CROP YIELD PREDICTION IN AGRICULTURE USING DATA MINING PREDICTIVE ANALYTIC TECHNIQUES

[1] P.Surya, [2] Dr. I.Laurence Aroquiaraj
[1] Ph.D Research Scholar, [2]Assistant Professor
[1]Department of Computer Science
[1] Periyar University, Salem, Tamilnadu, India

Abstract: Data Mining is emerging research field in Agriculture especially in crop yield analysis and prediction. As early into the growing season as possible, a farmer is focused in perceptive how much yield they about to expect. As with many other sectors the amount of agriculture data are increasing on a daily source. In our proposed work, collected agriculture dataset will be used to get crop yield prediction model using various regression techniques. Regression analysis was tested for the effective prediction or forecast of the agriculture yield for various crops in Tamilnadu state particularly in North Western zone of Tamilnadu. North western zone of tamilnadu state data consist four districts. The North western zone of Tamilnadu districts are Dharmapuri, Salem, Namakkal, Krishnagiri. By the analysis depends on the results of predictor model, in the north western zone, under the area having more cultivated crops are Tapiaco, Sugar cane, Ragi, Maize, Groundnut.

Key words: Agriculture, Yield Prediction, Classification, Linear Regression.

I. INTRODUCTION

Agriculture is the main stay of the State economy in that its role is fundamentally instrumental in terms of market, aspect and product contributions. India's economy mainly depends on agriculture yield growth and their related agro industry [12]. Area under agriculture comes to 61 per cent of the total geographical area of Tamil Nadu. Performance of agricultural sector mainly hinges on natural forces such as spatio-temporal distribution of rainfall, temperature, climate etc., with the result any deviation of monsoon from the normal pattern brings about enormous fluctuations in area and production. Crop yield has a direct impact on National and International economies annually and the yield predicted [13] plays a significant part in the food management and agriculture sector. Researchers have proposed statistical model based on time series data like agro- climatic weather variables, Agriculture crop yield that could suggest forecasting process for harvest.

Predictive modeling is a method that uses data mining and probability to forecast outcomes. Data Modeling for Prediction involves four stages namely historical data analysis (Descriptive), Data preprocessing, modeling of Data and Performance Estimation. In Data mining, a Classification technique gives a best solution for prediction process. Regression analysis, it observes the relation between a independent (predictor) and dependent (target) variables. This technique helps to estimate through time series data and finds the underlying effect among these variables [12].

Major contributions of this paper are as follows:

- To proposed a prediction methodology for crop yield in agriculture using data mining predictive techniques.

This paper is organized as follows: Section II describes the Literature Review needed for the research work. Section III discusses the proposed approach. Section IV describes the Methods and Materials required for the study. Dataset description discussed on Section V. Section VI was discussed with Results and analysis. Section VII concludes the work with possible furture enhancement.

II. LITERATURE REVIEW

This section discuss about various related works already done in data mining techniques using agriculture dataset. Most of the researchers focused on the problem for yield prediction.

D Ramesh , B Vishnu Vardhan, researchers reviewed the Data mining techniques and found out that there are several algorithms and techniques being applied in agricultural domain particularly for yield prediction based on Rainfall dataset. Multiple Linear Regression (MLR) is the method used to model the linear relationship between a dependent variable and one or more independent variable(s)[1]. Researchers focused on Yield prediction based on Rainfall dataset, results comparison between MLR Technique and K-Means algorithm.

Jyotshna Solanki, Prof. (Dr.) Yusuf Mulge, observed the research studies on different data mining techniques in the field of agriculture. Data mining techniques are used in agriculture for prediction of problem, disease detection, optimizing the pesticide and so on [2]. Data mining techniques are used for disease detection, pattern recognition by using multiple applications. Data mining is close to determine the similarities between searching the precious business information from the massive information systems such as finding linked products in gigabytes of store scanner data or the mining a mountain for a vein of important dataset.

Georg Rub, Rudolf Kruse, Martin Schneider, and Peter Wagner, in their research study examine deal with wheat yield prediction. Neural networks are often for predicting wheat yield from cheaply-available in-season data. Once this prediction is possible, the industrial application is quite straightforward: use data mining with neural networks [3].

P.Revathi, Dr.M.Hemalatha, work described the remarkable of machine learning classifier in agriculture database and extracting knowledge [4]. Consistent prediction methods are consequently required to help planners and policy makers take strategic decisions to protect national interest.

Table 1: List of Related Works for crop yield prediction

.

