International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Evaluation of Machine Learning Algorithms for Crop Yield Prediction

Renuka, Sujata Terdal

Abstract: ***Agriculture plays a significant role in the growth of the national economy. It relay on weather and other environmental aspects. Some of the factors on which agriculture is dependent are Soil, climate, flooding, fertilizers, temperature, precipitation, crops, insecticides and herb. The crop yield is dependent on these factors and hence difficult to predict. To know the status of crop production, in this work we perform descriptive study on agricultural data using various machine learning techniques. Crop yield estimates include estimating crop yields from available historical data such as precipitation data, soil data, and historic crop yields. This prediction will help farmers to predict crop yield before farming. Here we are utilizing three datasets like as clay dataset, precipitation dataset, and production dataset of Karnataka state, then we structure an assembled data sets and on this dataset we employ three different algorithms to get the genuine assessed yield and the precision of three different methods. K-Nearest Neighbor(KNN), Support Vector Machine(SVM), and Decision tree algorithms are applied on the training dataset and are tested with the test dataset, and the implementation of these algorithms is done using python programming and spyder tool. The performance comparison of algorithms is shown using mean absolute error, cross validation and accuracy and it is found that Decision tree is giving accuracy of 99% with very less mean square error(MSE). The proposed model can exhibit the precise expense of assessed crop yield and it is mark like as LOW, MID, and HIGH.***

Keywords: ***Descriptive analytics, Agriculture, Machine learning, K-Nearest Neighbour, Support Vector Machine, Decision tree.***

I. INTRODUCTION

Agriculture is one of the important industrial sectors in India and the country's economy relies heavily on it for the sustainability of its rural areas. Due to some factors [1] such as climate change, unplanned rainfall, falling water levels, excessive use of pesticides etc., the level of agricultural production in India is declining. Most farmers do not achieve expected crop yield for a variety of reasons. To understand production levels, yield prediction is carried out which involves predicting the yield of the crop based on the existing data. Previously, crop yield estimates were based on farmer's specific crops and cultivation experience. There are many ways to enhance and improve crop yield and quality. Data mining techniques are also helpful for predicting crop yields. In general, data mining analyzes knowledge from various approaches and summarizes it as profitable information. Data Mining software [2] is an analytical tool that allows users to classify and summarize identified relationships as well as analyze data at various angles or dimensions. Technically, data mining is finding correlations or patterns of fields in large relational databases. All of these data can provide information between models, connections, or relationships. Knowledge can be transformed into historical patterns and knowledge of future trends. For example, a survey of agricultural products helps farmers to suggest and prevent future crop losses.

Many research have been conducted to develop an efficient method for yield prediction but focus have been always on statistical techniques and not much has been done in machine learning approach. The crop production depends on various factors [3] which change with every square meter and depends on:

1. Geography of region,
2. Weather (Temprature, humidity ,percipitation),
3. Soil type (saline, alkaline,sodic,non-alkaline),
4..Soilcomposition(ph,N,P,K,EC,OC,Zn,F).

Various subsets of these parameters are used in different prediction models for various crops. Prediction models are essentially two main types.

1. Statistical models, which use a single prediction function that includes all sample spaces.
2. Machine learning technology, a new technology for knowledge search that connects input and output variable models. Machine learning has the ability to learn the machine without defined computer programming, so it improves machine performance by detecting and characterizing the consistency and pattern of drive data. Machine learning can be classified into three categories according to the learning method –Supervised learning, Unsupervised learning and Reinforcement learning. In our project we are making use of supervised learning algorithms to predict crop yield. This type of algorithms helps to build most accurate and effective model because here, the learning data comes with labels or desired outputs and the objective is to find a general rule of mapping input to output. It involves building a machine learning model that is based on labeled samples.

The proposed system analyzes the application of supervised machine learning approaches in forecasting sugar cane yield of Karnataka region.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Renuka, Computer Science and Engineering, PDACE, Kalaburagi, India.

Dr. Sujata Terdal, Computer Science and Engineering, PDACE, Kalaburagi, India.

