معرفی الگوریتم‌های مدل رده‌بندی درختی و کاربرد آن در تعیین عوامل مؤثر بر ابتلا به سرطان مری در استان گلستان

ناصر بهنام پور[۱]، ابراهیم حاجی زاده[۲*]، شهریار سمنانی[۳]، فرید زایری[۴]

۱- دانشجوی دکتری تخصصی آمار زیستی، دانشکده علوم پایه پزشکی دانشگاه تربیت مدرس، تهران، ایران
۲- دانشیار گروه آمار زیستی، دانشکده علوم پایه پزشکی دانشگاه تربیت مدرس، تهران ، ایران
۳- دانشیار گروه داخلی دانشکده پزشکی، مرکز تحقیقات کبد و گوارش، دانشگاه علوم پزشکی گلستان، گرگان، ایران
۴- دانشیار گروه آمار زیستی، دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

چکیده

زمینه و هدف: *یکی از اهداف متداول در تحقیقات پزشکی، تعیین عوامل مؤثر بر رخداد رویداد مورد بررسی است. با توجه به اثر متقابل عوامل خطر، از مدل های رگرسیونی، تحلیل ممیزی و رده بندی استفاده می‌شود. استفاده از این مدل ها، نیازمند برقراری فرض هایی است که معمولاً در داده‌های پزشکی برقرار نیستند، لذا بایستی روش‌های جایگزین مورد استفاده قرار گیرد. با توجه به تکثر و تنوع عوامل مؤثر بر ابتلا به سرطان مری، هدف این مقاله تعیین این عوامل با استفاده از مدل رده بندی درختی است.*

روش بررسی: *داده‌های این مقاله حاصل یک تحقیق مورد-شاهدی است. گروه مورد، کلیه موارد قطعی سرطان مری شامل ۹۰ فرد مذکر و ۶۹ فرد مؤنث طی یک سال در استان گلستان است. به ازای هر مورد، دو شاهد در نظر گرفته شد. یک شاهد از خانواده و شاهد دیگر از همسایگان بیمار با همسان سازی بر روی متغیرهای سن، جنس، قومیت و محل سکونت انتخاب شد. داده‌ها با استفاده از مدل رده‌بندی درختی و با استفاده از نرم افزار R تجزیه و تحلیل شد. ضابطه جینی برای انتخاب بهترین افراز روی هر گره و مساحت زیر منحنی مشخصه گیرنده برای تعیین میزان صحت مدل رده بندی درختی استفاده شد.*

یافته‌ها: *نتایج مدل رده بندی درختی نشان داد که مواجهه با سی‌تی اسکن و اشعه ایکس رنگی (از عوامل محیطی- اجتماعی)، نشستن دست‌ها پس از اجابت مزاج و سابقه کشیدن سیگار از عوامل سبک زندگی، سابقه خانوادگی ابتلا به سرطان (از عوامل وراثتی) می‌توانند در ابتلا به سرطان مری مؤثر باشند.*

نتیجه‌گیری: *عدم نیاز به برقراری هیچ پیش فرضی برای مدل سازی و تفسیر آسان نتایج مدل‌های درختی، دو مزیت اساسی آن است که می‌تواند مورد توجه و استفاده پژوهش گران حوزه‌های مختلف پزشکی قرار گیرد.*

کلید واژه‌ها: *سرطان مری، مدل رده بندی درختی، گلستان*

*نویسنده مسئول: دکتر ابراهیم حاجی زاده
نشانی: دانشکده علوم پایه پزشکی دانشگاه تربیت مدرس، تهران، ایران
تلفن: ۰۲۱۸۲۸۸۴۵۲۴ پست الکترونیک: hajizadeh@modares.ac.ir

مقدمه

برای تبیین رابطه بین پیامد (متغیر وابسته) و مواجهه (متغیرهای توضیحی)، در شرایطی که تعداد متغیرهای توضیحی زیاد بوده و دارای تنوع فراوان باشند، روش‌های آماری مختلفی ابداع و توسعه یافته است. این روش‌ها متناسب با ماهیت متغیر وابسته و چگونگی متغیرهای توضیحی بسیار متنوع‌اند. در آمار کلاسیک، برای چنین شرایطی استفاده از مدل‌های رایج رگرسیونی متداول است. اما این مدل‌ها دارای ساختار پیچیده‌ای بوده و اغلب دارای پیش شرط‌های سخت گیرانه‌ای مانند برقراری توزیع نرمال و همگنی واریانس‌ها و ... هستند و عدم برقراری این پیش شرط‌ها، استفاده از این مدل‌ها را محدود می‌کند(۱). لذا نیاز به ابداع روش‌های جدیدی است که علاوه بر فائق آمدن بر شرایط فوق بتوانند با سرعت بالا و انجام محاسبات کمتر به نتایج قابل قبول دست یابند. در همین راستا، روش‌هایی تحت عنوان درخت تصمیم و مدل های درختی ابداع و توسعه یافته اند که می‌توانند بخش قابل توجهی از این نیازها را پوشش دهند.

