Citation: Tarımer, İ., & Karadağ, B. C. (2020). Comparison with Classification Algorithms in Data Mining of a Fuel Automation System's Sales Data. *Global Economics Review, V*(I), 1-10. doi:10.31703/ger.2020(V-I).20

URL: http://dx.doi.org/10.31703/ger.2020(V-I).20 **Pages:** 245 – 254 **DOI:** 10.31703/ger.2020(V-I).20

p-ISSN: 2521-2974 **e-ISSN:** 2707-0093 **L-ISSN:** 2521-2974 **Vol.** V, **No.** I (Winter 2020)

Comparison with Classification Algorithms in Data Mining of a Fuel Automation System's Sales Data

İlhan Tarımer[*] Buse Cennet Karadağ[†]

Abstract *This article deals with Otobil and pumps sales estimates at fuel stations. The fuel station data used in the study consists of 2384 data in total. Depending upon these data, classification procedures were performed on fuel station sales data using classification algorithms. In the study the classification algorithms that J48, Random Forest, KStar, Logistic Regression, IBk and Naive Bayes algorithms are used to compare the sales data estimations by using a software. The results obtained show that the accuracy rates of the J48 algorithm are more successful than others in general. It understands that these sales estimations shall encourage fuel station owners and association bodies to get more gainful.*

Key Words: Fuel Station, Accuracy, Classification Algorithms, Sales Data

JEL Classification:

Introduction

Turkey fuel sector, together with more than 13000 fuel station enterprises has an important role within the Turkish economy. The vital issues of fuel station entrepreneurs, which are important actors of the fuel sector, is the identification and prevention of wastages (Tarımer, İ., & Karadag, B. C). In the last 5 years, 7% growth, service 4 million vehicles at daily and total sales of 35 million tons of fuel with it is ranked as the 6th in Europe.

As a definition, data mining is the process of discovering new meaningful correlations, patterns and trends by using pattern recognition technologies along with statistical and mathematical techniques, through the elimination of data stacks stored in storage media (Silahtaroglu, G., Data Mining). Data mining; is the process of using previously discovered information based on a wide variety of data stored in data warehouses, making use of them to make decisions and to realize the action plan. At this point, data mining is not a solution, but a tool that supports the decision process in reaching a solution and provides the necessary information for a solution (Akgöbek, Ö., Cakir, F). The computer is responsible for determining the relationship, rules and features between the data. The aim is to detect previously undetected data patterns (Dener, M., Dörterler, M. Orman, A). Data mining models are to be analyzed beneath two separate categories that are named as definer and estimator. Estimator one aim

[*] Department of Information Systems Engineering, Muğla Sıtkı Koçman University, Menteşe / Muğla, Turkey. Email: itarimer@mu.edu.tr

[†] Department of Information Systems Engineering, Muğla Sıtkı Koçman University, Menteşe / Muğla, Turkey.

that it is to develop a modal via the data that their results are known. On the other hand, definer models, the patterns found already at current data to be used at guiding to guide is used (Tarımer, İ., & Karadag, B. C).

In this study, the performance criteria of the classification algorithms, which is one of the predictive models, will be compared using the data obtained from the system. Naive Bayes, J48, Random Forest and K star will be used as classification algorithms. There are many types of research and studies done in the literature comparing the success of classification algorithms in different fields.

Tarimer and Karadag (2020), in a data mining study based on fuel station automation data, estimated the number of fuel purchases for 2019 with the annual data they received from automation and the fuel station. The aim of this study is to support fuel stations located on the tour route in terms of warehouse management and sales forecasts and to help create a sales control scheme for the station. They used artificial neural network algorithms and multiple linear regression algorithms for estimation. As a result of the work, it was determined that the best predict algorithm is a radial-based function network, which is one of the artificial neural network algorithms. It is determined from the multiple linear regression algorithms that backward and stepwise linear regression estimates autogas, forward and stepwise linear regression unleaded gasoline, and stepwise linear regression method estimates the real values for diesel.

