ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2018, 19(3), pp. 732 - 747, DOI: 10.18038/aubtda.426946

AN ANALYSIS OF PESTICIDE USE FOR COTTON PRODUCTION THROUGH DATA MINING: THE CASE OF NAZİLLİ

Zehra BURDUR [1], Canan Eren ATAY [2, *]

[1] Department of Computer Engineering, The Graduate School of Natural and Applied Sciences, Dokuz Eylül University, İzmir, Türkiye
[2] Department of Computer Engineering, Faculty of Engineering, Dokuz Eylül University, İzmir, Türkiye

ABSTRACT

Data mining involves certain methods of obtaining or inferring meaningful and otherwise-unknown information from the data. These techniques and methods have been applied in the fields of health care, marketing, banking, and agriculture. With the increasing significance of precision agricultural practices, farmers have become inclined to be engaged in a more conscious strategy of agriculture. Farmers use pesticides to destroy a disease or other hazard on their plants. Nevertheless, it has been learned over time that pesticides have harmful effects on human health and the environment. Although many farmers are aware of the risks of excessive use of agricultural pesticides, they still use them to get a faster yield and to maximize financial gain or minimize financial loss. The cotton planted fields and used pesticides data received from Aydın Nazilli District Directorate of Agriculture was carefully organized and evaluated with the decision tree based classification algorithms in the SPSS Clementine program. C5.0 and Classification And Regression Tree (C&RT) decision tree algorithms that are the most preferred data mining methods are employed in this study. Thereby, it was observed that there exist some certain suggestive differences between the fertility obtained from the product and the pesticide used. Contrary to the conventional wisdom of most farmers, the excessive use of pesticides actually leads to a decrease in the yield obtained from the product, proportional to the measured dose.

Keywords: Data mining, Agriculture, Decision tree, C5.0 algorithm, C&RT algorithm

1. INTRODUCTION

Agriculture is an indispensable sector all over the world because it plays a major role in the survival of the country's population, the contribution to the national income and employment, raw materials and capital supply to other sectors, and direct and indirect influence on export. The agriculture sector has been out of the interest of the informatics sector although it has undertaken very important tasks in the economic and social development of the countries.

The contribution of agricultural production to any given country's economy is of great importance throughout the world. Programs designed to ensure that production is adequate and of high quality are intended to prevent the occurrence of major losses in time and labor power expended. A major problem in agriculture planning is the yield forecast or prediction. In the past, farmers applied traditional methods and relied on their own experience to estimate the yield they would likely obtain from their crops. Agricultural expansion and advancements during the past forty years have increased the quantity of food produced and concomitantly improved the quality of fresh food available worldwide [1]. Yet despite this ongoing overall improvement, inefficient agricultural practices persist in many parts of the world, largely because of a lack of modern tools and technologies in addition to lack of policies, laws, regulations and routine controls [2], [3].

A pesticide is any substance or mixture of substances intended to prevent, destroy, repel, or mitigate any pest (http://www.toxipedia.org. Accessed June 1, 2017). Over 1 billion pounds of pesticides are used in the United States (US) annually, and approximately 5.6 billion pounds are used worldwide [4]. Excessive use of pesticides is harmful in multiple ways. For example, the overuse of pesticides can

*Corresponding Author: canan@cs.deu.edu.tr

Received: 24.05.2018 Accepted: 29.07.2018

impose a financial burden on farmers. Also, excessive pesticide usage may lead to immunity in pests, which ultimately makes them more harmful to crops and less susceptible to mitigation. Worldwide, it is estimated that approximately 1.8 billion people engage in agriculture, most of them use pesticides to protect their produce (http://www.toxipedia.org. Accessed June 1, 2017).

Technological advancements in agriculture will lead to improvements in productivity. For this purpose, the raw data is transformed by researchers into useful information through data mining. Data mining is the process of discovering previously unknown and potentially valuable patterns in large datasets [5]. For predicition and classification, although data mining is used in many fields such as medicine, engineering, biology, finance, and heavy industry, its use in the field of agriculture has been and continues to be relatively limited. With variations — including climate conditions, soil types, effects of inputs such as fertilizers and pesticides, and a multiplicity of other such parameters — the forecast of crop yields has led to greater interest in the utility of data mining.

On the other hand, precision agriculture is a form of business in which the farmer correctly identifies a variable in potential usage of the land by using information technology that helps the farmer in the subdivision of the land, which is realized by the input application of the pertinent variables. Precision agriculture, by using data mining and information systems, aims to reduce input usage, avoid waste of resources, increase the yield of crops, and minimize the environmental pollution generated. Among the targets of precision agriculture are reduction in chemical costs (such as fertilizers and pesticides), reduction in environmental pollution, provision of high quality food products, establishment of records in agriculture, and ensuring a more efficient flow of information for management and aquaculture decision making.

Decision trees are data mining approaches that are frequently used in classification and estimation. Despite being capable of being used in classification of other methodologies like the nerve networks, the decision trees with their easy to make interpretations and ease of being understood provides advantage or decision makers [6]. Decision trees are among the most commonly used classification techniques because of their low cost, good reliability, easy to understand, interpret and could be integrated with data base. Decision trees are essentially rule extraction algorithms and are used when the rules for classifying data clusters can be easily understood by the user of the "if-then" type.

In this work, an attempt is made to show how the integration of agricultural data that includes yield and pesticide usage can be useful for the optimization of pesticide usage for cotton growing with data mining. The cotton planted fields and used pesticides data from Aydın Nazilli District Directorate of Agriculture was used and evaluated with the SPSS Clementine software; in these processes, C5.0 and C&RT decision tree algorithms were employed. The purpose of this research work is to determine 1- which harmful pests lower the crop yield more, 2- which agricultural chemicals are the most effective in combating these agricultural losses, 3- which chemical products reduce crop yield due to over-usage. The outcomes of the model should have many benefits for optimum agricultural pesticide usage management and farming practices in the future.

The remainder of this paper is as follows. Section 2 presents a brief review of related work. Section 3 gives a background of our work. Section 4 describes the structure and working of the chosen decision tree algorithm, through the C5.0 and C&RT applications. A discussion of our results and conclusions is presented in Section 5.

2. RELATED STUDIES OF DATA MINING TECHNIQUES IN AGRICULTURE

Impact of pesticides use in agriculture, their benefits and hazards explained in detail in the review article [7]. Pesticide usage is the most important issue in the agriculture. Abdullah A, Brobst, S, Pervaiz I, Umer M, Nisar A reported a negative correlation between pesticide usage and crop yield in Pakistan [8]. They have performed unsupervised clustering of data through “Recursive Noise Removal Heuristic” by Abdullah and Brobst in [9]. Through data intensive experiments, they have discovered many interesting patterns. Oakley, Zhang, and Miller pointed out that initial findings of ongoing research project to capture differences in pest management strategies and decision making among growers using the California Pesticide Use Reports (PUR) database [10]. They analysed the PUR for the best management practices to reduce pesticide use. Aktar, Sengupta, and Crowdhury pointed out the impact of pesticides [11]. Their aim is to determine the hazards and benefits of pesticide. They state primary benefits of pesticide including these titles; improving productivity, quality of food, protection of crop losses/yield reduction, transport, sport complex, building, vector disease control.

