PREDICTION OF VARIOUS CROP YIELDS IN ANDHRA PRADESH USING MULTIPLE LINEAR REGRESSION

[1]Dr. J. Rajendra Prasad, [2]S. Sai Kumar, [3]G.Venu Gopal,
[4]Sk.Jameela Shahin, [5]P.Sri Varshita, [6]V.Devi Priya, [7]V.Hemanth Sai
[1]Professor, [2,3]Assistant Professors, [4,5,6,7] IV B.Tech Students
Department of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada, AP, India.

Abstract: Data mining tools allow enterprises to predict future trends. Agriculture prediction is the toughest task for agricultural departments and farmers across the globe. Agriculture growth depends on different parameters, namely Water, Nitrogen, Weather, Soil characteristics, Crop rotation, Soil moisture, Surface temperature and Rain water etc. This application gives a brief analysis of crop yield prediction. In this application, we predicted production of some crops like Paddy, Maize, Groundnut, Sugarcane, and Greengram for all the districts of Andhra Pradesh using Multiple linear regression (MLR).

Index Terms **– Multiple Linear Regression (MLR), RStudio, Classification, Clustering, Prediction, Regression.**

I. INTRODUCTION

Generally, agricultural department stores the information about the different factors that leads to crop yield. It mainly focuses on production increment and productivity of agricultural crops using environment friendly science and technology. While ensuring increased net farm income to the farmers through various schemes, programmes and welfare measures. However, it does not investigate the root cause for the decrease in crop yield/agriculture risk. They store the information manually. This information is not used for future analysis. So, further prediction from previous data is required [1] [2] [3].

According to estimates of economic survey of A.P , 2016-17 , the food grains production was 156.85 lakh tones , and all seeds production was 24.62 lakh tones ,an increase from the previous year of 9.09% and 12.9% respectively. In the food grains segment –paddy, bajra , maize ,ragi and pulses production increased a little, where as jowar , other millets and wheat fell. In the oil seed segment , the production of groundnut and castor saw a drop while sesamum production improved.

During 2016-17,A.P had a 28% of rainfall deficit than normal .Further in the state, the total irrigated area to grow area is 50.38%. Apparently,the cro[ps during kharif season have sufficient water while farming in rabi season has suffered due to rainfall deficit, and lower level of water in reservoirs since Dec1 ,2016.

The consequences of 14.03% growth in the agriculture and allied sectors have to reflect on the upward economic mobility of the approximate 62% of the state's population dependent on these sectors.

II. PROPOSED SYSTEM

We are predicting the production of various crop yields such as Paddy, Maize, Groundnut, Sugercane and Greengram of 13 districts of Andhra Pradesh state. We have collected the information of various crop yields during the years 2016-17, 2017-18 and 2018-19 from Andhra Pradesh Agricultural Department. We predict the production of these cropping yields for the next year i.e.2019-20. We analyze the data by using R-Programming Language tool by applying Multiple Linear Regression (MLR) technique. Using this technique, we observed that we are getting results very fast, accurate and reliable. The result will be presented in a graphical format for understanding purpose. By using this result, the agricultural department will make a decision for future crop production.

III. DATA COLLECTION

A. Classification: - Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics [4].

B. Clustering: - Clustering is a data mining technique that makes a meaningful or useful cluster of objects that have similar characteristic using the automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes [5].

C. Prediction: - The prediction as its name implied is one of a data mining techniques that discover the relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in the sale to predict profit for the future if we consider a sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction [6].

D. Regression: - Regression is a data mining technique used to predict a range of numeric values (also called *continuous values*), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.

IV. LINEAR REGRESSION:

A familiar Statistical Data Analysis technique is Linear Regression. It was the earliest type of regression analysis to be studied meticulously and to be used expansively in realistic applications. This is because the models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters. It is used to modeling the linear relationship between a dependent variable and one or more independent variables [4] [5] [6].

Simple linear regression
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

Logistic regression
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Multiple Regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term. Multiple regression requires two or more predictor variables, and this is why it is called multiple regression.
Top of Form
Bottom of Form
The multiple regression equation explained above takes the following form:
$y = b1x1 + b2x2 + \ldots + bnxn + c.$
Here, bi's (i=1, 2…n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes.

