Available online: October 10, 2018

Commun. Fac. Sci. Univ. Ank. Ser. A1 Math. Stat.
Volume 68, Number 1, Pages 762–770 (2019)
DOI: 10.31801/cfsuasmas.469131
ISSN 1303–5991 E-ISSN 2618-6470
http://communications.science.ankara.edu.tr/index.php?series=A1

COMMUNICATIONS
SERIES A1

A DATA SCIENCE STUDY FOR DETERMINING FOOD QUALITY: AN APPLICATION TO WINE

A.E. OZALP AND I.N. ASKERZADE

Abstract. In this paper, wine quality is investigated based on physicochemical ingredients which include fixed acidity, volatile acidity, citric acid, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulphate and alcohol, by ANFIS (Adaptive Neuro Fuzzy Inference System) method and by random forest algorithm which is a powerful classification algorithm. Although this study specifically investigate the relation between physicochemical ingredients and the quality of wine, the results can be adaped to determination of the quality of any food product in terms of the ingredients.

1. INTRODUCTION

Producers are in a competition of making their production more competitive, since human's existential pleasure is depended mostly on consumption. This competition exists also in the food industry which is fundamental for human life. It is difficult to predict the quality of the food or to measure the consumer's tasting. Fuzzy logic is an effective method for complex concepts like human's emotions and feelings to reduce the complexity. According to the archeological records, wine has been in human's life for 9500 years [1]. Like the other foods and drinks, quality of wine also has been an important factor for people's preference of wine consumption. Although determining and controlling the quality of wine seems oenologists' and food engineers' field, in recent years, with the developments of data mining, data scientists start to contribute in the area of the quality control.

In this study, to decide the quality of red wine, a fuzzy logic method which is ANFIS and the random forest algorithm will be used. Although the concept of quality is heuristic and includes emotion, fuzzy logic methods are supposed to reduce the heuristic complexity. As another method for quality control, random forest algorithm is chosen, because it is the most successful classification algorithm according to Delgado [2]. Quality control will be done using UCI (University of

Received by the editors: November 23, 2017; Accepted: March 30, 2018.

2010 *Mathematics Subject Classification.* 68P01, 68Q25, 68T27,94D05, 03B52.

Key words and phrases. Adaptive Neuro Fuzzy Inference System, fuzzy logic, data science, random forest algorithm.

California, Irvine) machine learning repository's red wine quality dataset which is uploaded by Paulo Cortez (a website which has approximately 400 data sets) [13]. There are two dataset for red and white wine in the repository. In this paper, red wine quality is investigated.

In literature, the first study for food quality assessment using fuzzy logic is in 1995 [3]. Goel and Perrot give a review of previous research on food quality using fuzzy logic [4, 5]. In [6] and [7] fuzzy logic methods are used to classify the data sets. The mathematical basis of fuzzy logic is also developed in [8].

2. DATASET

UCI machine learning repository's red wine quality dataset is used to determine the quality of red wine. In the creation of the dataset, the physicochemical properties are given for 1599 red wines for which a number of Oenologists taste the wines and give a score between 0, to the worst, and 10, to the best, to each wine. The quality value of each wine is determined from the average of the scores given by the Oenologists for that wine. Inputs of the dataset include the quality value and the amount of fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates and alcohol. The aim of this study is to determine the quality of red wine by using the fuzzy logic and random forest algorithm for the physicochemical properties of red wine, and compare the results by the scores of Oenologists'.

Table 1 shows first ten data of wine quality data set. Table 2 [9] gives the statistics of wine quality data set.

Table 1. First ten data of data set.

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.70	0	1.9	76	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	98	25	67	0.9968	3.20	0.68	9.8	5
7.8	0.76	0.04	2.3	92	15	54	0.9970	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	75	17	60	0.9980	3.16	0.58	9.8	6
7.4	0.70	0	1.9	76	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	75	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.60	0.06	1.6	69	15	59	0.9964	3.30	0.46	9.4	5
7.3	0.65	0	1.2	65	15	21	0.9946	3.39	0.47	10.0	7
7.8	0.58	0.02	2.0	73	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.50	0.36	6.1	71	17	102	0.9978	3.35	0.80	10.5	5

Table 2. Statistics of data set.

