Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.8, No.7, 2018

A Machine Learning Framework to Predict Determinant Factors of Seeds

Tekalign Tujo Gurmessa Elifenesh Yitagesu Desta
Lecturer, Madda Walabu University, College of Computing, Department of Computer Science, Bale Robe, Ethiopia
PO Box 247, Bale Robe, Ethiopia

Abstract
In this paper, we audit the machine learning apparatuses for foreseeing determinant components of seeds. We depict this issue regarding Big Data, ANN, Hadoop and R. We consider Machine-learning techniques especially suited to forecasts dependent on existing information, yet exact expectations about the far off future are frequently on a very basic level unthinkable. Farming is an industry where recorded and current information flourish. This survey researches the various information sources accessible in the horticultural field and dissects them for utilization in Seed determinant factor Predictions. We recognized certain relevant information and researched techniques for utilizing this information to improve forecast inside the farming action.

Introduction

The utilization of innovation in agribusiness works has begun since the early piece of the twentieth century when the business moved from the steed attracted digger to motorized tractors. The presentation of plant heredities, synthetic sources of info and harvest administration frameworks has changed the business into innovation great and information-rich world [1]. The innovative advances that make up the present processing condition have added to examine huge information while information gathering isn't new ideas particularly with regards to open information accumulation. The main beginning of more productive portable advancements and the digitization of information have vast records possessed the capacity to be assessed and examined in an auspicious and more helpful way.

Huge Information Examination devices like Hadoop, R and machine learning calculations like ANN are utilized to help rural focus specialists. Huge information holds one of the keys for agriculturists control focuses to gather and process information progressively to help ranchers that settles on the best choices with respect to planting, preparing and reaping crops. In the present conditions, rural undertakings are fit for producing and gather a lot of information. So this Development in information estimate requires robotized technique to concentrate and examination vital data's. Keeping in mind the end goal to take the full favorable position of the dirt sort, dampness, moistness, atmosphere and so forth ranchers need to know precisely the kind of seeds for their trimming. Diverse locale in Ethiopia have fluctuating atmospheres thus it is essential to think about the natural components of these different regions. This assistance to pick the best areas for cultivating of various sort of seeds. Precipitation additionally changes from region to area and this huge affects cultivating in light of the fact that while too little or an excess of rain can execute crops, the best possible measure of rain prompts idealize trim yield.

[4] Created farming administration for basic and exact estimation methods to anticipate rice yields in the arranging procedure. An ANN [5], is a type of computerized reasoning which is made out of countless handling parts considered fake neurons or hubs that are interconnected by direct connections, called associations, and which coordinate to perform parallel circulated preparing (PDP) activity with a specific end goal to take care of a given issue. A subgroup of handling component [6] is known as a layer in the system. The most minimal layer is the information layer and the most astounding layer is the yielding layer. Between the most reduced and most noteworthy layer, there might be an extra layer(s) of units, called concealed layer(s). The upside of neural systems over customary programming lies [6] in their capacity to end up an answer for various issues that don't have an algorithmic arrangement or the accessible arrangement is too mind-boggling to ever be found. An ANN [7] is balanced for a particular application, for example, design acknowledgment or information characterization, through a preparation procedure.

The ANN demonstrating ending up exceptionally prominent in various regions of farming, uniquely, in the territories where straight measurable displaying ends up unsuccessful. The ANN is utilizing in the field of agribusiness to foresee the harvest yield, biomass generation, seeding dates, physical and physiological harming of seeds, natural issue substance in the dirty, soil dampness estimation, streamlined properties of products, estimation of sugar content in foods grown from the ground of yield assortments [8].

1. Big Data

Exponential growth in the use of information technology by organizations has resulted in the availability of a tremendous volume of data to knowledge workers.

According to[1], Big Data refers to datasets whose size is outside the ability of typical database software tools to capture, store, manage, and analyze. The concept of Big Data is therefore relative to the storage and processing capability of the prevalent technology of the time. The exponential growth of transistor density was predicted by Intel co-founder [2], states that the number of transistors on a chip doubles about every two years, resulting in the rapidly rising processing power and declining hardware costs. As described in [3], the 1980 US Census data would be considered as Big Data in the 1980s, where the IBM 3850 Mass Storage System with a capacity of 102.2GB was the monster storage device of its day. The [4] group, characterized Big Data by the three V's: volume, velocity, and variety. Other characteristics such as veracity and value have been added to the explanation by other researchers.

