International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
Doi : https://doi.org/10.32628/CSEIT2173168

Crop Yield Prediction using Different Machine Learning Techniques

Pallavi Shankarrao Mahore, Dr. Aashish A. Bardekar
Sipna College of Engineering and Technology, Amravati, Maharashtra, India

Article Info
Volume 7, Issue 3
Page Number: 561-569

Publication Issue :
May-June-2021

Article History
Accepted : 18 June 2021
Published : 25 June 2021

ABSTRACT

Cotton, popularly known as White Gold has been an important commercial crop of National significance due to the immense influence of its rural economy. Transfer of technology to identify the quality of fibre is gaining importance for crop yield is compared with Random forest, Support Vector Machine, Weather, K Nearest neighbor. , which shows better performance results for each selected weather parameters. Crop yield rate depends upon various parameters such as the geography of area, soil type, soil nutrients, soil alkaline, weather condition, etc. The combination of these parameters can be used for selection of suitable crops for a farm or land to gain maximum yield. In this manuscript, soil and weather parameters such as soil type, soil fertility, maximum temperature, minimum temperature, rainfall are used to identify suitable crops for specified farm or land.

Keywords : Agriculture, Crop Yield, Random forest, Support Vector Machine, Weather, K Nearest neighbor.

I. INTRODUCTION

Cotton is produced in over 50 countries worldwide, averaging 20 – 24 million metric tons per year. China, the United States, India and Pakistan are the largest cotton producers, accounting for approximately 65 percent of the world cotton production alone. Brazil, Uzbekistan and other countries with smaller annual cotton crops cover the remaining 35 percent. China, India and Pakistan are also the largest consumers of cotton, accounting for app. 60 percent of the worldwide cotton consumption. Thus, most of the cotton produced is being traded and exported as a commodity in an international market[1].

The quality of cotton, however, is highly variable making it difficult to determine its commercial value or price. Cotton quality is a function of its variety, growing conditions, harvesting and ginning. Growth conditions change every year depending on the environment (weather and soil). In addition, agricultural, harvesting and ginning methods used for cotton production vary widely in different countries around the world. All these factors attribute to a wide range of cotton qualities available in the international cotton market. In cotton spinning, raw material costs make up 50-70% of the overall yarn manufacturing costs. Cotton purchasing is the highest risk for a spinner, and it is often based on trust gained over generations between cotton buyer (mill owner) and seller (merchant). Other stakeholders in the cotton supply chain are cotton seed breeders, producers, and ginners. All have a high interest in an objective method of assessing the quality of cotton. Cotton

classification provides this objective assessment of cotton quality, and it is the basis for determining the cotton price [2].

II. LITERATURE REVIEW

Mann, M. L., Warner, J. M., & Malik, A. S. et al. [1] has proposed another information combination strategy to Ethiopia which consolidates both remotely detected information (RSD) and agrarian overview information for a considerable beneficiary of specially appointed imported nourishment help. RSD is gotten close to mid-season for foreseeing significant harvest misfortunes and found at least 25% yield misfortunes at town level because of dry spell for five essential grain crops. Additionally, in the wake of establishing 81% exactness of harvest misfortunes at mid to late September creators accepted that these models can be utilized for future advancement for remotely detected information like Harmonized Landsat Sentinel (HLS) which is high goals for checking and foreseeing crop yields.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. et al. [2] has done an audit which is predominantly focuses and talked about on AI methods, yield estimation and, accuracy nitrogen the board. The survey exhibits the techinque of back proliferation significance and its precision of harvest yield expectation for various vegetation lists. They especially exhibits that guassian procedure are valuable for foreseeing and finding various qualities of plant leaves. Creators likewise audit the significance of M5-Prime Regression Trees are most appropriate device for finding numerous yield expectations. At last, this survey additionally exhibits on Fuzzy Cognitive Map (FCM) which will be utilized for crop yield expectation for model and portrayal of master information.

