International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 3, No. 3, pp. 201-216
doi: 10.25147/ijcsr.2017.001.1.34
https://stepacademic.net

Position Paper

Data Mining with R: An Applied Study

Burcu Durmuş
Rectorate Unit, Mugla Sitki Kocman University
(Department of Statistics, Mugla Sitki Kocman University)
burcudurmus@mu.edu.tr

Öznur İşçi Güneri
Department of Statistics, Mugla Sitki Kocman University
oznur.isci@mu.edu.tr
(corresponding author)

Date received: July 1, 2019
Date received in revised form: October 18, 2019
Date accepted: November 22, 2019

Recommended citation:

> Durmuş, B., & Güneri, Ö. İ. (2019). Data mining with R: An applied study. *International Journal of Computing Sciences Research*, *3*(3), 201-216. doi: 10.25147/ijcsr.2017.001.1.34

Abstract

Purpose – The aim of this study is to analyze different classification algorithms with R programming and to determine the accuracy rates. It also encourages the use of the R program by giving readers the opportunity to experiment.

Method – For the purposes mentioned above, different data sets were obtained from the UC Irvine Machine Learning Repository (2019), which was suitable for classification. After preparing data set and R program for data mining, performance evaluation was made with classification algorithms (J48, Random Forest, Naïve Bayes). The 'accuracy' criterion was taken into consideration when interpreting the results.

Results – At the end of the study, the accuracy rates were determined for three data sets. Looking at the "wine" data, the performance of all three algorithms is quite successful. The results of the other two data sets (lenses and liver) are parallel. Only the 'liver' dataset gave a slightly lower accuracy than expected with the Naïve Bayes algorithm (0.55).

Conclusion – In this study, performance comparison of algorithms has been made within the scope of data mining with R program. The accuracy rate was taken as a criterion. All codes are given with their outputs in order to be an example especially for young researchers or students. It is thought that this study can be a source for other researchers, will encourage the use of R and the researchers or students will try new papers by trying the codes.

Recommendations – In subsequent studies, a similar study can be done by developing the given codes. Or how to make classification analysis in R with different algorithms can be examined.

Keywords – data mining, R program, J48, random forest, Naïve Bayes

INTRODUCTION

One of the biggest problems that arise today with developing technology is the information stored. Technological advances in data storage methods cause the stored information to grow exponentially. The fact that the Internet is an integral part of daily life affects this growth. E-mails, web page records, training notes, market sales data, hospital records and results, bank transactions, social media accounts, sports competitions are produced in every second Almost all of this data is recorded electronically. The answer to the question of how, when and for which stored information will be used is within the scope of data mining. Data mining is the process of conducting analyses by obtaining useful information from a large stack of data. With this analysis, it is aimed to reveal meaningful information and relations and to reveal data patterns. Data mining also allows for new estimates based on historical data. Data mining methods used for these purposes are classified under three main headings as classification, clustering and association analysis.

In this study, classification method which is frequently encountered in literature is discussed. The classification method is one of the most up-to-date alternative methods that offer more practical and faster solutions than many other algorithms. In the classification method, a model is established with the help of various algorithms based on common features, differences, ratings or groupings within the data. In order to construct models within the scope of classification, algorithms based on many different theories have been developed. The theoretical structure of each of these algorithms is mathematically different. If you want to look roughly, the statistical basis of these algorithms is; decision trees, regression analysis, logistic functions and extensions, bayes theory, neural networks.

Algorithms used for classification purposes allow significant improvements in many areas such as early diagnosis of disease in medical science (Rokny et al., 2017; Kumar &

Sahoo, 2011; Kurt & Ensari, 2017), analysis of student achievement in education (Alom & Courtney, 2018; Fernandes et al., 2019), classification of plants or animals in biology (Salvador et al., 2018; Celik et al., 2017; Koc, Eyduran, & Omer, 2017), detection of spam in information (Abdulhamid et al., 2018), document classification (Rajvanshi & Chowdhary, 2017).