از طرفی، می‌دانیم سرطان یک بیماری با عامل ناشناخته است که عوامل بسیاری می توانند در ابتلا به آن مؤثر باشند و سرطان مری نیز از این قاعده مستثنی نیست. مطالعات سبب شناختی نشان داده است که از مجموع عوامل مربوط به سبک زندگی؛ الکل، نوشیدن چای داغ، سیگار، زمینه ژنتیکی، رژیم غذایی حاوی روغن جامد، قند و شکر، انواع شیرینی و دسر، نمک، انواع ترشی و نوشابه گازدار و کنسرو، خطر ابتلا به سرطان مری را افزایش می‌دهند. همچنین تحرک کافی، مصرف میوه، سبزی، زیتون، لبنیات کم چرب و ماهی خطر ابتلا به سرطان مری را کاهش می‌دهند. همچنین آشالازی، مری بارت، پرتو تابی، سن بالای چهل سال، ریفلاکس و... نیز خطر ابتلا به سرطان مری را افزایش می‌دهند. لذا با توجه شیوع بالای سرطان مری در استان گلستان و اهمیت شناسایی عوامل مؤثر بر ابتلا به این سرطان، و با توجه به تکثر و تنوع این عوامل و عدم برقراری پیش فرض های استفاده از مدل‌های رایج آماری، در این مقاله ضمن معرفی مدل رده‌بندی درختی و بعضی الگوریتم‌های آن، عوامل مؤثر بر ابتلا به سرطان مری بر اساس این مدل شناسایی و معرفی شده است.

درخت تصمیم و مدل رده بندی درختی:

درخت تصمیم یکی از روش‌های ناپارامتری رده‌بندی کردن است که با توجه به نوع متغیر وابسته به دو دسته رده بندی درختی (Classification Tree) برای متغیر رسته‌ای و رگرسیون درختی (Regression Tree) برای متغیر پیوسته تقسیم می‌شود. رده‌بندی درختی در راستای روش‌هایی نظیر تحلیل ممیزی (تابع تشخیص) (Discriminate function Analysis) و رگرسیون لجستیک (Logistic Regression) است. در این روش مجموعه‌ای از شرط‌های منطقی (Logical if-then conditions) به صورت یک الگوریتم با ساختار درختی (Algorithm Tree-building) برای رده بندی یا پیش بینی یک پیامد به کار می‌رود (۲).

با توجه به این که شاخص‌های متنوع و روش‌های گوناگونی برای تعیین درخت تصمیم معرفی شده است، الگوریتم های متنوع و گوناگونی نیز ارائه شده است که مهمترین و شناخته شده‌ترین الگوریتم ها به ترتیب عبارتند از:

۱. الگوریتم CHAID
(Chi-squared Automatic Interaction Detector)

این الگوریتم توسط کاس در سال ۱۹۸۰ برای استفاده در مورد متغیرهای کیفی معرفی شد که می تواند برای متغیرهای کمی گروه بندی شده نیز استفاده شود. در هر گره، می توان بیش از دو تقسیم نیز داشت. در این روش از مقدار P-Value آماره کای- دو مربوط به آزمون استقلال جداول توافقی استفاده می شود. از بین متغیر های موجود، متغیری که دارای P-Value کوچک تری باشد در مرحله اول برای تقسیمات روی یک گره در نظر گرفته می شود. ضعف این الگوریتم عدم توانایی آن در ایجاد بهینه ترین تقسیمات ممکن بر اساس متغیر های موجود است (۳).

۲. الگوریتم CART
(Classification and Regression Tree)

این روش که موجب تشکیل یک درخت تصمیم با تقسیمات دوتایی می‌گردد، توسط بریمن و همکارانش در سال ۱۹۸۴ به طور کامل معرفی شد. این روش برای متغیرهای کمی طراحی گردیده ولی قابل استفاده برای هر نوع متغیری است. بر اساس این الگوریتم، نرم افزار آماری تحت نام CART نیز ساخته شده است که از شناخته شده ترین برنامه‌ها است. در این روش و برای متغیر پاسخ کیفی، شاخص جینی (Gini Index) به عنوان معیاری برای انتخاب متغیرهای مناسب، معرفی شده است.

در معرفی مدل درختی با تقسیمات دوتایی می‌توان از شاخص‌های دیگری نظیر آنتروپی نیز استفاده نمود. مزیت شاخص جینی نسبت به آنتروپی و شاخص‌های دیگر، سرعت بالاتر آن در انجام محاسبات است. مدل CART را می‌توان به عنوان یکی از شناخته شده‌ترین الگوهای رده بندی به منظور تشخیص و پیشگویی در علوم پزشکی بر شمرد.

در مدل CART هرس کردن درخت رده بندی بر اساس Cost-Complexity صورت می‌پذیرد و بررسی دقت درخت معرفی شده به کمک نمونه آزمون معرفی می‌گردد.

یکی از ایرادات مطرح برای مدل CART اریبی این مدل در انتخاب متغیرها است. علاوه بر این، در متغیرهای کیفی با تعداد سطوح بیش از دو، نتایج حاصل گیج کننده خواهد بود. چون ممکن است چند سطح یک متغیر به یک گره تعلق بگیرد که این باعث می شود نتوان تفسیر ساده ای از نتایج ارائه نمود(۴) .