The dataset is collected from several online sources such as Kaggle.com and data.gov.in,
it consist of rainfall dataset, yield dataset and soil dataset. The SVM, KNN and Decision trees have been used for yield prediction. This paper is organized as follows : Section II presents the related work ,whereas the proposed method is discussed in Section III. Then, the experimental results on agricultural data are discussed in Section IV. Finally, conclusion is given in Section V.

II. RELATED WORK

Machine learning in Agriculture is a Novel field, a great deal of work has been done in field of Agriculture utilizing Machine learning. There are diverse guaging philosophies created and assessed by the specialists everywhere throughout the world in the field of farming or related sciences. Agricultural scientists in Pakistan have demonstrated that endeavors of harvest yield amplification through expert pesticide state strategies have prompted a hazardously high pesticide use. These examinations have revealed negative relationship between's pesticide use and harvest yield [5]. In their investigation they have explained that how data mining incorporated farming information including irritation exploring, pesticide utilization and meteorological information are helpful for streamlining of pesticide use. Topical data identified with agribusiness which has spatial properties was accounted for in one of the study[6]. Their research went for perceiving patterns in farming creation with references to the accessibility of information assets. K-means method turned into applied to carry out gauges of the contamination in the air[7],the k- nearest neighbor become connected for mimicking day by day precipitations and other climate elements [8],and numerous ability changes of the weather situations are dissected utilizing SVM[9]. Statistics mining techniques are often used to have a look at soil qualities. As example, the k-means method is used for segmenting soils in mixture with GPS-based technology [10]. A decision tree classifier for agriculture information turned into proposed [11].This new classifier uses new facts expression and can address each entire records and in entire records. Inside the test,10-fold cross validation technique is used to check the dataset, horse-colic dataset and soybean dataset. Their results showed the proposed selection tree is capable of classifying all styles of agriculture records. A yield prediction version turned into proposed in one of the take a look at [12] which makes use of data mining techniques for category and prediction. This model worked on enter parameters crop name, land location, soil type, soil ph, pest information, climate, water stage, seed type and this model anticipated the plant boom and plant diseases and therefore enabled to select the nice crop based on climate information and required parameters.There are few research works about sugarcane yield prediction which can be associated with our work. Sugarcane yield prediction technique with use of Random forest [13] became proposed in one of the survey, the features used in this study consist of biomass index, climate statistics (e.g., rainfall) and yields from previous years. Two predictive tasks are provided in [13]: (i) the category problem for predicting whether or not the yield can be above or underneath the found median yield, and (ii) the regression hassle for predicting the yield estimates in two distinct time intervals. In addition, support vector system[14] for rice crop yield prediction become proposed, the dataset used in this method are precipitation ,minimum, maximum and common temperature, place, evapotranspiration and manufacturing. The sequential minimal optimization classifier is implemented on the dataset. The dataset is processed through WEKA tool to build the set of rules on the current dataset. The results were generated in python by using SVM algorithm. In [15] based on the C4.5 algorithm, decision tree and decision rules have been developed, in their study they have developed a website called Crop Advisor: This an interactive website for discovering the affect of weather and crop production by using C4.5 algorithm[15]. This gives the idea of how different climatic parameters impact the growth of the crop. The selections were made based on the area under the chosen crop. The information regarding the associated years climatic parameters like rainfall, high and low temperature, wet day frequency where collected. The id3 algorithm [16] were developed to get good quality and improved Tomato crop yield which is implemented in PHP platform and uses csv as data sets. The features used in this study include area, production of tomato crop, temperature and humidity.

III. PROPOSED SYSTEM

In the proposed system, we use supervised learning to form a model, which provides predicted cost of crop yield and corresponding production order. The proposed system is described in following stages such as dataset collection, preprocessing step, feature selection and applying machine learning modules as shown in figure 1.

A. Dataset Collection: Data is collected from a variety of sources and prepared for data sets. And this data is used for descriptive analysis. Data is available from several online abstract sources such as Kaggle.com and data.gov.in. We will use an annual summary of crops for at least 10 years. The data sets used in this paper are soil dataset, rainfall dataset and crop yield data.

