Prediction and Analysis of Crop Yield using Machine Learning Techniques

Manoj G S[1], Prajwal G S[2], Ashoka U R[3], Prashant Krishna[4], Anitha P[5]

Department of Information Science and Engineering
JSS Academy of Technical Education, Bengaluru, Karnataka.

Abstract **- India being agricultural dependent country the economic status of the country is completely and partially dependent on this. Agricultural yield is affectedby the organic, economic and seasonal causes. Estimation of agricultural output is a big challenging task for this country as of the population status taking in consideration. In recent days, the ppl growing these products and such products are very much unstable to be produced due to the sudden weatherly environmental reasons and lack of ground hydro resources. The main objective is to collect data that can be stored and analyzed for forecasting the crop yield. For prediction of crop yield machine learning techniques are implemented. This helps the farmers to choose the best suitable crop. Also, this paper aims at bringing an enhancement in the field of agriculture by achieving better results in predicting crop yields. With the use of machinelearningtechniqueswithproperoptimizations,astatisticalmodelisbuilttoprovideaccurateandprecise decision. The output of this work would help farmers pick most suitable crops to be grown depending on the factors like season and area available with least possible chances oflosses.**

Keywords: Agriculture, prediction, optimization

1. INTRODUCTION

Agriculture is the main occupation in India and economy of the country is entirely depended on it for rural based existence [6]. Because of certain components like atmosphere changes, unpredicted precipitation, decline of water level, utilization of pesticides unnecessarily and so on. The degree of farming in India is diminished. The primary goal of this exploration workis to give a technique with the goal that it can perform illustrative examination on crop yield creationinacompellingway.Albeit,afewexaminationsuncoveredmeasurabledataaboutthe farming in India, hardly any examinations have researched crop forecast dependent on the memorable climatic and creation information.

The accuracy of 87% is achieved from the system and high correlation is seen between yield of crop and the climate. Agriculture inputs like chemicals, pest, soil quality and many more inputs were not considered for change in agriculture from field to field. This model is going to help farmers to make better decisions as to decide which crop to plant. Based on the season's climate it will help farmers to make important decisions, such as import, export, pricing, marketing before the crop is harvested [1].Crop production was influenced by the various economy, season and biology pattern [2]. Catastrophic changes in the patterns may cause a immerse loss for farmers. These lose can be avoided by implementing smart farming methodology that is incorporating technology inday-to-dayfarming.Thesemodelmainlyfocusesonweatherforecasting,croptypeplantation,crop prediction,andcropcostforecasting.Statisticalagriculturedatasetisconsideredforthismodel. Then it is pre-processed and classified into training and testing data. Support Vector Machine and Random Forest algorithms are used for good accuracy. The final output is to predict the yield of crop

Agricultureinthiscountryassumesasignificant job in economy and work [6]. The basic trouble present among Indian ranchers are they don't settle on the best possible crop dependent on the dirt necessities. Along these lines the profitabilityisinfluenced.Thisissueoftheranchershasbeensettledthroughaccuracyfarming. This technique is described by a dirt database gathered from the ranch, crop gave by farming specialists, accomplishment of parameters, for example, soil by soil testing lab datasets[6].

Inthiswork,multiplelinearregression,decisiontreeregression,polynomialregression isusedtodepictthecropoutputforvarioustypesofcropsacrossthestatesofIndiaandk-means clustering algorithm to classify the states of the country low, average and high production clusters. Machine learning technique for crop yield prediction helps farmers to track the soil quality, depending on the approach-based application of data mining [8][9]. Also, soil qualitycan be predicted for different crops, so that crop suitable for cultivation by soil type andoptimizes the crop yield by recommending effective fertilizer. The program aims to help farmers grow proper crops to achieve better yields[8].

II. LITERATURE SURVEY

Many applications are available for farmers to predict the yield of crop based on the climatic conditions [1]. Machine Learning algorithms were used to predict the crops.Random forest algorithmisusedforthefiveclimaticparameterstotrainthemodelbutotheragricultureinputs like soil quality, pest, chemicals used, etc. are not considered. The model was trained by 200 decisiontreestoconstructrandomforest.10-foldcrossvalidationwasusedforaccuracy of the trained model.

and classify the crop yield as best bio condition, good bio condition, poor bio condition. It is difficult to achieve smart farming in developing country because many of the farmers are illiterate and unaware of the technology. The project is now a web based so, in future this project is aim to develop an android and iOS application[2].

