DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

AGRICULTURAL DECISION SUPPORT SYSTEM USING DATA MINING FOR FARMERS

by

Büşra BOSTANCI

August, 2018

İZMİR

AGRICULTURAL DECISION SUPPORT SYSTEM USING DATA MINING FOR FARMERS

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
Büşra BOSTANCI

August, 2018
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"AGRICULTURAL DECISION SUPPORT SYSTEM USING DATA MINING FOR FARMERS "** completed by **BÜŞRA BOSTANCI** 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

....

(Jury Member) (Jury Member)

Prof. Dr. Latif SALUM

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank to my thesis advisor Asst. Prof. Dr. Canan Eren ATAY, my research supervisor, for her help, useful suggestions and guidance during this thesis study. I thank to the Izmir Menemen Provincial Directorate of Agriculture for sharing this agricultural dataset with us. I also thank to my friend, Zehra BURDUR KURT who shared her knowledge and encourage me throughout the study. Finally, I like to show my gratitude to thank Beyza BOSTANCI, Zekiye BOSTANCI, Hasan Basri BOSTANCI and also my friends and colleagues, for their constant support, unconditional love and encouragement to boost motivation throughout the process of writing my thesis.

Büşra BOSTANCI

AGRICULTURAL DECISION SUPPORT SYSTEM USING DATA MINING FOR FARMERS

ABSTRACT

The estimation of agricultural yield is a challenging and essential task for every farmer. Since the very old times, agriculture has always been the most important means of livelihood both in Turkey and all around the world. There are many factors that directly affect the efficiency in agriculture such as climatic features, use of water resources, proper and timely use of pesticides and fertilizers. Computer-based systems are needed to transform agriculture data into tangible information. Data mining involves certain methods of obtaining or inferring meaningful and otherwise-unknown information from the data. With the increasing significance of precision agricultural practices, farmers have become inclined to be engaged in a more conscious strategy of agriculture.

In this study, barley crop data received from İzmir Menemen Provincial Directorate of Agriculture was carefully organized and evaluated with the classification algorithms in the SPSS Clementine software. CHAID and CR&T algorithms were employed and major factors that affect crop yield were defined. Based on these, a decision support system has been developed for farmers to forecast both harvest season and crop yield.

Keywords: Decision Support System, data mining, decision tree, CHAID algorithm, CR&T algorithm, agriculture

VERİ MADENCİLİĞİ İLE ÇİFTÇİLER İÇİN KARAR DESTEK SİSTEMİ GELİŞTİRİLMESİ

ÖZ

Tarımsal verimin tahmini, her çiftçi için zorlu ve önemli bir görevdir. Tarım geçmişten günümüze kadar hem Türkiye hem de bütün dünyada birçok insan için en önemli geçim kaynağı olmuştur. İklimsel özellikler, su kaynaklarının kullanımı, tarımsal ilaçlar ve gübrelerin doğru ve zamanında kullanılması gibi tarımın etkinliğini doğrudan etkileyen birçok faktör vardır. Tarımsal verileri anlamlı bilgiye dönüştürme sürecinde bilgisayar destekli sistemlere ihtiyaç vardır. Veri madenciliği, verilerden anlamlı ve başka türlü bilinmeyen bilgiler elde etmek veya çıkarmak için belirli yöntemler içerir. Hassas tarım uygulamalarının artan önemi ile çiftçiler daha bilinçli bir tarım stratejisi ile meşgul olmaya eğimli hale geldi.

Bu çalışmada, İzmir Menemen Tarım il müdürlüğünden alınan Arpa ürünün ekim verileri dikkatle düzenlenmiş ve SPSS Clementine yazılımındaki sınıflandırma algoritmaları ile değerlendirilmiştir. CHAID ve CR & T algoritmaları kullanılmış ve ürün verimini etkileyen ana faktörler tanımlanmıştır. Buna dayanarak, çiftçilerin hem hasat mevsimini hem de ürün verimini tahmin etmeleri için bir karar destek sistemi geliştirilmiştir.

Anahtar Kelimeler: Karar Destek Sistemleri, veri madenciliği, karar ağacı, CHAID algoritması, CR&T algoritması, tarım

CONTENTS

Page

LIST OF FIGURES

Page

CHAPTER ONE
INTRODUCTION

Decision support systems are computer-based systems that help people make decisions and reduce the length of deciding period. Today, there are studies focusing on decision support systems in numerous fields. This can be caused by the huge volume of data available in various fields, problems that cannot be resolved, as well as by the importance of the predictions to be made in advance.

Today, studies in the field of agriculture have gained higher significance. Rapid growth that is prevalent in today's world population can be given as a reason for this situation. Agricultural resources should be used properly to prevent the famine that might be caused by the rapid population growth and uneven consumption (Geronimo et al., 2014). Making use of the agricultural data in a more conscious way:

- Crop yield is boosted,
- Potential hazards are predicted, and costs are reduced,
- Resources such as water, fertilizers and pesticides are used more consciously,
- Harvest date and approximate amount of yield are predicted in advance.

The field of agriculture possess a huge volume of data to accomplish all these goals. This data can be converted into information by utilizing information technology (Laurance, Sayer, Cassman, 2014). This in turn allows farmers to have easier control of irrigation, pesticide and fertilizer application dates, as well as to predict crop yield and harvest date in advance.

Agricultural data is quite complex. To analyze the data, interdisciplinary knowledge that includes math, statistics, crop agronomy, computer hardware and software is needed (Masters et al., 2013). To analyze the data correctly, all the data should be converted into a hierarchical table in a database. Data mining techniques are then needed for interpreting and turning this raw data into useful information (Fayyad, Piatesky–Shapiro & Smyth, 1996). Following the conversion of the data into useful information by making use of data mining techniques, computer-assisted decision support systems are developed. This allows farmers to make right decisions and predictions in advance in agriculture.

It has been observed that the information regarding the crops that farmers planted in the past years and the unfortunate conditions (such as weather, pests, natural disasters) had an influence on farmers' decision-making process. Farmers thus cannot make decisions because of possible natural events, even when needed. Though decision support systems have been being used in many different fields extensively, their use in agriculture has just started to become a common practice (Taechatanasat & Armstrong, 2013). Created by making use of data mining in agriculture, these systems function as information system applications that allow farmers to make the best decisions.

In line with the purpose of this thesis, a decision support system has been developed for farmers' use. To develop this decision support system, data analyses were conducted primarily based on data mining techniques.

The data used in this thesis were made available by the İzmir Menemen Provincial Directorate of Agriculture. The data retrieved on an annual basis relates to barley product, and encompasses various features pertaining to the product. In the light of this data, the weather data made available by the Turkish State Meteorological Service was combined on Excel datasheet for analysis. This data includes 1077 records. Various data mining analyses were applied on SPSS Clementine in accordance with the data used in the study. Together with the results obtained from this analysis, a windows forms applications in ASP.NET was designed on Visual Studio 2013. For this decision support system, Microsoft SQL Server 2012 was used as the database to record the information concerning farmers and the application. Finally, a decision support system was developed for farmers to track crop yield and harvest season in advance.

The chapters of this thesis are organized as follows. An introduction to decision support systems in agriculture and the purpose of the thesis are described in Chapter 1. Chapter 2 presents a literature review about agriculture focusing on decision support systems and data mining. Chapter 3 has summary of the basics of the decision support system and data mining. The technologies used in this study are explained in Chapter 4. Chapter 5 describes the data sets, data mining applications and the development of a prototype decision support application in this study. Finally, the study results are presented and ideas for future work are presented in Chapter 6.

CHAPTER TWO
LITERATURE REVIEW / RELATED WORKS

Recently, there has been a worldwide increase in agricultural data analysis and in the number of the studies focusing on decision support systems. The following is a summary of the relevant studies.

2.1 Data Mining Studies in the Field of Agriculture

S. Hari Ganesh & Jayasudha (2015) state that data mining is very popular in the field of agriculture, but they point out to the data mining in agricultural soil datasets. Because they state that this area is a new research field. They provided web-based answers for the soil testing research facilities. Additionally, they have provided farmers with free messages to give them expert advice. This advice is related to soil testing code and the fertilizers needed for their crop. This study is based on classification, and contains what is present in soil. The study generates a report according to the results retrieved (Ganesh & Jayasudha, 2015).

S. Veenadhari & B. Mishra & CD Singh (2011) point out that soybean productivity is influenced by climatic parameters. They did an analysis based on Bayesian classification for the decision tree induction technique. They found an existing correlation between climate data and soybean crop yield and then concluded that soybean fertilization confirms the effectiveness and correctness of these variables on crop productivity (Veenadhari, Mishra & Singh , 2011).

This article was created out of a study by Luis R.Salado and NavarroaThomas R.Sinclair, and based on the results of the study that was carried out in Argentina. The objective of this work was to investigate through simulations, potential differences in temporal soil water status among rotations over five years. Simulated soybean yields were poorly correlated with both the amount of soil water at sowing and the rainfall during the cropping period. These results highlight the importance of temporal distribution of rainfall on final yield. (Salado & Sinclair, 2009)

R. Benedetti A & R. Catenaro A & F. Piersimoni B focused on the data regarding the impact of rainfall on crop yield. They observed significant and positive correlation between yield and rainfall. This work was done for groundnut, pearl miIlet and sorghum products (Benedetti, Catenaro & Piersimoni , 2002).

In their study, A. Urtubia & J.R. Perez-Correa & A. Soto & P. Pszczolkowski intends to find early signs of undesirable fermentation behavior for winemakers. They used a database that covers periodic measurements of 29 components that include sugar, alcohols, organic acids and amino acids. They found an association between problematic fermentations and the specific patterns found by the datamining tools (Urtubia, Perez-Correa, Soto & Pszczolkowski, 2007).

2.2 Decision Support System Studies in the Field of Agriculture

P. Taechatanasat & L.Armstrong points out that special data is required for creating an agricultural decision support system. This paper presents an overview of the agricultural data required for decision making in agriculture. In addition, several important systems are designed for collecting data, and they give ideas for development (Taechatanasat & Armstrong, 2013).

Cornell University, the University of Wisconsin, and the USDA-ARS Dairy Forage Research Center initiated a joint project in January 2002 to study nutrient management tools, research applications, and educational efforts for dairy farms in New York and Wisconsin(Rotz, Muck & Rasmussen, 2004). They developed new software tools. This tools to address a range of nutrient management decisions, from optimizing dairy herd and feed management, matching soil and cropping systems, and manure and fertilizer management, to assisting land use planners in issues related to dairy herd expansion.

Akın, Yıldırım & Çakan give brief that why the decision support systems are not used by poor farmers, contrary to large-scale farms. Sustainability in agricultural production and animal husbandry have been emphasized. The goal of the project is to design a system that will boost the efficiency of poor farmers by predicting the risks, and to develope an easy-to-use computer system to this end (Akın, Yıldırım & Çakan, 2014).

This article was created out of a study by Ma Qingyuan, Chen Zhenghua, Zhang Chao, Yang Zhen, and based on the results of the study that was carried out in Nong Gong Shang Modern Farm, Shanghai, China between 2003 and 2005. They highlight that the system that allows them to make decisions will also help the farmers. The purpose of the system here is to answer questions such as when, where and how much irrigation, fertilizer and pesticide should be applied on each plot (Qingyuan, Zhenghua, Chao & Zhen, 2007).

Zhiqing Zhu, Rongmei Zhang, Jieli Sun points out that progress in agriculture will only be through sensitive farming practices. They developed the GIS-based agriculture expert system. This study also covers the data mining and Web technologies that will apply to the agricultural expert decision system (Zhu, Zhang & Sun, 2009).

Kornkanok Phoksawat, Massudi Mahmuddin proposes three points to be covered in researches (Phoksawat & Mahmuddin, 2016). These research topics are;

1. To determine the knowledge domain of intercropping planning system,

2. To develop an ontology-based knowledge prototype aimed at addressing the DSS of the planning intercropping system,

3. To design an optimization model used for decision-making for the intercropping system. They developed a decision support system that includes the ontology-based knowledge and multi-objective optimization model. This system has been developed to show the maximum income and to minimize the cost for cultivating the plants.

CHAPTER THREE
DATA MINING/DECISION SUPPORT SYSTEMS

This section describes data mining, decision support systems and some important concepts.

3.1 Knowledge, Data and Information

Knowledge take place between more raw forms of meaning such as data and information and a more complex form of meaning such as wisdom. Knowledge is formed by the use of information combined with talents, skills, ideas, interpretations, perceptions and motivations of the person. In order to better define the knowledge, it is necessary to establish the differences between the meaning of some similar concepts and the meaning of knowledge.

The data is a raw fact. The measurement of the data is obtained by means of counting, testing, observation or research. A single data does not have a function and a meaning. The data become meaningful when they are grouped, sorted and summarized or processed by hand or by computer and converted into information. In this way, they have the power to explain the context they belong to.

Information is data that has been interpreted so that it has meaning for the user. In other saying, information is data that has been identified with relational links. The quality of the information depends on the data that is clear and meaningful, that it contains correct values and presentation in an understandable way.

Knowledge is a combination of information, experience and insight that may benefit the individual or the organization. In the personal sense, knowledge is edited information. And it usually consists of a combination of experiments and experiences.

At the highest level of DIKW (Data-Information-Knowledge-Wisdom) pyramid, wisdom stand for the events that include all information for us to discern what is right or wrong. In order to help us to correctly understand the causes of events and to help me choose the right or the most beautiful, we can make the right decisions at the stages of comprehension and reasoning, when knowledge is systematically processed, observed and experimentally reshaped.

To summarize, there are records of data processing. Information is collective of processed meaningful data. The information answers "who", "what", "where", "when" questions. Knowledge is the application of data and information. Knowledge answers "how" question. Wisdom is knowing what competencies will be used for what purpose.

The graph below reflects the learning journey whereby we progressively transform the raw, unfiltered facts and symbols into information, knowledge, and eventually into intelligence and wisdom.

The DIKW Pyramid in in Figure3.1 shows the visualized version of this hierarchy.

