Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/13330
DOI URL: http://dx.doi.org/10.21474/IJAR01/13330

RESEARCH ARTICLE

AGRICULTURAL DATA ANALYSIS

Shobana S.[1] and Dr. M. Sujithra[2]

1. Msc Data Science, Department of Data ScienceFebruary 2020Coimbatore Institute of TechnologyCoimbatore.
2. Assistant ProfessorDepartment of Data ScienceCoimbatore Institute of TechnologyCoimbatore.

Manuscript Info

Manuscript History
Received: 29 June 2021
Final Accepted: 30 July 2021
Published: August 2021

Key words:-
Data Mining, Accomplishing, Agriculture, Environment, Variability

Abstract

In agriculture sector where farmers and agribusinesses have to make innumerable decisions every day and intricate complexities involves the various factors influencing them. An essential issue for agricultural planning intention is the accurate yield estimation for the numerous crops involved in the planning. Data mining techniques are necessary approach for accomplishing practical and effective solutions for this problem. Agriculture has been an obvious target for big data. Environmental conditions, variability in soil, input levels, combinations and commodity prices have made it all the more relevant for farmers to use information and get help to make critical farming decisions. This paper focuses on the analysis of the agriculture data and finding optimal parameters to maximize the crop production using Machine learning techniques like random forest regressor and Linear Regression. Mining the large amount of existing crop, soil and climatic data, and analysing new, non-experimental data optimizes the production and makes agriculture more resilient to climatic change.

Introduction:-

The project focuses on analysing agricultural system data. The data set consists of the crop yield and the crop details on monthly as well as yearly basis. This project is used to analyse the productivity parameters to solve the main problems faced by farmers. Then identify the bottlenecks, and the provides the best possible solutions to it. This project objective can be accomplished using spark, pyspark, MLlib. The output is generated in the form of models about good farming solutions. The outcome consists of the information like climate, growing crops on various factors like demand, change in production rate, future trend. This project is about creating the crop yield and crop details in order to help the farmers to overcome the problems faced by them. By making use of spark functions such as pyspark, MLlib, sql querying, and performing the structured data analysis using pyspark. Analysis part of the project involves analysing the structured dataset using pyspark, then using MLlib and sql queries the problems faced by the farmers will be taken and the solutions will be given in the form of models. Considering the seasons such as Kharif, Rabi and the crops that will give the production for the whole year. Then by considering the different crops such as Ragi, Rice, Safflower, Urad, Wheat, Arecanut, Banana, coriander, potato, sugarcane, sunflower, sweet potato, Tobacco, Turmeric, Arhar, bajra, castor seed, cotton, cowpea and dry chillies analysis is done for improving the production rate and reducing the problems faced by the farmers.

Corresponding Author:- Shobana S.
Address:- Msc Data Science, Department of Data ScienceFebruary 2020Coimbatore Institute of TechnologyCoimbatore.

Data Set Description

In this project the dataset consists of seven attributes such as state, district, crop year, season, crop, area and production. Also, this dataset consists of 246092 columns, which is a structured dataset in the form of csv file. The first step is that to the file by converting the csv file into pandas form and calculate the cost price index (CPI). The Full form of CPI is Consumer Price Index. A consumer price index (CPI) measures changes in the price level of market basket of consumer goods and services such as transportation, food and medical care etc. purchased by households. It is calculated by taking price changes for each item in the predetermined basket of goods and averaging them.Formula to calculate CPI is:

$$CPI_t = \frac{C_t}{C_0} * 100$$

Methodology:-

The following languages are used in this project for analysing the data.

1. Spark
2. Pyspark
3. MLlib

Spark

Apache Spark is a data processing framework that can quickly perform processing tasks on very large data sets, and can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools. These two qualities are key to the worlds of big data and machine learning, which require the marshalling of massive computing power to crunch through large data stores. Spark also takes some of the programming burdens of these tasks off the shoulders of developers with an easy-to-use API that abstracts away much of the grunt work of distributed computing and big data processing.

