AGRICULTURAL CROP YIELD PREDICTION USING DATA MINING

A report submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Department of Computer Science and Engineering

by

D. Adarsh Srivatsa (CS14B1010)
Mohammed Imran (CS14B1020)
Sumit Kumar Singh (CS14B1029)

DEPARTMENT OF
COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY PUDUCHERRY
KARAIKAL – 609609
NOVEMBER 2017

BONAFIDE CERTIFICATE

This is to certify that the project work entitled "**Agricultural Crop Yield Prediction Using Data Mining**" is a bonafide record of the work done by

D. Adarsh Srivatsa (CS14B1010)
Mohammed Imran (CS14B1020)
Sumit Kumar Singh (CS14B1029)

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** of the **NATIONAL INSTITUTE OF TECHNOLOGY PUDUCHERRY** during the year 2017 - 2018.

Dr.Narendran Rajagopalan
Assistant Professor
Project Guide

Dr. B.Surendiran
Head of the Department
Computer Science and Engineering Dept.

Project viva-voce held on _____

Internal Examiner

External Examiner

ABSTRACT

Food is basic source of energy for any living being. We majorly depend on agricultural produces as our main source of our food resource. The current problem faced by most of the farmers is failure of crops due to various climatic factors. So the need of the hour is to give some solution to the farmers based on scientific prediction of crop growth based on climatic conditions so that they can plan their cultivation accordingly to produce maximum yield.

This project is about predicting the crop growth pattern using the data available from the past years based on the climatic factors like temperature and rainfall using Data Mining techniques and simple Machine Learning Algorithms. The output is proposed to be in the form of visual representations like graphs which can be easier to comprehend.

ACKNOWLEDGEMENT

We would like to show our kind regards towards our respected Director, **Dr. Yog Raj Sood** for permitting us to undertake this project work.

We would like to thank our project guide, Prof. **Dr.Narendran Rajagopalan**, for his constant motivation and guidance during the project. We want to genuinely convey our thanks to **Dr.B.Surendiran,** Head of the Department and all the faculties of our Computer Science and Engineering department for their motivation in various reviews throughout the course of the project phase-I.

We would like to thank the project review members for their valuable suggestion throughout the period of project.

We are at the dearth of words to express gratitude to our wonderful parents for their unconditional support both financially and emotionally. I thank our parents for inculcating the dedication and discipline to do whatever we undertake well.

We have been fortunate to have friends who cherish us despite our eccentricities. By their remarks comments or compliments and unavoidable questions, we were able to make our project reviews better each time. Thank you all for making it possible for us to reach the final stage of our endeavor.

We would also like to thank all our sources, mentioned in the references, and our friends who helped us by providing mental and logistical support. Last but not the least we would like to thank God Almighty**.**

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 GENERAL INTRODUCTION

Agriculture is the major source of food for people. Yield prediction is very important in agriculture as the price of the crop depends on the total yield of that crop in that period of time. Every farmer is interested in knowing, how much yield he is about expect from his agricultural field. In the past, yield prediction was performed by considering farmer's previous experience on a particular crop. This prediction did not have any scientific methodologies and did not include the probabilities of climatic factors affecting the yield. 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 OBJECTIVES

To predict the crop growth pattern using the data available from the past years based on the climatic factors like temperature, pressure and rainfall using Data Mining techniques and simple Machine Learning Algorithms. The output is proposed to be in the form of visual representations like graphs.

1.3 MOTIVATION

Food is basic source of energy for any living being. We majorly depend on agricultural produces as our main source of our food resource. The current problem faced by most of the farmers is failure of crops due to various climatic factors. So the need of the hour is to give some solution to the farmers based on scientific prediction of crop growth based on climatic conditions so that they can plan their cultivation accordingly to produce maximum yield.

1.4 ORGANIZATION OF THESIS

Chapter 1 consists of Introduction to the research work along with problem statement, objectives, motivation, and organization of the thesis.
Chapter 2 of this document summarizes a literature review in the field of Data Mining.
Chapter 3 presents the hardware and software requirements.
Chapter 4 presents the implementation details of the project.
Chapter 5 summarizes the thesis work and concludes along with the future direction of work.