MachinelearningmodelbasedonConvolutionalNeuralNetworks(CNNs)ispresentedforthe yield prediction [3]. The main objective is to check crop and weed detection and also yield prediction. Convolutional Neural Networks (CNNs) are used in this analysis to create a crop yield prediction model based on the Normalized Difference Vegetation Index (NDVI) and RGB data acquired from Unmanned Aerial Vehicles (UAVs). The effect on predictive effectivenessofvariousaspectsoftheCNNsuchasselectionofthetrainingalgorithm,network size, regularization strategy, and tuning of the hyper parameters was evaluated. The results indicatethatintheearlystagesofgrowth,thebestperformingmodelcanpredicttheyieldwith a mean absolute error of 484 kg based solely on RGB images [3]. At later growth of point,the modelforRGBimagesreturnedhighererrorvalues.TheCNNsoftwareworkedslightlybetter with RGB data than with the NDVI data. The proposed system is not trained on a larger set of features like (climate and soil) along with time series image data to tune the trained model for accuracy.

Descriptionanalysisistheinitialandunderlyingconditionofexamination[4].Itisaprocedure wherein we can comprehend what occurred before and we can also realize that past is the best indicatorofthingstocome[4].Descriptionanalysisisappliedinthehorticultureoragriculture related creation area for sugarcane harvest to discover productive harvest yield estimation. Three datasets like Soil dataset, Rainfall dataset, and Yield dataset. Consolidated dataset is formedandbasedonjoinedset,somedirectedmethodsareappliedtolocatetherealevaluated cost and the precision of a few strategies. Also, three directed procedures are utilized like K- Nearest Neighbor, Support Vector Machine, and Least Squared Support Vector Machine. It is a near investigation that tells the precision of preparing proposed model and blunder rate. The precision of preparing model ought to be higher and mistake rate ought to be least. The proposedmodelcangivetherealexpenseofassessedcropyieldanditisnameasLOW,MID, and HIGH[4].

There are three datasets named as Soil dataset, Rainfall dataset, Yield dataset. These datasetsincorporateafewparameterswhichareusefultoknowthestateofharvestsandgroup the information into independent classes by performing directed preparing on the dataset that are gathered from farming area. This framework has the ability to perform both the characterization just as relapse. In the characterization step the information is groupedintothreeclasses(low,mid,andhigh),thoughinrelapse stepthegenuineexpenseofyield creation isassessed.Weutilizedthreesignificantcalculationsofmanagedlearning,forexample,KNN, SVM and LS-SVM to prepare and construct a model. This framework is work for organized dataset. In future we can actualize information free framework moreover. It implies organization of information whatever, our framework should work with sameproficiency.

Honest strive has been made to concentrate on utilization of information mining proceduresin the farming field [5]. Strategies and many calculations are made and utilized. In this module informationminingmethodsarebuild,whichutilizespastdatalikesoiltype,soilpH,ESP,EC ofaspecificdistricttogivebetterharvestandyieldestimationforthatdistrict.Thismodelcan beutilizedtoselectthemostastoundingharvestsforthedistrictandfurthermoreitsyieldthere by improving the qualities and addition of cultivating too. This helps ranchers to choose the harvest they might want to plant for the inevitable year. Expectation will help the related ventures for arranging the coordination’s of theirbusiness.

There is a thorough investigation of the agricultural land soil informationutilizingJ48 calculationandforecasttechniques.hereitisexhibitedacharacterizationcalculationcalledJ48 (C4.5) utilizing Weka device. J48 is straightforward classifier to make a choicetree,however it gave the best outcome in the analysis. According to the dirt example given to labfortesting andeditingdesigntheframeworkwillsuggestreasonablecompost.Itdevisestofabricate FertilizerRecommendationFrameworkwhichcanbeusedviablybytheSoilTestingLabs[5].

Variousarrangementtechniquestoarrangetheliverillnessinformationalcollection[6].