Pyspark

PySpark is a Python API for Spark released by the Apache Spark community to support Python with Spark. Using PySpark, one can easily integrate and work with RDDs in Python programming language too. There are numerous features that make PySpark such an amazing framework when it comes to working with huge datasets. Key Features of PySpark are Real-time computations: Because of the in-memory processing in the PySpark framework, it shows low latency. Polyglot: The PySpark framework is compatible with various languages such as Scala, Java, Python, and R, which makes it one of the most preferable frameworks for processing huge datasets. Caching and disk persistence: This framework provides powerful caching and great disk persistence. Fast processing: The PySpark framework is way faster than other traditional frameworks for Big Data processing. Works well with RDDs: Python programming language is dynamically typed, which helps when working with RDDs.

Spark Sql

Spark SQL allows you to use data frames in Python, Java, and Scala; read and write data in a variety of structured formats; and query Big Data with SQL. Spark SQL is Spark's interface for working with structured and semi-structured data. Structured data is considered any data that has a schema such as JSON, Hive Tables, Parquet. Schema means having a known set of fields for each record. Semi structured data is when there is no separation between the schema and the data.

Spark SQL provides three main capabilities for using structured and semi structured data:

1. It provides a Data Frame abstraction in Python, Java, and Scala to simplify working with structured datasets. Data Frames are similar to tables in a relational database.
2. It can read and write data in a variety of structured formats (e.g., JSON, Hive Tables, and Parquet).
3. It lets you query the data using SQL, both inside a Spark program and from external tools that connect to Spark SQL through standard database connectors (JDBC/ ODBC), such as business intelligence tools like Tableau.

Mllib

MLlib is Spark's machine learning (ML) library. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as:

1. ML Algorithms: common learning algorithms such as classification, regression, clustering, and collaborative filtering
2. Featurization: feature extraction, transformation, dimensionality reduction, and selection
3. Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
4. Persistence: saving and load algorithms, models, and Pipelines
5. Utilities: linear algebra, statistics, data handling, etc.

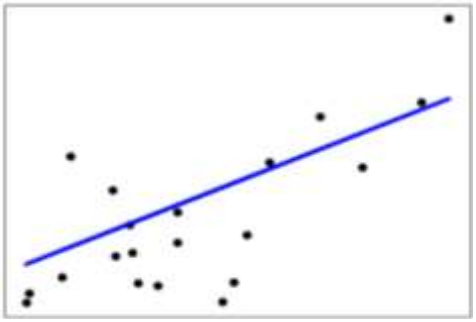
Machine learning algorithms

In this project the following machine learning algorithms are used:

1. Linear regression
2. Random forest regressor

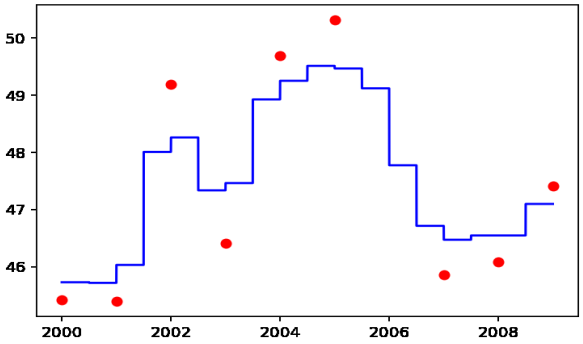
Linear regression

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). This is the graphical representation of linear regression.



Random Forest Regressor

Random forest is a bagging technique and not a boosting technique. The trees in random forest are run in parallel. There is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the **mode** of the or **classes (classification) mean prediction (regression)** of the individual trees. A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which **aggregates many decision trees**, with some helpful modifications: The number of features that can be split on at each node is limited to some percentage of the total (which is known as the **hyperparameter**). This ensures that the ensemble model **does not rely too heavily on any individual feature**, and makes **fair use of all potentially predictive features**. Each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents **overfitting**. The above modifications help prevent the trees from being too highly correlated. This is the graphical representation of random forest regression.

