Data Mining: Classification: Decision-Tree: C4. 5

AKİF ÇAKAR
COMPUTER ENGINEERING
DOKUZ EYLUL UNIVERSITY
İZMİR, TURKEY
akif.cakar@ceng.deu.edu.tr

Abstract*—Data mining is the useful tool to discovering the knowledge from large data. Different methods & algorithms are available in data mining. Classification is most common method used for finding the mine rule from the large database. Decision tree method generally used for the Classification, because it is the simple hierarchical structure for the user understanding & decision making. Various data mining algorithms available for classification based on Artificial Neural Network, Nearest Neighbour Rule & Baysen classifiers but decision tree mining is simple one. C4.5 algorithm have been introduced by J.R Quinlan which produce reasonable decision trees. The purpose of this article is to make studies on this algorithm. First, the original version is discussed and then the effect on the data sets and how to make differences.

Keywords—Data mining; classification algorithm; decision tree; C4.5 algorithm.

I. INTRODUCTİON

Data analysis is a structure used at all times. Although the name of this structure is often different, the logic used is the same. It is aimed to interpret the results of the data obtained in all systems and use them in accordance with the results obtained. Unlike the old methods, there are too many data sets in current structures. This requires computer help and efficient algorithms. These algorithms, which are developed according to the difference in need, are the most common classification techniques. Decision trees are the sub-headings of this. With this structure, the data is transformed into a tree logic structure with branches from top to bottom using the required algorithm. This makes the data set more meaningful. When looking at the decision tree, estimates and assumptions can be made on the data obtained or the data to be obtained. This is more accurate than estimation and deduction based on meaningless data. Over time, these requirements have been developed according to data sets, one of which is the C4.5 tree algorithm. This algorithm is developed and modified structure of the existing ID3 [4] algorithm and meets the needs.

Formulas are mainly used in the structure. Theories of Shannon is at the base of the ID3 algorithm and thus C4.5. Entropy Shannon is the best known and most applied. It first defines the amount of information provided by an event: the higher the probability of an event is low, the more information it provides is great. [2];

A. Shannon Entropy

If we are given a probability distribution P = (p1, p2, … , pn) and a sample S then the Information carried by this distribution, also called the entropy of P is giving by [10]:

$$\boldsymbol{Entropie(P)} = -\sum_{i=1}^{n} \boldsymbol{p_i} \times \boldsymbol{log(p_i)} \tag{1}$$

B. The gain information G (p, T)

We have functions that allow us to measure the degree of mixing of classes for all sample and therefore any position of the tree in construction. It remains to define a function to select the test that must label the current node.

It defines the gain for a test T and a position p

$$\boldsymbol{Gain(p,T)} = \boldsymbol{Entropie(p)} - \sum_{j=1}^{n} (\boldsymbol{p_j} \times \boldsymbol{Entropie(p_j)}) \tag{2}$$

where values (pj) is the set of all possible values for attribute T. We can use this measure to rank attributes and build the decision tree where at each node is located the attribute with the highest information gain among the attributes not yet considered in the path from the root.

These mathematical structures form the main structure of the algorithm. By referring to these structures, the algorithm is created in other steps.

II. RELATED WORKS

Today, a lot of data is obtained from every field. This amount of data is so large that it cannot be catch using the human eye. In order to achieve this, meaningful conclusions are drawn from these data sets and their use is made in case of necessity by making statistics and interpretation of them. In this sense, data mining can be used in all fields and to draw meaningful conclusions or estimates from the data. Significant data relationships within big data have increased the need for this field. So it became a rapidly developing area.[1]

After the main title obtained, infrastructure was created according to the needs. One of these is the classification technique. These techniques are needed to extract information from the data set and make it understandable. Artificial intelligence is a computational process of discovering patterns in large data sets that include methods of machine learning, statistics and intersection of database systems. By subdividing the data obtained in this structure, we obtain a meaningful result.[2]

Decision trees are the most powerful approaches in knowledge discovery and data mining. It includes the technology of research large and complex bulk of data in order to discover useful patterns. This idea is very important because it enables modelling and knowledge extraction from the bulk of data available. All theoreticians and specialist are continually searching for techniques to make the process more efficient, cost-effective and accurate. Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern recognition.[3]

The C4.5 tree can be considered as an improved version of the ID3 [4] tree. The biggest difference of the C4.5 tree from the ID3 tree is that it uses normalization. In other words,

entropy calculation is made on ID3 tree and decision points are determined according to this value. In C4.5 tree, entropy values are kept as a ratio. It is also possible to move subtree to different levels according to the frequency of access on the tree. C4.5 tree Unlike the approach of ID3 tree, C4.5 tree is prunning.[5]

- All features are checked at each step.
- The normalized information gain of each feature is calculated.
- The feature that gives the best information acquisition is carried as a decision in the decision tree.
- A sub-list is then created under this new decision node to construct the sub decision tree.

