SAMPLE SIZE EFFECT ON CLASSIFICATION PERFORMANCE OF MACHINE LEARNING MODELS: AN APPLICATION OF CORONARY ARTERY DISEASE

M. Kivrak, F.B. Akcesme, and C. Çolak

Abstract—Cardiovascular diseases are among the most common causes of death due to their widespread prevalence. Accurate and timely diagnosis of coronary artery disease, one of the fatal cardiovascular diseases, is very important. Angiography, an invasive method, is an expensive and special method used to determine the disease and can cause serious complications. Therefore, cheaper and more efficient data mining methods are used in the diagnosis and treatment of cardiovascular diseases. As an alternative approach, by establishing clinical decision support systems using data modeling and analysis methods such as data mining, errors and costs can be reduced by providing clinicians with computer-aided diagnosis, and patient safety and clinical decision quality can be significantly increased. In this study, the data set on the open-source access website was used to classify cardiovascular disease and consists of patient records of 14 variables created by the Cleveland clinic. Also, machine learning methods (C5.0 Decision Tree, Support Vector Machine, Multilayer Perceptron, and Ensemble Learning)were used to determine the risk of coronary artery disease by deriving 1000 and 10000 data sets from the cardiology data set obtained from original 303 patient records. Performance evaluation of models is compared in terms of accuracy, specificity, and sensitivity. In trying to determine the most successful model in estimating the risk of coronary artery disease, the results are presented comparatively.

Keywords—*Cardiovascular Diseases, Sample Size, Data Mining, Ensemble Learning*.

1. INTRODUCTION

THE Cardiovascular diseases (CVD) are caused by pathologies in the heart and blood vessels, and coronary artery disease (CAD), heart failure, cardiac arrest, ventricular arrhythmias, sudden heart death, ischemic stroke, transient ischemic attack, subarachnoid and intracerebral hemorrhage, abdominal aortic aneurysm, can result in diseases and congenital heart diseases [1].

CVD can cause myocardial infarction, heart failure, and sudden heart death. Nuclear screening, echocardiography, electrocardiogram (ECG), non-invasive (non-invasive) procedures such as exercise stress test, and invasive (interventional) procedures such as angiography are required for the diagnosis of coronary artery disease [2]. For this reason, the angiography diagnostic method, which is an invasive method, is used as a determinant in the definitive diagnosis of coronary artery diseases and in determining the severity of the disease. However, angiography procedure is a diagnostic method that requires a high cost and advanced technical expertise [3]. As an alternative approach, by establishing clinical decision support systems using data modeling and analysis methods such as data mining, errors and costs can be reduced by providing clinicians with computer-aided diagnosis, and patient safety and clinical decision quality can be significantly increased [4].

This study aims to classify cardiovascular disease and consisted of patient records of 14 variables created by the open-source dataset of the Cleveland Clinic. Besides, machine learning methods (C5.0 Decision Tree, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Ensemble Learning) were used to determine the risk of coronary artery disease by deriving 1000 and 10000 data sets from the cardiology data set obtained from original 303 patient records. Performance evaluation of models is compared in terms of accuracy, specificity, and sensitivity. In order to determine the most successful model in estimating the risk of coronary artery disease, the results are presented comparatively on the open-sourced heart dataset.

2. MATERIAL AND METHOD

2.1. Data Set

The dataset used for the analysis was obtained from http://archive.ics.uci.edu/ml/datasets/statlog+(heart) [5]. The data set contains the original 303 heart disease data and 14 variables. In the original 303 heart disease dataset, 1000 and 10000 datasets were derived from the dataset that showed similar distributions from the dataset due to the binomial distribution of the target variable (glass) and the normal, binomial and uniform distribution of the explanatory variables. These variables are class, age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, painloc, oldpeak, the slope of the peak exercise ST segment, number of major vessels (colored vessels) and thal. The detailed explanations of the variables are given in Table I.

2.2. Knowledge Discovery in Databases (KDD)

In the process of KDD; data selection (heart dataset), data preprocessing (extreme and missing value analyses), data

Mehmet Kıvrak, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: mehmetkivrak83@gmail.com)

F. Berat Akçeşme, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, University of Health Sciences, Istanbul, Turkey, (e-mail: farukberat.akcesme@sbu.edu.tr)

Cemil Çolak, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: cemilcolak@yahoo.com)

Manuscript received May 03, 2020; accepted May 30, 2020.
Digital Object Identifier:

 ISSN: 2548-0650

transformation(normalization, etc.), data mining and evaluation, and interpretation of the results were performed.

