International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

Volume 9, Issue 11 - November 2021- Pages 1-14

# AGRICULTURE CROP YIELD PREDICTION USING ARTIFICIAL NEURAL NETWORK

*1.Prof Sangamesh S K, 2. Prof Pavitra M G, Prof S T Halakatti*
**Department of Computer Science and Engineering,**
**RTE Society's Rural Engineering College Hulkoti**

## 1.INTRODUCTION

Agriculture is the basic source of food supply in all the countries of the world whether under-developed, developing or developed. Besides providing food, this sector has contributions to almost every other sector of a country. According to the Bangladesh Bureau of Statistics (BBS), 2017, about 17 % of the country's Gross Domestic Product (GDP) is a contribution of the agricultural sector, and it employs more than 45% of the total labor force. In light of the decreasing crop production and shortage of food across the world, one of the crucial criteria of agriculture now-a-days is selecting the right crop for the right piece of land at the right time. Therefore, in our research we have proposed a method which would help suggest the most suitable crop(s) for a specific land based on the analysis of the data of previous years on certain affecting parameters using machine learning. In our work, we have implemented Random Forest Classifier, Gaussian Naïve Bayes, Logistic Regression, Support Vector Machine, k-Nearest Neighbor, and Artificial Neural Network for crop selection. We have trained these algorithms with the training data and later these were tested with test dataset. We then compared the performances of all the tested methods to arrive at the best outcome.

For a country, one of the most crucial aspects of its development circles around its capacity to produce food. For decades, agriculture has been associated with the production of essential food crops. The rate of urbanization at present is by-far the most superior aim of our civilization. In doing this, we are ignorantly diminishing our capacity for agriculture; especially in terms of land and fertility. As the amount of land will not be increasing in thisera of urbanization and globalization, we will have to focus on making the most of what we have. Due to this issue, we have to device newer ways to farm arable lands and extract the absolute most from these limited land resources. In this age of technology and data- science, if implemented properly, the agricultural sector may also be greatly affected. It is true that a farmer is the best decider of crop selection and crop cultivation. However, machine learning techniques can be applied in this field for far greater precision and stability of selection. In this research, we have attempted to come up with a few techniques that will lead us to choose suitable crops based on specific state, specific district, season, and some other environmental aspects. By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to

overcome the situation faced by us. In India there are several ways to increase the economic growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining also useful for predicting the crop yield production.

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is an analytical tool that allows users to analyze data from many different dimensions or angles, categorize, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The patterns, associations, or relationships among all this data can provide information. Information can be converted into knowledge about historical patterns and future trends. For example, summary information about crop production can help the farmers identify the crop losses and prevent it in future.

Crop yield prediction is an important agricultural problem. Each and Every farmer is always trying to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield is primarily depending on weather conditions, pests and planning of harvest operation. Accurate information about history of crop yield is an

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

important thing for making decisions related to agricultural risk management. This research focuses on evolution of a prediction model which may be used to predict crop yield production. The proposed method uses data mining technique to predict the crop yield production based onthe association rules.

Data Mining is emerging research field in crop yield analysis. Yield prediction is avery important issue in agricultural. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. The yield prediction is a major issue that remains to be solved based on available data. Data mining techniques are the better choice for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. This research proposes and implements a system to predict crop yield from previous data. This is achieved by applying association rule mining on agriculture data. This research focuses on creation of a prediction model which may be used to future prediction of crop yield. This paper presents a brief analysis of crop yield prediction using data mining technique based on association rules for the selected region. The experimental results show that the proposed work efficiently predict the crop yield production.

Achieving maximum crop yield at minimum cost is one of the goals of agricultural production. Early detection and management of problems associated with crop yield indicators can help increase yield and subsequent profit. By influencing regional weather patterns, large-scale meteorological phenomena can have a significant impact on agricultural production. Predictions could be used by crop managers to minimize losses when unfavorable conditions may occur. Additionally, these predictions could be used to maximize crop prediction when potential exists for favorable growing conditions. Prediction of crop yield mainly strategic plants such as wheat, corn, rice has always been an interesting research area to agro meteorologists, as it is important in national and international economic programming.

Dry farming crop production, apart from relationship to the genetic of cultivator, adipic terms, effect of pests and pathology and weeds, the management and control quality during the growing season and etc. is severely depend to climatic events. Therefore, it is systems which can predict the more accuracy using meteorological data. Nowadays, thereare a lot of yield prediction models, that more of them have been generally classified in twogroup: a) Statistical Models, b) Crop Simulation Models (e.g. CERES). Recently, application of Artificial Intelligence (AI), such as Artificial Neural Networks (ANNs), Fuzzy Systems and Genetic Algorithm has shown more efficiency in dissolving the problem. Application of them can make models easier and more accuracy from complex natural systems with many inputs. In this research it has been tried to develop a various crop yield prediction model using it can be used to estimate crop production in long or shortterm and also with enough and useful data can get an ANNs.

