International Journal of Modern Agriculture
ISSN: 2305-7246
Volume 10 Issue 2, 2021

Analysis and Prediction of Crop Yield using Data Mining and Machine Learning Algorithms

Jasper Varun T[1], Nagadithya Bathala[2], Chaitra Nayak J[3]

[1]*Electronics and Communication Engineering, REVA University, (INDIA)*

[2]*Electronics and Communication Engineering, REVA University, (INDIA)*

[3]*Electronics and Communication Engineering, REVA University, (INDIA)*

ABSTRACT

Agriculture plays a vital role in India's economy and majority of India's population is dependent on agriculture but farmer's still face a lot of challenges and issues both in terms of productivity and cost management due to the varying climatic conditions and lack of modern agricultural knowledge. In this work we have built a simulation model to predict the yield of crops in each district of Karnataka, by considering various parameters like rainfall, season, crop year, district and area of the land in which the crop is to be grown. WEKA Tool is used for data pre-processing. Python is used for data analysis and jupyter notebook is used as a tool for applying the machine learning algorithms for predicting the crop yield. The algorithm that performed best in simulation was later implemented using python in jupyter notebook
.

Keywords: Data Preprocessing, Aggregation, Dimensionality Reduction, Feature subset selection, Random forest regression.

1. INTRODUCTION

Farmers getting to know the yield rate of crops well before cultivation has a lot of benefits such as they can use their land size effectively, calculate their profits, avoid losses and can plan their expenditure on the crops effectively [2]. Thus it overall improves the agricultural technique in India.

Data mining is the process of retrieving required information from a large dataset. Data mining process includes the following stages such as Extracting, Transforming loading data in a repository and Data management [4]. Data collection is the first stage of data mining in which the required data is collected from various resources and arranged in a systematic order. Later the collected dataset is pre-processed to get the required information, the following pre-processing techniques [1] are used: Feature subset selection, Dimensionality reduction, aggregation and data transformation.

Feature subset selection is the process of selecting relevant and required features according to our objectives from the available dataset. After performing this process, we make use of dimensionality reduction, in whichthe unwanted and irrelevant tuples and columns are deleted to reduce the size of the dataset, therefore analyzing a small datasetis less time consuming and efficient. After dimensionality reduction, Aggregation is performed. It is the process of combining two or many features into one. By doing so the variability of the data is reduced. After aggregation the final pre-processing

technique used is Data Transformation. It is the process of converting data from one type to another, for example converting categorical data type into continuous data type.

The processed dataset can now be used for applying Machine Learning algorithms. In our work, WEKA Tool is used for data pre-processing and to obtain the simulation results of Machine Learning algorithms. Since, the target variable is of continuous type, regression algorithms must be used for prediction [1]. Out of many algorithms available, the most popular one's for this type of problem are chosen, such as Linear Regression, K-Nearest Neighbor(KNN), Decision Tree and Random Forest[4]. These following algorithms where simulated using WEKA Tool and the results were obtained.

2. RELATED WORK

In [1] they have proposed a machine learning approach which aims at predicting the best yielded crop for a particular region by analyzing various atmospheric factors like rainfall, temperature, humidity etc., and land factors like soil pH, soil type including past records of crops grown and they finally conclude by saying Random forest gives the better yield prediction as compared to Polynomial Regression and Decision Tree. In [2] different supervised techniques with hybrid approach have been analyzed and it can suggest farmers with the right crop to be grown according to the conditions and they conclude by saying that hybrid machine learning algorithm works better than the existing supervised classification techniques.

In [3] they focus on predicting the yield of the crop by applying various machine learning techniques and they finally conclude that random forest regressor algorithm has the highest yield prediction accuracy. In [4] they have built a system to analyze large dataset and to predict the crop yield, in this work they have concluded saying Random Forest algorithm out performs Multiple Linear Regression algorithm in every performance statistics and Random forest is a efficient machine learning algorithm for crop prediction.

3. PROBLEM DEFINITION

With the increase in food consumption due to rapid raise in population, it is very important for farmers to produce crops with high yield without affecting their profit percentage, thus our work is to design, develop and implement a simulation model to predict the yield of crops with the help of suitable machine learning algorithms using different parameters such as temperature, rainfall and location. Using which the yield of a particular crop in a particular region can be known well in advance and the appropriate crop can be grown and harvested.

