A HYBRID APPROACH FOR CROP YIELD PREDICTION USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

A HYBRID APPROACH FOR CROP YIELD PREDICTION USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

Sonal Agarwal[1] **and Sandhya Tarar**[2]

School of ICT, Gautam Buddha University, Greater Noida, India

Sonal04agarwal@gmail.com[1] and tarar.sandhya@gmail.com[2]

Abstract. Agriculture is defined as the science and art of cultivating the flora and fauna. Farming in India is ranked as second around the globe and occupies 60.45% of Indian land. The Indian economy, dominatingly, depends upon farming along with agro-industry things. The soil ingredients (like Nitrogen, Phosphorous, Potassium), crop rotation, soil clamminess, atmospheric and surface temperature, precipitation, etc, play an efficient role in cultivation. The present evidence related to this field includes a model which is incorporated with ML algorithms (Random Forest, Decision Tree, Artificial Neural Network) to determine best crop. In this paper, the proposed model is enhanced by applying deep learning techniques and along with the prediction of crop, a clear information is achieved regarding the amounts of soil ingredients needed with their expenses separately. It provides a better accuracy than the existing model. It analyzes the given data and help the farmers in predicting a crop which in return help in gaining profits. The climatic and soil conditions of land are taken into consideration to predict a proper yield. The objective is to present a python based system that uses strategies smartly to anticipate the most productive reap in given conditions with less expenses. In this paper, SVM is executed as Machine Learning algorithm while LSTM and RNN are used as Deep Learning algorithms.

Key Words: Agriculture, Crop Prediction, Machine Learning, Deep Learning, SVM, LSTM, RNN

1. Introduction

Cultivation is among the huge occupations experienced in the nation. An enormous improvement is achieved in the nation financially by performing various farming activities. Therefore, it is referred to as the broadest money earning method. In India, 60.45% of the land is used for cultivation. It results in fulfilling the necessities of around 1.2 billion people. The process of modernizing agribusiness is vast in the present era. Hence, the farmers are moving towards advantage and achieve greater profits in less expenses [1, 2]. The informational indexes are analyzed with the help of Data Scientific (DA) by which the inferences about the data they contain can be reached, along with the guide of specified software and framework. Traditionally, the yield was expected on the basis of a rancher's understanding on a specific land and harvest [3, 4, 5]. As the conditions are changing bit by bit, the farmers focus on building up of a regularly expanding number of harvests. Being it as the current situation, many farmers need more data about the new yields. They are completely unaware of the profits they get after the cultivation [6, 7]. Likewise, the profitability of farm can be incremented by

having a good understanding and estimation of crop execution in natural conditions [8, 9]. In this paper, the proposed framework requires the information of the area of client. The constituents of the soil such as Nitrogen, Phosphorous and Potassium are achieved from the area [10, 11]. The two more datasets are contemplated in the handling part which include crop and feature datasets, taken from the website kaggle.com, have distinct information, taken as static information [12]. The static information represents the creation of harvest and data identified about various yields obtained from different government sites [13, 14]. The proposed structure uses Artificial Intelligence (AI) and calculates for making expectation like Multiple Linear Regression to recognize the model among information. Further, it is processed as indicated including conditions. Thus, it will provide the best feasible reaps as demonstrated by given biological conditions. Hence, this system simply needs the area of the customer and suggest various beneficial yield. It provides a decision to the farmer about which harvest to develop.