## 1.1 Problem Statement

Agriculture is the main occupation of the majority of population. The farmers of thedistrict rely heavily on agriculture for earning their livelihood. The development ofagriculture depends on various aspects such as type of soil, relief, vegetation, climatic conditions, attitudes of different social groups of farmers to agriculture, use of irrigation, HYV seeds, fertilizer, pesticides and insecticides, use of mechanical tools and implements,as well as proper scientific rotation of crops by which production be enhanced. The impact of these aspects of agriculture varies in different areas of the district. There are distinct variations in the magnitude of these concepts both over space and time. To have real understanding of the nature of agricultural development, scientific investigation and evaluation of different aspects of development become highly necessary. Keeping these points in view, the department of agriculture research has been selected as the study area because there has been significant development in agriculture in the district in the post- independence era. The level of agricultural development is not the same in all district whichis inhabited by various social groups of people. This is because they live in different geographical areas and their attitudes to agriculture are different. The five social groups among various groups, viz. the indigenous Hindus, the indigenous

## 1.2 OBJECTIVES

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

Our project is to predict the maximum yield of the crops produced at minimum cost. Early detection and management of problems associated with crop yield indicators can help increase yield and subsequent profit. By influencing regional weather patterns, large-scale meteorological phenomena can have a significant impact on agricultural production.

- To study the socio-spatial and temporal variations of agricultural land use pattern.
- To investigate the pattern of agricultural productivity, intensity of cropping, crop diversification and rotation of crops.
- To assess the contribution of various social groups to the agricultural changes in the region and examine the controlling factors behind such changes.

## 2. LITERATURE SURVEY

Machine learning is the branch of computer science which is used to build algorithms which exhibit self-learning property i.e. learning which is done by the machine itself hence the term "Machine Learning". It is viewed as one of the significant areas under Artificial Intelligence. To show intelligence machine needs to interpret and analyze the input. After analyzing it the result data apart from simply following the instructions on that data. This is the thing that machine learning algorithms do. Machine learning centers on the development of computer programs that can get data and utilize it to learn for themselves. The way toward learning starts with perceptions on information, for example, direct experience, or instruction, in order to look for patterns in data. It helps to make better decisions in the future based on the examples that we give. The essential point is to permit the computer learn automatically without human intercession or help and regulate actions consequently. It is a major field in computer science which is being utilized in various forms of progressive technological development programs all around the world. It is regarded as the future of engineering and Artificial Intelligence [12].

In another research, the author Enfield D.B [6] utilizes Machine learning in different applications in Indian agriculture. In this paper the various applications of machine learning techniques in agriculture have been listed such as Crop Selection and Crop Yield Prediction, Weather Forecasting, Smart Irrigation System, Crop Disease Prediction, Deciding the Minimum Support Price. These techniques will enhance the productivity of fields along with a reduction in the input efforts of the farmers. Along with the advances in machines and technologies used in farming, useful and accurate information about different matters also plays a significant role in it [5]. Machine learning provides many effective algorithms which can identify the input and output relationship in crop selection and yield prediction. To this issue of crop selection using machine learning some researches use Markov logic of crop rotations for early crop mapping. A Markov logic network (MLN) is a probabilistic logic which applies the ideas of a Markov network to first-order logic, enabling uncertain inference. The obtained results show that the proposed approach is able to predict the crop type of each field, before the beginning of the crop season, with an

accuracy as high as 60%, which is better than the results obtained with the previous approaches based on remote sensing imagery [10].

The authors S. Ying-xue, X. Huan, and Y. Li-jiao in their research work describes a crop model as a computer program that mathematically describes and models the principles of harvest growth and can be utilized to quantitatively and dynamically clarify the procedure of product development, improvement, yield, and response to ecological change[13]. Crop models can be classified as harvest factual models and crops imulation models (or crop growth models) in light of the fundamental numerical technique for the displaying. Furthermore, in another paper, the authors used Auto-regressive Integrate Moving Average (ARIMA) and The Support Vector Machines (SVMs). Various examinations were performed on the advancement of SVM and the most accuracy show by utilizing statistical criteria was additionally selected. The proposed model of Support

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

Vector Machine (SVM) can conjecture nonlinear or linear forecasting function upon kernelfunction. These are a few of the ideas which have been additionally connected to increase better estimating of farming utilizing machine learning. An UchooBoost Algorithm is utilized for testing with exactness farming. It is supervised learning based on algorithm. The best characteristic of UChooBoost is that it can be applied for an extended data expression and works on compounding hypotheses which leads to improve algorithm performance [9].

Agriculture planning performs a widespread role in financial boom and food security of agro-based country. Selection of crop(s) is a crucial trouble for agriculture planning. It relies upon different parameters which includes production rate, market price and government approaches. Using statistics methods or machine learning techniques numerous scientists research prediction of yield rate of crop, prediction of weather, soil classification and crop classification. If there is in excess of one alternative to plant a product at any given moment utilizing constrained land asset, at that point determination ofharvest is a riddle. However, this paper proposed a method named Crop Selection Method(CSM) can be applied to solve this crop selection problem and maximize net yield rate of crop over season and subsequently achieve maximum economic growth of the country. Crop selection method refers to a method of selecting crop(s) over a specific season depending upon various environmental as well as economic factors for the maximum benefit. These factors are precipitation levels, average temperature, soil type, market prices and demand, prevailing farm conditions, crop or varietal adaptability, resistance to pests and diseases, farming system, available technology etc. Right choice in the selection of

crop or crops to be grown, particularly perennial types, will eventually convert into a successful cultivating venture [11]. This task can be completed using Classification algorithms of WEKA. Waikato Environment for Knowledge Analysis (Weka) is a site of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. In some researches, WEKA Classifiers and regression methods are used to precisely predict the most suitable crop(s) to be grown in a particular season [11]. There are many more features such as humidity, soil nutrition value, pH etc., which are included in the training dataset.

