Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Analysis of agricultural crop yield prediction using statistical techniques of machine learning

Janmejay Pant [a], R.P. Pant [a], Manoj Kumar Singh [a], Devesh Pratap Singh [b], Himanshu Pant [a]

[a] Graphic Era Hill University, Bhimtal, India
[b] Graphic Era Deemed to be University, Dehradun, India

ARTICLE INFO

Article history:
Received 13 January 2021
Received in revised form 26 January 2021
Accepted 31 January 2021
Available online xxxx

Keywords:
Machine learning
Yield prediction
Agriculture
Accuracy
Feature selection
Correlation

ABSTRACT

Agriculture plays a crucial role in Indian economy. Crop yield is main component of food security as human population is increasing in a drastic way. One of the most important problems of agriculture is crop yield prediction. Agriculture yield depends on the various factors such as weather situation (rain, humidity, temperature etc.), information about pesticides. Apart from these factors exact information about the crop yield history is an essential concept for making predictions and controlling agriculture risk. Earlier yield prediction was performed by considering the farmer's experience on a particular field and crop. In this study machine learning is used to predict four popular yields which are mostly cultivated all over India. Once the crop yield is site specifically predicted, the inputs such as fertilizers could be applied variably according to the expected crop and soil needs. In our study we use Machine Learning approaches to develop a trained model to identify the patterns among data and it is used for crop prediction. In this study the prediction of four most cultivated yields in India is considered by applying machine learning. These crops include: Maize, Potatoes, Rice (Paddy) and wheat.

Selection and peer-review under responsibility of the scientific committee of the International Conference on Technological Advancements in Materials Science and Manufacturing.

1. Introduction

Crop yield prediction has the great importance in food production. Policy makers rely on accurate predictions to make timely import and export decisions to strengthen national food Security [1,2]. Yield prediction is also beneficial for the farmers for making decisions. Due to various complex factors, crop yield prediction is a challenging task. Crop yield prediction [3] may also support to policy or decision makers in agriculture sector. It also helps to identify the relevant features which naturally affect the crop yield. In earlier days the crop yield prediction was carried out by the experience of the farmers [4]. In the recent time farmers are required to cultivate more and more crops because the conditions and situations are changing rapidly. Still farmers don't have enough knowledge about the new crops and they are not aware about the environmental conditions which effect crop production. Such types of issues can be resolved by crop prediction. In India agriculture is highly affected by rainwater and it is unpredictable. Crop growth is also depends on the various constraints like soil micro and macro parameters, soil moisture, temperature etc [5]. A big part of Indian economy depends on agriculture. About 60% of land is used for crop production in India. In recent era various technology are used to make a good crop production that may lead farmers of our country towards profit [6]. Maintaining the crop yield production with good quality is a key challenge for the farmers while they have limited resources and environmental constraints. The productivity of crops can be predicted by using the machine learning technology and it is a best solution to enhance the crops productivity. A machine learning model can be developed for crop yield prediction analysis. These Ml models determine the response of crops towards the factors which are responsible for good crop production. Different ML models have been used for crop yield prediction [4].Crop yield prediction analysis requires a model of how crops respond to soil factors. Various data mining techniques have been used for crop yield prediction. An expert system [7] has been developed for the various agricultural tasks. Many computational intelligence techniques have been applied in the field of agriculture [8]. Qualitative and quantities data is handled by neural network [8]. Data mining techniques [9] have been applied in agriculture data set for wheat

https://doi.org/10.1016/j.matpr.2021.01.948
2214-7853/© 2021 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the International Conference on Technological Advancements in Materials Science and Manufacturing.

yield prediction. Some data models have been proposed to achieve accuracy [10] in yield prediction. KNN and SVM [10] is used to predict yield production. Soil is one of the important factors in crop production. In this regard data mining techniques are used in soil to identify the relationship among soils using clustering in WEKA tool [11]. Gideon O Adeoye [12] proposed model to emphasize on physical factors of soil, soil nutrients to predict maize crops. Naïve Bayes, K-Nearest Neighbour [12] are used to generate classification rules through which Crop prediction can be identified. This paper [13] describes the different regression methods for yield data set. A comparative study was made for different algorithm. This paper [14] proposed a prediction model for crop using k-means. Crop production is also depends on soil fertility [15]. Soil fertility is affected by many factors like air, water, organic matter and nutrients. Due to some factors, like use of fertilizers, pesticides, insecticides, cultivation etc., and the fertility of soil are regularly deceasing in recent times [15,16].

