مقاله پژوهشی
Original Article

کاربرد و مقایسهٔ روش‌های یادگیری ماشینی رندم فارست و درخت کلاس‌بندی- رگرسیونی در مطالعهٔ وابستگی ژنتیکی در بیماران مبتلا به شریان‌های کرونری

آتوسا مددکار[۱]، *مسعود کریملو[۲]، مهدی رهگذر[۲]، سید حمید جمال‌الدینی[۲]، رضا مظفری[۳]

Application and Comparison of Random Forest and CART in Genetic Association Study in Coronary Artery Disease

Madadkar A.[1], Karimlo M.[2], RahgozarM.[2], Jamaldini SH.[2], Mozaffari R[3]

مجله مطالعات ناتوانی، دوره ۳، شماره ۲، پاییز ۱۳۹۲، ۸-۱

MEJDS, Vol 3, No. 2, Autumn 2013, 1-8

پذیرش: ۱۳۹۲/۰۵/۱۵

دریافت: ۱۳۹۲/۰۴/۱۳

Accepted: 8/6/2013

Received: 7/4/2013

چکیده

هدف: در مطالعات ژنتیک، برگزیدن تعداد معدودی از پلی‌مورفیسم تک‌نوکلئوتیدی که نسبت به سایر پلی‌مورفیسم‌ها ارتباط بیشتری با بیماری مورد نظر دارند، امری ضروری است. این تحقیق با هدف مقایسهٔ دو روش یادگیری ماشینی مبتنی بر درخت رندم فارست و درخت کلاس‌بندی- رگرسیونی، به‌منظور پیش‌بینی افراد مبتلا به بیماری شریان‌های کرونر با استفاده از پلی‌مورفیسم تک نوکلئوتیدی انجام شد.

روش‌بررسی: به‌منظور انجام این پژوهش، از مراجعین بیمارستان قلب و شریان‌های شهید رجایی تهران در فاصله سال‌های ۱۳۹۰ و ۱۳۹۱ تمام‌شماری به‌عمل آمد. اطلاعات ۱۴۱ فرد مبتلا به بیماری شریان‌های کرونر و ۸۳ شاهد توسط مرکز تحقیقات ژنتیک این بیمارستان گردآوری شد. از طریق روش تعیین توالی، ژنوتیپ پلی‌مورفیسم تک‌نوکلئوتیدی انتخاب‌شده در جایگاه ژن **LDLR** و پروموتر ژن **PCSK9** هر فرد مشخص شد. همچنین به‌منظور کشف ارتباط میان پلی‌مورفیسم تک‌نوکلئوتیدی و بیماری شریان‌های کرونر، از دو روش یادگیری ماشینی: درخت کلاس‌بندی- رگرسیونی و رندم فارست استفاده شد. اعتبارسنجی مدل بر اساس اعتبارسنجی ده قسمتی انجام شد. مدل‌ها پس از برازش، با استفاده از سه ملاک حساسیت و ویژگی و خطا ارزیابی شدند. تحلیل داده‌ها با استفاده از نرم افزار **R(2.15.0)** صورت گرفت.

یافته‌ها: با توجه به ملاک‌های معرفی‌شده، درخت کلاس‌بندی- رگرسیونی نسبت به روش رندم فارست، عملکرد موفق‌تری داشت. این روش حساسیت و ویژگی و خطایی معادل ۰/۵۷۵ و ۰/۶۸۴ و ۰/۲۶۴ داشت. همچنین مدلی به‌منظور کلاس‌بندی مراجعین جدید ارائه شد.

نتیجه‌گیری: هر چند بر اساس ملاک‌های مقایسه‌شده دو روش تفاوت چندانی با یکدیگر نداشته و در نهایت هر دو روش در تحلیل پلی‌مورفیسم‌های تک‌نوکلئوتیدی توصیه می‌شوند. اما روش درخت کلاس‌بندی- رگرسیونی به‌عنوان اولین انتخاب پژوهشگر به‌منظور کشف ارتباط میان پلی‌مورفیسم و بیماری مورد نظر توصیه می‌شود.

کلیدواژه‌ها: پلی‌مورفیسم تک‌نوکلئوتیدی، بیماری شریان‌های کرونر، روش‌های یادگیری ماشینی، اثر متقابل.

Abstract

Objective: In the studies of genomics, it is essential to select a small number of single nucleotide polymorphisms (SNPs). That is more significant than the others for the association studies of disease susceptibility. Data mining technology provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in disease prediction and clinical diagnosis. In this study, our goal was to compare two machine learning methods using genetic factor and single nucleotide polymorphisms.

Methods: In order to perform the data analysis, a total of 141 patients and 83 controls in the genetics' section of Shahid Rajaee's heart center. The blood samples to draw conclusions about the LDLR and PCSK9 genes' SNPs was used. Also, the random forest and CART was used in order to discover the relationship between CAD and SNPs. These models were assessed by using four criteria including: sensitivity, specificity, precision and error. Data analysis was performed by SPSS (16.0) and R (2.15.0).

Results: CART had the better performance than Random Forest. Sensitivity, specificity, precision and error were 0.893, 0.506, 0.250 and 0.754 relatively. We introduced an algorithm to classify the high risk and low risk cases.

Conclusion:CART is suggested in order to assess the relationship between CAD and SNPs.

