Comparative Analysis of Machine Learning Models for Crop's yield Prediction

Zaheer Ud Din Babar[1], Riaz UlAmin[1], Muhammad Nabeel Sarwar[1], Sidra Jabeen[2] and Muhammad Abdullah[2]
mirzazaheeruldinbaber@gmail.com riazulamin@gmail.com mnabeelsarwar96@hotmail.com, Jabeen.sidra98@gmail.com abdullah1521.cs@gmail.com
University of Okara, Punjab, Pakistan
Corresponding Author: Riaz UlAmin riazulamin@gmail.com

Abstract
In light of the decreasing crop production and shortage of food across the world, one of the crucial criteria of agriculture nowadays is selecting the right crop for the right piece of land at the right time. First problem is that How Farmers can predict the right crop for cultivation because famers have no knowledge about prediction of crop. Second problem is that which algorithm is best that provide the maximum accuracy for crop prediction. Therefore, in this research Author proposed a method that would help to select the most suitable crop(s) for a specific land based on the analysis of the affecting parameters (Temperature, Humidity, Soil Moisture) using machine learning. In this work, the author implemented Random Forest Classifier, Support Vector Machine, k-Nearest Neighbor, and Decision Tree for crop selection. The author trained these algorithms with the training dataset and later these algorithms were tested with the test dataset. The author compared the performances of all the tested methods to arrive at the best outcome. In this way best algorithm from the mention above is selected for crop prediction.

Keywords:
Spinach, Humidity, Standard deviation, Logistic Regression

1. Introduction

Agriculture is the backbone of Pakistan's economy. It plays a very important role in the development of Pakistan. In the total labor of Pakistan, 48% of labor is directly involved with agriculture. So, it is the main source of income of the major part of economic papulation. In Pakistan 70% of the population is relating to directly or indirectly with agriculture, also it is the main source of production of food around the world, this it's also the main source of providing raw materials to the industry sector of Pakistan. The total GDP of Pakistan's agriculture contribution is 25% that is higher than other sectors' contributions [1].

Some fewer models are working in real-time to address the above problems. All machine learning models are working but we face issues to select the best machine learning model. For this purpose, I collected the dataset from different spinach fields and preprocess it and then trained machine learning models and compare their results & suggest the best model that predicts crop type that is suitable according to environmental conditions and provided results with maximum accuracy.

Machine learning model are used to predict suitable crop according to environmental factor (Temperature, Soil Humidity, Soil pH). Each machine learning model provided maximum accuracy and predicted Crop that should be cultivated to get maximum yield. But Question is that how farmer can select the best model that provided maximum accuracy. In the reference to above question Author Purpose a method in which machine learning models are trained on the dataset and after training these models tested and get their results after result comparison, model that provided maximum accuracy considered the best model for crop prediction.

This paper answer following main problems that farmers faced in the farmland for crop prediction. Research questions are given below

1. What types of machine learning models are available and how can select the best model for crop selection?
2. What type of dataset is used and what are the main parameters or elements that must be a part of the dataset?

2. Literature Review

Authors portray a way to deal with anticipate millet crop yield expectation, which can be done by taking high dimensional datasets. By utilizing Random Forest Classifier, we acquired 99.74% of exactness in ascertaining the millet crop yield forecast by taking different info fields like soil, min temp, max temp, moistness, precipitation, and so forth [2].

Author fostered a model of 'CROP SELECTION USING IOT AND MACHINE LEARNING' which assists farmers with choosing the best crop for their farmland. In this paper, Author utilize the KNN calculation to choose suitable yields for the farmland as indicated by the farmer land's climatic conditions. The utilized dataset comprises boundaries like harvest name, temperature, stickiness, and soil ph. In this paper, Author might want to assist farmers

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.47

with discovering crops that are more reasonable for their territory and to expand the yield. The KNN calculation will yield conceivable 5 harvests and the rancher can choose his great yield from that [3].

The task principally centers around determining valuable bits of knowledge on crop-yield expectation, climate anticipating, crop type ranch, and harvest cost gauging. The measurable rural dataset is embraced for trial investigation. The information is preprocessed and arranged into preparing and testing information. Then, at that point, appropriate order techniques like Support Vector Machine (SVM) and Random Forest are utilized for better characterization results[4].

