DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

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

by
Zehra BURDUR

August, 2018
İZMİR

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

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
in Partial Fulfillment of the Requirements for the Degree of Master of Science
in Computer Engineering Program

by
Zehra BURDUR

August, 2018
İZMİR

M.Sc THESIS EXAMINATION RESULT FROM

We have read the thesis entitled **"AN ANALYSIS OF PESTICIDE USE FOR COTTON PRODUCTION THROUGH DATA MINING: THE CASE OF NAZİLLİ"** completed by **ZEHRA BURDUR** under supervision of **ASST.PROF. CANAN EREN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Asst. Prof. Dr. Canan EREN

Supervisor

.....

Asst. Prof. Dr. Sevda ÖZKU

(Jury Member)

.....

Doç. Dr. Mehmet Ünlütürk

(Jury Member)

.....

Prof. Dr. Latif SALUM

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank to my thesis advisor Asst. Prof. Dr. Canan Eren, my research supervisor, for her help, useful suggestions and guidance during this thesis study.

I thank to the Aydın Nazilli District Directorate of Agriculture for sharing this agricultural dataset with us.

I also thank to my friend, Büşra BOSTANCI who shared her knowledge and encourage me throughout the study.

Finally, I like to show my gratitude to thank my family and also my friends and colleagues, for their constant support, unconditional love and encouragement to boost motivation throughout the process of writing my thesis.

Zehra BURDUR

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

ABSTRACT

Data mining involves certain methods of obtaining or inferring meaningful information unknown within the data. Besides of the fact that it is used in the fields of health, marketing, banking; it is also utilized in agriculture. With the increasing significance of precision agriculture practices, farmers have become inclined to be engaged in a more conscious agriculture. Farmers use pesticides to destroy a disease or hazard on plants. Nevertheless, in some researches, it has been revealed that pesticides have harmful effects on human health, environment and plant. Although many farmers are aware of the risks of excessive use of agricultural pesticides, they still use them to get a faster yield and avoid financial loss. The goal of this study 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. The cotton planted fields and used pesticides data received from Aydın Nazilli District Directorate of Agriculture was organized, it was evaluated with the help of classification algorithms in SPSS Clementine software. C5.0, Classification And Regression Tree (C&RT) and Chaid 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.

Keywords: Agriculture, pesticide, decision tree, C5.0 algorithm, C&RT algorithm, Chaid algorithm, SQL

VERİ MADENCİLİĞİ İLE PAMUK ÜRETİMİ İÇİN ZİRAİ İLAÇ KULLANIMININ ANALİZİ: NAZİLLİ ÖRNEĞİ

ÖZ

Veri madenciliği verilerden bilinmeyen anlamlı bilgiler edinme veya tahmin etme yöntemlerini içermektedir. Sağlık, pazarlama, bankacılık alanlarında kullanıldığı gibi tarım alanında da kullanılmaktadır. Hassas tarım uygulamalarına artan önem ile birlikte çiftçiler daha biliçli tarım uygulamaya yönelmişlerdir. Çiftçiler üründe oluşan hastalığı ya da zararlıyı yok etmek için zirai ilaç kullanmaktadırlar. Çeşitli araştırmalarda zirai ilaçların insan sağlığına, çevreye ve bitkiye zararı ortaya koyulmuştur. Pek çok çiftçi zirai ilacın gereğinden fazla kullanımının zararlarının farkında olsa da daha hızlı verim alabilmek ve mali anlamda zarar etmemek için zirai ilaç kullanmaktadır. Bu çalışmanın amacı, pamuk yetiştiricileri için maksimum verim oranı için uygulanacak pestisit miktarını belirten bir model oluşturmaktır. Nihai amaç, veriler üzerinde karar ağacı tabanlı sınıflandırma algoritmaları ile modelleme ve sonuçların gözlemlenmesidir. Aydın Nazilli İlçe Tarım Müdürlüğünden alınan veriler düzenlendikten sonra Spss Clementine yazılımında sınıflandırma algoritmaları ile değerlendirilmiştir. Çalışmada veri madenciliği yöntemlerinde en çok tercih edilen C5.0, C&RT ve Chaid karar ağacı algoritmaları kullanılmıştır. Üründen elde edilen verim oranı ile kullanılan zirai ilaç dozu arasında anlamlı farklar gözlemlenmiştir.

Anahtar Kelimeler: Tarım, zirai ilaç, karar ağacı, C5.0 algoritması, C&RT algoritması, Chaid algoritması, SQL

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 General

Data mining is a process of information that allows a large amount of stored data to be interpreted and made prospective estimates. As in our country, the contribution of agricultural production to a country's economy is of great importance in the world. Programs designed to ensure that production is adequate and of high quality prevent the occurrence of major losses in time and labor power. Data mining is one of the most common techniques having been used in recent years to make decisions quickly and accurately. For predicition and classification, although data mining is used in many fields such as medicine, engineering, biology, finance, industry, its use in the field of agriculture is relatively limited.

The major problem in agriculture is the yield forecast or prediction. In the past, farmers applied traditional methods and relied on their own experience to estimate the yield of the crops they would obtain. It is now easier to estimate the yield thanks to the improvements in the inferring of meaningful information from available data. With variations including climate conditions, soil types, effects of inputs such as fertilizers and pesticides, and a multiplicity of such parameters, the forecast of crop yields has led to data mining.

Agriculture is an indispensable sector all over the world including Turkey 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.

Although farmers know that they obtain crops in different quantities from different parts of the fields or that they have different soil types, they can not process

this knowledge to be more productive. For this reason, they traditionally apply the same amounts of inputs in fertilizers and pesticides by which the plants grow in the field regardless of size.

On the other hand, precision agriculture is a form of business in which the farmer correctly identifies a variable in 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 variables. Precision agriculture, by using data mining and information systems, aims to reduce input usage, avoid waste of resources, increase the yield from crop, and minimize the environmental pollution. Among the targets of precision agriculture are;

- Reduction of chemical costs such as fertilizers and pesticides
- Reduction of environmental pollution
- Providing high quality with high quality products
- Establishment of record in agriculture
- Ensuring a more efficient flow of information for management and accurate decisions.

1.2 Purpose

The aim of this study is to create a model for cotton growers that will indicate the optimum amount of pesticides to apply for the maximum yield rate. This study intends to design and verify a classification model through the usage of decision trees.

Original data was obtained from the Aydın, Nazilli District Directorate of Agriculture. The yield obtained every year and dose of the pesticide used in the crop were not properly maintained. Firstly, some necessary process were taken such as 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. 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 18 attributes (variables) were achieved.

Therefore the database named Agriculture was created using SQL managament studio 2012. Agriculture consist of 6 tables called Agriculture type, Field, Harmful, Pesticide, Production, Villages. Agricultural data was kept into in excel was converted csv format and make ready for SPSS Clementine software. C5.0, C&RT, Chaid decision tree data mining algorithms have been applied for classification using SPSS Clementine software. Thereby, it was observed that there exist some certain suggestive differences between the fertility obtained from the product and the pesticide used.

1.3 Organization of the Thesis

The thesis organized as follows. Chapter 2 provides information on development of data mining, definitons of data mining, application area of data mining, process of data mining and data mining techniques. Chapter 3 summarizes related studies of data mining in the field of agriculture, precision agriculture, agricultural pest control and provides pest and diseases of various crop and their active ingredients. Chapter 4 describes the data set used, database and its tables, and used algorithms are explained in chapter 4. Chapter 5 provides some statistics, the implementation of data mining and analysis results are presented. Chapter 6 contains some concluding remarks.

CHAPTER TWO
DATA MINING

2.1 Description of Data Mining

Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules (Berry & Linoff, 2004). Data mining can be defined in many different ways. A simple definition of data mining is extraction of previously unknown data from large amount of dataset. It is also about processing data and identifying patterns so that you can decide or judge. Data mining has a similar meaning to Knowledge Discovery In Database (KDD).

Data mining can be used in many functional areas such as financial, marketing organization, yield prediction in agriculture, communication. In information technology, data mining techniques integrate several disciplines. Disciplines as neural networks, image and signal processing, pattern recognisation, database technology, artificial intelligence.

Some of the data mining definitions are as follows:

- Data mining has been called exploratory data analysis, among other things. Data mining refers to the analysis of the large quantities of data that are stored in computers. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused (Olson & Delen, 2008, p. 5).

- "Data mining is the process of discovering interesting patterns and knowledge from large amounts of data." (Han, Kamber & Pei, 2012a, p. 8).

- Data mining consists of the (semi-) automatic extraction of knowledge from data. This amount of stored in databases continues grow fast. With an extraction, this large amount of stored data contains valuable hidden knowledge, which could be used to improve the decision-making process of

an organization. However, the number of human data analysts grows at a much smaller than the amount of stored data. Thus, there is a clear need for (semi-)automatic methods for extracting knowledge from (Freitas, 2009).

2.2 Development Of Data Mining

At any stage of everyday life, the concepts of data and information are frequently confronted. These concepts, which are often pronounced in everyday use, are undoubtedly the building blocks of the world of information. These concepts need to be addressed correctly.

Although the data word means real in Latin, it does not always show any concrete reality. The word used as data means plural "datum". In the conceptual sense, data are all recorded events, situations, ideas. They are basically known as raw, unprocessed, raw records. A data that gains a different dimension by processing will be preserved for a different purpose when it is recorded for later use. Information, potentially in the form of knowledge, is associated with the data, arranged, interpreted, processed. The transformation of information into knowledge occurs through the perception of the individual, his assimilation and conclusion. Wisdom is the point to be reached and is at the peak of these concepts. It is a phenomenon that occurs when information is collected and synthesized by a person. Personal qualities such as talent and experience are wisdom elements.

Nowadays wisdom is the most valuable asset. The fact that information is so valuable has been a key driver for the development of information technology. It has become inevitable for computers to be actively involved in knowledge management and production. There is a burst of information when it is looked at daily. The fact that the environment is full of information brings with it information.

The existence of an effective communication environment such as the Internet fuels this situation. Almost everyone at the macro level makes a contribution to this data base and makes use of it. However, this brings with it some problems. It is

necessary to extract the necessary information from such a large amount of data. At this stage, the concept of data mining is confront.

First, hierarchical and network data models with simple data models were developed. Hierarchical data models are data models that have a tree structure, a root on the basis of which each node has one node on top of it and n nodes on the other node. The network data models were a data model in which the type of record and the types of links were determined, and the types of records were determined by the relationship types. Any element in the network data model could interact with one another. However, establishing multiple relationships was not the issue. In the hierarchical data model this was even more limited. Therefore, they can not fully meet the needs of users. In line with these needs, Enhanced Data Models have been developed. These are known as Entity - Relationship, Relational and Object - Oriented data models. Today, relational data model is the most frequently used. Object-oriented data models are still in development.

Data mining has emerged conceptually in the 1960s as computers began to be used to solve data analysis problems. This process is called data dredging instead of data mining, data fishing. In the 1990s, the name of data mining was introduced by computer engineers. The purpose of this group was to emphasize the use of algorithmic computer modules to evaluate data rather than traditional statistical methods. From this point on, scientists are beginning to take various approaches to data mining. At the root of these approaches were disciplines and concepts such as statistics, machine learning, databases, automation, marketing, research.

After the 1990s, statistics moved to a common platform with data mining. Statistics was a collection of methods that provided an assessment and analysis of data over time. Data mining and statistics have been put into a close working relationship in the process of extracting, extracting, analyzing, and preparing the data. Apart from this, data mining, databases and machine learning discipline went together. The concept of machine learning, which is the foundation of today's artificial intelligence studies, is that computers generate new transactions by making certain transactions. After the 1980s, however, the approach changed, and the

machines were built to produce estimation algorithms in more specific terms. This situation did not want to combine the concepts of machine learning with practical statistical analysis under data mining. In the age when information is so valuable, data mining is a crucial step in the way to reach knowledge.

2.3 Data Mining Aplication Areas

Today, data mining is very important in the business world where information needs to be quickly accessed and decided. Researchers have researched large and scattered data sets and concentrated on data mining to meet the need for reliable information. Some data mining application areas are given below.

Banking:

- Customer profile creation and prediction of changes in card usage
- Creation of loyal customers to the bank
- Detection of fraud cases in card usage
- Covert correlation between different financial indicators

Marketing:

- Determination of consumption habits of consumers by determining their consumption habits
- Determination of relations between demographic characteristics of consumers Estimation of response to e-mail campaigns
- Estimation of reactions to a product to be marketed by market analysis

Medicine:

- Estimating and characterizing patient behaviors
- Identification of successful medical therapies on different diseases

- Estimation of potential disease hazards by examining regions in the light of demographic and historical data

Health Services and Insurance:

- Analysis of money to be paid on insurance policy
- Estimating which customers will buy a new insurance policy
- Identifying behavioral patterns of risky customers defining fraud behavior

2.4 Problems Encountered in Data Mining

In database with large amounts of data, major problems may arise. For this reason, data mining systems in small data clusters, simulated environments, may operate incorrectly in database with large amounts, incomplete, loud, empty, waste, contradictory or indefinite data clusters. For this reason, these problems need to be solved while data mining systems are being prepared. The problems that can be encountered in data mining applications include:

Waste Data: Waste data is unnecessary attributes in the set of samples used to get the desired result in the problem. This can lead to confusion during many operations.

Empty Data: A null value in a database can be any value that is not in the primary key. An empty value is an unequal value, including the value itself, including its definition.

Dynamic Data: Organizational on-line databases are dynamic and content changes constantly. This creates important problemss for information discovery methods.

Missing Data: It originates from the size or nature of the data set. When there are missing data, you need to:

- Records or records containing missing data can be removed.

- The average of the variable can be used in place of missing data.
- The most appropriate value can be used based on the available data.

Missing data yields significant problems in statistical analysis. Because statistical analyzes and related packet programs that allow these analyzes to be performed have been developed for situations where all of the data exist.

Handling Different Types Of Data: Real life applications require processing on different types of data, such as data not only of symbolic or categorical data types but also integer, fractional numbers, multimedia data, and geographic information, as in machine learning.

Noisy and Outliers: This is known as non-systematic fatal noise during data entry or data collection. Many qualities in large databases may be wrong. It also includes faults due to faulty measurement during data collection. As a consequence of these mistakes, many of the qualities may be wrong and due to these flaws, the data mining may not reach its full potential. Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set. (Tan, Steinbach & Kumar, 2004).

Limited information: Databases are generally prepared for purposes other than data mining, such as providing features or qualities that enable simple learning tasks. Therefore, some features may not be available to facilitate the learning task.

Database size: Database sizes are growing at a great speed. The database algorithm has been developed so that it can handle many small samples. Extreme care is required to allow the same algorithms to be used in large samples hundreds of times larger.

2.5 The Process of Data Mining

Data mining can be defined as the process of obtaining new information in the light of the available data. It is not possible to access the desired information from large data sets in one step. Data mining methods have improved since traditional methods take time.

Data mining is also commonly known as knowledge discovery process. Some preprocessing is applied before data mining techniques are applied. To be able to apply data mining methods, data held in data warehouses or databases must pass through certain steps. These steps required to implement a successful data mining exercise are given below. As (Hui & Jha, 2000) state process of knowledge discovery, data selection, data cleaning and transformation, data mining, pattern evaluation and knowledge presentation.

The basic steps are illustrated in Figure 2.1

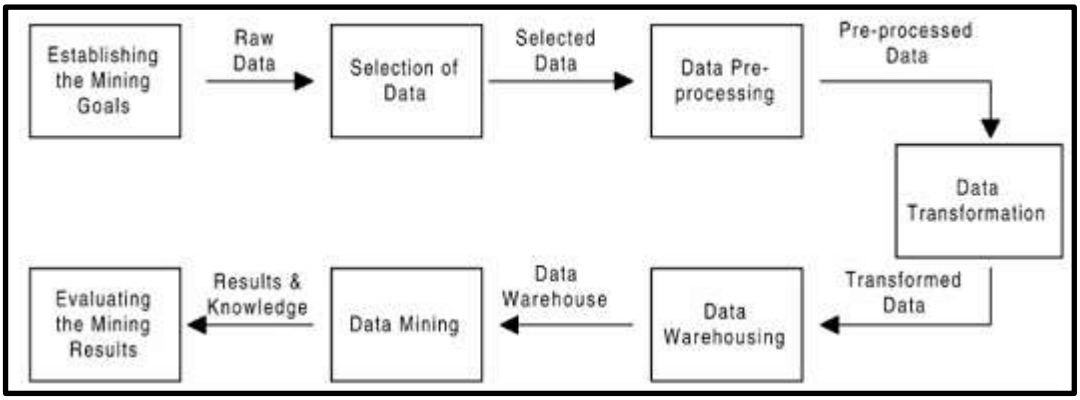


Figure 2.1 Steps of a data mining process (Hui & Jha, 2000)

The first phase of data mining is collecting data. The data is stored in different environments such as a database or data warehouse. Once data collection is complete, it is divided into two as a test and analysis data set.