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., ... & Reichert, G. et al. [3] This examination depends on Integrated Canadian CropYield Forecaster (ICCYF) for three fundamental harvests at different spatial and fleeting scales shows that the exactness of the conjecture is improved over the croping season when all the more ongoing information is accessible. This gives a lead time of around multi month before har-vest and around 3–4 months before the official last arrival of the review results from Statistics Canada, which is frequently made openly accessible in December. Since the official consequences of this examination originate from the last arrival of the four yearly yield overviews. Given provincial contrasts in ICCYF quality crosswise over Canada, the discoveries were inside or superior to anything the watched yield change as detailed by Statistics Canada. It is in this way possible that, notwithstanding the may assets gave by Statistics Canada to direct four harvest yield studies during the developing season, the ICCYF may have the option to finish.

Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. et al. [4] In this paper Machine learning approaches have been utilized to create crop yield forecast models for the Canadian Prairies. This investigation concentrated on determining grain, canola and spring wheat utilizing remotely detected vegetation records and thought about MODIS-NDVI, MODIS-EVI and AVHRRNDVI viability. Mkhabela et al. (2011) results have been affirmed as it has been demonstrated that MODIS-NDVI is a successful harvest y indicator yield.

III. WORKING METHODOLOGY

Crop yield is a very useful information for farmers . It is very beneficial to know the yield which results in reduction in loss. In the past the yield prediction is done by experienced farmers. The proposed system also works in a similar way. It takes the previous information and uses it to predict the future yield.

The crop yield mainly depends on weather and pesticides. This prediction is proportional to the accuracy on information provided. Therefore, the proposed system predicts the yield and decreases the loss

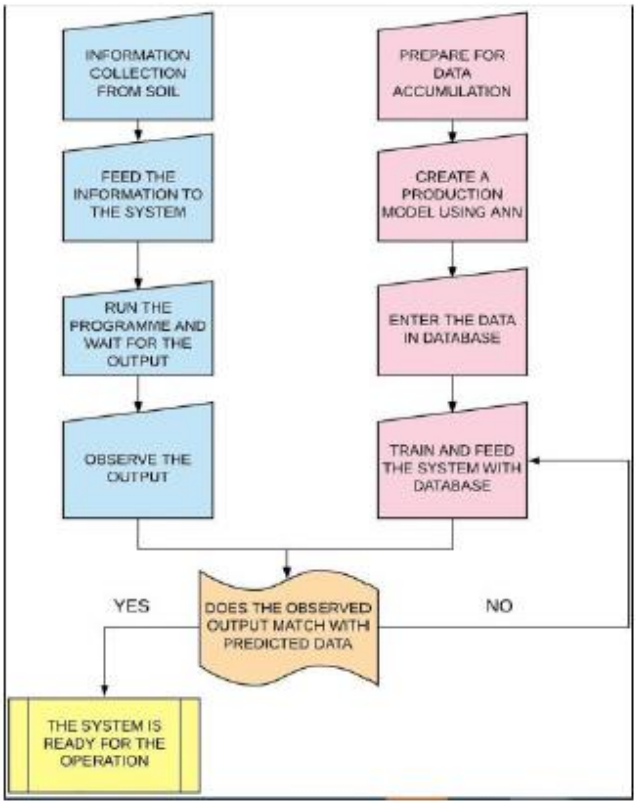


Fig 1 : Workflow of proposed system

The anticipated system acts as experienced farmer. But, with more accuracy and considers many other factors. Factors like soil condition, weather prediction, yield. The more increase in accuracy results in more profit in crop yield. To increase accuracy the data has to be perfect. With all the information provided the proposed system process all the data using data mining methods and predicts the harvest yield. With this forecast the farmer will be able to know his requirements.

IV. IMPLEMENTATION

Random Forest Machine Learning Algorithm

Random forest is a supervised learning algorithm. As the name suggests, this algorithm creates a forest and using precision techniques, makes it random. The "forest" it builds, is an ensemble of Decision Trees, which are mostly trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction[6].