4. PROPOSED METHODOLOGY

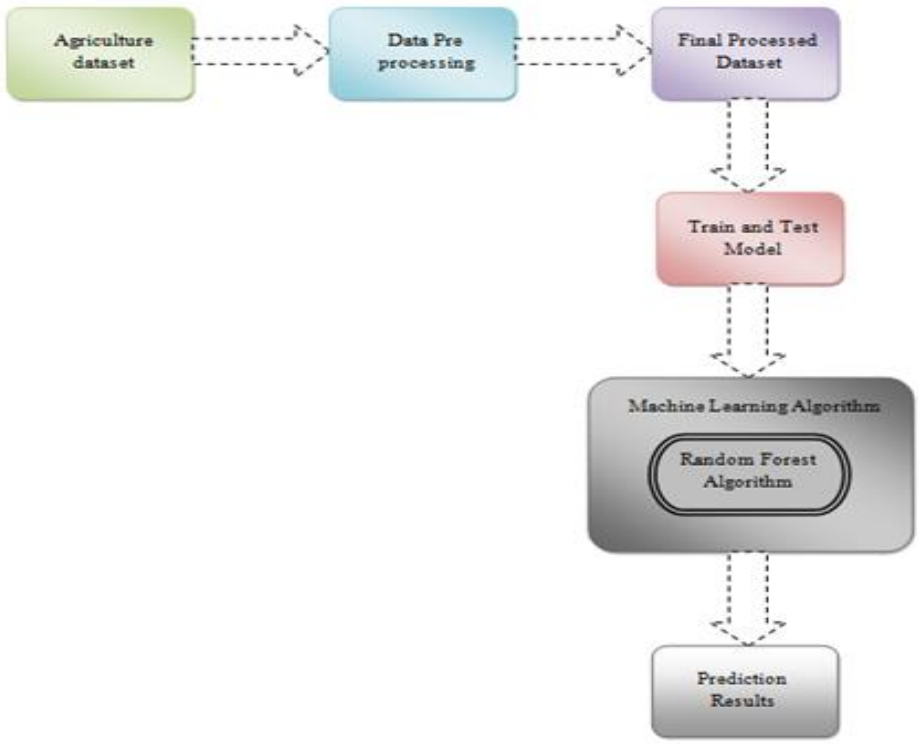

Fig 1: Block diagram of crop yield prediction model.

4.1 Agriculture Dataset

The first stage is collection of data, it is one of the primary needs of a machine learning project. In data collection , data is collected from various resources according to the project requirements and it is later preprocessed to obtain the final dataset on which the machine learning algorithm[3] can be applied.

For this work we have collected two separate datasets, the first dataset contains the crop yield along with different features such as district name, season name, area and crop name, the total number of tuples present in the dataset are 2,46,792 with 7 attributes, the second dataset contains the information regarding the rainfall in each district for the past 10 years and the total number of tuples present in this dataset are 190 along with 15 attributes. The crop dataset and rain dataset was downloaded from different data sources, a few tuples of both the dataset are shown below.

State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Andaman and Nicobar	NICOBARS	2009	Kharif	Arecanut	1254	2000
Andaman and Nicobar	NICOBARS	2009	Kharif	Other Kha	2	1
Andaman and Nicobar	NICOBARS	2009	Kharif	Rice	102	321
Andaman and Nicobar	NICOBARS	2009	Whole Ye	Banana	176	641
Andaman and Nicobar	NICOBARS	2009	Whole Ye	Cashewnu	720	165
Andaman and Nicobar	NICOBARS	2009	Whole Ye	Coconut	18168	65100000
Andaman and Nicobar	NICOBARS	2009	Whole Ye	Dry ginger	36	100
Andaman and Nicobar	NICOBARS	2009	Whole Ye	Sugarcane	1	2
Andaman and Nicobar	NICOBARS	2009	Whole Ye	Sweet pot	5	15
Andaman and Nicobar	NICOBARS	2009	Whole Ye	Tapioca	40	169
Andaman and Nicobar	NICOBARS	2010	Kharif	Arecanut	1254	2061
Andaman and Nicobar	NICOBARS	2010	Kharif	Other Kha	2	1