۳. الگوریتم های ID۳ و C۴,۵
(Induction of Decision Trees)

الگوریتم ID۳ توسط کوئینلن در سال ۱۹۸۶ معرفی گردید. این الگوریتم برای معرفی و ساخت درخت رده بندی با تقسیمات چندتایی در هر گره بسیار مناسب است و این الگوریتم برای متغیرهای کیفی طراحی گردید، ولی می‌توان از آن برای مجموعه‌ای از متغیرها، چه کیفی و چه عددی استفاده کرد. ملاک تصمیم گیری در این الگوریتم بر اساس شاخص آنتروپی است که به کمک آن شاخص‌های Information Gain و Gain Ratio محاسبه می شود. با توجه به بعضی از ضعف های الگوریتم ID۳ ، کوئینلن در سال ۱۹۹۳ آن را اصلاح و تحت الگوریتم C۴,۵ معرفی نمود. این الگوریتم نسبت به ID۳ اریبی کمتری دارد و برای مشاهدات با مقادیر گمشده مناسب است.

نتایج سریع، مختصر و مفید و قابل اطمینان این دو الگوریتم، را به عنوان یک روش قابل قبول برای رده‌بندی مشاهدات تبدیل نموده که در علوم پزشکی مورد استفاده قرار می گیرند (۵).

۴. الگوریتم QUEST
(Quick Unbiased Efficient Statistical Trees)

این الگوریتم در سال ۱۹۹۷ توسط لو و شی برای متغیرهای پاسخ اسمی طراحی شد. درخت رده بندی حاصل از این الگوریتم نظیر مدل CART دارای تقسیمات دوتایی بوده و ملاک تصمیم برای انتخاب متغیرها با استفاده از مقدار P-Value مربوط به آماره F آزمون ANOVA برای متغیرهای کمی و P-Value آماره کای- دو مربوط به جداول توافقی برای متغیر های کیفی صورت می پذیرد. این الگوریتم با توجه به این که از مقدار P-Value برای تصمیم گیری استفاده می نماید، موجب تشکیل درختی نااریب از متغیرها می گردد. این الگوریتم ضمن حفظ دقت برآورد در مدل CART، از سرعت بالاتری در معرفی یک درخت رده بندی نسبت به آن برخوردار است (۶).

۵. الگوریتم CRUISE
(Classification Rule with Unbiased Interaction Selection and Estimation)

این الگوریتم در سال ۲۰۰۱ توسط کیم و لو معرفی گردید که می تواند یک درخت رده بندی با تقسیمات چندتایی معرفی نماید. این الگوریتم به خوبی روش هایی نظیر CART و QUEST عمل می کند ولی با توجه به این که از تقسیمات چند تایی بهره می برد، از سرعتی بالاتری برخوردار است و درخت کوچک تری نیز با استفاده از این الگوریتم معرفی می‌گردد. درخت معرفی شده در این روش نااریب بوده و طوری طراحی گردیده که با وجود مقادیر گمشده برای داده ها نیز، به خوبی عمل نماید (۷).

روش بررسی:

داده‌های این مقاله حاصل یک تحقیق مورد-شاهدی است. گروه مورد، کلیه موارد قطعی سرطان مری طی سال ۱۳۸۸ استان گلستان است که به ازای هر مورد، دو شاهد در نظر گرفته شده است. یک شاهد از خانواده و شاهد دیگر از همسایگان بیمار که با همسان سازی بر روی متغیرهای سن، جنس، قومیت و محل سکونت انتخاب شدند. اطلاعات مورد نظر با استفاده از پرسشنامه شامل اطلاعات دموگرافیک، فرهنگ تغذیه‌ای و سبک زندگی که روایی و پایایی آن بر اساس یک مطالعه مقدماتی تأیید شده بود و با مراجعه حضوری به محل زندگی بیماران و شاهدان جمع آوری شد.

داده ها با استفاده از مدل رده بندی درختی و با استفاده از نرم افزار R تجزیه و تحلیل شد. برای دستیابی به مدل مناسب، الگوریتم های ID۳ ، C۴,۵،CHAID،CART ، QUEST و CRUISE مورد مطالعه قرار گرفت و از الگوریتم رده‌بندی درختی برای تعیین عوامل موثر بر ابتلاء به سرطان مری استفاده شد.

از ضابطه جینی برای انتخاب بهترین افراز روی هر گروه و از مساحت زیر منحنی مشخصه محرکه گیرنده ROC Curve Receiver Operating Characteristic Curve برای تعیین میزان صحت مدل رده بندی درختی استفاده شد. مقادیر بیش از ۸۰ درصد مساحت زیر منحنی نشان دهنده توان بالای مدل در رده بندی ممیزی و مقادیر بین ۷۰ تا ۸۰ درصد بیانگر قابل قبول بودن مدل تشخیصی است(۸).

