

COMPARISON OF DIFFERENT DECISION TREE MODELS IN CLASSIFICATION OF ANGINA PECTORIS DISEASE

I. Balıkcı Cicek, Z. Kucukakcali, and E. Guldogan

Abstract— Aim: The aim of this study is to classify Angina pectoris disease and compare the estimates of the methods by applying J48 and Random Forest methods, which are among the decision tree models, on the open access angina pectoris data set.

Materials and Methods: In the study, the data set named "Project Angina Data Set" was obtained from <https://www.kaggle.com/snehal1409/predict-angina>. In the data set, there are a total of 200 patients in whom angina pectoris was evaluated. Decision tree models J48 and Random Forest methods were used to classify angina pectoris disease.

Results: From the applied models, from the performance values obtained from the J48 method, the accuracy was 0.868, balanced accuracy 0.868, sensitivity 0.895, specificity 0.842, positive predictive value 0.85, negative predictive value 0.889 and F1-score 0.872. From the performance values obtained from the Random Forest method, the accuracy was 0.921, balanced accuracy 0.921, sensitivity 0.895, selectivity 0.947, positive predictive value 0.944, negative predictive value 0.9 and F1-score 0.919.

Conclusion: Considering the findings obtained from this study, it has been shown that the decision tree models used give successful predictions in the classification of angina pectoris disease.

Keywords— Classification, decision trees, J48, Random Forest, angina pectoris.

1. INTRODUCTION

CARDIOVASCULAR diseases rank first among the causes of death in developed and developing countries [1]. In a study in which death events in our country between 2009-2016 were examined epidemiologically, it was reported that cardiovascular diseases took the first place among the causes of death in all years [2]. Among the cardiovascular diseases, deaths due to ischemic heart diseases take the first place [3]. The most important symptom of ischemic heart disease is angina pectoris (AP). Angina pectoris is described as a clinical symptom characterized by discomfort or pain in the chest, jaw, shoulder, back and arm [4]. In AP pathophysiology, an increase in the oxygen demand of the myocardium at the cellular level

or a decrease in the oxygen level presented in the myocardium is the cause of angina.


Although decreased oxygen delivery is often found responsible as a result of narrowness in the coronary arteries, abnormal increases in oxygen demand such as increased heart rate, uncontrolled hypertension, and increased myocardial contractility can also lead to angina [5]. Although it increases with age, angina pectoris occurs between 0.1-20% in the general population between the ages of 45-74. Angina pectoris is thought to be present in 20,000-40,000 people per million populations, especially in most European countries [5]. Decision trees are one of the most used methods in classification problems. Decision trees are easier to construct, understand and interpret compared to other methods. In addition to these, another advantage of decision trees is that they produce successful models. In order to classify in the decision trees method, a tree is created from the data we have and the records in the dataset are applied to this tree, and the classification process of the records takes place according to the result. In other words, a data that we do not know which class belongs to according to the decision trees obtained from the database, is predicted according to the rule set created when we come to it [6]. J48 is a decision tree algorithm based on the very popular C4.5 algorithm developed by J. Ross Quinlan [7]. J48 Algorithm; Based on Information Gain Theory, it has automatic process capability to select relevant properties from data. It is an iterative algorithm that divides samples from where the information gain is best. The J48 can do an effective pruning to cut off not meaningful, in other words, weak branches [8]. The Random Forest (RF) method was proposed by Breiman in 2001 by developing the Bagging method, which envisions combining the decisions of multiple, multivariate trees each trained with different training sets, instead of generating a single decision tree. This method uses the bootstrap technique in the process of creating different sub-training sets and uses random feature selection during the development of trees [9].

The aim of this study is to compare the classification success of angina pectoris by applying J48 and Random Forest methods, which are among the different decision tree methods, on the angina pectoris dataset.


2. MATERIAL AND METHODS

2.1. Dataset

In the study, the dataset named "Project Angina Data Set" was obtained from <https://www.kaggle.com/snehal1409/predict-angina> in order to examine the working principle of J48 and Random Forest methods [10]. There are 100 (50.0%) no, 100 (50.0%) yes total 200 patients in the dataset used. The variables in the dataset and their descriptive characteristics of the variables are given in Table I.

