Wheat Yield Prediction using Data Mining

Manan Kawatra, Jagdeep Singh, Sachin Bagga

Abstract: ***These days peasants are facing problems in producing the yield of any crop due climatic changes which are unpredictable. Therefore, to provide improvement in agriculture a well organized and approach is needed to predict the yielding of crop on the basis of which planning can be done. The prediction must be good enough to support the peasants to take appropriate actions to get the apt amount of the crop by selling it. This work aims to provide prediction with high accuracy and less RMSE. It deals with the prediction of the wheat crop yielding using data mining techniques. It helps in predicting crop yield using previous data with the help of parameters like Rainfall and Temperature. It consists of clustering which has been done with the help of k-means, extracting features with independent component analysis, instance selection using moth flame optimization and the prediction and classification has been performed with Linear Discriminant Analysis. The technique is applied MATLAB. The performance is evaluated using mean square error rate and prediction of yielding of wheat is also calculated. The error must be less to get more accuracy for the predicting rate.***

Keywords: Data-mining, Linear Discriminant Analysis (LDA), K-means clustering ,moth flame optimization, wheat yielding.

I. INTRODUCTION

Data mining means the method of extracting information from a raw volume of data. A system uses the scanning of huge volume of data and patterns without using any type of human interference [1].

Given below are areas where this is applied

- Analyze the finance related data
- Retail related process
- Telephonic-communication
- Analyze of Bio-related Data
- Other Scientific Applications
- Intruder Detecting

In generalized language. It contains four major relationships [2]. They are:

- Clustering.
- Classification.
- Association.
- Sequential Pattern.

Prediction of any natural phenomenon like crop yield needs data with respect to its season of event and nature, in light of logical examination. Manual investigation creates irregularities because of a few components like weakness, logical inconsistency of individual's recognitions, and so forth. Soft Computing like K-implies fuzzy rationale, neutral figuring, and so forth can be connected to a wide assortment of worldwide applications as they can deal with vulnerabilities superior to conventional strategies. Predictive models take notable data and can anticipate the future qualities with not so much cost but rather more rapidly. They can offer help for human choices, making them more proficient or at times; they can be utilized to mechanize whole basic leadership forms. Presently a day, Farmers are facing problems to deliver the yield on account of erratic climatic changes. This technique can provide greater efficiency in various climatic conditions. Here are the objectives of the technique proposed.

a. To perform clustering using K-means for extracting information from the data
b. To implement moth flame approach for optimization and Linear Discriminant Analysis for prediction.
c. Evaluate and compare the performance in terms of error rate and accuracy of prediction

II. RELATED WORK

Shastry and hegde (2015) proposed the technique of Adaptive Neuro Fuzzy Inference System (ANFIS), Multiple Linear Regression (MLR) and Fuzzy rationale (FL). These are used to get the prediction for the wheat yielding taking in consideration extractable soil water, biomass as attributes. The result of the expectation techniques will help horticulture organizations in furnishing agriculturists by significant data regarding those attributes add to get more wheat yield. It has found that among all the models ANFIS technique performs better as it is having less error in comparison with others.

Choudhury and Jones (2014) proposed the comparison of the yield predictions by Simple Exponential Flattening, Double Exponential Flattening, Damped-Trend Linear Exponential Smoothing, and ARMA simulations undergone unconnectedly to every region. The ARMA simulations is declared to be extra robust time-series prototypes than the smoothing methods for forecasting production in the study.

Davide Cammarano et al. **(2013)** reviewed an outline of the present crop yield prognostication strategies and early warning systems for the world strategy to boost agricultural and rural statistics across the world. Totally different sections describing simulation models, remote sensing, yield gap analysis, and strategies to yield prognostication compose the manuscript.

Manuscript published on November 30, 2019.
* Correspondence Author

Manan Kawatra*, Information Technology, Guru Nanak Dev Engineering College, Ludhiana, India. Email: manankawatra26@gmail.com
Prof. Jagdeep Singh, Information Technology, Guru Nanak Dev Engineering College, Ludhiana, India. Email: jagdeepmalhi@gndec.ac.in
Prof. Sachin Bagga, Information Technology, Guru Nanak Dev Engineering College, Ludhiana, India. Email: sachinbagga@gndec.ac.in.