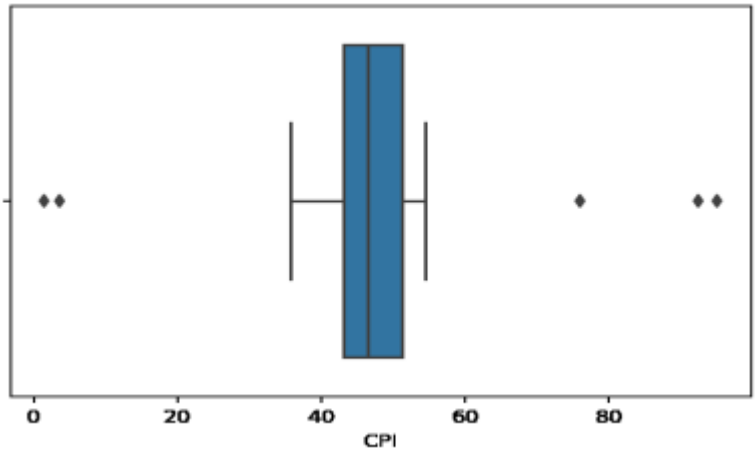


Results and Discussions:-

The following results are obtained in prediction of productivity of crops based on some of the factors such as season, crop year, crops, area for cultivation, production and consumer price index (CPI).

Box plot for cpi

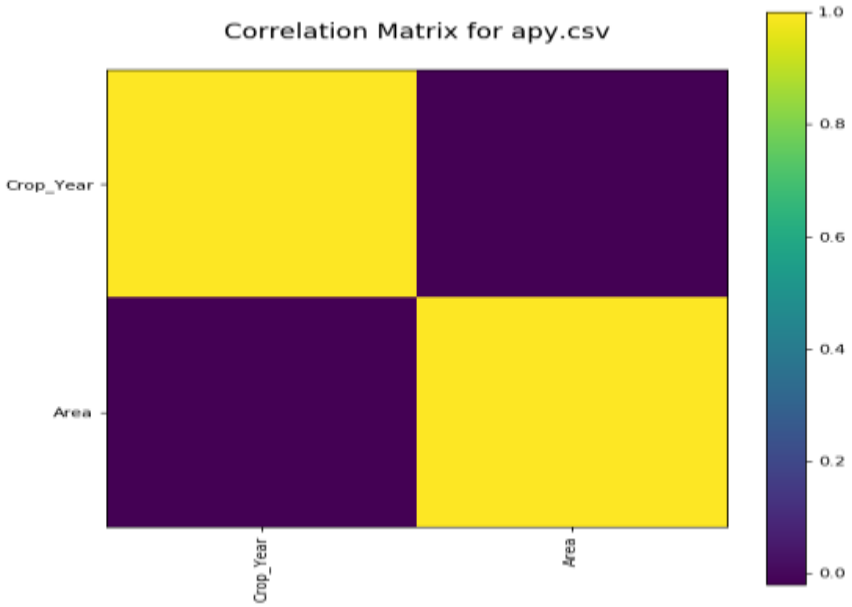
Boxplots are a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile(Q3), and "maximum"). **median (Q2/50th Percentile)**: the middle value of the dataset. **first quartile (Q1/25th Percentile):** the middle number between the smallest number (not the "minimum") and the median of the dataset. **Third quartile (Q3/75th Percentile):** the middle value between the median and the highest value (not the "maximum") of the dataset. **Interquartile range (IQR):** 25th to the 75th percentile. **Whiskers (shown in blue)outliers (shown as green circles)"maximum":** Q3 + 1.5*IQR**"minimum":** Q1 -1.5*IQR



This box plot is the representation of consumer price index, which is the important factor that is based on the production and area. Consumer price index is mainly based on transport, medical care, food etc. but in this project the consumer price index is based on area and production of crops.

