RESEARCH ARTICLE WWW.IJCCTS.ORG

Implement A System For Crop Selection And Yield Prediction Using Random Forest Algorithm

K.P.Uvarajan[1*], K.Usha[2]

[1]Assistant Professor, Department of Electronics and Communication Engineering,
KSR College of Engineering, Tiruchengode.

[2]M.E Final year, Department of Electronics and Communication Engineering, KSR College of Engineering, Tiruchengode.

Keywords:
Climate,
crop yield,
Indian Agriculture,
Machine Learning Techniques,
Random Forest Algorithm

DOI: 10.31838/IJCCTS.12.01.02

Received: 06.01.24
Revised: 05.02.24
Accepted: 02.03.24

Abstract

Agriculture is the bone that plays important part in the frugality of India. India is an agrarian country and its frugality largely grounded upon crop product. It'll allow policy makers and growers to take effective marketing and storehouse way to prognosticate crop yields before in their crop. The design will allow growers to capture the yields of their crops before civilization in the field of husbandry and therefore help them make the necessary opinions. Preparation of such a system with a use and the machine learning algorithm an also be distributed. The results attained are granted access to the planter. and yet there are colorful styles or protocols for similar veritably data analytics in crop yield vatic nation, and were suitable to prognosticate agrarian productivity with guidance of all those algorithms. It utilizes a random forest algorithm. By probing similar problems and issues similar as rainfall, temperature, moisture, downfall, moisture, its overcome using random forest technique, acceptable results an inventions resolve the situation. In countries like India, indeed in the agrarian sector, as there are numerous types of adding profitable growth. In addition, the processing is useful of ratiocinating the product of crop yields.

How to cite this article: El-Saadawi E, Abohamama AS, Alrahmawy Mf (2024). Implement A System For Crop Selection And Yield Prediction Using Random Forest Algorithm. International Journal of communication and computer Technologies, Vol. 12, No. 1, 2024, 21-26

Introduction

For numerous pattern bracket problems, an advance number of features used don't inescapably restate into a advanced bracket delicacy. In some cases the performance of algorithms devoted to speed an prophetic delicacy of the data characterization an indeed drop. Thus, point selection an serve as preprocessing tool of enormous significance previous to working the bracket problems. the purpose of the point selection is to reduce the maximum number of inapplicable features while maintaining an respectable bracket delicacy. A good point selection system an reduce the cost of point dimension, an increase classifier effectiveness an bracket delicacy. point selection if as considerable significance in pattern bracket, data analysis, multimedia information reclamation, medical data processing, machine literacy, and data mining operations. PSO is used to apply a point selection, and SVMs with the one-versus-rest system were used as observer for the PSO fitness function for five multiclass problems taken from the literature. The results reveal that our system illustrated a better delicacy than the bracket styles they were compare to.

Data Mining

Data mining (the analysis step of the" knowledge discovery in databases" process, or KDD), a field at the crossroad of computer wisdom and statistics, is the process that attempts to discover patterns in large dada sets. It utilizes styles at the crossroad of artificial intelligence, machine literacy, statistics, an database systems the overall thing of the data mining process is to prize information from a dataset an transfigure it into an accessible structure for further use away from the raw analysis step, it involves data base and data operation aspects ,data preprocessing, an conclusion considerations, post-processing of discovered structures, visualization, an online updating. Generally, data mining(occasionally called data or knowledge discovery) is the process of assaying data from different perspectives and recapitulating it into useful information that can be used to increase profit, cut costs, or both. At a mining software one of a number of logical tools for assaying data. It allows druggies to dissect data from numerous different confines or angles, classify it, and epitomize the connections linked. Technically, data mining is the process of chancing correlations or patterns among dozens of field in large relational database.

Data mining techniques

There are several major data mining ways have been develop and used in data mining systems lately including association, bracket, clustering, ratiocination an succession patterns.

Association

Association is one of the best known data mining fashion. In association, a pattern is discovered grounded on a relationship of a particular item on other particulars in the same sale. For illustration, the association fashion is use in request hand basket analysis to identify what products that guests constantly buy together. Grounded on this data businesses can have corresponding marketing crusade to vend further products to make further profit.

Classification

Classification is a classic data mining fashion grounded on machine literacy. Principally bracket is used to classify each item in a set of data into one of predefine set of classes or groups. Classification systems makes us of fine ways similar as decision trees, direct programming, neural network an statistics. In bracket, make the software that can learn how to classify the data particulars into groups, illustration, can apply classification in operation that “given all once records of workers who left the company, prognosticate while current workers are presumably to leave in the future.” Groups that are “leave” and “stay”.