Farheen et al. (2016) proposed K-NN method for creating groups and making calculations from vast quantity of data. They have used geospatial investigation for crop produce prediction. GSA technique was practical to the extracted statistics to identify designs in the field. Once designs and connections are learned, previous knowledge or arrangements will be adapted to optimize produce and manufacture costs, and minimalism environmental influence.

Holzman et al. (2014) proposed the crop yielding concept based on moisture content in the soil. High and low soil moisture is one of the primary restricting elements influencing crops efficiency. In this way, assurance of the connection between them is pivotal for sustenance security and support importing– sending out methodologies. The point of this work was to break down the inclination of temperature vegetation dryness list (TVDI) to estimate maize yield.

Iizumi et al. (2014) proposed the significance of ENSO to worldwide harvest creation. The checking and expectation of atmosphere actuated varieties in harvest yields, generation and fare costs in real food producing areas have turned out to be critical to empower national governments in import-subordinate nations to guarantee supplies of reasonable sustenance for buyers. In spite of the fact that the El Niño/Southern Oscillation (ENSO) regularly influences seasonal temperature and precipitation, and in this way crop yields in numerous districts, the general effects of ENSO on worldwide yields are dubious. Here it presents a worldwide guide of the consequences of ENSO on the production of real harvests and measure the effects on the worldwide average yielding irregularities.

Jiaxuan et al. (2017) presents a technique which is of low cost and accurate to get the prediction of crop yielding by openly obtainable distant detecting set of the data. Their method provides improvement in current techniques in thrice conductance. They have used forego hand-crafted topographies usually useful in the remote sensing data, also give a technique founded on modern illustration learning philosophies.

Perpetua et al. (2016) proposed brief proportional study of numerous papers that works with numerous techniques used to forecast the crop harvest. From the statistics that is willingly available, the mining methods of data give a all-inclusive picture about crop yielding approximately. Dissimilar data mining methods that remain in use for the crop yield approximation are K-Means, K-Nearest neighbour.

S Shirdhonkar et al. (2017) dealt with usage of various data mining methodologies will hinder yield of Maharashtra, India. To survey this field, 27 areas of Maharashtra were chosen on the arrangement of available information from transparently given Indian Administration chronicles distinctive troposphere and collect impediments. Precipitation, slightest temperature, typical contamination, outrageous temperature, circumstance trim evapo-transpiration, age and generation for the blustery season from June to November were the limitations decided for the preparation for the presences 1998 to 2002 by the creator.

Shreya S. et al. (2016) Associate in Nursing economical approach in field data mining to extract helpful information and provides prediction. varied approaches are enforced up to now square measure worked either for crop prediction. Crop prediction model aiding farmers to require correct call. This so helps in rising quality of farming and generate higher revenue for farmers. ancient bunch algorithms akin to k-Means, improved rough k-Means and-means++ makes the tasks difficult because of random choice of initial cluster center and call of variety of clusters. changed K-Means rule is thereby accustomed improve the accuracy of a system because it achieves the top-quality clusters duet initial cluster central choice.

Wu et al. (2015) aims at providing a novel technique to forecast the crop production based on big-data examination technology, which varies with traditional approaches in the structure of management data and in the resources of modeling. The technique can deal with full use of the current massive agriculture applicable datasets and still exploited with the capacity of data increasing rapidly, due to big-data approachable processing construction.

III. PROPOSED WORK

Literature review shows that **m**ost of the related works have not considered factors like rainfall etc. Different crops were considered, and various combinations were applied. But optimization techniques were not applied in the work so error rate was high. Their PSNR is low and RMSE is higher in comparison. So the proposed work is there to provide optimum solution. Below is the proposed framework of the technique.

