ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Analysis of Agricultural Data Using Big Data Analytics

K. Ravisankar, K. Sidhardha, Prabadevi B*
Department of Information Technology and Engineering, VIT University, Vellore.
***Corresponding author: E-Mail: prabadevi.b@vit.ac.in**

ABSTRACT

Agriculture has been the backbone of the Indian economy for many years. But in the past few years, Agriculture in India has lost its dominance over world Agriculture development. There are many reasons for the pitfall of agriculture in India. Considering the facts that Industrial sector is providing better pay rate compared to Agriculture and also minimized government support, the Agriculture has taken a big downfall. Moreover considering the fact that most of the farmers in Agriculture are illiterates and unaware of the newest technologies is also the reason for the pitfall of Agriculture.

KEY WORDS: Big Data, Agriculture, Data Analytics, Algorithms, Seasons, Economy.

1. INTRODUCTION

As the advancement in the technology is increasing in a broader way the usage of its technology to Agriculture sector is very limited. Agriculture is the backbone of the Indian economy. But in last few years, agriculture has taken down a huge steep due to many reasons. Lack of knowledge of seasons, crops and price are also one of them i.e. not knowing which season for which crop and which crops are receiving the higher payment for the government. Our project deals with the analysis of agricultural data which helps in better understanding of agriculture in India.

The outcome of the crops depends on the several factors like rainfall, season, temperatures, and also the price announced by the government. In the project, we will categorize the crops based on the season, based on Minimum price announced by the government, based on temperature. The data also allows the user to visualize a different kind of data. The data also contains different crops planted in Area (Hectares) and Production (Tonnes). The project also visualizes the different categories of data for the better of understanding of Agriculture in India.

Big data: A dataset or heap of data will be called as a big data when the amount of data in the file or system is more than the amount of data any one processor can handle, in these situations, we will handle the data using big data Techniques.

There are many types of big data techniques and tools available, but we will concentrate more on the standard common MapReduce and the YARN MapReduce. Initially, we will store the data which has been collected by and want to be analyzed in the hdfs storage system. In a hdfs storage system, the data will be stored in the form of blocks, which next will be separated into different clusters of blocks. There are many nodes (or) parts of big data analytics tool like Hadoop like name node, data node and resource schedulers like resource manager, node manager and app scheduler.

Existing System: In the existing studies, the variety of methods is implemented to classify the agriculture data. In the existing system, only a few of many factors available are considered for the agricultural data classification. Many of the existing systems are implemented using Data Mining techniques where the accuracy of classifying the data is lower when compared to the big data analytics. The existing systems do not consider the dependency between the various factors affecting the crop production. And more over the existing does use the old data which is not up to date. Thus the classification and prediction of data can't be true and accurate. Thus using up to date data with the "Big Data analytics" helps in accurate classification and prediction of data

Proposed System: Instead of using the data mining techniques we will use the Map reduction Techniques of Big Data Analysis. The Big data Analytics provides following advantages compared to data mining techniques

Cost Reduction: Big data technologies like Hadoop and cloud-based analytics can provide substantial cost advantages. While many comparisons are there between big data technology and traditional architectures, for example :(data warehouses and marts in particular) are difficult because of differences in functionality, a price comparison alone can suggest order-of-magnitude improvements.

- Faster and Better decision making by using distribute solutions.
- Extensibility and More Reliable way of analyzing data.

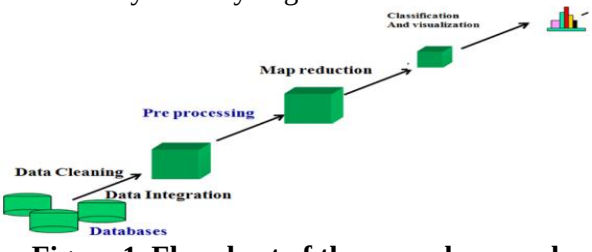

Figure.1. Flowchart of the procedure used

Module description:

Dataset collection: Here in this process, we collect all the required dataset. Regarding the datasets the initial description and the point to be remembered is about the attributes. The dataset regarding the attributes which suits the project must be analysed such that the entire results depend on the dataset collected and attributes containing in it. The data set collected are

- crops vs. seasons
- crops vs. price(various years)
- seasons vs. temperature(various years)

Data pre-processing: In datasets collected may contain various null values, inconsistent values and datasets may also be in different formats. In this process of the pre-processing, we will remove null values and inconsistent values. The null values reduce the accuracy of prediction of the data. And inconsistent data reduces and confuses the algorithm. In order to pre-process the data, we will be using the Weka tool (or) R studio tools. The pre-processed data helps in more accurate prediction of the data.

Data integration: It involves combining data residing in different sources and providing users with a unified view of these data. This process of integrating becomes compulsory in a variety of situations, which include both commercial occasions like (when two similar companies need to merge their databases) and scientific repositories. The data integration is done using the "R studio".

Map reduction: Map reduction technique in the Hadoop tool is used to get the data that's only required. Map reduction technique reduces the amount of data needed to be processed by the classification algorithm is reduced.

