Estimation of Survival According to Body Mass Index (BMI), Hypertension, Diabetes and Heart Disease with Optimizable Decision Trees

N. Hakime Nogay and H. Selcuk Nogay

Abstract— **Non-communicable chronic diseases such as cardiovascular diseases and diabetes and the risk factors of these diseases are becoming an increasing health and development problem in the world. Non-communicable chronic diseases are among the most important causes of death according to the World Health Organization (WHO). The prediction of death or survival is very important in terms of contributing to scientific studies for the earlier diagnosis of non-communicable chronic diseases. Today's developing world, where technology and artificial intelligence can be used in every field, enables the prediction of survival in chronic diseases to be realized with many machine learning methods. In order to know which artificial intelligence or machine learning method is the most effective, it will be very useful to make applications with the methods used and even with the subclasses of the same method and to compare the classification results obtained from the applications with each other. In this study, survival in chronic diseases was estimated by using decision tree methods in four different structures designed by training with body mass index taken from individuals with chronic diseases and other hospital records. The highest accuracy rate was obtained with the optimizable decision trees (ODT) method, which is the simplest model among these models, which allows the most optimal selection of hyperparameters.**

Index Terms—**Survival, Coarse, Fine , Medium, Optimizable Decision Tree, Body Mass Index,**

I. Introduction

Among the machine learning methods, the decision tree method is one of the most preferred methods by many researchers for problems that can be solved by binary classification [1]. Deaths from chronic non-communicable diseases (such as cardiovascular diseases, diabetes, cancer, etc.) are a growing problem of global health in middle- and high-income countries.

NALAN HAKIME NOGAY, is with Department of Nutrition and Diatetic, Erciyes University, Kayseri, Turkey,(e-mail: nalannogay@erciyes.edu.tr).

https://orcid.org/0000-0002-9435-5755

HIDIR SELCUK NOGAY, is with Department of Electrical, Kayseri University, Kayseri, Turkey (e-mail: nogay@kayseri.edu.tr).

https://orcid.org/0000-0001-9105-508X

Manuscript received December 16, 2021; accepted January 30, 2022.
DOI: 10.17694/bajece.1037645

According to WHO, six out of 10 causes of death in 2018 were chronic noncommunicable diseases. Again, according to WHO statistics, cardiovascular diseases alone caused the death of 17.65 million people in 2015 [2]. Studies have shown that an appropriate cardiovascular health profile is associated with a lower risk of developing other chronic diseases [3]. In a study investigating the risks of premature death from noncommunicable chronic diseases, high systolic blood pressure, high body mass index and risks related to dietary intake emerged as the main risk factors among both women and men [4]. Detection of risk factors for deaths from chronic diseases may contribute to the development of new health policies to reduce the burden of these diseases [5]. Considering the scientific studies examined, it is seen that there is a need for more machine learning algorithm-based studies for the prediction of death or survival in chronic diseases. In this study, the survival of individuals with chronic diseases was predicted by decision tree methods, which is one of the most popular machine learning algorithms.

The rest of the study is organized as follows; In the second section, technical information about the decision tree method used in the study is given. In the third part of the study, explanations about the preferred method and data set are given in order to carry out the study. In this section, the stages of the study are presented. In the results section, the results and graphics obtained from the study are given. In the last section, the interpretation of the results is given.

II. Theoretical Framework

Among the machine learning methods used for classification, the decision tree method continues to maintain its popularity thanks to its comprehensible and simple rules, easy interpretability and intelligibility [6]. In order to classify or categorize a data with the decision tree method, it is necessary to train the decision tree model and then subject it to the classification process. During the training of the model, a training data is used to create and train the model. A decision tree is obtained by using model classification rules trained with training data. In the second stage, which is the classification stage, test or validation data is used. Test or validation data tests the accuracy or reliability of the decision tree. If the obtained accuracy rate has reached the desired level, the rules are used to classify the next data row. Training data plays a key role in building the tree [7, 8]. A decision tree model has nodes, branches, and leaves, and each attribute is represented by a node. In decision trees, the top part of the tree is called the root, as opposed to a normal tree view. The branches are between the root and leaves [8, 9]. In order to create a decision tree, conclusions are drawn from the clues obtained from the data reserved for training. With these clues, decision rules are created to classify the data at the root node of the decision tree. At the root node, which is the first node of the tree, questions are started to be asked based on the attributes in the data in order to classify the data and create the tree structure. In this way nodes, branches and leaves are created. Test data from the root of the tree is applied to the tree so that the created decision tree can produce predictions with each data set and test the results. This new dataset, which is used for testing starting from the root, is sent to the child nodes or branches according to the results from the testing process. This process continues like an iteration until it reaches a certain last leaf on the tree. There is a separate path, or a separate decision rule, from the root to each leaf of the tree. Figure 1 shows a decision tree structure consisting of two-dimensional attribute values belonging to two classes. In the figure, the "column_9" attribute values; column_9 < values a, b, c, and d represent the threshold values for branching, and 1 and 0 represent the class labels. Variables in the tree structure can be one or more [10-12].

