Effectual Machine Learning for Indian Agriculture Seasonal Crop Using REPTREE Classification

Ms. Mythili
Head and Associate Professor,
Department of Information Technology/Computer Technology,
Hindusthan College of Arts and Science,
Coimbatore, India.

Abstract —Data mining plays a pivot role in predicting various aspects of the agriculture domain. Agriculture is the indispensable drape for the country growth. This paper proposes REPTREE method for the classification of the agriculture data. Especially the paper address the comprehensive evaluation of season wise production of various agro products and by-products. Many ongoing research focus on agriculture automation and effective analysis to perform effective and efficient agriculture. REPTREE classification is a fast decision tree learning method. This method performs the classification-regression tree using the information gain and variance. It further prunes the tree to get optimal solution. Results were promising as evident from the experiments.

Keywords— Agriculture, Decision tree, REPTREE, Pruning, Data Mining

I. INTRODUCTION

In today's conditions agricultural enterprises are capable of generating and collect large amounts of data. Growth in data size requires automated method to extract necessary data. By applying data mining technique it is possible to extract useful knowledge and trends. Knowledge gained in this manner, may be applied to increase work efficiency and improve decision making quality[1].

Information technology has become an integral part of our daily life. Techniques for managing data have become necessary and common in industry and services. Improvements in efficiency can be achieved in almost every aspect of business. This is especially true for agriculture, in order to modernize and better apply GPS technology. Agricultural companies in addition to reaping the fruits on the fields have started collecting large amounts of data. Large amount of information about soil and crop properties, which enables higher operational efficiency, is often contained in these data – in order to find this information it is necessary to apply adequate techniques[1].

II. LITERATURE REVIEW

Machine learning deals with the erection and study of systems that learns from data.

Mr. Narsi Reddy Gayam stated in his research learning —A study of crop yield distribution and crop insurance‖ which takes the input data from INDIA relating sugarcane and Soybean. He discovered that proposition of predictability of crop yields. The intensive data qualitatively analyzed by using Lilliefore method, here he considered unfounded hypothesis are normally distributed. The actual results indicate the considerations of the hypothesis in all cases are not true. Hence he concludes crop yield are not normally distributed [2]. The result found by Mr.NR Reddy is very much useful to estimate risk management implicated in sugarcane and soybean crops.

Dr. Bharat Misra, et al., [3] observed the research studies on application of data mining techniques in the field of agriculture. Some of the techniques, such as ID3 algorithms, the k-means, and the k-nearest neighbor, artificial neural networks and support vector machines applied in the field of agriculture were presented.

III. TECHNICAL PERSPETIVES OF WORKING METHODOLOGY

RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree. Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances. The example of REP Tree algorithm is applied on UCI repository and the confusion matrix is generated. [4] [5] [6]

Entropy measures the homogeneity of the samples and calculates the impurity of the arbitrary collection of samples. Given a collection S, combining positive and negative examples of same target concept, the entropy of S relative to this Boolean classification is as follows:

$$Entropy(S) = -p \oplus log_2\ p \oplus -p \ominus log_2 p \ominus$$

where $p \oplus$ the proportion of positive samples in S and $p \ominus$ is the proportion of negative samples in S.

```
Build_Tree (T, a_split) {
    Calculate IG(T,a) for each attribute a.
    Find the split point if the attribute is numeric type.
    Find the split attribute a_max with maximum IG among the
    attributes.
    If IG(T,a_max) > IG(T, a_split) {
    For all v∈ val(a_max) {
    T={x∈T|x_amax = v}
    Build_Tree(T, a_max)
    } } }
```

Table 1 . Pseudocode of REPTREE

Entropy is the measure of the impurity in the given collection of training samples whereas information gain is the measure of effectiveness of an attribute in classifying the training data. It is the expected reduction in entropy caused by partitioning the samples according to the attribute. Information gain Gain (S, A) is defined as

$$Gain\ (S,A) = Entropy(S) - \sum_{v\ \in\ values\ (A)} \frac{|S_v|}{|S|}\ Entropy\ (S_v)$$

where values(A) is the set of all possible values for attribute A. S_v is the subset of S for which attribute A has values V(S_v) = $\{s \in S | A(S) = v\}$.

