www.ijiarec.com

ISSN:2348-2079

Volume-9 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

An efficient crop yield prediction using random forest algorithm

S.Ajithkumar[1], Dr.M.Somu[2], Dr.V.Sharmila[3], Dr.A.Rajivkannan[4]

[1]PG Student, Dept. of C.S.E, K. S. R. College of Engineering (Autonomous), Tiruchengode, Tamilnadu, India

[2,3,4]Professors, Dept. of C.S.E, K. S. R. College of Engineering (Autonomous), Tiruchengode, Tamilnadu, India

ABSTRACT

India is an agricultural country and its economy is largely based upon crop productivity and rainfall. For analyzing the crop productivity, rainfall prediction is require and necessary to all farmers. Rainfall Prediction is the application of science and technology to predict the state of the atmosphere. It is important to exactly determine the rainfall for effective use of water resources, crop productivity and pre planning of water structures. Using different data mining techniques it can predict rainfall. Data mining techniques are used to estimate the rainfall numerically.Agriculture has the largest contribution in the GDP of our country. But still the farmer's don't get worth price of the crops. It is mostly happens due to improper irrigation or inappropriate crops selection or also sometimes the crop yield is less than that of expected. By analyzing the soil and atmosphere at particular region best crop in order to have more crop yield and the net crop yield can be predict. One suitable explanation behind this is the deficiency of adequate decision making by farmers on yield prediction. There isn't any framework in location to suggest farmer what plants to grow. The proposed machine learning approach aims at predicting the best yielded crop for a particular region by analyzing various atmospheric factors like rainfall, temperature, humidity etc., and land factors like soil pH, soil type including past records of crops grown. Finally our system is expected to predict the best yield based on dataset we have collected.This prediction will help the farmers to choose appropriate crops for their farm according to the soil type, temperature, humidity, water level, spacing depth, soil PH, season, fertilizer and months. This prediction can be carried out using Random Forest classification machine learning algorithm.

Keywords: Crop Productivity,Machine Learning Approach, Crop Selection, atmospheric factors.

INTRODUCTION

Analysis of time series data is one of the important aspects of modern research in the domain of knowledge discovery. Time series data is collected over a specific period of time such as hourly, daily, weekly, monthly, quarterly or yearly. Data mining techniques can use this data to predict upcoming situations in various domains such as climate change, education, and finance etc. These techniques can be used to extract hidden knowledge from time series data for future use. Weather forecasting is very beneficial but challenging task. Weather data consists of various atmospheric features such as wind speed, humidity, pressure and temperature etc.

Data mining techniques have the capacity to extract the hidden patterns among available features of past weather data and then these techniques can predict future weather conditions by using extracted patterns. Rainfall is a complex atmospheric process, which depends upon many weather related features. Accurate and timely rainfall prediction can be helpful in many ways such as planning the water resources management, issuance of early flood warnings, managing the flight operations and limiting the transport & construction activities. Accurate rainfall prediction is more complex today due to climate variations. Re searchers consistently have been working to predict rainfall with maximum accuracy by optimizing and integrating data mining techniques. Data mining algorithms are classified

Author for correspondence:
PG Student, Dept. of C.S.E, K. S. R. College of Engineering (Autonomous), Tiruchengode, Tamilnadu, India

assupervised and un-supervised. Supervised methods get trained first with pre-classified data (training data) and then classify the input data (test data). Un-supervised methods on the other hand do not require any training; instead of pre-classified data these techniques use algorithms to extract hidden structure form un-labeled data. It has been observed from latest research that for high accuracy, researchers prefer the integrated techniques for the rainfall prediction. To reflect the latest research, this study provides a systematic literature review by focusing on latest papers, which are published in last five years (2013-2017).