In decision trees, each variable is divided into at least two sub-variables. A threshold value of “c” can be set to maximize differences or reduce similarities [13-15]. In the decision tree design process, criteria and rules are determined by using some techniques for the formation of branches and nodes. Some of these techniques are; information gain, information gain ratio, Gini index, Towing rule, and Ki-Square probabilistic table statistics. CHAID, QUEST, ID3, C4.5, C5.0, and CART algorithms are some of the algorithms created using these techniques.

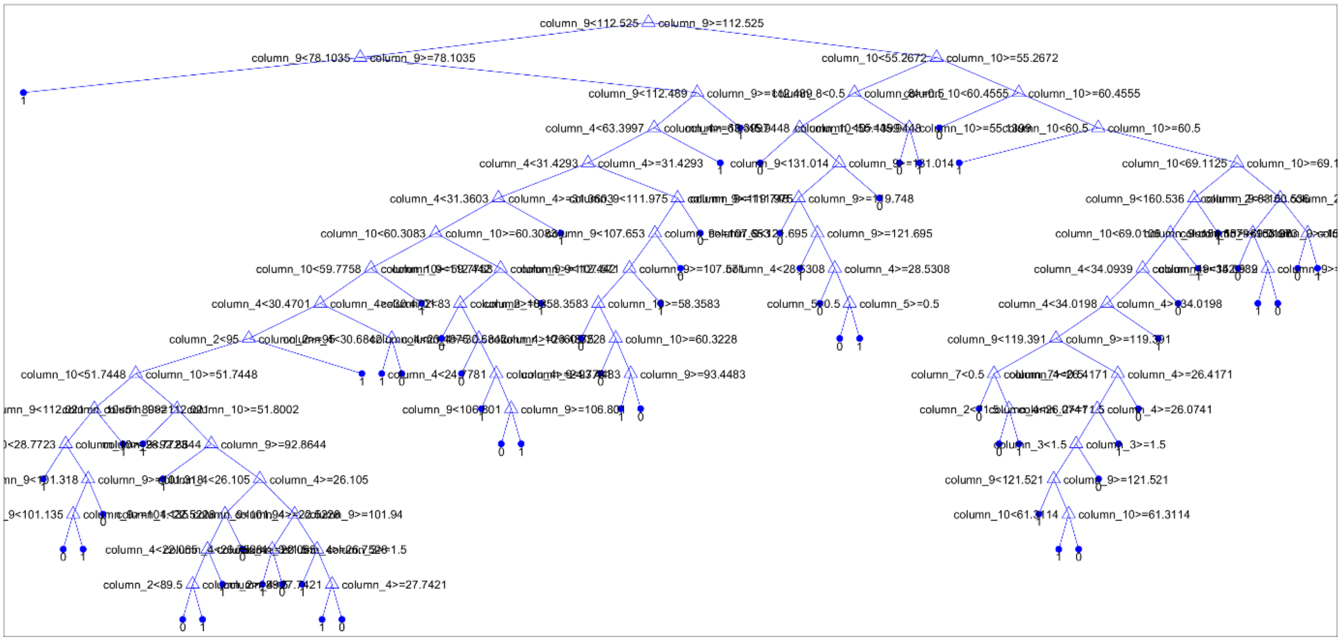


Fig. 1. Fine DT

In the information gain method, entropy rules are used to represent disorder in a system. ID3 and C4.5 algorithms can be given as examples of algorithms using the information gain approach [15]. In this study, the Gini index technique was used while creating the decision tree branches and nodes. In the Gini Index, hypothetically, all variables are continuous. The possibility that each variable can be divided into as many possible categories is considered [15]. The Gini Index is calculated as follows for the T data set with n categories and N samples.

