DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5784

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A PREDICTION SYSTEM FOR FARMERS TO ENHANCE THE AGRICULTURE YIELD USING COGNITIVE DATA SCIENCE

N. Muthurasu
Department of Computer Science
SRM Institute for Science and Technology
Vadapalani, Chennai, India

M. Narayanan Ramanathan
Department of Computer Science
SRM Institute for Science and Technology
Vadapalani, Chennai, India

Sahithyan S
Department of Computer Science
SRM Institute for Science and Technology
Vadapalani, Chennai, India

M. Aravind
Department of Computer Science
SRM Institute for Science and Technology
Vadapalani, Chennai, India

RamanagiriBharathan A
Department of Computer Science
SRM Institute for Science and Technology
Vadapalani, Chennai, India

Abstract: This paper gives an idea about how to discover additional insights from precision agriculture data through big data approach. Big data analytics in agriculture applications provide a new insight to give advance weather decisions, improve yield productivity and avoid unnecessary cost related to harvesting, use of pesticide and fertilizers. There are number of numerical weather models and algorithms that have been developed and enforced to predict the weather forecasting.

Keywords: Agriculture yield; Regression models; Predictive Models; Tensor flow; Big Data; Weather forecasting; Data Analytics; Predictive Analytics; NLP; Chatbot

I. INTRODUCTION

India ranks second worldwide in farm output. Agriculture yield of India constituted for 17.32% of the GDP (Gross domestic product) in 2016. Strenuous efforts are taken by government and other organization to improve the yield of agriculture in India. The Implementation model is done for a particular region of southern India. The Thanjavur district lies in the southern eastern part of Tamil Nadu. This district majorly lies on the Kaveri river belt and is considered to be one of the finest region for agriculture. This territory is one of the major rice producing region and is considered as the rice bowl of Tamil Nadu. The agriculture yield depends on various factors, especially where the water resources are scarce, agriculture is mostly dependent on precipitation.

In agriculture, crop yield is a measurement of the amount of agricultural production harvested per unit of land area and the seed generation of the plant itself. The major causes of low productivity of agricultural yield are Human factors, such as lack of training and efficiency of farmers, Traditional methods of cultivation, Problems of soil, pollution, inadequate irrigation facilities, unreliable monsoons, inefficient predicative models, fertilizers and pesticide and improper marketing strategies etc. Crop yield has a direct impact in the national economy as it constitutes 50% of the workforce and plays vital part due to food management. The prediction of yield of a particular crop before harvest is very significant for farmer as they will be well informed about the yield and can plan their future ideas.

The main factors affecting crop yield are the inputs and weather [1]. A crop prediction model assist in estimating crop yield that is expressed as function of precipitation, weather condition and nutrients present in soil. In developing a good probability model to predict crop outcome, predictive modelling is taken which involves 4 stages namely Descriptive analysis, Data treatment, Data modelling and estimation performance. One of the predictive modelling technique is regression analysis, it observes the relationship between a dependent (target) and independent variable (s) (predictor). This technique helps to forecast through time series data and finds the underlying effect among these variables. Regression analysis analyses the data and a line or curve is fitted using the data points, thereby helps in minimizing the distance of the corresponding data points from the line/ curve.

II. ARCHITECTURE

The predictive model architecture consists of major data source which depends on eleven factors that are considered. These factors are average temperature, cloud cover, Diurnal Temperature, Ground Frost Frequency, Maximum Temperature, Minimum temperature, potential

evapotranspiration, Precipitation, Reference crop Evapotranspiration, Vapor Pressure and wet day frequency. These data are collected from India Water Portal and Open Government Data Platform India. These data are collected in comma separated value (CSV) format and then preprocessed for noise reduction and imputation and do refinement using Excel. After cleaning and segregation of the data, then it is again converted CSV format. After the classification of data, a predictive algorithm is performed using deep learning neural network to determine the yield of the crop for that a particular year. The performance of them will be compared and tested. This result can be queried using web deployed chat-bot using Natural Language Processing (NLP) [2]. These results benefits data analysts or data scientists or market researchers to calculate and eliminate the group of variables used for developing models for prediction.

