1. ANALYSIS OF CROP YIELD PREDICTION USING DATA ANALYTICS TECHNIQUES

ABSTRACT:

Agrarian sector in India is facing rigorous problem to maximize the crop productivity. More than 60 percent of the crop still depends on monsoon rainfall. Recent developments in Information Technology for agriculture field has become an interesting research area to predict the crop yield. The problem of yield prediction is a major problem that remains to be solved based on available data. Data Mining techniques are the better choices for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. This paper presents a brief analysis of crop yield prediction using Multiple Linear Regression (MLR) technique and Density based clustering technique for the selected region i.e. East Godavari district of Andhra Pradesh in India.

METHODOLOGY:

- In this paper the statistical method namely Multiple Linear Regression technique and Data Mining method namely Density-based clustering technique were take up for the estimation of crop yield analysis.

Multiple Linear Regression:

- A regression model that involves more than one predictor variable is called Multiple Regression Model. Multiple Linear Regression (MLR) is the method, used to model the linear relationship between a dependent variable and one or more independent variables.
- Multiple Linear Regression (MLR) technique is based on least squares and probably the most widely used method in climatology for developing models to reconstruct climate variables from tree ring services.

Density-based Clustering Technique:

- The primary idea of Density-based clustering techniques is that, for each point of a cluster, the neighborhood of a given unit distance contains at least a minimum number of points.

- In these approaches, a given cluster continues to grow as long as the number of objects in the neighborhood which exceeds some parameter.However, this idea is based on the assumption that the clusters are in the spherical or regular shapes.

Advantages:

i. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value.

ii. The second advantage is the ability to identify outliers, or anomalies. For example, while reviewing the data related to management salaries, the human resources manager could find that the number of hours worked, the department size and its budget all had a strong correlation to salaries, while seniority did not.

Dis-Advantages:

i. Two examples of this are using incomplete data and falsely concluding that a correlation is a causation.

ii. When reviewing the price of homes, for example, suppose the real estate agent looked at only 10 homes, seven of which were purchased by young parents. In this case, the relationship between the proximity of schools may lead her to believe that this had an effect on the sale price for all homes being sold in the community.

2. A Novel Approach using Big Data Analytics to Improve the Crop Yield in Precision Agriculture

ABSTRACT:

Agriculture is the main work field in India. Farming industry adopts less innovative technology compared to other industries. Information and Communication Technologies provides simple and cost effective techniques for farmers to enable precision agriculture. The work propose a state of the art model in agriculture field which will guide the rural farmers to use Information and Communication technologies (ICT) in agriculture fields. Big data analytics is used to improve the crop yield. It can be customized for precision agriculture to improve the quality of crops which improves the overall production rate.

METHODOLOGY:

The process of using technology in farming requires deep knowledge of agricultural practices, biology, and chemistry. Many parameters has to be taken into consideration and investigated in depth when designing a system that should improve cultivation procedures by making the whole process more effective . IoT that can be used for precision agriculture in real time. This architecture is divided into two modules: Data Collection Module and Data Processing Module.

Data Collection Module:

Internet of Things is useful in managing the environment from remote location. The sensor nodes used in the sensor networks can sense field parameters like moisture level in the soil, temperature, and pH level. Different types of sensors like temperature sensor, Humidity sensor and soil moisture sensors are used to collect real time environmental data. Historical data about temperature and rainfall statistics are collected from standard data sets. Data collected from the farmers are integrated with this data.

Data Processing Module:

Since sensor network data is in unstructured data format Hadoop is a suitable platform to process unstructured data. Agriculture data contains large amount of historical data which has to be combined with sensor network data so it is possible to expect large volume of data for which Hadoop provides high scalability. Hadoop network has two major parts Hadoop Distributed File System and MapReduce programming paradigm.

Advantage and Dis-Advantage:

No methodology can be considered better than the others. Every company has to carefully evaluate its goals. Then, after a careful analysis, you will be able to pick and use the best method to reach your goal.

