International Journal of Engineering Research and General Science Volume 4, Issue 1, January-February, 2016
ISSN 2091-2730

Determining Bug Epidemic Patterns Using Decision Tree Algorithm

Roland A. Calderon
Southern Luzon State University
Lucena Campus
calderon_roland@yahoo.com

Dr. Maria Amelita E. Damian
Dr. Menchita F. Dumlao
Dr. Shaneth C. Ambat
AMA University

Dr. Geraldin B. *Dela Cruz*
Tarlac College of Agriculture

ABSTRACT- Data mining is the process of extracting useful and vital information from large sets of data that involves learning in a practical and non-theoretical sense (Witten et.al, 2011). Research topic involves the application of data mining techniques in agriculture field for predicting future trends such as bug epidemic. This acquires algorithms for describing hidden patterns frequently from large sets of data. Applications of appropriate data mining techniques are advised to apply to deliver productive decision making (Tripathy et.al, 2012). This application is known as Insect Epidemiology Data Mining (IEDM).

IEDM is an additional exploration of Discrete Mathematics and Theoretical Computer Science (DIMACS) that aims to provide an opportunity to develop and test problem instances and other methods of testing and comparing performance of algorithms (Abello and Cormode, 2004).

Researchers nowadays are able to offer number of information on agricultural-related activities which then can analyze the potential data mining and collect the relevant information. Data mining in agriculture requires historical data that helps to create automated prediction and analysis of various trends and behaviors to provide relative solutions.

However, data mining in agriculture can be used to improve the agricultural production such as rice crop and awareness in bug infestations. Discovering patterns in massive data sets will enable to predict possible outcomes by using this technique. This will be helpful for the agricultural farmers in forewarning about bug epidemic and enables to identify the factors that influence the pest population density. Farmer benefits will learn more about ecological and sustainable management options for pest epidemic and can apply pest control strategies to reduce crop loss. Technically, data mining in agriculture is a popular in a nontrivial process useful in classification applications since it resembles human reasoning and can be easily understood

Keywords— data mining, agriculture data mining, decision tree algorithm, bugs, epidemic, prediction

INTRODUCTION

The research concentrated on analyzing the prediction of bug's epidemic. Bugs epidemic is a widespread outbreak in the rice field in a severe form in the region. Every year, thirty seven percent (37%) is the estimated average of farmer's rice crops that suffer due to diseases and bugs species outbreak (Norton et al. 2010). Based from the study of Joshi et.al. (2007), invasive species such as Rice Black Bug(RBB) infest rice plants at all growth stages from maximum tilling to ripening stage.

According to Barrion et. al (2007), invasive bugs could not emigrated by flight or displaced by wind but they invade through insects swarming on boats that operate from island to island. Different bugs inhabit majority in different areas like rain fed and irrigated wetland environments, nearby woodlands, extensive weedy areas near rice fields, wild grasses near canals and staggered rice planting since they have constant food supply for the entire season.

Controlling of pests becomes difficult since the presence of host such as grasses and broadleaves are available regularly (PhilRice, 2000; Catindig and Heong, 2003). Application of insecticide should be avoided by the Agricultural institution because it affects the health and natural balance. Bureau of Plant Industry (BPI, 2015) stated that the practice of inorganic pesticide to RBB would also destroy its natural enemies. For this reason, farmers are advised to avoid spraying inorganic pesticides to control bugs.

Several fake chemicals emerge on the market which leads in development of insecticide resistance and pest resurrection. Authorities' outcome to limit this pest is disappointment and challenging them to further investigation (Sepe and Demayo, 2014).

The expansion of bug's habitat and weather are very important in determining the precise epidemiology of outbreak. This could give a certain pattern in bug's prediction. Predictive modeling could offer great beneficial to the farmers to enhance the bug prediction.

The main objective of this study is to explore the use of IEDM in addressing the infestation of bugs in the Philippine rice fields. The bug's damage covers several parts of the Philippines' result in 15–23% yield loss including Laguna, Cavite and Quezon. By abusing broad-spectrum, non-selective synthetic pesticides resulted to environmental degradation, immediate economic losses and damage the population of natural enemies of the pest.

This shows that agricultural farmers could probably have economic distress on the declined income from rice production. Hence, the occurrence of bugs in the rice field under study believed that plain cultivation of wetland habitats and host plants, and the lack of indigenous natural enemies are considered imbalanced.

