I.J. Information Technology and Computer Science, **2022, 1, 1-8**
Published Online February 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2022.01.01

Investigation of the Effect of Normalization Methods on ANFIS Success: Forestfire and Diabets Datasets

Mesut. Polatgil
Şarkışla Faculty of Applied Sciences / computer technologies, Sivas, 58070, Turkey
E-mail: mesutpolatgil@cumhuriyet.edu.tr

Received: 02 August 2021; Revised: 09 October 2021; Accepted: 16 November 2021; Published: 08 February 2022

Abstract: Machine learning and artificial intelligence techniques are more and more in our lives and studies in this field are increasing day by day. Data is vital for these studies. In order to draw meaningful conclusions from the available data, new methods are proposed and successful results are obtained. The preparation of the obtained data is very important in the studies to be carried out. Data preprocessing is very important in the preparation of data. The most critical stage of the data preprocessing process is the scaling or normalization of the data. Machine learning libraries such as scikit-learn and programming languages such as R provide the necessary libraries to scale data. However, it is not known exactly which normalization method will be applied and which will yield more successful results. The success of these normalization methods has been investigated on many different methods, but such a study has not been done on the adaptive neural fuzzy inference system (ANFIS). The aim of this study is to examine the success of normalization methods on ANFIS in terms of both classification and regression problems. So, for studies using the Anfis method, guidance will be provided on which normalization process will give better results in the data preprocessing stage. Four different normalization methods in the scikit-learn library were applied on the Diabets and Forestfire datasets in the UCI database. The results are presented separately for both classification and regression. It has been determined that min-max normalization in classification problems and working with original data in regression problems are more successful.

Index Terms: Data normalization, data scaling, Anfis, classification, regression, scikit-learn.

I. Introduction

With the development of technology, the use of computers and mobile phones has increased, and as a result, there has been a serious increase in the amount of data. The increase in the amount and speed of data has brought about the processing of this data and the extraction of meaningful information. Machine learning and artificial intelligence techniques have gained increasing importance in the processing of this data in recent years.

Fuzzy Logic, proposed by Zadeh in 1965, has been an important step in artificial intelligence techniques. In the classical set concept, an element either completely belongs to a set or it does not. Fuzzy logic softened this sharp distinction and suggested that an element may belong to more than one set. The fuzzy logic approach has been successfully applied in many fields such as health [1], education [2], finance [3]. Another important method such as fuzzy logic is artificial neural networks. This method, which was inspired by the human brain, has provided important developments in the field of artificial intelligence.

Another important method such as fuzzy logic is artificial neural networks. This method, which was inspired by the human brain, has provided important developments in the field of artificial intelligence.

The ANFIS method has emerged with the combination of two important different methods such as fuzzy logic and artificial neural networks. Since there is no learning from data in the fuzzy logic approach, and there is no exact rule about how the number of layers and nodes should be determined in artificial neural networks, ANFIS has been widely used. Successful application results have been obtained in many different fields [4,5,6].

The ANFIS method is a method that works with data. Therefore, preparing the data is a very important step. One of the most important points of this stage, known as data preprocessing, is the data normalization process. The data normalization process ensures that the features in the data set are brought to a certain range, regardless of the mean and distribution range. Thus, different mean and distribution states are eliminated. Thus, features that can take very large values and features that have small values will have similar importance when the algorithm is run. There are many normalization methods in the literature, mainly min-max and z-score methods [7]. There is no study examining the success of normalization methods on ANFIS. Therefore, the question of which normalization method is more suitable

for the studies on Anfis becomes an important problem. With the results obtained from the study, it is expected to show which method can be more efficient in the normalization phase to studies using Anfis.

In addition, due to the wide variety of normalization methods and the use of different normalization methods in the studies, the methods found in the scikit-learn library, which is widely used, were used in this study. These methods are; MinMaxScaler, RobustScaler, StandardScaler, and MaxAbsScaler are methods.

1.1. Related Works

Normalization processes, which constitute an important step of the data preprocessing stage, have been examined in terms of many methods in the literature. The success of the z-score and min-max methods on the KNN algorithm was examined on the IRIS dataset and it was determined that the min-max normalization was more successful [8].

Studies have been conducted to examine the effects of normalization methods on the performance of artificial neural networks (ANN). In these studies, the effect of the normalization process on artificial neural networks was analyzed together with feature selection and different sample sizes. Studies have shown that different normalization procedures and sample sizes lead to differences in study results [9,10,11].

