Eurasian Journal of Business and Management, 6(2), 2018, 33-47
DOI: 10.15604/ejbm.2018.06.02.004

EURASIAN JOURNAL OF BUSINESS AND MANAGEMENT

www.eurasianpublications.com

CLASSIFICATION OF HOUSEHOLD USERS USING INFORMATION TECHNOLOGIES BASED ON C5.0 ALGORITHM

Guner Gozde Teksin
Corresponding Author: Istanbul Commerce University, Turkey
Email: ggteksin@gmail.com

Munevver Turanli
Istanbul Commerce University, Turkey
Email: mturanli@ticaret.edu.tr

Abstract

The use of machine learning and decision tree methods with respect to modelling of big data has been increasing gradually. The data mining has focused on some methods for obtaining the useful information from big data which has not been recognized. The data mining and statistics concentrates upon the identification of structure. The concept of machine learning which evolved out of data mining as a need consists of models and algorithms which analyzes big data and may make significant deductions from data by benefitting from statistics and software. It provides convenience to modelling the big data that the decision tree methods are nonparametric methods. In this paper, the classification tree was created based on C5.0 algorithm by using R programming language.

Keywords: Big Data, Data Mining, Machine Learning, Decision Trees, Classification Tree, C5.0 Algorithm, Cookies

1. Introduction

Along with the development of technology, there are many meaningful and insignificant data. It becomes challenging to access desired data as the data number and volume increase. This gives rise to think the efficiency of data in terms of studies and the access to desired information becomes costly. Therefore, the concept of big data has evolved, and this concept pioneered the formation of miscellaneous novelties. Along with that the information technologies have taken part in all fields of life; the data sizes have showed increases. The increases on data sizes have made the access to significant data difficult and increased the importance paid to data analysis. The big data also means that miscellaneous data in terms of relationship exist together. If we examine the big data as significant data and insignificant data, the effort of accessing to significant data will become more precious. Therefore, it is possible to refer the structure consisting of significant data as data mining. However, the analysis and interpretation of data mining has gone beyond the manpower. This has led to the need of qualified labor force. This need has made contribution to the emergence and development of ''data mining'' concept. The data mining is discovering

significant data from the data mining being the structure where there are significant and insignificant data, together. The computer algorithms were developed for analyzing and interpreting the significant data. These computer algorithms developed are called as machine learning. These concepts may be explained as analyzing the data and making estimations from significant data.

In the data mining, there is no limitation for the data to be analyzed. The data concerning to health, the research data, the data obtained from internet, the bank data, economic data may be the subject of data mining. The use of internet, specifically in the recent years, has given a significant increase to digital data amount.

In the recent years, one of issues increasing the digital data amount is cookies. The cookies are text files where the user information and passwords are stored at browsers when a website is visited during internet use. However, these text files contain several information such in big data. The cookies were first started to be used by Netscape in 1994. Their intended use is to analyze whether the user has revisited the websites that s/he entered. Today, it is aimed to obtain more information about the user behavior while not straying from intended use of cookies. When a website is visited, the user name and password is entered and the remember me button is clicked, the user name and password is not entered during revisit. The reason of this is that the information is stored by cookies. However, these text files may pose a risk in terms of personal data while providing convenience. The viruses spread on computer may access these data with miscellaneous software and steal the personal data. When the online advertisements are clicked, the display of similar advertisements is conducted by means of cookies. Some websites are using the cookies only for tracking. The cookies may be deleted from internet browsers and the software preventing the cookies may be used. The aim of this study is to investigate the profiles of individuals who are using and not using the software for preventing the cookies being the structures, which track the activities conducted on internet, based on the data of Turkish Statistical Institute Households Technology Use Survey 2016.

In this study, the information concerning to big data, data mining, machine learning, decision trees and C5.0 algorithm was given. In the application part, the individuals who are using and not using the software for preventing the cookies tracking the activities conducted on internet, were classified based on independent variables.

2. Definition and Components of Big Data

The big data is the datasets which conform to a specific order or not, show a fast increase and which are stored at the database. Along with the proliferation of use of digital media, the attention to big data has also increased. Along with the increase of data size every passing day, the access to significant and big data has become both costly and challenging in terms of time.

Although the size of data makes the access to significant information difficult, it is important in terms of specifically making prediction about future based on previous data. In this case, primarily the data are collected, and the data is accessed by investigating and analyzing relationships of data between each other.

The data directs not only the companies but also social structure. The need of new computer algorithms for processing new data is shown among the causes of this case. The concepts of data mining, machine learning and decision tree have become important along with the increase of data sizes every passing day.

The process of data mining started along with its use in computer inventory in the beginnings of 1950s. The data mining allows for accessing the desired information among big-scaled data. Therefore, many software and package program are used. Some criteria shall be provided for referring the stored data as big data. These criteria are characterized with 3V in 2000s. However, the big data are recently characterized as 6V. These components are named as follows: variety, velocity, volume, verification, value and volatility.

Variety: The data produced has no single structure and the data shall be interconvertible since they consist of formats obtained from miscellaneous environments.

Velocity: The production of big data has been gathering pace every passing day. These data reach to unbelievable dimensions in a brief time. Along with the fast increase of data, the

number and speed of transaction has been showing increase. Both the software and hardware speed are directly proportional to the increase of data.