CHAPTER 2

LITERATURE REVIEW

2.1 DATA MINING

Data mining is the *process* of discovering interesting patterns and knowledge from *large* amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data mining as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**. The knowledge discovery process is an iterative sequence of the following steps:

1. **Data cleaning:**
   to remove noise and inconsistent data
2. **Data integration:**
   where multiple data sources may be combined
3. **Data selection:**
   where data relevant to the analysis task are retrieved from the database
4. **Data transformation:**
   where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations
5. **Data mining:**
   an essential process where intelligent methods are applied to extract data patterns
6. **Pattern evaluation:**
   to identify the truly interesting patterns representing knowledge
7. **Knowledge presentation:**
   where visualization and knowledge representation techniques are used to present mined knowledge to users

2.2 LITERATURE SURVEY

Satya Priya and Ryosuke SHIBASAKI, Center for Spatial Information Science, University of Tokyo proposed National Spatial Crop Yield Simulation Using Gis-Based Crop Production Model (2001). This paper evaluated spatial variability of crop production often due to different soil conditions, weather conditions and agricultural practices within a target-region. Also, the study demonstrated model applicability in evaluating an impact of climate changes over major cereal crops productivity at national level taking spatial variability into account.

Kuljit Kaur and Kanwalpreet Singh Attwal, Punjab University, Patiala, India proposed Effect of Temperature and Rainfall in Paddy Yield using Data Mining. This paper used Data Mining with Apriori Algorithm and WEKA Tool for analysis of daily temperature

and rainfall on paddy yield to predict the paddy yield and to analyze the effect of temperature and rainfall on the paddy yield.

D.R. Mehta, AD. Kalola, D.A Saradava and AS. Yusufzai, Main Dry Farming Research Station, GAU, Targhadia, India proposed Rainfall Variability Analysis And Its Impact On Crop Productivity- A Case Study. This paper proposed that the distribution of rainfall within the crop period is more important than the total amount of rainfall in a season. The results of prediction models revealed that various regression curves were found as best fit for different crops with seasonal rainfall.

CHAPTER 3

SYSTEM REQUIREMENT SPECIFICATION

3.1 HARDWARE REQUIREMENTS

This project requires a computer with good processing speed and enough RAM to work with the datasets.

3.2 SOFTWARE REQUIREMENTS

The project requires the following softwares

1. Python installed
2. Python libraries like
    i. Numpy
    ii. Scipy
    iii. Matplotlib
    iv. Pandas etc
3. MySql

CHAPTER 4

IMPLEMENTATION

4.1 AREA UNDER STUDY

For this project some of the districts of TamilNadu has been chosen some of which include Nagapattinam, Thanjavur, Dharmapuri etc. and the crop productions for these places have been obtained from the Government website for the past 15 years. The Annual Rainfall and Pressure in these regions have been obtained using Python code to scrape it from the online server.

The crops which are on study are Rice, Cotton, Coconut, Sugarcane which are the ones predominantly available in the the areas which we have selected for study.

4.2 PROCESS INVOLVED

The various process involved in mining the required data are as follows :

1. The crop production data was obtained from the website of TamilNadu. It contained the production of all the crops in all the districts as a single file as in Fig. 4.1, Fig.4.2.

3236	Tamil Nadu	KANCHIPURAM	Sunflower	9	5
3237	Tamil Nadu	KANCHIPURAM	Urad	575	684
3238	Tamil Nadu	KANCHIPURAM	Bajra	60	30
3239	Tamil Nadu	KANCHIPURAM	Groundnut	38030	8982
3240	Tamil Nadu	KANCHIPURAM	Horse-gram	7	11
3241	Tamil Nadu	KANCHIPURAM	Jowar	52	24
3242	Tamil Nadu	KANCHIPURAM	Maize	20	4
3243	Tamil Nadu	KANCHIPURAM	Moong(Green Gram)	195	242
3244	Tamil Nadu	KANCHIPURAM	Ragi	1166	294
3245	Tamil Nadu	KANCHIPURAM	Sunflower	2	2
3246	Tamil Nadu	KANCHIPURAM	Urad	840	984
3247	Tamil Nadu	KANCHIPURAM	Arhar/Tur	133	138
3248	Tamil Nadu	KANCHIPURAM	Banana	9262	223
3249	Tamil Nadu	KANCHIPURAM	Cashewnut	54	191
3250	Tamil Nadu	KANCHIPURAM	Coconut	33900000	3097
3251	Tamil Nadu	KANCHIPURAM	Dry chillies	35	130
3252	Tamil Nadu	KANCHIPURAM	Gram	15	27
3253	Tamil Nadu	KANCHIPURAM	Small millets	1	2
3254	Tamil Nadu	KANCHIPURAM	Sugarcane	179286	1731
3255	Tamil Nadu	KANCHIPURAM	Sweet potato	106	5
3256	Tamil Nadu	KANCHIPURAM	Tapioca	3980	133
3257	Tamil Nadu	KANNIYAKUMARI	Banana	120433	5282
3258	Tamil Nadu	KANNIYAKUMARI	Small millets	22060	26779
3259	Tamil Nadu	KANNIYAKUMARI	Banana	124600	4597
3260	Tamil Nadu	KANNIYAKUMARI	Black pepper	10	73