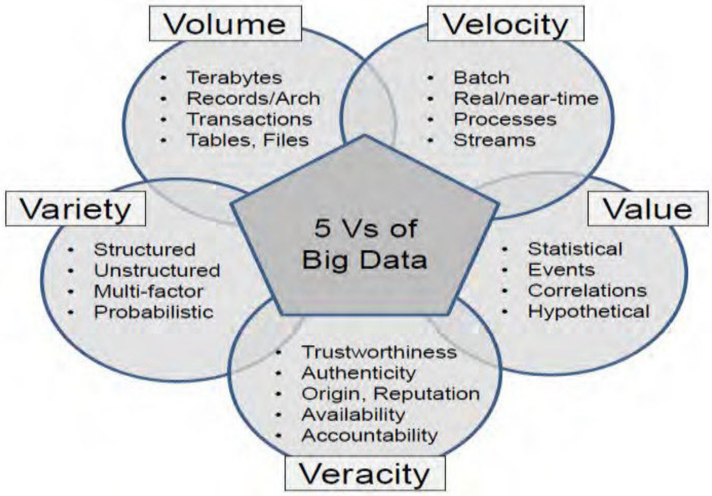

Figure 1: The 5'v of Big Data [5]

a) Volume
Refers to the size of data that has been created from all the sources. This presents the most challenges of IT structures. This is the feature that comes to most people's minds when they think of big data. Many companies already have large amounts of archived data in the form of logs, but do not have the capacity to process that data. The benefit gained from the ability to process large amounts of formation is the main attraction of big data analytics.

b) Velocity
Refers to the speed at which data is generated, stored, analyzed and processed. An emphasis is being put recently on supporting real-time big data analysis. It also refers to the increasing speed at which this data is created, so increasing speed at which the data can be processed, stored and analyzed by relational databases. Velocity refers to the speed at which new data is generated and the speed at which data moves around.

c) Variety
The next aspect of big data is its variety, refers to the different types of data being generated. It is common now that most data are unstructured and cannot be easily categorized or tabulated. Big data is not always structured data and it is not always easy to put big data into a relational database. This means that the category to which big data belongs to is also a very essential fact that needs to be known by the data analysis dealing with a variety of structured and unstructured data greatly increases the complexity of both strong and analyzing big data.

d) Variability
Refers to how the structure and meaning of data constantly change especially when dealing with data generated from natural language analysis.

e) Value
Value is the most important feature of big data. Though, the potential value of the big data is vast. It is all well and good having access to big data. It also refers to the possible benefit of big data can offer a business based on good big data collection, management, and analysis.

2. Basics of Artificial Neural Networks
The field of Artificial Neural Networks (ANNs) or Neuro computing or Connectionists theory (CT) [6] was born out of a research done by McCulloch and Pitt in 1943 on a simple model of a neuron. Following this were further attempts to generate knowledge from the study of the central nervous system of a human being and application of such knowledge to the design and implementation of systems. Artificial neural networks [7] are networks based on mathematical calculations which try to imitate the working principles of the networks that are found in the nerve cell of the central nervous system of an animal or human being during the design and implementation

of artificial systems. This imitation is a gross cell-by-cell (neuron-by-neuron, element-by-element) imitation.

A neuron transmits input [6][7] to the cell body or soma of another neuron through the synaptic connections to the dendrites. The output signal or information (consisting of nerve impulses or electrical triggers) from a cell is transmitted through the connection wire (axon) to the synapses of other neurons.

3. Types of Topologies in Artificial Neural Network

Connecting two or more artificial neurons[8] getting an artificial neural network is useful to solve real-life problems because single artificial neuron has almost no usefulness in solving real-life problems. Artificial neurons are the basic building blocks of an artificial neural network which are useful to solve complex real-life problems in a non-linear, distributed, parallel and local way. The way that individual artificial neurons are interconnected is called topology, architecture or graph of an artificial neural network. The architecture [9] is inspired by the working principle of the cerebral cortex of the brain. The two most widely used architectures of ANN are described below:

3.1 Feed-Forward Artificial Neural Networks

Artificial neural network with feed-forward topology[8] is called Feed-Forward artificial neural network and as such has only one condition: signal must flow from input to output in only one direction. However back loop is not allowed. Any number of layers, the type of transfer function used in an individual artificial neuron or the number of connections between individual artificial neurons. The simplest feed-forward artificial neural network is a single perceptron that is only capable of learning linear separable problems. However simple multi-layer feed-forward artificial neural network for purpose of analytical description (sets of equations (3), (4) and (5)) is shown in Figure 4 below.

$n1 = F1\ (w1x1+b1)$

$$\left.\begin{aligned} n2 &= F2\ (w2x2+b2)\\ n3 &= F2\ (w2x2+b2)\\ n4 &= F3\ (w3x3+b3)\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

$$\left.\begin{aligned} m1 &= F4\ (q1n1+q2n2+b4)\\ m2 &= F5\ (q3n3+q4n4+b5)\\ y &= F6\ (r1m1+r2m2+b6)\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (2)$$

$$y = F_6 r_1 (F_4 [q_1 F_1 (w_1 x_1 + b_1) + q_2 F_2 (w_2 x_2 + b_2) + b_4]) + .. + r_2 \left(F_5 \left[q_3 F_2 (w_2 x_2 + b_2) + q_4 F_3 (w_3 x_3 + b_3) + b_5\right]\right) + b_6 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (3)$$