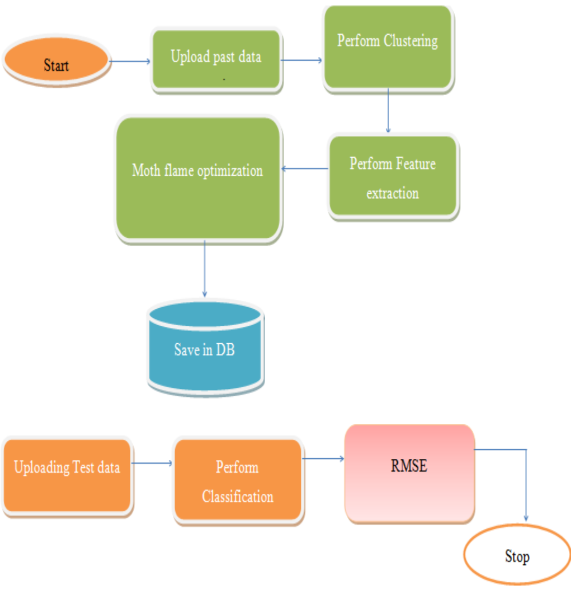


Figure.1. The Proposed Framework

A. Clustering using K-means:

K-means has been considered as a technique for vector quantization which is invented from signal dispensation used for cluster exploration in data mining process. K-means deals with the partition to n opinions gets k collections for every reflection goes to the cluster having adjacent average, his deals with the partitioning of the data into meaningful information.

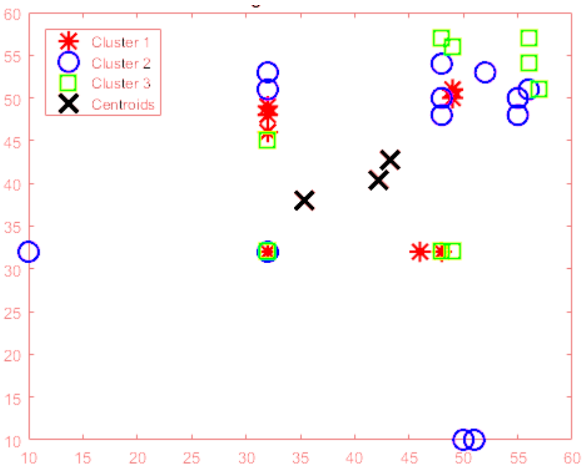


Figure2: K-means separation using clustering

The deals with the problem of N-P hard though, there are well-organized heuristic procedures that are frequently employed and meet quickly to the best level. These are frequently comparable to the maximization procedure for combinations of Gaussian distributions using iterative approach used by k-means procedure and Gaussian mixture approach. Moreover, they use cluster midpoints to prototype the information. K-means tends to discover clusters of comparable three-dimensional amount, while the expectation apparatus allows collections to have dissimilar shapes.

B. Linear Discriminant analysis:

It is a simplification of Linear Discriminant, a process deals with the statistics, configuration recognition and learning to catch a linear mixture of features that symbolizes or splits two or more programmes of events. The subsequent grouping may be recycled as a linear and more commonly, for reduction of dimensions before later arrangement.

LDA is related to examination of variance and regression examination, which tries to direct one variable which is dependent as a group of additional attributes or capacities. Though, ANOVA deals with the categorical variables and continuous variable, while discriminant examination has incessant independent scenario and a definite dependent variable. Logistic regression is more alike to discriminant analysis than analysis of variance is, as they clarify a categorical mutable by the standards of incessant independent solutions. These methods are better in solicitations where it is not sensible to undertake that the sovereign variables are usually distributed, which is an important supposition of the LDA technique.

