Quest Journals
Journal of Research in Applied Mathematics
Volume 7 ~ Issue 1 (2021) pp: 17-30
ISSN(Online) : 2394-0743 ISSN (Print): 2394-0735
www.questjournals.org

Research Paper

The comparison of the model performances of Naive Bayes, C4.5 and C5.0 algorithms: Implementation on fish consumption habits

Şenol Çelik
Bingol University, Faculty of Agriculture, Department of Animal Science, Biometry and Genetic Bingol, Turkey
Correspondence: Şenol Çelik (senolcelik@bingol.edu.tr)

ABSTRACT
This study was conducted to implement some data mining methods to examine the fish consumption habits of individuals. A total of 466 individuals were implemented the scale in this study. In addition to demographic and socioeconomic characteristics of individuals such as gender, marital status, monthly income and monthly food spend, fish consumption frequency, consumption amount, the reason for preference, and fish species preference were also examined. Fish species and consumption frequency classification was made according to the socioeconomic and demographic characteristics and the reason for fish consumption. Naive Bayes and Decision Tree (C4.5 and C5.0) algorithms of data mining were used for classification processes. It was considered that the classification error rate was low and the classification accuracy rate was high in the comparison of the classification model formed.
Considering the comparison of fish species and fish consumption frequency classification, it was found that the classification model formed with the C5.0 algorithm (error rates as 43.8% and 35.6% and accuracy rates as 56.2% and 64.4%, respectively) was a better classifier than the other model.
KEYWORDS: *Data mining, decision tree, naive bayes, confusion matrix.*

Received 23 Jan, 2021; Revised: 04 Feb, 2021; Accepted 07 Feb © The author(s) 2021.
P*ublished with open access at www.questjournals.org*

I. INTRODUCTION

Naive Bayes, C4.5 and C5.0 algorithms are among very important classification techniques in data mining.

Data mining constitutes useful and valuable information in the big databases by combining techniques in the machine learning, statistics and database fields (Ching and Michael, 2002). Data mining is the process of elimination of big amount of packed data in the storage mediums using statistical and mathematical techniques and pattern recognition technologies and the discovery of new significant correlations, patterns and trends (Larose, 2005). Data mining is the process of discovering significant patterns and rules in big amount of data (Linoff and Berry, 2011). Data mining is the entirety of all activities used to find new, hidden or unexpected patterns within data (Marakas, 2003).

Data mining process includes various steps. These are as follows: problem statement, data collection, tailoring the data for analysis, the implementation step of data mining methods, performance evaluation of the results obtained, and the evaluation of the results of the method with the best performance (Kantardzic, 2011).

C4.5 focuses on classification problem in decision trees. Two procedure steps are generally followed to form the most suitable tree structure. The first one is to for the tree structure with the training data set and the second is to carry put the pruning procedure in the tree structure formed. As the number of qualifications increases in the data sets, tree to be structured forms unnecessary nodes. This situation called overlearning negatively affects the success rate. The pruning procedure is usually implemented to eliminate the negative effects due to overlearning (Quinlan, 1986: Quinlan, 1993).

C5.0 algorithm is a classification algorithm based on binary or more splits. The information gain is addresses here as a splint standard. A decision tree that splits to two or more following each decision node is formed with this algorithm. Entropy is calculated in the determination step of the decision nodes in the decision tree, and the information gains of the qualifications identified as input in the training data set are determined accordingly. After this procedure, the decision node is formed with the qualification with the highest gain (Pandya and Pandya, 2015).

This study was conducted to compare the results by implementing the Naive Bayes, C4.5 and C5.0 algorithms in the evaluation results of a survey.

II. MATERIAL AND METHOD

Material

The population of the study included people who lived in Bingöl, Turkey and neighboring provinces. However, sampling was made as it was almost impossible to reach the entire population and obtain information about them in terms of time and cost. A survey study was carried out in February 2020 mostly in Bingöl, Turkey (Van, Elazığ, Diyarbakır and Muş), and neighboring provinces and the survey was implemented to 506 people. However, 40 individuals who were determined to have not consumed fish through the survey were excluded and the remaining 466 surveys were analyzed.

Method

The Naive Bayes classification is based on the Bayes theorem. In this classification y_j is $j \in \{1,2,\dots,J\}$ and is a discrete variable that represents one of the J number of classes. X feature is expressed with the $X=(x_1, x_2,\dots, x_n)$ feature vector with n number of features. According to the Bayes theorem, the posterior probability $p(y_j|X)$ for the yj value is expressed as follows (Deng et al., 2015).

$$P(y_j|X) = \frac{P(X|y_j)P(y_j)}{P(X)}$$

When the multiplication of the conditional probabilities of all features is calculated to predict the Y' target class, an equation is obtained for the Naive Bayes classifier.

$$Y' = argmax\, y_j \frac{P(y_j)\prod_{i=1}^{n}(P(X = x_i|j_j)}{\sum_{j=1}^{j} P(y_j)\prod_{i=1}^{n}(P(X = x_i|j_j)}$$

The probabilities are calculated by implementing this equation for each class, and then, the one with the biggest probability among the resulting values is selected as the target class (Deng et al., 2015).

$$P(y_j) = \frac{|N_j|}{|N|}$$

In the last equation, the total number of the training qualifications in the class label is $|N|$, and the number of training qualifications in the y_j class is $|N_j|$ (Bai and Nie, 2004).

C4.5 algorithm is one of the classifier of data mining where it is easy to understand and interpret the decision tree, and to process data rapidly. The class structures in this algorithm are the prediction of the classes of the data without classes within the previously determined categorical data structures, the presentation of the information visually in the form of a tree, and the easy modeling of the correlations between the variables (Biggs et al., 1991).

Decision trees look like a tree from top to bottom in the root, node and branch structure, and in this structure, the root and each node report a question while the branches separated from the nodes report the answers to that question (Loh and Shih, 1997).