Entropy (s) denotes entropy of original collection. The second term represents expected value of entropy after s is partitioned using the attribute A. The expected entropy in the second term is the sum of entropies of each subset S_v weighted by fraction $\frac{|S_v|}{|S|}$ that belongs to S_v. Gain (S,A) is therefore expected reduction in entropy caused by knowing value of attribute A.

The work uses the idea of cross validation from last saved tree. Data is divided into a training set and a testing set and cross-validation is applied to prune the tree. At each pair of leaf nodes with a common parent, evaluate the error on the testing data, and monitor whether the testing sum of squares would shrink if those two nodes are removed and made their parent a leaf. This is repeated until pruning no longer improves the error on the testing data. Pruning is superior to arbitrary stopping criteria because it directly checks whether the extra capacity (nodes in the tree) credits by improving generalization.

Considerations:

seed -- The seed used for randomizing the data.

minNum -- The minimum total weight of the instances in a leaf.

numFolds -- Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules.

numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model.

batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.

debug -- If set to true, classifier may output additional info to the console.

noPruning -- Whether pruning is performed.

spreadInitialCount -- Spread initial count across all values instead of using the count per value.

doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).

maxDepth -- The maximum tree depth (-1 for no restriction).

minVarianceProp -- The minimum proportion of the variance on all the data that needs to be present at a node in order for splitting to be performed in regression trees.

initialCount -- Initial class value count.

IV. Experimental Results

The problem is implemented using WEKA under 32 bit Vista operating system. Experiments are conducted on a laptop with Intel(R) CoreTM 2 Duo 2.00 GHz CPU, and 3 GB of RAM. The values of parameters of the proposed algorithm are selected based on some preliminary trials. The selected parameters gave the best results concerning both the solution quality and the computational time needed to reach this solution.

Attributes of the dataset is as follows:

@relation apy

@attribute State_Name {'Andaman and Nicobar Islands','Andhra Pradesh','Arunachal Pradesh',Assam,Bihar,Chandigarh,Chhattisgarh,'Dadra and Nagar Haveli',Goa,Gujarat,Haryana,'Himachal Pradesh','Jammu and Kashmir ',Jharkhand,Karnataka,Kerala,'Madhya

Pradesh',Maharashtra,Manipur,Meghalaya,Mizoram,Nagaland,Odisha,Puducherry,Punjab,Rajasthan,Sikkim,'Tamil Nadu','Telangana ',Tripura,'Uttar Pradesh',Uttarakhand,'West Bengal'}

@attribute District_Name {NICOBARS,'NORTH AND MIDDLE ANDAMAN','SOUTH ANDAMANS',ANANTAPUR,CHITTOOR,'EAST GODAVARI',GUNTUR,KADAPA,KRISHNA, UTTAR',HOOGHLY,HOWRAH,JALPAIGURI,MALDAH,'MEDINIPUR EAST','MEDINIPUR WEST',MURSHIDABAD,NADIA,PURULIA}