Clustering

Clustering is a data mining fashion that makes meaningful or useful luster of objects that have analogous characteristic using automatic fashion. Different from bracket, clustering fashion also define the classes and put objects in them, while in bracket objects are assigned into predefine classes. To make the conception clearer ,can take library as an illustration. In a library as an illustration. In a library, books have a wide range of motifs available. The challenge is how to keep those books in a way that compendiums can take several books in a specific content without hassle.

Literature Review

[1] Gabreiel M, Alves, Paulo E. Crunivel, in perfection husbandry an increase of data and information has been observe an new approaches to ameliorate knowledge are now needed. Thus, studies an big data are being conducted to fin innovative results as a means to dissect large data sets. In this work, we present a big data terrain for agrarian soil analysis from reckon tomography[CT] images. Our structure is planned in three layers source; big data terrain, an operations. We use Hadoop frame in the alternate sub caste to reuse CT images and bandy how the 3D reconstruction is performed. Another operation in the structure is the statistical analysis of soil analysis system to gain and understand of the problems related to agrarian lands.

[2] Tyrone T. Lin, Chung-shiao, Hsieh this paper substantially explores when the agrarian assiduity faces grain crop price oscillations and natural climate changes, it’ll take which position of price of grain crops and what probability of climate changes for developing a dynamic grain crop gyration model. In the former paper, the authors introduce the mixed strategy of game proposition to construct a 2-player game. In consideration of the pursuit of the maximization of their own interests, the decision-timber of dynamic grain crop gyration is the main focus of the former paper, and it’ll be extended to a multiple stable dynamic grain crop gyration strategy cycle, and now the authors develop a stationary Markov process as the base for a final decision. Marko chain is a system constantly use in decision-timber and is a model simple to be bandied.

[3] Niketa Gandhi, Lesia J. Armstrong, Owaizpetkar, Food product in India is largely dependent on cereal crops including rice, wheat and colorful beats. The sustainability an productivity of rice growing areas is dependent of suitable climatic condition. Variability in seasonal climate conditions can have mischievous effect, with incidents of failure reducing product. Developing better ways to prognosticate crop productivity in different climatic conditions and help planter and other stakeholders in better decision

making in terms of agronomy and crop choice. Machine literacy ways and be used to ameliorate ratiocination of crop yield under different climatic scripts. This paper presents the review on use of similar machine learning fashion for Indian rice cropping areas. This paper discusses the experimental results attained by applying SMO classifier using the WEKA tool on the dataset of 27 sections of Maharashtra state, India. The dataset considered for the rice crop yield ratiocination was source from intimately available Indian government records. The parameter considered for the study were rush, minimal temperature, average temperature, maximum temperature and reference crop evapotranspiration, area, product and yield for the Kharif season(June to November] for the times 1998 to 2002. For the present study the mean absolute error[MAE], root mean squared error[RMSE], relative absolute error[RAE] and root relative square error[RRSE] were calculate. The experimental results shoe that the performance of other ways on the same dataset was much better compared to SMO.

[4] Viya V.Polprof, S.M.PatilThe term big data,refers to vastly substantial data whose volume, variability, and haste make it veritably laborious to manage, process or anatomized. To dissect this vastly substantial kind of data Hadoop will be employed. Still, processing is veritably time-consuming. To resolve the dilemma & to diminishment replication time one result is to executing the job incompletely, where an approximate, early result becomes available to the use, completion of job. Proposed system gives a more nascent hart reduce armature that warrants data to be divided for easier & early processing. This isn't time consuming an amends system application for batch jobs as well. Propose system presents a more nascent interpretation of the Hadoop Map reduce frame that fortifies on- process aggregation, which warrants & avails druggies to get early results of a job as it's calculating. It'll estimate this fashion exerting authentic world datasets and operations and endeavor to amend the systems performance in terms of perfection and time, also the combiner will get execute later chart function & before reducer. Rather of processing complete train on-process aggregation divides the train into number of blocks which helps to gives the result in places. Dividing the train into number of datasets helps to give result to the stoner. The ideal of the proposed fashion is to amend the performance of Hadoop Map reduce for effective & easy immensely big data processing time.

Existing System

A goo point selection system and reduce the cost of point dimension, and increase classifier effectiveness and bracket delicacy, point selection is of considerable significance in pattern bracket, data analysis, multimedia information reclamation, medical data processing, machine literacy, and data mining operations. PSO is used to apply a point selection, and KNN with the one-versus-rest system were used as observers for the PSO fitness function for five multiclass problems taken from the literature. The results reveal that our system illustrated a better delicacy than the bracket styles they were compare to.