At training situation multitude decision trees are made and the output will be divided based on number of classes i.e., classification, prediction of class i.e., regression. The number of trees is proportional to accuracy in prediction. The dataset includes factors like rainfall, perception, temperature and production. This factor in dataset is used for training. Only two third of the dataset is considered. Remaining dataset is used for experimental basis. A. Datasets The dataset consists of factors like temperature, rainfall, humidity, ph. The datasets have been obtained. The data set has large number of instance or data that have taken from the past historic data. It includes many parameters or features like the temperature, humidity, rainfall temperature, land type etc.

Random forest algorithm Random Forest is a ML algorithm. At training situation multitude decision trees are made and the output will be divided based on number of classes i.e., classification, prediction of class i.e., regression. The number of trees is proportional to accuracy in prediction. The dataset includes factors like rainfall, perception, temperature and production. These factors in dataset is used for

training. Only two-third of the dataset is considered. Remaining dataset is used for experimental basis [7].

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. We have implemented random forest in classification, since classification is sometimes considered the building block of machine learning, and the simple fact that it was necessary for us to use that aspect of the algorithm based on the type of research we have up taken. Figure shows random forest depiction.

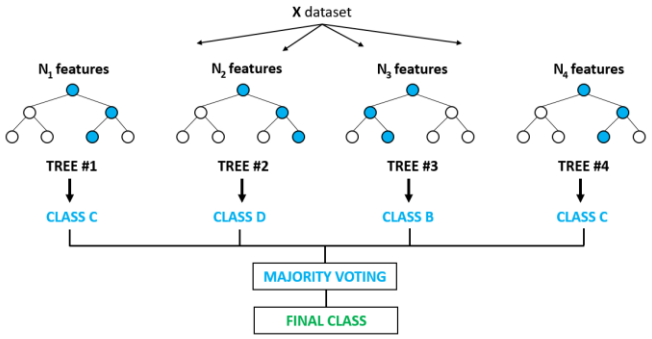


Fig 2. Random Forest Depiction

Fig. 2 Random Forest Depiction has nearly the same hyper parameters as a decision tree or a bagging classifier. Fortunately, we do not have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. As mentioned previously, with Random Forest, we can also deal with Regression tasks by using the Random Forest Regression. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node [8].

Support Vector Machine Algorithm

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which is a very useful technique for data classification. However, this learning algorithm can also be used for regression challenges. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

Support Vector Support vector regression is the natural extension of large margin kernel methods used for classification to regression analysis. It retains all the properties that characterize maximal margin algorithms of support vector machines such as duality, sparseness, kernel and convexity. It has become a powerful technique for predictive data analysis with many applications in varied areas of study like biological contexts, drug discovery, civil engineering, sunspot frequency prediction, image tracking, image compression etc., [6] [7].

The problem of regression is that of finding a function which approximates mapping from an input domain to the real numbers on the basis of a training sample. This refers to the difference between the hypothesis output and its training value as the residual of the output, an indication of the accuracy of the fit at this point. To decide how to measure the importance of this accuracy, as small residuals may be inevitable even need to avoid large ones. The loss function determines this measure. Support vector regression performs linear regression in the feature space using Ɛ-insensitive loss function

$$L_\varepsilon(y, f(\mathbf{x},\omega)) = \begin{cases} 0 & \text{if } |y - f(\mathbf{x},\omega)| \le \varepsilon \\ |y - f(\mathbf{x},\omega)| - \varepsilon & \text{otherwise} \end{cases}$$

The empirical risk is:

$$R_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(y_i, f(\mathbf{x}_i, \omega))$$

It is well known that SVM generalization performance (estimation accuracy) depends on a good setting of metaparameters parameters C, Ɛ and the kernel parameters. Selecting a particular kernel type and kernel function parameters is usually based on application-domain knowledge and also should reflect distribution of input (x) values of the training data. Parameter C determines the trade off between the model complexity and the degree to which deviations larger than Ɛ are tolerated in optimization formulation [8].