E. Kaya et al. Classified the Parkinson dataset using Naive Bayes, k-Nearest Neighbor (k-EYK), C4.5 Decision Trees Algorithm and DVM algorithms. In this study, the effects of parsing on the pre-classification data on the success of the classifications were observed. According to the study, the parsing process on the Parkinson dataset gave good results for all classifier algorithms used, but the DVM algorithm achieved the best result (Kaya E., O. Fındık, Babaoglu İ., Arslan A).

The flow plan of the paper is as follows: The design and interface images of the system are given in section 2. The method followed has been emphasized in section 3. In this section, the data are withdrawn to the database created through software automation; the comparison has been made by using data mining classification algorithms over these data. The test results are given at the end of this section. Conclusions are taken part in the 4th section.

Comparison of Classification Algorithms Using Data Sets and Results Obtained

In this section, a fuel station sales data is given, data mining analyzes are made based on classification algorithms depending upon these data. The estimation performances regarding all methods were measured, and the results are given in the table as for comparison.

Sales Dataset

The data used in this study are the data that is received from the established system and the data collected from the fuel station. The data set contains totally 2384 data as from price, litre, label, sales type and fuel type variables of the collected data. Some part of the dataset used here is seen in Table 1.

Table 1. Some Data of the Dataset

Price	Liter	Label	Sales Type	Fuel Type
6.4	42.76	34	OTOBIL	FS Diesel
6.4	43.97	34	OTOBIL	FS Diesel
6.4	44.53	34	Pump	FS Diesel
6.4	46.00	34	OTOBIL	FS Diesel
6.4	46.22	34	OTOBIL	FS Diesel
3.78	32.84	13	Pump	Auto Gas
3.78	34.37	48	Pump	Auto Gas
6.44	1.00	48	Pump	Diesel
6.44	1.51	45	Pump	Diesel
6.4	53.38	06	Pump	FS Diesel
6.4	54.69	34	OTOBIL	FS Diesel
7	2.79	48	OTOBIL	Unleaded
7	2.79	48	Pump	Unleaded
6.4	65.00	34	Pump	FS Diesel
…	……	………	…….	………

There are two different types of sales in the data set: Pump and Otobil. The pump sales-type represents the sales made directly by the personnel in charge. Otobil, on the other hand, is a system that enables vehicles to purchase fuel automatically from stations, without paying by cash or credit card, and delivers filling information to the customer in electronic environments, including the mileage of the vehicle.

Methods Used in the Study

The flow diagram of the study for comparison of classification algorithms is given, as shown in Fig. 1.