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2 \qquad (1)$$

Where pj is the relative frequency of class j in T. If the T data set is divided into two classes as T1 and T2 and the data numbers of these classes are N1 and N2, the Gini Index is calculated as follows.

$$gini_a(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2) \qquad (2)$$

Where ‘a’ is the number of separations, and the variable with the lowest Gini value is selected.

When a very complex structure emerges in the decision trees, the parts of the decision tree that do not affect the classification accuracy can be removed. This process which is simplifying the decision tree is called pruning [16-20]. In Figure 2, pruned branches are shown for the Fine Decision Tree (FDT) model.

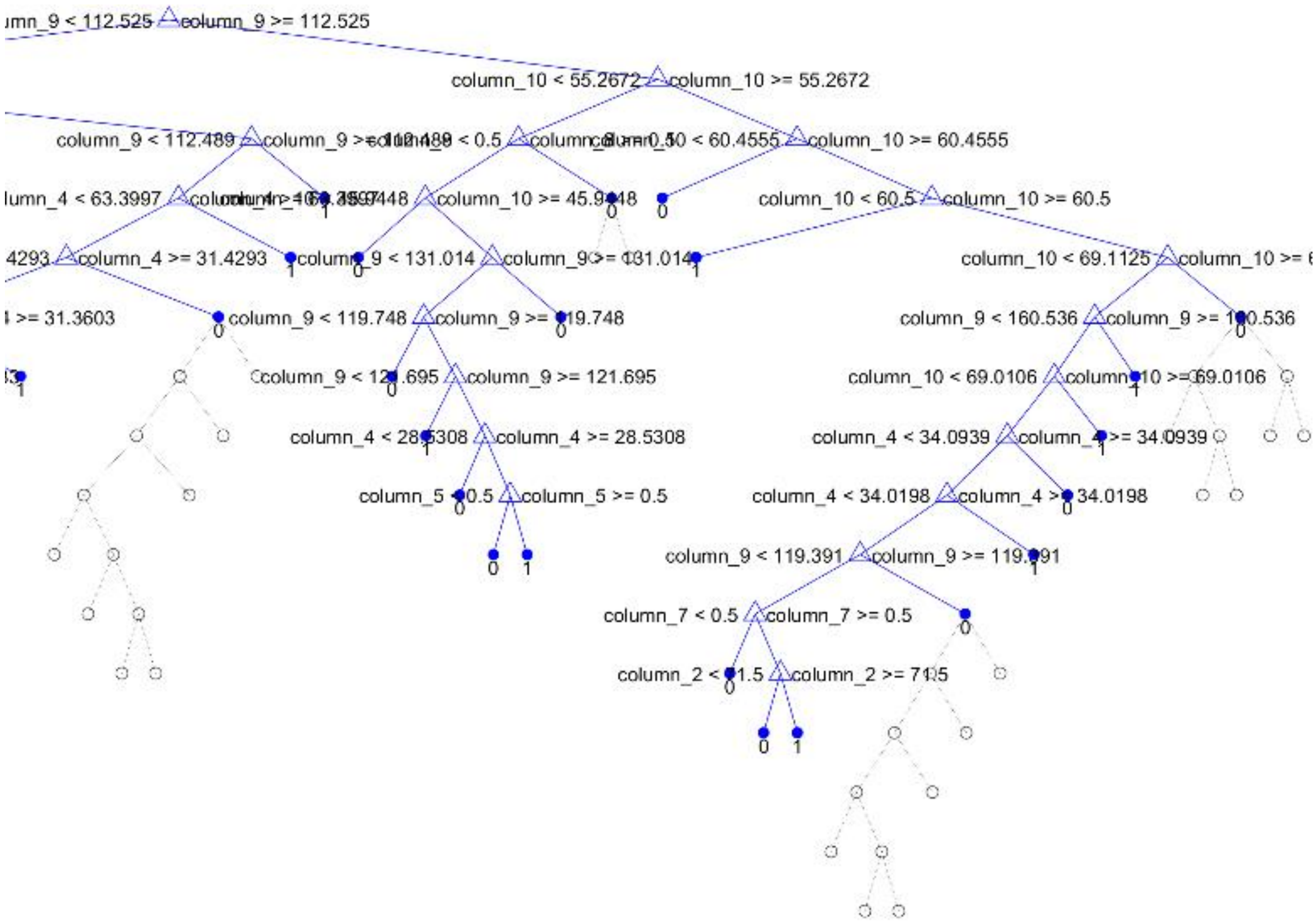


Figure 2. Right part of the fine DT model

III. Methodology

This study was carried out in the "Classification Learner Toolbox" environment under the "Machine Learning and Deep Learning" tab in the MATLAB R2021a package program. Figure 3 shows the Medium Decision Tree (MDT) model used in the study. The internal structure of the Optimizable Decision Tree (ODT) model designed in the study is shown in Figure 4. In the decision tree model, the “Optimizable” option was selected, which automatically selects the most optimal selection of hyperparameters and gives us minimum error curves at the end of this process. Therefore, the decision tree model proposed in the study is the ODT model. The Gini Index technique was preferred in determining the rules. Training and test data are needed to create the decision tree model. The data set used in the study was obtained from the data analysis study performed by Jingmin et al. to make a more accurate prediction of mortality among heart patients admitted to intensive care units [21]. In the created data set, body mass index (BMI), hypertensive, atrial fibrillation, diabetes, hyperlipidemia, systolic blood pressure, diastolic blood pressure, and survival status of 1177 participants were taken to be used for this study. It was recognized that BMI data of some patients were not recorded. So these unsaved rows were removed from the data set, the most recently used data set in the study consisted of 962 rows and 10 columns. The data set used in the study can be accessed from the https://doi.org/10.5061/dryad.0p2ngf1zd web address. Table 1 presents the summary of the data set used in the study. In addition, the definitions of the variables used for the training and test data in the study are shown in Table 1. 'Respond' in the dataset is survival. Survival is represented numerically as 0 and death as 1.

In addition to the ODT method in the study, in order to show the effectiveness of the proposed ODT method, estimation was made with FDT, MDT and CDT in the MATLAB environment. In the decision tree models (Figure 1), the numbers at the end of each branch represent the survival status. In order to increase the reliability of the study and test its accuracy, the 5-fold cross-validation method was used.