This study includes the identification of bug's occurrence that will help predict bugs epidemic in the rice field. The data to be processes will include bug's classification, rice stages and ecology, climate such as temperature and humidity, lunar cycle and soil status. CRosss-Industry Standard Process for Data Mining (CRISP-DM) will also be used as a guide for this study.

STATEMENT OF THE PROBLEMS

Based on the study of Tripathy et.al (2012), there is an irregularity in the occurrence of pest in the rice field. The purpose of this study is to design and develop a device that will cater primarily on helping to monitor rice bug in order to predict the bug infestation.

Specifically, the study required responses and answers the following problems:

1. How effective the light trapping device in terms of bug counting?
2. How effective is decision tree algorithm in predicting bug outbreak?
3. What data model can be created in predicting bug epidemic?

Scope and Limitation of the Study

In connection with the relationship of agriculture and technology, "Rice Field Insect Light Trap (RFILT)" device will used on the study. This device will limit on the operation such as bug classifying manually, bug counting and rice field temperature using thermometer. It will be used in low-land and high-land rice field near Mt. Banahaw de Lucban. RFILT operates every night from 7:30pm in the evening up to11:00 midnights which will be placed in the rice field. The study shows that these sap-sucking bugs are strongly attracted to high intensity light. It is a prospective system appropriate for collecting the data on different parameters relating to temperature and environment.

In the technical part of the device, it has a minimal cost which contains Arduino 168 Microcontrollers that acts as a brain which is perfectly suited to agriculture where decisions are to be made at micro-climatic level; two (2) 12V power sources that consume low-energy; an infra-red sensor that is used to count and sense bug; and 20w-LED light. C++ codes were used to control the device since it runs on real-time operation. ISO 9126 software quality assessment tool was also used to measure the quality of software and hardware evaluated by ten (10) experts from different related departments in the Southern Luzon State University.

The three (3) flying bugs are the main target of this research since (1) Rice Black Bug is identified as invasive species; (2) Rice Bug considered as pest found in all rice environments and (3) Rice Grain Bug recognized as the latest insect pest of rice (Convention on Biological Diversity, 2001).

The Huey helicopter of the Philippine Air force will be used to observe the aerial view of different rice field areas in Quezon. Agricultural farmers and rice crop technicians will be interviewed to acquire some past experiences related to rice field bugs.

The study also focused on the historical data obtained from different municipalities and cities in CALABARZON from the year 2005 up to 2015. Historical data are delimited within ten (10) years. Some data were obtained from the Office of the Department of Agriculture-Bureau of Plant and Industry (BPI), Agriculture department in University of the Philippines-Los Baños Laguna, and Philippine Rice Research Institute (PhilRice). Most retrieved data are obtained from Internet historical data such as lunar cycle. Data sets are consisting of five hundred eighty seven (587) records.

The study will focus on Computer Science and Agriculture courses only since integration of technology are now applied in the new curriculum courses.

The research data will be evaluated using WEKA or SPSS to prevent mistakes on computation. Decision Tree will also be used to determine the rice field bug prediction and verify the accuracy rates. Data mining techniques that create models can be applied to other similar invasive pest.

The study will utilize CRISP–DM techniques which contain Business Understanding where the primary goal is to determine the business objectives; Data Understanding described as collecting, describing and verifying of data; Data Preparation where the

primary keys are selecting, cleaning, and constructing of data; Modeling where the study used Decision Tree to produce data model for the project; Evaluation where the results are evaluated and reviewed; and Deployment where the primary goals are planning and monitoring of the deployment of results.

The research output contains the Predictive Analytics of Bug's Epidemic, data models for the prediction of bug's epidemic and prototype device of insects.

Related Literature and Studies

Agriculture Data Mining

Nowadays, data mining in agriculture are new in research topic because of rapid growing amount of data available from different areas which can be used effectively. Agricultural data virtually are being harvested along with the crops and are being stored in databases. With the ever-increasing amount of information about the farms, farmers are not only harvesting in terms of agriculture output but also a large volume of data.

Today, agricultural sectors particular in rice production suffered greatly from political turmoil and agronomic problems from pest since 1970s (Joshi et.al 2007). Surveying the pest and evaluating the density of the pest population in fields as a way of monitoring rice pest population is very important agricultural decisions.

With the advent of Information Technology, Data mining in agriculture is relatively a different research field. They have used techniques to find relationships between ecology and insect to predict the target episodes and developed solving complex agricultural problems.