This paper shows the most ideal method of crop selection and yield prediction in the least expense and exertion. Artificial Neural Network is considered hearty devices for displaying and forecast. This calculation intends to improve yield and expectation, just as, support vector machine, Logistic Regression, and random forest calculation are additionally thought to be in this examination for contrasting the precision and blunder rate. Besides, these algorithms utilized here are simply to perceive how well they performed for a dataset that is over 0.3 million. We have gathered 46 boundaries, for example, – greatest and least temperature, normal precipitation, stickiness, environment, climate, and sorts of land, kinds of synthetic manure, sorts of soil, soil structure, soil creation, soil dampness, soil consistency, soil response and soil surface for applying into this forecast cycle [5].

3. Methodology

In this research, all algorithms and data processing codes are implemented using the python programmable language. A highly efficient python integrated development environment (IDE) anaconda Spyder is used in this research. With Spyder, most of the programming code is run on Jupiter notebook during the initial data processing phase. This research consists of all the algorithms that are mentioned above has led to propose a comparison between all of them. There are the following steps that we have followed in this research.

- Dataset collection
- Data visualization
- Data Pre-processing in the form of data cleaning and feature extraction
- Data Pre-processing in the form of feature selection, feature scaling
- Data splitting into train and test data
- Fitting the algorithms
- Testing the accuracy of the model
- Data post-processing in the form of performance metrics

The block diagram of the model is given below.

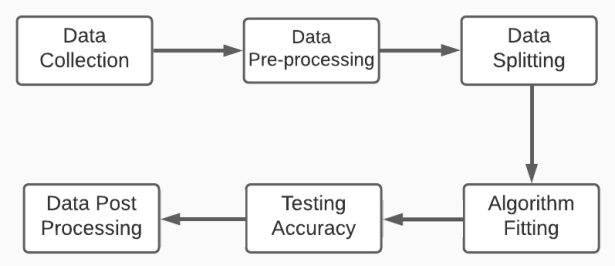

Figure 1: Proposed model architecture.

3.1 Dataset Collection

For research, it is necessary to have a proper dataset to work upon, and it is very difficult to find a legible and reliable dataset, and it took a lot of time and effort to collect the suitable dataset. In this research, all data is collected real-time by using different sensors and then preprocess. Approximately dataset is collected in three (3) months. The dataset contains four (4) columns and 2200 rows. The columns were “Temperature”, “Humidity”, “Soil Ph.”, “Label”. First three contain data in numeric form and 4th column contain string data that is very important because it contains crop name. this research is especially on crops so in the label column only the target crop name is mentioned. The image given below represents the dataset.

Table 1: Dataset values

	Temperature	Humidity	ph	label
0	20.879744	82.002744	6.502985	Spinach
1	21.770462	80.319644	7.038096	Spinach
2	23.004459	82.320763	7.840207	Spinach
3	26.491096	80.158363	6.980401	Spinach
4	20.130175	81.604873	7.628473	Spinach

3.2 Data pre-processing

After data collection by using different sensors, separate data attributes that are used in this research, and remove unnecessary data. Arrange data in rows crops-wise and Rewrite missing values. A descriptive form of data is given below.

Table 2: Dataset values after pre-processing

Attributes	***Temperature***	Humidity	ph
Count	2200.00000	2200.0000	2200.0000
mean	25.616244	71.481779	6.469480

std	5.063749	22.263812	0.773938
Min	8.825675	14.258040	3.504752
25%	22.769375	60.261953	5.971693
50%	25.598693	80.473146	6.425045
75%	28.561654	89.948771	6.923643

3.3 Data splitting

Data splitting is an important process in which a dataset is splitting into training and testing data. This process is very useful in any machine learning process as the main idea of machine learning depends on training and testing data and finding the accuracy of the machine-provided result. The algorithms were trained and apply this algorithm to test the set and measure the accuracy of the machine. Dataset is divided into 80:20 ratio, thus 80% of data is chosen as training set and the remaining 20% as the test set. There are many built-in python tool kits for splitting data such as pandas, Keras, sci-kit-learn, etc. In this research sci-kit-learn library is used for the machine learning approaches because of its built-in libraries.