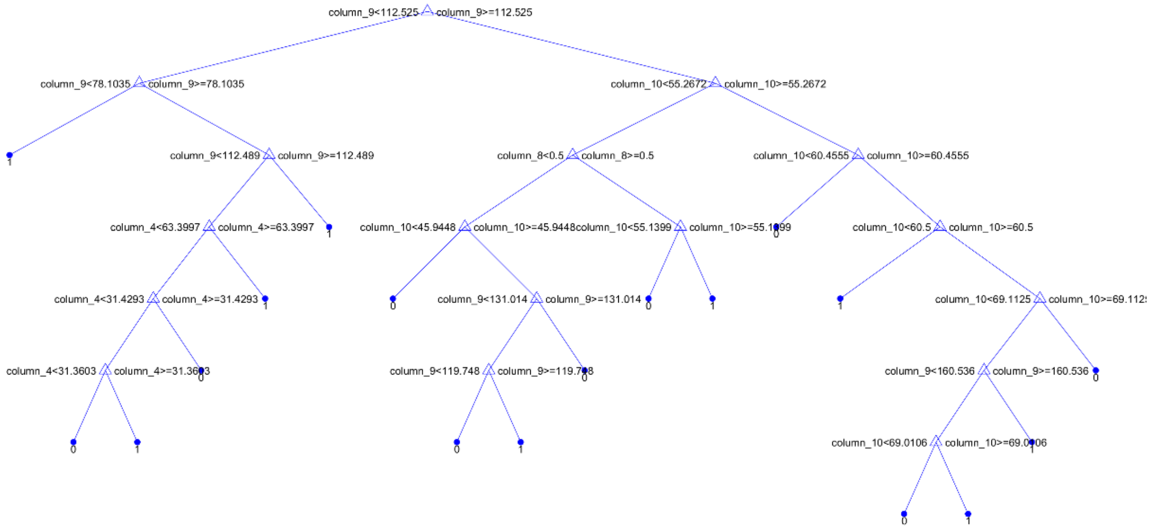


Figure 3. Medium DT model

TABLE I
SUMMARY OF THE DATA SET

Data	Variable Label	Variable Definition	Min	Max
Predictors	Column_ 2	Age	19	99
	Column_ 3	Gender (1:F, 2:M)	1	2
	Column_ 4	BMI	133.46	104.9
	Column_ 5	Hypertensive	0	1
	Column_ 6	Atrial fibrillation	0	1
	Column_ 7	Diabetes	0	1
	Column_ 8	Hyperlipemia	0	1
	Column_ 9	Systolic blood pressure	75	203
	Column_ 10	Diastolic blood pressure	247.36	107
Respond	Column_ 1	Outcome	0	1

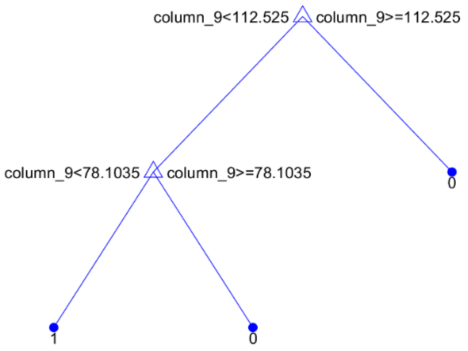


Figure 4. Structure of a decision tree consisting of two-dimensional attribute values

IV. RESULTS

The results obtained from the decision tree classifiers designed and applied in the study are presented in Table 2. According to the results in Table 2, the highest accuracy rate was obtained with the ODT method. Estimates of survival and death among individuals participating in the study can be examined through the confusion matrices in Figure 5 obtained for each of the four models.

TABLE II
THE RESULTS OF APPLICATIONS