Epidemiology Data Mining

Epidemiology is defined as observational science that concerns itself with finding and explaining patterns of health and disease in populations, usually of humans, but also populations of animals, insects and plants (Abello etal. 2004). Data mining involves by the used of techniques to find patterns in data using automated or machine assisted. The goal is to identify fundamental problems that can help from efficient computation, statistical, mathematical models that can aid in the processing and understanding of combined epidemiological data. Limitation experienced in conventional methods using statistical techniques of data analysis where certain situations are commonly encountered.

Epidemic in Data mining is mainly focused on determining the precise epidemiology of bug's outbreak in the rice field. Discovering patterns in massive data sets are required in the bug issues which can help in determining the spread and infestation of bugs.

Decision Tree

A Decision Tree is a powerful tool for classification and prediction intended to facilitate decision making in sequential decision problems. It represents rules that can readily be expressed so that user can be understood or directly used in a database where records dropping into a particular category may easily be retrieved.

Decision Tree is a classifier in the figure of a tree structure where the root node serves as the topmost node. A node with outgoing edges is called an internal node which denotes a test on an attribute. Each branch signifies a result of the test and each leaf node holds a class label.

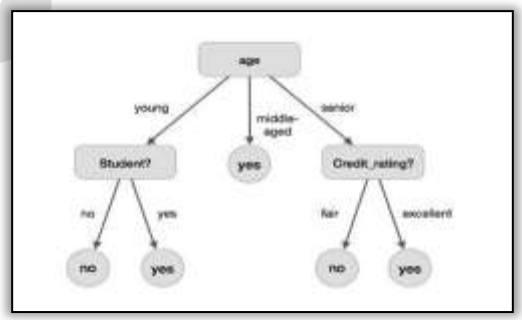

Figure 3. The concept of *buy_computer* whether a customer is likely to buy a computer or not

The Decision Tree can be easily converted to either three analyses: {1} Classification tree when the predicted outcome is the class to which the data belongs; {2} Regression tree when the predicted outcome can be considered a real number such as price of a house; and lastly {3}CART is an acronym from the words Classification And Regression Trees analysis used to refer to both of the two procedures (Moore, 2001).

Methods and Techniques Used

The researcher used Decision Tree Algorithm to predict future patterns. The created model combines and analyzes multiple sets of data to determine the pest infestation. The study used Classification Tree since it is designed for prediction problem where the dependent variable that needs to be predicted with greater accuracy based on several independent variables is nominal and ordinal. SPSS and WEKA are utilized to prevent mistakes and generate data model.

Instrument of the Study

The researcher used observation as a technique to gather data that involves watching and recording how the device functions and progress all throughout the given time. Bugs specimens are collected within a 3-month period at irregular intervals from different locations in Quezon Province. The collected specimens will be analyzed for manual identification and examination. Results seen on these materials will be an indicator to formulate the predictions of the study. (Barrio et.al, 2007)

The study also used the acquired basic data and information coming from agricultural farmers, rice crop technicians and related agricultural offices. It also concentrated on the historical data obtained from different municipalities and cities.

ISO/IEC 9126 software quality model was also used to evaluate the prototype device in order to meet the quality engineering requirements. The software criteria were based on the following characteristics: Efficiency, Functionality, Maintainability, Reliability, Portability, and Usability.

Figure 2. ISO/IEC 9126 Software Quality Model

External quality is used to measure the characteristics of the prototype. Functionality, Reliability, Usability are the main focused characteristics of the prototype (Azuma 2004), whereas the remaining characteristics are difficult to measure by the experts. The functionality requirements provide decision criteria that contribute in deciding the priority of each function when the software product

is used under specific condition. It focused on accuracy and interoperability. The reliability requirements focused on the recoverability and fault tolerance. Finally, the usability requirements focused on operability and learnability of the product (Kim et.al, 2014).

Table 1 shows the sub characteristics per criteria which are used by the researcher in developing indicators or questionnaires.

Table 1. ISO/IEC 9126 Sub Characteristics

The Four–point Likert Scale for the responses was used in rating the system. Arithmetic average was used to evaluate the overall performance of the device.

Table 2. Likert Scale

Range	Description
4	Strongly Agree (SA)
3	Agree (A)
2	Disagree (D)
1	Strongly Disagree (SD)

The respondents are composed of ten experts with a relatively high level of skill or knowledge in the study. They consist of seven (7) agriculturists and farmers which the device can be optimized its capacity and the rest are related in-line Professors.

Population and Sample of the Study

The records of the study are consisted of different occurrences of bug's infestation from different municipalities and cities from the 2005 up to 2015. The instance of data comprise of five hundred eighty seven (587) records. The instance of data was partitioned to training sets and test sets. Subsequently, it process using decision tree tool to generate data models which were used for bug prediction.