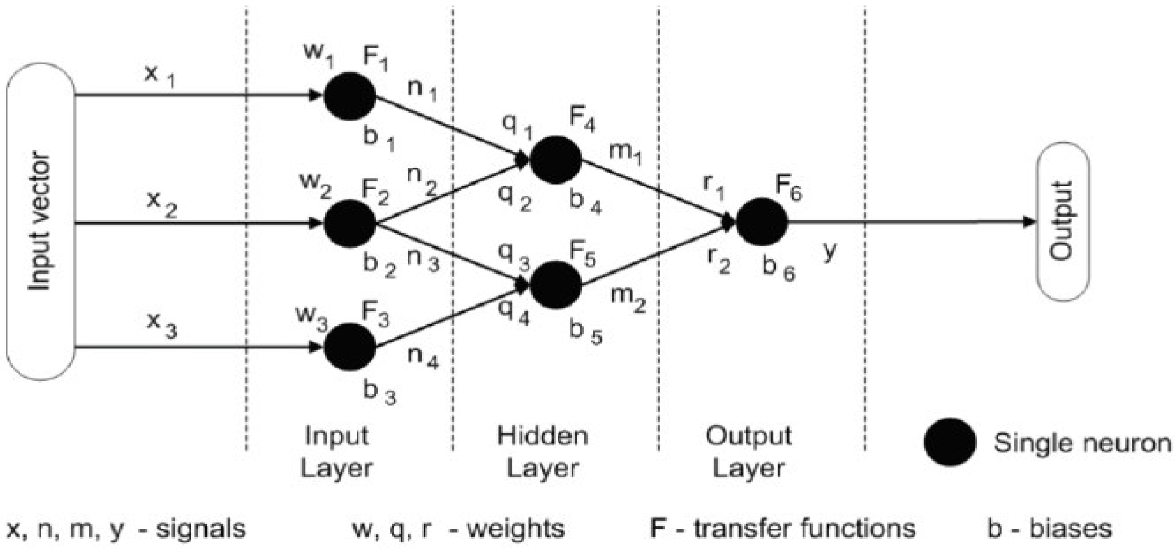

Figure 2: Feed-forward Artificial neural network [8]

As we can see in Figure 4 above and corresponding analytical description with sets of equations (1), (2) and (3) the simple feed-forward artificial neural network can be led to relatively long mathematical descriptions where artificial neural networks' parameters optimization problem solving by hand is impractical. Although analytical description can be used on any complex artificial neural network in practice we use computers and specialized software that can help us build, mathematically describe and optimize any type of artificial neural network. In addition to this [9] states that in feed-forward ANN the output of any layer does not affect that same or preceding layer.

3.2 Recurrent Artificial Neural Networks

Artificial neural network with the recurrent topology[10], is called recurrent artificial neural network. It is similar to a feed-forward neural network, however, there are no limitations regarding back loops as well as the number of layers, type of transfer function used in individual artificial neuron or number of connections between individual artificial neurons. In these cases signal is no longer transmitted only in one direction but it is also transmitted backward. This creates an internal state of the network which allows it to exhibit dynamic temporal behavior. Recurrent artificial neural networks can use their internal memory to process any sequence of inputs. Figure 5 below shows small Fully Recurrent artificial neural network and complexity of its artificial neuron interconnections. The most basic topology of the recurrent artificial neural network is a fully recurrent artificial network where every basic building block (artificial neuron) is directly connected to every other basic building block in all direction.

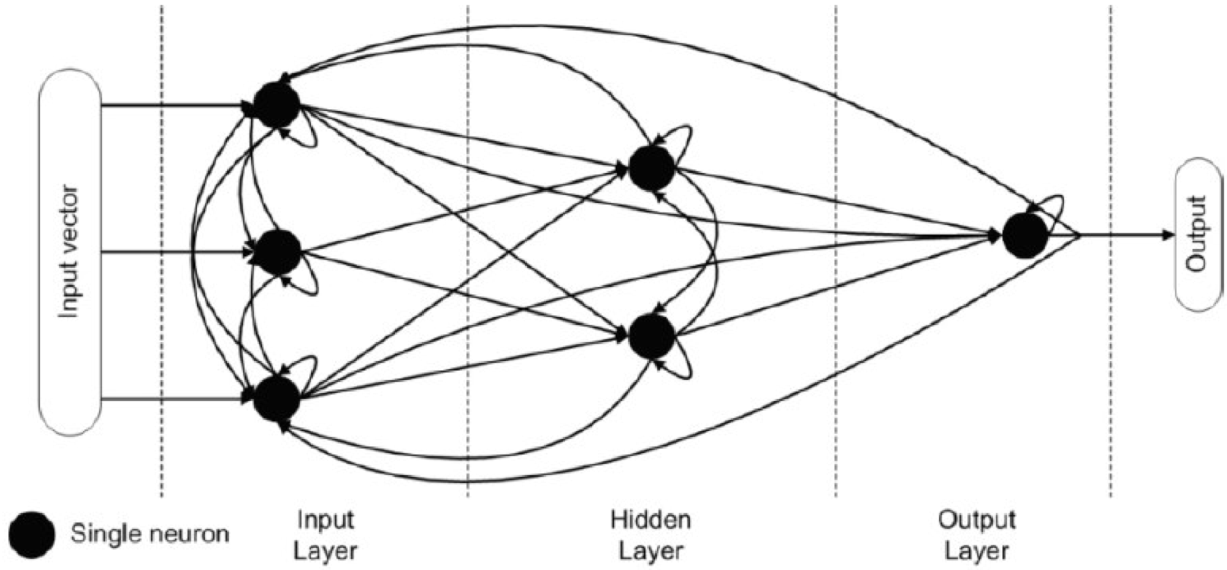

Figure 3: Fully recurrent artificial neural network [8]

In addition to this[9] states that in recurrent ANN the output of a neuron is fed back into itself as input.

