نمایش دانش پنهان از یک مجموعه داده سرطان پستان بومی با استفاده از روش رگرسیون و دسته‌بندی

هادی لطف نژاد افشار [۱]، لیلی رحمت نژاد [۲]، بهلول رحیمی [۱]، حمیدرضا خلخالی [*۳]،

۱. گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران
۲. گروه مامایی، دانشکده پرستاری و مامایی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران
۳. گروه آمار زیستی، مرکز تحقیقات ایمنی بیمار ، دانشکده پزشکی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران

* **عهده دار مکاتبات:** ایران، ارومیه، دانشگاه علوم پزشکی ارومیه. دانشکده پزشکی، مرکز تحقیقات ایمنی بیمار، گروه آمار زیستی

Email: khalkhali@umsu.ac.ir

چکیده

زمینه: استفاده از روش‌های معمول کشف دانش مانند درخت‌های تصمیم‌گیری در زمینه سرطان پستان مورد مطالعه قرار گرفته است. نمایش ارتباط نامکشوف بین داده‌ها در قالب: بصری و ساختاربندی شده، از دلایل محبوبیت و استفاده از درخت‌های تصمیم‌گیری هستند. در مطالعه حاضر، الگوریتمی از این دسته که در مطالعات چاپ شده قبلی استفاده نشده است، به کار رفته است.

روش‌ها: مجموعه داده‌ای که شامل اطلاعات ۵۶۹ بیمار بین سال‌های ۸۷ و ۹۰ می باشد، استفاده شده است. روش مدیریت مقادیر مفقوده مجموعه داده جایگذاری متعدد بود. از نرم افزار IBM statistics 21 برای ایجاد مدل و جایگذاری مقادیر مفقوده استفاده شد. مدل ایجاد شده توسط معیارهای: صحت، حساسیت و ویژگی مورد ارزیابی قرار گرفت.

یافته‌ها: مدل درخت تصمیمی با هفده گره ایجاد کرد. از نُه گره، مجموعه ای از قوانین معنی‌دار بالینی که در قالب اگر و آنگاه بودند تولید شد. قوانین نشان دادند که مهمترین متغیر برای پیش‌بینی احتمال زنده بودن بیماران مبتلا به سرطان پستان، متغیر مرحله بیماری است. عملکرد مدل ایجاد شده بر طبق معیارهای (حساسیت، ویژگی و صحت) به ترتیب عبارت بودند از: ۹۳/۵، ۵۳ و ۸۰/۳ ٪.

نتیجه‌گیری: مدل مطالعه حاضر به عنوان اولین مدل ایجاد شده در زمینه احتمال زنده بودن بیماران مبتلا به سرطان پستان، قوانین کاربردی نامکشوف را از یک مجموعه داده نه چندان بزرگ آشکار کرد.

کلید واژه: سرطان پستان، بقا، یادگیری ماشین، تحلیل رگرسیونی

مقدمه:

پیش‌بینی بقای سرطان ها، بحث مهمی در حوزه داده کاوی است[۱]. چون سرطان پستان بعد از سرطان پوست، دومین سرطانی است که بیشترین شیوع را دارد و همچنین دومین سرطانی است که منجر به فوت بیماران می شود، لذا پیش‌بینی بقای سرطان پستان یک زمینه پر اهمیت در حوزه داده کاوی محسوب می‌شود[۳، ۲]. بقای سرطان پستان توسط روش‌های متعدد داده کاوی مورد بررسی قرار گرفته است. الگوریتم‌های درخت تصمیم‌گیری نیز که جزء الگوریتم‌های دسته‌بندی داده کاوی هستند، توسط برخی از پژوهش‌گران برای پیش‌بینی بقای سرطان پستان استفاده شده‌اند. در مطالعات انجام شده[۱۰-۴، ۱] در زمینه پیش‌بینی بقای بیماران مبتلا به سرطان پستان، ۶۰ ماه بعد از تشخیص سرطان، الگوریتم‌های درخت تصمیم‌گیری مانند: C5 یا j48، ID3 و CHAID (CHi-squared Automatic Interaction Detection) به همراه سایر الگوریتم‌های دسته بندی داده کاوی به کار رفته‌اند. در تمام مطالعات ذکر شده، به غیر از مطالعه[۶] از مجموعه داده SEER (Surveillance Epidemiology and End Results) استفاده شده است. مجموعه داده SEER اطلاعات مربوط به سرطان‌های شایع و

