

# ASSOCIATIVE CLASSIFICATION APPROACH CAN PREDICT PROSTATE CANCER BASED ON THE EXTRACTED ASSOCIATION RULES

I. Balıkcı Cicek, Z. Küçükakcalı and C. Colak

**Abstract— Aim:** This study aims to classify the diagnosis status of prostate cancer and determine the related factors by applying the associative classification method, one of the data mining methods, to the open-access prostate cancer data set.

**Materials and Methods:** In the current study, an open-access data set named "Prostate Cancer" is used for classification. The performance of the associative classification model is evaluated using the classification performance metrics such as sensitivity, selectivity, accuracy, balanced accuracy, negative predictive value, positive predictive value, and F1-score.

**Results:** According to the prostate cancer classification results obtained from the associative classification model, the accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score values were obtained as 0.968, 0.789, 0.9, 0.879, 0.938, 0.882 and 0.923, respectively.

**Conclusion:** In the analysis of the open-access data set, the proposed associative classification model has distinctively successful results in classifying prostate cancer on the performance metrics.

**Keywords—** Prostate cancer, data mining, association rules, classification, associative classification.

## 1. INTRODUCTION

PROSTATE cancer is a type of cancer characterized by the abnormal division of cells in the prostate gland. Prostate cancer is the second most common cancer in men and the fifth leading cause of death worldwide [1]. However, it is the most common cancer diagnosed in men over middle age, both in developed and developing countries. While the probability of developing prostate cancer in a male between the ages of 0-39 is 0.01%, it is 2.58% for a man between the ages of 40-59, and 14.7% between the ages of 60-79. Besides, the probability of developing prostate cancer in a man during his lifetime is around 17.8% [2]. The risk factors of prostate cancer include familial predisposition, advanced age, race, genetics, diet, environmental factors, and hormonal factors [3]. Because it is seen too much, the correct management of the treatment,

diagnosis, and follow-up process of prostate cancer is important for the patient and the doctor, as well as for national health policies [4].

Data mining is an interdisciplinary field that acts as a bridge in many areas such as statistics, data visualization, and pattern recognition, database technology, machine learning, and artificial intelligence [5]. Data mining can be defined as discovering hidden relationships and patterns in data. It focuses on the methods required to define a valid, useful, and new model [6]. In general, data mining models are divided into two as descriptive and predictive. One of its descriptive models is association rules [7].

Data mining methods that analyze the co-occurrence of events are called association rules. Association rules express the co-occurrence of events together with certain possibilities. Association rules are methods that help reveal a seemingly unrelated relationship between data [8]. Association rules analysis are aimed at finding which variables are "together". Association rules analysis uses data mining methods to reveal hidden relationships between data [9].


Associative classification is a data mining model and is based on the logic of integrating the association rule model and classification. Associative classification is a type of classification approach that is created with a set of rules obtained by association rule mining to create classification models. The presence of the target/response variable on the right side of the rule obtained in the associative classification algorithm makes it easier for the user to understand and interpret it [10].

In the present study, it is aimed to classify prostate cancer as benign malignant by applying the associative classification model using association rules on open access prostate cancer data set, and to determine the rules related to the diagnosis of prostate cancer.


## 2. MATERIAL AND METHODS

### 2.1. Dataset

In the study, in order to examine how the associative classification method works and to evaluate the model, the open-access data set named "Prostate Cancer Data Set" was obtained from the address <https://www.kaggle.com/sajidsaifi/prostate-cancer> [11]. There are 100 patients diagnosed with prostate cancer in this open access data set. Of the patients diagnosed with cancer, 38 (38%) were diagnosed as benign, and 62 (62%) were diagnosed as malignant. The variables and the descriptive properties of the variables in the relevant data set 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) 

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

Manuscript received Sep 19, 2020; accepted Nov. 5, 2020.  
Digital Object Identifier:

TABLE I  
VARIABLES IN THE DATA SET AND THEIR DESCRIPTIVE PROPERTIES

Variable	Variable Description	Variable Type	Variable Role
Diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)	Qualitative	Dependent/ Target
Radius	Mean distances from the center to perimeter points	Quantitative	Independent / Predictor
Texture	The standard deviation of gray-scale values	Quantitative	Independent / Predictor
Perimeter	Mean size of the core tumor	Quantitative	Independent / Predictor
Area	-	Quantitative	Independent / Predictor
Smoothness	Mean of local variation in radius lengths	Quantitative	Independent / Predictor
Compactness	(mean of perimeter) <sup>2</sup> / (area - 1)	Quantitative	Independent / Predictor
Symmetry	-	Quantitative	Independent / Predictor
Fractal dimension	mean for "coastline approximation" - 1	Quantitative	Independent / Predictor