2. Literature Review

Sheenoy et al. represented a paper that places an answer for the decrement in cost of transportation. The IOT-based methodology is used to decrease quantity of agents and middle hops between the clients and the ranchers that further supports the rancher. The paper ends up being the inspiration for the research work. The paper executes mechanisms that are integrated and provides a prediction-based mechanism that advise for crops which yield maximum profit [15]. Monali et al. said, to make prediction of the crop yields, an analysis of crops is made and based on it, they are classified. The classification is performed on the basis of data mining algorithms [16]. Various rules of classification such as K-Nearest Neighbor, Naive Bayes are taken care of in this paper. These rules are studied and recognized that will be exact for dataset used in this research work [17]. Abdullah et al. provided a smartphone based application that calculates the pH values for soil, humidity and temperature progressively. A microcontroller block, communication block and sensing block are used by the system. Sensors are utilized in ranch in order to set a communication link with cellphones using Bluetooth continuously. This paper gives methods for remote investigation of soil through different procedures. It urged us to search for different methods through which the information can be passed on that will be taken from sensors for development and in the end producing the yield [18]. Hemageetha et al. gave various data mining techniques such as Association Rule Mining, Classification, Clustering, Market-based analysis, Decision Trees. It totally wraps the data mining idea. In this paper, different data mining algorithms, for example, K-Means, Naïve Bayes classifier, J48 are discussed [19]. Likewise, the soil classification depending on Genetic algorithm, Naïve Bayes, Association Rule Mining. In the end, the clustering in database of soil is covered. It supported us in comprehension along with various data mining algorithms analysis. While building up the task of this research work, it ends up to be very beneficiary. It helps in dataset mining acquired from remotely used sensors [20]. Nagini et al. displayed an Explorative data study shown in this paper along with an explaination of creating a number of predictive models is given. The different regression techniques are used on a sample dataset so as to recognize and examine their properties separately. The methods explained in the paper are Linear, non-Linear, Multiple-Linear, Polynomial, Ridge and Logistic regression [21]. A comparative study of a number of algorithms in data analytics is achieved. It further helps to take a better decision of the best suitable algorithm for the proposed structure [22]. Awanit et al. gave a framework to predict the crop production for the present year is proposed in this paper. An algorithm of data mining called as K-Means, is used to get the production of crop. In this, a mechanism is also applied to predict the crop in the form of fuzzy logic. A set of rules are used on a particular land for cultivation, precipitation, and creation of crops. It is referred to as a rule-based prediction logic, also termed as Fuzzy Logic. The process of using K-Means to study the datasets is explained in this paper [23]. As a set of rules are used in fuzzy logic form, these rules are again used for prediction of the crop which will maximize the profit on the basis of cost of crops in the past years and present information of soil and weather [24].

3. Proposed Framework

In the proposed framework, the machine learning and deep learning techniques are executed in order to predict the best crop production. An experiment is done on a crop dataset by the proposed model. The crop is chosen on the basis of the current atmosphere, the soil along with its constituents as the climatic and soil parameters are taken into consideration. Deep learning is used to achieve numerous successful calculations as it is used to get the best suitable crop in case a number of options available. By using this technique, crops are predicted accurately. The SVM algorithm is implemented under machine learning while LSTM and RNN are executed under deep learning technique as shown in figure 1.

3.1. Architecture of Proposed Model

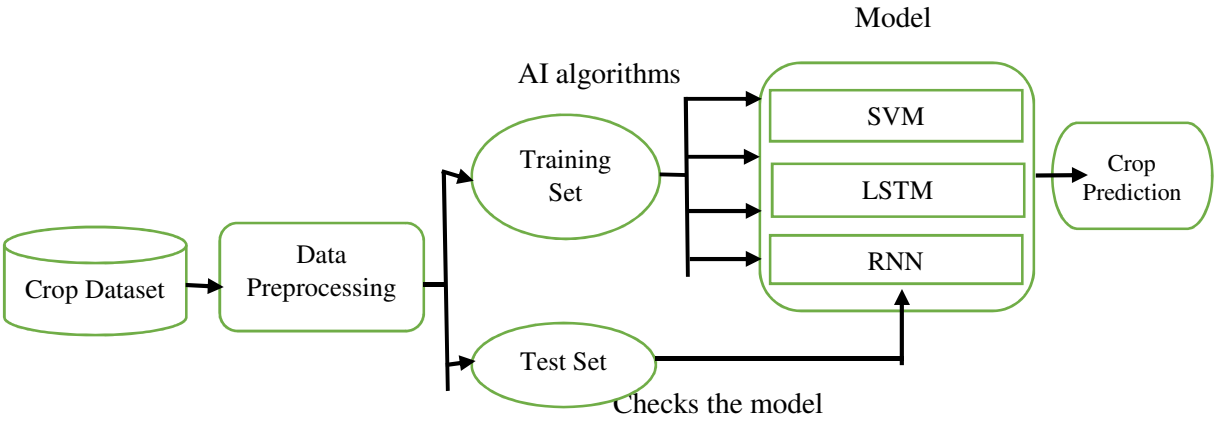

Figure 1. Architecture of proposed model used to predict crops.

