Comparison of Artificial Neural Networks & Decision Trees Methods on Health Insurance Data

Pelin KASAP[1] Burçin Şeyda ÇORBA[2] Tuba ÇELEBİ[3]
[1,2,3]Department of Statistics
[1,2,3]Ondokuz Mayis University, Turkey

Abstract— In this study, the insurance expenditures of the people who have health insurance are evaluated. We present Business understanding and Data understanding stages of The Cross-Industry Standard Process for Data Mining (CRISP-DM) process to investigate attributes influencing health insurance. In Data preparation stage of CRISP-DM process, Artificial Neural Networks (ANNs) method, C4.5 and Classification and Regression Trees (CART) Algorithms are used. When the applied methods are compared, it is shown that C4.5 is the most appropriate model with high accuracy rate.

Key words: Artificial Neural Networks Model; CART Algorithm; C4.5 Algorithm; CRISP-DM; Data Mining

I. INTRODUCTION

One of the aims of the institutions that collects data from different sources is to transform the data, which are useless by themselves, to useable information. Technological developments in recent year makes collecting and storing vast amount of data easier. Developments in data collecting tools and data base technology requires vast amount of data to be collected and analyzed in information stores [8]. The aim of the Data Mining (DM) is to find significant information which cannot be found using conventional methods among vast amount of data [17].

DM is a process of discovering various models, summaries, and derived values from a given collection of data [13]. The general experimental procedure adapted to DM problems involves the following steps:

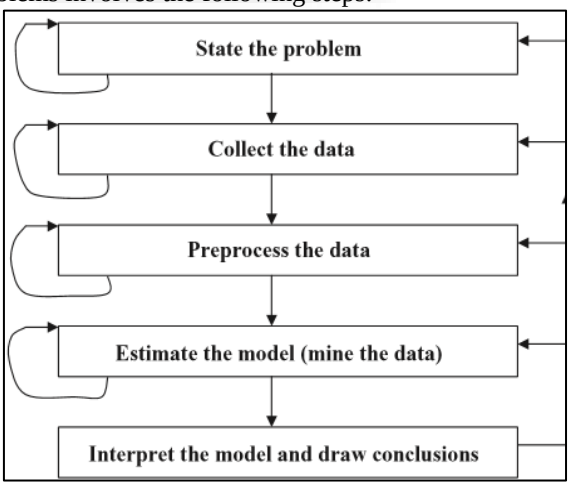

Fig. 1: The Data Mining Process [13]

All phases, separately, and the entire data - mining process, as a whole, are highly iterative, as shown in Figure 1. A good understanding of the whole process is important for any successful application. No matter how powerful the data - mining method used in step 4 is, the resulting model will not be valid if the data are not collected and preprocessed correctly, or if the problem formulation is not meaningful [13].

Several different algorithms are used to perform various tasks in DM. These algorithms try to find the appropriate model for the data. The algorithms examine the data and select the model that best suits its properties. DM tasks are divided into two categories as "predictive" and "descriptive" models [9].

In the predictive models, it is aimed to develop a model based on the known data and to estimate the result values for the unknown datasets by using this model [4]. Descriptive models explore the hidden common features and relationships in the data [1]. The purpose of descriptive models is to identify patterns and relationships in large datasets, to relate them to data [13]. Methods like classification, regression and time series analyses are predictive methods while methods like clustering, summarization and association rule mining are descriptive methods.

Both predictive and descriptive methods are supported by DM techniques.

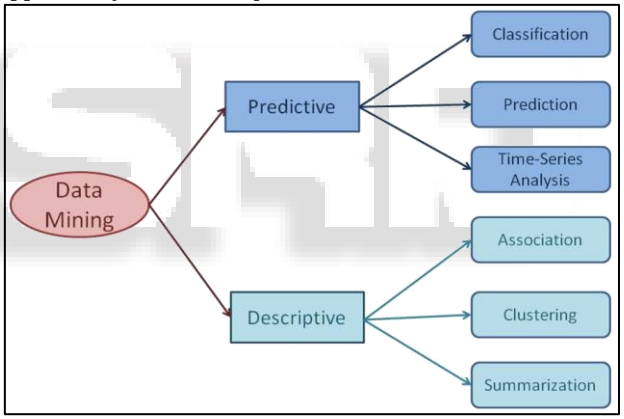

Fig. 2: Data Mining Models [31]

DM techniques are used to find forms in large datasets [12]. Classification method makes use of mathematical techniques such as decision trees, linear programming, artificial neural networks and statistics [20].