C4.5 tree algorithm is a system based on formulas. Therefore, it may have different interpretations as a result of changes to be made on it. Such situations can often arise as a result of the logic on the particular set of data. These can be considered as the development of the algorithm. Such structures [6] [7] [8] have been previously discussed and realized. Minor differences in operation without disturbing the main structure affect the accuracy of the result.

III. Materials and Methods

J. Ross Quinlan originally developed ID3 [11] at the University of Sydney. He first presented ID3 in 1975 in a book, Machine Learning, vol. 1, no. 1. ID3 is based off the Concept Learning System (CLS) algorithm. The basic CLS algorithm over a set of training instances C. ID3 is a supervised learning algorithm, [12] builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. ID3 algorithm builds tree based on the information (information gain) obtained from the training instances and then uses the same to classify the test data. ID3 algorithm generally uses nominal attributes for classification with no missing values.[12] This methodology can be explained more clearly with similar structures of ID3 and C4.5. The next step is to apply it on the code.

The pseudo code of this algorithm is (Given a set of attributes not target C1, C2, ..., Cn);

```
1: Create a root node N;
2: IF (T belongs to same category C)
       {leaf node = N;
          Mark N as class C;
          Return N;
       }
3: For i=1 to n
       {Calculate Information_gain (Ai);}
4: ta= testing attribute;
5: N.ta = attribute having highest information_gain;
6: if (N.ta == continuous )
       { find threshold;}
7: For (Each T in splitting of T)
8:       if (T is empty)
             {child of N is a leaf node;}
          else
             {child of N= dtree T)}
10: calculate classification error rate of node N;
11: return N;
```

Fig. 1. Pseudocode of C4.5 algorithm

The C4.5 is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. This accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rule's precondition if the accuracy of the rule improves without it.

The C4.5 algorithm makes a better classification of the given properties. In this way, the given data set is transformed into a tree structure in the most efficient way through all the features. C4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviors.

- A possibility to use continuous data.
- Ability to use attributes with different weights.
- Pruning the tree after being created.
- Using unknown (missing) values.

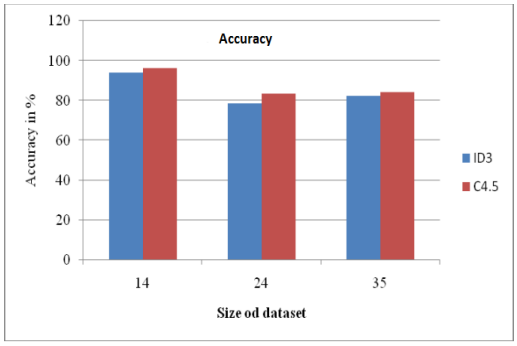

Fig. 2. Comparison of Accuracy for ID3 & C4.5 Algorithm

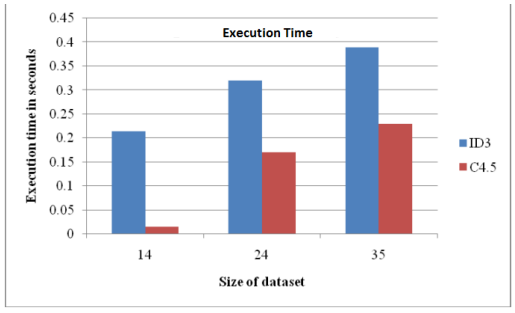

Fig. 3. Comparison of Execution Time for ID3 & C4.5 Algorithm

We can use this algorithm which we have determined the general working structure, on many systems. Even if there is no general difference on these systems, the speed of the system itself will affect the results slightly. Some of those;

- Python
- RapidMiner
- R Language
- Anaconda
- Tensorflow
- Keras
- Java
- Apache Spark

- Knime
- Weka
- Azure Machine Learning Studio
- Orange

A. *Example*

Table I Data Set S

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sun	Hot	85	Low	No
D2	Sun	Hot	90	High	No
D3	Overcast	Hot	78	Low	Yes
D4	Rain	Sweet	96	Low	Yes
D5	Rain	Cold	80	Low	Yes
D6	Rain	Cold	70	High	No
D7	Overcast	Cold	65	High	Yes
D8	Sun	Sweet	95	Low	No
D9	Sun	Cold	70	Low	Yes
D10	Rain	Sweet	80	Low	Yes
D11	Sun	Sweet	70	High	Yes
D12	Overcast	Sweet	90	High	Yes
D13	Overcast	Hot	75	Low	Yes
D14	Rain	Sweet	80	High	No

In this example we are going to detail the calculation of information gain for an attribute of continuing value.[12]

Gain (S, Humidity) =?

