ANALYTICS BASED ON ARTIFICIAL NEURAL NETWORK: A CASE STUDY BASED ON IOWA CORN YIELD FORECASTING

By

NAVEEN SINGIREDDY

Bachelor of Technology in Computer Science & Engineering

Jawaharlal Nehru Technological University

Hyderabad, Telangana, India

2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July, 2015

ANALYTICS BASED ON ARTIFICIAL NEURAL NETWORK: A CASE STUDY BASED ON IOWA CORN YIELD FORECASTING

Thesis Approved:

Dr. K. M. George

Thesis Adviser

Dr. Blayne Mayfield

Dr. Douglas Heisterkamp

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. K.M. George for his continuous support of my research study, for his patience and motivation. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank my thesis committee Dr. Blayne Mayfield and Dr. Douglas Heisterkamp for their insightful comments and encouragement.

I would like to thank SST Software, Stillwater, Oklahoma who partially supported this research work.

My sincere thanks also goes to Mr. Todd Pugh from SST Software, who provided me an opportunity to join their team as research assistant and for his support and help in providing data and to sort the variables.

I would like to thank my classmate and co-research assistant Pawan for his support.

Last but not the least, I would like to thank my parents, sisters, brother-in-laws and friends for supporting me spiritually throughout writing this thesis and my life in general.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: NAVEEN SINGIREDDY

Date of Degree: JULY, 2015

Title of Study: ANALYTICS BASED ON ARTIFICIAL NEURAL NETWORK: A CASE STUDY BASED ON IOWA CORN YIELD FORECASTING

Major Field: COMPUTER SCIENCE

Abstract: This work is an application of computing techniques to solve real world problems. Agriculture has been playing a crucial role in the growth of world economy and has been associated with production of basic food demands. The World's food demand is increasing day by day with the population but the cultivable land which is a limited resource is not increasing in proportion with the food demand. In order to maximize the production and improvise the quality of crop, various firms have come up with variety of fertilizes, genetic seeds, modern equipment, etc. While the technological advancement is still going on, there is another area of interest where most of the technological firms are now focusing, which is predicting the approximate crop yield in the different geographical locations.

In order to address the above problem, we proposed a model to understand the effects of various crop yield influencing parameters such as field location, soil properties, availability water in the ground, ground slope and climate conditions such as rainfall and temperature. We proposed this ANN model to predict corn yield of IOWA region with these influencing parameters which will help the various agricultural firms to give the recommendations to their growers to maximize the crop yield. According to our knowledge, this is the first implementation of ANN model to predict corn yield of IOWA region with these set of data features. In this work, the ANN model produced more consistent yield prediction and it resulted in r^2 of 0.70 and RMSE of around 1750.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Artificial Neural networks (ANN) is a black box model owing to the complex mathematical calculations involved in its algorithm. It is primarily used only for prediction purposes and not for explanatory purposes. But using the Decision tree with neural networks i.e. by modeling the predicted value of the neural networks with the independent variables using decision trees, 95% of the neural network model can be explained. Artificial neural networks have been used to solve practical problems in several areas such as robotics, medical sciences to diagnose several cancers problems, game playing and decision making, many financial applications [9]. It is also used in agricultural data analysis and modeling which is the topic of our research.

In this study, we are concerned with the dependency of input parameters on forecast results in ANN models. Our objective is to identify the significant variables that affect the outcome for tuning forecast performance. We use data on crop yield of Iowa for the analysis. We use Iowa as the test case region as it is one of the major states of agricultural production in United States. Between the Mississippi River on the east and the Missouri River on the west, Iowa is home to some of the most fertile top soil in the world. Its land areas can be divided into three main regions; the Young Drift Plains which cover most of the northern and central parts of Iowa, the Driftless Area parallel to the Mississippi River in the northeast, and the Dissected Till Plains in the southern area of the state. The fertile lands of Iowa makes the state, the number one corn producer in the United States [14]. The primary analysis of this study is conducted by developing an ANN model to predict the crop yield of Iowa region using an Artificial Neural Network (ANN) and understanding the importance of various data features involved in this ANN model. The results are compared against previous studies.

This area of application is chosen for our study because of its practical implications and need. Agriculture, for decades, had been associated with the production of basic food crops. Agriculture also plays a crucial role in the growth of an economy as it is contributing around 6% of overall world GDP [26]. The World's food demand is increasing day by day with the population but the cultivable land which is a limited resource is not increasing in proportion with the food demand. Under this circumstances improving the production efficiency or crop yield in agricultural field turned out to be very important in meeting the food demand. Agricultural system is very complex since it deals with large data situation which comes from a number of factors. Various firms have come up with variety of fertilizes, genetic seeds, modern equipment, etc. to maximize the production and improvise the quality of crop. While the technological advancement is still going on, there is another area of interest where most of the technological firms are now focusing, which is predicting the approximate crop yield in the different geographical locations. There have been considerable advances in agricultural production and crop models are increasingly used in agricultural field to improve production efficiency [4].

SST (Site-Specific Technologies) is a privately owned company based out of Stillwater, Oklahoma with branch offices in Oklahoma City, Tulsa, Illinois, and Iowa, as well as Brazil and Australia. Since its inception in 1994, SST has been at the forefront of precision agriculture as one of the pioneers of the industry. SST has decision support tools that help farmers efficiently and profitably run their businesses and it provides a site-specific technology to improve the crop yield. SST has huge amount of historical yield data of their clients of various locations. In this work, models will be developed using SST historical yield data at multiple locations soil type data, climate data which are field specific rainfall and temperature data, availability water in the ground and ground slope data to predict the crop yield.

Crop yield is influenced by many factors such as field location, soil type, water availability in the ground, rainfall and temperature. Understanding the effects of these factors on the crop yield can be very helpful in improving the crop yield. Predictive modeling provides an efficient methodology to understand the correlation between the dependent variable (crop yield) with several other independent variables and also to predict the crop yield under different conditions. Until today, some of the crop models constructed for this purpose are based on linear methods such as multiple linear regression. But these procedures were not sufficient owing to the complex relationships between the independent variables and the crop yield. To overcome the draw backs of multiple linear regression and to establish complex relation among the variables, many of crop models used non-linear approaches such as artificial neural networks.

We have large data set with attributes such as location (Country, State, County and Field), rainfall data, temperature data, water availability in the ground, ground slope, soil type properties (sand, silt, clay, organic matter and cation exchange capacity), field centroid latitude and longitude, crop season, crop type, crop acres (hectares), crop hybrid, soil types description with sub categories, relational maturity and crop yield per hectares (min, max and avg) and per acres((min, max and avg). However we have used some data preparation techniques to reduce the dimensionality, avoid the missing values and to normalize the numeric data into uniform range by ignoring some outliers [16].