The effect of normalization techniques on classification success was investigated both in brain tumor images and in open-to-use datasets. As a result of the examination, it was determined that the normalization techniques were quite effective on the classification success. ANN, SVM and KNN methods were used in the studies [12,13].

In the classification study on video data, it was stated that normalization methods were effective on the results using ANN method [14].

When the studies are examined, it is seen that the classification successes are generally examined on the ANN method. These examinations revealed the necessity of examining the estimation success of the normalization methods on the ANFIS success in addition to the classification success.

1.2. Anfis

ANFIS is an artificial intelligence method proposed by Jang, (1993). Fig. 1 shows the ANFIS structure [15,16].

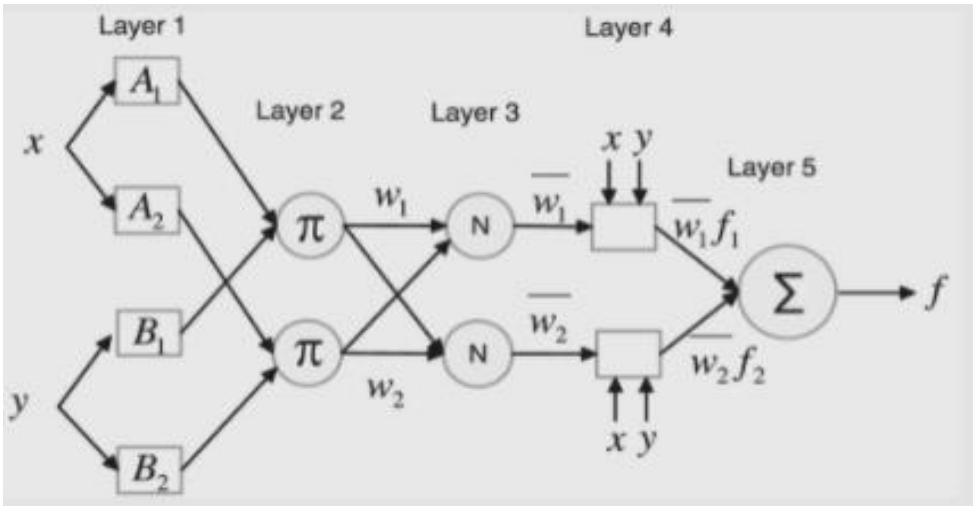

Fig.1. ANFIS structure

ANFIS has a structure consisting of five different layers. In the first layer, each input data is assigned to the clusters determined for the attribute with the degree of membership. If the number of clusters of attributes is high, there will be a serious increase in the number of rules to be produced in the system. For example, if the system has five features and each features has two sets, there will be a total of thirty-two rules. The system generates a prediction by running the rules created as if-then. Automatically created rules are updated according to the values in the data set and the system takes its final form.

2. Method

In the study, the effect of normalization methods on the ANFIS method was examined. Two different datasets were used because it was desired to examine both classification and prediction success. So, for studies using the Anfis method, guidance will be provided on which normalization process will give better results in the data preprocessing stage.

2.1. Datasets

Within the scope of the study, 2 datasets that are open to access in the UCI database were used [17,18]. Datasets

used are shown in Table 1. The reason for choosing two different datasets is that one is suitable for the classification and the other is suitable for the regression problem.

Table 1. Datasets used in the study

Dataset	No. of samples	No. of attributes	Data type
Diabets	768	8	Classification
Forestfire	517	12	regression

2.2. Normalization Methods

Within the scope of the study, 4 different normalization methods were used. The libraries offered by the scikit-learn library were taken into consideration in the selection of these normalization methods.

A. Min-max Normalization

Min-max scaling: In this method given in (1), the data is reduced to the range of 0-1. There is no change in the distribution of the data.

