Düzce University Journal of Science & Technology, 6 (2018) 455-468

Düzce University
Journal of Science & Technology

Open Source Data Mining Programs: A Case Study on R

Fatih KAYAALP [a*], Muhammet Sinan BAŞARSLAN [b]

[a] *Computer Engineering Department, Düzce University, Düzce, TÜRKİYE*
[a] *School of Advanced Vocational Studies, Doğuş University, İstanbul, TÜRKİYE*
** Corresponding Author: fatihkayaalp@duzce.edu.tr*

ABSTRACT

The processes on the way from raw data to meaningful information is called data mining. The data is processed by applying various methods of data mining in order to extract hidden information among raw data. The processed raw data becomes usable in the next steps of data mining. There are many open source and commercial applications to be used in data mining and data processing. In this study, information about data mining programs are given, and a case study on the R program. The R program has been chosen because it has a large preference rate among the users as shown by various graphs.

Keywords: *Data Mining, Open Source Programs, R, Churn Analysis*

Açık Kaynak Kodlu Veri Madenciliği Programları: R'da Örnek Uygulama

ÖZET

Ham verilerden anlamlı bilgilere geçiş sürecine veri madenciliği denir. Veri, ham veriler arasında gizli bilgileri çıkarmak için çeşitli veri madenciliği yöntemleri uygulanarak işlenir. İşlenmiş ham veriler, veri madenciliğinin bir sonraki aşamasında kullanılabilir hale gelir. Veri madenciliği ve veri işlemede kullanılmak üzere birçok açık kaynak ve ticari uygulama vardır. Bu çalışmada veri madenciliği programları hakkında bilgi verilmiş ve R programı üzerinde bir vaka çalışması sunulmuştur. R programı, çeşitli grafiklerle de gösterildiği üzere kullanıcılar arasında büyük bir tercih oranına sahip olması dolayısıyla seçilmiştir.

Anahtar Kelimeler: *Veri Madenciliği, Açık Kaynak Programlar, R, Kayıp Müşteri Analizi*

Received: 24/11/2017, Revised: 07/01/2018, Accepted: 05/02/2018

I. INTRODUCTION

Nowadays, access to knowledge is not the result of long struggles as in the old times. Instant access and sharing of information have become possible. The important thing here is whether this information is useful to be used. At this point, government and commercial institutions' use of the acquired information for the good of humanity is of great importance. For example, it is now much easier for commercial enterprises to offer their all range of products to the people. Through e-commerce, customers now do shopping from groceries to clothing stores and to large or small firms without leaving their homes. A customer-focused business entity, such as a bank or a telecommunications company, wants to maintain the continuity of business with its customers and does not want to lose customers. For this reason, they keep customer records and information. However, this information must be stable, in other words it needs processing for customer acquisition and preventing customer churn. These operations are generally called data mining. Nowadays, besides the information gathered with interaction during trading as in the example given above, data are obtained by means of various data collection methods.

Advances in data collection tools and database technologies require vast amounts of information to be stored and analyzed in information repositories. In line with the advances in computer technology, data mining methods and programs aim to put vast amounts of data into effective and efficient use. To combine knowledge and experience, it is necessary to use software developed for Data Mining. Data mining became compulsory due to ever growing data records (GB/hour), satellite and remote sensing systems, space observations through telescopes, developments in gene technology, scientific computations, simulations, models, and data mining [1].

Programs are needed to implement data mining applications. Access to applications used for data mining and acquiring information about these applications can be found on the Internet, in various books or articles. There is a variety of applications in this regard, such as IBM SPSS Modeler 15, Excel, SAS, Angoss, KXEN, SQL Server, MATLAB, RapidMiner (YALE), WEKA, R, Orange and KNIME. Table 1 shows these commercial and open source applications.

Table 1. *Commercial and Open Source Data Mining Programs[2]*

Commercial (Closed Source Code)	**Open Source Coded**
SPSS Clementine	Orange
Excel	RapidMiner
SPSS	WEKA
SAS	R
Angoss	Keel
KXEN	Knime
MS SQL Server	Tanagra
MATLAB	Scriptella ETL
Oracle Modules	jHepWork
…	Elki…

In this study, open source data mining programs are compared and a sample application is developed in R. A brief information on data mining is given in the second section. A comparison about open source data mining programs is presented and the R language is explained briefly in the third section. In the fourth section, a case study with R is presented. The fifth and final section presents the results.

II. Data Mining

There are various definitions of data mining in literature. It can be called as the task of obtaining a "valuable" information among vast amounts of data [3]. Data mining, in other words knowledge discovery in databases (KDD), is the process of extracting information that is potentially useful and understandable that has never been discovered before in vast amounts of data. Data analysis techniques such as backend database management systems, statistics, artificial intelligence, machine learning, parallel and distributed processes are also called data mining [4][5].