Information gain rate is used as the feature selection standard in the C4.5 algorithm. Thus, the feature with the highest information gain rate is selected (Duru, 2016; Şatır et al., 2016). The C4.5 algorithm has important advantages. These are that it is used both in categorical and numerical data sets, it is used when there are lacking attribute values in the training data, and the unwanted values in the training data set are eliminated (Ture et al., 2009).

The C4.5 algorithm operates based on the concept of entropy which is defined as the uncertainty of a system and the measure of randomness. In the entropy given in the following equation, the aim is to calculate the amount of information needed to classify the data set (Kavzoğlu and Çölkesen, 2010).

$$Entropy(S) = -\sum_{i=1}^{n} p_i\, log_2(p_i)$$

In this equation, p_1, p_2,…,p_n are the probabilities of the target variable's classes in the data. S is the data set while n is the number of different values that qualification can have. The probability values of a class is calculated as $p_i = C_i/|S|$. C_i are the classes of the data set (Kavzoğlu and Çölkesen, 2010).
The information gain of each qualification is calculated to determine the distinctive qualification in the decision tree methods. The information gain of an A feature for a S sample is calculated according to the following equation (Bahety, 2014; Larose and Larose, 2014).

$$Gain(A,S) = Entropy(S) - \sum_{i=1}^{n} \frac{S_i}{S} Entropy(S_i)$$

The information gain is calculated for each feature determined according to this equation. The feature with the highest information gain is defined as the root. These processes require the following for each node; samples should belong to the same class and samples should not separate into features that can be split. It continues until one of the conditions that all features are represented by the suitable class is met (Adhatrao, 2013; Bahety, 2014).
The split information of the S attribute is calculated as follows (Quinlan, 1993).

$$Split(A,S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} log_2\left(\frac{|S_i|}{|S|}\right)$$

The ratio of information gain to split information expresses how much information the splinting of the relevant qualification will provide. Thus, the information gain is calculated for each qualification and the tree structure is split based on the highest information gain.

The pruning procedure is extremely important for the formation of the decision tree structure. Pruning is the procedure performed to stop splitting so that the tree does not grow anymore after reaching a certain size during the formation of the tree structure (Quinlan, 1987).

The C5.0 algorithm is ideal for big databases and is the higher level of the C4.5 algorithm. C5.0 algorithm is also known as the boosting trees. While the information gain is used for the process of splitting into branches, the tree pruning is based on the error rate for each leaf. The C5.0 algorithm is much faster than C4.5 and uses the memory more efficiently. Although the results of the C5.0 are the same as those of the C4.0 algorithm, C5.0 obtains more proper decision trees in form (Çalış et al., 2014). Additionally, it produces better results in terms of minimizing the decision trees and producing decision rules (Shahnaz, 2006).

The decision tree is formed by maximizing the information gain and entropy values of the qualifications to split the decision nodes in the classification made with the C5.0 algorithm (Alpaydın, 2013; Silahtaroğlu, 2013). The steps used to form the decision tree are as in the C4.5 algorithm.

R program was used for the data analysis, and in the evaluation of the models formed with the algorithms, the comparisons were made according to the root mean square error (RMSE), mean absolute error (MAE), root relative squared error (RRSE) and relative absolute error (RAE) criteria. These equations and the terms included in these equations are given below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - E_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|O_i - E_i|^2}$$

$$RAE = \frac{\sum_{i=1}^{n}|O_i - E_i|}{\sum_{i=1}^{n}|E_i - \bar{E}_i|}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}|O_i - E_i|^2}{\sum_{i=1}^{n}|E_i - \bar{E}_i|^2}}$$

O_i is the observed value and E_i expected value.
Accuracy measures determined based on the confusion matrix such as the accuracy and error rates are used to evaluate the success of the models (Coşkun and Baykal, 2011).

$$\text{Correctly Classified} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error\ rate = \frac{TP + FN}{TP + TN + FP + FN}$$

Here, TP: True positive, TN: True negative, FP: False positive, FN: False negative.

III. RESULTS AND DISCUSSION

Socioeconomic and demographic characteristics related to the household fish consumption are presented in Table 1.

Table 1. Socioeconomic and demographic characteristics of fish consumers

Sex	**Frequency**	**Percent (%)**
Man	301	59.5
Woman	205	40.5
Total	506	100
Martial		
Married	185	36.6
Single	321	63.4
Total	506	100
Income		
1500-3000 TL	228	45.1
3000-4000 TL	148	29.2
4000-5000 TL	90	17.8
5000 TL+	40	7.9
Total	506	100
Food cost		
500-1000 TL	305	60.3
1000-1500 TL	115	22.7
1500-2000 TL	65	12.8
2000 TL+	21	4.2
Total	506	100
Consumption frequent		
Several times a month	94	18.6
1 per month	261	51.6
Several times a year	87	17.2
Does not consume	64	12.6
Total	506	100
Consumption amount		
0-1 kg	340	67.2

1-2 kg	108	21.3
2 kg+	58	11.5
Total	506	100
Consumption form		
Choking	30	5.9
In the oven	128	25.3
Grid	108	21.3
No consume	62	12.3
Frying in Oil	178	35.2
Total	506	100
Reason for preference		
Nutritious	94	18.6
Delicious	136	26.9
Healthy	182	36.0
No consume	59	11.7
Cheap	35	6.9
Total	506	100
Type of fish		
Alabalık	137	27.1
Hamsi	191	37.7
İstavrit	23	4.5
Levrek	42	8.3
Palamut	54	10.7
Somon	19	3.8
No consume	40	7.9
Total	506	100

Fish species preference

Which fish species will be chosen changes based on the economic and social conditions of the individuals. The chosen fish species are black sea salmon, anchovy, horse-mackerel, sea bass, bonito and salmon. The scale was implemented to 506 people. 40 people who did not consume fish were excluded from the evaluation and the analysis was made with the remaining 466 people.