The paper stresses the requirement for precision since it relies upon the dataset what's more, the realizing calculation. Characterization calculations such as Naïve Bayes, ANN, ZeroRand VFIwereutilizedtoarrangethesesicknessesandlookattheadequacy,rectificationrateamong them. The presentation of the models was contrasted and precision and also takes computational time. It was presumed that all the classifiers with the exception of naive bayes demonstrated improved prescient execution. Multilayer perceptron shows the most elevated exactness among the proposed calculations. This paper’s work would assist ranchers with increasing profitability in farming, forestall soil corruption in developed land, furthermore, diminish synthetic use in crop creation and productive utilization of water assets. This paper’s futureworkisfocusedonanimprovedinformationalcollection withenormousnumberoftraits and likewise executes yield expectation[6].

Anewapproachtocropyieldpredictionisimplemented basedontherelationshipbetweenthe Multi Linear Regression (MLR) and Artificial Neural Network (ANN) [7]. For this research workahybridMLR-ANNmodelwasproposedforeffectivecropyieldprediction.Theweights andbiasofinputandhiddenlayerareinitializedrandomlyinconventionalANNmodel.Instead of random weights and bias initialization, this hybrid MLR-ANN model initializes the input layer weights and bias by using the MLR coefficients and bias. The prediction accuracy ofthe hybrid model is compared with the models ANN, MLR, Support Vector Regression (SVR),k- Nearest Neighbor (KNN), and Random Forest (RF) using performance metrics. The computationaltimewascalculatedforboththehybridMLR-ANNandconventionalANN.The findings show that the proposed MLR-ANN hybridmodel provides greater precision than the traditional models. It finds the near optimum minimum of error and increases the accuracy of the prediction. Using supervised and unsupervised learning algorithms, such as BPN (Back Propagation Network) and Kohonen Self Organizing Map(Kohonen's SOM) are used for prediction of soil quality. Dataset is then trained through network learning. The system uses unsupervisedandsupervisedmachinelearningalgorithmsand deliversthebestaccuracy-based results. The results of the two algorithms will be compared and the one which gives the best and precise output will be chosen. The system will according to the usage for each algorithm, as it is help to lessen the farmers' difficulties. This resultsinprovidingthefarmerswithefficientinformationneededtoobtainhighyieldandthus maximize profits[8].

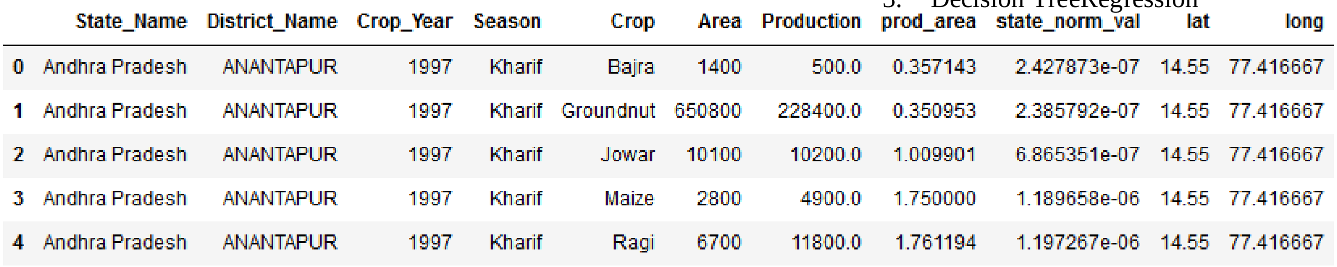
III IMPLEMENTATION

1. Dataset Description

The dataset which has been used in this project is collected from the Government agricultural website,India.Thisdatasetcontains44,397rowsofdataconsistingof11columnsofattributes each attribute describes the right information sufficient to predict data and also classify according to the purpose of usage. The dataset contains 12 states information including their 84 districts of data. The dataset gives us the precise data from the year 1997 to2014.

Fig1shows a small sample of the dataset used. The above mentioned11columnsall the column data as shown in the figure of the snap shot description including Latitude and longitude of that region respectively. In this work, data is pre- processed.