	FDT	MDT	CDT	ODT
Accuracy (%)	81.1	86.2	87.8	88.1
Prediction speed (obs/sec)	20000	67000	80000	68000
Training time (sec)	34.154	0.4704	0.3466	32.441
Maximum number of splits	100	20	4	1

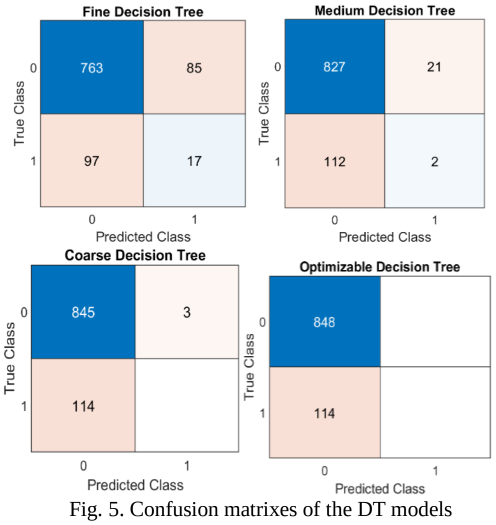


Fig. 5. Confusion matrixes of the DT models

According to the binary classification results obtained from the decision tree models, which method produces more successful results can be understood from the ROC curves shown in Figure 7. Among the four different decision tree methods used, the minimum classification error curve in Figure 8 was obtained in the ODT method that we applied and proposed in our study.

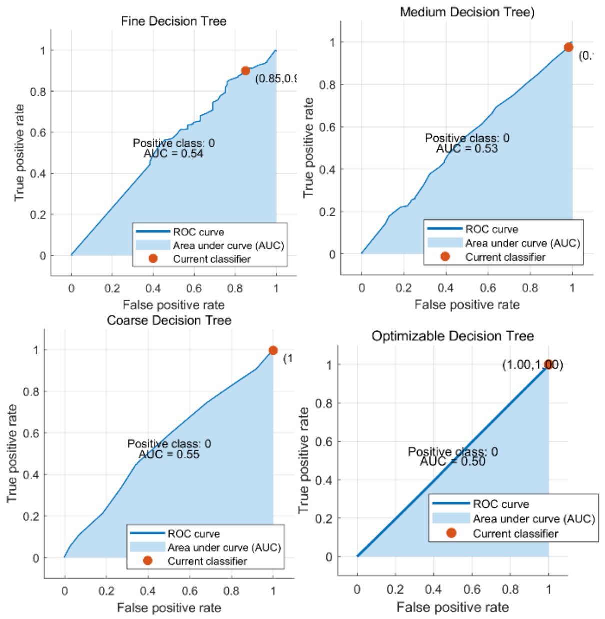


Figure 7. ROC curves of the DT models

The minimum classification error curves in Figure 8 give the difference between the observed actual minimum classification error and the estimated minimum classification error. In addition, the most optimal value point of the hyperparameters obtained as a result of the optimization is marked on this curve.