Fig 2. Crop dataset

State	District	Year	January	February	March	April	May	June	July	August	Septembe	October	Novembe	Decembe	Annual Total
Karnataka	Bagalkote	2013	1.8	0	0	11.6	103.9	123.9	75.2	32.8	142.9	57.4	0.6	0	550.1
Karnataka	Bagalkote	2014	0.8	0.2	2.5	44.7	49.7	68.1	117.6	71.1	76.9	124.3	0.1	0	556
Karnataka	Bagalkote	2015	0	0	9.1	9.5	85.9	98.1	29.7	22.4	122.3	37.5	13.9	0	428.4
Karnataka	Bagalkote	2016	0	0	0.2	3	52.9	195.4	43.5	108.7	267.3	38.6	4.2	0.1	713.9
Karnataka	Bagalkote	2017	0	16.7	94.2	10.3	37.5	46.5	28.3	60.2	135.4	60.3	27.2	0.3	516.9
Karnataka	Bagalkote	2018	0	0	2.9	14.8	66.3	119	41.7	107.3	195.6	220.1	28.8	31.2	827.7
Karnataka	Bagalkote	2019	13.4	0.1	0	10.6	57.4	109.7	89.6	149.3	64.2	58.3	83.2	0.4	636.2
Karnataka	Bangalore	2013	5.1	12.9	16.7	65.7	234.3	52.5	205.4	62.1	203.3	210.2	17.5	0	1085.7
Karnataka	Bangalore	2014	0.7	3.1	9.4	91.9	122.9	219.1	129.1	163.5	161	324.5	97.4	8.9	1331.5
Karnataka	Bangalore	2015	0	0	62.7	11	72.5	96.6	37.2	31.1	45.6	75.6	79.7	1.7	513.7
Karnataka	Bangalore	2016	0	1.3	0	46.3	76.6	57.1	100.8	133.6	166	184.2	18.6	20.3	804.8

Fig 3. Rainfall dataset

4.2 Data Preprocessing

There are a number of data mining preprocessing techniques availabe for preparing the data for analysis purpose, some of the preprocessing techniques[2] which we have used are:

- Feature Subset Selection
- Dimensionality Reduction
- Aggregation
- Data Transformation

Feature Subset Selection:

Feature Subset Selection is the process of selecting a set of attributes or features which are relevant for the model creation.

In this paper, the objective is to predict the yield of different crops which are suitable for different districts and various seasons in the state of Karnataka, therefore out of the 33 states available in the dataset, we have selected only the state of Karnataka and by doing so the dimensionality of the dataset has been reduced significantly from two lakh forty six thousand to twenty one thousand. Further in the process only the three most popular crops from each district has been selected, further reducing the size of data to 2503. The sample dataset after performing feature subset selection is as shown below.

	State_Name	District_N	Crop_Year	Season	Area	Crop	Production
2	Karnataka	BAGALKO	2009	'Kharif '	37419	Maize	130280
3	Karnataka	BAGALKO	2009	'Kharif '	51307	Sunflowe	29116
4	Karnataka	BAGALKO	2009	'Rabi '	12094	Maize	42223
5	Karnataka	BAGALKO	2009	'Rabi '	62717	Sunflowe	35641
6	Karnataka	BAGALKO	2009	'Summer	1509	Maize	5691
7	Karnataka	BAGALKO	2009	'Summer	4087	Sunflowe	3052
8	Karnataka	BAGALKO	2009	'Whole Ye	6986	Onion	65436
9	Karnataka	BAGALKO	2010	'Kharif '	38948	Maize	153176
10	Karnataka	BAGALKO	2010	'Kharif '	6732	Onion	39517

Fig 4. Dataset after performing feature subset selection

Dimensionality Reduction:

Dimensionality Reduction is the process of deleting irrelevant or less important attributes in the dataset before model creation.

In this dataset, we have deleted the attribute 'State Name' as it has less importance during model creation, the dataset after performing dimensionality reduction is as shown below.

District_N	Crop_Year	Season	Crop	Area	Productio
BAGALKOT	2009	Kharif '	Maize	37419	130280
BAGALKOT	2009	Kharif '	Sunflower	51307	29116
BAGALKOT	2009	Rabi '	Maize	12094	42223
BAGALKOT	2009	Rabi '	Sunflower	62717	35641
BAGALKOT	2009	Summer	Maize	1509	5691
BAGALKOT	2009	Summer	Sunflower	4087	3052
BAGALKOT	2009	Whole Yea	Onion	6986	65436
BAGALKOT	2010	Kharif '	Maize	38948	153176
BAGALKOT	2010	Kharif '	Onion	6732	39517
BAGALKOT	2010	Kharif '	Sunflower	48722	26039
BAGALKOT	2010	Rabi '	Maize	14743	61920