B.A. Smith et al [3] discuss year-round air temperature prediction models were developed for prediction horizons of 1 to 12 h using Ward-style ANNs. These models were intended for use in general decision support. The ANN design modifications described herein provided increased accuracy over previously developed, winter specific models during thewinter period. It was shown that models that included rainfall terms in the input vector weremore accurate than those that did not.

D.L. Ehret et al [5] introduce all crop attributes responded in much the same way to individual climatic factors. Radiation and temperature generally induced strong positive responses while RH produced a negative response. In the NN models, radiation and temperature were still prominent, but the importance of CO2 in predicting a crop response increased. One advantage of these automated systems is that they offer continuous information across a range of timescales. Furthermore, these systems can readily be used in commercial greenhouses so the derived NN models are relatively easy to deploy to a commercial setting where they can subsequently be improved over time. In this paper cropprediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. Parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity. For that purpose, we are used artificial neural network (ANN). This project shows the ability of artificial neural network technology to be used for the approximation and prediction of crop yields at rural district.

B. J I ET AL [2] developed agricultural management need simple and accurate estimation techniques to predict rice yields in the planning process. The necessity of the present study were to: (1) identify whether artificial neural network (ANN) models could effectively predict rice yield for typical climatic conditions of the mountainous region, (2) evaluate ANN model performance relative to variations of developmental parameters and(3) compare the effectiveness of multiple linear regression models with ANN models.

In this paper describes the development of artificial neural network models as an alternate and more accurate technique for yield prediction.

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

A Ghazi Zadeh, A Fahim, M El–Gindy International Journal of Vehicle Design 18(2), 132-193, 1997 Recent developments in the application of the artificial neural networks(NN) and fuzzy logic (FL) have attracted the attention of many researchers in the area of vehicle dynamics and control. Neural networks are able to emulate the solution of different classes of nonlinear algebraic equations and differential transfer functions. Fuzzy logic interface systems can map those functions that have no equivalent mathematical model or whose mathematical models are very complicated. A large number of studies have been published on the application of neural networks and fuzzy logic interface systems to vehicle dynamics and control. In this paper, an extensive literature survey of more than forty papersand reports, published during the last five years, has been conducted. Reviewed papers cover different subjects including: vehicle motion control, driver modelling, tyre modelling, braking control, suspension control, steering system, transmission control, and engine control. This literature review is part of an ongoing research project related to the application of neural networks and fuzzy logic to vehicle dynamics and control.

I Usu, F.Ursu, T.Sireteanu, CW Stammers Shock and Vibration Digest 32 (1), 3-10, 2000. This paper is concerned with the synthesis of control laws for semi active suspension systems employing artificial intelligence. A review is made of a simple, 2- degree-of-freedom, quarter car model with passive, active, and semi active control. An active linear quadratic Gaussian controller and a semi active, balance logic derived controller are then used to develop two artificial intelligence based controllers—a semi active neuro-controller and a semi active fuzzy logic controller. This paper focuses on the advancement of balance semi-active logic using variable dry friction and the development of fuzzy semi-active controllers. The concerns in view are twofold: the reduction in cost ofthe control system and the anti-chattering nature of the logic. The development is from an engineering perspective and attempts to reduce the well-known schism between theoreticians and users of feedback controls.

# 3.METHODOLOGY

In this paper crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. Parameters liketype of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity. For that purpose, we are used artificial neural network (ANN). This project shows the ability of artificial neural network technology to be used for the approximation and prediction of crop yields at rural district.

## 3.1 NEED OF CROP PREDICTION

Prediction of crop yield mainly strategic plants such as wheat, corn, rice has alwaysbeen an interesting research area to agro meteorologists, as it is important in national and international economic programming. Dry farming crop production, apart from relationshipto the genetic of cultivator, adaphic terms, effect of pests and pathology and weeds, the management and control quality during the growing season and etc. is severely depend to climatic events. Therefore, it is not beyond the possibility to acquire relations or systems which can predict the more accuracy using meteorological data. Nowadays, there are a lotof yield prediction models, that more of them have been generally classified in two group

a) Statistical Models, b) Crop Simulation Models (e.g. CERES). Recently, application of Artificial Intelligence (AI), such as Artificial Neural Networks (ANNs), Fuzzy Systems and Genetic Algorithm has shown more efficiency in dissolving the problem. Application of them can make models easier and more accuracy from complex natural systems with many inputs. In this research it has been tried to develop a wheat yield prediction model using ANNs. If we design a network which correctly learn relations of effective climatic factors on crop yield, it can be used to estimate crop production in long or short term and also with enough and useful data can get a ANNs model for each area. Furthermore, using ANNs can find the most effective factors on crop yield. Therefore, some factors that their measurements are difficult and cost effective can be ignored. In this the effect of climatic factors on wheat yield has only been applied.