یافته ها:

از مجموع ۱۵۹ بیمار مبتلا به سرطان مری، ۶۹ بیمار مؤنث و ۹۰ بیمار مذکر بودند. میانگین و انحراف معیار سن ابتلا در بیماران مذکر ۱۳/۷۵±۶۳/۶۸ و در بیماران مؤنث ۱۱/۳۴±۶۳/۰۷ سال و سن ابتلا در ساکنین روستا پایین تر از ساکنین شهر تعیین شد (جدول شماره ۱). همچنین میزان بروز خام و استاندارد شده سنی به ترتیب ۹/۳۴ ، ۱۴/۲۳ در هر صد هزار نفر تعیین شد.

جدول شماره ۱: شاخص‌های مرکزی و پراکندگی سن ابتلا به سرطان‌های دستگاه گوارش به تفکیک جنس و محل سکونت

متغیر		تعداد	میانه	میانگین	انحراف معیار
جنس	مذکر	۹۰	۶۵	۶۳/۶۸	۱۳/۷۵
	مؤنث	۶۹	۶۴	۶۳/۰۷	۱۱/۳۴
محل سکونت	روستا	۱۰۲	۶۵	۶۲/۹۷	۱۳/۹۶
	شهر	۵۷	۶۵	۶۴/۲۱	۱۰/۲۲
کل		۱۵۹	۶۵	۶۳/۴۲	۱۲/۷۳

در بررسی عوامل موثر بر ابتلاء به سرطان مری با استفاده از الگوریتم رده بندی درختی، نتایج به شرح زیر تعیین شد:

ریشه درخت رده بندی بر اساس سابقه مواجهه با CT اسکن تشکیل شده است، به طوری که خطر ابتلا به سرطان مری در افراد دارای سابقه تماس با CT اسکن و اشعه X (عکس رنگی) و سابقه سرطان در خانواده یا بستگان و سابقه کشیدن سیگار در بالاترین حد قرار می گیرد. اما در افراد بدون سابقه مواجهه با CT اسکن، نشستن دست ها با صابون یا مایع دستشویی پس از اجابت مزاج خطر ابتلا به سرطان مری را افزایش می دهد (شکل شماره ۱).

در بررسی کارایی مدل با استفاده از مساحت زیر منحنی مشخصه محرکه گیرنده (ROC Curve) برای تعیین میزان صحت مدل رده بندی درختی، مقدار مساحت ۸۶ درصد تعیین شد که نشان دهنده توان بالای مدل رده بندی درختی در تعیین عوامل موثر بر ابتلا به سرطان مری است.

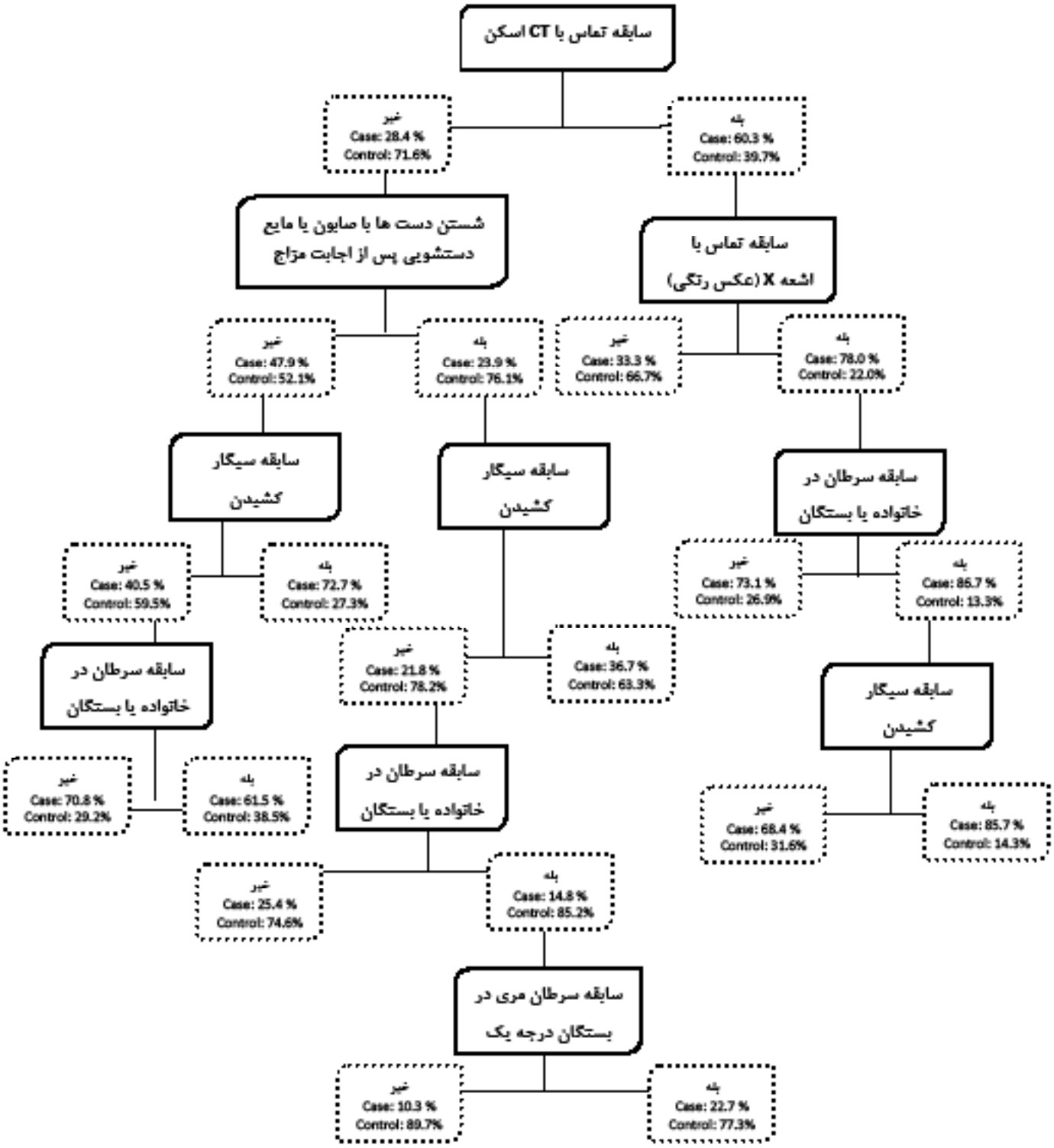
بحث و نتیجه گیری:

درخت تصمیم یکی از روش‌های داده کاوی (Data Mining) است و داده کاوی قادر به کشف و استخراج دانش جدید از داده (حتی داده های گذشته نگر) است. نحوه پیش پردازش داده‌ها و هم چنین متغیرهای منتخب، تأثیر قابل توجهی در کشف دانش دارد(۹). درخت تصمیم در مسائلی کاربرد دارد که بتوان آن‌ها را به صورتی مطرح نمود که پاسخ واحدی به صورت یک دسته یا کلاس ارائه دهند. برای مثال می توان درخت تصمیمی ساخت که به این سوال پاسخ دهد: آیا مریض به سرطان پستان مبتلاست؟(۱۰).

در این مطالعه با استفاده از مدل رده‌بندی درختی، متغیرهای سابقه تماس با CT اسکن، سابقه تماس با اشعه x (عکس رنگی)، سابقه سرطان در خانواده یا بستگان، سابقه کشیدن سیگار و نشستن دست ها با صابون یا مایع دستشویی پس از اجابت مزاج به عنوان عوامل تأثیرگذار بر ابتلا به سرطان مری آشکار شدند که در بعضی از مطالعاتی که تجزیه و تحلیل داده ها با استفاده از مدل های رایج آماری مانند رگرسیون لجستیک انجام شده است نیز، مورد تأیید قرار گرفته‌اند.

مطالعات مختلفی با استفاده از مدل رده بندی درختی به منظور بررسی پیش‌گویی کننده های مؤثر بر ابتلاء به بیماری ها انجام شده است که تنها اندکی از آن‌ها در زمینه سرطان مری است. لذا در این بخش، از مطالعاتی که با استفاده از مدل رده‌بندی درختی برای سایر بیماری‌ها صورت گرفته است نیز استفاده شده است.

در مطالعه Silvera و همکاران در سال ۲۰۰۴ که به بررسی نقش عوامل خطر تغذیه ای بر ابتلا به سرطان مری و معده با استفاده از مدل رده بندی درختی انجام شده است. رفلکس مری- معده ای به عنوان یکی از مهم‌ترین عوامل خطر ابتلا به آدنوکارسینومای کاردیا و غیر کاردیا است. همچنین مصرف گوشت قرمز، میوه‌هایی غیر از مرکبات، چای سیاه و سبزیجات خام به عنوان عوامل خطر در ابتلا به سرطان مری در این مدل مطرح بوده است(۱۱).



شکل شماره ۱: مدل رده بندی درختی برای تعیین عوامل مؤثر بر ابتلا سرطان مری

در مطالعه Silvera و همکاران در سال ۲۰۱۴ عوامل خطر سبک زندگی و تغذیه‌ای در بیماران مبتلا به سرطان معده مورد بررسی قرار گرفت. در افراد مبتلا به اسکوآموس سل کارسینومای مری، سیگار کشیدن به عنوان اصلی‌ترین عامل خطر و پس از آن رفلکس ازوفاژیال، درآمد، نژاد، میوه‌های غیر مرکبات و میزان انرژی دریافتی به عنوان عامل خطر مطرح گردیدند(۱۲).

Valera و همکاران در سال ۲۰۰۶ بر اساس مدل رده‌بندی درختی به بررسی پیشگویی کننده‌های ابتلا به سرطان کولورکتال پرداختند که، ۱) شاخص تکثیر، ۲) متاستاز پاتولوژیک گره‌های لنفاوی ۳) سایز تومور به عنوان زیر گروه های سازنده مدل رده بندی درختی برای خطر مرگ شناخته شدند(۱۳).