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

x': normalized attribute
x_{min} : Minimum value of the attribute
x_{max} : Maximum value of the attribute

B. Z-Score Normalization

In this method given in (2), the feature values are subtracted from the mean and normalized by dividing by the standard deviation. Although there is no limit to the attribute value range, it is usually seen that the values are in the range of -3 and +3.

$$x' = \frac{x_i - u_i}{\sigma} \tag{2}$$

x': normalized attribute
μ_i: Average of the attribute
σ: Standard deviation of the attribute

C. Robust Normalization

In this method given in (3), it is more useful especially when there is an outlier in the data. Because it uses the median value instead of the mean and the quartile range instead of the value range. The quartile range is known as the IQR (Interquartile Range) and is scaled according to the difference between the 1st and the 3rd Quartile range [19].

$$x' = \frac{x_i - medyan}{p_{75} - p_{25}} \tag{3}$$

x': normalized attribute
$medyan$:The median of the attribute
p_{75}: 3rd quartile of the attribute
p_{25}: 1st quartile of attribute

D. Maxabs Normalization

In this method given in (4), each feature in the data is obtained by dividing its absolute maximum value [20].

$$x' = \frac{x_i}{max|x_i|} \tag{4}$$

3. Findings

3.1. Normality Test Results

The assumption of normal distribution gains importance as the z-score normalization method will be applied to the features in the data sets within the scope of the study. Therefore, the features were tested for normality and the results are shown in Table 2 and Table 3.

Table 2. Normality test results for the diabetes dataset

Shapiro-Wilk	Kolmogorov Smirnov
1,61E-05	0.0
1,99E+05	0.0
1,58E-12	0.0
1,75E-07	0.0
7,92E-19	3,83E-172
1,84E+01	0.0
2,48E-12	1,35E-187
2,40E-09	0.0
1,29E-22	4,57E-163

Table 3. Normality test results for Forestfire dataset

Shapiro-Wilk	Kolmogorov Smirnov
6,71E-14	0
1,88E-17	0
2,20E-25	0
1,95E-15	4,73E-235
7,89E-35	0
5,65E-10	0
1,11E-24	0
9,43E-21	0
0,000126	0
7,80E-13	0
2,49E-09	0
1,26E-44	4,83E-120
1,34E-42	4,83E-120

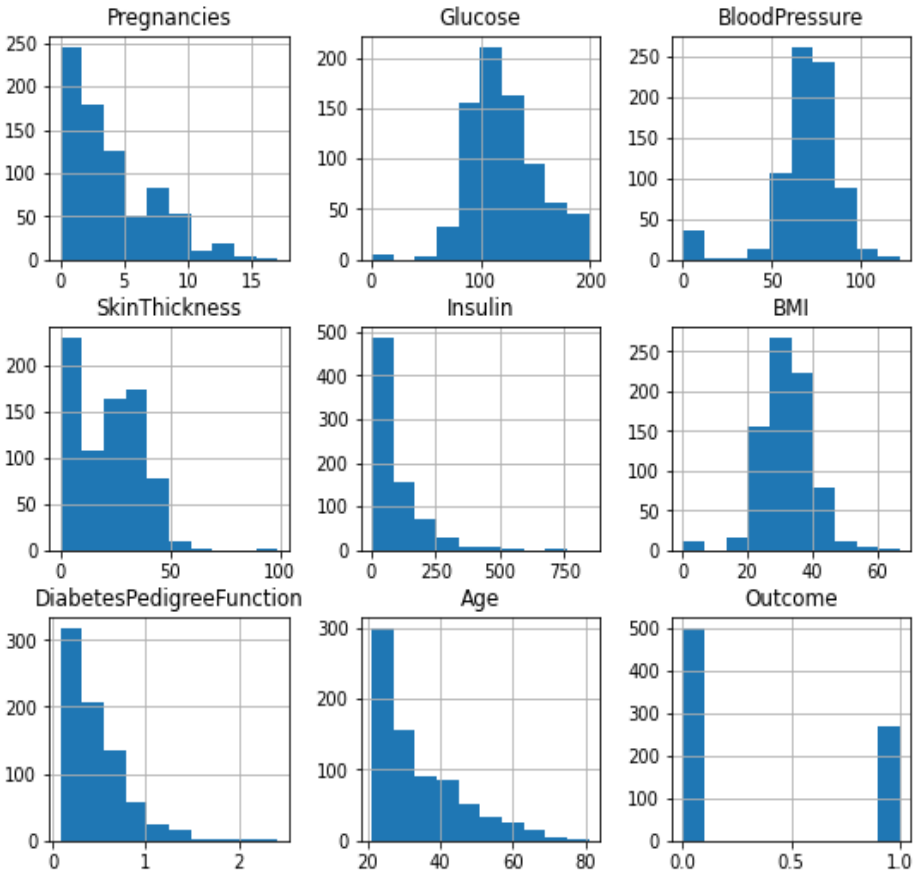

Fig.2. Diabetes dataset features histograms

In addition, the distributions of the features in the datasets used in the study are shown in Fig. 2 and 3.