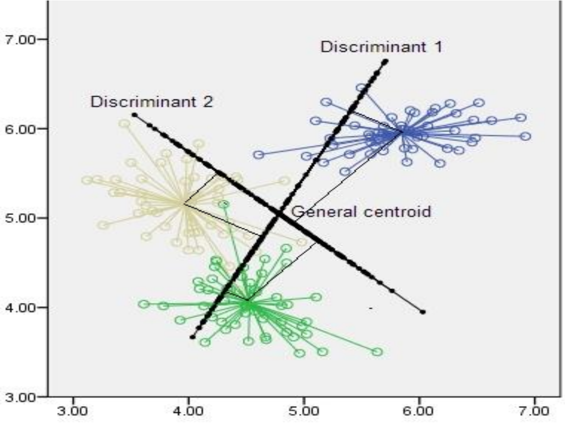


Figure3: LDA analysis

LDA very closely connected to principal component analysis and analysis of factor in that they appear for linear mixtures of solutions which are the best solutions of the data. LDA openly efforts to model the alteration among the information passed. PCA does not receipt into explanation any change in class, and issue analysis builds the feature mixtures based on alterations rather than resemblances. Discriminant study is also dissimilar from factor study which is not an interdependent method: a difference among independent solutions and dependent solutions which is also called criterion variables which areended.LDA mechanisms when the capacities made on sovereign variables for each opinion are nonstop measures. When selling with definite independent variables, the corresponding method is discriminant study.

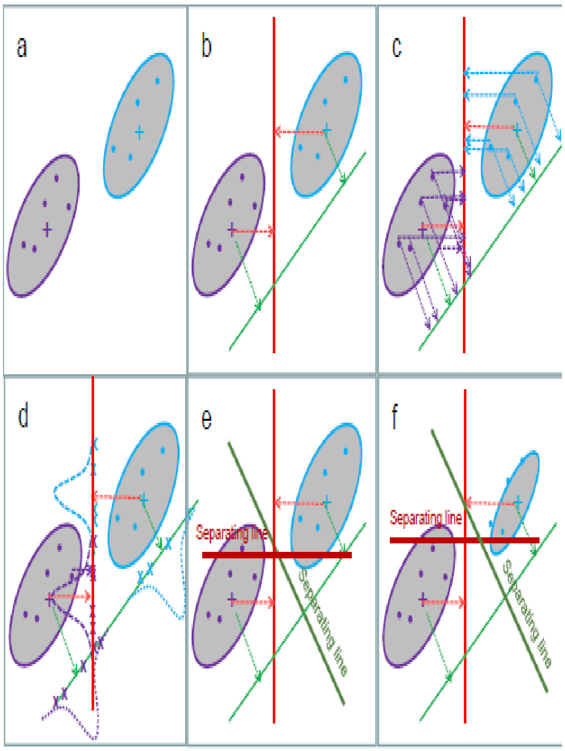


Figure4: LDA classification

C. Independent Component analysis:

ICA is a proficient arithmetical and computational strategy for enlightening concealed impacts that inspire gatherings of arbitrary variables, measurements, or signs.

It depicts a regenerative perfect for the experiential multivariate statistics, which is naturally accepted as a vast record of samples. In the prototypical model, the information is relied upon to be Linear mix of some unidentified dormant information, and the blending course of action is likewise unidentified. The concealed factors are normal non-Gaussian and similarly autonomous and are known as the free segments. These segments, otherwise called the sources that should be possible by ICA are one of the productive information to be naturally accessible.

ICA is identified with part study and factor examination. ICA is a more powerful system, achieved of disclosure the components or establishments when exemplary methodologies slump totally.