4. Artificial Neural Networks Types

The most significant class of neural networks for solving real-world problems are[9] multilayer perceptron, radial basis function, and Self-Organizing Map. Each type of artificial neural network has their own unique characteristics as described below.

4.1 Multilayer Perceptron

The most popular form of neural network architecture is[9] the multilayer perceptron (MLP). MLP is [23] normally composed of several layers of nodes. In the MLP the first layer is an input layer where external information is received whereas the last layer is an output layer where the problem solution is obtained. The input layer and output layer are connected by one or more intermediate layers called the hidden layers. The nodes in adjacent layers are usually fully connected by acyclic arcs from the input layer to the output layer.

4.2 Radial Basis Function Networks

RBF networks typically have [9] only one hidden layer with any number of units. The connections exist between the input layer and the hidden layer, and between the hidden layer and the output layer. It uses radial combination functions in the hidden layer, based on the squared Euclidean distance between the input vector and the weight vector. It also uses exponential or softmax activation functions in the hidden layer, in which case the network is a Gaussian RBF network. MLPs are said to be distributed-processing networks because the effect of a hidden unit can be distributed over the entire input space. However Gaussian RBF networks are said to be local-processing networks because the effect of a hidden unit is usually concentrated in a local area centered at the weight vector. Radial basis networks can need[12] more neurons than standard feed forward back propagation networks, but often they can be designed in a fraction of the time it takes to train standard feedforward networks. They work best when many training vectors are available.

4.3 Kohonen Neural Network

The self-organizing map also called Kohonen neural network [10] is an artificial neural network that is related to feed-forward networks but it needs to be told that this type of architecture is fundamentally different in arrangement of neurons and motivation. A common arrangement of neurons is in a hexagonal or rectangular grid. The self-organizing map is different in comparison to other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. They use unsupervised learning paradigm to produce a low-dimensional, discrete representation of the input space of the training samples, called a map what makes them especially useful for visualizing low-dimensional views of high-dimensional data. Such networks can learn to detect regularities and correlations in their input and adapt their future responses to that input accordingly.

5. Working Principles of ANN and Back Propagation Algorithm

An artificial neuron is [16] a simple mathematical model (function) which is important for building ANN. Such a model has three simple sets of rules: multiplication, summation, and activation. At the entrance of artificial neuron, the inputs are weighted what means that every input value is multiplied with individual weight. In the middle section of an artificial neuron is sum function that sums all weighted inputs and bias. At the exit of an artificial neuron, the sum of previously weighted inputs and bias is passing through activation function that is also called transfer function. Figure 6 below shows the three sets of rules which are important for the operation of ANN.