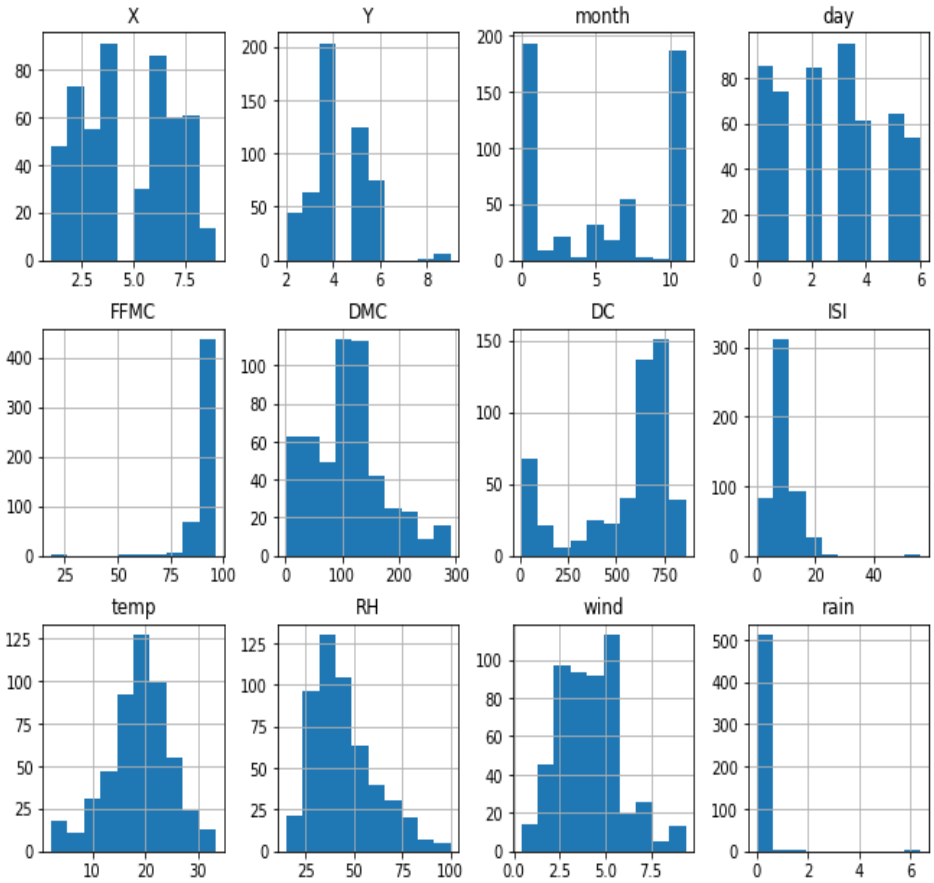

Fig.3. Forestfire dataset attributes histograms

3.2. Diabetes Dataset

The classification criteria obtained as a result of the classification made with ANFIS for the Diabetes dataset are given in Table 4.

Table 4. Classification results for the Diabetes dataset

	Precision	Sensivity	Accuracy	F1 score
orjinal	0.734	0.783	0.821	0.758
minmax	0.754	0.766	0.827	0.7603
standart	0.75	0.75	0.821	0.75
maxabs	0.75	0.75	0.821	0.75
robust	0.733	0.733	0.809	0.733

In the classification results for the Diabetes dataset, the min-max normalization process gave the best result in the F1 criterion. It is noteworthy that other normalization methods give worse results than the results obtained with the original data.

The graphical representation of the classification criteria given in Table 1 is also given in Fig. 4.