- ■ Disadvantages
- It doesn't classify the unlabeled data
- It take the further time for training

Proposed System

Random timber is a principally supervised literacy algorithm that's used for both group as well as retrogression. Random timber algorithm creates decision trees on different data samples and also prognosticate the data from each subset and also by advancing gives better the result for the system. Random forest used the bagging system to trained the data. Principally, the bagging system is a admixture of studying different models an increase the final result of the system. Foe getting high delicacy we used the random forest algorithm which gives delicacy which predicate but model and factual outgrowth of prediction in the dataset. In the arbitrary timber which beaters the decision tree from a sample of data an trees gives the ratiocination from each family and select the stylish result but voting which gives better delicacy for the model. It gives optimum results for the system.

Random forest works in two-phase first is to produce the arbitrary timber by combining N decision tree, and second is to make prognostications for each tree created in the first phase.

The working process can be explained in the below way and illustration

Step-1 select arbitrary K data point from the training set.

Step-2 figure the decisions trees associate with the named data points [subsets].

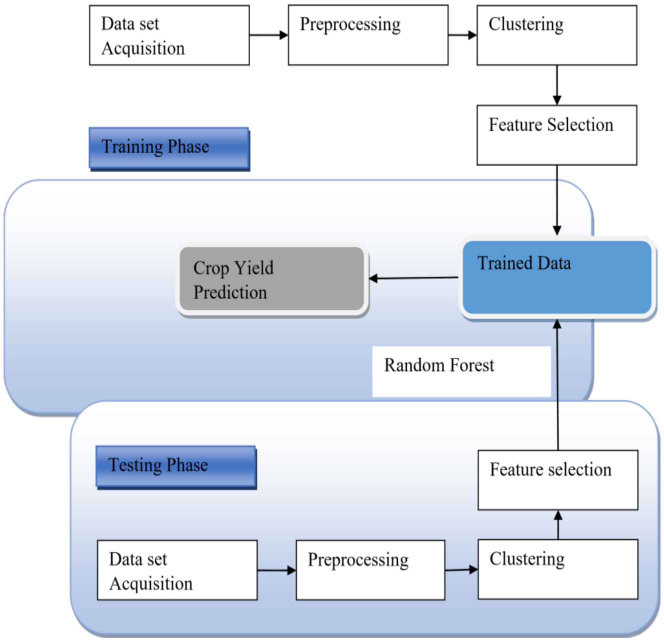
Step-3 choose the number N for decisions trees that you want to make.

Step-4 reprise step 1&2

- Advantages
- It takes lower training time as compared to other algorithm
- It predicts affair with high delicacy, indeed for the largest dataset it runs efficiently.
- It can also maintain delicacy when a large proportion of data is missing.

Design and Development Phase

The propose system is intended to help the farmers and researchers to understand the crop field and to choose a suitable crop for it. Various parameters are considered from soil to atmosphere for predicting the suitable crop. Soil parameters such as type, ph level, iron, copper, manganese, sulphur, organic carbon, potassium, phosphate, nitrogen.



Module Details

The Product will perform the following functions

- Dataset Acquisition
- Preprocessing
- Clustering
- Feature Selection
- Classification

Dataset Acquisition

- In this module is used to upload the weather details.
- It contains the 'Year', 'Rainfall', 'Area of Sowing', 'Yield', 'Fertilizers' (Nitrogen, Phosphorous and Potassium) and 'Production'.

Preprocessing

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results.

Clustering

Clustering is a technique in data mining to find interesting patterns in a given dataset .The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters information's into k groups, where k is considered as an input parameter.

It then assigns each information's to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then more computed and the process begins again. The k means algorithm is one of the simplest clustering techniques and it is commonly used in medical data and related fields. K-Means algorithm is a divisive, unordered method of defining clusters.

Feature Selection

Feature selection is the process of selecting a subset of relevant, useful features for use in building an analytical model. Feature selection helps narrow the field of data to just the most valuable inputs, reducing noise and improving training performance.