## 3.2 PROPOSED WORK

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

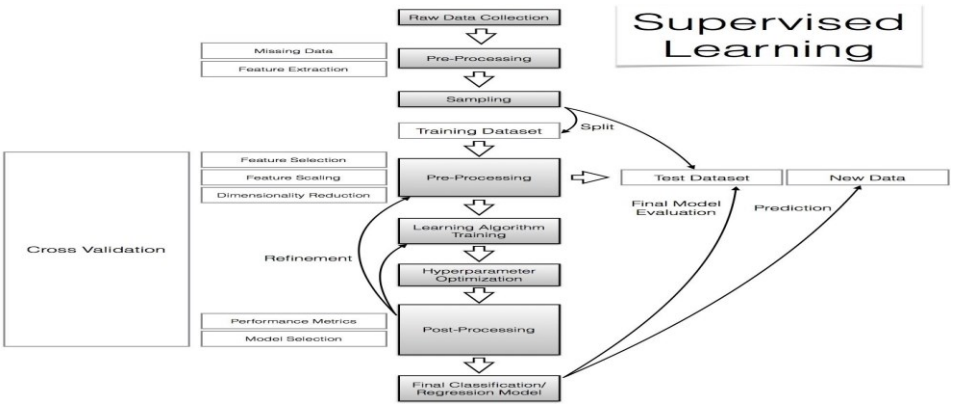
### 3.2.1 Supervised Learning

Supervised machine learning is the search for algorithms that reason from externallysupplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise modelof the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown. It simply is the task of learning a function that maps an input to an output based on input-output pairs of examples.A supervised learning algorithm analyzes the training data produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

Supervised learning models have some advantages over the unsupervised approach, but they also have limitations. The systems are more likely to make judgments that humans can relate to, for example, because humans have provided the basis for decisions. However, in the case of a retrieval-based method, supervised learning systems have trouble dealing with new information. If a system with categories for crops and fruits is presented with a flower, for example, it would have to be incorrectly lumped in one category or the other. Ifthe AI system was generative, however, it may not know what the flower is but would be able to recognize it as belonging to a separate category.

### 3.2.2 Data Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify a particular crop as to the possibility of its production given a set of features, as per our research. A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts crop selection could be developed based on observed data for many crop selections over a period of time. In addition to the crop data, the data might track a number of features. A specific crop would be the target, the other attributes would be the predictors, and the data for each crop would constitute a case. Classifications are discrete and do not imply order. Continuous, floating point valueswould indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm. The simplest
type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, rice or no rice. Multi-class targets havemore than two values: for example, low, medium, high, or unknown possibility of rice production. In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.



International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

**Figure 3.1 Block diagram of our model.**

Multiple pathways can be taken for a research of this sort. The field of machinelearning is vast enough to accommodate a great number of ways this type of predictionwork can be done.

For our model, we have implemented all the algorithms and data processing codes using the Python programmable language. The Integrated Development Environment (IDE) that we have used is Anaconda Spyder, during our initial data processing phases. Ourmodel consists of all the aforementioned algorithms which has led us to propose a comparison between all of them, based on performance. There are 8 steps that we have followed in our model:

- Dataset collection.
- Data visualization.
- Data pre-processing in the form of data cleaning and feature extraction.
- Data splitting into train and test sets.
- Fitting the algorithm.
- Parameter tuning (only for Artificial Neural Network).
- Testing the accuracy of the model.
- Data post-processing in the form of performance metrics.

## Data Collection

For a research of this sort, it is crucial to have an available dataset to work upon. It is very difficult to find legible and reliable datasets of this sort. It took us a lot of time and effort to find one suitable for us. The dataset that we finally found contained all the accurate features that we really wanted. The features were perfect for a research of this sort. There was a total of 12 columns and around 250,000 rows. The columns were “State Name”, “District Name”, “Crop Year”, “Season”, “Area”, “Rainfall”, “Humidity”, “Temperature”, “Previous Year’s Rainfall”, “Previous Year’s Humidity”, “Previous Year’s Temperature” and “Crop”. 4 of the columns contained data which were in string notation. The rest of thecolumns contained data which were numerical. The “State Name” column had the names of a number of important states in the country. The “District Name” column had names ofdistricts in those states. The “Crop Year” column had the crop years for the past 19 years. “Season” contained 4 different seasons. “Area” had area data in meters squared. “Area”, “Rainfall”, “Humidity”, “Temperature”, “Previous Year’s Rainfall”, “Previous Year’s Humidity”, and “Previous Year’s Temperature” all contained data respective of the namesjust mentioned. Finally, the “Crop” column, our most important column, contained 135 different crops which were grown in the places mentioned in the respective columns. All this data is for India, and were taken from the Indian Government Agricultural website upon email request. A research in the field of machine learning, especially deep learning within machine learning, requires heavy usage of computational power. In our research, we
had to use Kera’s and Tensor flow as a back-end engine which actually supports the Scikit-learn library. Scikit-learn is basically run by Kera’s. Furthermore, as our dataset contained a massive number of rows, algorithms such as the Artificial Neural Network algorithm, required high computational power.

## Data Pre-Processing

One of the primary tasks we completed was to convert all of our string data into numerical data. In order to do this, we converted all the string features into dummy variables. This greatly increased our column number. We then cleaned our data in a very singular pattern. We had a small portion of null values in the production column. Due to the miniature amount, we dropped the fields off of the dataset. This did not affect much, asthe amount was minimal. We also performed feature extraction on the dataset. Some of ourdata were categorical and some were continuous numeric. This type of mixed data always causes problems to the algorithms. Hence, we performed standard feature scaling on all ofthe data to bring them into a common scale.