همچنین اطلاعات اجتماعی و جمعیت‌شناختی جامعه ایالات متحده را در بر می‌گیرد[11]. نمایش بصری درخت‌های تصمیم‌گیری و قاعده‌مند کردن آنها در قالب قوانین اگر...آنگاه، به راحتی امکان پذیر است[12]. تحلیل درخت تصمیم‌گیری ایجاد شده و قوانین منتج از آن، درک واضح‌تری از الگوهای پنهان و ارتباط بین متغیرهای مجموعه داده در اختیار پژوهش‌گران قرار می‌دهد. این در حالی است که از میان مطالعات فوق‌الذکر، تنها مطالعه[8]، درخت تصمیم‌گیری و قوانین آن را گزارش و تحلیل کرده است. در پژوهش‌های انجام شده درباره پیش‌بینی بقای سرطان پستان، علاوه بر اینکه مدلی منتج از مجموعه داده بومی ایران وجود ندارد، بلکه هیچ مدلی توسط الگوریتم درخت‌های رگرسیون و دسته‌بندی نیز ایجاد نشده است. بنابراین، پتانسیل بالای این الگوریتم در پیش‌بینی و کشف داده از مجموعه داده های بالینی و همچنین شفافیت و وضوح بالایی که در ارائه دانش استخراجی به کاربران نهایی دارد، پژوهشگران مقاله حاضر را به ضرورت انجام مطالعه ای درباره کاربرد الگوریتم درخت‌های رگرسیون و دسته‌بندی در حوزه سرطان پستان ترغیب نمود. الگوریتم درخت‌های رگرسیون و دسته‌بندی یکی از الگوریتم‌های اصلی دسته‌بندی است که با روش تقسیم‌بندی بازگشتی (recursive partitioning) کار می کند و توسط Breiman [13] ایجاد شده است. این الگوریتم، داده‌های آموزشی را به روش بازگشتی، به مجموعه‌ای از بخش‌ها تقسیم بندی می‌کند، به طوریکه، مقادیر متغیر پاسخ یا پیش‌بینی شونده این بخش‌ها مشابه یکدیگر باشد. سپس، در هر بخش، متغیرهای پیش‌بینی کننده را بررسی می‌کند تا از میان آنها، بهترین جدا کننده آن بخش را پیدا کند. همه جدا کننده‌ها در هر مرحله تقسیم‌بندی، دودویی است و فرایند جدا سازی با ایجاد دو زیر گروه شروع می‌شود و به همین ترتیب ادامه می‌یابد تا بالاخره، یکی از معیارهای متوقف سازی باعث توقف فرایند شود. چون، بیشتر متغیرهای مجموعه داده مطالعه حاضر، گسسته و طبقه‌ای هستند، لذا از میان سایر الگوریتم‌های دسته‌بندی، الگوریتم درخت‌های رگرسیون و دسته‌بندی برای تجزیه و تحلیل انتخاب شد. هدف مطالعه حاضر، به کارگیری داده کاوی برای شناسایی الگوی پنهان و قوانین مهم از میان مجموعه داده مربوط به وضعیت بقای سرطان پستان در یک مرکز تحقیقاتی شهر ارومیه بین سال‌های ۸۷ و ۹۰ بود.

مواد و روش‌ها:

داده‌ها متعلق به مرکز تحقیقاتی و درمانی امید بود که یک موسسه خیریه برای حمایت از بیماران سرطانی آذربایجان غربی می‌باشد. متوسط پذیرش سالانه این مرکز ۷۰۰۰ بیمار است. بخش‌های خدماتی این مرکز عبارتند از: کلینیک تخصصی سرطان، رادیوتراپی، شیمی درمانی، فیزیک پزشکی، کلینیک دندانپزشکی، کلینیک چشم پزشکی و کلینیک روانپزشکی. بعد از مشاوره با متخصصان موسسه و مطالعه پژوهش‌های قبلی در حوزه پیش‌بینی بقای سرطان پستان[1,4-10]، متغیرهای مهم درباره بیماران مونث مبتلا به سرطان پستان از پرونده‌های پزشکی کاغذی استخراج شدند. جهت رعایت جنبه‌های اخلاقی، همه اطلاعات مربوط به بیماران بدون نام و نام خانوادگی بیماران استخراج شدند.

مطالعه توصیفی می‌باشد که به صورت مقطعی در سال ۱۳۹۴ انجام شد. مجموعه داده‌ای که در مطالعه حاضر تحلیل شد، در برگیرنده اطلاعات بقا، درمانی و جمعیت شناختی ۵۶۹ بیمار (با میانگین سنی ۴۸/۶ سال) بین سال‌های ۸۷ و ۹۰ بود. برای جلوگیری از تورش نمونه‌گیری، فقط اطلاعات پرونده‌هایی که با معیارهای ورودی مطالعه، مطابقت داشتند، لحاظ شدند. اطلاعات پرونده‌هایی که با معیارهای مطالعه حاضر مطابقت نداشتند، عبارت بودند از: بیماران با جنسیت مذکر، بیمارانی که کمتر از ۶۰ ماه مورد پی‌گیری قرار گرفته و هنوز زنده بودند و بیمارانی که کمتر از ۶۰ ماه مورد پیگیری قرار گرفته بودند و علت مرگ آنها، سرطان پستان نبود.