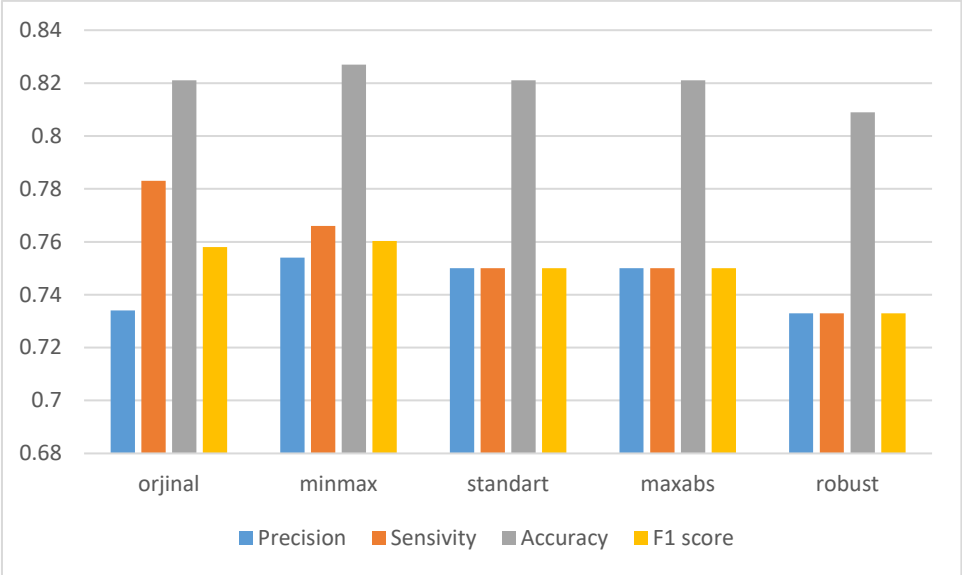

Fig.4. Classification results for the diabetes dataset

3.3. Forestfire Dataset

A regression estimation was performed for the Forestfire dataset. It is aimed to predict the fire zones in the data set. In order to compare the results of different methods over the same values, the fire zones variable was scaled with min-max normalization, and the input variables were subjected to other normalization processes. Thus, all methods could be compared based on estimating the same output values. The estimation results are given in Table 5 for mean square error (mse), root mean square error (rmse) and mean absolute error(mae).

Table 5. Forestfire dataset estimation results

	mse	rmse	mae
orjinal	0.046	0.215	0.092
minmax	0.073	0.274	0.104
standart	0.065	0.256	0.105
maxabs	0.078	0.281	0.104
robust	0.071	0.267	0.110

According to the regression estimation results for the Forestfire dataset, the best results were obtained with the original dataset. The graphical representation of the estimation results is given in Fig. 5.

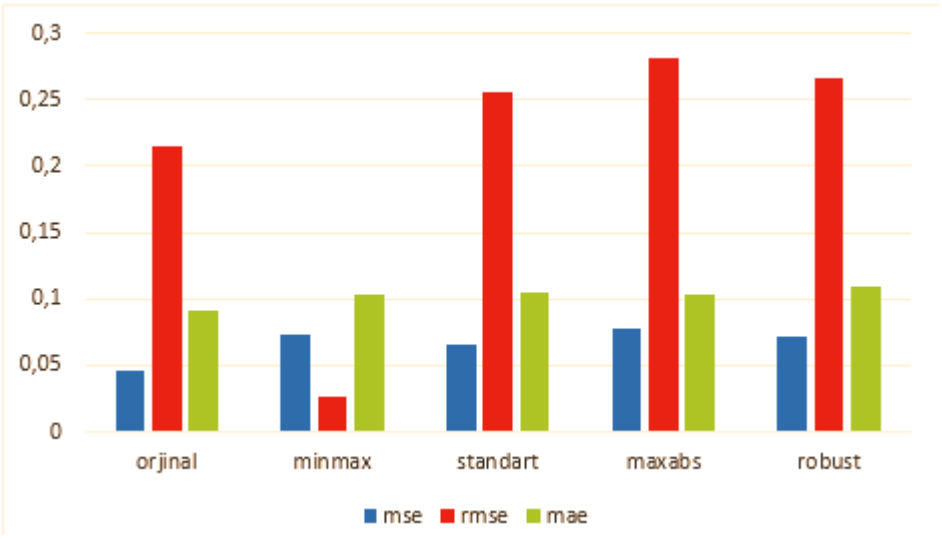

Fig.5. Prediction results for the Forestfire dataset

4. Discussion

There is no study comparing the normalization methods with the ANFIS method, but different algorithms and normalization methods have been used. Therefore, the results were compared with these studies.