3.4 Algorithm Fitting

The most crucial part of the model was to fit the algorithm with the data. All the algorithms were easily fitted as the programming of this part was comparatively easy. Simple method callings were all that was required. The algorithms, upon being implemented, processed all the data using all the internal calculations. Data frames were created and could be viewed in the variable explorer. Because the data had been split into training and testing datasets, the algorithm could start the core process: learning. The machine-learned from the train set. This learning was to be used later while predicting from the test set. Fitting is like training.

3.5 Testing accuracy

To test the accuracy, implemented different methods on different algorithms based on the requirement. Some were direct accuracy-check method calls from scikit-learn libraries. While in some other algorithms, implemented manual accuracy checks, again based on the algorithm itself. The accuracy check is crucial in understanding the viability of the algorithms and the research itself. Very low accuracy in all the algorithms would mean the entire research was a dead end. It would mean this method altogether is not viable for this research. Low accuracy in a few algorithms and high accuracy in the others would mean the ones with the low accuracy are not efficient in this model, but the others are. We would have discarded the low accuracy yielding algorithms. An Effective Model is a model which predicts the testing data most accurately as compared to other models and hence, can be deployed successfully.

3.6 Data post-processing

After all the accuracy has been considered, a few other data processes can still be implemented. This part of the model is not necessary for the primary target of the research, but we still used it for certain confirmation purposes. We implemented a method that would create a confusion matrix. A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if we have more than two classes in our dataset. Calculating a confusion matrix can give us a better idea of what our classification model is getting right and what types of errors it is making. Fig. shows the confusion matrix by implementing a Decision tree for crop prediction.

4. Result and Analysis

In this paper four different machine learning algorithms are implemented and perform their comparisons. Our target is to establish the best performing algorithm for this field of work, based on our data.
After getting results to form all algorithms now the important part of the research is that to the comparison between algorithms and select the best one for spinach cultivation. So that farmers can use this algorithm to predict crops. The random forest has 88.4% accuracy, the K-Nearest Neighbors algorithm has 84.5454% accuracy, the Support Vector Machine (SVM) algorithm has 75% accuracy, Decision Tree (DT) algorithm has 77.95% accuracy on the dataset. In the mention algorithms, Random Forest has maximum accuracy among others, so Ramadan Forest is the best algorithm for spinach crop selection. Table show an accurate comparison.

Table 3: Model accuracy

Model	Accuracy
Random Forest (RF)	88.4%
Support Vector Machine (SVM)	75%
K-Nearest Neighbors (KNN)	84.54%
Decision Tree (DT)	77.95%

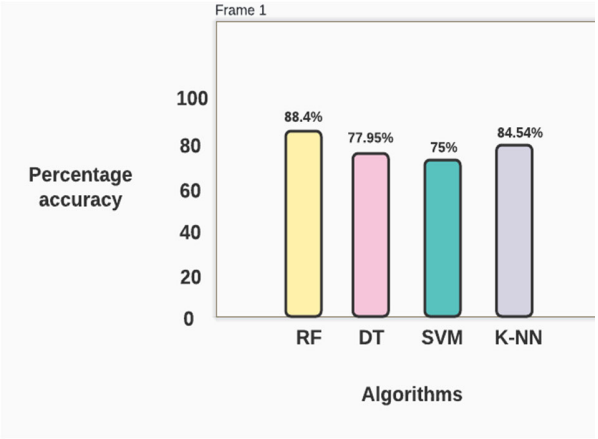

Figure 2: Graphical representation of results

4.1 Confusion matrix

For the evaluation of the machine learning model confusion matrix is the best technique. This technique summarizes the performance of an algorithm. Confusion matrix calculation gives a better idea about model accuracy and the types of errors that it's making. In classification, a problem confusion matrix is a summary of the production results. In this number of correct and incorrect predictions are summarized with count value. This is the main point of working on the confusion matrix. It shows detail about classification model is confused when it makes predictions, also it provides detail about the type of errors in the model.