B. Preprocessing step: This step is a very important step in machine learning. Preprocessing consists of inserting the missing values, the appropriate data range, and extracting the functionality. The kind of the dataset is critical to the analysis process. In this paper we have used isnull() method for checking null values and lable Encoder() for converting the categorical data into numerical data.

C. Feature Selection: Feature extraction should simplify the amount of data involved to represent a large data set. The soil and crop characteristics extracted from the pre-treatment phase constitute the final set of training. These characteristics include the physical and chemical properties of the soil. Here, we have used RandomForestClassifier() method for feature selection. This method selects the features based on the entropy value i.e., the attribute which is having more entropy value is selected as important feature for yield prediction.

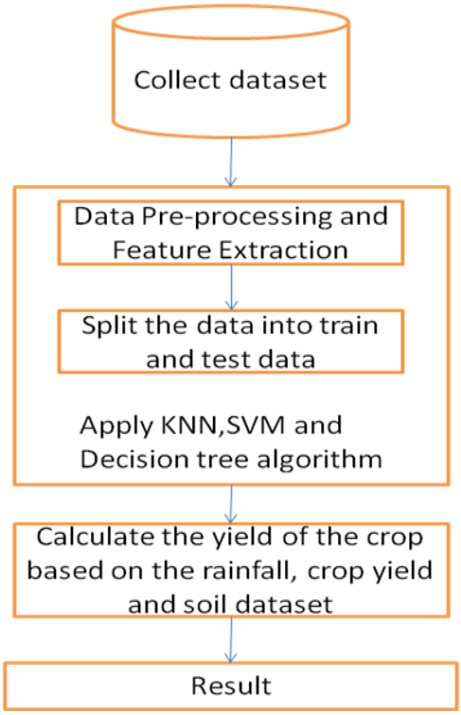

Fig.1 System Architecture

D. Split the Dataset into Train and Test Set: This step includes training and testing of input data. The loaded data is divided into two sets, such as training data and test data, with a division ratio of 80% or 20%, such as 0.8 or 0.2. In a learning set, a classifier is used to form the available input data. In this step, create the classifier's support data and preconceptions to approximate and classify the function. During the test phase, the data is tested.The final data is formed during preprocessing and is processed by the machine learning module.

***E. Applying Machine Learning modules*:** In our project we have used three different supervised machine learning algorithms for crop yield prediction which is given as follows

i. KNN Algorithm

KNN is a nonparametric supervised learning technique that uses training sets to segment data points into given categories. In simple classifications, the word collects information from all educational cases and similarities based on the new case. Look at the training for the most similar (neighbor) K cases and predict the new instance (x) by summarizing the output variables for these K cases. Classification is the class value mode (or most commonly). A flow diagram of the KNN algorithm is shown in Figure 2.

Flow chart:

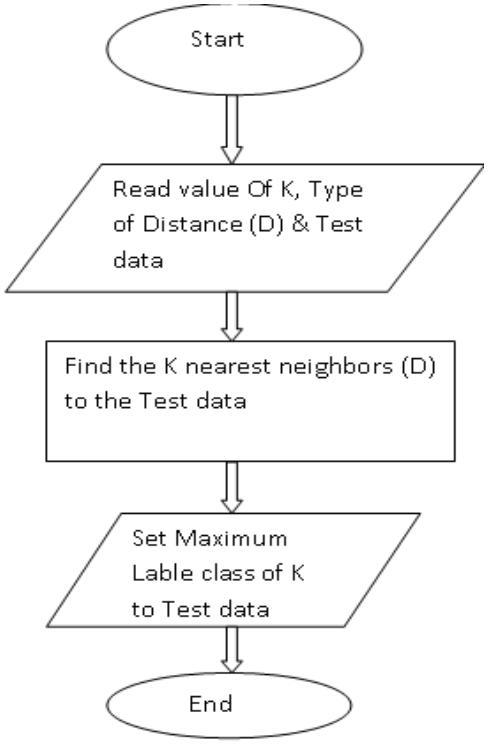

Fig.2 Flow chart for KNN algorithm

ii. Support Vector Machines

SVM divides the given data into decision surface. Decision surface is further divide the data into hyper plane of two classes. Training points defines the supporting vector which defines the hyper plane. Probably, a hyper plane with the greatest distance to the closest learning data point typically has better margins and larger errors because of the larger margins, the generalization of classifiers is weak. The flow chart for SVM is given in the figure 3, it shows the steps involved in SVM algorithm.