The second step is data cleaning and data transformation. Data clearing is performed to remove inappropriate or incorrectly entered data in the data.

Transforming the data into different formats and values is also part of this process. Data mining algorithms give more successful results in different data types.

The third step is the data mining in which the data mining model is built according to the characteristics of the data. There are more than one algorithms for the purpose of the project. So the purpose of the project to be done must be very well defined. The most accurate resultant algorithm is used on the data. Programs are needed to implement Data Mining applications. In this scope; SPSS Clementine, SPSS, SAS, Angoss, Kxen, Matlab as commercial, RapidMiner (yale), Weka, R, Orange, Knime as open source many programs have been developed.

The final step is pattern evaluation and knowledge presentation. The main purpose is to find the patterns and use the model for evaluation, the trained model in the data set and its new cases are used. New situations can be predicted in the light of the training model.

2.6 Data Mining Techniques

According to Yethıraj N. G. (2012), data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. (Yethıraj, 2012). Deciding which tehcnique to use is the most important part of the data mining process. Data mining method is widely used around the world for processing data via usage of many classification, regression, clustering, association rules on data attributes and instances is widely used. Only some popular algorithms mentioned in this section.

2.6.1 Predictive Models

Classification and regression methods are types of predictive models for constructing models of predicting trends in future data.

2.6.1.1 Classification

Classification is one of the most popular types of data mining. What it basically does is examine the qualities of a new object and assign that object to a predefined class. Classification approaches normally use predefined classes for model generation to classify data collected from data databases. Common classification techniques are neural networks, decision trees, k-nearest neighbor, genetics algorithms, support vector machines, fuzzy set approach, bayesian classification.

2.6.1.1.1 Neural Networks. A more recent approach to data mining, neural networks are generally used for prediction, classification, data association, data interpretation and data filtering. Neural networks are popular since neural networks are able to learn the computer by taking the human brains as examples, making meaningful information from the data and giving decision making ability by producing new information. Neural network consist of a great number of units called neurons are lined up in layers. The number of neurons in the input layer corresponds to the variables that will be used to feed the neural network and should be the most relevant variables to the problem in question (Haykin 1999). The schematic diagram of a neural network is illustrated in Figure 2.2.

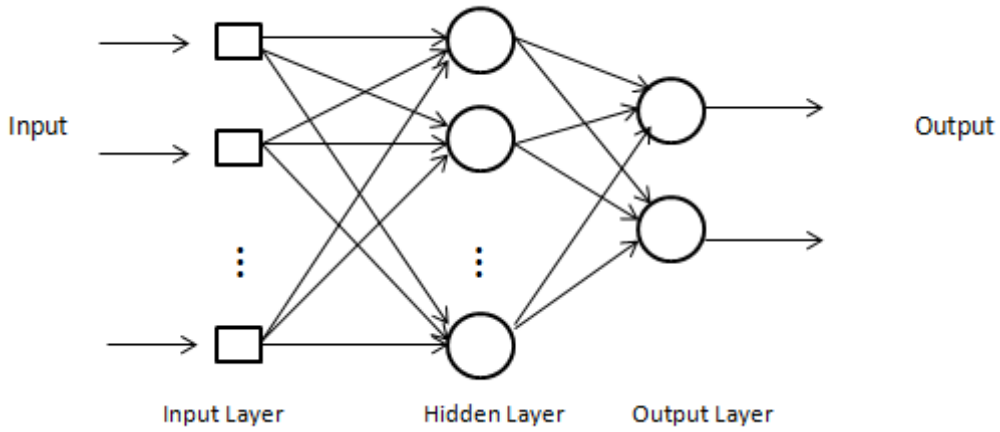


Figure 2.2 Schematic diagram of a neural network

The technique can be divided into two as supervised learning and unsupervised learning.

In supervised learning, both the input and the desired output information are entered into the network. After each attempt, the network responds correctly with its output by comparing their weights up to the level at which they can be accepted. It will iterate.

In unsupervised learning, there is no target vector. Input vector system applied. The system can change itself to produce a compatible output when this input vector is applied to the system.

2.6.1.1.2 Decision Trees. Decision tree is one of the most frequently used in classification and predication in data mining. It can be used in many areas such as production, financial analysis, biology. Because of its simplicity and speed, the high amount of data can be processed in a short time and it becomes more preferable when the amount of data increases according to the alternative methods.

It can be used to process both numerical and classical data. Most machine learning algorithms are useful in numerical applications or useful for classification problems.

Decision tree learning can also be used in both two areas. A sample decision tree structure is illustrated in Figure 2.3.

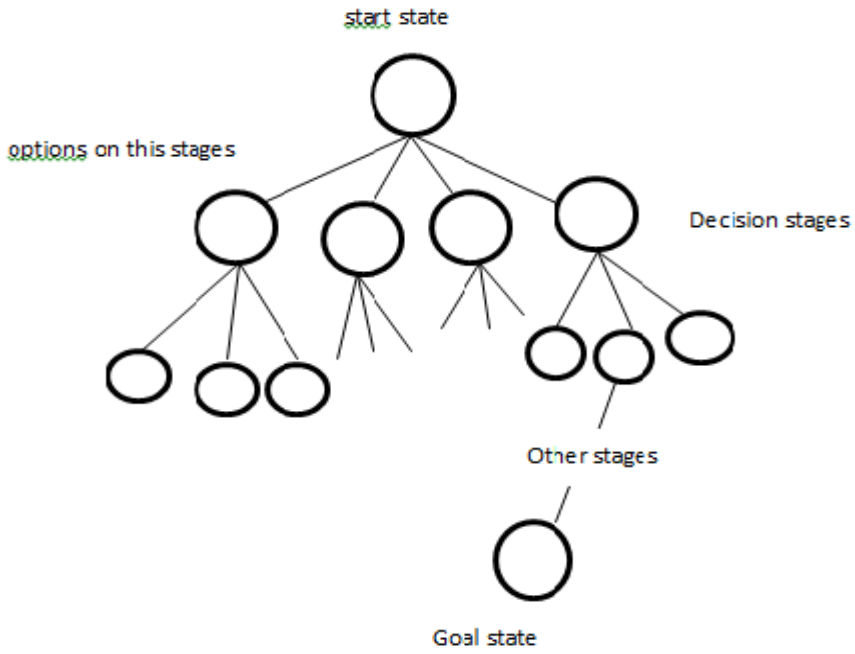


Figure 2.3 Sample decision tree structure

In decision tree learning, a tree structure is created, class labels at the level of the leaves of the tree, and the processes on the properties by means of the arms that go to these leaves and come out from the beginning. During decision tree learning, learned information is modeled on a tree. All the interior nodes of this tree represent the input of one. In decision tree learning, when a tree is learned, the cluster on which it is trained is divided into subsets according to various properties, which is repeated recursively and continues until there is no effect on the prediction of the repetition. This process is called recursive partitioning. For building decision trees methods include several statistical algorithms such as C&RT (Classification and Regression Trees), Chaid (Chi-Squared Automatic Interaction Detection), Quest (Quick, Unbiased, Efficient, Statistical Tree), C4. 5, C5.0.

2.6.1.1.3 K-Nearest Neighbor. K-nearest neighbor algorithm (K-NN) is simple and easy to implement, it is common in classification due to its powerful and useful

learning process. K-nearest neighbor algorithm is applied in a wide range of fields such as machine learning and data mining. K-NN algorithm are particularly preferred for classification applications with advantages such as being easy, analytically trackable, adaptable to local information.

In k-nearest neighbor algorithm, the samples included in the training set are indicated by n dimensional numerical qualities. All training samples are held in a n-dimensional sample space so that each sample represents a point in n-dimensional space. When an unknown sample is encountered, the new sample class label is assigned according to the majority vote of the class labels of the nearest neighbors by determining the closest k samples from the training set.

2.6.1.2 Regression

Regression analysis is used to determine the relationship between two or more variables that have a cause-and-effect relationship between them, and to estimate or predict them by using this relation. In many cases, it is possible to find cause-effect relationships in nature. In data mining independent variables are attributes already known and response variables are what we want to predict. Regression is used for analysis to determine the mathematical expression of the correlation between feed amount-milk amount, study period- received note, fertilizier-yield, age- lenght.

The regression analysis, in which a single independent variable is used, is called a one-way regression analysis, and the regression analysis in which more than one independent variable is used is called a multivariate regression analysis.

2.6.1.2.1 Linear Regression. Linear regression analyis is a model with a dependet and an independet variable in the form of $Y = \alpha + \beta X + \varepsilon$. In this form Y is assumed that it is dependent (result) variable and has a certain error. X is assumed to be an independent (causal) variable and measured without error. α is constant and is the value that Y takes when X = 0. β is the regression coefficient, which expresses the amount of change in X units per unit in Y units, as opposed to 1 unit change in X

units. ε is assumed random error and the mean is normal distribution with zero variance σ^2.

Linear regression aims to fit a straight line to predict behavior and this line is used to forecast a value for one variable given that the other.

2.6.1.2.2 Logistic Regression. In the known linear regression analysis, dependent and independent variables are specified numerically. A dependet variable attribute is specified as a categorical variable the relationship between independent variables or variables is sought through logistic regression. The goal of the logistic regression method is to find the simplest model that can predict the result of the dependent variable. Whether or not the model obtained as a result of the logistic regression analysis is appropriate tested using chi-square test.

2.6.2 Descriptive Models

Descriptive analyzes are considered to be a part of data mining. Many methods (predictive models) can be defined as dependent on the type, scale, or distribution of the data that make up the dataset (eg, linear regression models). This makes descriptive analysis an important step in the initial stages of data mining projects.

2.6.2.1 Clustering

Clustering is the process of grouping data according to their similarities. Most clustering methods use distances between data. The purpose of the clustering analysis is to classify the ungrouped data according to similarities and to help the researcher obtain summarized information. Clustering algorithms are divided into two general classes, one hierarchical and one non-hierarchical. There are two subclasses of the hierarchical class. These; Agglomerative and divisive approaches. The non-hierarchical class is divided into four sub-classes: partitioning, density-based, grid-based, and other approaches. Non-hierarchical clustering methods include the k-means method.

Samples in the same cluster are more similar to each other. The samples in different clusters are less similar to each other. Clustering is accomplished by learning without an educator. Applications such as the discovery of different groups of customers in the grocery stores and the shopping patterns of these groups, the classification of similar genes according to plant and animal classifications and functions in biology, the separation of groups according to types, values and geographical locations of houses in city planning are typical clustering applications. Clustering can also be used to classify documents for disambiguating information on the Web.

2.6.2.1.1 K Means Clustering. The k-Means Clustering Algorithm is one of the most used algorithms in data mining. K-means is the simplest algorithm developed to group a certain number of clusters over a given set of data. The algorithm starts with taking the objects, determining the number of clusters and determining the starting centroids. It assigns the most appropriate group to each object and calculates the number of assigned mass centers after each assignment. The newly formed group is compared to the past group. If there is no change in the group, end the algorithm, otherwise return to previous step. The cycle goes on until no change occurs. Unclustered data and clustered data are displayed in Figure 2.4.

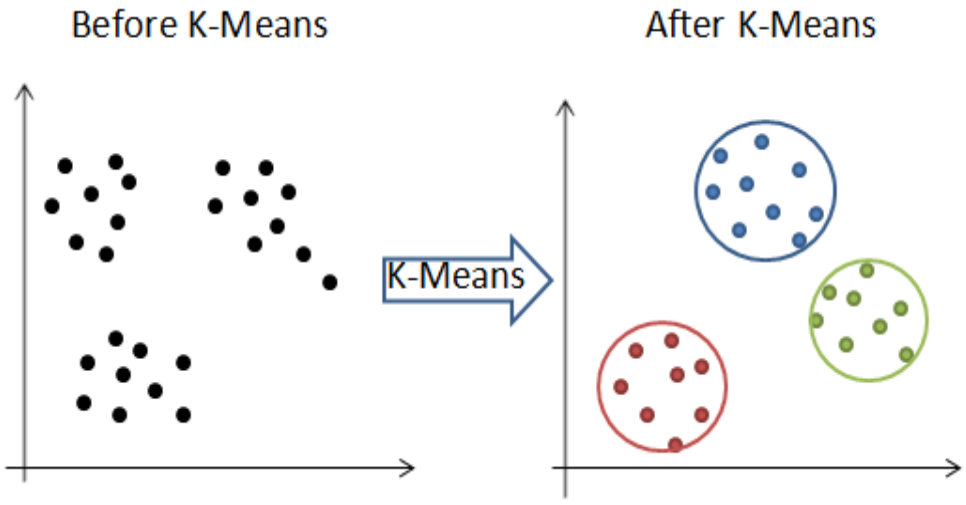


Figure 2.4 Clustering result from the k-means algorithm

2.6.2.1.2 Kohonen Maps. Kohonen map is a neural network based approach. Kohonen nets consist of input and output neurons. The number of input neurons determines the variable number in the data set. Each of the output neurons represents a cluster. Self-organizing maps have a different structure than standard clustering algorithms. The main difference is that instead of performing this operation on separate clusters, the data points associated with each other can be assigned to groups.

2.6.2.1.3 Agglomerative Clustering. Agglomerative clustering is a hierarchical clustering method starts with the assumption of each record as a cluster. The distance between records is calculated using three methods that are frequently used in clusters. The method of calculating the distance between the closest cluster members is single linkage method, calculating the maximum distance between cluster members is complete linkage method, and the third method is calculating the average of each record.

2.6.2.1.4 Divisive Clustering. The divisive approach, also called the top-down approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds (Han, Kamber & Pei, 2012b, p. 449).

2.6.2.2 Association Rules

Association rules is a method of data mining that analyzes past data from large data clusters and finds relationships between them, promoting future decisions by demonstrating the probability of events occurring together. This descriptive method focuses almost exclusively if -then rules on catogorical data by sorting data discover. Relational analysis can establish a meaningful and strong correlation between various variables such as gender and educational status. A relationship can also be established between customer age and income level and purchasing attitudes and behaviors. For example; If there is a strong relationship between having a bank

account and having a private pension insurance, a proposal can be made for private pension insurance for other customers with salary accounts.

2.6.2.2.1 Market Basket Analysis. Market basket analysis is common example of association rules. The rules of association are used to determine which products are sold together in markets. (e.g. if it is thought that who buys pencil, also buy notebook). Today, this method is actively used in many websites. For example, It proposes the products that are likely to be purchased together with this product by way of the purchased product.

Measures of support and confidence are used to establish the relationship between the items. The greater the support and confidence dimensions, the more likely it is that the rules of association are so strong. A value called support number is used to calculate these measures. Apriori, carma, sequence, GRI the most used algorithm in association rules method.

2.6.2.2.2 Apriori Algorithm. Apriori algorithm was developed by Agrawal and Srikant in 1994. In Data Mining, apriori is the most known and used algorithm in association rules algorithms. The name of the algorithm comes from the use of the prior knowledge of common objects, that is, from the "prior" point of taking information from the previous step. Apriori is the first algorithm that uses support-based pruning. With the basic approach, set of one or more items k-item set (e.g. 3-item set: {Honey, Milk, Bread}). The Apriori Algorithm requires an k-items frequent set of itemset to find the (k + 1) items frequent itemset. In this algorithm is, If the k-item set provide the minimum support criterion, then the sub-clusters of this cluster also provide the minimum support criterion. In the rule of association, the association between items is calculated by the criteria of support and confidence. The Support criterion specifies how often the relationship between the items in the field is. Support for product X, with different products X and Y, is the ratio of product X in all purchases. Support for X and Y products is the possibility that X and Y are all in one shop together.

$$\text{Support (X)} = \text{Number of X} / \text{Total Number of Shopping} \quad (2.1)$$

$$\text{Support (X, Y)} = \text{(X, Y)} / \text{Total Number of Shopping} \quad (2.2)$$

The confidence criterion tells us the probability of product Y being associated with product X. The confidence of the rules obtained is directly proportional to the values of support and confidence.

$$\text{Confidence (X, Y)} = \text{Number of Shopping with Number (X, Y)} / \text{X} \quad (2.3)$$

$$\text{Confidence (X} \Rightarrow \text{Y)} = \text{Support (X, Y)} / \text{Support (X)} \quad (2.4)$$

Each rule is expressed by a value of support and conficence. Let me give a simple example about Apriori algorithm, A⇛B [support = 2%, confidence = 60%] shows the 2% support value for the association rule indicates that A and B products are sold together in 2% of all the analyzed purchases. 60% confidence level indicates that 60% of customers who buy product A also buy product B at the same exchange.

