ORIGINAL ARTICLE

Comparing the Results of Logistic Regression Model and Classification and Regression Tree Analysis in Determining Prognostic Factors for Coronary Artery Disease in Mashhad, Iran

Zahra Bami[1],
Nasser Behnampour[2],
Bahram Sadeghpour Gildeh[3],
Majid Ghayour Mobarhan[4],
Habibollah Esmaily[5]

[1] MSc Student in Biostatistics, Faculty of Health, Golestan University of Medical Sciences, Gorgan, Iran
[2] Assistant Professor, Department of Biostatistics and Epidemiology, Faculty of Health, Golestan University of Medical Sciences, Gorgan, Iran
[3] Professor, Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran
[4] Professor, Department of Nutrition, School of Medicine, Metabolic Syndrome Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[5] Professor, Department of Epidemiology and Biostatistics, School of Health, Neonatal Research Center, Mashhad University of Medical Sciences Mashhad, Iran

(Received April 27, 2020 ; Accepted Junary 12, 2021)

Abstract

Background and purpose: Understanding of the risk factors for cardiovascular artery disease, which is the leading cause of death worldwide, can lead to essential changes in its etiology, prevalence, and treatment. The aim of this study was to compare the results of logistic regression model and Classification and Regression Tree Analysis (CART) in determining the prognostic factors for coronary artery disease in people living in Mashhad, Iran.

Materials and methods: The present case-control study used the cohort data of Mashhad stroke and heart atherosclerotic disorder (MASHAD STUDY), 2009. The prognostic factors for coronary artery disease were determined by CART and Logistic regression models using R and Stata 14. Then, the efficiency of the models was compared by computing the area under the performance characteristic curve (AUC). All patients with coronary artery disease were considered as the case and for each case, three controls were selected.

Results: According to Logistic model, prognostic factors for coronary artery disease included age, history of myocardial infarction, diabetes, history of hyperlipidemia, and family history of heart disease (father and brother). The CART algorithm showed age, history of myocardial infarction, history of hypertension, depression, physical activity level, and body mass index as prognostic factors for coronary artery disease in people in Mashhad.

Conclusion: Myocardial infarction and age were common prognostic factors for coronary artery disease according to the models applied. According to the efficiency of logistics model, binary multiple logistic regression model is suggested to be used in identifying the factors affecting coronary artery disease, if there is no interaction between the predictors.

Keywords: CART algorithm, Logistic regression, coronary artery disease, MASHAD STUDY

J Mazandaran Univ Med Sci 2021; 31 (195): 1-11 (Persian).

*** Corresponding Author: Nasser Behnampour -** Faculty of Health, Golestan University of Medical Sciences, Gorgan, Golestan, Iran (E-mail: behnampour@goums.ac.ir)

پژوهشی

مجله دانشگاه علوم پزشکی مازندران
دوره سی و یکم شماره ۱۹۵ فروردین سال ۱۴۰۰ (۱۱-۱)

مقایسه نتایج مدل رگرسیون لجستیک و الگوریتم CART در تعیین عوامل پیش‌آگهی دهنده ابتلا به بیماری عروق کرونر درشهر مشهد

زهرا بامی[۱]
ناصر بهنام پور[۲]
بهرام صادق پورگیلده[۳]
مجید غیورمبرهن[۴]
حبیب ا...اسماعیلی[۵]

چکیده

سابقه و هدف: درک عوامل خطر بیماری‌های قلبی- عروقی که مهم‌ترین علت مرگ در تمام دنیا است، می‌تواند تغییرات مهمی در روش‌های پیشگیری، اتیولوژی و درمان آن ایجاد نماید. هدف این مطالعه مقایسه عملکرد دو مدل رگرسیون لجستیک و الگوریتم CART در تعین عوامل پیش‌آگهی‌دهنده بر ابتلا به بیماری عروق کرونر در ساکنین شهر مشهد است.

مواد و روش‌ها: در این مطالعه مورد- شاهد از داده‌های مطالعه کوهورت (MASHAD STUDY: Mashhad Stroke and Heart Atherosclerotic Disorder) که در سال ۲۰۰۹، انجام شده بود، استفاده و عوامل پیش‌آگهی‌دهنده بر ابتلا به بیماری عروق کرونر با دو مدل رگرسیون لجستیک و الگوریتم CART، با نرم‌افزارهای ۱۴ Stata و R تعیین شد. کارایی دو مدل با سطح زیر منحنی مشخصه عملکرد (AUC) مقایسه شد.تمامی افراد مبتلا به بیماری عروق کرونر به عنوان مورد و به ازای هر مورد، سه شاهد در نظر گرفته شد.

یافته‌ها: رگرسیون لجستیک نشان داد سابقه سکته قلبی، ابتلا به دیابت، سابقه ابتلا به چربی خون، سن و سابقه بیماری عروق کرونر در پدر و برادر از عوامل پیش‌آگهی‌دهنده بر ابتلا به بیماری عروق کرونر در مشهد هستند. الگوریتم CART نیز، سن بالا، سابقه سکته قلبی، سابقه فشارخون، افسردگی، سطح فعالیت شبانه‌روزی و شاخص توده بدنی را به عنوان عوامل پیش‌آگهی‌دهنده تعیین کرد.

استنتاج: عوامل پیش‌آگهی‌دهنده مشترک حاصل از دو مدل، سابقه سکته قلبی و سن بود. با توجه به کارایی بهتر مدل لجستیک، می‌توان پیشنهاد کرد در صورت عدم وجود اثر متقابل در متغیرهای پیش بین، برای شناسایی عوامل موثر بر ابتلا به بیماری عروق کرونر از مدل رگرسیون لجستیک چندگانه باینری استفاده شود.