In agriculture, crop yield forecast is a very important problem. Raghuveer, Yogesh, and Shwetha stated that it is quite significant to predict crop yield in advance for market dynamics. They present some of the techniques such as k-means, the k-nearest neighbor, and decision tree in the field of agriculture [12]. Küçükönder, Vursavuş, and Üçkardeş determined the effect of mechanical properties such as strength of shell tearing point on tomato color, energy at shell tear point and hardness of shell rupture of tomato by the advisory learning algorithms of data mining method [13]. The data mining algorithms were applied to the K-Star, Random Forest and Decision Tree (C4.5) algorithms for the classification of mechanical properties and the error variance criterion, root mean square error (RMSE), mean absolute error. The mean absolute error (MAE), the relative relative squared error (RRSE) and the relative absolute error (RAE) are low and the classification accuracy is high. Ramesh and Vardhan pointed out that the problem of yield forecast is a major problem that remains unsolved [14]. Different data mining techniques are employed to evaluate the agriculture for estimating the future years' crop production. Their study presents a brief analysis of crop yield forecast using "Multiple Linear Regression (MLR)" technique and "Density-Based Clustering" technique for the selected region in East Godavari district of Andhra Pradesh in India.

Calıs, Kayapınar and Cetinyokus stated that decision trees are one of the data mining approaches widely used for classification and forecasting in [15]. In their study a survey was taken on computer and internet security and it was aimed to make an inference for people have different demographics characteristics by using decision trees. C5.0, C&RT, Chaid and Quest algorithms were employed and was found the accuracy rate of these algorithms. Then the decision tree is created with the algorithm giving the highest accuracy rate. As in our study, C5.0 algorithm was used when making rules with the decision tree because of its accuracy rate. Budak and Budak analyzed farm management and pest management practices on the basis of information gathered from 100 farmers by using a standardized questionnaire in the East Mediterranean region of Turkey in [16]. To the best of our knowledge, there isn't any research project that informs cotton growers as to the optimum amounts of pesticides to apply for the maximum yield rates for the growth of cotton in Turkey using data mining algorithms.

3. AGRICULTURAL PEST CONTROL

As the world population increases, arable lands are gradually diminishing mostly due to erosion and the opening of industrial facilities and roads. Countries are obliged to increase the amount of agricultural products by applying modern techniques and inputs. One of these techniques is called 'Agricultural Pest Control' and it is aimed to protect the plants from diseases, eliminate weeds, and increase overall agricultural production. Agricultural pest control is a mechanism that has some justifiable functionality, including the main goals to protect the plants, increase agricultural production, and improve the quality of agricultural work.

In some parts of Turkey, as well as in other countries, the nondirected usage of pesticides is excessive. Irrefutably, the misusage and overusage of pesticides has proven negative effects on human health, animals, and the environment. Admittedly, the degree of damage for each disease or pest must be identified on a case-by-case basis, and then the application of the proper type of active pesticide in the proper dose should be decided by considering the best methods of controlling agricultural pests and other factors. Well known pesticides can be categorized according to the types of pests they destroy:

Insecticides – insects, Herbicides – plants, Rodenticides - rodents (rats and mice), Bactericides – bacteria, and Fungicides – fungi.

Whereas pesticides are called by their trademarked names, the active substance that compose them may also be inferred by the chemicals included. In the market, there are different trademarks for a single active substance offered by different biochemical companies. The names of the pests and diseases of cotton, as well as the active substances of the pesticides applied, are shown in Table 1 below.

The reasons behind the use of pesticides can be summarized as: its low cost, the resultant higher economic income, and the increasing tendency to protect highly-valued products, such as fruits and vegetables. The final effects of pesticides vary depending on their formulation, the way of applying them, and the period of time of their usage. As stated in [17], the majority of farmers are aware of the long-term harmful effects of pesticides, but they generally lack the knowledge about the adverse effects of pesticides; thus they are unable to transform their awareness into practice.

In general, pesticides should be considered poisonous for humans, animals, and the environment in general, because all living beings in any ecosystem are directly or indirectly affected by these practices. The penetration of pesticides through the mouth, respiratory system, and the skin is thought to have a direct impact on the human body and its health. More precisely, pesticides can be absorbed by the soil indirectly for a variety of causes, such as rain, wind and they can directly reach any seeds laying dormant in that soil, leading to various harmful effects.

Table 1. Pest and active ingredients

Pest and Diseases	Active Ingredients	Pest and Diseases	Active Ingredients
Cochineal Insect	Hexythiazox	Seed Root Decay	Fludioxonil26 g/I+Metalaxyl-m 10 G/L
	Etofenprox 50g/I		Quintozene (PCNB) % 18
	Abamectan 18g/I		Pencycuron% 20+Captan % 50
	Spiromesifon 240g/I	Leaf Worm	Chlorpyrifos Ethyl % 25
Cotton Aphid	Acetamiprid %20		Emamectin-benzoate %5
	Pymetrozine %25		Flufenoxuron 50 g/l
	Flonicamed %50		Lufenuron 50g/l
Leaf Fleas	Acetamiprid %20	Green Worm	Indoxacarb %30
	Tau-Flovalinate 240g/I		Pyridalyl 500g/I
	Dimethoate 400g/I		Spinosad 480g/I
Tobacco Thrips	Imidacloprid+90 G/L Beta-Cyfluthrim210g/l		Flubendiamde%20
Weed	Pendimethalin450 G/L	White Fly	Buprofezin 400g/I
Green Worm	Indoxacarb %30		Bifenthrin 100g/I
	Pyridalyl 500g/I		Pyriproxyfen 100g/I
	Spinosad 480g/I	Cutworm	Chlorpyrifos Ethyl % 25
	Flubendiamde%20		Cypermethrin

On the other hand, in order to overcome the inevitable decline in pesticide efficiency against various types of targeted pests, a usage of higher doses may be required over time; therefore it may lead to both increased costs and reduced crop yields for the farmers employing them. In addition, certain factors such as a pesticide's chemical structure, its physical properties and type of formulation, its mode of application, and the climate and agricultural conditions also influence the performance of the pesticides in the respective environment.

4. DATA MINING ALGORITHMS FOR APPLICATION IN COTTON YIELDS

The goal of this paper is to create a model for cotton growers that will indicate the optimum amount of pesticides to apply for the maximum yield rate. The ultimate aim is the modeling decision tree based classification algorithms on the data and the observation of the results. This paper intends to design and verify a classification model through the usage of decision trees. This section explains data description, data pre-processing, and the classification model creation using SPSS Clementine software decision trees with C5.0 and C&RT algorithms.

4.1 Data Description and Data Pre-processing

For our study, we have selected the province of Aydın, Nazilli primarily because it is the second largest producer of the cotton crop in Turkey (www.tie.gov.tr. Accessed June 1, 2017). Aydın, Nazilli is located in the Aegean section of Turkey, as depicted in Figure 1. The district, shown in Figure 2, is in the geographic coordinates 28 degrees 19 minutes east of the Greenwich meridian and 37 degrees 54 minutes north latitude.

Original data was obtained from the Aydın, Nazilli District Directorate of Agriculture. In this first step we performed the necessary data cleaning, standardization, and correlation. Data cleaning techniques allow us to fill in missing values, smooth noisy data, identify outliers, and correct inconsistencies in the data. The original data matrix included 49.468 observations. Related data includes the records for the production of cotton crop in parcels throughout the Nazilli district. After choosing cotton related data, 1284 observations (objects and data) described by 14 attributes (variables) were achieved.