Apriori then generates larger itemsets recursively using the following steps :

- Determination of the minimum number of support and minimum confidenc value
- Finding the support value of each item in the item set
- Disable items with lower support than the minimum support value
- Establishment of bilateral associations by taking into consideration the single associations obtained
- Extracting items that are lower than the minimum support value
- Creation of triple associations
- Elimination of the minimum support value from tripartite associations
- Removal of association rules from tripartite associations

CHAPTER THREE
DATA MINING IN THE FIELD OF AGRICULTURE

3.1 Introduction

Data mining is a process of information that allows a large amount of stored data to be interpreted and made forward-looking estimates. As in our country, the contribution of agricultural production to the country's economy is of great importance in the world. Programs designed to ensure that production is adequate and of high quality prevent the occurrence of major losses in terms of time and labor power. One of the most common techniques used in recent years for situations that need to make decisions quickly and accurately is data mining. Data mining is used in many fields such as medicine, engineering, biology, finance, industry for prediction and classification, but its use in the field of agriculture is limited.

The major problem in the field of agriculture is yield prediction. In the past, farmers were applying to traditional methods and their own experience to estimate the yield of the crops they obtained. It is now easier to estimate the yield with improvements in the extraction of meaningful information from stored data. The prediction of crop yields has led to data mining, with variations in climate conditions, soil types, effects of inputs such as fertilizers and pesticides, and a multiplicity of such parameters, with an estimated yield.

3.2 Related studies of Data Mining Techniques in Agriculture

Pesticide usage is the most important issue in the agriculture. In their paper, Abdullah, Brobst, Pervaiz, Muhammed Umer & Nisar (2003) also pointed out the pesticide usage. Their study has reported a negative correlation between pesticide usage and crop yield in Pakistan. They have performed unsupervised clustering of data through “Recursive Noise Removal Heuristic” by Abdullah and Brobst. Through data intensive experiments, they have discovered many interesting patterns (Abdullah, Brobst, Pervaiz, Muhammed Umer & Nisar, 2003).

As Urtubia, Pe´rez-Co, Soto & Pszczo´lkowski (2006) state that wine is a widely produced product all around the world. The fermentation process of the wine is very important since it has an impact on both the productivity of wine-related industries and the quality of wine. If we were able to predict how the fermentation is going to be at the early stages of the process, we could interfere with the process in order to guarantee a regular and smooth fermentation. Fermentation is nowadays studied with the help of using different techniques, such as the k-means algorithm and a technique for classification based on the concept of biclustering. It should be noted that these works are different from the ones in which a classification of different kinds of wine is performed (Urtubia, Pe´rez-Co, Soto & Pszczo´lkowski, 2006).

Oakley, Zhang & Miller (2006) point 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. They analysed the PUR for the best management practices to reduce pesticide use. For prunes in Sutter and Yuba counties to identify on-farm innovation to reduce pesticide use. Surprisingly, their results indicated a lower number in fungicide users and an even higher number in growers using no herbicides in both counties. Their paper also highlights the effectiveness of using research correlation between researchers and community organizations to explore alternative pesticides (Oakley, Zhang & Miller, 2006).

Aktar, Sengupta & Crowdhury (2009) pointed out that impact of pesticides. Their aim is 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. Also they state hazards of pesticide including these titles; direct impact on humans, impact through food commodities, impact through food commodities, impact of enviroment, soil contamination, effect on soil fertility (Aktar, Sengupta & Crowdhury, 2009).

In agriculture, crop yield forecast is a very important problem. Raghuveer, Yogesh & Shwetha S (2014) state 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, decision tree in the field of agriculture (Raghuveer, Yogesh & Shwetha S, 2014).

Küçükönder, Vursavuş & Üçkardeş (2015) pointed out a study in order to determine the effect of the mechanical properties such as maximum force at the skin rupture point, energy at the skin rupture point and the skin firmness on color maturity of tomato by supervised learning algorithms of data mining. In their study, a total of 88 tomato samples were used, and color measurements for each tomato in 4 different equatorial regions were performed and a total of 352 color measurement units were used (Küçükönder, Vursavuş & Üçkardeş, 2015).

D Ramesh & B Vishnu Vardhan (2015) point out that agriculture sector in India is facing some rigorous problems in maximizing the crop productivity. More than 60 percent of the production efficiency still depends on the monsoon rainfalls. Recent developments in Information Technology for agriculture field has become an interesting research area to predict the crop yield. Based on the available data, the problem of yield forecast is a major problem that remains unsolved. Data mining techniques appear as a better choice for this purpose. In agriculture, 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 (Ramesh & Vardhan, 2015).

3.3 Precision Agriculture

Agriculture is an indispensable sector all over the world as it is in Turkey, because it can survive the life of the country's population, the contribution of national income and employment, raw materials and capital supply to other sectors, and

direct and indirect influence of exports. 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 as much as the day-to-day.

Although farmers know that they take crop in different quantities from different parts of the fields, or that they have different soil types, they can not evaluate this knowledge as prolific. For this reason, they traditionally apply in the same amount inputs such as fertilizers and pesticide, which the plant grows in the field as a whole regardless of its size.

Precision agriculture is a form of business that occurs when the farmer correctly identifies a variability in the land using information technology, and it is the meaning of the farmer and the subdivision of the land, which is realized by the input application according to this variable. Precision agriculture aims to reduce input usage, to avoid waste of resources, to increase the yield from crop, to minimize the environmental pollution caused by the production using of data mining and information systems. Among the targets of precision agriculture are;

- Reduction of chemical costs such as fertilizers and medicines
- Reduction of environmental pollution
- Providing high quality and high quality products
- Establishment of registration in agriculture
- Ensuring a more efficient flow of information for management and aquaculture decisions.

3.4 Agricultural Pest Control

As the world population is increasing day by day, arable lands are gradually diminishing 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 the weeds, and

increase the agricultural production. Being a mixture of substances, pesticides are applied to plants to struggle with the diseases and weeds causing losses in cultivation and production.

In the contrary case, farmers would have to cultivate the crops left over from the remaining disease and the pests. Therefore, it is necessary to start a agricultural pest control in order to protect plants from the damage of diseases, pests and weed. In short, agricultural pest control is a mechanism that has some functions like to protect the plants, increase agricultural production and improve the quality of work. The use of pesticides is assumably the most important component that has increased the quality of the agricultural production since 1940s. On the other hand, it is a fact that in some regions and parts of Turkey as well as in other countries, the unconscious usage of pesticide is excessive. Irrefutably, the misusage of pesticides has some certain negative effects on human health, animals and environment. The degree of damage for each disease or pest must be identified and to apply the type of active pesticide the proper dose should be decided by considering the method of control of agricultural pest and time. 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
- Fungicides – fungi

In terms of consumption of agriculture, Turkey is seen to have a very low consumption compared to developed countries in Table 3.1.

Table 3.1 Agricultural consumption as active ingredient by countries (Kantarci, 2007)

Countries	Active Ingredient Consumption
Turkey	0.47
Greece	4.41
Italy	5.25
Spain	3.09
Portuguese	8.36
France	4.24
Germany	2.42
England	3.57
Holland	10.23
Austria	2.06
Denmark	1.18
Hungary	1.75

Table 3.2 Agricultural consumption part according to crops (Dağ et al., 2000)

Crop	Area(kha)	Value	Rate(%)
Cotton	430	51.400	20.4
Cereals	14.750	47.900	19.7
Vegetable	780	41.700	16.2
Fruit	120	32.800	12.7
Wine	480	19.900	7.7
Citrus	80	17.600	6.8
Tobacco	180	7.900	3.1
Rice Plant	60	6.800	2.6
Legume	1970	6.500	2.5
Sunflower	580	2.700	1.0
Others		16.800	6.5
Total		252.000	100

Considering the use of pesticide on the basis of the grown product, it is seen that a large part of the agriculture consumed is used in cotton. Cotton has a part of 20.4 % as shown in Table 3.2.

According to tons of pesticides use statics from Ministry of Food and Livestock in Turkey and shown in Table 3.3 below.

Table 3.3 Pesticides use in Turkey (Tarim, 2017)

	Insecticides	Fungicides	Herbicides	Acaricides	Rodenticides	Other	Total
2006	7.628	19.900	6.956	902	3	9.987	43.376
2007	21.046	16.707	6.669	966	51	3.277	48.716
2008	9.251	17.863	6.177	737	351	5.613	39.992
2009	9.914	17.396	5.961	1.533	78	2.302	37.184
2010	7.176	17.546	7.452	1.040	147	5.344	38.705
2011	6120	18.124	7.407	1.062	421	6.978	40.112
2012	7.264	15.525	7.351	859	247	8.766	40.012
2013	7.741	16.248	7.336	858	129	7.128	39.439
2014	7.586	16.674	7.794	1.513	149	6.007	39.722
2015	8.117	15.984	7.825	1.576	197	5.327	39.026

3.4.1 Pest and Diseases of Various Crop and Their Active Ingredients

Whereas pesticides are called by their trademarks, the active substance may also be inferred by the chemicals included. In the market, there are different trademarks of a single active substance by different companies. In chapter 5, the data of cotton crop will be analysed with some data mining algorithms. Besides, some pesticides for cotton will also be introduced in that chapter. The names of the pests and diseases of some crop and the active substances of the pesticides in tables. Pests and diseases of tomatoes and active ingredients are depicted in Table 3.4.

Table 3.4 Pests and diseases of tomatoes and the active ingredients (Sorhocam, 2017)

Pest and Diseases	Active Ingredients
Tomato Bacterial Wilt Disease And Cancer	Fenhexamid % 50
	Bacillus sublitis
	Imazalil 500 g/l
	Pyrimethanil 300 g/l
Tomato Bacterial Spot Disease	Azoxystrobin 250 g/l
	Captan % 50
	Cymoxanil+Mancozeb %5+%45
	Dimethomorph+Mancozeb %9+% 60
Tomato Leaf Blight Disease	Propineb % 70
	Tebunoconazole 250 g/l
	Maneb % 80

The names of the pests and diseases of cotton and the active substances of the pesticides applied are shown in the Table 3.5 below.

Table 3.5 Pests and diseases of cotton and the active ingredients (Sorhocam, 2017)

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

For olives some pests, diseases and active ingredients are shown in Table 3.6.

Table 3.6 Pests and diseases of olives and the active ingredients (Sorhocam, 2017)

Pest and Diseases	Active Ingredients
Olive fruit fly	Cyfluthrin, 50 g/
	Trichlorfon, %80
	Formothion, 336 g/l
Olive moth	Diflubenzuron
	Diazinon, 185 g/l
	Fenthion, 525 g/l
Olive cochineal	Cochineal Malathion , 190 g/l
	Malathion , 650 g/l

3.5 Harmful Effects of Pesticides

The reasons behind the use of pesticides can be uttered as its low cost and hereby high economic income, the incresing tendency to protect high valued products such as fruits and vegetables. The final effects of pesticides vary interdepending on their formulation, the way of applying and the period of time. Nevertheless, farmers not being aware of the harmful effects of the pesticides are seen using incorrectly excessive amounts of pesticides. However, this phenomenon absolotuley causes some adverse effects in precision agriculture. Nevertheless, it is expected that the farmers who are aware of the harmful effects of the pesticides may behave differently in terms of the selection of the pesticides, application way and application dose. As stated in 'Fruit-growers' perceptions on the harmful effects of pesticides and their reflection on practices: The case of Kemalpasa, Turkey." (Işın & Yıldırım, 2006), whereas the majority of farmers are aware of the harmful effects of pesticides, they are lack of knowledge about the adverse effects of pesticides; thus they are unable to transform their awareness into practice. The reason for the use of unsuitable and/or excessive amounts of pesticides is thought to be, in addition to lack of knowledge, concern on the part of producers about the effectiveness of pesticides and their desire to ensure the yield and quality of their produce (Işın & Yıldırm, 2006).

In general, pesticides should be considered to be poisonous for humans, animals and the environment because all living beings in the ecosystem are directly or indirectly affected from these practices. The penetration of pesticides through the mouth, respiration and skin is thought to have a direct impact on human body. More precisely, pesticides can be absorbed by the soil indirectly for a variety of reasons such as rain, wind, etc., and they can directly reach the seed with harmful effects.

On the other hand, in order to overcome the decline in pesticide efficiency against some various types of invulnerable pests, a usage of higher dose may be required; therefore it may lead to both increased cost and reduced crop yield for the farmers. In

addition, certain factors such as pesticide's chemical structure, its physical properties and type of formulation, mode of application, and climate and agricultural conditions also influence the performance of the pesticides in the environment.

CHAPTER FOUR
USED TECHNOLOGIES

In this chapter, the technologies that are used during application development process are described. The data in the database, its related tables and database diagram are given. In addition to structure related to the database, SPSS Clementine software as a data mining analysis program to be used in the application is also mentioned.

4.1 Database

In this chapter, the technology that is used during application development process will be analyzed; and the avaliable data in the database, its related tables and database diagrams will be examined. In brief, the aim is to create a database to use for farmers. To start with, the name of database formed within the scope of implementation created on SQL Server Management Studio 2012 is “Agriculture”. Each table comprises of a table ID. Once the data is tabulated, the table will have an auto-incrementing number. Thereby, there wouldn’t be any repetition in data entry. In addition to the structure related to the database, SPSS Clementine as a data mining analysis software will also be used in the application. The database diagram is illustrated in the following Figure 4.1.

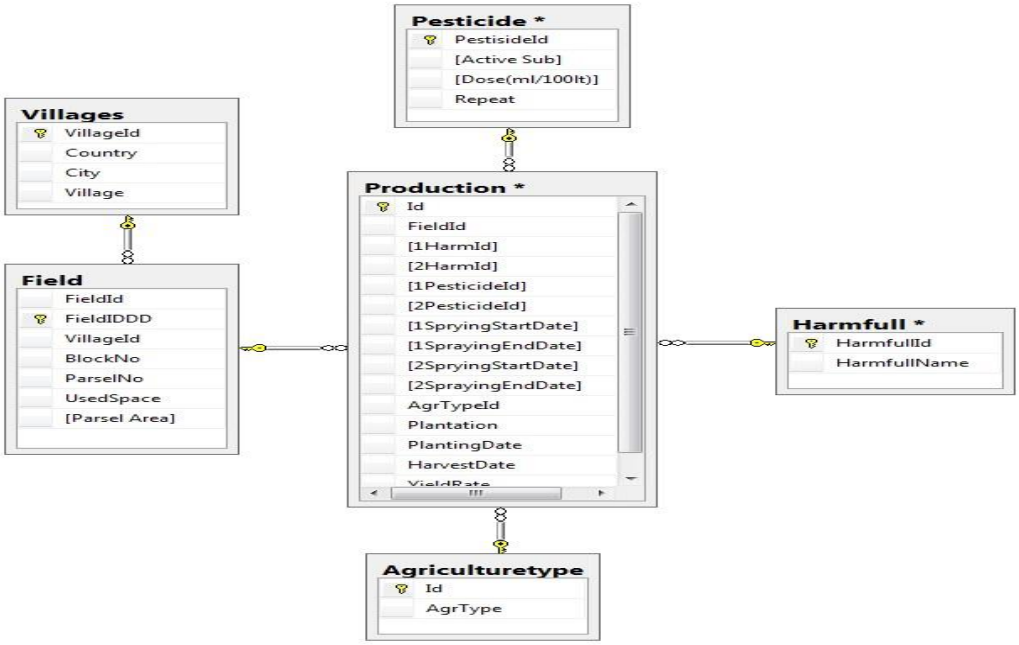


Figure 4.1 Database diagram

4.2 Microsoft SQL Server Management Studio 2012

SQL Server Management 2012 is a relational database management system published by Microsoft. The data in the relational database system is kept in table form. In addition, the tables may be in relation with each other (Microsoft SQL Server Management Studio 2012, n.d.).

SQL (structured query language) is a query language used to query information in relational databases. SQL basically works at a logical level with data. So, in order to be able to select several records from a table, a condition is selected that can select those records. All records that match the runtime come in one step, which can be displayed to the user or sent to another SQL or application. SQL deals with how records come individually and where the database is physically located and how it is stored**.**

However, it is necessary for the user to inform the program about which process to process. At this point, we are talking about the SQL program that allows the user to agree on the database program, and we call database query for every operation the

user makes with each SQL command. SQL commands allow data querying, adding, changing and deleting records to a table, creating, modifying and deleting database objects, controlling access to the database and its objects, ensuring database integrity and consistency**.**

In this title, it will be given how to display data in the form of tables and diagrams using SQL management studio 2012. SQL Server Management Studio is an integrated environment for infrasturture of SQL.