We must now sort the attribute values in ascending order, the set of values is as follows:

{65, 70, 70, 70, 75, 78, 80, 80, 80, 85, 90, 90, 95, 96}

we will remove values that are repeated:

{65, 70, 75, 78, 80, 85, 90, 95, 96}

Table II GAIN CALCULATION FOR THE ATTRIBUTE CONTINUOUS HUMIDITY USING C4.5 ALGORITHM

	65		70		75		78		80		85		90		95		96	
interval	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
Yes	1	8	3	6	4	5	5	4	7	2	7	2	8	1	8	1	9	0
No	0	5	1	4	1	4	1	4	2	3	3	2	4	1	5	0	5	0
Entropy	0	0.961	0.811	0.971	0.721	0.991	0.65	1	0.764	0.971	0.881	1	0.918	1	0.961	0	0.94	0
Info(S,T)	0.892		0.925		0.8950		0.85		0.838		0.915		0.929		0.892		0.94	
Gain	0.048		0.015		0.045		0.09		0.102		0.025		0.011		0.048		0	

Gain (S, Humidity) = 0.102

Suppose the unknown value of D12 day for visibility attributes.

Info(S)= -8/13*log2 (8/13)-5/13* log2 (5/13)= 0.961

Info (Outlook, S) = 5/13*Entropy (SSun)

+ 3/13* Entropy(Sovercast)

+ 5/13* Entropy(SRain)

= 0.747

Entropy (SSun) =-2/5* log2 (2/5) –3/5* log2 (3/5)= 0.9710

Entropy (SOvercast) =-3/3*log2 (3/3) –0/3* log2 (0/3)=0

Entropy (SRain) =-3/5* log2 (3/5) –2/5* log2 (2/5)= 0.9710

Gain (Outlook) = 13/14 (0.961 – 0.747) = 0.199

When a case of S with the known value is assigned to the subsets Si, the probability belonging to Si is 1, and in all other subsets is 0.

Fractionation of the set S using the test on the attribute visibility. A new wi weight is equal to the probability in this case: 5/13, 3/13 and 5/13, because the initial value (Table 2) w is |S1| = 5+5/13, |S2| = 3 +3/13, and |S3| = 5+5/13.

IV. RESULTS

C4.5 was theoretically explained and exemplified in the preceding sections. In this part, analysis and tests will be performed on a data set using this algorithm and different data mining tools. Currently, many systems produce data and this data accumulation results in high data sets. By selecting the appropriate algorithm for these data sets and using data preparation techniques, useful information in the data is obtained. We will apply this with a real-life example.

A. *Explanation of Dataset*

No.	Name
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	poutcome
17	y

Fig. 4. Bank Marketing Dataset's Attributes

The dataset[13] that selected contains information about whether customers who have a bank account but who do not have a credit card require a credit card as a result of telephone marketing. This information includes categorical and numerical data. This data determines our descriptive features. On the other hand, we have a target feature to see if they accept the credit card. (y is the target feature.) We have features that can capture the difference between customers thanks to the age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign,

pdays, previous and poutcome properties of a customer. There are ‘17’ feature and ‘45211’ data row in the dataset.

B. *Data Preparation Techniques*

Before working on our data, data preparation techniques was performed. These operations are applied to avoid situations that affect data mining results and to obtain clearer results. By following the steps below, the correct, incomplete, meaningless and incorrect information in the data was corrected.

- Gather data
- Discover and assess data
- Clean and validate data
- Transform and enrich data
- Store data

Corrections are often the replacement of missing, inaccurate or meaningless information with the most repetitive data.