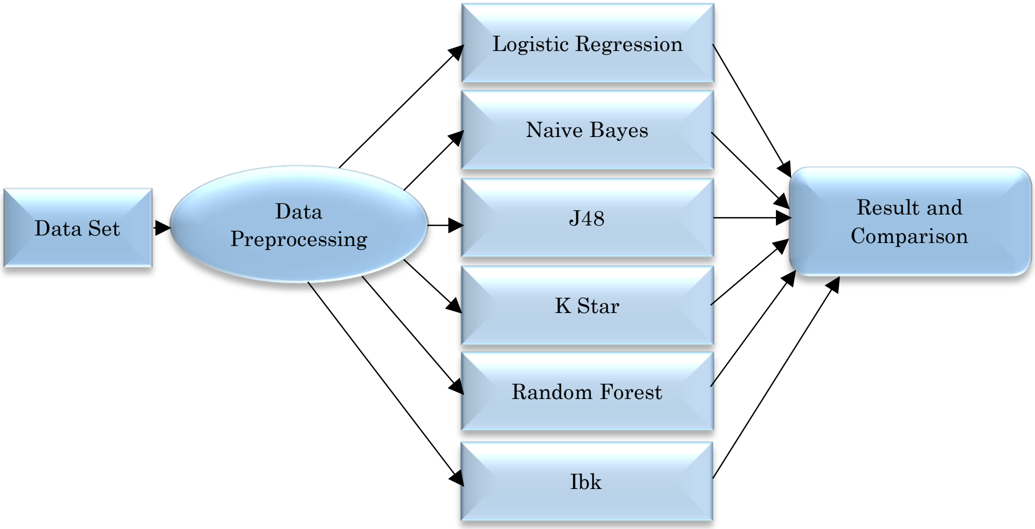


Figure 1. Flow Diagram for Comparison of Classification Algorithms

In this study, the Weka program, which is developed with the language of Java programming at Waikato University in New Zealand, was used for the analyzes. Weka is open-source software for data mining under the GNU General public license and stands for the Waikato Environment for knowledge analysis.

Weka is basically a data mining program developed in Java and distributed as open source by Waikato University, where machine learning algorithms and requirements such as data pre-processing are presented together. Weka software uses the "ARFF" (Attribute Relationship File Format) format as the file extension.

The Weka Program used for data mining analysis supports several file formats such as csv, arff, c4.5 libsvm, xarff. Therefore, the data set in Excel format has been converted into an arff file. In the study, the models were created by using J48, Random Forest, Naive Bayes, Logistic Regression, IBk and KStar algorithms; the sales-type was estimated, and the performance measures of the models were compared.

J48

The J48 algorithm developed by Quinalan is a C4.5 decision tree developed to classify nonlinear and small size data. Decision tree method is important in solving classification problems. With this method, a tree is created to model the classification process. After the tree has been created, the classification process takes place by applying it to each data group in the database. The missing values are ignored when creating the tree. Thus, estimation is performed using the remaining data. The basic idea in the J48 method is to classify using the rules produced by decision trees Quinlan J. R., (1994). Kolahkaj M., Khalilian M., (2015). The structure of a decision tree is shown in Fig. 2.

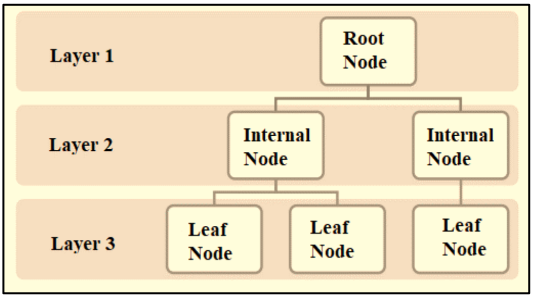


Figure 2. The structure of a decision tree

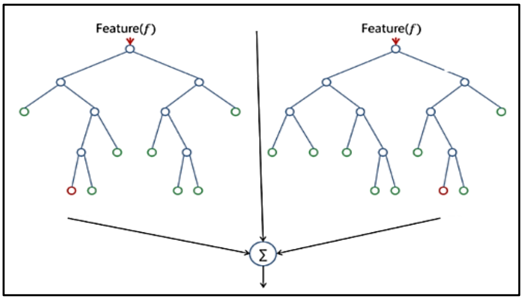


Figure 3.Random Forest Structure

Random Forest

Decision trees are one of the most used algorithms in classification problems. Decision trees are easier to structure and understand compared to other methods. In this technique (random forest algorithm), a tree is created for classification; then, each of records in the database is applied to this tree, and this record is classified according to the result. It can be said that it basically consists of two steps: The first is the establishment of the tree, and the second is the classification by applying the data to the tree one by one.

Naive Bayes

The Naive Bayes classifier is a statistics-based and highly efficient classification an algorithm based on Bayes' theorem Güldal, H., Cakici, Y., (2017). This algorithm can run on unbalanced datasets. The way the algorithm works calculates the probability of each state for an element and classifies it according to the highest probability value. With a little training data, he can do very successful jobs. If a value in the test set has an unobservable value in the training set, it gives 0 as a probability value, that is, it cannot predict (Paquin F., Rivnay J., Salleo A., Stingelin, N., Silva C.,).