Data mining is the search for relationships and rules in vast amounts of data using computer programs which allow us to make predictions about the future. If we assume that the future, at least the near future, will not be too different from the past, the rules drawn from the past will be valid in the future and will allow us to make the right predictions for the future [6].

The first steps of information discovery in databases were taken by the 1990 and data mining has become a new standard widely used in line with the new technologies [7]. Processing raw data to make it usable, that is discovery of knowledge requires a certain process. The process involves the steps described in Figure 1.

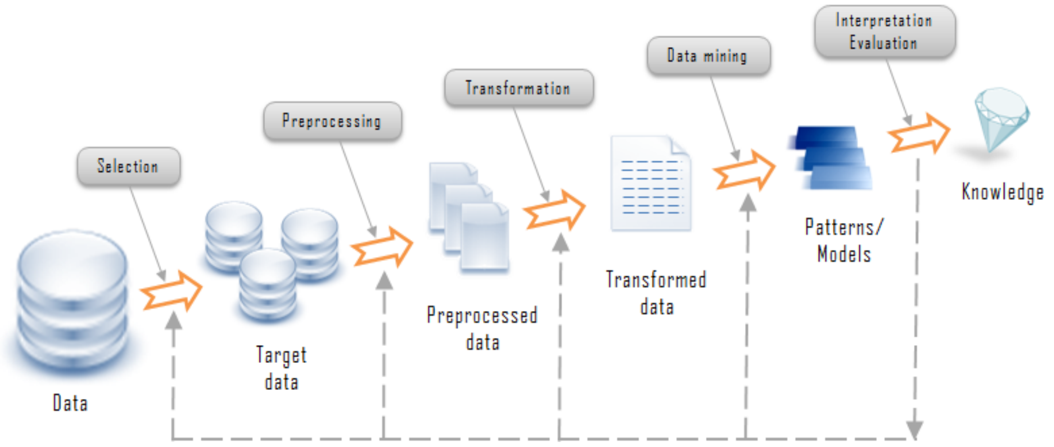

Figure 1. *Data Mining Process[4]*

III. Comparison of R and Other Open Source Data Mining Programs

Today it is necessary to use computer programs to make data mining applications due to the huge size of data. Computer programs are offered as commercial and open source (free) programs. This section provides a comparison of open source and freely available programs and an overview of the R program to be used in the study.

It is a program developed for graphics, statistical calculations and data analysis. It's a GNU project similar to the S language. It was developed in the Department of Statistics at Auckland University in New Zealand by Robert Gentleman and Ross Ihaka. It is also known as R & R. R is superior to the S language due to its different applications. It features linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering, etc. R can run on Windows, MacOS X and Linux systems [3][8].

As shown in Table 2, the R language does not lag behind other selected programs. In addition, thanks to its packages, R can be expanded according to the desired operations and R has advantages with its statistics efficiency as well as package diversity compared to others.

Table 2. *Comparison of Open Source Data Mining Softwares [2]*

Parameter	Keel	Knime	Orange	R	RapidMiner (YALE)
Data Mining Algorithms	Yes	Yes	Yes	Yes	Yes
Machine Learning Packages	Yes	Yes	Yes	Yes	Yes
Statistical Calculation	Yes	Yes	Yes	Yes	Yes
Data Analysis	Yes	Yes	Yes	Yes	Yes
Preprocessing	Yes	Yes	Yes	Yes	Yes
Attribute Selection	Yes	Yes	Yes	Yes	Yes
Visualization	Yes	Yes	Yes	Yes	Yes
GUI	Yes	Yes	Yes	Yes	Yes
Expandability	Yes	Yes	Yes	Yes	Yes
Flexibility	Yes	Yes	Yes	Yes	Yes
Ease of Use	Yes	Yes	Yes	Yes	Yes
Error-free Operation	Yes	Yes	Yes	Yes	Yes
Documentation	Yes	Yes	Yes	Yes	Yes
Scripting	Yes	Yes	Yes	Yes	Yes
Additional Packages	Yes	Yes	Yes	Yes	Yes
Data Import/Export	Yes	Yes	Yes	Yes	Yes
Supported File Formats	.dat, .arff, .csv, .xml, .txt, .prn, .xls, .dif, .html	.arff, .csv	.tab, .basket, .names, .data, .txt, .xls (.arff and .csv only reads)	.r, .txt, .ods, .csv, .xml	.sml, .srff, .stt, .bib, .clm, .cms, .cri, .csv, .dat, .ioc, .log, .matte, .mode, .obf, a bar, one pair, .res, .sim, .thr, .wgt, .wls, .xrff, .arff
Working with Databases	Yes (SQL Data-bases)	Yes (Oracle, MS SQL Server, PostgreSQL, MySQL, Access, ODBC, JDBC)	Yes (MySQL)	Yes (Informix, Oracle, Sybase, DB2, MS SQL Server, MySQL, PostgreSQL, MS Access, ODBC)	Yes (Oracle, MS SQL Server, PostgreSQL, MySQL, JDBC, Sybase, Access, IBM DB2, Ingres, Text Files)
Working with Excel	Yes (with import)	No	No	Yes	Yes