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

Feature scaling is extremely important due to the fact that, most algorithms feature a lot of internal calculation. Additionally, feature selection was done to reduce over fitting issues.

## Data Splitting

Data splitting is the process of splitting the dataset into training and testing data. This process is very useful for any machine learning process as the main idea of machine learning depends on training and testing data and finding the accuracy of the machine given result. In our research, we divided our dataset because we trained our algorithms on the testdataset where the particular crop had its data in there. Here the algorithms were trained using that data. We deduced that data having 1 to be yes, and 0 to be no. The algorithms were trained and we will apply the trained algorithm to our test set and measured the accuracy of the machine. The datasets are usually divided into an 80:20 ratio. However, forour model the dataset was split into both 80:20 and 60:40 ratio. Thus, the 80%/60% of theselected data was chosen as training set and the remaining 20%/40% was test set. There aremany built in python tool-kits for splitting data such as, 'pandas', 'keras', 'scikit-learn' etc.although we used "scikit-learn" for the machine learning approaches in this research because of its built-in libraries.

## Algorithm Fitting

The most crucial part of the model was to fit the algorithm with the data. All the algorithms were easily fitted as the programming of this part was comparatively easy. Simple method callings were all that were required. The algorithms, upon being
implemented, processed all the data using all the internal calculations. Data frames were created and could be viewed in the variable explorer. Because of the fact that the data hadbeen split into training and testing datasets, the algorithm could start the core process: learning. The machine learned from the train set. This learning was to be used later on whilepredicting from the test set. Fitting is similar to training. For example, let us assume that we have measured the production for a group of crops and decided that your model would be a normal distribution. Then determining the mean and variance of the normal distribution that best explains our observed data is called fitting: we are determining the parameters' mean $\mu$ and variance $\sigma$. Suppose we have an algorithm that estimates $\mu$ and $\sigma$ given our data. We can ask our algorithm to run "n" iterations where each iteration takes the same amount of time but more iterations yield slightly more accurate estimates of the parameters $\mu$ and $\sigma$. "n" is a value we supply that may influence the estimates and is calleda hyper-parameter.

## Testing Accuracy

To test the accuracy, we implemented different methods on different algorithms based on requirement. Some were direct accuracy-check method calls from scikit-learn libraries. While in some other algorithms, we implemented manual accuracy checks, againbased on the algorithm itself. In Random Forest of instance, mean was calculated. In Artificial Neural Network, despite calling an accuracy-check method, all the accuracy fromall the epochs were taken in for mean value of accuracy. The accuracy check is crucial in understanding the viability of the algorithms and also the research itself. A very low accuracy in all the algorithms would mean the entire research was a dead end. It would mean this method altogether is not viable for this research. A low accuracy in a few algorithms and a high accuracy in the others would mean the ones with the low accuracy are not efficient in this model, but the others are. We would have discarded the low accuracy yielding algorithms. However, in our case, all the algorithms yielded a very high accuracy. Although this issue did cause us a few problems later on when comparing the accuracy amongst all the algorithms, the fact that the accuracy is high on all, was a strong point in determining the methods to be viable in this field of research. An Effective Model is a model which basically predicts the testing data most accurately as compared to other models and hence, can be deployed successfully. Testing accuracy of k-NN is shown in the

## Data Post-Processing

After all the accuracy have been taken into account, a few other data processes can still be implemented. This part of the model is not necessary for the primary target of

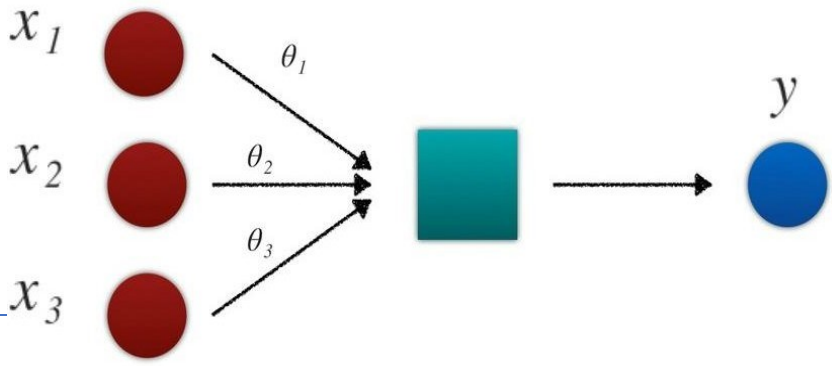
International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

the research, but we still used it for certain confirmation purposes. We implemented a method which would create a confusion matrix. A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if we have more than two classes in our dataset. Calculating a confusion matrix can give us a better idea of what our classification model is getting right and what types of errors it is making. The methods to reduce any over fitting issues that were implemented earlier in the data pre-processing stage, can even be implemented here instead the other stage. However,it was extremely necessary for us to use those in the model, any form of over fitting or under fitting would actually render the whole model lack of any real use.