V. RSTUDIO

RStudio is a new open-source integrated development environment (IDE) which is available for free. R is also a programming language for statistical computing and graphics. It has interesting features for both new and experienced R developers including code completion execute from source searchable history and support for authoring Sweave documents. The version of RStudio available today is a beta (v0.92) and is released under the GNU AGPL license. RStudio was founded by JJ Allaire. He was the creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio.

RStudio is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset. RStudio contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. RStudio is extensible and is being used for solving real-world data mining problems. RStudio is easy to use and to be applied at several different levels. R is open source software issued under the GNU General Public License. R is a GNU package. The source code for the R software environment is written primarily in C, R and FORTRAN. Pre-compiled binary versions of R are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available. RStudio is available in two editions:

1. **RStudio Desktop**
2. **RStudio Server**

1. RStudio Desktop: The program is run locally as a regular desktop application in operating systems such as Windows, macOS, and Linux.

2. RStudio Server: Allows accessing RStudio using a web browser while it is running on a remote Linux server.

RStudio provides an interface to the most common version control operations including managing changelists, diffing files, committing, and viewing history. While these features cover the basic everyday use of Git and Subversion, you may also occasionally need to use the system shell to access all of their principal functionality.

RStudio includes functionality to make it more clear-cut to use the shell with projects under version control. This includes:

- On all platforms, you can use the Terminal to open a new system shell with the working directory already initialized to your project's root directory.
- On Windows when using Git, the Shell command will open Git Bash, which is a port of the bash shell to Windows specially configured for use with Msys Git (this behavior may be disabled by using the standard Windows command prompt instead using Options -> Version Control).
- On Windows when using Subversion, RStudio opens a shell with a PATH configured to use a version of ssh.exe which ships with RStudio (required for svn+ssh connections).
- When running over the web, RStudio provides a web-based shell dialog.

Overall, R Studio is a free and open source Integrated Development Environment (IDE), for R, a programming language for statistical computing and Graphics. This is written in C++ programming language and uses the QT framework for its Graphical User Interface (GUI).

Main features of R Studio include:

Syntax highlighting, code completion and smart indentation

- Quickly jump to function definitions
- Extensive package development tools
- Easily manage multiple working directories using projects
- Interactive debugger to diagnose and fix errors quickly

R Studio download and installation:

- Download R studio from https://www.rstudio.com/products/rstudio/download/
- After download completes, open the RStudio and create file by using File->new->R script

VI . SOURCE CODE

```
        data<- read.csv("C:\\bar charts\\PADDY18-19.csv")
data
str(data)
install.packages("GGally")
library(GGally)
fit2 <- lm(PROD.IN.TONES ~AREA.IN.HA + ANNUAL.RAINFALL , data = data)
summary(fit2)
M<-c("Srikakulam","Vizianagaram","Visakhapatnam",
    "East Godavari","West Godavari",
    "Krishna","Guntur",
    "Prakasham","Nellore",
    "Chittoor","Kadapa",
    "Ananthapuram","Kurnool")
table<-c(data$AREA.IN.HA + data$ANNUAL.RAINFALL)
barplot(table,names.arg=M,xlab =" AREA.IN.HA + ANNUAL.RAINFALL",ylab = "PROD.IN.TONES",col="blue",main="Plot",border="blue", cex.names = 0.65,
      cex.lab   = 0.65,las=2)
predict(fit2,data.frame(AREA.IN.HA=73,ANNUAL.RAINFALL=678.9))
```

