Machine Learning Enabled Crop Recommendation System for Arid Land

BATOOL ALSOWAIQ, NOURA ALMUSAYNID, ESRA ALBHNASAWI, WADHA ALFENAIS
SURESH SANKARANARAYANAN
Department of Computer Science
College of Computer Science and Information Technology
King Faisal University
Al Hofuf, KINGDOM OF SAUDI ARABIA

Abstract: The agriculture industry plays a significant role in the economy of many countries, and the population is regarded as an essential profession. To increase agricultural production, crops are recommended based on soil, weather, humidity, rainfall, and other variables which are beneficial to farmers as well as the nation. This paper explores the use of “machine learning” algorithms to recommend crops in for Arid land based on features selected from tropical climate where crops grow effectively. Five “machine learning” models have been validated for recommendation of crops for arid land which resulted in “Random Forest” topping as the best model.

Keywords: Crop recommendation, Machine Learning, Random Forest. Humidity, rainfall

Received: October 24, 2022. Revised: May 19, 2023. Accepted: June 20, 2023. Published: July 10, 2023.

1. Introduction

Food and agriculture are two of the most important sources of livelihood for most people in the world. With the escalation of climate change, it is becoming increasingly important for farmers to optimize their production in order to maximize their yields and reduce their costs. However, the traditional methods of selecting and recommending crops are insufficient in determining the best possible crop for a given environment. Choosing the wrong crops will always result in lower productivity. Families who rely entirely on agriculture have a difficult time surviving. There has been good amount of work done in employing machine learning models like “SVM, Naïve Bayes, Decision Tree, Chaid, LEM2” algorithm for selecting the best crop. Most of the work employing machine learning have been done for one or two crops and mostly focussed for tropical climatic condition like India and other countries. The reason being most of agriculture production is dependent of availability of fertile soil and tropical climatic condition.

So accordingly, we here have focussed on validation different machine learning model like “SVM, Naïve Bayes, KNN, Decision Tree, Random Forest” in terms of accuracy and error towards developing an Machine Learning enabled crop recommendation from Arid land point of view. Though the vision of Kingdom of Saudi Arabia is to promote Agriculture in this region towards increasing the yield and productivity, there is need to develop crop recommendation system driven by machine learning as a first step which would benefit the Arid land. Secondly, we have chosen crop related dataset with features like “NPK, pH, Temperature, Rainfall, Humidity” which helps in the growth of crop in fertile soil with good climatic condition. So, this dataset has been beneficial for developing a Machine Learning enabled Crop recommendation which would benefit in selecting the best crop based on real time data received from Arid land based on model trained for crop recommendation.

This system would provide farmers with the opportunity to consider various agricultural benefits, such as improved yields, reduced costs, and improved decision-making processes. The rest of the paper is organized as follows. Section II discusses the various work carried out towards crop recommendation using machine learning. Section III discussed the Crop Recommendation using Machine Learning. Section IV discussed the performance of machine learning model implemented for crop recommendation with results. Section VI tabulates the results of the model followed by discussion. Section VII gives the concluding remarks and future work.

2. Literature Review

There has been some good amount of work carried out pertaining to crop recommendation using Machine Learning algorithms.

Authors in the paper [1] developed a crop recommendation system using ensemble technique with major voting using machine learning models “Random Tree, CHAID, K-NN and Naïve Bayes”. The authors in this work have resulted in an accuracy of 88%. The model has been proved to be adaptive and utilized by farmers through a GUI developed as Web portal. This study provided great insight into the growth of the agriculture sector by minimizing the wrong choices made by farmers and increasing their productivity. The work has drawback where they have focused on crops pertaining to district in city in India and no other crops. In addition, they have achieved an accuracy of 88% for limited dataset. There has been no correlation shown among the features used for crop classification towards recommendation which is very important. There has been no error computed in terms of precision and recall for showing how accurately classification is done in terms of True Positive and True Negatives.

The research in [2] describes the development of an assembling technique of three machine learning models which are "Random Forest, Naïve Bayes and Linear SVM" towards crop recommendation resulting in an improved crop productivity. The system is based on soil specific physical and chemical characteristics such as "NPK, pH", soil type, pores in soil , average rainfall, sowing season and temperature of the surface. This has been focused on two crops which are "Kharif and Rabi" resulting in an accuracy of 99.91%. The challenge in this work is limited crop used for recommendation though dataset of these crops is large. Secondly, though the accuracy achieved is higher, there has been no correlation seen among features for crop recommendation. Also, they have worked as ensemble method for classification using all four models. There has been no error computed in terms of precision and recall for showing how accurately classification is done in terms of True Positive and True Negatives.

