ASSESSMENT OF ASSOCIATIVE CLASSIFICATION APPROACH FOR PREDICTING MORTALITY BY HEART FAILURE

Z. Kucukakcali, I. Balikci Cicek, E. Guldogan, and C. Colak

Abstract— *Aim:* This study aims to predict mortality status by heart failure and to determine the related factors by applying the relational classification method, one of the data mining methods, on the open-access heart failure data set.

Materials and Methods: In this study, the associative classification model has been applied to the open-access data set named “Heart Failure Prediction”. The performance of the model was evaluated by accuracy, balanced accuracy, sensitivity, selectivity, positive predictive value, negative predictive value, and F1-score.

Results: Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values obtained from the model were 0.866, 0.819, 0.688, 0.951, 0.868, 0.865 and 0.767 respectively.

Conclusion: The findings obtained from this study showed that successful results were obtained in the study performed with the associative classification model on the heart failure data set. Also, certain rules regarding the disease to be used in preventive medicine practices were obtained with the proposed model.

Keywords— Heart failure, classification, association rules, relational classification.

1. INTRODUCTION

HEART failure (HF) is a medical syndrome caused by cardiac structural or functional abnormalities, accompanied by typical symptoms and signs that occur due to decreased cardiac output (CO) and/or increased intracardiac pressure at rest, stress, and effort. These symptoms are shortness of breath, pretibial edema, weakness; findings are increased jugular venous pressure, pulmonary rales, and peripheral edema. [1]. HF causes serious mortality and morbidity and poses a serious burden to the healthcare system worldwide. Heart failure is seen in one in 10 people aged 75 and over in developed countries. In the 2025 program of the World Health Organization (WHO), it has been stated that the burden of HF disease is a potential target to be reduced. In a study carried out at our country, the incidence of HF in Turkey was determined to be 2.9%. Besides, it has been shown that CG affects 1.5 million people, and 3 million people are at risk soon [2, 3].

Data mining can be defined simply as the discovery of useful information hidden in data. Data mining enables researchers to make effective and informed decisions with techniques offered by different disciplines such as artificial intelligence, machine learning, statistics, and optimization. It also enables revealing hidden, implicit, beneficial relationships, patterns, relations, or trends that are difficult to reveal with classical methods. Data mining is the search for the relations and rules that will allow us to make predictions about the future of a large amount of data using computer programs [4]. Associative classification is a branch of scientific work, known as data mining. Associative classification combines the association rule and classification, two known methods of data mining, to create a model for predictive purposes. In other words, associative classification is a type of classification approach that is created with a set of rules obtained by the association rule mining to create classification models. One of the important advantages of using a classification based on association rules according to classical classification approaches is that the output of an associative classification algorithm is represented by simple if-then rules, making it easier for the users to understand and interpret it [5].

This study aims to predict mortality status by heart failure and to determine the related factors by applying the associative classification method, one of the data mining methods, on the open-access heart failure data set.

2. MATERIAL AND METHODS

2.1. Dataset

In the study, the associative classification model, which is a data mining method that combines classification and association rules methods, has been applied to an open-access data set named "Heart Failure Prediction" [6].

There are 299 patients in the data set used. 96 (32.1%) of these patients died after a certain period of follow-up. Explanations about the variables and their properties in the data set are given in Table I.

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

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

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

✉ **Cemil COLAK**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr)

Manuscript received August 12, 2020; accepted Sep. 29, 2020.
Digital Object Identifier:

TABLE I
EXPLANATIONS ABOUT THE VARIABLES IN THE DATASET AND THEIR PROPERTIES

Variable	Variable Description	Variable Type	Variable Role
age	Decrease of red blood cells or hemoglobin	Quantitative	Predictor
anemia	Level of the CPK enzyme in the blood (mcg/L) (boolean)	Qualitative	Predictor
creatinine_phosphokinase	Level of the CPK enzyme in the blood (mcg/L)	Quantitative	Predictor
diabetes	If the patient has diabetes (boolean)	Qualitative	Predictor
ejection_fraction	Percentage of blood leaving the heart at each contraction (percentage)	Quantitative	Predictor
high_blood_pressure	If the patient has hypertension (boolean)	Qualitative	Predictor
platelets	Platelets in the blood (kiloplatelets/mL)	Quantitative	Predictor
serum_creatinine	Level of serum creatinine in the blood (mg/dL)	Quantitative	Predictor
serum_sodium	Level of serum sodium in the blood (mEq/L)	Quantitative	Predictor
sex	Woman or man (binary)	Qualitative	Predictor
smoking	If the patient smokes or not (boolean)	Qualitative	Predictor
time	Follow-up period (days)	Quantitative	Predictor
death_event	If the patient deceased during the follow-up period (boolean)	Qualitative	Output