واژه های کلیدی: الگوریتم CART، رگرسیون لجستیک، بیماری عروق کرونر، MASHAD STUDY

مقدمه

زندگی انسان‌ها وابسته به عملکرد مناسب قلب است و اگر عملکرد قلب مناسب نباشد، روی سایر قسمت‌های بدن از قبیل ذهن، کلیه و غیره اثر خواهد گذاشت(۱). بیماری‌های قلبی- عروقی مهم‌ترین عامل

مولف مسئول: ناصر بهنام پور- گرگان: دانشگاه علوم پزشکی گلستان، دانشکده بهداشت E-mail: behnampour@goums.ac.ir
۱. دانشجوی کارشناسی ارشد آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی گلستان، گرگان، ایران
۲. استادیار، گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی گلستان، گرگان، ایران
۳. استاد، گروه آمار ریاضی، دانشکده علوم پایه، دانشگاه فردوسی مشهد، مشهد، ایران
۴. استاد، گروه تغذیه، دانشکده پزشکی، مرکز تحقیقات سندرم متابولیک، دانشگاه علوم پزشکی مشهد، مشهد، ایران
۵. استاد، گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، مرکز تحقیقات نوزادان، دانشگاه علوم پزشکی مشهد، مشهد، ایران
تاریخ دریافت : ۱۳۹۹/۲/۸ تاریخ ارجاع جهت اصلاحات : ۱۳۹۹/۲/۱۰ تاریخ تصویب : ۱۳۹۹/۱۰/۲۳

مرگ و میر در دنیا می‌باشند. در سال ۲۰۱۷ سازمان جهانی بهداشت اعلام کرد سالانه در دنیا ۱۲ میلیون مرگ بر اثر بیماری‌های قلبی رخ می‌دهد. بیش‌ترین علت مرگ و میر در ایران (۳۸ درصد) ناشی از ابتلا به بیماری قلبی و عروقی است و این میزان مرگ بر اثر بیماری قلبی رقم بسیار بالایی محسوب می‌شود(۲). رشد چشمگیر این بیماری و اثرات و عوارض آن و هزینه‌های بالایی که بر جامعه وارد می‌کند، باعث شده است که نظام سلامت در پی اجرای برنامه‌هایی برای پیشگیری، شناسایی زود هنگام و درمان موثر آن باشد(۳).

از طرف دیگر، حجم داده‌های پزشکی روزبه‌روز در حال افزایش است و پزشکان معمولاً اطلاعات ارزشمندی را در خصوص بیماری‌ها و ارتباط آن‌ها با دیگر عوامل ایجادکننده بیماری‌ها به دست می‌آورند(۴). اما این مجموعه داده‌های خام به خودی خود ارزشی ندارند و برای معنی بخشیدن به این داده‌ها باید آن‌ها را تحلیل و تبدیل به اطلاعات بهتر کرد(۵). با توجه به شیوع بیماری‌های قلبی- عروقی در سراسر جهان، استفاده از روش‌های جدید در تحقیقات زیست پزشکی بسیار مورد توجه قرار گرفته است. داده‌کاوی ابزاری است که برای حصول به چنین دانشی ما را یاری می‌کند. یکی از زمینه‌های پرکاربرد داده‌کاوی در علم پزشکی، استفاده از تکنیک‌هایی برای ایجاد مدل‌های پیشگویی‌کننده، جهت شناسایی افراد در معرض خطر برای کاهش عوارض ناشی از بیماری است(۶). در چند سال اخیر، درخت تصمیم و الگوریتم‌های متنوع آن، در مطالعات پزشکی مورد استفاده قرار گرفته است(۷). از آنجا که هدف این مطالعه شناسایی عوامل موثر در ایجاد بیماری است و مدل رگرسیون لجستیک برای چنین اهدافی کاربرد دارد، عوامل پیش‌آگهی‌دهنده بیماری بر اساس مدل رگرسیون لجستیک و الگوریتم CART شناسایی و کارایی دو روش بر اساس سطح زیر منحنی مشخصه عملکرد مقایسه شده است.

مواد و روش ها

داده‌ای این تحقیق مورد - شاهدی، حاصل مطالعه کوهورتی است که تحت عنوان MASHHAD-STUD نام‌گذاری و در سال ۲۰۰۹ در شهر مشهد آغاز و در سال ۲۰۱۶ با ویزیت همه افرادی که مشکوک به بیماری عروق کرونر بوده‌اند، پایان یافت. در مطالعه کوهورت مورد نظر در مجموع ۱۱۰۰۰ نفر ثبت نام کردند. پس از بررسی شرایط افراد بر اساس معیارهای ورود به مطالعه، تعداد ۹۷۶۱ نفر سالم تشخیص داده شده و در مطالعه باقی ماندند. سپس اطلاعات جمعیت شناختی، اندازه‌گیری‌های تن‌سنجی، داده‌های شیوه زندگی شامل عادات مصرف سیگار، سطح تحصیلات، پرسشنامه اضطراب بک (Beck Anxiety Inventory)، پرسشنامه افسردگی بک (Beck Depression Inventory) و همچنین پرسشنامه مربوط به عوامل موثر بر ابتلا به بیماری عروق کرونر برای این افراد، جمع‌آوری شد.

همه افراد باقی مانده در طرح، برای بررسی وضعیت قلب و عروق تحت کنترل قرار گرفتند. دو بار طی سال‌های ۲۰۱۴- ۲۰۱۱ از طریق تلفن و یک بار در فاصله زمانی ۲۰۱۶- ۲۰۱۵ پی‌گیری شدند. در این مرحله، از ۷۶۸ فردی که نشانه‌های Cardiovascular Disease (CVD) را بّه‌صورت خوداظهاری گزارش نمودند درخواست شد که با همراه داشتن اسناد پزشکی، در معاینه پزشکی شرکت کنند. پس از انجام معاینات لازم، تنها برای ۲۳۵ نفر از آن‌ها بیماری عروق کرونر مورد تایید قرار گرفت(۸). لذا در این مطالعه، حجم گروه بیمار شامل تمامی۲۳۵ نفری است که در مطالعه پیشگفت بیماری عروق کرونر مورد تایید قرار گرفته بود و به ازای هر بیمار، سه نفر به صورت تصادفی از بین ۹۴۸۲ فرد سالم که دارای اطلاعات ثبتی کامل بودند، انتخاب شد.

