Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018)
IEEE Xplore Compliant - Part Number:CFP18J06-ART, ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7

Use of Data Mining in Crop Yield Prediction

Shruti Mishra **Priyanka Paygude** **Snehal Chaudhary** **Sonali Idate**

Department of Information Technology,

Bharati Vidyapeeth (Deemed to be University), India, College of Engineering, Pune

Abstract- Agriculture is the most important sector that influences the economy of India. It contributes to 18% of India's Gross Domestic Product (GDP) and gives employment to 50% of the population of India. People of India are practicing Agriculture for years but the results are never satisfying due to various factors that affect the crop yield. To fulfill the needs of around 1.2 billion people, it is very important to have a good yield of crops. Due to factors like soil type, precipitation, seed quality, lack of technical facilities etc the crop yield is directly influenced. Hence, new technologies are necessary for satisfying the growing need and farmers must work smartly by opting new technologies rather than going for trivial methods. This paper focuses on implementing crop yield prediction system by using Data Mining techniques by doing analysis on agriculture dataset. Different classifiers are used namely J48, LWL, LAD Tree and IBK for prediction and then the performance of each is compared using WEKA tool. For evaluating performance Accuracy is used as one of the factors. The classifiers are further compared with the values of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Relative Absolute Error (RAE). Lesser the value of error, more accurate the algorithm will work. The result is based on comparison among the classifiers.

Keywords-Data Mining, classification, J48, LWL, LAD Tree, IBK..

I. INTRODUCTION

Data Mining is the process of analyzing, extracting and predicting the meaningful information from huge data to extract some pattern. This process is used by companies to turn the raw data of their customer to useful information. The process of Data Mining includes first selection of data followed by pre-processing of data and then transforming the data to get patterns which can then be used to predict useful insights. Pre processing includes finding outliers and detecting missing values whereas transformation finds the correlation between objects.

Applying the data mining techniques on historical climate and crop production data several predictions can be made on the basis of knowledge gathered which in turn can help in increasing crop productivity. Decision Support System (DSS) has to be implemented for the farmers to prevent the overheads of decisions about the soil and crop to be cultivated. DSS is a software system that helps the analysts to predict or identify useful information from a raw dataset, documents or business models to analyze a problem and solve it by making decisions".

This system would help farmers to make important decisions which were earlier taken by using inefficient trivial methods or by guessing. The prediction system will be implemented by using data mining techniques. Previous researches [1-3] depict the application of data mining techniques in the agricultural sector.

This work includes several sections as follows – Section II describes all the previous work which were accomplished by several researchers. The motivation behind this paper is discussed in the Section III. For experimentation the dataset is used which is described in Section IV. Several algorithms are used for analysis namely Classification algorithms namely – J48 and LAD Tree and Lazy Learner algorithms namely – IBK and LWL which are discussed in Section V. WEKA tool is used for analysis. Section VI presents the experimentation steps in WEKA and also depicts the confusion matrix for each of the classifier. Performance measurements and their general definitions are elaborated in Section VII. Terms like RMSE, MAE, RAE, Sensitivity, Specificity and accuracy are defined. The performance of each classifier is evaluated through factors namely Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Relative Absolute Error (RAE). Accuracy is also compared which is depicted in Table 2 and Table 3 of Section VIII. In the end the Conclusion along with the future work is discussed.