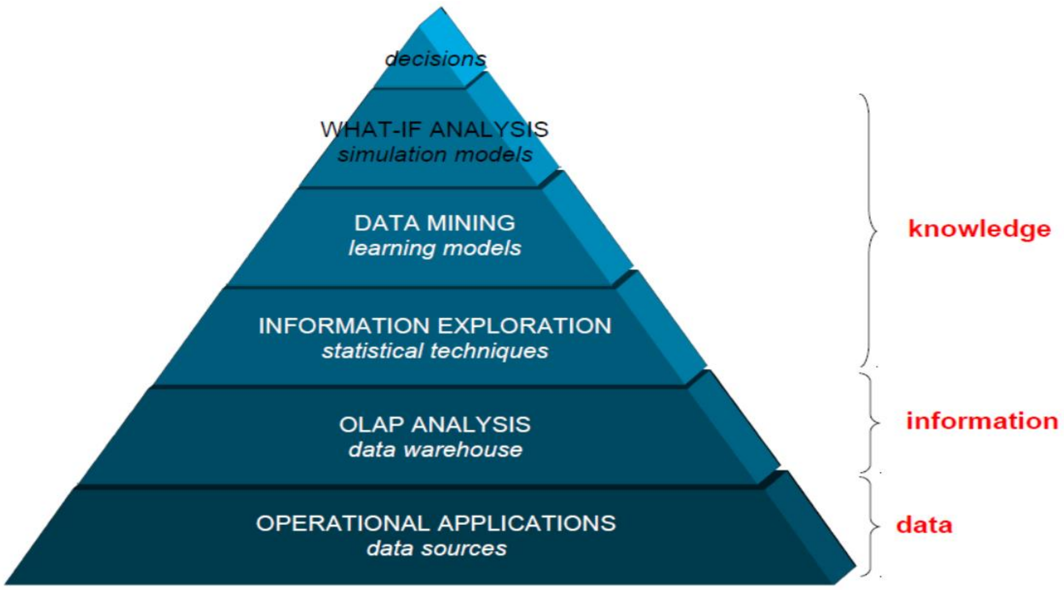


Figure 3.1 Summarize of data processing

3.2 Decision/ Decision Making

Decision is intentional and reflective choice in response to perceived needs (Kleindorfer et al., 1993). Decision-making is decision maker's choice of one alternative or a subset of alternatives among all possible alternatives with respect to her/his goal or goals (Evren & Ülengin, 1992).

The steps of scientific decision-making are as follows.

Identification of the problem: First, it is checked whether there is a problem that needs to be solved. If there is a problem, its dimensions should be addressed. After revealing the problem, we look at where this problem originated. It is necessary to investigate how to remove the resources that make up the problem.

Designing Solutions: In order for a problem to be a problem that requires a decision, it must be a solution in some aspects. If there is only one solution to a problem, there is nothing to decide. It tries to increase the solution preferences by going through different problems in a certain way.

Identification of Possible Consequences: After applying the solution paths one by one, the results that can be encountered must be estimated. Taking all factors into consideration, a result close to the truth can be obtained.

Comparison of the results: The solution with the most additive solution is to be chosen. By using the resources in the best way, the solution path that leads to the least damage must be selected.

Standardize Successful Improvements: Once you have successfully solved the problem, standardize the steps you have taken and the solutions you have received, as well as the process you have reached the solution. The process is repeatable and try to improve it constantly.

3.3 Data Mining

Nowadays, the need to convert data into information has emerged because of the rapid growth of the data in every field. For this reason, the definition of Data Mining has emerged. Data Mining is a search for relationships and rules that will allow us to make predictions about the future from a large amount of data. (Knowledge Discovery in Databases) There are many good definitions of the term “Data Mining” but the core concept behind all of them is the same. IBM defines Data Mining as, "the process of extracting previously unknown, valid and actionable information from large databases and then using the information to make crucial business decisions" (Piatetsky-Shapiro et al., 1996).

As it is understood from these definitions, it is possible through data mining studies to reveal the relations between the data and make estimations when necessary. This means that data mining is a process in which all data generated in an institution is evaluated as the process of revealing hidden or intellectual information that may or may not exist in the future using specific methods. From this point of view, it can be said that data mining has an important place for decision support systems of institutions.

Data mining studies include a number of technical approaches such as classification, relationship building, clustering, regression, data summarization, analysis of changes, determination of deviations.

If we group the processes of data mining, we can distinguish three important groups as the preparation of the data first, then the application of the data mining applications and finally the analysis of the data.

The stage of preparation of the data: The phase in which the data is selected by the specialists of the subject, cleaned and finally pre-processed. This is the longest period in the data mining process.

Data mining algorithms phase: The phase in which a given data mining algorithm is used and all valuable information particles are defined using the prepared data.

Data analysis phase: The stage in which data mining results are evaluated to understand the discovered information, if any, by comparing the facts found with data mining algorithms.

3.3.1 History of Data Mining

The initial stage of data mining is up to the first digital computer ENIAC (Electrical Numerical Integrator And Calculator). The efficient use of computers begins with the storage of data. Computers were originally developed to do complex calculations and have begun to be used for data storage operations, as required by users' needs.

Due to this reason, the databases go out to sea. With the increase of the databases, it was necessary to expand the hardware environment in which these data would be stored. This way, the concept of data warehouse has emerged. The data was kept in physical drives because they wanted to hide for a long time. Along with this process, the organization, organization and management of the growing databases began to become more difficult. In this phase, the concept of data modeling emerged.

The first models developed are Hierarchical and Network data models. Hierarchical data models have tree structure. It has a root in its base, which is a data model that has 1 node at the root and n nodes at the bottom. Network data bodellites are such that any node can interfere with another node. The type of record is the type of data, the types of links, and the types of records that the entity determines the relationships of the links. It was not a matter of establishing a multi-relationship. In the hierarchical data model this was even more limited. Therefore, they cannot 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. Relational data model is the most frequently used one today. Object-oriented data models are still in improvement.

First, the concept of data mining came out in the 1960s. By resolving the data analysis problems, it became clear that it was possible to reach the desired data. This was done first with data search and data capture names.

In the 1990s, the name of Data Mining was introduced by computer engineers. The purpose of this name change was to put an emphasize on the evaluation of the data by algorithmic computer modules rather than traditional statistical methods. After this stage, scientists began 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, and research. With the rapid growth of the computer, it is possible to do statistical surveys that were not possible before. After these works, the statistics started to be considered as a partner with data mining. Data mining and statistics are commonly used in the process of extracting and analyzing the information from a lot of data stacks and preparing the data for use.

3.3.2 Data Mining Application Areas

Data mining can find the application area in any environment where the data is generated intensively. Some application areas can be summarized as follows.

Health: One of the most widely used application areas of data mining is the field of medicine and health. There is a wide range of applications such as the diagnosis of a heart attack, preliminary diagnosis for cancer, determination of various risks and priorities by using health data.

Banking and Finance: In banking sector, data mining is used in areas such as credit and credit card fraud estimates, risk appraisal, and profit analysis.

Internet: The online data is growing. Internet mining can be described as the discovery of useful information on the internet in short.

Shopping: The most commonly used data mining approach in this area is basket analysis. The purpose of basket analysis is to find relationships between products. Knowing this relationship is used to increase the profit in business.

Social Network: Nowadays, with the spread of the internet, the applications that are being used for communication and entertainment have increased. These applications, which can be defined as social media, have given people easy access to a great deal of information online. By analyzing the data in this area, useful information about people can be obtained.

3.3.3 Data Mining Steps

Certain processes must be applied to the data held in the databases in order for the operations of data mining to be carried out. A successful data-mining project consists of specific steps. These steps, shown in Figure 3.2, are briefly explained below.

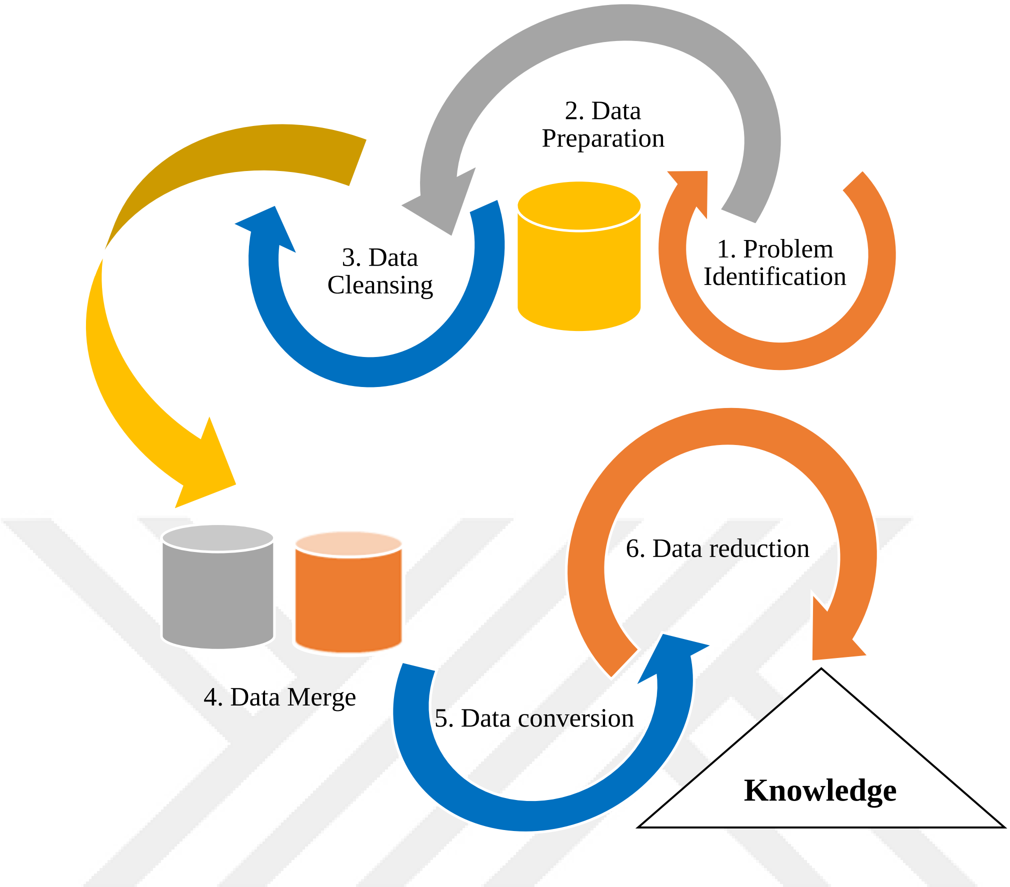


Figure 3.2 Data Mining Steps

Problem identification: One of the most important factors in data mining is determining the purpose of the project. The usefulness of the results of the problem needs to be analyzed.

Data preparation: This is one of the most time-consuming components of the data mining process. At this stage, information generated in information systems should be analyzed well and related to the problem. It is an important step for the analyst to measure the quality of the data at this stage. This step is necessary for the coupling of a large amount of data to a single database.

Data preparation includes the following steps in the data preprocessing process.

Data cleansing: Data cleansing is important for data mining to be successful. Errors must be corrected to produce useful and accurate results on the data. The missing information should not be taken into account, and should be deleted.

Data Merge: Data mining analysis needs a lot of data to get good results. Therefore, we must combine data appropriately by taking advantage of many data sources.

Data conversion: In the data used, some attribute types may not be suitable for the algorithm to be applied. 'Data conversion' process is performed in order to adapt the qualities to the algorithm. Normalization or qualification can be done on it.

Data reduction: Data mining will cause unnecessary operations to be performed if the data to be applied is too large and the result will contain data that will not be affected. This type of data needs to be cleaned. In this technique, data fusion, data cube, data compression, size reduction methods are available.

3.3.4 Data Mining Techniques

There are several important data mining techniques. The most important and most used data mining techniques are: association, classification, cluster, prediction and sequential patterns etc..

3.3.4.1 Association

Association is one of the best-known data mining technique. The rules of association are uncontrolled data mining. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. As an example, this technique is used in an analysis of the market basket, which is often used to identify products that customers buy together. In the direction of these data, they can earn more and sell more products by creating new campaigns.

The association rules are used to identify associations that occur concurrently, as shown in the examples below.

- When percentage of customers buy beer 75%, potato chips are also taken,
- When percentage of customers who buy low-fat cheese and lean yogurt buy 85%, diet milk also taken.

Figure 3.3 shows a Market basket analysis. This is a perfect example for illustrating association rule mining (Dhanabhakyam & Punithavalli, 2011).

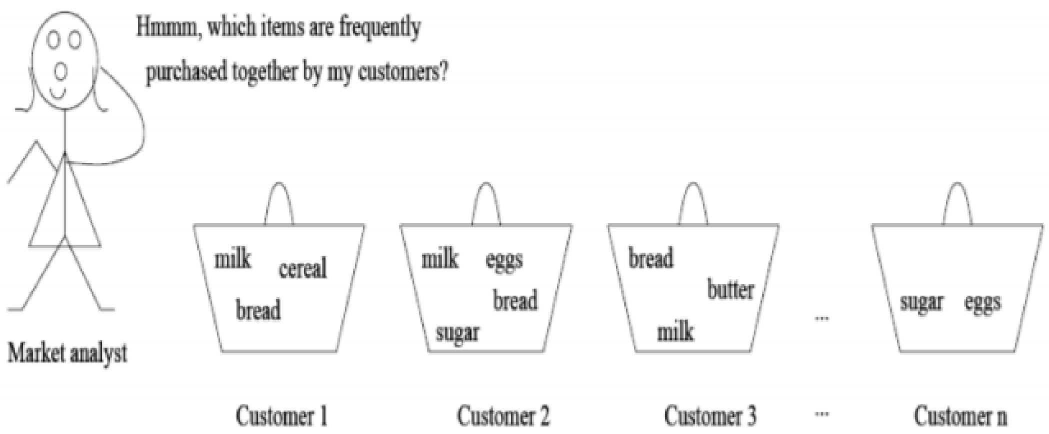


Figure 3.3 Example for illustrating association rule mining (Dhanabhakyam & Punithavalli, 2011)

Types of association rules: Different types of association rules are based on:

- Types of values handled
    - Boolean association rules
    - Quantitative association rules

- Levels of abstraction involved
    - Single-level association rules
    - Multilevel association rules

- Dimensions of data involved
    - Single-dimensional association rules
    - Multidimensional association rules

3.3.4.2 Classification

Classification is a predefined class assignment of an object by examination of its qualities. Class characteristics must be well defined. As the results are known in advance, they enter the classroom supervised learning group. For Example, teachers classify students' grades as A, B or C. Figure 3.4 shows a Classification analysis.