Flow Chart:

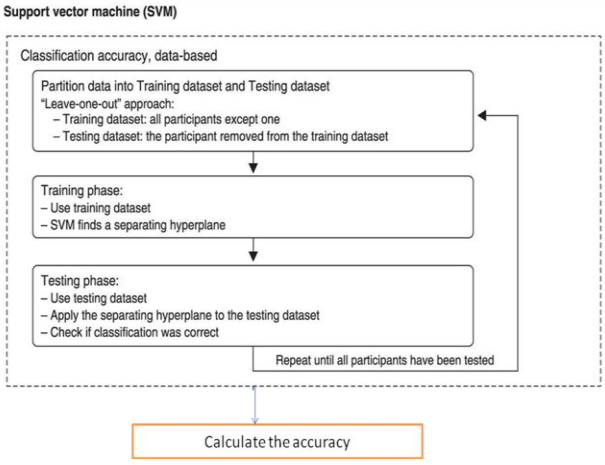

Fig.3 Flow chart for SVM algorithm

iii. Decision tree

Decision tree is a predictive model which works by checking condition at every level of the tree and proceeds towards bottom of the tree where various decisions are listed.

The condition depends on the application and the outcome might be in terms of decision. There are various types of Decision tree algorithms such as C4.5,CART and ID3 algorithm. In this paper we are using ID3 algorithm for model building. The factors like Rainfall, soil type and crop yield can be used as input to Prediction algorithm and the resultant will be in terms of the prediction of productivity.

The ID3 algorithm [15] starts with the original set S as the root node. Each iteration of the algorithm finds each unused attribute in set S and computes the entropy H (S) (or information gain IG (S)) of this attribute. Then select the attribute with the smallest entropy value (or the largest information gain). The set S is then divided into the selected attributes. The algorithm is continually reproduced in each subset, considering only those attributes that were not previously selected. The figure 4 shows the flow diagram of the ID3 algorithm.squared error, accuracy and cross validation which are used to estimate the efficiency of the method as shown in the table 1.

The formula for calculating Accuracy and Mean Squared Error(MSE) are:

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

Machine learning algorithms have been applied individually using the Cross Validation techniques with 10 folds and accuracy of prediction has been observed for each of them. In this paper the accuracy for SVM was calculated for two different kernels i.e, svm _rbf and svm_linear among these two RBF kernel was showing more error rate. The decision tree is giving more accuracy with very less MSE. The table 2 shows cross validation runs for different algorithms.

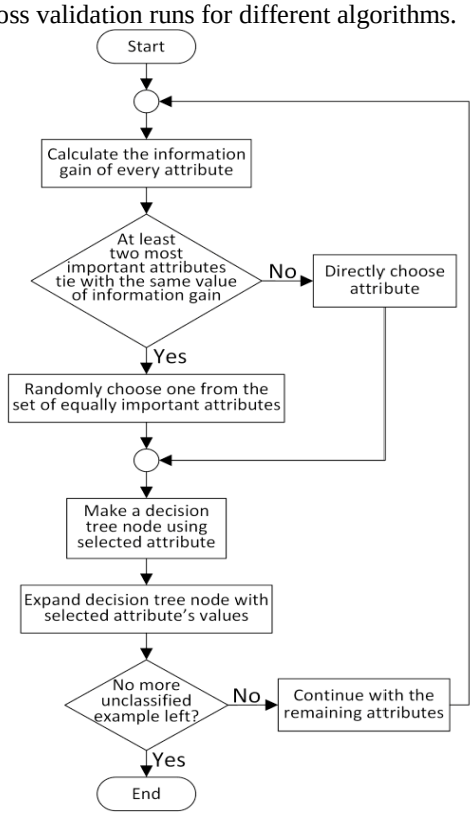

Fig.4 Flow chart for ID3 algorithm

Once the model has been trained efficiently it is tested on the Testing dataset which is different from the Training data in sample values.