Fig.1: Snapshot of dataset description



	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production	prod_area	state_norm_val	lat	long
0	Andhra Pradesh	ANANTAPUR	1997	Kharif	Bajra	1400	500.0	0.357143	2.427873e-07	14.55	77.416667
1	Andhra Pradesh	ANANTAPUR	1997	Kharif	Groundnut	650800	228400.0	0.350953	2.385792e-07	14.55	77.416667
2	Andhra Pradesh	ANANTAPUR	1997	Kharif	Jowar	10100	10200.0	1.009901	6.865351e-07	14.55	77.416667
3	Andhra Pradesh	ANANTAPUR	1997	Kharif	Maize	2800	4900.0	1.750000	1.189658e-06	14.55	77.416667
4	Andhra Pradesh	ANANTAPUR	1997	Kharif	Ragi	6700	11800.0	1.761194	1.197267e-06	14.55	77.416667

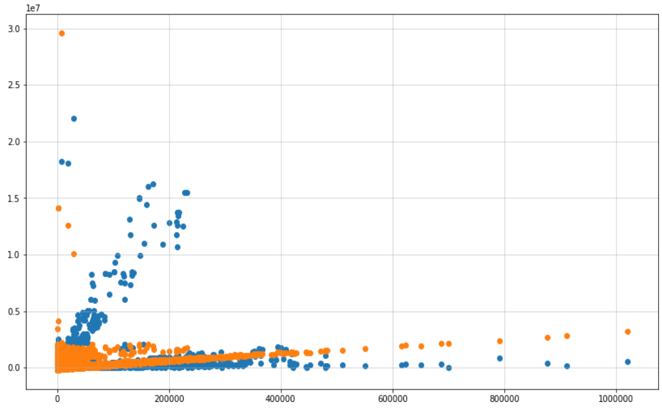
checked for Null values and dropped down some of the columns, which were not required for the algorithm toperform.

2. Multilinear Regression

Multilinear regression is used in this project to predict the yield, for this the algorithm uses many self-dependent variables to predict the result of the respective variable. This is a useful modelforsearchingthecorrelationbetweenthetwoparameters, independentvariables(usedto make predictions) and the dependent variable (the values to bepredicted).

Fig 2: Area vs Production graph using Multilinear Regression

Fig 2 shows the graph for production versus area. Itcan be observed that prediction cannot be achieved as the graph doesn't show quite a straight line with the actual value and R2 score accuracy was found to be28%.



3. Decision TreeRegression

Decision tree algorithm is used to build a regression or a classification model in the form of a tree structure. This is done by breaking a dataset into smaller subsets and at the same time, associated decision tree is developed incrementally. The final tree consists of root node, decision nodes and leaf nodes. since the accuracy score obtained by the multiple linear regression algorithm is very less, decision tree regression has been used to achieve a better accuracy.