Figure 4.1 The screenshot of Crop Production data 1

5115	Tamil Nadu	NAGAPATTINAM	Cotton(lint)	9850	2496
5116	Tamil Nadu	NAGAPATTINAM	Groundnut	5818	1812
5117	Tamil Nadu	NAGAPATTINAM	Jowar	30	14
5118	Tamil Nadu	NAGAPATTINAM	Maize	15	3
5119	Tamil Nadu	NAGAPATTINAM	Moong(Green Gram)	34551	39811
5120	Tamil Nadu	NAGAPATTINAM	Urad	40958	46929
5121	Tamil Nadu	NAGAPATTINAM	Arecanut	1	1
5122	Tamil Nadu	NAGAPATTINAM	Banana	23882	575
5123	Tamil Nadu	NAGAPATTINAM	Cashewnut	661	1693
5124	Tamil Nadu	NAGAPATTINAM	Coconut	44600000	3854
5125	Tamil Nadu	NAGAPATTINAM	Dry chillies	6	23
5126	Tamil Nadu	NAGAPATTINAM	Small millets		1
5127	Tamil Nadu	NAGAPATTINAM	Sugarcane	278902	2962
5128	Tamil Nadu	NAGAPATTINAM	Tapioca	2873	96
5129	Tamil Nadu	NAGAPATTINAM	Tobacco	237	128
5130	Tamil Nadu	NAMAKKAL	Banana	27016	1239
5131	Tamil Nadu	NAMAKKAL	Horse-gram	1630	3292
5132	Tamil Nadu	NAMAKKAL	Onion	49562	6851
5133	Tamil Nadu	NAMAKKAL	Sesamum	1190	1599
5134	Tamil Nadu	NAMAKKAL	Small millets	1540	2041
5135	Tamil Nadu	NAMAKKAL	Arhar/Tur	6030	9370
5136	Tamil Nadu	NAMAKKAL	Bajra	2240	1512
5137	Tamil Nadu	NAMAKKAL	Banana	42260	1280

Figure 4.2 The screenshot of Crop Production data 2

2. This data was later uploaded in a MySQL database and the refined data consisting of the crops in study was obtained. The query used was

   Select * from Crops where crops =”rice” or crops =”coconut” or crops=”sugarcane” or crops =”cotton” ;

   The file now contained only the crops that were under the study which include rice, coconut, cotton, sugarcane.

   The data obtained after performing the above was as shown in Fig.4.3, Fig.4.4

Tamil Nadu	DHARMAPURI	2010	Kharif	Cotton(lint)	7444	19187
Tamil Nadu	DHARMAPURI	2010	Kharif	Rice	22862	93207
Tamil Nadu	DHARMAPURI	2010	Whole Year	Sugarcane	18351	1611784
Tamil Nadu	DHARMAPURI	2011	Kharif	Cotton(lint)	8738	28121
Tamil Nadu	DHARMAPURI	2011	Kharif	Rice	27067	128226
Tamil Nadu	DHARMAPURI	2011	Rabi	Cotton(lint)	2290	6301
Tamil Nadu	DHARMAPURI	2011	Whole Year	Coconut	7836	1.24E+08
Tamil Nadu	DHARMAPURI	2011	Whole Year	Sugarcane	20305	2285338
Tamil Nadu	DHARMAPURI	2012	Kharif	Cotton(lint)	13662	41593
Tamil Nadu	DHARMAPURI	2012	Kharif	Rice	12003	53669
Tamil Nadu	DHARMAPURI	2012	Whole Year	Sugarcane	16102	1299947
Tamil Nadu	DHARMAPURI	2013	Kharif	Cotton(lint)	11462	33815
Tamil Nadu	DHARMAPURI	2013	Kharif	Rice	26689	132535
Tamil Nadu	DHARMAPURI	2013	Rabi	Cotton(lint)	1024	2641
Tamil Nadu	DHARMAPURI	2013	Whole Year	Coconut	5472	35400000
Tamil Nadu	DHARMAPURI	2013	Whole Year	Sugarcane	8433	615686
Tamil Nadu	DINDIGUL	1997	Whole Year	Cotton(lint)	7844	24770
Tamil Nadu	DINDIGUL	1997	Whole Year	Rice	27589	136520
Tamil Nadu	DINDIGUL	1997	Whole Year	Sugarcane	6826	5749000
Tamil Nadu	DINDIGUL	1998	Kharif	Cotton(lint)	12558	31723
Tamil Nadu	DINDIGUL	1998	Kharif	Rice	27278	93305
Tamil Nadu	DINDIGUL	1998	Whole Year	Sugarcane	7051	886010
Tamil Nadu	DINDIGUL	1999	Kharif	Cotton(lint)	7287	18306