There are several methods for measuring the validity of the classification model. Among these methods, accuracy, sensitivity, accuracy and error rate are the most popular. These criteria are calculated using the equations explained in the second section. These calculations are based on statistical calculations such as correct estimation rate, correct classification rate, wrong classification rate. Therefore, these criteria are known as the most commonly used criteria in the literature. In addition, the success of the model is explained by correctly classified observations and incorrectly classified observations. For this purpose, the information obtained from the test is indicated by the confusion matrix. Table 1 shows an example confusion matrix. In this table, 16 observations of class "a" were correctly estimated and 3 of those appearing in class "b" were incorrectly estimated. Information criteria can be easily calculated via the confusion matrix. For example, the accuracy rate can be easily calculated on the Table as the ratio of correctly classified observations to all observations.

Table 1. Confusion Matrix

		Predict		
		A	**b**	**c**
Real	**A**	16	3	0
	B	2	35	0
	C	1	0	57

Data mining analyses are generally performed with the help of programs on computer. There are many programs developed in the literature for classification analysis. The most preferred are open source programs such as Weka, SPSS, Knime, R, Oracle. In recent years, the widespread use of software in academic circles has increased the use of programs. The R program, which is frequently preferred in the field of statistics, is also affected by this increase. The R-programming language is user-friendly, providing advantages in many areas. R programming language, which has an important place in data mining, is used in the analysis of classification algorithms.

As mentioned, there are many algorithms and programs developed for classification purposes. Simultaneous analysis of all algorithms is not practical. It is also known that some algorithms are developed and prepare the ground for other algorithms. Therefore, it is more meaningful to consider algorithms that are based on newly developed and more robust mathematical foundations. Nevertheless, it is clear that there are many

algorithms to be examined. In this study, J48, Random Forest algorithms and Naïve Bayes algorithm based on probability tree structures are discussed. The study is based on three basic steps in order to investigate how the R program can be used to classify the data:

1. examining the theoretical information about the algorithms to be discussed in the study,
2. conducting classification analysis through R program, and
3. interpreting the results and evaluating the contribution of the study.

The first step is the material and method stage. At this stage, the structure of the algorithms and the model performance criteria are examined. The data sets used are also introduced at this stage. The second step involves the analysis of 3 different data sets (liver, lenses, wine) in the R program with the aforementioned 3 different classification algorithms. The procedures are explained step by step. The program outputs are given as they are in order to clearly see the results of the analysis. Also, in this step, all R codes are presented to the reader and they are given the opportunity to experiment. In the third and last step of the study, the algorithm results were compiled collectively with the help of tables. The compiled results were interpreted to explain the contribution of the literature and the study was completed by presenting suggestions.

LITERATURE REVIEW

When the literature is reviewed, it is seen that data mining emerged conceptually in the 1960s when computers were used to solve data analysis problems (Han, Kamber, & Pei, 2012). Data mining, which was called as data scanning in the first days, has reached the present term with the consideration of computer engineers. In the 1990s, traditional statistical methods were abandoned, and data analyzes were evaluated with the help of computer modules. However, these modules were very difficult to use and required significant data preparation (He, 2009). This has led researchers and especially computer engineers to develop new modules. Looking at the interfaces that can be used for today's data mining analysis, it is seen that some of them are developed commercially and some of them are offered as open source. SPSS, MATLAB, Oracle and Weka, R, Knime, RapidMiner are examples (Kaya & Özel, 2014). Many studies conducted under the name of data mining are available in the literature. The rapid development of computer technologies and the ease of data acquisition and storage increase the importance of data mining and thus push researchers to work on this issue.

Alfaro et al. (2013) conducted a study with adabag which is a classification package in R program. They showed applications for the three data sets in the literature and as a result discussed the similarities and differences of the three different algorithms.

Zhang (2016) conducted a classification study with Naïve Bayes. In this study, it has clearly explained how and with which packages the classification is used in R. Kızılkaya and Oğuzlar (2018) compared the performance of logistic regression and decision tree

controlled learning algorithms with R language. They stated that logistic regression yielded the most successful result according to sensitivity criterion.

Goswami et al. (2018), in their compilation studies on the application of data mining techniques, found that there are not enough resources for natural disaster detection especially in the Indian region. This study reveals the necessity of data mining in combating natural disasters. Çınar (2019) determined the performance of C5.0 and Gini classification algorithms in determining students' learning levels by using R language. C5.0 algorithm showed better results.