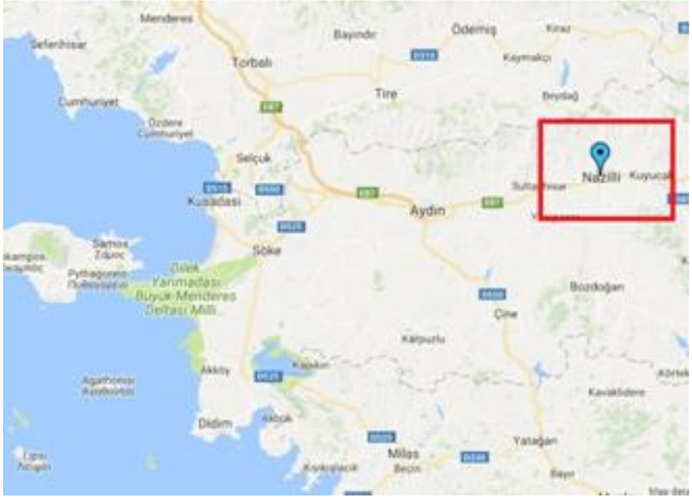


Figure 1. Map of Aegean section of Turkey

Figure 2. District of Nazilli

After cleaning the raw data, we created 6 tables in a cotton database named Agriculture with using SQL Managament Studio 2012. The Production table contains information on what is planted on each field, when and what pesticies have been sprayed, when the harvest was performed, and what was the yield rate. Since other tables are self explanatory, we have not detailed them here for the sake of space.

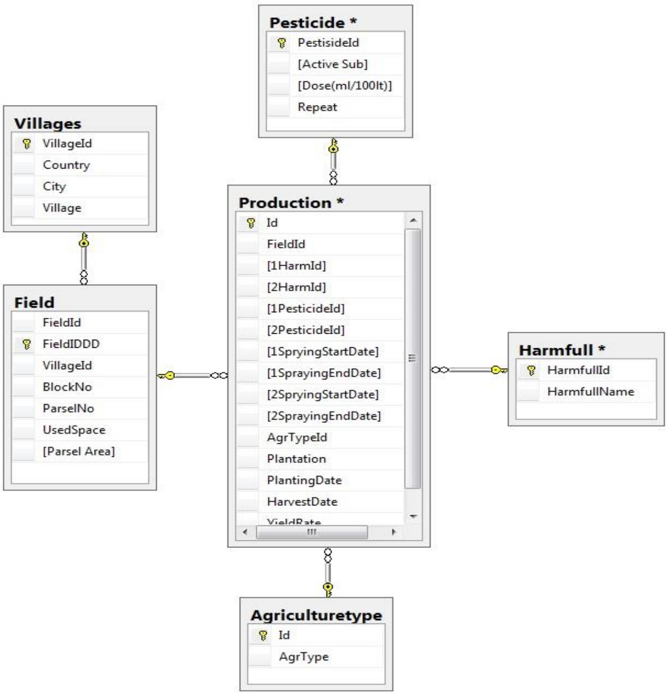
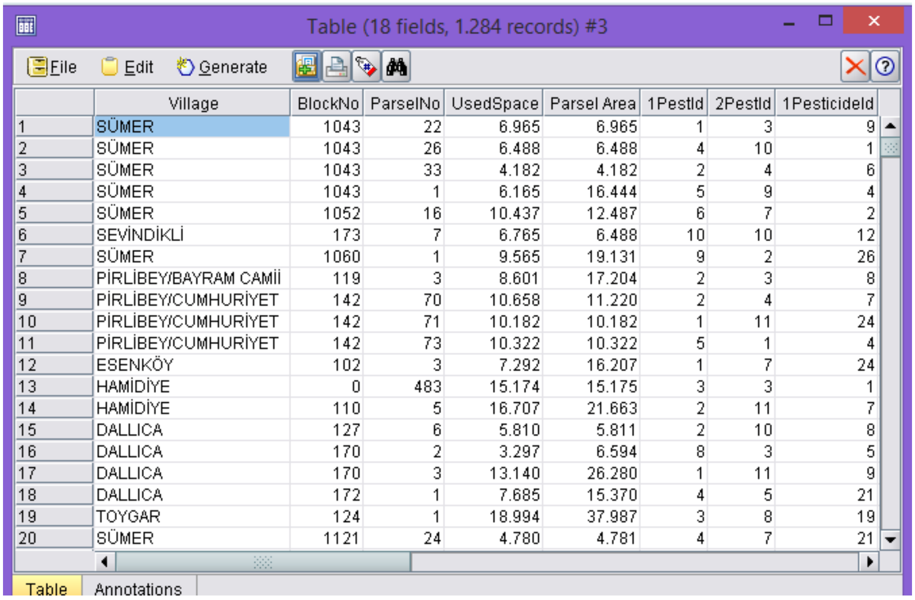


Figure 3. Cotton database diagram

The data includes parcel production certificate throughout the Nazilli district records of cotton crop. The records have totally 1284 field with the following data labels: village name, block no, parcel no, used space, 1. Pestid, 2. Pestid, 1. Pesticideid, 2. Pesticideid, 1. Pesticide dose, 2. Pesticide dose, 1. Spraying start date, 1. Pesticide end date, 2. Spraying start date, 2. Pesticide end date, planting date, harvest date, yield rate. This data set divided into 3 parts; 60% training set, 20 % validation set, and 20 % test set. While an instance of the cotton crop record is illustrated in Figure 4, distrubition of yield rate and doses are depicted in Figure 5.



	Village	BlockNo	ParselNo	UsedSpace	Parsel Area	1PestId	2PestId	1PesticideId
1	SÜMER	1043	22	6.965	6.965	1	3	9
2	SÜMER	1043	26	6.488	6.488	4	10	1
3	SÜMER	1043	33	4.182	4.182	2	4	6
4	SÜMER	1043	1	6.165	16.444	5	9	4
5	SÜMER	1052	16	10.437	12.487	6	7	2
6	SEVİNDİKLİ	173	7	6.765	6.488	10	10	12
7	SÜMER	1060	1	9.565	19.131	9	2	26
8	PİRLİBEY/BAYRAM CAMİİ	119	3	8.601	17.204	2	3	8
9	PİRLİBEY/CUMHURİYET	142	70	10.658	11.220	2	4	7
10	PİRLİBEY/CUMHURİYET	142	71	10.182	10.182	1	11	24
11	PİRLİBEY/CUMHURİYET	142	73	10.322	10.322	5	1	4
12	ESENKÖY	102	3	7.292	16.207	1	7	24
13	HAMİDİYE	0	483	15.174	15.175	3	3	1
14	HAMİDİYE	110	5	16.707	21.663	2	11	7
15	DALLICA	127	6	5.810	5.811	2	10	8
16	DALLICA	170	2	3.297	6.594	8	3	5
17	DALLICA	170	3	13.140	26.280	1	11	9
18	DALLICA	172	1	7.685	15.370	4	5	21
19	TOYGAR	124	1	18.994	37.987	3	8	19
20	SÜMER	1121	24	4.780	4.781	4	7	21