متغیرهای مورد بررسی

متغیرهای جنسیت، سن، مدرک تحصیلی، وضعیت شغلی، وضعیت تاهل، میزان فعالیت در شبانه روز، اندازه

دور کمر به دور باسن، نمره اضطراب، نمره افسردگی، سطح کلسترول، سطح HDL، سطح LDL، سطح تری‌گلیسرید، ابتلا به دیابت در حال حاضر، سابقه ابتلا به بیماری عروق کرونر در پدر، مادر، برادر، خواهر، سابقه ابتلا به فشارخون، پرفشاری خون، سابقه ابتلا به دیابت، افزایش چربی خون، سابقه پوکی استخوان، تغییر رژیم غذایی، سابقه سکته قلبی، مصرف سیگار، کاهش وزن در سه ماهه اخیر، شاخص توده بدنی، سابقه شکستگی استخوان به عنوان عوامل پیش‌بینی‌کننده بیماری عروق کرونر وارد مدل شد. پس از Merge کردن، داده‌ها، برای الگوریتم CART از پکیج RATTLE نرم افزار R3.5.3 استفاده شد(۹). همچنین متغیرهای مورد نظر با استفاده از مدل رگرسیون لجستیک نیز تجزیه و تحلیل شد. به منظور بررسی اثر هم‌زمان متغیرها و تعدیل اثر متغیرهای مخدوش‌کننده، رگرسیون لجستیک چندگانه در حضور همه متغیرها و به روش backward-selection در نرم‌افزار Stata 14 اجرا و نتایج بر اساس نسبت خطر تعدیل شده گزارش شد. همچنین سطح زیر منحنی مشخصه عملکرد برای سنجش میزان قدرت و دقت مدل محاسبه شد.

روش‌های تجزیه و تحلیل داده‌ها

داده کاوی

داده‌کاوی بدون داشتن فرضیه اولیه به کاوش و تحلیل داده‌ها می‌پردازد. روابط در داده‌کاوی معمولاً به صورت الگوها و مدل‌هایی از قبیل معادلات رگرسیونی، سری‌های زمانی، خوشه‌ها، رده‌بندی‌ها و گراف‌ها ارائه می‌شود. فرایند داده‌کاوی را می‌توان به سه مرحله، آماده‌سازی داده‌ها، یادگیری مدل و ارزیابی و تفسیر مدل تقسیم نمود. روش پیش‌بینانه از متداول‌ترین روش‌های یادگیری مدل در داده‌کاوی است. روش‌های طبقه‌بندی، رگرسیون و تشخیص انحراف و برخی از روش‌های دیگر یادگیری مدل، ماهیت پیش‌بینانه دارند(۱۰). در روش طبقه‌بندی مجموعه داده اولیه به دو مجموعه داده آموزشی(Train Dataset) و داده آزمایشی(Test Dataset) تقسیم می‌شوند. مجموعه داده‌های آموزشی برای ساختن مدل و مجموعه داده‌های آزمایشی برای اعتبارسنجی و محاسبه دقت مدل ساخته شده، استفاده می‌شود(۱۰).

برای تقسیم‌بندی، روش‌های مختلفی وجود دارد و دو روش Hold-out و K-fold Cross-validation از مهم‌ترین آن‌ها می‌باشد. در روش اول نسبت تقسیم مجموعه داده آزمایشی و داده آموزشی ۳۰ به ۷۰ است. در روش دوم کل مجموعه داده به ۵ یا ۱۰ یا k قسمت تقسیم می‌شود. یک قسمت برای مجموعه آزمایش و K-1 قسمت دیگر برای مجموعه آموزش مورد استفاده قرار می‌گیرد(۱۱).

الگوریتم CART

این روش که موجب تشکیل یک درخت تصمیم با تقسیمات دوتایی می‌شود توسط Breiman و همکاران در سال ۱۹۸۴ به‌طور کامل معرفی شد. این روش برای متغیرهای مستقل کمی طراحی شده است ولی قابل تعمیم و استفاده برای هر نوع متغیری می‌باشد(۱۱). در مدل CART هرس کردن درخت‌رده‌بندی بر اساس هزینه پیچیدگی (Cost-Complexity[۱]) انجام شده و دقت درخت معرفی شده به کمک نمونه آزمون بررسی می‌شود(۱۰). در پکیج Rattle از این معیار برای هرس کردن درخت استفاده می‌شود. روش تقسیم داده‌های آزمایشی و آموزشی در آن، 10-fold crossvalidation است(۹). بسته به نوع متغیر هدف، چندین اندازه ناخالصی (Impurity) برای پیدا کردن تقسیم‌گر در مدل CART وجود دارد. اگر متغیر هدف طبقه‌ای باشد شاخص جینی(Gini index) استفاده می‌شود. اگر متغیر هدف پیوسته باشد می‌توان روش کم‌ترین مربعات خطا (LSD) یا کم‌ترین قدر مطلق خطا (LAD) را استفاده کرد(۱۲). رگرسیون لجستیک: اگر متغیر پاسخ دو حالتی باشد، برای مدل‌سازی رابطه بین این متغیر و متغیرهای پیش‌بین می‌توان از رگرسیون لجستیک باینری استفاده

کرد.در مدل رگرسیون لجستیک باینری، تابع ربط لوجیت زیر مورد استفاده قرار می گیرد:

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log\pi - \log(1-\pi)$$

مدل رگرسیون لجستیک چندگانه (Multiple logistic regression model)، توانایی مدل‌بندی بین متغیر پاسخ دوحالتی و چندین متغیر پیش بین را دارد(۱۳).

یافته ها

از ۹۴۰ فرد مورد مطالعه، ۳۹۳ (۴۱/۸درصد) مرد و ۵۴۷ (۵۸/۲ درصد) زن بودند. محدوده سن افراد ۷۱-۳۳ سال بود. ۸۸۲ نفر (۹۳/۸ درصد) از آن‌ها متاهل، ۲نفر (۰/۲ درصد) مجرد، ۵۶ نفر (۶/۰ درصد) همسر از دست داده بودند (جدول شماره ۱).