K Star

The purpose of this algorithm, which is one of the learning algorithms used in data mining, is to use an unknown attribute in the test data set. For example, classification is made on the basis of comparison with the samples in the training data set that have been classified in the database but have not been revealed (Kücükönder H., Vursavus K.K., Üçkardes F.,).In addition, the K-star function can be used to classify data sets with a numeric or symbolic attribute value.

Logistic Regression

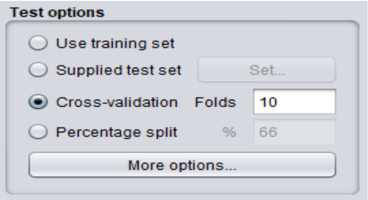
Logistic regression algorithm is like the regression problem in which the dependent variable is a categorical variable. It is widely used in linear classification problems. Although it is called regression, there is a classification here. The purpose of logistic regression is to find the most suitable model to describe the relationship between a set of independent (predictive or explanatory) variables related to the bidirectional characteristic (dependent variable = response or outcome variable). Logistic regression models have been widely used in the fields of biology, medicine, economics, agriculture and veterinary medicine and transportation in recent years (Bircan, H.).

IBk

It is the closest one to the K-Neighbor algorithm. This algorithm is used for classification, can choose the appropriate value of K-based neighbors with cross-validation, and can also weight distance (Aha D.W., Kibler, D., Albert M.K.,). This algorithm calculates the distance of each record in the database from other records when classifying.

Performances of the Generated Models

This algorithm was developed one as that is the 10-fold cross-validation method provided by the WEKA program was used as the test method (Figure 4). With this method, the data source is divided into 10 sections, and each section is used as a test set once and the remaining 9 sections as a learning set.



Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

Figure 4. Fold Cross-Validation Method

The basic concepts used when evaluating model performance are accuracy, precision, recall and F-measure. The success of the model is related to the number of samples assigned to the correct class and the number of samples assigned to the incorrect class. Accuracy is a measure of how often the classifier correctly predicts. Precision is a measure of how accurately it is predicted from all classes. It should be as high as possible. The recall is the ratio of the number of correctly classified positive samples to the total positive samples. The measure F is the recall and precision harmonic mean. It is a measure of how well the classifier is performing and is often used to compare classifiers. Performance measures of the algorithms used are in view Table 2.

Table 2. Performance Measures of the Algorithms

Algorithms	**Performance Measure (%)**			
	Accuracy	**Precision**	**Recall**	**F-Measure**
Logistic Regression	93.24	92.01	93.21	91.15
J48	93.07	91.70	93.10	91.40
Random Forest	91.48	90.50	91.50	90.90
Naive Bayes	91.73	91.10	91.70	91.40
Kstar	92.07	91.60	92.70	89.90
IBK	90.22	89.70	90.20	89.90

When the performance results of the algorithms are examined, it can be seen that the Logistic Regression algorithm correctly predicted with an accuracy rate of 93.24%. According to the accuracy criteria, after the Logistic Regression algorithm, J48, KStar, Naive Bayes, Random Forest and IBk algorithms, respectively. Examining the precision and recall rates alone will not give accurate results in determining the best performing algorithm. To determine the correct result, we need to look at the F measure rate. When the F measure rates are examined, it is seen that the best performing algorithms are J48 and Naive Bayes. These algorithms are followed by Logistic Regression, Random Forest, K Star and IBk algorithms, respectively.

The performance information of the results obtained as a result of the studies of the algorithms is expressed by the confusion matrix. In the confusion matrix, rows represent the real numbers of the samples in the test set, and the columns represent the numbers of the samples the model predicts. The structure of the confusion matrix is given in Figure 5.