Steps of implementation

Step 1: Load the crop dataset containing a number of parameters.

Step 2: Load the useful libraries and packages.

Step 3: Data preprocessing is performed.

Step 4: To prepare dataset, the data is divided in training set and testing set.

Step 5: After this, a model is constructed by applying machine learning (SVM algorithm) and deep learning (LSTM, RNN) techniques which in turn predicts the best suitable crop that needs to be grown.

Step 6: The test set checks the performance of model. If any garbage value is given as input, the model throws an error, 'value mismatch or wrong prediction'.

3.2. Dataset Description

A huge generic crop dataset containing agricultural parameters is taken to feed the model. Another dataset is taken as feature dataset. The datasets are collected from a website, named, kaggle.com. The size of the crop dataset is 7841 kb. The prediction parameters in this dataset includes temperature, rainfall, pH value, relative humidity, and area. There are a number of crops taken in this dataset like, wheat, rice, maize, millet, pea, pigeon pea, sugarcane, green gram, etc. A number of values are available for each and every prediction parameter for single crop. For instance, when taking crop as

wheat, any value can be given to the prediction parameters among a set of values available in the dataset, for wheat. It is same for the entire crops available in the dataset.

Crop	pH	Temperature	Rainfall	Relative H	Value
Wheat	5.5	13	49	40	0
Wheat	5.5	13	49	41	0
Wheat	5.5	13	49	42	0
Wheat	5.5	13	49	43	0
Wheat	5.5	13	49	44	0
Wheat	5.5	13	49	45	0
Wheat	5.5	13	49	46	0
Wheat	5.5	13	49	47	0
Wheat	5.5	13	49	48	0
Wheat	5.5	13	49	49	0
Wheat	5.5	13	51	40	0
Wheat	5.5	13	51	41	0
Wheat	5.5	13	51	42	0
Wheat	5.5	13	51	43	0
Wheat	5.5	13	51	44	0
Wheat	5.5	13	51	45	0
Wheat	5.5	13	51	46	0
Wheat	5.5	13	51	47	0
Wheat	5.5	13	51	48	0
Wheat	5.5	13	51	49	0
Wheat	5.5	13	53	40	0
Wheat	5.5	13	53	41	0
Wheat	5.5	13	53	42	0
Wheat	5.5	13	53	43	0

Figure 2. Crop Dataset.

In figure 2, a small part of the crop dataset is displayed. In this, crop is wheat, the pH value of soil is 5.5, the temperature contains a value of 13, the rainfall and relative humidity contain a number of values according to the weather conditions on different days, and a target value is assigned to every crop to label it.

3.3. Algorithms Used

3.3.1. Support Vector Machine (SVM)

Step 1: Import the required packages.

Step 2: Load input data.

Step 3: Choose the required number of features from the dataset.

Step 4: Plot SVM boundaries with the help of original data.

Step 5: Define a value for the regularization parameter.

Step 6: Finally, the object of SVM classifier is generated.

3.3.2. Long-Short Term Memory (LSTM)

Step 1: Define a neural network in Keras in the form of sequence of layers.

Step 2: Compile the network that requires various specified parameters.

Step 3: Fit the network that requires the specified training data, both an input patterns matrix X and a matching output patterns array y.

Step 4: Evaluate the network on the training data. Evaluation of the model on a test or validation set is done rarely.

Step 5: Make required predictions which can be achieved in a format given by the network's output layer.

3.3.3. Recurrent Neural Network (RNN)

Step 1: The network is provided with single time step of input.

Step 2: With the help of the previous state and the current input, compute the current state.

Step 3: The current state ht turns out to be ht-1 for the next time step.

Step 4: any number of time steps can be made depending on the problem. The information is joined from the entire previous states.

Step 5: After the completion of the entire time steps, the final current state is utilized which computes the output.