Keywords: SNPs, CAD, Machine Learning, Interaction

۱. دانشجوی کارشناسی ارشد آمار زیستی، گروه آمار و انفورماتیک، دانشگاه علوم بهزیستی و توانبخشی، تهران، ایران؛ ۲. دکترای آمار زیستی، دانشیار گروه آمار زیستی، دانشگاه علوم بهزیستی و توانبخشی، تهران، ایران؛ ۳. مرکز تحقیقات ژنتیک بیماری‌های قلب، مرکز آموزشی تحقیقاتی و درمانی قلب و عروق شهید رجایی، دانشگاه علوم پزشکی تهران، تهران، ایران. ***آدرس نویسنده مسئول:** تهران، اوین، بلوار دانشجو، بن‌بست کودکیار، دانشگاه علوم بهزیستی و توانبخشی، گروه آمار زیستی، *تلفن: ۰۲۱۲۲۲۱۸۰۱۴۶، *رایانامه: mkarimlo@yahoo.com

1. Master student of Biostatistics, Department of Biostatistics, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran; 2. Biostatistician, Associate Professor, Department of Biostatistics, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran; 3. Cardio genetics research center. Rajaie cardiovascular medical and research center, Tehran University of medical science, Tehran, Iran. ***Corresponding author's address:** Biostatistics department, University of Social Welfare and Rehabilitation Sciences, Kodakyar St., Daneshjoo Blv., Evin, Tehran, Iran. *Tel: +98(21) 22180146 *E-mail: mkarimlo@yahoo.com

مقدمه

پلی‌مورفیسم‌های تک‌نوکلئوتیدی، گوناگونی ژنتیکی در طول ژنوم هستند که حداقل در یک درصد از جمعیت رخ می‌دهند (۱-۳). بسیاری از پلی‌مورفیسم‌های تک‌نوکلئوتیدی تأثیری بر عملکرد سلول ندارند؛ اما برخی از آن‌ها می‌توانند استعداد ژنتیکی ابتلا به بیماری یا پاسخ به درمان فرد را تحت تأثیر قرار دهند. از همین رو، پلی‌مورفیسم‌های تک‌نوکلئوتیدی در مطالعات پزشکی و تشخیص بیماری ارزش بسیار زیادی دارند. پژوهشگران معتقدند که نقشه‌های پلی‌مورفیسم تک‌نوکلئوتیدی در یافتن ژن‌های مؤثر در ابتلا به بیماری‌های پیچیده‌ای از جمله: بیماری‌های قلبی-عروقی و دیابت و سرطان‌ها کمک شایانی خواهد کرد (۴). هنگام بررسی همزمان چندین پلی‌مورفیسم تک‌نوکلئوتیدی در یک بیماری شایع، معمولاً نقش هر پلی‌مورفیسم تک‌نوکلئوتیدی آنقدر کوچک است که این تأثیر با روش‌های کلاسیک آماری قابل تشخیص نیست (۵). فرض بر این است که پلی‌مورفیسم‌های تک‌نوکلئوتیدی، ریسک ابتلا به یک بیماری را تغییر می‌دهند؛ اما بسیار بعید است که تنها یک پلی‌مورفیسم تک‌نوکلئوتیدی مسئول بروز این بیماری باشد؛ لذا گفته‌شده که تعدادی پلی‌مورفیسم تک‌نکلئوتیدی، به‌صورت همزمان تغییرات مرتبط با خطر ابتلا به بیماری را تبیین می‌کنند. از طرف دیگر بسیاری از این اثرات به‌صورت ضربی با بیماری مورد نظر ارتباط دارند (اثر متقابل هر پلی‌مورفیسم با پلی‌مورفیسم دیگر یا اثر متقابل هر پلی‌مورفیسم با عوامل محیطی) و این امر تشخیص پلی‌مورفیسم‌های مرتبط با بیماری و در نتیجه تعیین پیش‌آگاهی ابتلا به بیماری را مشکل می‌سازد.

روش‌های یادگیری ماشینی با استفاده از اطلاعات مجموعهٔ آموزشی یا یادگیری، پیش‌بینی‌های آینده را بر اساس الگو یا قواعد یاد گرفته‌شده انجام می‌دهند. این روش‌ها در دههٔ اخیر، یکی از روش‌های مرجع در تحلیل پلی‌مورفیسم‌های تک‌نوکلئوتیدی به شمار می‌روند. از طرفی روش‌های یادگیری ماشینی بر مبنای درخت نه تنها تفسیر ساده‌ای دارند، توانایی تحلیل همزمان پلی‌مورفیسم‌های تک نکلئوتیدی را نیز دارند. در این مطالعه از داده‌های مرتبط با پلی‌مورفیسم‌های تک‌نکلئوتیدی در پیش‌بینی ابتلا به بیماری شریان‌های کرونری استفاده شد.

بیماری شریان‌های کرونری یکی از شایع‌ترین علت‌های مرگ در نیمکره غربی بوده (۶) و انتظار می‌رود که تا سال ۲۰۲۰ عمده‌ترین علت مرگ در جهان باشد (۷). این اختلال، بیماری پیچیده‌ای است که به علت عوامل ژنتیکی و محیطی متعدد و اثرات متقابل پیچیده میان ژن و محیط ایجاد می‌گردد (۸, ۹). تنها در ایالات متحده این بیماری سالانه مسئول ۱,۵ میلیون مورد جدید سکته قلبی، ۳۵۰۰۰۰ مورد جدید نارسایی قلبی و ۵۰۰۰۰۰ مرگ و ناتوانی است (۶).