KNN (k-Nearest neighbor) Machine Learning Algorithm

The k-Nearest neighbor methodology is wide used adopted thanks to its potency [9]. The key plan of the algorithmic rule is to categorize a brand new sample within the most frequent class of its nearest neighbor within the coaching set. This is often the foremost selection formula on the category labels of its neighbors. The k-nearest neighbor classification algorithmic rule may be divided into 2 phases: coaching section and testing section. KNN is similar to kernel methods with a random and variable bandwidth. The idea is to base estimation on a x th number of observations k which are closest to the desired point.

Suppose we have a sample **{X1,X2 ……..Xn}:**

For any fixed point x 2 Rq ; we can calculate how close each observation Xi is to x using the Euclidean distance $||x||$ = (x‟x) $^{1/2}$ this distance is

$$Di = ||x\text{-}Xi|| = ((x\text{-}X)' (x\text{-} Xi))^{1/2}$$

This is just a simple calculation on the data set. The order statistics for the distances Di are

$$0 \le D_{(1)} \le D_{(2)} \le \ldots\ldots\ldots \le D_{(n)}.$$

The observations corresponding to these order statistics are the “nearest neighbors” of x: The 1st nearest neighbor is the observation closest to x; the second nearest neighbor is the observation second closest, etc. This ranks the data by how close they are to x: Imagine drawing a small ball about x and slowly initiating it. As the ball hits the …first observation Xi this is the “…first nearest neighbor” of x: As the ball further initiates and hits a second observation, this observation is the second nearest neighbor.

The observations ranked by the distances, or “nearest neighbors”, are

$$\{X(1),X(2),X(3),\ldots X(n)\}$$

The kth nearest neighbor of x is X(k). For a given k; let

$$Rx = ||X_{(k)} - x|| = D_k$$

The Euclidean distance between x and X(k): Rx is just the Kth order statistic on the distances Di . When X is multivariate the nearest neighbor ordering is not invariant to data scaling. Before applying nearest neighbor methods, is therefore essential that the elements of X be scaled so that they are similar and comparable across elements.

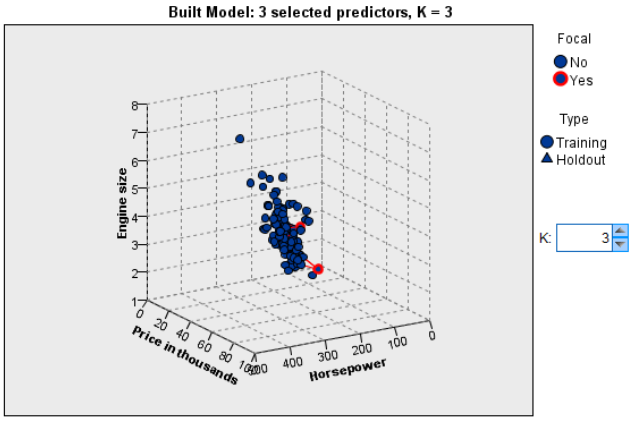


Fig 3. Projection of Feature using KNN, when K=3 for Crop Yield Dataset

The k value and KNN method shown in fig 3 is to determine the minimum points and radius value automatically. Using these methods crop data set is analyzed and determined the optimal parameters for the wheat crop production. Multiple linear regressions are used to find the significant attributes and form the equation for the yield prediction. This model is simple, does not required any sophisticated statistical tools, required data for crop growing periods, yield data for past years and provides marginally good prediction. Therefore it can be used for district, agro climatic zone and state level prediction. After analyzing the results of statistical methods and KNN found that the results of KNN are less accurate than the statistical methods, but not getting the high level accuracy these methods through. Hence, it must be further improved for more accuracy and lower errors. As the existing data mining algorithm KNN is not giving the satisfied accurate prediction results for the sugarcane crop yield. So, the designing and development of hybrid algorithm clustered KNN is done [10].