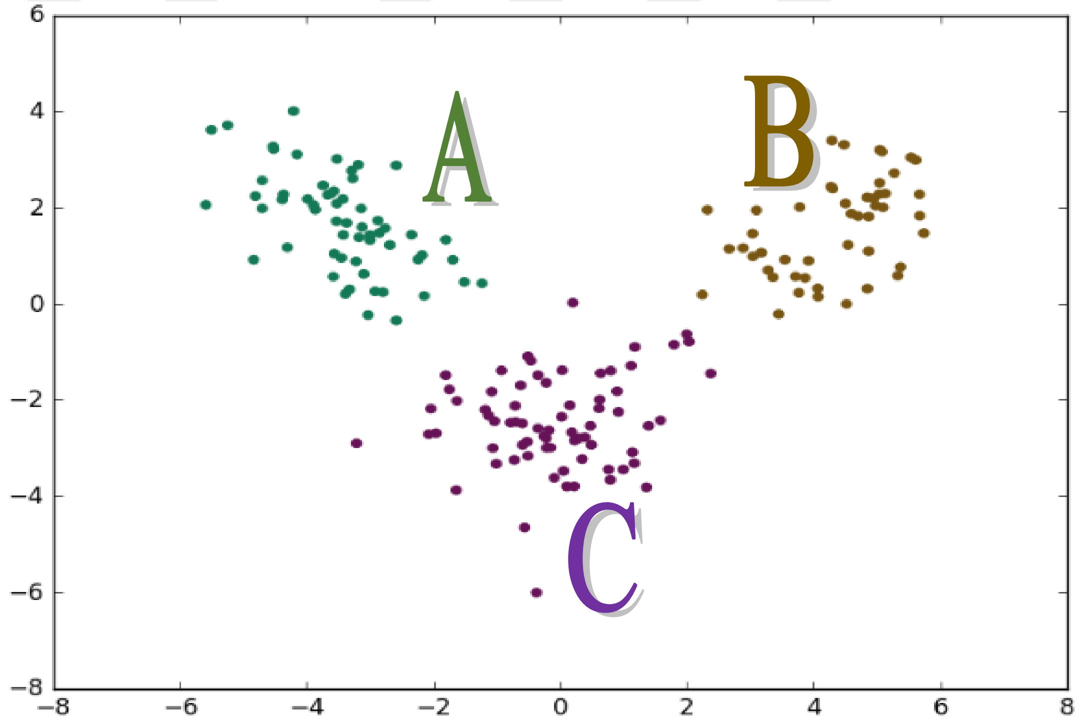


Figure 3.4 Example for illustrating classification

Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application like "given all past records of employees who left the company, predict which current employees will probably be leaving in the future." In this case, we divide the employee's records into two groups, which are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group. (Raval, 2012)

3.3.4.2.1 Decision Trees. Decision trees were first proposed by Bierman and Friedman in 1973. This technique is based on when there is an independent variable with the strongest association, the dataset is based on the logic of dividing it into two. In this way, a tree structure is created by continuing until the divisions are completed.

A decision tree is a structure that is used by dividing large quantities of records into very small groups of records by applying simple decision-making steps. With each successful partitioning process, the members of the result groups become much more similar to one another. In many classification problems where large databases are used and in decision trees containing complex or inaccurate information, decision trees are a useful solution. In figure 3.5 shows a sample decision trees.

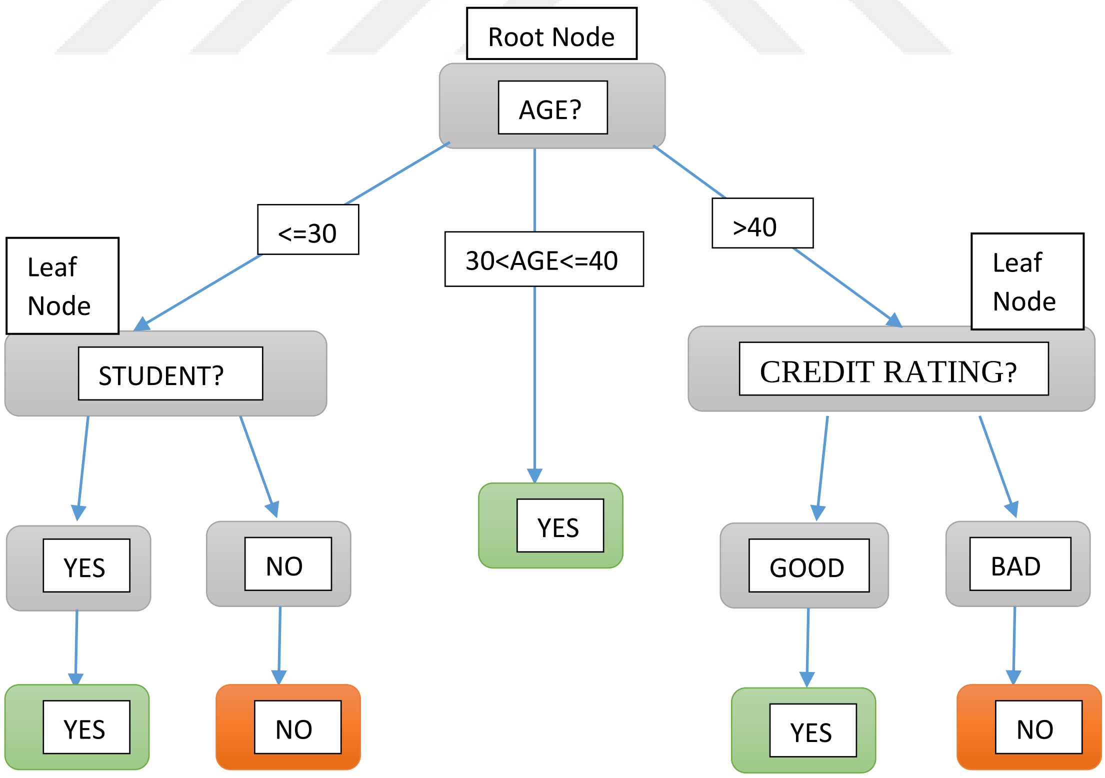


Figure 3.5 Sample decision tree diagram

Important algorithms used in Decision Trees are:

ID3: First developed by J. Ross at Sydney University, the ID3 algorithm uses entropy to find the most discriminating variable when classifying. Entropy has values between 0 and 1. The ID3 algorithm is based on 3 steps:

- Entropy (dispersion) values are calculated by taking into account properties that are not yet included in the syllabus.
- The entropy is sorted according to the values and the lowest value property is selected from among them.
- The decision of the selected feature is added to the decision.

CART (C & RT)(Classification and Regression Tree): The C & RT technique utilizes entropy like the ID3 algorithm to best distinguish decision trees. It is based on the principle of separating two trees from each decision node. In the CART algorithm, partitioning is performed by applying a certain criterion in a node. This takes into account the values for which all the qualities are present, and after all matches two divisions are obtained. Selection is performed on these divisions. Figure 3.6 shows a simple example for illustrating classification and regression tree analysis.

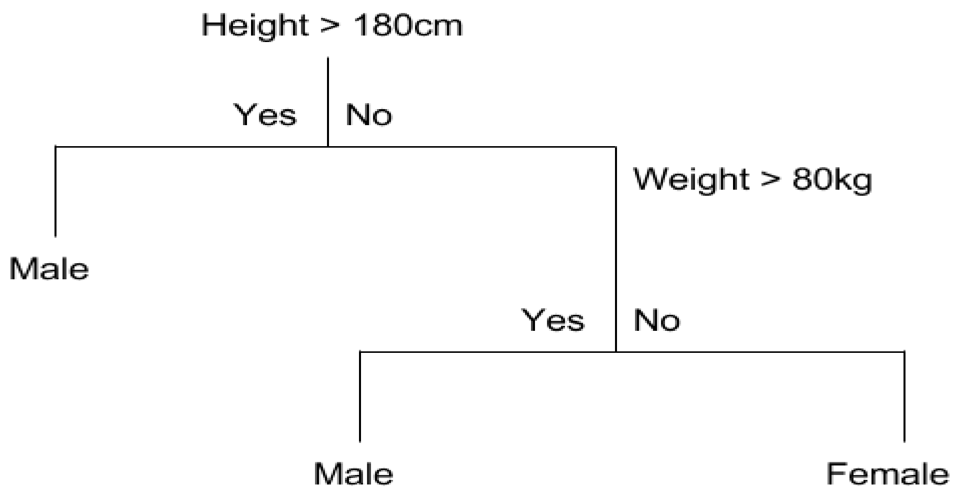


Figure 3.6 Example for illustrating classification and regression tree analysis

C4.5: First introduced by Quinlan in 1993, the decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. Authors of the Weka machine learning software describes the C4.5 algorithm as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date" (Witten, Frank & Hall, 2011).

CHAID (Chi-square Automatic Interaction Detection): This method was developed by Kaas in 1980. It uses the chi-square test to generate the optimal partitioning. CHAID creates the trees in which each node defines the partitioned state. To calculate the best segment, the predictor variables are combined to a statistically significant difference in a pair that matches the target variable.

The CHAID algorithm is widely used in decision trees algorithms for continuous and categorical variable typing. Chi-Square technique classifies the groups that differ according to their level of relationship. The leaves of the tree are branched not by the double but by the number of different structures. While C4.5 and CART derive from binary trees, CHAID derives from non-binary multiple trees.

QUEST: (Quick, Unbiased, Efficient, Statistical Trees) is a two-compartment decision tree algorithm developed by Wei-Yin Loh and Yu-Shan Shih for data mining and classification.

3.3.4.3 Regression

The mathematical representation of the relationship between any dependent variable and one or more independent variables is called regression analysis. There is a correlation coefficient between the dependent variable and the independent variables, expressed as a percentage of the power. In regression analysis, when the relation between variables is linear, it is expressed as Linear Regression Model, and when it is not linear, it is expressed as Non-Linear Regression Model.

3.4 Decision Support System

Decision support systems are computer based information systems that help people make decisions. They enable efficient use of data and models, enabling complex problems to be solved more easily. People make mistakes when making decisions on their own. These are the main ones.

- The problem cannot be expressed well.
- Making poor probability estimates.
- Excessive trust.
- Failure to correctly determine priorities.

These errors increase as the number of data items are increased. As a result, the development and prosperity of computer-aided decision support systems has increased over time.

Many different definitions have emerged to describe decision support systems (DSS).

Intended to assist decision makers in how they can make use of data and models to solve unstructured problems, interactive computer-based systems definition was first used by Scott Morton to define DSS (Gorry & Morton, 1971).

DSS is also defined as a system that utilizes a combination of individual intellectual resources and computer capabilities to help higher-quality decisions be made, according to another classic definition given by (Keen & Morton, 1978). The system functions as a computer based support system to help decision makers manage the process with semi-structured problems.

As defined by Bonczek and Whinston, DSS is defined as a computer-based system that encompasses three components: These components include a language system (a mechanism to provide communication between the user and other components of the DSS), a knowledge system (a repository of problem domain knowledge embodied in DSS as either data or procedures), and a problem-processing system (a link between the other two components), all of which interact with each other (Bonczek, et al., 1980).

The term DSS is “a situation where a ‘final’ system can be developed only through an adaptive process of learning and evolution”, according to Keen. Keen, therefore, defines a DSS as the product of a developing mental process that consists of components influencing one another i.e. the DSS user, the DSS builder, and the DSS itself (Keen, 1980).

In another definition provided by Turban in a more precise way, DSS is "an interactive, flexible, and adaptable computer-based information system, especially developed for supporting the solution of a non-structured management problem for improved decision making. It utilizes the data, provides an easy-to-use interface, and allows for the decision maker's own insights.“ (Turban, 1995).

The following is a decision support system layers, depicted in Figure 3.7.

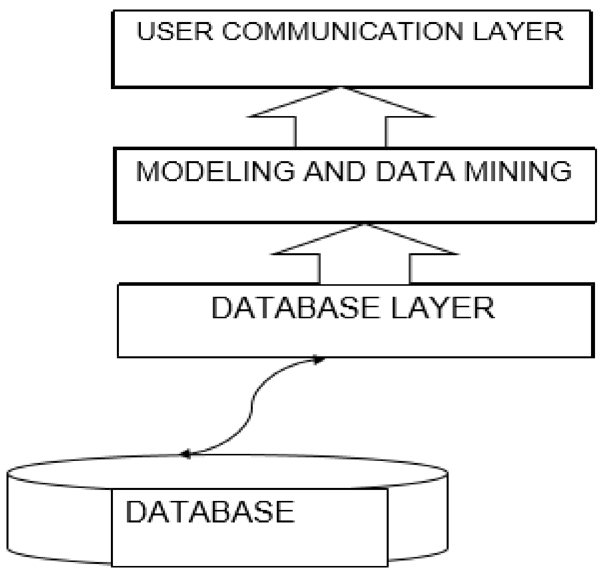


Figure 3.7 Decision support system layers

The application areas of decision support systems are as follows:

- Clinical applications and laboratory applications. In this group of applications, the aim is to make sure that the decisions made by experts on health issues are in line with the processes performed on the collected clinical data.
- Finance / Risk applications: In this group of applications, the aim is to solve various problems in finance and banking sectors.
- The DSS used in the agricultural sector helps with predictions and risk calculations, and thereby making decision-making process easier for farmers.
- DSS applications are also used in the scientific studies.
- The DSS can also be used in all types of construction decisions (e.g. railway, bridge, highway, building, dam construction), forest management calculations (fire management, determination of planting areas, erosion and risk identification).

In general, it is possible to say that the DSS makes the decision-making process easier. The benefits provided by the DSS are described below:

- It speeds up the decision-making process.
- It improves the decision maker's data access and detection process.
- Stability rate is higher.
- By thinking of many alternatives, you reach the most appropriate solution.
- It provides guidance on what to do in unexpected situations or in risky situations.
- Finding the best solution as soon as possible saves both time and money.
- It uses data sources in the best way.
- It enables the automation of the decision process.

CHAPTER FOUR
USED TECHNOLOGIES

In this chapter, the technologies that are used to develop applications to be used in this study are described. The database program used is explained and the tables used for the data in the database are given. In addition to, the data mining program that is used to analyze data is disclosed. Finally, the program used to create the Decision support system system is explained.

4.1 Database

Two separate databases were used in this study. One of them is used for data mining system and the other for decision support system.