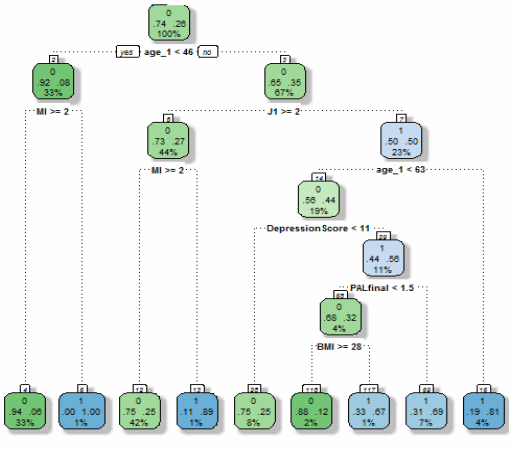
جدول شماره ۱: توزیع فراوانی صفات جمعیت شناختی افراد مورد بررسی به تفکیک مورد و شاهد

صفات جمعیت شناختی	گروه		تعداد (درصد)	سطح معنی داری
جنسیت	بیمار	مونث	۱۲۵ (۵۳/۲)	
		مذکر	۱۱۰ (۴۶/۸)	
	سالم	مونث	۴۲۲ (۵۹/۹)	۰/۰۷۳*
		مذکر	۲۸۳ (۴۰/۱)	
وضعیت تاهل	بیمار	متاهل	۲۱۶ (۹۱/۹)	
		همسر از دست داده	۱۸ (۷/۷)	
		مجرد	۱ (۰/۴)	
	سالم	متاهل	۶۶۵ (۹۴/۳)	۰/۳۰۶**
		همسر از دست داده	۳۸ (۵/۴)	
		مجرد	۲ (۰/۳)	
سطح تحصیلات	بیمار	بیسواد	۴۵ (۱۹/۱)	
		ابتدایی	۱۱۰ (۴۶/۸)	
		تا دیپلم	۵۹ (۲۵/۱)	
		دانشگاهی و حوزوی	۲۱ (۰/۹)	
	سالم	بیسواد	۹۴ (۱۳/۴)	۰/۰۰۳*
		ابتدایی	۲۷۴ (۳۶۹/۰)	
		تا دیپلم	۲۴۹ (۳۵/۴)	
		دانشگاهی و حوزوی	۸۶ (۱۲/۳)	

* کای- دو **: آزمون دقیق فیشر

توضیح و تشریح الگوریتم CART و مقادیر حاصل از آن:

ابتدا باید یادآور شد تعداد، احتمال و درصدهای ذکر شده در شکل فوق، برای افرادی است که اطلاعات آن‌ها در مورد کلیه متغیرهای باقی مانده در مدل درختی، کامل بوده است. در این مطالعه حجم نمونه شاهد سه برابر مورد بوده است، لذا در ابتدای مطالعه افراد شاهد ۰/۷۵ و افراد مورد ۰/۲۵ کل نمونه را تشکیل داده بودند، اما به دلیل عدم کامل بودن بعضی صفات برای نمونه ها، این نسبت در خروجی CART به ۰/۷۶ و ۰/۲۴ تغییر کرده است (تصویر شماره ۱).



تصویر شماره ۱: الگوریتم CART

نتایج حاصل از درخت رده‌بندی نشان می‌دهد که از میان تمام متغیرهای مورد بررسی، سن به عنوان مهم‌ترین عامل پیش‌بینی‌کننده ابتلا به بیماری عروق کرونر شناسایی و در ریشه قرار گرفته است. نقطه برش مدل برای سن ۴۶ سال می‌باشد. ۳۳ درصد افراد مورد بررسی کم‌تر از ۴۶ سال داشته‌اند، که در این افراد، نسبت شاهد ۰/۹۲ و مورد ۰/۰۸ بوده است، لذا می‌توان نتیجه گرفت که در سن کم‌تر از ۴۶ سالگی خطر ابتلا به بیماری عروق کرونر کم‌تر است. مدل نشان می‌دهد که در افراد زیر ۴۶ سال، سابقه سکته قلبی (MI) مهم بوده است، چون بلافاصله بعد از سن در مدل نمایان شده است. اگر این افراد سابقه سکته قلبی نداشته‌اند (MI<2)، نسبت شاهد ۰/۹۴ و مورد ۰/۰۶ بوده است، اما اگر سابقه سکته قلبی داشته‌اند (MI>=2)، نسبت شاهد ۰/۰ و مورد ۱/۰۰ بوده است. یعنی تمام کسانی که کم‌تر از ۴۶ ساله بوده و سابقه سکته قلبی داشته‌اند در گروه مورد قرار

گرفته‌اند. معنی آن این است که چنین افرادی به شدت در خطر ابتلا به بیماری عروق کرونر هستند. همچنین ۶۷ درصد افراد مورد بررسی در گروه سنی بزرگ‌تر یا مساوی ۴۶ قرار داشته‌اند. نتایج نشان می‌دهد که در این افراد سابقه فشارخون (J1) مهم بوده است، چون بلافاصله بعد از سن در مدل نمایان شده است، اگر این افراد سابقه فشارخون داشته (J1>=2) اما سن آن‌ها کم‌تر از ۶۳ سال باشد، نمره افسردگی اهمیت خواهد داشت و اگر نمره افسردگی این افراد بزرگ‌تر یا مساوی ۱۱ باشد، سطح فعالیت دارای اهمیت خواهد بود و اگر برای این افراد سطح فعالیتی بیش‌تر یا مساوی ۱/۵باشد، نسبت شاهد از ۰/۷۶ اولیه به ۰/۳۱ کاهش و نسبت بیمار از ۰/۲۴ اولیه به ۰/۶۹ افزایش خواهد یافت، که بیان‌کننده افزایش خطر ابتلا به بیماری عروق کرونری می‌باشد. همچنین نتایج حاصل از مدل رگرسیون لجستیک چندگانه نشان داد متغیرهای سابقه سکته قلبی، سابقه افزایش چربی خون، دیابت، سن و افرادی که پدر و برادرشان بیماری عروق کرونر دارند، از مهم‌ترین عوامل تاثیرگذار بر ابتلا به بیماری عروق کرونر می‌باشد(جدول شماره ۲). در این مطالعه، اصلی‌ترین عامل شناسایی شده براساس مدل لجستیک چندگانه، سابقه سکته قلبی با (OR =۵/۰۴۷) بوده است. همچنین دیابت (OR =۲/۲۷)، سابقه بیماری قلبی در برادر (OR =۱/۸۶۴)، سابقه بیماری قلبی در پدر (OR =۱/۸۴۶)، سابقه افزایش چربی خون (OR =۱/۸۱۰) و سن (OR =۱/۰۶۵)، از دیگر عوامل پر خطر در ابتلا به بیماری عروق کرونر در مدل لجستیک می‌باشد.

در مدل رگرسیون لجستیک، برای صفات طبقه بندی شده، برای محاسبه OR، عدم ابتلا به عنوان سطح رفرنس در نظر گرفته شده است.