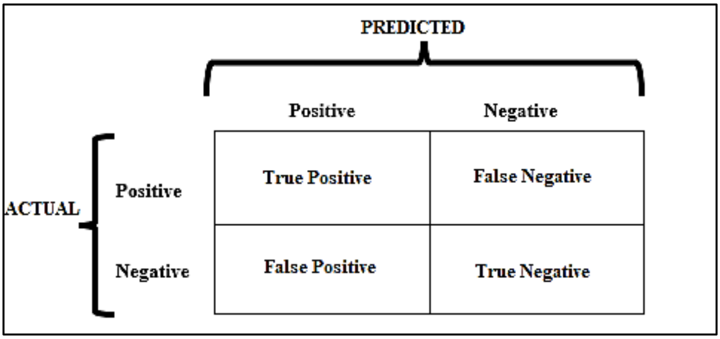
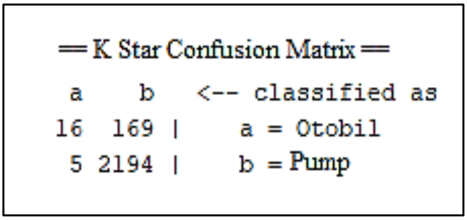


Figure 5. Structure of the Confusion Matrix

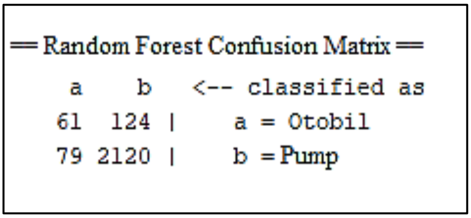
Confusion matrices are given, as shown in figures 6–11; they are the result of studies based upon the algorithms. When the 0matrices are examined, it is determined that the KStar algorithm makes 16 correct predictions and 169 incorrect predictions for Otobil. It has been determined that it has made 2194 correct and 5 false predictions for the pump. Random Forest Algorithm determines that it has made 61 correct predictions and 124 incorrect predictions for otobil. It has determined that 2120 correct and 79 false predictions for the pump are made. J48 Algorithm has determined that 39 correct predictions and 146 incorrect predictions for Otobil are made. It has been determined that it has made 2180 correct and 19 false predictions for the pump. Naive Bayes algorithm has determined that 71 correct predictions and 114 incorrect predictions for Otobil are made. It has determined that 2116 correct and 83 false predictions for the pump are made. IBk algorithm has determined that 57 correct predictions and 128 incorrect predictions for Otobil are made. It has determined that 2094 correct and 105 false predictions for the pump are made. Logistic Regression algorithm has determined that 39 correct predictions and 146 incorrect predictions for Otobil are made. It has determined that 2184 correct and 15 false predictions for the pump are made.



```
== K Star Confusion Matrix ==

  a    b   <-- classified as
 16  169 |     a = Otobil
  5 2194 |     b = Pump
```

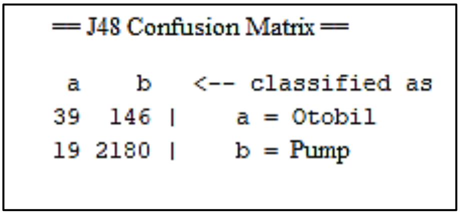
Figure 6. Confusion matrices of K Star



```
== Random Forest Confusion Matrix ==

  a    b   <-- classified as
 61  124 |     a = Otobil
 79 2120 |     b = Pump
```

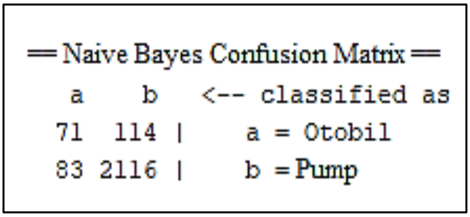
Figure 7. Confusion matrices of Random Forest



```
== J48 Confusion Matrix ==

  a    b   <-- classified as
 39  146 |     a = Otobil
 19 2180 |     b = Pump
```

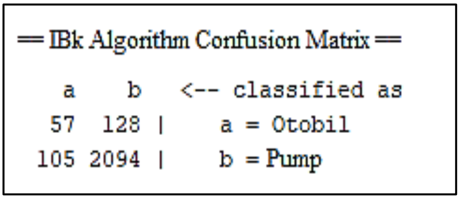
Figure 8. Confusion matrices of J48



```
== Naive Bayes Confusion Matrix ==

  a    b   <-- classified as
 71  114 |     a = Otobil
 83 2116 |     b = Pump
```