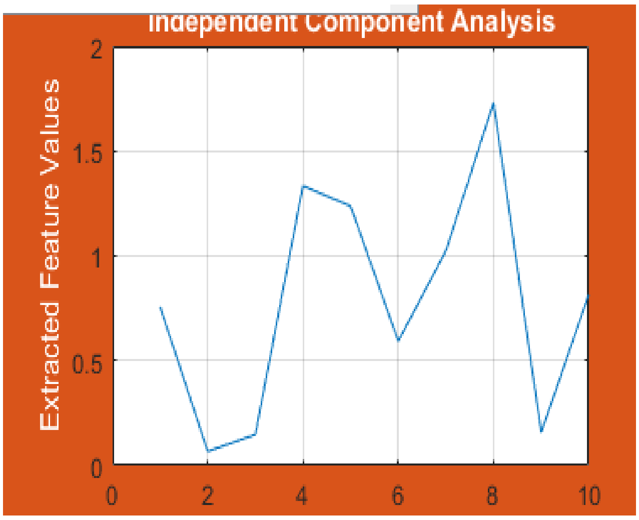


Figure5: Extracting Features using ICA

The data examined using ICA could create many dissimilar kinds of submission pitches, deals with digital imageries, text records, economic gauges and psychometric capacities. In many belongings, the capacities are assumed as a set of similar signs; the period blind separation is used to describe this tricky. Typical instances are combinations of concurrent signals that is picked up using numerous microphones, intelligence waves logged by multiple instruments, meddlesome signals incoming at a mobile handset, or parallel series gained from some manufacturing procedure

D. Moth flame Optimization:

The primary aim of this optimization is the working or flying technique of moths naturally which can be termed as transverse orientation. Moths are seen flying in night with a way or sequence that they maintain a specific inclination from moon, using an efficient criterion for travelling in line for travelling huge path. But they are caught in a useless path which is spiral in shape making a round again and again around lights which are artificial. This model uses this process in mathematics to make a technique to get an optimal solution**.**

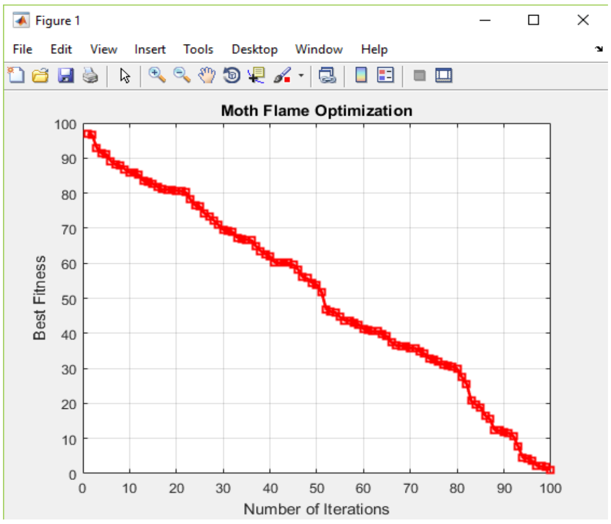


Figure6: The principal of moth flame optimization

IV. RESULTS

The experimental results are evaluated from the proposed framework in the form of RMSE in comparison with other approaches.

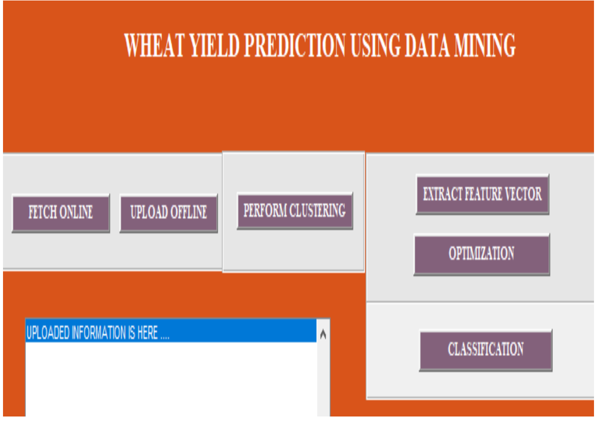
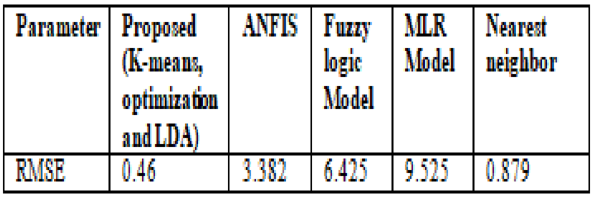


Figure7: The Main panel of the technique used

The above figure shows the main panel of the system used here. It includes clustering, extracting features, optimization and classification

Table 1: Comparison of the performance with proposed technology



Parameter	Proposed (K-means, optimization and LDA)	ANFIS	Fuzzy logic Model	MLR Model	Nearest neighbor
RMSE	0.46	3.382	6.425	9.525	0.879

The table shows the numeric comparison between various techniques which were used earlier with the proposed one. It depicts that proposed technique has lesser value of error as optimization has been done which were not included in previous works. This helps in giving appropriate results with less error rates.