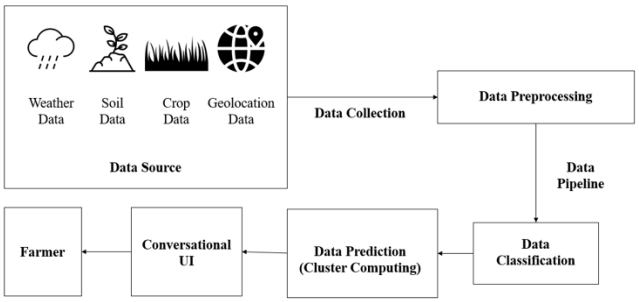

Figure 1. Architecture

III. PREDICTIVE ANALYTICS

A. *What is Predictive Analytics*

Predictive analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future. It uses a number of data mining. Predictive analytics allows organizations to become proactive, forward looking, anticipating outcomes and behaviors based upon the data and not on a hunch or assumptions. Prescriptive analytics, goes further and suggest actions to benefit from the prediction and also provide decision options to benefit from the predictions and its implications.

B. *Predictive Analytics in Agriculture*

As the advancement in the technology is increasing in a broader way the usage of its technology to Agriculture sector is very limited. Agriculture is the backbone of the Indian economy. But in last few years, agriculture has taken down a huge steep due to many reasons. Lack of knowledge of seasons, crops and price are also one of them i.e. not knowing which season for which crop, and which crops are receiving the higher payment for the government. Our project deals with the analysis of agricultural data which helps in better understanding of agriculture in India.

The outcome of the crops depends on the several factors like rainfall, season, temperatures, and also the price announced by the government. In the project, we will categorize the crops based on the season, based on Minimum price announced by the government, based on temperature. The data also allows the user to visualize a different kind of data. The data also contains different crops planted in Area (Hectares) and Production (metric ton). The project also visualizes the different categories of data for the better of understanding of Agriculture in India.

C. *Statstically Approach*

Statistical Analysis is the study of the collection, organization, analysis, interpretation and presentation of data. Statistical Analysis begins with the identification of process or population in consideration. The population is collection of observation of the process at various times known as at time series and data from each of the observation serves as a member of the overall group.Statistical Analysis - Descriptive statistics and Inferential statistics: Descriptive statistics summarize the population data in consideration by describing what was observed in the sample graphically or numerically. Numerical descriptors are mean and standard deviation for continuous data types. Frequency and percentage are more useful and used while describing categorical data[3].

Here are the statistically approaches we adopted to

- Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong naive independence assumptions[4].
- Support vector machines re supervised learning models with associated learning algorithms that analyze data and recognize patterns, are used for classification and regression analysis[5].
- Regression analysis Estimating the relationships among variables[4].
- Time series analysis: A sequence of data points, measured at successive points in time[4].

D. *Deep Learning Approach*

Predictive modelling is the process of creating, testing and validating a model to best predict the probability of an outcome. A number of modelling methods from machine learning, artificial intelligence, and statistics are available in predictive analytics software solutions for this task. The model is chosen on the basis of testing, validation and evaluation using the detection theory to guess the probability of an outcome in a given set amount of input data. Models can use one or more classifiers in trying to determine the probability of a set of data belonging to another set. There're different models available on the Modelling portfolio of predictive analytics software enables to derive new information about the data and to develop the predictive models. Each model has its own strengths and weakness and is best suited for particular types of problems.

A model is reusable and is created by training an algorithm using historical data and saving the model for reuse purpose to share the common business rules which can be applied to similar data, in order to analyze results without the historical data, by using the trained algorithm.

Most of the predictive modeling software solutions has the capability to export the model information into a local file in industry standard Predictive Modeling Markup Language, format for sharing the model with other Predictive Modeling Markup Language compliant applications to perform analysis on similar data [6].

The process involves running one or more algorithms on the data set where prediction is going to be carried out. This is an iterative processing and often involves training the model, using multiple models on the same data set and finally arriving on the best fit model based on the business data understanding [7].