Volume: The fast growth of data indicates the increase of data number required to be stored in the future. Therefore, the data stack evolved out depending on the increase of data brings a significant strategy on how it will be stored.

Verification: It is a significant data characteristic. The miscellaneous strategies may be developed depending on fast increase of data, but the safety of data is important in terms of all strategies. The following issues shall be investigated that: who produce, store and see the data and at time same whether they are right data or not.

Value: Obtaining significant and efficient data for studies following the stages of producing and processing data reflects the characteristic of data in terms of value (Aksoy *et al.* 2017).

Volatility: Since the data is a structure growing fast, it leads to a discussion of for how long it will work. Each data has a lifetime in big data. Thus, how soon the data to be obtained will become invalid shall be known.

The big data analysis is made in the countries where technology has been developing. It is used in important fields from defense industry to estimation of votes. Therefore, the data mining is needed for accessing and analyzing the significant data.

3. Concept of Data Mining and Intended Uses

The data mining means the deduction of potentially usable or significant data from unrecognized big data. This case contains many statistical researches such as clustering and determination of deviations on data. The data mining allows for investigating the relationships, changes and differences which occur on big data.

The aim of data mining is to investigate the data collected with some statistical methods. Within the process of data mining, the big-scaled data is made efficient and effective and the data patterns which have not been recognized on dataset previously, are evolved.

It is possible to see the data mining as the sequence of statistical methods. However, there are differences between statistics and data mining. In the data mining, the visually supporting qualitative models are created within the frame of rules of logic. Along with that the data mining is a field grounding on the human, it also uses the human and computer association as base.

If there are two or more variables and the relationship between these variables is desired to be explained, the data mining is used. Therefore, the ways of easy access to exact information was preferred.

The truth of requirement and necessity of accessing the right information came in sight in terms of the information pollution of our era and information produced along with the development of technology. The data mining allows for revealing the implicit information, which is hard to understand easily and accurately.

Today, the data mining is used in many fields. The data mining is used for determination of credit card fraud, life opportunities of credit card holders, deceptions on financial statements, sales and stock forecasting and investigation of factors affecting the Interbrand competition, evaluation of causes of profit margins, determination of sectoral risk profiles of stock exchange companies, performance analyses of GSM companies and cancerous cells.

3.1. Machine Learning in Data Mining

The machine learning is an artificial intelligence product which ensures automatic learning and development skill by benefitting from the experiences programming to systems in an explicit way. The machine learning allows for developing the computer algorithms which may access to data.

The aim of machine learning is to display the highest performance by the algorithm constituted from current data and computer algorithms used. The machine learning is based on the highest performance estimation. The machine learning algorithms carry the meaning of receiving necessary decisions in the data analysis. In the stage of taking decisions, the decision rules, decision trees and neural networks are used. In these methods, there are learning

strategies. These strategies are as follows: supervised machine learning, unsupervised machine learning, semi-supervised machine learning, reinforced machine learning and deep machine learning.

The supervised machine learning algorithms are the use of labeled samples for estimating the events in the future which have been learned in the past. In the supervised machine learning, the data interact with each other and in this model, it is aimed to produce the production of results obtained from analysis in a close interaction with the targeted results (Atalay and Celik, 2017)

In the unsupervised machine learning algorithms, it is tried to group the values which are close to each other by revealing the relationship between the data of which model has been constituted (Atalay and Celik, 2017). However, the unsupervised machine learning algorithms are not used when the train data is not classified.

The semi-supervised machine learning covers the supervised and unsupervised machine learning together. This is because the labeled and unlabeled data are used together in the train dataset for estimating the events in the future. The semi-supervised machine learning is the model constituted by the data providing a high level of unsupervised learning condition and low level of supervised learning condition.

The reinforced machine learning is the interpretation of data as good or bad based on the result of creation of model (Atalay and Celik, 2017). It is the selection of way of trial and finding error. In the deep machine learning, there are mathematically and statistically linear and nonlinear transformations. The data are analyzed with miscellaneous algorithms for formation of desired model.

However, the selection of right algorithm is a very challenging method. The selection of right algorithm can be ensured with predictor-corrector method. The selection of algorithm to be used vary by the type and volume of data to be analyzed, the results desired to be obtained from data and analysis techniques. If it is required to create a model for making estimation, it will be convenient to use the supervised learning algorithms.

3.2. Decision Trees in Machine Learning

The decision trees are a method used for both classification and estimation. In spite of many methods such as artificial neural networks, the decision trees are more preferred since they provide advantages for decision makers in terms of easy interpretation and understandability (Chien and Chen, 2008).

The structure of a decision tree consists of root, node, branch and leaf. They look like a reverse tree in respect of shape. Therefore, the bottom part is referred to as leaf while the upper part is referred to as root. The dataset qualities are referred to as node. The structure ensuring the internodal connection is referred to as branc (Gumuscu *et al.* 2016).

The decision trees play a role in the classification of both categorical and continuous data. In the decision tree techniques, the classification is made in two stages such as learning and classification. In the learning stage, the train data is used for creating the model. In the testing stage of model created, the test data is used. The classification rules and decision tree are constituted when the model is created with train data. In this stage, it is decided in which stages the tree will be divided. The classification rules and verification of decision tree are conducted thanks to test data.