@attribute Crop_Year numeric

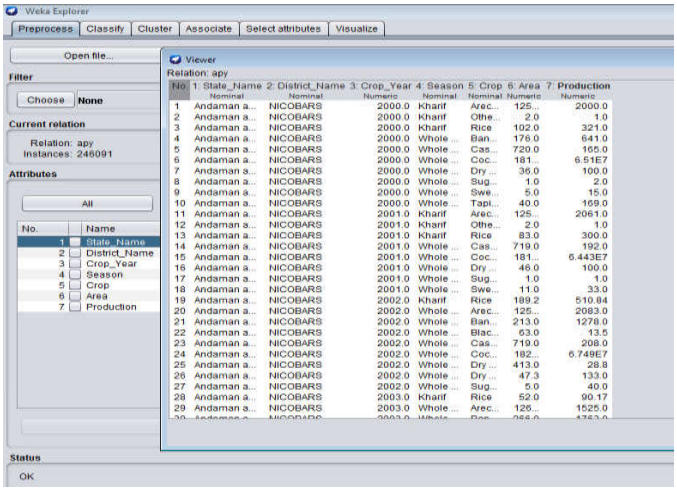

Figure 1 Indian Agriculture Seasonal Crop Prediction Preview

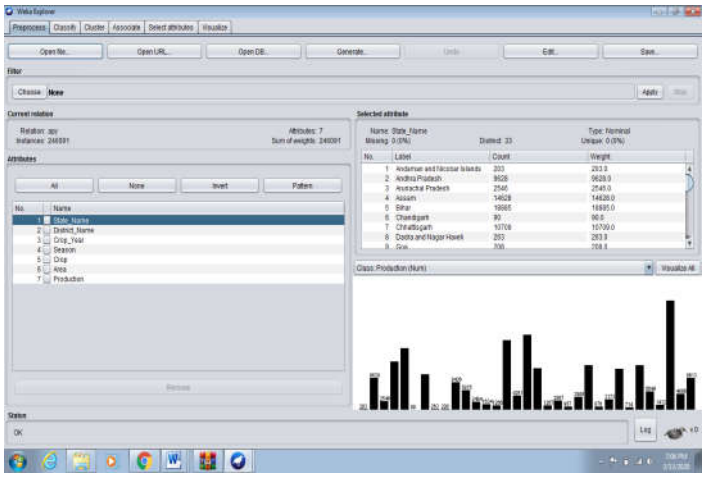

Figure 2 Data set statistical view

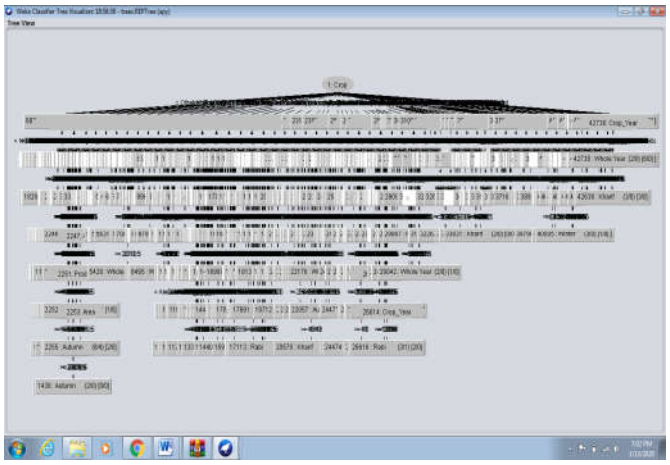

Figure 3 REPTREE generated for the dataset

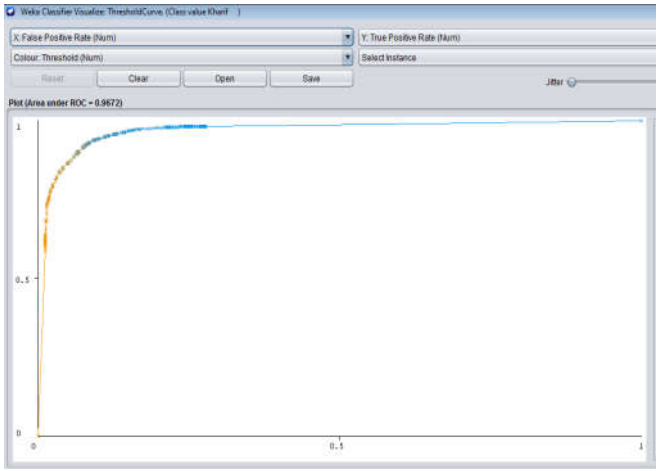

Figure 4 ROC of the KHARIF season

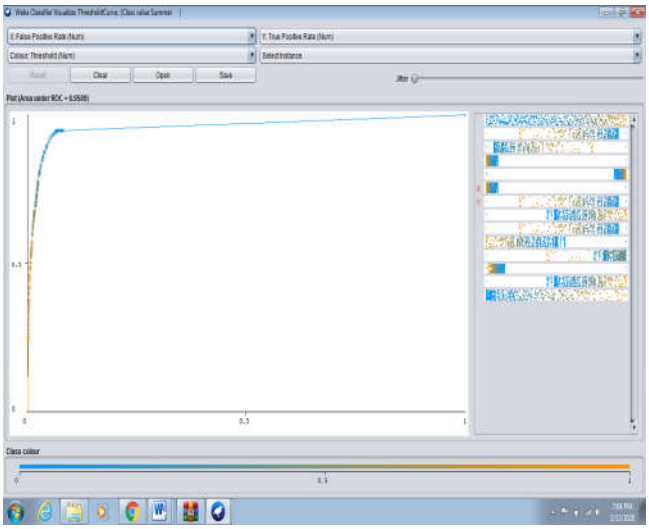

Figure 4 ROC of the SUMMER season

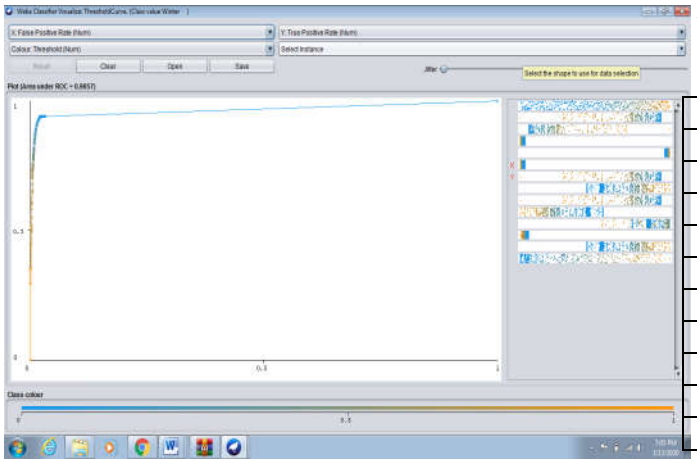