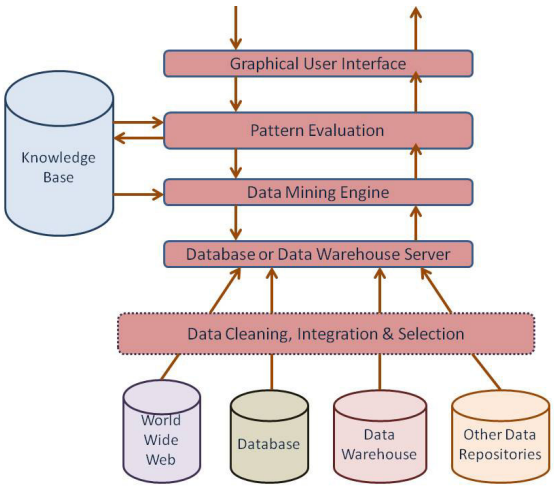
Three renowned online search libraries are selected for literature extraction: Elsevier, IEEE and Springer. Initially 4844 papers are extracted and then through a systematic research process 8 most relevant research articles are selected for critical review. Further organization of this paper is as follows. Section II elaborates the related work. Section III presents the research protocol, which is followed in this research. Section IV presents the review of shortlisted articles.

Section V discusses the review findings. Section VI finally concludes this study.Crop production is a complexphenomenon that is influenced by agro-climatic input parameters. Agriculture input parameters varies from field to field and farmer to farmer. Collecting such information on a larger area is a daunting task. However, the climatic information collected in India at every 1sq.m area in different parts of the district are tabulated by Indian Meteorological Department.

The huge such data sets can be used for predicting their influence on major crops of that particular district or place. There are different forecasting methodologies developed and evaluated by the researchers all over the world in the field of agriculture or associated sciences. Some of such studies are Agricultural researchers in Pakistan haveshown that attempts of crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide usage.

These studies have reported negative correlation between pesticide usage and crop yield. In their study they have shown that how data mining integrated agricultural data including pest scouting, pesticide usage and meteorological data are useful for optimization of pesticide usage. Thematic information related to agriculturewhich has spatial attributes was reported in one of the study their study aimed at discerning trends in agriculture production with references to the availability of inputs.

Data Mining Architecture



RELATED WORK

The graphs used in Galileo share some common features with k-d trees, but do not employ binary splitting and allow much greater fan-out as a result. Similar to Tries, identical attributes in a record can be expressed as single vertices, which simplify traversals and can reduce memory consumption. However, Galileo graphs support multiple concurrent data types, maintain an explicit feature hierarchy (that can also be reoriented at runtime), and employ dynamic quantization through configurable tick marks.

Mongo DB shares several design goals with Galileo, but is a document-centric storage platform that does not support analytics directly. However, Mongo DB has rich geospatial indexing capabilities and supports dynamic schemas through its JSON-inspired binary storage format, BSON. Mongo DB can use the Geohash algorithm for its spatial indexing functionality, and is backed by a B-tree data structure for fast lookup operations. For load balancing and scalability, the system supports sharding ranges of data across available computing and storage resources, but imposes some limitations on the breadth of analysis that can be performed on extremely large datasets in a clustered setting.

Facebook's Cassandra project is a distributed hash table that supports column-based, multidimensional storage in a tabular format. Like Galileo, Cassandra allows user-defined partitioning schemes, but they directly affect lookup operations as well; for instance, using the random data partitioner backed by a simple hash algorithm does not allow for range queries or adaptive changes to the partitioning algorithm at runtime. This ensures that retrieval operations are efficient, but also limits the flexibility of partitioning schemes. Cassandra scales out linearly as more hardware is added, and supports distributed computation through the Hadoop runtime. Predictive and approximate data structures are not maintained by the system itself, but could be provided through additional preprocessing as new data points are added to the system.

LITERATURE SURVEY

Indian Summer Monsoon Rainfall (ISMR) Forecasting using Time Series Data: a Fuzzy-Entropy-Neuro basedExpert System

Statistical analysis reported the dynamic nature of rainfall in monsoon, which could not be predicted effectively with mathematical and statistical models. So, the authors in this research recommended to use three techniques for this type of prediction: Fuzzy Set, Entropy and Artificial Neural Network. By using these three techniques, a forecasting model is developed to deal with the dynamic nature of the ISMR. In proposed model, fuzzy set theory is used to handle uncertainties which are inherited in dataset. The entropy computational concept was modified in this model and used to provide the input as a degree of membership in the entropy function. That entropy function was referred as Fuzzy Information-Gain (FIG). Then,each fuzzified rule was defuzzified using the ANN. The value of FIG of each fuzzy- set was then used as input into ANN. The proposed model wasnamed as "Fuzzy-Entropy-Neuro Based Expert System for ISMR Forecasting" because it is the integration of fuzzy set, entropy and ANN. To evaluate the performance of proposed model following accuracy measures were used: Standard Deviations (SDs), Correlation Coefficient (CC), Root Mean Square Error (RMSE) and Performance Parameter (PP). According to results the proposed model is effective and efficient in comparison with other existing models.