Many decision tree algorithms were developed for creating decision tree from datasets easily. These algorithms were examined as follows: ID3, C4.5, C5.0, CART, CHAID, SLIQ, SPRINT, MARS and QUEST. The decision tree algorithms aim to create the optimum decision tree structure while generally eliminating the error. The big trees created with the decision tree algorithms do not carry the property of optimum tree and they have a low success in terms of generalization.

4. C5.0 Algorithm and Optimum Tree Structure

Following 1970s, Quinlan (1986) was trying to develop tree-based models. In 1980s, these methods were transformed into classification tree methods. C4.5 algorithm was developed by Quinlan (1993). C5.0 algorithm which was developed by Quinlan (1993) is based on C4.5 and ID3 (Iterative Dichotomiser 3) algorithm (Quinlan, 1986).

C4.5 and C5.0 are advanced versions of ID3 decision tree algorithms (Li *et al.* 2009). Therefore, C5.0 algorithm is more preferred. C5.0 is one of classification techniques which started to be used in various fields. In general, the decision trees provide convenience in terms of understanding and interpretation. In C5.0 algorithm, the binary trees constitute, and the optimum tree constitutes following the pruning.

In C4.5 and C5.0 algorithms, the dependent variable has categorical (nominal/ordinal) structure. The independent variables may be categorical or continuous. There is no limitation with respect to independent variables. C5.0 algorithm is a nonparametric method. There is no need for assumptions specific to parametric methods such as normality, linearity, etc.

In C5.0 algorithm, the multiple branches are constituted from each node. The number of branches varies depending on the category number of estimator. The information gain is used as a distinguishing criterion. The pruning procedure depends on the error ratio on each leaf (Bounsaythip *et al*. 2001).

C5.0 algorithm is generally used in the big datasets. C5.0 algorithm uses the boosting algorithm for increasing the accuracy. Therefore, it is also referred to as boosting tree. C5.0 algorithm enables us to obtain more smooth decision trees in terms of shape since they are more developed compared to C4.5 algorithm (Calis *et al.* 2014). C5.0 algorithm is a memory-based algorithm and better than C4.5 algorithm in terms of that the studies may be more sufficient. At the same time, they yield better results in terms of being understandable and production of decision rules by minimizing the decision tree (Shahnaz, 2006).

ID3 uses the entropy measurement and information criteria as a distinguishing criterion in the nodes of C4.5 and C5.0 algorithms. The distinguishing criterion allows for dividing the tree nodes with train data. It is expected that the results of entropy measurements will be lower. The fields of which entropy measurement is low, are the ones of which decision is the highest. The balance of classes' possibilities will allow for high entropy. The high information gain will occur in case the differences between class possibilities are significant compared to values before division.

The possibilities in the number of k for x variable is referred to as p_1, p_2, p_3, … , p_k. The entropy measurement is made as follows: (Quinlan, 1993).

$$Entropy= H(X) = - \sum_{j=1}^{k} p_j \log_2 (p_j) \quad (1)$$

T subclusters are classified as T_1, T_2, T_3, … , T_k depending on X variable in the train dataset. The class shall be determined for each T. The weighted mean of information required to determine the classes is calculated as the weighted sum of entropies. The weighted mean of necessary information is calculated as follows:

$$H_S (T) = \sum_{i=1}^{k} p_i H_S (T_i) \quad (2)$$

The information gain is calculated as follows:

$$Information\ Gain\ (S) = H\ (T) – H_S\ (T) \quad (3)$$

In short, the information gain also means the differences of entropy measurement before and after division. The pruning is very effective and efficient in C4.5 and C5.0 algorithms. They may yield good results although their bases are not theoretically solid in terms of pruning. The crossvalidation technique is sued in these algorithms. The crossvalidation is the examination of accuracy and reliability of model created with train data, with the test data.

C5.0 algorithm uses the binominal confidence limit method as pruning method. This is because the binominal confidence limit method allows for determining whether these values will be estimated or not in case the missing values are taken in hand.

In C5.0 algorithm, there are many poor classificators. The merge of these poor classificators in a strong classificatory is referred to as boosting (increasing). The increasing is used for decreasing the prejudice and variance. The boosting algorithm bears a resemble to AdaBoost algorithm developed at the beginning of 1990s (Kuhn and Johnson, 2013).

In the tree and rule-based models, there are two ways of processing the categorical variable data. These two options are also available in C5.0 algorithm. The categories are processed either in the form of grouped or independent. In the grouped categories, each categorical variable is entered individually and, in this way, how the model will be divided is decided. The independent categories resolve the variable into two dummy variables and each of them is accepted as independent. In general, the grouped category approach is used when only one part of categories carries the high estimation quality. These two approaches have advantages. Primarily, the model is created in both ways and then, the results are decided. However, the disadvantages approach may be evaluated in terms of conformance to data.

5. Implementation

In the research, the dataset of Turkish Statistical Institute Households Technology Use Survey 2016 was used (TUIK, 2016). C5.0 algorithm was used with the train dataset obtained with the help of 13510 observation taken from this dataset and it was tried to create a significant tree. In the research, ‘’R’’ programming language 3.4.4 version was used and the optimum trees belonging to C5.0 algorithm were created. The variables used in the research are shown in Table 1.