ژن *LDLR* یکی از مهم‌ترین ژن‌های دخیل در افزایش LDL و بیماری شریان‌های کرونر است (۱۰-۱۴). همچنین طی مطالعات گوناگون ثابت شده که ژن *PCSK9* یکی از علل ابتلا به بیماری شریان‌های کرونر است (۱۵). در این مطالعه ارتباط پلی‌مورفیسم‌های تک‌نکلئوتیدی در دو ژن *LDLR* و *PCSK9* با بیماری شریان‌های کرونری مورد بررسی قرار گرفت. هدف مطالعه، تشخیص پلی‌مورفیسم‌های تک‌نوکلئوتیدی تاثیرگذار با استفاده از الگوریتم‌های رندم فارست و درخت کلاس‌بندی-رگرسیونی بوده؛ در ضمن این دو الگوریتم در نحوه عملکرد پیش‌بینی خود مقایسه گردیدند.

روش بررسی

داده‌های این مطالعه حاصل تمام شماری مراجعین بیمارستان قلب شهید رجایی تهران در فاصله دو سال ۱۳۹۰ و ۱۳۹۱ بوده. با توجه به معیارهای ورود و خروج و همچنین وجود گم‌شدگی در داده، ۲۲۴ فرد در مطالعه اصلی وارد شدند. به منظور بررسی میزان گرفتگی شریان‌های، از تمامی ۲۲۴ فرد آنژیوگرافی به عمل آمد و میزان گرفتگی شریان‌های در هر فرد ثبت گردید. گرفتگی شریان‌های بیش از ۵۰ درصد معادل با بیماری شریان‌های کرونر در فرد محسوب گردید. بنابراین از این تعداد، ۸۳ فرد، کنترل یا شاهد محسوب شده و ۱۴۱ فرد دیگر مورد یا مبتلا به بیماری شریان‌های کرونر بودند. با اخذ رضایت نامه آگاهانه

کتبی از داوطلبین شرکت‌کننده در مطالعه، نمونه خون بیماران جهت انجام آزمایش‌های ژنتیکی گرفته شد. پس از استخراج DNA و با استفاده از تکنیک تعیین و توالی یابی، پلی‌مورفیسم‌های مربوط به دو ژن *LDLR* و *PCSK9* مورد بررسی قرار گرفت. متغیر وجود بیماری شریان‌های کرونری در فرد، متغیر وابسته محسوب شده و پلی‌مورفیسم‌های تک‌نکلئوتیدی، جنس، سن، وزن، قد، فشارخون، تری گلیسیرید کل، کلسترول، LDL، HDL به عنوان متغیرهای مستقل یا پیش‌بین بالقوه در نظر گرفته شدند. تحلیل‌های این مطالعه با استفاده از نرم‌افزار R نسخه 2.15.2 انجام شده و سطح معناداری ۰,۰۵ در نظر گرفته شد.

برخی پلی‌مورفیسم‌های تک‌نکلئوتیدی دارای مقدار گمشده بودند، اما نرخ گم‌شدگی در هر پلی‌مورفیسم کمتر از ۲۰ درصد بود. از آن جایی که دو روش به کار برده شده در این مطالعه، در صورت وجود داده گمشده، قابلیت اجرا شدن ندارند، قبل از انجام تحلیل، مقادیر گمشده با استفاده از تابع rfImp در بسته RandomForest در نرم‌افزار R جانهی شده‌اند. در مجموع ۳۵ پلی‌مورفیسم تک‌نکلئوتیدی به منظور تحلیل چند متغیره در نظر گرفته شد. فهرست این ۳۵ پلی‌مورفیسم در جدول ۱ آمده است. هر پلی‌مورفیسم دارای سه سطح است: هموزیگوت شایع (نوع وحشی AA,) ، هتروزیگوت (Aa)، هموزیگوت واریانت (aa)؛ در مطالعه نیز هر پلی‌مورفیسم به صورت متغیری با سه سطح در نظر گرفته شد.

جدول ۱: فهرست ۳۵ پلی‌مورفیسم مطالعه، ژن و جایگاه ژنی آنان

ژن	ناحیه ژنی	فهرست پلی‌مورفیسم‌های تک‌نکلئوتیدی موجود در ناحیه ژنی مربوطه
	Promoter	rs1122608, rs1529729, rs4300767, rs10417578, rs10411252
	Exon2	rs2228671, rs3745677
	Exon9	rs1003723, rs1569372, rs5930
	Exon11	rs5929, rs4508523
LDLR	Exon12	rs688, rs1799898, rs7259278, rs17248882, rs2738447
	Exon13	rs5925, rs145643854, rs2569549, rs116959285
	Exon13	rs2304182, rs145293532, rs2738460, rs2304181, rs13306501
	3'UTR1	rs14158, rs6413504, rs2116897, rs3826810
	3'UTR2	rs1433099, rs2738466
PCSK9	Promoter	rs11206510, rs17192725, rs17111490

یادگیری ماشینی، شاخه‌ای از هوش مصنوعی، روشی علمی است برای طراحی و تعمیم الگوریتم‌هایی که به رایانه اجازه می‌دهند که رفتارها را بر اساس داده‌های تجربی مانند پایگاه داده‌ها، استنتاج نماید. تمرکز اصلی یادگیری ماشینی، تشخیص الگوهای پیچیده و تصمیم‌گیری هوشمندانه بر اساس داده است؛ تاکنون به منظور کشف اثرهای متقابل میان پلی‌مورفیسم‌های تک‌نوکلئوتیدی، روش‌های آماری متعددی، از جمله: روش‌های مارس (۸،۷)، شبکه‌های عصبی (۹)، کلاس بندی و درخت‌های رگرسیونی (۱۰)، کاهش بعد چند عاملی (۱۱)، رگرسیون منطقی (۱۲)، روش تقسیم بندی ترکیبی (۱۳) و تحلیل دوباره سازی (۱۴) استفاده شده است. در این مطالعه تنها به استفاده از دو روش رندم فارست و درخت تصمیم بسنده شده. قسمت بعد تنها به مرور اجمالی این دو روش می‌پردازد، جزئیات بیشتر مربوط به روش‌ها را می‌توان در منابع ارجاع داده شده، یافت.

