DATA MINING

Project On

Data Mining Techniques and Applications to Agricultural Data

Project by:

Venkata Narendra Babu Pratapaneni
Srikanth Javvaji
Richu Thakur

TABLE OF CONTENTS

FIGURE OF CONTENT

ABSTRACT

Data Mining is emerging field in Agricultural crop yield analysis and crop price prediction. There are various data mining techniques such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) which are used for very recent applications of Data Mining techniques. Yield prediction and crop price prediction are very important agricultural problems that remains to be solved based on the available data. These problems can be solved by employing Data Mining techniques. We will try to find suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities and also crop price prediction.

1. INTRODUCTION

1.1 Purpose of the project

Yield prediction is very important in agriculture as the price of the crop depends on the total yield of that crop in that period. Every farmer is interested in knowing, how much yield he is about expect. In the past, yield prediction was performed by considering farmer's previous experience on a particular crop. There is vast amount of data available in Indian agriculture. The data when becomes information is highly useful for many purposes. Data Mining can be used to analyze large data sets and establish useful classifications and patters in the data sets. The overall goal of the Data Mining process is to extract the information from a data set and transform it into understandable structure for further use.

1.2 Existing System

Agrarian sector face rigorous problem to maximize the crop productivity. More than 60 percent of the crop still depends on monsoon rainfall. In majority of the countries the farmers are not getting the expected crop yield due to several reasons. The agricultural yield is primarily depends on weather conditions. Rainfall conditions also influences the rice cultivation. Various other factors like climate, CO2 level, humidity, fertilizer, soil profile etc. affect the production. The farmers cannot predict the production based on the weather conditions of the year due to which they cannot take precautionary measures to improve the productivity. Due to which the yield is affected and the economy is also affected.

1.3 Proposed Idea

. In this context, the farmers necessarily require a timely advice to predict the future crop productivity and an analysis is to be made in order to help the farmers to maximize the crop production in their crops.

Recent development 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

2. RELATED WORKS

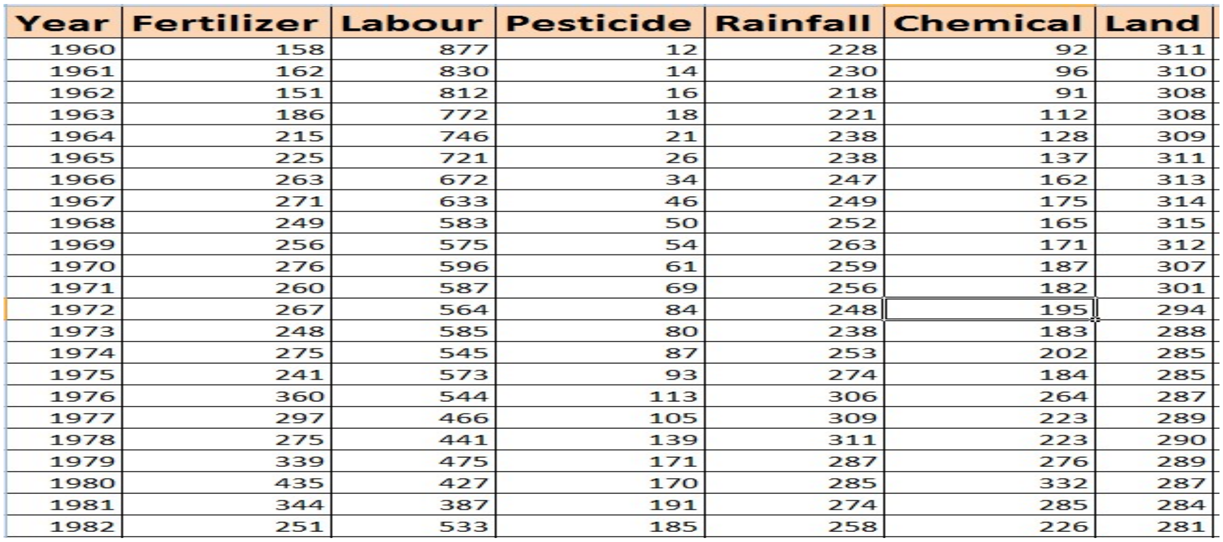
From studying research article [1] we identify that analysis of large amount of data which is stored for analysis can provide considerable chances of increasing efficiency and can have economic advantages.