Figure 4. An instance of the cotton crop records

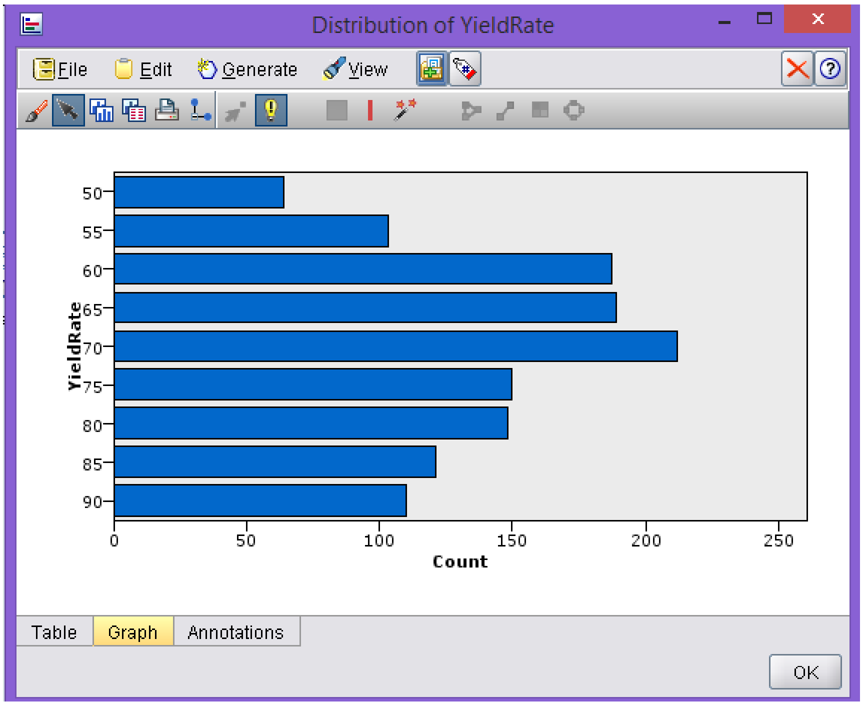


Figure 5. Distrubition of yield rate and doses for cotton

4.2 Decision Tree Algorithms

In general, a decision tree is an analytical tool that makes it possible to develop a variety of classification systems that can predict or classify future observations based on a set of predetermined decision rules. Consequently, a decision tree is a description of the splits found by the algorithm. It consists of a root, with any number of nodes, branches, and leaves (also known as terminals). Algorithms such as C4.5, C&RT, ID3, and QUEST, are used for the building of decision trees.

We have chosen to employ a decision tree based classification algorithm in conjunction with regression analyses because they make possible the prediction of the rate of yield and the pesticide dose that is used. The reason for using the regression analysis is to determine the relation between two or more variables having a cause-and-effect relationship, and to make an estimation or a prediction about the topic by using this relation. In many cases, given sufficient data, finding a cause-effect relationship in nature is quite likely. In this research, a regression model was used to explain the relationship between the variables of the yield rate and the desired doses for pesticide. Therefore, decision trees were built, and the C5.0 and C&RT algorithms were used. Concerning the use of the model of feature selection, ‘dose1’, ‘dose2’, and ‘blockNo’ appear to be important fields considering their values, as shown in Table 2.

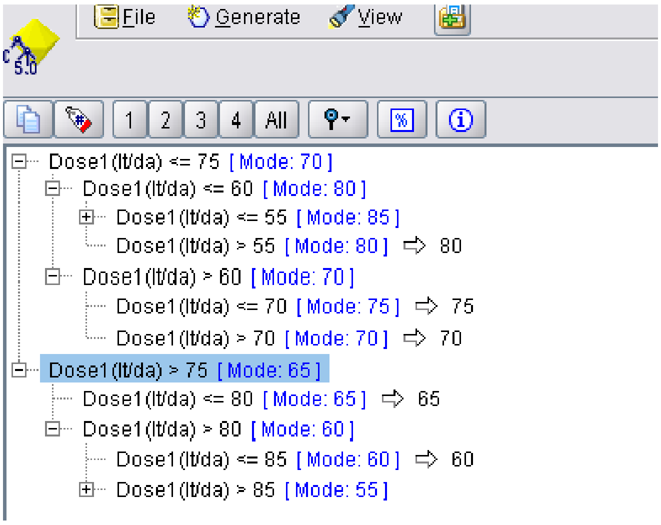
Table 2. Feature selection of yield rate

Field	Type	Importance	Value
Dose1(lt/da)	Range	Important	1,0
Dose2(lt/da)	Range	Important	1,0
Block No	Range	Important	0,961
Parcel Area	Range	Unimportant	0,79
Parcel No	Range	Unimportant	0,443
Used Space	Range	Unimportant	0,436
2SprayingStartDate	Set	Unimportant	0,377
2SprayingEndDate	Set	Unimportant	0,31
2PestId	Set	Unimportant	0,271
2PesticideId	Set	Unimportant	0,219
1PesticideId	Set	Unimportant	0,155
1SprayingStartDate	Set	Unimportant	0,081
1PestId	Set	Unimportant	0,077
HarvestDate	Set	Unimportant	0,044
1SprayingEndDate	Set	Unimportant	0,001
PlantingDate	Set	Unimportant	0,005
Village	Set	Unimportant	0,0

4.2.1 C5.0 Algorithm

A C5.0 method works by splitting the sample based on the attribute that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field; the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed.

When the C5.0 algorithm is applied and the target is designated as the yield rate, and after the agricultural pesticide is selected as the input, then the algorithm generates some rules and reveals the results. To interpret this process, it should be noted that the algorithm reveals two main rules, displayed below in Figure 6.



```
Dose1 (lt/da) <= 75 [ Mode: 70 ]
    Dose1 (lt/da) <= 60 [ Mode: 80 ]
        Dose1 (lt/da) <= 55 [ Mode: 85 ]
        Dose1 (lt/da) > 55 [ Mode: 80 ] ⇨ 80
    Dose1 (lt/da) > 60 [ Mode: 70 ]
        Dose1 (lt/da) <= 70 [ Mode: 75 ] ⇨ 75
        Dose1 (lt/da) > 70 [ Mode: 70 ] ⇨ 70
Dose1 (lt/da) > 75 [ Mode: 65 ]
    Dose1 (lt/da) <= 80 [ Mode: 65 ] ⇨ 65
    Dose1 (lt/da) > 80 [ Mode: 60 ]
        Dose1 (lt/da) <= 85 [ Mode: 60 ] ⇨ 60
        Dose1 (lt/da) > 85 [ Mode: 55 ]
```

Figure 6. The rules for pesticide dose by using C5.0 algorithm

The first rule is that the dose of the pesticide is equal to or less than 75, and the second rule ascertains that is more than 75. Assuming that the dosage of pesticide is equal to or less than 75, then the yield rate obtained turns out to be 70%. However, when the dose is equal to or less than 60, then the yield rate rises to 80%. On the other hand, if the dose is equal to or less than 55, then it can be seen that the yield rate reaches 85%. Yet, in a situation when the dose is determined to be more than 55, then the yield rate obtained decreases to 80%. In other respects, when the dosage of the pesticide is more than 60, then the yield rate turns out to be 70%. Under the condition that the dosage of the pesticide is equal to or less than 70, then a yield rate of 75% is obtained. Lastly, when the dose is more than 70, then the yield rate becomes 70%.

The second rule is the case of the pesticide dose being more than 75. When the dose is more than 75, then the yield rate appears to be 65%; but when the dose is equal to or more than 80, then the yield rate is 65%. However, if the dose is increased and becomes more than 80, then a rate of 60% in yield is achieved. In another case, in the circumstance that the dosage of pesticide is equal to or more than 85, then the yield rate again ends up at 60%. Finally, when the dose is more than 85, then the yield rate drops and settles at 55%. To sum up, it has been observed in the analysis of this data that, as a result of the excessive use of pesticides by farmers against various harmful pests, the yield obtained from the crops is actually reduced —the opposite of their presumed goal. For a deeper understanding, the algorithm is also depicted as a tree structure, as seen in Figure 7.