AUTHOR & YEAR	TITLE	METHODOLOGY	PROBLEM STATEMENT
G.M.Nasira, N.Hemageetha , 2012	Forecasting Model for Vegetable Price Using Back Propagation Neural Network	Neural networks	Vegetable price forecasting model
Youvrajsinh Chauhan, Jignesh Vania , 2014	Disease Prediction on Soil Micronutrients Analysis of Cotton By J48 Classification	Classification	Soil Micronutrients Analysis Prediction
Utkarsha P. Narkhede, K. P. Adhiya, 2014	A Study of Clustering Techniques for Crop Prediction - A Survey	Clustering	Crop prediction
Farah Khan, Dr. Divakar Singh, 2014	Knowledge Discovery on Agricultural Dataset Using Association Rule Mining	Association rule mining	Crop productivity enhancement

III. PROPOSED APPROACH

In the proposed method, initially the raw data set was collected and it is subjected to preprocess for noise removing (replacement of missing values) and computational methods. From that dataset, it is subjected to Feature selection for make a predictive modeling. In this proposed approach it is mainly focused on Regression Techniques. Various regression analysis should be performed and it was compared and tested. Regression analysis is a form of predictive modeling technique which investigates the association between a dependent (target) and independent variable**(s)** (predictor). This technique is used for forecasting, time series modeling and discovers the causal effect relationship between the variables. Regression analysis indicates the significant relationships between dependent variable and independent variable and it indicates the strength of impact of multiple independent variables on a dependent variable. These techniques are typically motivated by 3 metrics. They are number of independent variables, type of dependent variables and regression line pattern shape.

IV. METHODS & MATERIALS

4.1 Linear Regression

Linear Regression is one of the commonly used well-known modeling techniques in data mining concept, in which the dependent variable is to be taken as continuous, in other independent variables will be continuous or discrete, and regression line is linear [14]. Here to find the relationship two variable, one is dependent variable (Y) and other one variable that is independent (X) with best fit straight line is commonly called as regression line. The regression equation is shown in below,

$$Y=a + b*X + e \quad \text{------ (1)}$$

In (1) where ‘e’ is error term, ‘b’ is line slope and ‘a’ is intercept. This (1) equation can be used to predict the value of target variable based on given predictor variable(s). Linear Regression is very perceptive toward Outliers. It can terribly affect the regression line and forecasted values.

4.2 Logistic Regression

Logistic regression technique is used to find the probability of event of Success and event of Failure. Logistic regression will be used when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can be represent by this equation (2), (3),

Odds = p / (1-p) = probability (p) of event occurrence/ probability (p) of no event occurrence

$$\ln(\text{odds}) = \ln(p/(1-p)) \text{ --------(2)}$$

$$\text{logit}(p)=\ln(p1/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk \text{ ---------- (3)}$$

Here in this equation (2) where 'p' is the probability of presence of the characteristic of interest and in (3) 'logit' function for an associate function that can be used for chosen best way of binomial distribution (dependent variable).

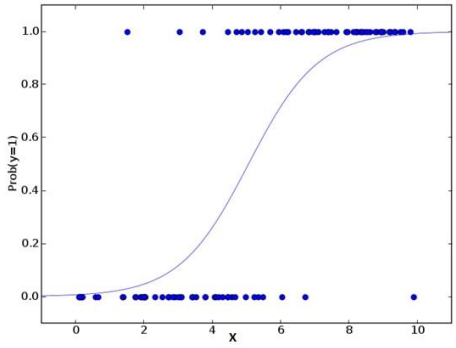

Fig. 1: Logistic Regression

4.3 Polynomial Regression

A regression of y on x may be a polynomial regression of y on x if the ability of independent variable power is greater than one. Such an equation is known as polynomial regression equation.

$$Y=a+b*X^2 \quad \text{---------- (4)}$$

In this polynomial regression technique, the best fit line is a curve line but not a straight line. That curves line fits into the data points.

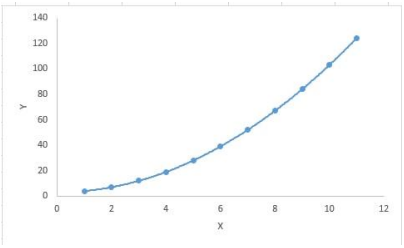

Fig.2: Polynomial Regression

4.4 Ridge regression

Ridge Regression could be a technique used once the data suffers from multi co-linearity (independent variables are extremely correlated). In multi co-linearity, even though the least squares estimate (OLS) are unbiased and variances are large, deviates the true value far by the observed value. In the regression estimates, by adding a degree of bias, ridge regression reduces the standard errors [14].