The Naive Bayes classification results that explain the prior and conditional probabilities of the factors that affect the fish species preferences are presented in Table 2. The prior probability was calculated the highest for "Anchovy" species (41%). The accurate classification rate of the Naive Bayes classification was found as 46.35%

Table 2. Naive Bayes classification

```
A-priori probabilities:
Y
  Alabalik      Hamsi   istavrit     Levrek    Palamut      Somon
0.29399142 0.40987124 0.04935622 0.09012876 0.11587983 0.04077253

Conditional probabilities:
          Sex
Y                 man     woman
  Alabalik 0.5766423 0.4233577
  Hamsi    0.6596859 0.3403141
  istavrit 0.3913043 0.6086957
  Levrek   0.3571429 0.6428571
```

```
  Palamut  0.6296296 0.3703704
  Somon    0.5789474 0.4210526

          Marital
Y            married    single
  Alabalik 0.4160584 0.5839416
  Hamsi    0.3298429 0.6701571
  istavrit 0.6956522 0.3043478
  Levrek   0.3809524 0.6190476
  Palamut  0.3888889 0.6111111
  Somon    0.2631579 0.7368421

          Income
Y           1500_3000  3000_4000  4000_5000 5000_10000
  Alabalik 0.37956204 0.40145985 0.16788321 0.05109489
  Hamsi    0.49214660 0.28795812 0.16230366 0.05759162
  istavrit 0.17391304 0.30434783 0.34782609 0.17391304
  Levrek   0.38095238 0.14285714 0.30952381 0.16666667
  Palamut  0.55555556 0.16666667 0.18518519 0.09259259
  Somon    0.42105263 0.21052632 0.10526316 0.26315789

          Food spend
Y           1000_1500  1500_2000  2000_3000   500_1000
  Alabalik 0.17518248 0.09489051 0.02189781 0.70802920
  Hamsi    0.21989529 0.13612565 0.04188482 0.60209424
  istavrit 0.08695652 0.56521739 0.04347826 0.30434783
  Levrek   0.42857143 0.04761905 0.09523810 0.42857143
  Palamut  0.29629630 0.12962963 0.00000000 0.57407407
  Somon    0.21052632 0.10526316 0.21052632 0.47368421

          Preference
Y           be_healty being_cheap  delicious nutritious
  Alabalik 0.41605839  0.08029197 0.24817518 0.25547445
  Hamsi    0.39790576  0.13612565 0.29842932 0.16753927
  istavrit 0.17391304  0.26086957 0.34782609 0.21739130
  Levrek   0.50000000  0.09523810 0.16666667 0.23809524
  Palamut  0.31481481  0.11111111 0.37037037 0.20370370
  Somon    0.31578947  0.10526316 0.52631579 0.05263158
```

The C4.5 algorithm results that show the factors related to the fish species preference are presented in Table 3 and Figure 1.

Table 3. Classification results of the C4.5 algorithm (fish species preference)

```
Food spend = 1000_1500
|   preference = be_healty
|   |   income = 1500_3000: Hamsi (17.0/6.0)
|   |   income = 3000_4000
|   |   |   sex = man: Palamut (3.0/1.0)
|   |   |   sex = woman: Alabalik (6.0/3.0)
|   |   income = 4000_5000
|   |   |   sex = man: Alabalik (5.0/3.0)
|   |   |   sex = woman: Levrek (6.0/2.0)
|   |   income = 5000_10000: Alabalik (5.0/3.0)
|   preference = being_cheap: Hamsi (16.0/7.0)
|   preference = delicious
|   |   income = 1500_3000: Hamsi (18.0/10.0)
|   |   income = 3000_4000
|   |   |   marital = married: Hamsi (2.0)
|   |   |   marital = single: Palamut (3.0)
|   |   income = 4000_5000: Alabalik (6.0/2.0)
|   |   income = 5000_10000: Levrek (2.0/1.0)
|   preference = nutritious: Levrek (17.0/8.0)
Food spend = 1500_2000
|   preference = be_healty
|   |   income = 1500_3000: Palamut (4.0/2.0)
|   |   income = 3000_4000
|   |   |   sex = man: Alabalik (5.0/2.0)
```

```
|   |   |   sex = woman: Hamsi (4.0/1.0)
|   |   income = 4000_5000: Hamsi (4.0/1.0)
|   |   income = 5000_10000: Hamsi (1.0)
|   preference = being_cheap: Hamsi (2.0/1.0)
|   preference = delicious
|   |   income = 1500_3000: Alabalik (6.0/4.0)
|   |   income = 3000_4000
|   |   |   marital = married: Alabalik (3.0/2.0)
|   |   |   marital = single: Hamsi (3.0)
|   |   income = 4000_5000: istavrit (7.0/1.0)
|   |   income = 5000_10000: istavrit (2.0/1.0)
|   preference = nutritious: Hamsi (22.0/10.0)
foodspend = 2000_3000
|   preference = be_healty: Levrek (8.0/4.0)
|   preference = being_cheap: Alabalik (2.0/1.0)
|   preference = delicious: Hamsi (9.0/3.0)
|   preference = nutritious: Alabalik (1.0)
foodspend = 500_1000
|   income = 1500_3000: Hamsi (137.0/77.0)
|   income = 3000_4000
|   |   preference = be_healty
|   |   |   sex = man: Hamsi (24.0/9.0)
|   |   |   sex = woman: Alabalik (20.0/8.0)
|   |   preference = being_cheap: Alabalik (7.0/4.0)
|   |   preference = delicious
|   |   |   sex = man: Hamsi (12.0/2.0)
|   |   |   sex = woman: Alabalik (7.0/3.0)
|   |   preference = nutritious: Alabalik (26.0/10.0)
|   income = 4000_5000: Hamsi (35.0/21.0)
|   income = 5000_10000
|   |   preference = be_healty: Levrek (2.0)
|   |   preference = being_cheap: istavrit (1.0)
|   |   preference = delicious: Palamut (3.0/1.0)
|   |   preference = nutritious: Alabalik (3.0/1.0)

Number of Leaves  :    41

Size of the tree :     59
```

The confusion matrix is presented in Table 4. The following calculation results were found: Correctly Classified Instances=53.86%, Kappa statistic=0.2961, Mean absolute error=0.1971, Root mean squared error=0.3139, Relative absolute error=81.9761%, Root relative squared error=90.63%.