This study promotes data mining using R and aims to analyze different classification algorithms with R programming and determine accuracy rates. In the current literature, there are many studies or applications about data mining and its applications. However, to the best of our knowledge, there are few studies that clearly show how the classification is made with the R program, which allows many analyzes especially in recent years. In this study, analysis steps are given in addition to the existing studies. Thus, it is thought that especially young researchers will gain habits such as experimenting, self-learning and reading the results.

METHODOLOGY

J48 Algorithm

One of the most well-known decision tree algorithms, J48 is the Weka equivalent of the C4.5 algorithms. The J48 algorithm is also known as ID3. In this algorithm, the entropy and information gain values for the target class are calculated using equations 1 to 3. The expected information needed to classify a tuple in D is given by (Han et al., 2012):

$$Entropy = -\sum_{i=1}^{m} p_i log_2(pi) \qquad \textit{Equation 1}$$

A kind of normalization is applied to the gain of knowledge by using the "split information" value defined similarly to entropy.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x log_2\left(\frac{|D_j|}{|D|}\right) \qquad \textit{Equation 2}$$

This value represents the potential information generated by dividing D into sections v corresponding to the v results of an A test. For each result, the number of tubes that achieve this result is taken into account. Gain ratio in this case:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \qquad \textit{Equation 3}$$

Here $p_1, p_2, \cdots p_m$ are probabilities with totals of 1 (Özdemir, 2018).

Entropy indicates the likelihood of an unexpected situation (Bhargava et al., 2013). Information gain values show information values calculated for each attribute from the entropy value. In J48 algorithm, decision tree is created starting from the variable that gives the highest information. Processing is complete when all variables are included in the tree (Patil & Sherekar, 2013).

Algorithm Steps:

- Entropy and related information gain values are calculated.
- Features that give the best information gain are added to the decision tree. The best feature creates the base node.
- After the calculation of all features and branching of the decision tree, the model installation is completed (Kaur & Chhabra, 2014).

Random Forest Algorithms

It is one of the most preferred decision tree algorithms for classification problems (Eraldemir, Arslan, & Esen, 2017). Multiple decision trees are generated for the classification process and then random forests are generated. Because of the high number of decision trees created, the classification success is high.

Algorithm Steps:

- The feature that provides the best classification is selected and the starting node is created.
- A training set is formed with a part of the data set. The remaining data is the test set.
- Trees are created with the number of variables to be used in each node and the numbers of trees in N. Variables are selected randomly at each node.
- When N trees are produced, the model is completed and the class of the new member is estimated (Akar & Güngör, 2012).

Naïve Bayes Algorithms

It is a probabilistic method based on Bayes' Theorem. It is named after the famous mathematician Thomas Bayes. In this method, probability values are calculated from the

observed properties and classification is made. It equalizes the probability value to "0" if there is an incalculable or unobservable value. Bayes' Theorem is expressed by Equation 4 (Odabaş, 2017).

$$P(X|C_i) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad \textit{Equation 4}$$

Model Performance Evaluation

There are many criteria used to evaluate model performance. Accuracy, error rate, precision, and sensitivity are the most important ones. In this study, accuracy values were taken into consideration. The accuracy value is calculated by the ratio of the number of correctly classified observations to the total number of observations. This criterion refers to the capability of the classifier. In other words, the fact that this criterion value is high and acceptable shows the applicability of the model in the classification of new observation values. Therefore, accuracy values were taken into consideration in the study. Thus, the predictor will be a good predictor for new observations (Han et al., 2012).

Data Sets

In this study, different data sets were obtained from UC Irvine Machine Learning Repository (2019) suitable for classification procedures. When selecting data sets, it was taken into consideration that they have different attribute characters and they do not contain missing data. Thus, it was predicted that the unpredictable values would not remain and that the model would be established healthier. Information on these data sets is given in Table 2.

Table 2. Information of Data Sets

	Sample Data Sets		
	Win	Lense	Liver
Number of instances	178	24	345
Number of attributes	13	4	7
Attribute	I, R	C	C, I, R
Missing value	No	No	No
Area	Phy	N/A	Life

APPLICATION WITH R

Before proceeding to the data mining stage with R, the packets given in Figure 1 must be available in R. Other required packages are installed automatically in these packages.

```
install.packages("plyr")
install.packages("caret")
install.packages("RWeka")
install.packages("partykit")
install.packages("randomForest")
install.packages("e1071")
```

Figure 1. R Packages Required for Classification Analysis

Data set and R program were prepared for the study. The data sets were then evaluated with the help of classification algorithms. The application consists of three steps. The first step is to prepare the data and transfer it to the program. The second step is to analyze with classification algorithms and the third step is to evaluate and compare the results. The classification analysis steps for the "wine" data set are described in detail in this section. Other data sets were analyzed with the same codes. The results are presented for discussion.