1.1. Methodology

In this research study machine learning approaches are used to make crop yield prediction. In our research work data from FAO and world data bank are used. We have used four machine learning algorithms. These algorithms are also compared to achieve most accurate crop prediction.

1.2. Data set

FAOSTAT (Food and Agriculture Organization of the United Nations) provides data related to food and agriculture. This repository contains data of two hundred countries. The final data set has the following input fields – Item collected, country, Yield, rain, pesticides and temperature. Crops yield of four most cultivated crops in India is collected from FAO data repository. The collected data country, item, year starting from 1990 to 2016 and yield value for these years. The climate factors which affect crop production include rain fall and temperature. Environmental factors such as pesticides and soil also affect crop growth. The information about rain fall per year in India is collected from world data bank. Pesticides used for each item is collected from FAO data repository. Average temperature for country India is gathered from world data bank repository.

1.3. Preprocessing

Once data is gathered from the repositories which is publically available. The very first task is data cleaning. Once data cleaning process is completed then data frames can be merged together based on common columns. Normalization is also required to maintain a common scale for all the attributes. The final data fame is expected to have item (crop) country, year, yield value, average rainfall, pesticides and average temperature as the final features. We have applied different machine learning algorithms for prediction. We have also compared these algorithms to deliver the best results. The following Machine Learning models are used in our study to predict the crop yield. A comparison is also made between these models.

- Gradient Boosting Regressor
- Random Forest Regressor
- SVM
- Decision Tree Regressor

We have developed the above models in python using the different libraries such as Pandas, Numpy, sklearn, Seaborn, matplotlib, OneHotEncoder, sklearn and Pre-processing by (MinMaxScaler).

After cleaning and exploring the relationship between the features, the final data frame that contains all the features that will be used for the prediction process can be seen below in the table (Table 1):

The data set has the final features as given:

- Area: country of production.
- Item: type of crop.
- Year: year of production.
- Hg/ha_yield: country's yearly production of the crop that year.
- Average_rain_fall_mm_per_year: Average amount of rain recorded that year.
- Pesticides_tonnes: Amount of pesticides used on the crop that year.
- Avg_temp: Average temperature recorded for that year.

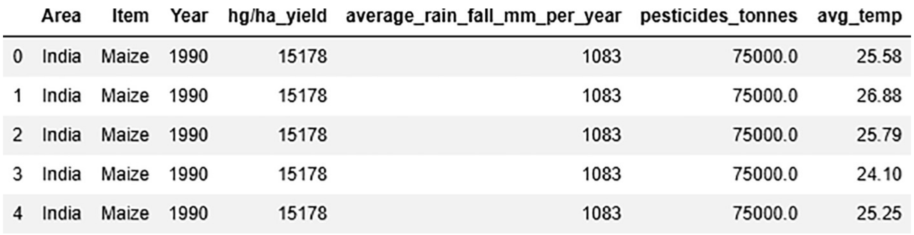
The final data frame is obtained by joining four different data frames from FAO and World Data bank to collect all required features. Correlation matrix is developed to demonstrate the relationship between the different features. The correlation can be visualized by correlation matrix as heat map. FF

It is now clear from the above figure that there is no relationship or correlation between variables. All features are independent. In our data frame two columns, items and countries have the categorical data values that contain label value rather than numeric value. Some machine learning algorithms require all the input and output variables as numeric instead of categorical values as these models cannot directly work on labeled data. A conversion of categorical data into numeric values is required. This conversion can be possible by one hot encoding process. In this process categorical data are converted into a form that could be provided to ML algorithms for getting the better prediction. One hot encoding is applied in our data frame to convert items and country columns into numeric array. This encoding creates a binary column for each category and returns a matrix with the results.

The preprocessed data is divided into two parts (training and testing) with the split factor 0.7. It means we have 70% data is for training and 30% data for the testing. The training data set is a primary data set that is used to train ML algorithms to learn for producing the exact predictions. However the test data is used to compile how well ML algorithm is trained with the training data set (Figs. 1-3).

Table 1
Sample of Final Data Set after cleaning.