İpek BALIKCI CICEK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (ipek.balikci@inonu.edu.tr) 

Zeynep KUCUKAKCALI, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (zeynep.tunc@inonu.edu.tr) 

Emek GULDOGAN, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (emek.guldogan@inonu.edu.tr) 

Manuscript received Sep 10, 2020; accepted Oct 18, 2020.
Digital Object Identifier:

TABLE I
VARIABLES IN PROJECT ANGINA DATASET AND DESCRIPTIVE PROPERTIES OF THESE VARIABLES

Variable	Variable Description	Variable Type	Variable Role
Status	Whether the woman has angina (no/yes)	Qualitative	Dependent/Target
Age	Age	Quantitative	Independent/Predictor
Smoke	Smoking status (current, ex-, non-smoker)	Qualitative	Independent/Predictor
Cig	Average number of cigarettes smoked per day	Quantitative	Independent/Predictor
Hyper	Hypertensive condition (absent, mild, moderate)	Qualitative	Independent/Predictor
Angfam	Family history of angina (no/yes)	Qualitative	Independent/Predictor
Myofam	Family history of myocardial infarction (no/yes)	Qualitative	Independent/Predictor
Strokefam	A family history of stroke (no/yes)	Qualitative	Independent/Predictor
Diabetes	Whether the woman has diabetes (no/yes)	Qualitative	Independent/Predictor

3. DECISION TREES

Decision trees, one of the prediction methods, is one of the popular and powerful methods of information discovery and data mining. Decision trees are a hierarchical and ordered way of displaying the rules in the data. Decision trees are a visual modeling method that shows the mass of information about the problem faced by the decision maker in a more understandable way, and presents the decision options and probabilistic situations in a certain order by sorting. In this context, it can be said that decision trees represent a hierarchical model that includes decisions and results. Thanks to its easy-to-understand graphical structure and rules, it is widely used in many areas [11]. Decision trees model, which is among the classification models in data mining, is a model with predictive value. Decision trees ask questions starting from the first stage to the final decision options and form their structure with the answers they receive to these questions, and rules (if-then rule) can be written with this tree structure [12].

3.1. J48

J48 developed by *Quinalan* is a C4.5 decision tree developed for the classification process of nonlinear and small size data. J48 is a decision tree that uses entropy concept knowledge to classify. It applies Quinlan's C4.5 algorithm to generate a pruned C4.5 tree. Decision making is done by dividing each attribute dataset into subsets to examine entropy differences. The highest normalized information gains, the attributes are selected [13].

3.2. Random Forest

In this algorithm, developed by *Breiman* in 2001, the purpose for the classifier is to combine the decisions of multiple trees, each trained in different training sets, rather than generating a

single decision tree. Random feature selection with the same distribution is used for different training sets. While creating decision trees, when determining the attribute at each level, firstly, some calculations are made in all trees and the attribute is determined, then the attributes in other trees are combined and the most used attribute is selected. After the selected attribute is included in the tree, the same processes are repeated at other levels. To start the algorithm, the number of variables used in each node and the number of trees to be developed must be determined by the user. Random Forest uses the CART (Classification and Regression Tree) algorithm to generate a tree. Nodes and branches are created in accordance with the features of this algorithm [14].

3.3. Performance Evaluation of Models

Performance criteria obtained by using the classification matrix given below were used in the performance evaluation of J48 and Random Forest methods.

TABLE II
CLASSIFICATION MATRIX FOR CALCULATING PERFORMANCE CRITERIA

		Real		
		Positive	Negative	Total
Predicted	Positive	True positive (TP)	False negative (FN)	TP+FN
	Negative	False positive (FP)	True negative (TN)	FP+TN
	Total	TP+FP	FN+TN	TP+TN+FP+FN