VII. RESULTS

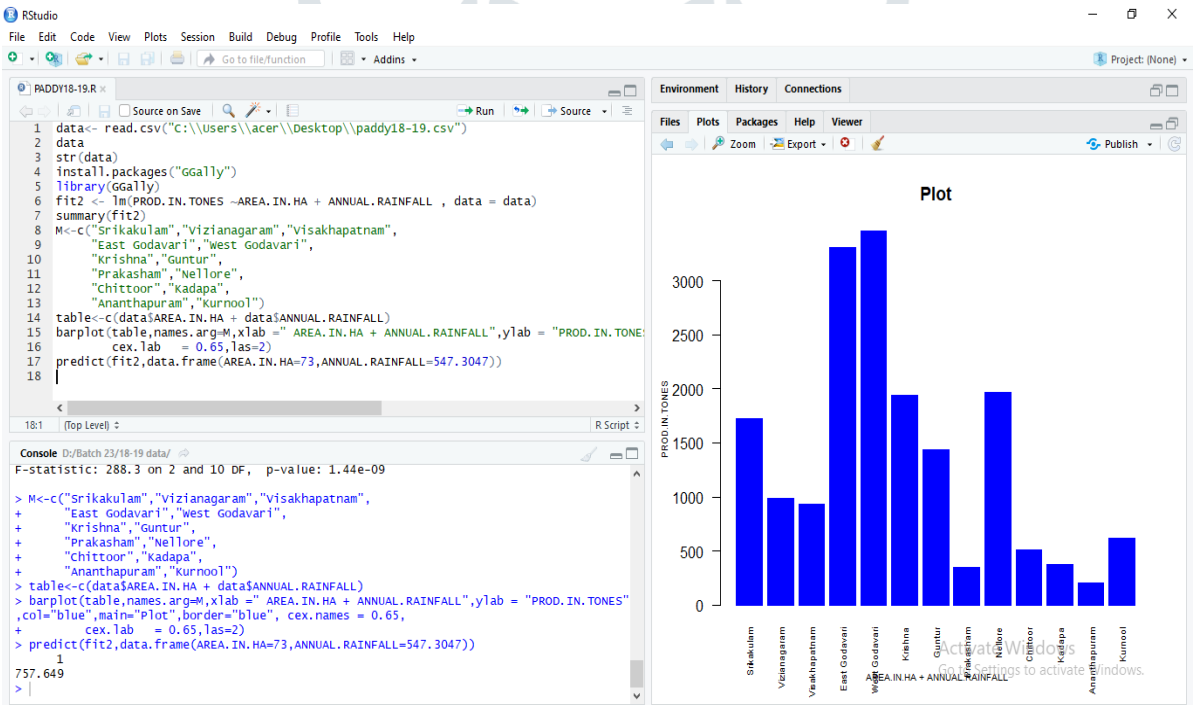

Fig. 1: Output Screen

Paddy

After analyzing the paddy crop yield production of the years 2016-17,2017-18 and 2018-19 of all 13 districts in the Andhra Pradesh and predicting the next year (2019-20) crop yield production of the paddy crop. The following table represents the prediction values of the paddy crop in Area, Production in Tones and Annual Rainfall.

DISTRICT	AREA.IN.HA	PROD.IN.TONES	ANNUAL.RAINFALL
Srikakulam	200	804.5137	1522.208
Vizianagaram	115	767.1586	869.7138
Visakhapatnam	109	764.5217	823.6554
East Godavari	383	884.9372	2926.991
West Godavari	400	892.4082	3057.49
Krishna	225	815.5005	1714.119
Guntur	167	790.0111	1268.887
Prakasham	42	735.0771	309.3361
Nellore	255	816.4002	1714.119
Chittoor	60	759.6439	447.5114
Kadapa	45	753.9708	332.3653
Ananthapuram	25	746.4067	178.8372
Kurnool	73	764.5605	547.3047