An Extensive Evaluation of Seven Machine Learning Methods for Rainfall Prediction in Weather Derivatives

The ultimate goal was to not bias the experiment to particular climate type or for particular geographic location. According to results the accumulating rainfall amounts can bring good results as compared to prediction using daily rainy data. While using the accumulated data, Support Vector Regression, Radial Basis Functions, and Genetic Programming overall performed well however Radial Basis Functions performed better then modern technique of "Markov chain". For all selected datasets, each technique used the same parameters so it was not guaranteed that the best possible set of parameters was used for all the techniques.

During the experiment, the researchers have noted a relationship between predictive accuracy and climatic attributes such as: volatile nature of rainfall, amount of maximum rainfall and the inter quartile range of rainfall. Moreover no significant difference was noted in algorithms' prediction error among the cities of both the continents (USA and Europe). Issue regarding the discontinuity in rainfall data was solved with the help ofaccumulated rainfall amounts.

A Novel Approach For Optimizing Climate Features And Network Parameters In Rainfall Forecasting

In authors presented a Deep Learning based architecture to predict the daily accumulated rainfall for next day. Proposed architecture consists of two techniques: Auto encoder Network and the Multilayer PerceptronNetwork. Auto encoder is an unsupervised network which performed the feature selection activity and the Multilayer Perceptron Network was assigned the classification and prediction tasks. Dataset for prediction was obtained from Institute de Studios

Ambient ales (IDEA) of Universal National de Colombia which is located inManizales Colombia. Dataset spanned from 2002 to 2013 and consisted of 47 weather attributes.

IDEA extracted the data from a meteorological station located in the central area of same city and stored in an environmental DWH. As ETL steps were performed on data so pre processing was not needed. Obtained 2952 data samples were classified into subsets for the purpose of training, validation and testing, with 70%, 15% and 15%, respectively. Normalization process was then performed to keep the values of data in to the range of 0 to 1 for better working.

Results of the experiment were compared with other methods such as: naive approach which predicts the accumulated rainfall of t − 1 for t, MLP with optimized parameters for training & validation set and with some other published techniques. Performance was evaluated in terms of measurement errors: Mean Square Error and Root Mean Square Error.

EXISTING SYSTEM

The advancement in information storage is providing vast amounts of data. A huge data set of crop database is extracted.The database contains measurements of soil data from various locations. In addition to the research establishes whether Soils are Classified Using various data mining techniques. Comparison was made between Naive bayes classification and analyse the most effective technique.

DISADVANTAGES

- Must have Knowledge on Bayesian probability or Bayesian methods.
- Time taken for the process is larger Based on the assumption that features have same statistical relevance.

