Development of a Decision Support System to increase the Tea Crops yield

Umer Iqbal, Muhammad Shahbaz, Adnan Khalid and Qamar-uz-zaman

***Abstract* – Pakistan is the largest consumer of tea, but not the biggest producer of it. To fulfill its needs Pakistan expends a large amount of foreign reserve to import tea. To save foreign reserves and fulfill tea needs, we have to make precise decisions at the right time to increase the tea yield to maximize the utilization of land suitable for it. The factors on the basis of which farmers, researchers and the government make decisions to increase tea crop yield are environmental and soil conditions such as soil PH level, humidity level and rainfall level as well as Plucking rounds. We can use these factors to make a right and timely decision by using the decision support system with data mining because it encompasses both classical statistics and modern machine learning techniques. To develop the Decision Support System with Data mining we use different methods for decision making, such as decision trees, naïve Bayes, neural network and linear regression. These techniques utilize the dataset for factors affecting the tea yield, which is collected from black tea farms located in Shinkyari, Mansehra to build and train a data mining model to discover hidden patterns and relationships. Once we get a stable data mining model for decision support system using different techniques, we can have accuracy through cross validation for each. In this regard neural network has higher accuracy followed by decision tree and then other techniques. In this study, decision support system is developed to find the best environmental conditions to maximize the tea plan production.**

***Index Terms* – Data mining, Tea, Production, Decision support system, Data modeling techniques**

I. INTRODUCTION

Tea is one of the main consumer beverages of the world. Pakistan is among the largest consumers of tea, but unfortunately is not a good producer. Pakistan was the 5th largest importer of tea and imported 120,345 metric tons of tea in 2010 [1]. Pakistan is importing black tea from 19 different countries and importing green tea in small quantities from 4 different countries. The annual consumption of tea in Pakistan is 1 kg per capita [2]. To save the foreign reserve, which are consumed in tea import, Pakistan needs to improve and encourage the local farmers for tea production. The production of tea started in 1973, under National Tea Research Institute NTRI. But still the local production of tea is very low compared to its demand.

Umer Iqbal and Muhammad Shahbaz is from Department of Computer science and Engineering, University of Engineering and Technology, Lahore, Pakistan. Adnan Khalid is from Department of Computer Science, Government College University, Lahore, Pakistan. Qamar uz Zaman is from National Tea Research Institute, Shinkyari, Manshera. Email: engr.umeriqbal@gmail.com. Manuscript received May 6, 2015; revised on August 15 and November 11, 2015; accepted on December 7, 2015.

To maximize local production and attract investors to invest in growing tea crops, NTRI is taking a lot of measures such as distributing free seeds and arranging awareness sessions among the farmers.

Decision Support System (DSS) is a software application which processes the data, presents it to the user and makes a decision according to that data but it doesn't bring complete information to fulfil requirements. To make a decision support system more effective and efficient we use data mining (DM) techniques. Data mining decision support system (DMDSS) utilizes attributes of temperature, soil PH, humidity and rainfall and plucking rounds to develop models which help in making decisions.[11, 12] Our research work facilitates many areas of tea crop growth, such as helping farmers to increase tea crop yields, helping researchers, aiding the Government, assisting policy makers, management, regulating tea price controlling agencies, encouraging banks to give loans, incentivizing investors and Importers etc.

Many decision support systems for tea industry have been designed, developed, deployed and utilized in domains of marketing, land selection and quality but very little work has been done on tea production such as to make an optimized decision to select a best suitable site for tea, the researchers has developed the framework which helps the management regarding land allocation. [3]. Secondly, researchers have proposed an information service platform based on GIS to farmers, investors, businessmen and government, which includes three parts. i) GIS platform business oriented e-commerce for tea crop. ii) To guide the farmer about production, production information service platform for gathering tea related data. iii) Decision making and planning for industries and government is through GIS Platform [4]. To estimate the overall quality of tea, the researcher uses the fuzzy neural network for tea classification on basis of smell and taste. [5] To improve the quality of tea smell, researchers have used electronic nose with neural networks to grade the tea quality. [6]

For the development of DMDSS, collected data is pre-processed first, and then loaded into a database. After the data loading, data mining models are developed using different techniques as discussed in section 2. These models are trained and tested by dividing the data into different parts such as training, testing and validation. After several experiments Neural Net has high accuracy and best performance.

The organization of the Paper is as follows; Section 2 is about methods used in DMDSS. Section 3 is the introduction of DSS for tea production. Section 4 explains data pre-processing. Section 5 gives details of data analysis through graphs. Section 6 explains the process flow for the

 ISSN – 1999-4974

development of DMDSS. Section 7 explains experimental results. And last section includes conclusion and future recommendations.

II. METHODS USED IN DMDSS

DM is a collection of interactive and iterative discovery processes. The basic purpose of DM is to the extract knowledge, association patterns and relationships from the dataset which is unknown, nontrivial and unformulated. The integration of the DSS with data mining can help the decision support system to discover the new rules and decisions and dig more useful information from the dataset. [7]

The Goal of the DMDSS is to develop a model which has high accuracy and reliable prediction and is easily understandable. It is divided into three major categories, clustering, associated rules and classification. [7] All the techniques of DM improve the decision power of the DSS are discussed below.

A. Naïve Bayes

It is a statistic classifier and predicts the probability that a sample belongs to a particular class. Bayesian classification is based on the Bayes theorem. Simple Naive is also known as a Naïve Bayesian Classifier. It is assumed that attribute value is independent of the other attributes for a given class that is why simple computation is involved and so we consider it as "naive". Bayesian Classifier has higher speed and accuracy than the decision trees and neural networks on large databases. [8] Bayes rule states that if you have data G and hypothesis A, then Formula is

$$P\,(A|G) = \left[\frac{P\left(\frac{G}{A}\right)P(A)}{P(G)}\right] \qquad (1)$$

P (G) = independent probability of G
P (A) = independent/ prior probability of A
P (G/A) = Conditional/likelihood Probability of G given on A
P (A/G) = Conditional/Posterior Probability of A given on G

B. Linear Regression

We often need a method that predicts one variable with respect to another. For the continuous value prediction of one variable, which is dependent on other independent or predictor variables we use regression and log linear to find an approximate dataset. The examples of how much gas cost with respect to external climatic parameters can explain this method.

In the Simple Regression Technique response variable Z with respect to predictor variable Y model the linear function. Data is fitted with a straight line through the model. The equation for linear regression is below

$$Z = a_0 + a_1 Y \qquad (2)$$

The a_o and a1 are the regression coefficients. The point where the line crosses the Z is a_o and a_1 is called interceptor. The change in Z with respect to Y is a_1 and is called the slope.

C. Neural Net

The Neural Net is basically inspired from the human brain and is used to solve a variety of problems such as control, memory, clustering, classification, optimization, content addressable and function approximation. [8] The structure of the neural network is a group of layers, which are composed of neurons. The structure has three parts, input layer which takes input from the external environment, output layer which gives result and intermediate or hidden layer which processes the data. The weighted sum of all the input processing elements is computed through the summation function. The formula is given below.

$$Zj = \sum_{k=1}^{n} X_k W_{jk} \qquad (3)$$

D. Decision Tree

Number The Decision tree is a classification technique applied on the give data set to generate a tree model or set of rules. It is an inverted tree, on the top of the tree is a root and bottom has leaves and intermediate is the nodes. The datasets used for decision trees are divided into two parts. One used for the model building is called training data set and the other used to test the model is called testing dataset.

Entropy is the Theoretical approach to find the split goodness. The amount of information in the attribute is measured through it. P (E) is a probability of Event E will happen and E is from 1 to N, where is a number of classes.

$$\text{Entropy (I)} = \sum_{E=0}^{n}(-P(E)\log_2 P(E)) \qquad (4)$$

Iterative Dichotomiser 3 (ID3) method invented by Ross Quinlan is credited as a pioneer effort in the family of decision tree building algorithms [2]. ID3 further uses various methods to decide the attribute to which the tree will split. These methods are called attribute selection measures. Some attribute selection measures are Information Gain, Gini Index and Gain Ratio [8].

The expected information needed by ID3 algorithm to assign a class to an arbitrary tuple in data (D) is represented by the following formula

$$\text{Info}(D) = \sum_{i=1}^{n}[P_i \log_2(p_i)] \text{ Where } P_i = \frac{|C_{i,D}|}{|D|} \qquad (5)$$

Pi represents the probability for an arbitrary tuple in data (D) being labelled as class Ci. Info (D) represents the average amount of information required to recognize the class label of a tuple in D, is known as the entropy of D.

III. DECISION SUPPORT SYSTEM FOR TEA PRODUCTION

Pakistan is an agricultural country, which produces a variety of crops in large quantity for its population, but is not producing the tea to fulfil its population needs. Pakistan has established NTRI to get better quality and quantity of tea but it still has not reached the target to produce the tea locally to fulfil the country's need. To get this target, we contribute our

 ISSN – 1999-4974

efforts for development of DMDSS for the tea plant to improve yield minimizing effects such as attacks by different pests on tea leaves, climatic changes, unsuitable soil, lack of farmer's scientific approach etc.

Researchers usually take decision for tea plantations based on their experience that may be flawed. It is hard to make correct decision just on an experience basis. To make decision more accurate and effective we use data driven decision making approach. For data driven approach, we utilize the historical data collected from tea gardens, integrating these into dataset using modern information management techniques such as data warehousing followed by data mining techniques to develop a model upon which we make decisions.

IV. DATA PRE-PROCESSING

NTRI is doing research in multiple fields of tea such as land suitable for tea field, pest controlling, tea yield, tea processing, agro-meteorology and other related areas [2]. All of the above are documented either in hand written register or excel sheet. From these documents we collect 7 years (2007–2014) data and make it available in digital form (Microsoft excel sheet) for further processing.

From the collected data, eight attributes were selected on the basis that they affect tea production directly or indirectly. These attributes include tea garden's minimum temperature, maximum temperature, soil PH, labour training, per rainfall, humidity, labour cost, usage of pesticide and plucking rounds in the month. This data is in continuous format i.e. having numeric values. To explore the effective relationship between these selected attributes, they are converted into discrete format, i.e. in labels. We chose appropriate labels for a range of numeric values in each attribute.

After discretization of the dataset, it was observed that it has less redundancy or insignificant information. To remove redundant values, we selected distinct rows and neglected all those rows which had null values for all columns. We got an expert opinion of NTRI researchers for real values in a row, which has two values for same month.

It was analysed that the two attributes labour training and usage of pesticide has no effect on tea yield as their values are constant for all rows, so they are not included in the dataset. Another attribute per labour cost was also neglected because it has an indirect effect on tea plant production. Consequently, only five independent attributes (temperature, Humidity, rainfall, Plucking Rounds and soil PH) and one class attribute (tea yield) were selected for development of DMDSS.

V. DATA ANALYSIS THROUGH GRAPHS

After data pre-processing, we analyses data to get hidden patterns, relationships and trends in data by using simple graphs (pie chart, line graph, bar chart) and surface graph. Firstly, we use simple graph, to do analyses for dependent attribute (tea yield) against independent attributes (rainfall, humidity, temperature, and soil PH and plucking rounds). After this we utilize the surface graph to analyze effect of two independent attributes against dependent attribute. Surface graph help in analyzing the 3 attributes at a time, which help in deciding which factor has more effect on yield than others and their relation with each other.

VI. PROCESS FOR DEVELOPMENT OF DMDSS

For development of decision support system a number of process models are there such as CRISP-DM and SEMMA. [9, 10]. Our process of developing the DMDSS is hybrid of above two processes and we call it TDMDSS (Tea Data Mining, Decision Support System) Process as shown in Fig 1. The detail of its six steps is below.

First step understanding of tea domain, especially the factors affecting the tea crop yield.

After identification of factors, data related to these are collected and combined in one place to create the dataset.

Dataset needs modification due to some issues such as missing values, an outlier in the data, duplication, and noise. After modification, the dataset is transformed into the required format and data types.

After dataset modification we used it for exploration using the simple graphs and surface graphs to find the relationship, patterns, and trends.

After dataset modification, we develop data mining models for techniques such as decision tree, ID3, linear regression, Neural Net, Naïve Bayes. All of these are discussed in section 2.

After development of DM models, its rules, relation, patterns and results are evaluated. In evaluation process, we take help of tea crop domain experts and data mining expert to evaluate the quality of the models developed.

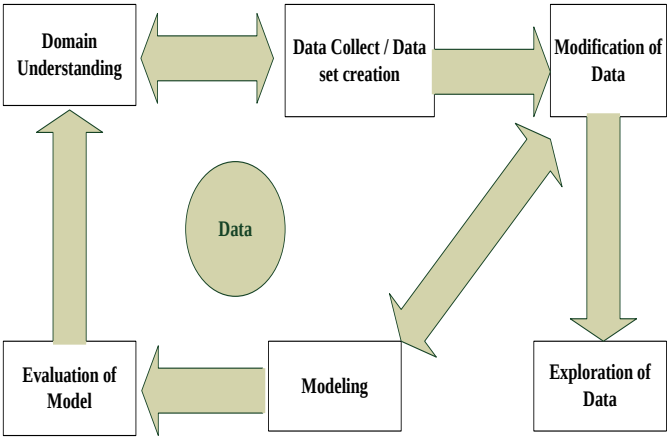

Fig 1: Process Flow of Tea DMDSS

VII. EXPERIMENTAL RESULTS

The basic purpose of these experiments is to evaluate performance and accuracy of the DM models for different techniques. This helps in decision making and prediction to increase the tea crop yield.

Rapid Miner software was used to train, test and validate different data mining models. For this we have divided the

ISSN – 1999-4974

data into different parts according to data mining techniques. The training data import data into rapid miner and then use different data mining techniques to train models. Train model is tested through testing data. Lastly, we use the cross validation technique to validate the model and its performance.

For building the Neural Net model for decision making we divide the data into three part training, testing and validation of 70, 20 and 10 percent respectively. Different experiments are performed by setting the number of hidden layers and the number of nodes in it. The optimal result found through 1 hidden layer of 8 nodes, learning rate 0.5, training cycles 100 and shuffle and normalize are enabled and momentum is 0.3334. Confusion matrix generates from models of decision tree give an accuracy of 75 %.

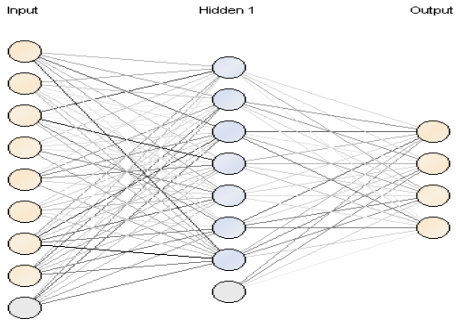

Fig 2: Neural Network

For Decision trees, tea plant data tuples were divided into 90% training tuples and 10% test tuples. It was seen that most accuracy of the model is based on the criterion of information gain and disable pruning. Setting the minimum split size to 21 and leaf size to 1, we get optimal result. Confusion matrix generates from models of decision tree give an accuracy of 70 %. In the below Fig 3 of Decision tree there are many terms used such as "PC" means plucking counts, PH1 means soil PH (Hydrogen Potential) up to 15 cm, PH2 means Soil PH at 15 to 30 cm below the soil, PH3 means Soil PH at 30 to 45 cm below the soil, Y1 means low yield, Y2 means Medium yield, Y3 means high yield, Y4 mean very high yield.

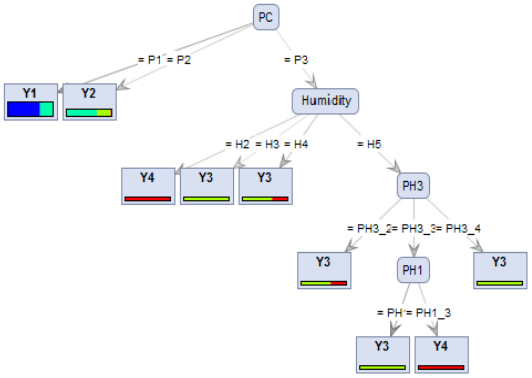

Fig 3: Decision Tree

We use Naïve Bayes to predict the classes of attributes. The dataset is divided into two parts training and testing of 80 and 20 percent respectively. Different experiments are performed, but optimal result is found by enabling the Laplace, local random seed is 1, sampling type to linear sampling, number of validation 2, Laplace correction and leave one out to false and create complete model, average performance only and use local random seed to true. There are 8 distributions for each yield. Yield Y1 has highest Probability of 0.495.Confusion matrix generates from models of Naïve Bayes give an accuracy of 68 %.

We use the linear regression to get the straight line to predict the value of the given data set. The dataset is divided into two parts training and testing of 80 and 20 percent respectively. Different experiments are performed, but optimal result is found by Feature selection in M5 Prime, Min Tolerance 0.05, and enabling Eliminate Collinear Feature.

Performance evaluation of data mining model is carried out through the Cross Validation process. Accuracy of models is checked through a dynamic number of folds. In this process data is divided into two part training and testing, training data is used to develop DM models, which is then tested using testing part, which has two parts apply model and performance evaluation. After validation, performances of multiple data mining models are given in table I. In this we can see that neural net has higher accuracy than others and it is followed by a decision tree. ID3 and Naïve Bayes have lower performance. It was observed that as the number of validations fold increases performance also increases.

TABLE 1: PERFORMANCE OF DATA MINING MODEL

Folds	Models			
	DT	ID3	Neural Net	Naïve Bayes
2	68	62.63	67.31	61.56
5	71.49	62.35	71.03	67.45
10	70.41	62.58	70.32	67.72
15	70.58	61.02	70.33	67.96
20	70.32	61.2	71.69	68.49
25	69.1	62.59	69.61	68.76
Average %	69.983	62.0616	70.0483	66.99

We also analyse data through basic graph and results we obtained are given below. Optimal tea yield each year is from May to August (Fig 4). Temperature between 20 ˚C and 30 ˚C has an optimal tea yield (Fig 5). PH of value between 4 and 5 has optimal yield (Fig 6).Humidity has very little effect on yield, the reason for this is Siren River is flowing on one side tea gardens, which creates required humidity for tea plants, Rainfall up to 250 mm has an optimal yield of tea (Fig 7). We see that higher plucking round results in higher tea yield and vice versa (Fig 8, 9).

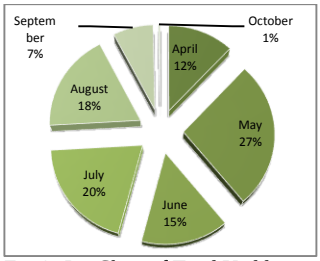

Fig 4 : Pie Chart of Total Yield per month

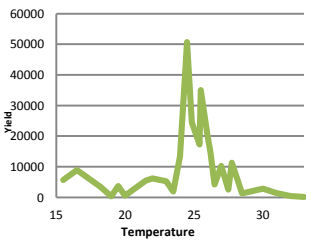

Fig 5: Line Graph of Temp on x-axis and Yield on y-axis

ISSN – 1999-4974

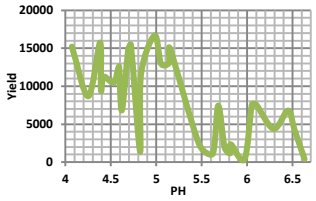

Fig 6: Line Graph of PH on x-axis and Yield on y-axis

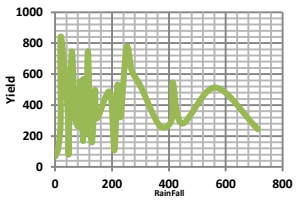

Fig 7: Line Graph of Rainfall on x-axis and Yield on y-axis

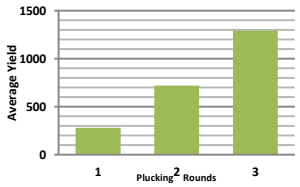

Fig 8: Graph showing Average Yield in plucking rounds

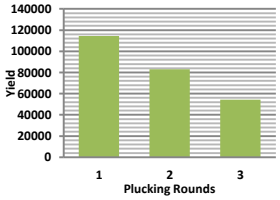

Fig 9: Graph showing Total Yield in plucking rounds

After Analysis of data through the basic graph we analyse the data through a surface graph in which we take 2 attributes at a time against yield of tea crop. The results we obtained are given below. By analysing the PH and Temperature against yield we get that if the temperature remains between 30 C and 25 ˚C a decrease in PH value of soil results in an increase of tea crop yield. Tea crop yield decrease as temperature a rising above 30 ˚C or below 25 ˚C (Fig 10). Temperature between 22 and 29 ˚C and rainfall between 0 and 100 has a maximum yield for tea crop. We observe that as the temperature value move up or below 25 ˚C and decrease in rainfall has a low yield of tea crop (Fig 11). The rainfall decrease with respect increase humidity results in increase yield of tea crop vice versa (Fig 12). As temperature and humidity increases, yield of the tea crop also increases and vice versa. Yield of the tea crop becomes optimal when humidity is above 85, which create feeling temperature is needed (Fig 13).