Table 1. Names of Variables Used in the Research

software_for_cookies	The use of software preventing the tracking of activities conducted over internet	1= Yes 2= No
information_about_cookies	Whether the individuals have information on cookies or not	1= Yes 2= No
cookies_disabled	Disabling the cookies for protecting the internet settings from cookies	1= Yes 2= No
cloud_usage	Storing the documents such as image, music, video or file on internet	1= Yes 2= No
age	The individuals in the range of 16-74 years	16-74 age range
gender	Gender of individual	1= Male 2= Female
working_status	Working status of individual	1= Yes 2= No
educational_status	Educational status of individual	1= Primary School 2= High School 3= University 4= Master 5= PhD 6= Uneducated

In the analysis to be made with C5.0 decision tree algorithm, the software_for_cookies was taken in hand as dependent variable. The variables of information_about_cookies, cookies_disabled, cloud_usage, age, gender, working_status ve educational_status were used as independent variables.

Table 2. Descriptive Statistics

Variable	n	%
software_for_cookies	Yes = 1384 No = 12126	Yes = 10% No = 90%
information_about_cookies	Yes = 3918 No = 9592	Yes = 29% No = 71%
cookies_disabled	Yes = 1963 No = 11547	Yes = 15% No = 85%
cloud_usage	Yes = 1842 No = 11668	Yes = 14% No = 86%
age	Min = 16 Max = 74	
gender	Female = 6158 Male = 7352	Female = 45% Male = 55%
working_status	Yes = 7258 No = 6252	Yes = 54% No = 46%
educational_status	Primary School = 6002 High School = 3823 University = 3122 Master = 294 PhD = 59 Uneducated = 210	Primary School = 44.4% High School = 28.2% University = 23% Master = 2% PhD = 0.4% Uneducated = 1%

In Table 2, 55% of 13510 persons who participated the research were male and 45% of them were female. 90% of individuals were not using a software preventing the tracking of activities conducted on internet. 71% of participants did not have information on cookies. 85% of individuals were not disabling the cookies for protecting their internet settings from cookies. 86% of individuals were storing the documents such as image, music, video or file on internet. The participants of study were in the age group of 16-74 years and the mean age was 35 years. 99.7% of research participants were literate. Moreover, 54% of individuals were working in any workplace.

In the machine learning algorithms, the entire dataset is not used. The aim of this is to create a model by using some part of dataset and to test the model created, with the remaining dataset. Therefore, the training data is used for creating the model while the test data is used for testing the model. However, it is important to use how much of the data in the training data. The use of less training data will give a rise to variance in the parameter estimations. A higher level of training data shall be used for obtaining high performance from analysis.

In the analysis, the training datasets were primarily created from the observations obtained from 13510 persons and the most convenient tree was decided. Therefore, 70% of dataset was considered as training data while 30% of dataset was used as test data for testing the model. Then, 80% of dataset was considered as training data and remaining 20% was used for testing the model. Finally, 90% of dataset was considered as training data and 10% was used for testing the model. In the final stage, the tree facilitating the examination of relationship between the variables constituting the dataset was examined as the optimum tree.

5.1. Findings Belonging to Classification Trees Created with C5.0 Application

Primarily, 70% of dataset was taken and the model was created with training data while the model created was tested with remaining 30%.

```
Class specified by attribute `outcome'

Read 9457 cases (9 attributes) from undefined.data

Decision tree:

cookies_disabled = No: No (8096/400)
cookies_disabled = Yes:
:...cloud_usage = No: No (940/363)
    cloud_usage = Yes:
    :...gender = Female:
        :...working_status = No: No (60/24)
        :   working_status = Yes:
        :   :...age <= 33: No (40/14)
        :       age > 33: Yes (24/8)
        gender = Male:
        :...educational_status in {Primary school,PhD}: No (42/17)
            educational_status in {University,Master,
            :                      uneducated}: Yes (169/72)
            educational_status = High School:
            :...working_status = No: Yes (33/12)
                working_status = Yes: No (53/24)
```

Figure 1. 1st Classification Tree Rule Sequence Created with C5.0 Algorithm

The nodes and branching points of tree created at the end of C5.0 algorithm by using R programming language are stated in Figure 1. An easier interpretation and understanding of tree are ensured with the rule sequence.

Table 3. Significant Variables in 1st Classification Tree Rule Sequence Created with C5.0 Algorithm

Variables	Degree of Effective Variable	Effect Percentages
cookies_disabled	1st Degree	100%
cloud_usage	2nd Degree	14.39%
gender	3rd Degree	4.45%
educational_status	4th Degree	3.14%
working_status	5th Degree	2.22%
age	6th Degree	0.68%

In Table 3, the degrees and percentages of variables effective in the creation of 1st classification tree are given. Accordingly, the independent variable being first degree effective on the status of using a software preventing the tracking of activities conducted on internet was analyzed as disabling the cookies in internet settings with 100% effect while the second degree independent variable was analyzed as storing the documents such as image, music, video or file on the internet with 14.39% effect, the third degree most effective independent variable was the gender of individuals with 4.45% effect, the fourth degree independent variable was educational background of individuals with 3.14% effect, the fifth degree independent variable was working status of individuals with 2.22% effect and the sixth degree independent variable was age of individuals with 0.68% effect.