There are numerous DM research articles on data in health and many other areas. Chae et al. (2001) have examined DM algorithms to provide policy information for hypertension management using the Korean Health Insurance Company database and to show how health outcomes estimates can be used. They have compared the performance of logistic regression and two decision tree algorithms that CHAID (Chi-squared Automatic Interaction Detection) and C5.0. According to their study, the CHIAD algorithm performed better than the logistic regression in predicting hypertension, and it is concluded that C5.0 has the lowest predictive power.

Kaur and Wasan (2006) briefly have examined the potential use of classification based DM techniques such as Rule based, decision tree and Artificial Neural Network to massive volume of healthcare data. They have presented a

case study of application of DM and analyze data of children with Diabetes mellitus and Diabetes insipidus. In their study has been applied the concept of classification method.

Anyanwu and Shiva (2009) have compared of performance of the commonly used decision tree algorithms using Statlog data sets. The Statlog data set (Michie et al, 1994) are a large scale data set including various disciplines like financial (Australian and German data sets); transportation (Vehicle data sets), science (Shuttle data set) and health (Heart data set). The experimental analysis showed that SPRINT and C4.5 algorithms have a good classification accuracy compared to other algorithms used in their study.

Tomar and Agarwal (2013) have explored the utility of various DM techniques such as classification, clustering, association, regression in health domain. In their paper, they have present a brief introduction of these techniques and their advantages and disadvantages. Their survey have highlight applications, challenges and future issues of DM in healthcare. Also, in their paper have discussed recommendations regarding the suitable choice of available DM technique.

Singh and Gupta (2014) have examined the three most commonly used decision tree algorithms, Iterative Dichotomizer 3 (ID3), C4.5 and Classification and Regression Trees (CART), to understand their use and scalability in different properties. They have explained the basic characteristic of these algorithms. They have stated the answer to which algorithm should be used to achieve the best result in a given dataset type.

Kasap and Çorba (2017) have applied the stages of the CRISP-DM process to investigate attributes influencing prices of housing and have used CART, C5.0 decision tree algorithms and Neural Networks model in modeling phase. It is deduced that C5.0 model is the most appropriate model with the highest validation rate.

In this study, decision trees such as CART and C4.5 and artificial neural networks (ANNs) are used as classification method on health insurance data. The outline of the paper is as follows: In Section 2, CRISP-DM process is presented and decision tree algorithms such as CART, C4.5 and ANNs model are introduced. In Section 3, an application of health insurance data are given. Finally, Conclusions are given in Section 4.

II. MATERIALS & METHODS

A. The Cross-Industry Standard Process for Data Mining Process

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) is a popular methodology for increasing the success of DM projects. According to CRISP-DM, DM process has a life cycle composed of six stages. The order of these stages can be changed, i.e. not fixed [7].

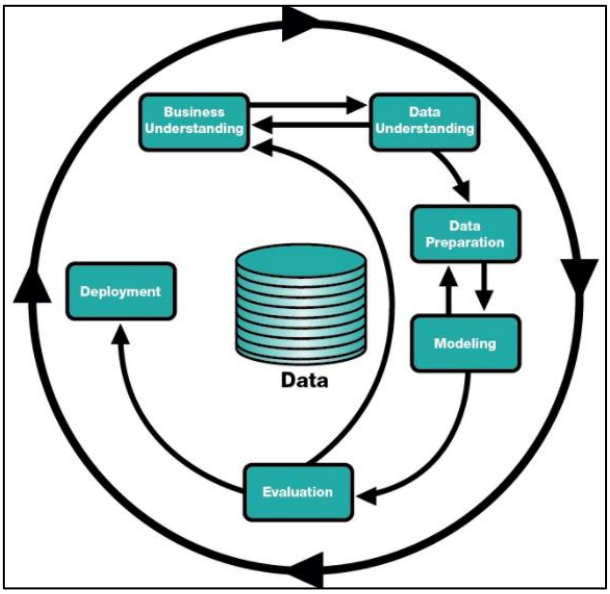

Fig. 3: The CRISP-DM Process Model [30]

B. Data Mining Techniques

DM techniques are used to find forms in large datasets. Forms in the database are described by relations between the features [12]. There are a few important DM techniques have been developed and used in DM projects recently including association, classification, clustering, prediction and sequential patterns etc., are used for knowledge discovery from databases [20].