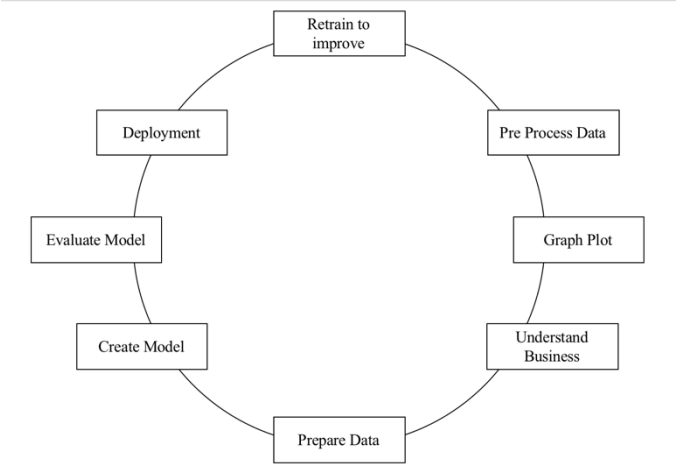

Figure 2. Deep Learning Process

IV. PROPOSED SYSTEM

Having built 132 Deep Learning Neural Networks using Tensor Flow to predict values for a given month and a year and making all 132 predicted values as input to the yield value prediction deep learning neural network. Making the above process automated and use of NLP to make it more user friendly.

A. *Sample Data*

	A	B	C	D	E	F	G
1	Year	Jan	Feb	Mar	Apr	May	Jun
2	1901	26.57	27.17	27.722	29.311	30.186	29.887
3	1902	25.605	25.392	27.809	29.892	30.515	30.106
4	1903	25.955	27.047	28.132	29.95	30.211	29.904
5	1904	25.058	25.479	27.326	29.937	30.103	29.367
6	1905	24.973	26.65	28.739	29.199	30.103	30.061
7	1906	25.965	27.457	27.934	30.501	30.884	29.834
8	1907	25.34	26.173	28.127	29.354	30.602	29.854
9	1908	25.463	25.766	27.435	30.626	30.401	29.825
10	1909	25.354	26.25	28.048	29.736	29.91	29.365
11	1910	25.518	25.931	27.32	29.401	30.797	29.675

Figure 3. Sample Dataset for Thanjavur District's Average Temperature between 1901-2002

B. Factor Prediction Neural Network Configuration

Properties	*Value*
Learning Rate	0.01
Training Epoch	50000
Number of Inputs	1
Number of Outputs	1
Layer 1 Nodes	50
Layer 2 Nodes	100
Layer 3 Nodes	50

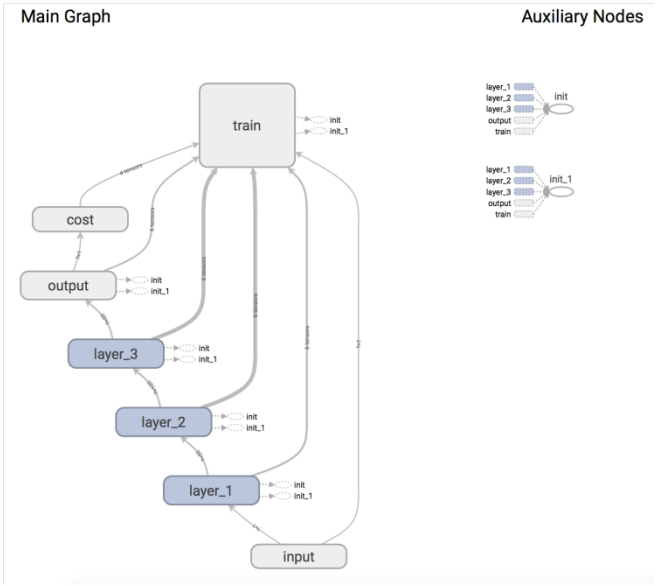

Figure 4. Deep Learning Neural Network Graph

C. Yield Prediction Neural Network Configuration

Properties	*Value*
Learning Rate	0.01
Training Epoch	50000
Number of Inputs	132
Number of Outputs	1
Layer 1 Nodes	500
Layer 2 Nodes	1000
Layer 3 Nodes	500