II. RELATED WORK

Different research in past explains that data mining techniques can help in building a system that could effectively Solve complex agriculture problems without less human intervention. [3-5] analyzed on agriculture data by using different techniques and later compared the performances. S. Pudumalar, E.Ramanujam, R.Harine, Rajashreeń, C.Kavyań, T.Kiruthikań, J.Nishań used ensemble model for precision agriculture [6]. R. Kumar, M.P. Singh, P Kumar and J.P. Singh proposed the use of data mining techniques for implementing Crop Selection Method (CSM) which tells about the sequence of crops to be planted [7]. Ahamed et al applied K-means Clustering technique to predict the rice yield in the areas of Bangladesh [5]. A. Mucherino, P. Papajorgj and P. M. Pardalos surveyed about different data mining techniques and how they can be useful in agriculture sector [1]. N. Gandhi and L. Armstrong studied and analyzed agriculture dataset for prediction of rice yield in humid subtropical climate zone and tropical wet and dry climate zone in India [3-5]. In [9] D. Ramesh and B. Vardhan worked on prediction of crop yield using Multiple Linear Regression (MLR) and Density based clustering techniques. U. Kumar Dey, Abdullah Hasan Masud, Mohammed Nazim Uddin in their study analyzed crop yield prediction by using Support Vector Machine (SVM), Multiple Linear Regression (MLR), AdaBoost and Modified Non Linear Regression [10]. In [11-14] study is related to the data mining techniques which are applied on agriculture dataset for analysis of different algorithms. [15-18] discusses about analysis of different algorithms and at the end comparing them through different factors such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Accuracy or factors like Average Rain in Area, Average Sowing in Area etc. In [19] the authors have discussed about a type of agriculture called Precision Agriculture. In [20], authors worked on gathering remote sensing data and then further using it to build indices for analyzing crop productivity. In [23], the study presents the comparison of classification methods like KNN, Bayesian Network, and Decision Tree.

III. MOTIVATION

India is the highest crop producing country in world competing U.S. and it make sense as it has to satisfy its huge population. Every year tones of crop is destroyed either due to climatic percussions or due to unawareness of the cultivation cycle. Every farmer wants to know about the amount of yield his farm will produce to get the estimates about his earnings. Indian farmers face a lot of challenges in making decisions about which farming technique to opt for and which crop should be selected for which climate. The agriculture sector directly affects the overall economic development of the country. Indian land is being acquainted with various soil types and experiences different climates, hence just need a little prior planning with the help of historical datasets. There have been many studies related to crop yield in past by many researchers but each has their own loophole. Hence, there is always a scope of improvement in this study as the classifiers can be modified to enhance the performance.

IV. DATA SET

The data set used in this experiment is taken from kaggle.com [25] which is the community of large number of data scientists which provides datasets for various analyses or to build models. The data set was organized in Microsoft Excel with columns as Sr.No, District Name, Year, Crop, Area and Production. There were various fields empty which were filled by approximate values. The file was in CSV (Comma Separated Values) format. There were 152 records of 7 districts of Maharashtra and for each district there were 17 different crops. After that preprocessing of data was performed in which certain columns were removed which not required in our study like Sr.No, District name and Year. Major factors used are:

1) Production: Total production (in tones) for each crop in different districts was used for the study.

2) Area: Area of cultivation (in Hectare) for different crops of different district was taken in account.

3) Seasons: There are two seasons for growing crops. These two seasons are named according to the season of harvest of the crop namely Kharif and Rabi.

V. METHODOLOGY USED

All the analysis of the data set was done using "WEKA (Waikato Environment for Knowledge Analysis)" [8]. It is open source software which is written in JAVA programming language and developed in University of Waikato, New Zealand. It is used for solving data mining and machine learning problems. It is licensed under GNU General Public License

It performs tasks like preprocessing, classification, regression, clustering and visualization. The data set has to be fed into the software and desired task is selected. It provides number of classifier for building models and solves analytical problems. It has the interactive Graphical User Interface (GUI) with all the options that are required for data analysis. The dataset used for processing using WEKA is stored in .arff (Attribute Relation file format).

A. Classification

Classification algorithms uses classifiers to classify a group of similar objects under one type and when a new object is introduced, prediction is made so as to put that object into one of the class. This technique helps in categorizing data in different classes. Classification comes under supervised algorithm. Several flavors under Classification are Naïve Bayes, K-Nearest Neighbor, Decision Tree; Support Vector Machine etc.The classification is done using 2 phases. First the training set builds the model that is required to answer our query and second the performance of the model is checked with the help of test data.

B. Classification Algorithms

The present study aims to use four different classifiers namely J48, LAD Tree, LWL and IBK.