Attribute(units)	Red wine		
	Min	Max	Mean
Fixed acidity(g(tartaric acid/dm^3)	4.6	15.9	8.3
Volatile acidity(g(acetic acid/dm^3)	0.1	1.6	0.5
Citric acid(g/dm^3)	0.0	1.0	0.3
Residual sugar(g/dm^3)	0.9	15.5	2.5
Chlorides(g(sodium chloride)/dm^3)	0.01	0.61	0.08
Free sulfur dioxide(mg/dm^3)	1	72	14
Total sulfur dioxide(mg/dm^3)	6	289	46
Density(g/cm^3)	0.990	1.004	0.996
pH	2.7	4.0	3.3
Sulphates(g(potasssium sulphate/dm^3)	0.3	2.0	0.7
Alcohol(vol%)	8.4	14.9	10.4

2.1. **Classification.** Decision tree is one of the most popular classification method in data mining. This classifier is a supervised learning method and it is built from a set of training examples. A decision tree has nodes, branches and leaves. The name of the top node is root node. All decision trees start from the root node. They grow top to bottom at each level. Nodes are connected by branches and leaves [10].

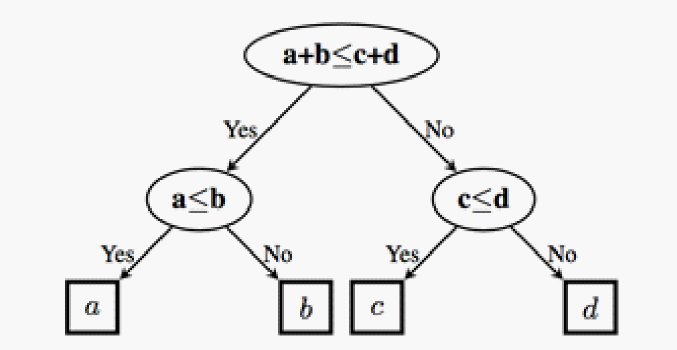


FIGURE 1. An example of decision tree.

Another algorithm for classification is Random Forest Algorithm. As it is understood from its name, random forest algorithm creates the forest with a number of trees. The random forest algorithm is developed by Leo Breiman and Adele Cutler [2]. This approach is a supervised, decision tree based algorithm. It is the most successful classification algorithm [2].

ANFIS is a network structure which based on Takagi –Sugeno fuzzy inference system. It suggested by Roger Jang in 1993 [11]. This system takes advantages of fuzzy logic's decision making feature and artificial neural networks' learning ability, and thus, it combines human's intelligence and artificial intelligence.

2.2. **ANFIS Architecture.** ANFIS has 5 layers which is shown in Figure 2.

A.: Layer 1: This layer is fuzzification layer.

$$\begin{aligned} O_i^1 &= \mu_{A_i}(x), \quad i=1,2 \\ O_i^2 &= \mu_{B_i}(y), \quad i=1,2 \end{aligned}$$

where x and y are input values, A_i and B_i are linguistic variables. A_i and B_i are degrees of membership of membership functions. For membership degree calculation, Gaussian and Bell membership functions are used.

B.: Layer 2: This layer is called the rule layer.

$$w_i = \mu_{A_i}(x) \times \mu_{B_i}(y), \qquad i = 1, 2.$$

C.: Layer 3: Third layer is the normalization layer.

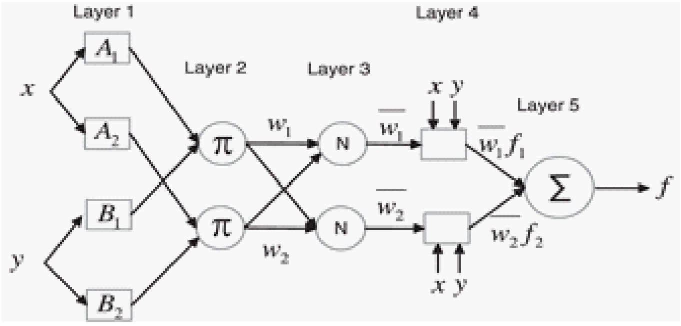


FIGURE 2. ANFIS architecture.

$$\overline{w}_i = \frac{w_i}{w_1 + w_2}, \qquad i = 1, 2.$$

D.: Layer 4. In this layer defuzzification is done.

$$O_i^3 = \overline{w}_i f_i = \overline{w}_i (p_i(x) + q_i(y) + r_i), \qquad i = 1, 2.$$

In this formula w_i is layers' output value, $\{p_i, q_i, r_i\}$ is the set of fuzzy inference system's posteriori parameters.