Table 4. Confusion Matrix

```
   a   b   c   d   e   f   <-- classified as
  56  76   0   4   1   0 |   a = Alabalik
  26 158   0   6   1   0 |   b = Hamsi
   5   8   8   1   1   0 |   c = istavrit
   4  17   0  20   1   0 |   d = Levrek
   6  35   1   3   9   0 |   e = Palamut
   5  12   1   1   0   0 |   f = Somon
```

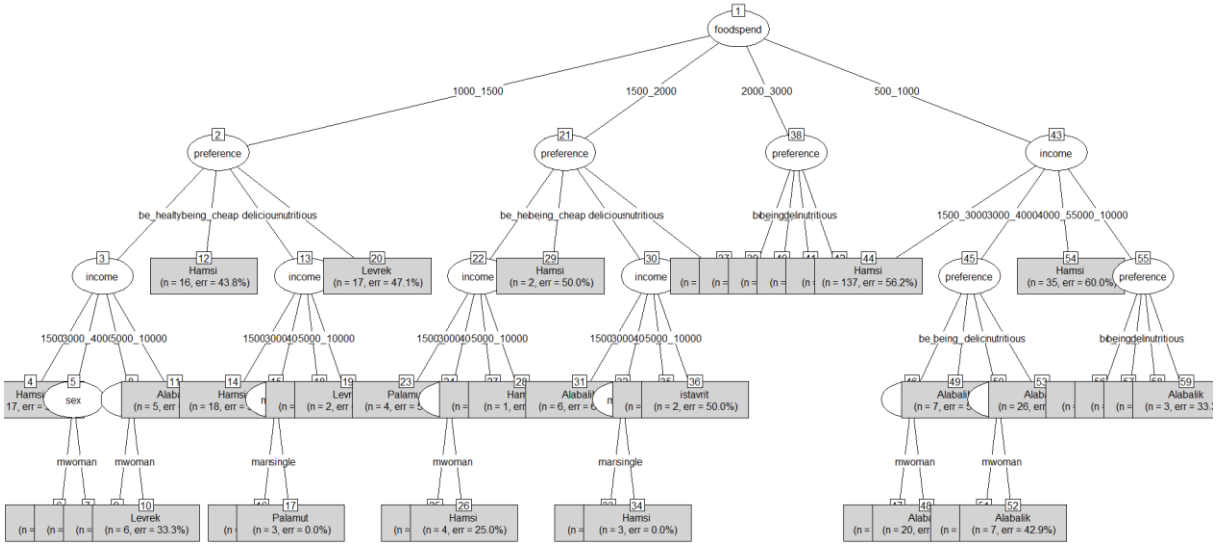

Figure 1. C4.5 algorithm

The C5.0 algorithm results regarding the factors that affected the fish species preference are given in Table 5. The confusion matrix is given in Table 6 and the C5.0 decision tree is given in Figure 2.

Table 5. Classification results of the C5.0 algorithm (fish species preference)

```
Food spend = 1500_2000:
:...marital = married:
:   :...preference in {be_healty,being_cheap,nutritious}: Hamsi (22/7)
:   :   preference = delicious: istavrit (11/4)
:   marital = single:
:   :...preference = being_cheap: istavrit (1)
:       preference = delicious:
:       :...income in {1500_3000,3000_4000,4000_5000}: Hamsi (8/3)
:       :   income = 5000_10000: istavrit (2/1)
:       preference in {be_healty,nutritious}:
:       :...income in {3000_4000,5000_10000}: Alabalik (7/2)
:           income = 4000_5000: Hamsi (7/4)
:           income = 1500_3000:
:           :...preference = be_healty: Palamut (3/1)
:               preference = nutritious: Alabalik (2/1)
foodspend in {1000_1500,2000_3000,500_1000}:
:...foodspend = 2000_3000:
    :...preference = be_healty: Levrek (8/4)
    :   preference in {being_cheap,nutritious}: Alabalik (3/1)
    :   preference = delicious: Hamsi (9/3)
    foodspend = 500_1000:
    :...income in {1500_3000,4000_5000}:
    :   :...marital = married: Alabalik (58/36)
    :   :   marital = single: Hamsi (114/55)
    :   income = 5000_10000:
    :   :...preference = be_healty: Levrek (2)
    :   :   preference = being_cheap: istavrit (1)
    :   :   preference = delicious: Palamut (3/1)
    :   :   preference = nutritious: Alabalik (3/1)
    :   income = 3000_4000:
    :   :...sex = woman: Alabalik (43/18)
    :       sex = man:
    :       :...preference in {be_healty,being_cheap,
    :           :              delicious}: Hamsi (38/12)
    :           preference = nutritious: Alabalik (15/6)
    foodspend = 1000_1500:
    :...preference = being_cheap: Hamsi (16/7)
        preference = nutritious: Levrek (17/8)
```

```
        preference in {be_healty,delicious}:
        :...income = 1500_3000: Hamsi (35/16)
            income = 5000_10000:
            :...preference = be_healty: Alabalik (5/3)
            :   preference = delicious: Levrek (2/1)
            income = 3000_4000:
            :...preference = be_healty:
            :   :...sex = man: Palamut (3/1)
            :   :   sex = woman: Alabalik (6/3)
            :   preference = delicious:
            :   :...marital = married: Hamsi (2)
            :       marital = single: Palamut (3)
            income = 4000_5000:
            :...sex = woman:
                :...preference = be_healty: Levrek (6/2)
                :   preference = delicious: Alabalik (1)
                sex = man:
                :...marital = married: Hamsi (5/2)
                    marital = single:
                    :...preference = be_healty: Palamut (1)
                        preference = delicious: Alabalik (4/1)
```