Development Methodology

CRISP-DM methodology was used in the development in creating the study. It helps the researcher to understand the data mining process and provide a road map while planning a data mining project. The study follows the six phases of a data mining process which indicate the researcher the importance of frequent dependencies between the phases. *(See figure 1)*

RESULTS AND DISCUSSIONS

Research Questions

Rice Field Insect Light Trap (RFILT) is one of the effective tools of management of the insect pests of the researcher as it mass traps both the sexes of insect pests and also substantially reduces the carryover pest population. This provides information related to insect distribution, abundance, flight patterns and helps in deciding the timing of the application of pesticides.

The researcher designed an ordinary light as an attractant and funnel to direct lured insects into a container. It is battery-operated and made up of different types of wood and metal products that are easily to modify. Counting of pests using sensor technology was also included. The aim of the device is to monitor rice pest population in particular areas suitable in early predicting pest outbreak.

In order to fully examine the device, the researcher selected ten (10) experts to validate the performance of the device based on the ISO/IEC 9126 standard. Six (6) of them came from the agricultural institution capable of conducting experiments and research on farms to improve production in crops while the rest are rice crop farmers who are responsible for preparing land for planting, caring for the crops and harvesting.

Expert's Assessment on the performance of the Rice Field Insect Light Trap (RFILT) Prototype

Efficiency was assessed using the following indicators: (1) The system responds quickly to the command; (2) The system uses appropriate program language; (3) The system performs according to specifications; and (4) The system provides efficient voltages to give signal to other components. Below is the experts' response for the acceptability of the software.

Usability	Experts										**MEAN**	Expert's Response Description
	A	B	C	D	E	F	G	H	I	J		
1. The user can use the system easily.	4	4	4	4	4	4	3	4	4	4	3.9	Strongly Agree
2. The device is presentable.	4	4	3	4	4	4	3	4	4	4	3.8	Strongly Agree
3. The system can easily be learned by the end – user.	4	4	3	4	3	4	4	3	4	4	3.7	Strongly Agree
4. The system remains to standard or regulation to usability.	4	4	3	4	3	4	3	3	4	4	3.6	Strongly Agree
	Gen. Weighted Mean										3.75	

Table 3. Weighted Mean and Description of the Usability of the Rice Field Insect Light Trap (RFILT) Prototype

Table 9 shows that the respondents graded the Usability of the Rice Field Insect Light Trap Prototype as "Strongly Agree" in terms of all indicators with a mean performance of 3.75.

System Criteria	**Weighted Mean**	**Expert's Response Description**
Efficiency	3.05	*Agree*
Functionality	3.23	*Agree*
Maintainability	3.08	*Agree*
Reliability	3.05	*Agree*
Usability	3.75	*Strongly Agree*
Overall Weighted Mean	**3.23**	**Agree**

Table 4 . Summary of the Weighted Mean of the Five (5) Criterias for Rice Field Insect Light Trap (RFILT) Prototype

Table 10 shows the summary results of the prototype based on ISO/IEC 9126 standard. Usability got the highest result 3.75, which means that the device was presentable and is easy to use. Functionality got second with the weighted mean of 3.23 showing that the device meets the expected output. Maintainability ranked as third with the score of 3.08 which reveals that provision for enhancement and faults can be easily diagnosed. Efficiency and Reliability got the same score of 3.05; this reveals that the device has the capability to provide desired performance relative to the amount of resources used under stated conditions and capability to maintain its level of performance under stated conditions for a stated period of time.

Overall, the Rice Field Insect Light Trap (RFILT) Prototype, based on the respondents' response, recorded an overall mean performance of 3.23 that has an interpretation of "Agree" and concluded that the device can be now used for operation.

In order to fully test the RFILT, the researcher conducted a five (5) day experiment to test the validity of the prototype. There were two methods to count the bugs. Automated counting using sensor-technology was administered to determine the bugs that entered into the funnel trapped into the container.

Performance Measure of the Algorithm

The confusion matrix is an effective evaluation tool for analyzing how well the classifier can recognize if the model is confusing two classes. A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is *n*-by-*n*, where *n* is the number of classes. The rows present the number of actual classifications in the test data. The columns present the number of predicted classifications made by the model.

The confusion matrix table illustrates a tabular display that evaluates the forecasting precision of a predictive model.