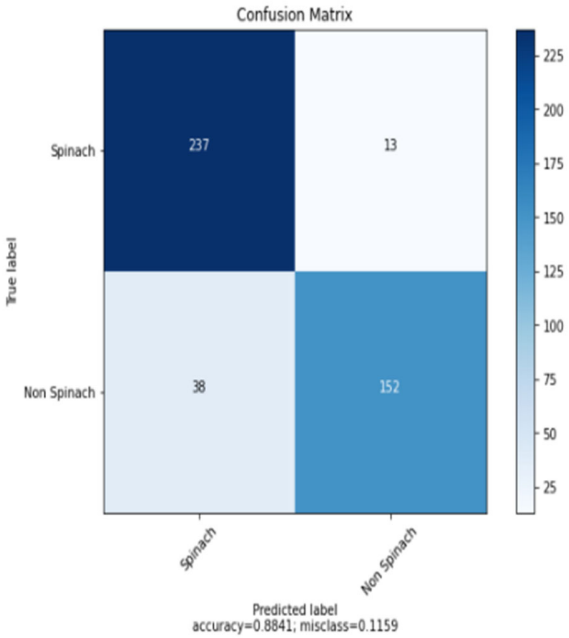

Figure 3: Confusion matrix of Random Forest

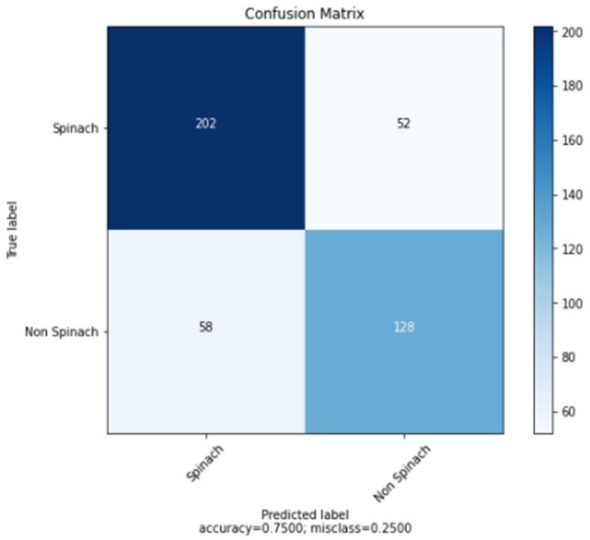

Figure 4: Confusion matrix of SVM

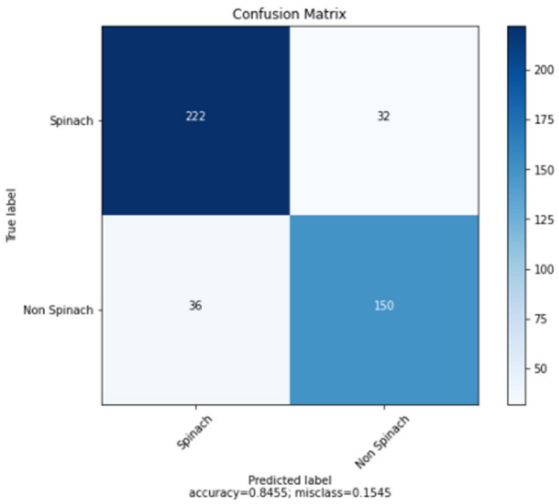

Figure 5: Confusion matrix of KNN

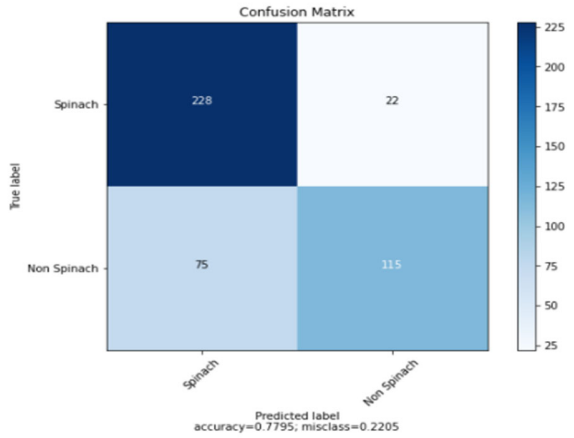

Figure 6: Confusion matrix of Decision Tree

5. Conclusion and Future Work

This Research can play a vital role in today's world of Agriculture. And agriculture fundamental aspect of modern civilization. With increasing world hunger and economic breakdown, the proper selection of crop emerges as a massive factor in this. This proposed model can predict the proper crop for a particular piece of land in a way that is very efficient. We have implemented 4 different types of machine learning algorithms in this research. All the accuracy of the models was carefully obtained through various methods and compared with each other. Using multiple algorithms helped to understand which algorithm is more suitable for this system. The crops can be predicted based on a very suitable set of features included in the dataset used. This research work found that, the cleaner the data, the better the accuracy of the result. The entire length of this research was very enjoyable, as Author was able to work in the field of machine learning. Some of the python library usage, algorithm fitting, and accuracy checking methods were very interesting in practicality. Author trust all algorithms and research work to efficiently work on any platform and any new type of data. The predictions made were solid and robust. Such strength in the model delights us, and Author hopes this keeps working over the years without issues.
In the future, this model would be implemented with a much more efficient dataset for a specific piece of land containing information such as different soil properties, different mineral percentages, etc. The accuracy values we obtained in the specific crop prediction part of our research were very poor to our target standards. In the future, implement ANN for crop prediction and check its viability on this.

References

[1] A. Zafar, S. J. I. J. o. A. R. i. A. Mustafa, Finance, and M. Sciences, "SMEs and its role in economic and socio-economic development of Pakistan," vol. 6, no. 4, 2017.
[2] B. M. Josephine et al., "Crop Yield Prediction Using Machine Learning," vol. 9, no. 02, 2020.
[3] S. S. Nair, C. Lueis, and V. Balachandran, "Crop Selection using IoT and Machine Learning."
[4] *T. Van Klompenburg, A. Kassahun, C. J. C. Catal, and E. i. Agriculture, "Crop yield prediction using machine learning: A systematic literature review," vol. 177, p. 105709, 2020.*
[5] *S. S. Dahikar, S. V. J. I. j. o. i. r. i. e. Rode, electronics, instrumentation, and c. engineering, "Agricultural crop yield prediction using artificial neural network approach," vol. 2, no. 1, pp. 683-686, 2014.*