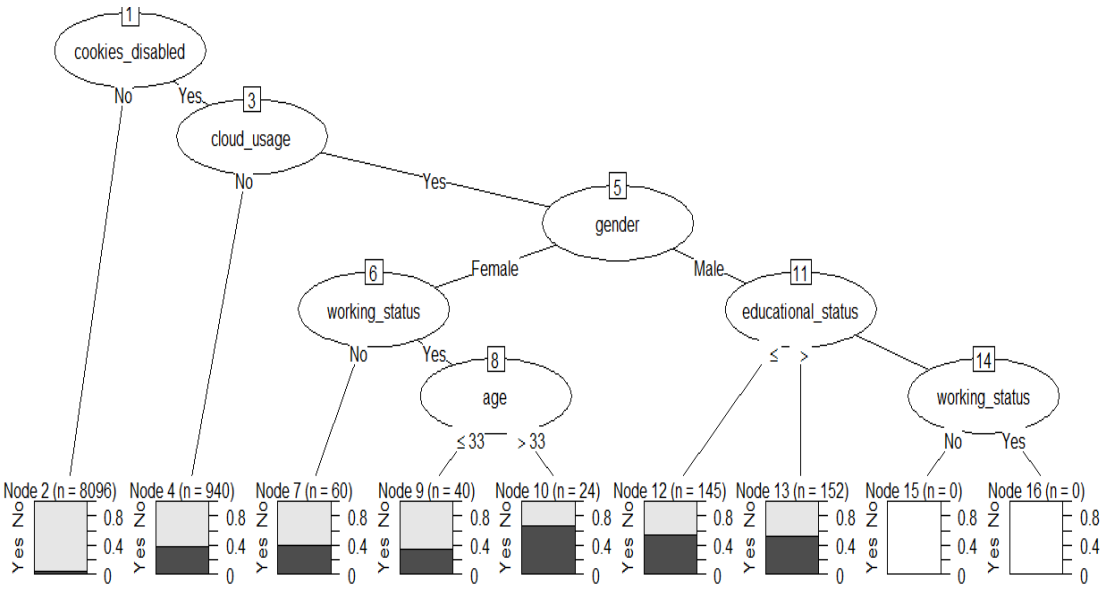

Figure 2. 1st Classification Tree Created with C5.0 Algorithm

The tree figure created by the classification tree given in Figure 1 and effective independent variables given in Table 3 are shown in Figure 2. 85% of individuals within 70% training data were not disabling the cookies in the internet settings. 0.05% of individuals who were not disabling the cookies in the internet settings, were not using software preventing the tracking of activities conducted on the internet. 15% of individuals within 70% training data were disabling the cookies in the internet settings and 70% of individuals were not storing the documents such as image, music, video or file on the internet and in other words, they were not using cloud storage. 61% of individuals who were not using cloud storage, were using a software preventing the tracking of activities conducted on the internet. The remaining 421 individuals were divided into branches in node 5 based on gender. 30% of 421 individuals were female while 70% were male. When the females were divided into branches in node 6 based on their working statuses, the female individuals were 124 persons and 60% of them, in other words, 48.3% were not employed in any work. 40% of non-employed women were not using a software preventing the tracking of activities conducted on the internet. The number of women employed in any work, was 64 and they were divided into new branches with the effect of age independent variable. The age of 40 (62.5%) of 64 working women was 33 years or below. 14 of women, in other words 35%, who were 33 years old or below and at the same time who were employed in any work, were not using a software preventing the tracking of activities conducted on the internet. 24 of 64 working women was above 33 years and 33.3% of these women were using a software preventing the tracking of activities conducted on the internet. The educational background of male individuals consisting of 297 persons were divided into branches based on node 11. The educational background of 145 of 297 male individuals were primary school and PhD. 40% of 145 male individuals were not using a software preventing the tracking of activities conducted on the internet. There were 152 males constituted by individuals who received education at the level of university and post graduate and who did not receive education. 42.6% of 152 males were using a software preventing the tracking of activities conducted on the internet. No comment was made at 3rd branch belonging to educational background in the tree. However, 36.3% of males whose educational backgrounds were high school and who were not employed were using a software preventing the tracking of activities conducted on the internet while 45.2% of males whose educational backgrounds were high school and who were employed were not using a software preventing the tracking of activities conducted on the internet based on the decision rules sequence table.

In 2nd classification tree, the model of which 80% of data set was taken as training data was created and the model was tested with remaining 20%, in other words test data.

```
Read 10808 cases (9 attributes) from undefined.data

Decision tree:

cookies_disabled = No: No (9220/461)
cookies_disabled = Yes:
:...educational_status in {Primary school,uneducated}: No (364/111)
    educational_status in {High School,Master,University,PhD}:
    :...gender = Female: No (436/166)
        gender = Male:
        :...working_status = Yes: No (611/266)
            working_status = No:
            :...information_about_cookies = No: No (21/6)
                information_about_cookies = Yes: Yes (156/53)
```

Figure 3. 2nd Classification Tree Rule Sequence Created with C5.0 Algorithm

The decision tree obtained by using C5.0 algorithm and 80% training data are given in Figure 3. The individuals' statuses of using a software preventing the tracking of activities conducted on the internet are classified by using this table.

Table 4. Significant Variables in 2nd Classification Tree Rule Sequence Created with C5.0 Algorithm

Variables	Degree of Effective Variable	Effect Percentages
cookies_disabled	1st Degree	100%
educational_status	2nd Degree	14.69%
gender	3rd Degree	11.32%
working_status	4th Degree	7.29%
information_about_cookies	5th Degree	1.64%

In Table 4, the degrees and percentages of variables effective in the creation of 2nd classification tree are given. Accordingly, the independent variable being first degree effective on the status of using a software preventing the tracking of activities conducted on internet was analyzed as disabling the cookies in internet settings with 100% effect while the second degree independent variable was educational background of individuals with 14.69% effect, the third degree most effective independent variable was the gender of individuals with 11.32% effect, the fourth degree independent variable was working status of individuals with 7.29% effect and the fifth degree independent variable was status of having information on cookies with 1.64% effect.