یکی از روش‌هایی که بر اساس آن می‌توان دو مدل را با یکدیگر مقایسه کرد استفاده از سطح زیر منحنی عملکرد است. بدین منظور مقادیر به‌دست آمده از سطح زیر منحنی مشخصه عملکرد دو مدل رگرسیون لجستیک و الگوریتم CART در جدول شماره ۳ ارائه شده است. این مقادیر بیانگر آن است که مدل رگرسیون لجستیک در این مطالعه نسبت به مدل CART کارایی بیش تری دارد.

جدول شماره ۲: عوامل پیش آگهی دهنده بر ابتلا به بیماری عروق کرونر بر اساس مدل رگرسیون لجستیک چندگانه

متغیر	OR	95% CI	سطح معنی داری
سابقه سکته قلبی	۵/۰۴۷	۲/۴۴۷-۱۰/۴۱۱	<۰/۰۰۰۱
دیابت	۲/۹۸۴	۱/۳۰۷-۲/۸۴۶	۰/۰۰۹
افرادی که برادرشان بیماری عروق کرونر دارند	۲/۶۸۹	۱/۵۱۶-۴/۷۶۹	<۰/۰۰۱
افرادی که پدرشان بیماری عروق کرونر دارند	۱/۷۵۰	۱/۰۷۵-۲/۸۴۶	۰/۰۲۴
سابقه افزایش چربی خون	۱/۶۶۱	۱/۰۷۵-۲/۵۶۶	۰/۰۲۲
سن	۱/۰۷۹	۱/۰۴۹-۱/۱۱۰	<۰/۰۰۱

جدول شماره ۳: مقایسه سطح زیر منحنی عملکرد مدل رگرسیون لجستیک و الگوریتم CART

نام مدل	مساحت سطح زیر منحنی مشخصه عملکرد
رگرسیون لجستیک	۰/۸۲۸
الگوریتم CART	۰/۷۳۰

بحث

در این مطالعه مدل درخت رده‌بندی با الگوریتم CART و رگرسیون لجستیک چندگانه برای شناسایی عوامل پیش‌آگهی‌دهنده ابتلا به بیماری عروق کرونر استفاده شد. در هر دو مدل کلیه متغیرهای بیان شده در بخش روش بررسی وارد مدل شده و از میان آن‌ها بر اساس اجرای مدل CART، به ترتیب عوامل سن، سابقه افزایش فشارخون، سابقه سکته قلبی، نمره افسردگی، میزان فعالیت فرد در شبانه روز و شاخص توده‌بدنی و بر اساس مدل رگرسیون لجستیک چندگانه سابقه سکته قلبی، دیابت، وجود بیماری عروق کرونر در برادر و پدر، سابقه افزایش چربی خون و سن به عنوان مهم‌ترین عوامل پیش‌آگهی‌دهنده ابتلا به بیماری عروق کرونر معرفی شدند. این نتایج هم‌سو با دیگر تحقیقات انجام شده در جهان می‌باشد، به‌عنوان مثال، Karaolis و همکاران (۲۰۱۰) در پژوهشی تحت عنوان "ارزیابی خطر وقوع عروق کرونر قلب براساس داده‌کاوی" با بررسی متغیرهای سن، جنس، استعمال سیگار، فشارخون سیستولیک، سابقه خانوادگی CHD، سابقه فشارخون و

دیابت، سطح کلسترول، تری‌گلیسیرید و گلوکز به عنوان متغیر مستقل و نوع بیماری (انفارکتوس میوکارد (MI)، مداخله عروق کرونر از راه پوست (PCI) و جراحی پیوند بای پس عروق کرونر (CABG) به‌عنوان متغیر وابسته با اجرای الگوریتم C4.5 بر روی مشاهدات، نشان داد عواملی چون سن، جنس، استعمال سیگار، فشارخون و کلسترول، عوامل خطر برای این بیماری می‌باشند. مقایسه متغیرهای ورودی در مدل، بیان‌کننده آن است که در مطالعه حاضر عوامل سن، سابقه فشار خون، سابقه خانوادگی CHD، دیابت، سطح کلسترول، تری‌گلیسیرید با مطالعه Karaolis و همکاران مشترک بوده است، به طوری که عوامل خطر شناسایی شده مشترک دو مدل عبارت‌اند از: سن، فشارخون(۱۴). صالحی و همکاران (۲۰۱۸) در بررسی عوامل خطرزای بیماری عروق کرونر با استفاده از درخت تصمیم با مخاطره رقیب و بر اساس بر اساس ۱۴ فاکتور، جنسیت، سن، نمایه توده بدنی، مصرف فعلی سیگار، سابقه خانوادگی ابتلا به بیماری عروق کرونر، سابقه PCI قبلی، سابقه عمل جراحی پیوند بای پس عروق کرونر، وضعیت ابتلا به دیابت، سطح کراتین سرم خون، میزان قند خون ناشتا، درصد کسر تخلیه بطن چپ، ابتلا به هایپرلیپیدمی، ابتلا به فشارخون بالا و نوع استنت‌های مورد استفاده، نشان داد چهار عامل قندخون، وضعیت دیابت، شاخص توده بدنی و سن بیش‌ترین تاثیر را در بیماری عروق کرونر، داشته است. مقایسه متغیرهای ورودی در مدل، بیان کننده آن است که در مطالعه حاضر عوامل سن، سابقه خانوادگی ابتلا به بیماری عروق کرونر، وضعیت ابتلا به دیابت، ابتلا به هایپرلیپیدمی، ابتلا به فشارخون بالا، با مطالعه صالحی و همکاران مشترک بوده است، به‌طوری که عوامل خطر شناسایی شده مشترک دو مدل عبارت‌اند از: ابتلا به دیابت، سن، شاخص توده بدنی(۱۵).