3. Crop Yield Prediction Using Random Forest Algorithm

ABSTRACT:

Most agricultural crops have been badly affected by the effect of global climate change in India. In terms of their output over the past 20 years. It will allow policy makers and farmers to take effective marketing and storage steps to predict crop yields earlier in their harvest. This project will allow farmers to capture the yield of their crops before cultivation in the field of agriculture and thus help them make the necessary decisions. Implementation of such a method with a web-based graphic software that is simple to use and the machine learning algorithm can then be distributed. The results obtained are granted access to the farmer. And yet there are various methods or protocols for such very data analytics in crop yield prediction, and we are able to predict agricultural productivity with guidance of all those algorithms. It utilizes a Random Forest Algorithm. By researching such problems and issues such as weather, temperature, humidity, rainfall, humidity, there are no adequate solutions and inventions to resolve the situation we face. In countries like India, even in the agricultural sector, as there are many types of increasing economic growth. In addition, the processing is useful for forecasting the production of crop yields.

METHODOLOGY:

A previously developed country relies on agriculture for its economic development. As the country's population increases, reliance on agriculture often increases and the country's subsequent economic process is affected. In this case, the rate of crop yields plays a major role in the country's economic development. There is therefore a need to raise crop yield rates. To solve this problem, some biological approaches (e.g. crop seed quality, crop hybridization, strong pesticides) and a few chemical approaches (e.g. fertilizer, urea, potash use) are used. A crop sequencing technique is needed in addition to those approaches to increase the web yield rate of the crop over the season. One of the current systems we have defined is the Crop Selection Method (CSM) for the seasonal realization of a net crop yield rate. We have taken CSM's example to show how it allows farmers to produce more yield.

(a) Seasonal crops. During a season, Crops can be planted throughout the season.

(b) Week through crops :during the year, crop are also cultivated. Oh, vegetables, a paddy, a tour, for starters.

ADVANTAGE:

Crop Selection is Proper Crop Selection is a Factor in Successful Crop

Farming. It is a requisite that must be undertaken before actually starting a farming venture. Even without a predetermined location and site of a farm, the crop to be grown can be decided based mainly on its marketability and profitability.

4. Machine learning based Pedantic Analysis of Predictive Algorithms in Crop Yield Management

ABSTRACT:

Predictive analytics is a statistical technique used to forecast and investigate the development from past chronological data or to extract the information from data. With the help of rising technologies like predictive analytics in data mining, machine learning combining with Internet of Things [IoT], the major challenges in crop yield can be solved and pave way to earn profit. Machine learning means the process of making the system to learn from the previous experiences that help in prediction. In this paper, an conjectural evaluation on diverse prediction algorithms like support vector machines (SVM), recurrent neural networks (RNN), K nearest neighbour regression (KNN-R), Naïve Bayes, BayesNet, support vector regression (SVR) etc., is done and its performance are described on the basis of error rates and accuracy level in crop yield. BayesNet shows the higher accuracy of about 97.53% and RNN has less percentage error rates that dominate other algorithms in harvest prediction.

METHODOLOGY:

Accurate Prediction using Long Short Term Memory & RNN:

There are two modules projected [17]. One is to predict the reap of paddy and other is to foresee the stipulate of rice. Two algorithms are used to perform the prediction from machine learning. 1. RNN 2. LSTM. The RNN [16] is explained as a directed graph and establishes a correlation between each node in a chronological manner. In RNN, dynamic temporal behavior has been recognized with the help of various inputs that are independent of each other. This feature helps RNN to produce best prediction results. In RNN, each node sends the message to the next node as an output feedback. LSTM [17][31] is another module which predicts the rice demand. It has the capability to learn the long term dependency between the elements. Approximation algorithms like Genetic algorithms, Iterated local search can also produce best results for resource optimization and planning issues.

Autoregressive Integrated Moving Average (ARIMA) model in crop yield :

Machine learning algorithms like Autoregressive Integrated Moving Average (ARIMA) model and KNN are used along with IoT and Image processing techniques [3]. By usage of this technique, the improvement in crop productivity and reduced the usage of chemical fertilizers has been shown. The input values like N, P, K, pH and temperature of the soil from these farm areas are unruffled. ARIMA uses the past values to predict the future time series value Y. It is a time series model which uses three parameters called p, d, and q parameters where p represents the differencing degree of integrated component, d is the parameter which represents the order of moving average in d times and q represents the number of slacks used in the model. Thus, KNN [11] [22] classifier classifies the inputted forecasted values and produces the output values against the nutrients provided.