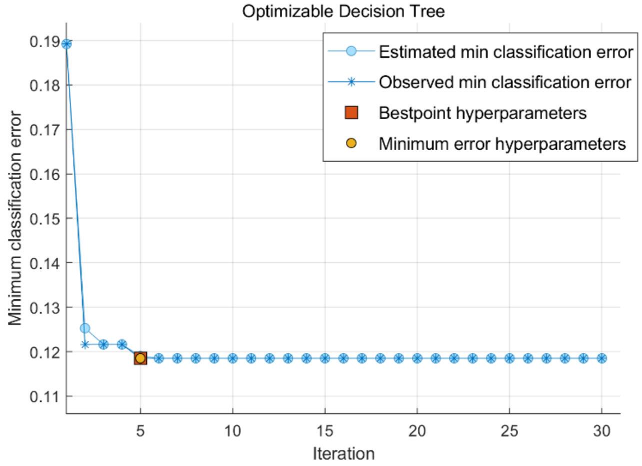


Figure 6. Minimum classification eror curve of the DT model

V. Conclusion and Discussion

When Table 2 is examined, it is clear that the ODT model has the highest accuracy rate. Among these four decision tree techniques, the simplest ones are CDT and MDT. The training times of these simple ones are also very short compared to other models. On the other hand, the most successful model in terms of prediction speed is FDT. The number of splits in decision trees is directly related to the number of nodes used in the tree. The low number of splits also gives us an idea about the complexity of the designed decision tree. In the ODT tree with the lowest number of splits of one, no branches are actually visible. However, we created a symbolic decision tree by creating branches for two nodes and for three nodes. The ODT model has only one split. However, despite the simplest structure, the highest accuracy rate was obtained. The CDT has only 4 split counts and still has the second highest accuracy rate of 87%. When the confusion matrices in Figure 5 are examined, it is seen that FDT produced the lowest survival estimate with 763, and ODT produced the highest survival estimate with 848. Looking at the ROC curves in Figure 8, it is seen that the AUC (Area Under the Curve) values are very close to each other. However, it is understood that the ODT method with 0.5 is more successful. As a result, survival of patients with heart disease can be predicted with high accuracy using decision tree models with 9 variables, including BMI, using some of the hospital data. If one of these decision tree models is ODT, it has been seen that the highest accuracy rate can be achieved with the most optimal decision tree hyperparameters. In the future, a more generalizable model can be designed with more datasets and deep learning methods, taking into account the frequency of food consumption, in which more variables can be added.

References

[1] B. Gupta, A. Rawat, A. Jain, A. Arora, N. Dhami, “Analysis of Various Decision Tree Algorithms for Classification in Data Mining”, International Journal of Computer Applications (0975 – 8887) Volume 163 – No 8, April 2017.

[2] S. J. Szydlowski, M. Luliak. Prevention of Disease-related Mortality from Chronic Non-communicable Diseases. CSWHI 2020; 11.2. 28 – 33; DOI: 10.22359/cswhi_11_2_06

[3] Z. Zhang, S. Jackson, R. Merritt, C. Gillespie, Q. Yang, “Association between cardiovascular health metrics and depression among U.S. adults: National Health and Nutrition Examination Survey”, Ann. Epidemiol. (2007-2014) 2019.

[4] D. C. Malta, B. B. Dunca, M. I. Schmidt, R. Teixeira, A.L.P. Ribeiro, M. S. Felisbino-Mendes et al. “Trends in mortality due to non-communicable diseases in the Brazilian adult population: national and subnational estimates and projections for 2030”. Popul Health Metr. 2020.18(Suppl 1).16.

[5] P. T. Istilli, L. H. Arroyo, R. A. Dias Lima et al. Premature mortality from chronic non-communicable diseases according to social vulnerability. Mundo da Saúde 2021,45: 187-194.