	District	Year	January	February	March	April	May	June	July	August	Septembe	October	Novembe	Decembe	Annual Total
2	Bagalkote	2013	1.8	0	0	11.6	103.9	123.9	75.2	32.8	142.9	57.4	0.6	0	550.1
3	Bagalkote	2014	0.8	0.2	2.5	44.7	49.7	68.1	117.6	71.1	76.9	124.3	0.1	0	556
4	Bagalkote	2015	0	0	9.1	9.5	85.9	98.1	29.7	22.4	122.3	37.5	13.9	0	428.4
5	Bagalkote	2016	0	0	0.2	3	52.9	195.4	43.5	108.7	267.3	38.6	4.2	0.1	713.9
6	Bagalkote	2017	0	16.7	94.2	10.3	37.5	46.5	28.3	60.2	135.4	60.3	27.2	0.3	516.9
7	Bagalkote	2018	0	0	2.9	14.8	66.3	119	41.7	107.3	195.6	220.1	28.8	31.2	827.7
8	Bagalkote	2019	13.4	0.1	0	10.6	57.4	109.7	89.6	149.3	64.2	58.3	83.2	0.4	636.2
9	Bangalore Rur	2013	5.1	12.9	16.7	65.7	234.3	52.5	205.4	62.1	203.3	210.2	17.5	0	1085.7
10	Bangalore Rur	2014	0.7	3.1	9.4	91.9	122.9	219.1	129.1	163.5	161	324.5	97.4	8.9	1331.5

Fig 5. Dataset after performing dimensionality reduction

Aggregation:

Aggregation is the process of combining two or more columns or two different datasets from different data sources to get a summarized dataset.

For the above dataset after performing dimensionality reduction, we have aggregated the season wise rainfall and then added the rainfall attribute to crop dataset, the dataset obtained after performing aggregation is as shown below.

	District	Year	Kharif_rainfall	Rabi_rainfall	summer_rainf	WholeYear_rainfall
2	Bagalkote	2013	77.075	0.6	59.85	45.84167
3	Bagalkote	2014	97.475	0.275	41.25	46.33333
4	Bagalkote	2015	52.975	3.475	50.65	35.7
5	Bagalkote	2016	114.525	1.075	62.875	59.49167
6	Bagalkote	2017	71.05	11.05	47.125	43.075
7	Bagalkote	2018	141.175	15	50.75	68.975
8	Bagalkote	2019	90.35	24.275	44.425	53.01667
9	'Bangalore	2013	170.25	8.875	92.3	90.475
10	'Bangalore	2014	194.525	27.525	110.825	110.9583

	District_N	Crop_Year	Season	Rainfall	Area	Crop	Production
2	BAGALKOT	2009	'Kharif '	92.08	37419	Maize	130280
3	BAGALKOT	2009	'Kharif '	92.08	51307	Sunflower	29116
4	BAGALKOT	2009	'Rabi '	7.96	12094	Maize	42223
5	BAGALKOT	2009	'Rabi '	7.96	62717	Sunflower	35641
6	BAGALKOT	2009	'Summer	50.98	1509	Maize	5691
7	BAGALKOT	2009	'Summer	50.98	4087	Sunflower	3052
8	BAGALKOT	2009	'Whole Ye	50.34	6986	Onion	65436
9	BAGALKOT	2010	'Kharif '	92.08	38948	Maize	153176
10	BAGALKOT	2010	'Kharif '	92.08	6732	Onion	39517

Fig 6. Dataset after performing aggregation

Data Transformation:

Data transformation is the process of converting data from one type into another type, it is the last step of data preprocessing

The purpose of data transformation is to transform categorical data types into numeric data type, so that the dataset is ready to be analyzed by regression algorithms. Data transformation was done using python in jupyter notebook with the help of sklearn preprocess library to convert the categorical data type into numeric data type, and the final dataset after performing all the preprocessing techniques is as shown below.

	District_Code	Crop_Year	Season_Code	Rainfall	Area	Crop_Code	Production
2	4	2009	0	92.08	37419	4	130280
3	4	2009	0	92.08	51307	8	29116
4	4	2009	1	7.96	12094	4	42223
5	4	2009	1	7.96	62717	8	35641
6	4	2009	2	50.98	1509	4	5691
7	4	2009	2	50.98	4087	8	3052
8	4	2009	3	50.34	6986	5	65436
9	4	2010	0	92.08	38948	4	153176
10	4	2010	0	92.08	6732	5	39517

Fig 7. Final dataset after applying preprocessing techniques

4.3 Machine Learning Algorithms

Machine learning algorithms are classified as

- Classification
- Regression
- Association
- Clustering

Since we need to use supervised learning type of algorithm and our target variable is of continuous type, we are using regression technique, out of many available algorithms[4] in regression, we simulated for Linear Regression, K-Nearest Neighbor (KNN), Decision tree and Random Forest algorithm using WEKA TOOL.On running the simulation, we found that Random Forest algorithm was a high accuracy rate compared to the other three and thus Random Forest algorithm was implemented in jupyter notebook using python and the results obtained are discussed below.