CHAPTER II

REVIEW OF LITERATURE

Artificial Neural Networks:

The Artificial Neural Network structure is based on the human brain's biological neural processes used to solve complex problems where it tries to imitate into mathematical models. Interrelationships of correlated variables that symbolically represent the interconnected processing neurons or nodes of the human brain are used to develop models. ANN models finds relationships by observing a large number of input and output examples to develop a formula that can be used for predictions [10]. The applications of Artificial Neural networks have been increasing in business. More and more development tools have emerged on the market. Many neural-net systems have been shown to work well in identifying patterns, learning from experience, reaching some conclusions, and making predictions [18]. There are six main characteristics of ANN technology: network structures, parallel processing ability, distributed memory, fault tolerance ability, collective solution, and learning ability. ANNs can be applied to many problems that are solved conventionally by statistical and management science techniques. In fact, the common characteristics enable ANNs to solve these problems better and faster than conventional techniques, even without human intervention. ANN Tasks can be classified as approximation, optimization, classification, prediction and generalization [18]. Generally, ANN is considered as black box that capable of producing accurate output data because it learns from the previous input data and try to match the actual output data [21].

One of powerful methodology in solving non-linear problems is Artificial Neural Network. Artificial Neural networks have been known as having excellent potential for improving prediction accuracy in geophysical data. However, there have been few studies which have demonstrated this potential using real data sets. The application of neural networks as fitting model is based upon several considerations. First, neural networks appear to perform better than other techniques and require no assumptions about the explicit parametric nature of distributions of the data. However, neural networks, once trained, are computationally more efficient [21].

Artificial Neural network models have been used for the purpose of prediction in various data analytics applications. ANN models have been developed to solve various practical problems in several areas such as for prediction of bankruptcy, Stock market prediction, robotics, medical sciences to diagnose several cancers problems, game playing and decision making, agricultural forecasting and weather forecasting applications.

Related Work and Input Selection:

The combination of advancement in technology and motivation to improve the production of crop yield in agriculture are becoming very challenging and interesting. It has been shown that several factors affected the crop performance either directly or indirectly. Most of the factors highlighted normally in researches are soil factors [1,2], weather and climatic conditions including temperature and rainfall, crop and fertilizer applications [2], availability water in the ground [17]. All these factors together formed a complex system for agriculture since it deals with a large set of data.

Crop modelling was initially viewed as a tool to understand the impact of influencing factors such as soil type, fertilizer, weather and climatic conditions on crop yield [4]. For a number of years, most crop models were based on linear methods which were constructed of linear or multiple linear regression, simulation, expert systems and correlation analysis. These methods assume a linear relationship between input parameters and crop yield. However, these techniques were neither sufficient nor comprehensive enough to show the relationships or interactions of parameters and crop yield [4].

Regression is widely accepted approach but predictive accuracy depends on the specific conditions of the input data. Simulation approaches requires physical relationship among the input parameters which are affecting the crop yields and also requires many biophysical inputs that often must be estimated rather than measured. Expert systems rely on human expertise and characterize yield by sets of logical rules, but the initial formation of rules requires a lot of real expert's knowledge not a group of people with different experiences and relies on the limits of the input data [8]. Wieland R et al. [7] compared the performance of the ANN and a similar fuzzy expert's system resulted in expert's systems cannot be better than ANN, because it cannot be trained like a neural network with random weights by back propagation if there is an error. So these complex situations need non-linear approaches such as neural network. Many of the researches had listed two approaches to predict crop yield. First using the normal traditional mathematical approaches and second using the applications of artificial intelligence.

Now a days, ANN has become a well-liked method to most researchers because of its ability of prediction, forecasting and classification in geology fields. Although ANN model consumes more time to be developed, it can produce more consistent yet accurate crop yield prediction rather than regression models [5].

From our above literature review, we have come to know that ANN models are more effective and capable models when it compares with other statistical or linear regression models. Now we review some work on artificial neural network which is our topic of interest in this thesis study. Kaul et al. [5] developed feed forward back-propagation ANN model to predict the corn and soybean yield prediction for Maryland using the rainfall and soil properties. ANN produced a better improved results than traditional statistical regression methods. The results produced in Maryland, ANN model for corn resulted in coefficient of determination (r^2) and Root Mean Squared Error (RMSE) of 0.77 and 1036 as compared to 0.42 and 1356 for regression, respectively.

Similar research performed by O'Neal et al. [8] to predict maize yield with local crop-stage weather data and yield data from 1901 to 1996 using a fully connected back-propagation ANN and regression models resulted in RMSE of 10.5% and 12.2%, respectively [12]. The other similar work by B. J I et al. [11], developed an ANN model to predict rice grain yield using historical yield data at multiple locations throughout Fujian, rainfall data and the weather variables were used for each location. Similar trend shows, ANN models consistently produced more accurate yield predictions than regression models. ANN rice grain yield models for Fujian resulted in r^2 and RMSE of 0.67 and 891 vs 0.52 and 1977 for linear regression, respectively [11]. So in our work we have considered rainfall and temperatures to predict the corn yield of Iowa region. In our work we have got daily rainfall and temperature of entire year so we calculated yearly (Jan- Sept) and seasonal (Mar – Sept) rainfall and temperature to see the best fit for the model.

In other research, Drummond et al. [9] applied a feed-forward neural network to estimate nonlinear relationship between soil parameters and crop yield. The estimated yield generated by ANN model tended to be very similar to the actual crop yield. From the above research we have come to know that soil properties play major role in predicting the crop yield.

The authors William F. Schillinger et al. [17] conducted a statistical predictive analysis on wheat yield data of Inland PNW with availability water and rainfall, and their statistical model resulted in r^2 of 0.73. In this study authors have shown us that availability water in the ground is also one of the parameter in predicting the crop yield.

In some other research Shearer S. A., et al used Soil Landscape Features and Soil Fertility Data to predict the crop yield [1]. In their work Slope of the land is also included in Soil Landscape Features data. So in our work we have considered slope in predicting the crop yield.

Artificial Neural Networks, from above considerable study has been proven to be more capable and effective model to predict the crop yield as it deals with a large set of data. It produces more accurate and better results when it compares with other statistical methods. Artificial intelligent system has brought artificial neural network (ANN) to become a new technology which provides assorted solutions for the complex problems in agriculture researches. It has been shown that, ANN can solve and develop improved models to predict the crop yield. From all the above study we have seen that many researchers considered soil properties, rainfall, temperature, land slope and availability water in the ground for predicting the crop yield using ANN models in different researches at different geographical regions. According to our knowledge, this is the first implementation of ANN model to predict corn yield of IOWA region with these set of data features.

Data Analytics and Dimensionality Reduction:

Data analytics is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and to verify or improve the existing models. There have been many machine learning algorithms and each of which has its own requirement in modeling. Artificial neural network is one such model and it requires all the input and target data type as numeric. In machine learning, dimension reduction is the process of reducing the number of input variables to model. There are different approaches in reducing the dimensionality such as stepwise iteration procedure, principle component analysis and generalized regression models.

Stepwise iteration procedure: It is a process of building a model by successively adding or removing variables. It is also known as trial and error method. There are three approaches in in this stepwise method and they are forward selection, backward elimination and bidirectional elimination. Forward selection starts with no variables in the model, test the model with variable addition that improves the model the most, and repeating this process until none improves the model. Backward elimination starts with all variables in the model test the model with variable elimination that improves the model the most by being eliminated, and repeating the process until no further improvement is possible. Bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded.

Principle component analysis (PCA): It has been used in data reduction and interpretation applications. PCA is a statistical procedure which converts a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principle components. The number of principle components is less than or equal to the number of original variables. Thus, it reduces the dimensionality of a data set involving a large number of variables. The principal components are ordered so that the first pc represent most of the variations in the original variables [25].

