 (1799)	1-1799 (1799)
		Testing (0.3)	1780-2570 (771)	1780-2570 (771)	1780-2570 (771)
4.	80-20	Training (0.8)	1-2056 (2056)	1-2056 (2056)	1-2056 (2056)
		Testing (0.2)	2057-2570 (514)	2057-2570 (514)	2057-2570 (514)

3.9 Parameter and Testing Methods

A parameter is a model that is used to produce a prediction, and it refers to the classification strategy that will be utilised, which will be Random Forest, Naïve Bayes, and k-Nearest Neighbour (k-NN). Testing methods refer to the tools that the assessment uses to back up the results of each methodology, whether true or untrue.

The method used is a confusion matrix that can indicate technique accuracy, performance, and experimental errors. Accuracy is calculated by dividing the total quantity of data in the data set by the number of correct predictions. Cohen's Kappa Statistic is a very useful, but under-utilised, metric. In machine learning, we may encounter a multi-class classification issue. In certain circumstances, measurements like accuracy or precision/recall do not offer a whole view of our classifier's performance. That's why kappa was used as performance measures because crop

recommendation dataset has a multi-class classification which are 23 types of class. kappa can be calculated using both the observed (total) accuracy and the random accuracy.

Mostly the best value for kappa and accuracy were 1.0 or 100%, the worst value was 0.0 or 0% [21]. This equation will be used during the training and testing model to find the accuracy of each algorithm used. The main measure of performance is evaluated in terms of accuracy and kappa from the confusion matrix of classification. The measures and experimental errors are computed by using equations that are described in the following:

Accuracy:

$$\text{Accuracy} = \frac{\text{TN+TN}}{\text{TP+TN+FN+FP}} = \frac{\text{TP+TN}}{\text{P+N}} \tag{8}$$

kappa:

$$\mathcal{K} = \frac{\text{p}_\text{o} - \text{p}_\text{e}}{1 - \text{p}_\text{e}} \tag{9}$$

Errors:

$$\text{Errors} = 1 - \text{accuracy} \tag{10}$$

4. Results and Discussion

The goal of the experiments is to compare the results of Random Forest, Naïve Bayes, and k-Nearest Neighbour (k-NN) for crop recommendation datasets. The study's outcomes are evaluated using two performance measures: accuracy and kappa, using four different types of cross validation: 50-50, 60-40, 70-30, and 80-20. The findings of three algorithms from this experiment, which were evaluated using two performance measures, and experimental errors are shown in Table 7.

Table 7 - Performances measures result

No.	Cross Validation (%)	Machine Learning	Error (%)	Accuracy (%)	kappa
1.	50-50	Random Forest	0.0358	0.9642	0.962
		Naïve Bayes	0.0382	0.9618	0.960
		k-NN	0.0787	0.9213	0.917
2.	60-40	Random Forest	0.0302	0.9698	0.968
		Naïve Bayes	0.0340	0.9660	0.964
		k-NN	0.0710	0.9290	0.925
3.	70-30	Random Forest	0.0233	0.9767	0.975
		Naïve Bayes	0.0298	0.9702	0.968
		k-NN	0.0674	0.9326	0.929
4.	80-20	Random Forest	0.0214	0.9786	0.977
		Naïve Bayes	0.0253	0.9747	0.973
		k-NN	0.0610	0.9339	0.930

Table 8 shows the results of performance measures mean of both evaluation metrics; accuracy and kappa of soil classification.

Table 8 - Performances measures mean

Machine Learning	Mean Accuracy (%)	Mean kappa
Random Forest	97.23	0.971
Naïve Bayes	96.82	0.966
k-Nearest Neighbour (k-NN)	92.92	0.925