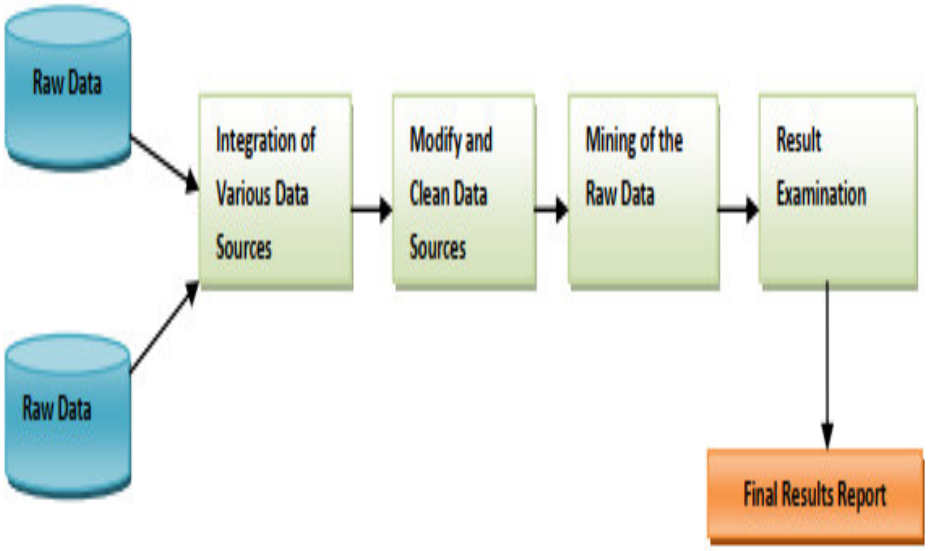
PROPOSED SYSTEM

This system proposes an approach to analyse large data set. This proposal gives a introduction to application ofanalysisin themassive data analysis in the field of agronomy. Data about weather, irrigation, and yield from several other sources (e.g. Meteorological station and irrigation-plan records).For past few decades are collected and analyzed to produce an output. Which has the highest productivity of each grain in their respective geographical conditions? Simultaneously, the data about weather, soil condition, moisture content, due factor etc. Are recorded. From these records the random forest model are trained to evaluate the perfect crop for the current geographical conditions.

Advantages

- Powerful and accurate,
- Good performance on prediction.
- User able to know the predicted crops values so, they can by plant more effectively. Can able to deploy different types of crops by selecting them in the same window.

System Model



METHODOLOGY

DECISION TREE GENERATION

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). In this post we'll learn how the random forest algorithm works, how it differs from other algorithms and how to use it.Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Algorithm Generate-Tree(TS; attribs, default)

Input: Ts, the set of training data sets attribs, set of attributes
default, default value for the goal predicate
Output: tree, a decision tree

1. if Ts is empty then return default
2. default -Majority -Value Ts
3. if Hai (Ts)= 0 then return default
4. else if attribs is empty then return default
5. else
6. best -Choose-Attribute (attribs,Ts)
7. tree -a new decision tree with root attribute best
8. for each value vi of best do
9. Tsi -{datasets in Ts as best = ki}
10. subtree- Generate-Tree (Ts attribs-best, default)
11. connect tree and subtree with a branch labelled ki
12. return tree

In Section 2, we discussed an algorithm that generates an unrealized training set, T’, and a perturbing set, Tp, from the samples in Ts. In this section, we use data tables T’ and Tp as a means to calculate the information content and information gain of Ts, such that a decision tree of the original data sets can be generated based on T’ and Tp.

MODULES

DATASET

In this research, two types of datasets are used: Weather data and Yield data. The Weather dataset used in this research has been collected from Department of Metrological Centre, Anand Agriculture University(AAU), Anand and Yield data are taken from 4 different Yearbooks of Directorate of Agriculture Gujarat State, Gandhinagar for four different district of Gujarat which are Jamnagar, Junagath, Rajkot, Amreli. A lot of pre-processing was required to handle missing values, noise and outliers. We considered

different 20 attributes for this research: rainfall, maximum and minimum temperature, Vapour Pressure, Relative humidity, Basic Sunshine, Evaporation Pressure, Soil Temperature at different depth, Wind Speed, irrigated area for all districts; and cultivated area for crop yield considered according to the districts. After the necessary formatting and pre-processing of the datasets, the finalized version of our data contains Total 4 districts for the time periods of 2006 to 2013.

INPUT VARIABLES AND MEASURES

From the vast initial dataset, we selected a limited number of important input variables that is 20, which have the highest contribution to agricultural production. All the inputs were considered for the eight-year periods of 2006 to 2013. There are various measures to different prediction models:

1. R Square [Coefficient of determination] is simply the square of the sample correlation coefficient (i.e., r) between the outcomes and their predicted values. The coefficient of determination ranges from 0 to 1.
2. RMSE [Root Mean Square Error] is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modeled.