First, to analyze data more easily in the SPSS program, used the Excel program. This program is one of the Microsoft Office programs, can be used like a database at the same time. A second database is used for the decision support system.The program used for this is MICROSOFT SQL SERVER MANAGEMENT STUDIO. In this way, the connection to the application has been made easier.

4.1.1 Excel

Excel is an object-oriented program developed by Microsoft. Prior to developing Microsoft Excel, the Lotus program provided similar needs. When the Lotus program could not meet the needs of the time, Microsoft decided to develop Excel, a new program. Excel is a spreadsheet program. Excel is an application program that allows you to keep all kinds of information and perform all calculations and analyzes. Excel 1.0 was introduced in 1982. This version has passed into history as the first Intel-based program (Excel, n.d.).

4.1.2 Microsoft SQL Server Management Studio

Microsoft SQL Server is a relational database management system developed by Microsoft. As a database server, it is a software product whose primary function is to store and retrieve data as requested by other software applications, be it those on the same computer or those running on another computer across a network (including the Internet). There are at least a dozen different editions of Microsoft SQL Server aimed at different audiences and for workloads ranging from small single-machine applications to large Internet-facing applications with many concurrent users. Its primary query languages are T-SQL and ANSI SQL (Microsoft SQL Server Management Studio, n.d.).

4.2 SPSS Clementine

Clementine is a mature data mining toolkit which aims to allow domain experts (normal users) to do their own data mining. Clementine has a visual programming or data flow interface, which simplifies the data mining process. Clementine applications include customer segmentation/profiling for marketing companies, fraud detection, credit scoring, load forecasting for utility companies, and profit prediction for retailers. SPSS Clementine was one of the very first general purpose data mining tools, and one of the most popular data mining packages (SPSS Clemantine, n.d.).

4.3 Visual Studio 2013

Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It is used to develop computer programs, as well as web sites, web apps, web services and mobile apps. Visual Studio uses Microsoft software development platforms such as Windows API, Windows Forms, Windows Presentation Foundation, Windows Store and Microsoft Silverlight. It can produce both native code and managed code. Visual Studio includes a code editor supporting IntelliSense (the code completion component) as well as code refactoring. The integrated debugger works both as a source-level debugger and a machine-level debugger.

Other built-in tools include a code profiler, forms designer for building GUI applications, web designer, class designer, and database schema designer. It accepts plug-ins that enhance the functionality at almost every level—including adding support for source control systems (like Subversion) and adding new toolsets like editors and visual designers for domain-specific languages or toolsets for other aspects of the software development lifecycle (like the Team Foundation Server client: Team Explorer).

Visual Studio supports 36 different programming languages and allows the code editor and debugger to support (to varying degrees) nearly any programming language, provided a language-specific service exists. Built-in languages include C, C++, C++/CLI, Visual Basic .NET, C#, F#, JavaScript, TypeScript, XML, XSLT, HTML and CSS. Support for other languages such as Python, Ruby, Node.js, and M among others is available via plug-ins. Java (and J#) were supported in the past (Visual Studio, n.d.).

4.3.1 ASP.NET

ASP.NET is an open-source server-side web application framework designed for web development to produce dynamic web pages. It was developed by Microsoft to allow programmers to build dynamic web sites, web applications and web services.
It was first released in January 2002 with version 1.0 of the .NET Framework, and is the successor to Microsoft's Active Server Pages (ASP) technology. ASP.NET is built on the Common Language Runtime (CLR), allowing programmers to write ASP.NET code using any supported .NET language. The ASP.NET SOAP extension framework allows ASP.NET components to process SOAP messages.

ASP.NET's successor is ASP.NET Core. It is a re-implementation of ASP.NET as a modular web framework, together with other frameworks like Entity Framework. The new framework uses the new open-source .NET Compiler Platform (codename "Roslyn") and is cross platform. ASP.NET MVC, ASP.NET Web API, and

ASP.NET Web Pages (a platform using only Razor pages) have merged into a unified MVC 6 (ASP.NET, n.d.).

CHAPTER FIVE
IMPLEMENTATION OF THE APPLICATION

5.1 Data Mining Applications

The purpose of this paper is to create a decision support system that will be of help for farmers in their agricultural practices by means of defining the factors that affect crop yield. In this study, first of all, factors that affect crop yield have been defined by using various data mining algorithms. In this section, the data used is explained, and results from the data mining algorithms applied on this data are also covered.

5.1.1 Data Source

The data made available by the İzmir Menemen Provincial Directorate of Agriculture regarding Barley product's characteristics as well as the weather data from the Meteorological Service Department constitute the overall data used in this study.

The data made available by the İzmir Menemen Provincial Directorate of Agriculture covers the amounts of barley crop yields of a 10-year period in Menemen, Nazilli and Karacabey, and its content is as follows: Variety of barley product, planting date, harvest date, yield, weight in hectoliter, weight by 1000 grams, protein ratio, sieve fraction, earing span, length, location. The weather data of a 10-year period made available by the Meteorological Service Department has been presented in the form of monthly totals or averages, and covers the following content: monthly total of solar heat, monthly average of humidity, monthly average of wind speed, monthly average of temperature, and monthly total of rainfall. These two types of data were combined and formed as a new set of data with new column headings to record all the data on the Excel platform. This data was achieved along with 1077 observations (objects and data) described by 20 attributes (variables). In other words, we generated a data matrix O with the dimensions of 1077x20 with the attributes shown in Figure5.1.

Id	Çeşit	Planting Date	Harvest Date	Yield (kg/da)	Hectoliter	1000 grain weight	Protein	Sieve ratio	Başaklanma Günü	Lenght	Location	İlk 3ayGüneş	Kalan Güneş	Nem	Rüzgar Hız	sıc. Ort	Toplam Ay	Toplam Yağmur	İlk Ay Yağmur	Son Ay Yağmur
1	1	17.11.2015	8.6.2016	432	70	56	14	91	104	123	Menem...	616	1254	62	3	15	8	446	41	13
2	2	17.11.2015	8.6.2016	501	61	46	12	75	104	105	Menem...	616	1254	62	3	15	8	446	41	13
3	3	17.11.2015	8.6.2016	469	65	42	13	78	101	110	Menem...	616	1254	62	3	15	8	446	41	13
4	4	17.11.2015	8.6.2016	571	68	40	13	57	105	115	Menem...	616	1254	62	3	15	8	446	41	13
5	5	17.11.2015	8.6.2016	598	65	42	13	79	100	111	Menem...	616	1254	62	3	15	8	446	41	13
6	2	17.11.2015	8.6.2016	440	59	47	12	73	103	104	Menem...	616	1254	62	3	15	8	446	41	13
7	1	17.11.2015	8.6.2016	404	65	56	14	86	104	125	Menem...	616	1254	62	3	15	8	446	41	13
8	4	17.11.2015	8.6.2016	441	64	40	12	72	104	112	Menem...	616	1254	62	3	15	8	446	41	13
9	3	17.11.2015	8.6.2016	596	50	37	13	64	100	109	Menem...	616	1254	62	3	15	8	446	41	13
10	2	17.11.2015	8.6.2016	513	57	50	11	76	104	106	Menem...	616	1254	62	3	15	8	446	41	13
11	1	17.11.2015	8.6.2016	454	67	54	14	89	103	125	Menem...	616	1254	62	3	15	8	446	41	13
12	4	17.11.2015	8.6.2016	488	63	41	12	75	102	116	Menem...	616	1254	62	3	15	8	446	41	13
13	3	17.11.2015	8.6.2016	590	63	39	13	56	101	111	Menem...	616	1254	62	3	15	8	446	41	13
14	2	17.11.2015	8.6.2016	401	58	52	12	76	103	105	Menem...	616	1254	62	3	15	8	446	41	13
15	1	17.11.2015	8.6.2016	501	68	55	13	92	105	124	Menem...	616	1254	62	3	15	8	446	41	13
16	4	17.11.2015	8.6.2016	483	62	44	13	80	106	115	Menem...	616	1254	62	3	15	8	446	41	13
17	3	17.11.2015	8.6.2016	567	65	40	13	54	102	108	Menem...	616	1254	62	3	15	8	446	41	13
18	2	23.11.2015	24.6.2016	387	50	47	14	67	103	105	Nazilli	712	1004	61	2	15	8	408	38	43
19	1	23.11.2015	24.6.2016	346	60	56	13	86	95	100	Nazilli	712	1004	61	2	15	8	408	38	43
20	4	23.11.2015	24.6.2016	339	58	45	11	82	96	102	Nazilli	712	1004	61	2	15	8	408	38	43
21	3	23.11.2015	24.6.2016	430	59	48	11	74	98	103	Nazilli	712	1004	61	2	15	8	408	38	43
22	4	17.11.2015	8.6.2016	455	61	37	13	52	102	101	Menem...	616	1254	62	3	15	8	446	41	13
23	6	17.11.2015	8.6.2016	393	57	42	15	61	105	100	Menem...	616	1254	62	3	15	8	446	41	13
24	5	17.11.2015	8.6.2016	472	62	46	17	49	103	104	Menem...	616	1254	62	3	15	8	446	41	13
25	2	5.12.2014	23.6.2015	270	51	46	13	84	88	121	Menem...	375	1059	65	3	14	7	622	186	71
26	1	5.12.2014	23.6.2015	427	67	41	12	94	104	109	Menem...	375	1059	65	3	14	7	622	186	71
27	3	5.12.2014	23.6.2015	336	66	32	12	66	101	110	Menem...	375	1059	65	3	14	7	622	186	71
28	4	5.12.2014	23.6.2015	320	64	40	12	85	102	110	Menem...	375	1059	65	3	14	7	622	186	71
29	2	5.12.2014	23.6.2015	269	55	37	13	79	89	120	Menem...	375	1059	65	3	14	7	622	186	71
30	1	5.12.2014	23.6.2015	213	62	40	13	88	102	110	Menem...	375	1059	65	3	14	7	622	186	71
31	3	5.12.2014	23.6.2015	411	63	28	13	59	102	109	Menem...	375	1059	65	3	14	7	622	186	71
32	4	5.12.2014	23.6.2015	338	59	30	13	79	102	111	Menem...	375	1059	65	3	14	7	622	186	71
33	2	5.12.2014	23.6.2015	301	51	37	11	76	88	122	Menem...	375	1059	65	3	14	7	622	186	71
34	1	5.12.2014	23.6.2015	278	64	40	11	89	100	109	Menem...	375	1059	65	3	14	7	622	186	71
35	3	5.12.2014	23.6.2015	417	62	27	12	55	101	110	Menem...	375	1059	65	3	14	7	622	186	71
36	4	5.12.2014	23.6.2015	332	63	30	12	77	102	109	Menem...	375	1059	65	3	14	7	622	186	71
37	2	26.12.2014	18.6.2015	213	54	41	15	88	88	121	Nazilli	355	1042	56	2	15	7	605	194	45
38	1	26.12.2014	18.6.2015	252	66	48	16	93	89	120	Nazilli	355	1042	56	2	15	7	605	194	45
39	3	26.12.2014	18.6.2015	329	65	37	13	69	105	109	Nazilli	355	1042	56	2	15	7	605	194	45
40	4	26.12.2014	18.6.2015	384	60	38	13	84	100	111	Nazilli	355	1042	56	2	15	7	605	194	45
41	2	26.12.2014	28.6.2015	305	60	41	14	52	95	109	Karaca...	363	891	65	2	15	7	560	136	22
42	1	26.12.2014	28.6.2015	109	63	47	15	77	90	112	Karaca...	363	891	65	2	15	7	560	136	22
43	3	26.12.2014	28.6.2015	461	65	31	13	61	110	115	Karaca...	363	891	65	2	15	7	560	136	22

İlk 3Ay Güneş=Total holar heat in the first 3 months
Kalan güneş=sum of amount of solar heat received until harvest time – sum of solar heat received for the first 3 months
Nem= monthly average of humidity
RüzgarHızı= Monthly average of wind speed
Sıc. Ort=Monthly average of temperature
Toplam Ay=Total month
İlk Ay Yağmur=Total rainfall in the first month
Son Ay Yağmur=Total rainfall in the last month

Figure 5.1 Details of the barley crop records

5.1.2 Data Properties

The details of the data type is shown in Figure 5.2. The data is listed by data type and data feature as follows:

Field	Type	Values	Missing	Check	Direction
Id	Range	[1,1077]		None	In
Çeşit	Set	1,2,3,4,5,6,7,...		None	In
Planting Date	Set	"10.2.2002","...		None	In
Harvest Date	Set	"10.6.2006","...		None	In
Yield (kg/da)	Range	[65,788]		None	Out
Hectoliter	Range	[50,71]		None	In
1000 grain weig...	Range	[3,61]		None	In
Protein	Range	[7,17]		None	In
Sieve ratio	Range	[8,107]		None	In
Başaklanma Gü...	Range	[43,113]		None	In
Lenght	Range	[11,140]		None	In
Location	Set	Karacabey,M...		None	In
İlk 3ayGüneş	Range	[196,712]		None	In
Kalan Güneş	Range	[400,1254]		None	In
Nem	Range	[55,74]		None	In
Rüzgar Hız	Range	[2,6]		None	In
sıc. Ort	Range	[12,17]		None	In
Toplam Ay	Range	[5,9]		None	In
Toplam Yağmur	Range	[133,622]		None	In
İlk Ay Yağmur	Range	[6,194]		None	In
Son Ay Yağmur	Range	[0,71]		None	In

İlk 3Ay Güneş=Total holar heat in the first 3 months
Kalan güneş=sum of amount of solar heat received until harvest time – sum of solar heat received for the first 3 months
Nem= monthly average of humidity
RüzgarHızı= Monthly average of wind speed
Sıc. Ort=Monthly average of temperature
Toplam Ay=Total month
İlk Ay Yağmur=Total rainfall in the first month
Son Ay Yağmur=Total rainfall in the last month

Figure 5.2 Details of the data type

Location-based distribution of the data is shown in Figure 5.3 its average yield-based distribution is as shown in Figure 5.4.