## 3.3 Algorithms for Proposed Model

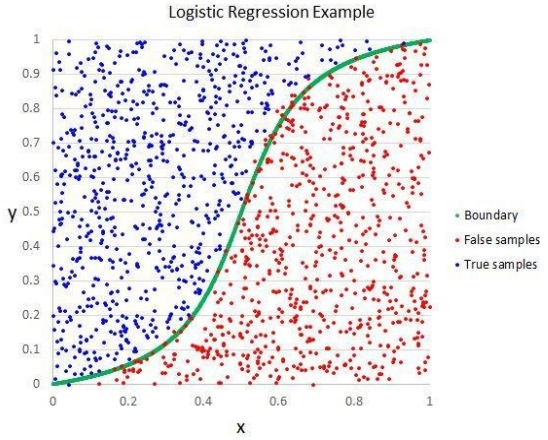
### 3.3.1 Logistic Regression

Methods involving regression are essential to any data analysis models which attempt to describe the association between a response variable and any number of predictor variables. Situations involving discrete variables constantly arise. For instance, the datasetwe have implemented has an outcome involving the presence or absence of a particular crop, given a set of features. Logistic regression analysis extends the techniques of multiple regression analysis to investigate and inquire situations in which the outcome is categorical, which is, taking on multiple values. This is a very basic branch of data science. Although the name suggests a regression technique, logistic regression is a statistical classification model which deals with categorical dependent variables. Classification is decision. To make an optima decision we need to assess the utility function, which implies that we need to account for the uncertainty in the outcome, i.e. a probability. Logistic regression is emphatically not a classification algorithm on its own. It is only a classification algorithm in combination with a decision rule that makes dichotomous the predicted probabilities of
the outcome. This is one of the very first algorithm any machine learning practitioner attempts when faced with a classification problem. The basic mechanism and output of this algorithm is similar to many other machine learning algorithms. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous or binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
This algorithm works with binary data, where either the event happens, represented by “1”,or the event does not happen, represented by 0. So given some feature “X”, it tries to find out whether some event “y” happens or not. So “y” can either be “0” or “1”. In the case where the event happens, “y” is given the value “1”. If the event does not happen, then “y” is given the value of “0”. For example, if “y” represents whether a particular crop among a huge variety of crops, then “y” will be “1” if the crop does grow or “y” will be “0” if it doesnot. This is known as Binomial Logistic Regression. There is also another form of Logistic Regression which uses multiple values for the variable “y”. This form of Logistic Regression is known as Multinomial Logistic Regression. Figure 3.3 shows a simple flowchart representation of logistic regression.



International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

**Figure 3.3 Flow Chart of Logistic Regression**




Logistic Regression uses the logistic function to find a model that fits with the datapoints. The function gives an "S" shaped curve to model the data. The curve is restricted between "0" and "1", so it is easy to apply when "y" is binary. Logistic Regression can then model events better than linear regression, as it shows the probability for "y" being "1" fora given "x" value. Logistic Regression is used in statistics and machine learning to predictvalues of an input from previous test data. Scatter plot classification of data by Logistic Regression is shown in Figure 3.4

**Figure 3.4 classification of data by Logistic Regression.**

## 3.3.2 Support Vector Machine (SVM)

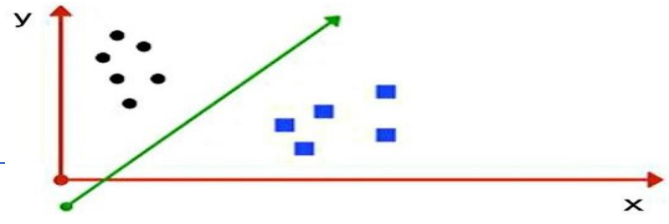
"Support Vector Machine" (SVM) is a supervised machine learning algorithm which is a very useful technique for data classification. However, this learning algorithm can also be used for regression challenges. A classification task usually involves separatingdata into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

## Linear SVM

Support Vector Machine classifier plots each data item with the value of each feature as a point in an n-dimensional space (where n is number of features) being the valueof a particular coordinate. SVM maps data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. Then, itperforms classification by finding the hyper-plane that differentiates the two classes very well. points can be categorized, even when the data are not otherwise linearly separable.
Then, it performs classification by finding the hyper-plane that differentiates the two classes very well.

When the data can be linearly separated in two dimensions, as shown in Figure 3.5,any machine learning algorithm tries to find a boundary that divides the data in such a way

that the mis-classification error can be minimized. Nevertheless, there can be several boundaries that correctly divide the data points as shown below in Figure 3.6.



International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

**Figure 3.5 Classifications of Two Classes Using Boundary.**

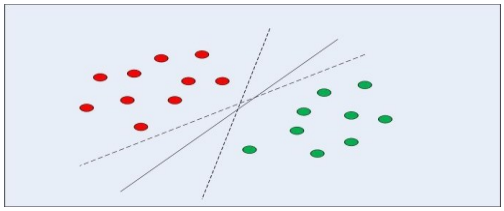


**Figure 3.6 Multiple Decision Boundaries.**

SVM is different from the other classifiers in the way that it chooses the decision boundary that maximizes the distance from the nearest data points of all the classes. This boundary has the maximum margin from the nearest points of the training class as well as the test class. As a result, SVM classifier does not only find a boundary; it finds the most optimal decision boundary. This boundary resulting from SVM is called the maximum margin classifier, or the maximum margin hyper plane. The nearest points from the hyper plane that maximize the distance between the decision boundaries are called support vectors.

## 3.4 Result Analysis

The core aim of our research was to establish a model that will efficiently predict a particular crop based on a set of features. During the course of this research, we have also successfully created a model that can predict the possibility of a particular crop to be able to be grown, given a set of features. In order to do this, as mentioned above, we have

implemented 6different machine learning algorithms. Amongst them, The SVM was done using 3 different aspects, and the KNN was done using 2. This ensured us the ability to bring a comparison between varieties of different algorithms. Our target was to establish the best performing algorithm for this field of work, based on our data.