Camp و همکاران در سال ۲۰۰۲ با استفاده از روش رده بندی درختی به بررسی عوامل خطر سرطان کولون پرداختند، نتیجه مطالعه نشان داد که استفاده از داروهای NSAID (سطح اول)، سابقه خانوادگی (سطح دو و سه)، بسیاری از پیش‌گویی کننده‌های تغذیه‌ای شامل رژیم غذایی غربی (سطح دو) و الگوی تغذیه ای محتاطانه (سطح ۴)، مصرف دانه‌های خوراکی (سطح ۳)، مصرف کلسیم (سطح ۴)، مصرف اسید فولیک (سطح ۴)، سطح لوتئین (سطح ۶)، فعالیت فیزیکی (سطح ۳)، BMI (سطح ۵) و سن (سطح ۴)، می‌توانند به عنوان عامل خطر ابتلا به سرطان کولون در نظر گرفته شوند(۱۴).

در مطالعه ساکی مالحی و همکاران در سال ۱۳۹۰ که در بیماران مبتلا به سرطان پستان انجام گرفت، مدل رده‌بندی درختی نشان داد که متغیرهای سابقه خانوادگی سرطان پستان، سابقه خانوادگی سرطان تخمدان، سابقه بیماری خوش‌خیم پستان، سن منارک، وضعیت قاعدگی، نداشتن فعالیت فیزیکی و استرس‌های ناشی از متارکه با همسر به عنوان عامل خطر سرطان پستان شناسایی شدند(۱۵).

ساکی مالحی و همکاران در سال ۱۳۹۱ به بررسی پیش‌گویی کننده‌های بقای بیماران در سرطان کولورکتال بر اساس مدل رده‌بندی و رگرسیون درختی پرداختند. بر اساس مدل درخت تصمیم، مرحله سرطان در زمان تشخیص، سن بیمار در زمان تشخیص، نوع مرفولوی تومور و درجه سرطان به عنوان پیش آگهی‌های مهم در بقای بیماران مبتلا به سرطان کولورکتال شناخته شد(۱۶).

در مطالعه Stephen و همکاران مدل رده بندی درختی به منظور تعیین پیش‌گویی کننده‌های سرطان‌های بدخیم پروستات استفاده شد، در این مدل، اطلاعات ۱۰۶۷ بیمار (نژاد بیمار، سابقه خانوادگی، وازکتومی، سن، آنتی ژن اختصاصی پروستات سرم و....) به منظور تشکیل یک ساختار پیشگویی کننده در عوامل اثرگذار بر سرطان بدخیم پروستات مورد استفاده قرار گرفت. مدل رده بندی درختی، آنتی ژن اختصاصی سرم ۰ng/ml/cc/۰۵۸ و کمتر از آن را به عنوان اولین گره مد نظر قرار داد. در مرحله بعدی تجزیه و تحلیل، گره بعدی آنتی ژن اختصاصی سرم بیشتر از ۰/۰۵۸ و کمتر یا مساوی ng/ml/cc ۰/۱۶۵ تعیین گردید. در گام بعدی بیمارانی که کمتر یا مساوی ۵۷/۵ سال سن داشتند نیز در معرض خطر کمتری برای بدخیمی سرطان پروستات در مدل تعیین گردیدند. در آخرین مرحله نیز بیماران بزرگتر از ۵۷/۵ سال بر مبنای حجم کلی پروستات بیشتر از ۲۲/۷ cc به دو زیر گروه تقسیم شدند و به عنوان های برگ های مدل درختی تعیین گردیدند(۱۷).

در مطالعه KIM و همکاران در سال ۲۰۱۰، از مدل رده بندی درختی در ارزیابی بیماران مبتلا به سرطان ریه مرتبط با شغل استفاده گردید، بر مبنای این مدل بهترین پیش گویی کننده بر ابتلا به سرطان ریه مرتبط با شغل، مواجهه با کارسینوژن‌های شناخته شده این بیماری بود. در دومین مرحله، عدم تشخیص سرطان به مدت ۸/۶ سال و یا بیشتر و سومین پیشگویی کننده سابقه مصرف سیگار کمتر از ۱۱/۲۵ بسته در سال تشخیص داده شد(۱۸).

در این مطالعه سابقه مواجهه با CT اسکن و اشعه X (عکس رنگی) از عوامل اثرگذار بر ابتلا به سرطان مری مطرح بوده است که یافته‌های سایر مطالعات آن را تأیید می‌نماید(۲۰و۱۹). اگر چه استفاده از اشعه‌های تشخیصی در شناسایی زودرس بیماری‌ها نقش بسیار ارزشمندی را داشته است، اما به نظر می‌رسد اثرات سوء استفاده دراز مدت از آن‌ها کمتر مورد بررسی و واکاوی قرار گرفته است.

بر مبنای یافته های این مطالعه، سابقه سرطان در خانواده یا بستگان به عنوان عامل خطر در ابتلا به سرطان مری تأیید شده است. در مطالعه Gao و همکاران در سال ۲۰۰۹ و Wu و همکاران در سال ۲۰۱۱ سابقه خانوادگی هر گونه سرطان و سابقه خانوادگی سرطان‌های دستگاه گوارش در ابتلا به سرطان‌های دستگاه گوارش مورد تأیید قرار گرفته است(۲۲ و۲۱). برخی عوامل ژنتیکی در کنار عوامل تشدید کننده محیطی می‌تواند منجر به ابتلا به برخی سرطان‌ها در نسل‌های متوالی در یک خانواده گردد.