Automatic Phenology Based Algorithm for Rice Detection:

Automatic rule based method is anticipated [5] [25] by collecting the sentinel 2 data based on the three factors to perceive the rice crops from other crops. The factors are Near Infrared reflectance during the fostering time, Red band reflectance during the reap time and Normalized Difference Vegetation Index (NDVI) Values. The major challenges faced while using SVM are: 1. Number of classes, 2. Identify the precincts of the classes, and 3. Select the textural and polar metric features. In this exploration, ML are time consuming and expensive, whereas rule based reduces the necessity of ground data that covers the temporal shift during cultivation, detects and classify the rice crops efficiently has been concluded.

5.Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis:

Abstract:

Soil is an important parameter affecting crop yield prediction. Analysis of soil nutrients can aid farmers and soil analysts to get higher yield of the crops by making prior arrangements. In this paper, various machine learning techniques have been implemented in order to predict Mustard Crop yield in advance from soil analysis. Data for the experimental set-up has been collected from Department of Agriculture Department, Talab Tillo, Jammu; comprising soil samples of different districts of Jammu region for Mustard crop. For the current study, five supervised machine learning techniques namely K-Nearest Neighbor (KNN), Naïve Bayes, Multinomial Logistic Regression, Artificial Neural Network (ANN) and Random Forest have been applied on the collected data. To assess the performance of each technique under study; five parameters namely accuracy, recall, precision, specificity and f-score have been evaluated. Experimentation has been carried out to make known the most accurate technique for mustard crop yield prediction. From experimental results, it has been predicted that KNN and ANN (among the undertaken ML techniques for the study) found to be most accurate techniques for mustard crop yield prediction.

MATERIALS AND METHODOLOGY :

The main objective of the current research work is to predict mustard crop yield from soil analysis using machine learning techniques. To attain the objective of the current research, experiments have been carried out on Matlab platform. After gaining insight of problem domain, discussion with farmers and soil chemists and reviewing literature; research problem has been framed out. For current research problem, real data has been collected from Soil Testing Lab, Directorate of Agriculture Department, Talab Tillo, Jammu. Data has been collected from mustard growing areas of various districts of Jammu Region under Model Village Programme 2019-20. This dataset consists of 5000 instances with 11 input parameters representing soil nutrient status of Jammu region and one output attribute (i.e. Class Label). The parameters of the dataset are Ph (ph value of soil), EC (electrical conductivity), OC (organic carbon), N (nitrogen), P (phosphorus), K (potassium), S (sulphur), Cu (copper), Fe (iron), Zn (zinc) and Mn (manganese) representing soil nutrients. The output attribute represents three classes for mustard crop yield namely low, medium and high. Out of total 5000 instances collected, 3666 falls in Low class, 958 falls in class Medium and 376 falls in class High. The first 15 instances of the dataset are presented in table I.

A. K-Nearest :

Neighbor Implementation KNN is used for both classification and regression problems. It is one of the simplest classification algorithms. It works by determination of the parameter k which is number of nearest neighbors. When there is new data point to classify, then its k_nearest neighbors is find out from the training data by calculating the distance between the input variable and the all the data points in the dataset. This distance is calculated using various measures such as Euclidean distance, Minkowski distance, Mahalanobis distance. The larger is k; the better is classification (Harrison 2018, Brownlee 2016). For the experimental study, KNN has been implemented with 10-fold cross validation and optimum value of k is found to be 25.

B. *Naïve Bayes Classifier Implementation* :

A Naive Bayes classifier is one of the classifiers in a family of simple probabilistic classification techniques in machine learning. It is based on the Bayes theorem with independence features. Each class labels are estimated through probability of given instance. It needs only small amount of training data to predict class label necessary for classification. Naïve Bayes is particularly effective for data sets containing multiclass predictors (Gandhi 2018, Shubam 2018). For the current study, Naïve Bayes has been implemented with 10-fold cross validation.