Table 6. Confusion Matrix

```
   (a)   (b)   (c)   (d)   (e)   (f)    <-classified as
  ----  ----  ----  ----  ----  ----
            75    56     1     4     1        (a): class Alabalik
            37   147           6     1        (b): class Hamsi
             4     8    10     1              (c): class istavrit
             9    12           20     1        (d): class Levrek
            15    24     2     3    10        (e): class Palamut
             7     9     2     1              (f): class Somon
```

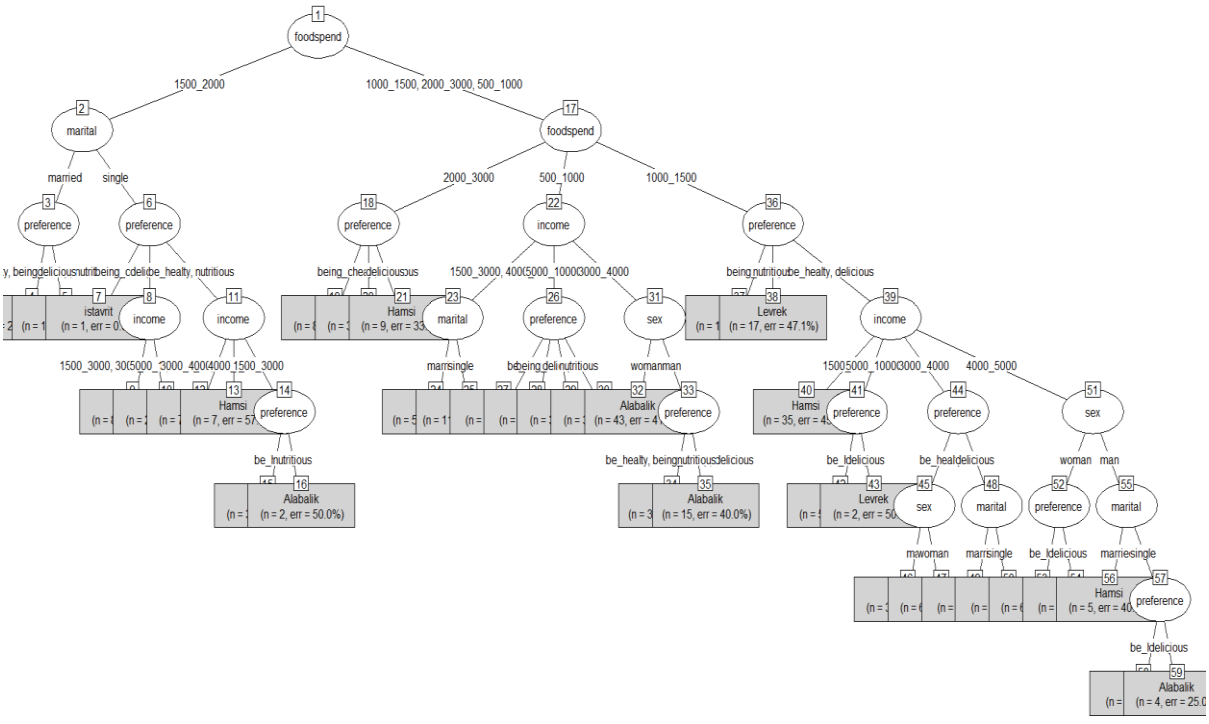

Figure 2. C5.0 decision tree structure

Since the classification error was 43.8% in the C5.0 algorithm, the accuracy rate was 56.2%. Compared with other algorithms, C5.0 algorithm had higher level of accuracy rate than the Naive Bayes and C4.5

algorithms. While the most effective factor regarding the preferred fish species was the monthly food spend in the C5.0 algorithm, and the most important second factor was the monthly income. The decision rules and results of all effective factors are given in Figure 2 in detail.

Fish consumption frequency

Fish consumption frequency changes based on the economic and social conditions of the individual as well as her palate for food. Consumption frequency has three options as "several times a month", "once in a month" and "several times a year". Of 506 individuals who participated in the survey, 466 consumed fish; thus, the statistical analysis was made accordingly.

The Naive Bayes classification results that explain the prior and conditional probabilities of the factors that affect the fish consumption frequency are presented in Table 2. The highest prior probability belonged to "several times a month" with the ratio of 55.79%. The accurate classification rate of the Naive Bayes classification was found as 59.87%.

Table 7. Naive Bayes classification

```
A-priori probabilities:
Y
     few_a_week several_a_month  several_a_year
      0.2017167       0.5579399       0.2403433

Conditional probabilities:
                  sex
Y                      man     woman
  few_a_week      0.5957447 0.4042553
  several_a_month 0.5846154 0.4153846
  several_a_year  0.5892857 0.4107143

                  marital
Y                    married    single
  few_a_week      0.4787234 0.5212766
  several_a_month 0.3961538 0.6038462
  several_a_year  0.2678571 0.7321429

                  income
Y                   1500_3000  3000_4000  4000_5000 5000_10000
  few_a_week      0.42553191 0.14893617 0.25531915 0.17021277
  several_a_month 0.36538462 0.38846154 0.19230769 0.05384615
  several_a_year  0.61607143 0.18750000 0.11607143 0.08035714

                  foodspend
Y                   1000_1500  1500_2000  2000_3000   500_1000
  few_a_week      0.27659574 0.15957447 0.06382979 0.50000000
  several_a_month 0.19230769 0.13846154 0.04615385 0.62307692
  several_a_year  0.26785714 0.10714286 0.01785714 0.60714286

                  preference
Y                   be_healty being_cheap  delicious nutritious
  few_a_week      0.37234043  0.09574468 0.40425532 0.12765957
  several_a_month 0.45000000  0.07307692 0.22307692 0.25384615
  several_a_year  0.25892857  0.24107143 0.35714286 0.14285714
```

The C4.5 algorithm results that show the factors related to fish consumption frequency are presented in Table 8 and Figure 3.