E.: Layer 5. This layer is called output layer.

$$O_i^4 = y = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{w_i f_i}, \qquad i = 1, 2.$$

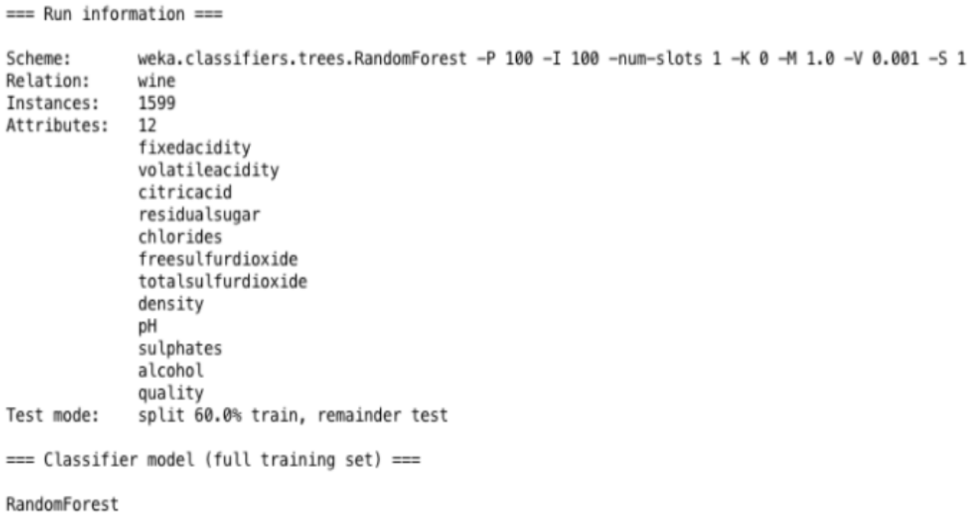
2.3. **Hybrid Learning Algorithm.** There are two parameter sets; S_1 states input parameter set S_2 states output parameter set. Total parameter set is $S = S_1 + S_2$. There are two states in hybrid learning algorithm which are backfeed and forward-feed. S1 is obtained by least squares method and S_2 is obtained by linearization [12]. Output is as follows:

$$\begin{aligned}
f &= \frac{w_1}{w_1 + w_2} f_1 + \frac{w_2}{w_1 + w_2} f_2 \\
f &= w_1 f_1 + w_2 f_2 \\
f &= (w_1 x)\, p_1 + (w_1 y)\, q_1 + (w_1)\, r_1 + (w_2 x)\, p_2 + (w_2 y)\, q_2 + (w_2)\, r_2
\end{aligned}$$