C. *Multinomial Logistic Regression Implementation* :

It is also called as multiclass classification. Target variable can take more than two value and values should be ordered. The multiclass logistic regression is similar to binary logistic regression, except the label (class) is now an integer in {1, 2, … C} where C is number of classes. Scores for all classes are calculated. It is implemented using Softmax function. Class with most votes is chosen for prediction (Martin, Nagesh 2019). For the experimental study, Logistic regression on all parameters for multiclass labels has been implemented in Matlab using function mnrfit and mnrval returns the predicted probabilities for the multinomial logistic regression model with predictors X, and the coefficient estimates, B.

D. Artificial Neural Network Implementation ANN :

Is one of the most used techniques for the prediction model. ANN is usually based on imitation of human brain; just like our brain it has neurons for transmitting one data to another. All the neurons are connected together in layers. The application of ANN is widely used in agriculture practices. It compares patterns nonlinear effect and underline concept of the relation between them and hence it is a kind of ML technique which has a vast memory. One of the disadvantages of ANN is that where the dataset is significantly different compared to trained data set (Chauhan 2019, Hardesty 2017). For experimental set up, ANN has been implemented with Scale Conjugate Gradient algorithm (trainscg) with 10-fold cross validation for mustard crop yield prediction. ANN model consists of 5 input layer neurons, 21 neurons in hidden layer and 3 output layer neurons.

E. Random Forest Implementation :

Random Forest is a supervised learning algorithm. It creates a forest and makes it somehow random. 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 its simplicity and the fact that it can be used for both classification and regression tasks which form the majority of current machine learning systems (Yiu 2019, Donges 2019). For the current study, fitrensemble has been applied on collected dataset with 10-fold cross validation

REFERENCE:

Camps-Valls G, Gomez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martin-Guerrero J D, Moreno J, "Support Vector Machines for Crop Classification using Hyper Spectral Data", Lect Notes Comp Sci 2652, 2003, pages : 134-141.

Wu Fan, Chen Chong, Guo Xiaoling, Yu Hua, Wang Juyun, Prediction of crop yield using big data, 2015 8th International Symposium on Computational Intelligence and Design

Araby, A. A., Abd Elhameed, M. M., Magdy, N. M., Abdelaal, N., Abd Allah, Y. T., Darweesh, M. S., ... & Mostafa, H. (2019, May). Smart IoT Monitoring System for Agriculture with Predictive Analysis. In 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST)

S.Veenadhari, Dr Bharat Misra, Dr CD Singh.2019."Machine learning approach for forecasting crop yield based on climatic parameters.”. International Conference on Computer Communication and Informatics

Brownlee, J. (2016). K Nearest Neighbors for Machine Learning. Retrieved March 23, 2020, from https://machinelearningmastery.com. Chauhan, N. S. (2019). Introduction to Artificial Neural Networks (ANN).

Retrieved March 26, 2020, from https://towardsdatascience.com. Donges, N. (2019). A Complete Guide to the Random Forest Algorithm. Retrieved March 30, 2020, from https://builtin.com/data-science/random-forest_algorithm.html. Gandhi, R. (2018).

Naïve Bayes Classifier. Retrieved March 25, 2020, from https://towardsdatascience.com. Hardesty. L. (2017). Explained: Neural Networks. Retrieved March 26, 2020, from https://news.mit.edu/2017/explained_neural-networks-deep-learning-0414. Harrison, O. (2018).

Machine Learning Basics with the K_Nearest Neighbors Algorithm. Retrieved March 23, 2020, from https://towardsdatascience.com.

Jayalakshmi, R. & Devi, M. S. (2019). Relevance of Machine Learning Algorithms on Soil Fertility Prediction using R. International Journal of Computational Intelligence and Informatics, 8(4), 193-199. Martin, K. Logistic Regression Models for Multinomial and Ordinal Variables. Retrieved March 27, 2020, from https://www.theanalysisfactor.com/logistic-regression_models-for -multinomial-and-ordinal-variables.html

Nagesh, S. (2019). Real world implementation of Logistic Regression. Retrieved March 27, 2020, from http://towardsdatascience.com. Priya, P., Muthaiah, U. & Balamurugan, M. (2018). Predicting Yield of the Crop Using Machine Learning Algorithm. International Journal of Engine

Athmaja S., Hanumanthappa M, “Applications of Mobile Cloud Computing and Big data Analytics in Agriculture Sector: A survey”, October 2016