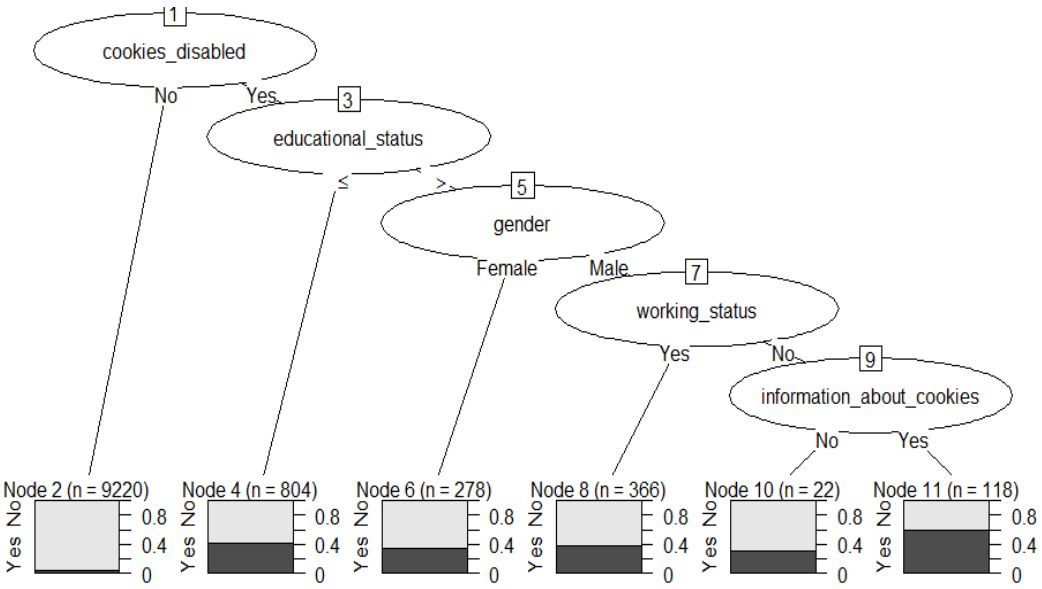

Figure 4. 2nd Classification Tree Created with C5.0 Algorithm

2nd classification tree was created with 80% training data. 85% of individuals were not disabling the cookies in the internet settings. 0.05% of individuals who were not disabling the cookies in the internet settings were not using a software preventing the tracking of activities conducted on the internet. 15% of individuals were disabling the cookies in the internet settings. The individuals who were disabling the cookies are divided into two branches in node 3.

The first branch of 3rd node is consisted of individuals who did not receive education and received education at the level of primary school. The individuals who did not receive education and received education at the level of primary school were 804 persons and 30.4% of 804 perso ns were not using a software preventing the tracking of activities conducted on the internet. The number of persons whose educational backgrounds were high school, undergraduate, postgrad uate and doctoral were 784 persons. When 784 persons were classified according to their ages, 278 were female and 38.1% of women stated that they were not using a software preventing the tracking of activities conducted on the internet. The remaining 506 individuals were male and wh en the males were classified based on their working conditions, 43.5% of 366 males who were e mployed were not using a software preventing the tracking of activities conducted on the interne t. The number of male individuals who were not employed was 140. 22 of 140 male individuals d id not have information on cookies and 28.5% of 22 male individuals were not using a software p reventing the tracking of activities conducted on the internet. 118 male individuals did not have i nformation on cookies and 34% of 118 male individuals were using a software preventing the tra cking of activities conducted on the internet.

In 3rd classification tree, 90% of dataset was taken as training data and the model was c reated and the model was tested with remaining 10%, which is test data.

```
Read 12159 cases (9 attributes) from undefined.data

Decision tree:

cookies_disabled = No: No (10375/505)
cookies_disabled = Yes:
:...cloud_usage = No: No (1225/459)
    cloud_usage = Yes:
    :...gender = Female: No (161/66)
        gender = Male:
        :...working_status = No: Yes (116/45)
            working_status = Yes:
            :...educational_status in {Primary school,High School,uneducated,
                :                      PhD}: No (104/46)
                educational_status in {Master,University}: Yes (178/84)
```

Figure 5. 3rd Classification Tree Rule Sequence Created with C5.0 Algorithm

The decision rules sequence obtained by using C5.0 algorithm and 90% training data are given in Figure 5.

Table 5. Significant Variables in 3rd Classification Tree Rule Sequence Created with C5.0 Algorithm

Variables	Degree of Effective Variable	Effect Percentages
cookies_disabled	1st Degree	100%
cloud_usage	2nd Degree	14.67%
gender	3rd Degree	4.60%
working_status	4th Degree	3.27%
educational_status	5th Degree	2.32%

In Table 5, the degrees and percentages of variables effective in the creation of 3rd classification tree are given. Accordingly, the independent variable being first degree effective on the status of using a software preventing the tracking of activities conducted on internet was analyzed as disabling the cookies in internet settings with 100% effect while the second degree independent variable was usage of cloud storage on the internet with 14.67% effect, the third degree most effective independent variable was the gender of individuals with 4.60% effect, the fourth degree independent variable was working status of individuals with 3.27% effect and the fifth degree independent variable was educational background of individuals with 2.32% effect.