1) Decision Trees

Decision trees are DM methodologies applied in many real world applications as a powerful solution to classification problems. There is a large number of decision tree induction algorithms described primarily in the machine learning and applied statistics literature [13]. Decision tree algorithms (ID3, C4.5, C5, CART, etc.) were originally intended for classification. A decision tree is a flow chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions [18;29]. During the late 1970s and early 1980s, Quinlan (1986), are searcher in machine learning, developed a decision tree algorithm known as ID3. Quinlan (2014) later presented C4.5 (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared. Breiman et al. (1984) published the book CART, which described the generation of binary decision trees [10].

a) Classification and Regression Trees Algorithm

The CART (Classification And Regression Trees) algorithm is a widely used statistical procedure for producing classification and regression models with a tree-based structure [10]. The 1984 monograph, CART, coauthored by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone (BFOS), represents a major milestone in the evolution of artificial intelligence, machine learning, nonparametric statistics, and DM [23;28]. The CART mechanism is intended to produce not one tree, but a sequence of nested pruned trees, each of which is a candidate to be the optimal tree.

An important practical property of CART is that the structure of its classification or regression trees is invariant with respect to monotone transformations of independent

variables. One can replace any variable with its logarithm or square root value, the structure of the tree will not change. One of the disadvantages of CART is that the system may have unstable decision trees. Insignificant modifications of learning samples, such as eliminating several observations, could lead to radical changes in a decision tree: with a significant increase or decrease in tree complexity are changes in splitting variables and values [13].

CART uses GINI Index to determine in which attribute the branch should be generated [19]. The Gini index is used to select the feature at each internal node of the decision tree [13].

The use of a cross-validated score function distinguishes CART from most other DM algorithms based on tree models [11].

b) C4.5 Algorithm

The most important part of the C4.5 algorithm is the process of generating an initial decision tree from the set of training samples. As a result, the algorithm generates a classifier in the form of a decision tree: a structure with two types of nodes - a leaf indicating a class, or a decision node specifying some tests to be carried out on a single attribute value, with one branch and a subtree for each possible outcome of the test [13].

The C4.5 algorithm uses the concept of entropy to find the most distinctive variable in the classification. Entropy is used to measure the uncertainty and randomness within a data set. The goal of C4.5 tree induction is to ask the right questions so that this entropy is reduced [28].

2) Artificial Neural Networks Model

Artificial Neural Networks (ANNs) is an abstract computational model of the human brain. Although the term ANNs is most commonly used, other names include "neural network", parallel distributed processing (PDP) system, connectionist model, and distributed adaptive system. ANNs are also referred to in the literature as neurocomputers.

A neural network, as the name indicates, is a network structure consisting of a number of nodes connected through directional links [13].

In most cases an ANNs is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data storing firms are harvesting information from datasets in the process known as DM. The difference between these data storages and ordinary databases is that there is actual manipulation and cross fertilization of the data helping users makes more informed decisions [21].

The model of neural network is divided into three types such as feed-forward network, feedback network and self-organization network [3].

III. APPLICATION

In this chapter of the study, a data set about health insurance will be examined. The stages of CRISP-DM process to investigate attributes influencing health insurance are completed. These stages are as follows:

A. Business Understanding

In this chapter of the study, a data set about health insurance will be examined. The variables of this data set are defined as below:

1) Register No

It shows the register number of the insuree; Payment date: This is the date which person, that received treatment, made his/her payment on; Service: It shows the service insuree got. Price: It shows the price taken of the service; State: It shows whether insuree is retired or working; Birth date: It shows the birth date of the person that received treatment; Gender: It shows the gender of the person that received treatment; Relation: It shows the relativeness level of the person that received treatment to the insuree.

B. Data Understanding

In total, 10,000 registered insurance data belong to 3,552 insuree. The reason for this is that the mother, spouse and children of the insuree can also use the same insurance so more than one person can come for one insurance. Descriptive variables of 8 variable belonging to 10,000 registries, frequency tables, summary graphs and cross tables are formed.

When the relationships of people that got service for insurance are examined it is seen that 45 times mother, 5 times father, 2690 times children, 991 times spouse and 6269 times himself/herself got serviced. Also, the youngest of them is a baby below 1 years old and the oldest one is 90 years old. When we examine gender distribution by total number of services, of the 10,000 people receiving total service, 37.09% were male and 62.91% were female.