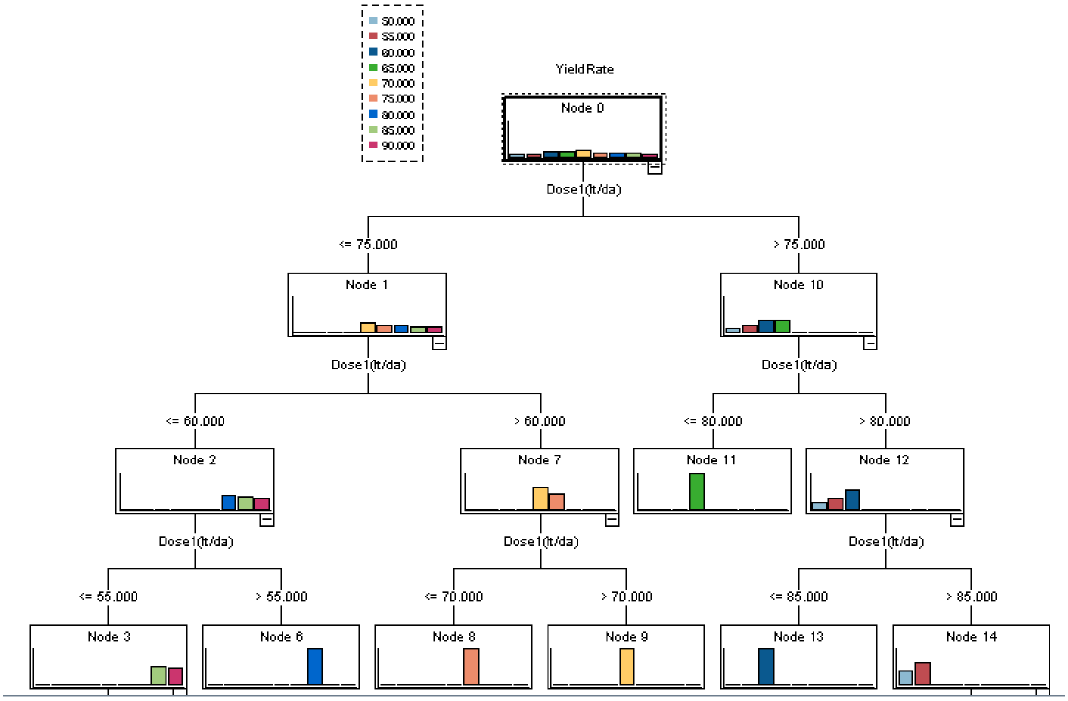
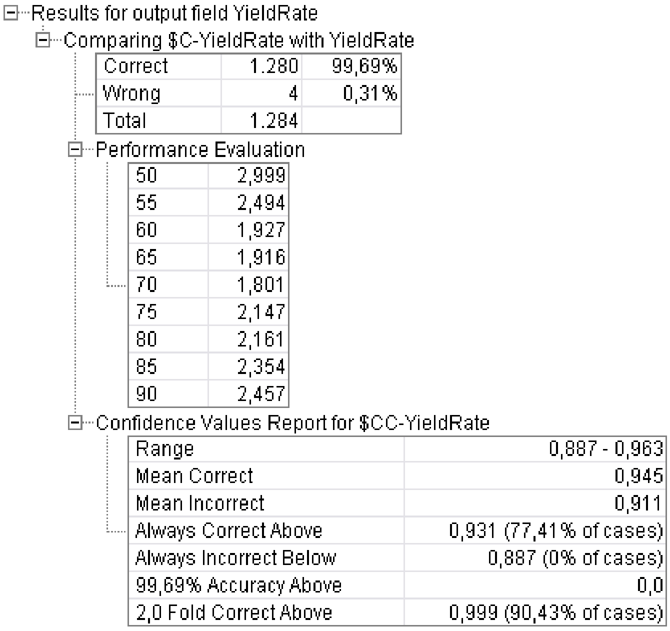


Figure 7. Structure of decision tree of the C5.0 algorithm

In terms of the accuracy of our chosen methodology, it is seen in Figure 8 that this algorithm has a high accuracy rate, namely, that of 99.69%. Performance evaluation and confidence values reports also shown in Figure 8. Some statistics was evaluated for C5.0 algorithm, as seen in Figure 9.



Results for output field YieldRate

Comparing $C-YieldRate with YieldRate

Correct	1.280	99,69%
Wrong	4	0,31%
Total	1.284	

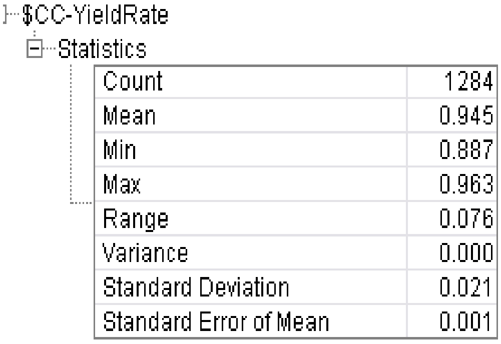
Performance Evaluation

50	2,999
55	2,494
60	1,927
65	1,916
70	1,801
75	2,147
80	2,161
85	2,354
90	2,457

Confidence Values Report for $CC-YieldRate

Range	0,887 - 0,963
Mean Correct	0,945
Mean Incorrect	0,911
Always Correct Above	0,931 (77,41% of cases)
Always Incorrect Below	0,887 (0% of cases)
99,69% Accuracy Above	0,0
2,0 Fold Correct Above	0,999 (90,43% of cases)

Figure 8. Analysis for the C5.0 algorithm



$CC-YieldRate

Statistics

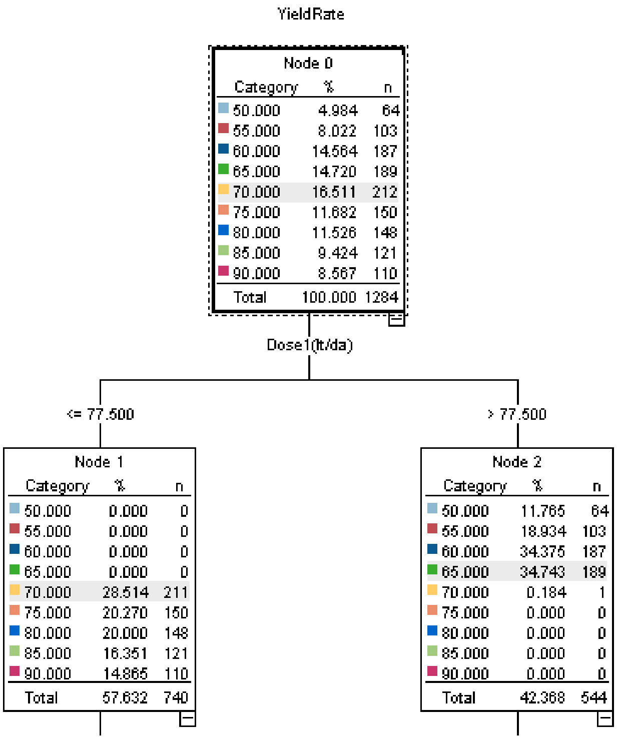
Count	1284
Mean	0.945
Min	0.887
Max	0.963
Range	0.076
Variance	0.000
Standard Deviation	0.021
Standard Error of Mean	0.001

Figure 9. Statistics of C5.0 algorithm

4.2.2 C&RT Algorithm

The C&RT tree is constructed by splitting subsets of the data set using all the descriptors as predictors, in order to create two child nodes for each repeated split, beginning with the whole data set. In this case, the Gini impurity index is employed to choose the best predictor; this method works by choosing a split at each node such that each child node is more pure than its parent node. The goal of this strategy is to produce subsets of the data which are as homogeneous as possible with respect to the class label. For each split, the respective predictive descriptor is evaluated to find the optimal cut point based on an improvement in the score or a reduction in the calculated impurity. Subsequently, the predictors are compared with one another, and the predictor with the best improvement is selected for the subsequent split. This process is repeated recursively until its stopping rule is triggered.