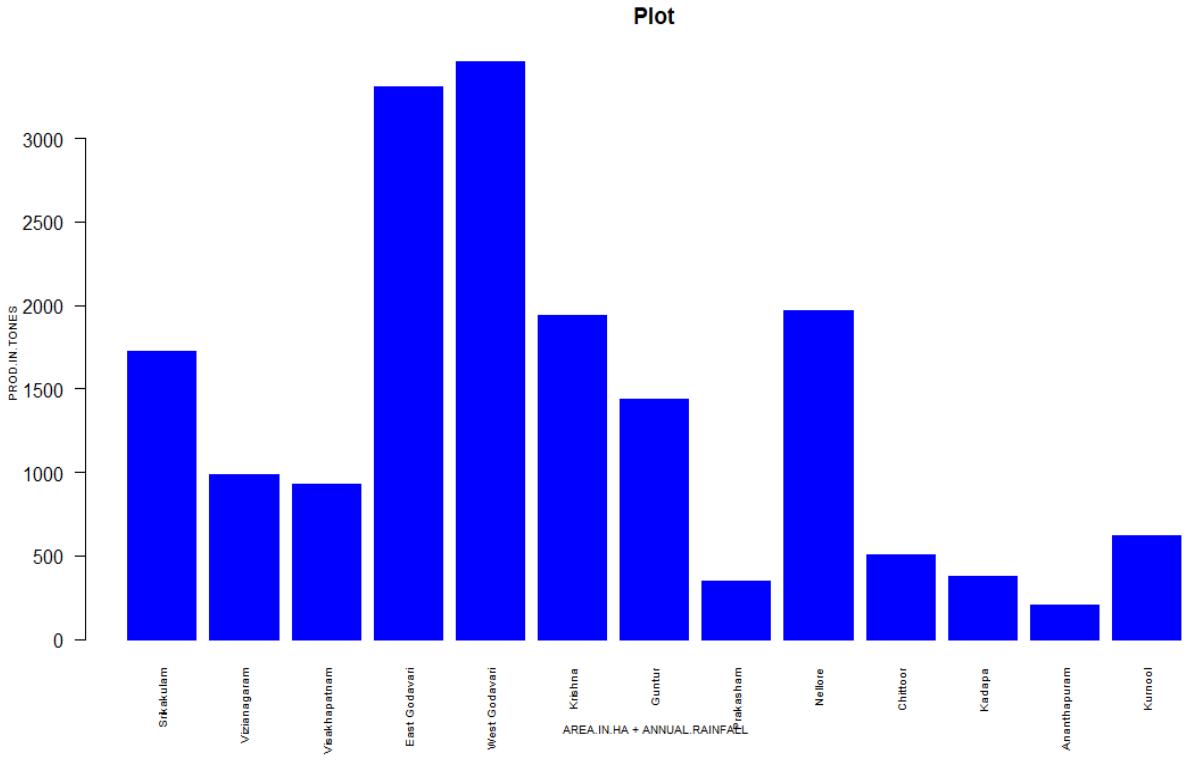

Fig. 2: Graph showing Estimated Prediction of Paddy crop for the year 2019-20

The Production of Paddy Crop for the year 2019-20 was highest in West Godavari district and lowest in Ananthapuram district.

Maize

After analyzing the Maize crop yield production of the years 2016-17,2017-18 and 2018-19 of all 13 districts in the Andhra Pradesh and predicting the next year (2019-20) crop yield production of the Maize crop. The following table represents the prediction values of the Maize crop in Area, Production in Tones and Annual Rainfall.

DISTRICT	AREA.IN.HA	PROD.IN.TONES	ANNUAL.RAINFALL
Srikakulam	22	1188	1522.208
Vizianagaram	35	977.9271	869.7138
Visakhapatnam	7	1016.017	823.6554
East Godavari	11	1609.185	2926.991
West Godavari	58	1560.432	3057.49
Krishna	26	1235.469	1714.119
Guntur	40	1082.738	1268.887
Prakasham	14	856.3729	309.3361
Nellore	1	1281.219	1714.119
Chittoor	2	917.7809	447.5114
Kadapa	3	883.0775	332.3653
Ananthapuram	22	804.4765	178.8372
Kurnool	35	885.8815	547.3047