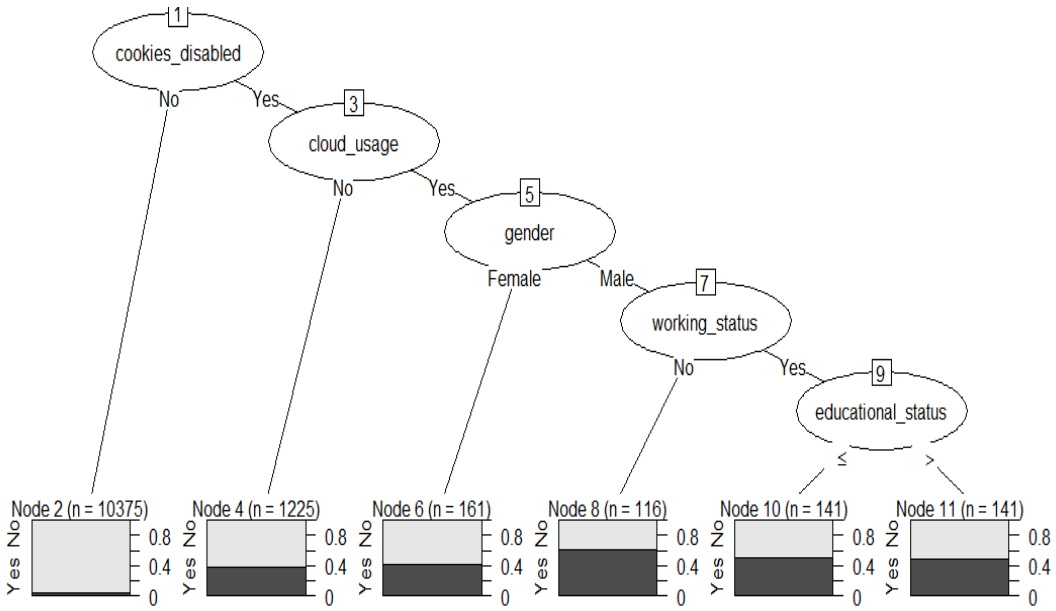

Figure 6. 3rd Classification Tree Created with C5.0 Algorithm

The tree was created with 90% training data in 3rd classification tree. 85% of individuals were not disabling the cookies in the internet settings. 0.05% of individuals who were not disabling the cookies in the internet settings were not using a software preventing the tracking of activities conducted on internet. 15% of individuals were disabling the cookies. Besides that the number of individuals disabling the cookies was 1784 persons, 1225 of them were not using cloud storage on the internet for personal purposes. 37.5% of individuals who were not using cloud storage were not using a software preventing the tracking of activities conducted on internet. 559 persons were using cloud storage and they were classified based on their genders in node 5. 161 of 559 persons were female and 41% of female were not using a software preventing the tracking of activities conducted on internet. The number of male individuals who were not employed was 116 and 38.75% of them were using a software preventing the tracking of activities conducted on internet. 282 male individuals who were employed were classified based on their educational backgrounds in node 9. Pursuant to node 10 consisting of individuals whose educational backgrounds were primary school, high school and doctoral, 44.2% of these individuals were not using a software preventing the tracking of activities conducted on internet. 47.2% of 141 individuals consisting of persons having university and master degree educational backgrounds were using a software preventing the tracking of activities conducted on internet.

In order to make a decision for optimum tree in classification trees consisted of 70%, 80% and 90% training data, the classification ratios shall be considered. In the stage of deciding the convenient tree, the tree of which faulty classification ratio was lowest was the tree of which right classification ratio was highest.

5.2. Classification Results of C5.0 Algorithm

Table 6. Classification Table of 1st Classification Tree Created with C5.0 Algorithm

Classification	Yes	No	Total
Yes	90	875	965
No	47	8445	8492
Total	137	9320	9457

When the classification tree of 1st classification tree created with 70% training data was examined, it was seen that 90 observations giving the response of yes was classified accurately.

There were 875 observations whose no responses were classified as yes. There were 47 observations who gave yes response but classified as no. 8445 observations who gave no response were classified accurately. Accordingly, 8535 observations in total were classified accurately. The error ratio in the classification of 1st classification tree was 9.7%.

Table 7. Classification Table of 2nd Classification Tree Created with C5.0 Algorithm

Classification	Yes	No	Total
Yes	103	1010	1113
No	53	9642	9695
Total	156	10652	10808

When the classification tree of 2nd classification tree created with 80% training data was examined, it was seen that 103 observations giving the response of yes was classified accurately. There were 1010 observations whose no responses were classified as yes. There were 53 observations who gave yes response but classified as no. 9642 observations who gave no response were classified accurately. Accordingly, 9745 observations in total were classified accurately. The error ratio in the classification of 2nd classification tree was 9.9%.

Table 8. Classification Table of 3rd Classification Tree Created with C5.0 Algorithm

Classification	Yes	No	Total
Yes	165	1076	1241
No	129	10789	10918
Total	294	11865	12159

When the classification tree of 3rd classification tree created with 90% training data was examined, it was seen that 165 observations giving the response of yes was classified accurately. There were 1076 observations whose no responses were classified as yes. There were 129 observations who gave yes response but classified as no. 10789 observations who gave no response were classified accurately. Accordingly, 10954 observations in total were classified accurately. The error ratio in the classification of 3rd classification tree was 10%.