1) J48

J48 is an extension of C4.5 which is used to generate a decision tree using C4.5 algorithm. Decision tree generated can be used for classification and hence also called statistical classifier. The main thing that must be kept in mind while using algorithm is that the database must be properly organized and information must be correct for proper analysis.

2) LWL

Locally Weighted Learning is a lazy learning algorithm. It uses an instance-based algorithm to assign instance weights. Being a lazy learner it defers the processing of data until it becomes necessary to give the results of a query.

3) LAD Tree

Logical Analysis of Data (LAD) is a rule-based machine learning algorithm based on ideas from Optimization and Boolean Function Theory. It is a classifier for binary target variable. This option is present under tree sub menu.

4) IBK

Instance Based K-nearest neighbor uses K-NN for classification. It is also a lazy learner i.e. it delays construction of classifier until classification time.

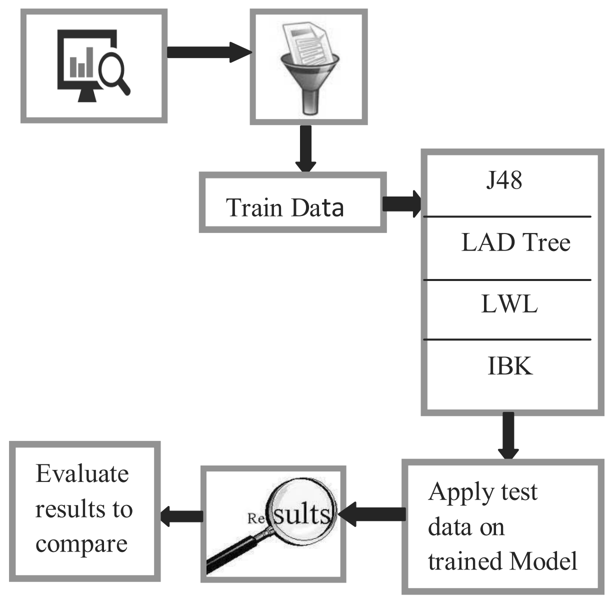
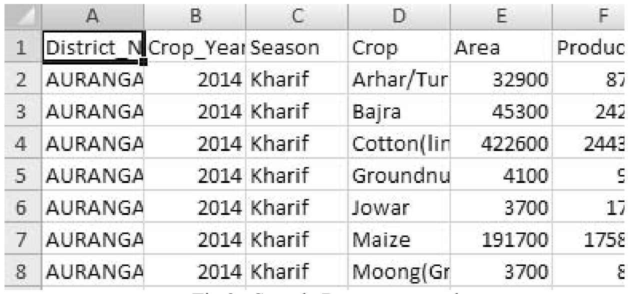


Fig. 1 Architecture of Prediction System

The above figure shows the entire architecture of Yield prediction system. The raw data or weather statistics are used, which are then cleaned and sorted. The classification techniques like J48, LAD Tree, IBK, LWL are then implemented over the trained data. The results of each algorithm is noted from WEKA and compared with each other. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Relative Absolute Error (RAE) values are taken into consideration for each case. Thereafter performance is measured using three factors namely Sensitivity, Specificity, and Accuracy.

VI. EXPERIMENT ON WEKA

The dataset is opened in WEKA tool. It is first pre processsed to remove unnecessary data columns like Distric _name and Crop_year.



	A	B	C	D	E	F
1	District_N	Crop_Year	Season	Crop	Area	Produc
2	AURANGA	2014	Kharif	Arhar/Tur	32900	87
3	AURANGA	2014	Kharif	Bajra	45300	242
4	AURANGA	2014	Kharif	Cotton(lin	422600	2443
5	AURANGA	2014	Kharif	Groundnu	4100	9
6	AURANGA	2014	Kharif	Jowar	3700	17
7	AURANGA	2014	Kharif	Maize	191700	1758
8	AURANGA	2014	Kharif	Moong(Gr	3700	8