Step 6: In order for updation in the weights, the error is back-propagated towards the network. Hence, RNN is trained.

4. Method Evaluation

The performance of the proposed model is measured so as to get the result. There are certain formulae that are used in order to get the accuracy of the result. These formulae are as follows:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

where TP is defined as, when the cases are predicted as positive and are positive in actual. FP is, when the cases are predicted as positive and are negative in actual. TN is, when the cases are predicted as negative and are negative in actual. FN is, when the cases are predicted as negative and are positive in actual.

5. Implementation and Result

The implementation of the research work gets started by loading the dataset of crops that has been gathered. It is proceeded with importsing the necessary libraries and packages and continued with performing data preprocessing. The data is spat into trained data and test data. Finally, a model is constructed in which required AI algorithms are utilized which in return will provide the best suitable crop that should be grown on a particular land.

5.1 Result Analysis

The research work gets implemented by using a crop dataset which is collected from a website, kaggle.com. It contains various crops, for example, wheat, rice, maize, millets, green gram, pea, pigeon pea, sugarcane and many more. It is included with a few prediction parameters like, pH value, temperature, rainfall, relative humidity, and area. For a predictive model, the machine learning and deep learning algorithms need two types of data, namely, Trained set and Test set. The Trained data is the collected survey data that has been accumulated from the past events. While the current survey data is the Test data.

After running the code, the response is represented graphically for each one of the parameters that have been taken initially. These parameters include fertilizer usage, pesticides, area, uv and water. The yield is predicted on the basis of the data for these parameters:

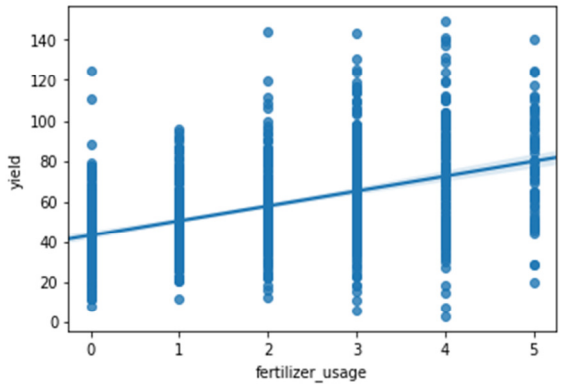

Figure 3. Yield vs Fertilizer usage.

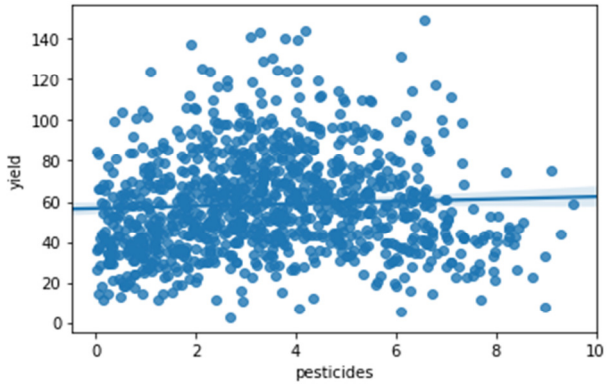

Figure 4. Yield vs Pesticides.

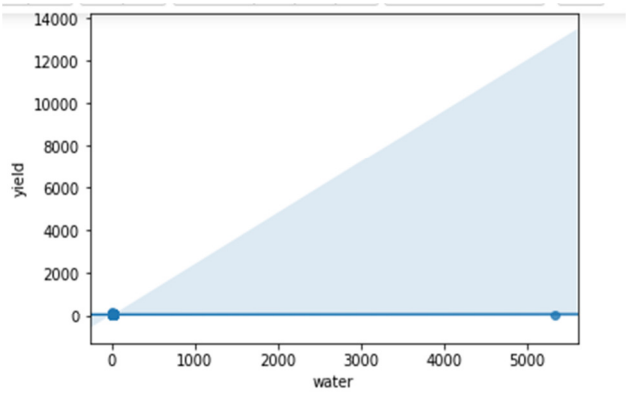

Figure 5. Yield vs Water.