Table 3 shows the top 10 programs and usage preferences rates of data mining programs in 2016 [9].

Table 3. Usage Preference Rates of Data Mining Program [9].

Program Name	Rate of Preferences
R	49%
Python	45.8%
SQL	35.5%
Excel	33.6%
RapidMiner	32.6%
Hadoop	22.1%
Spark	21.6%
Tableau	18.5%
KNIME	18.0%
Scikit-learn	17.2%

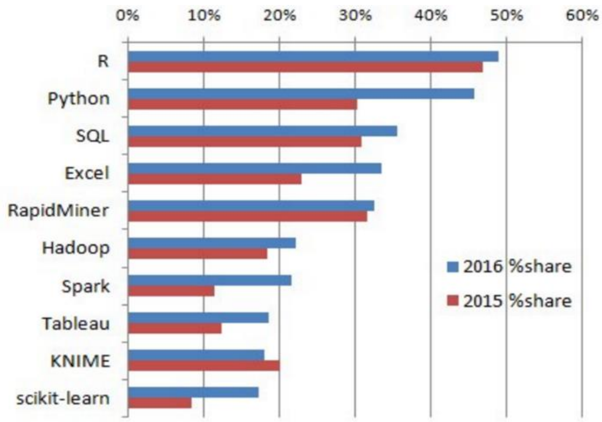

Figure 2. *Top 10 Data Mining Programs in 2016 compared to 2015 [10]*

As it is seen in Table 3 and Figure 2, the rate of preferences of the R program is increasing day by day since it is platform independent, open source, have numerous packages available and supported by a large user group on Internet. Because of this, an R-based application has been developed in this study.

IV. A CASE STUDY ON R

This study aims to make an application in customer conversion analysis by using a telecommunications data set which foresees to be separated from customer service subscription by a telecommunications company in Turkey. Among the open source programs, the R program is used. In this section, all the operations performed during the application are presented.

A. PROBLEM DEFINITION

A customer churn analysis study was conducted to estimate the churn rate by using customer data obtained from a telecommunications company. The data set used in the study includes information about 8000 customer records for a period of 6 months of call related data. In this study, a classification was performed with the decision trees for customer churn rates at the 6th month using C4.5 decision tree algorithm.

B.DATA PREPARATION AND DEFINITION

The telecommunications data set contains 8000 customer records and 22 attributes. Table 4 shows all the variables, formats and types related to telecommunications data set. When Table 4 is examined, it is seen that telecommunications data set has numerical and categorical attributes. A view of sample data is also given in Figure 3.

Table 4. *All variables, formats and types related to telecommunications data set*