Fig 2. Sample Dataset screenshot

Fig 3. Options in WEKA tool

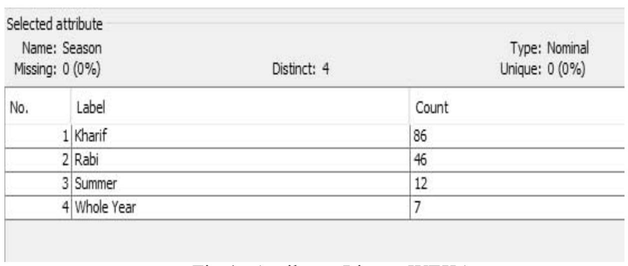


Fig 4. Attributes List on WEKA

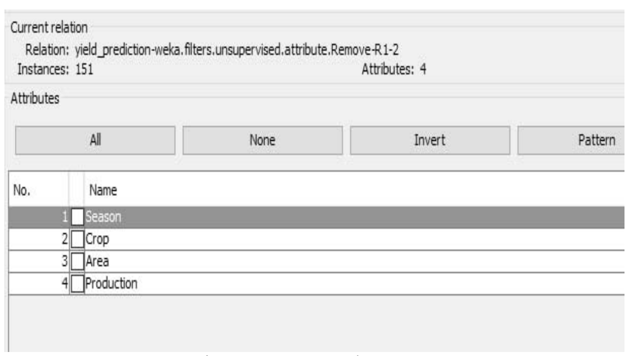


Fig 5. Preprocessed Dataset

Classify tab is selected to apply the instances on dataset to perform analysis. One by one we choose four techniques namely LWL, IBK, LAD Tree, and J48 and the results are displayed on the screen. Season attribute is selected as it has the nominal values. The values of RMSE, MAE, and RAE are noted.

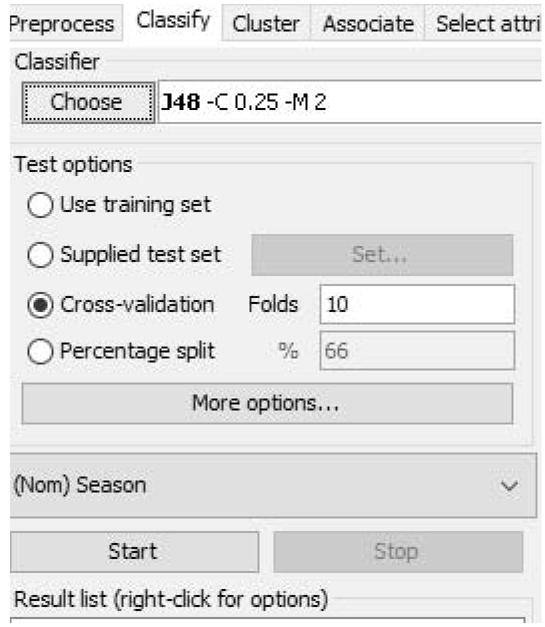


Fig 6 Selection of J48 classifier

For each classifier the confusion matrix is produced which helps in describing the performance of the classifier with the values present in it. The diagonal values depict the correctly identified instances for each attribute and all other values depict incorrectly identified instances.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 77  5  4  0 |  a = Kharif
 11 31  4  0 |  b = Rabi
  6  3  3  0 |  c = Summer
  0  0  0  7 |  d = Whole Year
```

Fig 7 Confusion Matrix for J48

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 82  0  4  0 |  a = Kharif
 36 10  0  0 |  b = Rabi
 11  0  1  0 |  c = Summer
  0  0  0  7 |  d = Whole Year
```

Fig 8 Confusion Matrix for LWL

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 73  7  6  0 |  a = Kharif
  6 36  4  0 |  b = Rabi
  3  3  6  0 |  c = Summer
  0  0  0  7 |  d = Whole Year
```

Fig 9 Confusion Matrix for IBK

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 86  0  0  0 |  a = Kharif
 38  8  0  0 |  b = Rabi
 12  0  0  0 |  c = Summer
  7  0  0  0 |  d = Whole Year
```

Fig 10 Confusion Matrix for LAD Tree

The matrices depict the values for the instances correctly classified and instances incorrectly classified for four different seasons namely Kharif, Rabi, summer and Whole year.