Fig. 3 Area vs Production graph using Decision tree regression

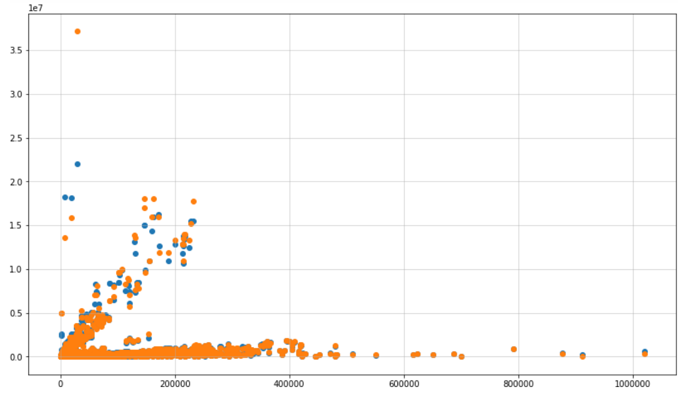


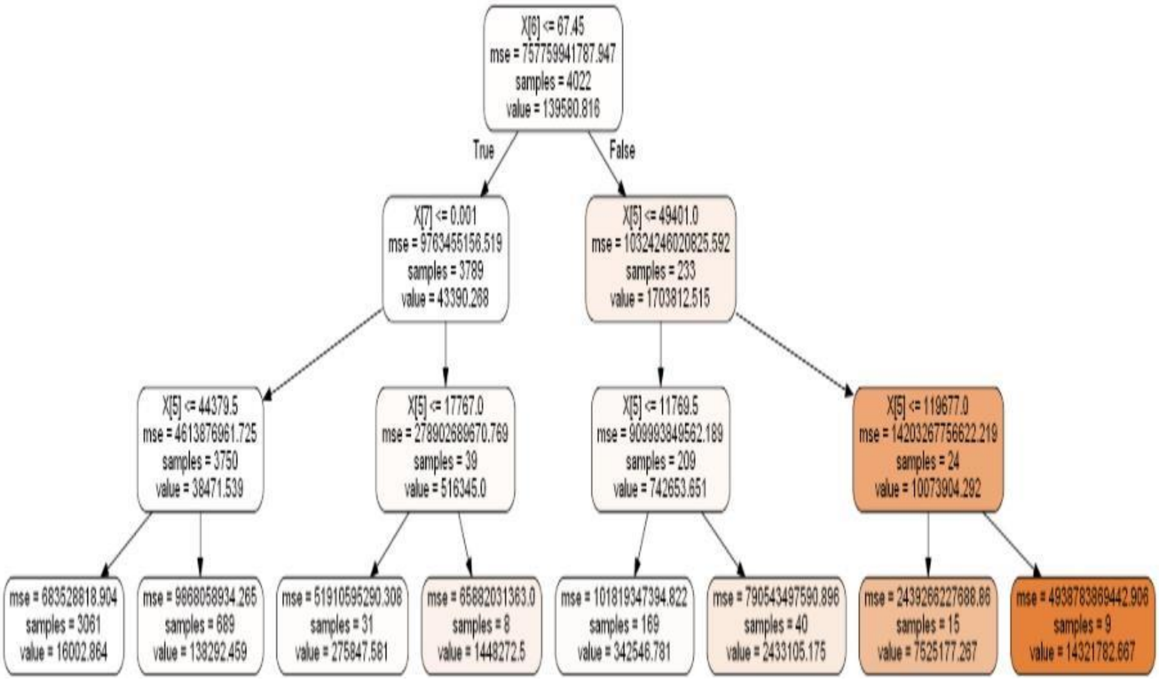
Fig. 3 shows the area vs production graph plotted using decision tree regression algorithm. It can be observed from the graph that the predicted values and the actual values are very close, unlike in multi linear regression and hence an accuracy score of 95.7% is obtained using the decision treeregression algorithm. Since the dataset is too large to construct a single decision tree for the entire dataset, separate decision trees have been generated for each of the states in India.

Fig. 4 shows the decision tree for the state Karnataka. Each decision node contains four different fields-attribute names, mean squared error, samples and production value. Each leaf node consists of three fields- mean squared error, samples and production value.

4. K-Means Clustering

Clustering is the method of dividing the points into the known proportion of batches, so that d points within the same batches are same and different from points in different batches. One of thebestclusteringisKMeansclustering. KMeans aims to split the observation into k clusters in which each observation belongs to the nearest mean cluster or cluster centroid. States are divided into three clusters based on the production, that is low production state, average production state, highproductionstate.Thereisacentroidforeachclusteringgroup,states are classified as low production state, average production state, high production state based on the mean distance between the centroid. From the result it is observed that states AP, BR, GJ, HR, MP, RJ, West Bengal are classified as low production state as their distance from the low production centroid is near compared to the other clustering centroid. Karnataka, Maharashtra and Punjab are classified as average production state for the year 1999. Tamil Nadu and Uttar Pradesh are classified as high production states for the year 1999.