5. RESULTS AND ANALYSIS

5.1 Pre-processed dataset

The dataset collected was preprocessed during different data mining preprocessing techniques as mentioned above and the tool which we have used for preprocessing is WEKA tool and Jupyter notebook. The final preprocessed dataset is as shown below.

	District_Code	Crop_Year	Season_Code	Rainfall	Area	Crop_Code	Production
2	4	2009	0	92.08	37419	4	130280
3	4	2009	0	92.08	51307	8	29116
4	4	2009	1	7.96	12094	4	42223
5	4	2009	1	7.96	62717	8	35641
6	4	2009	2	50.98	1509	4	5691
7	4	2009	2	50.98	4087	8	3052
8	4	2009	3	50.34	6986	5	65436
9	4	2010	0	92.08	38948	4	153176
10	4	2010	0	92.08	6732	5	39517

Fig 8. Final preprocessed dataset

5.2 Simulation Results

After obtaining the final dataset, different machine learning algorithms such as, Linear Regression, K-Nearest Neighbor, Decision Tree, and Random Forest algorithm were applied on the dataset, the simulation was done using WEKA tool and the simulation results of each algorithm is as shown below.

LINEAR REGRESSION ALGORITHM

Fig 9. Graph of actual yield vs predicted yield using Linear Regression

Linear Regression algorithm was applied using WEKA tool and the accuracy and error rate obtained through simulations are as shown below.

```
Correlation coefficient                  0.8982
Mean absolute error                  20564.466
Root mean squared error              32356.352
Relative absolute error                 51.1483 %
Root relative squared error             44.528  %
Total Number of Instances              851
```

Fig 10. Simulation results of Linear Regression algorithm

K-NEAREST NEIGHBOR (KNN) ALGORITHM

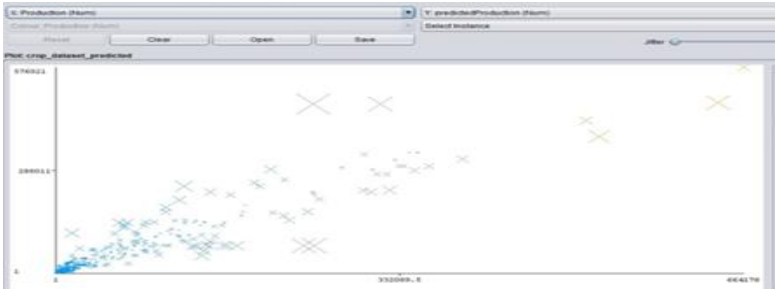

Fig 11. Graph of actual yield vs predicted yield using K -Nearest Neighbor

K-Nearest Neighbor algorithm was applied using WEKA tool and it is seen that it performs better than Linear Regression model, the accuracy and error rate obtained through simulations are as shown below.

Correlation coefficient	0.9466
Mean absolute error	9020.0029
Root mean squared error	23433.8294
Relative absolute error	22.4347 %
Root relative squared error	32.249 %
Total Number of Instances	851

Fig 12. Simulation results of K-Nearest Neighbor algorithm

DECISION TREE ALGORITHM

Fig 13. Tree view of Decision Tree model

Decision Tree algorithm was applied using WEKA tool and it is seen that it performs better than Linear Regression model and also K-Nearest Neighbor model, the accuracy and error rate obtained through simulations are as shown below.