'Wine' data prepared for analysis were transferred to R program. Information on the data was examined and the number of classes suitable for classification was determined (Torgo, 2011; Zhao, 2015. Figure 2 shows these operations.

```
> data <- read.csv(file.choose(), header = F)
> rownames(data) <- paste0("variable", 1:dim(data)[1])
> colnames(data) <- c("wine", "alcohol", "malic", "ash",
"alcali","magne","totalphe","flava","nonflava","proant","color","hue","od","proline")
> final_data <- as.factor(x = data[[1]])
> library(plyr)
> data$wine <- revalue(final_data,c("1"="Type1", "2"="Type2", "3"="Type3"))
 wine
 Type1:59
 Type2:71
 Type3:48
```

Figure 2. Transferring the Data Set to the Program and Categorizing the Data

The data set was divided into training set and test set. The training set is used to train the model according to the algorithm to be selected. In other words, the model is created with the help of the data in the training set. The test set measures the performance of the algorithm on the model. That is, these data are given to the established model as a new observation. Thus, it is checked whether the model performs successful classification. In the research, 70% of the data were used as training data and the rest as test data (Figure 3).

```
> library(caret)
Installing mandatory package: lattice
Installing mandatory package: ggplot2
Registered S3 methods overwritten by 'ggplot2':
  method        from
  [.quosures    rlang
  c.quosures    rlang
  print.quosures rlang
> set.seed(123)
> education_index <- createDataPartition( y = data$wine, p = 0.7, list = FALSE)
> education_data <- data[education_index,]
> test_data <- data[-education_index,]
```

Figure 3. Obtaining Training and Test Data Sets

Classification with J48 Algorithm

In this section, analysis is made with J48, which is one of the decision tree algorithms. The analysis steps for the training set calculated with Figure 3 are given in Figure 4. According to the flow chart of the algorithm, flava feature gives the highest information gain in tree formation. Therefore, it is determined as the first property. According to the results, the number of leaves is 5, the size of the tree is 9. When the Confusion Matrix is examined, there are two misclassified observations. Correct classification rate of the algorithm is 98.41%, Kappa statistical value is 0.97 and mean square root error is 0.10. Figure 5 shows the decision tree structure for the J48 algorithm.

The results for the test data set are in Figure 6. When the Confusion Matrix of the test data set is examined, it is seen that there is no misclassified observation.

Classification with Random Forest Algorithm

The classification results for this multi-tree algorithm are given in Figure 7. In the model created with the training set, four observations were misclassified. When the test set results are examined, it is seen that the algorithm has 1 accuracy rate for this data set.

```
> library(RWeka)
> classification1 <-J48(wine ~ . , data = education_data)
> show(classification1)

J48 pruned tree
------------------
flava <= 1.57
|   color <= 3.8: Type2 (8.0)
|   color > 3.8: Type3 (35.0/1.0)
flava > 1.57
|   proline <= 720: Type2 (39.0/1.0)
|   proline > 720
|   |   color <= 3.4: Type2 (3.0)
|   |   color > 3.4: Type1 (41.0)

Number of Leaves: 5
Size of the tree     :9

> summary(classification1)

=== Summary ===
Correctly Classified Instances         124         98.4127 %
Kappa statistic                        0.9759
Mean absolute error                    0.0206
Root mean squared error                0.1015
Relative absolute error                4.6882 %
Root relative squared error            21.6553 %
Total Number of Instances              126

=== Confusion Matrix ===
  a  b  c   <-- classified as
 41  1   0 |  a = Type1
  0 49   1 |  b = Type2
  0   0 34 |  c = Type3

> library(partykit)
Installing mandatory package: grid
Installing mandatory package: libcoin
Installing mandatory package: mvtnorm
> plot(classification1)
```