Attribute	Description	Data type
gender_flag	Gender	NOMINAL
Age	Age	NUMERICAL
age_of_line	Customer lifetime duration	NUMERICAL
tariff_type	Tariff type (Postpaid, Prepaid)	NOMINAL
device_type	Device type, Smartphone, Laptop etc.	NOMINAL
last_reload_year	Last reload date (for Prepaid subscribers)	NOMINAL
last_reload_amount_07	Last reloaded amount (for Prepaid subscribers)	NUMERICAL
mmo_count_07	Monthly number of conversations with subscribers (call)	NUMERICAL
mmo_duration_07	Monthly talk time with subscribers (call)	NUMERICAL
mmo_non_count_07	Number of conversations per month with other operator subscribers (call)	NUMERICAL
mmo_non_duration_07	Monthly talk time with other operator subscribers (call)	NUMERICAL
mmt_count_07	Monthly number of conversations with subscribers (in call)	NUMERICAL
mmt_duration _07	Monthly talk time with subscribers (in call)	NUMERICAL
mmt_non_count _07	Number of conversations per month with other operator subscribers (in call)	NUMERICAL
mmt_non_duration_07	Monthly talk time with other operator subscribers (in call)	NUMERICAL
call_distinct_07	Number of different people talked monthly	NUMERICAL
msmo_count_07	Number of SMSs per month	NUMERICAL
callcenter_count_07	Monthly CC Complaints call count	NUMERICAL
payment_07	Monthly payout amount, invoice for postpaid, total reload for prepaid	NUMERICAL
Unpaid_07	Number of unpaid invoices on time	NOMINAL
payment_type_07	Invoice payment method, automatic?	NOMINAL
Churn_2013_07	Has the subscriber churned?	BINARY

```
'data.frame':   8000 obs. of  22 variables:
 $ gender_flag        : int  2 2 2 3 3 2 3 2 2 2 ...
 $ age                : int  50 37 29 47 22 55 31 37 53 49 ...
 $ age_of_line        : num  2935 49 2935 2935 49 ...
 $ tariff_type        : int  1 2 1 1 1 1 2 1 1 1 ...
 $ device_type        : int  10 10 0 10 10 3 10 3 3 3 ...
 $ last_reload_year   : int  2014 1900 1900 2014 2014 2014 2014 2014 2013 1900 ...
 $ last_reload_amount : Factor w/ 31 levels "0","10","10,75",..: 23 1 1 14 14 14 17 14 2 1 ...
 $ mmo_count_07       : num  7 32 1 1 8 0 148 0 0 19 ...
 $ mmo_duration_07    : num  7 32 1 1 8 0 148 0 0 19 ...
 $ mmo_non_count_07   : num  3 15 29 98 61 117 136 23 0 1 ...
 $ mmo_non_duration_07: num  10 214 0 9168 1514 ...
 $ mmt_count_07       : num  54 27 9 93 139 12 29 19 0 48 ...
 $ mmt_duration_07    : num  7185 20040 2569 1993 0 ...
 $ mmt_non_count_07   : num  3 15 29 98 61 117 136 23 0 1 ...
 $ mmt_non_duration_07: num  3611 15330 8081 15238 6483 ...
 $ call_distinct_07   : int  39 2 33 0 32 0 46 17 0 0 ...
 $ msmo_count_07      : int  1644 35 29 2 1644 17 1644 5 0 16 ...
 $ callcenter_count_07: int  0 0 1 0 1 0 2 1 0 0 ...
 $ payment_07         : Factor w/ 507 levels "0","1","1,00E-06",..: 175 1 1 1 1 1 260 1 1 1 ..
 $ unpaid_07          : int  1 0 0 0 0 1 0 0 1 1 ...
 $ payment_type_07    : int  2 27 2 20 10 1 19 9 8 19 ...
 $ churn_2013_07      : Factor w/ 2 levels "H","E": 1 1 1 1 1 1 1 1 1 1 ...
```

Figure 3. *Types and general distributions of data types in R Studio*

Figure 4 shows a summary of the telecommunications data set.

```
> summary(aveamice7)
 gender_flag      age           age_of_line      tariff_type           device_type      last_reload_year mmo_avea_count_07 mmo_avea_duration_07 mmo_nonavea_count_07
 U:  13      Min.   :16.00   Min.   :  49   Kontörlü:5461   Mobil Tel      :3701   2014      :3980   Min.   :   0.00   Min.   :   0.00      Min.   :   0.00
 K:5758      1st Qu.:28.00   1st Qu.: 266   Faturalı:2539   Akıllı Telefon:3254   Bilinmeyen:2326   1st Qu.:   0.00   1st Qu.:   0.00      1st Qu.:   2.00
 E:2229      Median :36.00   Median : 705                   Bilinmeyen    : 786   2013      :1189   Median :  15.00   Median :  15.00      Median :  24.00
             Mean   :38.13   Mean   :1117                   Usb Modem     : 229   2012      : 275   Mean   :  40.63   Mean   :  40.63      Mean   :  63.87
             3rd Qu.:46.00   3rd Qu.:1476                   Tablet PC     :  17   2011      : 113   3rd Qu.:  54.00   3rd Qu.:  54.00      3rd Qu.:  82.00
             Max.   :73.00   Max.   :2935                   Modül         :  11   2010      :  77   Max.   :1049.00   Max.   :1049.00      Max.   :1155.00
                                                            (Other)       :   2   (Other)   :  40
 mmo_nonavea_duration_07 mmt_avea_count_07 mmt_avea_duration_07 mmt_nonavea_count_07 mmt_nonavea_duration_07 call_distinct_07 msmo_count_07  callcenter_count_07
 Min.   :     0          Min.   :  0.00    Min.   :     0       Min.   :   0.00      Min.   :     0          Min.   : 0.00    Min.   :   0   Min.   : 0.0000
 1st Qu.:   133          1st Qu.:  4.00    1st Qu.:   650       1st Qu.:   2.00      1st Qu.:   815          1st Qu.: 6.00    1st Qu.:   1   1st Qu.: 0.0000
 Median :  2782          Median : 23.00    Median :  3310       Median :  24.00      Median :  4564          Median :23.00    Median :  15   Median : 0.0000
 Mean   : 10564          Mean   : 42.62    Mean   : 10541       Mean   :  63.87      Mean   :  8989          Mean   :30.23    Mean   : 403   Mean   : 0.9441
 3rd Qu.: 13610          3rd Qu.: 59.00    3rd Qu.: 10469       3rd Qu.:  82.00      3rd Qu.: 11735          3rd Qu.:44.00    3rd Qu.: 172   3rd Qu.: 1.0000
 Max.   :322551          Max.   :629.00    Max.   :781720       Max.   :1155.00      Max.   :199968          Max.   :99.00    Max.   :1644   Max.   :26.0000

   payment_07     unpaid_07             payment_type_07 churn_2013_07
 0      :4156   Min.   :0.0000   Mobil Tel      :3701   H:7857
 20     : 708   1st Qu.:0.0000   Akıllı Telefon:3254   E: 143
 10     : 434   Median :0.0000   Bilinmeyen    : 786
 30     : 302   Mean   :0.2021   Usb Modem     : 229
 25     : 154   3rd Qu.:0.0000   Tablet PC     :  17
 40     : 120   Max.   :2.0000   Modül         :  11
 (Other):2126                    (Other)       :   2
```