محمدپور و همکاران (۲۰۱۱) در پژوهشی تحت عنوان "کاربرد شبکه عصبی مصنوعی جهت ارزیابی بیماری عروق کرونری قلب" با بررسی متغیرهای سن، شاخص توده بدنی،کراتینین، کلسترول تام، تری‌گلیسرید، سابقه مصرف سیگار، سابقه فشارخون، سابقه دیابت، سابقه بیماری عروق کرونر با اجراء مدل شبکه عصبی مصنوعی، مقادیر حساسیت و ویژگی را به ترتیب ۰/۹۶ و ۱ به‌دست آوردند. مقایسه متغیرهای ورودی در مدل، بیان کننده آن است که در مطالعه حاضر عوامل سن، کلسترول تام، تری‌گلیسرید، سابقه فشارخون، سابقه دیابت، سابقه بیماری عروق کرونر جزء عوامل مشترک با مطالعه محمدپور و همکاران است. با توجه به آن‌که خروجی مدل شبکه عصبی فاقد متغیرهای شناسایی شده به عنوان عوامل خطر می‌باشد، لذا مقایسه مدل‌های درختی و لجستیک چندگانه با مدل شبکه عصبی مفید و کاربردی نمی‌باشد. اما اگر همه مدل‌ها ورودی یکسانی داشته باشند، مقایسه سطح زیر منحنی مشخصه عملکرد (AUC) می‌تواند در شناسایی مدل کاراتر منطقی و مفید باشد. با توجه به موضوعات بیان‌شده تنها به بیان مقدار سطح زیر منحنی مدل لجستیک و مقدار حساسیت و ویژگی مدل درختی می‌پردازیم. در این مطالعه سطح زیر منحنی مدل لجستیک ۰/۸۲۸ و برای الگوریتم CART، ۰/۷۳ بوده است(۱۶). صباغ گل (۲۰۱۷) در بررسی تشخیص بیماری عروق کرونر با استفاده از درخت تصمیم c4.5 و بر اساس ۱۳ عامل بیماری، سن، جنسیت، درد قفسه‌سینه، فشارخون در زمان استراحت، کلسترول، قند خون ناشتا، نتایج نوار قلب در حال استراحت، ماکزیمم ضربان قلب، آنژین ناشی از ورزش، افسردگی ایجاد شده st در هنگام تست ورزش، شیب قطعه st در زمان ورزش، تعداد رگ‌ها در فلورسکوپی، اسکن تالیوم و یک متغیر وابسته بنام وجود بیماری عروق کرونر نشان داد سطح بالای کلسترول، جنسیت، سن بالا، بالا بودن ماکزیمم ضربان قلب، اسکن تالیوم بالاتر از ۳، نوار قلب غیرنرمال بیش‌ترین تاثیر را در ابتلا به بیماری عروق کرونر داشته است. مقایسه متغیرهای ورودی در مدل، بیان‌کننده آن است که در مطالعه حاضر عوامل سن، کلسترول، فشار خون با مطالعه صباغ

گـل مـشترک بـوده اسـت، بـه طـوری کـه عامـل خطـر شناسایی شده مشترک دو مدل عبارت از: سن اسـت(۶). Kurt و همکـاران (۲۰۰۸) در پژوهـشی تحـت عنـوان "مقایسه عملکرد رگرسیون لجستیک، درخت رده‌بندی، رگرسیون درختی، و شبکه عصبی بر پـیش‌بینـی بیمـاری عروق کرونر" پس از مقایسه عملکرد مدل‌ها، نـشان داد مدل شبکه عصبی با صـحت ۰/۷۸ درصـد بهتـرین مـدل است و متغیرهای سن، جنس، سابقه بیماری عـروق کرونـر در خانواد، وضعیت سیگار کشیدن، دیابـت، فـشارخون، کلـسترول، شـاخص تـوده بـدنی، شـاخص‌هـای قابـل اعتمادی در پیش‌بینی بیماری عروق کرونر است. مقایسه متغیرهای ورودی در مدل، بیان کننـده آن اسـت کـه در مطالعه حاضر عوامل سن، سابقه بیماری عروق کرونر در خانواد، دیابت، فشارخون از عوامل مشترک است(۱۷).

Mobley و همکـاران (۲۰۰۰) در پژوهـشی تحـت عنوان "پیش‌بینی تنگی شریان عروق کرونری بـه وسـیله شبکه عصبی مصنوعی" نشان داد که بـا توجـه بـه تعـداد داده‌هـای ورودی یـا عوامـل خطـر در بیمـاری عـروق کرونر، شبکه عصبی مصنوعی می‌تواند تبدیل به ابـزاری با ارزش در شناسایی بیمارانی که نیـازی بـه آنژیـوگرافی عروق کرونر ندارند، باشد(۱۸).

Soni و همکاران (۲۰۱۱) در پژوهشی تحت عنـوان "پیش‌بینی مبتنـی بـر داده کـاوی در تـشخیص پزشـکی: یک مطالعه مـروری در پـیش‌بینـی بیمـاری قلبـی" نـشان دادند که مدل ارائه شده توسط درخت تصمیم نسبت بـه مـدل بیـز و خوشـه‌بنـدی، دارای صـحت بـالاتری اسـت (۹۹/۲ نسبت بـه ۹۶/۵و ۸۸/۳) صـحت مـدل CART در این مطالعه ۰/۷۵ است(۱۹).

نشاطی تنها و همکاران (۲۰۱۵) در پژوهـشی تحـت عنوان" پـیش‌بینـی بـروز حملـهٔ حـاد قلبـی بـا اسـتفاده از رگرسیون لجستیک " و بر اساس ۲۸ فاکتور بالینی، سن، جنسیت، سیگار، سابقه سـکته قلبـی، دیابـت، فـشارخون، چربی خـون، درد شـدید قفـسه سـینه، درد سـمت چـپ قفسه سینه، درد سمت راست قفسه سینه، درد پشت، درد دست چپ، درد دست راست، تعریق سرد، تنگی تنفس، حالـت تهـوع، اسـتفراغ، بـی‌هوشـی، تـپش قلـب، درد فوق‌المعدی، سابقه بیماری عروق کرونـر، ورم انـدام‌هـا، خـواب آلـودگی، دلـشوره، سـردرد بیـان کـرد در مـدل رگرسیون لجستیک به روش Enter متغیرهای درد شدید قفسه سینه، درد پشت، تعریق سرد، تنگـی نفـس، حالـت تهوع و اسـتفراغ، و در مـدل درخـت تـصمیم بـر اسـاس الگوریتم CHAID، تنگی نفس، تپش قلب، ورم اندام‌ها، تعریق سرد، درد سمت چپ قفسه سینه، درد شدید قفسه سینه، سن بیش‌ترین تاثیر را در بروز حمله حاد قلبـی داشـته است. مقایسه متغیرهای ورودی در مدل، بیـان‌کننـده آن است که در مطالعه حاضر عوامل سن، سابقه سکته قلبی، دیابت، فشار خون، چربـی خـون، سـابقه بیمـاری عـروق کرونر با مطالعه نشاطی و همکاران مشترک بـوده اسـت، به‌طوریکه تنها عامـل خطـر شناسـایی شـده مـشترک دو مدل عبارت است از: سن(۲۰).