2.3. Classification Method

The most commonly used data mining methods on the analyzed datasets have been applied for the classification of CVD. Performance data obtained by using C5.0 Decision Tree, SVM, MLP, and Ensemble Learning classification methods were comparatively presented to the data sets (303, 1000, and 10000 sample sizes).

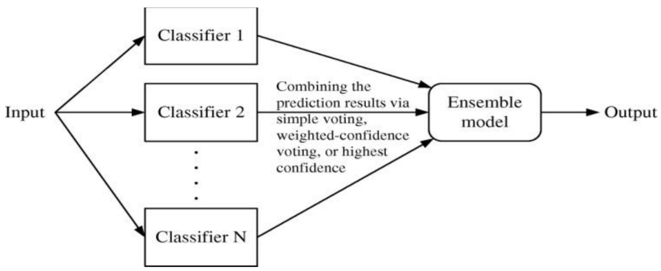

Fig.1. Classification Method and Ensemble Learning Algorithm.

2.3.1.C5.0 Decision Tree

The C5.0 Decision Tree is one of the methods for supervised learning in the form of a tree structure used for classification as well as regression in general. The aim is to build the tree structure that predicts the label of a target variable using the model created.[6]. The C5.0 algorithm uses the concept of knowledge gain and entropy to optimally separate nodes. When there are k probabilities for X variable (attribute) $P_1, P_2, P_3, \ldots\ldots P_k$ respectively, entropy for variable X is given in the equation below [7].

$$Entropy = H(X) = -\sum_{j=1}^{k} p_j log_2(p_j) \tag{1}$$

When the target attribute of the sub-clusters $T_1, T_2, T_3, \ldots\ldots T_k$in the training set is subdivided into sub-compartments, the weighted average of the information required to determine the class of each T is given as the weighted sum of entropies.

$$H_S(T) = \sum_{i=1}^{k} p_i H_S(T_\text{i}) \tag{2}$$

Information gain is calculated to perform the separation process. The C5.0 algorithm realizes the optimal separation process by determining the separation criterion that has the greatest information gain in each decision node. Information gain is given in the equation below[8].

$$IG\ (S) = H(T) - H_S(T) \tag{3}$$

2.3.2. Support Vector Machine (SVM)

SVM, which is accepted as the latest technology in pattern recognition, aims to increase the predictive performance by finding the Maximal Marginal Hyper Plane (MMH). Sequential Minimum Optimization (SMO) improves the training of the SVM classifier using polynomial nuclei. This generally replaces all missing values and converts the nominal properties to binary values[9,10].To find a decision boundary between the two classes, SVM tries to maximize the gap between classes, choosing linear separations in a property area. Classification of the k-core function points in space $x_i$isy_i, which varies between -1 and +1. If x' is a point with an unknown classification, the prediction classification y''is as in the equation below.

$$y' = Sign(\sum_{i=1}^{n} \alpha_i y_i K(X_i, X') + d \tag{4}$$

In the equation, K; core function, n; support vector number, α; adjustable weight and d are defined as bias. The classification process is linear in the number of support vectors [11].

2.3.3. Multilayer Perceptron (MLP)

The most widely used artificial neural network model today is the MLP network, which has also been extensively analyzed and many learning algorithms have been developed from it.[12].MLP is a feed-forward, fully artificial neural network model that maps input data sets to an appropriate output set by adjusting the weight between internal data nodes.

$$y = \emptyset(\sum_{i=1}^{n} W_i X + b) = \emptyset(W^T X + b) \tag{5}$$

Equality; W defines the weight vector, X the vector of inputs, b bias (bias), and ∅ activation function [13].

2.4. Ensemble Learning

Ensemble learning methods essentially aim to achieve the most accurate result by combining different methods. It can also be applied successfully in various machine learning systems such as feature extraction, error correction, unstable data, learning to deviate in non-stationary distributions, and confidence estimation."Bagging and Boosting" are the most commonly used algorithms for the training of ensemble classifiers. The most common unification rule used to combine individual classifiers is majority voting. The choice of the $W_c$class with the majority vote is as inequality [14,15].

$$\sum_{t=1}^{T} d_{t,c} = max_c \sum_{t=1}^{T} d_{t,c} \tag{6}$$

2.5. Performance Metrics

Accuracy (AC) is defined as the division of values incompatible eyes by the total number of observations and is indicated by equation 7.