ANN Model Selection:

There are many types of ANN, but looking at the all the above literature work, in many of the researches, authors developed a feedforward back propagation (FFBP) network by adjusting ANN parameters learning rate and number of hidden nodes. The model then can be "learnt" by training [3]. According to Alvarez [6], the learning process can be defined as "a process which consists in adjusting the weights associated to the transfer functions between neurons comparing ANN output with observed data". The most common training method is back-propagation. The back-propagation method is used to train the feed-forward neural network to minimize the error [1] where in this training the error represents the difference between calculated output and the target value [7]. The process is repeated until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal.

Artificial Neural Networks (ANN) are trained to solve specific problems. There are several steps involved in the ANN learning process. In the first step we have to determine the type of training data using various classification and clustering approaches. In the second step we need to gather a training data set that satisfactorily describe a given problem by using the normalization techniques if necessary. In the third step we need to describe training data set in a form understandable to a chosen artificial neural network. In the fourth step we do the learning and after the learning we can test the performance of learned artificial neural network with the test data set. Test data set consist of data that has not been introduced to artificial neural network while learning [29]. In this work, we propose to apply this methodology to corn yield prediction.

ANN Model Description:

Artificial neural networks can be used to develop empirically based agronomic models. The ANN structure is based on the human brain's biological neural processes used to solve complex problems where it tries to imitate into mathematical models. Interrelationships of correlated variables that symbolically represent the interconnected processing neurons or nodes of the human brain are used to develop models. ANN models finds relationships by observing a large number of input and output examples to develop a formula that can be used for predictions [10].

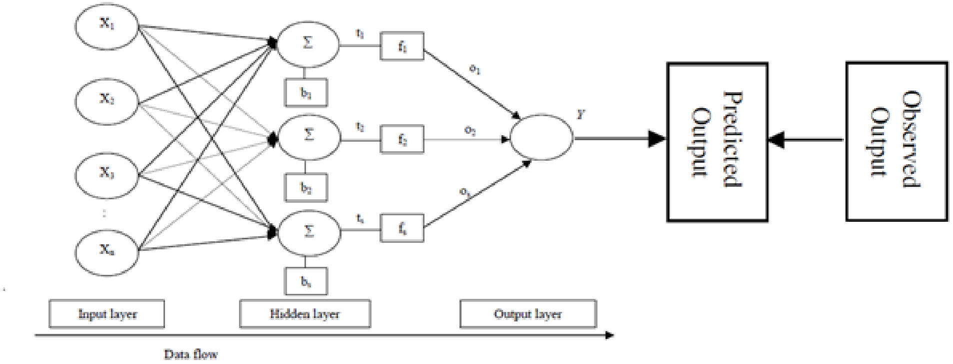

Figure 1. Layers and connections of a feed-forward back-propagating ANN

A minimum of three layers is required in a non-linear ANN model as shown in Figure 1: the input, hidden and output layers. The number of hidden nodes depends on specific problem of the study which can easily extend to more hidden layers [15].

The backpropagation algorithm is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals "forward", and then the errors are propagated backwards [15].

The input and output layers contain nodes that correspond to input and output variables, respectively [5].

w is connected to an output layer, where the outputs of units are connected to the inputs of the next via connection weight [3].

Data move between layers across weighted connections. A node accepts data from the previous layer and calculates a weighted sum of all its inputs, t:

$$t_i = \sum_{j=1}^{n} (w_{ij} x_j + b_i)$$

Where, n is the number of inputs, w is the weight of connection between node i and j, x is the input from node j, and b_i is a bias. In order to calculate the node output o_i, a transfer function f_i is then applied to the weighted value:

$$o_i = f_i(t_i)$$

The most popular transfer or activation function is sigmoidal function [5] for the hidden and output layers. Input layer commonly uses a linear transfer function to pass the information to hidden layers [5].

CHAPTER III

METHODOLOGY

Data Procuration and Preparation:

Data sources: The data for this thesis were obtained from the SST Software, which is a privately owned company based out of Stillwater, Oklahoma with branch offices in Oklahoma City, Tulsa, Illinois, and Iowa, as well as Brazil and Australia. Since its inception in 1994, SST has been at the forefront of precision agriculture as one of the pioneers of the industry. As we have seen data features in the previous section, we have features such as location (Country, State, County and Field), water availability in the ground, ground slope, soil type properties (sand, silt, clay, organic matter and cation exchange capacity), field centroid latitude and longitude, crop season, crop type, crop acres (hectares), crop hybrid, soil types description with sub categories, relational maturity and crop yield per hectares (min, max and avg) and per acres((min, max and avg) in the provided dataset for the years 2005 to 2014. However, as we have rainfall and temperature data only for 2011 to 2014. So we have extracted daily rainfall and temperature data using field centroid latitude and longitude from the same source for the years 2011 to 2014. So in this thesis work, we have considered data only for these four years.

Data Preprocessing and Dimensionality Reduction:

Each machine learning model will have its unique requirement in its input and output data type. In the same manner, an ANN model in our work requires all the input and target data type as numeric.

In our dataset we have soil description types as well as different numeric soil properties such as component, sand, silt, clay percentages, organic matter and cation exchange capacity based on their soil category types. High dimensionality is a problem to machine learning task [16] so we have ignored all the redundant soil description features in the dataset as we already have them in as numeric features. We have relation maturity (crop maturity duration) and hybrid type features in the dataset but most of the values are missing so we have ignored these features in our work. We also have available water in the ground 0 to 25CM and 0 to 100CM but both features have same correlation with the target variable and they both are highly correlated each other. So, in our work we are considering one of them.

We have also used excel linear correlation function and MATLAB Plot regression to see the linear correlation between target vs each input variable to understand the importance of the each input variable.

After all these data preprocessing, data features remained for our research are soil properties such as SAND, SILT, CLAY percentages, OM (Organic matter), CEC (Cation Exchange capacity), COMPONENT and daily rainfall and temperature, water availability in the ground.

We have daily rainfall and temperature data for each field centroid latitude, longitude and we have observations for subfield level (i.e. for each field centroid latitude, longitude we may have more than one sub field).

We have calculated yearly (Jan to Sept) average and seasonal (Mar to Sept) average rainfall and temperature values for each field based on field centroid latitude, longitude and we have joined temperature and rainfall data with other provided data features based on field centroid latitude, longitude. We named this data as “**dataset 1**”.

We have also calculated average values for all the data features by field level as we have field level rainfall and temperature data. We named this data as "**dataset 2**".

Crop year feature helps in separating the data for training, validation and testing by grouping the data for each year to better fit the model.

ANN Model Implementation:

ANN models find relationships by observing a large number of input and output examples to develop a formula that can be used for predictions [10]. A minimum of three layers is required in a non-linear ANN model as shown in Figure 1: the input, hidden and output layers. The number of hidden nodes which are depending on specific problem of the study which can easily extend to more hidden layers [15]. The backpropagation algorithm is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals "forward", and then the errors are propagated backwards [15]. The input and output layers contain nodes that correspond to input and output variables, respectively [5]. The input layer is used to distribute the inputs to a number of hidden layers, and the output of which is connected to an output layer, where the outputs of units are connected to the inputs of the next via connection weight [3]. Data move between layers across weighted connections. A node accepts data from the previous layer and calculates a weighted sum of all its inputs

We have used MATLAB neural network toolbox to build this ANN feed forward back propagation model. The three transfer functions of the neural network toolbox have been used in developing neural net model are TANSIG (tangent sigmoid transfer function), PURELIN (linear transfer function) and LOGSIG (log-sigmoid transfer function). The most popular transfer or activation function is sigmoidal function [5] for the hidden and output layers. Input layer commonly uses a linear transfer function to pass the information to hidden layers [5]. There are 10 inputs to the model and the target is crop yield.