یافته‌های این مطالعه هم راستا با سایر مطالعات، سابقه کشیدن سیگار را به عنوان یکی از عوامل اثرگذار بر سرطان مری تأیید

بر مبنای یافته‌های این مطالعه، نشستن دست ها با صابون یا مایع دستشویی پس از اجابت مزاج بر ابتلا به سرطان مری اثرگذار بوده است. در مطالعه Lee و همکاران در سال ۲۰۱۲ متغیرهای استفاده از آب های زیر زمینی (OR =۳/۴)، استفاده از آب تصفیه نشده فاضلاب (OR =۲/۸) و نشستن مکرر دست ها پس از اجابت مزاج (OR=۳/۵) بر افزایش بروز عفونت با هلیکوباکترپیلوری اثرگذار بوده‌اند (۲۶) و نقش هلیکوباکترپیلوری در بروز سرطان معده و زخم‌های معده مورد تأیید است(۲۷).

در این مطالعه، ضمن معرفی الگوریتم‌های مختلف مدل رده‌بندی درختی، بعضی از مؤلفه‌های سبک زندگی مانند بهداشت فردی، در معرض CT اسکن و اشعه X (عکس رنگی)، سابقه سرطان در خانواده و سابقه کشیدن سیگار به عنوان عوامل خطر شناسایی شدند که برای بعضی از عوامل فوق، در سایر مطالعات کمتر به آن پرداخته شده است و بایستی مورد توجه بیشتر قرار گیرند. دو مزیت اساسی استفاده از مدل‌های درختی نسبت به بسیاری از مدل‌های پیش بینی، سهولت در تفسیر نتایج و غیرخطی بودن آن است که علاوه بر ایجاد امکان استفاده آن در حالاتی که تعداد عوامل اثرگذار زیاد است، به دلیل سهولت در تفسیر نتایج، این روش می‌تواند علاوه بر متخصصین آماری، مورد استفاده پزشکان و پیراپزشکان نیز قرار گیرد.

تشکر و قدردانی:

نویسندگان بر خود فرض می‌دانند که از پرسنل شبکه بهداشت و درمان استان گلستان و بیماران و خانواده‌های ایشان و افراد شاهدی که در این مطالعه شرکت داشته‌اند، تشکر نمایند. این مقاله برگرفته از رساله دکتری تخصصی (PhD) رشته آمار زیستی دانشگاه تربیت مدرس است.

References:

1. Agresti A. An Introduction to Categorical Data Analysis. 2 edition. Hoboken, NJ: Wiley-Interscience; 2007. 400 p.

2. Lee SK. On Classification and Regression Trees for Multiple Responses and Its Application. J Classif. 2006 Jun 1; 23(1):123–41.

3. Kass GV. An Exploratory Technique for Investigating Large Quantities of Categorical Data. Appl Stat. 1980; 29(2):119.

4. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. 1 edition. New York, N.Y.: Chapman and Hall/CRC; 1984. 368 p.

5. Quinlan JR. Induction of decision trees. Mach Learn. 1986 Mar 1;1(1):81–106.

6. Loh W., Shih Y. Split Selection Methods for Classifcation Trees. Published in Statistica Sinica,. 1997; 7:815–40.

7. Kim H, Loh W. Classification trees with unbiased multiway splits. Journal of the American Statistical Association. 2001; 96:589–604.

8. DeMaris A. Regression with Social Data: Modeling Continuous and Limited Response Variables. John Wiley & Sons; 2004. 563 p.

9. Ghasem Ahmad L. Using Data Mining Techniques for Prediction Breast Cancer Recurrence. Iranian Journal of Breast Disease. 2013; 5(4):23–34.

10. Ghasem Ahmad L. Review top 7 Algorithms in Data Mining for Prediction Survivability, Diagnosis and Recurrence of Breast Cancer. Iranian Journal of Breast Disease. 2013; 6(1):52–61.

11. Silvera SAN, Yale University. Dietary factors and risk of subtypes of esophageal and gastric cancer. Diss Abstr Int.

12. Navarro Silvera SA, Mayne ST, Gammon MD, Vaughan TL, Chow W-H, Dubin JA, et al. Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. Ann Epidemiol. 2014 Jan; 24(1):50–7.

13. Valera VA, Walter BA, Yokoyama N, Koyama Y, Iiai T, Okamoto H, et al. Prognostic groups in colorectal carcinoma patients based on tumor cell proliferation and classification and regression tree (CART) survival analysis. Ann Surg Oncol. 2007 Jan; 14(1):34–40.

14. Camp NJ, Slattery ML. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). Cancer Causes Control CCC. 2002 Nov; 13(9):813–23.

15. Saki A, Hajizadeh E, Tehranian N. Evaluating the Risk Factors of Breast Cancer Using the Analysis of Tree Models. 2014 Jun 5; Available from: http://www.academia.edu/659307/Evaluating_the_Risk_Factors_of_Breast_Cancer_Using_the_Analysis_of_Tree_Models

16. Saki Malehi A, Hajizadeh E, Fatemi R. Evaluation of Prognostic Variables for Classifying the Survival In Colorectal Patients using The Decision Tree. Iran J Epidemiol. 2012 Sep 15; 8(2):13–9.