Researchers in [3] have worked towards deploying four machine learning models which are "SVM, Decision Tree, Naïve Bayes, Random Forest, KNN, Logistic Regression". This resulted in selection of appropriate crop using the machine learning model with a classification accuracy of 89.66% for "SVM" followed by "KNN, Random Forest, Naive Bayes, Decision Tree and Logistic Regression". Followed by the crop selected, the system would suggest the past that would affect the crop followed by recommendation towards pest control. This work has focused on different crops with their selection. The model has not achieved higher accuracy as the number of samples of each crop is less for some crops and high for some crops. It does not have a consistent number of samples of each crop. There is also possibility of data imbalance which has not been explored and correlation of soil features contributing for crop selection also not evaluated. There has been no error computed in terms of precision and recall for showing how accurately classification is done in terms of True Positive and True Negatives.

The research paper [4] proposes a methodology for improving agricultural crop yield by accounting for the soil micro and macronutrient levels to predict crop suitability. To achieve this, fuzzy logic and rough set rule induction were used to create rules for the dataset and evaluate different algorithms for their accuracy. The results of the evaluation found that the "LEM2" algorithm gave the highest prediction accuracy of 89% for the dataset without fuzzy logic, while the AQ algorithm showed better accuracy than others with 3 linguistic variables. The paper concluded that the proposed methodology could be used in all situations, as it could help farmers to determine the crop that best suits their soil type and could reduce soil erosion, as farmers could shift to lesser-water intensive crops when water availability is low. Though the work has considered 23 crops with 16 features for crop selection, there has been no usage of machine learning model deployed for selection of crops with higher accuracy.

3. Crop Recommendation Using Machine Learning

So, based on literature reviewed pertaining to machine learning for crop recommendation system based on soil features and climatic condition, we in this paper have worked on dataset collected pertinent towards tropical climate for crop recommendation for Arid land. The reason for choosing dataset pertaining to tropical climate condition is towards training the model with soil features for the recommendation of crop. The model trained and evaluated would be used for crop recommendation for Arid land based on features collected in real time. The availability of Arid land dataset for crop recommendation is not available as agriculture in Arid land is really challenging and is one of the primary focusses of Kingdom of Saudi Arabia. So before going into the results and analysis of machine learning model, we investigate methodology of proposed work pertaining to following models which are "Support Vector machine, Decision tree, Random Forest, K-Nearest Neighbor and Naïve Bayes". The details about the model are explained in brief.

3.1 Random Forest Algorithm

"Decision Tree" are created based on variety of samples and averaging and majority voting are employed for classification and regressing. It can handle categorical variables in the case of classification. Our project provides solutions to the multiclass problem. The principle of "Bagging" is employed for "Random Forest". "Random Forest" also employed an ensemble method known as "Bootstrap Aggregation". Each sample is trained independently producing results. The final decision is based on majority voting where the results of all models merged which is termed as aggregation Description

3.2 Decision Tree

In a "decision tree" algorithm, a root node provides the optimal split. It is represented by the predictor variable which helps to divide the data set into two or more subsets. Then the entropy or Gini Index is used to measure the homogeneity of a split. It measures the potential for information gain when splitting a node. Then, each node has a split of the predictor variable which yields the best homogeneity. Building a decision tree continues until all the nodes are pure, meaning all the nodes are homogeneous and belong to the same class. The process stops here, and the final tree is used to make predictions.

3.3 Naïve Bayes

The "Naïve Bayes" algorithm is a "supervised learning" method for classification problems. This method makes the prediction based on probability which we call it as "probabilistic classifier". The machine learning model is based on "Bayes Theorem" which calculates the likelihood of hypothesis of data.

3.4 Support Vector Machine

"Support vector machine (SVM)" is utilized for classification and regression. "Hyperplane" in N-dimensional space is found using this algorithm which classifies the data points. Hyperplane is used to essentially divide the input features into two classes. It becomes challenging when the number of features exceed three.

3.5 K-Nearest Neighbors

The "K-Nearest Neighbors (KNN)" algorithm is a simple, "supervised learning" algorithm used for classification and regression. The algorithm works by comparing an input sample to the k nearest neighbors in the training dataset. The k neighbors are determined based on their proximity to the input sample, typically using Euclidean distance as a measure of similarity. "KNN" is known as a non-parametric algorithm because it does not require assumptions about the underlying probability distribution of the data, making it a popular choice for many classification and regression problems.