4.3 Designing Database Tables

SQL Server Management 2012 is a relational database management system published by Microsoft. The data in the relational database system is kept in table form. In addition, the tables may be in relation with each other. In this title, the database named Agriculture was created. It consist of 1284 recors of field. Agriculture consist of 6 tables called Agriculture type, Field, harmful, Pesticide, Production, Villages.

First table is Agriculture type which was illustrated in the following Figure 4.2. This table provide information of farming applied in the field is dry or wet.

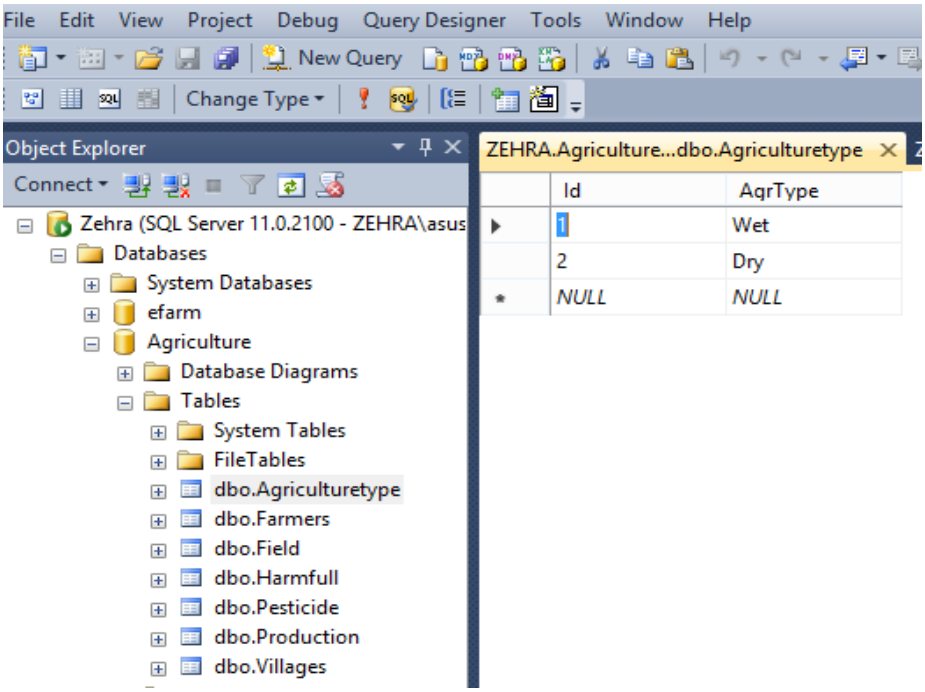


Figure 4.2 Records of agriculture type table

The second table's name is Field as displayed in Figure 4.3. The data includes totally 1284 fields records with the following data attributes: FieldId, FieldIDDD, VillageId, BlockNo, ParcelNo, UsedSpace, ParcelArea. As illustrated in the following figure.

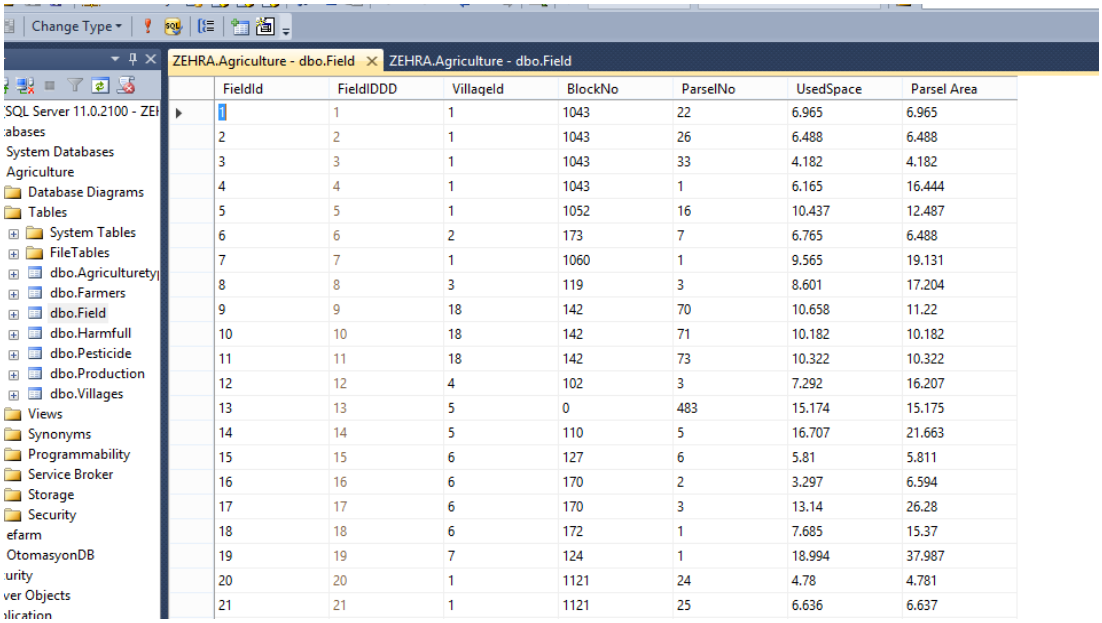


Figure 4.3 Records of field table

The thirt table is the name of Harmfull as shown in Figure 4.4. It includes 10 records pest and disease of cotton crop with the following data attributes: HarmfullId, HarmfullName.

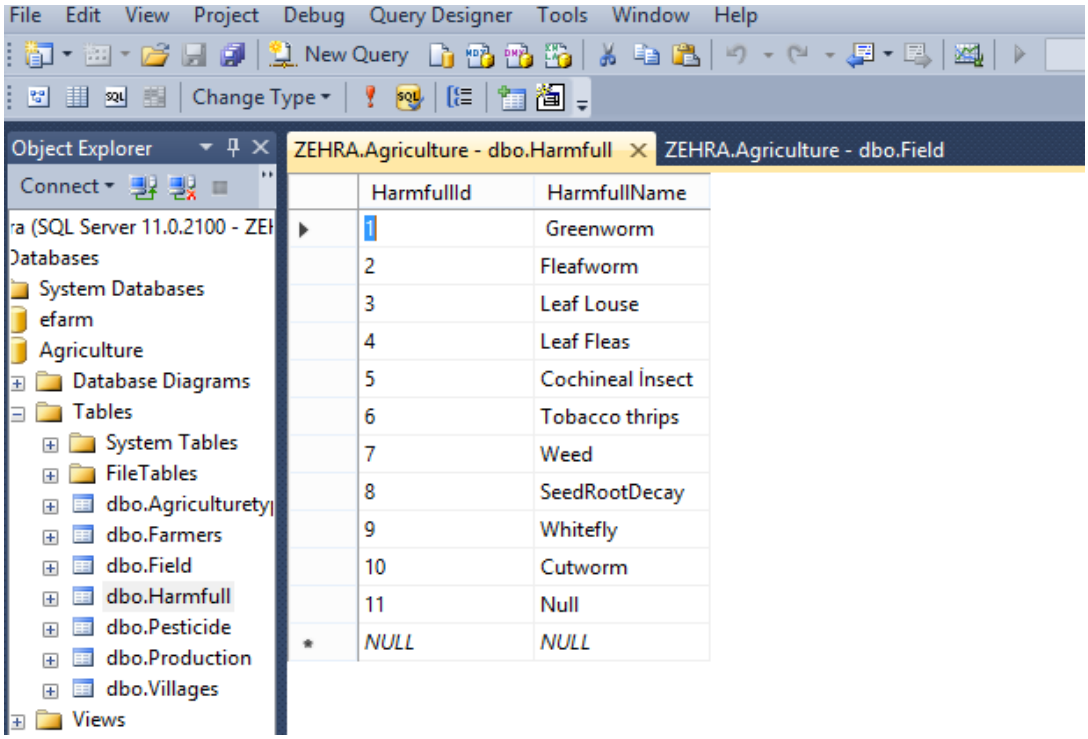


Figure 4.4 Records of harmfull table

The fourth table is concist of pesticide information as in Figure 4.5. It has 29 records of pesticide of cotton pest and disease active ingrediant with the following data attributes: PesticideId, Active Sub, Dose, Repeat.

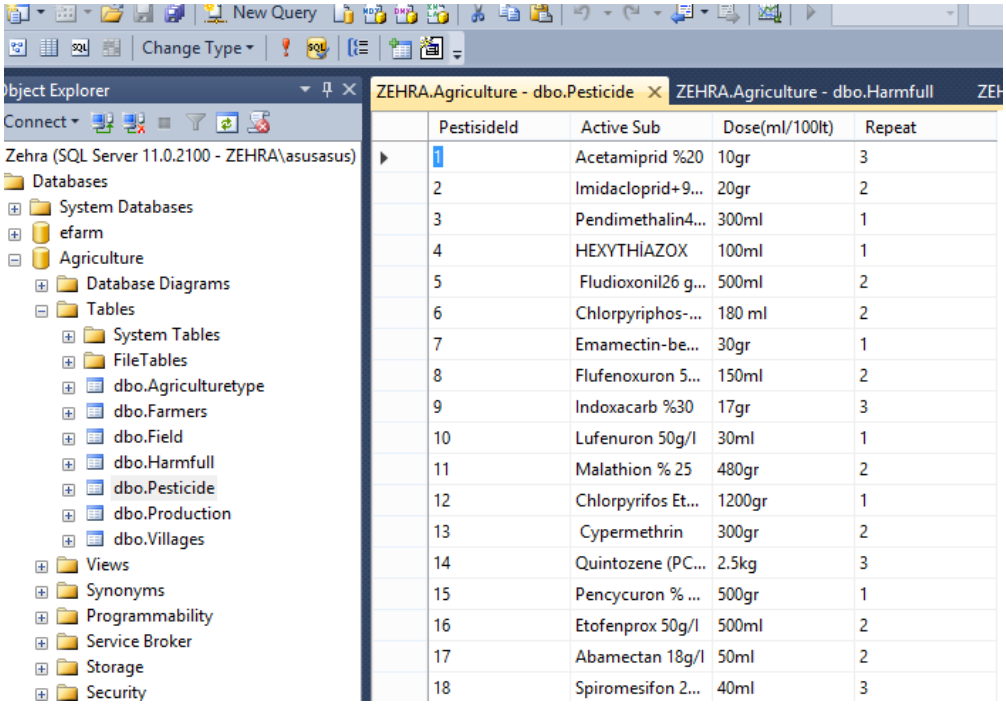


Figure 4.5 Records of pesticide table

The fifth table is production table. It includes records with the following data attributes: Id, FieldId, 1HarmId, 2HarmId, 1PesticideId, 2PesticideId, 1SprayingStartDate, 1SprayingEndDate, 2SprayingStartDate, 2SprayingEndDate, AgrTypeId, Plantation, PlantingDate, HarvestDate, YieldRate as depicted in following Figure 4.6.

ZEHRA.Agriculture - dbo.Field | ZEHRA.Agriculture - dbo.Production | ZEHRA.Agriculture - dbo.Pesticide | ZEHRA.Agriculture - dbo.Harmfull

FieldId	1HarmId	2HarmId	1PesticideId	2PesticideId	1SpryingStartD...	1SprayingEnd...	2SpryingStartD...	2SprayingEnd...	AqrTy
1	1	3	9	1	2016-06-28 00:0...	2016-07-02 00:0...	2016-07-03 00:0...	2016-08-14 00:0...	1
2	4	10	1	12	2016-05-14 00:0...	2016-05-28 00:0...	2016-05-25 00:0...	2016-06-08 00:0...	1
3	2	4	6	21	2016-06-06 00:0...	2016-06-19 00:0...	2016-04-20 00:0...	2016-05-18 00:0...	1
4	5	9	4	26	2016-07-18 00:0...	2016-07-23 00:0...	2016-07-20 00:0...	2016-08-16 00:0...	1
5	6	7	2	3	2016-06-26 00:0...	2016-07-12 00:0...	2016-04-20 00:0...	2016-05-15 00:0...	1
6	10	10	12	12	2016-04-13 00:0...	2016-04-27 00:0...	2016-04-25 00:0...	2016-06-02 00:0...	1
7	9	2	26	6	2016-05-05 00:0...	2016-05-17 00:0...	2016-07-03 00:0...	2016-07-10 00:0...	1
8	2	3	8	20	2016-06-17 00:0...	2016-06-25 00:0...	2016-05-01 00:0...	2016-05-29 00:0...	1
9	2	4	7	1	2016-04-01 00:0...	2016-04-15 00:0...	2016-04-20 00:0...	2016-04-23 00:0...	1
10	1	11	24	29	2016-06-09 00:0...	2016-06-19 00:0...	2016-05-25 00:0...	2016-06-01 00:0...	1
11	5	1	4	23	2016-04-29 00:0...	2016-05-12 00:0...	2016-07-03 00:0...	2016-07-10 00:0...	1
12	1	7	24	3	2016-06-08 00:0...	2016-06-14 00:0...	2016-07-03 00:0...	2016-07-10 00:0...	1
13	3	3	1	20	2016-05-09 00:0...	2016-05-16 00:0...	2016-05-25 00:0...	2016-06-01 00:0...	1
14	2	11	7	29	2016-04-23 00:0...	2016-04-30 00:0...	2016-04-20 00:0...	2016-04-27 00:0...	1
15	2	10	8	12	2016-06-26 00:0...	2016-07-08 00:0...	2016-05-01 00:0...	2016-05-08 00:0...	1
16	8	3	5	1	2016-04-06 00:0...	2016-04-19 00:0...	2016-04-20 00:0...	2016-04-27 00:0...	1
17	1	11	9	29	2016-05-09 00:0...	2016-05-19 00:0...	2016-04-25 00:0...	2016-06-02 00:0...	1
18	4	5	21	4	2016-05-11 00:0...	2016-05-20 00:0...	2016-05-01 00:0...	2016-05-29 00:0...	1

Figure 4.6 Records of production table

The last table is village table. It consist of 24 records of villages of Nazilli as displayed in Figure 4.7. This table consist villages name that produce cotton crop. It includes records with the following data attributes: VillageId, Country, City, Village.

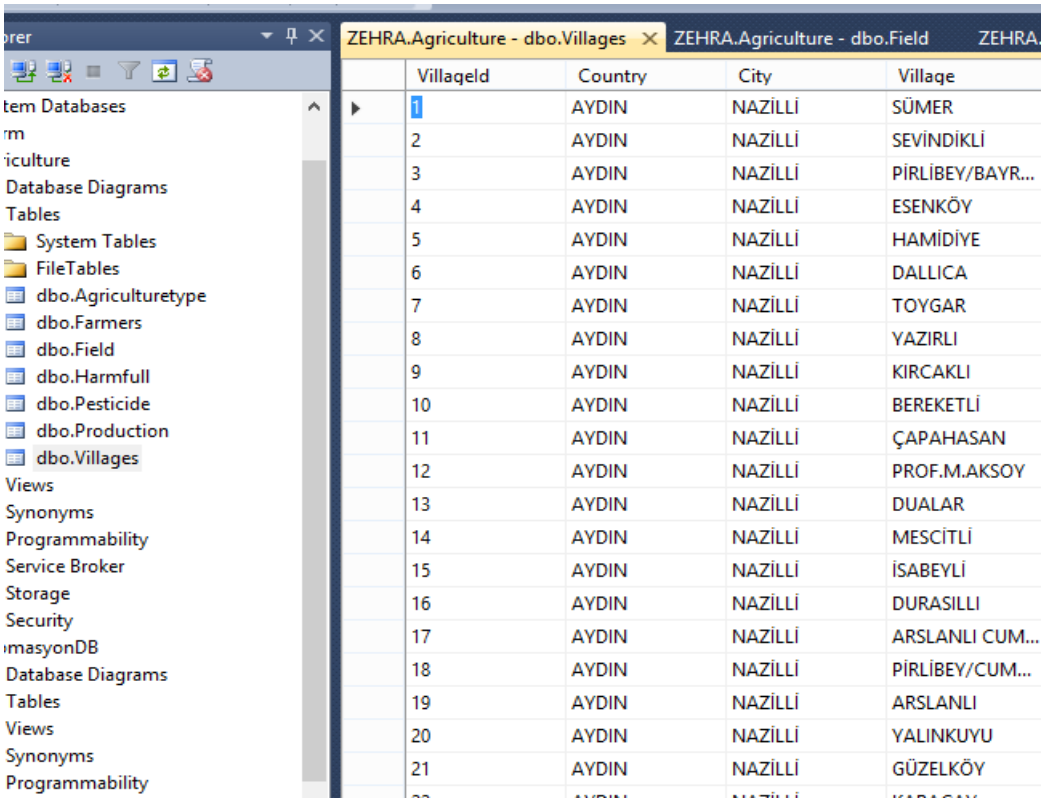


Figure 4.7 Records of village table

4.4 Selection of The Software

Choosing a suitable tool for the data mining process may seem like a major problem and should be done systematically. In market with the increasingly popularity of data mining, more technology tools appeared. Clementine of SPSS, Enterprise Miner of SAS Institute, and Intelligent Miner of IBM are the most preferred data mining tools (Durdu, 2012).