Equation of the linear regression is represented as

$$Y = a + b*X \quad \text{----- (5)}$$

Hence the equation included by an error term looks like:

$$Y=a + b*X + e \quad \text{------- (6)}$$

Where 'e' denotes an error term. Error term is the value required to correct for a prediction error between the observed and predicted value.

$$y = (a + y) = a + b1x11+ b2x22 +....+ e1 \quad \text{--- (7)}$$

for multiple independent variables equation (7) works out.

V. DATASET DESCRIPTION

The dataset was collected from website of agricultural government portal. It contains state wise, district wise, crop wise, season wise and year wise data on crop covered area and production of crop yield in India. In this dataset, the attributes of Area is measured in hectares and Crop production is measured in tones per hectare.

VI. RESULTS AND ANALYSIS

The result and analysis was done under the data refers to Tamilnadu state particularly focused on North Western zone of Tamilnadu. North western zone of tamilnadu state data refers to the district of Dharmapuri, Salem, Namakkal, Krishnagiri. It contains crop wise, season wise and year wise data on crop covered area and production. Area is measured in hectares and Crop production is measured in tones per hectare.

Linear regression predictor model for two attributes is given below

$$lr(\text{production in tons} \sim \text{Area in hectares} + \text{Crop, dataframe}) \quad \text{------(8)}$$

As this (8) it will predict the Production in Tons, when those values are assigned for Area in hectares and Crop.

Linear regression predictor model for three attributes is given below

lr(production in tons ~ Area in hectares + Crop+Year, dataframe) ------(9)

As this equation (9) it will predict the Production in Tons, when those values are assigned for Area in hectares, Crop and Year.

The summary values of area in hectares and production in tons are shown in below Table: 2.

Table: 2 Summary values of Production and Area

Summary Values	Min	1st Quart	Median	Mean	3rd Quart	Max
production	0	102	1230	944216	13466	646700000
Area	1	103	909	5957	5625	3675546

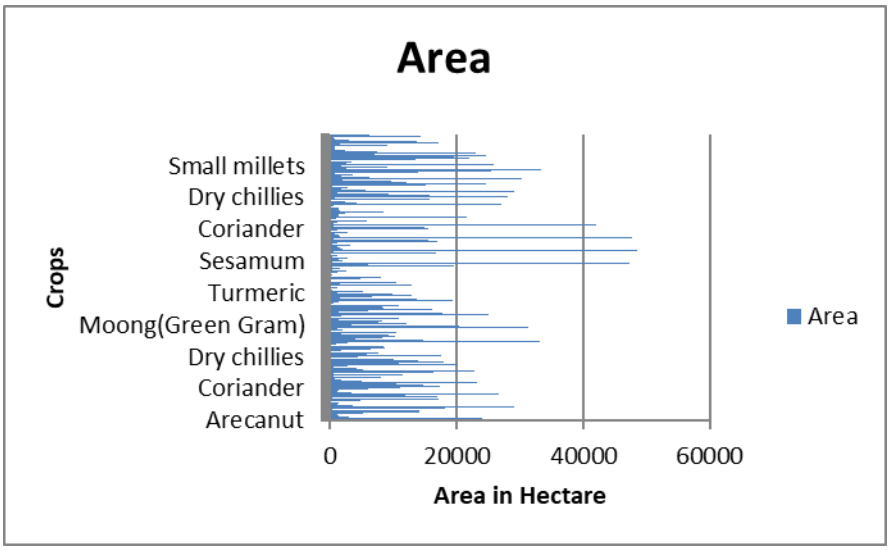

Fig: 3 Area in hectares vs crop

Fig 3 graph shows Area in hectares vs crop in which crops was taken along y-axis and area in hectare was taken along x-axis. In the north western zone, under the area having more cultivated crops are Tapiaco, Sugar cane, Ragi, Maize, Groundnut.