3. ASSOCIATIVE CLASSIFICATION

Rules of association are a type of unsupervised data mining that looks for the relationship between records in a data set. Association rules are the process of determining the events or features that occur together. Association rules are often expressed as if it happens, then this happens. Mostly used in descriptive data analysis, data preprocessing, determining discrete values, and finding trends and relationships [7]. Association rules are rules with support and confidence measurements in the form of "IF- precursor expression-, IF- successor expression" [8].

Association rules share many common features with classification. Both use rules to characterize regularities in a dataset. However, these two methods differ greatly in their goals. While classification focuses on prediction, association rules focus on providing information to the user. In particular, it focuses on detecting and characterizing unexpected relationships between data items. [9].

Associative classification is a data mining method that combines classification and association rules methods to make predictions. In other words, an associative classification is an approach that uses rules obtained with association rules to create classification models. Associative classification is a special association rule mining with the target/response/dependent/class variable to the right of the rule obtained. In a rule such as X →Y, Y must be the target / response / dependent / class variable. One of the principal benefits of using a classification based on association rules according to classical classification approaches is that simple if-then rules represent the output of an associative classification algorithm. This advance makes it easier for the user to understand and interpret the results [10].

3.1. Performance evaluation criteria

The classification matrix for the calculation of performance metrics is given in Table II.

TABLE II
THE METRICS OF MODEL'S CLASSIFICATION PERFORMANCE

		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

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Balanced accuracy = [[TP/(TP+FP))] + [TN/(TN+FN)]]/2

Sensitivity = TP/(TP+FP)

Specificity = TN/(TN+FN)

Positive predictive value = TP/(TP+FN)

Negative predictive value =TN/(TN+FP)

F-score = (2*TP)/(2*TP+FP+FN)

4.DATA ANALYSIS

Quantitative data are summarized by median (minimum-maximum), and qualitative variables are given by number and percentage. Normal distribution was evaluated with the Kolmogorov-Smirnov test. In terms of input variables, the existence of a statistically significant difference and the relationship between the categories of the output variable, "who died during follow-up" and "who did not die during follow-up", was examined using the Mann-Whitney U and Pearson Chi-square test. The values of $p<0.05$ were deemed statistically significant. In all analyzes, IBM SPSS Statistics 26.0 for the Windows package program was used.

5. RESULTS

Descriptive statistics related to the target variable examined in this study are presented in Table 3 and Table 4. A statistically significant difference exists between output variable classes in terms of age, ejection_fraction, serum_creatinine, serum_sodium, time variables. (p<0.001)

TABLE III

DESCRIPTIVE STATISTICS FOR QUANTITATIVE INPUT VARIABLES

Variables	death_event		P* value
	Survived patients	Dead patients	
	Median (min-max)	**Median (min-max)**	
age	60(40-90)	65(42-95)	**<0.001**
creatinine_phosphokinase	245(30-5209)	259(237-861)	0,684
ejection_fraction	38(17-80)	30(14-70)	**<0.001**
platelets	263000(25100-850000)	258500(47000-621000)	0,425
serum_creatinine	1(0,5-6,1)	1,3(0,6-9,4)	**<0.001**
serum_sodium	137(113-148)	136(116-146)	**<0.001**
time	172(12-285)	45(4-241)	**<0.001**

*: Mann Whitney U test

TABLE IV

DESCRIPTIVE STATISTICS FOR QUALITATIVE INPUT VARIABLES

Variables		death_event		P* value
		Survived patients	Dead patients	
anaemia	absence	120(59.1%)	50(52.1%)	0.252
	presence	83(40.9%)	46(47.9%)	
diabetes	absence	118(58.1%)	56(58.3%)	0.973
	presence	85(41.9%)	40(41.7%)	
high_blood_pressure	absence	137(67.5%)	57(59.4%)	0.170
	presence	66(32.5%)	39(40.6%)	
Gender	woman	71(35.0%)	34(35.4%)	0.941
	man	132(65.0%)	62(64.6%)	
smoking	absence	137(67.5%)	66(68.75%)	**0.827**
	presence	66(32.5%)	30(31.25%)	

***: Pearson's chi-square test**

The classification matrix of the associative classification model used to classify the heart failure dataset in this study is given below in Table V.