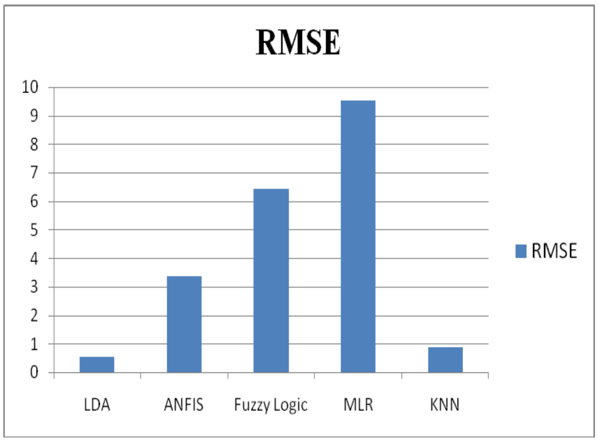
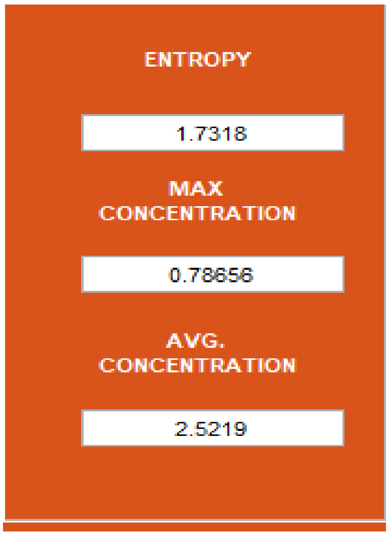


Figure8: RMSE Comparison

The Figure7 shows the graphical representation of the comparison of RMSE among different methods. This shows that RMSE is less for LDA and hence it is better than other methods.



ENTROPY

1.7318

MAX CONCENTRATION

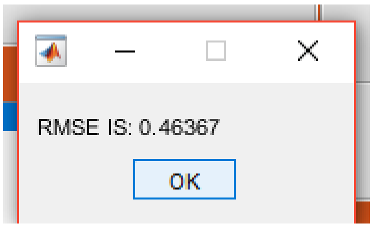
0.78656

AVG. CONCENTRATION

2.5219

Figure 9: Extracted values

The Figure8 depicts the values of the attributes extracted for example concentrations and entropy that provides the density of the content which undergoes the processing and the degree of randomness that can be used for testing. It is an important step for extracting features that can be used further for prediction.



RMSE IS: 0.46367

Figure10: RMSE

The above figure is showing RMSE that depicts that system can be used to attain low RMSE and can also depicts the yield of wheat required in coming time so that it will help in planning about improvement and estimating the production while cultivating.

V. CONCLUSION

Atmosphere and different natural conditions are changing in Indian landmass and the climatic conditions are unpredictable and these have been developing critical problems to the agriculture. Presently a day, farmers are facing problems to create the yield due to eccentric climatic changes. Crop yield predicting, which delivers evidence for decision making, is significant in various aspects of country's economy. On account of the noteworthiness of crop yield gauge, the assurance of this approach is to relate a few foreseeing strategies for surveying crop yield anticipating recreations.