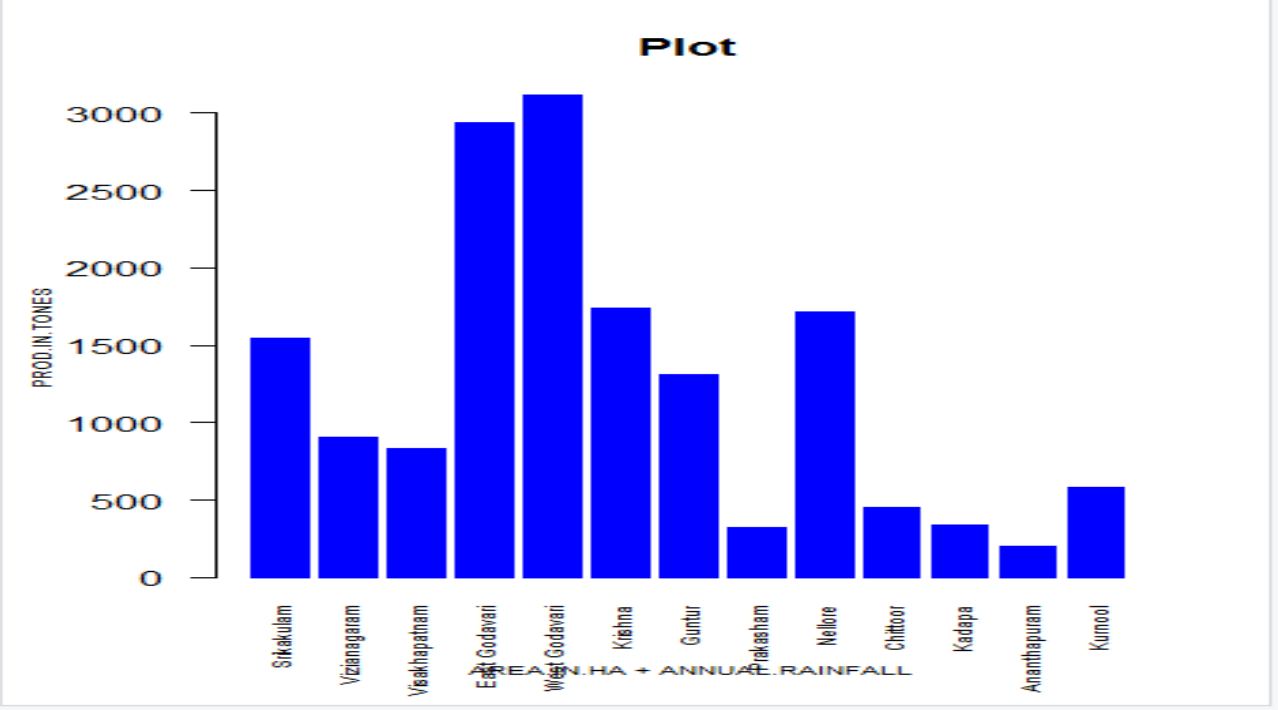

Fig .3: Graph showing Estimated Prediction of Maize crop for the year 2019-20

The Production of Maize Crop for the year 2019-20 was highest in West Godavari district and lowest in Ananthapuram district.

Groundnut

After analyzing the Groundnut crop yield production of the years 2016-17, 2017-18 and 2018-19 of all 13 districts in the Andhra Pradesh and predicting the next year (2019-20) crop yield production of the Groundnut crop. The following table represents the prediction values of the Groundnut in Area, Production in Tones and Annual Rainfall.

DISTRICT	AREA.IN.HA	PROD.IN.TONES	ANNUAL.RAINFALL
Srikakulam	9	1211.789	1522.208
Vizianagaram	2	1038.317	869.7138
Visakhapatnam	2	1025.167	823.6554
East Godavari	0	1629.315	2926.991
West Godavari	2	1662.912	3057.49
Krishna	2	1279.389	1714.119
Guntur	3	1150.448	1268.887
Prakasham	5	872.8428	309.3361
Nellore	14	1257.429	1714.119
Chittoor	137	670.7324	447.5114
Kadapa	52	793.408	332.3653
Ananthapuram	524	114.178	178.8372
Kurnool	92	781.5722	547.3047