Area	Item	Year	Hg/ha_yield	Average_rain_fall_mm_per_year	Pesticides_tonnes	Avg_temp
India	Maize	1990	15,178	1083	75000.0	25.58
India	Maize	1990	15,178	1083	75000.0	25.58
India	Maize	1990	15,178	1083	75000.0	25.79
India	Maize	1990	15,178	1083	75000.0	24.10
India	Maize	1990	15,178	1083	75000.0	25.25



	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
0	India	Maize	1990	15178	1083	75000.0	25.58
1	India	Maize	1990	15178	1083	75000.0	26.88
2	India	Maize	1990	15178	1083	75000.0	25.79
3	India	Maize	1990	15178	1083	75000.0	24.10
4	India	Maize	1990	15178	1083	75000.0	25.25

Fig. 1. Correlation Matrix as Heat Map.

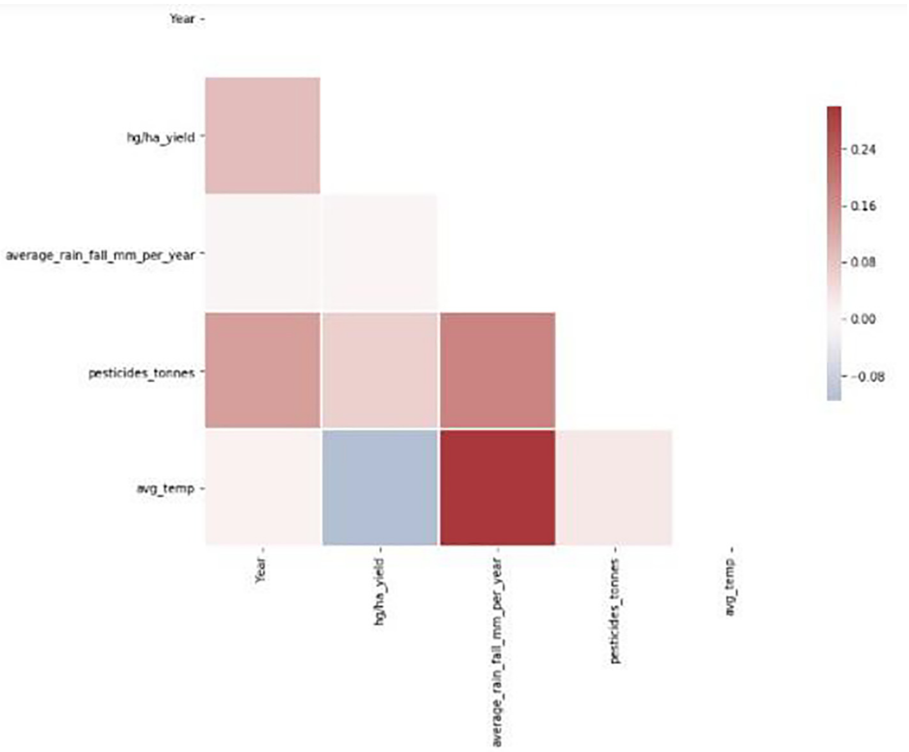
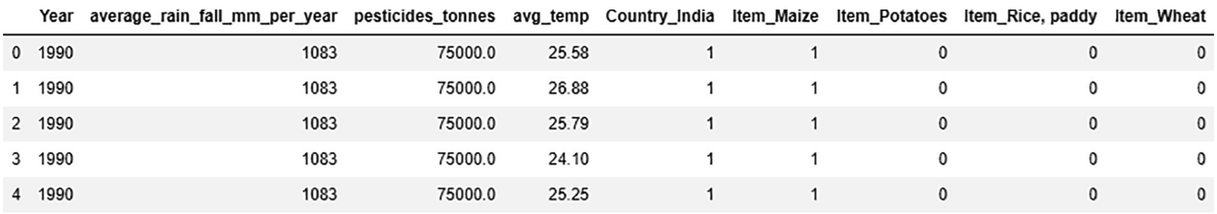


Fig. 2. Fame after one hot encoding.



	Year	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp	Country_India	Item_Maize	Item_Potatoes	Item_Rice, paddy	Item_Wheat
0	1990	1083	75000.0	25.58	1	1	0	0	0
1	1990	1083	75000.0	26.88	1	1	0	0	0
2	1990	1083	75000.0	25.79	1	1	0	0	0
3	1990	1083	75000.0	24.10	1	1	0	0	0
4	1990	1083	75000.0	25.25	1	1	0	0	0

Fig. 3. R^2 Score Value of Different Models.