The researchers have used [2] KMeans algorithm to forecast the pollution in the atmosphere, the K Nearest Neighbor was applied [3] for simulating daily precipitations and other weather variables and different possible changes of the weather scenarios are analyzed [4] using Support Vector Machines. Clustering techniques are found in grading [5] apples before marketing in agriculture. Weeds were detected [6] on precision agriculture.

The researchers worked [7] on rainfall variability analysis and its impact on crop productivity. The effect of observed seasonal climatic conditions such as rainfall and temperature variability on crop yield prediction was considered [8] through an empirical crop model. There are two approaches to investigate the impact of climate change on crop production which include the crop suitability approach and the production function approach [9]. Researchers were found that the yields of winter wheat are reduced when temperatures rise, due to the consequent reduction of the growth phases of the plant [10] and concluded that the complexity of a model was based on the level of detailed analysis [11] or it was less detailed with only estimations of moisture content [12].

3. OVERVIEW OF DATA

The data used for the project is obtained for the years from 1960 to 2005 for Missouri State. Each area in this collection is identified by the respective longitude and latitude of the region. The data are taken in input variables. The variables are 'Year', 'Rainfall', 'Fertilizers', ’Labor’, ‘Pesticide’, ‘Chemical’ and ’Land’. The attribute 'Year' specifies the year in which the data are available in Hectares. 'Rainfall' attribute specifies the average rainfall in the specified year in Centimeters. 'Land' attribute specifies the total area sowed in the specified year for that region in Hectares. 'Farm Output' attribute specifies the production of crop in the specified year in Metric Tons. 'Fertilizers' specify in Tons in the specified year.



Year	Fertilizer	Labour	Pesticide	Rainfall	Chemical	Land
1960	158	877	12	228	92	311
1961	162	830	14	230	96	310
1962	151	812	16	218	91	308
1963	186	772	18	221	112	308
1964	215	746	21	238	128	309
1965	225	721	26	238	137	311
1966	263	672	34	247	162	313
1967	271	633	46	249	175	314
1968	249	583	50	252	165	315
1969	256	575	54	263	171	312
1970	276	596	61	259	187	307
1971	260	587	69	256	182	301
1972	267	564	84	248	195	294
1973	248	585	80	238	183	288
1974	275	545	87	253	202	285
1975	241	573	93	274	184	285
1976	360	544	113	306	264	287
1977	297	466	105	309	223	289
1978	275	441	139	311	223	290
1979	339	475	171	287	276	289
1980	435	427	170	285	332	287
1981	344	387	191	274	285	284
1982	251	533	185	258	226	281

Fig 1: Input data

4. METHODOLOGY

The data mining method used for developing the project is Multiple Linear Regression technique. This technique has been used for prediction of crop yield analysis.

4.1 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. The dependent variable is sometimes termed as predictant and independent variables are called predictors.

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. This crop yield prediction model is presented with the use of Multiple Linear Regression (MLR) technique where the predictant is the Production and there are seven predictors namely are 'Year', 'Rainfall', 'Fertilizers’, ‘Labor’, ‘Pesticide’, ‘Chemical’ and ’Land’.

5. RESULTS AND DISCUSSION

In this project crop yield analysis is processed by implementing Multiple Linear Regression technique.

The exact value along with the corresponding estimated value using Multiple Linear Regression technique for 45 years’ interval of sample data about Missouri State is shown in the Table-1.

The estimated results using Multiple Linear Regression technique which are ranging between -13% and +19% for 45 years' interval.

Table-1: Exact production and estimated values using Multiple Linear Regression technique

Observation Year	Production(Exact)	45 years' interval	
		Production(Estimate)	Percentage difference
1995	16016	16163	-1
1996	18686	16967	17
1997	18948	16958	19
1998	18867	18011	8
1999	17701	17759	-0.6
2000	19180	18517	6
2001	19933	18193	17
2002	19020	18036	9
2003	18607	19948	-13
2004	21254	19566	16

The below chart shows actual and predicted production values from 1995-2004.