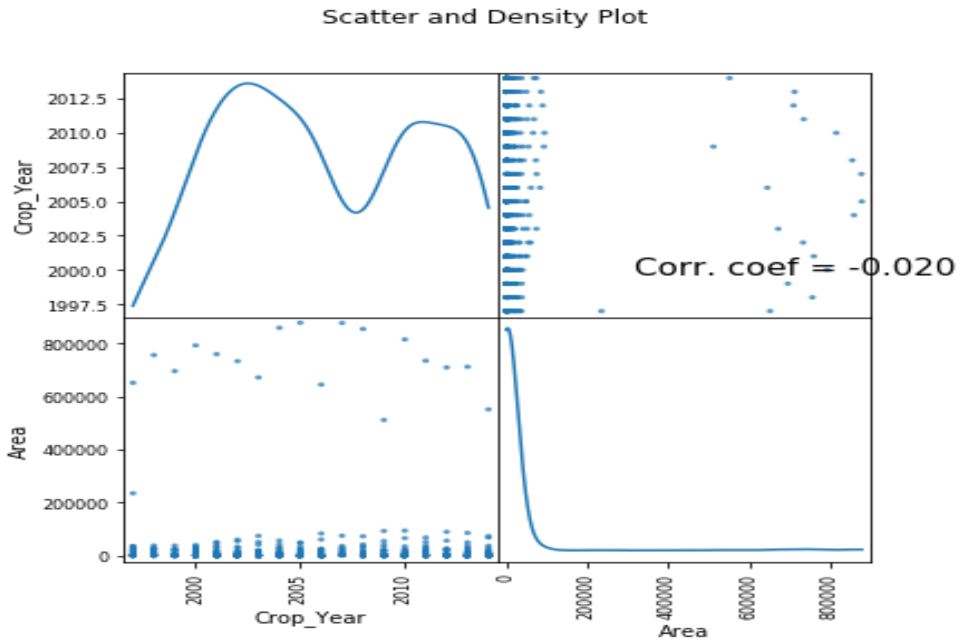
Correlation Matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.



Scatter And Density Plot

A density scatterplot; that is, a pattern of shaded squares representing the counts/percentages of the points falling in each square.



Summary Statistics

Summary statistics summarize and provide information about your sample data. It tells you something about the values in your data set. This includes where the average lies and whether your data is skewed. Summary statistics fall into three main categories:

1. Measures of location (also called central tendency).
2. Measures of spread.
3. Graphs/charts.

Measures of location

Measures of location gives the information about where the data is centred at, or where a trend lies. The following terms are some of the major measures of location.

1. Mean (also called the arithmetic mean or average).
2. Geometric mean (used for interest rates and other types of growth).
3. Trimmed Mean (the mean with outliers excluded).
4. Median (the middle of a data set).

Measures of spread

Measures of spread will give information about how the dataset is spread out and varied. The following terms are some of the major measures of spread.

1. Range (how spread out your data is).
2. Interquartile range (where the "middle fifty" percent of your data is).
3. Quartiles (boundaries for the lowest, middle and upper quarters of data.
4. Skewed (does your data have mainly low, or mainly high values?).
5. Kurtosis (a measure of how much data is in the tails).

Graphs or Charts

There are literally a greater number of ways to visualize the summary statistics graph. Some of the common ways to visualize the graphs are listed below.

1. Histogram.
2. Frequency Distribution Table.
3. Box plot.

4. Bar chart.
5. Scatter plot.
6. Pie chart.

This is the summary statistics for production

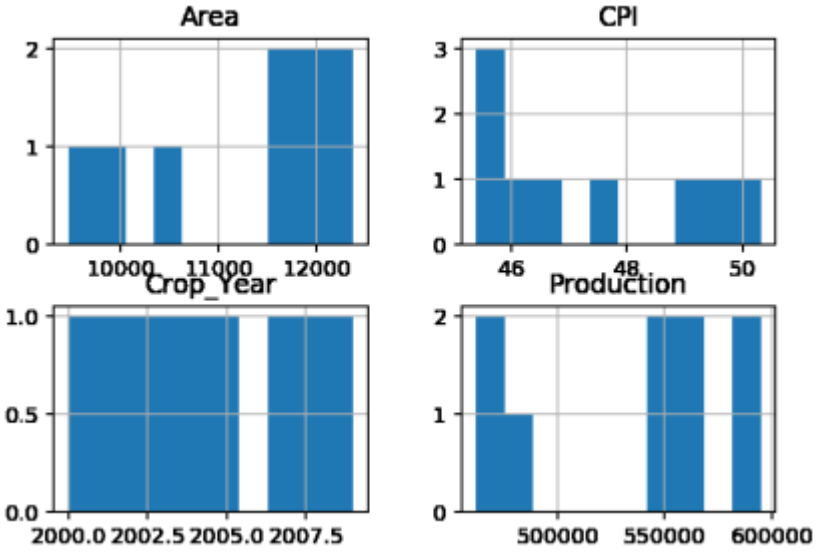
Mean	3553856.723
Standard Error	933537.9581
Median	376.5
Mode	1
Standard Deviation	13169173.5
Sample Variance	1.73427E+14
Kurtosis	15.45830801
Skewness	4.002035404
Range	71300000
Minimum	0
Maximum	71300000
Sum	707217487.9
Count	199