Figure 4. J48 Algorithm for Classification Steps and Decision Tree

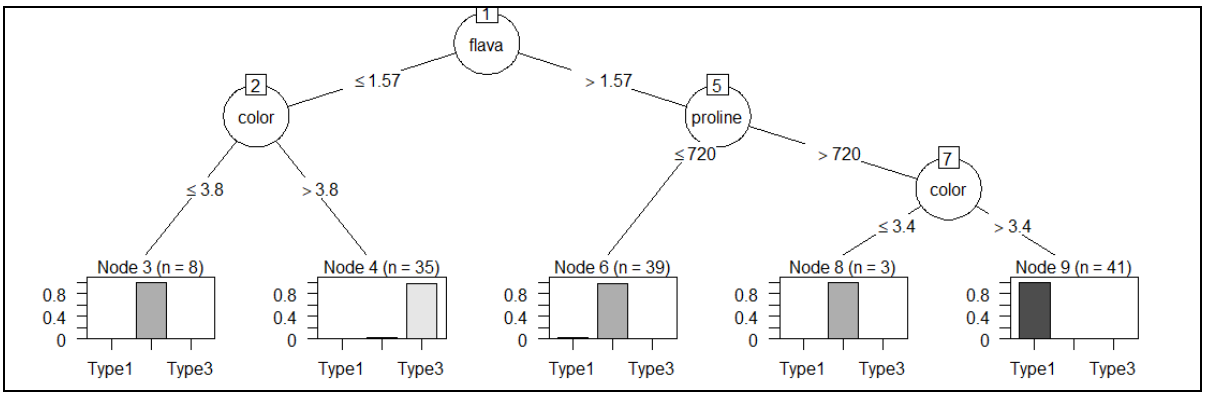

Figure 5. Decision Tree for J48 Algorithm

```
> p <- predict(classification1, test_data, type = "class")
> confusion.matrix <- table(test_data$wine, p, dnn = c("REAL", "PREDICT"))
> show(confusion.matrix)
          PREDICT
REAL    Type1 Type2 Type3
  Type1    17     0     0
  Type2     0    21     0
  Type3     0     0    14
> r <- nrow(confusion.matrix)
> c <- ncol(confusion.matrix)
> diagonal <- (function (x) x + (x-1)*c) (1:r)
> accuracy <- sum(confusion.matrix[diagonal]) / sum(confusion.matrix)
> show(paste("Accuracy = ", accuracy))
[1] "Accuracy =  1"
```

Figure 6. Accuracy Rate for Test Data

```
> library(randomForest)
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.
> forest <- randomForest(wine ~., data = education_data)
> show(forest)

Call:
 randomForest(formula = wine ~ ., data = education_data)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 3.17%
Confusion matrix:
      Type1 Type2 Type3 class.error
Type1    42     0     0        0.00
Type2     1    46     3        0.08
Type3     0     0    34        0.00

> predict_forest<- predict(forest, test_data, type = "class")
> confusion.matrix_forest <- table(test_data$wine, predict_forest, dnn = c("REAL",
"PREDICT"))
> show(confusion.matrix_forest)
             PREDICT
REAL    Type1 Type2 Type3
  Type1    17     0     0
  Type2     0    21     0
  Type3     0     0    14
> accuracy_forest <- sum(confusion.matrix_forest[diagonal]) /
sum(confusion.matrix_forest)
> show(paste("Accuracy = ", accuracy_forest))
[1] "Accuracy =  1"
```

Figure 7. Accuracy Detection for Random Forest Algorithm

Classification with Naïve Bayes Algorithm

This algorithm based on probability also performed very well. The required steps and results are shown in Figure 8. According to the model results, only 2 observations are not correct. The accuracy rate was calculated as 0.98%. In Figure 9, the accuracy values of the wine data set are collectively observed. J48, Random Forest, Naïve Bayes algorithms have accuracy rates of 1, 1, 0.98 respectively. It can be said that successful results were obtained in all three methods.

```
> library(e1071)
> Naive_Bayes_Model=naiveBayes(wine ~., data=data)
> NB_Predictions=predict(Naive_Bayes_Model,data)
> confusion.matrix_naive <- table(NB_Predictions,data$wine)
> confusion.matrix_naive

NB_Predictions Type1 Type2 Type3
      Type1        58     0       0
      Type2         1    70       0
      Type3         0     1      48

> accuracy_naive <- sum(confusion.matrix_naive[diagonal]) / sum(confusion.matrix_naive)
> show(paste("Accuracy = ", accuracy_naive))
[1] "Accuracy =  0.98876404494382"

> comparison <- data.frame(c(accuracy, accuracy_forest,accuracy_naive))
> colnames(comparison) <- "Accury"
> rownames(comparison) <- c("J48", "RandomForest", "NaiveBayes")
> show(comparison)
           Accury
J48        1.000000
RandomForest 1.000000
NaiveBayes   0.988764
```