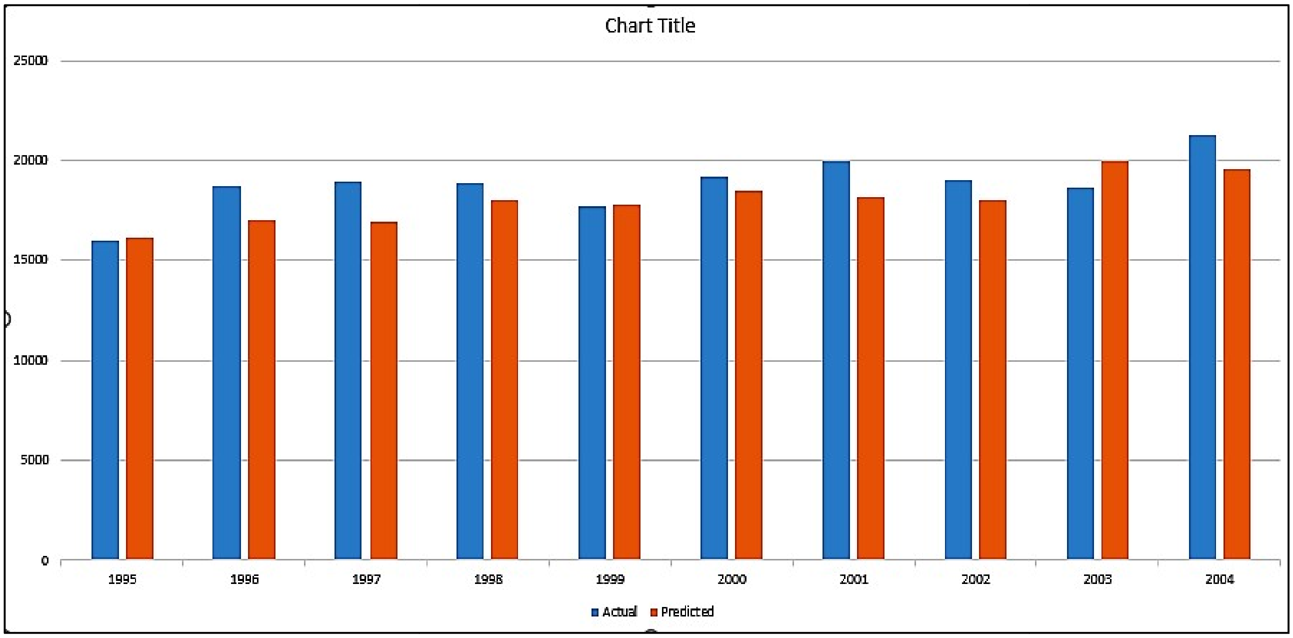


Fig 2 : Actual Vs Predicted production values

The below chart shows actual and predicted production over 45 years interval from 1960-2004.