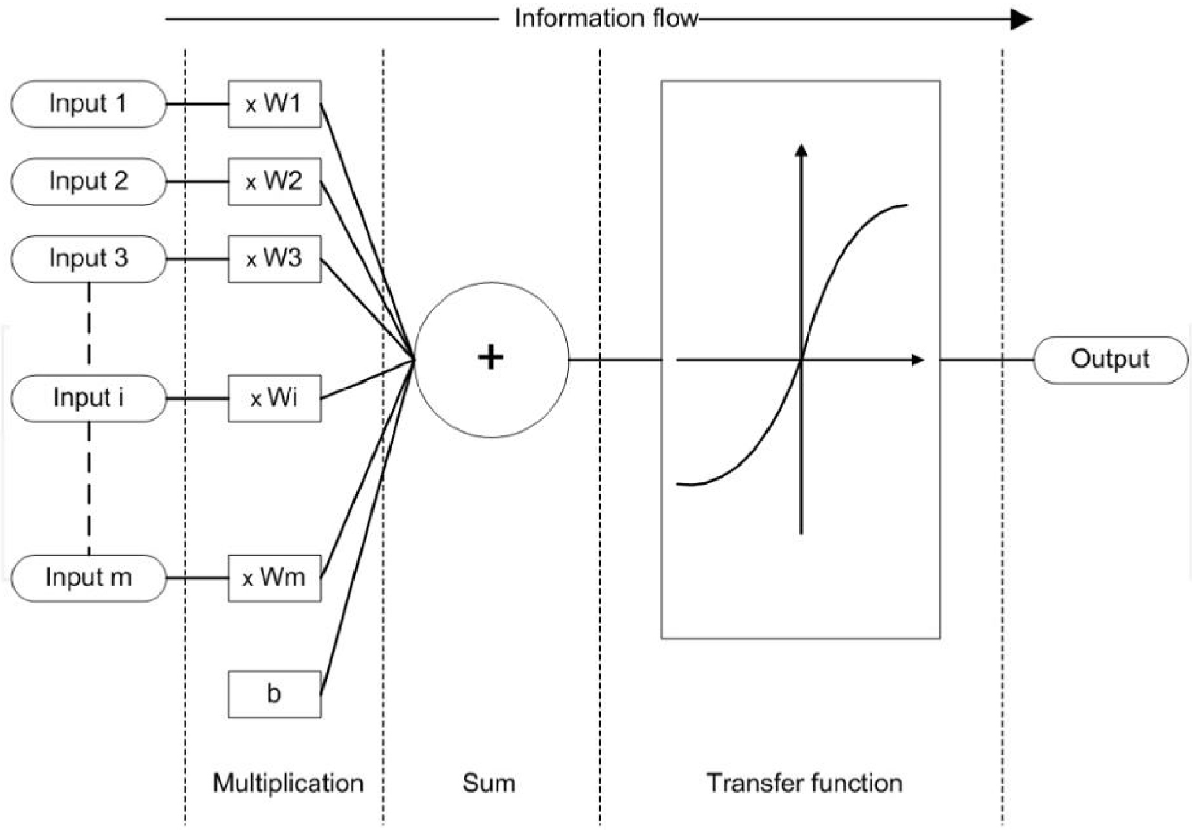

Figure 4: Working principle of an artificial neuron [8]

The artificial neural network has been trained [14] by exposing it to sets existing data where the outcome is known. Multi-layer networks use a variety of learning techniques; the most popular is the backpropagation algorithm. It is one of the most effective approaches to machine learning algorithm developed by David Rumelhart and Robert McLelland(1994). The signal flows from the direction of the input layer towards the output layer. A network is trained rather than programmed. Learning in artificial neural networks is typically accomplished using examples. This is also called "training" in artificial neural networks because the learning is achieved by adjusting the connection weights in artificial neural networks iteratively so that trained. The number of iterations of the training algorithm and the convergence time is variable depending on the weight initialization.

The algorithm includes two passes: a forward pass and a backward pass.

- ❖ In forwarding pass [15] an activity pattern is applied to the input nodes and it propagates through the network layer by layer. As a result, a set of outputs is produced as the actual response of the network.

The weights at the functional points of the network are fixed in the forward pass. In addition to this [16] also states that during the forward pass of the signal, according to the initial weights and activations function used, the network gives an output. That output is compared with the desired output. If both are not the same, an error occurs.

- During backward pass[15] the synaptic weights are all adjusted in accordance with an error-correction rule. The actual response is subtracted from the desired output to produce an error signal. The error signal is propagated backward through the network. The synaptic weights are adjusted to have actual output nearer to the desired output. The weight adjustment is done according to the generalized delta rule to reduce the error. In addition to this [16] also states that during a reverse pass, the error is back propagated and weights of hidden and output layer are adjusted.

6. Predictive Data Mining

[18] On their part say that predictive data mining involves using some variables or fields in the database to predict unknown or future values of other variables of interest.

Similarly, [19] states that predictive methods aim to describe one or more of the variables in relation to all the others; they are also called asymmetrical, supervised or direct methods. This is done by looking for rules of classification or prediction based on the data. These rules help us to predict or classify the future result of one or more response or target variables in relation to what happens to the explanatory or input variables. The main methods of this type are those developed in the field of machine learning such as the neural networks (multilayer perceptions) and decision trees but also classic statistical models such as linear and logistic regression models.

7. Big Data Analytics

According [1] Big Data Analytics is designed at making intelligence of data by applying efficient and scalable algorithms on Big Data. This includes the design of efficient and effective algorithms and systems to integrate the data and uncover the hidden values from data. It also includes methodologies and algorithms for automatic or mixed-initiative knowledge discovery and learning, transformation and modeling, predictions and explanations of the data. Developments in this area include new algorithms, methodologies, systems, and applications for knowledge discovery, understanding, and applications based on the Big Data. New computing paradigms are expected in new areas such as human computation, crowdsourcing, and sentiment analysis as well as data visualization technologies. The Big Data analytics architecture described in figure 7 below utilizes the massively parallel, distributed storage and processing framework as provided by Hadoop HDFS and MapReduce.

[1] Also described geospatial intelligence as using data about space and time to improve the quality of predictive analysis. For example, real-time recommendations of places of interest can be based on the real-time location from smartphone usage. This real-time information can be combined with batched analytics to improve the quality of the predictions. Other examples of real-time analytic applications include real-time trending of social media data, real-time Web clickstream analysis, algorithmic trading, and real-time M2M analysis.

Emerging technologies for Big Data real-time analytics include technologies for collection and aggregation of real-time data for Hadoop, in-memory analytic systems, and real-time analytics applications for processing of data stored in Hadoop. Real-time insight created by real-time analytics can be consumed by real-time operations and decision processes.