Figure 4.3 The screenshot of Crop Production data after processing

Tamil Nadu	NAGAPATTINAM	2009	Whole Year	Sugarcane	3958	288713
Tamil Nadu	NAGAPATTINAM	2010	Kharif	Cotton(lint)	261	763
Tamil Nadu	NAGAPATTINAM	2010	Kharif	Rice	156049	321506
Tamil Nadu	NAGAPATTINAM	2010	Whole Year	Sugarcane	3046	289076
Tamil Nadu	NAGAPATTINAM	2011	Kharif	Cotton(lint)	1	3
Tamil Nadu	NAGAPATTINAM	2011	Kharif	Rice	170042	577026
Tamil Nadu	NAGAPATTINAM	2011	Rabi	Cotton(lint)	316	917
Tamil Nadu	NAGAPATTINAM	2011	Whole Year	Coconut	4026	55400000
Tamil Nadu	NAGAPATTINAM	2011	Whole Year	Sugarcane	3079	266324
Tamil Nadu	NAGAPATTINAM	2012	Kharif	Cotton(lint)	675	1694
Tamil Nadu	NAGAPATTINAM	2012	Kharif	Rice	142210	201553
Tamil Nadu	NAGAPATTINAM	2012	Whole Year	Sugarcane	3421	319635
Tamil Nadu	NAGAPATTINAM	2013	Kharif	Rice	154750	605329
Tamil Nadu	NAGAPATTINAM	2013	Rabi	Cotton(lint)	2496	9850
Tamil Nadu	NAGAPATTINAM	2013	Whole Year	Coconut	3854	44600000
Tamil Nadu	NAGAPATTINAM	2013	Whole Year	Sugarcane	2962	278902
Tamil Nadu	NAMAKKAL	1997	Whole Year	Cotton(lint)	3429	7730
Tamil Nadu	NAMAKKAL	1997	Whole Year	Rice	23983	99150
Tamil Nadu	NAMAKKAL	1997	Whole Year	Sugarcane	9415	13608000
Tamil Nadu	NAMAKKAL	1998	Kharif	Cotton(lint)	2549	6418
Tamil Nadu	NAMAKKAL	1998	Kharif	Rice	24167	83225
Tamil Nadu	NAMAKKAL	1998	Whole Year	Sugarcane	9411	1505330
Tamil Nadu	NAMAKKAL	1999	Kharif	Cotton(lint)	3468	7390

Figure 4.4 The screenshot of Crop Production data after processing

3. Then the data for Rainfall and Pressure was obtained by scraping it from the website for the past 15 years using a Python code.
   The data obtained is as shown in Fig.4.5 to Fig.4.8