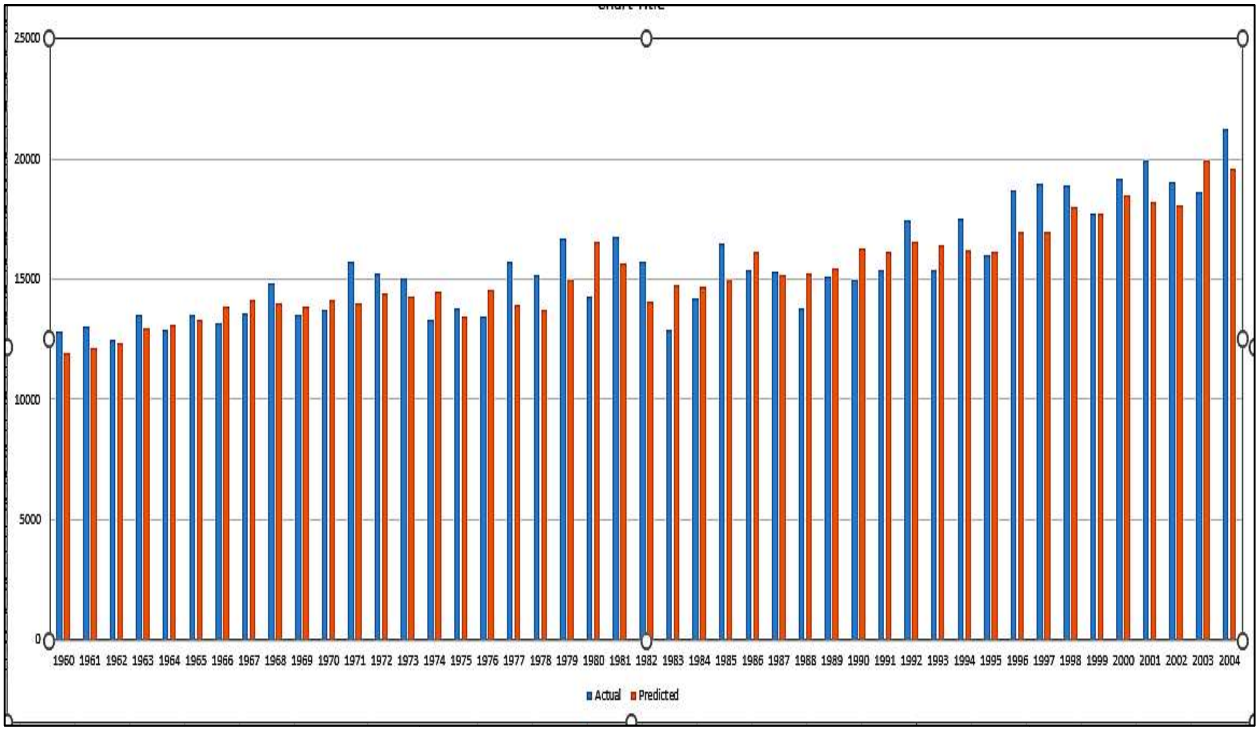


Fig 3: Actual Vs Predicted production over 45 years

6. CONCLUSION

Initially the statistical model Multiple Linear Regression technique is applied on existing data. In the subsequent work a comparison of the crop yield prediction can be made with the entire set of existing available data and will be dedicated to suitable approaches for improving the efficiency of the proposed technique. The predicted model is useful to predict the yield based on the forecasted conditions and sufficient measures can be taken timely to increase the productivity of the crop.

7. CONTRIBUTION REPORT

S.NO	ASSIGNMENT	NARENDRA	SRIKANTH	RICHU
1.	Report	Y	N	Y
2.	Presentation	N	Y	N
3.	Mathematical model	Y	N	Y
4.	Data collection	N	Y	N
5.	Coding	Y	Y	Y

OVERALL CONTRIBUTION

S.No.	Percentage Contributed			
1.	Narendra	Srikanth	Richu	**Total**
	33.33%	33.33%	33.33%	**100%**

8. REFERENCES

1. G Ruß, "Data Mining of Agricultural Yield Data : A Comparison of Regression Models", Conference Proceedings, Advances in Data Mining – Applications and Theoretical Aspects, P Perner (Ed.), Lecture Notes in Artificial Intelligence 6171, Berlin, Heidelberg, Springer, 2009, pages : 24-37.

2. Jorquera H, Perez R, Cipriano A, Acuna G, "Short Term Forecasting of Air Pollution Episodes", In: Zannetti P (eds) Environmental modeling , WIT Press, UK, 2001.

3. Rajagopalan B, Lall U, "A K-Nearest Neighbor Simulator for Daily Precipitation and Other Weather Variables", Wat Res Res 35(10), 1999, pages : 3089-3101.

4. Tripathi S, Srinivas V V, Nanjundiah R S, "Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach", J Hydrol, 2006, pages : 621-640.

5. Leemans V, M F Destain, "A Real Time Grading Method of Apples Based on Features Extracted from Defects", J. Jood Eng., 2004, pages : 83-89.

6. Tellaeche A, X P Burgos Artizzu, G Pajares, A Ribeiro, "A Vision-Based Classifier for Weeds Detection in Precision Agriculture through the Bayesian and Fuzzy K-Means Paradigms", Adv.Soft. Comp., 2008, pages : 72-79.

7. Mehta D R, Kalola A D, Saradava D A, Yusufzai A S, "Rainfall Variability Analysis and Its Impact on Crop Productivity - A Case Study", Indian Journal of Agricultural Research, Volume 36, Issue 1, 2002, pages : 29-33.

8. M Trnka, “Projections of Uncertainties in Climate Change Scenarios into Expected Winter Wheat Yields”, Theoretical and Applied Climatology, vol. 77, 2004, pages : 229-249.

9. M J Foulkes, “Raising Yield Potential of Wheat”, Journal of Experimental Botany, vol. 62, 2011, pages : 469-486.

10. G R Batts, “Effects Of CO2 And Temperature on Growth and Yield of Crops of Winter Wheat over Four Seasons”, European Journal of Agronomy, vol. 7, 1997, pages : 43-52

11. R J Brooks, "Simplifying Sirus : Sensitivity Analysis and Development of A Meta-Model for Wheat Yield Prediction", European Journal of Agronomy, vol. 14, 2001, pages : 43-60.

12. R V Martin, "Seasonal Maize Forecasting for South Africa and Zimbabwe Derived From an Agroclimatological Model", Journal of Applicable Meteorology, vol. 39, 2000, pages : 1473-1479.