Based on the classification results, the error ratio showed an increase as the training data increased. Therefore, the error ratio of 1st classification tree which was obtained with C5.0 algorithm and 70% training data was the lowest one. Thus, 1st classification tree was determined as optimum tree.

6. Conclusion

The use of data mining and machine learning algorithms in the modelling of big data has been increasing day by day. The artificial intelligence has become popular for processing the big data and it has been started to conduct many studies concerning to machine learning.

The machine learning allowed for development of computer algorithms and in the recent years, especially the data scientists have started to pay attention to use the machine learning algorithms. In the machine learning, the decision trees are used for both classification and estimation. It is seen that the decision trees are frequently used in the studies since both they depend on the assumptions and they are easy in terms of interpretation.

The reason of using C5.0 algorithm in this study was that it is applied to big datasets and it is easy in terms of interpretation. The binary trees are created in C5.0 algorithm. In C4.5 and C5.0 algorithms, the dependent variable has a categorical structure while the independent variable may consist of categorical or numerical data.

In the research, C5.0 decision tree algorithm codes were written in R programming language with the variables selected from the results of Turkish Statistical Institute Households Technology Use Survey 2016 and the trees were interpreted.

The model was created with 70%, 80% and 90% training data created from 13510 observation dataset and the model was analyzed with 30%, 20% and 10% test data. The decision

rules sequence created with C5.0 algorithm during analysis are primarily shown in Table 3, Table 5 and Table 7 and the convenience was provided for interpretation of tree. The significance levels and effect percentages of variables are shown in Table 4, Table 6 and Table 8. There were 6 effective variables in 1st classification tree created with C5.0 algorithm while there were 5 effective variables in 2nd classification trees and 5 effective variables in 3rd classification tree and the effect percentages of variables differ from each other.

1st classification tree created with 70% training data is shown in Figure 1 while 2nd classification tree created with 80% training data is shown in Figure 4 and 3rd classification tree created with 90% training data is shown in Figure 6. Accordingly, the thing which do not change in three trees is as follows: 85% of individuals a software preventing the tracking of activities conducted on the internet.

In general, 70% training data and 30% test data were used in the decision tree algorithms. In the research, the reason of using 80% training data and 90% training data was to investigate the change of faulty classification occurred on the trees as the training data increased on the trees. The classifications created by C5.0 algorithm are shown in Table 6, Table 7 and Table 8. Accordingly, 9.7% of 9457 observations in 1st classification tree were classified faulty while 9.95 of 10808 observations in 2nd classification tree was classified faulty and 10% of 12159 observations in 3rd classification tree was classified faulty. It was observed that the faulty classification ratio increased as the train data increased. Therefore, it was concluded that 1st classification tree was optimum tree.

References

Aksoy, B., Bayrakci, C., Bayrakci, E. & Uguz S. 2017. Usage of big data in institutions. *Suleyman Demirel University The Journal of Faculty of Economics and Administrative Sciences*, 22, pp. 1853-1878.

Atalay M., and Celik E., 2017. Buyuk veri analizinde yapay zeka ve makine ogrenmesi uygulamalari [Artificial Intelligence and Machine Learning Practices in Big Data Analysis] *Mehmet Akif Ersoy University Journal of Social Science Institute,* 9(22), pp. 161.

Bounsaythip, C. and Rinta-Runsala, E., 2001. Overview of data mining for customer behavior modeling. VTT Information Technology Research Report, version 1, p. 21.

Chien, C. F., and Chen. L. F., 2008. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, 34, pp. 280-290. https://doi.org/10.1016/j.eswa.2006.09.003

Calis, A., Kayapinar S., Cetinyokus T., 2014. An application on computer and internet security with decision tree algorithms in data mining. *Gazi University Journal of Industrial Engineering*, 25, pp. 2-19.

Gumuscu, A., Tasaltin R. and Aydilek B. I., 2016. C4.5 karar agaclarinda genetik algoritma ile budama [C4.5 Prunning in Decision Trees Based on Genetic Algorithm]. *Dicle University Institute of Science*, 5(2), pp. 77-80.

Kuhn, M. and Johnson, K., 2013. *Applied predictive modeling*. New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3

Li J., Fu A. W. and Fahey P., 2009. Efficient discovery of risk patterns in medical data. *Artificial Inteligence in Medicine*, 45(1), pp. 77-89. https://doi.org/10.1016/j.artmed.2008.07.008

Shahnaz, F., 2006. Decision tree based algorithms. In: M. W. Berry ed. 2006. *Lecture notes in data mining.* New Jersey: World Scientific Publisher. https://doi.org/10.1142/9789812773630_0007

TUIK., 2016. *Hanehalki bilisim teknolojileri kullanimi arastirmasi [Survey on the use of household information technologies] [online].* Available at: <http://www.tuik.gov.tr/PreTablo.do?alt_id=1028> [Accessed on 10 February 2018].

Quinlan, J. R., 1986. Induction of decision trees. *Machine Learning*, 1(1), pp. 81-106. https://doi.org/10.1007/BF00116251

Quinlan J. R., 1993. *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kauffman Publishers Inc.