[6] C. F. Chien, L. F. Chen, “Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry,” Expert Systems with Applications, vol. 34, 2008, pp. 280-290

[7] S. Tsang, B. Kao, K. Y. Yip, Wai-Shing Ho, and S. D. Lee, “Decision Trees for Uncertain Data”, IEEE Transactions on Knowledge and Data Engineering, vol. 23. 1, January 2011.

[8] L. Rokach, O. Maimon, “Data Mining with Decision Trees Theory and Applications”, 2nd edition, volume 81, World Scientific Publishing Co. Pte. Ltd. April 2014.

[9] J. R, Quinlan. “C4.5: Programs for Machine Learning”, Morgan Kaufmann, San Mateo, CA, 302, 1993.

[10] G. Dougherty, “Pattern Recognition and Classification, Springer New York Heidelberg Dordrecht London”, first edition, DOI 10.1007/978-1-4614-5323-9

[11] WY. Loh and Yu-Shan Shih, “Split Selection Methods for Classification Trees, Statistica Sinica”, vol. 7. 4 (October 1997), pp. 815-840

[12] M. A. Friedl, C. E. Brodley, “Decision tree classification of land cover from remotely sensed data”, Remote Sensing of Environment, 61, 1997, 399–409

[13] S. R. Safavian, D. Landgrebe, “A survey of decision tree classifier methodology”, IEEE Transactions on Systems Man and Cybernetics, 21, 1991, 660-674

[14] PN. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining” (First Edition) (March 25, 2006) Copyright 2006, Pearson Addison-Wesley.

[15] O. Maimon, L. Rokach, “Data Mining and Knowledge Discovery Handbook” , Springer; 2nd ed. 2010.

[16] N. Suneetha, Ch. V. M. Hari, V. Sunil Kumar, “Modified Gini Index Classification: A Case Study of Heart Disease Dataset”, International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 1959-1965

[17] L. Breiman, J.H. Friedman, R.A. Olshen and C. J. Stone. 1984, “Classification and Regression Trees” Monterey, CA: Wadsworth, 358

[18] L. E. Raileanu, and K. Stoffel, Theoretical Comparison between the Gini Index and Information Gain Criteria, Annals of Mathematics and Artificial Intelligence 41.1, 77-93. May 2004

[19] K. Teknomo, “Decision Tree Tutorial”, Revoledu.com Online edition, Last Update: October 2012

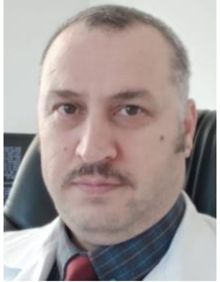
[20] J. Mingers, “An empirical comparison of pruning methods for decision tree induction”, Machine Learning, 4, 1989, 227–243

[21] Z. Jingmin, et al. “Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database”, Dryad, Dataset, 2021, https://doi.org/10.5061/dryad.0p2ngf1zd.

BIOGRAPHIES

NALAN HAKIME NOGAY Istanbul, Turkey, in 1977. He received the B.S., M.S. and Ph.D. degrees in nutrition and dietetics from Hacettepe University, Ankara, Turkey, in 1999, 2002 and 2009 respectively.

From 2009 to 2015, she was an Assistant Professor with the Nutrition ad Dietetic Department. Since 2015, she has been an Associate Professor with the Nutrition and Dietetic Department, Ankara Hacettepe University. She has more than 40 papers. Her research areas are nutrition in the autistic and disabled people, obesity and community nutrition.



HIDIR SELCUK NOGAY Isparta, Turkey, in 1975. He received the B.S. , M.S. and Ph.D. degrees in electrical education from the Kocaeli University, Marmara University and Marmara University respectively in 1999, 2002 and 2008.

From 1999 to 2008, he was a Research Assistant with the Deep Learning and Electrical Machine Laboratory. From 2009 to 2011, he was an Assistant Professor with the Electrical Engineering. From 2011 to 2015, he was an Associate Professor with the Electrical Engineering. Since 2016, he has been a Professor with the Electrical Engineering Department, Kayseri University. His research interests include deep learning applications in medicine, electrical power systems, electrical machinery and renewable energy.