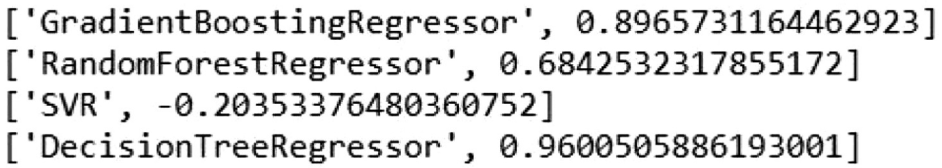
2. Experiments and results

In this research study we have trained four regressors models. Before deciding which model is best for crop yield prediction we need to evaluate each model. After the evaluation we compare and choose the best model that fits in our data set. In this work we have tried different models and approaches to solve the optimization problem then choose the best model which is suitable to fit that is not over fit and under fit. We have used Gradient Boosting Regressor, Random Forest Regressor, SVM and Decision Tree Regressor. These models can be compared through the Rooted Square value. The evolution matrix for the above four models is calculated based on R^2 (coefficient of determination) regression score function. The proportion of the variance for items column (crops) in the model can be represented by this coefficient. The R^2 score determines how data points fit in a curve or in line.

It is clear from the above results that decision tree Regressor has the high R^2 score i.e. 96%.

3. Model results and conclusions

Basically The R- Squared interprets that how well the models fit the observed data. For example R- Squared of 90% determines that 90% of data can be fit in the regression model. Higher R squared value defines better fitting for the model. From the above results it can be said that the decision tree Regressor fits the observed data



```
['GradientBoostingRegressor', 0.8965731164462923]
['RandomForestRegressor', 0.6842532317855172]
['SVR', -0.20353376480360752]
['DecisionTreeRegressor', 0.9600505886193001]
```

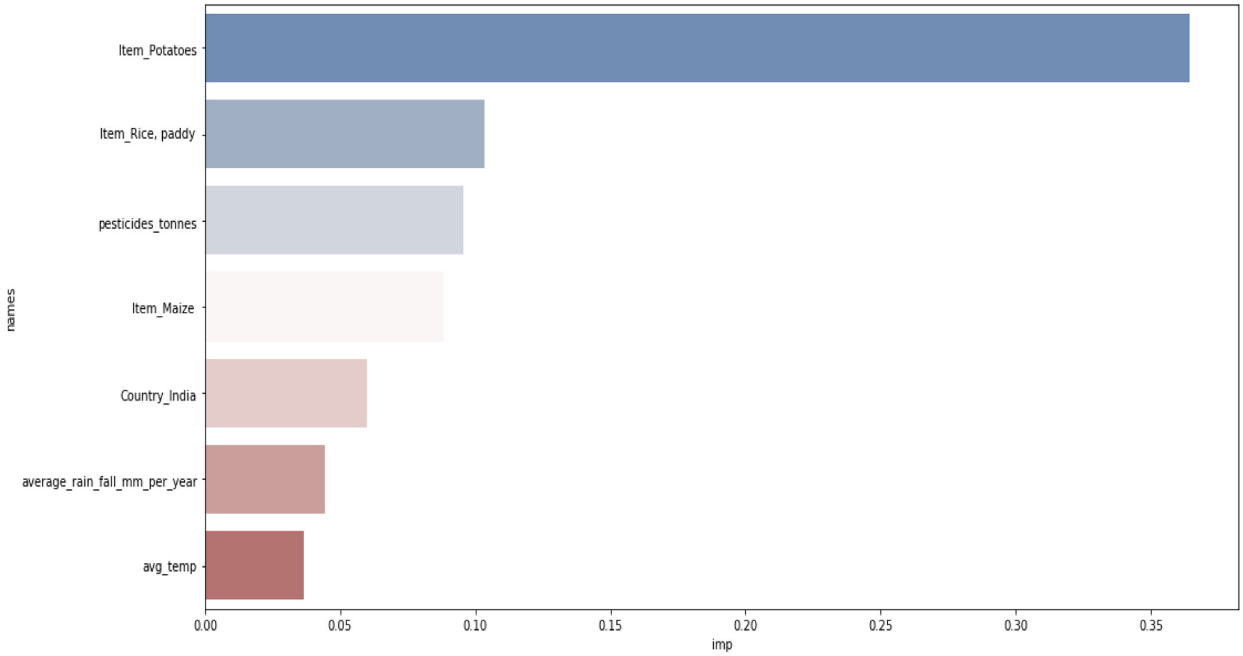
Fig. 4. Feature Importance.

to the highest score of 96%. The importance of a feature is obtained by the probability of that node Feature importance [19,20] is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature [15,17,18].We have obtained seven important features for the decision tree model.