Data classification and clustering: Data classification is the process (or) procedure of sorting and categorizing data into various types, forms which will help us more when we are analyzing the data. Data classification helps us a lot to do the separation of data and classification of data according to data set requirements for various business motives or personal objectives. It is mainly a data management process. Data classification helps in the prediction of the data. The tool used for the classification is the "R studio".

Visualization: The next module in the project will be the visualization of the classified data. The visualization helps in better understanding of the classified data. The Visualization is the process where the prediction of the data takes place.

Performance analysis: Roc is widely used for evaluating the performance of any application. The ROC curve shows how the percentage of performance by dividing the number of correctly classified positive cases varies with the number of incorrectly classified negative cases.

Here data is analyzed using the Naive Bayes classifier, the decision tree classifier. The analysis results are presented as ROC curves, and the risk factors are ranked and compared for the two classifiers.

Algorithms:

Simple Linear Regression: In basic straight relapse, we foresee scores on one variable from the scores on a moment variable. The variable we are foreseeing is known as the measured variable and is alluded to as Y. The variable we are constructing our forecasts with respect to is known as the indicator variable and is alluded to as X. At the point when there is stand out indicator variable, the forecasting technique is called straightforward relapse. In basic direct relapse, the subject of this segment, the forecasts of Y when plotted as a component of X shape a straight line.

The case information in Table.1 is plotted in Figure.1. You can see that there is a positive relationship amongst X and Y. In the event that you would foresee Y from X, the higher the estimation of X, the higher your expectation of Y.

Decision tree: Decision Trees (DTs) are a non-parametric regulated learning strategy utilized for order and relapse. The objective is to make a model that predicts the estimation of an objective variable by taking in straightforward choice guidelines derived from the information highlights.

Advantages:

- Requires little information planning. Different systems regularly require information standardization, sham factors should be made and clear values to be evacuated. Note however that this module does not bolster missing qualities.
- The cost of utilizing the tree (i.e., anticipating information) is logarithmic in the quantity of information focuses used to prepare the tree.
- Ready to handle both numerical and absolute information. Different strategies are generally represented considerable authority in breaking down data sets that have one and only kind of factor.

Usage of algorithms: We will use the simple linear regression to predict the agriculture parameter values like temperature, rainfall, moisture level and profit. A decision tree will be used to find out what crop is preferable at what state by using sample data.

Usage of Big data tools: We will use Hadoop to analyze data by importing the dataset into Hadoop hdfs cluster and then by suing hive we will find out the wanted data by using SQL queries and then we will cross check values. we got with data obtained from the algorithmic analysis.

Algorithms results:

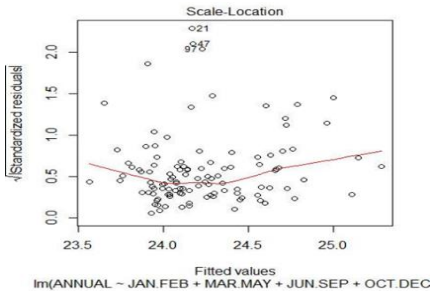

Figure.2. Scatter plot of the annual rainfall

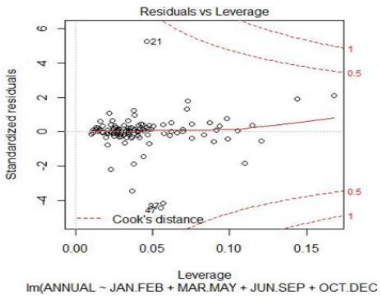

Figure.3. Residuals vs Leverage plot of annual rainfall

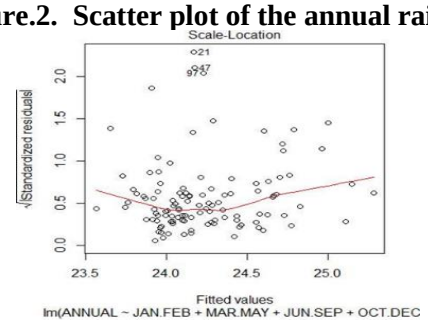

Figure.4. Scale location of annual Temperatures

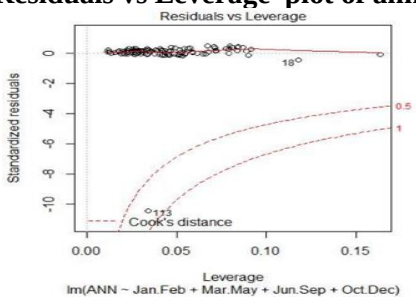

Figure.5. Residual vs Leverage plot of annual rainfall

Data visualization results:

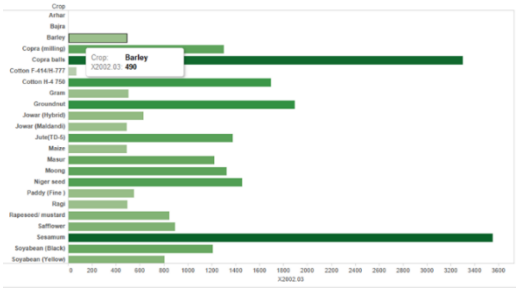

Figure.6. Bar chart showing the estimated price of various crops

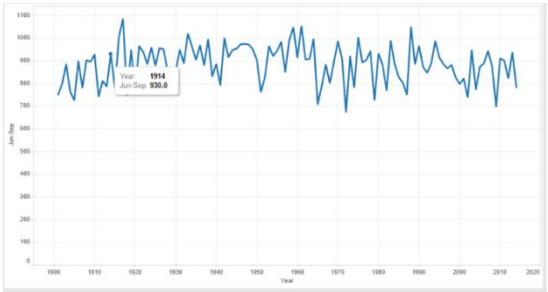

Figure 7. Line chart showing rainfall for season of June-September

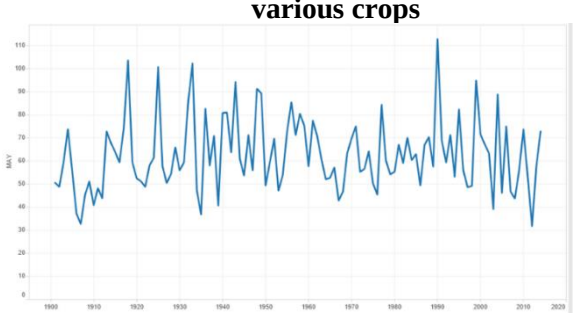

Figure.8. Line chart showing the Temperatures for May between 1903-2013

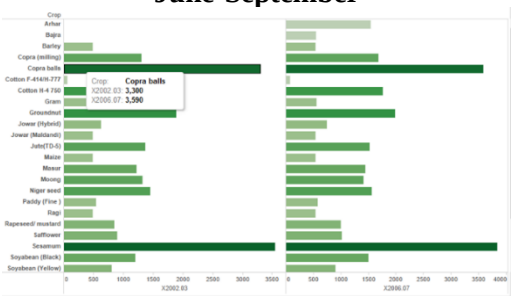

Figure.9. Bar chart used for comparing prices of various crops between 2003 and 2004

Scope of future work: There is a lot of scope for doing more work on agriculture data in future. Some of them will be listed below.

- Getting feedback from users to improve the user interface.
- Using text analytics to find out the opinions of common people.
- Creating an interactive discussion forum using new technologies like meteorjs.
- Using more new advanced algorithms for increased accuracy in predictions.
- The main aim is to create an ecosystem of agriculture-oriented software solutions to aid farming in every aspect.
- When the data size which we are going to handle reaches huge sizes ,then by using big data tools like Hadoop, spark and sandbox's can be used for more effective analysis.

- Our main intention is to make our product's availability high by communicating with government offices to allocate new centers with experts to help farmers more.
- Increasing the number of recommendations in fields like fertilizers and pesticide usage based upon the previous usage.
- Including IoT to encourage more e-farming.
- Using the cloud to store and retrieve data (azure).

2. CONCLUSION

Hereby we will conclude that all the data results which we were given (or0 predicted were solely done based on the previous year's data and we have kept some good amount of work in order to find the inner relationships between the agriculture parameters and we are hoping that we have done that. The main objective of this paper is to help the farmers or agriculture workers such that they can do agriculture more smartly in a much better calculated way.

REFERENCES

De S.S, Chattopadhyay G, Bandyopadhyay B, and Paul S, A neurocomputing approach to the forecasting of monthly maximum temperature over Kolkata, India using total ozone concentration as predictor, Comptes Rendus Geoscience, 343 (10), 2011, 664-676.

Garcia, Ted, and Taehyung Wang, Analysis of Big Data technologies and method-Query large Web public RDF datasets on Amazon cloud using Hadoop and Open Source Parsers, In Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on, IEEE, 2013, 244-251.

Kannan, Elumalai, and Sujata Sundaram, Analysis of trends in India's agricultural growth, The Institute for Social and Economic Change, Bangalore, India, Working paper 276, 2011.

Krishnankutty N, Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network, Current Science, 90 (6), 2006, 773.

Patel, Aditya B, Manashvi Birla, and Ushma Nair. Addressing big data problem using Hadoop and Map Reduce, In Engineering (NUiCONE), 2012 Nirma University International Conference on, IEEE, 2012, 1-5.

Taie, Mostafa Anwar, Ibrahim El-Faramawy, and Mohamed Elmawazini, Methods for Prediction, Simulation and Verification of Real-Time Software Architectural Design based on Machine Learning Algorithms, SAE Technical Paper, 2015.

Umachandran, Krishnan, and Debra Sharon Ferdinand-James, Affordances of Data Science in Agriculture, Manufacturing, and Education, In Privacy and Security Policies in Big Data, IGI Global, 2017, 14-40.

Xin, Ning Yu, and Li Yue Ling, How we could realize big data value." In Instrumentation and Measurement, Sensor Network and Automation (IMSNA), 2013 2nd International Symposium on, IEEE, 2013, 425-427.