Our data set is now ready. The next step is to determine the test techniques. In the test we will do on our data set, 90% of the data will be used for training and 10% of the data will be used for the test. This selection is performed randomly using a single step and another method, cross-validation. (Fold number was taken as 10) Evaluation metrics such as accuracy, precision, recall, F, ROC, MSE, RMSE, MAE, R2, which are appropriate to the algorithm are found using the data obtained as a result of these steps.[14]

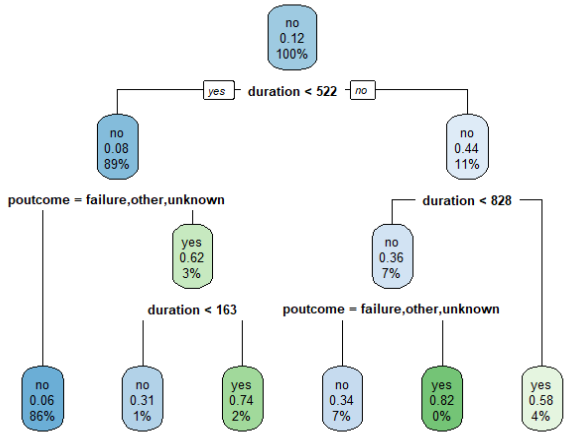

Fig. 5. The DT Model Of Dataset

C. *Selecting Data Mining/Machine Learning Tools*

We have provided all necessary steps for the testing phase. The next step is to decide which tool to use. Although different tools offer different interfaces, the algorithms running within them often produce similar results, since they are often the same. The most important difference for this is that a data mining tool provides meaningful and functional visualizations to the user. These factors were at the forefront of my choice. C4.5 algorithm was run on the data using four different data mining tools. In most data mining tools, the C4.5 algorithm is referred to as the J48 library. The tools used for this data analysis test are:

- Weka
- RapidMiner
- R
- Orange

D. *Implementing the Selected Dataset on the Tools via C4.5*

1) *RapidMiner Studio*

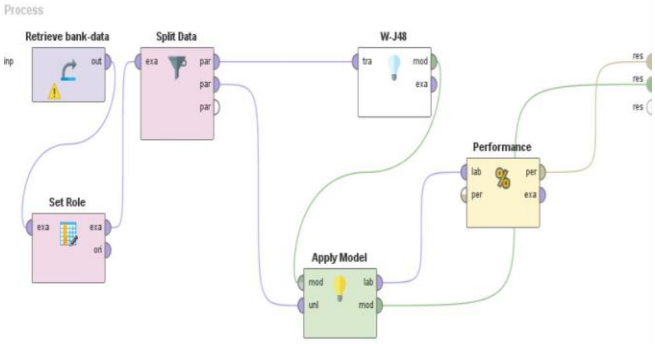

Fig. 6. RapidMiner Studio Schema

As a first, the data set was tested using RapidMiner Studio. The necessary steps for this test are as shown in the figure 6. After the data set is received, the target property is selected. Afterwards, test and training data are selected and separated according to the determined features. RapidMiner connects to the W-J48 package which corresponds to C4.5 and runs the model. In order to see the data more semantically and to see the test results, after the performance connection is made, it is connected to the result part and the structure is completed.

accuracy: 90.69%

	true no	true yes	class precision
pred. no	3829	258	93.69%
pred. yes	163	271	62.44%
class recall	95.92%	51.23%	

Fig. 7. The Model’s Result at RapidMiner

The result is a 90.08% success rate. The probability of predicting 'yes' is greater than the probability of predicting 'no'. Evaluation metrics were obtained using ConfisionMatrix.

Performance Vector (Performance)
Result not stored in repository.

```
PerformanceVector:
accuracy: 90.08%
ConfusionMatrix:
True:    no      yes
no:      11452   820
yes:     525     767
```

Fig. 8. The Confusion Matrix of Model using RapidMiner

The result shows that a successful selection of a tree predicts our dataset.

2) Orange

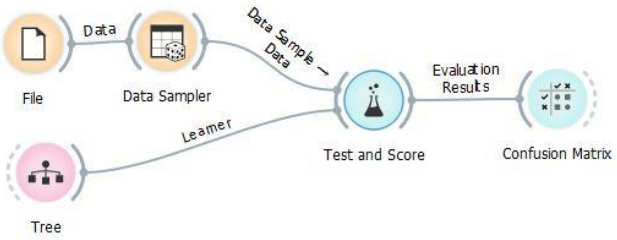

Fig. 9. Orange Schema

Another tool is Orange. In this tool, as in figure 9 structure was established. As with the other tool, the features of the data set are defined, the training data and the test data are determined and the required algorithm is selected and run to obtain the result.