IV. DISCUSSION AND RESULTS

Table 1.Comparision table for different parameters

Algorithms	Accuracy	MSE
Decision tree	99	0.01
KNN	58.078	0.603049
SVM_rbf	58.952	0.559123
SVM_linear	53.712	6.209897

This section shows the results obtained after implementation of the Machine learning algorithms on sugar cane crop data set of Karnataka state, India. The algorithms KNN, SVM and decision tree are applied on yield dataset, soil dataset and rainfall dataset. In this paper, we have used Spyder an open source, multi-platform integrated development (IDE) environment for Python scientific programming. The different parameters set for these algorithms were mean

Table 2. Cross validation runs for different algorithms

Cross validation	Decision tree	KNN	SVM_rbf	SVM_linear
1	0.992375	0.959368	-0.058233	0.9306726
2	0.945654	0.886403	-0.189782	0.8693865
3	0.814398	0.940818	-0.191459	0.9343228
4	0.896650	0.940682	- 0.140031	0.7896232
5	0.875382	0.872272	-0.197912	0.7903085
6	0.976996	0.956903	-0.136351	0.9011202
7	0.942567	0.976868	-0.116380	0.8950682
8	0.944686	0.962245	-0.147855	0.8929110
9	0.909315	0.876680	-0.102809	0.7334802
10	0.949203	0.956820	-0.144903	0.8934893

The following figure 5 and 6 shows comparison graph for accuracy and mean squared value for KNN, SVM_RBF, SVM_LINEAR and Decision tree. As shown in the graph the accuracy for decision tree is more and also it is showing less error rate.