The first value of the dose of pesticide started at 77.500in Figure 10. It turns out that if the dosage of the pesticide is less than 77.500, then the result is 57.632 in total. The yield of this pesticide is 70% at 211, 75% at 150, 80% at 148, 85% at 121, and 90% at 110. It should be noted that pesticide has the highest yield of 70% in the products. The tree continues to branch from the value of 72.5 or less. If the dose of pesticide is more than 77.500, then the total result is 42.368. It can be observed that pesticide has the highest value in the yield with a rate of 65%. The yield rate is 50% at 64, 55% at 103, 60% at 187, and 65% at 189.



Node 0

Category	%	n
50.000	4.984	64
55.000	8.022	103
60.000	14.564	187
65.000	14.720	189
70.000	16.511	212
75.000	11.682	150
80.000	11.526	148
85.000	9.424	121
90.000	8.567	110
Total	100.000	1284

Node 1

Category	%	n
50.000	0.000	0
55.000	0.000	0
60.000	0.000	0
65.000	0.000	0
70.000	28.514	211
75.000	20.270	150
80.000	20.000	148
85.000	16.351	121
90.000	14.865	110
Total	57.632	740

Node 2

Category	%	n
50.000	11.765	64
55.000	18.934	103
60.000	34.375	187
65.000	34.743	189
70.000	0.184	1
75.000	0.000	0
80.000	0.000	0
85.000	0.000	0
90.000	0.000	0
Total	42.368	544

Figure 10. The first branch in the decision tree for the variable of pesticide doses

For the second level branching, when the dosage of the pesticide is less than 72.500, it is found that the yield has the most share with 75%, as depicted in Figure 11. The tree continues branching when the yield rate is less than or equal to 72.500. The other part still proceeds, branching on the condition that the value is more than 82.500.

When the dosage of the pesticide is equal to or less than 65.000, then the total result is 29.595. The yield rate has the most shares with 80%. The yield rate is 85% at 121, and it is 90% at 110. The other part of the tree carries on branching with the value of more than 87.500 in Figure 12. It has a total value of 13.084. The yield rate is 55% at 103, 50% at 64, and 70% at 1.

The fourth branching continues with a value of 57.500. When the dosage of the pesticide is less than or equal to 57.500, then it generates a total value of 18.069. It can be seen that the yield rate is 85% at 121 and 90% at 110. In the other part of the tree, when the pesticide dose is less than or equal to 95.000, then the total resultant value is 8.255. As is seen in the Figure 13, the yield rate is 55% at 103 and 50% at 2.

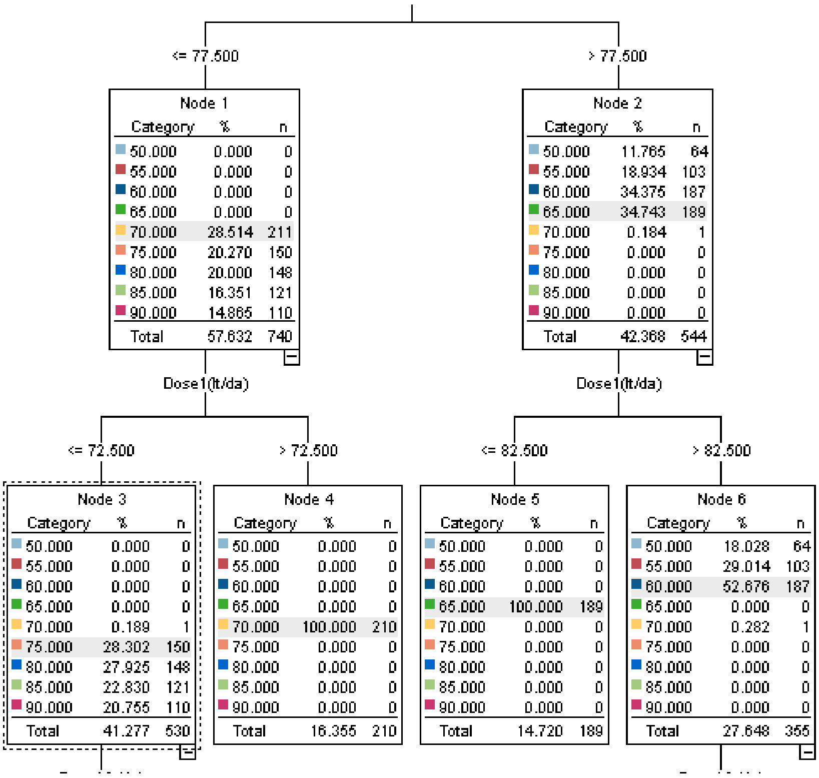


Figure 11. The second branch in the decision tree for the variable of pesticide doses

