European Journal of Business and Social Sciences, Vol. 6, No. 02, **May 2017.** **P.P. 54 - 62**
URL: *http://www.ejbss.com/recent.aspx-/*
ISSN: 2235 -767X

DETERMINATION OF FACTORS AFFECTING HAPPINESS LEVEL BY CLASSIFICATION TREE TECHNIQUE

Seda BAĞDATLI KALKAN Ph.D
Yasemin BAHAR YÜCEL [1]

ABSTRACT

Classification and regression tree techniques have been often used especially in the social sciences in recent years. Since these techniques do not require any statistical assumption, they make contribution to many researches. The data of life satisfaction questionnaire, applied by Turkish Statistical Institute (TSI) to 9397 people in Turkey in 2015, was used in this study. The aim of the study is to determine the factors that affect persons' happiness level by using the classification trees technique. The classification trees were created using Classification and Regression Tree (CART) and Chi-Squared Automatic Interaction Detector (CHAID) algorithms. The created trees were construed comparatively.

Key words: Classification and Regression Tree (CART) and Chi-Squared Automatic Interaction Detector (CHAID), Life Satisfaction, Happiness Level

[1] Master Student in İstanbul Commerce University, Statistics Department

1. INTRODUCTION

Classification and Regression Tree techniques are among commonly used methods of analysis especially in social sciences due to fact that they do not require any assumption, produce easy-to-understand results, and have different algorithms. The method is called as a regression tree when the dependent variable is continuous, and a classification tree when the dependent variable is categorical.

The main purpose of the Classification and Regression Tree technique is to divide the main data matrix (independent variables matrix) into the homogeneous subgroups according to the dependent variable. In creating subgroups, data is offered in a hierarchical order in the form of a branched tree. The tree indicates the dependent variables with best divisions in the intermediate nodes in the figure. The critical value of thedividing dependent variables are indicated in the branches of these nodes. The leaves indicate the values of the dependent variable. There are routes extending from the root node (first nodal point) to the leaves (last nodal point). Throughout these routes, the division between classes is maximized and the variation within each class is minimized (Özkan Y., 2016).

CART and CHAID techniques are among most commonly used decision trees algorithms. Both techniques allow for the inclusion of the dependent and independent variables in the analysis as continuous or categorical variables (Doğan, 2003). Besides, both techniques are not responsible for the assumptions of regression techniques (distributional assumptions, extreme values, etc.) since they are nonparametric methods (Nisbet, Elder, & Miner, 2009).

CART offers the independent variables, likely having an effect on the dependent variable, in the form of a tree structure, based on the importance level. CART generates a classification tree if the dependent variable is categorical, and a regression tree if the dependent variable is continuous (Breiman, Friedman, Olshen, & Stone, 1984).

The most important features of CART and CHAID algorithms can be summarized as having the ability of simultaneous model inclusion of continuous and categorical data , of comprehensibly indication of the independent variable (s) that are effective on dependent variables on a tree diagram, and of easier visual evaluation of tree diagram than other analysis results.

The purpose of this study is to determine the factors affecting persons' happiness level by the classification tree technique.TSI has been making an annual general life satisfaction research throughout Turkey since 2003. The research focuses on the level of happiness, sources of happiness, level of hope and level of satisfaction with public services.

The concept of happiness is a concept that has been considered for many years. According to the current Turkish dictionary of the Turkish Language Society, happiness is defined as feeling proud due to continuously achieving the whole desire. Happiness has been investigated since 300s B.C. From a philosophical point of view, it is seen that happiness is associated with good person and bad person distinction (Babaoğlu, 2008).

The concept of happiness, which has been intensively discussed in the world, has recently entered into the fields of economics as well as the field of medicine, psychology and sociology. From the point of view of those people who express that happiness is associated with the combination of the variables such as demographic, economic, physical environment, social environment and socio-economical status of our current country; there is a great relationship between happiness and satisfaction with life (Şeker, 2009).

Life satisfaction means the level of individual's positive attainment as a result of his/her complete assessment of his/her quality of life. Pleasure, which can also be expressed as life satisfaction, and life pleasure with enjoyment inclusive express how much an individual is pleased with his / her life (Güler & Emeç, 2006). As seen, happiness and life satisfaction are interpenetrated concepts, therefore, life satisfaction data were used in this study.