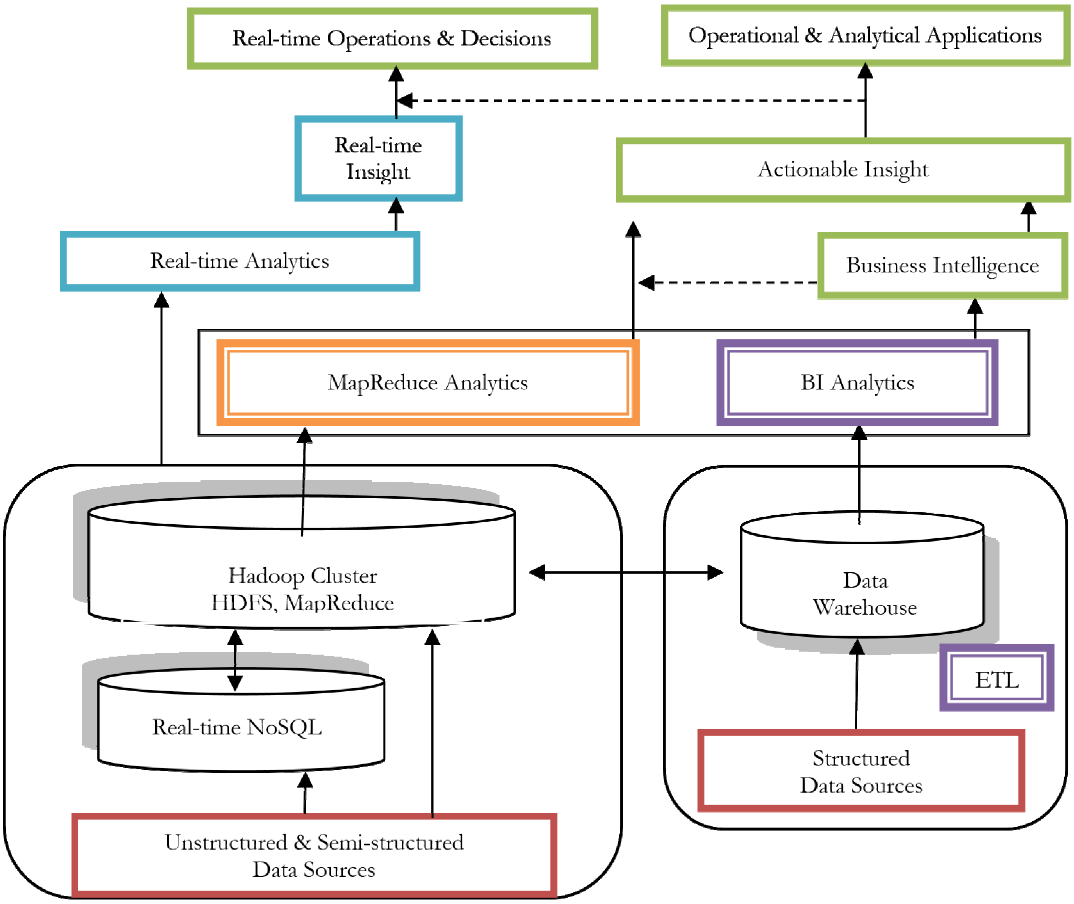

Figure 5: An Architecture for Big Data Analytics[1]

8. Big Data in Agriculture

There is a number of definitions exists about big data. Analyst [4] defines big data as " high velocity, high-volume, and high-variety information properties that demand cost-effective, innovative forms of information processing for improved insight and decision making." [20] Showed empirical research and determined that organizations performance is directly related to their ability to make data-driven decisions. This makes it is significant to understand the success factors required to implement big data techniques in organizations. So by using the Big Data Analytics and other applications to increase the agricultural yields and reduce the expenses of farming by taking accurate decisions.

It is important that agriculturists and the farming community, in general, understand the behavior of the climatic parameters that affect seeds and other plants. Knowledge of seed variability is vital in order to minimize losses and maximize production in agriculture.

9. Machine Learning Algorithms

Currently, non-cloud machine learning libraries lack scalability, not well verified and not built with production-quality, and they also lack open source community support. Therefore, Mahout has been created with scalable machine-learning algorithms all written using Map-Reduce paradigm, including 1) supervised learning algorithms such as neural networks, support vector machines, Naive Bayesian classifiers, decision trees, random forests, and logistic regression; 2) unsupervised learning algorithms including k-means, hierarchical clustering, self-organizing maps, fuzzy k-means, etc. Hundreds of ML algorithms have been tested by Mahout Libraries, Best ML algorithms are automatically selected by Algorithm Selection Optimizer to address the specific needs of users.

9.1 Machine Learning for Big Data Analytics

Machine learning techniques are very effective and relevant to many real-world applications such as network security, bioinformatics, healthcare, banking and finance, and transportations. The bioinformatics and health-related data are created and accumulated continuously, resulting in a large volume of data. Newer forms of big data, such as 3D imaging, genomics, and biometric sensor readings are also fueling this exponential growth. Future applications of real-time data, such as early detection of infections/diseases and fast application of the appropriate treatments (not just broad-spectrum antibiotics) could reduce patient morbidity and mortality. The ability to perform real-time analytics against such big stream data across all fields would help seed type prediction.

According to [21], there are two types of learning methods in machine learning, these are supervised and unsupervised learning methods. In supervised learning, a method learns from a set of objects with a class label, often called a training set. The learned knowledge is used to assign a label to unknown objects often called test objects. Whereas unsupervised learning methods do not depend on the availability of prior knowledge or training instances with class labels. Both of methods require preprocessing of datasets for actual results.

10. Hadoop Ecosystem & Tools

A big data platform involves a number of components and is often called a big data ecosystem. As [22] notes, a big data platform could be related to an oil refinery: an important number of inputs, output, processes, and the structures guides the process of moving, storing, and analyzing data.