Figure 4 ROC of the WINTER season

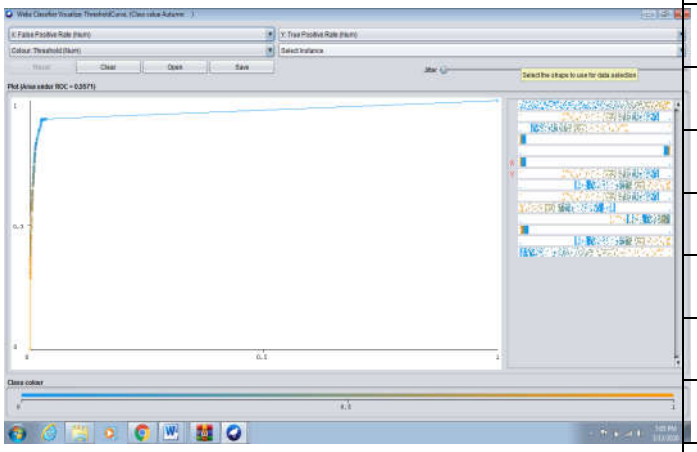

Figure 4 ROC of the AUTUMN season

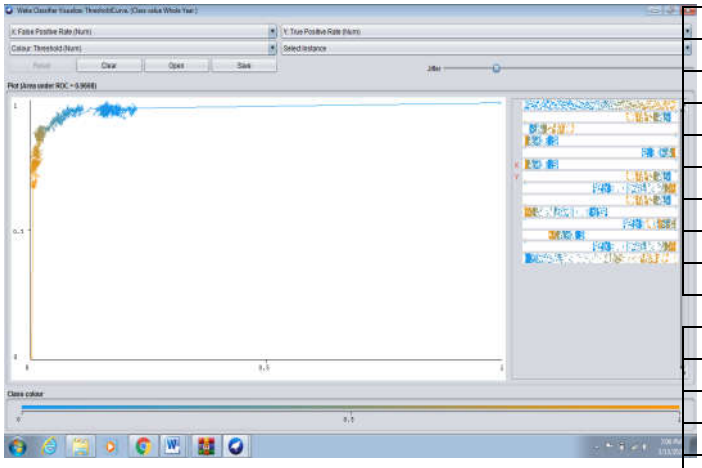

Figure 4 ROC of the WHOLE YEAR

V. FINDINGS AND DISCUSSION

The dataset used contains the seasonal production of agro direct and indirect products. The development environment that is used is WEKA. The dataset is preprocessed and subject to the REPTREE classifier. Results show that this proposed approach performs well.