When we examine the gender distribution of insured persons, of the 4,753 persons, 39.28% were male and 60.72% were female.

When we examine percentage distribution of insured persons by state variable, of the 4,753 people, 84.77% are employees, 14.92% are retired and 0.32% are retired employees. When we examine distribution of insured persons by relation variable, There are 4,753 people that benefited from insure. 64.42% of them is insured person, 25.31% of them is insured person's children, %9.66 of them insured person's spouse, 5.7% of them insured person's mother and 4% of them is insured person's father.

When we examine the price variable, the minimum value is 0 TL and the maximum value is 42,648.5 TL. The average price was found as 139.2 TL.

	Mother	Father	Children	Spouse	Himself/Herself
Working	97.55	139.46	99.48	114.22	142.55
Retired	2213.47	6002.05	95.67	123.52	202.96
Retired Employee			78.51	90.57	205.40

Table 1: Total Price by Status & Relation Variables

According to Table 1, the highest expenditure per capita was made by retired fathers with the amount of 6002.05 TL and the lowest expenditure was made by the children of retired employees with 78.51 TL. No expenditure was made for the parents of retired employees.

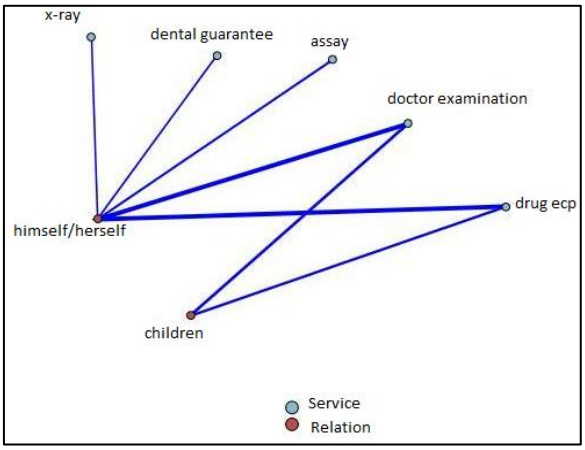

Fig. 4: The Relationship between Service & Relation Variables

According to Fig. 4, the insured person himself has come for the most doctor examination, drug ecp, assay, dental guarantee and x-ray. The children of the insured people has come for the doctor examination and the drug ecp.

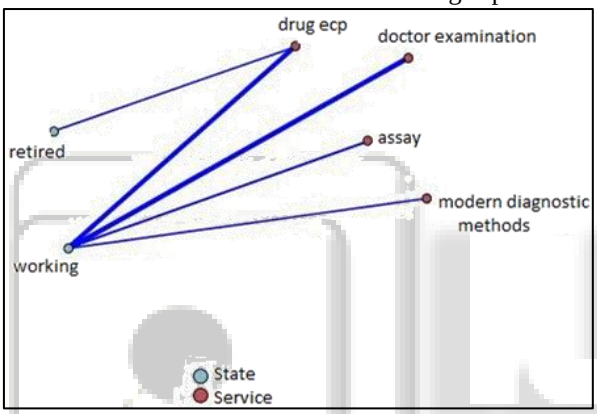

Fig. 5: Relationship between State & Service Variables

According to Fig. 5, when we look at the service situation of the insured persons, as employees have benefited from doctors' offices, drug ECPs, assay and modern diagnostic methods, mostly, retirees have benefited mainly from drug ECP.

C. Data Preparation

Data are examined one by one and inconsistences caused by varied reasons are tried to be removed. In this period, data of health insurance is made suitable for DM techniques. Cleaning and transforming of 10,000 data is conducted.

The data were filtered as "non-empty" and divided into 9 groups. The groups were determined as the response variable and the most appropriate data were assigned to the blank by using the classification methods CART, Artificial Neural Networks and C4.5. The response variable was selected as "group" and as input variables as "service, relation, status, gender".

Firstly, CART method was applied to the data. The accuracy rates of the model are as follows:

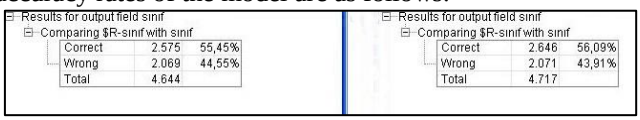

Fig. 6: Accuracy Ratio of Model Created by CART Method

According to Fig. 6, the average accuracy of CART model was found to be 55.77%.