درخت کلاس‌بندی-رگرسیونی: یکی از قدرتمندترین و در عین حال ساده‌ترین روش‌های ناپارامتری آماری به منظور کلاس‌بندی متغیر پاسخ است (۱۶). در مطالعات اکتشافی این روش یکی از روش‌های جایگزین برای رگرسیون چندگانه است. شمای این روش همانند نمودار گردشی است که در آن هر گره آزمونی برای متغیر پیش‌بین یا مشخصه فرد مورد مطالعه بوده و هر شاخه آن نتیجه آزمون و هر برگ یا گره انتهایی کلاس متغیر پاسخ را نشان خواهد داد.

فرض می‌شود داده شامل متغیر پاسخ $Y=(Y_1,Y_2,...,Y_n)$ و p متغیر پیش بین بالقوه $X=(X_1,X_2,...,X_p)$ است و $j=1,...,p$. یک درخت ابتدا با تعیین متغیر X_j از میان تمام متغیرهای بالقوه ساخته می‌شود، این متغیر «بهترین پیش‌بینی» را برای Y خواهد داشت. بعد از انتخاب این متغیر، افراد بر اساس مقادیر X_j به دو گره مختلف در درخت تقسیم می‌شوند. فرض می‌شود X_j یک متغیر نشانگر دو حالتی باشد که حضور یا عدم حضور ژنوتیپ خاص را در j- امین پلی‌مورفیسم تک‌نکلئوتیدی تحت بررسی، نشان می‌دهد. در این صورت افرادی که این ژنوتیپ خاص را در X_j دارند، به یک گروه و سایر افراد به گروه دیگر تقسیم می‌شوند. این فرایند در هر گره به وجود آمده تکرار می‌شود. مجموعه افراد با Ω نشان داده می‌شود. فرض می‌شود، تمام متغیرهای پیش بین بالقوه، دو حالتی هستند و «بهترین پیش بین» اول را با $X_{(1)}$ نشان داده می‌شود. بر اساس مقادیر $X_{(1)}$ افراد (مشاهدات) به دو گروه Ω_1 و Ω_2 تقسیم می‌شوند. که در آن

$$\Omega_1=\{i:X_{i(1)}=0\},\Omega_2=\{i:X_{i(1)}=1\},i=1,2,...$$.

گام بعدی الگوریتم ساختن درخت، تشخیص متغیر بعدی است که «بهترین پیش‌بینی» را برای متغیر Y داشته باشد، اما در داخل هر گروه Ω_1 یا Ω_2، زیر گروه‌های بعدی بر اساس مقادیر $X_{i(2)}$ و $X_{i(3)}$ تعریف می‌شوند. این فرآیند آنقدر به صورت بازگشتی ادامه می‌یابد تا به شرایط ملاک توقف برسیم. در آخر بعد از ساختن درخت ممکن است آن را هرس نماییم، یعنی بعضی شاخه‌های درخت را حذف می‌نماییم تا از بیش برازش شدن مدل جلوگیری نماییم (۱۷). به منظور اجرای این روش از تابع rpart در بسته rpart نرم‌افزار R استفاده شد.

الگوریتم رندم فارست: این الگوریتم اولین بار در سال ۲۰۰۱ توسط بریمن معرفی شد (۱۸). دوباره فرض می‌شود $X=(X_1,X_2,...,X_p)$ مجموعه‌ای است از p متغیر پیش‌بین بالقوه، $X_j=(X_{1j},X_{2j},...,X_{nj})^T$ ، و همچنین فرض می‌شود Y صفت تحت بررسی است (در این مطالعه ابتلا یا عدم ابتلا به بیماری شریان‌های کرونری) و در آن n تعداد افراد در نمونه است. همچنین b شماره درخت و B تعداد کل درختان است. مقدار اولیه b را برابر یک قرار می‌دهیم:

در گام k-ام $\underline{\theta}_k$ نمونه مستقل و با توزیع یکسان از مجموعه داده‌ها انتخاب می‌گردد، این مجموعه با نام مجموعه یادگیری یا آموزشی خوانده می‌شود. همچنین $\underline{x}$ نیز نمونه تصادفی از مجموعه متغیرهای پیش‌بین مورد مطالعه انتخاب می‌گردد. تابع پیش‌بین $h(\underline{x},\underline{\theta}_k)$ با استفاده از $\underline{x}$ و $\underline{\theta}_k$ ساخته می‌شود. گام‌ها B مرتبه تکرار شده تا به تعداد B درخت برسیم.