### 3. ASSOCIATIVE CLASSIFICATION MODEL

Association rules are an unsupervised data mining method that searches for relationships between data records. Association rules are used to detect useful, consistent, and interesting relationships from large data sets that cannot be deduced at first glance. Association rules, in other words, are useful for detecting frequently seen situations in the data set. The main purpose of association rule algorithms is to determine the probability of occurrence of two or more events together [12]. Relationships between data in association rules analysis are shown with IF-THEN expressions. "IF <if certain conditions are met>" is in the form "THEN <estimate the values of some attributes>" and are rules that are measures of support and trust. Association rules are an approach that supports future studies by analyzing past data and determining association behaviors in these data. One of the data mining models, association rules, is frequently preferred in terms of applicability in almost every field [13].

Associative classification is a supervised learning model that classifies based on the use of association rules. Association rules are derived using "if-then" clauses called precursor-successor. In associative classification models, the right side (successor) of association rules consists only of the categories of the class/response / dependent variable. Thus, the relevant associative classification model classifies the unlabeled test data set by using these association rules [14].

#### 3.1. Performance evaluation metrics

While comparing the classification performances, performance metrics such as sensitivity, specificity, accuracy, balanced accuracy, negative predictive value, positive predictive value, and F1-score were used.

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

TABLE II  
CLASSIFICATION MATRIX OF PERFORMANCE METRICS

		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

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

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

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

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

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

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

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

### 4. DATA ANALYSIS

Quantitative data were expressed as mean  $\pm$  standard deviation, median (minimum-maximum), and qualitative data as number (percentage). Conformity to normal distribution was assessed using the Shapiro-Wilk test. Whether there is a statistically significant difference between the "Benign" and "Malignant" groups, which are the categories of dependent/target variable (prostate cancer) in terms of independent variables, was examined using the Mann-Whitney U test and the independent-samples t-test. Values of  $p < 0.05$  were considered statistically significant. IBM SPSS Statistics 26.0 package program was used for all analyzes.

### 5. RESULTS

Descriptive statistics for the independent variables examined in this study are given in Table III. There is a statistically significant difference between the dependent/target variable groups in terms of the perimeter, area, compactness, symmetry, smoothness variables ( $p < 0.05$ ).

TABLE III  
DISTRIBUTION TABLE OF THE DEPENDENT/TARGET VARIABLE

Benign		Malignant	
Count	Percentage	Count	Percentage
38	38	62	62

The distribution table for the dependent/target variable in the data set is given in Table III.

TABLE IV  
DESCRIPTIVE STATISTICS FOR QUANTITATIVE INDEPENDENT VARIABLES

Variables	Diagnosis				p-value
	Benign	Malignant			
	Median (min-max)	Mean± Standard deviation	Median (min-max)	Mean± Standard dev.	
radius	18 (9-25)	-	16 (9-25)	-	0.09*
texture	17 (11-27)	-	18 (11-27)	-	0.450*
perimeter	78.5 (52-133)	-	104 (72-172)	-	<0.001*
area	458.5 (202-1326)	-	790.5 (371-1878)	-	<0.001*
compactnes	0.0785 (0.038-0.246)	-	0.1405 (0.051-0.345)	-	<0.001*
symmetry	0.182 (0.135-0.274)	-	0.193 (0.153-0.304)	-	0.013*
fractal_dimension	0.0635 (0.053-0.09)	-	0.063 (0.053-0.097)	-	0.963*
smoothness	-	0.099± 0.015	-	0.105± 0.014	0.049**

\*: Mann Whitney U test, \*\*: Independent samples t-test

The distribution table for the dependent/target variable in the data set is given in Table IV.

In this study, the classification matrix of the associative classification model used in classifying the prostate cancer dataset is given in Table V.