4. Performance of MAchine learning model for crop recommendation

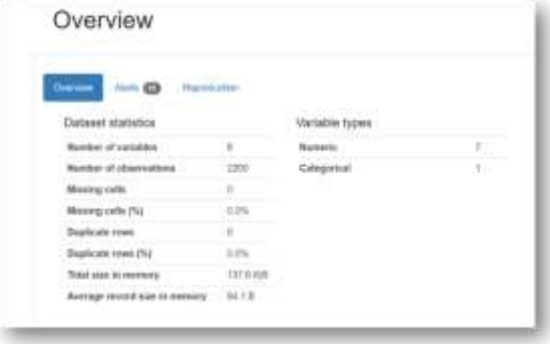
For our work, we have acquired a dataset as CSV file from Kaggle [5]. The dataset contains eight columns comprising of 2200 records, and 17600 fields. The column contain "N" which is the amount of "Nitrogen" content in soil, "P" which is the amount of Phosphorus content in soil, "K" which is the amount of "Potassium" content in soil, temperature which is in degree Celsius, humidity which is the relative humidity in %," "PH" value of the soil, rainfall which is rainfall in mm and label which is the classification's name. In each record, the datatype for each column is shown in **Table 1**. In the Crop Recommendation dataset, we have 22 crops which are as follows: "Rice, Maize, Jute, Cotton, Coconut, Papaya, Orange, Apple, Muskmelon, Watermelon, Grapes, Mango, Banana, Pomegranate, Lentil, Blackgame, Mung Bean, Moth Beans, Pigeon Peas, Kidney Beans, Chickpea, Coffee"

Table 1 The datatype of the dataset's features.

-N (Nitrogen) -P (Phosphorous) -K (Potassium)	-temperature -humidity -PH -rainfall	Label
int64	float64	STRING

4.1 Data Visualization and analysis

For visualizing the dataset content and characteristic we used Python's Pandas profiling library that automates exploratory data analysis. It generates a dataset profile report that provides valuable insights. With the profile report, we can know which variables to use and which ones to drop. The following is some illustration and discussion of the generated report about the dataset:



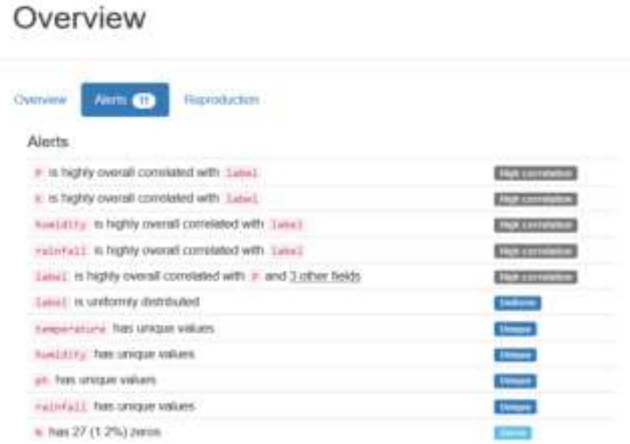
Overview

Dataset statistics	
Number of variables	8
Number of observations	2200
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	137.6 KiB
Average record size in memory	64.1 B

Variable types	
Numeric	7
Categorical	1

Fig. 1 Detailed overview of the dataset-1.

The first generator is the overview, which shows the information about the statistics of the dataset. As shown in **Fig.1**, the number of columns is 8, the number of rows is 2200, missing values are 0 etc.



Overview

Alerts

P is highly overall correlated with label	High correlation
K is highly overall correlated with label	High correlation
humidity is highly overall correlated with label	High correlation
rainfall is highly overall correlated with label	High correlation
label is highly overall correlated with P and 3 other fields	High correlation
label is uniformly distributed	Uniform
temperature has unique values	Unique
humidity has unique values	Unique
ph has unique values	Unique
rainfall has unique values	Unique
N has 27 (1.2%) zeros	Zeros

Fig. 2 Detailed overview of the dataset-2.

Fig. 2 shows the features that strongly affect the label (crop) which are P, K, humidity, and rainfall. It also presents feature N which mostly contains zero since it has 27 rows but this does not affect the dataset.

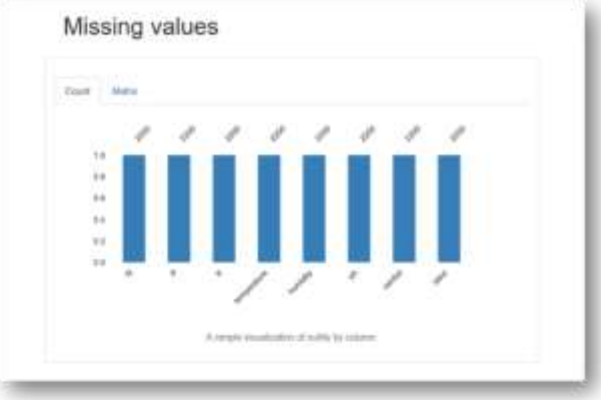


Fig. 3 Missing values View.