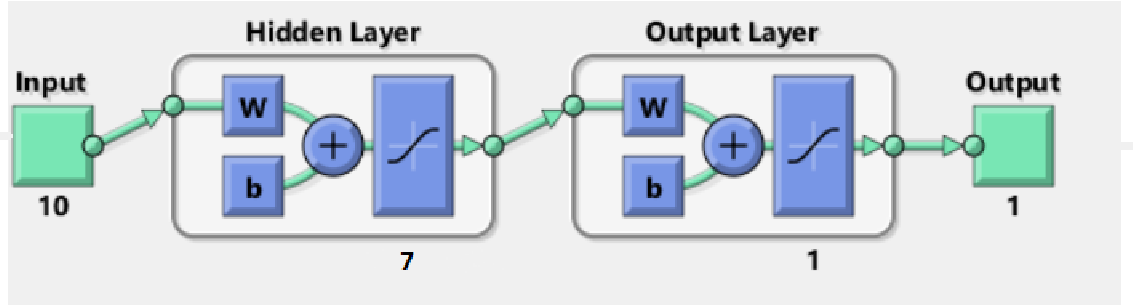

Figure 2. Implementation of FFBP ANN Model using mat lab nntool

Only one hidden level is used in this thesis work for ANN modeling since the performance of the more accurate model that relies on more than one hidden level does not guarantee good results [23]. On the contrary, the model with only one hidden level normally gives better results. As shown in the Figure 2 initially we have 10 inputs to the model and one target output, we tried with the different combinations of hidden neurons to train the model. The very less number of neurons in the hidden layer resulted in slow fitting and the more number of neurons resulted in over fitting the model. As elaborated by Al-Shalabi et al. [24] the most appropriate method to determine optimal design is still through trial and error method. For the choice of the number of neurons in the hidden node is recommended not to be greater than 75% of the number of the input nodes [22]. The training function that we have implemented in our model is "trainlm" (Levenberg-Marquardt) which is usually faster and more consistent than Bayesian Regularization (trainbr) and Scaled Conjugate Gradient (trainscg).

Sample code of the model:

```
x = input';
t = target';

% Choose a Training Function
% 'trainlm' is usually fastest.
trainFcn = 'trainlm';  % Levenberg-Marquardt

% Create a Fitting Network
hiddenLayerSize = 7;
net = fitnet(hiddenLayerSize,trainFcn);

% Setup Random Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 60/100;
net.divideParam.valRatio = 20/100;
net.divideParam.testRatio = 20/100;

% Setup Sequential Division of Data for Training, Validation, Testing
net.divideParam.trainInd = TrainIndices;
net.divideParam.valInd = ValidIndices;
net.divideParam.testInd = TestIndices;

% Train the Network
[net,tr] = train(net,x,t);

% Test the Network
```

```
y = net(x);
e = gsubtract(t,y);
performance = perform(net,t,y)

% View the Network
view(net)
```

Neural Network Stepwise Iteration Analysis:

Stepwise Iteration is a process of building a model by successively adding or removing variables. It is also known as trial and error method. There are three approaches in in this stepwise method and they are forward selection, backward elimination and bidirectional elimination. We tried with different combination of inputs to the model with stepwise forward selection and backward elimination of variables. We have used the same model outlined in the methodology section. All the inputs to the ANN model are normalized to the interval [0, 1] and output is always renormalized by the mat lab neural network tool box itself**.** For every analysis, the model is trained 3 times with same data and average result obtained from those 3 times are considered as final result.

Principle Component Analysis (PCA):

PCA has been used in data reduction and interpretation applications. PCA is a statistical procedure which converts a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principle components. The number of principle components is less than or equal to the number of original variables [25]. Thus, it reduces the dimensionality of a data set involving a large number of variables. The principal components are ordered so that the first variables represent most of the variations in the original variables [25]. The below are the steps to calculate principle components.

i. Let X denotes a set of raw scores of size N. Then the average of X is defined as avg(X)= (∑X)/N

ii. The deviation from mean is defined as $X_d = x – avg(X)$ where $x \epsilon X$.

iii. The correlation between two variables X and Y is given by

$$r_{XY} = \frac{\sum X_d Y_d}{\sqrt{\sum X_d{}^2}\sqrt{\sum \; Y_d{}^2}}$$

iv. The correlation matrix of n variables is defined as an n x n symmetric matrix R whose diagonal elements are 1 and the (i, j) th entry is the correlation between ith and jth variable. This Correlation matrix is of the form $r_{xy.}$

v. If A is a n x n symmetric matrix, then the real number λ is called eigenvalue of A if and only if there is a non-zero vector V in R^n for which AV = λV. Any such vector is called eigenvector associated with the eigenvalue λ.

vi. The first PC is the eigenvector associated with the largest eigenvalue.

To calculate the eigenvalue and eigenvector we use matlab function princomp(A) which gives us eigenvectors for each eigenvalue. The first PC gives the order of the most significant variables based on vector values. Let n be the total number of initial variables, "w_{1i}" be the weightage of the ith variable in 1st PC, "w_{2i}" be the weightage of the ith variable in 2nd PC and so on and "i_n" be the ith variable. Then the first dimension can then be calculated as

$$V_1 = w_{11}*i_1 + w_{12}*i_2 + w_{13}*i_3 + \ldots\ldots\ldots\ldots + w_{1n}*i_n$$

Similarly second dimension can be calculate as

$$V_2 = w_{21}*i_1 + w_{22}*i_2 + w_{23}*i_3 + \ldots\ldots\ldots\ldots + w_{2n}*i_n$$

So on you have to compute the values for all the principle components.

The reduced dimensions are then given as an input to Neural Network to verify its accuracy and to compare it with the neural network Stepwise Iteration method.

Same as ANN, all inputs to PCA analysis are normalized to the interval [0, 1] and output is always renormalized by the mat lab neural network tool box itself**.** We have used mat lab mapminmax function to normalize the inputs. The formula implemented by the function is: y = (ymax-ymin)*(x-xmin)/(xmax-xmin) + ymin; where y represents the normalized value, x the input value. The max and min represent the highest and lowest values among x and y. The mapminmax function allows us to specify the range of y as an argument to it. Furthermore, no assumptions on the distribution of data is necessary and examination of our dataset did not produce any outliers.

CHAPTER IV

FINDINGS

In order to analyze the performance of the proposed model in this thesis work, we have done some experiments with the datasets using feed-forward back-propagating ANN.