VII. PERFORMANCE EVALUATION

A. Factors used for performance measurements are:

1) Specificity: It is defined as percentage of incorrectly classified instances. It is True Negative Rate (TNR)

2) Sensitivity: It is defined as percentage of correctly classified instances. It is True Positive Rate (TPR).

3) Accuracy: It is defined as the overall success rate of the classifier.

***4) RMSE*:** It is defined as the difference between the values predicted by the model and the actual values noted

***5) MAE*:** It is another factor in statistics which measures the difference between two continuous variables.

***6) RAE*:** This measure gives the total absolute error between the variables.

True Class	Positive	Negative	Total
Positive	TP	FN	TP+FN
Negative	FP	TN	FP+TN
Total	TP+FP	FN+TN	TP+FP+FN+TN

Table 1 Performance Measure

B. General Definitions:

1) True Positive (TP) depicts the number of instances where system detects for a condition when it is really present.

2) True Negative (TN) depicts the number of instances where system does not detect a condition when it is absent.

3) False Negative (FN) depicts the number of instances where system does not detect a condition when actually it is present.

4) False Positive (FP) depicts the number of instances where system detects a condition when it is really absent.

Following are the equations that calculate the sensitivity (TPR) and specificity (TNR):

$$\text{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$$

$$\text{TNR} = TN/(FP + TN)$$

Accuracy can be calculated by:

$$(\mathrm{TP} + \mathrm{TN}) / (\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN})$$

VIII. EXPERIMENTAL RESULTS AND ITS ANALYSIS

Different classifier gives different results on same data set. The result of errors for all the four classifiers is presented in Table 1. The percentage of accuracy is presented in Table 2 for all the four classifiers.

Algorithm	**RMSE**	**MAE**	**RAE (%)**
J48	0.2773	0.1101	37.9755
LWL	0.3209	0.2213	76.3471
LAD Tree	0.4127	0.1997	68.8888
IBK	0.3057	0.104	35.8648

Table 2 Error Values

Algorithm	Accuracy (%)
J48	78.145
LWL	66.225
LAD Tree	62.251
IBK	80.794

Table 3 Accuracy Percentage

According to the results obtained from the WEKA tool LAD Tree has comparatively higher values of errors and hence its accuracy is the lowest. Further, IBK has obtained the highest accuracy among all. The values obtained are not fixed; they can be changed if pruning is further done by decreasing the values of confidence factor for each classifier. The result also depends upon the type and nature of data set.

IX. CONCLUSION AND FUTURE WORK

The errors for different classifiers can be minimized if the dataset is pruned further by decreasing the confidence factor. Lesser the errors more accurate will be the analysis. IBK achieves highest accuracy whereas; LAD Tree has the lowest accuracy. The information that we acquired after analysis can be combined in a form that is useful to the farmers for early prediction and decision making process.

With the help of this information the percentage of loses and unsatisfactory yield will decrease as the management of the whole process can be done with the help of real statistics.

In future we can use real time weather and soil datasets which will be gathered personally by equipments or the datasets can be acquired from trusted websites like indiaagristat.com or india.gov.in/data-portal-india. To further modify the model we can combine different classifiers to build one single model which is called Ensemble. By doing this we can achieve a level of performance which could not be achieved by single algorithm. Also, the nature of Dataset affects the analysis therefore, more cleaned and pre processed can be used for better results.

REFERENCES

[1] Mucherino A, Papajorgji P, Pardalos PM: A survey of data mining techniques applied to agriculture. Oper Res. 2009, 9 (2): 121-140.

[2] Cunningham, S. J., and Holmes, G. (1999). Developing innovative applications in agriculture using data mining. In the Proceedings of the Southeast Asia Regional Computer Confederation Conference,1999.

[3] N.Gandhi and L.J. Armstrong, "Applying data mining techniques to predict yield of rice in Humid Subtropical Climatic Zone of India", Proceedings of the 10th INDIACom-2016, 3rd 2016 IEEE International Conference on Computing for Sustainable Global Development, New Delhi, India, 16th to 18th March 2016.