The purpose of this study is to determine the factors that affect persons' happiness level by classification tree technique.The factors affecting happiness levels were determined using CART and CHAID algorithms, and the results were analysed comparatively.

2. METHOD

2.1 Classification Tree:

The main purpose of the classification studies is to create a correct classifier and reveal the underlying estimator structure of the problem.Hence, in order to determine in which class an observation will be included, it is necessary to understand that which variables or interactions between variables will be useful to solve the problem. The most important criterion in the classification process is not only to produce an accurate classifier for the data set, but also to shed light on the data in the estimator structure and understand them. In other words, the previous experiences of competent persons for the related data is also important. Although these two criteria generally are equally important, occasionally one has a greater influence than the other (Yohannes & Webb , 1999).

The classification tree consists of three stages. These are "the creation of the tree", "pruning" and "the selection of the most appropriate tree structure". Classification and regression tree algorithm based on the principle of creating maximum level homogeneous subclasses (trees), determines the maximum possible number of sub-trees in the "tree creation" section. But, it is necessary to select the trees, having important relations with the dependent variable, among the subtrees. Therefore, the second part of the algorithm, the "pruning" module, comes into play. Thus, after the pruning, the classification tree can be obtained by "the selection of the most appropriate tree structure".

In the classification tree method, Gini impurity measure is used when the decision is taken on whether the binary dependent variable is pure.

For a *t* -node, the Gini impurity index (*g(t)*) is determined by the following formula.

$$g(t) = \sum_{j \neq i} p\left(\frac{j}{t}\right) p\left(\frac{i}{t}\right)$$

Here, *i* and *j* are dependent variable categories.

Since the dependent variables in the models for species distribution consist of binary (on-off) categories, the equation for the index is as follows.

$$g(t) = 2p(1/t)p(2/t)$$

When all renewed codes in the node belong to only one category, the index value equals zero. Each variable in the independent variable set is evaluated to find the independent variable that will make the best estimate of a node, and the variable with the best value is selected. For any *t*-node; *s*, the candidate discriminator of a node, performs both the right side discrimination (t_R) and the left side discrimination (t_L) (Özkan K., 2012).

CART ALGORITHM

Cart Algorithm is the continuation of CARY Morgan and Sonquist's AID (Automatic Interaction Detection) named decision tree algorithm and was proposed by Breiman et al. in 1984 (Breiman, Friedman, Olshen, & Stone, 1984)**.** The CART algorithm, which can accept both numerical and nominal data types as input and predictive variables, can be used as a solution in classification and regression problems (Akçapınar Sezer, Bozkir, Yağız, & Gökçeoğlu, 2010).

Although the CART algorithm is not easily accepted due to having a tree structure, the number of its usage has increased in recent years.CART has a number of advantages when algorithms using its tree structure is reviewed. Its greatest advantage is that it does not require assumptions about independent variable values due to not being a parametric method. Therefore, as well as the variables can be numerical, they can also be categorical. By this means, the researcher will have gained time since not dealing with the conversion process for analysis. Even though the problem discussed contains hundreds of possible arguments, CART algorithm is an algorithm with the capability of searching all possible variables that can be divided (Oğuzlar, 2004).

The main purpose of the method is to create an error-free dataset classifier that yields the estimator structure for the problem anaylsed. The purpose of classification is to determine to which class a future value can be assigned, through the characterized tree. For this purpose, it is determined which variables or interactions between variables are necessary to estimate the best result.

In other words, the CART algorithm both finds the appropriate variable in tree branching, and also identifies how this variable can be divided into two groups in case of having more than two types of values (Silahtaroğlu, 2016).The separation process in this algorithm is performed according to the gini, towing index calculations for categorical dependent variables, the least squares index calculations for continuous variables (Akpınar, 2000).