V. RESULTS

D. *Linear Regression*

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

Regression	*Metrics*		
	Coefficient	*Intercept*	*Accuracy score*
LR	0.00786195	16.56081794	0.39231505008674633

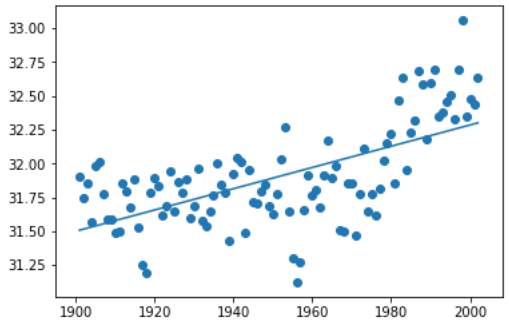

Figure 5. Maximum Temperature vs Year in Linear Regression

The accuracy percentage of the linear regression model is found to be very low and unreliable.

E. Bayesian Ridge Regression

Bayesian linear regression allows a fairly natural mechanism to survive insufficient data, or poor distributed data. It allows you to put a prior on the coefficients and on the noise so that in the absence of data, the priors can take over. More importantly, you can ask Bayesian linear regression which parts (if any) of its fit to the data is it confident about, and which parts are very uncertain (perhaps based entirely on the priors).

Regression	*Metrics*		
	Coefficient	*Intercept*	*Accuracy score*
BRR	0. 00774687	16.785406927	0.39223098533533796

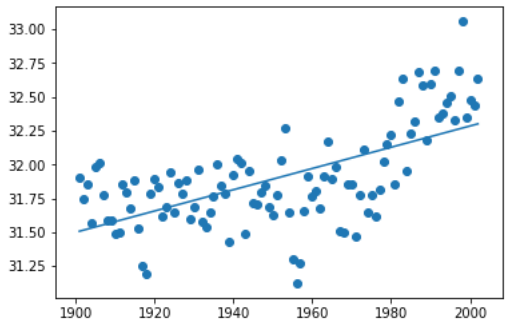

Figure 6. Maximum Temperature vs Year in Bayesian Ridge Regression

The above score achieved from Bayesian regression is similar to the score achieved from the linear model. Hence this Regression is also proved to be inefficient for this data set.

F. Huber Regression

The Huber Regressor optimizes the squared loss for the samples where |(y - X'w) / sigma| < epsilon and the absolute loss for the samples where |(y - X'w) / sigma| > epsilon, where w and sigma are parameters to be optimized. The parameter sigma makes sure that if y is scaled up or down by a certain factor, one does not need to rescale epsilon to achieve the same robustness.

Regression	*Metrics*		
	Coefficient	*Intercept*	*Accuracy score*
HR	0.0163423	0.00638954616843	-0.064328675981744077

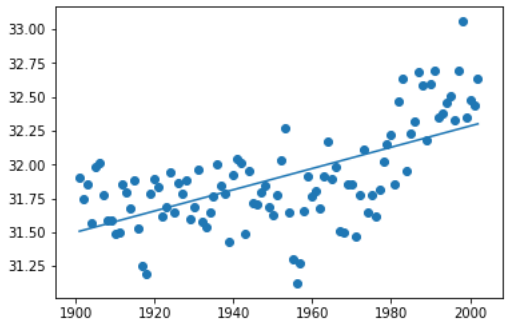

Figure 7. Maximum Temperature vs Year in Huber Regression

The efficiency value is found to be in negative and hence the score achieved by the Huber regression is unreliable.