SPSS Clementine uses a methodology called CRISP-DM (Cross-Industry Standard Process for Data Mining) that has six phases for data mining project are given below as defined in (SPSS Inc. Clementine 12.0 User's Guide, 2012).

Understanding what to do or research: The aims and requirements of the project should be articulated. These goals and constraints are appropriate to the problem definition of the data mining. Preparations should be initiated to achieve these goals.

Data Understanding: In this phase , the data is collected. Analyzing the data to get preliminary information about the data. The structure of the data is evaluated.

Data Preparation: The goal in this phase is to prepare the first unprocessed data for the data set to be used. The appropriate variables for analysis should be selected. The required conversions should be made according to the required variables. Missing data must be regulated.

Modelling: In this phase, depending on what you are aiming for, the appropriate data mining technique is chosen. If it is thought that the results found in the obtained model are incorrect and inconsistent, then the first step is reversed and the other variables present are added to the model.

Evaluation: The resulting outcomes are discussed from the point of view of the problem or business owner and are analyzed for suitability. If the model is appropriate, the next step is passed. In this step, the project can be stopped because of inadequate results and the objects used for analysis can be reviewed again.

Deployment: If project results are to be used on a continuous basis, the model we find is reported.

There are many open source and commercial software developed in data mining. As a result, data mining software is divided into open source and commercial software. While commercial software Examples are SPSS Clementine, Matlab, MS SQL Server, Excel, and open source software are WEKA, Orange, R, RapidMiner**.**

Diagram of the model in SPSS Clementine software format is presented in Figure 4.8.

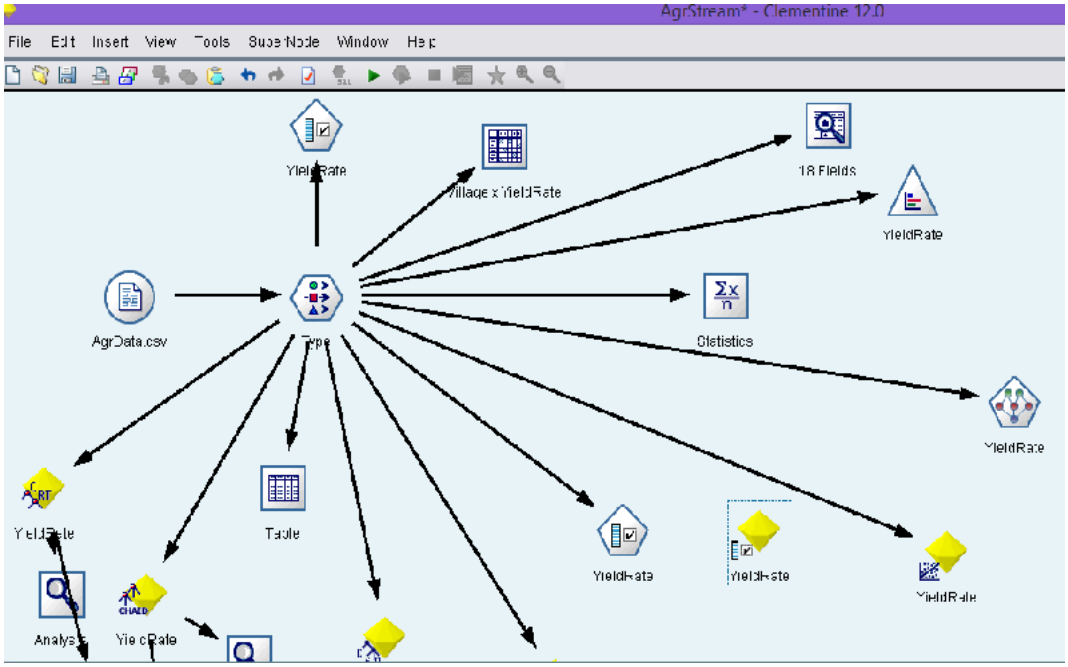


Figure 4.8 Model diagram in SPSS Clementine software

4.5 Used Algorithms

In this chapter, various classification algorithms will be mentioned for the model to be created on chapter 5. These algorithms are C&R tree, C5.0, Chaid. Along with that Chaid, C&R tree, C5.0, are the algorithms used to build decision trees.

4.5.1 C&RT Algorithm

C&RT algorithm known as Cart is a nonparametric statistical tool that uses decision trees to solve classification and regression problems using both categorical and continuous variables. The tree structure is used to transform the heterogeneous data set into a homogeneous and smaller structure. There is the advantage of working with noisy data. It is only a disadvantage to decide according to the binary situation. There are areas where certain rules are used such as the creation of rules to predict future events, the separation of various cases into certain categories such as high, moderate, low risk groups, and the identification of potential members of a particular class. The decision tree is a tree-like structure consisting of roots, branches and leaves, separated from branches that begin with a root covering all the observations

and divide the data into subgroups as the tree descends downward. Each node is a node in the tree growing from the root to the branch, the non-homogeneous nodes are called child nodes and the homogeneous nodes are called parent nodes. It is the decision tree algorithm obtained by using the binary tree structure that two child nodes are formed from the parent node. The fact that the first node, the so-called root, is different will also change the way to be traced when reaching the extreme leaf and thus the classification. There are three basic steps as creating the maximum tree, determining the depth of the tree, and applying the test data grossly. It is an algorithm based on dividing all arguments into subgroups. The best independent variable is selected using the variance of gini, twoing, chi-squared deviation in the purity and change measure. Using gini and twoig algorithms, a maximum tree is created to obtain the best pure partition state.

First, if we look at the gini algorithm, the goal for the best partitioning situation is to get the largest set of data at each step. It is also isolated for the part we are not interested in after splitting. On the other hand, the twoing algorithm, provides a more balanced structure for handling the 50% of the parent and child nodes. Since the C&R tree algorithm also works on the continuous type data, it contains a method called regression.

The second step is to determine the depth of the tree. The minimum number of records that a node will have is n. Generally, a value of 10% of the sample data set is selected, and if n is less than this value, the algorithm will stop working. Choosing a larger value accelerates the algorithm's operation, but a decision tree that makes the wrong decision can occur. If there is no value that can be classified, or if the remaining values belong to a class, the loop algorithm is terminated and the decision tree is created.

4.5.2 C5.0 Algorithm

It is the improved version of the C4.5 algorithm used for large data sets. It allows to obtain more smooth trees in form. Tree is created by multiple branches from each node. The predictor's category number equals the number of branches. A single classifier combines multiple decision trees. The information gain is used for the sorting operation. Entropy is used for each attribute to determine the most discriminating qualities. Entropy shows the likelihood of uncertainty, randomness, and unexpected state of occurrence.

With p_i referring to the probability of a given class as the outcome for each of m possible classes for the target variable, information entropy is defined as

$$I_E(p) = -\sum_{i=1}^{m} p_i \log p_i = info(p) \tag{4.1}$$

Information gain refers to the difference in information entropy before and after the split. So for a given split S with k partitions, n_i equal to the number of samples that would result in the i^{th} partition from S.

$$info_{before}^{S} = -\sum_{i=1}^{m} p_i \log p_i \tag{4.2}$$

$$info_{after}^{S} = info_{i\frac{n_i}{n}} \tag{4.3}$$

So the information gain is calculated as

$$\text{gain(S)}=info_{before}^{S} - info_{after}^{S} \tag{4.4}$$

4.5.3 Chaid Algorithm

Chaid is an effective statistical technique for partitioning purposes. As in other decision trees algorithms, a target variable (dependent variable) and two independent variables are used to describe the target variable. While other algorithms are derived from binary trees, Chaid derives multiple trees.

One of the most used decision tree algorithms other than Cart is Chaid. Chaid (Chi-squared Automatic Interaction Detector) is a method that uses chi-square statistics for the detection of optimal partitions. Evaluates all values of a potential predictor variable using the significance of a statistical test as a criterion. The target combines all values that are evaluated homogeneously with respect to the variable or the bound variable to be of the same meaning, and evaluates all other values as heterogeneous. Next, each node provides a group of homogeneous values of the selected variable, by selecting the best predictor variable according to the first branch form of the decision tree. This process continues until the tree is fully grown. The statistical test used depends on the level of measurement of the target variable.

CHAPTER FIVE
DEVELOPMENT PROCESS

5.1 Introduction

In this section, information about attributes of the agricultural data to be used for applications and the modeling and the results of the algorithms to be applied on these data will be mentioned. The modeling of the various classification algorithms on the data and the observation of the results were aimed. The reason for choosing the classification algorithms as the algorithm is that there is a prediction between the yield rate and the pesticide dose used. In addition to the classification method, classification and regression analysis was performed on the data. This is because regression analysis is used to determine the relation between two or more variables that have a cause-and-effect relationship between them, and to make estimation or prediction about that topic by using this relation. In many cases, it is possible to find cause-effect relationships in the nature. In this thesis, a regression model is used to explain the relationship between the yield rate and the pesticide doses desire variables used.

In this chapter, various classification algorithms will be mentioned for the model to be created on chapter 5. Along with that C5.0, C&RT, Chaid are the algorithms used to build decision trees.

5.2 Information About Attributes Used in The Application and Some Statistics

The data includes parcel production certificate throughout the Nazilli district records of cotton crop. The data is about agriculture. 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. As illustrated in the following figure,