The authors B. J I et al [11] used climate conditions to predict the grain yield models for Fujian. So in our work we have considered rainfall and temperatures to predict the corn yield of Iowa region. In other research, Drummond et al. [9] applied a feed-forward neural network to estimate nonlinear relationship between soil parameters and crop yield. The estimated yield generated by ANN model tended to be very similar to the actual crop yield. From the above research we have come to know that soil properties play major role in predicting the crop yield. In some other work, the authors William F. Schillinger et al. [17] conducted a statistical predictive analysis on wheat yield data of Inland PNW with availability water and rainfall. In their study authors shown us that availability of water in the ground is also one of the parameter in predicting the crop yield. In some other research Shearer S. A. et al used Soil Landscape Features and Soil Fertility Data to predict the crop yield [1]. In their work Slope of the land is also included in Soil Landscape Features data. So in our work we have considered slope in predicting the crop yield.

After all the data processing, data analysis and from related work, the data features remained for our research are soil properties such as SAND, SILT, CLAY percentages, OM (Organic matter), CEC (Cation Exchange capacity), COMPONENT and daily rainfall and temperature, Slope of the ground, Availability water in the ground.

We have made two datasets out of given data. First dataset has the subfield level observations and the second has the field level observations however the temperature and rainfall data that we have is field level in both the datasets.

We tried to find the correlation between each input data features towards crop yield however we could not see any good correlation in the sub filed level. For that reason we have made the second dataset to find the field level observations from the sub filed level dataset just by averaging the each sub field observations of the fields.

The total observations that we have in the dataset 1 and dataset 2 are

Crop Season Year	2011	2012	2013	2014
Sub Field level total observations (dataset 1)	35000	55000	57000	30000
Field level total observations (dataset 2)	4382	5917	6510	3142

Table 1. Iowa corn yield number of observations by year

As shown in the model section under methodology, we have used mat lab neural network tool to develop the ANN model. As there have been three transfer functions built in the neural network toolbox of mat lab, we tried all three combinations of transfer functions in the hidden and output layer. However, we could see the same pattern as per the authors Kaul M et al. [5] the most popular transfer or activation function is sigmoidal function for the hidden and output layers i.e. the network with sigmoid transfer function has been a perfect fit of the model. As we have 10 inputs to the model and one target output, we tried with the different combinations of hidden neurons to train the model. The very less number of neurons in the hidden layer resulted in slow fitting and more number of neurons resulted in over fitting the model. As per the authors Bailey et al. [22], for the choice of the number of neurons in the hidden node is recommended not to be greater than 75% of the number of the input nodes. We have tried running the model with yearly, seasonal rainfall and temperature combination, however, we could see the better performance of the model with yearly rainfall and temperature data.

As we know that coefficient of determination is a number that indicates how well data fit a model or how well observed outcomes are replicated by the model or it is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable [28]. One use of the coefficient of determination is to test the goodness of fit of the model. It is expressed as a value between zero and one. A value of one indicates a perfect fit, and therefore, a very reliable model for future forecasts. A value of zero, on the other hand, would indicate that the model fails to accurately model the dataset. Root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model and the values actually observed.

A data set has *n* observed values marked $y_1 ... y_n$ and each associated with a predicted values $f_1 ... f_n$. If $\bar{y}$ is the mean of the observed data:

$$R^2 \equiv 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (f_i - y_i)^2$$

RMSE is just a square root of MSE

In order to test the model performance across all the years, we have considered two years data for training, one year data for validation and remaining one year data for testing. We have repeated this experiment with all the four years combinations with train, validate and test.

SAND, SILT, CLAY percentages, OM (Organic matter), CEC (Cation Exchange capacity), COMPONENT and Yearly rainfall and temperature, ground slope, Availability water in the ground are the model inputs in this experiment. We tried with different combinations of inputs to the model. We have used the same model which is outlined in the model section in the methodology and maintained the number of neurons in the hidden layers as not more than 75% of the number of the input nodes as per the authors Bailey et al. [22].

Neural Network Stepwise iteration analysis results:

We tried with different combinations of inputs to the model and below are the findings of coefficient of determination (r^2) and RMSE with different input combinations for the 2012 data

Input Variables	r^2	**RMSE**
Rainfall(Y)	0.5121	1853
Temperature(Y)	0.4732	1897
Rainfall(S)	0.4748	1901
Temperature(S)	0.4516	1942
Sand	0.4157	1921
Silt	0.3896	1948
Clay	0.3217	2058
OM	0.3692	2037
CEC7	0.2146	2163
Ground water100	0.3974	2032

Slope	0.3068	2097
Component	0.2763	2126
Rainfall, Temperature(Y)	0.5934	1812
Rainfall, Temperature(S)	0.5681	1836
Sand, Silt	0.4867	1907
Clay, Sand	0.4369	1923
Clay, Silt	0.4091	1937
Sand, Silt, Clay	0.4883	1911
Sand, OM	0.4576	1937
Clay, OM	0.4385	1928
Silt, OM	0.3981	1974
Clay, OM, Silt	0.4502	1907
Clay, OM, Sand	0.4832	1898
Sand, Silt, Clay, OM	0.4963	1901
Clay, OM, Silt, CEC7	0.4578	1927
Clay, OM, Silt, Component	0.4982	1883
Sand, Silt, Clay, OM, Component	0.5137	1869
Sand, Clay, OM, Silt, CEC7	0.5081	1878
Sand, Clay, OM, Silt, Component,CEC7	0.5193	1863
Rainfall, Temp, Ground water	0.6168	1817
Rainfall, Temp, Ground water, Slope	0.6293	1811
Rainfall, Temp, Ground water, Slope, Component	0.6281	1819
Rainfall, Temp, Ground water, Slope, Sand, Silt, Clay	0.6749	1793

Rainfall(Y), Temp(Y), Ground water, Slope, Sand, Silt, Clay, OM	0.6981	1759
Rainfall(S), Temp(S), Ground water, Slope, Sand, Silt, Clay, OM	0.6673	1801
Rainfall, Temp, Ground water, Slope, Sand, Silt, Clay, OM, Component	0.6892	1784
Rainfall, Temp, Ground water, Slope, Sand, Silt, Clay, OM, CEC7	0.6927	1773
Rainfall, Temp, Ground water, Slope, Sand, Silt, Clay, OM, Component, CEC7	0.6979	1768

Table 2: Findings for neural network stepwise iteration method

As shown in the Figure 2 and Table 2, we have run the model against each year individually with different combinations of the inputs, we could see the better performance of the model with SAND, SILT, CLAY percentages, OM, Slope, Ground water, Rainfall and Temperature as input to the model. Even if we pass CEC and Component along with other inputs to the model, no difference in the model performance was observed. So we have just ignored these two features i.e. CEC and Component. The below is the ANN model that shows the number of inputs to the model, number of hidden layers, number of hidden neurons and output.

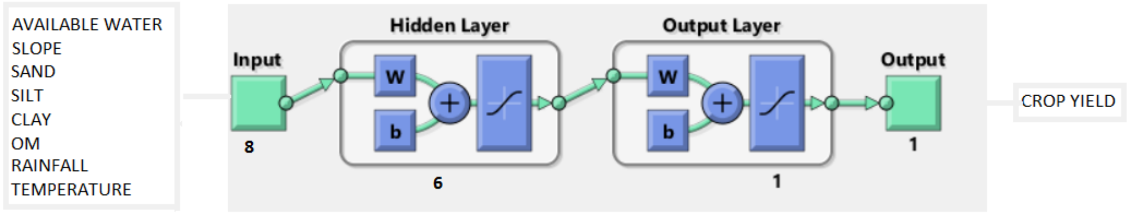

Figure 3. FFBP ANN Model using mat lab nntool with inputs and outputs

As shown in Figure 3, the number of inputs to the model are 8 so as per the authors [22] the number of hidden neurons should be less or equal to 75% of the number of inputs to the model i.e. in this model number of hidden neurons should be less than or equal to 6. The transfer function that we have implemented in both hidden and output layers of this above model is TANSIG (tangent sigmoid transfer function) as per the authors [5].