Table 8. Classification results of the C4.5 algorithm (fish consumption frequency)

```
Income = 1500_3000
|   preference = be_healty: several_a_month (74.0/34.0)
|   preference = being_cheap: several_a_year (29.0/12.0)
|   preference = delicious
|   |   foodspend = 1000_1500: several_a_year (18.0/8.0)
|   |   foodspend = 1500_2000: several_a_year (6.0/2.0)
|   |   foodspend = 2000_3000: several_a_month (2.0)
|   |   foodspend = 500_1000
|   |   |   sex = man: several_a_month (26.0/13.0)
|   |   |   sex = woman
```

```
|   |   |   |   marital = married: several_a_month (5.0/2.0)
|   |   |   |   marital = single: several_a_year (13.0/5.0)
|   preference = nutritious: several_a_month (31.0/12.0)
income = 3000_4000: several_a_month (136.0/35.0)
income = 4000_5000
|   preference = be_healty: several_a_month (33.0/12.0)
|   preference = being_cheap
|   |   marital = married: several_a_year (3.0/1.0)
|   |   marital = single: several_a_month (7.0/3.0)
|   preference = delicious
|   |   marital = married: few_a_week (13.0/3.0)
|   |   marital = single: several_a_month (10.0/4.0)
|   preference = nutritious: several_a_month (21.0/6.0)
income = 5000_10000
|   preference = be_healty: few_a_week (12.0/6.0)
|   preference = being_cheap
|   |   sex = man: few_a_week (4.0/1.0)
|   |   sex = woman: several_a_year (2.0)
|   preference = delicious
|   |   sex = man: several_a_month (10.0/5.0)
|   |   sex = woman: few_a_week (2.0)
|   preference = nutritious: several_a_month (9.0/5.0)

Number of Leaves  :    22
Size of the tree :     33
```

The relevant confusion matrix is given in Table 9. The following calculation results were found: Correctly Classified Instances=63.73%, Kappa statistic=0.2946, Mean absolute error=0.3363, Root mean squared error=0.4101, Relative absolute error=85.40%, Root relative squared error=92.45%.

Table 9. Confusion Matrix (for fish consumption frequency)

```
   a   b   c   <-- classified as
  21  64   9 |   a = few_a_week
   8 233  19 |   b = several_a_month
   2  67  43 |   c = several_a_year
```

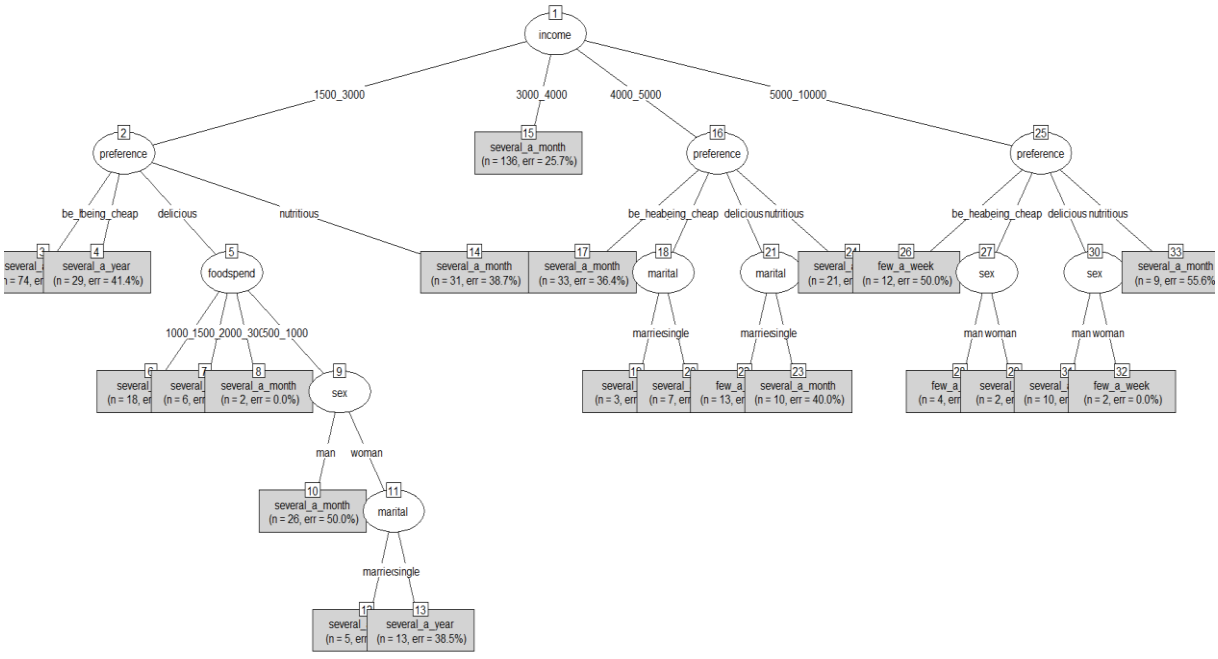

Figure 3. C4.5 algorithm to determine the frequency of fish consumption

The C5.0 algorithm results regarding the factors that affected the fish species preference are given in Table 10. The confusion matrix is given in Table 11 and the C5.0 decision tree is given in Figure 4.