V. RESULT ANALYSIS

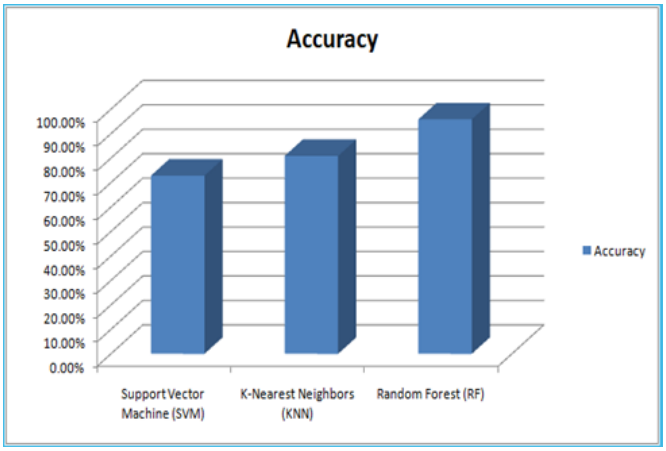
Analysis of project is done based on three parameter which are described as below.

Training Model Accuracy

Here we had train the model using the training data set it is generated from the actual data set dividing it into the 80% of training data set and 20% of these data set 80% of data set is given to the different machine learning algorithms. After completion of the trainings data set is given to the trained model in that model is is tested with test data set it will produces the accuracy of the different model which is shown below.

Table 1 : Training Model Accuracy

Model Performance	
Algorithm	Accuracy
Support Vector Machine (SVM)	72.8643%
K-Nearest Neighbors (KNN)	80.9045%
Random Forest (RF)	95.8543%



We analyze that support vector machine algorithm is produced 72% of accuracy shown in Table 1 which is least among the three algorithms. In k nearest neighbor algorithm produced 80% of accuracy which is comparatively acceptable as compared to support vector machine algorithm.

The best result is generated by random forest algorithm which has given around 96% of accuracy which is highest among these three algorithms.

Dataset Analysis

These complete dataset has 8 parameters like moisture, rainfall, average humidity, mean temperature, max temperature, min temperature, alkaline, sandy and the predicted crop yield value. It contains nearly 5000 records. Table 2 shows the description of agriculture dataset .The Table 2 shows the crop yield predicted value and the crop yield and demand predicted output.

This dataset consists of factors like temperature, rainfall, moisture, humidity, alkaline, sandy . The

datasets have been obtained from the Kaggle website and other different websites. The data set has instance or data that have taken from the past historic data. It includes 8 parameters or features like the temperature, rainfall, moisture, humidity, alkaline, sandy etc

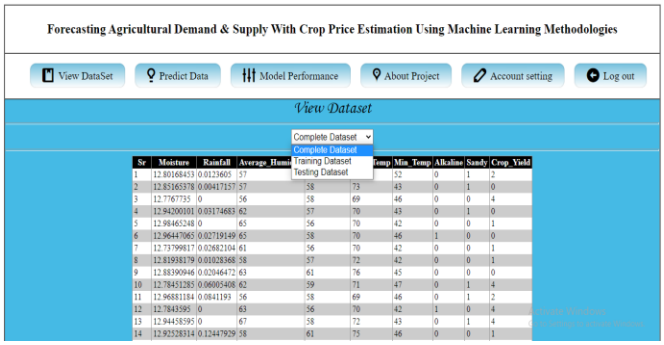


Table 2: Complete Dataset

Confusion Matrix Analysis

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

Confusion Matrix

Algorithm	Matrix
Support Vector Machine (SVM)	[[171 18 17 0 47] [0 142 0 0 0] [5 5 94 0 22] [0 0 13 0 0] [59 7 23 0 173]]
K-Nearest Neighbors (KNN)	[[187 16 5 0 45] [1 135 1 1 4] [7 4 100 0 15] [1 0 2 9 1] [26 12 10 1 213]]
Random Forest (RF)	[[230 17 3 0 3] [0 137 0 1 4] [0 0 126 0 0] [1 0 0 12 0] [0 4 0 0 258]]

Table 3: Confusion Matrix

Following Table 3 shows the confusion matrix of the different algorithm gives us the information about the accuracy for different inputs given to the model and generated output produced by the machine learning models are also confusion matrix of the random forest algorithm is generated the better result among the other two algorithms.