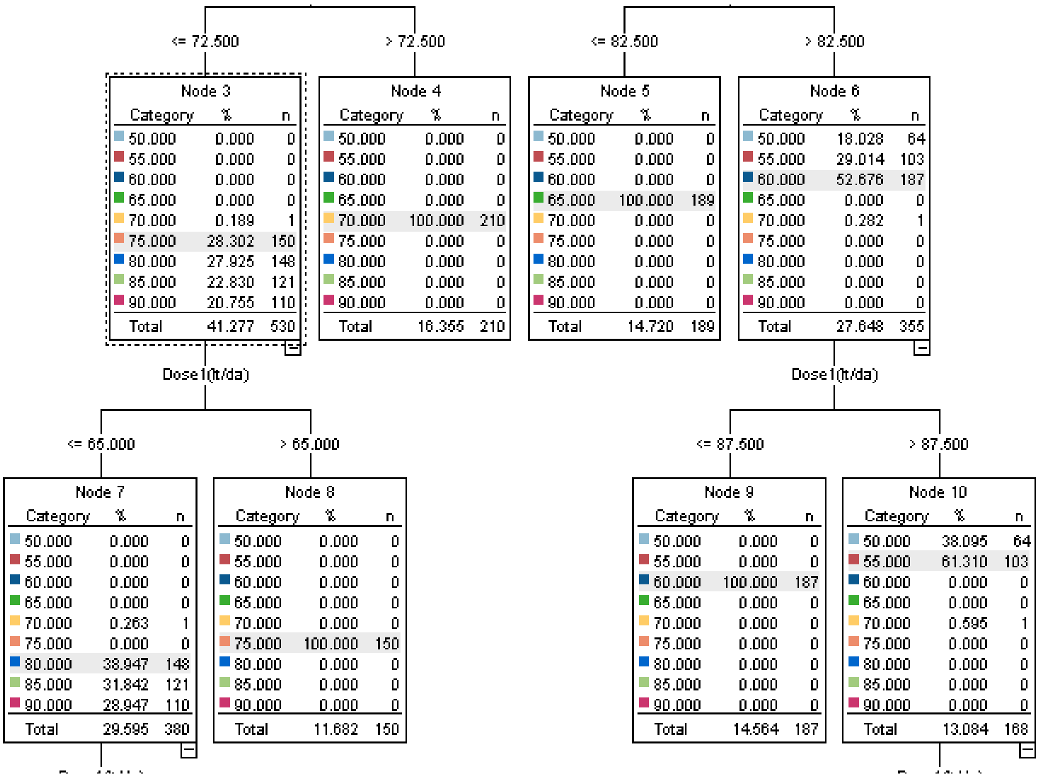


Figure 12. The third in C&RT algorithm for the variable of pesticide doses

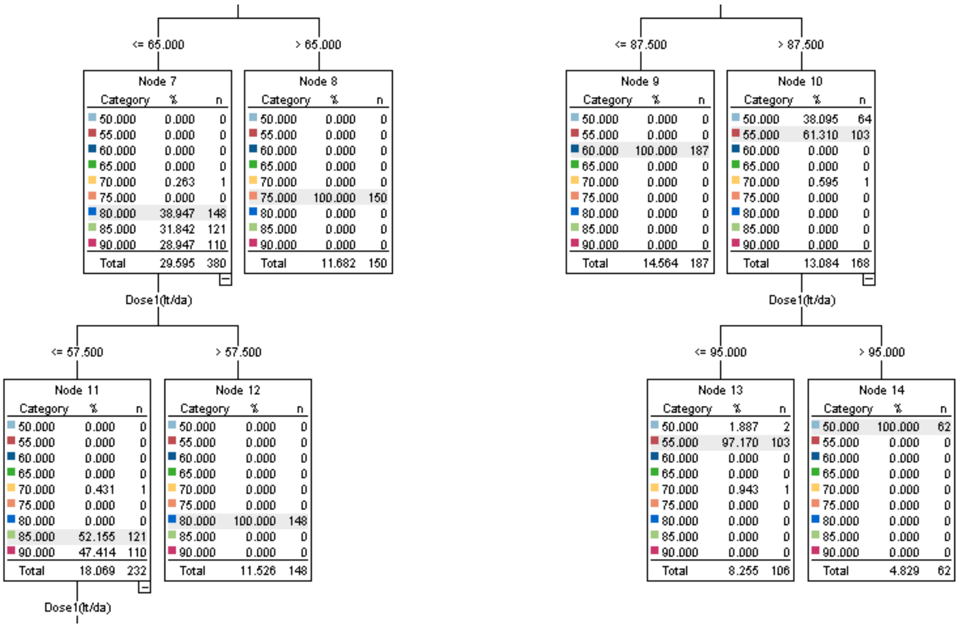


Figure 13. The fourth branch in the decision tree for the variable of pesticide doses

The tree continues branching at the value of 52.500 in Figure 14. When the dosage of the pesticide is more than 52.500, then the total result is 9.502. The yield rate is 85% at 121 and 70% at 1. If the dosage of the pesticide is less than or equal to 52.500, then the total value reaches 8.567. Moreover, the yield rate is 90% at 110, which is indicated in the last tree class.

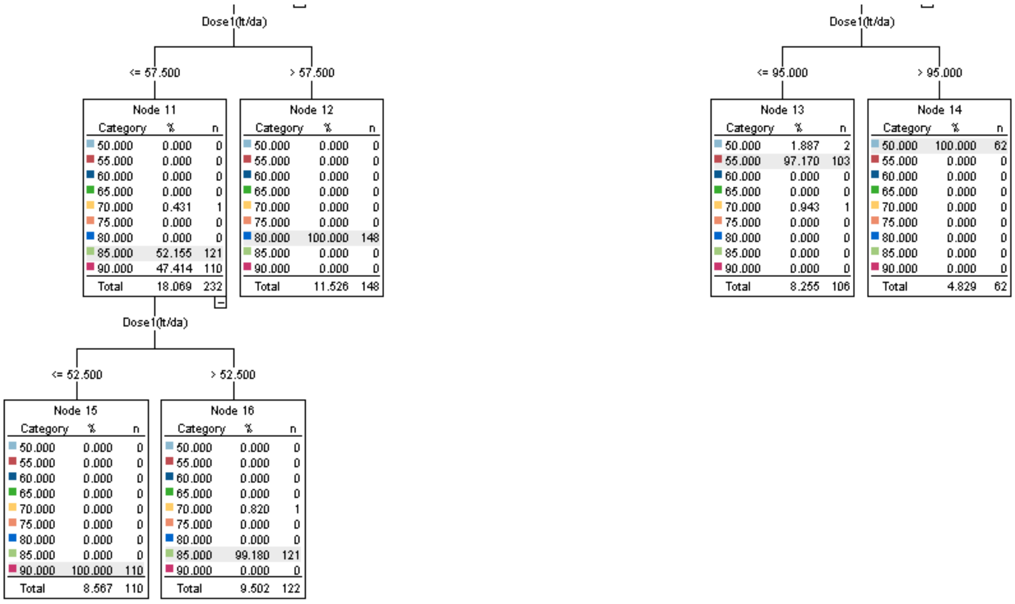
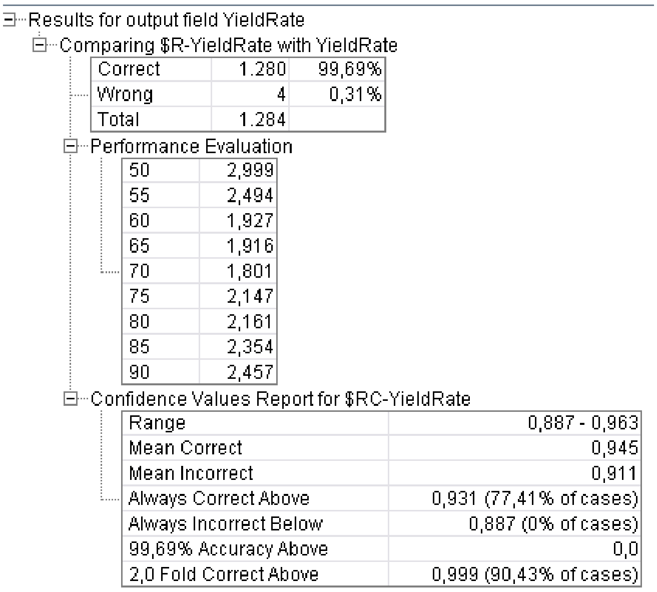


Figure 14. The last branch in the decision tree for the variable of pesticide doses

Finally, and most importantly, the yield rate is reduced due to the high doses of pesticide. In the last branch of the tree, the value for the dose used for harmful pests is calculated to be 52.500, and in the cases in which the value is less than this, then it is found to have the highest rate with 90%.

In terms of the accuracy of our chosen methodology, it is seen in Figure 14 that this algorithm has a high accuracy rate, namely, that of 99.69%. Performance evaluation and confidence values reports also shown in Figure 15. Some statistics was evaluated for C&RT algorithm as seen in Figure 16.