=== Stratified cross-validation ===		
=== Summary ===		
Correctly Classified Instances	215263	87.4729 %
Incorrectly Classified Instances	30828	12.5271 %
Kappa statistic	0.8243	
Mean absolute error	0.0516	
Root mean squared error	0.1758	
Relative absolute error	21.6553 %	
Root relative squared error	50.9126 %	
Total Number of Instances	246091	

=== Detailed Accuracy By Class ===						
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.917	0.078	0.882	0.917	0.899	0.833	0.967	0.943	Kharif
0.877	0.024	0.917	0.877	0.897	0.866	0.970	0.933	Whole Year
0.686	0.007	0.657	0.686	0.671	0.665	0.957	0.722	Autumn
0.878	0.039	0.895	0.878	0.886	0.844	0.970	0.928	Rabi
0.690	0.018	0.707	0.690	0.698	0.679	0.959	0.748	Summer
0.765	0.007	0.743	0.765	0.754	0.748	0.966	0.778	Winter
Weighted Avg.	0.875	0.048	0.875	0.875	0.875	0.829	0.968	0.916	

=== Confusion Matrix ===						
a	b	c	d	e	f	<-- classified as
87950	2518	481	3345	1247	410	a = Kharif
4832	50242	3	1740	285	203	b = Whole Year
321	84	3394	83	571	496	c = Autumn
4656	1564	97	58801	1711	158	d = Rabi
1730	228	643	1661	10241	338	e = Summer
207	142	544	100	430	4635	f = Winter

| State_Name = Tamil Nadu : Whole Year (35/0) [10/0]
| State_Name = Telangana : Whole Year (0/0) [0/0]
| State_Name = Tripura : Whole Year (0/0) [0/0]
| State_Name = Uttar Pradesh : Whole Year (0/0) [0/0]
| State_Name = Uttarakhand : Whole Year (0/0) [0/0]
| State_Name = West Bengal : Whole Year (0/0) [0/0]
Crop = Sapota : Kharif (27/0) [12/0]
Crop = Cabbage
| Crop_Year < 2007.5 : Whole Year (96/0) [48/0]
| Crop_Year >= 2007.5
| | District_Name = NICOBARS : Rabi (0/0) [0/0]

		District_Name = NORTH AND MIDDLE ANDAMAN : Rabi (0/0) [0/0]
		District_Name = SOUTH ANDAMANS : Rabi (0/0) [0/0]
		District_Name = ANANTAPUR : Kharif (1/1) [3/1]
		District_Name = CHITTOOR : Kharif (3/3) [3/0]
		District_Name = EAST GODAVARI
		Area < 22.5 : Kharif (2/0) [1/0]
		Area >= 22.5 : Rabi (3/0) [0/0]
		District_Name = GUNTUR
		Area < 38.5 : Rabi (3/0) [0/0]
		Area >= 38.5 : Kharif (2/0) [1/0]
		District_Name = KADAPA : Rabi (0/0) [1/0]
		District_Name = KRISHNA : Kharif (4/3) [2/0]
		District_Name = KURNOOL : Kharif (5/2) [1/1]
		District_Name = PRAKASAM : Rabi (3/1) [2/1]
		District_Name = SPSR NELLORE : Rabi (0/0) [0/0]
		District_Name = SRIKAKULAM : Kharif (4/2) [2/1]
		District_Name = VISAKHAPATANAM
		Area < 115.5 : Rabi (2/0) [1/0]
		Area >= 115.5 : Kharif (2/0) [1/0]
		District_Name = VIZIANAGARAM
		Area < 94 : Rabi (3/1) [1/0]
		Area >= 94 : Kharif (2/0) [0/0]
		District_Name = WEST GODAVARI : Rabi (2/1) [1/0]

VI. CONCLUSION

Data mining is inevitable in all domains in our day to day life. Agriculture is the indispensable drape for the country growth. This paper proposes REPTREE method for the classification of the agriculture data. Especially the paper address the comprehensive evaluation of season wise production of various agro products and by-products. Many ongoing research focus on agriculture automation and effective analysis to perform effective and efficient agriculture. REPTREE classification is a fast decision tree learning method. This method performs the classification-regression tree using the information gain and variance. It further prunes the tree to get optimal solution. Results were promising as evident from the experiments.