In the Fig. 4 it is mentioned that seven features have the good importance and crop potatoes has the maximum importance to make decisions for a model. Rice paddy has good importance in the observed data set and obtained second highest importance. After rice pesticides has good effect and it is third highest important feature. After that maize, rainfall and temperature having the importance in sequence way.

The box plot is implemented for each crop yield. It is clear now that potatoes are highest production in India as per the predictive model. Then rice, wheat and maize have the production score.

Decision tree [18] Regressor gives the best accuracy for crop prediction as compare to other algorithms we have used for predic-



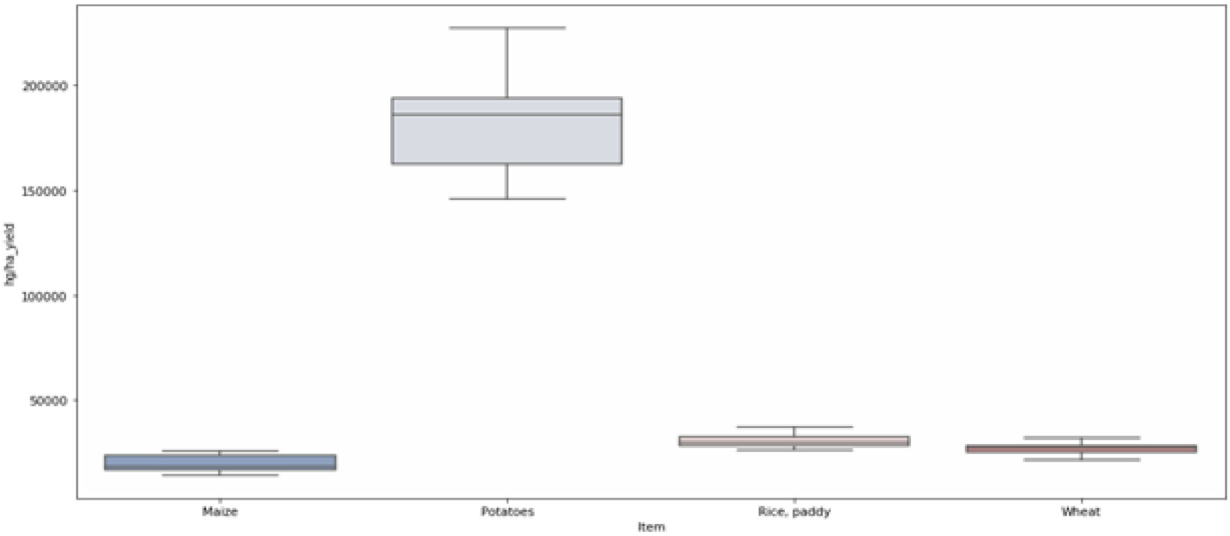


Fig. 5. Box Plot for each Crop Yield.

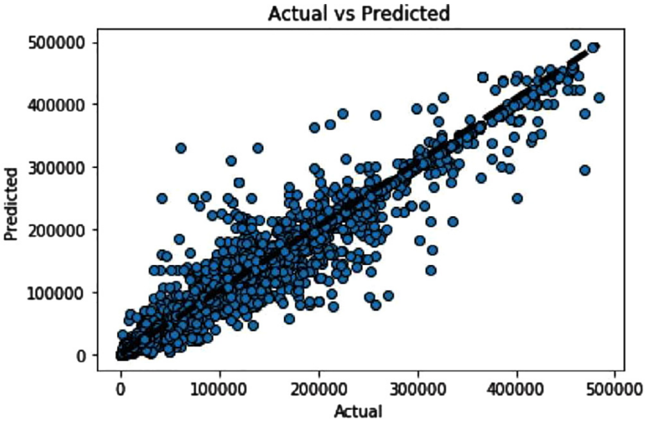


Fig. 6. Visualization of prediction.

tion. In our study Decision tree obtained 96% score as accuracy to predict crop. It is clear from the results that potatoes is the more suitable crop in India for production as it is predicted with the highest importance. The feature with the highest importance is become root node of the decision tree. In our data set potatoes item is the root node as it has maximum importance. The figure below determines the prediction which is visualized as a straight line. It shows the goodness of the model which is neither over fit nor under fit. Figs. 5 and 6 also determines the R square score that describes a good fitting model to make crops yield predictions for India.