This is the summary statistics for crop year

Mean	2004.944724
Standard Error	0.23940903
Median	2005
Mode	2010
Standard Deviation	3.37727997
Sample Variance	11.40602
Kurtosis	-1.089821126
Skewness	0.258727718
Range	10
Minimum	2000
Maximum	2010
Sum	398984
Count	199

Histogram Comparison

In this project histogram comparison is done for visualizing the dependencies of the factors crop year, area, CPI and production. Because the mentioned factors are dependent of each other.



Grahpical Representation of Random Forest Regressor

The graphs mentioned here are graphical representation of random forest regressor. First graph i.e. figure1 represents the graph before prediction between two factors year and production.

Then the next graph i.e. figure 2 represents the graph after prediction between the above-mentioned factors and as the final representation i.e. figure 3 represents the comparison graph between the actual graph before and after prediction.

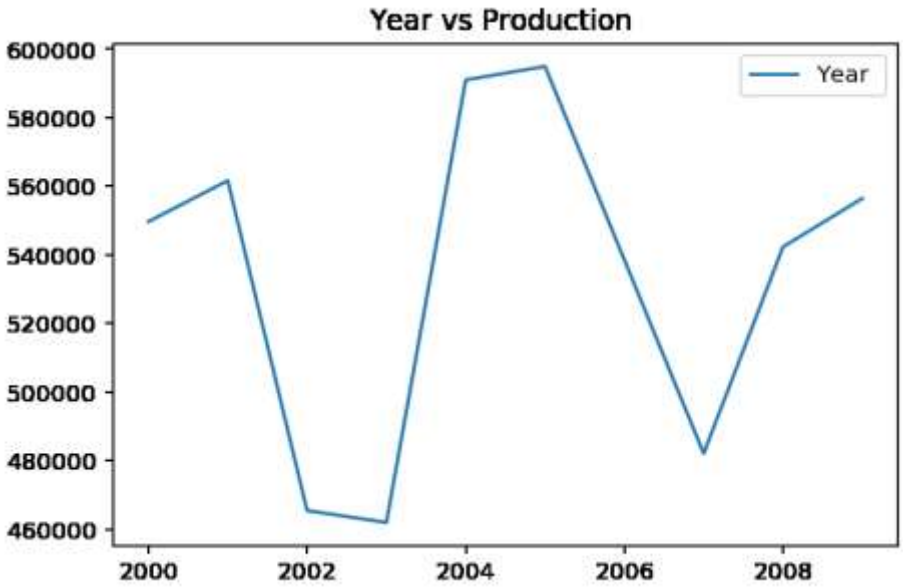


Figure1:-actual graph before prediction.