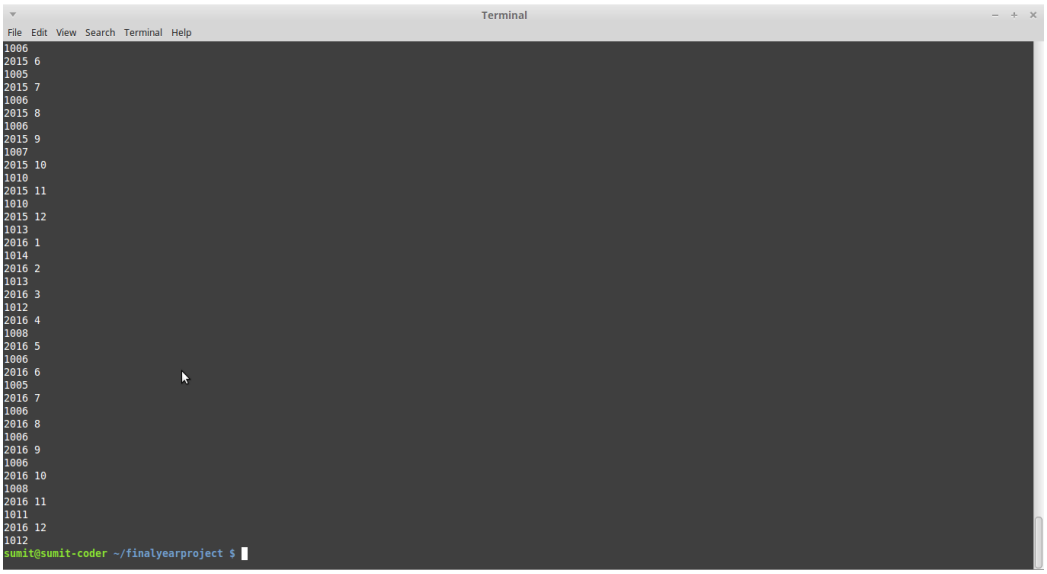


Figure 4.5 The screenshot of Pressure data scraping

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
2	1010	1010	1010	1010	1010	1010	1010	1010	1010	1014	1012	1013	1014	1014	1014	1014
3	1010	1010	1010	1010	1010	1010	1010	1010	1010	1013	1011	1011	1012	1012	1013	1013
4	1010	1010	1010	1010	1010	1010	1010	1010	1010	1011	1010	1010	1011	1012	1012	1011
5	1010	1010	1010	1010	1010	1010	1010	1010	1010	1009	1009	1008	1008	1009	1010	1008
6	1010	1010	1010	1010	1010	1010	1010	1010	1010	1006	1007	1006	1005	1007	1007	1006
7	1010	1010	1010	1010	1010	1010	1010	1010	1010	1006	1007	1006	1005	1005	1006	1006
8	1010	1010	1010	1010	1010	1010	1010	1010	1010	1007	1006	1005	1006	1007	1007	1006
9	1010	1010	1010	1010	1010	1010	1010	1010	1010	1006	1007	1007	1007	1007	1007	1006
10	1010	1010	1010	1010	1010	1010	1010	1010	1008	1008	1008	1008	1007	1008	1008	1007
11	1010	1010	1010	1010	1010	1010	1010	1010	1010	1008	1009	1010	1009	1010	1011	1008
12	1010	1010	1010	1010	1010	1010	1010	1010	1011	1010	1011	1011	1011	1011	1010	1011
13	1010	1010	1010	1010	1010	1010	1010	1010	1013	1010	1011	1012	1012	1012	1013	1011
14																
15																
16																
17																
18																
19																
20																

nagapattinam | pudukkottai | ramanathapuram | sivagang ...