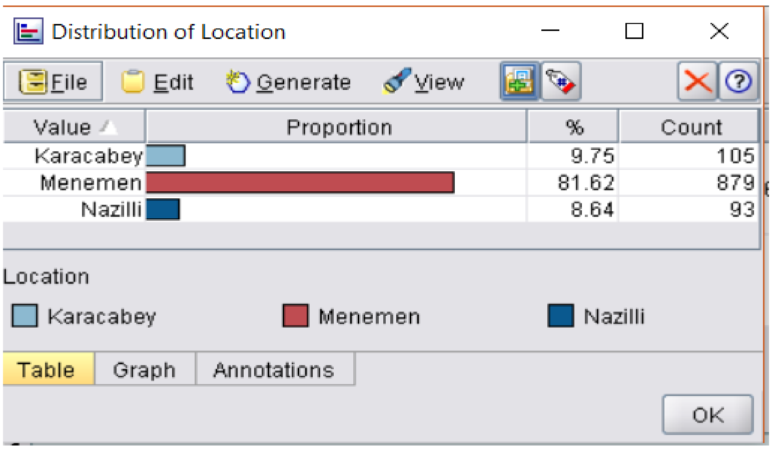


Figure 5.3 Details in location-based distribution of the data

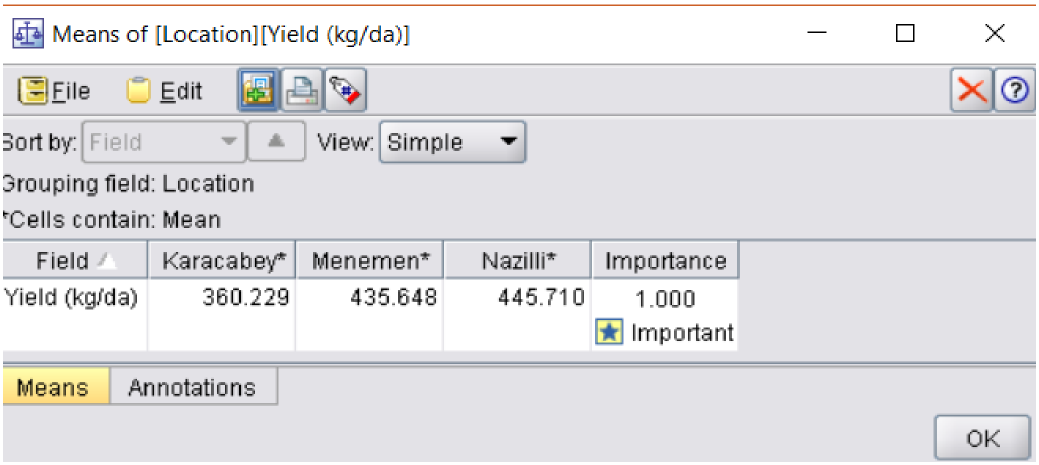


Figure 5.4 Means of yield in location-based

Its reported yield-based distribution and its reported location-based distribution are as shown in the figure 5.5.

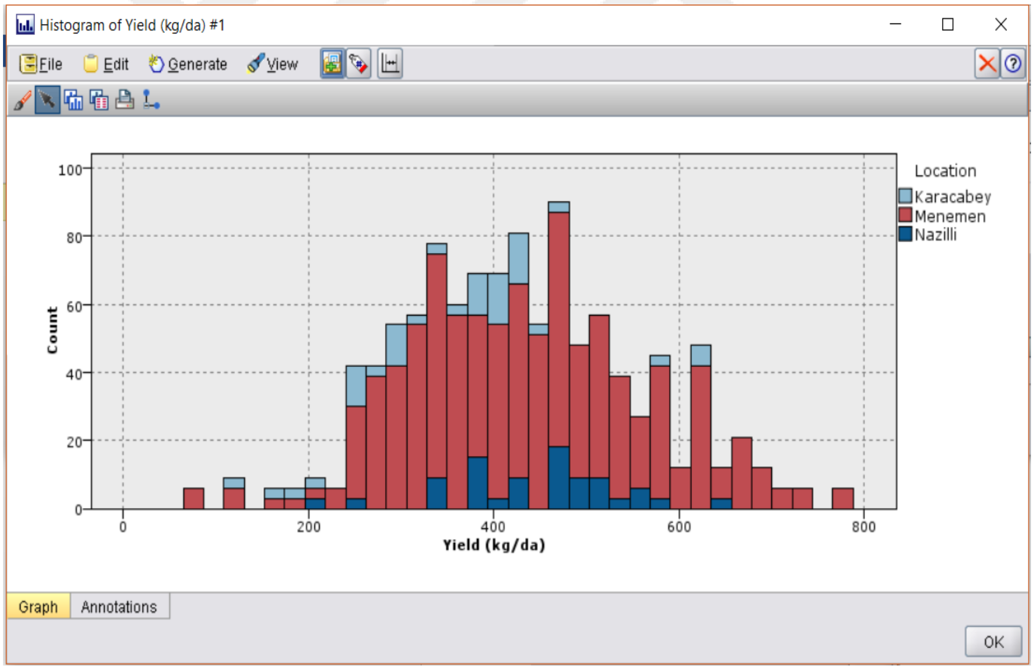


Figure 5.5 Yield-based distribution in location-based

The data mining that was carried out on the data was intended to find out the factors that have the highest influence on the weather-based and location-based yields, the two that are influenced most. Because, in this case, it is possible to calculate crop yields based on weather conditions.

First, a feature selection was applied to comprehend the significance of the factors that are affecting crop yield. As shown in the figure 5.6, all weather-related factors except humidity affect crop yield. Our aim is to determine the factors that have an effect on crop yield by applying various algorithms on our data in various ways, and to conclude with a prediction on crop yield by interpreting the results retrieved.

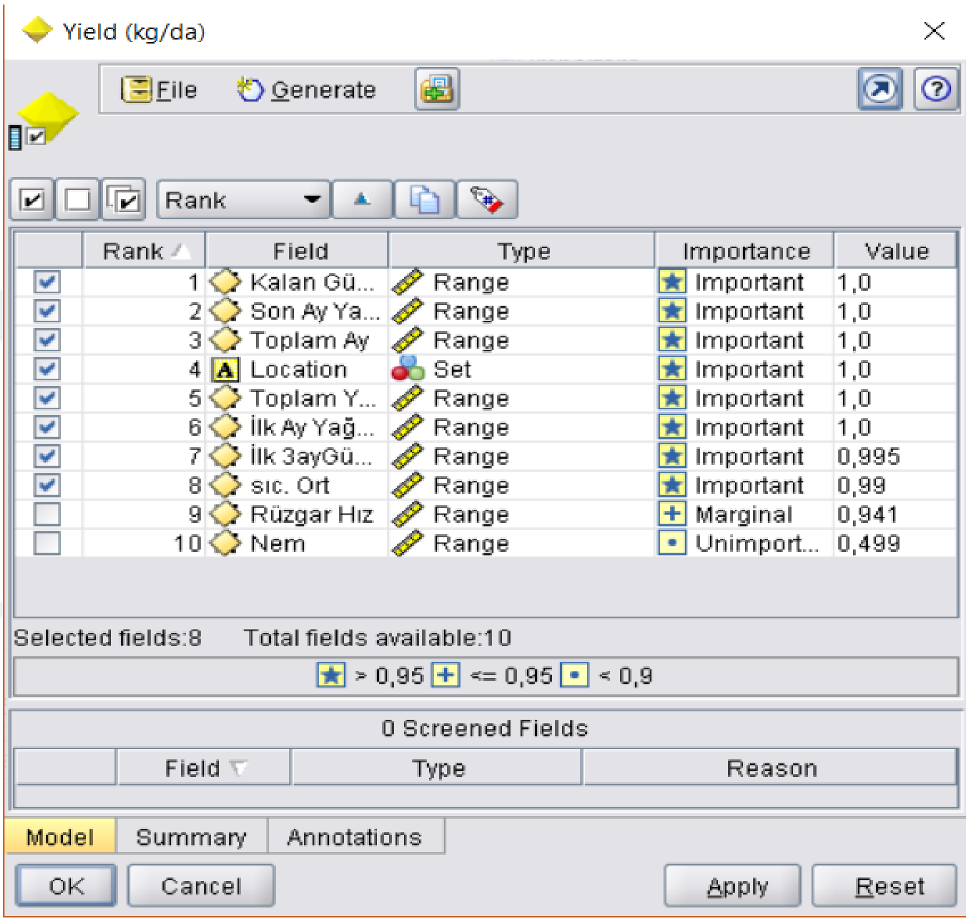


Figure 5.6 Significance of the factors that are affecting crop yield

Then, based on the impressions from the studies inspected, it was observed that the data structure matches with CHAID Algorithm and C&R TREE Algorithm. These analysis structures were employed on this data respectively.

5.1.3 CHAID Algorithm

In CHAID algorithm; target value is selected as yield, input values are selected as the weather data, and various rules are then formed accordingly. These rules are subsequently interpreted to yield results from them. The rules shown in Figure 5.7 below were used to interpret this process.

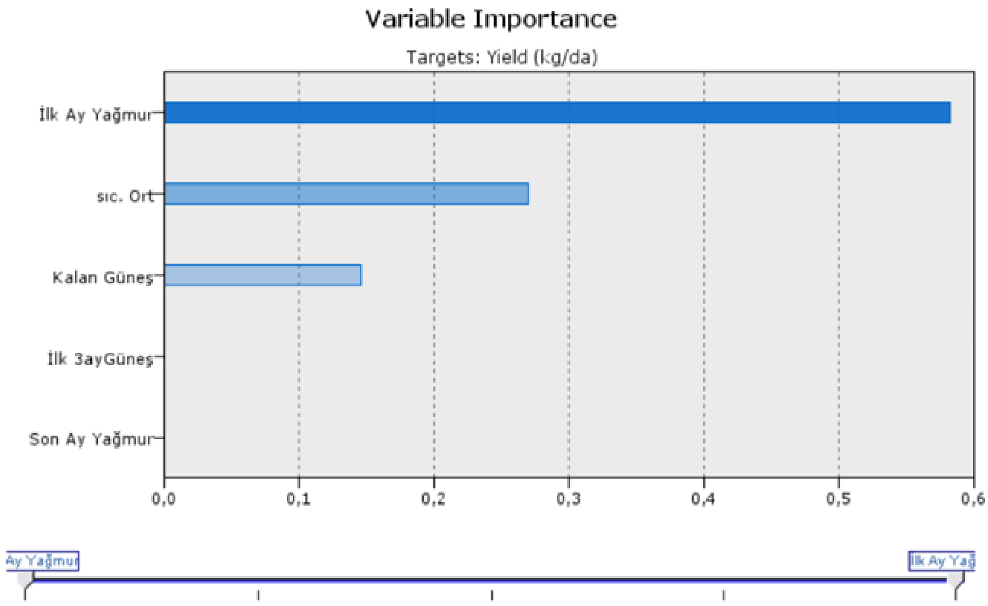
İlk Ay Yağmur=Total rainfall in the first month
İlk 3Ay Güneş=Total holar heat in the first 3 months
Sıc. Ort=Temperature average
Son Ay Yağmur=Total rainfall in the last month

Figure 5.7 The rules by using CHAID algorithm

This process reveals the factors that have the highest effect on crop yield based on this algorithm. It is observed that the most important of these factors is the total rainfall for the 1-month period following the crop planting.

If the amount of the rainfall during the first month is below 6 mm, the average crop yield is 507.909. If the amount of the rainfall during the first month is higher than 6 mm and less than 20 mm, the average crop yield is observed to be 417.628. This case reveals another factor that affects the crop yield. This factor is the total amount of sunlight that the crop gets during the first 3-month period. If this amount is less than 235, the average crop yield is observed to be 489.167; and if it is between 235 and 481, the average crop yield is then 244.8. If this amount is higher than 481, the average crop yield is observed to be 417.846. If the amount of rainfall is between 20 mm and 31 mm in the first month the average crop yield is 510.951. If the rainfall for the first month is between 31 and 39, then another factor affecting crop yield is revealed in this case. This factor is the average temperature for the period between planting and harvesting. If the average temperature is below 12°C, the average crop yield is 328.029; and if it is between 12°C and 14°C, the average crop yield is observed to be 346.235. If the rainfall for the first month is between 39 mm and 56 mm, the average crop yield is 43.358, and thereby revealing another factor that affects crop yield. This factor is the total amount of sunlight that the crop gets during the first 3-month period. If this amount is less than 386, the average crop yield 405.638; and if it is higher than 386 the crop yield is then 488.45. If the rainfall during the first month is between 56 mm and 85 mm, the average crop yield is 641.5; and if this amount of rainfall is between 85 mm and 98 mm the average crop yield is observed to be 423.029. In this case, there is another factor affecting crop yield, which is the total amount of sunlight that the crop gets during the first 3-month period. If this amount is below 352, the average crop yield is observed to be 372.214, and if it is higher than 352, the average crop yield is 458.6. If the rainfall for the first month is above 98 mm, the average crop yield 306.659, which in turn reveals 2 factors that affect crop yield respectively. The first of these factors is the amount of rainfall during the one-month period preceding the crop’s harvest. If this amount is below 29 mm, the average crop yield 282.375; and if it is above 29 mm the average crop yield is 335.8. This case reveals another factor that affects the crop yield. This factor is the total amount of sunlight that the crop gets after the first 3-month period. If this amount is below 995, the average crop yield is 406.5; and if it is higher than 995, the average crop yield is 318.125.

The factors affecting crop yield are given in order of importance in Figure 5.8.



İlk Ay Yağmur=Total rainfall in the first month
İlk 3Ay Güneş=Total holar heat in the first 3 months
Sıc. Ort=Temperature average
Son Ay Yağmur=Total rainfall in the last month

Figure 5.8 Variable importance by using CHAID algorithm

If applied again with partition taken based on the location, CHAID algorithm allows different rules to form; and these rules can be interpreted to get new results. The rules shown in Figure 5.9 below were used to interpret this process.

Toplam Ay <= 7 [Ave: 260,667, Effect: -99,562] ⇨ 260,667
Toplam Ay > 7 [Ave: 394,692, Effect: 34,464]
  İlk 3ayGüneş <= 352 [Ave: 372,214, Effect: -22,478] ⇨ 372,214
  İlk 3ayGüneş > 352 [Ave: 420,917, Effect: 26,224] ⇨ 420,917

Toplam Ay=Total month
İlk 3Ay Güneş=Total holar heat in the first 3 months

Figure 5.9 The rules by using CHAID algorithm

If partition is changed to location-based, the most important factor is observed to be the period that spans from planting date to harvest date. The factors affecting crop yield within this approach are given in order of importance in Figure 5.10.