3. APPLICATIONS AND RESULTS

A. Random forest application in WEKA (Waikato Environment for Knowledge Analysis)

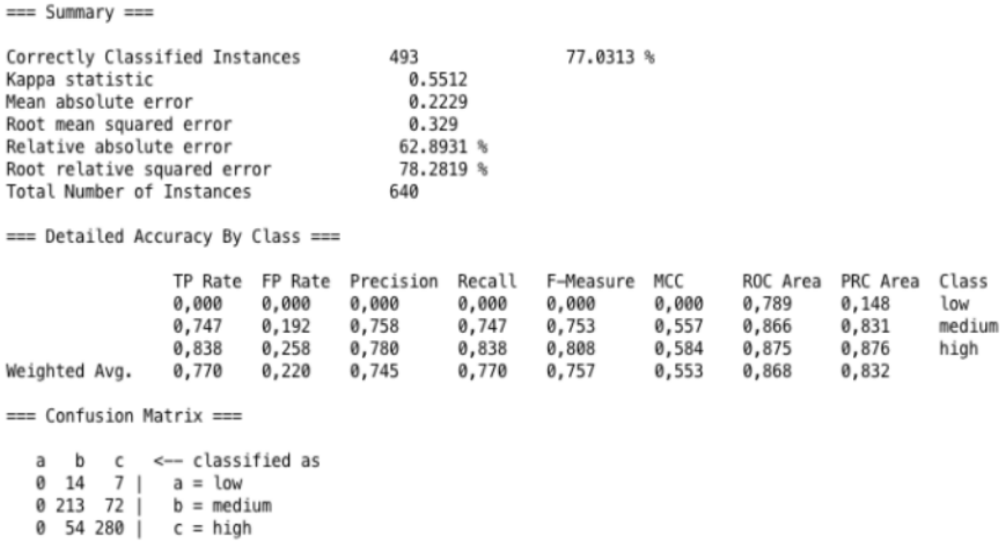
To make the classification more clear in data set, data which have output value 3 and 4 will be classified as low, 5 as medium and 6,7,8 as high.



```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:     wine
Instances:    1599
Attributes:   12
              fixedacidity
              volatileacidity
              citricacid
              residualsugar
              chlorides
              freesulfurdioxide
              totalsulfurdioxide
              density
              pH
              sulphates
              alcohol
              quality
Test mode:    split 60.0% train, remainder test

=== Classifier model (full training set) ===

RandomForest
```

FIGURE 3. Data set information.

60% of data is used for training, and remaining 40% is used for testing the data. Obtained result are presented in Figure 4.



```
=== Summary ===

Correctly Classified Instances         493               77.0313 %
Kappa statistic                          0.5512
Mean absolute error                      0.2229
Root mean squared error                  0.329
Relative absolute error                 62.8931 %
Root relative squared error             78.2819 %
Total Number of Instances              640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    0,000      0,000    0,000      0,000    0,789     0,148     low
                 0,747    0,192    0,758      0,747    0,753      0,557    0,866     0,831     medium
                 0,838    0,258    0,780      0,838    0,808      0,584    0,875     0,876     high
Weighted Avg.    0,770    0,220    0,745      0,770    0,757      0,553    0,868     0,832

=== Confusion Matrix ===

   a   b   c   <-- classified as
   0  14   7 |   a = low
   0 213  72 |   b = medium
   0  54 280 |   c = high
```

FIGURE 4. Weka Random Forest result.

Algorithm's classification success rate in the case for our data is 77.0313%.

B. ANFIS application in MATLAB

For modelling membership functions, Parallel Coordinates Data Visualizer was used. Data set is not changed for this application.

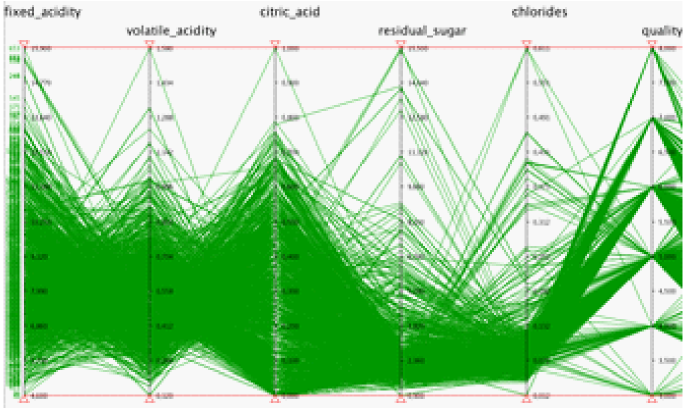
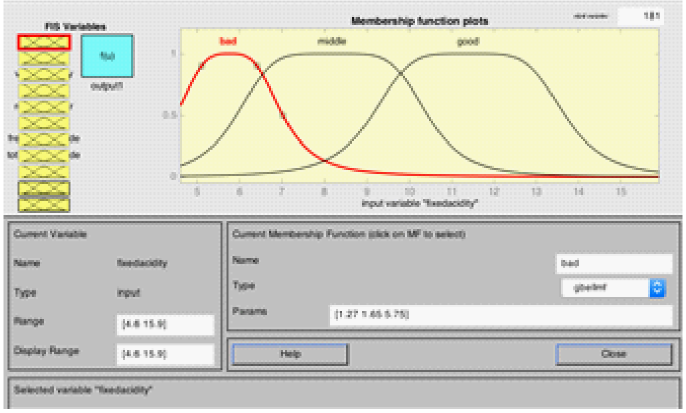


FIGURE 5. Parallel Coordinates data visualizer's figure of data set.