[4] N. Gandhi and L. Armstrong, "Rice Crop Yield forecasting of Tropical Wet and Dry climatic zone of India using data mining techniques", IEEE International Conference on Advances in Computer Applications (ICACA), pp. 357-363, 2016.

[5]] N.Gandhi and L.J. Armstrong, "A review of the application of data mining techniques for decision making in agriculture", 2016 2nd InternationalConference on Contemporary Computing and Informatics (ic3i).

[6] S.Pudumalar, E. Ramanujam, R. Harine Rajashreeń, C. Kavyań, T. Kiruthikań, J. Nishań, 'Crop Recommendation System for Precision Agriculture', 2016 IEEE Eighth International Conference on Advanced Computing (ICoAC).

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

[8] WEKA 3: Data Mining Software in Java, Machine Learning Group at the University of Waikato, Official Website: http://www.cs.waikato.ac.nz/ml/weka/index.html, accessed on 12nd October 2017.

[9] D. Ramesh and B. Vardhan, "Analysis of crop yield prediction using data mining techniques", International Journal of Research in Engineering and Technology, vol. 4, no. 1, pp. 47-473, 2015.

[10] Umid Kumar Dey, Abdullah Hasan Masud, Mohammed Nazim Uddin, "Rice yield prediction model using data mining", International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, Cox's Bazar, Bangladesh.

[11] A. Ahamed, N. Mahmood, N. Hossain, M. Kabir, K. Das, F. Rahman, R. Rahman, "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh", 16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 1-6, 2015.

[12] A.B. Mankar and M.S. Burange, "Data Mining-an evolutionary view of agriculture", ", International Journal of

Application or Innovation in Engineering and Management, Vol 3, No 2 pp .102-105, March 2014

[13] Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, “Analysis of agriculture data using data mining techniques: application of big data”, Jouranl of Big Day, Springer Open.

[14] H. Patel and D. Patel, “A Brief Survey on Data Mining Techniques applied to Agriculture Data” International Journal of Computer Applications, Vol. 95, No. 9, pp. 6-8, June 2014.

[15] R.A.Medar and V.S.Rajpurohit, “A Survey on Data Mining Techniques for Crop Yield Prediction”, International Journal of Advance Research in Computer Science and Management Studies, Vol.2, No. 9, pp. 59-64, September 2014.

[16] M.C.S.Geetha, “A Survey on Data Mining Techniques in Agriculture”, International Journal of Innovative Research in Computer and communication Engineering, Vol. 3, No. 2, pp. 887-892, February 2015.

[17] A.B.Mankar and M.S.Burange, “Data Mining – An Evolutionary View of Agriculture”, International Journal of Application or Innovation in Engineering and Management, Vol. 3, No. 3, pp. 102-105, March 2014.

[18] R.Kalpana, N.Shanthi and S.Arumugam, “A Survey on Data Mining Techniques in Agriculture”, International Journal of Advances in Computer Sciences and Technology, Vol. 3, No. 8, pp. 426-431, August 2014.

[19] Anshal Savla, Parul Dhawan, Himtanaya Bhadada, Nivedita Israni, Alisha Mandholia , Sanya Bhardwaj (2015), ‘Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture', Innovations in Information,Embedded and Communication systems (ICIIECS).

[20] Aakunuri Manjula, Dr.G .Narsimha (2015), ‘XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction’, Conference on Intelligent Systems and Control (ISCO)

[21] Data Mining – (Classifier /Classification- Function), https://gerardnico.com/wiki/data_mining/classification, last accessed on 6th October 2017.

[22]Data Mining Techniques, https://www.ibm.com/developerworks/library/ba-data-mining-techniques/index.html, last accessed on 7th October 2017.

[23] Bhuvana, Dr.C.Yamini (2015), ‘Survey on Classification Algorithms in Data mining.’ International Conference on Recent Advances in Engineering Science and Management

[24]Top 10 Data Mining Algorithms, https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html, last accessed on 9th October 2017

[25] Onkar Kadam, Last updated Sept’ 2017, ‘Crop Data Analysis’ https://www.kaggle.com/onkarkadam/crop-data-analysis