17. Spurgeon SEF, Hsieh Y-C, Rivadinera A, Beer TM, Mori M, Garzotto M. Classification and regression tree analysis for the prediction of aggressive prostate cancer on biopsy. J Urol. 2006 Mar; 175(3 Pt 1):918–22.

18. Kim T-W, Koh D-H, Park C-Y. Decision tree of occupational lung cancer using classification and regression analysis. Saf Health Work. 2010 Dec;1(2):140–8.

19. Parkin DM, Darby SC. 12. Cancers in 2010 attributable to ionising radiation exposure in the UK. Br J Cancer. 2011 Dec 6; 105(S2):S57–S65.

20. Bauer S, Gusev BI, Pivina LM, Apsalikov KN, Grosche B. Radiation exposure due to local fallout from Soviet atmospheric nuclear weapons testing in Kazakhstan: solid cancer mortality in the Semipalatinsk historical cohort, 1960-1999. Radiat Res. 2005 Oct; 164(4 Pt 1):409–19.

21. Wu M, Zhang Z-F, Kampman E, Zhou J-Y, Han R-Q, Yang J, et al. Does family history of cancer modify the effects of lifestyle risk factors on esophageal cancer? A population-based case-control study in China. Int J Cancer J Int Cancer. 2011 May 1; 128(9):2147–57.

22. Gao YT, Hu N, Han X, Giffen C, Ding T, Goldstein A, et al. Family history of cancer and risk for esophageal and gastric cancer in Shanxi, China. BMC Cancer. 2009 Aug 5; 9(1):269.

23. Freedman ND, Abnet CC, Leitzmann MF, Mouw T, Subar AF, Hollenbeck AR, et al. A prospective study of tobacco, alcohol, and the risk of esophageal and gastric cancer subtypes. Am J Epidemiol. 2007 Jun 15; 165(12):1424–33.

24. Gao YT, McLaughlin JK, Gridley G, Blot WJ, Ji BT, Dai Q, et al. Risk factors for esophageal cancer in Shanghai, China. II. Role of diet and nutrients. Int J Cancer J Int Cancer. 1994 Jul 15; 58(2):197–202.

25. Zeka A, Gore R, Kriebel D. Effects of alcohol and tobacco on aerodigestive cancer risks: a meta-regression analysis. Cancer Causes Control CCC. 2003 Nov; 14(9):897–906.

26. Lee YY, Ismail AW, Mustaffa N, Musa KI, Majid NA, Choo KE, et al. Sociocultural and dietary practices among Malay subjects in the northeastern region of Peninsular Malaysia: a region of low prevalence of Helicobacter pylori infection. Helicobacter. 2012 Feb; 17(1):54–61.

27. Lee YY, Mahendra Raj S, Graham DY. Helicobacter pylori infection--a boon or a bane: lessons from studies in a low-prevalence population. Helicobacter. 2013 Oct; 18(5):338–46.

Original article

The Introduction and Application of Classification Tree Model for Determination of Risk Factor for Esophageal Cancer in Golestan Province

Nasser Behnampour[1], Ebrahim Hajizadeh[2*], Shahriar Semnani[3] ,Farid Zayeri[4]

1. PhD student in Biostatistics, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
2. Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
3. Gastroenterology and Hepatology Research Center, Golestan University of Medical Sciences, Gorgan, Iran
4. Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Abstract

Background & objective: *One of the common purposes of medical research is Determination of effective factors on the occurrence of event. Due to the interaction of risk factors regression models, discriminant analysis and classification procedures used. Uses of these models require making the assumption which in the medical data isn't usually established. Therefore, alternative methods must be used. According to diversification of risk factors for of esophageal cancer, the purpose of this article is the Introduction and application of classification and regression tree for determination of risk factor for esophageal cancer in Golestan province.*

Methods: *Data of this article gathered from case-control study. Case group contain all confirmed cases of esophageal cancer that consist of 90 male and 60 female subjects in Golestan province during one year. Two control groups were considered for each case. Control groups were selected from family of patients and neighbors and matched for age, sex, ethnic and place of residence. Data was analyzed with classification and regression tree model and by using of R software. Gini criterion was used for selection of best splitting in each node and ROC surveyed accuracy of CRT model.*

Results: *Results of Classification tree model showed that exposure to CT and X-ray dye (socio-environmental factors), unwashed hands after defecation, history of smoking (lifestyle factors) and family history of cancer (ethnic factors) can be effective in esophageal cancer occurrences.*

Conclusion: *Tree models don't require the establishment of no default for making model and feasibility of tree models results` interpretation are two essential beneficiary of these models which can use in medical sciences.*

Key words: *Esophageal cancer, Classification tree model, Golestan*

*Corresponding author: Ebrahim Hajizadeh (PhD)
Address: Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University Tehran, Iran.
Phone: 02182884524 Email:hajizadeh@modares.ac.ir