All input variables and output variable is used for modelling the wine quality. Figure 6 shows the fixed acidity's membership functions. For all 11 input variables Bell membership function is used because of its being more successful than others.



Current Variable

Name fixedacidity

Type input

Range [4.6 15.9]

Display Range [4.6 15.9]

Current Membership Function (click on MF to select)

Name bad

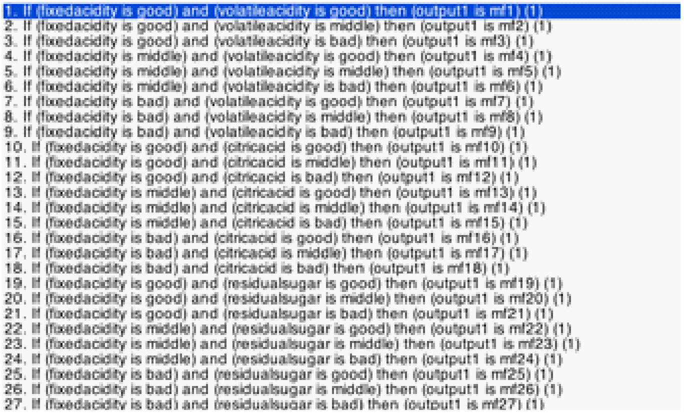
Type gbellmf

Params [1.27 1.65 5.75]

Selected variable "fixedacidity"

FIGURE 6. Fixed acidity membership functions.

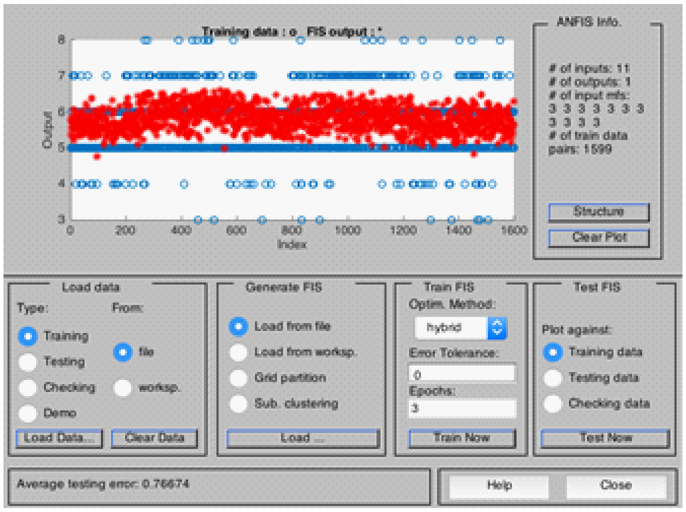
Figure 7 shows the first 27 rules. In total 495 rule was used.



```
1. If (fixedacidity is good) and (volatileacidity is good) then (output1 is mf1) (1)
2. If (fixedacidity is good) and (volatileacidity is middle) then (output1 is mf2) (1)
3. If (fixedacidity is good) and (volatileacidity is bad) then (output1 is mf3) (1)
4. If (fixedacidity is middle) and (volatileacidity is good) then (output1 is mf4) (1)
5. If (fixedacidity is middle) and (volatileacidity is middle) then (output1 is mf5) (1)
6. If (fixedacidity is middle) and (volatileacidity is bad) then (output1 is mf6) (1)
7. If (fixedacidity is bad) and (volatileacidity is good) then (output1 is mf7) (1)
8. If (fixedacidity is bad) and (volatileacidity is middle) then (output1 is mf8) (1)
9. If (fixedacidity is bad) and (volatileacidity is bad) then (output1 is mf9) (1)
10. If (fixedacidity is good) and (citricacid is good) then (output1 is mf10) (1)
11. If (fixedacidity is good) and (citricacid is middle) then (output1 is mf11) (1)
12. If (fixedacidity is good) and (citricacid is bad) then (output1 is mf12) (1)
13. If (fixedacidity is middle) and (citricacid is good) then (output1 is mf13) (1)
14. If (fixedacidity is middle) and (citricacid is middle) then (output1 is mf14) (1)
15. If (fixedacidity is middle) and (citricacid is bad) then (output1 is mf15) (1)
16. If (fixedacidity is bad) and (citricacid is good) then (output1 is mf16) (1)
17. If (fixedacidity is bad) and (citricacid is middle) then (output1 is mf17) (1)
18. If (fixedacidity is bad) and (citricacid is bad) then (output1 is mf18) (1)
19. If (fixedacidity is good) and (residualsugar is good) then (output1 is mf19) (1)
20. If (fixedacidity is good) and (residualsugar is middle) then (output1 is mf20) (1)
21. If (fixedacidity is good) and (residualsugar is bad) then (output1 is mf21) (1)
22. If (fixedacidity is middle) and (residualsugar is good) then (output1 is mf22) (1)
23. If (fixedacidity is middle) and (residualsugar is middle) then (output1 is mf23) (1)
24. If (fixedacidity is middle) and (residualsugar is bad) then (output1 is mf24) (1)
25. If (fixedacidity is bad) and (residualsugar is good) then (output1 is mf25) (1)
26. If (fixedacidity is bad) and (residualsugar is middle) then (output1 is mf26) (1)
27. If (fixedacidity is bad) and (residualsugar is bad) then (output1 is mf27) (1)
```