Figure 4. *Overview of the data set*

C.DATA CLEANING

Data cleaning is the most critical and time-consuming step in the data mining process. There are missing values and outliers in the data set. Various existing packages can be used in R program to determine and analyze missing values. In this study, the missing data are obtained by using mice package, which has various predefined expectation algorithms [11]. Outlier values were solved by the boxing method.

D.MODELLING AND ASSESSMENT

The next step after data cleaning and transformation is the modeling step. Different models are tested on the dataset and the model with the highest accuracy is selected. In this study, a model was created using the C4.5 decision tree classification algorithm. Rpart [11], Rweka [12], Partykit [13], and Caret [14] packages were used to construct the model. To use these packages, they have to be called from the relevant library. After this, the data set is divided into two parts as training and test data.

Dividing the data set into training and test data sets is important in terms of getting the best validation and learning outcomes. Methods such as k-fold cross validation, bootstrap and hold out are used to achieve this separation [15].

In this study, accuracy and error rate, diagnostic superiority ratio and f-measure were used as performance measures in addition to dividing training and test groups into different percentages for performance evaluation of decision tree and classification model. The measurement performance metrics used for performance evaluation will be briefly described in the subheading.

1. Performance Metrics

Evaluation of the model created by classification algorithms is carried out by various methods. One of these methods is the confusion matrix [16]. The actual values and the values predicted by the classification algorithm are shown in Table 3. Performance evaluation criteria of classification algorithms are shown in Table 5 below [16-18].

Table 5. *The Confusion Matrix*

		Actual Result		
		Yes	No	Toplam
Predicted Result	Yes	True Positive (tp)	False Positive (fp)	tPoz
	No	False Negative (fn)	True Negative (tn)	tNeg
	Total	poz	neg	m

Accuracy of the model generated by the classification algorithms according to Table 5 is given by Eq. (1) [17].

$$\boldsymbol{Accuracy} = \frac{d_p + d_n}{m} \tag{1}$$

The error rate equation is shown in Eq. (2) [17].

$$\boldsymbol{Error\ Rate = 1 - Accuracy} \tag{2}$$

The supremacy of the predicted positive class is called the ratio of the negative class supremacy [94]. Equation (3) shows the diagnostic superiority ratio [17].

$$\boldsymbol{Diagnostic\ Superiority\ Ratio} = \frac{LR+}{LR-} \tag{3}$$

The F-measure equation is shown in Eq. (4) [17].

$$\boldsymbol{F - measure} = \frac{2 * PPV * TPR}{PPV + TPR} \tag{4}$$

The dataset is divided into two parts for training and testing of the proposed model. There are several methods to be used for this partitioning. In this study, holdout methods were used. In hold out method, the test and training data sets are separated with a specific ratio. Hold out separations used in the study are shown in Figure 5.