## 3.5 Accuracy Analysis

As mentioned before, we have extracted accuracy from 9 different processes. Only the ANN was run on one instance only. Apart from that, all the other algorithms were tested on a variety of instances of the same dataset. We took 5 samples of our primary dataset. The samples had an increasing number of rows starting from 2000 and ending at 11000. We ran each algorithm on these samples in 2 separate train test splits. Initially, we used the 60%-40%train test ratio. Eventually, we changed that to 80%:20%. Both gave us fair results, but the latter gave us a better accuracy, which led us to finalize the model on that. Two separate outcomes were extracted from our algorithms with slight tweaking. We could both predict a particular crop and the percentage chance of a specific crop to be able to be grown, given a particular set of features. However, the former method fetched bad accuracy levels, which we found highly unconvincing. The highest accuracy we got was from the KNN (K=optimal) algorithm, which was 81.3%, on a set of 11000 features. In similar features, but with a specific crop as the dependent variable, all the algorithms gave strong accuracy levels. The lowest accuracy was given by Random Forest Classifier which was still 92.30%. The highest accuracy was again given by the KNN (K=optimal) algorithm. The ANN was tested only once with 11000 features, and only for the specific crop model. We trained and tested the network up to 1000 epochs. The accuracy that we obtained was 96.95%. All the accuracy levels in this part of the thesis indicated to a strong viability of our research in this field. In the (Table. 3.1) and (Table. 3.2) below we have shown

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

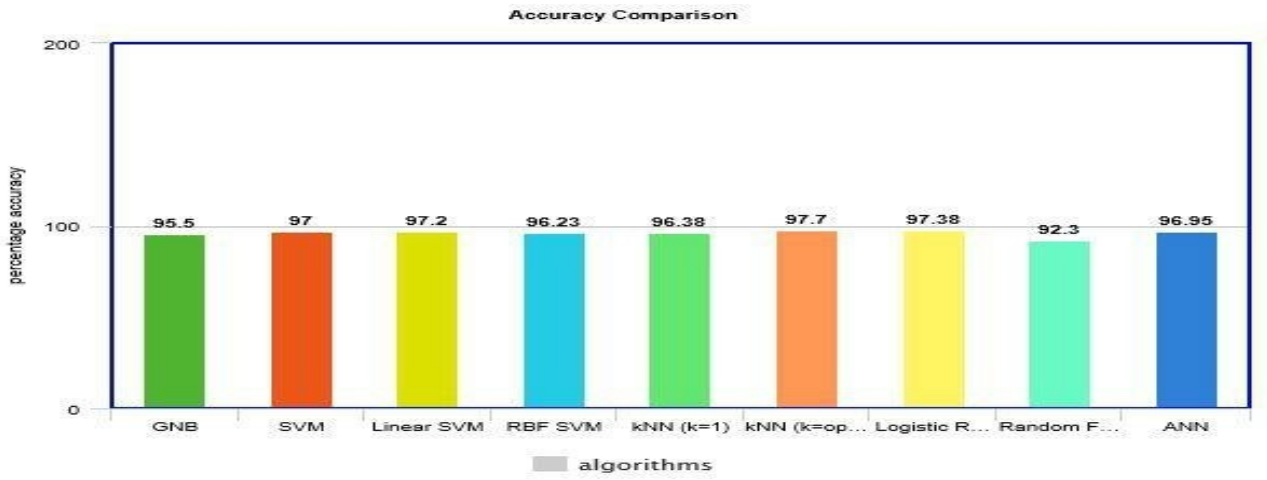
the comparison among all the classifiers we have implemented for specific crop possibility prediction and crop prediction respectively.

**Table 3.1 Accuracy comparison for crop possibility**

| Data | GNB | SVM | Linear SVM | RBF SVM | KNN (k=1) | KNN (k=optimal) | LR | RF |
|---|---|---|---|---|---|---|---|---|
| 2000 | 70.18 | 70.50 | 65.44 | 67.13 | 65.44 | 66.01 | 62.58 | 30.04 |
| 4000 | 70/88 | 49.92 | 49.64 | 58.34 | 38.84 | 48.80 | 57.50 | 44.37 |
| 6000 | 55.97 | 63.12 | 67.50 | 62.92 | 54.02 | 56.46 | 52.52 | 45.41 |
| 8000 | 58.79 | 63.52 | 58.90 | 64.79 | 57.33 | 60.0 | 43.75 | 45.84 |
| 11000 | 56.2 | 63.7 | 56.8 | 68.8 | 61.38 | 81.3 | 66.46 | 55.0 |

**Table 3.2 Accuracy comparison for crop prediction**

**Figure 3.11 Graphical representation of accuracy among all classifier's comparison for cropprediction.**




| Data | GNB | SVM | Linear SVM | RBF SVM | KNN (k=1) | KNN (k=optimal) | LR | RF | ANN |
|---|---|---|---|---|---|---|---|---|---|
| 2000 | 80.61 | 95.07 | 92.97 | 94.38 | 90.44 | 95.50 | 95.22 | 95.07 | |
| 4000 | 84.26 | 94.95 | 96.35 | 94.53 | 93.26 | 95.65 | 96.07 | 95.16 | |
| 6000 | 89.04 | 97.28 | 97.09 | 97.84 | 96.16 | 96.44 | 95.78 | 95.78 | |
| 8000 | 85.36 | 95.27 | 95.59 | 95.74 | 95.59 | 96.81 | 95.31 | 93.71 | |
| 11000 | 95.5 | 97.0 | 97.2 | 96.23 | 96.38 | 97.70 | 97.38 | 92.30 | 96.95 |

In Figure 3.11 we have shown the graphical representation of accuracy comparisonamong all classifiers for crop prediction.