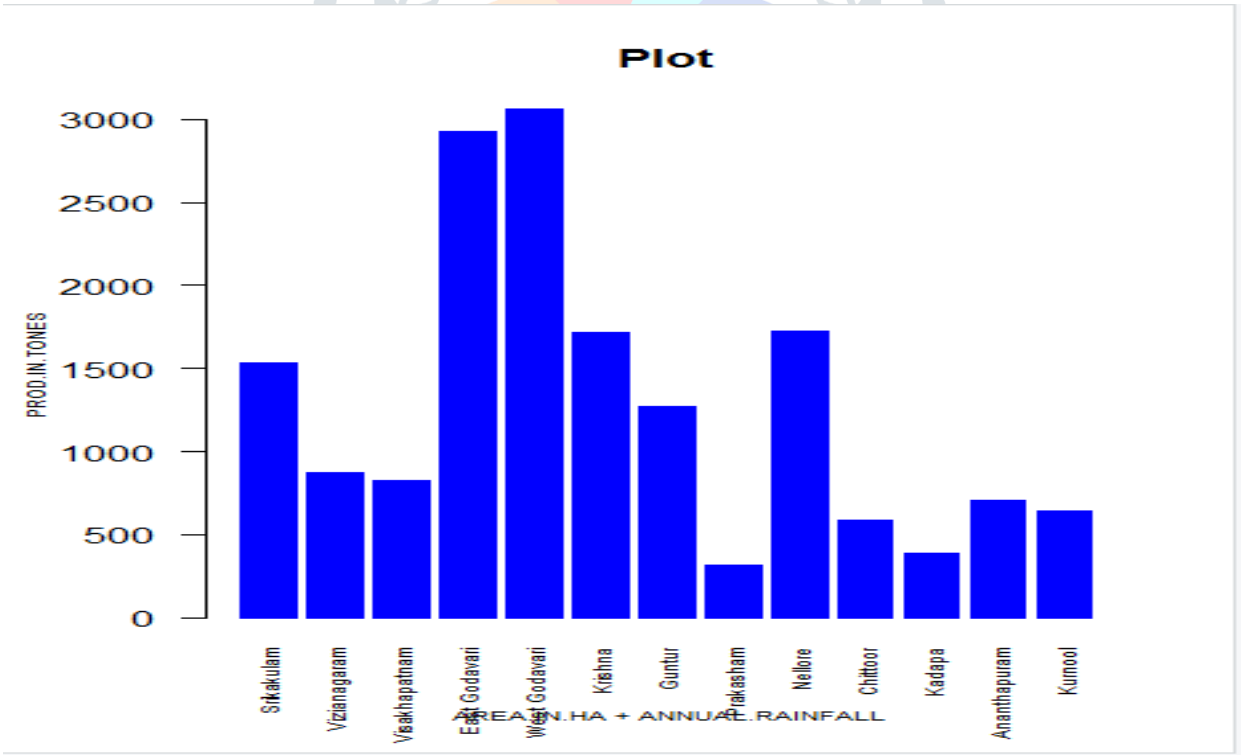

Fig . 4: Graph showing Estimated Prediction of Groundnut crop for the year 2019-20

The Production of Groundnut Crop for the year 2019-20 was highest in West Godavari district and lowest in Prakasham district.

Sugarcane

After analyzing the Sugarcane crop yield production of the years 2016-17, 2017-18 and 2018-19 of all 13 districts in the Andhra Pradesh and predicting the next year (2019-20) crop yield production of the Sugarcane crop. The following table represents the prediction values of the Sugarcane crop in Area, Production in Tones and Annual Rainfall.

DISTRICT	AREA.IN.HA	PROD.IN.TONES	ANNUAL.RAINFALL
Srikakulam	5	1219.109	1522.208
Vizianagaram	13	1018.187	869.7138
Visakhapatnam	34	966.6076	823.6554
East Godavari	9	1612.845	2926.991
West Godavari	10	1648.272	3057.49
Krishna	12	1261.089	1714.119
Guntur	0	1155.938	1268.887
Prakasham	0	881.9928	309.3361
Nellore	1	1281.218	1714.119
Chittoor	20	884.841	447.5114
Kadapa	0	888.5675	332.3653
Ananthapuram	0	844.7363	178.8372
Kurnool	1	948.1012	547.3047