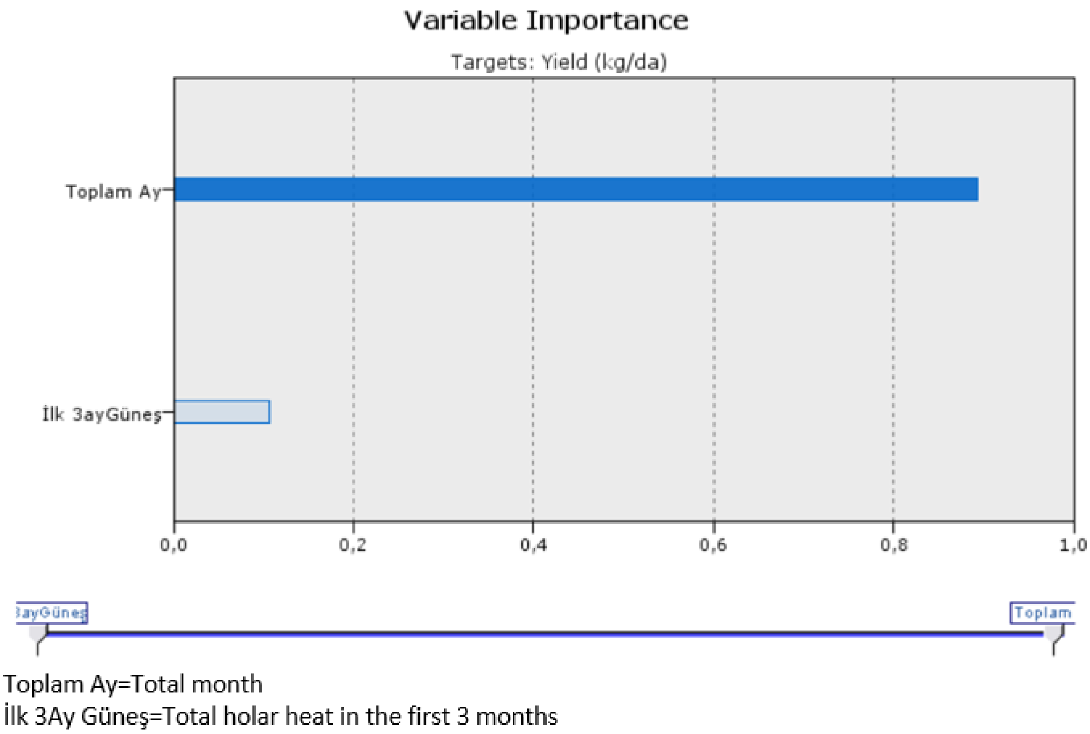


Figure 5.10 Variable importance by using CHAID algorithm

Based on these two separate approaches employed in CHAID algorithm, the major factors affecting crop yield are as follows in order of importance.

- Total months from crop planting to harvesting
- Total amount of solar heat that the crop gets within the first 3 months after being planted
- Total amount of rainfall within the first months following crop planting

5.1.4 C&R Tree Algorithm

In C&R Tree algorithm; target value is selected as yield, input values are selected as the weather data, partition is done on the basis of location, and various rules are then formed accordingly. These rules are subsequently interpreted to yield results from them. The rules shown in Figure 5.11 and fiure 5.12 below were used to interpret this process.

- Kalan Güneş <= 806 [Ave: 394,692, Effect: 34,464]
  - İlk 3ayGüneş <= 359 [Ave: 372,214, Effect: -22,478] ⇨ 372,214
  - İlk 3ayGüneş > 359 [Ave: 420,917, Effect: 26,224] ⇨ 420,917
- Kalan Güneş > 806 [Ave: 260,667, Effect: -99,562]
  - İlk 3ayGüneş <= 376,500 [Ave: 280,5, Effect: 19,833] ⇨ 280,5
  - İlk 3ayGüneş > 376,500 [Ave: 244,8, Effect: -15,867] ⇨ 244,8

Kalan güneş=sum of amount of solar heat received until harvest time – sum of solar heat received for the first 3 months
İlk 3Ay Güneş=Total holar heat in the first 3 months

Figure 5.11 The rules by using C&R Tree algorithm

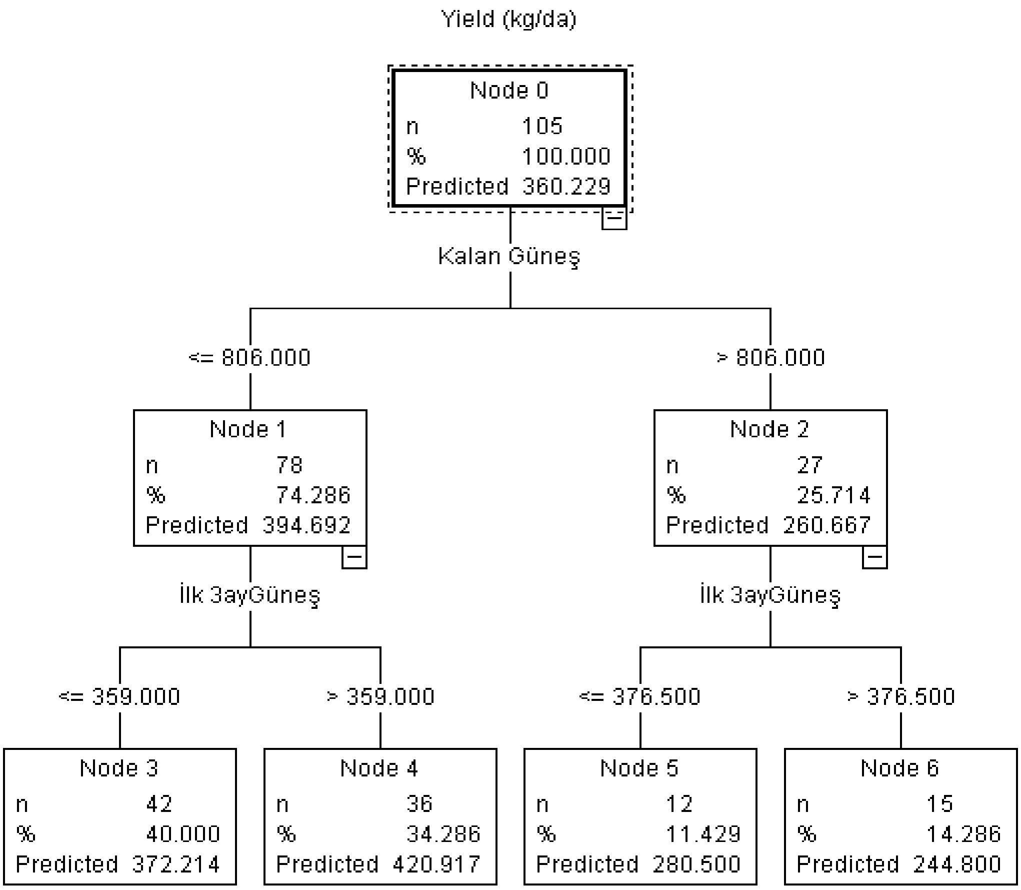


Figure 5.12 Structure of decision tree of the C&R Tree algorithm

According to C&R Tree Algorithm, the two most important factors affecting crop yield are the total solar heat that the crop gets within the first three months following crop planting, and the amount of solar heat that it gets starting this 3-month period until harvest. It is also observed that these two factors are proportional to each other. If this amount is less than 806, the second factor is observed to be the amount of the sunlight that the crop gets during the first 3-month period. If this amount is below 359, the average crop yield 372.214, or 420.917 if this amount is higher than 359. If the rest of the sunlight amount is higher than 806, again the second factor affecting yield is observed to be the amount of the sunlight that the crop gets during the first 3-month period. In this case, if this value is lower than 376.500 the average crop yield is 280.5, and if it is below 376.500 the average amount of crop yield is 244.8.

According to C&R Tree Algorithm, the two most important factors affecting crop yield are the total solar heat that the crop gets within the first three months following crop planting, and the amount of solar heat that it gets starting this 3-month period until harvest. It is also observed that these two factors are proportional to each other.

Based on this approach, if the ratio of total solar heat that the crop gets within the first three months following crop planting to total solar heat that it gets in the period following this 3-month period does not double it, then crop yield deteriorates. If it doubles, then the crop yield is boosted.

The CHAID Algorithm and C&R Tree Algorithm analyses prove that the factors affecting crop yield are as follows in order of importance:

- Total amount of solar heat that the crop gets within the first 3 months after being planted
- Total amount of rainfall within the first months following crop planting
- Ratio of the total solar heat that the crop gets within the first three months following crop planting to the total solar heat that it gets in the period following this 3-month period
- Total months from crop planting until harvest time

To test the accuracy of the data mining techniques we used, we reduced our data by 15% and applied the data mining algorithms again. In this case as well, the same effect was observed, but the factors influencing the efficiency were not changed. For the C & R Tree algorithm, the situation looks like this Figure 5.13.

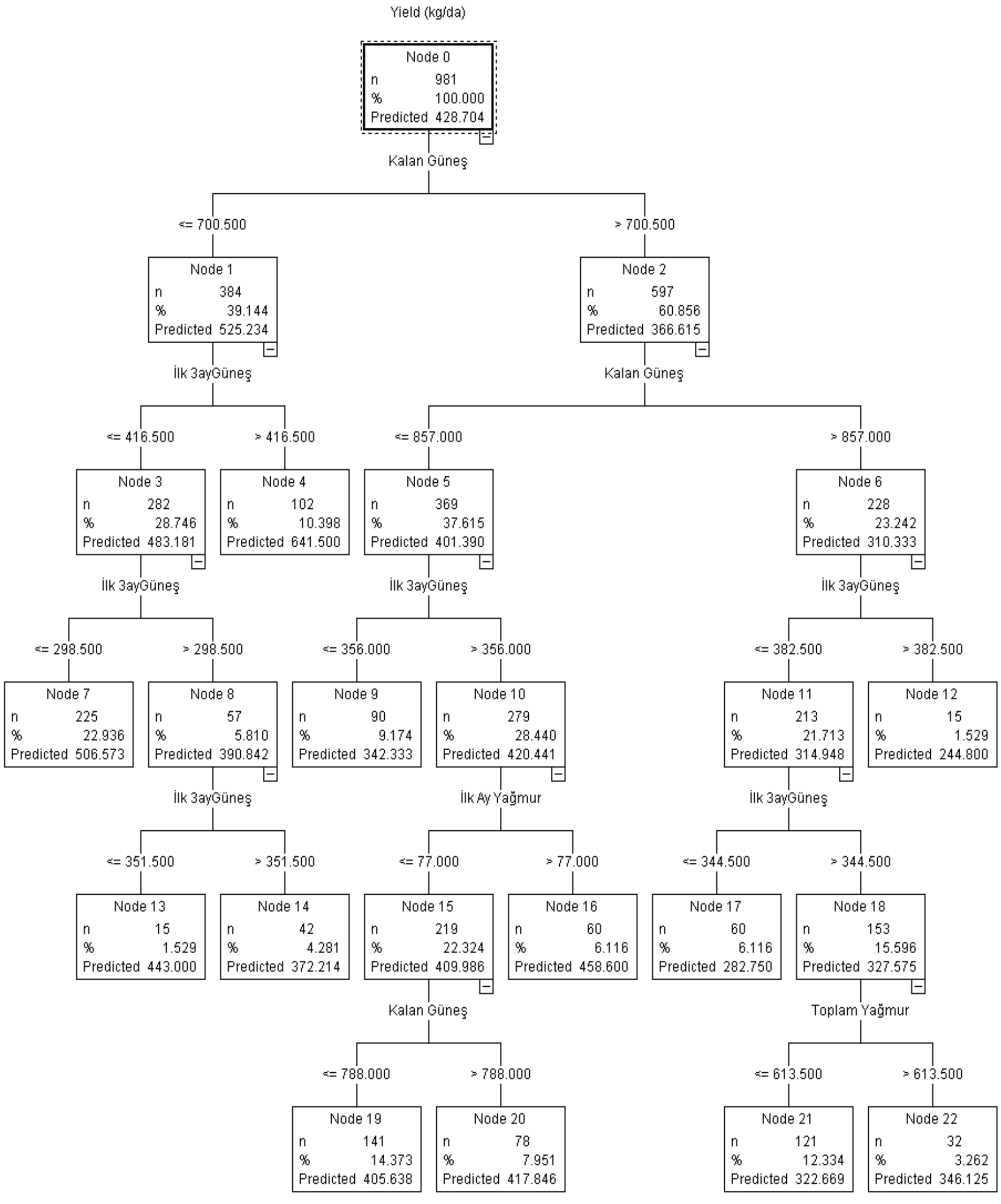


Figure 5. 13 Structure of decision tree of the reapplied C&R Tree algorithm

Similar results for the Chaid algorithm are shown in Figure 5.14.

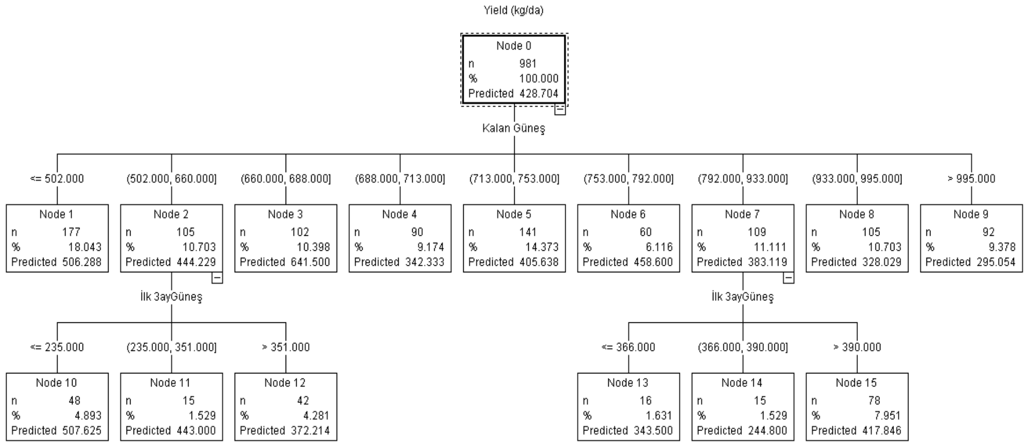


Figure 5. 14 Structure of decision tree of the reapplied using Chaid algorithm

Finally, in terms of the accuracy of our C&R Tree algorithm, it is seen in Figure 5.15 that this algorithm has a high accuracy rate, namely, that of 86.91%. In terms of the accuracy of our Chaid algorithm, that is show similar result.