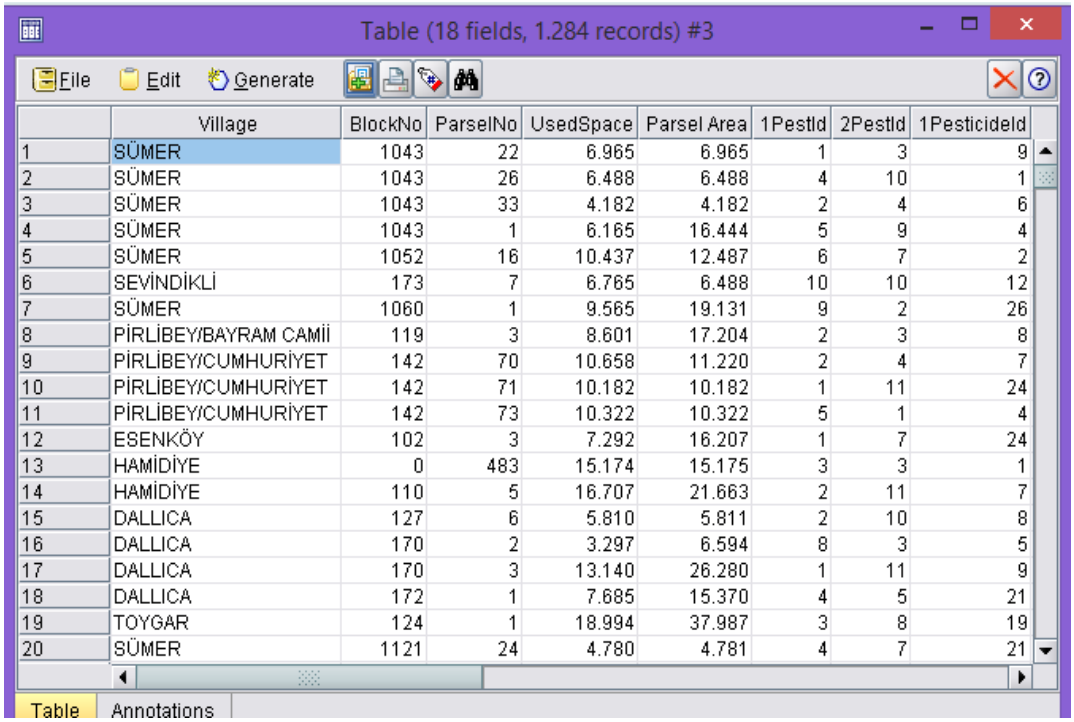


Table (18 fields, 1.284 records) #3

File Edit Generate

	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

Table Annotations

Figure 5.1 Details of the cotton crop records

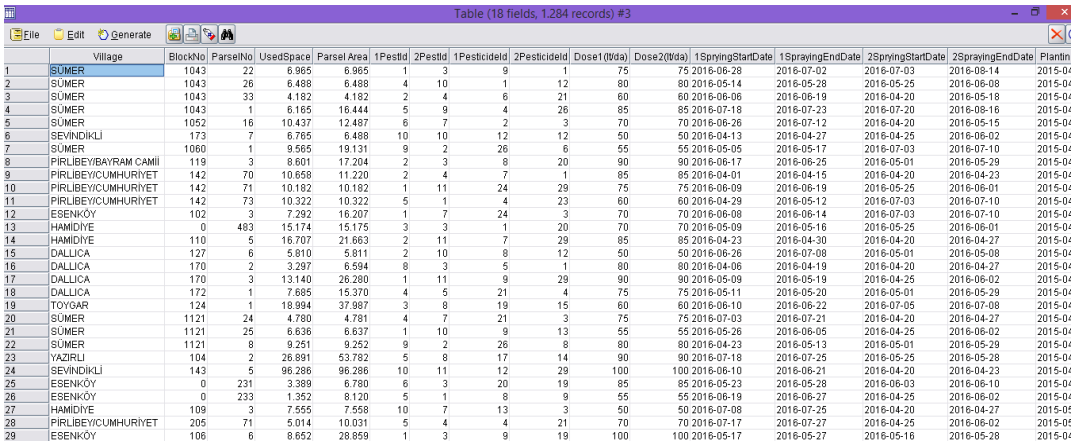


Table (18 fields, 1.284 records) #3

File Edit Generate

	Village	BlockNo	ParselNo	UsedSpace	Parsel Area	1PestId	2PestId	1PesticideId	2PesticideId	Dose1(lt/da)	Dose2(lt/da)	1SpryingStartDate	1SprayingEndDate	2SpryingStartDate	2SprayingEndDate	Plantin
1	SÜMER	1043	22	6.965	6.965	1	3	9	1	75	75	2016-06-28	2016-07-02	2016-07-03	2016-08-14	2015-04
2	SÜMER	1043	26	6.488	6.488	4	10	1	12	80	80	2016-05-14	2016-05-28	2016-05-25	2016-06-08	2015-04
3	SÜMER	1043	33	4.182	4.182	2	4	6	21	60	60	2016-06-06	2016-06-19	2016-04-20	2016-05-18	2015-04
4	SÜMER	1043	1	6.165	16.444	5	9	4	26	85	85	2016-07-18	2016-07-23	2016-07-20	2016-08-16	2015-04
5	SÜMER	1052	16	10.437	12.487	6	7	2	3	70	70	2016-06-26	2016-07-12	2016-04-20	2016-05-15	2015-04
6	SEVİNDİKLİ	173	7	6.765	6.488	10	10	12	12	50	50	2016-04-13	2016-04-27	2016-04-25	2016-06-02	2015-04
7	SÜMER	1060	1	9.565	19.131	9	2	26	6	55	55	2016-05-05	2016-05-17	2016-07-03	2016-07-10	2015-04
8	PİRLİBEY/BAYRAM CAMİİ	119	3	8.601	17.204	2	3	8	20	90	90	2016-06-17	2016-06-25	2016-05-01	2016-05-29	2015-04
9	PİRLİBEY/CUMHURİYET	142	70	10.658	11.220	2	4	7	1	85	85	2016-04-01	2016-04-15	2016-04-20	2016-04-23	2015-04
10	PİRLİBEY/CUMHURİYET	142	71	10.182	10.182	1	11	24	29	75	75	2016-06-09	2016-06-19	2016-05-25	2016-06-01	2015-04
11	PİRLİBEY/CUMHURİYET	142	73	10.322	10.322	5	1	4	23	60	60	2016-04-29	2016-05-12	2016-07-03	2016-07-10	2015-04
12	ESENKÖY	102	3	7.292	16.207	1	7	24	3	70	70	2016-06-08	2016-06-14	2016-07-03	2016-07-10	2015-04
13	HAMİDİYE	0	483	15.174	15.175	3	3	1	20	70	70	2016-05-09	2016-05-16	2016-05-25	2016-06-01	2015-04
14	HAMİDİYE	110	5	16.707	21.663	2	11	7	29	85	85	2016-04-23	2016-04-30	2016-04-20	2016-04-27	2015-04
15	DALLICA	127	6	5.810	5.811	2	10	8	12	50	50	2016-06-26	2016-07-08	2016-05-01	2016-05-08	2015-04
16	DALLICA	170	2	3.297	6.594	8	3	5	1	80	80	2016-04-06	2016-04-19	2016-04-20	2016-04-27	2015-04
17	DALLICA	170	3	13.140	26.280	1	11	9	29	90	90	2016-05-09	2016-05-19	2016-04-25	2016-06-02	2015-04
18	DALLICA	172	1	7.685	15.370	4	5	21	4	75	75	2016-05-11	2016-05-20	2016-05-01	2016-05-29	2015-04
19	TOYGAR	124	1	18.994	37.987	3	8	19	15	60	60	2016-06-10	2016-06-22	2016-07-05	2016-07-08	2015-04
20	SÜMER	1121	24	4.780	4.781	4	7	21	3	75	75	2016-07-03	2016-07-21	2016-04-20	2016-04-27	2015-04
21	SÜMER	1121	25	6.636	6.637	1	10	9	13	55	55	2016-05-26	2016-06-05	2016-04-25	2016-06-02	2015-04
22	SÜMER	1121	8	9.251	9.252	9	2	26	8	80	80	2016-04-23	2016-05-13	2016-05-01	2016-05-29	2015-04
23	YAZIRLI	104	2	26.891	53.782	5	8	17	14	90	90	2016-07-18	2016-07-25	2016-05-25	2016-05-28	2015-04
24	SEVİNDİKLİ	143	5	96.286	96.286	10	11	12	29	100	100	2016-06-10	2016-06-21	2016-04-20	2016-04-23	2015-04
25	ESENKÖY	0	231	3.389	6.780	6	3	20	19	85	85	2016-05-23	2016-05-28	2016-06-03	2016-06-10	2015-04
26	ESENKÖY	0	233	1.352	8.120	5	1	8	9	55	55	2016-06-19	2016-06-27	2016-04-25	2016-06-02	2015-04
27	HAMİDİYE	109	3	7.555	7.558	10	7	13	3	50	50	2016-07-08	2016-07-25	2016-04-20	2016-04-27	2015-05
28	PİRLİBEY/CUMHURİYET	205	71	5.014	10.031	5	4	4	21	70	70	2016-07-17	2016-07-27	2016-04-25	2016-06-02	2015-05
29	ESENKÖY	106	6	8.652	28.859	1	3	9	19	100	100	2016-05-17	2016-05-27	2016-05-16	2016-05-28	2015-04

Figure 5.2 Records of cotton crop

In the database, pesticideId and active substances records represent as shown in the following Table 5.1.

Table 5.1 Representation of pesticide name

PestisideId	Active Sub.
1	Acetamiprid %20
2	Imidacloprid+90 G/L Beta-Cyfluthrim210g/l
3	Pendimethalin450 G/L
4	Hexythiazox
5	Fludioxonil26 g/I+Metalaxyl-m 10 G/L
6	Chlorpyriphos-ethyl 480 g/l
7	Emamectin-benzoate %5
8	Flufenoxuron 50 g/l
9	Indoxacarb %30
10	Lufenuron 50g/l
11	Malathion % 25
12	Chlorpyrifos Ethyl % 25
13	Cypermethrin
14	Quintozene (PCNB) % 18
15	Pencycuron % 20+Captan % 50
16	Etofenprox 50g/I
17	Abamectan 18g/I
18	Spiromesifon 240g/I
19	Pymetrozine %25
20	Flonicamed %50
21	Tau-Flovalinate 240g/I
22	Dimethoate 400g/I
23	Pyridalyl 500g/I
24	Spinosad 480g/I
25	Flubendiamde%20
26	Buprofezin 400g/I
27	Bifenthrin 100g/I
28	Pyriproxyfen 100g/I
29	null

In the database, pestId and pesticideId records represent pest name and pesticide name as shown in the following Table 5.2. PestId substitute for pest names as depicted in the following Table 5.2.

Table 5.2 Representation of pest name

PestId	Pest Name
1	Greenworm
2	Fleafworm
3	Leaf Louse
4	Leaf Fleas
5	Cochineal İnsect
6	Tobacco thrips
7	Weed
8	SeedRootDecay
9	Whitefly
10	Cutworm
11	Null

Before starting the application with a data, it should be apply dialog setting and read values for specified fields. As illustrated in the following figure 5.3, the Yield Rate is set as the target value and it is selected to be "Out". Direction should be set "In" for all of the other fields, though.

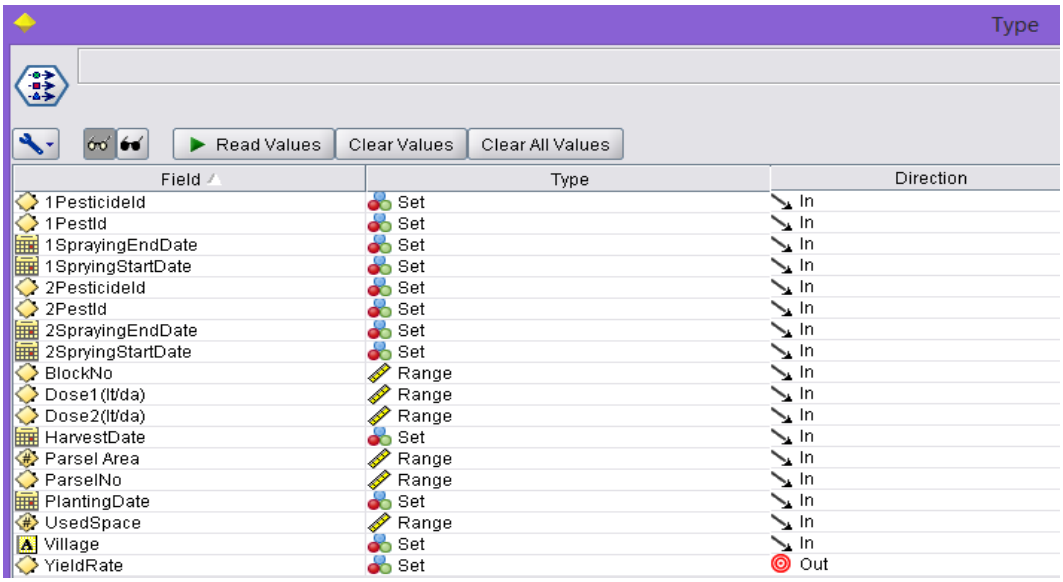
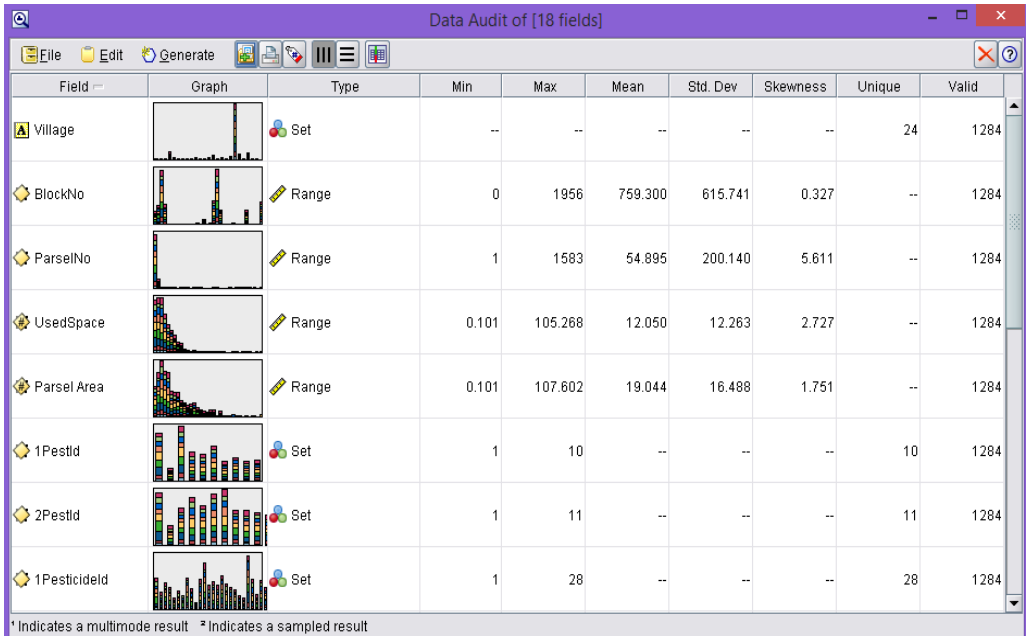


Figure 5.3 Data type of fields

The data audit browser is displayed with descriptive statistics for each field. In this statistics shows that each field has 1284 valid value. As shown in figure 5.4, the fields have max, min, mean, std. dev., skewness, unique statistics.

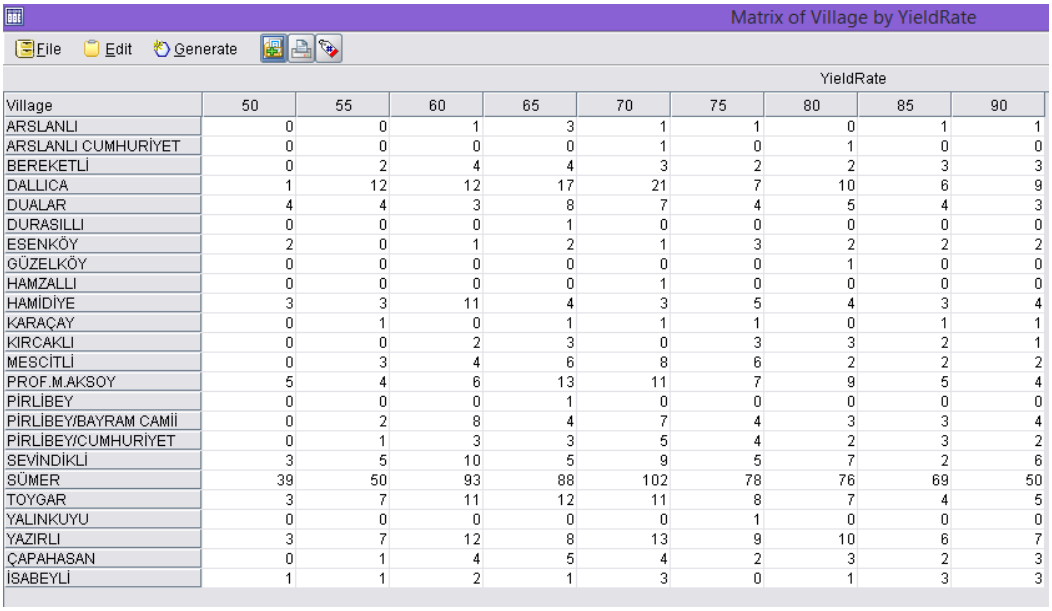


Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Village		Set	--	--	--	--	--	24	1284
BlockNo		Range	0	1956	759.300	615.741	0.327	--	1284
ParselNo		Range	1	1583	54.895	200.140	5.611	--	1284
UsedSpace		Range	0.101	105.268	12.050	12.263	2.727	--	1284
Parsel Area		Range	0.101	107.602	19.044	16.488	1.751	--	1284
1PestId		Set	1	10	--	--	--	10	1284
2PestId		Set	1	11	--	--	--	11	1284
1PesticideId		Set	1	28	--	--	--	28	1284

[1] Indicates a multimode result [2] Indicates a sampled result

Figure 5.4 Data audit of fields

Figure 5.5 shows that counts of cross tabulation using unsorted rows and columns village and yield rate.



	YieldRate								
Village	50	55	60	65	70	75	80	85	90
ARSLANLI	0	0	1	3	1	1	0	1	1
ARSLANLI CUMHURİYET	0	0	0	0	1	0	1	0	0
BEREKETLİ	0	2	4	4	3	2	2	3	3
DALLICA	1	12	12	17	21	7	10	6	9
DUALAR	4	4	3	8	7	4	5	4	3
DURASILLI	0	0	0	1	0	0	0	0	0
ESENKÖY	2	0	1	2	1	3	2	2	2
GÜZELKÖY	0	0	0	0	0	0	1	0	0
HAMZALLI	0	0	0	0	1	0	0	0	0
HAMİDİYE	3	3	11	4	3	5	4	3	4
KARAÇAY	0	1	0	1	1	1	0	1	1
KIRCAKLI	0	0	2	3	0	3	3	2	1
MESCİTLİ	0	3	4	6	8	6	2	2	2
PROF.M.AKSOY	5	4	6	13	11	7	9	5	4
PİRLİBEY	0	0	0	1	0	0	0	0	0
PİRLİBEY/BAYRAM CAMİİ	0	2	8	4	7	4	3	3	4
PİRLİBEY/CUMHURİYET	0	1	3	3	5	4	2	3	2
SEVİNDİKLİ	3	5	10	5	9	5	7	2	6
SÜMER	39	50	93	88	102	78	76	69	50
TOYGAR	3	7	11	12	11	8	7	4	5
YALINKUYU	0	0	0	0	0	1	0	0	0
YAZIRLI	3	7	12	8	13	9	10	6	7
ÇAPAHASAN	0	1	4	5	4	2	3	2	3
İSABEYLİ	1	1	2	1	3	0	1	3	3

Figure 5.5 Matrix of village by yield rate

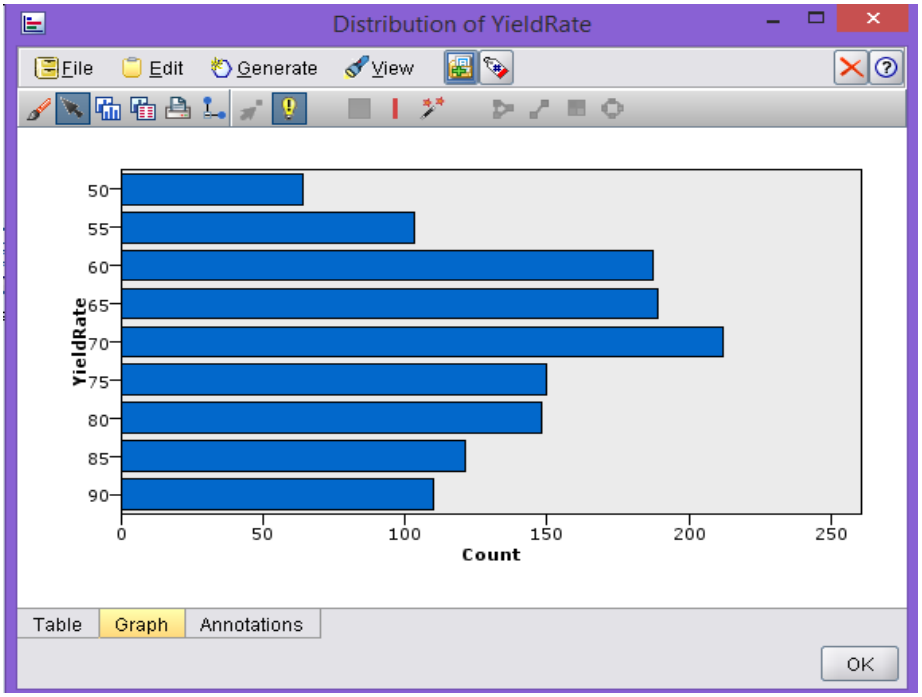
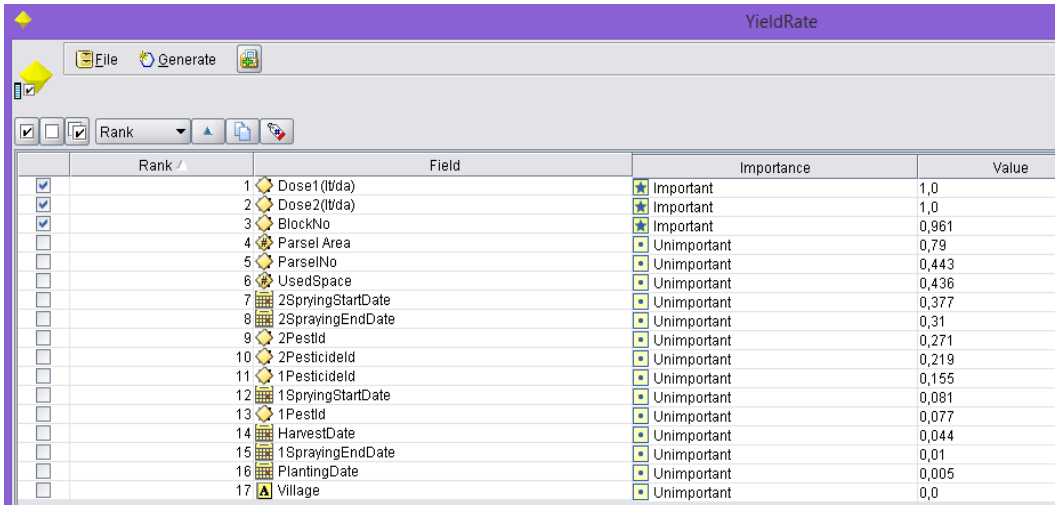


Figure 5.6 Distribution of yield rate

When using modeling feature selection, dose1, dose2, blockNo are important fields with their values as shown in the figure 5.7.