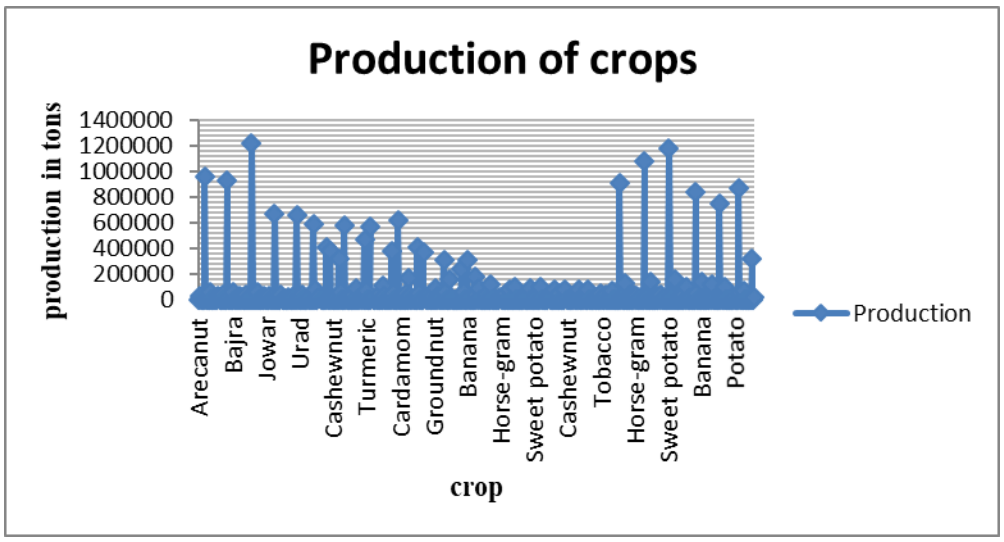

Fig: 4 productions in tons vs crop

Fig 4 graph shows the production in tons vs crop in which productions in tons was taken along y-axis and crop was taken along x-axis. Sugar cane and Tapioca has the highest crop yield production rate in North West zone of tamilnadu. Rice, Banana and Maize crops has the second highest yield production rate compared to the other crops. Ragi, Turmeric, Coconut, Cotton, Jowar are having next highest yield of production rate.

VII.CONCLUSION

This paper dealt with various regression techniques for agriculture crop yield prediction. Our proposed work mainly focused on to get predictor model by using regression techniques. Predictor formula is most useful in the crop prediction of Agriculture crop Production in Tons. Highest rate of yield production in tamilnadu state particularly in North Western zone is sugarcane and tapioca. Banana, Maize, Ragi, Turmeric, Coconut, Cotton, Jowar are having next highest yield of production rate compared to the other crops depends on the results of predictor model.

REFERENCES

[1] D Ramesh, B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013, ISSN (Print) : 2319-5940 ISSN (Online) : 2278-1021.

[2] Jyotshna Solanki, Prof. (Dr.) Yusuf Mulge, "Different Techniques Used in Data Mining in Agriculture", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, May 2015 ISSN: 2277 128X.

[3] Georg Rub, Rudolf Kruse, Martin Schneider, and Peter Wagner, "Data Mining with Neural Networks for Wheat Yield Prediction", Springer-Verlag Berlin Heidelberg 2008.

[4] P.Revathi, Dr.M.Hemalatha, "Categorize the Quality of Cotton Seeds Based on the Different Germination of the Cotton Using Machine Knowledge Approach", International Journal of Advanced Science and Technology Vol. 36, November, 2011.

[5] Yethiraj N G, "Applying Data mining techniques in the field of Agriculture and Allied Sciences", International Journal of Business Intelligent, Vol 01, Issue 02, December 2012, ISSN: 2278-2400.

[6] P.Surya, Dr. I.Laurence Aroquiaraj ,M.Ashok Kumar, "The Role Of Big Data Analytics In Agriculture Sector : A Survey" International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST) Vol. 2, Special Issue 10, March 2016, ISSN 2395-695X (Print) ISSN 2395-695X (Online).

[7] Raorane A.A, Kulkarni R.V., "Review- Role of Data Mining in Agriculture", Raorane A.A. et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 270 – 272.

[8] G.M.Nasira, N.Hemageetha, "Forecasting Model for Vegetable Price Using Back Propagation Neural Network", International Journal of Computational Intelligence and Informatics, Vol. 2: No. 2, July - September 2012.

[9] Farah Khan, Dr. Divakar Singh, "Knowledge Discovery on Agricultural Dataset Using Association Rule Mining", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 5, May 2014.

[10] Utkarsha P. Narkhede, K. P. Adhiya, "A Study of Clustering Techniques for Crop Prediction - A Survey", American International Journal of Research in Science, Technology, Engineering & Mathematics, 2014.

[11] Youvrajsinh Chauhan, Jignesh Vania, "Disease Prediction on Soil Micronutrients Analysis of Bt Cotton By J48 Classification", International Journal of Engineering Development and Research,Volume 2, Issue 2,ISSN: 2321-9939, 2014.

[12] S.Nagini, Dr. T.V. Rajini Kanth, B.V. Kiranmayee, "Agriculture Yield Prediction Using Predictive Analytic Techniques", IEEE, 2016.

[13] J. M. Hayes and W. L. Decker, "Using NOAA AVHRR Data to Estimate Maize Production in the United States Corn Belt," International Journal of Remote Sensing, Vol. 17, No. 16, 1996, pp. 3189-3200.doi:10.1080/01431169608949138

[14] https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression.