A COMPARISON OF CLASSIFIERS PERFORMANCE FOR PESTICIDE USAGE AND TOXICITY CATEGORIZATION TO CONTROL EXCESSIVE PRACTICES

***Mrs. G. Jaya Laksmi (Ph.D), **Dr. Suneetha Manne, ***Ms. M. Lavanya**

**Assistant Professor, IT Department, VR Siddhartha Engineering College, Vijayawada, India.*

***Professor & Head, IT Department, VR Siddhartha Engineering College, Vijayawada, India.*

****M. Tech Scholar, IT Department, VR Siddhartha Engineering College.*

ABSTRACT

The increasing amount and complexity of data in toxicity prediction based on intelligent methods for mining the data triggered different classifiers to be applied in toxicity prediction. Around 70 percent of our population is directly engaged in agriculture .The goal of this research to identify the high toxic pesticides used by farmers for crop production .To discover this knowledge ,data mining classification techniques has been implemented to differentiate pesticides usage and compared to analyze the best classifier. By this, awareness is created to avoid highly toxic pesticides which are creating acute and chronic health hazards in humans.

Index Terms*— pesticides, toxicity, classification, data mining, agriculture.*

INTRODUCTION:

India is an agricultural country. About 70% of our population depends upon agriculture. 33% of our National income originates from agriculture. The development of agriculture has much to do with the financial benefit of our country. The agriculture part of India has involved nearly 43 % of India's topographical region. The total Share of Agriculture, as far as rate of Gross Domestic Product is 13.9 percent among 2013-14. For the twelfth Plan (2012-17), a development focus of 4 percent has been set for the Agriculture Sector. According to the fourth Advance Estimates of Production of grains for 2014-15, total food grain creation is assessed to be 264.77 Million Tons. Indian farming population is in crisis and facing a huge number of problems for increasing the efficiency of crop production. One of the major tasks in agriculture production is pest management.

India’s utilization of pesticides is 76 per cent as against the world average of 44 per cent.

Excessive use of pesticides is harmful in several ways. On one hand, farmers have to pay more for the pesticides, while on the other, increased pesticide usage making more harmful to the

crops [1]. Without having proper guidance, the farmers are applying wrong, untimely, unnecessary, and excessive pesticides. It causes several problems for farmers, environment and public health.

Pesticide toxicity is the measure of the ability of a pesticide to cause injury. Toxicity represents the kind and extent of damage that can be done by a chemical. In other words, if you know the toxicity of a pesticide, you know "how poisonous" it is [2]. For modeling the ecological data like agriculture, the data mining techniques are used.

DATA MINING IN PESTICIDE TOXICIY CLASSIFICATION:

Data mining is the process of discovering interesting, knowledge such as patterns, associations, changes, anomalies and significant structures, from large amount of data stored in database, data warehouses or other information repositories [3].Data mining has been popularly treated as a synonym of knowledge discovery in database although some researchers view data mining as an essential step of knowledge discovery [4].

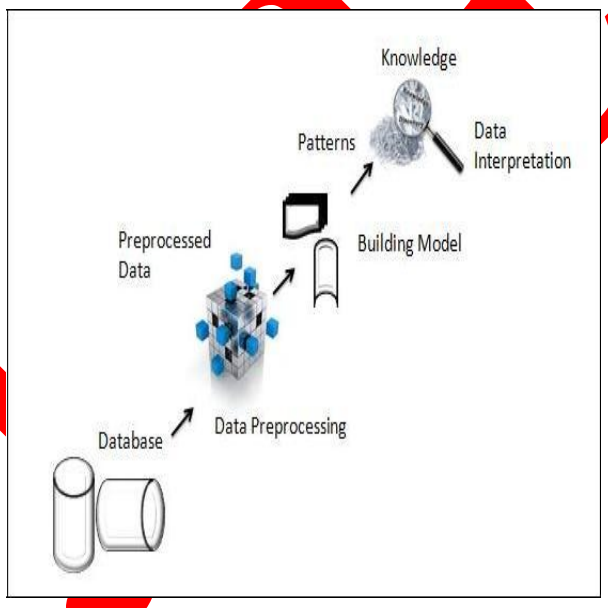

Figure1: An overview of the steps comprising the KDD process

In general, a knowledge discovery process consists of an iterative sequence of the following steps:

- Data cleaning
- Data Integration

- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

The data mining classification techniques like decision tree, neural networks, Support vector machine etc., and helps to classify the toxicity levels in pesticides applied on crops for good crop production based on LD50 value of a pesticide. This pesticide toxicity classification helps to classify the toxic levels in pesticides and to indicate the less usage of high toxic pesticides and adopt alternative solutions to avoid human health hazardous.

RELATED WORK:

A study conducted by [Rabiaimtiaz, Dr. Malik Sikandar Hayatkhiyal, Dr. Shahid Khalil, Aihabkhan, and Dr. Ahsan Abdullah] on Effect of Pesticides on Human Life using Visual Data Mining reported that Chernoff faces are used to deduce the data in understandable form and the clustering of agricultural data is performed by Chernoff faces method. This paper explores the concept of multidimensional data[5] .Other study conducted by [Frank Lemke, Emilio Benfenati, John Adolf Muller] on Prediction of Acute Toxicity of Pesticide Residues reported that Implements a concept for developing alternative tools for toxicity modeling and prediction of chemical compounds to be used for evaluation and authorization purposes of public regulatory bodies to help minimizing animal tests, costs, and time associated with registration and risk assessment processes[6]. [Jyothy S T, Deepu Kumar T L, Tata Elxsi L, Dr. Andhe Pallavi] Analysis of algorithms in Pesticide Toxicity Classification. This paper proposed the algorithms tell whether the pesticide will belong to a level ranging from highly toxic level to non toxic level. By knowing the toxicity level of a pesticide, proper pesticide can be used for infected plant [7]. **[**Emilio Benfenati, Paolo Mazzatorta1, Daniel Neagu, and Giuseppina Gini] Combining classifiers of pesticides toxicity through a neuro-fuzzy approach. The objective of this exploration is to apply neuro-fluffy systems to give a change in joining the consequences of five classifiers connected in harmfulness of pesticides [8]. [Francisco Prieto Garcia, Sandra Y. Cortés Ascencio, John C. Gaytan Oyarzun, Alejandra Ceruelo Hernandez and Patricia Vazquez Alavarado] Pesticides: classification, uses and toxicity. This paper proposes the measures of exposure and genotoxic risks. The principle strategies utilized in the appraisal of pesticides introduction are the historical backdrop of presentation, the specialist's assessment and natural and organic observing [9]. **[**Elena Lo Piparo

and Andrew Worth] Review of QSAR Models and Software Tools for anticipating Developmental and Reproductive Toxicity. This report gives a cutting edge survey of accessible computational models for formative and regenerative lethality, including Quantitative Structure-Activity Relationship (QSARs) and related estimation techniques, for example, choice tree methodologies and master frameworks [10].

[M.G. Hill, P.G. Connolly P. Reutemann, D. Fletcher] The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. This paper proposes the models with as few as four traits gave valuable figures, and plantation administration qualities were the most imperative determinants of model anticipating precision [11]. [Ankur Omer, Poonam Singh, N. K. Yadav, R. K. Singh] An Overview of Data Mining Algorithms in Drug Induced Toxicity Prediction. This paper proposed the machine Learning, one of the powerful in silico data mining techniques has evolved as the most efficient and powerful tool for exploring new insights on combinatorial relationships among various experimental data generated [12]. [Gongde Guo, Daniel Neagu, Xuming Huang, and Yaxin] An examination concerning the blend of distinctive classifiers for danger forecast. These characterization techniques included in creating classifiers for blend are picked as far as their speak to capacity and assorted qualities which incorporate the Instance-based Learning calculation (IBL), Decision Tree learning calculation (DT), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Multi-Layer Perceptions (MLPs) and Support Vector Machine (SVM)[13].

DATA SET:

Pesticide usage is the best approach to increase the crop productivity. There are different parameters consider to classify the toxicity of a pesticide. Parameters are particularly crop name, pest name, season, type, pesticide name, preferred dose/acre, applied dose /acre, LD50 value of a pesticide and predefined Ld50(recommended by WHO),toxic level and category. On the basis of these parameters classification can be applied to know the toxicity levels of a pesticide. The data to classify the toxicity is directly collected from the farmers. The collected raw data is used to derive the parameters we need. Each attribute is described in the following table:

Attribute name	Attribute description	Values	Data Type
Crop name	The name of the particular crop.	Paddy.	Nominal
Pest name	The name of the particular pest that affects the crop.	Blast, sheath blight, stem borer etc.	Nominal
Season	Pest occurs in season	Kharif, rabi	Nominal
Type	Type of pest	Fungal, Insect, Virus, etc	Nominal
Pesticide name	Name of the pesticide	Coragen, malathion, etc.	Nominal
Preferred dose/acre	Preferred dose recommended by the IPM	50kg/acre etc.	Float
Applied dose/acre	Applied dose(used by farmers)	80kg/acre etc.	Float
LD50	LD50 value of a particular pesticide	5000mg/kg,180mg/kg etc.	Double
PredefinedLD50	PredefinedLD50 value given by WHO	<50,>=50-500,>500-5000.	Double
Toxic level	Toxicity of a pesticide	Low,high, moderate	Nominal

Figure 2: Pesticide Data set

The above table shows the input for classifying by some classification techniques. The table contains crop name means which type of crop paddy, wheat etc. Another attribute is season. The time of yield production e.g. khariff or Rabi.Pest name, either fungal or viral pests are occurred for

crops. The pesticides are applied in prescribed dose but farmers used excess amount of pesticides than suggested.LD50 value of a pesticide can be calculate by using the chemical composition and LD50 value of each value. Then take the predefined LD50 values which are recommended by WHO. Based on the data, the classification algorithms classifying the pesticides as low, moderate and highly toxic.

PROPOSED SYSTEM OVERVIEW:

Till now research is not done in the pesticide toxicity classification using various data mining techniques in India to avoid high-toxic pesticides usage and enrich farmers to use proper pesticide used for infected plant. The classification algorithms like decision tree, k-Nearest Neighbor helps to classify the toxicity levels in pesticides applied on crops for good crop production.

CLASSIFICATION ALGORITHMS:

Following section explain classification algorithms like Random Forest Decision tree Classifier, K-nearest neighbor.

Random Forest Decision tree Classifier:

The Random Forests algorithm [14] is able to classify large amounts of data with accuracy. Random Forests are a collection of learning method for classification and regression that construct a number of decision trees at training instant and outputting the class that is the form of the classes output by individual trees. Among all the decision tree classifiers Random Forest classifier is the best algorithm because:

It is one of the most accurate learning algorithms available. For many data sets, it gives highly accurate results. It runs efficiently on large databases. It can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification.

Random Forest uses the Information gain to construct decision trees.

Information Gain:

To compute the expected information gain by using the following formula:

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$\text{InfoA}(D) = \sum_{i=1}^{v} (|Dj|/|D|) * \text{Info}(Dj)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{InfoA}(D)$$

K-Nearest Neighbor Classifier:

"K-Nearest Neighbors classifier [15] is one of the simplest classifier that discovers the unknown data point using previously known data points and classified data. The k-nearest neighbors' algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority choose of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. K is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor.

K-nearest neighbor discovers the unknown data point by computing the distance between unknown data point and other data points by using Euclidean distance:

$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}$$

K-Means Clustering Algorithm:

According to Eulcidian distance formula, the distance calculated between only two data points. It cannot be possible for three data points. Then, by using k-means clustering algorithm for grouping the similar data points. Then applying the K-nearest neighbor algorithm for classifying the pesticides as low or high toxic.

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters [16].

CALCULATING LD50 VALUE OF A PESTICIDE:

The pesticide toxicity can be measured by using the factor LD50 or Lethal Dose50. The LD50 is the dose of a pesticide that will kill half of a group of test animals from a single exposure. The LD50 is expressed in milligrams per kilogram of body weight of the pest. The data mining classification algorithms classifies the pesticide toxicity based on LD50 value of the pesticide.

Calculated LD50=

Calculated oral or dermal LD50 =

$$\frac{100}{\frac{(\%\text{ chemical }\#1)}{(\text{LD50 chemical }\#1)} + \frac{(\%\text{ chemical }\#2)}{(\text{LD50 chemical }\#2)} + \frac{(\%\text{ chemical }\#3)}{(\text{LD50 chemical }\#3)}}$$

Calculated oral or dermal LD50 for a pesticide= $\frac{100}{\sum_{x=1}^{n} \frac{\%A_x}{T_{A_x}}}$

Where: %Ax = weight percent of each component in the Pesticide

Ax= the acute oral or dermal LD50 of the corresponding component

Σ = sum of all the math divisions for $\frac{\%A_x}{T_{A_x}}$

DATA FLOW DIAGRAM:

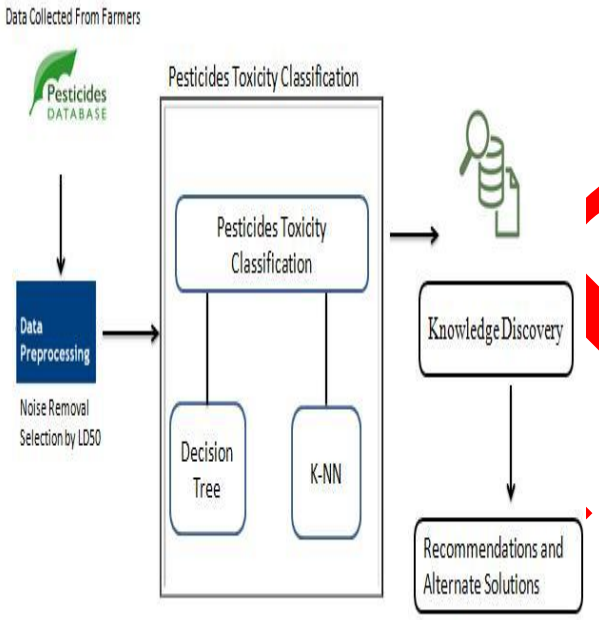

Figure 3: Overview for Pesticide Toxicity Classification

Taking pesticides data as input from farmers and upload data into database in a structure format. Extracting the required data from the database based on the selection of LD50 value. Applying various classification techniques decision tree [Random Forest] and k-nearest neighbour algorithm on the data extracted. The result will gives the information about the pesticide either toxic or non toxic and to indicate the less usage of high toxic pesticides and adopt alternative solutions to avoid human health hazardous.

EXPERIMENT AND RESULTS:

The data can be directly collected from the farmers. The Random forest, decision tree classifier and k-nn algorithms are implemented using Java programming.

	A	B	C	D	E	F	G	H	I	J
1	Crop	Disease	Season	Type	Pesticide name	Prefered dc	applied do	chemical composition	chemical in pesti	lue of chemical
2	paddy	blast	kharif	fungal	capton	2g	5g	1.)N-trichloromethyl-4-cyclohexane-1,	45	8400
3								2.)polyvinyl alcohol	5	20
4								3.)rhodamine B dye	2	300
5								4.)water	48	
6										
7	paddy	blast	rabi	fungal	thiram	2g	5g	1.)tebuconazole	0.6	1700
8								2.)tetramethylthiuram disulfide	20	560
9										
10	paddy	blast	kharif	fungal	carbendizm	1.5g	3g	1.)2-aminobenzimidazole	50	825
11								2.)5-methyl carbamate	50	13
12										
13	paddy	blast	rabi	fungal	edifenphos	500ml	1000ml	1.)o,o-dimethyl s-phenyl phosphrothio	0.3	8
14								2.)diphenyl sulphide	2.5	200
15								3.)toluene	1.5	636
16										
17	paddy	blast	kharif	fungal	beam	250g	500g	1.)tricyclozole 5-methyl-1,2,4-triazole	75	50
18								2.)sodium ligno sulphonate	3	5000
19								3.)sodium lauryl sulphate	3	1288
20								4.)silica	5	22500

Figure 4: Raw data set

After collecting the raw data from the farmers, preprocessing the data by selection of LD50. The required data contains attributes as crop name, pest name, season, type, pesticide name, preferred dose/acre, and applied dose/acre, LD50 value of a pesticide, predefined LD50, toxic level and category. Pesticide data set consists of total 500 instances. Toxic level is a class label which categorized as low, moderate, high. Java implementation for calculate LD50 value:

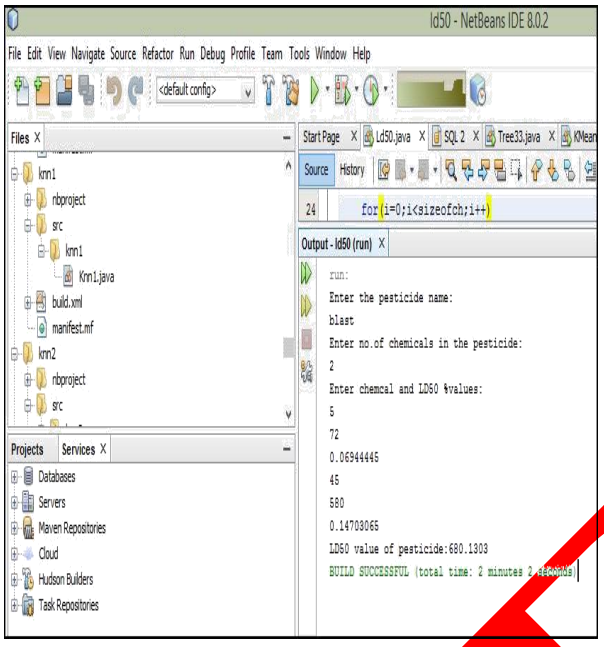

Figure 5: LD50 value of Pesticide

After calculating the LD50 value of a pesticide and take the necessary attributes for pesticide toxicity classification.

	A	B	C	D	E	F
1	Crop	Disease	Pesticide	LD50	toxic level	
2	paddy	blast	capton	381.64	moderate	
3	paddy	blast	thiram	2772.6	low	
4	paddy	blast	carbendiz	25.59	high	
5	paddy	blast	edifenphc	1909.9	low	
6	paddy	blast	beam	66.52	moderate	
7	paddy	blast	propicona	3629.18	low	
8	paddy	blast	tricyclozol	155.39	moderate	
9	paddy	brown spc	metominc	983.6	low	
10	paddy	brown spc	contof	2483.3	low	
11	paddy	brown spc	carbendiz	25.59	high	
12	paddy	brown spc	zineb	3000	low	
13	paddy	brown spc	mancozeb	8696.03	low	
14	paddy	bacterial l	plantomyc	1040.28	low	
15	paddy	bacterial l	steptocycl	474.26	moderate	
16	paddy	bacterial l	copper ox	931.83	low	
17	paddy	sheath bli	carbendiz	25.59	high	
18	paddy	sheath bli	propicona	3629.18	low	
19	paddy	sheath bli	hexacona:	2744.14	low	
20	paddy	sheath bli	validamyc	3333	low	
21	paddy	sheath rot	carbendiz	25.59	high	
22	paddy	false mut	thiram	2772.6	low	
23	paddy	false smut	carbendiz	25.59	high	
24	paddy	false smut	propicona	3629.18	low	
25	paddy	false smut	chlorothal	9259.25	low	
26	paddy	false smut	copper ox	931.83	low	

Figure 6: Required Data set

After completing the attribute selection the datamining classification like random forest, k-nearest neighbor are applied on the data.

Results obtained by Random forest classifier by using Java:

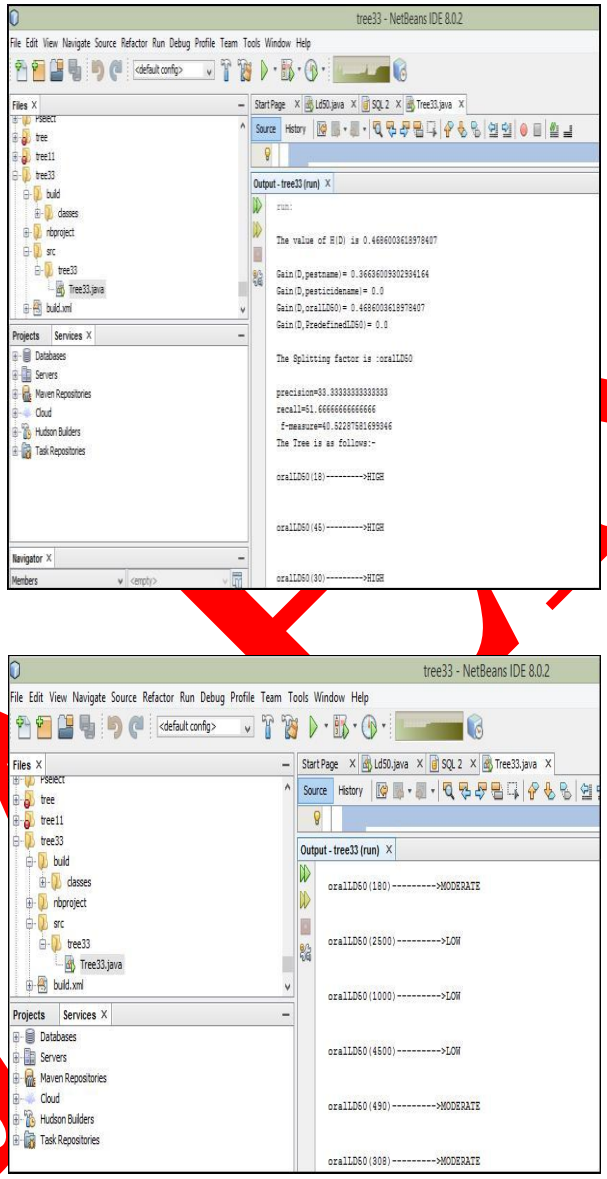

Figure7: Output for Random Forest classifier

Figure 7 shows the java implementation of Random Forest algorithm for classifying the pesticides as low, moderate and high.

Results obtained by K-means clustering algorithm by using Java:

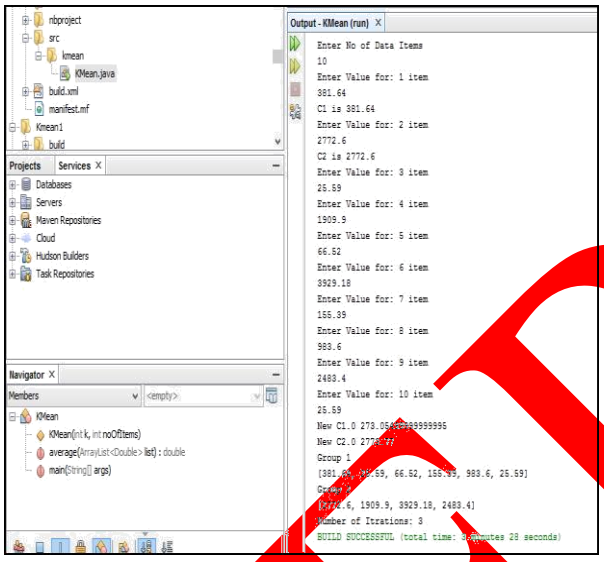

Figure 8: Output for K-Means Algorithm

Figure 8 show that the K-means clustering algorithm divides the data into two clusters. Now, applying the K-nearest neighbor classification algorithm on clustered data.

Results obtained by K-nearest neighbor classifier by using Java:

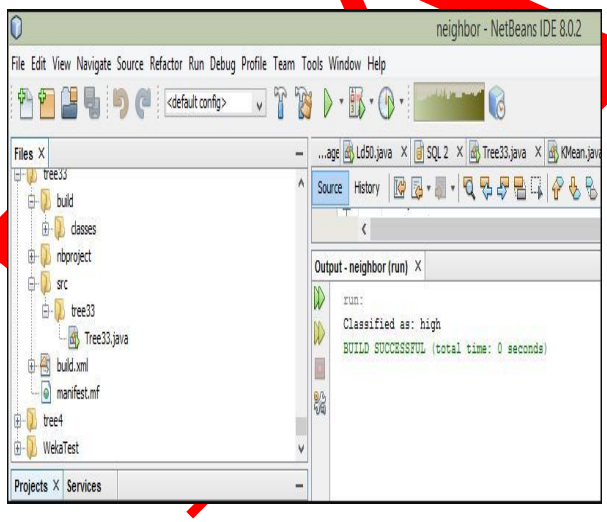

Figure 9: Output for K-nearest neighbor Results of Random forest algorithm are compared with other algorithms such as K-nearest neighbors. Information of all algorithms is summarized in following table.

Table 1: Comparative result of classifier

Evolution Criteria	RANDOM FOREST	K-NN
Accuracy	93.33%	86.67%
Precision	95.23%	91.66%
Recall	93.94%	87.88%
F-measure	94.58%	89.73%

The following graph shows the result:

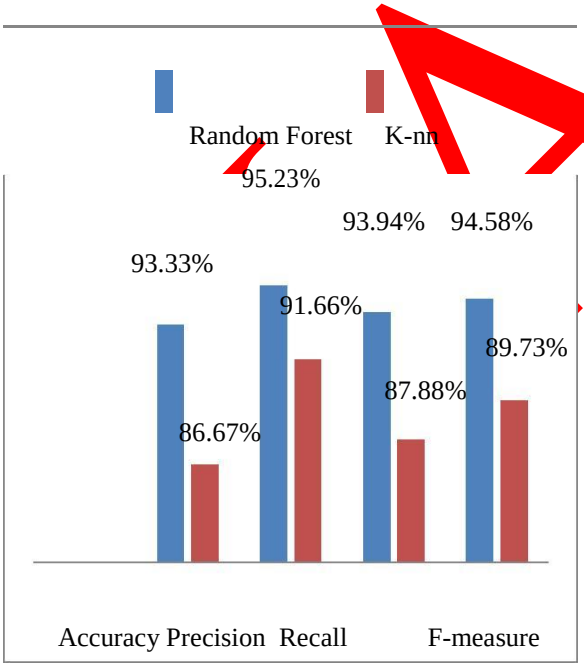

Figure 10: Graph for Classifiers Performance

From table we can say that Random forest algorithm gives better result among the others.

CONCLUSION AND FUTURE WORK:

The results of pesticide toxicity classification differentiates pesticides into high, moderate and low toxics applied on crops for good production and this classification helps to avoid highly toxic pesticides. The results are represented to farmers and suggested to use only low toxic pesticides and recommend considering herbal and bio- pesticides to minimize acute health problems in humans. Pesticide data exists in a complex form which is difficult to be understood by simple users. To facilitate the decision making process of farmers, the data is selected by removing noisy data.

For further work in addition to the farmer survey data, data set may include pesticides residues and input parameters from agriculture research centers to analyze the toxicities of pesticides more accurately.

REFERENCES:

[1] Abdullah A., Brobst S., Pervaiz I., Umer M., and Nisar A., "Learning Dynamics of Pesticide Abuse through Data Mining", Proceedings of Australasian Workshop on Data Mining and Web Intelligence, 2004.

[2] Pesticide Safety Education Program (PSEP):http://psep.cce.cornell.edu/Tutorials/c ore-tutorial/module04/index.aspx.

[3] [Ester M., Kriegel H.-P., and Sander J. 1997 "Spatial Data Mining: A Database Approach", Proc. 5th Int. Symp. on Large Spatial Databases, Berlin, Germany, pp. 47-66].

[4] [A.Padmapriya,N.Subitha,Clustering algorithm for Spatial Data Mining: An Overview,International Journal of Computer Applications (0975 – 8887) Volume 68– No.10, April 2013].

[5] [Rabiaimtiaz, Dr.Malik Sikandar Hayatkhiyal, Dr.Shahid Khalil, Aihabkhan, and Dr.Ahsan Abdullah,"Effect of Pesticides on Human Life using Visual Data Mining"].

[6] [Frank Lemke, Emilio Benfenati, John Adolf Mulle, "Data driven Modeling and Prediction of Acute Toxicity of Pesticide Residues"].

[7] [Jyothy S T, Deepu Kumar T L, Tata Elxsi L, Dr. Andhe Pallavi,"Comparison of Two Discriminant analysis algorithms in Pesticide Toxicity Classification"].

[8] **[**Emilio Benfenati, Paolo Mazzatorta1, Daniel Neagu, and Giuseppina Gini ,"Combining classifiers of pesticides toxicity through a neuro-fuzzy approach"].

[9] [Francisco Prieto Garcia, Sandra Y. Cortés Ascencio, John C. Gaytan Oyarzun, Alejandra Ceruelo Hernandez and Patricia Vazquez Alavarado, "Pesticides: classification, uses and toxicity"].

[10] [Elena Lo Piparo and Andrew Worth, "Review of QSAR Models and Software Tools for

anticipating Developmental and Reproductive Toxicity"].

[11] [M.G. Hill, P.G. Connolly P. Reutemann, D. Fletcher, "The use of data mining to assist crop protection decisions on kiwifruit in New Zealand"].

[12] [Ankur Omer, Poonam Singh, N. K. Yadav, R. K. Singh," An Overview of Data Mining Algorithms in Drug Induced Toxicity Prediction"].

[13] [Gongde Guo, Daniel Neagu , Xuming Huang , and Yaxin,"An examination concerning the blend of distinctive classifiers for danger forecast"].

[14] [Vrushali Bhuyar, "Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District"].

[15] http://www.saedsayad.com/k_nearest_neighbors.htm.

[16] http:// home.deib.polimi.it/matteucc/Clust ering/tutorial_html/kmeans.html.