Figure 4.6 The screenshot of Pressure data after scraping

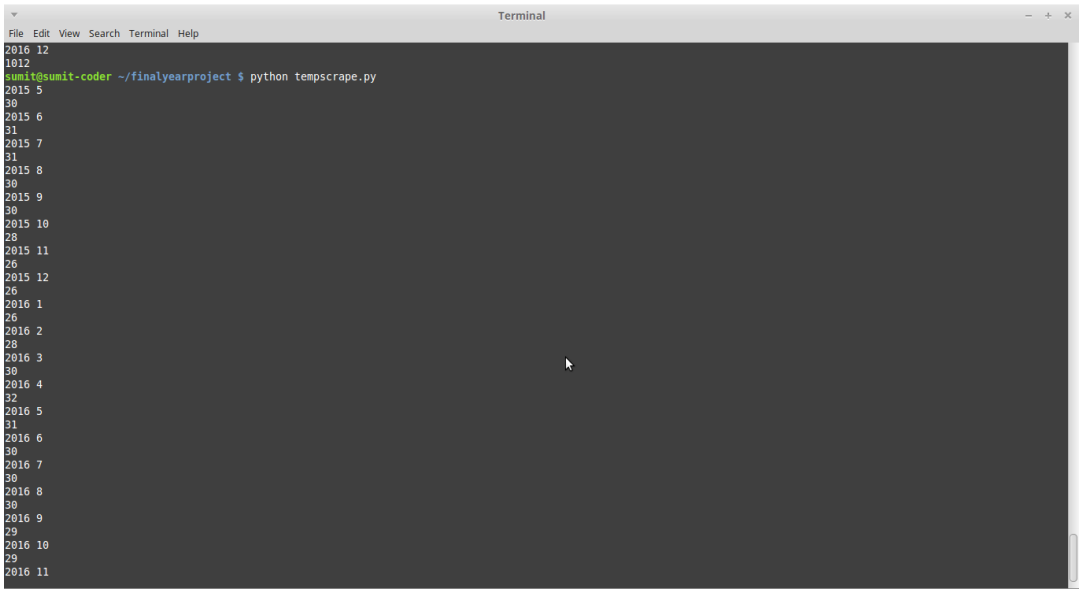
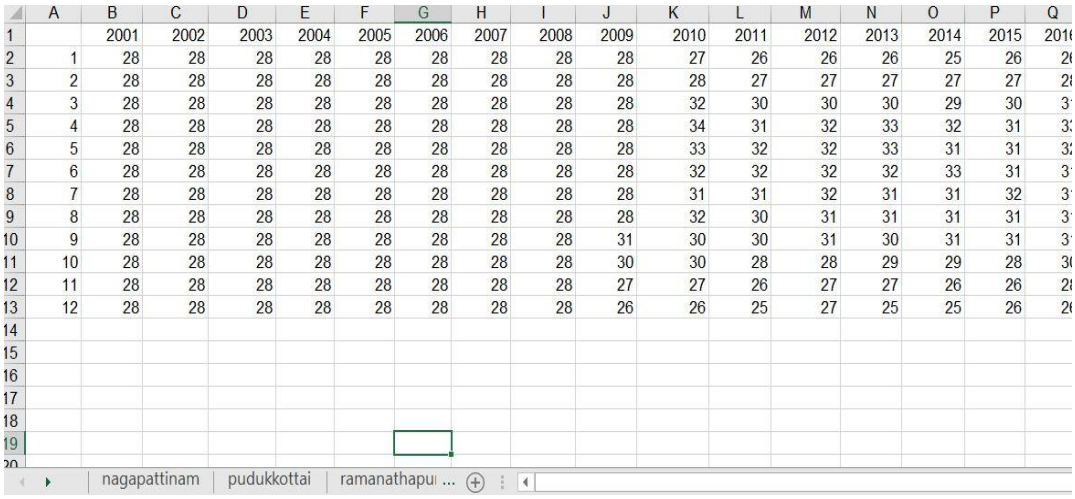


Figure 4.7 The screenshot of Temperature data scraping



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	201
2	1	28	28	28	28	28	28	28	28	28	27	26	26	26	25	26	2
3	2	28	28	28	28	28	28	28	28	28	28	27	27	27	27	27	2
4	3	28	28	28	28	28	28	28	28	28	32	30	30	30	29	30	3
5	4	28	28	28	28	28	28	28	28	28	34	31	32	33	32	31	3
6	5	28	28	28	28	28	28	28	28	28	33	32	32	33	31	31	3
7	6	28	28	28	28	28	28	28	28	28	32	32	32	32	33	31	3
8	7	28	28	28	28	28	28	28	28	28	31	31	32	31	31	32	3
9	8	28	28	28	28	28	28	28	28	28	32	30	31	31	31	31	3
10	9	28	28	28	28	28	28	28	28	31	30	30	31	30	31	31	3
11	10	28	28	28	28	28	28	28	28	30	30	28	28	29	29	28	3
12	11	28	28	28	28	28	28	28	28	27	27	26	27	27	26	26	2
13	12	28	28	28	28	28	28	28	28	26	26	25	27	25	25	26	2
14																	
15																	
16																	
17																	
18																	
19																	
20																	

nagapattinam | pudukkottai | ramanathapu ...

Figure 4.8 The screenshot of Temperature data after scraping

4. The next step includes the analysis of the data obtained from mining.

CHAPTER 5

CONCLUSION

5.1 RESULT

The data required for the project have been successfully mined from various sources and the primary processes of Data Mining which includes Data Cleaning, Data Integration, Data Selection and Data Transformation have been done successfully and is ready to be utilized for the further procedures of analyzing the data.

5.2 FUTURE WORKS

The future works in this project include Analysis of Data obtained so far. Multiple Linear Regression Algorithm is planned to be used for the analysis part.

REFERENCES

1. S. Priya, R. Shibasaki, " National Spatial Crop Yield Simulation Using Gis-Based Crop Production Model" , Ecological Modelling, vol.36, pp.113-129 January 2001.

2. 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.

3. 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.

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