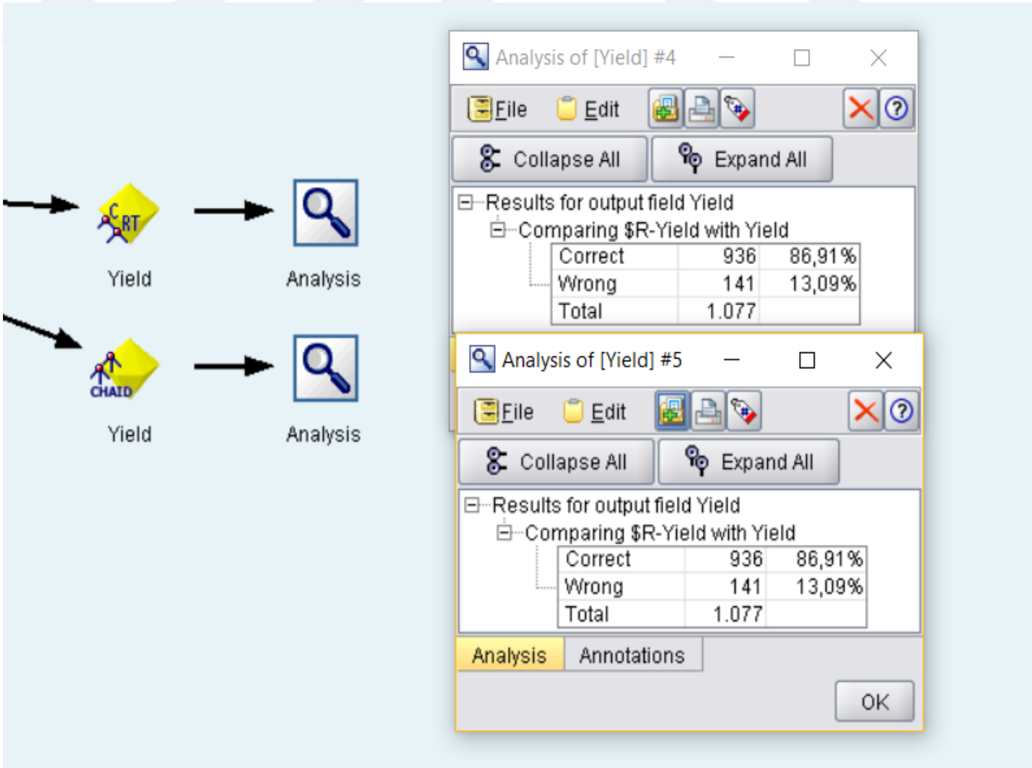


Figure 5. 15 Analysis for the C&R Tree algorithm and Chaid algorithm

To help users to make decisions, after obtaining necessary information from the Meteorological Service Department, a decision support system has been created pursuant to these factors.

5.2 Implementation of Decision Support System

5.2.1 Datebase in Decision Support System

The tables used in the Decision support system system developed is as seen in Figure 5.16.

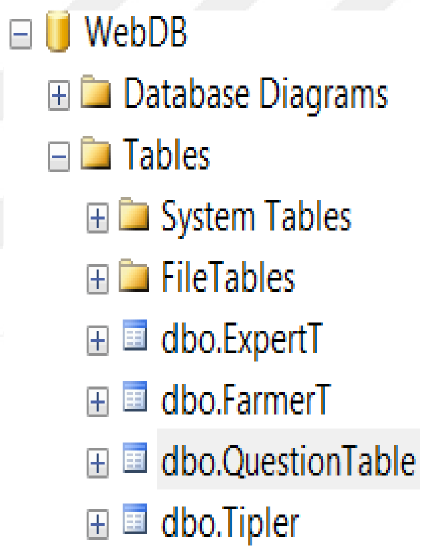


Figure 5.16 The tables used in decision support system in the Microsoft SQL SERVER

The database system created consists of four tables. The first table i.e. Experts Table has been developed for experts on the system to give answers to the questions directed by farmers. The design of the expert table is as shown in Figure 5.17. The Expert Table is made up of an ID number that varies as primary key, and of T.R. identity number, TypeID that varies depending on whether it is a Farmer's or an Expert's, as well as of a number allocated for Expert, Expert name, its Address, phone number and e-mail address.

ASUS-PC.WebDB - dbo.ExpertT

Column Name	Data Type	Allow Nulls
ExpertId	int	☐
ExpertTC	nvarchar(50)	☐
TipId	int	☐
ExpertNo	nvarchar(50)	☐
ExpertName	nvarchar(50)	☐
Adress	nvarchar(50)	☐
Telephone	nvarchar(50)	☐
Mail	nvarchar(50)	☐

Figure 5.17 The Expert Table used in decision support system in the Microsoft SQL SERVER

The second table i.e. Farmer Table has been developed for farmers registered on the system. The design of the farmer table is as shown in Figure 5.18. The Farmer Table is made up of an ID number that varies as primary key, and of T.R. identity number, TypeID that varies depending on whether it is a Farmer's or an Expert's, as well as of a number allocated for Farmer, Farmer name, its Address, phone number and e-mail address.

ASUS-PC.WebDB - dbo.FarmerT

Column Name	Data Type	Allow Nulls
FarmerId	int	☐
FarmerTC	nvarchar(50)	☐
TipId	int	☐
FarmerNo	nvarchar(50)	☐
FarmerName	nvarchar(50)	☐
Adress	nvarchar(50)	☐
Telephone	nvarchar(50)	☐
Mail	nvarchar(50)	☐

Figure 5.18 The Farmer Table used in decision support system in the Microsoft SQL SERVER

The third table i.e. Question Table is made up of the questions asked by the farmers on the system. The design of the Question Table is as shown in Figure 5.19. The Question Table is made up of an ID number that varies as primary key, type of the questioner, name of the Farmer asking question, its phone number, the number allocated for the Farmer asking question, the question asked by the Farmer, the answer given by Expert, name and phone number of the Expert giving the answer.

ASUS-PC.WebDB - dbo.QuestionTable

Column Name	Data Type	Allow Nulls
QuestionId	int	☐
TipId	int	☐
QFName	nvarchar(50)	☐
QFTelephone	nvarchar(50)	☐
QFNumber	nvarchar(50)	☐
QuestionM	nvarchar(MAX)	☑
QuestionS	nvarchar(MAX)	☑
SFName	nvarchar(50)	☑
SFTel	nvarchar(50)	☑

Figure 5.19 The Question Table used in decision support system in the Microsoft SQL SERVER

The fourth table i.e. Type Table has been determined based on the characteristics of the individuals registered on the system. The design of the type table is as shown in Figure 5.20. In upgraded versions of the system, new types can be incorporated into the system. The Type Table is made up of an ID number that varies as primary key, and a Type Name that specifies the Type itself.

ASUS-PC.WebDB - dbo.Tipler

Column Name	Data Type	Allow Nulls
TipId	int	☐
TipName	nvarchar(50)	☑

Figure 5.20 The Tip Table used in decision support system in the Microsoft SQL SERVER

5.2.2 Application in Decision Support System

Decision support systems are computer based information systems that help people make decisions. They enable efficient use of data and models, enabling complex problems to be solved more easily. People make mistakes when making decisions on their own.

In this study, to help farmers get maximum amount of crop yield in future, a decision support system has been developed based on farmers' past product info (harvest, yield, etc.), as well as on meteorological data. The decision support system was developed in Visual Studio using C# programming language and Asp.net software architectural. The page architecture used in the Decision support system is shown in Figure 5.21.

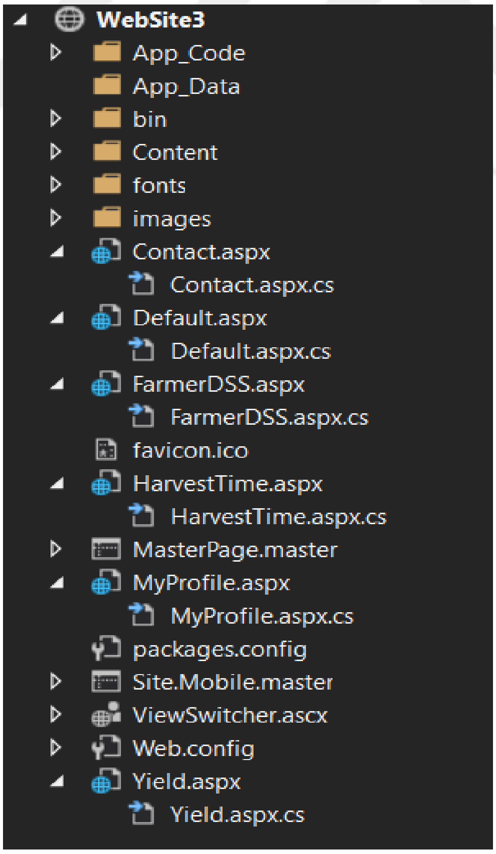


Figure 5.21 Farmer Decision Support System Asp.NET architectural

The decision support system developed can be used for Menemen, İzmir as majority of the data used relates to this region. Meanwhile; as the recent weather forecast data made available by the Meteorological Service Department covers 2017-2018, the system can only be used for this period. Implementation helps farmers in three main areas: These areas are as follows:

1) Harvest time forecast: The crop planting date info provided by users and the weather data from the Meteorological Service Department are combined and used to estimate the harvest time that will bring the highest crop yield.

2) Yield Forecast: The crop planting date and harvest date info provided by users and the weather data from the Meteorological Service Department are combined and used to forecast crop yield amount.

3) Farmers' Problems: The problems that farmers face and the questions directed by them to get the information they want are answered by experts.

When a farmer opens the program, an entry screen is displayed. On the upper part of this page is user log in, and in the middle part is an area that points to 4 different pages Figure 5.22.

Figure 5.22 Farmer decision support system home page

Figure 5.23 Farmer decision support system our service page

This entry page shown above and the Our Services page can be accessed by users without logging into the system as shown in Figure 5.23. Farmers are however required to log in to access the other three pages. A warning screen is displayed as shown in Figure 5.24 if the farmer is not logged in.

Figure 5.24 Farmer decision support system warning page

After the farmer has logged in, the Contact Us//Expert page is displayed. On this page, farmers can ask their questions and access necessary information that they need to contact page manager. This page is as shown in Figure 5.25.

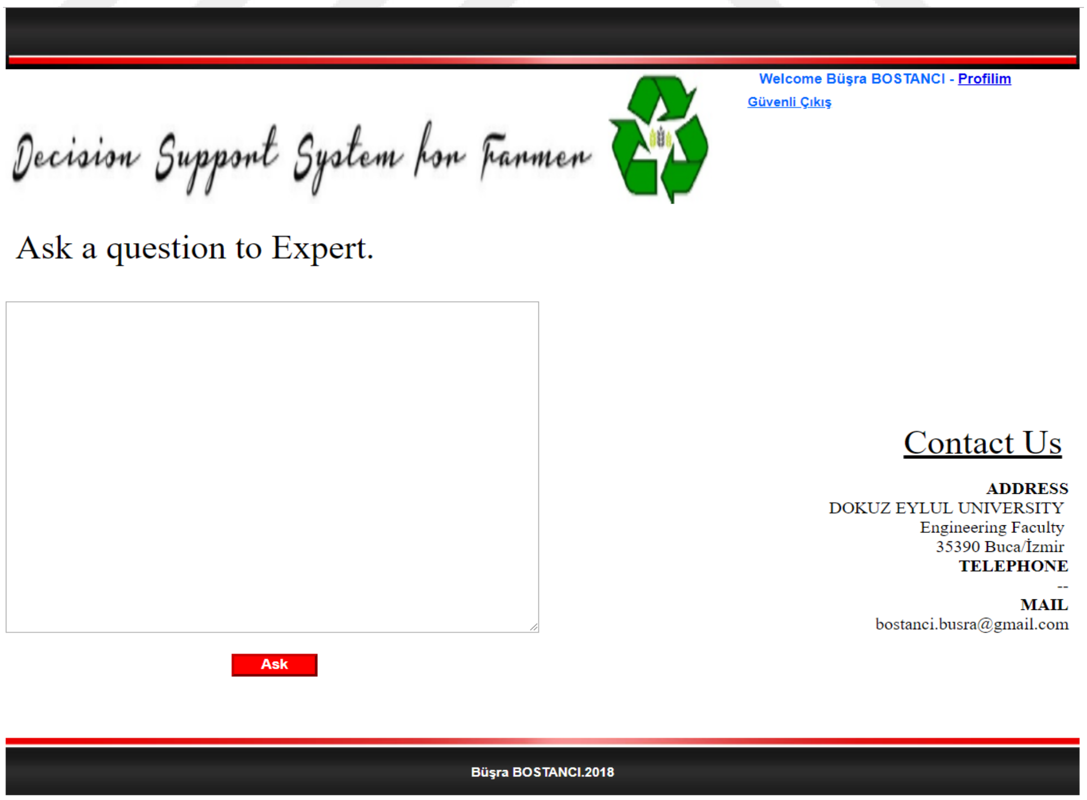


Figure 5.25 Farmer decision support system contact and question page

Farmers can see the answers clicking My Profile on the right upper side of the page after their questions have been answered by experts. My Profile page is as shown in Figure 5.26. This page shows the farmer's questions. They can select one question among them and see the answer to that question if that question has already been answered.