Figure 8. Accuracy Detection for Naïve Bayes Algorithm

```
> comparison <- data.frame(c(accuracy, accuracy_forest,accuracy_naive))
> colnames(comparison) <- "Accury"
> rownames(comparison) <- c("J48", "RandomForest", "NaiveBayes")
> show(comparison)
           Accury
J48        1.000000
RandomForest 1.000000
NaiveBayes   0.988764
```

Figure 9. Accuracy Rate Results for All Three Algorithms

RESULTS

In this study, the stages of data mining classification algorithms are shown by using R over the "wine" data set frequently used in the literature. The main purpose of the study is to encourage readers to analyze with R and to present the application of basic classification algorithms. For this purpose, three commonly used algorithms in the literature have been selected: J48, Random Forest, Naïve Bayes. Two points are emphasized when selecting algorithms. First, the algorithm structures are based on different mathematical foundations. The second is the frequency of researchers choosing this algorithm. Throughout the study, the analysis of "wine" data was explained to the reader. However, in the background of the study, "liver" and "lenses" data sets were also analyzed. When selecting these datasets, it was considered that they do not contain missing data and that they have different attribute characters.

After the data sets were ready for analysis, the application phase of the study was started. First, 70% of the data set was identified as training set and 30% as test data (Figure 3). Then the actual analysis was done. For this, j48, Random Forest and Naïve Bayes algorithms were selected. Data sets were classified with the help of these algorithms. The 'accuracy' criterion was chosen to interpret the results. This criterion is chosen because it gives accuracy as mentioned before. Obviously, it shows the ratio of how accurately a new observation is predicted.

In the last step of the study, the results were examined and comments were made. The accuracy rates for the three data sets are given in Table 3. Looking at the "wine" data, the performance of all three algorithms is quite successful. The results for the other two data sets (Lenses and Liver) are in the parallel. Only the "liver" dataset gave a slightly lower accuracy than expected with the Naïve Bayes algorithm (0.55).

Table 3. Comparison of Algorithm Results

	Data Sets		
Algorithms	Wine	Lenses	Liver
J48	1	1	0.67
Random Forest	1	1	0.74
Naïve Bayes	0.98	1	0.55

DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

Classification analysis is of great importance in both statistical and interdisciplinary analysis for reasons such as discovery of connections in the data set, identification of relationships, patterns between features, and recognition of the data set for meaningful analysis. For this reason, many researchers have made studies in this field. When the

studies were examined, it was thought that the number of data mining analyzes with R program was low. For this reason, in this study, performance comparison of algorithms has been made within the scope of data mining with R program. The accuracy rate was taken as a criterion. All codes are given with their outputs in order to be an example especially for young researchers or students.

It is thought that this study can be a source for other researchers, will encourage the use of R and the researchers or students will try new papers by trying the codes. In subsequent studies, a similar study can be done by developing the given codes. Or how to make classification analysis in R with different algorithms can be examined.

REFERENCES

Abdulhamid, S. M., Shuaib, M. Osho, O., Ismaila, I., & Alhassan, J.K. (2018). Comparative analysis of classification algorithms for email spam detection. *International Journal of Computer Network and Information Security*, *10*(1), 60-67.

Akar, Ö., & Güngör, O. (2012). Rastgele Orman Algoritması Kullanılarak Çok Bantlı Görüntülerin Sınıflandırılması. *Journal of Geodesy and Geoinformation, 1*(2), 139-146.

Alfaro, E., Gamez, M., & Garcia, N. (2013). Adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*, *54*(2), 1-35.

Alom, B. M. M., & Courtney, M. (2018). Educational data mining: A case study perspectives from primary to university education in Australia. *International Journal of Information Technology and Computer Science*, *2*, 1-9.