Individual Effect of Attributes on Yield

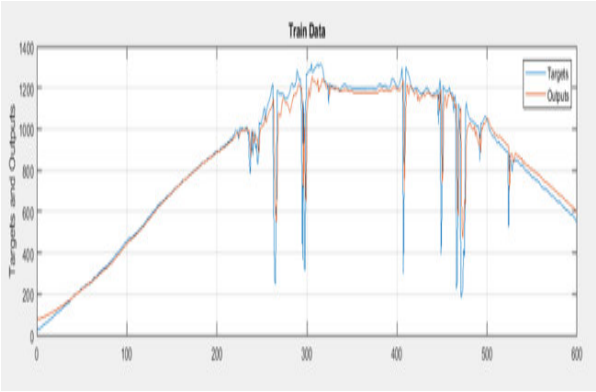
To find the individual attribute effect on yield correlation is perform on the data. In addition, values of "r" and "R square" are generated. From that, we can see that which attribute affect more on yield. From the values of R Square, it clearly sees that above values of 0.5 depends more on yield and are important attribute to grow of plant so we have considered it.

Combined Effect of Attributes on Yield

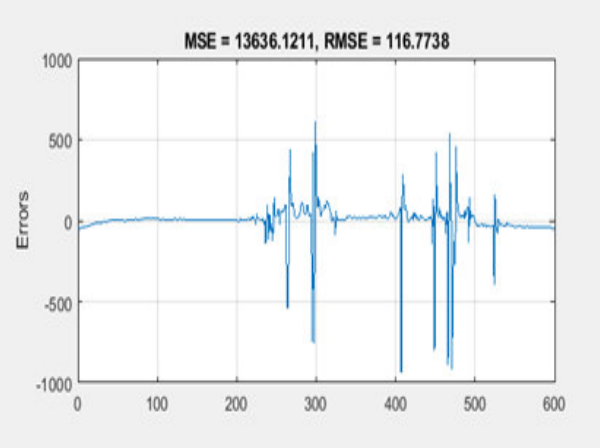
We have taken 4 different alternatives from which the different models accuracy measures can be found.

A. Environmental Attributes: Rainfall, Maximum Temperature, Minimum Temperature, Basic Sunshine, Relative humidity (2 times), Vapor Pressure(2 times), Evaporation Pressure.
B. Soil Attributes: Temperature from the different depth of Soil (6).
C. Abiotic Attributes: Water Content (2), Density (2), Wind Speed.
D. Area Central Attributes: Total area from which the yield is produced and total production.