Hadoop ecosystem includes a set of tools that function near MapReduce and HDFS (the two main Hadoop core components) and help the two store and manage data, as well as perform the analytic tasks. As there is an increasing number of new technologies that surround Hadoop, it is important to realize that certain products may be more appropriate to fulfill certain requirements than others. Hadoop core technologies provide a fault tolerance mechanism to store large datasets. *Hadoop Distributed File System (HDFS)* is where data are stored. Data files are fragmented into blocks and are distributed over the servers. It is designed to run on several clusters and to be resilient to failures since it makes several copies of its data blocks. MapReduce, one the other hand, is a paradigm to process data. It was the first programming method to develop applications in Hadoop, comprising of two programs written in Java: *Mappers*, to extract data from HDFS and put into maps, and *Reducers* to aggregate the results produced by the mappers.

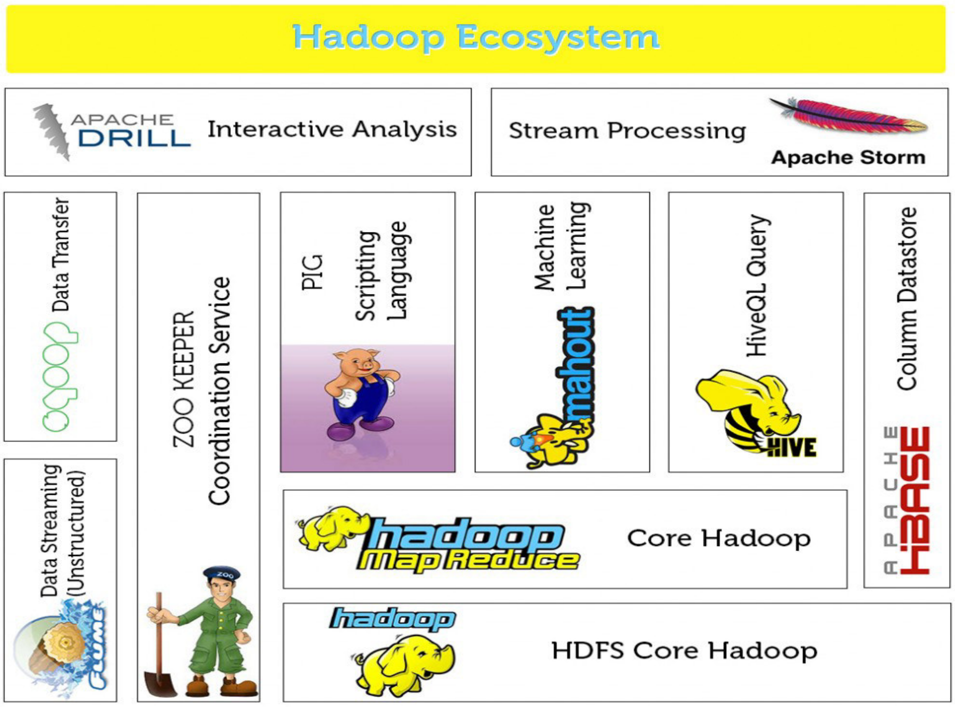

Figure 6: Hadoop Ecosystem [23]

11. Predictive Modeling

While we build our model we used predictive models method since it is an iterative process in which a model is created from an initial hypothesis and then refined until it produces a valuable business outcome or discarded in favor of another model with more potential [24]. The other advantage of using Predictive analytics is, it is also useful to incorporate the model for more analytical insights. Tasks we have performed for predictive model development is depicted in figure 9 below.

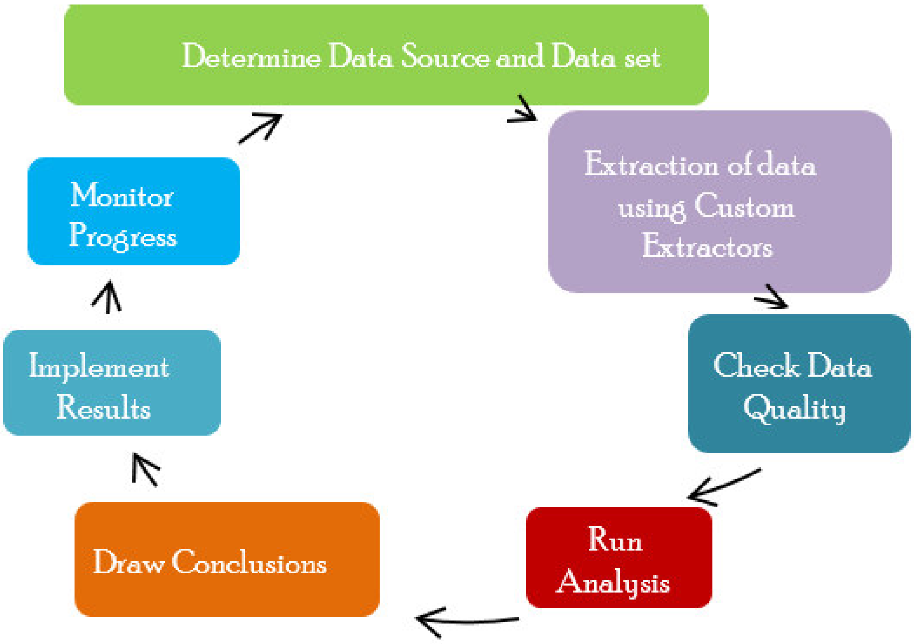

Figure 7: Predictive Analytics Process [24]