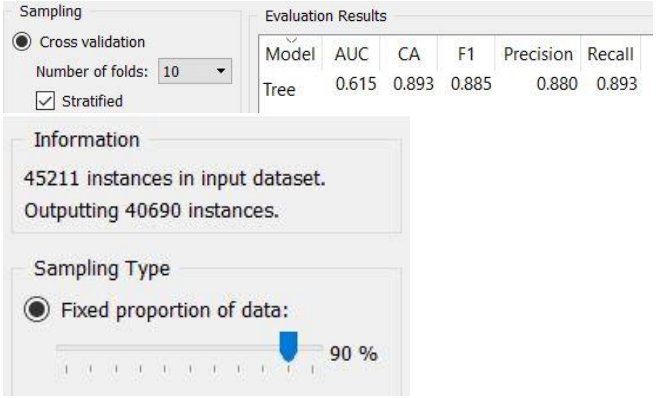

Fig. 10. Adjustments for Splitting Data

In Figure 10, the model is run after the necessary adjustments are made for the training and test data. The results show that the data is modeled with a success rate of 88%. At the same time the value obtained in the test result is shown in Figure 11.

Fig. 11. The Model's Result at Orange

3) R Studio

This data mining tool has more code-weighted structure than others. For this, a structure is established as shown in figure 12. The data taken from the computer in the line of code is converted into test and training data within the range of desired values. The model is then created and tested using the J48 library, which corresponds to C4.5.

```
library(caTools)
library(RWeka)

setwd('C:\\Users\\akifc\\Desktop')
options(max.print=999999)

data <- read.csv("bank-full.csv")
spl = sample.split(data$y, SplitRatio = 0.9)

dataTrain = subset(data, spl==TRUE)
dataTest = subset(data, spl==FALSE)

resultJ48 <- J48(as.factor(y)~., dataTrain)
dataTest.pred <- predict(resultJ48, newdata = dataTest)
table(dataTest$y, dataTest.pred)


summary(resultJ48) # calls evaluate_Weka_classifier()
```

Fig. 12. R Code for The C4.5 Algorithm

Thanks to this code structure, our decision tree is formed and we get the results in figure 13.

```
=== Summary ===

Correctly Classified Instances       38319               94.173  %
Kappa statistic                          0.6912
Mean absolute error                      0.0946
Root mean squared error                  0.2175
Relative absolute error                 45.7851 %
Root relative squared error             67.6671 %
Total Number of Instances            40690

=== Confusion Matrix ===

     a     b   <-- classified as
 35220   710 |     a = no
  1661  3099 |     b = yes
```

Fig. 13. The Confusion Matrix and Result of the Model at R

As shown in Figure 12, our structure was modeled with a success rate of 94.173% and produced results. Unlike other tools, the rate of predicting 'yes' is high. As a result, modeling has achieved more success. In addition, evaluation metrics are shown.

4) Weka

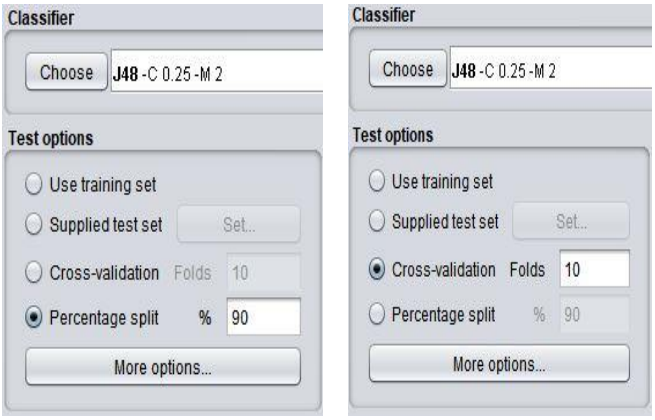

Fig. 14. Adjustments for Splitting Data

Another data mining tool is weka. As shown in Figure 14, after selecting the data set, adjustments are made for test and training data and the model is run. Like other tools, this tool is based on model selection and does not require a code extension.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40831               90.3121 %
Kappa statistic                          0.4839
Mean absolute error                      0.1269
Root mean squared error                  0.2773
Relative absolute error                 61.4259 %
Root relative squared error             86.2833 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,959    0,519    0,933      0,959    0,946      0,488    0,843     0,947     no
               0,481    0,041    0,609      0,481    0,537      0,488    0,843     0,486     yes
Weighted Avg.  0,903    0,463    0,895      0,903    0,898      0,488    0,843     0,893