Correlation coefficient	0.9694
Mean absolute error	7260.4893
Root mean squared error	18003.6237
Relative absolute error	18.0584 %
Root relative squared error	24.7761 %
Total Number of Instances	851

Fig 14. Simulation results of Decision Tree model

RANDOM FOREST ALGORITHM

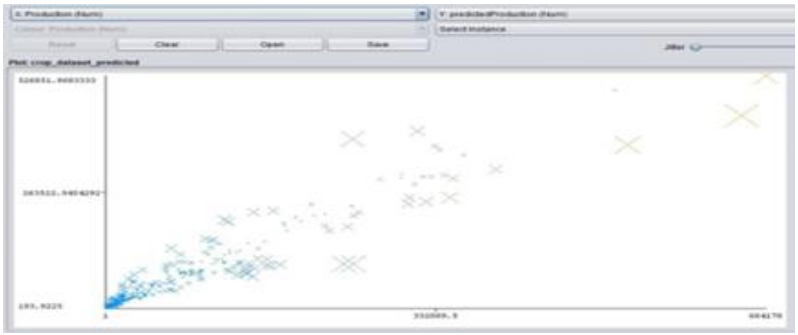

Fig 15. Graph of actual yield vs predicted yield using Random Forest algorithm

Random Forest algorithm was applied using WEKA tool and it is seen that it performs better than Linear Regression model, K-Nearest Neighbor modeland Decision Tree model, the accuracy and error rate obtained through simulations are as shown below.

```
Correlation coefficient                  0.9632
Mean absolute error                   7883.9046
Root mean squared error              20083.457
Relative absolute error                 19.609  %
Root relative squared error             27.6383 %
Total Number of Instances              851
```

Fig 16. Simulation results of Random Forest model

Since Random Forest Algorithm showed high accuracy compared to the other three, we have implemented Random Forest algorithm in jupyter notebook using python language and the output is as shown below.

```
regressor = RandomForestRegressor(n_estimators=20, random_state=0)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)

from sklearn import metrics

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 7.820569612590803
Mean Squared Error: 372.34043725802087
Root Mean Squared Error: 19.296124928545133
```

```
regressor.score(X,y)
```

```
0.9898192438948791
```

Fig 17. Performance measurements of Random Forest model using python

```
enter district code by referring to the table :

      District_Name District_Code
0   Bangalore_Rural             0
1   Bangalore_Urban             1
2    Dakshin_Kannad             2
3      Uttar_Kannad             3
4           Bagalkot             4
5            Belgaum             6
6            Bellary             7
7              Bidar             8


4
enter year :
2021
enter season code by referring to the table:

  Season_Name Season_Code
0      Kharif           0
1        Rabi           1
2      Summer           2
3  Whole Year           3

0
enter area of land in heactres :
245
enter crop code from below table:

   Crop_Name Crop_Code
0      Maize         4
1      Onion         5
2  Sunflower         8

4
Rainfall in mm : 107.09124999999997
Production in tonnes:[743.8]
```

Fig 18. Final output to end user in jupyter notebook

6. CONCLUSION

In the future a better user interface would be developed for the model so that anyone from a remote location can access it and additional attributes can be added to the dataset to get even more accurate predictions. Thus in this paper we have discussed about the data collection from various resources and how to preprocess the collected data using different preprocessing techniques, later using WEKA tool different regression algorithms were applied and the simulation results were analyzed, on analyzing the results, we can conclude that Random Forest algorithm has the highest accuracy rate compared to the other algorithms. With the help of the crop yield prediction model, a farmer can increase his profitability and reduce his losses.

REFERENCES

[1]Sangeeta and Shruthi G "Design And Implementation Of Crop Yield Prediction Model In Agriculture", International Journal Of Scientific & Technology Research Volume 8, Issue 01, January 2020.

[2]Ratnmala Nivrutti Bhimanpallewar And Manda Rama Narasingarao, "Alternative Approaches Of Machine Learning For Agriculture Advisory System", 10th International Conference On Cloud Computing, Data Science & Engineering (Confluence) IEEE 2020.

[3] PSA MEMORY OPTIMIZATION METHOD FOR HEARTBEAT AND EMG MONITORING FOR PROSTHETICS, Avnip Deora, Dr.Anil Kumar, Dr.Pawan Kumar, International Journal Of Advance Research In Science And Engineering http://www.ijarse.com IJARSE, Volume No. 10, Issue No. 01, January 2021 ISSN-2319-8354(E).

[4]Aruvansh Nigam, Saksham Garg, Archit Agrawal and Parul Agrawal "Crop Yield Prediction Using Machine Learning Algorithms", 2019 Fifth International Conference on Image Information Processing (ICIIP).

[5]R.Karthikeyan, M.Gowthami, A.Abhishhek and P.Karthikeyan "Implementation Of Effective Crop Selection By Using The Random Forest Algorithm", International Journal Of Engineering & Technology, 7 (3.34) (2018) 287-290.