$$AC = \frac{TP+TN}{TP+TN+FN+FP} \tag{7}$$

Sensitivity is the ability of the test to distinguish patients from real patients and is indicated by equation 8.

$$Sensitivity = \frac{TP}{TP+FP} \tag{8}$$

Specificity is the ability of the test to distinguish robots from real robots and is indicated by equation 9 [16].

$$Specificity = \frac{TN}{\text{TN+FN}} \tag{9}$$

3. RESULTS

3.1. Model Development

In data sets of 303, 1000, and 10000; Due to the low performance of the model, the feature selection model was applied to the data set. Variables 0.8 and above were determined as important contributing variables, while 0.6 and above variables were determined as marginal contributing variables. After the optimization process, data sets were divided into two as 70 % training and 30 % testing. Data analysis was performed by using the IBM SPSS Modeler Version 18.0 package program.

3.2. Evaluation of the Models

After the model development, the evaluation metrics calculated within the scope of the investigation of how the sample size affects the model performance by using different classification methods are shown in Table II. For n = 303, the highest accuracy rate in the train data set was 77.2 %, while the group was ensemble learning, while the lowest classifier was MLP with 60.7 %.

TABLE I

THE DETAIL EXPLANATION OF THE VARIABLES IN THE DATASET

Variables	***Explanation***
Class	Target(0: healthy,1: disease)
Age	age
Gender	gender*(1=male, 0=female)*
Chest pain type	chest pain type *(1=angina, 2=atypical angina, 3=non-anginal pain, 4=asymptomatic pain*)
Resting blood pressure	resting blood pressure
Serum cholesterol	serum cholesterol in mg/dl
Blood sugar	fasting blood sugar, (*fasting blood sugar > 120 mg/dl) (1 = true; 0 = false*)
Electrocardiographic results	resting electrocardiographic results (0=normal,1= having ST-T wave abnormality, 2= showing probable or definite left ventricular hypertrophy by Estes' criteria)
Max heart rate	maximum heart rate achieved
Pain lock	exercise induced angina (*1 = yes; 0 = no)*
Oldpeak	Oldpeak= *ST depression induced by exercise relative to rest*
ST-segment	the slope of the peak exercise ST segment
Vessels	number of major vessels
Thal	Thal(A thalliumstress test; thal: 3 = normal; 6 =

In the test data set after model training, the highest classifier was again ensemble learning with 76.7 %, while the lowest was C5.0 with 63.3 %. For n = 1000, the highest accuracy rate in the train data set was 95.4 %, while the group was ensemble learning, while the lowest classifier was MLP with 66.7 %. In the test data set after model training, the highest classifier was again ensemble learning with 96.8 %, while the MLP was the lowest with 62.4 %. For n = 10000, the highest accuracy rate in the training data set was MLP with 94.2 %, while the lowest classifier was C5.0 with 86.7 %. After model training, the highest classifier was again MLP with 100 % in the test data set, while SVM was the lowest with 96.3 %.

TABLE II

MODEL PERFORMANCE METRICS

Train (n=303)	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Time (Second)**
SVM	69.7	63.7	62.6	5
C5.0	75.6	68.8	63.6	4
MLP	60.7	54.9	52.8	6
Ensemble	77.2	68	70.5	7
Test (303)				
SVM	73	70.6	75	2
C5.0	63.3	69.7	100	1
MLP	67.4	72.6	66.7	3
Ensemble	76.7	71.7	81.2	5
Train (n=1000)				
SVM	86	79.5	78.2	15
C5.0	94.1	88.6	83.8	12
MLP	66.7	63.5	58.9	13
Ensemble	95.4	87.8	87.2	11
Test (n=1000)				
SVM	90.9	81	83.3	8
C5.0	95.2	89.9	89.5	7
MLP	62.4	55.7	65	9
Ensemble	96.8	92.6	89.7	6
Train (n=10000)				
SVM	90.4	82.6	82.9	34
C5.0	86.7	84.5	81.7	28
MLP	94.2	88.6	86.2	44
Ensemble	90.5	86.4	82.1	23
Test (n=10000)				
SVM	96.3	90.3	90.2	17
C5.0	96.7	93.3	91	12
MLP	1	1	1	38
Ensemble	99.3	98.5	98.6	11

4. CONCLUSION

Diagnosis and treatment of a serious disease, such as cardiovascular diseases, is a very difficult problem and requires many pretreatment experiments and important datasets. The success of the models to be used when applying different classification methods can only be measured by proving the performance. In this study, increasing the sample size in the data sets positively contributes to the model performance, it was determined that an ensemble learning algorithm is an approach that can be suggested in three data sets in general.