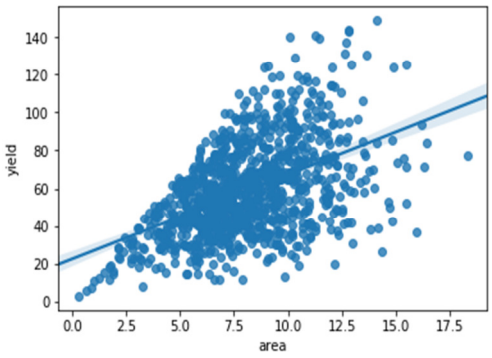

Figure 6. Yield vs Area.

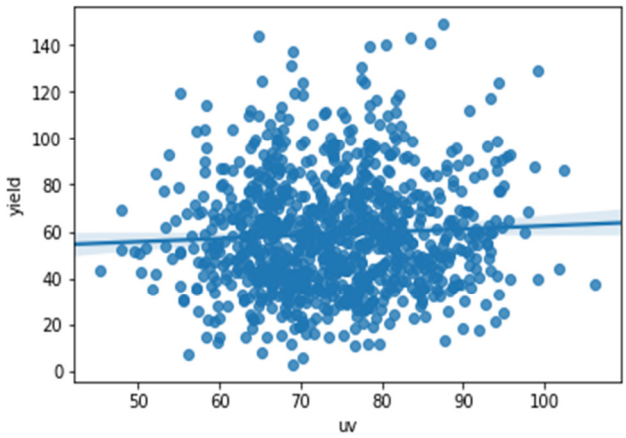

Figure 7. Yield vs uv.

The graphs drawn above are displaying the yield with respect to the features taken i.e., fertilizer_usage, pesticides, water, area, uv. The SVM classifier is used in plotting these graphs. The data points in a scatter plot are plotted on a particular place that coordinate values i.e., the x-value and the y-value. It means a data point is an intersection of x-coordinate and y-coordinate. In figure 3, the yield is determined w.r.t. fertilizer_usage. In figure 4, the yield is calculated against pesticides. In figure 5, the yield is determined by analyzing water. In figure 6, the yield is decided on the basis of area. In figure 7, the yield is calculated against uv. In all these cases, the values of features are taken from the feature dataset and hence, the yield is calculated.

5.2. Performance Evaluation

There are certain parameters on which the dataset is taken into analysis. These parameters are temperature, rainfall, location of soil, relative humidity and area. These are the features with the help of which the dataset is analyzed and the better yield is predicted. There are certain crops that are taken for yielding on a particular piece of land. These crops include wheat, rice, maize, pea, pigeon pea, green gram, potato, sugarcane, soyabean, etc.

Table 1. Performance Analysis.

Algorithm	**Features**	**Crops**	**Accuracy**
DT, ANN, RF (used in existing paper)	Temperature, Rainfall, Location, Soil	Wheat, Rice, Soyabean	93%
LSTM, RNN, SVM (used in proposed work)	Temperature, Rainfall, pH value, relative humidity, area	Wheat, Rice, Maize, Millets, Pea, Pigeon Pea, Green Gram, Soyabean, Sugarcane	97%

In table 1, it is shown that there are certain features on the basis of which the dataset is analyzed over a set of crops. When applying Artificial Neural Network (ANN), and Random Forest algorithms, the accuracy is calculated as 93%. Whereas, when applying Long-Short Term Memory (LSTM), Recurrent Neural Network (RNN) and Support Vector Machine (SVM) algorithms, the accuracy is calculated as 97%. Therefore, it can be said clearly that along with Machine Learning algorithm, the Deep Learning algorithms play a vital role in predicting the yield with better accuracy.