Fig. 4 Decision Tree for production in Karnataka

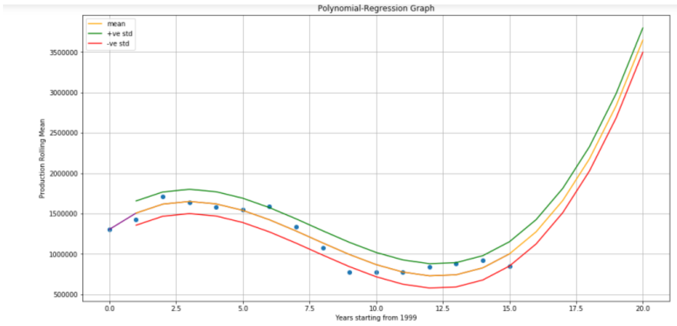


5. Rolling Mean and Poly Regression

Rolling Mean is a calculation for analyzing futuristic predictions by using the fluctuation in the series and graphs. In our project, we take the state name as input from the user andapplyrollingmeanandpolynomialregressionalgorithmsforthatstate.Forthegivenstate, we calculate the total production for every year by adding the production from every crop and every district. Here, we have taken the mean for 3 years which means the value for the third year will be the average of first, second and the third year's total production. And the rolling meanforfirsttwoyearswillbenull.Therollingmeanvaluesforeachyeararestoredandthen we apply polynomial regression on the rolling mean values. In the polynomial regression we havetakenthepolynomialdegreeasthreebecausechoosingalowervaluewillbeunder-fitting and taking a greater number will beover-fitting.

Poly regression algorithm is used to predict the rolling mean values for the next five years. The graph for the poly regression is shown in Fig 5. Here, the blue dots represent the rollingmeanvaluesfrom1999to2014.Noticetherearenodatapointsafter2014asthedataset only contains values till 2014. The values after 2015 are predicted using the polynomial regression algorithm. The orange color line is used to represent the mean. The green and red lines are used to depict the range for the calculated mean. The state name, district name, the crop producing the highest yield in a district and the polynomial regression graph for the state is displayed to the user as theoutput.

Fig. 5 Polynomial-Regression graph.



IV.Comparison analysis between the Multilinearregression and Decision Tree Regressor

Comparison between Multiline arregression and Decision Tree Regressor is done to find the best fit for prediction model for this work. The r2 score in decision tree regressor was found out to be 95.7% as compared with the multilinear regression. So, Decision tree regressor is considered as the best fit for the prediction of yield. From this it is came to know decisiontree regressor is easy to interpret than multilinear regression.

Multilinear regression is good when the relationship between variables are straight or linear. In this dataset the data being highly variable because of which decision tree regressor is used for further prediction andanalysis.

V. CONCLUSION

Crop yield prediction has been a challenging issue for farmers since many years. This work mainly focuses on analyzing the production of crop yield in India from 1999 to 2014, and to predict the yield for the next 5 years using the machine learning techniques. After seeing the results,wecanconcludethatthedecisiontreeregressoroutscore stheotheralgorithmsused,in termsofaccuracy.Perfect depictionat the varioushorticultural outputswill definitely be favorable to the people practicing thistoimprove this output beneficially.

REFERENCES

[1] Abhijeet Pandhe, Praful Nikam, Vijay Pagare, PavanPalle,Prof. DilipDalgade Crop Yield Prediction based on ClimaticParameters,2019.

[2] Sowmitri B S 1, Hemanth Harikumar 2, R MeeraRanjani 3, Prathibha D 4 Smart Farming Crop Yield Prediction using Machine Learning,2018

[3] Petteri Nevavuorib, Nathaniel Narraa, TarmoLippinga, Crop yield prediction with deep convolutional neural networks, 2019.

[4] Arun Kumar, Naveen Kumar, Vishal Vats, Efficient crop yield prediction using machine learning algorithms,2018.

[5] Pooja M C, Sangeetha M, Shreyaswi J Salian, Veena Kamath, MithunNaik, Implementation of CropYield Forecasting Using Data Mining,2018.

[6] Ramesh A. Medar and Vijay S. Rajpurohit, A Survey of data mining techniques for crop yield prediction, IJARCSMS, Volume 2, Issue 9, September2.

[7] P.S.Maya Gopal, R.Bhargavi A novel approach for efficient crop yield prediction,2019.

[8] Neelambika B Hiremath, Dr. Dayannada P, “Machine Learning Techniques for Analysis of Human Genome Data”, International Journal of Smart Education and Urban Society, IGI Global, Volume 10, Issue 1, Article 5, 2018, DOI:10.4018/IJSEUS.2019010105.

[9] RushikaGhadge,JuileeKulkarni,PoojaMore,SacheeNene,PriyaRLPre dictionofCrop Yield using Machine Learning,2018