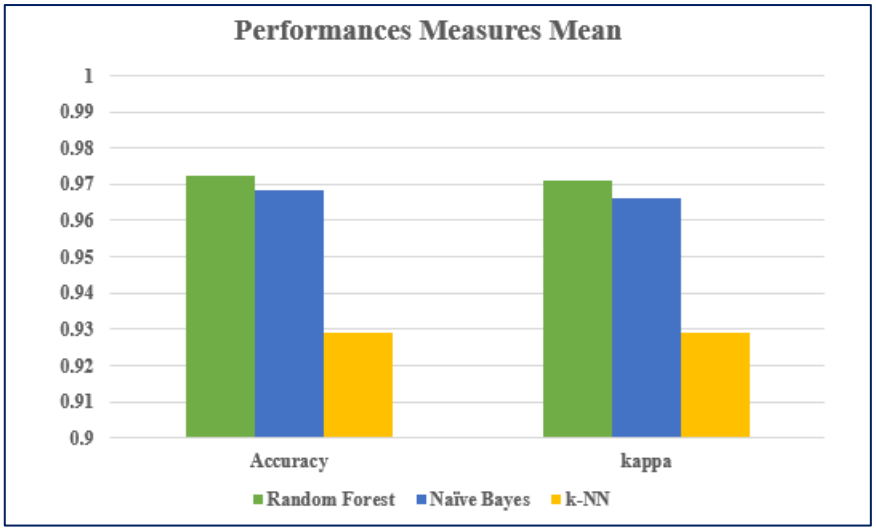

Fig. 3 - Performances measures mean of three machine learning

According to the data shown in Figure 3, Random Forest has the highest mean accuracy of 97.23% and the highest mean kappa of 0.971. Meanwhile, among these two algorithms, k-NN had the lowest accuracy and kappa, with 92.92 percent and 0.925, respectively. With 96.82 percent accuracy and 0.966 kappa, Nave Bayes comes in second. According to the results of the study, Random Forest surpassed the other two algorithms in classifying crop recommendations across all evaluation parameters, including accuracy and kappa. Table 9 tabulates the comparison among the algorithm used in this paper with the algorithm that used in the previous work.

Table 9 - Comparison of proposed work results with those of prior studies

Paper	Data Size	Algorithm (s)	Accuracy (%)
[17]	5000	Naïve Bayes	72.33
		k-Nearest-Neighbor (k-NN)	88.67
		Multinomial Logistic Regression	80.24
		Random Forest	94.13
		Artificial Neural Network (ANN)	76.86
[18]	400	Naïve Bayes	69.23
		Decision Tree	53.84
		Random Forest	53.84
		k-Nearest-Neighbor (k-NN)	84.61
[19]	60	Decision Tree	80.84
		k-Nearest-Neighbor (k-NN)	73.56
		Naïve Bayes	72.90
		Ensemble Classifier	84.14
Proposed Work	2570	Random Forest	97.23
		Naïve Bayes	96.82
		k-Nearest-Neighbor (k-NN)	92.92

This study used 2570 data size and found that Random Forest has the highest accuracy of 97.23 percent among Naïve Bayes and k-Nearest Neighbour (96.82 percent and 92.92 percent), as shown in Table 9. Based on a comparison with the results of a study from [17] that used 5000 data sizes, it also shows that Random Forest is proven to outperform other algorithms with an accuracy of 94.13%. But in this study, the experimental results show that Naïve Bayes has less accuracy with 72.33%. Meanwhile, k-NN is in second place with an accuracy of 88.67%. Random Forest came in bottom with 53.84 percent accuracy, followed by Naïve Bayes with 69.23 percent in the study [18]. With an accuracy of 84.61, k-NN outperforms both of these methods.

Despite this, the data set used by [18] is smaller than that employed by this study and [17]. Random Forest, based on the comparison findings, provides improved accuracy when the data size employed is larger, and it is also used to categorise large datasets in many applications [22]. Machine learning is an efficient approach for soil categorization, according to the study, discussion, and analysis of the data presented in this paper. In comparison to the other two classifiers, Nave Bayes and k-NN, Random Forest was the best performer in terms of accuracy when the findings and performance were analysed. Random Forest is quite capable and may be the best choice for investigations involving agricultural soils.

5. Conclusion

This study examines crop suggestions based on soil categorization utilizing the Random Forest, Nave Bayes, and k-Nearest Neighbor classification methods (k-NN). Each algorithm's total accuracy results are compared. This study looks into the techniques that can provide the best results in terms of recommending the best crop for a specific soil. The following is a suggestion that can be used with the proposed algorithm: (i) A comparison of the various techniques can lead to a more thoroughly analyzed outcome. (ii) Use the most recent dataset. (iii) At least one dataset should contain three different algorithms, or one algorithm should contain three different datasets.

Acknowledgment

I would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support.