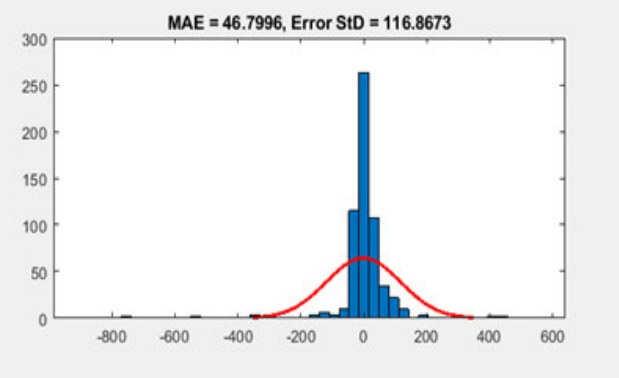
RESULTS



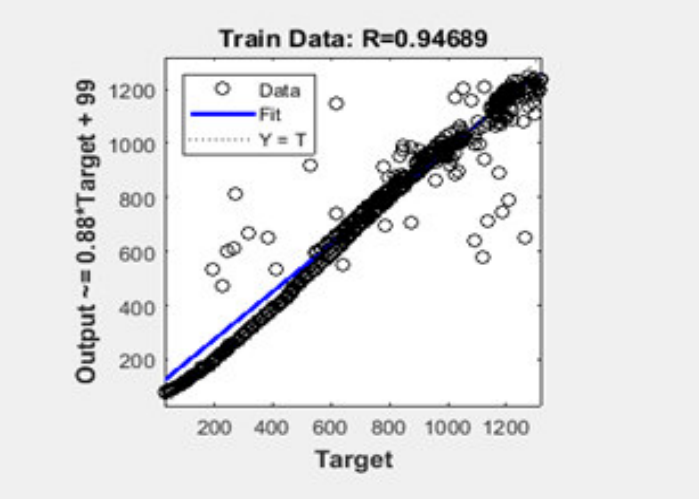
ALL DATASET



MSE AND RMSE CALCULATION

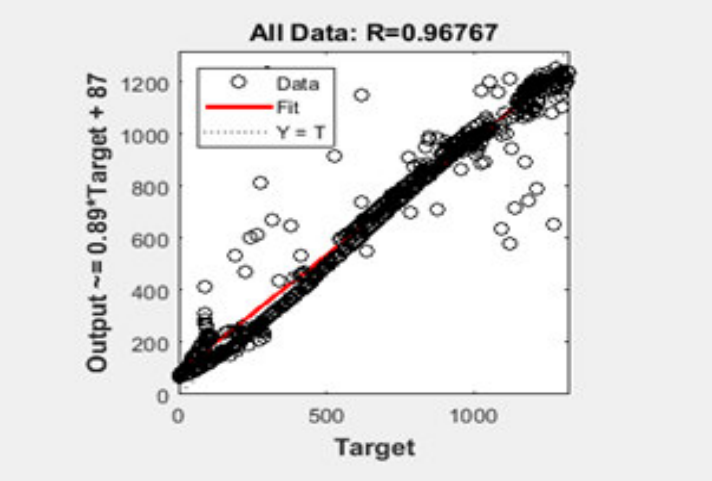


MAE CALCULATION



TRAINIG DATASET

TESTING DATASET



PERFORMANCE EVALUATION

CONCLUSION

As per the proposed system it presents a deep learning framework for crop yield prediction using remote sensing data. It allows for real-time forecasting throughout the year and is applicable worldwide, especially for developing countries where field surveys are hard to conduct. Was the firsttouse modern representation learning ideas for crop yield prediction, and successfully learn much more effective features from raw data than the hand-crafted features that are typically used. It is been proposed as random forest algorithm based on histogramsand present a Deep Gaussian Process framework that successfully removes spatially correlated errors, whichmight inspire other applications in remote sensingand computational sustainability.

REFERENCES

[1]. Ponce-Guevara, K. L., Palacios- Echeverria, J. A.,Maya Olalla, E.,Dominguez Limaico, H. M., Suarez- Zambrano, L. E., Rosero Montalvo,P.D., Alvarado-Perez, J. C. (2017). GreenFarm- DM: A tool for analyzing vegetable crops data from a greenhouse using data mining techniques (First trial). 2017 IEEE Second Ecuador Technical Chapters Meeting(ETCM).

[2]. Jheng, T.-Z., Li, T.-H., Lee, C.-P. (2018). Using hybrid support vector regression to predict gricultural output. 2018 27th Wireless and OpticalCommunication Conference (WOCC).
[3]. Manjunatha, M., Parkavi, A. (2018). Estimation of Arecanut Yield in VariousClimatic Zones of Karnataka using Data Mining Technique: A Survey. 2018 International Conference on Current Trends To wards Con- verging Technologies (ICCTCT).
[4]. Shakoor, M. T., Rahman, K., Rayta, S.N., Chakrabarty, A. (2017). Agriculturalproduction output prediction using Supervised Machine Learningtechniques.2017 1st International Conference on Next Generation Computing Applications (NextComp).
[5]. Grajales, D. F. P., Mejia, F., Mosquera, G. J. A., Piedrahita, L. C.,Basurto,C. (2015).Crop-planning, making smarter agriculture with climate data. 2015 Fourth InternationalConference on Agro-Geoinformatics (Agro-Geoinformatics).
[6]. [6] Shah, P., Hiremath, D., Chaudhary, S. (2017). Towards development of spark based agricultural information system including geo-spatial data. 2017 IEEE International Conference on Big Data (BigData).
[7]. Afrin, S., Khan, A. T., Mahia, M., Ahsan, R., Mishal, M. R., Ahmed, W.,Rahman, R. M. (2018). Analysis of Soil Properties and Climatic Data to Predict Crop Yields and Cluster Dierent Agricultural Regions of Bangladesh. 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS).
[8]. Sekhar, C. C., Sekhar, C. (2017). Productivity improvement in agriculture sector using big data tools. 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC).
[9]. Sahu, S., Chawla, M., Khare, N. (2017). An ecient analysis of crop yield prediction using Hadoop framework based on random forest approach. 2017 International Conference on Computing, Communication and Automation (ICCCA).