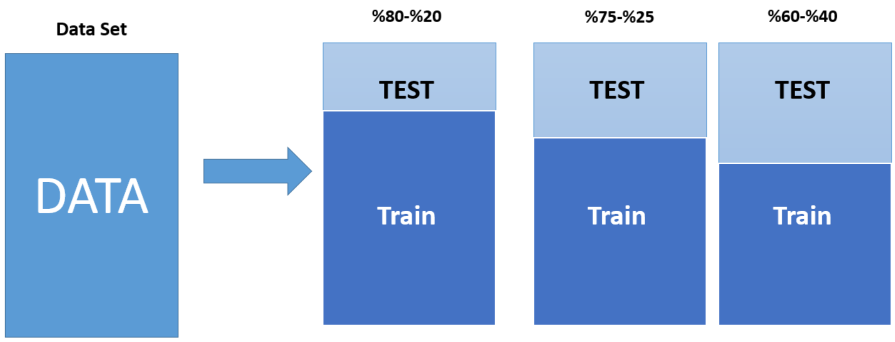

Figure 5. *Test and training set separation with the hold out method.*

2. C4.5 Decision Tree Algorithm Application

In this section, applying C4.5 decision tree method which is one of the classification methods is given. The libraries of Rpart, Rweka, Partykit and Caret packages mentioned in the previous section must be installed and called. Figure 6 shows the installation and call of the corresponding libraries.

```
#Upload libraries
install.packages("rpart")
install.packages("RWeka")
install.packages("partykit")
install.packages("caret")
install.packages("ROCR")

#Invoking libraries
library(rpart)
library(RWeka)
library(partykit)
library(caret)
library(ROCR)
```

Figure 6. *Library installation and call*

Figure 7 shows the 60% training and 40% test set separations. The operation codes shown in Figure 7 are also used for separations at different percentages. The performance measurement by changing the percentages is called as the hold out method.

```
#Random Separation of Training and Test Set
set.seed(2016)
randomayir60 <- createDataPartition(y = aveamice7c45$churn_2013_07, p = .60, list = FALSE)
trainSET60 <-mice7c45[randomayir60,]
testSET60 <-mice7c45[-randomayir60,]
```

Figure 7. *Separation of data set as training and test set through the hold out method*

The codes for decision tree model, the overview of the model and the graphical visualization of the tree are shown in Figure 8 respectively.

```
#Creation of model with training set
DT_model60 <-J48(churn_2013_07~., data=trainSET60)
print(DT_model60)
summary(DT_model60)
plot(DT_model60)
```

Figure 8. *Setting up the model with the training set*

Figure 9 shows a screenshot of the tree generated by the C4.5 decision tree algorithm seen in Figure 7. It is unlikely to print the model by looking at the tree pattern. Figure 10 shows a screenshot of the separation obtained by printing the tree.