Classification

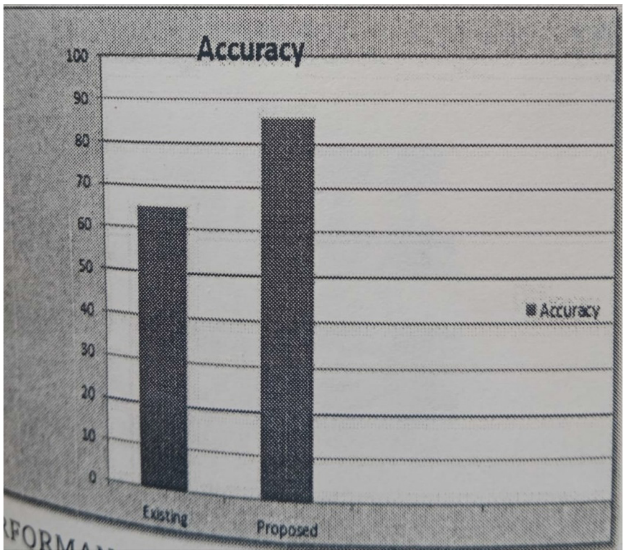
In this module, implement classification algorithm to classify the data, finally predict the yield production using Random Forest Classification. Random forests are the aggregation of tree predictors in such a way that each tree depends on the values of a random subset sampled independently and with the same distribution for all trees in the forest. Random Forest used the bagging method to trained the data which increases the accuracy of the result. For getting high accuracy we used the Random Forest algorithm which gives accuracy which predicate by model and actual outcome of predication in the dataset. The predicted accuracy of the model is analyzed 91.34%. Fig.2 shows the flowchart of random forest model for crop yield prediction.

Pseudocode of the Proposed System:

1. In this basically we first randomly selected the k to feature out of the total m feature in the model.
2. Using the best split point choose the k feature and calculate the node d.
3. So we used the split method, split the nodes into the daughter node.
4. Repeat 1 to three steps until l number of nodes has been reached
5. Build forest by repeating steps 1 to 4 for n number times to make n number of trees.

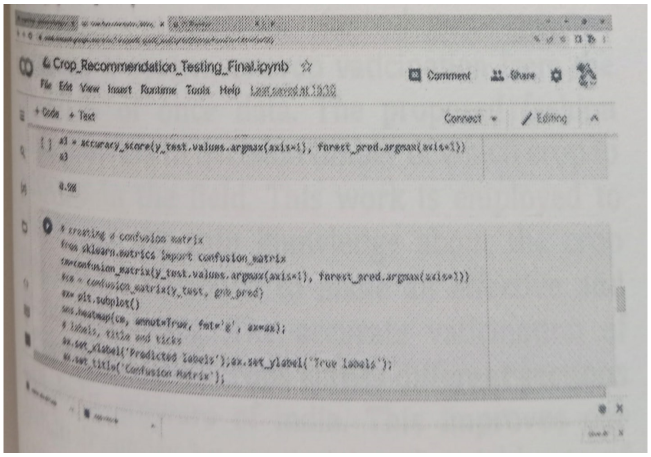
To perform prediction using the trained random forest algorithm uses the below pseudocode:

1. In this, we took the test features and every random decision tree to predicate the output and anticipated the outcome which is stored
2. And then we calculated the vote which is given by every decision tree for each predicated outcome.
3. Finally, we considered high voted predicated outcome which gives the final prediction from the random forest algorithm.



Performance Evaluation

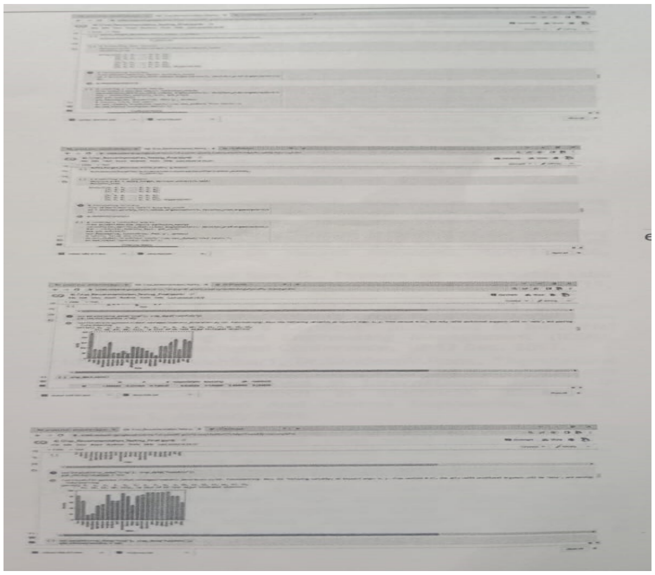
In this above graph is representing the comparison of the existing and proposed system classification algorithms. In KNN algorithm is provide the accuracy is higher than the other classification method.