Tsien و همکاران (۱۹۹۸) در پژوهشی تحـت عنـوان "اسـتفاده از درخـت تـصمیم و رگرسـیون لجـستیک در تشخیص سکته قلبی"و بر اساس ۴۵ متغیـر ورودی ماننـد سن، جنس، سـابقه سـکته قلبـی در خـانواده، درد سـمت چپ قفسه سـینه، درد سـمت راسـت قفـسه سـینه، تنگـی نفس،... و با برازش مدل رگرسیون لجستیک و الگوریتم C4.5 دریافت که درخت طبقه‌بندی با ۸۱ درصد صحت، عملکرد بهتری دارد، براساس شـاخص FT متغیرهـای مهـم مدل عبارت اسـت از: ارتفـاع جدیـد st، مـوج Q، طـول مدت، موج T، درد بازوی راسـت، سـن، افـسردگی St، ایسکمی قبلی، سابقه خانوادگی بیمـاری عـروق کرونـر، Crackles اسـت. مقایـسه متغیرهـای ورودی در مـدل، بیان‌کننده آن اسـت کـه در مطالعـه حاضـر عوامـل سـن، سابقه سکته قلبی در خانواده با مطالعه Tsien و همکاران مشترک بوده است، به‌طوری که عوامل خطر شناسایی شده مشترک دو مدل عبارت اسـت از: سـن و سـابقه بیمـاری عروق کرونر در خانواده(۲۱). صفدری و همکاران (۲۰۱۴) در پژوهشی تحت عنوان «درخت تصمیم و شبکه عصبی

در پیشگویی ابتلا به انفارکتوس قلبی» نشان داد که فاکتورهای فشارخون بالا، کلسترول بالا، سن بالا و مصرف سیگار بیش ترین تاثیر را در ابتلا به انفارکتوس قلبی دارند. صحت مدل ایجاد شده با استفاده از درخت تصمیم ۹۳/۴ گزارش شد. در این مطالعه، فاکتورهای بالینی وارد شده در مدل، بیان نشده است. عوامل خطر شناسایی شده مشترک دو مدل عبارت است از: فشارخون بالا، کلسترول بالا و سن(۴).Purusothaman و Krishnakumari (۲۰۱۵) درپژوهشی تحت‌عنوان "بررسی تکنیک‌های داده‌کاوی در پیش‌بینی خطر بیماری قلبی" با مرور مقالات مرتبط با بیماری قلبی که بر روی مشاهدات مختلفی صورت گرفته بود، دقت رویکرد ترکیبی را (۹۶ درصد) گزارش کرد، هدف این مطالعه، کمک به یافتن بهترین مدل برای مطالعات دیگر است(۲۲).

Zabah و همکاران (۲۰۱۸) در پژوهشی تحت عنوان "تشخیص بیماری عروق کرونر با استفاده از روش مبتنی بر عصبی فازی" و با بررسی متغیرهای اصلی و زمینه‌ای شامل جنسیت، مصرف سیگار، فشارخون بالا، دیابت، سابقه خانوادگی بیماری قلبی، سابقه سکته قلبی، نتیجه تست ورزش، نتیجه اکو، سن، کراتینین، کلسترول، تری گلیسیرید و اجراء الگوریتم‌های شبکه عصبی و منطق فازی نشان داد که اگر چه شبکه‌های عصبی به تنهایی می‌توانند به پیش‌بینی بیماری کرونری قلب بپردازند، اما با ترکیب شبکه عصبی و منطق فازی و تشکیل شبکه نرو فازی دقت محاسبات به طور قابل توجهی افزایش می‌یابد(۲۳). لذا با توجه به مقالات مروری فوق نتایج زیر قابل تامل است.

مقایسه مطالعات بر اساس مدل

در پژوهش نشاطی و همکاران(۲۰) و در مقایسه مدل لجستیک و الگوریتم CHAID، مدل لجستیک مطلوب شناخته شد. در پژوهش Tsien و همکاران(۲۱) و در مقایسه مدل لجستیک و درخت تصمیم، مدل درخت تصمیم پیشنهاد شد. در پژوهش صفدری و همکاران(۴) و در مقایسه مدل شبکه عصبی و درخت تصمیم، مدل درخت تصمیم پیشنهاد شد. در پژوهش Soni و همکاران(۱۹) و در مقایسه مدل بیز، درخت تصمیم و خوشه‌بندی، مدل درخت تصمیم پیشنهاد شد. در پژوهش صباغ(۶) و کارولیس(۱۴) استفاده از الگوریتم C4.5 مورد توجه قرار گرفت. در پژوهش صالحی و همکاران (۲۰۱۸) استفاده از درخت مخاطره رقیب مورد استفاده قرار گرفت. در پژوهش Kurt و همکاران(۱۷) و در مقایسه مدل لجستیک، درخت تصمیم و شبکه عصبی، مدل شبکه عصبی پیشنهاد شد. در پژوهش Zabah و همکاران(۲۳) و در ترکیب شبکه عصبی و منطق فازی، شبکه نروفازی پیشنهاد شد. در پژوهش Purusothaman و همکاران(۲۲) و در مقایسه مدل درخت تصمیم، قواعد انجمنی، رویکرد ترکیبی، شبکه بیزی، ماشین بردار پشتیان و شبکه عصبی، رویکرد ترکیبی، پیشنهاد شد. در پژوهش محمدپور و همکاران(۱۶) تاکید به‌سزایی بر روی انتخاب متغیرهای ورودی در مدل شد. در پژوهش Mobley و همکاران(۱۸) مدل شبکه عصبی برای شناسایی بیماران قلبی که نیاز به آنژیوگرافی ندارند مناسب شناخته شد.

مقایسه مطالعات بر اساس یافته ها

در مطالعات آماری زمانی قابلیت مقایسه مدل‌ها با یکدیگر وجود دارد که متغیرهای ورودی یکسانی وارد مطالعه شود، اما در مطالعات فوق تنها در بعضی موارد، تعدادی کمی از متغیرهای ورودی یکسان است. لذا مقایسه بین این مطالعات می‌تواند منجر به نتایج گمراه کننده‌ای شود. اما آن‌چه که در نیمی از مطالعات مشاهده شد، وجود عوامل مشترکی چون دیابت، سن، کلسترول، فشارخون, استعمال سیگار، سابقه بیماری عروق کرونر در خانواده است.