REFERENCES

[1] B. MiloviC1 and V. RadojeviC2,Application of Data Mining in Agriculture Bulgarian Journal of Agricultural Science, 21 (No 1) 2015, 26-34 Agricultural Academy

[2] Rainfall variability analysis and its impact on crop productivity Indian agriculture research journal 2002 29,33.,8) SPRS Archives XXXVI-8/W48 Workshop proceedings: Remote sensing support to crop yield forecast and area estimates GENERALIZED SOFTWARE TOOLS FOR CROP AREA ESTIMATES AND YIELD FORECAST by Roberto Benedetti A, Remo Catenaro A, Federica Piersimoni B

[3] S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh, —Data mining Techniques for Predicting Crop Productivity – A review articlel, International Journal of Computer Science and technology, march 2011.

[4] Shivnath Ghosh, Santanu Koley, ”Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks”. International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 2, 292 297 ISSN: 2321-8169

[5] Kumar, Rakesh, M.p. Singh, Prabhat Kumar, and J.p. Singh. "Crop Selection Method to Maximize Crop Yield Rate Using Machine Learning Technique." 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM) (2015). Web

[6] Gutierrez D. D. (2015). Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R. Basking Ridge, NJ: Technics Publications.

[7] Forrest, S.; Perelson, A.S.; Allen, L.; Cherukuri, R. (1994). "Self-nonself discrimination in a computer" (PDF). Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Los Alamitos, CA. pp. 202–212.

[8] Jump up^ Timmis, J.; Neal, M.; Hunt, J. (2000). "An artificial immune system for data analysis". BioSystems. 55 (1): 143–150.doi:10.1016/S0303-2647(99)00092-1. PMID 10745118.

[9] Jump up^ Greensmith, J.; Aickelin, U. (2009). "Artificial Dendritic Cells: Multi-faceted Perspectives" (PDF). Human-Centric Information Processing Through Granular Modelling: 375–395.

Appendix – A
Sample Data

@relation apy

@attribute State_Name {'Andaman and Nicobar Islands','Andhra Pradesh','Arunachal Pradesh',Assam,Bihar,Chandigarh,Chhattisgarh,'Dadra and Nagar Haveli',Goa,Gujarat,Haryana,'Himachal Pradesh','Jammu and Kashmir ',Jharkhand,Karnataka,Kerala,'Madhya Pradesh',Maharashtra,Manipur,Meghalaya,Mizoram,Nagaland,Odisha,Puducherry,Punjab,Rajasthan,Sikkim,'Tamil Nadu','Telangana ',Tripura,'Uttar Pradesh',Uttarakhand,'West Bengal'}

@attribute District_Name {NICOBARS,'NORTH AND MIDDLE ANDAMAN','SOUTH ANDAMANS',ANANTAPUR,CHITTOOR,'EAST GODAVARI',GUNTUR,KADAPA,KRISHNA, UTTAR',HOOGHLY,HOWRAH,JALPAIGURI,MALDAH,'MEDINIPUR EAST','MEDINIPUR WEST',MURSHIDABAD,NADIA,PURULIA}

@attribute Crop_Year numeric

@attribute Season {'Kharif ','Whole Year ','Autumn ','Rabi ','Summer ','Winter '}

@attribute Crop {Arecanut,'Other Kharif pulses',Rice,Banana,Cashewnut,'Coconut ','Dry ginger',Sugarcane,'Sweet potato',Tapioca,'Black pepper','Dry chillies','other oilseeds',Turmeric,Maize,'Moong(Green Gram)',Urad,Arhar/Tur,Groundnut,Sunflower,Bajra,'Castor seed',Cotton(lint),Horse-gram,Jowar,Korra,Ragi,Tobacco,Gram,Wheat,Masoor,Sesamum,Linseed,Safflower,Onion,'other misc. pulses',Samai,'Small millets',Coriander,Potato,'Other Rabi pulses',Soyabean,'Beans & Mutter(Vegetable)',Bhindi,Brinjal,'Citrus Fruit',Cucumber,Grapes,Mango,Orange,'other fibres','Other Fresh Fruits','Other Vegetables',Papaya,'Pome Fruit',Tomato,'Rapeseed &Mustard',Mesta,Cowpea(Lobia),Lemon,'Pome