Secondly, ANNs method was applied to the data. The accuracy rates of the model are as follows:

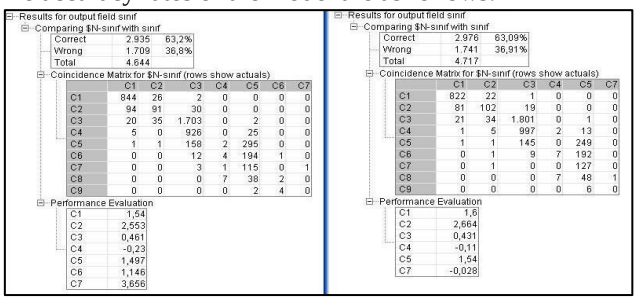

Fig. 7: Accuracy Ratio of Model Created by ANNs Method

According to Fig. 7, the average accuracy of ANNs model was found to be 63.145%.

Finally, C4.5 method was applied to the data. The accuracy rates of the model are as follows.

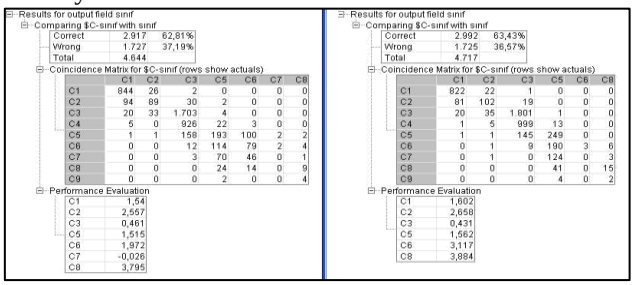

Fig. 8. Accuracy Ratio of Model Created by C4.5 Method

According to Fig. 8, the average accuracy of C4.5 model was found to be 63.12%.

When the accuracy rates of the applied methods are compared, the values of the ANNs and C4.5 models are approximately the same and so the importance of the variables in the model has examined to decide.

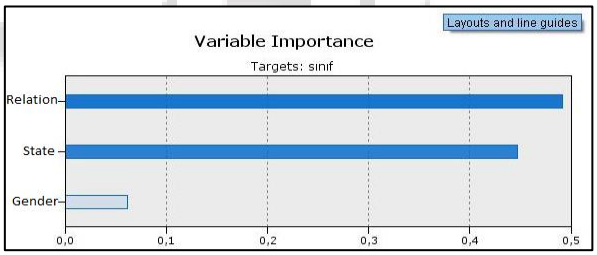

Fig. 9: Importance Level of Variables in the Model According to the ANNs Method

According to Fig. 9, the significance of the variables in the model was 0.492 for the relation, 0.446 for the state and 0.062 for the gender.

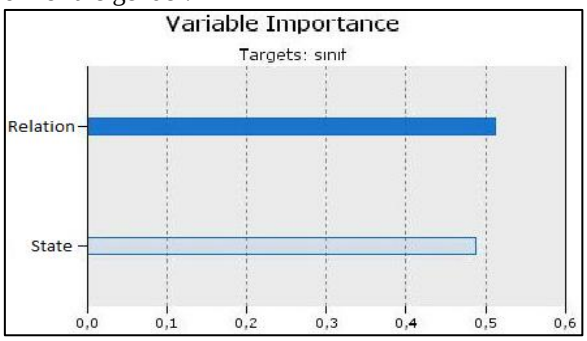

Fig. 10: Importance Level of Variables in the Model According to the C4.5 Method

According to Fig. 10, the significance of the variables in the model was 0.5118 for relation and 0.4882 for the state.

When the significance levels of the variables were compared for the methods, C4.5 method was chosen to assign the most appropriate data to the "vacancies". The data were filled with C4.5 method and adapted to the modeling phase.

D. Modeling

The model is applied to the entire set of data we are working on. The answer variable is "service" and input variable is "price, status, gender, relation, age". Decision tree model CART, Artificial Neural Networks and C4.5 were applied.

E. Evaluation

The data set which we work on is divided to two part. While the half of the data is used to form the model the other half is used to test the model. The data set which is used to form the model in the first stage is used to test the model in the second stage and the data set which is used to test the model in first stage is used to form the model in the second stage. Obtained results are given below.