The performance of the model r^2 - Coefficient of determination and RMSE- Root-mean-square error with above set of 8 inputs to the model for a 2012 data is shown in Figure 4 and Table 3.

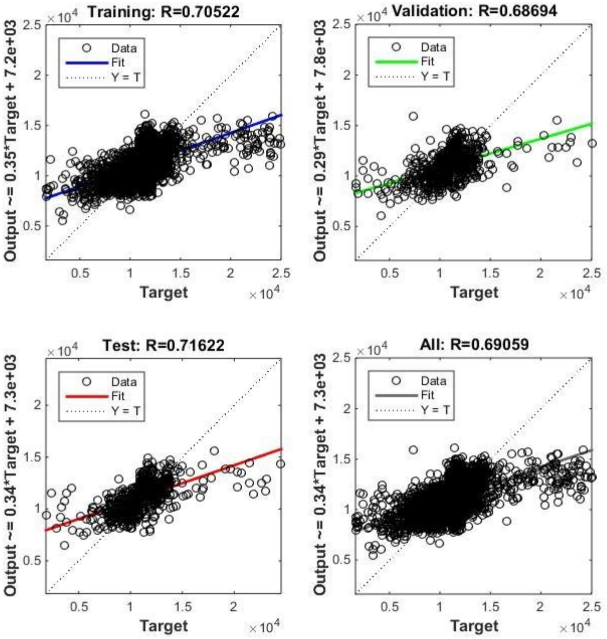

Figure 4. r^2 for the trained model using 2012 data

The table below shows the results of neural network model with above mentioned 8 set of input variables for 2011 to 2014 data

Crop Season Year	Training	Validation	Test	All	RMSE
2011	0.69341	0.70512	0.69677	0.69983	1741
2012	0.70522	0.68694	0.71622	0.69059	1802
2013	0.71834	0.69568	0.70647	0.70158	1757
2014	0.68856	0.69023	0.69495	0.69326	1826

Table 3. Model performance r^2 and RMSE for every year through 2011 to 2014

As shown in the Table 3 when we ran the model with different combination for each year individually, the above mentioned 8 set of parameters input to the model, explained on average of 70% of the data.

As per the first experiment, in order to see the model performance with throughout the years. We have consider two years data for training, one year data for validation and remaining one year data for testing and repeated this experiment with all the four years combinations with train, validate and test.

Crop Season Year	Training	Validation	Test	All	RMSE
Tr -2011 & 2012	0.67458	0.66369	0.66271	0.66783	2291
V - 2013					
T - 2014					
Tr -2011 & 2013	0.6816	0.67594	0.68936	0.68197	2138
V - 2012					

T - 2014					
Tr -2012 & 2013	0.67928	0.67753	0.66935	0.67264	2239
V - 2014					
T - 2011					
Tr -2013 & 2014	0.66502	0.66102	0.65376	0.65986	2316
V - 2011					
T - 2012					
Tr -2011 & 2014	0.65967	0.66724	0.64869	0.65127	2328
V - 2013					
T - 2012					
Tr -2012 & 2014	0.67034	0.67348	0.66524	0.66763	2294
V - 2013					
T - 2011					

Tr - Train

V - Validation

T - Test

Table 4. Model performance r^2 and RMSE with combination of data w.r.t. years

Principle component analysis results:

We performed PCA using matlab function princomp(). We have used mapminmax to normalize the inputs. This PCA function gives us two results, which are eigenvector and eigenvalues. Table 7 and Table 8 shows the eigenvector and eigenvalues for our input variables for 2012 data

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11

Water25	-0.2263	-0.0044	0.1557	0.0726	-0.0226	-0.4248	0.2317	-0.1399	0.8095	0.0988	-0.0002
Water100	-0.2889	-0.0079	0.1656	-0.0125	-0.0568	-0.5328	0.4239	-0.318	-0.566	-0.0247	0.0031
Slope	-0.1086	-0.045	-0.402	-0.0321	-0.0998	0.2022	0.7229	0.4916	0.0201	0.0896	0.0002
Component	0.0385	0.0171	0.2274	0.6829	0.642	0.1631	0.1966	0.0129	-0.0473	0.0214	-0.001
Sand	0.6226	0.0557	-0.0341	-0.0515	0.0414	-0.2351	0.1402	-0.0176	0.0019	0.1513	0.7111
Silt	-0.616	-0.0772	-0.2	0.228	-0.0253	-0.056	-0.3306	0.2102	-0.0561	0.021	0.6014
Clay	-0.1986	0.0198	0.3992	-0.2757	-0.0375	0.5613	0.2686	-0.3147	0.0927	-0.3097	0.3642
OM	0.0512	0.0479	0.5082	-0.1111	0.0102	-0.2541	-0.0386	0.6713	-0.0236	-0.455	0.0043

CEC7	-0.1182	0.0305	0.4742	-0.1897	-0.0185	0.1259	-0.0372	0.2075	-0.0938	0.8092	-0.0066
Rainfall	-0.1469	0.0084	-0.2124	-0.5894	0.7545	-0.1191	-0.0387	0.0145	0.0251	-0.0023	0.0001
Temperature	-0.0854	0.9924	-0.0791	0.0293	-0.0259	0.0073	-0.0018	0.0036	-0.0017	0.0004	-0.0003

Table 5. PCA Eigenvectors

We then computed the variance vector for these eigenvalues λ. Table 7 list the variance for the inclusion of each PC.

$$\text{Variance}_i = \frac{\sum_{i=1}^{n} \lambda_i}{\sum \lambda_i}$$

Principle Component	Eigenvalue
PC1	0.0634
PC2	0.0354
PC3	0.0179
PC4	0.0139
PC5	0.011

PC6	0.0063
PC7	0.004
PC8	0.0019
PC9	0.0011
PC10	0.0007
PC11	0

Table 6. Eigenvalues for Principal Components

Principle Component	Variance
PC1	0.4072
PC2	0.6348
PC3	0.7502
PC4	0.8396
PC5	0.9102
PC6	0.9504
PC7	0.9758
PC8	0.9882
PC9	0.9952
PC10	1
PC11	1

Table 7. Variance of PC in dimensions

From the above Table 7 we observe that inclusion of up to 9 PCs will contribute to the variance of approximately 99%. Which means reducing our variables to 6 dimensions using PC1 through PC9 will contribute to the total variance of approximately 99%. We have calculated 6 dimensional data for prediction i.e. V1, V2, V3, V4, V5, V6, V7, V8 and V9 corresponding to PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8 and PC9.

The PCs contain the weights of variables as a vector which helps us in finding the order of most significant variables. The higher the weight corresponding to attribute in PC for the highest eigenvalue, higher is the significance of that variable. The below Table 8 shows the order of significance for all the input variables.