Table 10. Classification results of the C5.0 algorithm (fish consumption frequency)

```
preference in {be_healty,nutritious}: several_a_month (275/92)
preference in {being_cheap,delicious}:
:...income in {4000_5000,5000_10000}:
    :...income = 5000_10000: few_a_week (18/10)
    :   income = 4000_5000:
    :   :...marital = single: several_a_month (17/7)
    :       marital = married:
    :       :...preference = being_cheap: several_a_year (3/1)
    :           preference = delicious: few_a_week (13/3)
    income in {1500_3000,3000_4000}:
    :...foodspend = 1000_1500: several_a_year (32/14)
        foodspend = 2000_3000: several_a_month (3)
        foodspend = 1500_2000:
        :...marital = married: few_a_week (4/2)
        :   marital = single: several_a_year (9/2)
        foodspend = 500_1000:
        :...preference = being_cheap:
            :...sex = man: several_a_year (19/6)
            :   sex = woman: several_a_month (10/3)
            preference = delicious:
            :...sex = man: several_a_month (38/18)
                sex = woman:
                :...income = 1500_3000:
                    :...marital = married: several_a_month (5/2)
                    :   marital = single: several_a_year (13/5)
                    income = 3000_4000:
                    :...marital = married: several_a_year (2)
                        marital = single: several_a_month (5/1)
```

Table 11. Confusion Matrix (belong to frequency of fish consumption)

```
  (a)   (b)   (c)    <-classified as
 ----  ----  ----
   20    66     8    (a): class few_a_week
   10   230    20    (b): class several_a_month
    5    57    50    (c): class several_a_year
```

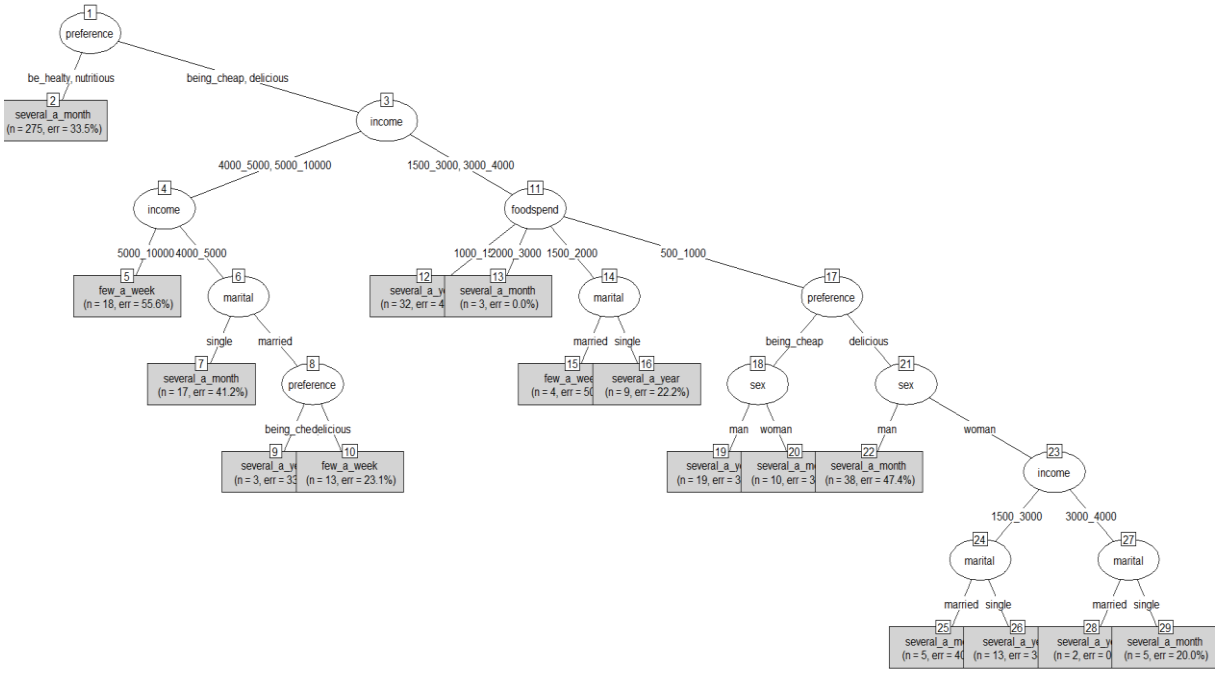

Figure 4. C5.0 decision tree structure for the frequency of fish consumption

Since the classification error was 35.6% in the C5.0 algorithm, the accuracy rate was 64.4%. Compared with other algorithms used in this study, C5.0 algorithm had higher level of accuracy rate than the Naive Bayes and C4.5 algorithms. The most effective factor in detecting the fish consumption frequency was the reason for fish preference followed by the monthly income. The decision rules and results of all effective factors are given in Figure 4 in detail. For instance, those with a monthly income between 5000-10000 Turkish Liras among those who prefer fish based on low price and taste and with a monthly income between 4000-5000 and 5000-10000 Turkish Liras consume fish several times a week (Node 1, Node 3, Node 4 and Node 5). Similarly, single individuals with a monthly income between 4000-5000 Turkish Liras among those who prefer fish based on low price and taste and with a monthly income between 4000-5000 and 5000-10000 Turkish Liras consume fish several times a week (Node 1, Node 3, Node 4, Node 6 and Node 7). Married individuals with the same conditions and preference based on taste consume fish several times a week or a year (Node 1, Node 3, Node 4, Node 6, Node 8, Node 9 and Node 10). The results stated in the other rules and nodes can be interpreted similarly.

Similar to this study, R program was used in a study on data mining classification algorithms. the study was conducted with the C5.0 and Gini algorithms, and it was reported that C5.0 algorithm had higher performance (Çınar, 2019). Küçükönder et al. (2015) made the classification of some mechanical features about the color ripeness of tomato with the K-Star, Random forest and C4.5 classification algorithms. The authors stated that the classification model formed with the K-Star algorithm produced better results. Another study used the logistic regression, C5.0, CART and Support Vector Machine methods for the classification of the return on equity, and the highest accurate classification success belonged to the CART algorithm (Yakut and Gemici, 2017).