These results show that the min-max normalization method is more successful in classification studies. These results are similar to studies conducted on different methods in the literatüre [8,10,11,21]. There are also studies on the fact that different normalization methods are more successful in classification problems [13]. There are also studies in the literature showing that the original data set is more successful in classification studies. Among these, the feature normalization study for the detection of brain tumors showed that working with original data was more successful [12].

In the regression estimation results made for the Forestfire dataset, the use of the original dataset gave the best results. There are studies in the literature that give very close estimation results with the normalized data set of the original data set in regression problems [22]. However, there are also studies in which normalization methods reveal different success results in terms of mse and mae criteria [23]. It is thought that these results are due to the data sets and algorithms used.

5. Conclusion

In the literature, the success of different normalization methods on different algorithms has been examined. This study was carried out to investigate the effect of the success of normalization methods on ANFIS. The results for both classification and regression methods are presented in the study. So, for studies using the Anfis method, guidance will be provided on which normalization process will give better results in the data preprocessing stage.

The study revealed that four different normalization methods had no major effect on ANFIS in classification problems, and that working with the original data set was more effective in regression problems.

One of the most important limitations of the study is that outlier detection was not made in the data set and strong scaling algorithms were not used against outlier data. The effects of such algorithms on the results can be examined in future studies by including these algorithms in the study [24,25].

Studies examining the success of normalization methods have generally focused on classification problems. In this study, both classification and estimation were considered. Therefore, this study is important in that it indicates the necessity of considering both classification and estimation in studies in the field of data preprocessing, such as normalization methods and outlier detection.

References

[1] S. Mohsin, Fuzzy Logic based Health Care System using Wireless Body Area Network Prakashgoud Patil, International Journal of Computer Applications. 80 (2013) 975–8887.

[2] M.A.S. Machado, T.D.R.G. Moreira, L.F.A.M. Gomes, A.M. Caldeira, D.J. Santos, A Fuzzy Logic Application in Virtual Education, Procedia Computer Science. 91 (2016) 19–26. https://doi.org/10.1016/J.PROCS.2016.07.037.

[3] N. Hachicha, B. Jarboui, P. Siarry, A fuzzy logic control using a differential evolution algorithm aimed at modelling the financial market dynamics, Information Sciences. 181 (2011) 79–91. https://doi.org/10.1016/J.INS.2010.09.010.

[4] T. Vasileva-Stojanovska, M. Vasileva, T. Malinovski, V. Trajkovik, An ANFIS model of quality of experience prediction in education, Applied Soft Computing. 34 (2015) 129–138. https://doi.org/10.1016/J.ASOC.2015.04.047.

[5] L. Sarangi, M. Narayan Mohanty, S. Patnaik ITER, Design of ANFIS Based E-Health Care System for Cardio Vascular Disease Detection, (2017). https://doi.org/10.1007/978-3-319-49568-2_63.

[6] S. Barak, J.H. Dahooie, T. Tichý, Wrapper ANFIS-ICA method to do stock market timing and feature selection on the basis of Japanese Candlestick, Expert Systems with Applications. 42 (2015) 9221–9235. https://doi.org/10.1016/J.ESWA.2015.08.010.

[7] T. Jayalakshmi, A. Santhakumaran, Statistical Normalization and Back Propagationfor Classification, International Journal of Computer Theory and Engineering. (2011) 89–93. https://doi.org/10.7763/IJCTE.2011.V3.288.

[8] Amit Pandey, Achin Jain,"Comparative Analysis of KNN Algorithm using Various Normalization Techniques", International Journal of Computer Network and Information Security(IJCNIS), Vol.9, No.11, pp.36-42, 2017.DOI: 10.5815/ijcnis.2017.11.04

[9] P. Cihan, O. Kalipsız, E. Gökçe, Hayvan Hastalığı Teşhisinde Normalizasyon Tekniklerinin Yapay Sinir Ağı Ve Özellik Seçim Performansına Etkisi, Turkish Studies. 12 (2017) 59–70. https://doi.org/10.7827/TurkishStudies.11902.

[10] S. Yavuz, M. Deveci, İstatiksel Normalizasyon Tekniklerinin Yapay Sinir Ağın Performansına Etkisi, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi. 0 (2015) 167–187. https://dergipark.org.tr/en/pub/erciyesiibd/78019 (accessed May 25, 2021).