Rank	Field	Importance	Value
1	Dose1(lt/da)	Important	1,0
2	Dose2(lt/da)	Important	1,0
3	BlockNo	Important	0,961
4	Parsel Area	Unimportant	0,79
5	ParselNo	Unimportant	0,443
6	UsedSpace	Unimportant	0,436
7	2SpryingStartDate	Unimportant	0,377
8	2SprayingEndDate	Unimportant	0,31
9	2PestId	Unimportant	0,271
10	2PesticideId	Unimportant	0,219
11	1PesticideId	Unimportant	0,155
12	1SpryingStartDate	Unimportant	0,081
13	1PestId	Unimportant	0,077
14	HarvestDate	Unimportant	0,044
15	1SprayingEndDate	Unimportant	0,01
16	PlantingDate	Unimportant	0,005
17	Village	Unimportant	0,0

Figure 5.7 Feature selection of yield rate

5.3 Implematations

5.3.1 Implementation of C5.0 Algorithm

Considering the accuracy rate of the algorithm, it is seen that the algorithm has a high accuracy ratio with 99.69 per cent in Figure 5.8.

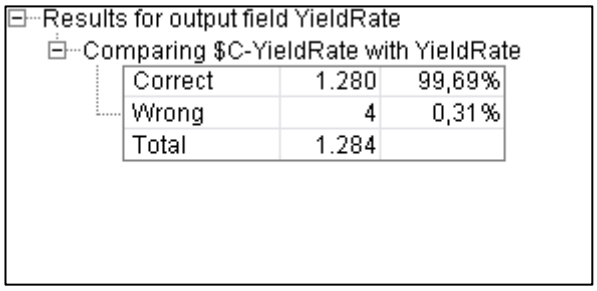


Figure 5.8 The rate of accuracy for the c5.0 algorithm

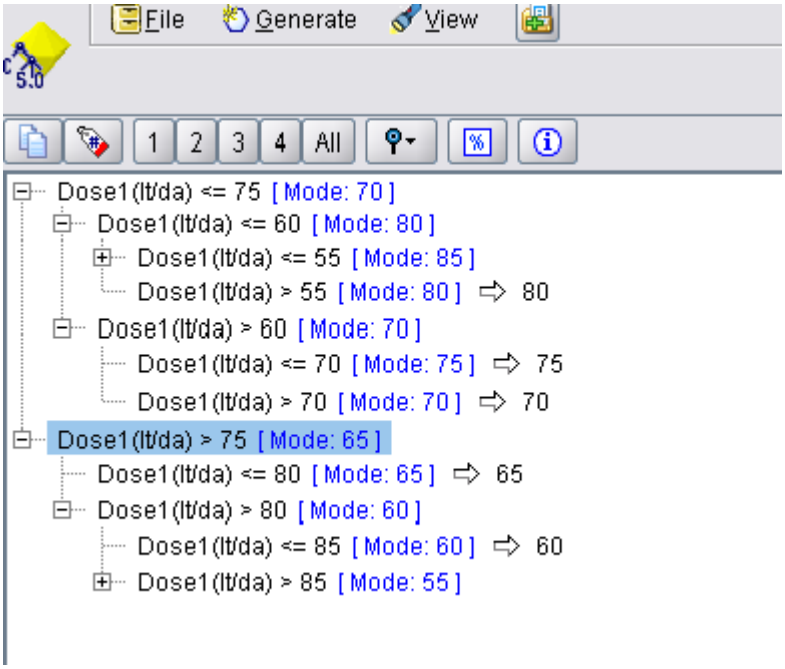


Figure 5.9 The rules for pesticide dose by using c5.0 algorithm

As shown above Figure 5.9, 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, the algorithm generates some rules and reveals the results. To interpret this process, it can be uttered that the algorithm discloses two main rules.

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

The second rule is the case of being more than 75 in dose. While the dose is more than 75, the yield rate appears as 65%, but when the dose is equal and more than 80, the yield rate is 65%. However, if the dose is increased and becomes more than 80, a rate of 60% in yield is achieved. In another case, on condition that the dosage of pesticide is equal and more than 85, the yield rate again ends up in 60%. Nevertheless, when the dose is more than 85, the yield rate drops and becomes 55%. To sum up, it has been observed in the analysis that as a result of the excessive use of pesticides by farmers against various harmful pests, the yield obtained from the crops is reduced. For further understanding, the algorithm is also depicted as a tree structure which can be seen below Figure 5.10.

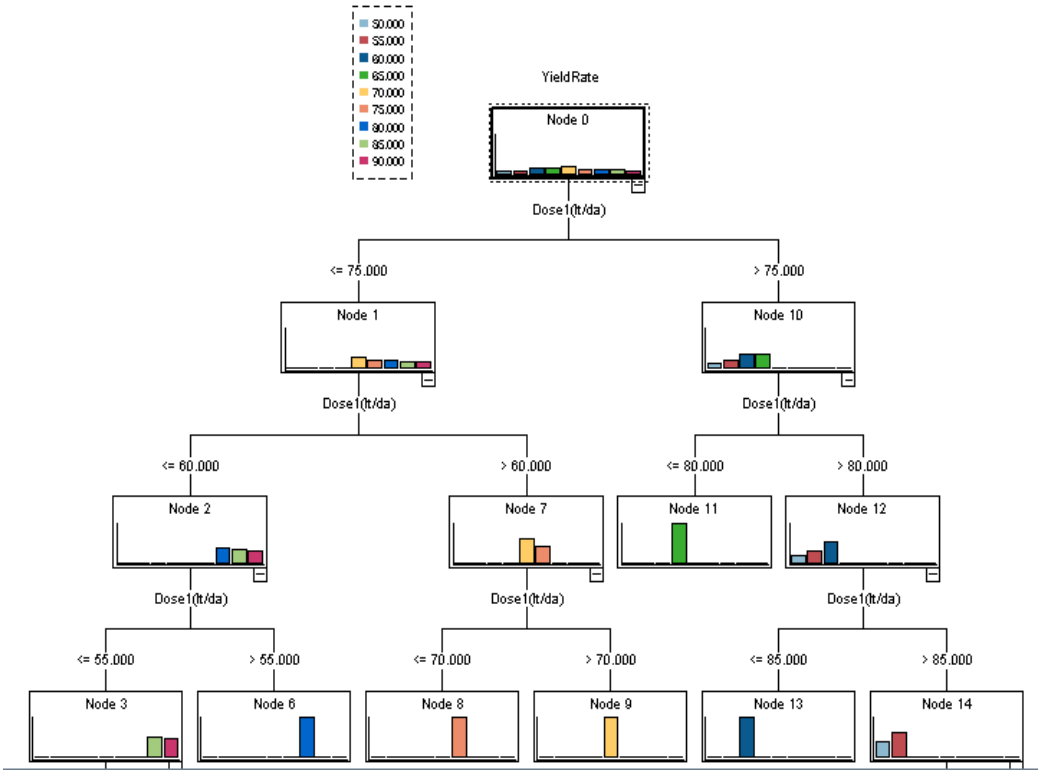


Figure 5.10 Structure of decision tree of the c5.0 algorithm

In conclusion, by utilizing the C5.0 algorithm, it can be observed which harmful pests lower the yield more and which agricultural chemicals is more effective. Accordingly, pesticide names and pest types are represented in numbers. Analysis is done after id equivalents are entered into the data type section. Since the process is requested to be determined as a rule, an if-then equation is formed, but also the same results can be obtained by creating a tree structure.

5.3.2 Implementation of C&RT Algorithm

Considering the accuracy rate of the algorithm, it is seen that the algorithm has a high accuracy ratio with 99.69 per cent in Figure 5.11.

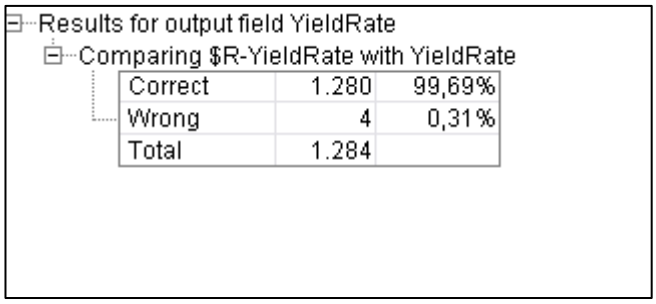


Figure 5.11 The rate of accuracy for the C&RT algorithm

It can be mentioned that the difference between this algorithm and C5.0 is that it does regression.

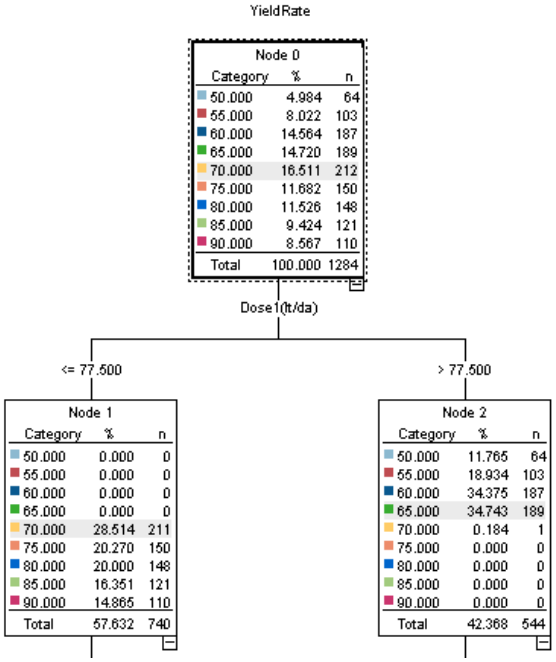


Figure 5.12 The first branch in C&RT algorithm for the variable of pesticide doses

Above, an analysis of the algorithm is formed in a tree structure as displayed in Figure 5.12. It is seen that the first value of the dose of pesticide started at a value of 77.500. If the dosage of the pesticide is less than 77.500, the result appears as 57.632 in total. It is seen that the yield of this pesticide is 70% in 211, 75% in 150, 80% in 148, 85% in 121 and 90% in 110. It is seen that pesticide has the highest yield of 70% in the products used with dozen. The tree continues to branch from the value of 72.5 and less. If the dose of pesticide is more than 77.500, the result is 42.368 in total. It can be observed that pesticide has the highest value in the yield with a rate of

65%. The yield rate is 50% in 64, 55% in 103, 60% in 187, 65% in 189. The tree continues to branch from the value of more than 82.500 and less than equal 82.500.

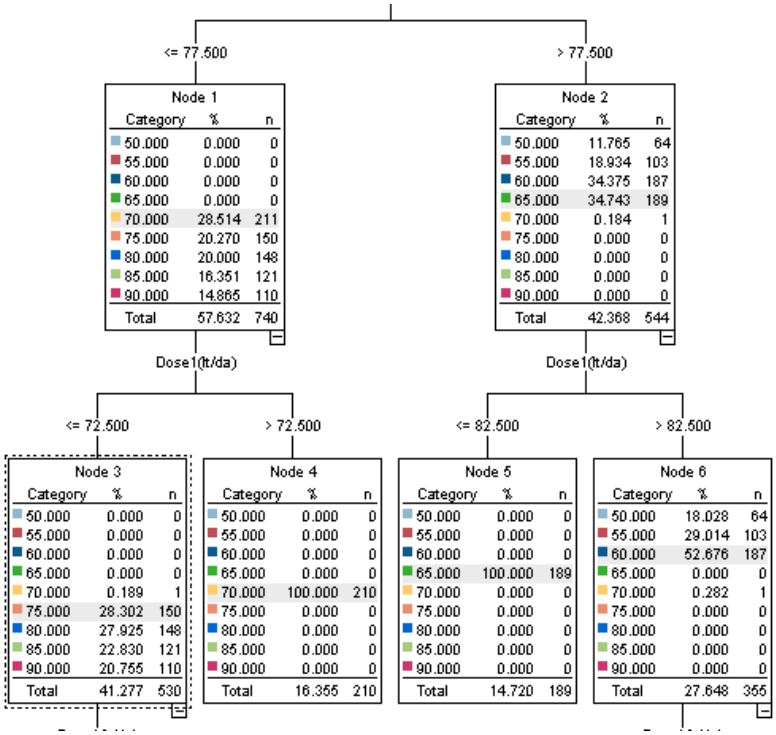


Figure 5.13 The second branch in C&RT algorithm for the variable of pesticide doses

For the second level branching, when the dosage of the pesticide is less than 72.500, it is seen in Figure 5.13 that the yield has the most share with 75%. The tree continues branching when the yield rate is less than or equal to 72.500. And the other part still proceeds branching on the condition that the value is more than 82.500.

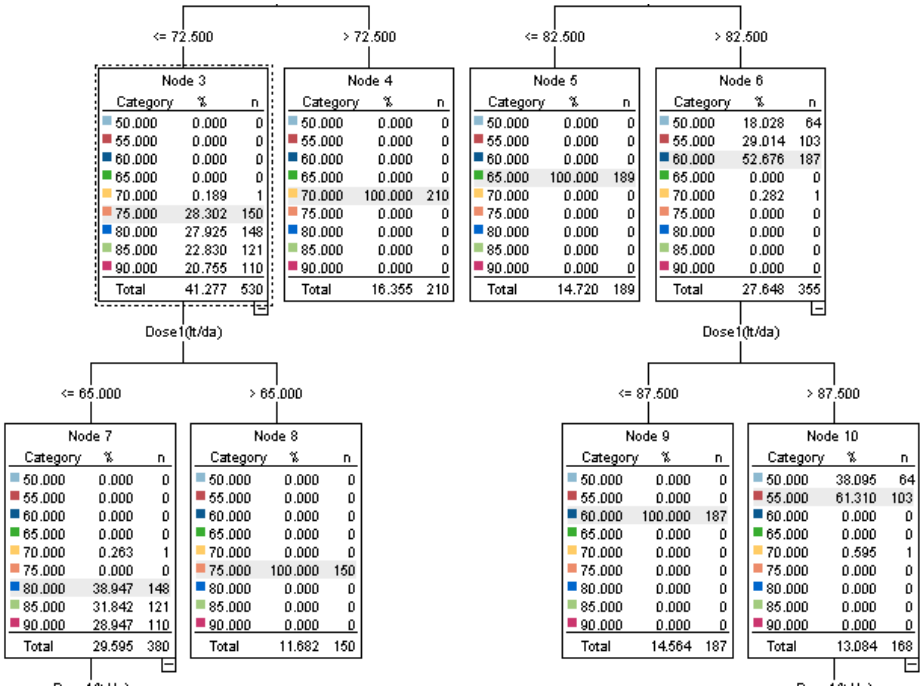


Figure 5.14 The third in C&RT algorithm for the variable of pesticide doses

When the dosage of the pesticide is equal or less than 65.000, the total result is 29.595. The yield rate has the most share with 80%. The yield rate is 85% in 121, and it is 90% in 110. The other part of the tree carries on branching with the value of more than 87.500 in Figure 5.14. It has a value of 13.084 in total. The yield rate is 55% in 103, and 50% in 64, 70% in 1.