References

[1] R. Weil and N. Brady, Elements of the nature and properties of soils, 15th ed.
[2] Lombardozzi, D. L., Lu, Y., Lawrence, P. J., Lawrence, D. M., Swenson, S., Oleson, K. W., Wieder, W. R., & Ainsworth, E. A. (2020). Simulating agriculture in the Community Land Model Version 5. Journal of Geophysical Research: Biogeosciences, 125(8). https://doi.org/10.1029/2019jg005529 .
[3] Atkinson, J., Davison, L., & Springman, S. (2000, May). Soil description and classification. Soil Classification. Retrieved from https://environment.uwe.ac.uk/geocal/SoilMech/classification/default.htm
[4] Singh, R. K., Sinha, V. S., Joshi, P. K., & Kumar, M. (2020). Modelling Agriculture, Forestry and Other Land Use (AFOLU) in response to climate change scenarios for the SAARC Nations. Environmental Monitoring and Assessment, 192(4). https://doi.org/10.1007/s10661-020-8144-2 .
[5] Bronick, C. J., & Lal, R. (2005). Soil Structure and management: A Review. Geoderma, 124(1-2), 3–22. https://doi.org/10.1016/j.geoderma.2004.03.005 .
[6] Lin, B. B. (2011). Resilience in agriculture through crop diversification: Adaptive Management for Environmental Change. BioScience, 61(3), 183–193. https://doi.org/10.1525/bio.2011.61.3.4 .
[7] J. Mitchell and K. Soga. (2005). Fundamentals of soil behavior, 3rd ed. Hoboken, NJ: Wiley.
[8] Food and Agriculture Organization of the United Nations. (2020, November). Soil classification. Soil classification | FAO SOILS PORTAL | Food and Agriculture Organization of the United Nations. Retrieved from http://www.fao.org/soils-portal/data-hub/soil-classification/en/.
[9] Butler B. E. and B. Butler, Soil Classification for Soil Survey. Oxford: Clezedon Press.
[10] IBM Cloud Education. (2020, September). What is machine learning? IBM. Retrieved 2021, from https://www.ibm.com/cloud/learn/machine-learning .
[11] Hamzah, N. F., Nawi, N. M., & Hezam, A. A. (2021). The Analysis Performance of Heart Failure Classification by Using Machine Learning Techniques. *Journal of Soft Computing and Data Mining*, *2*(2), 98-108.
[12] Leonel, J. (2019, October 9). Classification methods in machine learning. Medium. Retrieved 2021, from https://medium.com/@jorgesleonel/classification-methods-in-machine-learning-58ce63173db8
[13] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.
[14] Prabhakaran, S. (2018, November 4). How naive Bayes algorithm works? (with example and full code): ML+. Machine Learning Plus. Retrieved 2021, from https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/ .
[15] Zhang, Z. (2016). Introduction to machine learning: K-Nearest Neighbors. Annals of Translational Medicine, 4(11), 218–218. https://doi.org/10.21037/atm.2016.03.37
[16] Cover, T., & Hart, P. (1967). Nearest neighbor Pattern Classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/tit.1967.1053964 .
[17] Pandith, V., Kour, H., Singh, S., Manhas, J., & Sharma, V. (2020). Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. Journal of Scientific Research, 64(02), 394-398. doi:10.37398/jsr.2020.640254.
[18] Taher, K. I., Abdulazeez, A. M., & Zebari, D. A. (2021). Data mining classification algorithms for analyzing soil data [Abstract]. Asian Journal of Research in Computer Science, 17-28. doi:10.9734/ajrcos/2021/v8i230196.
[19] Motia, S., & Reddy, S. (2021). Ensemble classifier to support decisions on Soil Classification, IOP Conference Series: Materials Science and Engineering, 1022(1), 012044. doi:10.1088/1757-899x/1022/1/012044
[20] Wirth, R. & Hipp, Jochen. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.
[21] McHugh, M. L. (2012). Interrater Reliability: The kappa statistic. Biochemia Medica, 276-282. doi:10.11613/bm.2012.031

[22] M. Zakariah, "Classification of large datasets using Random Forest Algorithm in various applications: Survey", International Journal of Engineering and Innovative Technology (IJEIT), vol. 4, no. 3, pp. 189-198, 2014.