Conclusion

The entry of Machine Learning in Agriculture has provoked to take activities to misuse its respect and application in Seed Prediction. This paper depicted the attributes of Machine Learning and exhibited a Framework for Machine Learning. Enormous Data innovation goes astray from customary information administration SQL-based RDBMS approaches as it manages information with high volume, speed and assortment. This paper outlines the meteorological information stockpiling and in addition examination stage dependent on Hadoop system with various calculations for forecast. This stage depends on conveyed record framework HDFS which fuses appropriated database Hbase, information distribution center administration and productive inquiry handling device Hive, information movement instrument sqoop. The best information mining expectation calculation relapse likewise incorporated into the framework. It is conceivable to utilize Artificial Intelligence (AI) to create models that can be utilized in expectation of seed determinant factors.

Reference

[1] J. O. Chan, "An Architecture for Big Data Analytics," 2013, vol. 13, no. 2, pp. 1–14.
[2] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, pp. 114–117, 1965.
[3] A. Jacobs, "The pathologies of big data," *Commun. ACM*, vol. 52(8), pp. 36–44, 2009.
[4] I. Gartner, "Gartner says solving 'big data' challenge involves more than just managing volumes of data [Press release].," 2011. [Online]. Available: http://www.gartner.com/ newsroom/id/1731916. [Accessed: 27-Jun-2017].
[5] A. Inshwarappa, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology," in *International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015)*, 2015, pp. 319–324.
[6] Prof. K.Suzuki, "Diagnosing Skin Diseases using an Artificial Neural Network," in *Artificial Neural Networks- Methodological Advances and Biomedical Applications*, Shanghai, InTech, p. 374.
[7] D. Graupe, "principles of artificial neural networks," *World Sci. Publ. Co. Pte. Ltd.*, 2007.
[8] K. Suzuki, *ARTIFICIAL NEURAL NETWORKS METHODOLOGICAL ADVANCES AND Edited by Kenji Suzuki*.
[9] K. Girish, "Artificial Neural networks and its applications," *IARI*, vol. 110, 2012.
[10] Prof. K.Suzuki, "Introduction to Artificial Neural Networks," in *Artificial Neural Networks-Methodological Advances and Biomedical Applications*, p. 374.
[11] Q. K. Al-Shayea and I.S. H. Bahia, "Urinary System Diseases Diagnosis Using Artificial Neural Networks," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 10 No.7.
[12] M. Beale et al., "Neural Network Toolbox™ User's Guide," 2006.
[13] D. M. M. H.S. N. Murthy, "ANN Model to Predict Coronary Heart Disease Based on Risk Factors,"

Bonfring Int. J. Man Mach. Interface, vol. 3, No.2.
[14] C. V. S. D. valley, "Diagnosis Chest Diseases Using Neural Network and Genetic Hybrid Algorithm," *Int. J. Eng. Res. Appl. ISSN 2248*, vol. 5, no. 1(Part 2), pp. 20–26, 2015.
[15] S.Radhimeenakshi and G.M.Nasira, "Prediction of Heart Disease using Neural Network with Back Propagation," *Int. Conf. Comput. Intell. Syst.*, vol. 04, no. Special Issue, pp. 1166 – 1169, 2015.
[16] Dr. T.P. Shah and P.J. Shah, "Connectionist Expert System for Medical Diagnosis using ANN– A case study of skin disease Scabies," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 8, pp. 227–230, 2013.
[17] Mr. A.Vishwa, "pre-diagnosis of lung cancer using feed forward neural network and backpropagation algorithm," *Int. J. Comput. Sci. Eng.*, vol. 3 No.9, pp. 3313–3319, 2013.
[18] P. Fayyad, Usma, Piatetsky-Shapiro, G., and Smyth, "From Data Mining to Knowledge Discovery an overview.," in *In advances in knowledge discovery and data mining*, 1996.
[19] Giudici P., *Applied Data mining Statistical Method for Business and Industry*. The University of Pavia. Italy.
[20] H. H. Brynjolfsson, E., Hitt, L. M. & Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance" 2011. [Online]. Available: http://ssrn.com/abstract=1819486. [Accessed: 10-Feb-2013].
[21] D. K. Bhattacharyya and J. K. Kalita, "Network anomaly detection: A machine learning perspective," CRC Press, 2013.
[22] H. Inc., "Apache Hadoop: the big data refinery," 2012. [Online]. Available: http://info.hortonworks.com/Big-Data-Refinery-White-Paper.html.
[23] G. Bharadwaja Kumar, "An encyclopedic overview of 'big data' analytics," *Int. J. Appl. Eng. Res.*, vol. 10, no. 3, pp. 5681–5705, 2015.
[24] P. M. S. P. Babu, "Big Data and Predictive Analytics in ERP Systems for Automating Decision Making Process."