## FUTURE TRENDS

In the future, we hope that this model would be implemented with much more efficient dataset for a specific piece of land containing information such as different soil property, soil pH, different

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

mineral percentage etc. So that no agricultural plot be wasted by harvesting less efficient crop. We want this model to be used worldwide for the furtherdevelopment of the agricultural sector. This can be a field of interest for both researchers and entrepreneurs. In this research, we have developed a model that will predict suitable crop or predict the viability of a specific crop for a particular plot of land. Our plan is to build on this research and improve this model; mostly by adding further algorithms based on other aspects of machine learning. For future work, we also want to build a platform forall the farmers who will be using this model, to share the predictions on their land with other farmers all over the land. This will let a farmer in one country to know the prospect of farming in another part of the world; down to a specific geographical unit. The accuracy values we obtained in the specific crop prediction part of our research was very poor to ourtarget standards.

In the future, we hope to implement ANN for crop prediction and check its viabilityon this. In addition to that, we desire to take this model into the mobile phone platform. Android and iOS applications will be a part of it. Nowadays, even some the poorest farmers are seen to be using such devices. A mobile application can be of help to them as well. One of our most ambitious goals is to improve our model so that it can give a string of crop sequencing predictions. This will enable us to derive predictions for not just the next year, but of multiple years ahead. This time we have only implemented supervised learning. In the future, we plan to use unsupervised and reinforced learning as well. That will give a new dynamic to our research. We are looking for a better world in a sense that we hope our model to be able to successfully raise the standard of farming and agriculture. We hope, whatever shapes this model take, it will stay user friendly and properly welcomed by the potential users.

## CONCLUSION

We believe this model can play a very essential part in today's world. Agriculture is a fundamental aspect of modern civilization. With increasing world hunger and economybreakdown, the proper selection of crop emerges as a massive factor in this. Our proposedmodel can predict the proper crop for a particular piece of land in a way that is very efficient. We have implemented 6 different types of machine learning and deep learning algorithms in this research. All the accuracy of the models was carefully obtained throughvarious different methods, and compared with each other. Using multiple algorithms helpedto understand which algorithm is more suitable for this system. The crops can be predicted based on a very suitable set of features included in the dataset used. From this research work, we found that, the cleaner the data, the better the accuracy of the result. The entire length of this research was very enjoyable, as we were able to work in the field of machinelearning and deep learning. Some of the python library usage, algorithm fitting, and accuracy checking methods were very interesting in practicality. We trust all our algorithmsand research work to efficiently work on any platform and any new type of data. The predictions made were solid and robust. Such strength in the model delights us, and we hope this keeps working over the years without issues.

## REFERENCES

[1]. Aggarwal Sachin (2001). Application of Neural Network to Forecast Air Quality Index. Thesis submitted in partial fulfillment of requirements for a degree in Bachelor of Technology, April 2001.

[2]. B. J I ET AL Artificial neural networks for rice yield prediction in mountainous regions. Journal of Agricultural Science (2007), 145, 249–261.

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395
Volume 9, Issue 11 - November 2021- Pages 1-14

[3]. B.A. Smith et al Artificial Neural Networks for Automated Year-round Temperature Prediction. Computers and Electronics in Agriculture 68 (2009) 52–6.

[4]. Cheng, B. and Titterington. D. M. (1994). Neural networks: A review from a statistical perspective Statistical Science, 9: 2-54.

[5]. D.L. Ehret et al, Neural network modeling of greenhouse tomato yield, growth and water use from automated crop monitoring data. Computers and Electronics in Agriculture 79 (2011) 82–89.

[6]. Enfield, D. B., 1996. Relationships of inter-American rainfall to tropical Atlantic and Pacific SST variability. Geophysical Research Letters 23(23): 3305-3308.

[7]. Everingham, Y. L., R. C. Muchow, R. C. Stone, and D. H. Coomans, 2003. Using southern oscillation index phases to forecast sugarcane yields: a case study for Northeastern Australia. International Journal of Climatology 23(10): 1211-1218.

[8]. Handler, P, 1990. USA corn yields, the El Niño and agricultural drought: 1867-1988.

International Journal of Climatology 10(8): 819-828.

[9]. Hansen, J. W., J. W. Jones, C. F. Kiker, A. W. Hodges, 1999. El Niño-Southern Oscillation impacts on winter vegetable production in Florida. Journal of Climate 92-102.

[10]. Hansen, J. W., A. W. Hodges, and J. W. Jones, 1998. ENSO Influences on agriculture in the southeastern United States. Journal of Climate 11(3): 404-411.

[11]. Haykin. S, 1999. Neural Networks: A Comprehensive Foundation (Second Edition).

Upper Saddle River, NJ: Prentice Hall.