فرض می‌شود $P=\sum_{k=1}^{E} I(h(\underline{x},\underline{\theta}_k))=1$ و $Q=\sum_{k=1}^{E} I(h(\underline{x},\underline{\theta}_k))=0$ ، اگر $P>Q$ باشد، رندم فارست پیش‌بینی می‌نماید که $\underline{x}$ به کلاس ۱ تعلق دارد (در این مطالعه ابتلا به بیماری شریان‌های کرونری) و برعکس، اگر $P<Q$ باشد، رندم فارست پیش‌بینی می‌نماید که $\underline{x}$ به کلاس ۰ تعلق دارد (در این مطالعه عدم ابتلا به بیماری شریان‌های کرونری است).

متداول است که در مسائل کلاس‌بندی، مقدار B را بزرگ‌تر از ۵۰۰، تعداد مجموعه یادگیری را ۲/۳ ام حجم کل نمونه و حجم مجموعه $\underline{x}$ را برابر مجذور تعداد متغیرهای مورد بررسی، در نظر بگیرند. در این مطالعه تعداد درختان برابر ۳۰۰۰ و تعداد متغیرهای به تصادف انتخاب شده در تشکیل هر درخت، برابر ۷ در نظر گرفته شد. به منظور اجرای این روش از تابع randomForest در بسته randomForest نرم‌افزار R استفاده شد.

اعتبار مدل‌های بدست آمده با استفاده از سه ملاک سنجیده شدند، این چهار ملاک عبارتند از: حساسیت، ویژگی و خطا.

حساسیت: برابر است با تعداد مثبت حقیقی بر روی جمع تعداد مثبت حقیقی و منفی کاذب. بازه این معیار از صفر تا یک است (۱۹, ۲۰) و به وسیله ذیل محاسبه می‌شود. هرچه این مقدار به یک نزدیک‌تر باشد، مطلوب‌تر است.

$$\text{Sensitivity} = \text{حساسیت} = \frac{TP}{TP+FN}$$

ویژگی برابر است با تعداد منفی حقیقی بر روی جمع تعداد منفی حقیقی و مثبت کاذب. بازه این معیار از صفر تا یک بوده (۱۹, ۲۰) و به وسیله ذیل محاسبه

می‌شود. هرچه این مقدار به یک نزدیک‌تر باشد، مطلوب‌تر است.

$$\text{ویژگی} = \text{Specificity} = \frac{TN}{TN + FP}$$

خطا برابر است با مجموع تعداد مثبت کاذب و منفی کاذب بر روی تعداد کل افراد. بازه این معیار از صفر تا یک بوده (۲۰) و به وسیله ذیل محاسبه می‌شود. هرچه این مقدار به صفر نزدیک‌تر باشد، مطلوب‌تر است.

$$\text{خطا} = \text{Error} = \frac{FP + FN}{TP + TN + FP + FN}$$

یافته‌ها

گام نخست تحلیل داده‌ها، بررسی مشخصات نمونه مورد نظر است، در ابتدا فراوانی مشخصات افراد مورد مطالعه در دو گروه مورد و شاهد بررسی شدند، هدف بررسی وجود اختلاف معنادار میان افراد بیمار و سالم در متغیرهای کمی و کیفی مطالعه بود.

جداول ۲ و ۳ به ترتیب مشخصات جمعیت شناختی افراد شرکت‌کننده در مطالعه را نشان می‌دهند. جدول ۲، متغیرهای کمی مطالعه و جدول ۳ متغیرهای کیفی مطالعه را نشان می‌دهد. از آنجایی که تنها سعی بر آن بوده که مدل پیش‌بینی بر اساس پلی‌مورفیسم‌های تک‌نکلئوتیدی ساخته شود، تنها از متغیرهایی در ساختن مدل استفاده شده که که اثر اغتشاش‌گر بر بیماری مورد نظر داشتند.

جدول ۲: مقایسه سطوح متغیرهای کمی مطالعه در افراد مبتلا به بیماری شریان‌های کرونر و افراد سالم

مشخصات افراد مورد مطالعه (واحد اندازه‌گیری متغیر)	مورد (تعداد کل=۱۴۱)		شاهد (تعدادکل=۸۳)		آماره t مستقل	p-مقدار احتمال
	میانگین	انحراف معیار	میانگین	انحراف معیار		
سن (سال)	۴۹/۹۷	۱۰/۱۰	۵۳/۷۰	۱۱/۷۷	-۲/۵۰	۰/۰۱۳
نمایه توده بدنی (کیلوگرم بر مجذور متر)	۲۷/۶۵	۴/۲۰	۲۷/۱۵	۵/۳۳	۰/۷۶	۰/۴۴۵
فشارخون سیستولی (میلی‌متر جیوه)	۱۲/۴۷	۲/۱۹	۱۲/۴۸	۱/۸۲	-۱/۱۷	۰/۲۴۰
فشار خون دیاستولی (میلی‌متر جیوه)	۷/۶۹	۱/۳۴	۸/۳۱	۷/۵۵	-۰/۹۳	۰/۳۵۵
مقدار LDL (میلی‌گرم بر دسی لیتر)	۹۸/۸۰	۴۰/۷۱	۹۶/۹۵	۳۵/۸۳	۰/۳۳	۰/۷۴۲
مقدار HDL (میلی‌گرم بر دسی لیتر)	۴۰/۵۹	۱۵/۸۷	۳۷/۷۸	۱۶/۶۷	-۱/۴۴	۰/۱۵۲
مقدار VLDL (میلی‌گرم بر دسی لیتر)	۲۳/۸۶	۲۹/۰۲	۲۵/۱۰	۱۶/۸۳	-۱/۳۴	۰/۰۹۰
کلسترول کل (میلی‌گرم بر دسی لیتر)	۱۵۸/۰۵	۴۲/۲۰	۱۶۹/۹۴	۴۶/۱۲	-۱/۵۱	۰/۱۳۴
تری‌گلیسیرید (میلی‌گرم بر دسی لیتر)	۱۶۶/۷۲	۸۹/۴۳	۱۶۸/۷۷	۷۷/۲۹	۱/۴۱	۰/۰۸۰