Data mining means of getting the meaningful data through a data set in a way which can be is helpful to attain more accuracy and small ER estimations. Farming is very important in Indian economy so security and some degree of accuracy should be there so that investment can give proper output in earlier literatures the researchers did not get low value of RMSE which can be considered as primary aspect of this technique. So, this technique consists of a combination of many methods to get lesser value of. It consists of clustering, optimization and classification the performance evaluation has been done with the help of RMSE and PSNR and this approach helps in achieving high PSNR and less RMSE that shows greater prediction rates. So, the developed system can get achieve low ER in terms of MSE and more PSNR which deals with the less prediction errors using data mining approach. The developed system can get RMSE of 0.46 with high signal to noise ratio. The PSNR should be more which depicts system helps to get less error rate in harsh environments

VI. FUTURE SCOPE

In future Scope it may include many other optimization techniques or lesser PSNR can be tried to achieve by using other algorithms combinations. So that more comparisons can be done, and more reliability can be attained and best suited technology can be made. More climatic parameters can be used to give best result. Dependency of the yield on each climatic factor can be there that is whether the yield is directly dependent to that factor or inversely

REFERENCES

1. Choudhury, J.,2014, "Crop yield prediction with use of time series model" Journal of Economics and Economic Education Research, Vol. 15, Issue 3, pp. 53-65.
2. Davide,Cammarano, ElisabettaCarfagna., 2004," Review of Crop Yield Forecast Methods, Early Warning Systems" Vol. 9, Issue 2, pp. 27-32.
3. Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data "International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, pp 15-20
4. Fischer, A., Byerlee, and Greg , E., 2014, "Crop yielding and global food security." ACIAR: Canberra Vol. 3, Issue 2, pp.341-352.
5. Holzman, M E., Rivas, R. and Piccolo, MC, 2014, "Estimation of soil moisture and the relationship with crop yield by use of temperature and vegetation." International Journal of Applied Earth Observation and Geoinformation, Vol. 28, Issue 4, pp.181-192.
6. Jay, L., lLapira, E., Behrad , B. and Kao, H.,2013, "Recent advances and trends in manufacturing systems in big data ",Manufacturing Letters, Vol.1 , Issue.1, pp. 38-41.
7. Jiaxuan,., Low, M.,Lobell, D. and Ermon, S.,2017, "Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing ", AAAI, Vol.7, Issue6, pp. 4559-4566
8. M. Hauben, V. Patadia, and D. Goldsmith., 2006, "What counts in data mining?", vol. 29, no. 10, pp. 827-832.
9. Naushina, Prof. R. V.,2016, "Annual Crop Yield Prediction and Recommend Planting of Different Crops by Using Data Mining Technique ", IJIRC, Vol. 4, issue 10, pp.176-184.
10. Perpetua, N,Shruthi, B.S., 2016, "Comparative Study of Data Mining Techniques in Crop Yield Prediction" , IJARCCE, Vol. 5,Issue 2,pp. 341-350.
11. A. Shastry, H. A. Sanjay, and M. Hegde.,2015, "A parameter based ANFIS model for crop yield prediction", (IACC), 2015 IEEE International, pp. 253-257. IEEE.
12. N.A. Hessling 1992, "Relationship between the Weather and the Yielding of Wheat in the Argentine Republic", Monthly Weather Review 50, Issue 6, pp. 302-308.
13. Paredes, P., Rodrigues and Pereira, L.S., 2014, "Partition evapo-transpiration, yield prediction of maize under various irrigation management strategies." Agricultural, Vol. 9, Issue 4, pp.27-39.
14. Wu, F., GuoXiaoling, CC., Hua, Y. and Juyun, W., 2015, "Prediction of Crop Yield Using Big Data ", Computational Intelligence and Design (ISCID), Vol. 1, Issue 3, pp. 255-260.
15. Iizumi, Toshichika, Jing-JiaLuo, Andrew J. Challinor, Gen Sakurai, Masayuki Yokozawa, Hirofumi Sakuma, Molly E. Brown, and Toshio Yamagata.,2014, "Impacts of El Niño Southern Oscillation on the global yields of major crops."Nature communications Vol.3, Issue5, pp.3712.
16. S.S. Bhanose, K. A. Bogawar, A. G. Dhotre, and B. R. Gaidhani.,"Crop and Yielding Prediction Model", International journal 1, no. 1.
17. W. Hugh, and J. Robert, 2001, "Data warehouses stages of growth", Information Systems Management; vol. 18, no. 3, pp. 42-51.