TABLE V

CLASSIFICATION MATRIX FOR THE ASSOCIATIVE CLASSIFICATION MODEL

Prediction	**Reference**		
	Survived patients	Dead patients	**Total**
Survived patients	193	30	223
Dead patients	10	66	76
Total	203	99	299

The values for the metrics of the classification performance of the model are given in Table 6. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values obtained from the model were 0.866, 0.819, 0.688, 0.951, 0.868, 0.865 and 0.767 respectively.

TABLE VI

VALUES FOR THE CLASSIFICATION PERFORMANCE METRICS OF THE MODEL

Metric	**Value**
Accuracy	0.866
Balanced accuracy	0.819
Sensitivity	0.688
Specificity	0.951
Positive predictive value	0.868
Negative predictive value	0.865
F1-score	0.767

Table 7 shows the first 5 of the association rules used by the classification algorithm. As expressed in Table 7, when anaemia=0,ejection_fraction=[27.5,80),serum_sodium=[136,148),time=[73.5,285) are considered, the patient's survival probability is 98.7 %.

TABLE VII

Left-hand side rules	**Right-hand side rules**	**Support**	**Confidence**	**Freq.**
{anaemia=0,ejection_fraction=[27.5,80),serum_sodium=[136,148),time=[73.5,285)}	{death_event=0}	0.247	0.987	74
{age=[40,71),diabetes=0,ejection_fraction=[27.5,80),serum_creatinine=[0.5,1.81),serum_sodium=[136,148),time=[73.5,285)}	{death_event=0}	0.214	0.985	64
{age=[40,71),ejection_fraction=[27.5,80),serum_creatinine=[0.5,1.81),sex=1,time=[73.5,285)}	{death_event=0}	0.294	0.978	88
{age=[40,71),anaemia=0,ejection_fraction=[27.5,80),platelets=[1.28e+05,8.5e+05),serum_creatinine=[0.5,1.81),time=[73.5,285)}	{death_event=0}	0.288	0.977	86
{age=[40,71),diabetes=0,ejection_fraction=[27.5,80),platelets=[1.28e+05,8.5e+05),serum_creatinine=[0.5,1.81),time=[73.5,285)}	{death_event=0}	0.261	0.975	78

If age =[40,71),diabetes = 0, ejection_fraction=[27.5,80), serum_creatinine=[0.5,1.81),serum_sodium=[136,148),time=[73.5,285) are considered, the patient's survival probability is 98.5 %.
As age=[40,71),ejection_fraction=[27.5,80),serum_creatinine=[0.5,1.81),sex=1,time=[73.5,285) are considered, the patient's survival probability is 97.8 %. If age=[40,71),anaemia=0,ejection_fraction=[27.5,80),platelets=[1.28e+05,8.5e+05),serum_creatinine=[0.5,1.81),time=[73.5,285) are considered, the patient's survival probability is 97.7 %. age=[40,71),diabetes=0,ejection_fraction=[27.5,80),platelets=[1.28e+05,8.5e+05),serum_creatinine=[0.5,1.81),time=[73.5,285) are considered, the patient's survival probability is 97.5 %.

6. DISCUSSION

Cardiac failure is the final stage of all forms of cardiac disease, a health problem that is increasing in prevalence and incidence, affecting at least 23 million people worldwide. Heart failure is still one of the most common cardiovascular diseases in the world, and similar clinical results are seen in our country. In recent years, the incidence of heart failure has continued to increase all over the world, and death rates are still at very high levels. Advances in the treatment of cardiovascular disorders increase the survival and lifespan of individuals. Therefore, the follow-up and treatment of patients with heart failure is becoming more important and remains an open area for research and new developments [11, 12].