Results for output field YieldRate

Comparing $R-YieldRate with YieldRate

Correct	1.280	99,69%
Wrong	4	0,31%
Total	1.284	

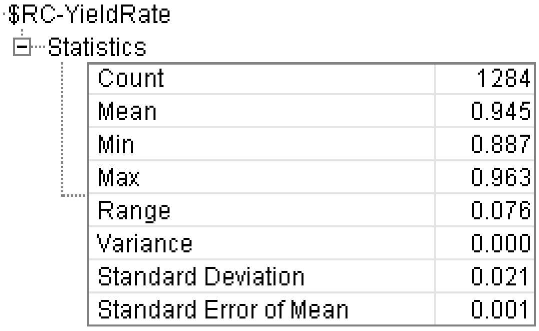
Performance Evaluation

50	2,999
55	2,494
60	1,927
65	1,916
70	1,801
75	2,147
80	2,161
85	2,354
90	2,457

Confidence Values Report for $RC-YieldRate

Range	0,887 - 0,963
Mean Correct	0,945
Mean Incorrect	0,911
Always Correct Above	0,931 (77,41% of cases)
Always Incorrect Below	0,887 (0% of cases)
99,69% Accuracy Above	0,0
2,0 Fold Correct Above	0,999 (90,43% of cases)

Figure 15. Analysis for the C&RT algorithm



$RC-YieldRate

Statistics

Count	1284
Mean	0.945
Min	0.887
Max	0.963
Range	0.076
Variance	0.000
Standard Deviation	0.021
Standard Error of Mean	0.001

Figure 16. Statistics of C&RT algorithm

4.3 Comparison of Results

The reason C5.0 and C&RT algorithms chosen because their accuracy rate was found to have a higher value for each dependent variables than for the other algorithms. By comparing these algorithms which belongs to decision trees shows classifications characteristics on real data set and success rates of these two methods. In our application, C5.0 and C&RT algorithms have 99,69% predictive accuracy percentage, as depicted in Figure 8 and 15, respectively.

In conclusion, by utilizing the C5.0 algorithm, we can ascertain which harmful pests lower the crop yield more and which agricultural chemicals are the most effective in combating these agricultural losses. Accordingly, the pesticide names and pest types are represented in the figures and data provided here. Our analysis is done after the identity equivalents are entered into the data type section. Since the process is set up to be determined as a rule, then an if-then equation is formed; but also the same results can be obtained by creating a tree structure.

As a result of our application of the C&RT algorithm, the analysis shows a reduced crop yield from the chemical products due to over-usage. However, it also indicates that the yield should be higher if in each case the recommended pesticide is used at its recommended dose.

5. CONCLUSION

In general, data mining is used in many critical areas of human endeavor, such as education, marketing, banking, and medicine. In this particular study, we have applied some sound techniques of data mining to the field of agriculture, which has in turn allowed us to scrutinize and evaluate data in an unprecedented manner. Within the scope of this study, the data received from the Aydın Nazilli District Directorate of Agriculture was organized for analysis. Then the SPSS Clementine software was used in order to analyze the data with the help of a decision tree algorithm, which is one of the classification methods utilized in data mining. In this study, the well-regarded C5.0 and C&RT algorithms were used to demonstrate pesticide abuse. Some meaningful results were observed when the pesticide doses and corresponding yield rates were selected as the input variable and as the target variable, respectively. Once the rules generated by the C5.0 algorithm about the pesticide dose were examined, two rules emerged, in which the dosage of the pesticide is more than 75, or it is less than or equal to 75. When the results obtained from the C&RT decision tree were examined, it was observed that the tree started branching from the value of 77.500 for the pesticide dose. Taking both algorithms into consideration, it can safely be concluded that the excessive levels of the pesticide dose leads to a decrease in the yield obtained from the product. It can be recommended that the use of pesticides should be reduced — at least to their recommended levels — considering their detrimental effects on human health, the environment, and the crop yields effected by these man-made products. In this regard, farmers should be educated as to how the use of pesticides should be optimized.

It is imperative that all parties involved in agriculture in general, and cotton in particular, understand that pesticide usage can be reduced, thereby benefitting the farmers as well as the ecosystem. This can best be achieved by determining the conditions in which the usage is optimum and then trying to discern the factors that lead farmers to excessive pesticide usage. In this paper, we have shown how data mining can be successfully applied for this purpose. This work clearly shows that the factors that can be controlled by the farmer can be modeled quite well into a data mining problem such that we can obtain answers to the most common questions, by revealing patterns of interest in otherwise unorganized data.

REFERENCES

[1] De Geronimo E, Aparicio VC, Barbaro S, Portocarrero R., Jaime S, & Costa JL. Presence of pesticides in surface water from four sub-basins in Argentina. Chemosphere 2014; 107: 423-431.

[2] Laurance WF, Sayer J, Cassman K. Agricultural expansion and its impacts on tropical nature. Trends in Ecology & Evolution 2014; 29: 107-116.

[3] Masters WA, Djurfeldt AA, De Haan C, Hazell P, Jayne T, Jirström M, Reardon T. Urbanization and farm size in Asia and Africa: implications for food security and agricultural research. Global Food Security 2013; 2: 156-165.

[4] Donaldson D, Kiely T, Grube A. Pesticide's industry sales and usage 1998-1999 market estimates. US Environmental Protection Agency, Washington (DC) 2002: Report No. EPA-733-R-02-OOI.

[5] Fayyad U, Piatesky–Shapiro G, Smyth P. Data mining to knowledge discovery in databases. AI Magazine, 1996. pp. 50-67.

[6] Chien C.-F., Chen L.-F., "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry", Expert Systems with Applications, vol. 34, pp. 280-290, 2008.

[7] Aktar MW, Sengupa D, Chowdhury A. Impact of pesticides use in agriculture: their benefits and hazard. Interdiscip Toxicol 2009; 2: 1-12.

[8] Abdullah A, Brobst, S, Pervaiz I, Umer M, Nisar A. Learning dynamics of pesticide abuse through data mining. Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalization 2003; 32: 151-156.

[9] Abdullah A, Brobst S. Clustering by recursive noise removal. In Proc. Atlantic Symposium on Com Biology and Genome Informatics; 2003; USA, pp. 973-977.

[10] Oakley E, Zhang M, Miller R. Mining pesticide use data to identify best management practices. Renewable Agriculture and Food Systems 2006; 22: 260–270.

[11] Aktar W, Sengupta D, Chowdhury A. Impact of pesticides use in agriculture: their benefits and hazrds. Interdiscip Toxicol 2009; 2: 1–12.

[12] Raghuveer K, Yogesh M J, Shwetha S. Data mining in agriculture: a review AEIJMR 2014; 2: 2348 – 6724.

[13] Küçükönder H, Vursavuş K, Üçkardeş F. Determining The Effect of Some Mechanical Properties on Color Maturity of Tomato with K-Star, Random Forest and Decision Tree (C4.5) Classification Algorithms. "(article in Turkish with an abstract in English)". Türk Tarım- Gıda Bilim ve Teknoloji Dergisi 2015; 3: 300-306.

[14] Ramesh D, Vardhan B V. Analysis of Crop Yield Prediction Using Data Mining Techniques. International Journal of Research in Engineering and Technology 2015; 4: 470-473.

[15] Calıs A, Kayapınar S, Cetinyokus T. An Application on Computer and Internet Security with Decision Tree Algorithms in Data Mining. Endüstri Mühendisliği Dergisi, 2014; 25(3): 2-19.

[16] Budak F, Budak DB. Farm level analysis of pesticide use in cotton production in East Mediterranean region of Turkey. Journal of Environmental Biology, 2006; 27(2): 299-303.

[17] Isin S, Yildirim I. Fruit-growers' perceptions on the harmful effects of pesticides and their reflection on practices: The case of Kemalpasa, Turkey. Crop protection 2007; 26: 917-922.