The Missing values view shows how many missing values are in each feature, as in **Fig. 3**. The dataset has 2200 rows which means no missing values in our dataset.

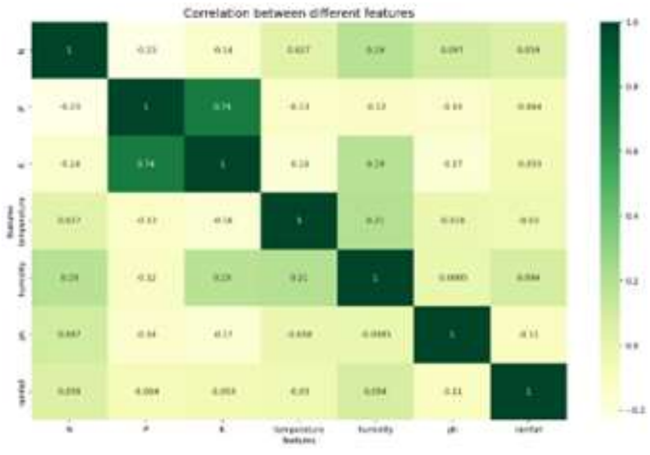


Fig. 4 Correlation View.

Fig. 4 demonstrates the correlation view which illustrates the relationship between all features together, a positive number indicates a strong relationship, while a negative number indicates a light relationship.
Now based on visualization of dataset with correlation of features, there is need to evaluate the machine learning models. So, towards this, dataset is split into training and testing. The performance of machine learning is enhanced by splitting into training and testing which is a basic step of data preprocessing. We take 80% of dataset for training and remaining 20% for testing.

After training the machine learning model using a trained set of data, it is important to use evaluation metrics to measure the performance of the ML model. For a different set of machine learning algorithms, there exist many evaluation measures. We will discuss only the Classification Evaluation Matrix since our selected dataset 'Crop Recommendation' belongs to this category There are different approaches for this type of evaluation matrix [6] such as:

- "Confusion matrix
- Accuracy
- Precision
- Recall
- Specificity
- F1 Score"

A. Confusion Matrix

It is a tabular summary of several true and false predicates made by the ML classifier model to measure its performance. It is considered a useful tool for visualizing the measurement of Accuracy, Precision, Recall, and "AUC-ROC" curve. It simply shows n-dimensional, where n indicates the number of classes in the dataset. It consists of four parts[8]:

1) True positive (TP): It indicates the situation where both the actual value and the predicted value by the ML model are positive.
2) True negative (TN): It indicates the situation where both the actual value and the predicted value by the ML model are negative.
3) False positive (FP): It indicates a situation where the actual value is negative, but the predicted value by the ML model is positive.
4) False negative (FN): It indicates the situation where the actual value is positive, but the predicted value by the ML model is negative.

4.2 Accuracy

It calculates the ratio of the correct prediction made by the ML model over the total number of instances evaluated. It is a valid technique when a dataset is balanced, in which the proportion of all instances of each class is quite similar. However, it is invalid in the opposite situation where the proportion of the total number of instances per class is far from each other. The following is the accuracy formula:

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

4.3 Precision

It is an approach that deals with an imbalanced dataset. It calculates s ratio of correct positive predictions to the total number of instances classified in the positive class. The following is the precision formula:

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

4.4 Recall

It can be called Sensitivity is a helpful measurement when a dataset is imbalanced where the minority class is positive. It calculates the ratio of correct positive predictions to the dataset's overall number of positive instances. The following is the recall formula:

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

4.5 Specificity

It is the opposite of Sensitivity in which the minority class of an imbalanced dataset is negative. It calculates the ratio of correct negative predictions to the overall number of negative instances in the dataset. The following is the specificity formula:

$$Specificity = \frac{TN}{TN + FP}$$

4.6 F1-Score

It is useful where it is important to avoid false positives and negatives. This metric represents the harmonic mean between recall and precision values. The following is the F1-Score formula:

$$F1 - Score = \frac{2(Precision\ \times Recal)}{Precision + Recal}$$

5. |Result and Discussion

We have evaluated the four machine learning models which are "SVC, Naïve Bayes, Decision tree, Random Forest and KNN" Algorithm towards crop recommendation of different crops based on features correlated. The models have been evaluated in terms of precision, recall, F1 score and accuracy. In addition, the accuracy of model towards training and testing tabulated too in **Table 2** and **Table 3** respectively**. Fig.5** shows the model's performance.