TABLE V  
CLASSIFICATION MATRIX FOR THE ASSOCIATIVE CLASSIFICATION MODEL

Prediction	Reference		
	Bening	Malignant	Total
<b>Bening</b>	30	2	32
<b>Malignant</b>	8	60	68
<b>Total</b>	38	62	100

The values of the classification performance metrics for the associative classification model are shown in Table 6. The sensitivity obtained from the model was calculated as 0.968, selectivity 0.789, accuracy 0.9, balanced accuracy 0.879, negative predictive value 0.938, positive predictive value 0.882, and F1-score 0.923.

Table 7 shows the association rules used by the classification algorithm. As expressed in Table 7, when perimeter=[52, 87.5) and compactness=[0.038, 0.097) are considered, the probability of a male not having prostate cancer is 96.6%. Similarly, as texture=[17.5, 27) and perimeter=[87.5, 172) are taken into account, the probability of a man getting prostate cancer is 96.3%, and when area=[575, 1.88e+03) and smoothness=[0.0905, 0.143) are regarded, the probability of a man with prostate cancer is 95.7%. If area=[575, 1.88e+03) and compactness=[0.097,0.345) are considered, the probability of a man with prostate cancer is 95.7 %.

TABLE VI  
VALUES OF THE CLASSIFICATION PERFORMANCE METRICS OF THE MODEL

Metric	Value
Sensitivity	0.968
Specificity	0.789
Accuracy	0.9
Balanced accuracy	0.879
Negative predictive value	0.938
Positive predictive value	0.882
F1-score	0.923

The other rules generated from the classification based on the association rules model can be interpreted as the rules described earlier (Table VII).

TABLE VII  
ASSOCIATION RULES USED BY THE CLASSIFICATION ALGORITHM

Left-hand side rules	Right-hand side rules	Sup.	Conf.	Freq.
{perimeter=[52, 87.5), compactness=[0.038,0.097)}	{diagnosis=Benign}	0.28	0.966	28
{texture=[17.5,27), perimeter=[87.5,172)}	{diagnosis=Malignant}	0.26	0.963	26
{area=[575,1.88e+03), smoothness=[0.0905,0.143)}	{diagnosis=Malignant}	0.45	0.957	45
{area=[575,1.88e+03), compactness=[0.097,0.345)}	{diagnosis=Malignant}	0.44	0.957	44
{texture=[11,16.5), perimeter=[87.5,172), compactness=[0.097,0.345)}	{diagnosis=Malignant}	0.2	0.952	20
{perimeter=[87.5,172), smoothness=[0.0905,0.143)}	{diagnosis=Malignant}	0.47	0.94	47
{area=[575,1.88e+03)}	{diagnosis=Malignant}	0.5	0.926	50
{texture=[11,16.5), compactness=[0.097,0.345)}	{diagnosis=Malignant}	0.25	0.926	25
{radius=[9,24.5), texture=[17.5,27), compactness=[0.097,0.345)}	{diagnosis=Malignant}	0.25	0.893	25

## 6. DISCUSSION

Prostate cancer is a disease that can start anywhere in the prostate gland, slow in the first 5-10 years, then grow rapidly and can spread to other organs [15]. Prostate cancer is an important cause of illness and death among men. Prostate cancer-initiating causes of prostate cancer are not fully known yet, factors such as genetic factors, chronic inflammation, and infection, high-fat diet, smoking, alcohol use, obesity [16]. The global burden of prostate cancer is among the top five cancers in terms of both incidence and mortality. Therefore, it is possible to prevent the progression of the disease and to apply alternative treatment protocols by diagnosing prostate cancer in the early stages [1].

Data Mining is the analysis and summarization of large amounts of data by various methods in order to detect unexpected or unpredictable relationships and obtain understandable and useful information [17]. One of the most important areas of achieving attractive results in data mining studies is association rules mining. Association rules are mostly

used in explanatory data analysis, data preprocessing, determining discrete values, finding trends, and relationships. The relationships between the obtained rules and object or object groups can be determined, and they can be used as a guide in decisions to be made and definitions to be made [18]. Associative classification is a data mining model that classifies by combining association rules and classification methods. The associative classification model gives high accuracy results compared to traditional methods. The main advantage of associative classifiers is that they are easy to interpret. Other models found in data mining complicate the interpretability of results due to a large number of unnecessary and contradictory rules. Associative classification, on the other hand, provides great success in interpreting and classifying results by trimming unnecessary rules [10].