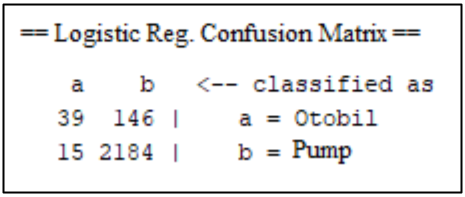
Figure 9. Confusion matrices of Naive Bayes



```
== IBk Algorithm Confusion Matrix ==

   a    b   <-- classified as
  57  128 |     a = Otobil
 105 2094 |     b = Pump
```

Figure 10. Confusion matrices of Ibk



```
== Logistic Reg. Confusion Matrix ==

  a    b   <-- classified as
 39  146 |     a = Otobil
 15 2184 |     b = Pump
```

Figure 11. Confusion matrices of Logistic Regression

When the numbers of the values predicted by the algorithms are analyzed, it is seen that the best performance for the Otobil sales type is that the Naive Bayes algorithm. It is seen that the best performance for pump sales type is KStar algorithm.

Graphs of true and false data predicted by algorithms in the confusion matrix are shown in figures 12–17. In these figures, those represented by x represent correctly classified data, and places represented by squares represent misclassified data.

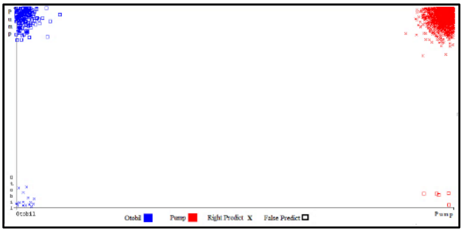


Figure 12. KStar Algorithm

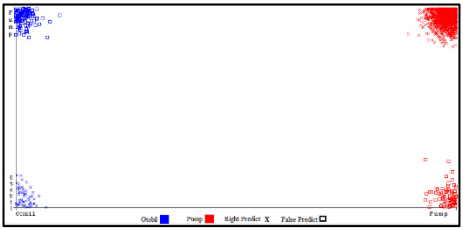


Figure 13. Random Forest Algorithm

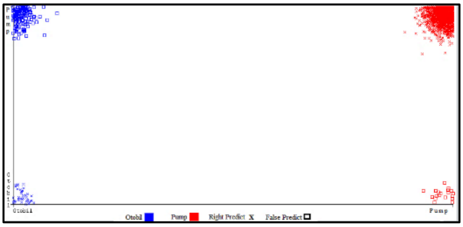


Figure 14. J48 Algorithm

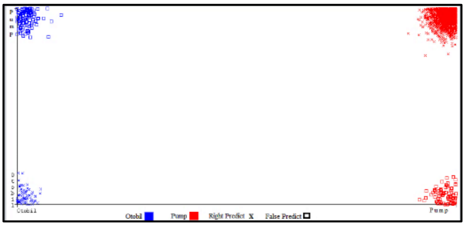


Figure 15. Naive Bayes Algorithm

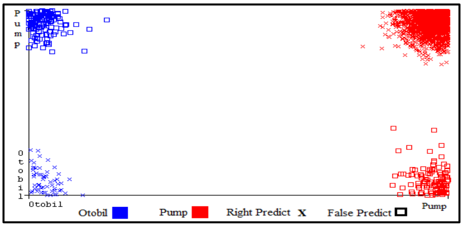


Figure 16. IBk Algorithm

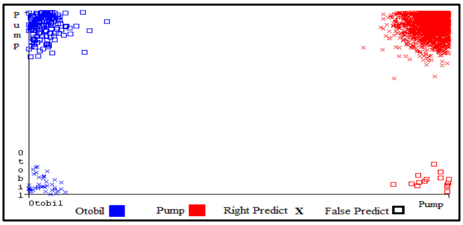


Figure 17. Logistic Regression Algorithm

In the above graphs, while the X-axis shows the actual sales-type data, the Y-axis shows the estimated sales-type data. When the graphs are analyzed, the data marked with x in blue and red indicate that the types of sales that were realized as otobil and pump were estimated correctly by the algorithm used. The blue squares indicate that the type of sales realized as otobil is estimated as a pump by the algorithms used. The red squares show that the type of sales realized as a pump is predicted otobil by the algorithms used.