CHAID ALGORITHM

CHAID (Chi-Squared Automatic Interaction Detector) algorithm was developed by G.V Kass in 1980. It is a very effective statistical method for dividing or tree creation. CHAID evaluates all independent variables using a statistical test as a branching criterion. It combines statistically homogeneous (unvarying) values in reference to dependent variable values, and it doesn't apply any operation to those that are different. Then, it selects the best independent variable to form the first branch of the tree, the branches of which includes homogeneous values, and this process continues in a repetitive manner until the tree is fully grown.The statistical test to be used as the branching criterion vary according to the level of measurement of the dependent variable. If the dependent variable is continuous, F test is used; if categorical, chi-square test is used. The first branch of the tree is formed with the variable having the smallest p value (Atılgan, 2011). The most important difference between CHAID algorithm and other algorithms is its derivation of multiple trees instead of binary trees (Türe, Tokatlı, & Kurt, 2009). With this method of analysis, continuous and categorical data can be simultaneously included in the same model (Kayri & Boysan, 2007). In other words, it is not necessary for all dependent and independent variables to be measured with the same type of scale (Koyuncugil & Özgülbaş, 2008). Therefore, CHAID analysis removes parametric and nonparametric distinction and features statistically semiparametric in the method algorithm (Kayri & Boysan, 2007).

In the CHAID algorithm, chi-square statistics are used, especially since the interrelationships and interactions of the independent variables are matters. Chi-square test statistics discuss the dependence between variables (Kayri & Boysan, 2007). As well as the CHAID algorithm,identifies the relationship between dependent variables and independent variables, it also reveals the interactions of independent variables (Kayri & Boysan, 2008).

3. APPLICATION

In this study, the interrelationships between the independent variables affecting happiness and the relationship between the dependent variable and the independent variables were investigated separately using the Classification Tree technique and the CART and CHAID algorithms.

Happiness was included as a dependent variable and the other 24 variables as independent variables in the study. First, the CHAID algorithm, and then the CART algorithm and classification trees were created, and the results were evaluated.At first, it is necessary to decide which tree is the most suitable in the classification tree. For each of two algorithms, three different trials were conducted in order to determine the appropriate tree. Firstly, 30% of the data set was taken as a test sample, 70% of that as a training sample. Secondly, % 50 of the data set was taken as a test sample, 50% of that as a training sample. Finally, % 70 of the data set was taken as a test sample, 30% of that as a training sample. In the models created with 3 different ratios, it was attempted to determine the appropriate tree through different node numbers. It was decided that the tree with highest correct classification rate is the most suitable tree.

It is seen that the most suitable tree created by the CHAID algorithm was the one when 50% of the data set was taken as a test sample, 50% of that as a training sample.