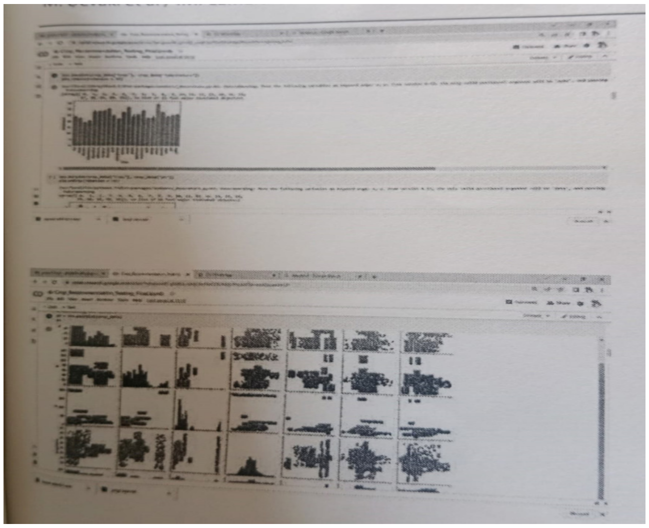

Crop_Recommendation_Testing_Final.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code
+ Text
Connect
Editing

```
a3 = accuracy_score(y_test.values.argmax(axis=1), forest_pred.argmax(axis=1))
a3
```

```
0.98
```

```
# creating a confusion matrix
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test.values.argmax(axis=1), forest_pred.argmax(axis=1))
#cm = confusion_matrix(y_test, gnb_pred)
ax= plt.subplot()
sns.heatmap(cm, annot=True, fmt='g', ax=ax);
# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
```





Conclusion

This system focuses on the prediction of crop and calculation of its yield with the help of machine learning techniques. Several machine learning methodologies used for the calculation of accuracy. Random Forest classifier was used for the crop prediction for chosen district. Implemented a system to crop prediction from the collection of past data. The proposed technique helps farmers in decision making of which crop to cultivate in the field. This work is employed to search out the gain knowledge about the crop that can be deployed to make an efficient and useful harvesting. The accurate prediction of different specified crops across different districts will help farmers of India. This improves our Indian economy by maximizing the yield rate of crop production.

References

1. Gabriel M. Alves, Paulo E. Cruvinel, "Big Data environment for agricultural soil analysis from CT digital images", published in Semantic Computing (ICSC), IEEE Tenth International Conference, Feb 2016.
2. Pallavi V. Jirapure, Prof. Prarthana A. Deshkar "Qualitative data analysis using Regression method for Agricultural data", published in Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference, 29 Feb/March 2016.
3. Mohammed Zakariah, "Classification of large datasets using Random Forest Algorithm in various applications: Survey" published in International Journal of Engineering and Innovative Technology (IJEIT), Volume 4, Issue 3, September 2014.
4. Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., and Jain, A. K., "Dimensionality Reduction Using Genetic Algorithms," IEEE Trans. Evolutionary Computation, vol. 4, no. 2, pp. 164-171, July 2000.
5. Narendra, P.M. and Fukunage, K., "A Branch and Bound Algorithm fo Feature Subset Selection," IEEE Trans. Computers, vol.6, no. 9, pp. 917-922, Sept. 1977.
6. Pudil, P., Novovicova, J., and Kittler, J., "Floating Search Methods in Feature Selection," Pattern Recognition Letters, vol.15, pp. 1119-1125, 1994.
7. Roberto B., "Using mutual information for selecting features in supervised neural net learning," IEEE Transactions on Neural Networks, 5(4):537-550, 1994.
8. Zhang, H. and Sun, G.. Feature selection using tabu search method. Pattern Recognition, 35: 701-711, 2002.
9. Tyrone T. Lin, Chung-Shiao. Hsieh, "A Decision Analysis for the Dynamic Crop Rotation Model with Markov Process's Concept" published in Industrial Engineering and Engineering Management (IEEM), IEEE International Conference, 10-13 Dec. 2013.
10. Snehal S.Dahikar1, Dr.Sandeep V.Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach" published in International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering, Vol. 2, Issue 1, January 2014.
11. Wu Fan, Chen Chong, Guo Xiaoling, Yu Hua, Wang Juyun, "Predictionof crop yield using big data", Published in: Computational Intelligence and Design (ISCID), 2015 8th International Symposium, 12-13 Dec. 2015.
12. Snehal S. Dahikar, Sandeep V. Rode, Pramod Deshmukh, "An Artificial Neural Network Approach for Agricultural Crop Yield Prediction Based on Various Parameters", published in International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 4, Issue 1, January 2015.
13. Uvarajan, K.P. and Gowri Shankar, C., 2020. An integrated trust assisted energy efficient greedy data aggregation for wireless sensor networks. Wireless Personal Communications, 114, pp.813-833