Figure 7. Fuzzy inference system's first 27 rules.

Before training, model's average testing error is 0.76674. Average testing error calculated as follows: Firstly model's calculation for outputs are subtracted from data set's outputs, then absolute values of these numbers are divided by 1599 (the total numbers of data in the set (Fig.8)).



# of inputs: 11
# of outputs: 1
# of input mfs:
3 3 3 3 3 3 3
3 3 3 3
# of train data
pairs: 1599

Type: ● Training ○ Testing ○ Checking ○ Demo

From: ● file ○ worksp.

● Load from file
○ Load from worksp.
○ Grid partition
○ Sub. clustering

Optim. Method: hybrid

Error Tolerance: 0

Epochs: 3

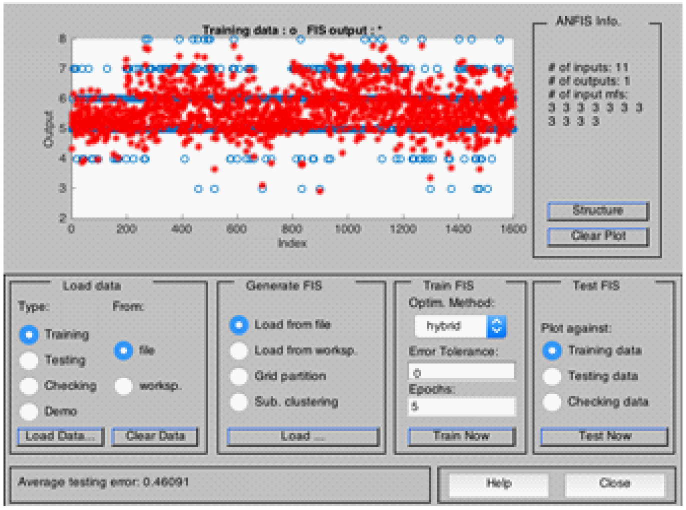
Plot against: ● Training data ○ Testing data ○ Checking data

Average testing error: 0.76674

Figure 8. Before training average testing error.

After training, model's average testing error is 0.46091 (Fig. 9).

To clarify inference system's success, error tolerance should be specified. As Cortez's paper [9] 0.5 and 1 is chosen for error tolerance for this paper too. With 1



ANFIS Info.
# of inputs: 11
# of outputs: 1
# of input mfs:
3 3 3 3 3 3 3
3 3 3 3

Load data
Type: ● Training ○ Testing ○ Checking ○ Demo
From: ● file ○ worksp.