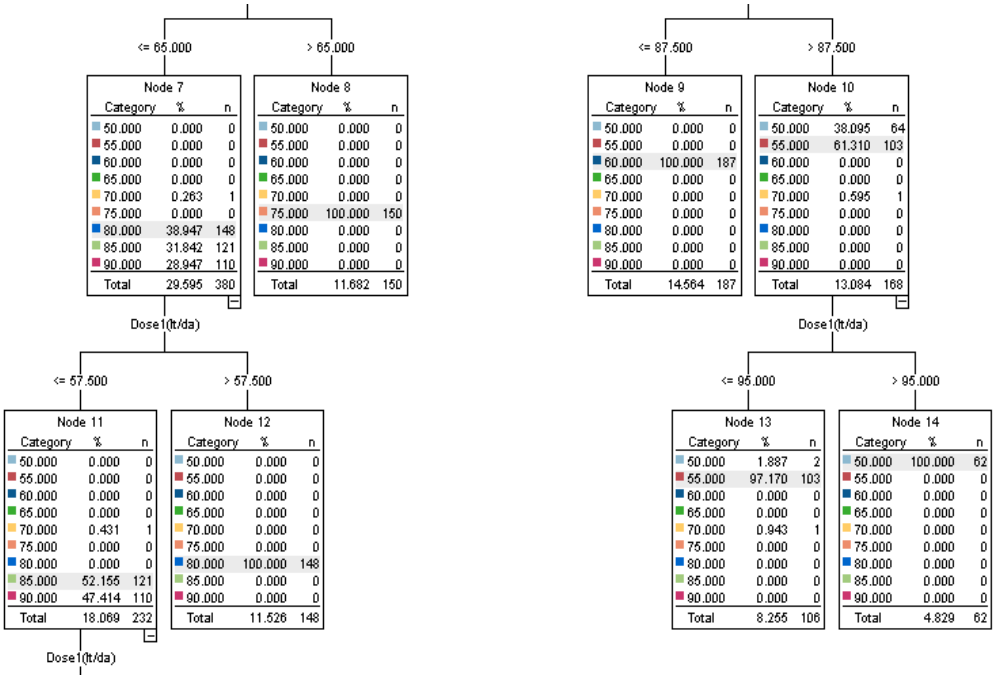


Figure 5.15 The fourth branch in C&RT algorithm for the variable of pesticide doses

The fourth branching continues with a value of 57.500 as shown in Figure 5.15. When the dosage of the pesticide is less than or equal to 57.500, it generates a total value of 18.069. It can be seen that the yield rate is 85% in 121, and 90% in 110. In the other part of the tree, when the pesticide dose is less than or equal to 95.000, the total value results in 8.255. As it is seen in the figure, the yield rate is 55% in 103, 50% in 2. The tree continues branching at the value of 52.500. When the dosage of the pesticide is more than 52.500, the total result is 9.502. The yield rate is 85% in 121, 70% in 1. If the dosage of the pesticide is less than or equal to 52.500, the total value becomes 8.567. Moreover, the yield rate is 90% in 110, which is indicated in the last tree class.

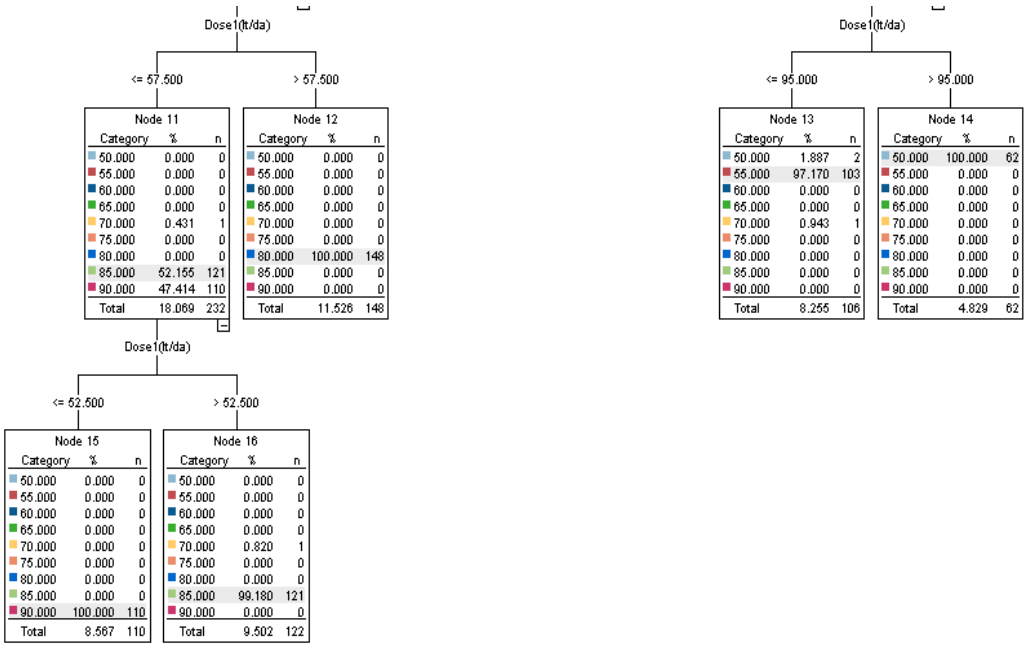


Figure 5.16 The last branch in C&RT algorithm for the variable of pesticide doses

Finally, as the tree structure demostrates in Figure 5.16, 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 case of a value less than this, it is found to have the highest rate with 90%. That is to say, the result of the analysis shows a reduced yield from the product due to overdose, however it also indicates that the yield will be higher if the recommended pesticide is used at the recommended dose.

5.3.3 Implementation of Chaid Algorithm

Considering the accuracy rate of the algorithm, it is seen that the algorithm has a high accuracy ratio with 99.69 per cent in Figure 5.17.

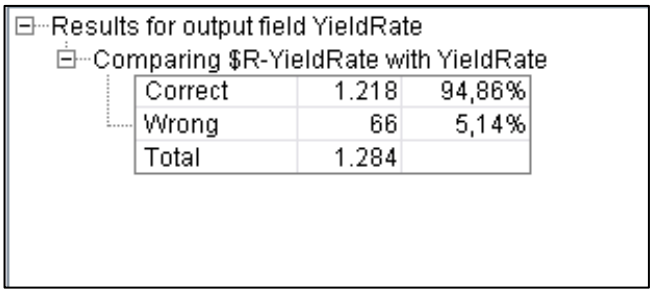


Figure 5.17 The rate of accuracy for the chaid algorithm

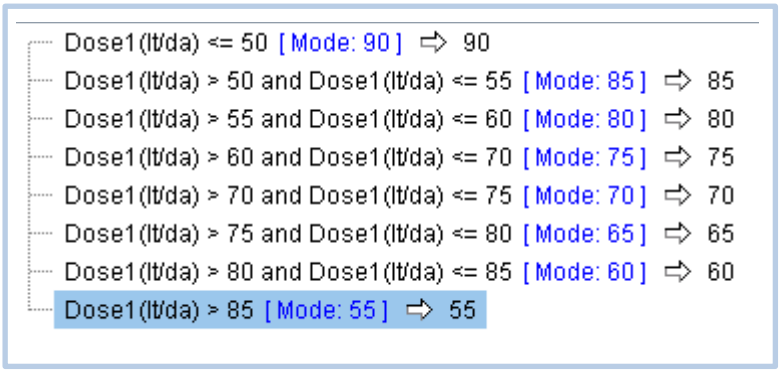


Figure 5.18 The rules for pesticide dose by using chaid algorithm

As shown above Figure 5.18, when the Chaid algorithm is applied and the target is designated as the yield rate, and after the agricultural pesticide is selected as the input, the algorithm generates some rules and reveals the results. To interpret this process, it can be uttered that the algorithm discloses eight main rules.

The first rule is that the dose of the pesticide is equal or less than 50, the yield rate obtained turns out to be 90%. However, when the dose is equal or less than 55 and more than 50, the yield rate rises to 85%. On the other hand, if the dose is equal or less than 60 and more than 55 , the yield rate obtained decreases to 80%. Yet, in a situation when the dose is determined as more than 60 and equal and less than 70, the yield rate obtained decreases to 75%. Under the condition that the dosage of the pesticide is equal or less than 75 and more than 70, this time 70% of yield rate is obtained. Again, when the dose is more than 75 and equal and less than 80, the yield rate becomes 65%. While the dose is more than 80 and equal and less than 85, the yield rate appears as 60%. However, if the dose is increased and becomes more than 85, a rate of 55% in yield is achieved.

For further understanding, the algorithm is also depicted as a tree structure which can be seen below Figure 5.19.

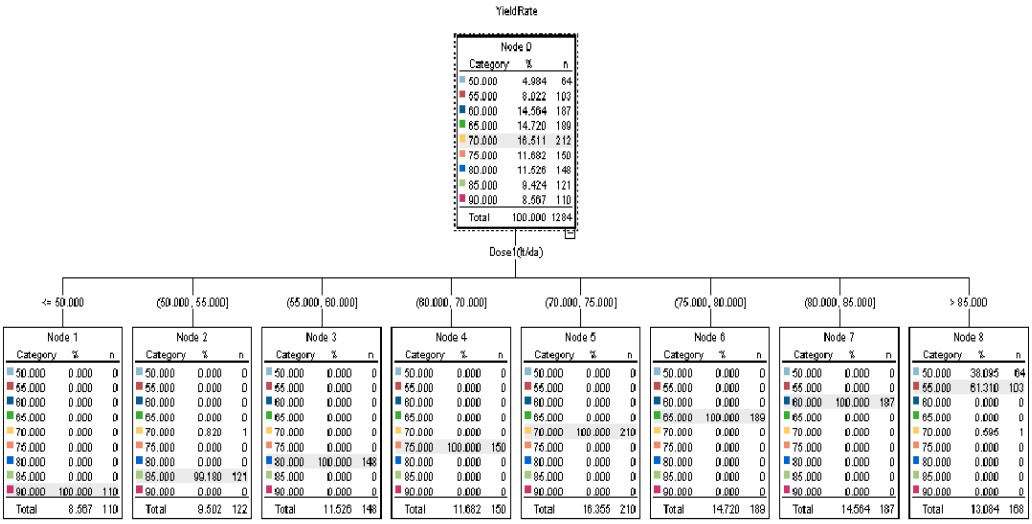


Figure 5.19 Structure of decision tree of the chaid algorithm

CHAPTER SIX
CONCLUSION

With recent developments, the increase in the amount of data has made essential the processing of data and the extraction of meaningful information. Data mining is used in many areas such as agriculture, education, marketing, banking and medicine.

Yield prediction is a very important agricultural problem. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmers experience on particular field and crop (Raghuveer, Yogesh & Shwetha, 2014).

In this study, the application of data mining in the field of agriculture has been scrutinized. Within the scope of the study, primarily, the data received from the Aydın Nazilli District Directorate of Agriculture was organized. Additionally, SQL Database has been created in aid of farmers. Thus, it is aimed to reduce the problems about keeping data in agriculture. Then SPSS Clementine software was used in order to analyze the data with the help of decision tree algorithm which is one of the classification methods in data mining. 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.

In this study, C5.0, C&RT and Chaid algorithms were used to create a model for cotton growers that will indicate the optimum amount of pesticides to apply for the maximum yield rate. Some meaningful results were observed when pesticide dose and yield rate were selected as the input and as the target variable, respectively. Once the rules generated by the C5.0 algorithm about the pesticide dose were examined, two rules arose, in which the dosage of the pesticide is more than 75, and 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. When the results obtained from Chaid algorithm were examined, 8 branch arose. It was observed that the dosage of pesticide is less than or equal to 55

the yield rate obtained 90%. While the dose is more than 85 the yield rate appears as 55%.

Taking these algorithms into consideration, it has been understood that the excessive use 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 considering its effects on human health, environment and yield rate obtained from product. In this regard, farmers should be educated and the use of pesticides should be minimized. .

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 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 thesis, we have shown how data mining can be successfully applied for this purpose. This study 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.

Literature studies also reveal the damage of agricultural chemicals in certain situations. Recent studies by agriculture researchers in Pakistan have shown that attempts of crop yield maximization through pro-pesticide state policies have led to a dangerously high pesticide usage. These studies have reported a negative correlation between pesticide usage and crop yield in Pakistan. Hence excessive use (or abuse) of pesticides is harming the farmers with adverse financial, environmental and social impacts (Umer & Nisar, 2003). As a result of application of the C&RT, Chaid and C5.0 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. 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.

As a future work, different and more efficient classification method can be implemented, better prediction of yield rate could be possible. Based on to this work decision support system can be developed for farmers and agricultural engineer.

REFERENCES

Abdullah, A., Brobst, S., Pervaiz, I., Umer, M., & Nisar, A. (2003). 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, 32,* 151-156.

Aktar, W., Sengupta, D., & Chowdhury, A. (2009). Impact of pesticides use in agriculture: their benefits and hazards. *Interdisc Toxicol, 2* (1), 1-12.

Aydın / Nazilli District Directorate of Agriculture, (n.d.). Cotton planted fields and used pesticides data.

Berry, M., & Linoff, G. (2004*). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management* (2nd. ed.). John Wiley & Sons.

Dağ, S., Akçay, T., Gündüz, A., Kantarcı, M., & Şişman, N. (2000). *Türkiye'de Tarım İlaçları Endüstrisi ve Geleceği.* Türkiye Ziraat Mühendisliği V. Teknik Kongresi, Ankara. 933-957.

Durdu, M. (2012). *Application of data mining in customer relationship management: Market basket analysis in a retailer store.* M.Sc. Thesis, Dokuz Eylül University, İzmir.

Freitas, A. A. (2009). *A survey of evolutionary algorithms for data mining and knowledge discovery*. Verlag NY: Springer.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). NJ: Prentice Hall.

Han, J., Kamber, M., & Pei, J. (2012a). *Data mining concepts and techniques.* (3rd ed.). USA: Elsevier.

Han, J., Kamber, M., & Pei, J. (2012b). *Data mining concepts and techniques.* (3rd ed.). USA: Elsevier.

Hui, S. C., & Jha, G. (1999). Data mining for customer service support. *Information & Management, 38* (2000), 1-13.

Işın, S., & Yıldırım, İ., (2006). Fruit-growers' perceptions on the harmful effects of pesticides and their reflection on practices: The case of Kemalpasa, Turkey. *Crop Protection, 26,* 917-922.

Kantarcı, M. (2007). *Globak BKÜ pazarı ve AR-GE.* Tarım ilaçları Kongre ve Sergisi, 25-26 Ekim2007, Ankara. TMMOB Kimya ve Ziraat Mühendisleri Odaları. Bildiri Kitabı, 13-23.

Küçükönder, H., Vursavuş, K., & Üçkardeş, F. (2015). K-star, rastgele orman ve karar ağacı (C4.5) sınıflandırma algoritması ile domatesin renk olgunluğu üzerinde bazı mekanik özelliklerin etkisinin belirlenmesi. *Türk Tarım- Gıda Bilim ve Teknoloji Dergisi, 3* (5), 300-306.

Microsoft SQL Server Management Studio (n.d.). Retrieved March 12, 2017, from https://www.tutorialspoint.com/ms_sql_server/ms_sql_server_quick_guide.htm

Oakley, E., Zhang, M., & Miller, R. (2006*).* Mining pesticide use data to identify best management practices. *Renewable Agriculture and Food Systems, 22* (4), 260–270.

Olson, David L., & Delen, D. (2008). *Advanced data mining techniques.* Verlag Berlin Heidelberg : Springer.

Urtubia, A., Pe´rez-Co J. R., Soto, A., & Pszczo´lkowski, P. (2006). Using data mining techniques to predict industrial wine problem fermantations. *Food Control. 18* (12), 1512-1517.

Tarım, (2017). *Pesticides use in Turkey*. Retrieved April 24, 2017, from http://www.tarim.gov.tr/Konular/Bitki-Sagligi-Hizmetleri/Zirai-Mucadele

Raghuveer, K., Yogesh, M. J., & Shwetha, S. (2014). Data mining in agriculture: A review. *AEIJMR, 2*, 2348 – 6724.

Ramesh, D., & Vardhan, B. V. (2015). Analysis of crop yield predictıon using data mining techniques. *International Journal of Research in Engineering and Technology*, *4*, 470-473.

Sorhocam, (2017). *Pests and diseases of tomatoes and the active ingredients.* Retrieved July 20, 2017, from https://www.sorhocam.com/konu.asp?sid=986&zirai-mucadelede-kullanilan-etken-maddeler-ilaclar.html

Sorhocam, (2017). *Pests and diseases of cotton and the active ingredients*. Retrieved July 20, 2017, from https://www.sorhocam.com/konu.asp?sid=986&zirai-mucadelede-kullanilan-etken-maddeler-ilaclar.html

Sorhocam, (2017). *Pests and diseases of olives and the active ingredients.* Retrieved July 20, 2017, from https://www.sorhocam.com/konu.asp?sid=986&zirai-mucadelede-kullanilan-etken-maddeler-ilaclar.html

SPSS Inc. Clementine 12.0 User's Guide, 2002. (2017). Retrieved April 8, 2017, from https://tr.scribd.com/document/16066537/Text-Mining-for-Clementine-12-0-User-s-Guide

Tan, P., Steinbach, M., & Kumar, V. (2004). *Introduction to data mining*. USA: Pearson.

Yethıraj, N. G. (2012). Applying data mining techniques in the field of agriculture and allied sciences. *International Journal of Business Intelligents*, *1*, 2278-2400.