Variables	Absolute weightage
Sand	0.6226
Silt	0.616
Water100	0.2889
Water25	0.2263
Clay	0.1986
Rainfall	0.1469
CEC7	0.1182
Slope	0.1086
Temperature	0.0854
OM	0.0512
Component	0.0385

Table 8. Input variables significance order as per PC1

Now we tried these reduced dimensions values V1 to V9 which are calculated from principal component analysis as an input to the neural network. The below Table 9 shows the r^2 and RMSE to predict yield. Each variable is first measured for individually, and then combined to another one by one. All these values are measured against output variable crop yield

Variable	r^2	RMSE
V1	0.4636	1912
V2	0.4972	1887
V3	0.4481	1901
V4	0.4539	1919
V5	0.4803	1892
V6	0.4065	1957
V7	0.4107	1944
V8	0.4263	1935
V9	0.4128	1943
V1, V2	0.5319	1862
V1, V2, V3	0.5673	1843
V1,V2, V3, V4	0.6194	1819
V1, V2, V3, V4, V5	0.6526	1801
V1, V2, V3, V4, V5, V6	0.6641	1792
V1, V2, V3, V4, V5, V6, V7	0.6685	1785
V1, V2, V3, V4, V5, V6, V7, V8	0.6711	1778
V1, V2, V3, V4, V5, V6, V7, V8, V9	0.6784	1765

Table 9. r^2 and RMSE of reduced PC dimensions to nine

The results in the above Table 9 shows that the reduced dimensions for 2012 data from PCA as input to the neural network measured r^2 and RMSE of 0.6784 and 1765 respectively. Similar experiment with all four years data together from 2011-2014 measured r^2 and RMSE of 0.6348 and 2348.

Now if you compare these results with stepwise iteration ANN results, both resulted in r^2 and RMSE of 0.7 and 1750 respectively. But in the case of stepwise iteration ANN, it measured more accurate than principle component analysis. Using PCA, you can just reduce the variable dimensions. This principle component analysis also helps in finding the significance of variables by using the first principle component or Eigen vector of ordered absolute weights (as shown in the Table 8) and try with different combinations of input weights of first principle component to the ANN model as shown in the table 2.

Decision tree analysis results:

Artificial Neural networks (ANN) is a black box model owing to the complex mathematical calculations involved in its algorithm. It is primarily used only for prediction purposes and not for explanatory purposes. But using the Decision tree with neural networks i.e. by modeling the predicted value of the neural networks with the independent variables using decision trees, 95% of the neural network model can be explained [13]. So we tried using mat lab to generate decision tree with those same 8 set of inputs variables and predicted value from neural network as a target and the below Figure 5 explains the model for the 2012 year data.

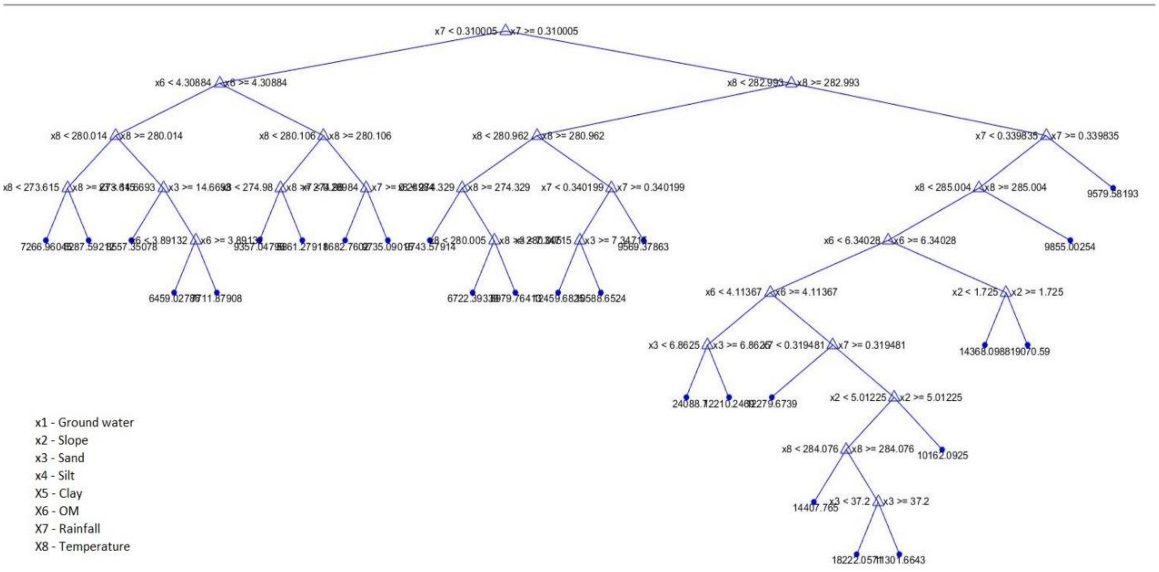

Figure 5. Decision tree explanation for the predicted values from the neural network

Initially, with actual data set, it generated a decision tree with 357 levels but using prune function that provided in the mat lab, Figure 5 is the decision tree that by setting the pruning levels to 20

We used mat lab predictor importance function to see the importance of each input variable and below is the table that shows the variable importance using decision tree with predicted values of neural net as a target values.

Rainfall	Temperature	Sand	Slope	Silt	Ground water	Clay	OM
2.2448	1.4282	0.6403	0.4339	0.387	0.3653	0.3362	0.2713

Table 10. Variable importance using decision tree

Also we have measured r^2 and RMSE of 0.91 and 250 respectively for decision tree with predicted values of neural net as a target values. Here, the value r^2 is indicating that we can explain about 91 % of the neural net model using this decision tree. This decision tree will help in explaining a neural net model using a tree spilt condition, mean and variance of predicted values and also to understand the variable importance order.

I have also tried the decision tree analysis on year 2012 actual data with all 11 variables as input to the model and yield from the same dataset as target and the below Table 11 shows the order of the variable importance.

Rainfall	2.1626
Temperature	1.441
OM	0.7494
Sand	0.4154
Slope	0.392
Ground water 100CM	0.3109
Ground water 25CM	0.3087
Silt	0.3038
Clay	0.2704
Component	0.1176
CEC	0.0968

Table 11. Variables importance using decision tree for 2012 data

As shown in the Table 11, if you see the variable importance, component and CEC are less importance than other variables and we have seen a same kind of pattern in the neural network model that we have ignored component and CEC as input to the model since their absence doesn't make any difference in the model results. In this experiment, decision tree model with 2012 as input, resulted in r^2 and RMSE of 0.63 and 1852 respectively.

Rainfall	389.7494
Temperature	335.9536
OM	136.727
Slope	115.1285
Sand	111.2099
Ground water 100CM	104.4266
Silt	95.7094
Ground water 25CM	95.3304
Clay	95.1794
CEC	36.7306
Component	28.0256

Table 12. Variables importance using decision tree for all four years 2011-14 data

Same as above experiment with 2012 data, as shown in the Table 12, all four years 2011-14 data as input to the decision tree resulted in r^2 and RMSE of 0.59 and 2483 respectively, and CEC and component are less important in predicting the crop yield. If you compare the decision tree model results with neural network model, neural networks are produced more accurate results than decision tree.