Generate FIS
● Load from file
○ Load from worksp.
○ Grid partition
○ Sub. clustering

Train FIS
Optim. Method: hybrid
Error Tolerance: 0
Epochs: 5

Test FIS
Plot against: ● Training data ○ Testing data ○ Checking data

Average testing error: 0.46091

Figure 9. Fixed acidity membership functions.

error tolerance, 1532 of 1599 data (95.80%) are correctly classified; with 0.5 error tolerance, 1178 data (73.67%) are correctly classified. Here, 1 error tolerance means that if the absolute value of the difference between model's output and data set's output is less than or equal to 1, classification is accepted as true. 0.5 error tolerance means that if the absolute value of the difference between model's output and data set's output is less than or equal to 0.5, the classification is accepted as true.

4. CONCLUSION

This paper suggests two techniques for classification of wine quality. First one is random forest algorithm, with this algorithm 77.0313% classification success rate is achieved. It should not be forgotten that for this application data set is changed. The other one is ANFIS application. ANFIS is a system that have both fuzzy logic and artificial neural networks advantages. This application seems very satisfying when comparing with Cortez's SVM (Support Vector Machine) regression model [9]. With 0.5 and 1 error tolerance SVM regression model accuracy is 62.4% and 89% respectively [9]. In this study, with 0.5 and 1 error tolerance, we obtain the accuracy as 73.67% and 95.80% respectively.

References

[1] Jackson, R. S., Wine Science Principles and Applications, 3th Ed. San Diego, USA 2008.
[2] Delgado, F.M., Cernadas, E., Barro, S., Do we need hundreds of classifiers to solve real world classification problems?, *Journal of Machine Learning Research* 15, (2014) 3133-3181.

[3] Verma, B. Application of fuzzy logic in post harvest quality decisions. Proceedings of the National Seminar on Post harvest Technology of Fruits. Bangalore, India: University of Agricultural Sciences, 1995.

[4] Goel, P., Goel, S., Bhatia, S. Food Quality Assessment Using Fuzzy Logic. *2nd International Conference on Computing for Sustainable Global Development* (INDIACom), 2015.

[5] Perrot, N., Ioannou, I., Allais, I., Curt, C., Hossenlopp, J., Trystram, G., Fuzzy concepts applied to food quality control: A review, *Fuzzy Sets and Systems*, 157, (2010) 1145-1154.

[6] Askerzade, I.N., Mahmood, M., Control the extension time of traffic light in single junction by using fuzzy logic, *International Journal of Electrical & Computer Sciences IJECS–IJENS*, 10-2, (2010) 48-55.

[7] Ziasabounchi, N., Askerzade, I., ANFIS based classification model for heart disease prediction, *International Journal of Electrical & Computer Sciences IJECS–IJENS*,14-2, (2014), 7-12.

[8] Amrahov, Şahin Emrah, Askerzade, I.N., Strong solutions of the fuzzy linear systems, arXiv preprint arXiv:1107.2126, 2011.

[9] Cortez, P., Cerderia, A., Almeida, F., Matos, T., Reis, J., Modelling wine preferences by data mining from physicochemical properties, *Decision Support Systems,* 47, (2009) 547-533.

[10] Negnevitsky, M., A guide to Intelligent Systems, 2nd edition. Edinburg Gate, Harlow.

[11] Jang, J.S.R., Sun, C.T., Mizutani, E., Neuro Fuzzy and Soft Computing, Prentice Hall International, Inc. 1997.

[12] Askerbeyli, İ., Abbuljabar, J.S., Using fuzzy logic methods for carbon dioxide control in carbonated beverages. *International Journal of Electrical & Computer Sciences IJECS-IJENS* Vol: 11-3, (2011), 196-202.

Current address: A.E. OZALP: Hacettepe University, Faculty of Sciences, Dept. of Mathematics, Ankara, TURKEY

E-mail address: emre@gmail.com

ORCID Address: http://orcid.org/0000-0002-9170-3013

Current address: I.N. ASKERZADE: Ankara University, Faculty of Engineering, Dept. of Computer Engineering, Ankara, TURKEY

E-mail address: Iman.Askerbeyli@eng.ankara.edu.tr

ORCID Address: http://orcid.org/0000-0003-4466-8128