در این مطالعه، اصلی‌ترین عامل شناسایی شده بر اساس مدل لجستیک چندگانه، سابقه سکته قلبی بوده

است (OR =۵/۰۴۷). همچنین عواملی چون دیابت (OR =۲/۹۸۴)، ابتلا به بیماری عروق کرونر در برادر (OR =۲/۶۸۹)، ابتلا به بیماری عروق کرونر در پدر (OR =۱/۷۵۰)، افزایش چربی خون (OR= ۱/۶۶۱) و سن (OR =۱/۰۷۹)، از دیگر عوامل پر خطر در ابتلا به عروق کرونر در مدل لجستیک چندگانه می‌باشد.

در مدل CART نیز، اولین عامل سن با نقطه برش ۴۶ و ۶۳تشخیص داده شد. پس از آن، سابقه افزایش فشار خون، سابقه سکته قلبی، نمره افسردگی، سطح فعالیت فرد و شاخص BMI از مهم‌ترین عوامل خطر در ابتلا به بیماری عر وق کرونر تشخیص داده شد.

نتایج این پژوهش نشان داد هرچند استنباط بر اساس الگوریتم CART بسیار ساده‌تر و تفسیر آن قابل درک‌تر است، اما اگر سطح زیرمنحنی عملکرد (AUC)، ملاک تصمیم‌گیری باشد، مدل رگرسیون لجستیک دارای کارایی بیش‌تری می‌باشد. لذا می‌توان از میان این دو روش، مدل رگرسیون لجستیک چندگانه برای شناسایی عوامل پیش آگهی‌دهنده مورد استفاده قرار گیرد. همچنین با توجه به این که طی چند دهه اخیر، بیماری‌های قلبی- عروقی از جمله شایع‌ترین علل مرگ و میر در جوامع توسعه یافته و یا در حال توسعه بوده است، لذا با شناسایی عوامل خطر، امکان اجرای برنامه‌های غربالگری هدفمند فراهم شده به طوری که می‌توان انتظار داشت کارایی این برنامه‌ها برای پیشگیری از ابتلا به بیماری عروق کرونر قابل توجه باشد.

سپاسگزاری

این مقاله قسمتی از پایان نامه کارشناسی ارشد آمار زیستی می‌باشد و نویسندگان مقاله از دانشگاه‌های علوم پزشکی گلستان و مشهد تشکر و قدردانی می‌نمایند. کد اخلاق این مطالعه IR.GOUMS.REC.1398.256 می‌باشد.

References

1. Chaitrali D, Sulabha A. A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering and Technology 2012; 3(3): 30-40.
2. Mazaheri S, Ashoori M, Bechari Z. A Model to Predict Heart Disease Treatment Using Data Mining. Journal of Payavard Salamat 2017; 11(3): 287-296 (Persian).
3. International Conference on Researches in Science & Engineering 31 Aug. 2017. (Persian).
4. Safdari R, Ghazi saeedi M, Gharooni M, Nasiri M, Arji G. Comparing performance of decision tree and neural network in predicting myocardial infarction. Journal of Paramedical Sciences & Rehabilitation 2014; 3(2): 26-37 (Persian)
5. Subbalakshmi G, Ramesh K, Rao MC. Decision support in heart disease prediction system using naive bayes. Indian Journal of Computer Science and Engineering (IJCSE). 2011; 2(2): 170-176.
6. SABBAGH GH. Detection of coronary artery disease using C4. 5 decision tree. Journal of Healh and Biomedical Informaics 2017; 3(4): 284-299 (Persian).
7. Aviva P, Caroline S. Medical statistics at a glance. 3rd ed: Blackwell Science Ltd, Padstow, UK; 2000.
8. Sadabadi F. Investigation of the association between heat shock protein 27 gene and antibody variations with new coronary event mortality and morbidity in an Iranian cohort. PHD of molecular medicine thesis in Medicen School, Mashhad University of Medical Sciences, 2019.

9. Williams G. Data mining with Rattle and R: The art of excavating data for knowledge discovery: Springer Science & Business Media; 2011.
10. Mehrbakhsh Z. Comparison of logistic regression and classification models to determine the factors affecting the incidence of esophageal cancer in Golestan. Master of Science thesis in Health School, Mashhad University of Medical Sciences; 2015.
11. Breiman L. Classification and regression trees: Routledge; 2017.
12. Azimi M. Data Mining With Decision Tree. Master of Mathematical Sattistics thesis, Tabriz, University Of Tabriz; 2012.
13. Agresti A. An Introduction to Categorical Data Analysis 2007.
14. Karaolis M, Moutiris JA, Pattichis CS, editors. Assessment of the risk of coronary heart event based on data mining. 2008 8th IEEE International Conference on BioInformatics and BioEngineering; 2008: IEEE.
15. Salehi E, Hagizadeh E, Alidoosti M. Evaluation Risk Factors of Coronary Artery Disease Through Competing Risk Tree. J Arak Uni Med Sci 2018; 21(4): 18-29 (Persian).
16. Mohammadpour Tahamtan R, A, Esmaeili M, H, Ghaemian A, Esmaeili J. Application of Artificial Neural Network for Assessing Coronary Artery Disease. J Mazandaran Univ Med Sci 2012; 22(86): 9-17 (Persian).
17. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert systems with applications 2008; 34(1): 366-374.
18. Mobley BA, Schechter E, Moore WE, McKee PA, Eichner JE. Predictions of coronary artery stenosis by artificial neural network. Artif Intell Med 2000; 18(3): 187-203.
19. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications 2011; 17(8): 43-48.
20. Neshati Tanha A, Soleimani P. Prediction of Acute Heart Attack using Logistic Regression (Case Study: A Hospital in Iran). Advances in Industrial Engineering 2016; 50(11): 109-119.
21. Tsien CL, Fraser H, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. Medinfo. 1998; 98.
22. Purusothaman G, Krishnakumari P. A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology 2015; 8(12): 1.
23. Zabah I, Koohjani Z, Maroosi A, Layeghi K. Diagnosis of Coronary Artery Disease using Neuro-fuzzy-based method. Journal of Torbat Heydariyeh University of Medical Sciences 2018; 6(3): 48-59 (Persian).