Table 2 Comparison between the performance scores of the modes.

Models	Accuracy	Precision	Recall	F1
SVC	98.36%	98.36%	98.63%	98.40%
Decision Tree	98.18%	98.18%	98.25%	98.20%
Random Forest	99.45%	99.45%	99.60%	99.50%
KNN	98.00%	98.14%	98.10%	98.10%
Naïve Bayes	99.09%	99.10%	99.30%	99.15%

Table 3 Models' training and testing accuracy.

Model	Train accuracy	Test accuracy
SVC	97.88 %	98.36%
Decision Tree	100%	98.18%
Random Forest	100%	99.45%
KNN	98.79 %	98.00%
Naïve Bayes	99.58 %	99.09%

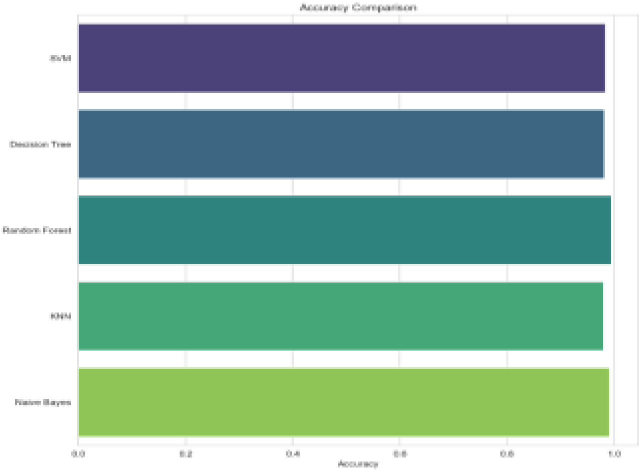


Fig.3 Comparison plot between the models based on the Accuracy.

6. Conclusion and future work

The rapid advancement of technology has seen the emergence of machine learning algorithms, which could help revolutionize agricultural production. Selecting the wrong crops leads to lower productivity. Also, families have a difficult time surviving if they rely entirely on this income. This paper discussed the possibilities of employing machine learning algorithms to develop a crop recommendation system, which helps farmers make informed decisions on which crops to grow based on factors such as climate and soil quality. This paper have validated five machine learning algorithms which are "SVC, Decision Tree, Random Forest, KNN and Naïve Bayes" towards crop recommendation which resulted in Random forest with highest accuracy of 99.45 %, followed by Naïve Bayes, SVC, Decision Tree and KNN respectively. We chose the best model, which is Random Forest based on accurate results because the Crop Recommendation dataset is balanced which calculates the ratio of a correct prediction made by the ML model over the total number of instances evaluated. Moreover, we tested the overfitting and underfitting and we observe that the Random Forest model is generalized.

In future, this work can explore different approaches such as the more sophisticated machine learning algorithms like Deep Neural with optimization for better accuracy of the recommendation system. The system can be integrated with Crop yield prediction for providing better productivity. Additionally, the system could introduce larger datasets with various features such as market prices, weather forecasts, etc. to improve the system's predictive accuracy.

References

[1] S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop recommendation system for precision agriculture," 2016 Eighth International Conference on Advanced Computing (ICoAC), Chennai, India, 2017, pp. 32-36, doi: 10.1109/ICoAC.2017.7951740.

[2] N. H. Kulkarni, G. N. Srinivasan, B. M. Sagar and N. K. Cauvery, "Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique," *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, Bengaluru, India, 2018, pp. 114-119, doi: 10.1109/CSITSS.2018.8768790.

[3] A. Kumar, S. Sarkar and C. Pradhan, "Recommendation System for Crop Identification and Pest Control Technique in Agriculture," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0185-0189, doi: 10.1109/ICCSP.2019.8698099.

[4] A. M. Rajeswari, A. S. Anushiya, K. S. A. Fathima, S. S. Priya and N. Mathumithaa, "Fuzzy Decision Support System for Recommendation of Crop Cultivation based on Soil Type," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 2020, pp. 768-773, doi: 10.1109/ICOEI48184.2020.9142899.

[5] Ingle, A. (2020) Crop recommendation dataset, Kaggle. Available at: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset (Accessed: February 11, 2023).

[6] "A Guide to Evaluation Metrics for Classification Models | Deepchecks", Deepchecks, 2022. [Online]. Available: https://deepchecks.com/a-guide-to-evaluation-metrics-for-classification-models/. [Accessed: 27- Apr- 2022].

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.