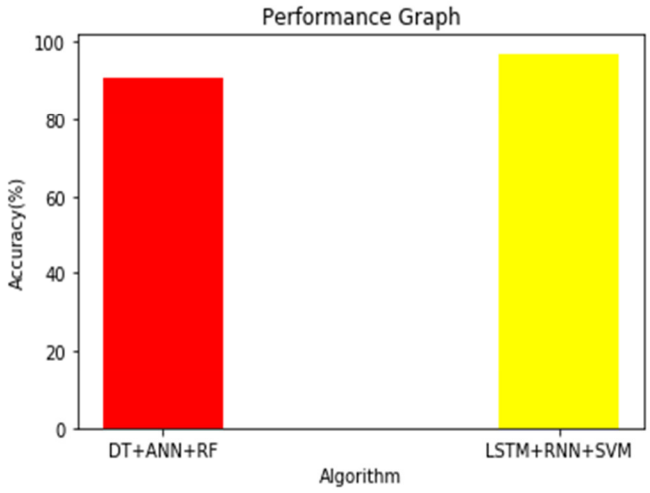

Figure 8. Performance Graph displaying accuracy level.

In figure 8, the accuracy is shown graphically for both the approaches. The first bar i.e. the red bar has shown the accuracy as 93% when ANN and Random Forest algorithms are used. Whereas, the second bar i.e., the yellow bar has shown the accuracy as 97% when applying LSTM, RNN and SVM algorithms together. Hence, the second result gives better accuracy.

6. Conclusion and Future Scope

The proposed model is constructed by using AI algorithms to reduce the farmers' problems of getting losses in their farms due to lack of knowledge of cultivation in different soil and weather conditions. The model is created by using machine learning (SVM) and deep learning (LSTM, RNN) techniques. The model predicts best crops that should be grown on land with less expenses among a number of crops available after analyzing the prediction parameters. To the best of studies, there is no such work in existence that uses the same techniques in predicting the crops. Hence, it is concluded that there is an enhancement in the accuracy of this research work when compared to the existing work that used another techniques for prediction of crops. The accuracy is calculated as 97%. It has a vast extension in future and can be actualized and interfaced with a flexible and multi-skilled application. The farmers need to be educated and hence, will get a clear information regarding best crop yield on their mobiles. With this, even if the rancher is at home, the work can be managed at that particular instant of time, without facing any kind of loss ahead. The progress in the agribusiness field will be extremely appreciable which will further result in helping the farmers in production of crops.

References

[1] Bendre M R, Thool R C and Thool V R September 2015 Big Data in Precision agriculture *NGCT*

[2] Suma N, Samson S R, Saranya S, Shanmugapriya G and Subhashri R February 2017 IOT Based Smart Agriculture Monitoring System *IJRITCC*

[3] Dhivya B, Manjula, Bharathi S and Madhumathi March 2017 A Survey on Crop Yield Prediction based on Agricultural Data *International Conference in Modern Science and Engineering*

[4] Ravichandran G, Koteeshwari R S 2016 Agricultural Crop Predictor and Advisor using ANN for Smartphones *IEEE*

[5] Kitchen N R, Lundb S T D, Sudduth K A and Buchleiter G W 2003 Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems *Agronomy Journal* **95** 483-495

[6] Ezrin M H 2009 Relationship between apparent electrical conductivity of paddy soil and rice yield, M.S. Thesis Institute of Advance Technology, Universiti Putra Malaysia

[7] Pantazi X E, Tamouridou A A, Alexandridis T K, Lagopodi A L, Kashefi J and Moshou D 2017 Evaluation of hierarchical self-organising maps for weed mapping using UAS multispectral imagery *Comput. Electron. Agric.* **139** 224–230

[8] Zhang M, Li C and Yang F 2017 Classification of foreign matter embedded inside cotton lint using short wave infrared (SWIR) hyperspectral transmittance imaging *Comput. Electron. Agric.* **139** 75–90

[9] Ming J, Zhang L, Sun J and Zhang Y April 2018 Analysis models of technical and economic data of mining enterprises based on big data analysis *IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* pp. 224–227

[10] Šťastný J, Konečný V and Trenz O 2011 Agricultural data prediction by means of neural network vol 7 pp 356–361

[11] Veenadhari S, Misra B and Singh C D March 2011 Data mining Techniques for Predicting Crop Productivity – A review article *International Journal of Computer Science and technology*

[12] Maduskar V and Kelkar Y February 2012 Survey on Data Mining *International Journal of Emerging Technology and Advanced Engineering* Website: www.ijetae.com ISSN 2250-