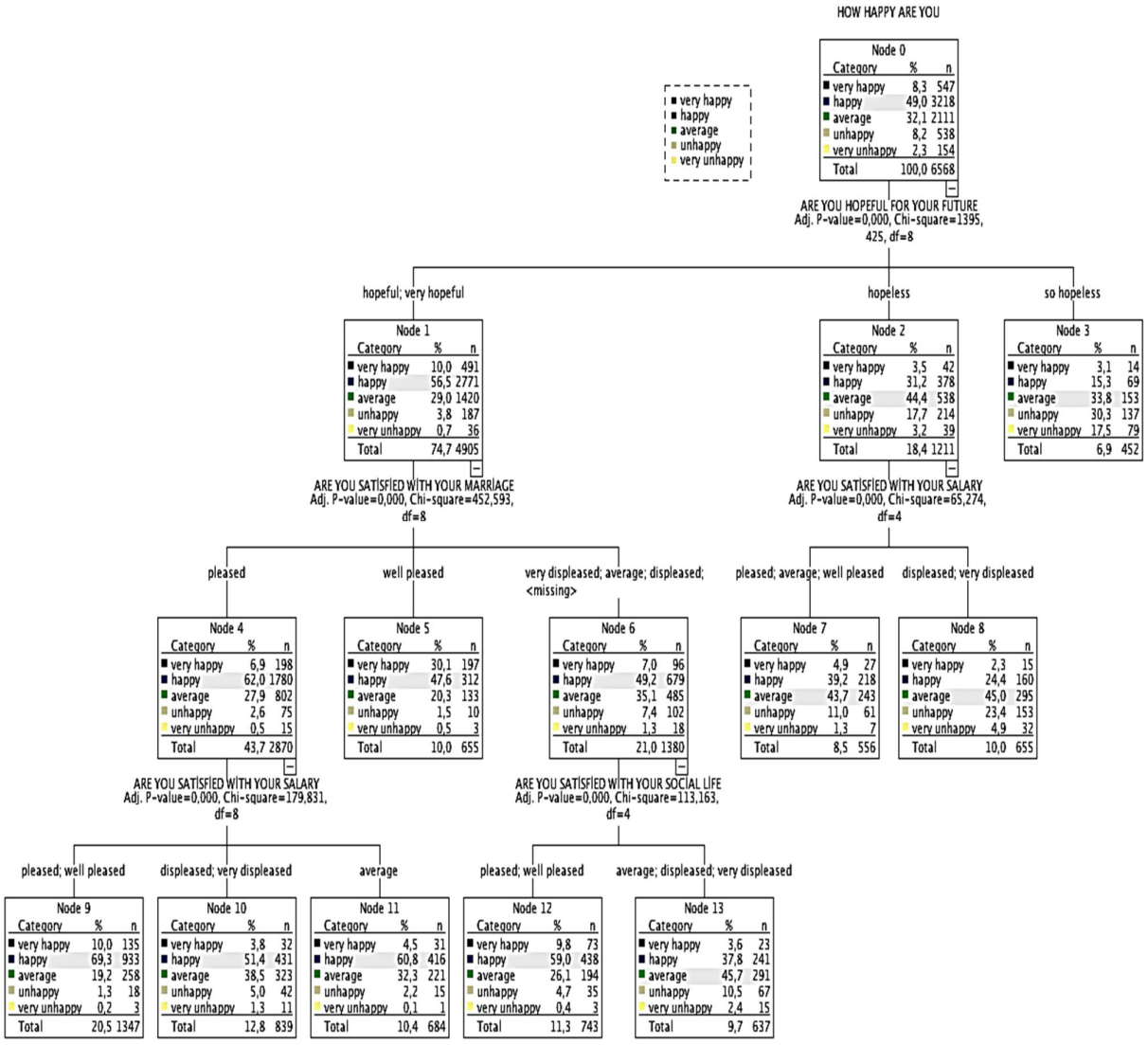

Table 1: CHAID Tree Diagram

When Table 1 is examined, it is concluded that the variables affecting happiness are being hopeful for the future, being satisfied with marriage, salary and social life, and that there is a interrelationship between these variables.

It was concluded that the most important variable affecting happiness is being hopeful for the future ($p=0.00$, $\chi^2=1395.425$,$sd=8$). This variable was classified as three nodes. According to this result, it is seen that about 57% of those who are hopeful and very hopeful for the future are happy. As hope level drops down, happiness rates also decrease.

It was concluded that the variable of being hopeful for the future is influenced by the variable of marital satisfaction ($p=0.00,\chi^2=452.593,sd=8$).It is seen that 62% of those who are very satisfied with their marriage are happy.

It was concluded that the variable of marital satisfaction is influenced by the variable of salary satisfaction ($p=0.00,\chi^2=179.831,sd=8$). It is seen that 69 % of those who are satisfied and very satisfied with their salary are happy.

It was concluded that the variable of marital satisfaction influence the variable of social life satisfaction of people with categories of medium-level, being unsatisfied, being absolutely unsatisfied ($p=0.00,\chi^2=113.163,sd=4$).

It was concluded that if the people who are medium-level satisfied, being unsatisfied, and being absolutely unsatisfied with their marriage, are satisfied with their social life ; %59 of them are happy.