Model	2012		2011-2014	
	r^2	RMSE	r^2	RMSE
Neural network stepwise iteration	0.7	1750	0.66	2200
Principle component analysis	0.67	1765	0.63	2348
Decision tree analysis	0.63	1852	0.59	2483

Table 13. Results comparison of three models

As shown in the Table 13, when one compares the results of neural network stepwise iterative model with other two models which are principle component analysis and decision tree analysis, neural network model produced more accurate results with less error than principle component and decision tree analysis.

CHAPTER V

CONCLUSION

The purpose of this thesis study is to find the significance of different input variables provided in the datasets using an artificial neural network model and predict the crop yield of IOWA region. We have used different approaches such as stepwise iteration method using neural networks, principle component analysis and decision tree analysis to find the significance of input variables. We found that weather variables rainfall and temperature are very important and they have huge impact on predicting crop yield. The soil properties such as sand, silt, clay, organic matter and ground slope, ground water are also very significant in predicting the crop yield.

We have implemented a decision tree model with predicted values of neural network as target, to explain the neural network and to see the variable importance. We have also compared the ANN stepwise iteration method results with principle components through ANN results and decision tree analysis, neural networks produced more accurate results r^2 and RMSE of 0.7 and 1750 respectively.

We found that learning rate, number of hidden nodes, and the training function had an effect on ANN model development and the accuracy of ANN crop yield predictions.

If there is heavy floods in the test data region, then that weather data distribution makes big difference from training data distribution. As a future work, in this case, we can try a domain adaptation technique if the training data distribution is different from test data distribution [27].

REFERENCES

[1] Shearer S.A, Burks T.F, Fulton J.P, Higgins S.F, Thomasson J.A and Mueller T.G, (2000), "Yield prediction using a neural network classifier trained using soil landscape features and soil fertility data", Proceeding of Annual International Meeting, Midwest Express Center, Milwaukee, Wisconsin, 2000.

[2] Hornick B.S, (1992), "Factors affecting the nutritional quality of crops", American Journal of Alternative Agriculture 7 (1992) 63-68.

[3] Marchant J.A and Onyango C.M, (2002), "Comparison of Bayesian classifier with multilayer feed-forward neural network using example of plant/weed/soil discrimination", Computers and Electronics in Agriculture 39 (2002) 3-22.

[4] Hammer G.L, Kropff M.J, Sinclair T.R and Porter J.R, (2002), "Future contribution of crop modeling: From heuristics and supporting decision making to understand genetic regulation and aiding crop improvement", European Journal of Agronomy 18 (2002) 15-31.

[5] Kaul M, Hill R.L and Walthall C, (2005), "Artificial neural network for corn and soybean prediction", Agricultural System 85 (2005) 1-18.

[6] Alvarez R, (2009), "Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach", European Journal of Agronomy 30 (2009) 70-77.

[7] Wieland R and Mirschel W, (2008), "Adaptive fuzzy modelling versus artificial neural network", Environmental Modelling & Software 23 (2008) 215-224.

[8] O'Neal M.R, Engel B.A, Ess D.R, Frankenberger J.R, (2002), "Neural network prediction of maize yield using alternative data coding algorithms", Biosystems Engineering 83 (1) (2002) 31-46.

[9] Drummond S, Joshi A and Sudduth K.A, (1988), “Application of neural network: Precision farming, in: Proceedings of Neural Networks”, IEEE International Joint Conference, Anchorage, AK, USA, 1998.

[10] PACHEPSKY Y.A, TIMLIN D AND VARALLYAY G, (1996), “Artificial neural networks to estimate soil water retention from easily measurable data”. Soil Science Society of America Journal 60 (1996), 727–733.

[11] J I B, SUN Y, YANG S AND WAN J, (2007), “Artificial neural networks for rice yield prediction in mountainous regions”. Journal of Agricultural Science 145 (2007), 249–261.

[12] Siti Khairunniza-Bejo, Samihah Mustaffha and Wan Ishak Wan Ismail, (2014), “Application of Artificial Neural Network in Predicting Crop Yield”. Journal of Food Science and Engineering 4 (2014) 1-9.

[13] Setiono, Rudy, and Huan Liu. "Understanding neural networks via rule extraction." IJCAI. Vol. 1. 1995.

[14] http://www.netstate.com/states/geography/ia_geography.htm.

[15] Carlos Gershenson, “Artificial Neural Networks for Beginners”.

[16] http://horicky.blogspot.com/2012/05/predictive-analytics-data-preparation.html.

[17] Schillinger, William F., Steven E. Schofstoll, and J. Richard Alldredge, (2012),” Predicting Wheat Grain Yields Based on Available Water”. Washington State University Extension, 2012.

[18] Li, Eldon Y. "Artificial neural networks and their business applications". Information & Management 27.5 (1994): 303-313.

[19] Gavrilov, A. V., and V. M. Kangler. "The use of Artificial Neural Networks for Data Analysis." The Third Russian-Korean International Symposium on Science and Technology.-Novosibirsk: NSTU. 1999.

[20] Nik Haslinda, Abdul Halim and Salwani Abdullah, 2013. “Forecasting the General Examination Results Using Back Propagation”. Asian Journal of Information Technology, 12: 1-6.

[21] Yusoff, Nor Eleena, et al. "Application of back propagation neural network and ANFIS in forecasting university program". Science and Social Research (CSSR), 2010 International Conference on. IEEE, 2010.

[22] Bailey, David, and Donna Thompson. "How to develop neural-network applications". AI expert 5.6 (1990): 38-47.

[23] Mohamad Zin, R., et al. "Predicting the performance of traditional general contract projects: a neural network based approach". 6th Asia-Pacific Structural Engineering and Construction Conference, 2006: C78-C86.

[24] Al Shalabi, Luai, Zyad Shaaban, and Basel Kasasbeh. "Data mining: A preprocessing engine". Journal of Computer Science 2.9 (2006): 735-739.

[25] George, K. M. "An application of PCA to rank problem parts." Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology. ACM, 2012.

[26] https://www.quandl.com/collections/economics/agriculture-share-of-gdp-by-country

[27] Jiang, Jing. "A literature survey on domain adaptation of statistical classifiers." URL: http://sifaka.cs.uiuc.edu/jiang4/domainadaptation/survey (2008).

[28] http://www.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html

[29] Wilamowski, Bogdan M. "Neural network architectures and learning." Industrial Technology, 2003 IEEE International Conference on. Vol. 1. IEEE, 2003.

VITA

Naveen Singireddy

Candidate for the Degree of

Master of Science

Thesis: ANALYTICS BASED ON ARTIFICIAL NEURAL NETWORK: A CASE STUDY BASED ON IOWA CORN YIELD FORECASTING

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for the Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in July, 2015.

Completed the requirements for the Bachelor of Technology in Computer Science and Engineering at Jawaharlal Nehru Technological University, Hyderabad, Telangana/India in 2010.

Experience:

Worked as Graduate Teaching and Research Assistant at Oklahoma State University, Stillwater, Oklahoma for 2 years

Worked as Software development engineer at Broadridge Financial Solutions, Hyderabad, India for 3 years

Worked as Software analyst intern at N2N Services, Hyderabad, India for 3 months