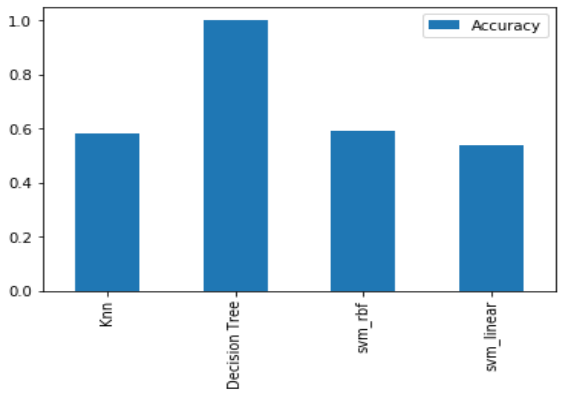

Fig.5 Comparison graph for Accuracy

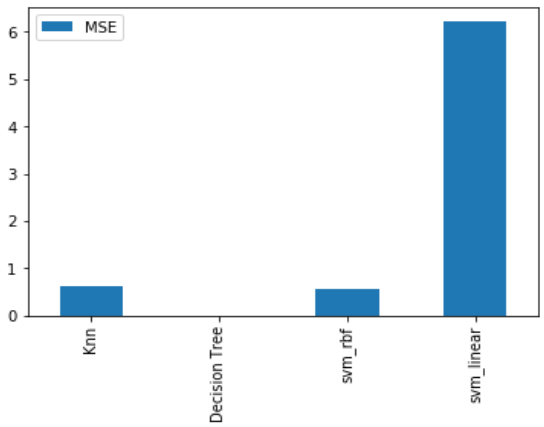

Fig.6 Comparison graph for MSE

V. CONLUSION

Different machine learning algorithms have been implemented on agricultural data to evaluate the best performing method. In this paper, we used three different supervised learning algorithms, such as SVM, KNN, and decision trees. This study provides information on how to apply data analysis to the data set of sugar cane crops. There are three sets of data: soil data sets, rainfall data sets, and crop yield data sets. This data set consists of a variety of parameters that are useful for identifying status of crops and conducting supervisory training on data sets collected from agriculture domain to divide information into multiple classes. This paper shows the comparison of three different algorithms like, decision tree, KNN and SVM. These algorithms were used to train the 0.8 or 80 percentage of the input data and are tested with the remaining 0.2 or 20 percentage of test dataset and results of the algorithms were compared based on accuracy and mean square error. Here, the decision tree algorithm is giving more accuracy of 99% and also the mean square error for this algorithm is very less. This system will help to reduce the problems faced by farmers and will serve as an intermediary to provide farmers with the information they need to earn high profits and maximize profits.

REFERENCES

1. Arun Kumar, Naveen Kumar, Vishal Vats.**"**Efficient crop yield prediction using machine learning algorithms", IJRET Volume: 05 Issue: 06, June-2018, pp 3151-3159
2. P.Priya, U.Muthaiah ,M.Balamurugan**. "** Predicting yield of the crop using machine learning Algorithm", IJESRT et al., 7(4)**:** April-2018, pp 2277-2284
3. Vaneesbeer Singh, Abid Sarwar, Vinod Sharma. **"**Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach", IJARCS 8 (5), May-June 2017, pp 1254-1259.
4. A.S. Ponraj, Vigneswaran T. "Machine Learning Approach for Agricultural IoT",IJRTE, Volume-7, Issue-6, March 2019,pp 383-391.
5. Abdullah, A., Brobst, S, Pervaiz.I., Umer M.,and A.Nisar, "Learning dynamics of pesticide abuse through data mining", Proceedings of Australian Workshop on Data mining and Web Intelligence, New Zealand, January .2004,pp 63-68.
6. Kiran Mai,C., Murali Krishna, I.V, an A.VenugopalReddy, " Data Mining o f Geospatial Database for Agriculture Related Application", Proceedings of Map India,New Delhi, 2006,pp 83-96.
7. Jorquera H, Perez R, Cipriano A, Acuna, "Short term forecasting of air pollution episodes",In. Zannetti P (eds) Environmental Modeling 4. G(2001) WITPress, UK,pp 221-237.
8. Rajagopalan B. Lall, " A k- nearestneighbor daily precipitation and other weather variables." U (1999) WatResResearch35(10) :3089 – 3101.
9. Tripathi S, Srinivas VV, Nanjudiah , "Down scaling of precipitation for climate change scenarios: a support vector machine approach", RS (2006),J. Hydrology 330-337.
10. Verheyen, K., Adrianens, M. Hermy and S.Deckers(2001). High resolution continuous soil classification using morphological soil profile descriptions. Geoderma, 101:31-48.
11. Jun Wu, Anastasiya Olesnikova, Chi- Hwa Song, Won Don Lee (2009).The Development and Application of Decision Tree for Agriculture Data. IITSI, pp 16-20.
12. R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques," 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-4.
13. Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber. "Accurate prediction of sugarcane yield using a random forest algorithm", Agronomy for Sustainable Development, vol. 36, no. 2, 2016.
14. N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, 2016, pp. 1-5. S. Veenadhari, B. Misra and C. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters," 2014 International Conference on Computer Communication and Informatics, Coimbatore, 2014, pp. 1-5.
15. CH. Vishnu Vardhan chowdary, Dr.K.Venkataramana, "Tomato Crop Yield Prediction using ID3", March 2018,IJIRT Volume 4 Issue 10 pp,663-62.

AUTHORS PROFILE

Renuka Mtech student in Computer Network and Engineering at Poojya Doddappa Appa College of Engineering, Kalaburagi, Karanataka. Her Research interest is in the area of Computer Networks and Machine Learning.

Dr. Sujata Terdal M.Tech, Ph.D, Associate Professor, Computer Science and Engineering Department have teaching experience of 23 years. Areas f research are Mobile Ad Hoc Networks and Wireless Networks. She has published number of research papers in International and National journals and conferences.