Conclusions

The area that arises due to the inaccessibility of information due to the increasing amount of data forces scientists to use computer programs at Data Mining applications. In this study, a free software Weka tool was used to create models, and the classification process was performed on the data set obtained through the system using different classification algorithms. Performance values of algorithms were examined, and Logistic Regression algorithm was found to be more successful on the predictions than other algorithms in general. It has been seen that accuracy, precision and Recall rates are measured as % 93.24, % 92.01, and % 93.21 respectively. That points out that these rates obtained in the study are superior predictions when comparing them in terms of entire classification algorithms.

When the confusion matrix and graphs are analyzed for sales types, it is seen that the best prediction algorithm for Otobil sales type is obtained by means of the Naive Bayes algorithm. It is also mentioned that the best prediction algorithm for the pump sales type is obtained by means of the the KStar algorithm.

For the future works, we foresee to collect much more data when creating data sets and to use more algorithms in the classifications. Additionally, owners and associations of fuel stations in the country would be encouraged to what types of fuel sales could be more lucrative for them by using monthly and annual data.

References

Aha D.W., Kibler, D., Albert M.K., (1991). Instance-based learning algorithms. *Mach Learn 6***,** 37–66, https://doi.org/10.1007/BF00153759 .

Akgöbek, Ö., Cakir, F. (2009). An Expert System Design in Data Mining. Academic Informatics'09, XI. *Academic Informatics Conference Papers*, 809–813. https://doi.org/ISBN: 978-605-60504-1-1.

Bircan, H. (2004). Logistic Regression Analysis: An Application on Medical Data. Kocaeli University. *Journal of Social Sciences, (8)*, 185–208.

Dener, M., Dörterler, M. Orman, A. (2009). Open Source Data Mining Programs: Sample Application in WEKA. Academic Informatics '09 - XI. *Academic Informatics Conference Papers*, 787–796.

Güldal, H., Cakici, Y., (2017). Analysis of Course Management System Software Users' Interactions Using Classification Algorithms, *21*(4), pp. 1355–1367).

Kaya E., O. Fındık, Babaoglu İ., Arslan A. (2011). Effect of Discretization Method on the Diagnosis of Parkinson's Disease. *International Journal of Innovative Computing, Information and Control, 7*(8), 4669–4678.

Kolahkaj M., Khalilian M., (2015). A recommender system by using classification based on frequent pattern mining and J48 algorithm. *2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran*, pp. 780-786.

Kücükönder H., Vursavus K.K., Üçkardes F., (2015). Determining the Effect of Some Mechanical Properties on Color Maturity of Tomato with K-Star, Random Forest and Decision Tree (C4.5) Classification Algorithms. *Turkish Journal of Agriculture - Food Science and Technology, 3*(5), 300. https://doi.org/10.24925/turjaf.v3i5.300-306.261 .

Paquin F., Rivnay J., Salleo A., Stingelin, N., Silva C., (2015). Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors. *Journal of Mater. Chem. C, 3,* 10715–10722. https://doi.org/10.1039/b000000x .

Pradeep K. R., Naveen N. C., (2016). Predictive analysis of diabetes using J48 algorithm of classification techniques. *2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida*, pp. 347-352.

Quinlan J. R., (1994). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.

Silahtaroglu, G., Data Mining. (2016). *Papatya Publishing House, Istanbul / Turkey.*

Tarımer, İ., & Karadag, B. C. (2020). A Data Mining Case Study Over Fuel Sales Automation Data. *Bilecik Şeyh Edebali University Journal of Science, 7*(1), 282–296. https://doi.org/10.35193/bseufbd.611749 .