[11] G. Aksu, C. Oktay Güzeller, M.T. Eser, The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model, International Journal of Assessment Tools in Education. 6 (2019) 170–192. https://doi.org/10.21449/ijate.479404.

[12] B.K. Singh, N.I.T. Raipur, K. Verma, A.S. Thoke, Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification, International Journal of Computer Applications. 116 (2015) 975–8887.

[13] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, Applied Soft Computing. 97 (2020) 105524. https://doi.org/10.1016/J.ASOC.2019.105524.

[14] A. Ali, N. Senan, The Effect of Normalization in Violence Video Classification Performance, IOP Conference Series: Materials

Science and Engineering. 226 (2017) 012082. https://doi.org/10.1088/1757-899X/226/1/012082.
[15] J.S.R. Jang, ANFIS: Adaptive-Network-Based Fuzzy Inference System, IEEE Transactions on Systems, Man and Cybernetics. 23 (1993) 665–685. https://doi.org/10.1109/21.256541.
[16] J.S. Roger Jang, C.T. Sun, Functional Equivalence Between Radial Basis Function Networks and Fuzzy Inference Systems, IEEE Transactions on Neural Networks. 4 (1993) 156–159. https://doi.org/10.1109/72.182710.
[17] I.S. Sitanggang, S. Roseli, L. Syaufina, Information Technology and Computer Science, Information Technology and Computer Science. 9 (2018) 13–20. https://doi.org/10.5815/ijitcs.2018.09.02.
[18] S. Islam Ayon, M. Milon Islam, Information Engineering and Electronic Business, Information Engineering and Electronic Business. 2 (2019) 21–27. https://doi.org/10.5815/ijieeb.2019.02.03.
[19] R. Vaitheeshwari, V. Sathieshkumar, Performance analysis of epileptic seizure detection system using neural network approach, ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings. (2019). https://doi.org/10.1109/ICCIDS.2019.8862158.
[20] C. Chen, N. Kitbutrawat, S. Kajita, H. Yamaguchi, T. Higashino, Modeling BLE propagation above the ceiling for smart HVAC systems, Proceedings - 2019 15th International Conference on Intelligent Environments, IE 2019. (2019) 68–71. https://doi.org/10.1109/IE.2019.00012.
[21] S.V. Khond, Effect of Data Normalization on Accuracy and Error of Fault Classification for an Electrical Distribution System, Https://Doi.Org/10.1080/23080477.2020.1799135. 8 (2020) 117–124. https://doi.org/10.1080/23080477.2020.1799135.
[22] K. Manimekalai, A. Kavitha, MISSING VALUE IMPUTATION AND NORMALIZATION TECHNIQUES IN MYOCARDIAL INFARCTION, ICTACT JOURNAL ON SOFT COMPUTING. (2018) 3. https://doi.org/10.21917/ijsc.2018.0230.
[23] J. Hyma, P. Reddy, A. Damodaram, Performance analysis of Heterogeneous Data Normalization with a New Privacy Metric, Dat a Mining Journal of Comput Er Science IJCSIS Journal of Comput Er Science IJCSIS Sept Ember. (2018). https://sites.google.com/site/ijcsis/ (accessed October 9, 2021).
[24] S. Kappal, Data Normalization Using Median Median Absolute Deviation MMAD based Z-Score for Robust Predictions vs. Min – Max Normalization, London Journal of Research in Science: Natural and Formal. (2019). https://research.journalspress.com/index.php/science/article/view/444 (accessed October 9, 2021).
[25] X.H. Cao, I. Stojkovic, Z. Obradovic, A robust data scaling algorithm to improve classification accuracies in biomedical data, BMC Bioinformatics 2016 17:1. 17 (2016) 1–10. https://doi.org/10.1186/S12859-016-1236-X.

Authors' Profiles

Mesut. Polatgil He is a lecturer in computer science in the Faculty of Applied Sciences. He received his Ph.D. degree in Numerical methods from the University of Sivas Cumhuriyet. He has more than 8 years of teaching and research experience in information systems development and machine learning. His research interests include machine learning, numerical methods and data science.

How to cite this paper: Mesut. Polatgil, "Investigation of the Effect of Normalization Methods on ANFIS Success: Forestfire and Diabets Datasets", International Journal of Information Technology and Computer Science(IJITCS), Vol.14, No.1, pp.1-8, 2022. DOI: 10.5815/ijitcs.2022.01.01