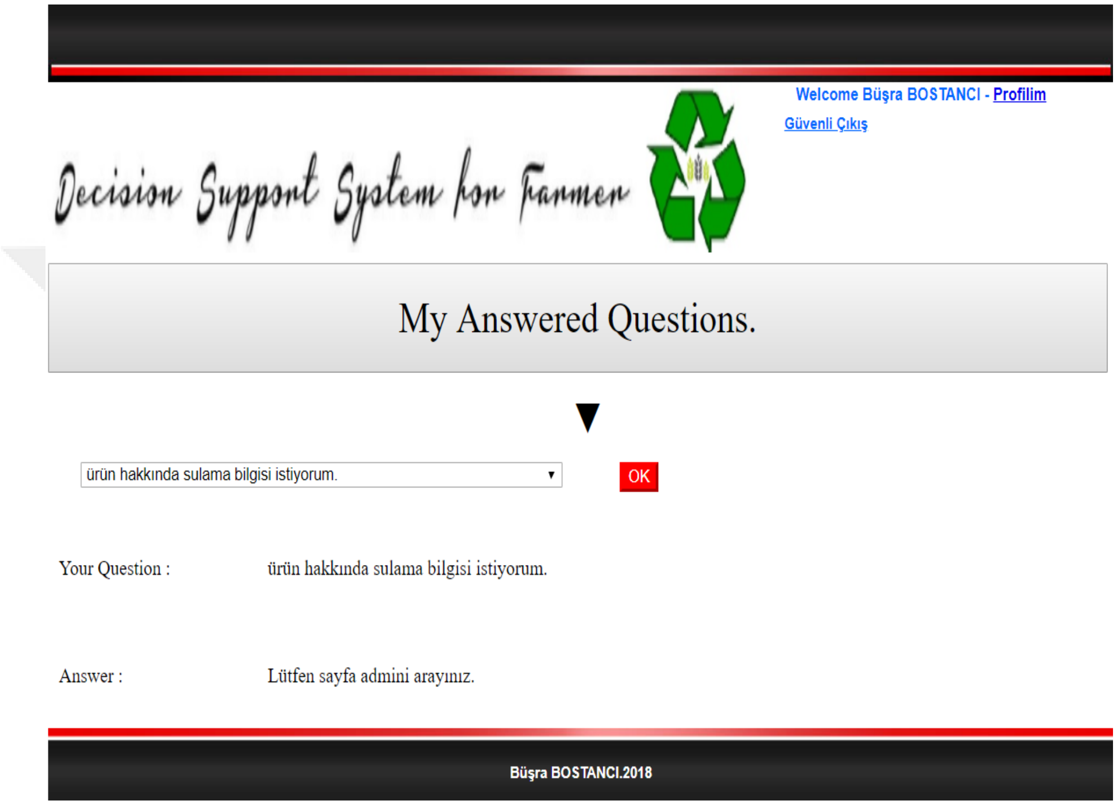


Figure 5.26 Farmer decision support system profile page

The other two main pages are as shown below respectively:

1) Harvest time forecast page:

On this page, farmers can view the forecasted harvest time based on the planting date provided by the farmers. Currently, this project runs only for the barley product in İzmir's Menemen district. Farmers can get the information regarding the total amount of solar heat that their crop has received so far, harvest time and the remaining amount of solar heat. This information is created based on the meteorological data available and the results from the data mining, as well as the crop planting date info provided by the farmers themselves. A warning screen is displayed as shown in Figure 5.27 if the optimal harvest time is missed.

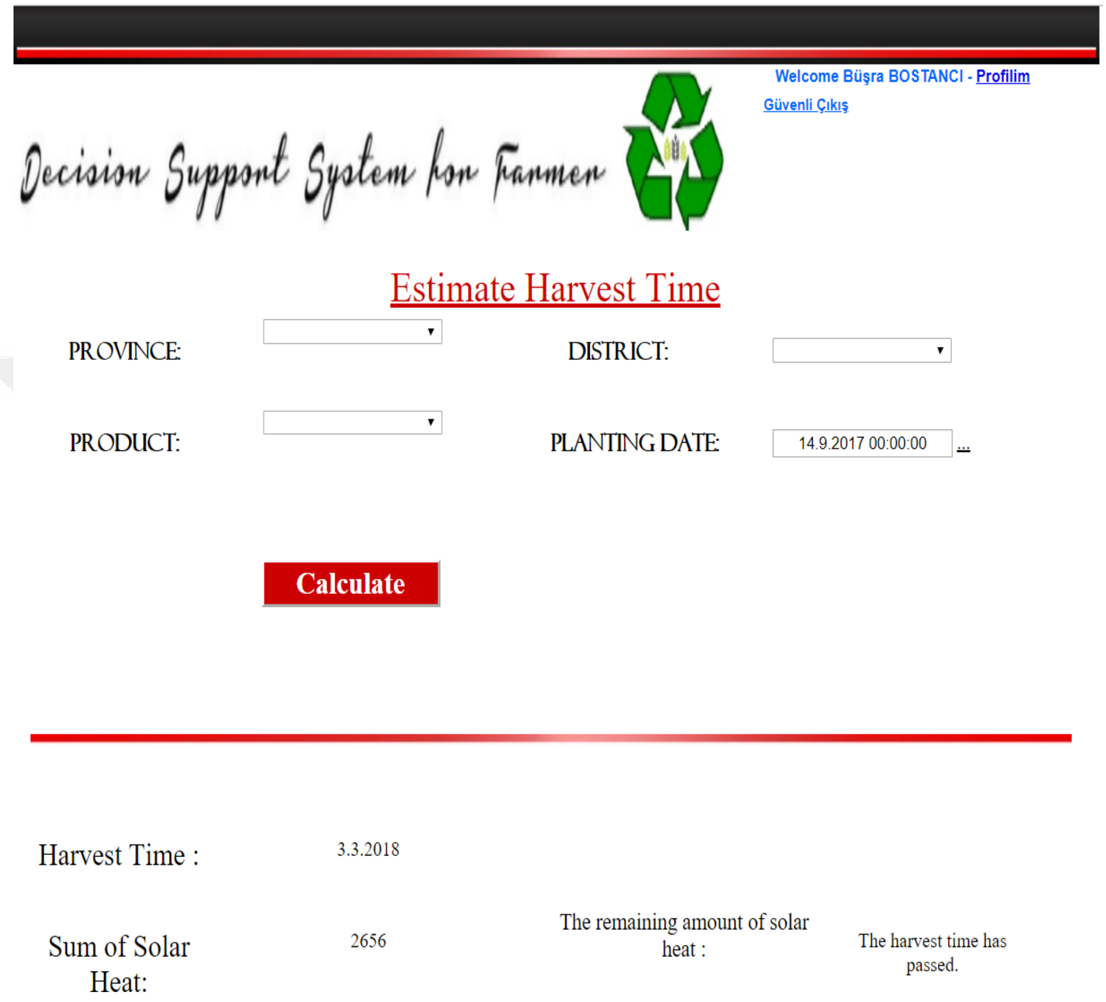


Figure 5.27 Farmer decision support system estimate harvest time page

2) Crop yield calculation page:

On this page, farmers can view the results created based on their planting date and harvest time. Currently, this project runs only for the barley product in İzmir's Menemen district. Farmers are informed about the approximate crop yield that they might get after harvesting their crop. This approximate crop yield info is created based on the planting date info and harvest time info provided by the farmers themselves. This page is as shown in Figure 5.28.

Figure 5.28 Farmer decision support system estimate yield page

CHAPTER SIX
CONCLUSION AND FUTURE WORK

6.1 Conclusion

There is a huge volume of data that is ready to be converted into useful information for the field of agriculture. Farmers need higher amounts of crop yield, lower costs, and minimal damage so they can maximize their gaining which can be achieved through more conscious use of this data at the time of decision-making. In this study, we have applied some sound techniques of data mining to the field of agriculture, which has allowed us to evaluate and examine data for the farmers. Within the scope of this study, the data received from the İzmir Menemen Provincial Directorate of Agriculture was organized for analysis. Then the SPSS Clementine software was used in order to analyze the data with the help of CHAID and CR&T decision tree algorithms to demonstrate the optimum harvesting time for barley yield. This analysis was carried out to determine the factors that have an impact on crop yield.

To test the accuracy of the data mining techniques we used, we reduced our data set by 15% and reapplyed the data mining algorithms used. That seen that the results we obtained in this case are similar to the results of our analysis for the entire data set.This analysis was carried out to determine the factors that have an impact on crop yield. When we test the validity of the algorithms that used, that seen that our two tests are 86.91% accurate. To help farmers make better decisions, an application for a decision support system has been designed in accordance with the results retrieved.

Some meaningful results were observed when the optimum harvesting time and corresponding yield rates were selected as the input variable and as the target variable, respectively. Total amount of solar heat and rainfall that the crop gets within the first three months and the first month following crop planting, respectively, affects crop yield the most.

We have shown how data mining can be successfully applied for this purpose. This work clearly shows that the factors that can be controlled by the farmer can be modeled quite well into a data mining problem such that we can obtains answers to the most common questions, by revealing patterns of interest in otherwise unorganized data.

6.2 Future Work

The study is projected to continue with new tools to be incorporated to the current version of the system. Among these tools are the ones that will help farmers keep record of irrigation, fertilization and pest control dates and lead them for the next step they will take, thereby further maximizing crop yield. Such system design will allow researchers to access more and new data pertaining to each and every agricultural year so they can use this new data for analyses and further improvement, while also helping farmers plan every detail of their needs beforehand and maximize their crop yields.

REFERENCES

Akın, T., Yıldırım, C., Çakan, & H. (2014). Tarım ve Hayvancılıkta Bilişim Tabanlı Karar Destek Sistemleri. Akademik Bilişim Konferansı Bildirileri, 659-663.

ASP.NET (n.d.). Retrieved June 17, 2018, from http://en.wikipedia.org/wiki/ASP.NET.

Benedetti, A., Catenaro, A., & Piersimoni, B. (2002). Remote sensing support to crop yield forecast and area estimates generalızed software tools for crop area estımates and yıeld forecast. *Indian Agriculture Research Journal*, *8*, 29-33.

Bonczek, R.H., Holsapple, C.W., & Whinston, A.B. (1980). The Evolving Roles of Models in Decision Support Systems. *Decision Sciences, 11*(2).

Breiman, L., Friedman, J., Olshen, R. & Stone C. (1984) Classification and Regression Trees. *Wadsworth Int. Group.*

De Geronimo, E., Aparicio, V.C., Barbaro, S., Portocarrero, R., Jaime, S., & Costa, JL. (2014). Presence of pesticides in surface water from four sub-basins in Argentina. *Chemosphere 107*, 423-431.

Dhanabhakyam, M., & Punithavalli, M. (2011). A Survey on Data Mining Algorithm for Market Basket Analysis. *Global Journal of Computer Science and Technology*, *11* (11), 22-28.

Excel (n.d.). Retrieved June 17, 2018, from http://en.wikipedia.org/wiki/EXCEL.

Evren, R., & Ülengin, F., (1992). Yönetimde Karar Verme. *İstanbul Teknik Üniversitesi Yayınları,* N.1478.

Fayyad, U., Piatesky–Shapiro, G., & Smyth, P. (1996). Data mining to knowledge discovery in databases. *AI Magazine*, 50-67.

Gorry, A.G., & Morton, S.S. (1971). Framework for Management Information Systems. *Sloan Management Review, 13*(1).

Ganesh, S.H., & Jayasudha. (2015). Data Mining Technique to Predict the Accuracy of the Soil Fertility. *International Journal of Research in Engineering and Technology, 4* (7)*, 33*0-333.

İzmir/ Menemen Provincial Directorate of Agriculture, (n.d.) - *Amounts of barley crop yields of a 10-year period.*

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *App. Statist, 29* (2), 119-127.

Keen, G.P., & Morton, M.S. (1978). *Decision Support Systems:An Organizational Perspective.*

Keen, P.G. (1980). Decision support systems: a research perspective. In: *Decision support systems : issues and challenges.* Oxford, New York.

Kleindorfer, P., Kunreuther, H., & Schoemaker, P. (1993). *Decision Sciences: An Integrative Perspective.* Cambridge: Cambridge University Press.

Laurance, W.F., Sayer, J., Cassman, K. (2014). Agricultural expansion and its impacts on tropical nature. *Trends in Ecology & Evolution 29*, 107-116.

Masters, W.A., Djurfeldt, A.A., De Haan, C., Hazell, P., Jayne, T., Jirström, M. et al. (2013). Urbanization and farm size in Asia and Africa: Implications for food security and agricultural research. *Global Food Security 2*, 156-165.

Microsoft SQL Server Management Studio (n.d.). Retrieved June 17, 2018, from http://en.wikipedia.org/wiki/SQL_Server_Management_Studio.

Phoksawat, K., & Mahmuddin, M. (2016). Ontology-Based Knowledge and Optimization Model for Decision Support System to Intercropping. *IEEE*.

Piatetsky-Shapiro, G., Brachman, It., Khabaza, T., Kloesgen, W., & Simoudis, E. (1996). An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications, *In Proceedings of KDD-96, Menlo Park, CA: AAAI Press.*

Rotz, C.A. & Coiner, C.U. (2004) *Integrated Farm System Model: Reference Manual. USDA Agricultural Research Service, University Park, PA.*

Qingyuan, M., Zhenghua, C., Chao, Z., & Zhen, Y. (2007). Management Decision-Making Support System of Percision Agriculture Based on CNCS. *IEEE*, 819-822.

QUINLAN, J.R. (1986). Induction of Decision Trees, *Machine Learning, 1*, 81-106.

Salado-Navarro, L.R., Sinclair, T.R., (2009). Crop rotations in Argentina: analysis of water balance and yield using crop models. *Agric. Syst.* 102, 11–16.

SPSS Clementine (n.d.). Retrieved June 17, 2018, from http://www.the-data-mine.com/Software/SPSSClementine.

Taechatanasat, P., & Armstrong, L. (2013). Decision Support System Data for Farmer Decision Making, Edith Cowan University Research Online ECU Publications.

Turban, E. (1995). *Decision Support and Expert Systems: Management Support Systems.* N J, Englewood Cliffs.

Turkey Meteorological Service Department, (n.d.) - *The weather data of a 10-year period (Harmonized dates with Barley products).*

Urtubia, A., Perez-Correa J.R., Soto, A., & Pszczolkowski, P. (2007). Using data mining techniques to predict industrial wine problem fermentations. *Food Control*, *18*, 1512– 1517.

Veenadhari, S., Mishra, B., & Singh, C.D. (2011). Soybean Productivity Modelling using Decision Tree Algorithms. *International Journal of Computer Applications*, *27* (7), 11-15.

Visual Studio (n.d.). Retrieved June 17, 2018, from http://en.wikipedia.org/wiki/Microsoft_Visual_Studio.

Witten, I.H., Frank, E., Hall, M.A. (2011). Data Mining In *Practical machine learning tools and techniques* (3rd ed.) (191) Morgan Kaufmann, San Francisco.

Zhu, Z., Zhang, R., & Sun, J. (2009). Research on GIS-based Agriculture Expert System. *World Congress on Software Engineering IEEE*, 252-255.