Granet',Sapota,Cabbage,'Peas (vegetable)','Niger seed','Bottle Gourd',Sannhamp,Varagu,Garlic,Ginger,'Oilseeds total','Pulses total',Jute,'Peas & beans (Pulses)',Blackgram,Paddy,Pineapple,Barley,Khesari,'Guar seed',Moth,'Other Cereals & Millets','Cond-spcs other',Turnip,Carrot,Redish,'Arcanut (Processed)','Atcanut (Raw)','Cashewnut Processed','Cashewnut Raw',Cardamom,Rubber,'Bitter Gourd','Drum Stick','Jack Fruit','Snak Guard','Pump Kin',Tea,Coffee,Cauliflower,'Other Citrus Fruit','Water Melon','Total foodgrain',Kapas,Colocosia,Lentil,Bean,Jobster,Perilla,'Rajmash Kholar','Ricebean (nagadal)','Ash Gourd','Beet Root',Lab-Lab,'Ribed Guard',Yam,Apple,Peach,Pear,Plums,Litchi,Ber,'Other Dry Fruit','Jute & mesta'}

@attribute Area numeric

@attribute Production numeric

@data

'Andaman and Nicobar Islands',NICOBARS,2000,'Kharif ',Arecanut,1254,2000

'Andaman and Nicobar Islands',NICOBARS,2000,'Kharif ','Other Kharif pulses',2,1

'Andaman and Nicobar Islands',NICOBARS,2000,'Kharif ',Rice,102,321

'Andaman and Nicobar Islands',NICOBARS,2000,'Whole Year ',Banana,176,641

'Andaman and Nicobar Islands',NICOBARS,2000,'Whole Year ',Cashewnut,720,165

'Andaman and Nicobar Islands',NICOBARS,2000,'Whole Year ','Coconut ',18168,65100000

'Andaman and Nicobar Islands',NICOBARS,2000,'Whole Year ','Dry ginger',36,100

Sikkim,'WEST DISTRICT',2015,'Kharif ',Rice,3016,5851

Sikkim,'WEST DISTRICT',2015,'Kharif ','Small millets',711,744

Sikkim,'WEST DISTRICT',2015,'Kharif ',Soyabean,546,520

Sikkim,'WEST DISTRICT',2015,'Kharif ',Urad,914,811

Sikkim,'WEST DISTRICT',2015,'Rabi ',Barley,12,11

Sikkim,'WEST DISTRICT',2015,'Rabi ','Rapeseed &Mustard',625,540

Sikkim,'WEST DISTRICT',2015,'Rabi ',Wheat,20,21

'Tamil Nadu',ARIYALUR,2008,'Kharif ',Rice,24574,?

'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Arhar/Tur,209,?

'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Bajra,565,?

'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Banana,190,?

'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Cashewnut,31113,?

'Tamil Nadu',ARIYALUR,2008,'Whole Year ','Castor seed',27,?

'Tamil Nadu',ARIYALUR,2008,'Whole Year ','Coconut ',335,?

'Tamil Nadu',ARIYALUR,2008,'Whole Year ',Coriander,460,?

'Tamil Nadu',COIMBATORE,2013,'Rabi ','Rapeseed &Mustard',3,1

'Tamil Nadu',COIMBATORE,2013,'Rabi ',Urad,153,131

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Arecanut,1817,3686

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Arhar/Tur,544,526

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Banana,7412,324506

'Tamil Nadu',COIMBATORE,2013,'Whole Year ','Black pepper',128,26

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Cardamom,808,63

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Cashewnut,103,29

'Tamil Nadu',COIMBATORE,2013,'Whole Year ','Coconut ',84531,1212000000

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Coriander,138,48

'Tamil Nadu',COIMBATORE,2013,'Whole Year ','Dry chillies',481,131

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Gram,1162,811

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Sugarcane,1170,121181

'Tamil Nadu',COIMBATORE,2013,'Whole Year ','Sweet potato',2,42

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Tapioca,340,10174

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Tobacco,100,159

'Tamil Nadu',COIMBATORE,2013,'Whole Year ',Turmeric,1203,6472