G. Support Vector Regression

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis

Regression	*Metrics*		
	Coefficient	*Intercept*	*Accuracy score*
SVR	0.0163423	0.00638954616843	-0. 21506065704930322

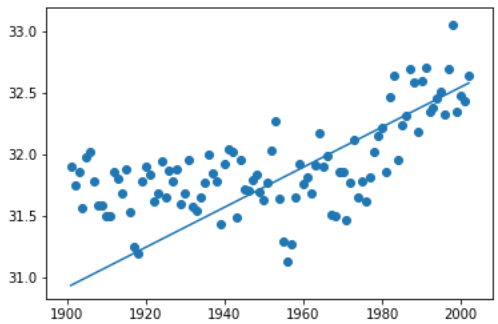

Figure 7. Maximum Temperature vs Year in Support Vector Regression

The Accuracy score achieved from the support vector Regression is also found to be in negative and hence we propose a Deep Learning Neural Network to arrive at a reliable predictive solution.

H. Deep Learning Neural Network

Deep learning is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.

Deep learning models are loosely related to information processing and communication patterns in a biological nervous system, such as neural coding that

attempts to define a relationship between various stimuli and associated neuronal responses in the brain.Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics and drug design, where they have produced results comparable to and in some cases superior to human experts.

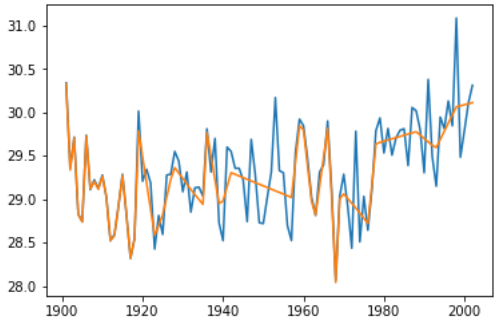

Figure 8. Maximum Temperature vs Year in Deep Learning Model – Actual vs Predicted

Regression	***Metrics***	
	Cost	***Accuracy score***
DNN	0.01011510007083416	92.4599868774

CONCLUSION

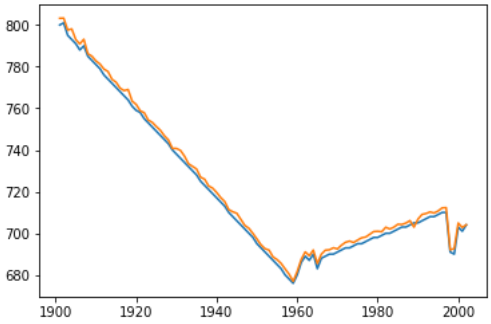

Figure 9. Yield per Hectare vs Year - Actual vs Predicted

We found out that the regression model is unreliable and inefficient. Hence, we conclude that Deep Learning Neural Network is the best fit for crop yield estimation because of its accuracy and reliability. It is only neural networks that keeps improving its accuracy for every prediction. We have built a chat bot to make the app more usable for the farmers, making data science easily available for common man.

V. REFERENCES

[1] Bendre, M. R., et al. "Big Data in Precision Agriculture: Weather Forecasting for Future Farming." *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 2015, doi:10.1109/ngct.2015.7375220.

[2] Devi, Manmita, and Mohit Dua. "ADANS: An Agriculture Domain Question Answering System Using Ontologies." *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, doi:10.1109/ccaa.2017.8229784.

[3] Jayanthi, D., and G. Sumathi. "Weather Data Analysis Using Spark — An in-Memory Computing Framework." *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2017, doi:10.1109/ipact.2017.8245142.

[4] Bose, Pritam, et al. "Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, 2016, pp. 6563–6573., doi:10.1109/tgrs.2016.2586602.

[5] Shah, Purnima, et al. "Towards Development of Spark Based Agricultural Information System Including Geo-Spatial Data." *2017 IEEE International Conference on Big Data (Big Data)*, 2017, doi:10.1109/bigdata.2017.8258336.

[6] Pacheco, Anna, et al. "The Impact of National Land Cover and Soils Data on SMOS Soil Moisture Retrieval Over Canadian Agricultural Landscapes." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 11, 2015, pp. 5281–5293., doi:10.1109/jstars.2015.2417832.

[7] Nagini, S., et al. "Agriculture Yield Prediction Using Predictive Analytic Techniques." *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016, doi:10.1109/ic3i.2016.7918789.