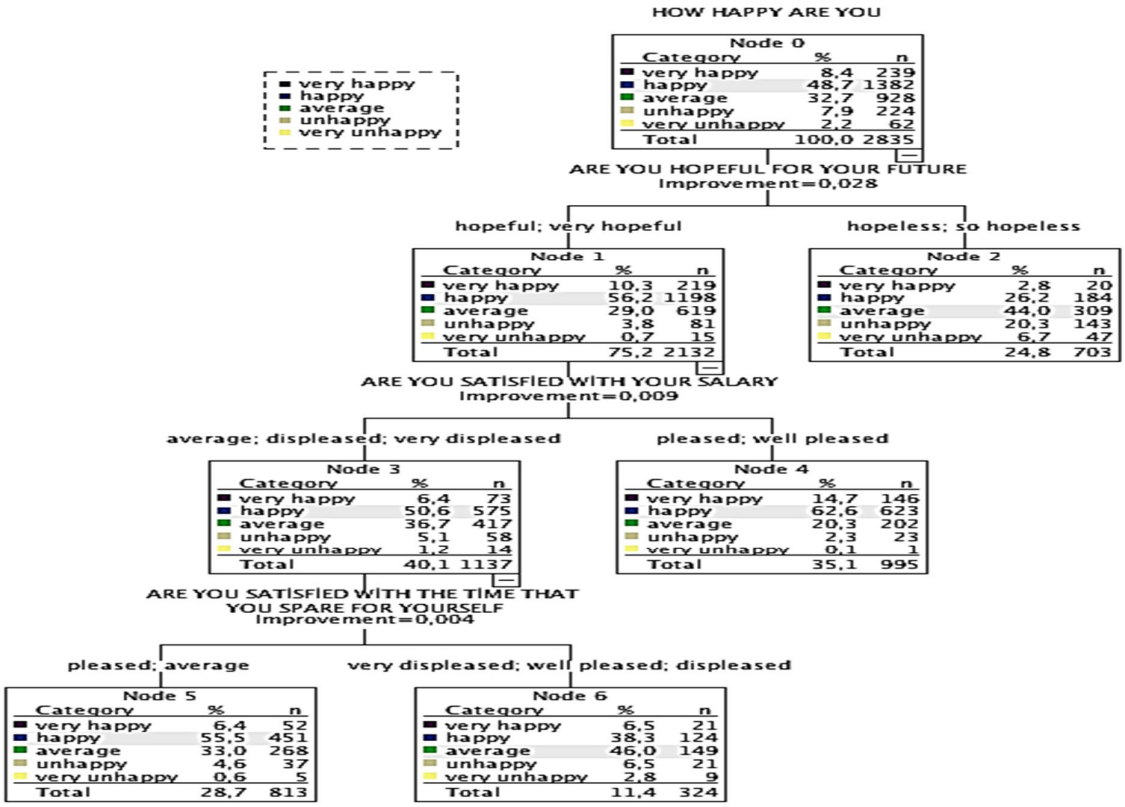

Table 2: CART Tree Diagram

The CART algorithm and the variables affecting happiness were reviewed in Table 2.

When Table 2 is examined, it is concluded that the variables affecting happiness are the variables of being hopeful for the future, being satisfied with salary and being satisfied with social life and that there is a interrelationship between these variables. It has been concluded that the most important variable affecting happiness is the variable of being hopeful for the future. It is seen that about 56% of those who are very hopeful and hopeful for the future are happy. Happiness levels decrease as hope level drops.

It has been concluded that the variable of being hopeful for the future influences the variable of salary satisfaction. 62 % of those who are satisfied and very satisfied with their salary are happy.

If the people who are medium-level satisfied, being unsatisfied, and being absolutely unsatisfied with their salary, are satisfied with the time they spared for themselves, %56 of them are happy.

4. CONCLUSION

In this study, the factors affecting the happiness levels of people were determined by classification trees in the direction of the data of the life satisfaction questionnaire conducted by TSI. Two different trees were created by using the CHAID and CART algorithms in classification trees.

The level of happiness in the study was taken as a dependent variable in the study. 24 independent variables were selected by taking competent persons' opinion. The Independent variables, considered to have statistically significant effect on the happiness variable, are indicated in the classification trees generated by both algorithms. According to the classification tree created by CHAID algorithm, it was concluded that the variables affecting happiness are being hopeful for the future, being satisfied with marriage, salary, and social life, and that there is an interrelationship between these variables. According to the classification tree created by the CART algorithm, it was concluded that the variables affecting happiness are being hopeful for the future, being satisfied with salary and social life, and that there is an interrelationship between these variables.

49% of the people who participated in the study are observed to be happy. It was concluded that the most important variable affecting happiness in both trees is being hopeful for the future.While the variable of being hopeful for the future is influenced by the variable of being satisfied with salary in the CART algorithm; it is influenced by the variable of being satisfied with marriage in the CHAID algorithm.