=== Confusion Matrix ===

     a     b   <-- classified as
 38289  1633 |     a = no
  2747  2542 |     b = yes
```

Fig. 15. The Confusion Matrix and Result of the Model at Weka

After the model is run, the result in figure 15 is obtained. As we have seen here, similar to other tests, we achieved 90.3121% success. The logic common to all the different steps we take is this; In a model, a decision tree is created with the data allocated for education. Then, the decision tree created using the data given for the test is tested and the accuracy of the model is calculated by comparing the predicted results with the actual results.

We tested the same data in four different data mining tools with the same characteristics. As a result, we have obtained models and results. When the results are examined, the logic and results of the different tools working on this model are similar. However, it has obtained a better model by achieving a different result than the others. This tool is R Studio, as can be understood from the data. In fact, although many use the J48 library for C4.5, such a difference has been achieved. There can be many reasons for this. Some of them; randomly selected while the data is allocated for training and testing. This randomness affects the result. Another possibility is that R optimizes the J48 in itself, resulting in better modeling. Different reasons like this can affect the result. In general, the results were obtained close to each other and the data was modeled.

V. Conclusion

In data mining, classification algorithms form a decision-making algorithm by transforming data into a tree structure. One of these algorithms, C4.5, is an improved version of the ID3 algorithm. These two algorithms were developed by 'J.R Quinlan.. There is a fundamental difference between these two algorithms, which are similar in general structure. ID3 determines the decision tree through categorical properties. The C4.5 algorithm can create a decision tree, even if it contains numeric data. It is an efficient difference. By breaking the numerical properties according to the range criteria, the data in the data set distributes between these ranges and creates a decision algorithm. The basis of this structure is entropy process. Thus, a data is obtained by calculating the entropy of a property that affects the result. With this data, the algorithm determines the characteristics of the branches of the tree. A feature with the highest information gain is the most logical choice. This structure recursively repeats until a single decision is made in each branch. Thanks to the decision tree created, the results are obtained by following the characteristics of future data on the tree. Thus, the estimation is realized.

There are many data mining tools. These tools, which have the same general structure, may have small differences and different structures as development logic. For the C4.5 algorithm, the J48 library is generally used. Thus, this decision tree algorithm can be applied to the desired data set. As a result of the applied algorithm, a result is obtained about how successful the modeling is. The height of this result increases the success rate and more accurate predictions can be made. At the same time, finalizing the data using data preparation techniques increases the success rate for the algorithm. Both the accuracy of the data and the correct implementation of the algorithm affect the success rate.

References

[1] Hand, D. J. (2006). Data Mining. Encyclopedia of Environmetrics, 2.(

[2] Kesavaraj, G., & Sukumaran, S. (2013, July). A study on classification techniques in data mining. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[3] Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6).

[4] Jin, C., De-Lin, L., & Fen-Xiang, M. (2009, July). An improved ID3 decision tree algorithm. In 2009 4th International Conference on Computer Science & Education (pp. 127-130). IEEE.

[5] Korting, T. S. (2006). C4. 5 algorithm and multivariate decision trees. Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil..

[6] Mazid, M. M., Ali, S., & Tickle, K. S. (2010, February). Improved C4. 5 algorithm for rule based classification. In Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases (pp. 296-301). World Scientific and Engineering Academy and Society (WSEAS).

[7] Xiaoliang, Z., Hongcan, Y., Jian, W., & Shangzhuo, W. (2009, December). Research and application of the improved algorithm C4. 5 on decision tree. In 2009 International Conference on Test and Measurement (Vol. 2, pp. 184-187). IEEE.

[8] Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications, 117(16), 18-21.

[9] Benjamin Devéze & Matthieu Fouquin, DATAMINING C4.5 – DBSCAN, PROMOTION 2005, SCIA Ecole pour l'informatique ettechniques avancées.

[10] Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. International Journal of Advanced Computer Science and Applications, 4(2), 0-0.

[11] J.R. QUINLAN, Induction of Decision Trees, 1986, Machine Learning 1:81-106.

[12] Ankur Shrivastava and Vijay Choudhary ,Comparison between ID3 and C4.5 in Contrast to IDS Surbhi Hardikar, VSRD-IJCSIT, Vol. 2 (7), 2012, 659-667.

[13] Bank Marketing Dataset, 2017, Sandeep Verma - https://www.kaggle.com/skverma875/bank-marketing-dataset

[14] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1