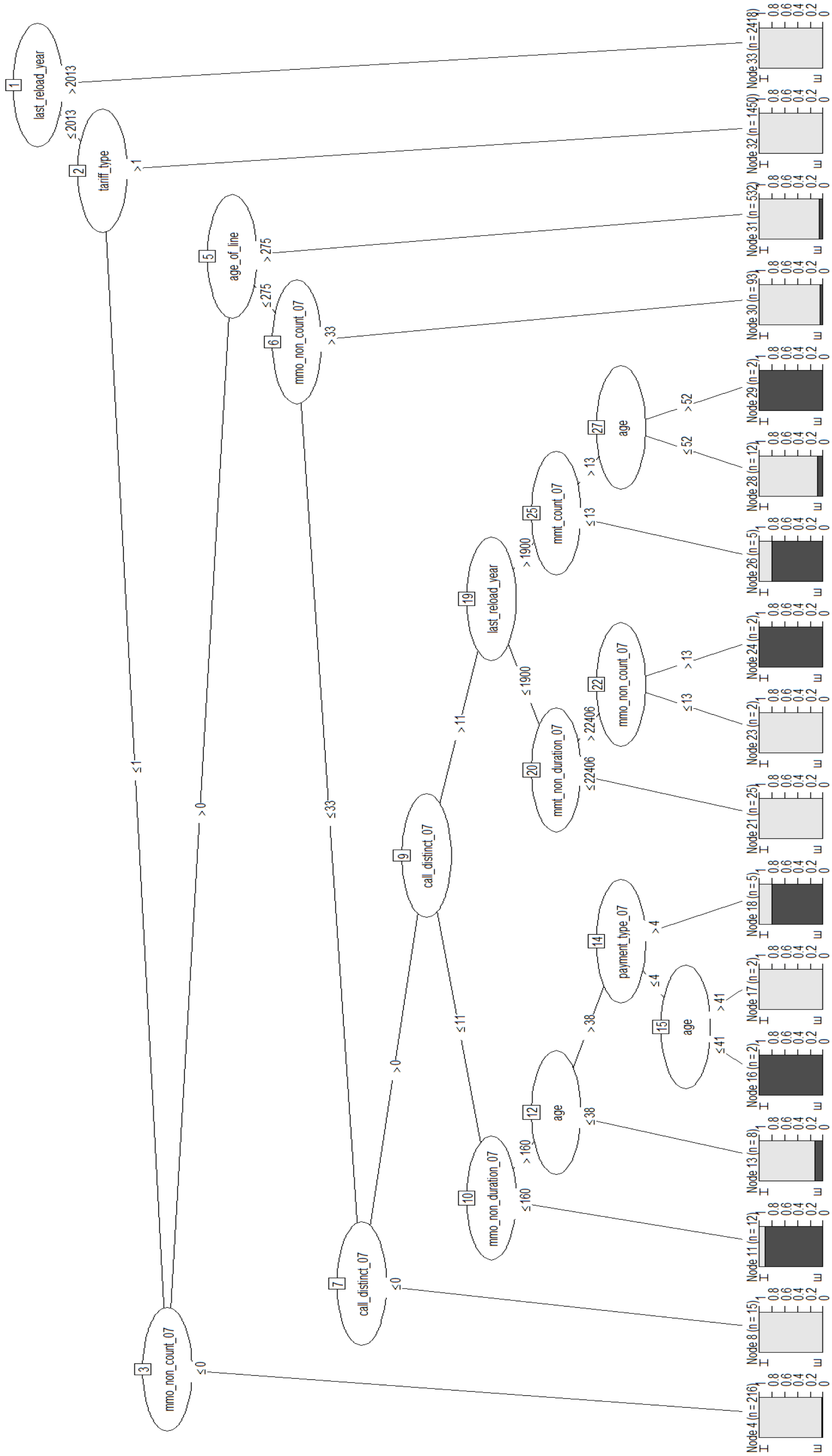

***Figure 9**. Image of the C4.5 tree*

```
> print(DT_model60)
J48 pruned tree
------------------

last_reload_year <= 2013
|   tariff_type <= 1
|   |   mmo_non_count_07 <= 0: H (216.0/3.0)
|   |   mmo_non_count_07 > 0
|   |   |   age_of_line <= 275
|   |   |   |   mmo_non_count_07 <= 33
|   |   |   |   |   call_distinct_07 <= 0: H (15.0)
|   |   |   |   |   call_distinct_07 > 0
|   |   |   |   |   |   call_distinct_07 <= 11
|   |   |   |   |   |   |   mmo_non_duration_07 <= 160: E (12.0/1.0)
|   |   |   |   |   |   |   mmo_non_duration_07 > 160
|   |   |   |   |   |   |   |   age <= 38: H (8.0/1.0)
|   |   |   |   |   |   |   |   age > 38
|   |   |   |   |   |   |   |   |   payment_type_07 <= 4
|   |   |   |   |   |   |   |   |   |   age <= 41: E (2.0)
|   |   |   |   |   |   |   |   |   |   age > 41: H (2.0)
|   |   |   |   |   |   |   |   |   payment_type_07 > 4: E (5.0/1.0)
|   |   |   |   |   |   call_distinct_07 > 11
|   |   |   |   |   |   |   last_reload_year <= 1900
|   |   |   |   |   |   |   |   mmt_non_duration_07 <= 22406: H (25.0)
|   |   |   |   |   |   |   |   mmt_non_duration_07 > 22406
|   |   |   |   |   |   |   |   |   mmo_non_count_07 <= 13: H (2.0)
|   |   |   |   |   |   |   |   |   mmo_non_count_07 > 13: E (2.0)
|   |   |   |   |   |   |   last_reload_year > 1900
|   |   |   |   |   |   |   |   mmt_count_07 <= 13: E (5.0/1.0)
|   |   |   |   |   |   |   |   mmt_count_07 > 13
|   |   |   |   |   |   |   |   |   age <= 52: H (12.0/1.0)
|   |   |   |   |   |   |   |   |   age > 52: E (2.0)
|   |   |   |   mmo_non_count_07 > 33: H (93.0/5.0)
|   |   |   age_of_line > 275: H (532.0/33.0)
|   tariff_type > 1: H (1450.0/17.0)
last_reload_year > 2013: H (2418.0/1.0)

Number of Leaves  :     17

Size of the tree :      33
```

Figure 10. *Print of the model and tree separation*

A sample rule obtained from Figure 10: if age_of_line c <= 274 and mmo_non_count_07 <= 24 then E (customer churn).