جدول ۳: مقایسه سطوح متغیرهای کیفی مطالعه در افراد مبتلا به بیماری شریان‌های کرونر و افراد سالم•

مشخصات		مورد (تعداد کل=۱۴۱)	شاهد (تعدادکل=۸۳)	آماره آزمون کای دو	p-مقدار احتمال
جنسیت	زن	(۲۲/۷) ۳۲	(۶۸/۷) ۵۷	۴۶/۱۲	<۰/۰۰۱
	مرد	(۷۷/۳) ۱۰۹	(۳۱/۳) ۲۶		
وضعیت سیگار کشیدن	سیگاری	(۶۵/۲) ۷۳	(۶۰/۶) ۶۶	۰/۵۱	۰/۴۷۶
	غیر سیگاری	(۳۴/۸) ۳۹	(۳۹/۴) ۴۳		
تحصیلات	بی‌سواد	(۵۳/۶) ۳۰	(۴۶/۶) ۲۶	۳/۷۸	۰/۲۸۶
	ابتدائی	(۶۵/۹) ۵۸	(۳۴/۱) ۳۰		
	دیپلم	(۷۰/۲) ۴۰	(۲۹/۹) ۱۷		
	بالاتر از دیپلم	(۶۶/۷) ۱۲	(۳۳/۳) ۶		
قومیت	فارس	(۶۳/۹) ۵۵	(۳۶/۱) ۳۱	۲/۶۷	۰/۶۱۳
	ترک	(۶۷/۸) ۳۸	(۳۲/۲) ۱۸		
	بلوچ، لر، کرد	(۷۱/۰) ۲۲	(۲۹/۰) ۹		
	شمالی	(۶۰/۹) ۱۴	(۳۹/۱) ۹		
	سایر	(۵۲/۰) ۱۳	(۰/۴۸) ۱۲		

• اعداد به صورت تعداد (درصد ستونی) گزارش شده‌اند.

با توجه به جدول ۲ تنها متغیر پیوسته معنادار، متغیر سن است (p=۰/۰۱۳). در جدول ۳ نیز تنها متغیر گسسته معنادار، متغیر جنس است (p<۰/۰۰۱). سایر متغیرهای پیوسته و یا گسسته مورد نظر، ارتباط معناداری با بیماری نداشتند.

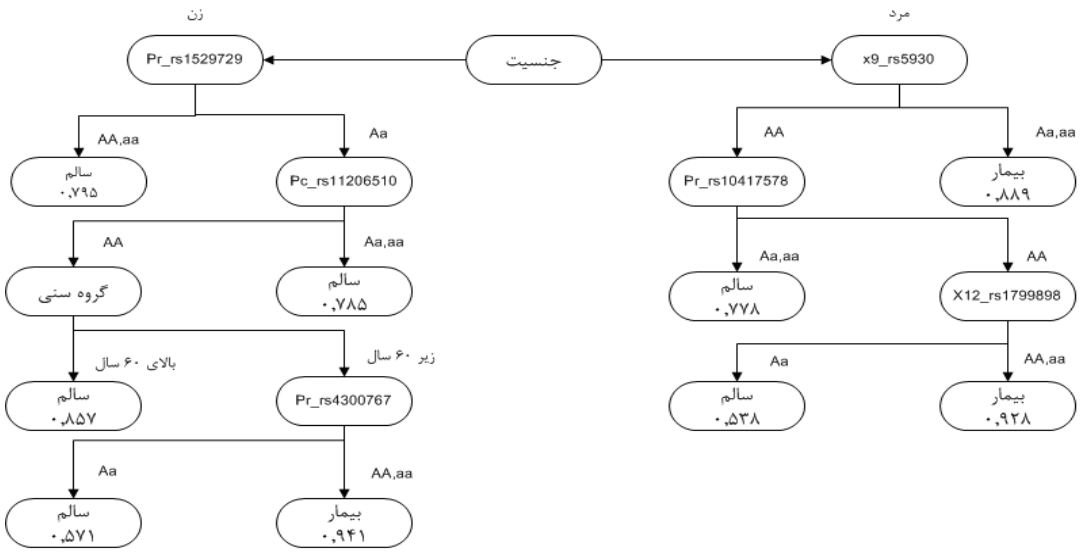
بنابراین با توجه به نتایج جداول ۲ و ۳ نیاز به تعدیل نسبت به دو متغیر سن و جنس وجود داشت. از آن‌جایی که اکثر روش‌های مورد بررسی در مطالعه، تنها متغیرهای مستقل رسته‌ای را می‌پذیرند، متغیرهای مورد نظر به صورت رسته‌ای وارد مدل شدند. با توجه به مقالات پیشین مرتبط با بیماری قلبی، یک رده‌بندی رایج سن، افراد زیر ۴۰ سال، افراد میان ۴۰ تا ۴۹ سال، افراد میان ۵۰ تا ۵۹ سال و افراد بالای ۶۰ سال است (۲۱). بنابراین متغیر سن با رده‌های فوق و متغیر جنس در تحلیل اصلی وارد شدند.