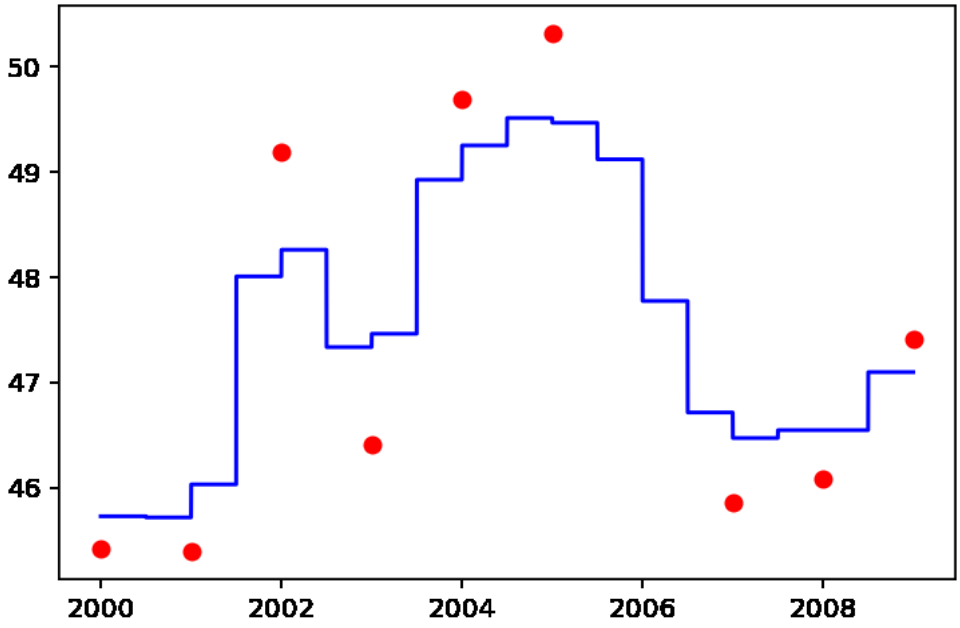


Figure2:-Actual graph after prediction.

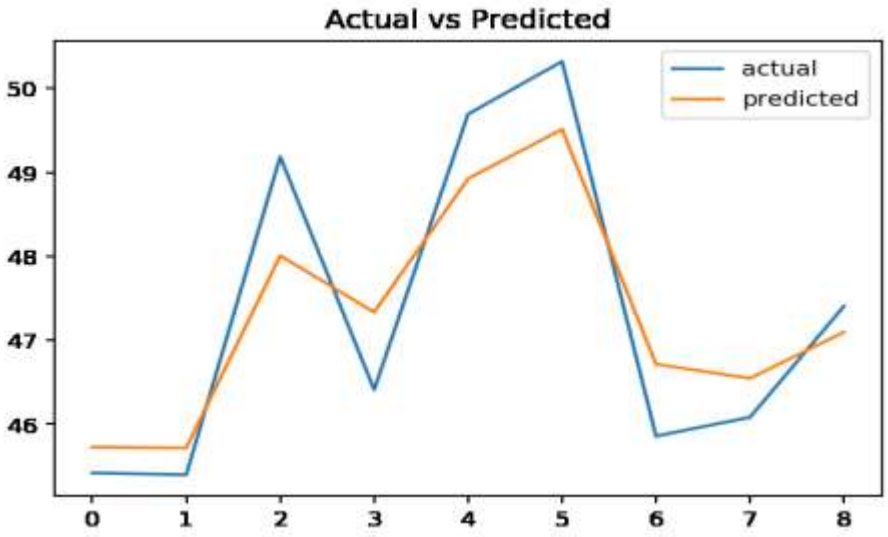


Figure3:-Actual graph before prediction vs actual graph after prediction.

By observing the third graph i.e. figure3 actual values are slightly greater than predicted values. This gives the information that the production rate is maximum in the year 2006 in the actual graph but in the predicted graph the production rate is maximum in the year 2005. Also, the production rate is minimum in the year 2002-2004 in the actual graph but in the predicted graph the production rate is minimum in the year 2000-2002.

Inference

In this project, one of the major problems faced by the farmers i.e. productivity parameters is predicted by making use of Big data concepts such as apache spark, pyspark, spark sql and MLlib. Implementation of machine learning concepts such as linear regression and random forest regressor in spark gives the above-mentioned results. Considering the results, the conclusion is given based on mean squared error of linear regression and random forest regressor. Because mean squared value is quite enough to predict the suitable model for the data taken for analysis