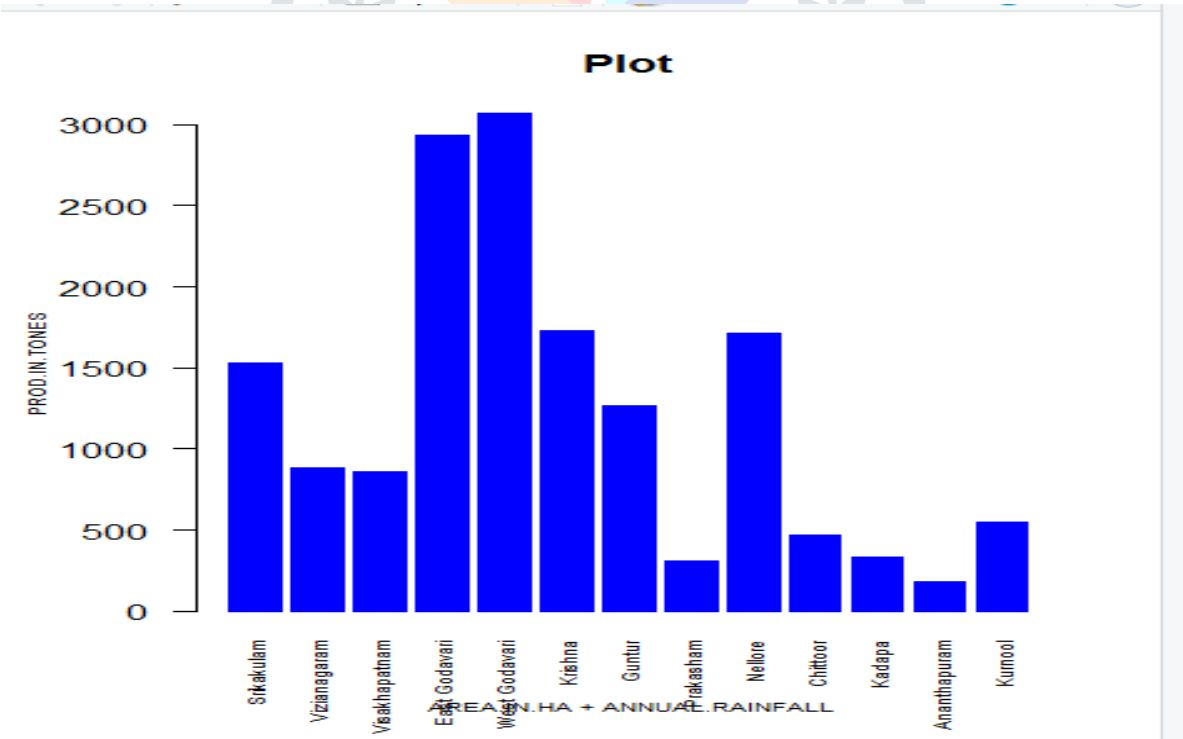

Fig. 5: Graph showing Estimated Prediction of Sugarcane crop for the year 2019-20

The Production of Sugarcane Crop for the year 2019-20 was highest in West Godavari district and lowest in Ananthapuram district.

Greengram

After analyzing the Greengram crop yield production of the years 2016-17, 2017-18 and 2018-19 of all 13 districts in the Andhra Pradesh and predicting the next year (2019-20) crop yield production of the Greengram crop. The following table represents the prediction values of the Greengram crop in Area, Production in Tones and Annual Rainfall.

DISTRICT	AREA.IN.HA	PROD.IN.TONES	ANNUAL.RAINFALL
Srikakulam	34	1166.04	1522.208
Vizianagaram	13	1018.187	869.7138
Visakhapatnam	3	1023.337	823.6554
East Godavari	17	1598.205	2926.991
West Godavari	14	1640.952	3057.49
Krishna	11	1262.919	1714.119
Guntur	41	1080.908	1268.887
Prakasham	5	872.8428	309.3361
Nellore	2	1279.389	1714.119
Chittoor	1	919.6109	447.5114
Kadapa	1	886.7374	332.3653
Ananthapuram	5	835.5863	178.8372
Kurnool	4	942.6112	547.3047