The performance criteria to be used in the performance evaluation of the models in this study are given below.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

$$\text{Balanced accuracy} = [(TP/(TP+FP)) + (TN/(TN+FN))]/2$$

$$\text{Sensitivity} = TP/(TP+FP)$$

$$\text{Specificity} = TN/(TN+FN)$$

$$\text{Positive predictive value} = TP/(TP+FN)$$

$$\text{Negative predictive value} = TN/(TN+FP)$$

$$\text{F1-score} = (2*TP)/(2*TP+FP+FN)$$

4. DATA ANALYSIS

Quantitative data are expressed as mean \pm standard deviation, median (minimum-maximum), and qualitative data as number (percentage). Conformity to normal distribution was evaluated by the Kolmogorov-Smirnov test. In terms of independent variables, whether there is a statistically significant difference between the "no" and "yes" groups, which are the categories of the dependent / target variable (status), and whether there is a relationship, Mann-Whitney U test, Pearson chi-square test, Continuity Correction test and Fisher's Exact test. It was examined using the chi-square test values of $p < 0.05$ were considered statistically significant. IBM SPSS Statistics 26.0 package program was used for all analyzes.

For the validity of the model, a 10-fold cross-validation method was used. In the 10-fold cross-validation method, all data is divided into 10 equal parts. One part is used as a test set and the

remaining 9 parts are used as a training dataset and this process is repeated 10 times.

5. RESULTS

Descriptive statistics of independent variables examined in this study are given in Table 3. According to the findings in Table 3; there is a statistically significant difference between the dependent / target variable groups in terms of age and cig variables ($p < 0.05$).

According to the findings in Table 4; there is a statistically significant relationship between the smoke, hyper and myofam variables and the dependent / target variable (status) groups ($p < 0.05$).

TABLE III

DESCRIPTIVE STATISTICS FOR QUANTITATIVE INDEPENDENT VARIABLES

Variables	Status		p-value *
	no	yes	
	Median(min-max)	Median(min-max)	
age	49 (29-74)	57 (29-73)	<0.001
cig	0 (0-30)	12 (0-40)	<0.001

*: Mann Whitney U test

TABLE IV

DESCRIPTIVE STATISTICS FOR QUALITATIVE INDEPENDENT VARIABLES

Variables		Status		p-value
		no	yes	
		Number (%)	Number (%)	
smoke	current	22 (22)	61 (61)	<0.001*
	ex	14 (14)	26 (26)	
	non-smoker	64 (64)	13 (13)	
hyper	absent	83 (83)	67 (67)	0.022*
	mild	14 (14)	23 (23)	
	moderate	3 (3)	10 (10)	
angfam	no	94 (94)	85 (85)	0.065**
	yes	6 (6)	15 (15)	
myofam	no	88 (88)	47 (47)	<0.001**
	yes	12 (12)	53 (53)	
strokefam	no	94 (94)	94 (94)	1**
	yes	6 (6)	6 (6)	
diabetes	no	97 (97.97)	94 (94.9)	0.445***
	yes	2 (2.02)	5 (5.1)	

*: Pearson chi-square test, **: Continuity Correction test, ***: Fisher's Exact test

Classification matrix of J48 and Random Forest models are given in Table 5 and Table 6, respectively.

TABLE V

CLASSIFICATION MATRIX OF THE J48 MODEL

Prediction	Reference		
	no	yes	Total
no	17	3	20
yes	2	16	18
Total	19	19	38

TABLE VI

CLASSIFICATION MATRIX OF THE RANDOM FOREST MODEL

Prediction	Reference		
	no	yes	Total
no	17	1	18
yes	2	18	20
Total	19	19	38

Table 7, the values of performance criteria calculated from models created to classify Angina pectoris disease in the test stage are given below.

TABLE VII

PERFORMANCE CRITERIA VALUES CALCULATED FROM CREATED MODELS IN THE TESTING PHASE

Performance Metrics	Model	J48	Random Forest
		Value	Value
Accuracy (%)		86.8	92.1
Balanced accuracy (%)		86.8	92.1
Specificity (%)		84.2	94.7
Sensitivity (%)		89.5	89.5
Positive predictive value (%)		85.0	94.4
Negative predictive value (%)		88.9	90.0
F1-score (%)		87.2	91.9

In Figure 1, the values of the performance criteria obtained from J48 and Random Forest methods are plotted.