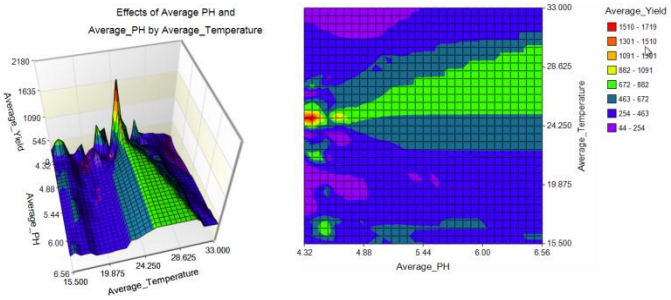

Fig 10: Surface Graph - Effect of PH and Temperature on Yield

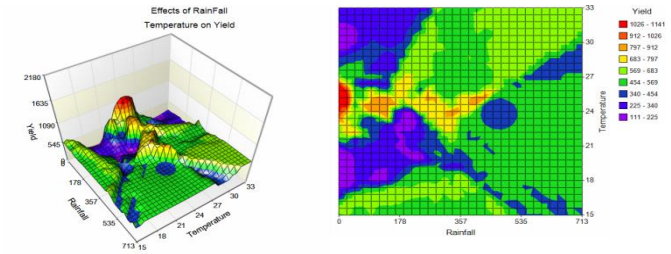

Fig 11: Surface Graph - Effect of Rainfall and Temperature on Yield

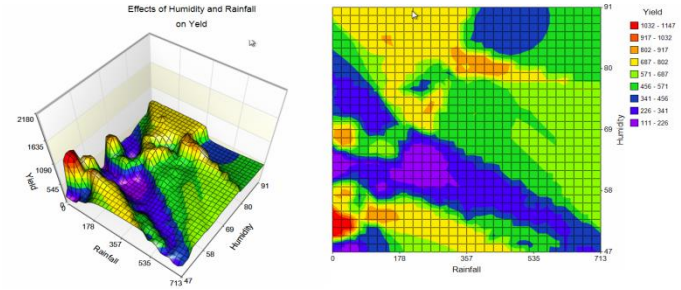

Fig 12: Surface Graph - Effect of Rainfall and Humidity on Yield

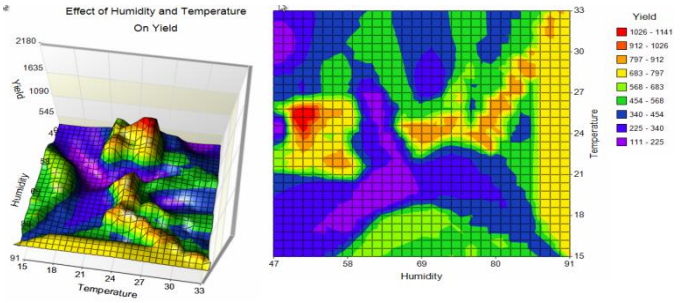

Fig 13: Surface Graph- Effect of Humidity and Temperature on Yield

VIII. CONCLUSION

The collected data from tea farms is pre-processed and then divided into different parts according to DM technique. Training data is used to train the DM models. Later on this model was tested on test data tuples. Lastly validates data through cross validation techniques. It was observed that DM models predicted 70% of accuracy. Model with such accuracy can be successfully deployed in the real time environment to get maximum tea production. We have also observed in section 7 that how data mining techniques discover the hidden patterns and relationships between factors like soil conditions such as soil PH level, humidity level and rainfall level, Plucking rounds that affect the tea yield. Moreover, these data mining models allow tea farmers to better understand the effects of different attributes on tea production. This model can be improved by taking many other parameters into consideration that affect tea plants such as pesticide usage, pest types and daylight duration. We can collect more detailed data by using wireless sensor network. We can create an online portal so that farmers, investors, planners and others can get help with it.

REFERENCES

[1] F.S. Hamid , "Tea in Pakistan", International Society of Tea Science, Page 21- 24,1996

[2] Tavakkoli-Moghaddm, R., Siadat, A. ; Kaboli, A. "A decision framework for location-allocation problems: A case study in tea industry", Management of Innovation and Technology, 4th IEEE International Conference, 1061 – 1065, Sept 2008

[3] Peng Lv, Feng Jing, Jinzhuang Xue, Yunfei Wang and Weihong Cui, "Development of a GIS-based Public Information Service Platform for Chinese Tea Industry" , Multimedia Technology (ICMT), International Conference, 1-4 ,Oct 2010

[4] Runu Banerjee (Roy) , Angiras Modak , Sourav Mondal, Bipan Tudu , Rajib Bandyopadhyay and Nabarun Bhattacharyya, "Fusion of electronic nose and tongue response using fuzzy based approach for black tea classification" , International Conference on Computational

ISSN – 1999-4974

Intelligence: Modeling Techniques and Applications (CIMTA), 605-622, 2013.

[5] S. Borah , E. L. Hines, M. S. Leeson, D. D. Iliescu, M. Bhuyan and J. W. Gardner, "Neural network based electronic nose for classification of tea aroma " , Springer Science+Business Media, 7-14, Dec 2007

[6] Zhidan Wu and Yue Yang , "Research and Design of Decision Support System based on Data Mining and Web Technology", Management and Service Science (MASS), International Conference, 1-3 , Aug-2010

[7] Jiawei Han, Micheline Kamber, Jian Pei "Data Mining Concepts and Techniques, Third Edition", Morgan Kaufmann Publishers, June 2011.

[8] Chen Wei-Chou,Hong Tzung-Pei,Jeng Rong, "A framework of decision support systems for use on the World Wide Web," Journal of Network and Computer Applications, 1-17, Oct 1999.

[9] Shearer C, "The CRISP-DM Model: The New Blueprint for Data Mining", Journal of Data Warehousing, 13-22. Nov 2000

[10] Sair, Sarwar, Fayaz Ahmed, Abdul Waheed, Qamar uz zaman, "Study of determindation of Neutrient status of NTRI tea garden soil " journal of Science and Technology Development, 39-43, 2011

[11] Abdul Waheed, F.S.Hamid, Habib Ahmad, Sohail Aslam, Naseer Ahmad, Ahmad Akbar, "Different climatic data observation and its effect on tea crop" , Journal of Material and Environment Science, 299-308, 2013.

 ISSN – 1999-4974