V. CONCLUSION

As a result of the study, the training and test sets of the model are compared with 60%-40%, 75%-25% and 80%-20% divisions, respectively. Table 3 shows the criteria of accuracy, error, superiority and F-measure, which are performance evaluation criteria for these distinctions. In some studies, only the accuracy criterion is sufficient to compare the models generated by the classification algorithms. In addition to the accuracy; error, diagnostic odds ratio and F-measure are also included.

Table 6. *Accuracy, Error, Diagnostic odds ratio and F-measure values*

	Accuracy			**Error**			**DOR (Diagnostic Odds Ratio)**			**F-measure**		
Hold Out Percentages	60	75	80	60	75	80	60	75	80	60	75	80
C4.5	0.978	0.977	0.981	0.021	0.022	0.018	40.59	38.88	77.54	0.989	0.988	0.990

In the study conducted on the customer dataset, C4.5 Decision of the customers with the decision tree was estimated. This work has been implemented on R program from open source data mining programs.

The performance criteria has achieved high performances in all holdout distinctions of the classification model established by C4.5 decision tree algorithm. Due to the fact that the decision tree is performing well with the nominal data types seen in previous studies. Since the purpose of this study is to carry out a sample work on R program, a more comprehensive study will be carried out by diversifying the classification algorithms in later studies.

This study was carried out to show the capabilities of R program needed for data mining processes. As seen in the tables in Section 3, the use of R has increased significantly in the last two years and does not lag behind the others.

The aim of the churn analysis which was the case study of the paper is to help the Customer Care Management stuff of the company to detect the possible churners. The results of the the classification model established by C4.5 decision tree algorithm shows that the churners can be detected with 0.977 to 0.981 accuracies. The next stage after detecting the possible churners, the company would offer attractive proposals to persuade these customers not to churn.

Telecommunications data set is used within the scope of the study. The missing data and outliers in the data set are resolved and prepared for processing with the help of the continuously evolving packages provided with the R program. A model, based on prediction with C4.5 decision tree algorithm was established as a result.

VI. REFERENCES

[1] M. Dener, "Açık Kaynak Kodlu Veri Madenciliği Programları:WEKA'da Örnek Uygulama" presented at 11th Acad. Inform.Conf., 2009, Şanlıurfa, Turkey, 2009.

[2] M. Kaya, S. A. Özel, "Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması", presented 14th Acad. Inform. Conf., 2014, Mersin, Turkey, 2014.

[3] Ö. Yalçın, *Veri Madenciliği Yöntemleri*, 2nd Ed., Papatya Yayıncılık, 2013.

[4] J. Han, M. Kanber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[5] M. J. Berry, G.S. Linoff, *Mastering Data Mining*, John Wiley&Sons, New York, 2004.

[6] L. B. Ayre, *Data Mining For Information Professionals*, June, 2006.

[7] E. Alpaydın, Zeki Veri Madenciliği: Ham Veridan Altın Bilgiye Ulaşma Yöntemleri, 2000.

[8] Anonymous, *R* (12, December 2016) [Online]. Access: https://en.0wikipedia.org/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvUl8ocHJvZ3JhbW1pbmdfbGFuZ3VhZ2Up

[9] P. Gregory, Top Analytics, Data Science Software – Kdnuggets Software Poll Results, [Online]. Erişim: https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

[10] Anonymous, (05, January 2018) [Online]. Access: https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html, 2018.

[11] T. Therneau, B. Atkinson, B. Ripley, (20, February 2017) *Rpart Packages*. [Online]. Access: https://cran.r-project.org/web/packages/rpart/rpart.pdf. 2017.

[12] T. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, A.Zeleis, (11, January 2016). *RWeka Packages*, [Online]. Access: https://cran.r-project.org/web/packages/RWeka/RWeka.pdf, 2016.

[13] T. Horthon, A. Zeileis, (10, January 2016). *Partykit Packages*, [Online]. Access: https://cran.r-project.org/web/packages/partykit/ partykit.pdf, 2016.

[14] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, C. Candan, (2016, 12 November). *Caret: Classification and Regression Training*. [Online]. Access: https://cran.r-project.org/pub/R/web/packages/caret/caret.pdf, 2016.

[15] Ş. Özdemir, *Eğitsel Veri Madenciliği Çalışması: Lise Öğrencilerinin Okula Devamlılık Durumlarının Öngörülmesi*, R ile Veri Madenciliği (Balaban-Kartal), Çağlayan Kitabevi, 2016.

[16] N. Japkowicz, "Performance evaluation for learning algorithms," International Conference on Machine Learning, Edinburg, Scotland, 2012.

[17] M. Clark, "An Introduction to machine learning with Applications in R," Lecture Notes, University of Notre Dame, 2015.

[18] P. Flach, "The many faces of ROC analysis in machine learning," *ICML Tutorial*, 2004.