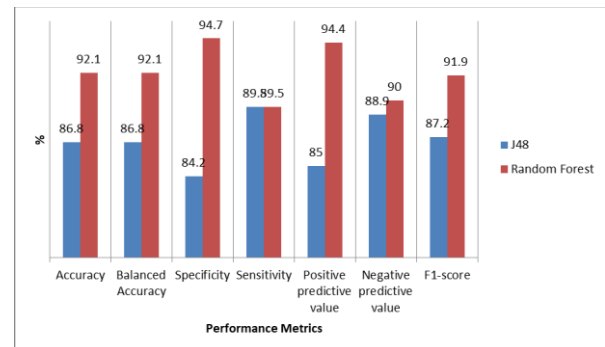


Fig.1. Performance metric values obtained from J48 and Random Forest methods in the testing stage

6. DISCUSSION

Angina pectoris (AP) was first described by Heberden in 1772, and the most common cause is coronary atherosclerosis [15]. Discomfort in the chest caused by ischemic atherosclerotic coronary artery disease associated with impaired coronary blood flow is called angina pectoris. It mostly spreads to the neck, left shoulder, and medial side of the left arm and lasts no longer than 10-15 minutes. Angina pectoris responds immediately to short-acting nitrates [16, 17]. The prevalence of angina in both sexes increases rapidly with age. From 0.1-1% in women aged 45-54, to 10-15% in women aged 65-74; it increases in 2-5% for men aged 45-54 and to 10-20% for men aged 65-74. Accordingly, it can be calculated that in most European countries, 20 000-40 000 people per million of the general population have angina [18]. Decision trees are a

method that is frequently used in classification because it is easier to structure and understand when compared to other classification methods [19]. Decision trees are the most widely used methods among classification models because of their easy interpretation, easy integration with database systems, and good reliability. These methods have predictive and descriptive properties [20]. Decision trees decide which class the new data belongs to, based on the old data, by subtracting rules. The decision tree acts in line with the questions asked and the answers received, and creates rules by combining the answers to the questions. We can say that the tree formed is a set of rules consisting of many "if-then" [21]. When creating decision trees, it is important what algorithm is used. Because the shape of the tree created according to the algorithm used can change. Different tree structures give different classification results. The fact that the first node forming the root node is different will change the way to be followed when reaching the farthest leaf, thus the classification [22]. There are many different decision tree algorithms. Of these, J48 is information-based and has automatic process capability to select relevant features from data. In addition, it is the algorithm with the highest classification success according to algorithms such as Naive Bayes, ID3, Logistic Regression [23]. Random Forest is an algorithm that aims to increase the classification value by generating more than one decision tree during the classification process [24].

According to the findings in this study; the performance criteria obtained from the J48 method, accuracy was 0.868, balanced accuracy 0.868, sensitivity 0.895, specificity 0.842, positive predictive value 0.85, negative predictive value 0.889, and F1-score 0.872. The performance criteria obtained from another method, the Random Forest method, the accuracy was 0.921, balanced accuracy 0.921, sensitivity 0.895, specificity 0.947, positive predictive value 0.944, negative predictive value 0.9 and F1-score 0.919. When these classification performances were compared, Random Forest method gave more successful estimation results compared to J48 method.

As a result, the decision trees methods used have produced quite successful results in the study with the angina pectoris dataset.

REFERENCES

- [1] A. K. Şimşek and Ş. E. Alpar, "Akut Koroner Sendromlu Hastalarda Sağlıklı Yaşam Davranışlarının Kazandırılması," *Türk J Cardiovasc Nurs*, vol.11, pp. 31-36, 2020.
- [2] K.Y. Emlk and A. E. Önal, "2009-2016 Yıllarında Türkiye'deki Ölümlerin Epidemiyolojik Yönünden İncelenmesi ve Ölüm Bildirim Sisteminin Önemi," *İstanbul Tıp Fak. Dergisi*, vol.82, pp.25-26, 2019.
- [3] B. Bayrak, S. Oğuz, S. Arslan, B. Candar, S. Keleş, B. Karagöz, et al., "Miyokard İnfarktüsü Geçirmiş Hastalarda Algılanan Stresin Belirlenmesi," *Türk J Cardiovasc Nurs*, vol.10, pp.129-137, 2019.
- [4] G. A. Diamond, "A clinically relevant classification of chest discomfort," *Journal of the American College of Cardiology*, vol. 1, pp.574-575, 1983.
- [5] K. Soyulu, "Kararlı angina pektoris," *Deneysel ve Klinik Tıp Dergisi*, vol. 29, pp. 117-121, 2012.
- [6] G. Silahtaroğlu, *Veri madenciliği: Kavram ve algoritmaları*: Papatya, 2013.
- [7] H. Nizam and S. S. Akın, "Sosyal medyada makine öğrenmesi ile duyu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması," *XIX. Türkiye'de İnternet Konferansı*, 2014.
- [8] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *Proceedings of International*

- Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, 2013..
- [9] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.
- [10] Available: <https://www.kaggle.com/snehal1409/predict-angina>
- [11] S. K. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data mining and knowledge discovery*, vol.2, pp.345-389, 1998.
- [12] R. Lior, *Data mining with decision trees: theory and applications*, World scientific, vol. 81, 2014.
- [13] G.I. Salama, M. Abdelhalim, and M.A.-e. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis," in *2012 Seventh International Conference on Computer Engineering & Systems (ICCES)*, 2012, pp.180-185.
- [14] B. Daş and İ. Türkoğlu, "DNA dizilimlerinin sınıflandırılmasında karar ağacı algoritmalarının karşılaştırılması," *Elektrik-Elektronik-Bilgisayar ve Biyomedikal Müh. Semp. (ELECO 2014)*, pp.381-383, 2014.
- [15] C. Sliub, "Angina pectoris and coronary heart disease. RO Brandenburg (ed.) *Cardiology: Fundamentals and Practice*. New York," Year Book Medical Publ, 1987.
- [16] C. Man, Z. Dai, and Y. Fan, "Dazhu hongjingtian preparation as adjuvant therapy for unstable angina pectoris: A meta-analysis of randomized controlled trials," *Frontiers in Pharmacology*, vol.11, p.213, 2020.
- [17] T. F. Imran, R. Malapero, A. H. Qavi, Z. Hasan, B. de la Torre, Y.R. Patel, et al., "Efficacy of spinal cord stimulation as an adjunct therapy for chronic refractory angina pectoris," *International journal of cardiology*, vol.227, pp.535-542, 2017.
- [18] D.K. Hancı, "Perkütan Koroner Girişim Yapılan Hastalarda Stent Balonu Dilatasyonu Sırasında Angina Gelişme Sıklığı, Yeri, Karakteri ve Angina Lokalizasyonunun Koroner Anatomi İle İlişkisi," *Kardiyoloji Uzmanlık Tezi*, Kocaeli Üniversitesi, Tıp Fakültesi, 2020.
- [19] E. Atılğan, "Karayollarında meydana gelen trafik kazalarının karar ağaçları ve birliktelik analizi ile incelenmesi," *Hacettepe Üniversitesi İstatistik Anabilim Dalı, Yüksek Lisans Tezi*, 2011.
- [20] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*: Elsevier, 2011.
- [21] M. Akman, "Veri madenciliğine genel bakış ve random forests yönteminin incelenmesi: sağlık alanında bir uygulama," master's thesis, Ankara University, Ankara, 2010.
- [22] G. Silahtaroğlu, "Veri madenciliği," *Papatya Yayınları*, İstanbul, 2008.
- [23] A. Altınkardeş, H. Erdal, F. Baba, and A.S. Fak, "ABPM Ölçümü Olmaksızın Karar Ağaçları Algoritması ile Non-Dipper/Dipper Öngörüsü," *IX. Ulusal Tıp Bilişimi Kongresi*, Antalya, Türkiye, 15-17 Kasım 2012.
- [24] Ö. Akar and O. Güngör, "Classification of multispectral images using Random Forest algorithm," *Journal of Geodesy and Geoinformation*, vol. 1, pp. 105-112, 2012.

BIOGRAPHIES

İpek BALIKÇI ÇİÇEK obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

Zeynep KUCUKAKCALI obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

Emek GÜLDOĞAN obtained his BSc. degree in Computer Engineering from Middle East Technical University in 2001. He received MSc. degree in biostatistics and medical informatics from the Inonu University in 2005, and Ph.D. degrees in biostatistics and medical informatics from the Inonu University in 2017. He is currently working as an assistant professor of the Department of Biostatistics and Medical Informatics at Inonu University and as the information processing manager at Turgut Özal Medical Center. His research interests are cognitive systems, data mining, machine learning, deep learning.