Association rules, one of the descriptive models of data mining, are methods that analyze the coexistence of events. These relationships are based on the coexistence of data elements and express the co-occurrence of events together with certain possibilities. Classification analysis is one of the basic methods of machine learning and is used by a large scientific community. Classification is an estimation process that assigns each observation in the dataset to the predetermined classes under certain rules [13]. Associative classification makes classification by combining two common data mining methods, association rules, and classification methods. In recent years, association rules methods have been successfully used to create correct classifiers in associative classification [5].

In this study, the associative classification model, one of the data mining methods, was applied to the data set named “Heart Failure Prediction”, which is an open-source data set. For this purpose, different factors (explanatory variables) that may be associated with heart failure (dependent variable) were estimated with the relational classification model, and rules were obtained. According to the experimental results, from the performance metrics obtained from the model, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values obtained from the model were 0.866, 0.819, 0.688, 0.951, 0.868, 0.865 and 0.767 respectively.

In a study conducted with the same data set, the results were compared using ten different machine learning methods. According to the results of this study, the highest accuracy was obtained as 0.74 with the Random Forest model [6]. In this study, an accuracy of 0.866 was obtained, and rules about the disease were also obtained.

As a result, the associative classification model used produced successful results in the study conducted with the heart failure data set. Besides, certain rules regarding the disease to be used in preventive medicine practices have been obtained with this model.

REFERENCES

[1] P. Ponikowski, A. Voors, S. Anker, H. Bueno, J. Cleland, A. Coats, et al., "Authors/Task Force Members; Document Reviewers (2016) 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC," Eur J Heart Fail, vol. 18, pp. 891-975, 2016.

[2] G. Fonarow, K. Adams Jr, W. Abraham, C. Yancy, and W. Boscardin, "ADHERE Scientific Advisory Committee Study Group and Investigators. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis," Jama, vol. 293, pp. 572-80, 2005.

[3] M. B. Yılmaz, A. Çelik, Y. Çavuşoğlu, L. Bekar, E. Onrat, M. Eren, et al., "Snapshot evaluation of heart failure in Turkey: Baseline characteristics of SELFIE-TR," Turk Kardiyoloji Dernegi arsivi: Turk Kardiyoloji Derneginin yayin organidir, vol. 47, pp. 198-206, 2019.

[4] H. Akpınar, "Veri tabanlarında bilgi keşfi ve veri madenciliği," İstanbul Üniversitesi İşletme Fakültesi Dergisi, vol. 29, pp. 1-22, 2000.

[5] F. A. Thabtah, "A review of associative classification mining," Knowledge Engineering Review, vol. 22, pp. 37-65, 2007.

[6] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, vol. 20, p. 16, 2020/02/03 2020.

[7] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," 1996.

[8] D. T. Larose and C. D. Larose, Discovering knowledge in data: an introduction to data mining vol. 4: John Wiley & Sons, 2014.

[9] N. Ye, The handbook of data mining: CRC Press, 2003.

[10] F. Thabtah, "A review of associative classification mining," The Knowledge Engineering Review, vol. 22, pp. 37-65, 2007.

[11] G. S. Bleumink, A. M. Knetsch, M. C. Sturkenboom, S. M. Straus, A. Hofman, J. W. Deckers, et al., "Quantifying the heart failure epidemic: prevalence, incidence rate, lifetime risk and prognosis of heart failure: The Rotterdam Study," European heart journal, vol. 25, pp. 1614-1619, 2004.

[12] A. Mosterd and A. W. Hoes, "Clinical epidemiology of heart failure," Heart, vol. 93, pp. 1137-1146, 2007.

[13] İ. Perçın, F. H. Yağin, E. Güldoğan, and S. Yoloğlu, "ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine," in 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1-5.

BIOGRAPHIES

Zeynep KÜÇÜKAKÇALI 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.

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.

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.

Cemil ÇOLAK obtained his BSc. degree in statisticsfrom Ondokuz Mayıs University in 1999. He received MSc. degree in Biostatistics from the Inonu University in 2001, and Ph.D. degree in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. His research interests are cognitive systems, data mining, reliability, and biomedical system, 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, artificial intelligence, data mining, analysis.