If the people who are not satisfied with their salary in the CART algorithm are not satisfied with the time they spared for themselves, their happiness levels are low. If the people who are not satisfied with their marriage in the CHAID algorithm are not satisfied with their social life, their happiness levels are low.

As a result, the tree creation and development methods of CART and CHAID algorithms are different from each other, differences between these two trees can be seen. Although, the nodes in the CHAID algorithm can be divided into more than two classes; the classes in the CART algorithm are divided only in binary. Therefore, the CHAID algorithm allows for attaining more comprehensive results; the CART algorithm allows for attaining more general results in the classification tree. The researcher can make his own decision on which algorithm is more appropriate.

References

1. Akçapınar Sezer , E., Bozkır, A. S., Yağız, S., & Gökçeoğlu, S. (2010). Karar Ağacı Derinliğinin CART Algoritmasında Kestirim Kapasitesine Etkisi:Bir Tünel Açma Makinasının İlerleme Hızı Üzerinde Uygulama. *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu.* Kayseri.
2. Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 1-22.
3. Atılgan, E. (2011). Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi İle İncelenmesi. *Yüksek Lisans Tezi*. Hacettepe Üniversitesi İstatistik Anabilimdalı.
4. Babaoğlu, H. (2008, 05 09). *VATAN*. 04 25, 2017 tarihinde m.gazetevatan.com adresinden alındı
5. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. *Chapman and Hall.*
6. Doğan, İ. (2003). Holştayn Irkı İneklerde Süt Verimine Etki Eden Faktörlerin Chaid Analizi İle İncelenmesi. *Ankara Üniversitesi Veterinerlik Fakültesi Dergisi*, 65-70.
7. Güler, B. K., & Emeç, H. (2006). Yaşam Memnuniyeti ve Akademik Başarıda İyimserlik Etkisi. *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi*, 129-149.
8. Kayri, M., & Boysan, M. (2007). Araştırmalarda Chaid Analizinin Kullanımı ve Baş Etme Stratejileri İle İlgili Bir Uygulama. *Hacettepe Üniversitesi Eğitim Bilimleri Dergisi*, 133-149.
9. Kayri, M., & Boysan, M. (2008). Bilişsel Yatkınlık İle Depresyon Düzeyleri İlişkisinin Sınıflandırma ve Regresyon Ağacı İle İncelenmesi. *Hacettepe Üniversitesi Eğitim Bilimleri Dergisi*, 168-177.
10. Koyuncugil, A. S., & Özgülbaş, N. (2008). İMKB'de İşlem Gören Kobilerin Güçlü ve Zayıf Yönleri:Chaid Karar Ağacı Uygulaması . *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi*, 1-21.
11. Nisbet, R., Elder , J., & Miner, G. (2009). Handbook of statistical analysis and data mining applications. *Elsevier*.
12. Oğuzlar, A. (2004). CART Analizi ile Hanehalkı İşgücü Anketi Sonuçlarının Özetlenmesi. *İktisadi ve İsari Bilimler Dergisi*, 79-90.
13. Özkan, K. (2012). Sınıflandırma ve Regresyon Ağacı Tekniği (SRAT) ile Ekolojik Verinin Modellenmesi. *Süleyman Demirel Üniversitesi Orman Fakültesi Dergisi*, 31-52.
14. Özkan, Y. (2016). *Veri Madenciliği Yöntemleri.* İstanbul: Papatya Yayıncılık.
15. Şeker, M. (2009). Mutluluk Ekonomisi. *Sosyoloji Konferansları Dergisi*.
16. Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları.* Papatya Yayıncılık.
17. Türe, M., Tokatlı, F., & Kurt, Ü. (2009). Using Kaplan-Meirer Analysis Together With Decision Tree Methods (C&RT,CHAİD,QUEST,C4.5 and ID3)In Determining Recurrence-Free Survial of Breast Cander Patient. *Expert Systems With Applications*, 2017-2026.
18. Yohannes , Y., & Webb , P. (1999). Classificatiob and regression trees, cart: A user manual for identifying indicators of vulnerability to famine and chronic food insecurity. *Microcomputers in Policy Research*.