Table 5. Confusion Matrix Table

		Bugs Epidemic	
		Yes	**No**
Predicted	**Yes**	True Positive	False Positive
	No	False Negative	True negative

The table above shows how to maximize the correctness of classified in instances. For binary classification scenarios, the misclassification rate gives the overall model performance with respect to the exact number of categorizations in the training data.

To determine the accuracy level of the confusion matrix table the equation were used:

$$\boldsymbol{Accuracy\ (ACC) = \frac{TN + TP}{TP + FP + TN + FN}}$$

ACC/Accuracy signifies the amount of the total number of bug predictions that were correct. True Positive (TP) signifies the amount of actual outcomes of bug outbreak accurately classified as predicted bug's outbreak and True Negative (TN) refers to the number of bug infested accurately classified as predicted bug infested status.

To answer the last research question, Decision Tree was applied to create data model which includes all predictor variables. The rules set shown below predicts the probability of bug epidemic. A target value of 1 means rice field are in the stage of infestation; 2 means rice field are in the stage of outbreak condition.

Table 6 Rules for contains 5 Rules

CHAID Results (Note: 1=Infested, 2=Outbreak)

RULE SETS (CHAID Model)

IF (Lunar cycle != "First Quarter" AND Lunar cycle != "New Moon") AND

(Rice Phase = "Vegetative Stage") THEN

Node = 3

Prediction = 2 (Outbreak)

Probability = 1.000000

IF (Lunar cycle != "First Quarter" AND Lunar cycle != "New moon") AND (Rice phase != "Vegetative Stage" AND Rice Phase != "Reproduction Stage" AND Rice Phase != "Resting Stage") AND (Temperature <= "32 to 38")

THEN

Node = 6

Prediction = 2 (Outbreak)

Probability = 0.828571

IF (Lunar cycle != "First Quarter" AND Lunar cycle != "New moon") AND (Rice phase != "Vegetative Stage" AND Rice phase != "Reproduction Stage" AND Rice phase != "Resting Stage") AND (Temperature > "32 to 38 ")

THEN

Node = 7

Prediction = 2 (Outbreak)

Probability = 0.972603

IF (Lunar cycle != "First Quarter" AND Lunar cycle != "New moon") AND (Rice phase = "Reproduction Stage" OR Rice phase = "Resting Stage")

THEN

Node = 5

Prediction = 1 (Infested)

Probability = 0.527132

IF (Lunar cycle = "First Quarter" OR Lunar cycle= "New moon")

THEN

Node = 2

Prediction = (Infested)

Probability = 0.992424

Decision Tree nodes

In order to visualize the generated model, Decision tree nodes are used in the prediction of bugs to describe several possible paths representing deliberate actions or choices, followed by events with different chances of occurrence.

The nodes shown below are the diagram which predicts the probability of bug epidemic.