According to the results of the analysis, the average accuracy rate of the applied CART model is 55.77%; the average accuracy rate of the applied ANNs model is 63.145%; the average accuracy rate of the applied C4.5 model is 63.12%. According to this result, the values of the ANNs and C4.5 models are approximately the same. However, considering the significance of variables for ANNs and C4.5, C4.5 method is chosen as the most suitable model.

IV. CONCLUSIONS

In this study, ANNs method and decision trees were applied to extract the information stored in the health insurance data. CRISP-DM process containing 6 steps of DM were applied to the data and the results have been achieved. While classifying in modeling phase after the first three stages, decision tree algorithms such as CART, C4.5 and ANNs model have applied. When the results were evaluated according to accuracy rates and the significance of variables, C4.5 model is chosen as the most suitable model.

REFERENCES

[1] Adriaans and Zantinge, Data Mining, Longman, Harlow: Addison Wesley, 1996.
[2] Anyanwu, M. N., & Shiva, S. G. (2009). Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security, 3(3), 230-240.
[3] Arif, M., Alam, K. A., & Hussain, M. (2015). Application of data mining using artificial neural network: survey. International Journal of Database Theory and Application, 8(1), 245-270.
[4] Berry, M. & Linoff, G., S., (2002), The Art and Science of Customer Relationship, Industrial Management & Data Systems.
[5] Breiman, L. J. Friedman, R. Olshen, and C. Stone, 1984: Classification and Regression Trees.
[6] Chae, Y. M., Ho, S. H., Cho, K. W., Lee, D. H., & Ji, S. H. (2001). Data mining approach to policy analysis in a health insurance domain. International journal of medical informatics, 62(2-3), 103-111.
[7] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.
[8] Dener, M., Dörterler, M., & Orman, A. (2009). Açık Kaynak Kodlu Veri Madenciliği Programları: Weka'da Örnek Uygulama. Akademik Bilişim, 9, 11-13.
[9] Dunham, M. H. (2006). Data mining: Introductory and advanced topics. Pearson Education India.
[10] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
[11] Hand, D., Mannila, H., & Smyth, P. (2001). Principles of data mining. 2001. MIT Press. Sections, 6, 2-6.
[12] Hegland, M. (2001). Data mining techniques. Acta numerica, 10, 313-355.
[13] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.
[14] Kasap,P and Çorba, B.Ş. (2017). The Review of Attributes Influencing Housing Prices using Data Mining Methods, International Journal of Sciences: Basic and Applied Research (IJSBAR), Vol. 34, No. 1, pp 155-165.
[15] Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. Journal of Computer science, 2(2), 194-200.
[16] Michie, D., Spiegelhalter, D.J. and Taylor, C.C. Machine Learning, Neural and Statistical Classification, Ellis Horwood, 1994.
[17] Özkan M. & Boran, L. (2014). Veri Madenciliğinin Finansal Kararlarda Kullanımı, Çankırı Karatekin Üniversitesi İİBF Dergisi, 4(1), 59-82.
[18] Parashar, H. J., Vijendra, S., & Vasudeva, N. (2012). An efficient classification approach for data mining. International Journal of Machine Learning and Computing, 2(4), 446.
[19] Patil, N., Lathi, R. and Chitre V. 2012. Comparison of C5 & CART Classification algorithms using pruning technique, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol.1, Issue 4.
[20] Raval, K. M. (2012). Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 2(10).
[21] Singh Y. and Chauhan, A.S. 2009. Neural networks in data mining, Journal of Theoretical and Applied Information Technology, Vol.5, No.1, pp.37-42.
[22] Singh, S., & Gupta, P. (2014). Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. International Journal of Advanced Information Science and Technology (IJAIST), 27(27), 97-103.
[23] Steinberg, D., & Colla, P. (2009). CART: classification and regression trees. The top ten algorithms in data mining, 9, 179.
[24] Taylor, C., Michie, D., & Spiegalhalter, D. (1994). Machine Learning, Neural and Statistical Classifiers.
[25] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.
[26] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[27] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.
[28] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.
[29] Vanitha, A., & Niraimathi, S. (2013). Study on decision tree competent data classification. International Journal of Computer Science and Mobile Computing (IJCSMC), 2, 365-370.
[30] http://crisp-dm.eu/home/about-crisp-dm
[31] www.wideskills.com