در این مطالعه اعتبارسنجی دو مدل توسط اعتبارسنجی متقابل ۱۰ قسمتی انجام شد. در جدول ۴ حساسیت، ویژگی، خطا، دقت، و ضریب همبستگی متیو در دو روش بر اساس میانگین دو قسمت ارائه شده‌اند. به صورت طبیعی، روشی مطلوب است که به صورت همزمان حساسیت و ویژگی آن بالاتر و خطا پایین‌تر باشد. با توجه به جدول ۳، مقادیر حساسیت، ویژگی و خطا به نظر می‌رسد که درخت کلاس‌بندی-رگرسیونی به طور کلی عملکرد موفق‌تری نسبت به رندم فارست دارد. برای این روش مقادیر حساسیت، ویژگی و خطا به ترتیب برابر ۰/۵۷۵، ۰/۶۸۴ و ۰/۲۶۴ برآورد شده‌اند.

جدول ۴: میانگین و انحراف معیار حساسیت، ویژگی و خطا روش‌های رندم فارست و درخت کلاس‌بندی-رگرسیونی در پیش‌بینی ابتلا به بیماری افراد

روش	حساسیت (sensitivity)		ویژگی (Specificity)		خطا (Error)	
	میانگین	انحراف معیار	میانگین	انحراف معیار	میانگین	انحراف معیار
رندم فارست	۰/۵۵۳	۰/۳۲۴	۰/۶۶۷	۰/۲۵۶	۰/۲۸۵	۰/۱۲۲
درخت کلاس‌بندی-رگرسیونی	۰/۵۷۵	۰/۲۷۰	۰/۶۸۴	۰/۲۵۶	۰/۲۶۴	۰/۱۳۰

شکل ۱ مدل پیشنهاد شده توسط درخت کلاس‌بندی-رگرسیونی برای پیش‌بینی بیماری شریان‌های کرونر با استفاده از پلی‌مورفیسم‌های تک‌نکلئوتیدی و متغیرهای جنس و رده‌های سنی است.



شکل ۱: درخت حاصل از درخت کلاس‌بندی-رگرسیونی. اعداد داخل آخرین شاخه، مقدار احتمال تخصیص یافتن مشاهده به کلاس تعلق یافته است. در این شکل نمادهای AA، Aa و aa به ترتیب اشاره به هموزیگوت شایع، هتروزیگوت و هموزیگوت واریانت بودن ژنوتیپ پلی‌مورفیسم مورد نظر دارند. هنگامی که در این مدل متغیری در زیر شاخه متغیر دیگر قرار گیرد، میان دو متغیر اثر متقابل وجود دارد.

با توجه به شکل ۱، در صورتی فردی مبتلا به بیماری تشخیص داده می‌شود که:

- مرد باشد و پلی‌مورفیسم rs5930 او از نوع هموزیگوت شایع (نوع وحشی) نباشد (با احتمال ۰/۸۸۹).
- یا مرد باشد و هر سه پلی‌مورفیسم rs5930، rs10417578 و rs1799898 او از نوع هموزیگوت شایع (نوع وحشی) باشد (با احتمال ۰/۹۲۸).
- یا زن بوده و رده سنی او بین ۴۰ تا ۵۹ سال، پلی‌مورفیسم rs1529729 او از نوع هتروزیگوت، هر دو پلی‌مورفیسم rs11206510 و rs4300767 او از نوع هموزیگوت شایع (نوع وحشی) باشد (با احتمال ۰/۹۴۱).

بحث

با توجه به آمار مرگ و میر و ناتوانی ناشی از شریان‌های کرونر، امکان غربالگری و تشخیص افرادی که از نظر ژنتیکی مستعد این بیماری باشند، می‌تواند در کاهش هزینه‌های درمان از مسیر پیشگیری و کنترل بروز بیماری در افراد با خطر بالا تأثیرگذار باشد. با توجه به مدل پیشنهادی، می‌توان با کنترل عوامل محیطی خطرزا موجب کاهش بروز بیماری در افراد سالمی که دارای استعداد ژنتیکی ابتلا به بیماری تشخیص داده می‌شوند، شد. این مدل اثرات متقابل میان پلی‌مورفیسم‌ها را نیز به ترتیب توالی که در درخت آمده‌اند، مشخص نمود.

از دیدگاه یادگیری ماشینی سایر مقالات مشابه نتایج متفاوت دارند. در سال ۲۰۰۴ لونتا و همکاران (۲۲)، نتیجه آن‌ها این است که در مسائل مرتبط با کشف اثر متقابل پلی‌مورفیسم‌های تک‌نکلئوتیدی، الگوریتم رندم فارست عملکرد بهتری نسبت به روش‌های استاندارد آماری از جمله رگرسیون لجستیک دارد.