In this study, an associative classification model, which combines the association rule and classification, which is one of the data mining methods, was applied to the data set named "Prostate Cancer", which is an open-access data set [11]. In this context, different factors (independent variables) that may be associated with a prostate cancer diagnosis (the dependent variable) are estimated with the associative classification model, and association rules have been found. According to the results of the associative classification analysis, the performance metrics obtained from the model were obtained as sensitivity 0.968, selectivity 0.789, accuracy 0.9, balanced accuracy 0.879, negative predictive value 0.938, positive predictive value 0.882, and F1-score 0.923.

In a study using the same data set, the accuracy results obtained with different machine learning methods were compared. According to the results of the current study, the highest accuracy was obtained as 0.80 with the k-Nearest Neighbor and Naive Bayes Classification models. In this study, an accuracy of 0.9 was obtained, and the rules related to the disease were also obtained [19].

As a result, the associative classification model proposed as a result of the analysis of the open-access data set produces distinctively successful predictions in classifying prostate cancer according to performance metrics.

#### REFERENCES

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, pp. 394-424, 2018.
- [2] A. Jemal, R. C. Tiwari, T. Murray, A. Ghafoor, A. Samuels, E. Ward, et al., "Cancer statistics, 2004," *CA: a cancer journal for clinicians*, vol. 54, pp. 8-29, 2004.
- [3] A. Jemal, A. Thomas, T. Murray, and M. Thun, "Cancer statistics, 2002," *Ca-A Cancer Journal for Clinicians*, vol. 52, pp. 23-47, 2002.
- [4] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, pp. 7-30, 2016.
- [5] S. Özkes, "Veri madenciliği modelleri ve uygulama alanları," 2003.
- [6] A. K. Pujari, *Data mining techniques: Universities press*, 2001.
- [7] H. Akpınar, "Veri tabanlarında bilgi keşfi ve veri madenciliği," *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, vol. 29, pp. 1-22, 2000.
- [8] L. Boney, A. Tewfik, and K. Hamdy, "Minimum association rule in large database," in *Proceedings of Third IEEE International Conference on Computing*, 2006, pp. 12-16.

- [9] S. Vinodh, N. H. Prakash, and K. E. Selvan, "Evaluation of leanness using fuzzy association rules mining," *The International Journal of Advanced Manufacturing Technology*, vol. 57, pp. 343-352, 2011.
- [10] M. Azmi, G. C. Runger, and A. Berrado, "Interpretable regularized class association rules algorithm for classification in a categorical data space," *Information Sciences*, vol. 483, pp. 313-331, 2019.
- [11] Available: <https://www.kaggle.com/sajidsaifi/prostate-cancer>
- [12] Y. Köse, "Değerli müşterilerde ürün kategorileri arasındaki satış ilişkilerinin veri madenciliği yöntemlerinden birliktelik kuralları ve kümeleme analizi ile belirlenmesi ve ulusal bir perakendecide örnek uygulama," *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü*, 2015.
- [13] A. S. Albayrak and S. K. Yılmaz, "Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama," *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, vol. 14, 2009.
- [14] F. A. Thabtah, "A review of associative classification mining," *Knowledge Engineering Review*, vol. 22, pp. 37-65, 2007.
- [15] T. Kahvecioğlu and F. E. Güneş, "İnek sütü ve prostat kanseri ilişkisi," *Sağlık ve Yaşam Bilimleri Dergisi*, vol. 1, pp. 44-49, 2019.
- [16] A. Ahlbom, P. Lichtenstein, H. Malmström, M. Feychting, N. L. Pedersen, and K. Hemminki, "Cancer in twins: genetic and nongenetic familial risk factors," *Journal of the National Cancer Institute*, vol. 89, pp. 287-293, 1997.
- [17] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," 1996.
- [18] A. S. Kumar and R. Wahidabanu, "Data Mining Association Rules for Making Knowledgeable Decisions," in *Data Mining Applications for Empowering Knowledge Societies*, ed: IGI Global, 2009, pp. 43-53.
- [19] Available: <https://www.kaggle.com/alihantabak/prostate-cancer-predictions-with-ml-and-dl-methods>.

#### 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 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.

**Cemil ÇOLAK** obtained his BSc. degree in statistics from 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.