4. Conclusion and future scope

In this research work different machine algorithms are used to predict crop yield in India. We have used the data set for making prediction for four primary crops such as potatoes, rice, wheat and maize. The decision tree Regressor achieves highest accuracy to predict crop yield. Out of four crops, which are mentioned above, the prediction score of potatoes is excellent. . The model's predictions can be enhanced in future by adding some more relevant features.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Horie, M. Yajima, H. Nakagawa, Yield forecasting, Agric. Syst. 40 (1992) 211–236, https://doi.org/10.1016/0308-521X(92)90022-G.

[2] S. Khaki, L. Wang, Crop Yield Prediction Using Deep Neural Networks, Front. Plant Sci. 10 (2019) 621, https://doi.org/10.3389/fpls.2019.00621.

[3] V. Lamba, V.S. Dhaka, Wheat yield prediction using artificial neural network and crop prediction echniques (A Survey), Int. J. Res. Appl. Sci. Eng. Technology 2 (2014) 330–341.

[4] S. Manimekalai, K. Nandhini, Crop Yield Prediction from Soil Parameters through Neupper Rule Established Algorithm, Int. J. Eng. Technol. 7 (3.34) (2018) 908–912.

[5] D.S. Zingade, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar, Chandan Mehta," Crop Prediction ystem using Machine Learning, Int. J. Adv. Eng. Res. Dev. Specl. Issue Recent Trend. Data Eng. 4, (5), Dec., 2017.

[6] https://en.wikipedia.org/wiki/Agriculture.

[7] J.P., Rao, Expert System in Agriculture, http://www.manage.gov.in/managelib/faculty/PanduRanga.htm, 1992.

[8] A. Schultz, R. Wieland, G. Lutze, Neural networks in agroecological modelling – stylish application or helpful tool?, Comput Electr. Agric 29 (2000) 73–97.

[9] Georg Rub, Rudolf Kruse, Martin Schneider and Peter Wagner, "Data Mining with Neural Networks for Wheat Yield Prediction", Research and Development in Intelligent Systems XXVI, Incorporating Applications and Innovations in Intelligent Systems XVII, Peterhouse College, Cambridge, UK, 15-17 December 2000.

[10] D. Ramesh, B Vishnu Vardhan, Data mining technique and applications to agriculture yield data, Int. J. Adv. Res. Comput. Commun. Eng. 2 (9) (2013).

[11] Shweta Taneja, Rashmi Arora, Savneet Kaur, Mining of Soil Data Using Unsupervised Learning Technique, Int. J. Appl. Eng. Res. 7 (11) (2012), ISSN 0973-4562.

[12] M.R. Bendre, R.C. Thool, V.R. Thool, "Big Data in Precision agriculture",Sept, 2015 NGCT.

[13] N.Suma, Sandra Rhea Samson, S.Saranya, G.Shanmugapriya, R.Subhashri, IOT Based Smart Agriculture Monitoring System, Feb 2017 IJRITCC.

[14] N.Heemageetha, A survey on Application of Data Mining Techniques to Analyze the soil for agricultural purpose, 2016 IEEE.

[15] Janmejay pant, Pushpa Pant, Ashutosh Bhatt, Himanshu Pant, Nirmal Pandey, Feature Selection towards Soil Classification in the context of Fertility classes using Machine Learning, Int. J. Innovat. Technol. Explor. Eng. (IJITEE) ISSN: 2278-3075, 8 12, 2019.

[16] S.J. Reashma, A.S. Pillai, ‖Edaphic factors and crop growth using machine learning—A review, International conference on intelligent sustainable systems (ICISS), IEEE 2017 (2017) 270–274.

[17] Towardsdatascience.com.

[18] J.R.OtukeI, T.Blaschke ,"Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms", Int. J. Appl. Eart Observat. Geoinformat., Vol. 12, (1), 2010, S27-S31.

[19] Neelam Singh, Neha Garg, Janmejay Pant, Document clustering using feature selection based on multiviewpoint and link similarity measure, Int. J. Comput. Appl. 5 (3) (2014) 1151–1155.

[20] Janmejay Pant, Bhaskar Pant, Amit Juyal, Comparative Study of Different Models before Feature Selection and AFTER Feature Selection for Intrusion Detection, Int. J. Comput. Appl., 98, (14), 2014.