IV. CONCLUSION

This study was carried out to determine the factors that affected the predicted classification related to the fish species preference and fish consumption frequency of people who lived in Bingöl, Turkey and neighboring provinces in 2020. Analyses were made using the consumption habits of the individuals and the Naive Bayes, C4.5 and C5.0 algorithm methods. Considering the analysis results of the C4.5 and C5.0 algorithms, the most important factors that affected the classification prediction for the fish species were monthly food spend, income, and the reason for preference. As a result of the C4.5 algorithm, the most important factors that affected the fish consumption frequency were monthly income and the reason for consumption while the most important factors that affected the consumption frequency in the C5.0 algorithm were the reason for consumption, monthly income and monthly food spend. In the study on C4.5 and C5.0 algorithms and fish species preference, 59 rules were created for each. In the classification of the fish consumption frequency, 33 rules were created in C4.5 algorithm and 29 rules were created in C5.0 algorithm.

When analysis methods were examined in terms of model performance, C5.0 algorithm was found to have better performance than Naive Bayes and C4.5 algorithms. In conclusion, data mining and classification algorithms are expected to be useful for the further stages of the evaluation of the consumers' behaviors.

ACKNOWLEDGEMENT

The survey was implemented by some students who study in the Department of Plant Protection and Field Crops in the Faculty of Agriculture at Bingöl University in Turkey. I would like to thank the students named Ayşenur Saidoğlu, Ezgi Günay, Fırat Yetkin, Gamze Matur, Murat Kaymazalp, Tümay Şimşek, Berfin Polat and Halil İbrahim Kandemir who participated in the implementation of the survey.

REFERENCES

[1]. Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., Honrao, V. 2013. Predicting Students' Performance using ID3 and C4. 5 Classification Algorithms. arXiv preprint arXiv:1310.2071.
[2]. Alpaydın, E. 2013. Yapay öğrenme. Boğaziçi Üniversitesi, İstanbul.
[3]. Bahety, A. 2014. Extension and Evaluation of ID3–Decision Tree Algorithm. Entropy (S), 2(1): 1-8
[4]. Bai, J., Nie, J. Y. 2004. Using language models for text classification. In: Proceedings of the Asia Information Retrieval Symposium (AIRS)
[5]. Bulut, F. 2016. Çok Katmanlı Algılayıcılar ile Doğru Meslek Tercihi. Anadolu University Journal Of Science And Technology–A Applied Sciences and Engineering, 17(1): 97- 109.
[6]. Ching, W. K., Michael, K. P. 2002. Advances in Data Mining and Modeling, World Scientific, Hong Kong.
[7]. Çalış, A., Kayapınar, S., Çetinyokuş, T. 2014. Veri Madenciliğinde Karar Ağacı Algoritmaları
[8]. ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama. Endüstri Mühendisliği Dergisi, 25(3-4): 2-19.
[9]. Çınar, A. 2019. Veri madenciliğinde sınıflandırma algoritmalarının performans değerlendirmesi ve R dili ile bir uygulama. Marmara Üniversitesi Öneri Dergisi, 14(51):
[10]. 90-111.
[11]. Coşkun C, Baykal. A. 2011. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması, Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, Malatya, 2 - 4 Şubat 2011 İnönü Üniversitesi, 51-58.
[12]. Deng, H., Sun, Y., Chang, Y., Han, J. 2015. Probabilistic Models for Classification. C.C. AGGARWAL (Eds.), Data Classification Algorithms and Applications, CRC Press, New York, USA.
[13]. Kantardzic, M. 2011. Data Mining Concepts, Models, Methods, and Algorithms. A John Wiley and Sons, Inc., Second Edition, USA.
[14]. Kavzoğlu T, Çölkesen İ. 2010. Karar Ağaçları ile Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örneği, Harita Teknolojileri Elektronik Dergisi 2(1): 36-45.
[15]. Küçükönder, H. 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. Türk Tarım-Gıda Bilim ve Teknoloji Dergisi, 3(5): 300-306.
[16]. Larose, D. T. 2005. Discovering Knowledge in Data: An Introduction in Data Mining, Wiley, USA.
[17]. Larose, D. T., Larose, C. D. 2014. Discovering Knowledge in Data an Introduction to Data Mining, New Jersey: John Wiley and Sons.
[18]. Linoff, G. S., Berry, M. J. A. 2011. Data Mining Techniques for Marketing, Sales and Customer Relationship Management, Wiley, Canada.
[19]. Marakas, G. M. 2003. Decision Support Systems in the 21st Century, Prentice Hall, USA.
[20]. Pandya, R., Pandya, J. 2015. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. International Journal of Computer Applications (0975– 8887), 117(16):18-21.
[21]. Quinlan, J. R. 1986. Induction of decision trees. Machine Learning, 1: 81-106
[22]. Quinlan, J. R. 1987. Simplifying decision trees. International Journal of Man-Machine Studies, 27: 221-234.
[23]. Quinlan, J. R. 1993. C4.5 Programs for Machine Learning: Morgan Kaufmann.
[24]. Shahnaz, F. 2006. Decision Tree Based Algorithms, Michael W. Berry (Ed.), Lecture Notes in Data Mining, USA: World Scientific Publisher.
[25]. Silahtaroğlu, G. 2013. Veri madenciliği. Papatya Yayıncılık Eğitim, İstanbul.
[26]. Şatır, E., Azboy, F., Aydın, A., Arslan, H., Hacıefendioğlu, Ş. 2016. Veri İndirgeme ve Sınıflandırma Teknikleri ile Glokom Hastalığı Teşhisi. El-Cezeri Journal of Science and Engineering, 3(3): 485-497.
[27]. Ture, M., Tokatli, F., Kurt, I. 2009. Using Kaplan–Meier analysis together with decision tree methods (CART, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Systems with Applications, 36(2): 2017-2026.
[28]. Yakut, E., Gemici, E. 2017. LR, C5.0, CART, DVM Yöntemlerini Kullanarak Hisse Senedi Getiri Sınıflandırma Tahmini Yapılması ve Kullanılan Yöntemlerin Karşılaştırılması: Türkiye'de BIST'de Bir Uygulama. Ege Akademik Bakış, 17(4): 461-479.