ACKNOWLEDGMENT

The study was reported as oral presentation in1st International Data Science Congress in Health on 05-06 December 2019.

REFERENCES

[1] Wong, N. D. (2014). *Epidemiological Studies of CHD and the Evolution of Preventive Cardiology.* Nature Reviews. Cardiology, 11(5), 276.

[2] Verma, L.,Srivastava, S., Negi, P. C. (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *Journal of Medical Systems*, 40(7), 1-7.

[3] Alizadehsani, R.,Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., &Boghrati, R. (2012). Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms. In Data Mining Workshops (ICDMW), *2012 IEEE 12th International Conference on (pp. 9-16). IEEE.*

[4] Srinivas, K.,Rani, B. K., &Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering* (IJCSE), 2(02), 250-255.

[5] E. Smirnov, I. Sprinkhuizen-Kuyper, and G. Nalbantov, "Unanimous voting using support vector machines," in *BNAIC-2004: Proceedings of the Sixteenth Belgium-Netherlands Conference on Artificial Intelligence*, 2004, pp. 43-50.

[6] Nicholas, E. (2008). Introduction to Clementineand Data Mining. Brigham Young University

[7] Larose, D.T., and Larose, C.D. (2014) *Discovering Knowledge In Data An Introduction To Data Mining*, New Jersey: John Wiley&Sons.

[8] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

[9] Platt, J. C. (1999). Fasttraining of support vector machines using sequential minimal optimization, advances in kernel methods. Support Vector Learning, 185-208.

[10] Azuaje, F. (2006). Wittenih, frank e: Veri madenciliği: Pratik makine öğrenim araçları ve teknikleri. Biyomedikal mühendislik çevrimiçi , 5 (1), 1-2.

[11] Valdimir, V. N.,&Vapnik, N. (1995). Thenature of statistical learning theory.

[12] Rosenblatt, F. (1958). T*wo theorems of statistical separability in the perceptron*. United States Department of Commerce

[13] Miller, D. J.,& Pal, S. (2007). Transductive methods for the distributed ensemble classification problem. *Neural computation*, 19(3), 856-884.

[14] Zhang, C.,&Ma, Y. (Eds.). (2012) *Ensemble machine learning: methods and applications*. Springer Science& Business Media.

[15] Polikar, R. (2012). Ensemble learning. *In ensemble machine learning*, Springer, Boston, MA, 1-34.

[16] Alpar, R. (2016). Uygulamalı istatistik ve geçerlik-güvenirlik: *spor, sağlık ve eğitim bilimlerinden örneklerle.* Detay Yayıncılık.

.

BIOGRAPHIES

Mehmet Kıvrak obtained his BSc degree in statistics from Dokuz Eylul University (DEU) in 2001. He received the BSc. and MSc. diploma in StatisticsfromDokuz Eylul Universityin 2001 and 2006 respectively, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics ofInonu University in 2017. He accepted as an expert statisticianTurkish Statistical Institute in 2009. His research interests are data mining, cognitive systems, reliability and genetics and bioengineering, and signal processing. His current research interests are genetics and bioengineeringand data mining.

F.Berat Akçeşme obtained his BSc degree in biological sciences and bioengineeringfrom the International University of Sarajevo in 2004. He received the BSc. and MSc. diploma in biological sciences and bioengineering from the International University of Sarajevoin 2004 and 2009 respectively, and Ph.D. degrees in Genetics and Bioengineering of the same university in 2012. He accepted as a Postdoctoral Researcher by the department of biostatistics and medical informatics faculty of medicine, university of health sciences in 2012. His research interests are cognitive systems, reliability and biomedical system, and genetics, and bioengineering. In 2017, he joined the Department of Biostatistics and Medical Informatics Faculty of Medicine, University of Health Sciences as an assistant professor, where he is presently anassistant professor. He is active in teaching and research in the general genetics and bioengineering modeling, analysis.

Cemil Çolak obtained his BSc. degree in statisticsfrom Ondokuz Mayıs University in 1999. He received MSc. diploma in statistics from the InonuUniversity in 2001, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. He accepted as a Postdoctoral Researcher by the department of biostatistics and medical informatics of Inonu University in 2007. His research interests are cognitive systems,data mining, reliability, and biomedical system, and genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a professor, where he is presently a professor. He is active in teaching and research in the general image processing and data mining modeling, analysis.