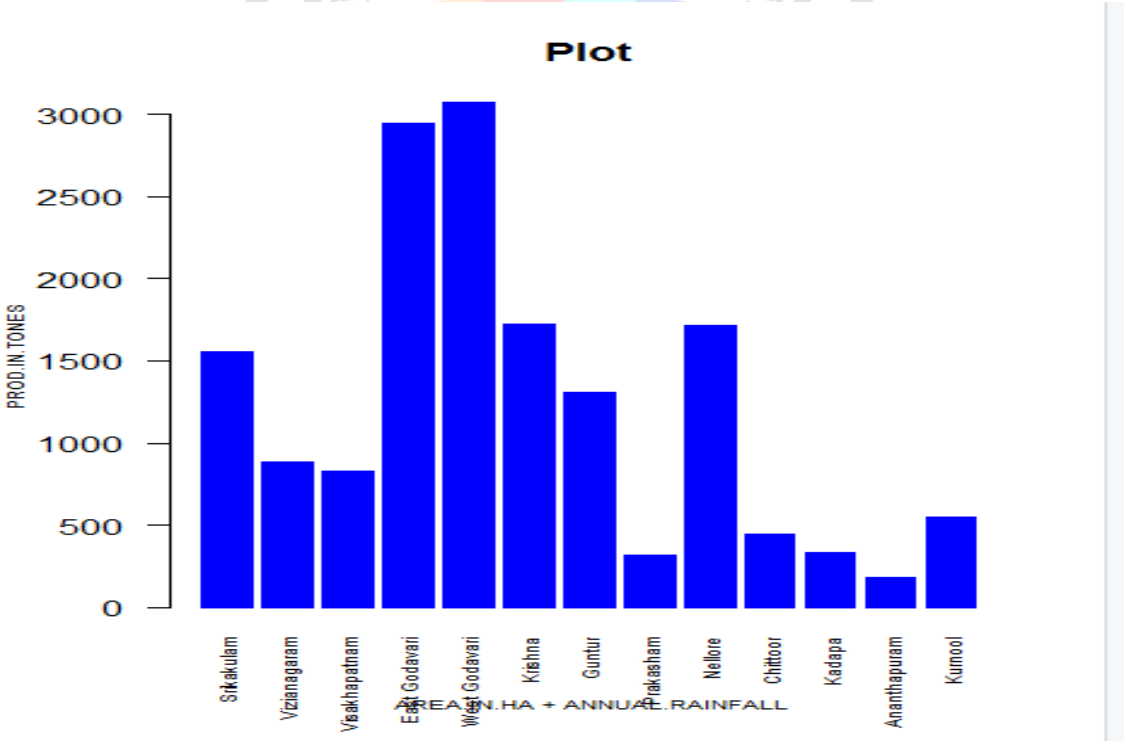

Fig. 6 : Graph showing Estimated Prediction of Greengram crop for the year 2019-20

The Production of Greengram Crop for the year 2019-20 was highest in West Godavari district and lowest in Ananthapuram district.

CONCLUSION

As humans, we cannot predict the future production. Taking this problem into consideration we have found a solution Prediction of various Crop Yields in Andhra Pradesh using Multiple Linear Regressions using Data Mining tool R.

The positive signs would be an increase in the farm incomes, the economic status of the dependent families, and regular debt repayment by the farmers. With the government spending large budget on the farm loan waiver , free electricity supply, fertilizer subsidy, availability of credit etc., It is important to track the farm-level incomes to target and support the most vulnerable sectors involved in the agricultural and allied sectors.

REFERENCES

[1]. https://www.investopedia.com/terms/m/mlr.asp
[2]. https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/
[3]. Duan Yan-e ,―Design of Intelligent Agriculture Management Information System Based on IOT‖,IEEE,4th,.Fourth International reference on Intelligent Computation Technology and Automation,2011.
[4]. Jiawei and MichelineKamber. Simon Fraser University ―Data Mining Concepts & Techniques‖ 2000.
[5]. Abdullah, A., Brobst, S., M.Umer M. 2004. "The case for an agri data ware house: Enabling analytical exploration of integrated agricultural data". Proc. of IASTED International Conference on Databases and Applications. Austria. Feb
[6]. Janet Kaaya,‖Role of information technology in agriculture‖,Proceedings of FoA Conference, Volume 4,1999.