Conclusion:-

In this project, one of the major problems faced by the farmers i.e. productivity parameters is predicted by making use of Big data concepts such as apache spark, pyspark, spark sql and MLlib. Implementation of machine learning concepts such as linear regression and random forest regressor in spark gives the above-mentioned results. Considering the results, the conclusion is given based on mean squared error of linear regression and random forest regressor. Because mean squared value is quite enough to predict the suitable model for the data taken for analysis. The mean squared error value of linear regression is 0.33 and mean squared error value for random forest regressor is 0.73 since the error value of linear regression is close to zero, neglect this machine learning algorithm. Random forest regressor is the model that suits for the data involved in this project. By making use of random forest regressor the maximum and minimum production rate is predicted based on seasons, crops and area. Farmers should cultivate the appropriate crops that suits the season and the area allotted for cultivation. For example, if a farmer is planting black pepper, the rabi crop in kharif season or he is planting rice the kharif crop in kharif season but in small area, then both the situations leads to loss of production. If a farmer wish to increase production then the farmer must choose the right crop that is suitable for the season and the amount of area allotted for cultivation. In this project future prediction includes the pesticides and natural manure that is given for crops and prediction of suitable manure for increasing the growth of plants and maximizing the production rate is the extension of this project.

Reference:-

1. H. Chen, W. Zhang, and L. Fan. (2011).“Methodology of crop breeding progress and prospect,” *Bulletin of Science and Technology*, vol. 27, no. 1, pp. 61–63, View at: Google Scholar
2. D. Chun-shui and C. Zhou. (2013). “Advances in modern data-driven breeding technologies,” *Journal of Maize Sciences*, vol. 21, no. 1–8, pp. 1–2, View at: Google Scholar
3. T. M. Li, J. Y. Chen, and D. D. Yan. (2014). “Analysis of application prospect of big data,” in *Proceedings of the Academic Annual Conference of Sichuan Communication Association*, pp. 67–69,View at: Google Scholar

4. H. T. Teng. (2008). "Exploration on digital maize breeding," *Chinese Agricultural Science Bulletin*, vol. 12, no. 24, pp. 495–498, View at: Google Scholar
5. L. J. Fang, W. D. Wang, B. Wang, C. Y. Ye, Q. Y. Shu, and H. Zhang.(2006). "Crop breeding-related data and application of big data technologies in crop breeding," *Journal of Zhejiang University (Agriculture & Life Sciences)*, vol. 42, no. 1, pp.
6. Perrier, A. 1985. Updated evapotranspiration and crop water requirement definitions. Crop Water Requirements, eds. Perrier, A., and C. Riou, pp. 885-887. Paris: INRA.
7. 2. Jensen, M. E., R. D. Burman, and R. G. Allen (eds.). 1990. Evapotranspiration and Irrigation Water Requirements. ASCE Manuals and Reports on Engineering Practices No. 70, New York.
8. 3. Doorenbos, J., and W. O. Pruitt. 1975. Guidelines for Predicting Crop Water Requirements, FAO Irrigation and Drainage Paper 24, Rome: FAO.
9. 4. Monteith, J. L. 1965. Evaporation and environment. 19th Symposia of the Society for Experimental Biology, Vol. 19, pp. 205-234. Cambridge, UK: Cambridge University Press.
10. 5. Monteith, J. L. 1985. Evaporation from land surfaces: Progress in analysis and prediction since 1948. Advances in Evapotranspiration, pp. 4-12. St. Joseph, MI: American Society of Agricultural Engineers.
11. Asare-Kyei, D., Forkuor, G., and Venus, V., 2015. Modeling Flood Hazard Zones at the Sub-District Level with the Rational Model Integrated with GIS and Remote Sensing Approaches. *Water*.
12. ASCH, 2015. *Crop pests and diseases: A manual on the most important pests and diseases of the major food crops grown by smallholder farmers in Africa*. Nairobi, Kenya: Africa Soil Health Consortium.
13. Awo, M.A., 2012. *Marketing and Market Queens: A Study of Tomato Farmers in the Upper East Region of Ghana*. LIT Verlag Münster.
14. Baffes, J., 2005. The 'Cotton Problem'. *The World Bank Research Observer*, 20 (1), 109–144.
15. Banful, A.B., 2011. Old Problems in the New Solutions? Politically Motivated Allocation of Program Benefits and the 'New' Fertilizer Subsidies. *World Development*, 39 (7), 1166–1176.
16. Barry, B., Kortatsi, B., Forkuor, G., Gumma, M.K., Namara, R., Rebelo, L.M., van den Berg, J., and Taube, W., 2010. Shallow Groundwater in the Atankwidi Catchment of the White Volta Basin: Current Status and Future Sustainability. *Research Report. International Water Management Institute*, (139), 1–22.