VI.CONCLUSION

This paper presented the various machine learning algorithms for predicting the yield of the crop on the basis of temperature, rainfall, season and area. Experiments were conducted on Indian government dataset and it has been established that Random Forest gives the highest yield prediction accuracy. By combining rainfall, temperature along with other parameters like season and humidity, land type can be made. Results reveal that Random Forest is the best classifier when all parameters are combined. These will not only help farmers in maintaining the right crop supply to grow but also in cost management also it can be helpful.

VII. FUTURE WORK

This research work can be enhancing to the high level by building a recommender system of

agriculture production and distribution for farmer. By which farmers can make their own decision like which season which crop should sow so that they can get better profit. This system works for structured dataset or database. In coming years try applying data independent system also that meanas the format may be whatever, our system should work with same accuracy.

VIII. REFERENCES

[1]. Mann, M. L., Warner, J. M., & Malik, A. S. (2019). Predicting high-magnitude, low-frequency crop losses using machine learning: an application to cereal crops in Ethiopia. Climatic Change, 154(1- 2), 211-227.

[2]. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and electronics in agriculture, 151, 61-69.

[3]. Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., ... & Reichert, G. (2015). Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agricultural and Forest Meteorology, 206, 137-150.

[4]. Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agricultural and forest meteorology, 218, 74-84.

[5]. Rao, K. R., & Josephine, B. M. (2018, October). Exploring the Impact of Optimal Clusters on Cluster Purity. In 2018 3rd International Conference on Communication and Electronics Systems (ICCES) (pp. 754-757). IEEE.

[6]. R. Furter, H. Ghorashi, A. Schleth, “The Role of Cotton Classification in the Textile Industry”, Uster Technologies AG, Tekstil teknologi, pp 100 – 105.

[7]. Y. Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri, “Python NLT K Sentiment Inspection using Naïve Bayes Classifier” International Journal of Recent T echnology and Engineering, ISSN: 2277-3878, Volume-8, Issue-2S11, Sep 2019 .

[8]. D. Srinivasa Rao, Ch. Ramesh Babu, Y. J. Nagendra Kumar, N. Rajasekhar, T . Ravi, “Medical Image Fusion Using T ransform Based Fusion T echniques”, International Journal of Recent T echnology and Engineering, Volume-8 Issue-2 ISSN: 2277-3878

[9]. Srikanth Bethu, V Sowmya, B Sankara Babu, G Charles Babu, Y. Jeevan Nagendra Kumar, “Data Science: Identifying influencers in Social Networks”, Periodicals of Engineering and Natural Sciences, ISSN 2303-4521 Vol.6, No.1, pp. 215˜228 .

[10]. Raj, J. S., & Ananthi, J. V. (2019). RECURRENT NEURAL NETWORKS AND NONLINEAR PREDICT ION IN SUPPORT VECTOR MACHINES. Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40.

[11]. Y. Jeevan Nagendra Kumar, Dr. T . V. Rajini Kanth, “GISMAP Based Spatial Analysis of Rainfall Data of Andhra Pradesh and T elangana States Using R”, International Journal of Electrical and Computer Engineering (IJECE), Vol 7, No 1, February 2017, Scopus Indexed Journal, ISSN: 2088-8708

[12]. B Sankara Babu, A Suneetha, G Charles Babu,Y.Jeevan Nagendra Kumar, G Karuna, “ Medical Disease Prediction using Grey Wolf optimization and Auto Encoder based Recurrent Neural Network”, Periodicals of

Engineering and Natural Sciences, June 2018
ISSN 2303-4521 Vol.6, No.1, pp. 229~240

Cite this article as :

Pallavi Shankarrao Mahore, Dr. Aashish A. Bardekar, "Crop Yield Prediction", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7, Issue 3, pp.561-569, May-June-2021. Available at
doi : https://doi.org/10.32628/CSEIT2173168
Journal URL : https://ijsrcseit.com/CSEIT2173168