در سال ۲۰۱۱ چن و همکاران (۲۳)، عملکرد روش‌های رگرسیون منطقی، رندم فارست و رگرسیون لجستیک بیزی را بررسی نمودند. از نظر آن‌ها با توجه به بررسی‌های انجام شده ، هیچ‌کدام از روش‌های مورد بررسی برتری نسبت به دیگری ندارد. در سال ۲۰۱۲ آپستیل- گودارد و همکاران (۵)، در مقاله‌ای مروری خود، عملکرد روش‌های رندم فارست، کاهش بعد چند عاملی، شبکه‌های عصبی و ماشینی‌های برداری پشتیبان را بر روی داده‌های پلی‌مورفیسم تک‌نکلئوتیدی بررسی کردند، از نظر آن‌ها از میان روش‌های به کار گمارده شده، هیچ روشی به تنهایی نمی‌تواند بهترین گزینه ممکن برای کشف اثرها باشد. همچنین در این سال، یو و همکاران (۲۴)، عملکرد رگرسیون لجستیک، رگرسیون منطقی، درخت کلاس‌بندی-رگرسیونی و رندم فارست را بررسی نمودند. در این مطالعه، رندم فارست بهترین عملکرد را در میان روش‌های به کار برده شده داشت. هرچند با توجه به مشابهت این مطالعه، با مطالعه حاضر، نتایج بدست آمده تطابق ندارند. هرچند در مطالعه یو، چندین ژن مورد بررسی قرار گرفته است.

نتیجه‌گیری

این مطالعه یکی از نخستین مطالعات بر مبنای مقایسه روش‌های یادگیری ماشینی در مطالعات ژنتیکی است. در پژوهش حاضر، درخت کلاس‌بندی-رگرسیونی عملکرد نسبتاً بهتری در ارائه پیش آگهی در بیماری شریان‌های کرونر نشان داد. این روش نه تنها پلی‌مورفیسم‌های تأثیرگذار در بیماری مورد نظر را معرفی می‌نماید، بلکه مدل درختی را برای طبقه‌بندی افراد به افراد مستعد به بیماری و افراد با خطر کمتر در مواجه به این بیماری معرفی می‌نماید.

هرچند که تشخیص پلی‌مورفیسم‌های تاثیرگذار توسط مدل، محققان را در یافتن پلی‌مورفیسم‌های مرتبط با بیماری شریان‌های کرونری کمک می‌نماید، اما مدل پیش‌بینی معرفی شده نیازمند بررسی‌های بیشتر است و تعمیم نتایج به صورت کلی کاوش‌های بیشتر را در این زمینه می‌طلبد.

تشکر و قدردانی

نگارندگان مقاله بر خود لازم میدانند که از کلیه همکاران مرکز مطالعات ژنتیک بیماری قلب دانشگاه علوم پزشکی تهران تشکر نمایند. همچنین از تمامی بیماران شرکت کننده که انجام این تحقیق بدون وجود آن‌ها میسر نبود، قدردانی می‌شود.

References

1. Vali U, Brandstrom M, Johansson M, Ellegren H. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genet. 2008;9:8.
2. Vignal A, Milan D, SanCristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol. 2002 May-Jun;34(3):275-305.
3. Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol. 2006 Mar 10;356(5):1263-74.
4. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 Feb 15;409(6822):928-33.
5. Upstill-Goddard R, Eccles D, Fliege J, Collins A. Machine learning approaches for the discovery of gene-gene interactions in disease data. Brief Bioinform. 2011 May 18.
6. Association AH. Heart and Stroke Statistical Update. 2013; Available from: http://www.americanheart.org/statistics/pdf/HSSTATS2001_1.0.pdf.
7. Lopez AD, Murray CC. The global burden of disease. Nat Med. 1998.
8. Luis AJ. Atherosclerosis. Nature. 2000:233-41.
9. Ross R. Atherosclerosis is an inflammatory disease. Am Heart J. 1999 Nov;138(5 Pt 2):S419-20.
10. Hegele RA, Emi M, Wu LL, Hopkins PN, Williams RR, Lalouel JM. Clinical application of deoxyribonucleic acid markers in a Utah family with hypercholesterolemia. Am J Cardiol. 1989 Jan 1;63(1):109-12.
11. Pimstone SN, Sun XM, du Souich C, Frohlich JJ, Hayden MR, Soutar AK. Phenotypic variation in heterozygous familial hypercholesterolemia: a comparison of Chinese patients with the same or similar mutations in the LDL receptor gene in China or Canada. Arterioscler Thromb Vasc Biol. 1998 Feb;18(2):309-15.
12. Hobbs HH, Brown MS, Goldstein JL. Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. Hum Mutat. 1992;1(6):445-66.
13. Pullinger CR, Kane JP, Malloy M, J. Primary Hypercholesterol: genetic causes and treatment of five monogenic disorders. Expert Rev Cardiovasc. 2003;1:107-19.
14. Anderson R.G, Goldstein J, Brown M. From cholestrol hemostasis to new paragdigms in membrane biology. Cell Biol. 2003;13:534-9.
15. Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, Devillers M, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet. 2003 Jun;34(2):154-6.
16. Breiman L. Classification and regression trees. Belmont, Calif.: Wadsworth International Group; 1984.
17. Foulkes AS. Applied Statistical Genetics with R for population-based association studies: Springer; 2009.
18. Breiman L. Random Forest. Machine Learning. 2001.
19. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ: British Medical Journal. 1994;308(6943):1552.
20. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16(5):412-24.
21. Ockene JK, Hosmer DW, Williams JW, Goldberg RJ, Ockene IS, Raia 3rd T. Factors related to patient smoking status. American journal of public health. 1987;77(3):356-7.
22. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet. 2004;5:32.
23. Chen CC, Schwender H, Keith J, Nunkesser R, Mengersen K, Macrossan P. Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. IEEE/ACM Trans Comput Biol Bioinform. 2011 Nov-Dec;8(6):1580-91.
24. Yoo W, Ference BA, Cote ML, Schwartz A. A Comparison of logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Enviromental Interactions. Internationl Journal of Applied Science and Technology. 2012;2(7).