Zaheer Ud Din Babar received the BS and MS degrees, from University of Okara in 2018 and 2021. He has been a Lecturer at ILM College Renala since 2018. His research interest includes Image processing, Machine learning and Internet of Things, and their application related to agriculture.

Muhamad Nabeel Sarwar received his B.S degree in Computer Science from The Bahauddin Zakariya University Multan (BZU), Pakistan, in 2019. His Final Year Project "Fake news detection using machine learning". He earned his M.S degree in Computer Science from University of Okara, Punjab, Pakistan. His research interest includes machine learning, deep learning, natural language processing and computer vision. He has been part of research projects such as lungs cancer detection, early diagnosis of covid-19 and sentimental analysis. During his MS degree He was research assistant at university of Okara. Currently he is teaching as visiting faculty at university of okara Pakistan alongside his business setup as entrepreneur.

Riaz UlAmin (Associate Professor, Faculty of Computing, University of Okara) has been serving in various departments for over 17 years. He has over 30 publications and completed several funded research projects. His core research expertise is in the field of distributed systems, System Security in general and Digital forensics in particular. He has worked for industry providing viable solution using different process mining techniques and tools such as DISCO and PROM. While working with different forensic tools and approaches he has an extensive exposure to different ISO standards such as ISO 17799 and COBIT. Currently, he is leading research group who is investigating digital forensics to support local context with problem in several dimensions such as complexity, diversity, consistency and correlation; volume and universal time lining.

Sidra Jabeen earned her bachelor's degree in computer science with distinction from COMSATS institute of information technology, Lahore Pakistan in 2013. Her final year project" Enhancing IEEE 802.15.7 using OFDM and WDM" won grass root research initiative award in 2013. She is currently a student of master in computer science at University of Okara. Her areas of research interest are computer communication and networking, machine learning, Artificial intelligence and practical application of computer science in education. She is working as Secondary school educator in Punjab school education department since 2015 in okara. Her pervious publications are in local conference as well as in international journal of science and technology.

Muhammad Abdullah Phd scholar, Zhengzhou University, chins. MSCS from university of agriculture Faisalabad. I have expertise in machine learning, deep learning and NLP. I have a good interest in NLP and its implications in different fields.