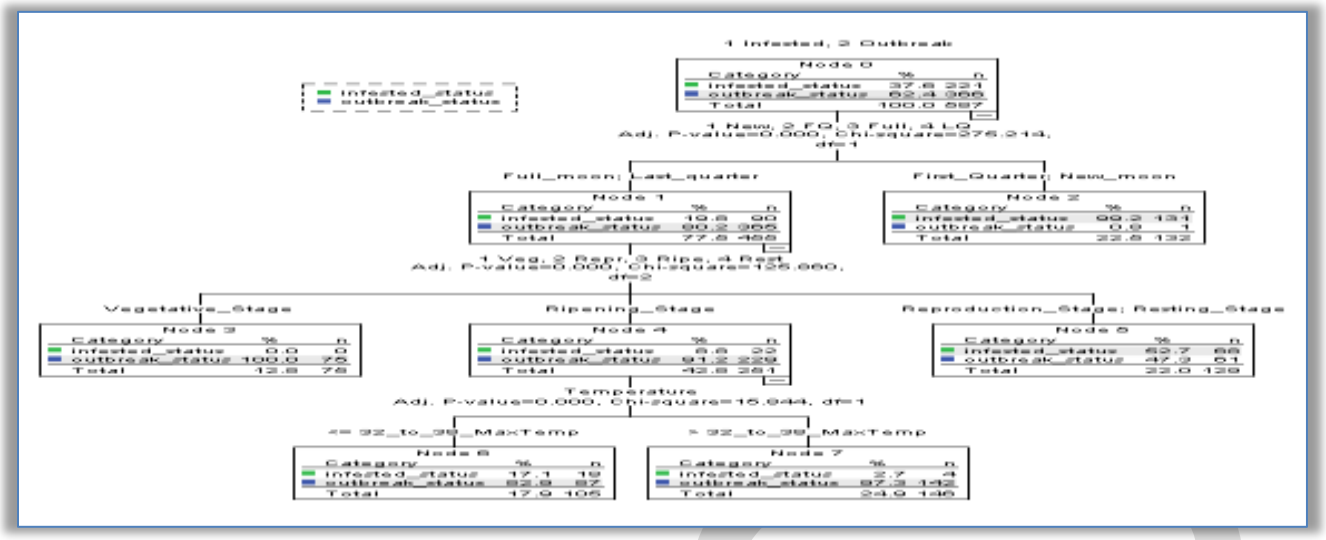

Table 7. Decision Tree diagram of Bug epidemic probability

Table 30 reveals the CHAID method which shows that the Lunar Cycle level is the best predictor of epidemic status. For the Lunar Cycle category, Lunar level is the significant predictor of Epidemic status followed by Vegetative level. In Vegetative stage level, 100% resulted in outbreak status. Since there are no child nodes below it, this is considered a terminal node.

For the Ripening stage, the next best predictor is temperature. Over 82% bugs occurred in the outbreak status if the temperature is lesser or equal to 32 to 38 temperatures while 97.3% if the temperature greater than to 32 temperatures. For Reproduction and Resting stage, 52.7% bugs occurred in the infested status and this is also considered a terminal node.

REFERENCES:

1. Abdullah, A., et al., (2004*), Learning Dynamics of Pesticide Abuse through Data Mining*, Australian Computer Society, Inc. Darlinghurst, Australia
2. Abello, J. & Cormode, G., (2004), *Data Mining and Epidemiology*, DIMACS Report, Center for Discrete Mathematics and Computer Science Rutgers University, Piscataway NJ
3. Abello, J. & Cormode G., (2006), *Report on DIMACS Tutorial on Data Mining and Epidemiology*, DIMACS Center, CoRE Building, Rutgers University
4. Al-Daajeh, S., (2009), *Balancing Dependability Quality Attributes Relationships for Increased Embedded Systems Dependability*, School of Engineering, Blekinge Institute of Technology SE – 372 25
5. Azuma, M., *Applying ISO/IEC 9126-1 Quality Model to Quality Requirements Engineering on Critical Software*, Department of Industrial and Management Systems Engineering, 2004
6. Barrion, A.T, Joshi R.C., and Sebastian, L.S., (2007*),"Rice Black Bugs, Taxonomy, Ecology, and Management of Invasive Species"*, Philippine Rice Research Institute
7. Batista, G. et al., (2011), SIGKDD *Demo: Sensors and Software to Allow Computational Entomology, an Emerging Application of Data Mining*, Bill and Melinda Gates Foundation
8. Frias, P.M., (2012, September 8), *Rice Grain Bug: New Insect Pest Eyed in Caraga*. Retrieved from http://ati.da.gov.ph
9. Hsu, William H., (2001), *Knowledge Discovery and Data Mining in Databases*, Machine Learning and Pattern Recognition
10. *Tropical Asian Irrigated Rice*, Annu. Rev. Entomol, Hanoi Vietname, 45:549–574
11. PhilRice (2015), *Diskarteng Wais Laban sa Peste*, Vol 2 Number 2, Philippine Rice
12. PhilRice (2006), Management of the Rice Black Bug- revised edition, Philippine Rice Research Institute, *Rice Technology Bulletin p2*
13. Ponweera P.A.D.M.D., Premaratne S.C.(2011), Enhancing Paddy Cultivation in Sri Lanka through a Decision Support System, *International Journal of Emerging Technology and Advanced Engineering* , Volume 1, Issue 2

14. Pratheepa, M, et al., (2011), A decision tree analysis for predicting the occurrence of the pest, Helicoverpa armigera and its natural enemies on cotton based on economic threshold level, *Current Science, vol. 100, no. 2, 25*
15. Ramamurthy V. etal. (2010), *Efficiency Of Different Light Sources In Light Traps In Monitoring Insect Diversity, Division of Entomology*, Indian Agricultural Research Institute, Munis Entomology & Zoology, p109-114
16. Ramesh, D., & Vardhan, B., (2013), Data Mining Techniques and Applications to Agricultural Yield Data, *International Journal of Advanced Research in Computer and Communication Engineering*
17. Zaïane, O. R., (1999), *Principles of Knowledge Discovery in Databases - Introduction to Data Mining*, Department of Computing Science, University of Alberta