Bhargava, N., Sharma, G., Bgargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *International Journal of Advanced Research inComputer Science and Software Engineering, 3*(6), 1111-1119.

Celik, S., Eyduran, E., Karadas, K., & Tariq, M. M. (2017). Comparison of predictive performance of data mining algorithms in predicting body weight in Mengali Rams of Pakistan. *Revista Brasileira de Zootecnia ,46*(11), 863-872.

Çınar, A. (2019). Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi ve R Dili ile Bir Uygulama. *Marmara Üniversitesi Öneri Dergisi*, *14*(51), 90-111.

Eraldemir, S. G., Arslan, M. T., & Esen, Y. (2017). Comparison of random forest and J48 decision tree classifiers using HHT based features in EEG. In *International Advanced Researches & Engineering Congress*, Osmaniye, Turkey.

Fernandes, E., Holanda, M., Victorino, M., Vinicius, B., Carvalho, R., & Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*, 335-343.

Goswami, S., Chakraborty, S., Ghost, S., Chakrabarti, A., & Chakraborty, B. (2018). A review on application of data mining techniquesto combat natural disasters. *Ain Shams Engineering Journal, 9*, 365-378.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques.* USA: Morgan Kaufmann Publishers.

He, J. (2009). Advances in data mining: History and future. In *2009 Third International Symposium on Intelligent Information Technology Application* (Vol. 1, pp. 634-636). IEEE.

Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications*, *9*(22), 13-17.

Kaya, M., & Özel, S. A. (2014). Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması. In *Akademik Bilişim'14,* Mersin, Turkey.

Kızılkaya, Y.M., & Oğuzlar, A. (2018). Bazı Denetimli Öğrenme Algoritmalarının R Programlama Dili ile Kıyaslanması. *Karadeniz*, *37*, 90-98.

Koc, Y., Eyduran, E., & Omer, A., (2017). Application of regression tree method for different data from animal science. *Pakistan Journal of Zoology*, *49*(2), 599-607.

Kumar, Y., & Sahoo, G. (2011). Predication of parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers. *Indian Journal of Medical Sciences*, *65*(6), 231-242.

Kurt, M. S., & Ensari, T. (2017). Destek Vektör Makineleri ve Çok Katmanlı Algılayıcılar ile Diyabet Teşhisi. In *2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE. doi: 10.1109/EBBT.2017.7956757

Odabaş, Ö. (2017). *Veri Madenciliği Teknikleri ile Telekom Sektöründe Ayrılan Müşteri Analizi* (unpublished manuscript). İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

Özdemir, S. (2018). *Veri Madenciliği* (unpublished manuscript). Gazi Üniversitesi, Ankara, Turkey. Retrieved from http://w3.gazi.edu.tr/~suatozdemir/teaching/dm/

Patil, T.R., & Sherekar, S. S. (2013). Performance analysis of Naïve Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications,* *6*(2), 256-261.

Rajvanshi, N., & Chowdhary, K. R. (2017). Comparison of SVM and Naïve Bayes text classification algorithms using Weka. *International Journal of Enginerring Research and Technology*, *6*(9), 141-143.

Rokny, H.A., Sadroddiny, E., & Scaria, V. (2017). Machine learning and data mining techniques for medical complex data analysis. *Neurocomputing*, *1*, 7-17.

Salvador, C., Martins, M.R., Vicente, H., & Caldeira, A. T. (2018). A data mining approach to ımprove ınorganic characterization of *Amanita ponderosa* mushrooms. *International Journal of Analytical Chemistry*, 2018, 1-18. doi: 10.1155/2018/5265291

Shuaib, M., Osho, O., Ismaila, I., & Alhassan, J. K. (2018). Comparative analysis of classification algorithms for email spam detection. *International Journal of Computer Network and Information Security*, *10*(1), 60-67.

Torgo, L. (2011). *Data mining with R learning with case studies*. USA: Chapman & Hall/CRC, Taylor & Francis Group.

UC Irvine Machine Learning Repository. (2019). Retrieved from https://archive.ics.uci.edu/ml/index.php

Zhang, Z. (2016). Naïve Bayes classification in R. *Annals of Translational Medicine*, *4*(12), 241-245.

Zhao, Y. (2015). *R and data mining: Examples and case studies*. Retrieved from http://www.RDataMining.com.