برای شناسایی داده‌های ناقص (داده های مفقوده و پرت) و توسعه مدل، داده‌ها وارد نرم افزار IBM SPSS Statistics ویرایش ۲۱ شدند. داده‌های ناقص تاثیر منفی بر عملکرد مدل دارند[14]. در مجموعه داده مطالعه حاضر، هیچ داده پرتی شناسایی نشد. ولی برخی از متغیرها، مقادیر مفقوده داشتند. توزیع داده‌های مفقوده در جدول یک و الگوی عددی آنها در جدول

دو نشان داده شده است. جدول دو، تعداد مقادیر مفقوده رکوردها را در بین متغیرها نشان می‌دهد. به عنوان مثال، در مجموعه داده مطالعه حاضر ۴۲۳ رکورد در هیچ متغیری مقادیر مفقوده نداشته‌اند یا به عبارت دیگر فاقد مقادیر مفقوده بوده‌اند، در حالی که، در ۳۰ رکورد، سه متغیر: گیرنده‌های ER، PR و Her2 فاقد داده بوده اند.

معمول‌ترین روش برای مدیریت داده‌های مفقوده، حذف آنهاست[۱۵]. مجموعه داده مطالعه حاضر، تقریباً مجموعه داده بزرگی نیست. بنابراین، حذف داده‌های مفقوده آن باعث کاهش هر چه بیشتر حجم مجموعه داده می‌شد. برای جلوگیری از ریزش داده‌ها، داده‌های مفقوده با روش جایگذاری متعدد جایگزین شدند. در این روش، داده‌های مفقوده با مقادیر چندگانه یا متعدد جایگذاری می‌شوند. چون تخمین دقیق مقادیر مفقوده از نظر علمی غیر ممکن است، لذا، برای مدیریت این عدم قطعیت، مقادیر چندگانه ایجاد می‌شوند[۱۶].برای انجام روش جایگذاری متعدد در نرم افزار IBM SPSS Statistics ویرایش ۲۱، الگوی داده‌های مفقوده، قبل از جایگذاری مورد تحلیل قرار گرفت. تحلیل نشان داد که داده‌های مفقوده به صورت تصادفی توزیع شده‌اند (شکل یک) و از لحاظ آماری می‌توان جایگذاری متعدد را اجرا کرد.هیستوگرام شکل یک، مقادیر مفقوده را به صورت نزولی نشان می‌دهد و در سمت راست محور ایکس(X) شکل یک، همه متغیرهای مجموعه داده (به غیر از متغیر هدف) قرار گرفته اند. همچنین اعدادموجود در سمت راست نشان دهنده نسبت مقادیر مفقوده در بین متغیرها می باشد. که ترتیب قرار گرفتن متغیرها بدین ترتیب است که متغیری که هیچ مقدار مفقوده یا کمترین مقدار را دارد در منتهی‌الیه سمت راست محور و متغیری که بیشترین مقدار مفقوده را دارد در منتهی‌الیه سمت چپ محور ایکس است. در محور وای(Y) نیز الگوی مقادیر مفقوده و نامفقوده قرار دارند. به عنوان مثال الگوی آخر شکل یک، هیچ مقدار مفقوده‌ای ندارد، در حالی که الگوی اول، موارد زیادی را در بر می‌گیرد که فقط در رابطه با متغیر Her2، مقادیر مفقوده دارند.چون در قسمت پایین و راست شکل (یک)، هیچ مقدار مفقوده‌ای دیده نمی‌شود و همچنین در قسمت فوقانی و چپ شکل (یک)، هیچ مقدار نامفقوده ای نیست، بنابراین، الگوی مقادیر مفقوده و نامفقوده از یکنواختی برخوردار است و لذا می‌توان جایگذاری متعدد را اجرا کرد. بعد از اجرای جایگذاری متعدد، پنج مجموعه دادهِ بدون مقادیر مفقوده تشکیل شدند.

مطابق با اهداف مطالعه حاضر، از الگوریتم درخت‌های رگرسیون و دسته بندی برای ایجاد مدل استفاده شد. این الگوریتم از یک فرایند بازگشتی دودویی (binary recursive) استفاده می‌کند. فرایند بازگشتی دودویی، زیر مجموعه‌های یک مجموعه داده کامل را که شامل همه متغیرهای پیش‌بینی کننده می‌شود، به دو گره فرزند تقسیم می‌کند. این تقسیم به صورت متداوم برای ایجاد گره‌های بیشتر انجام می‌گیرد. تعدادی از معیارهای اندازه‌گیری ناخالصی مانند: جینی (Gini)، توینگ (twoing) و توینگ ترتیبی (ordered twoing) برای انتخاب بهترین متغیر پیش‌بینی کننده مورد استفاده قرار می‌گیرند[۷-۹]. زیر مجموعه‌های تولید شده حتی‌الامکان باید نسبت به متغیر هدف متجانس باشند. به عبارت دیگر، توزیع متغیر هدف در آنها همگن باشد. در مطالعه حاضر، معیار جینی که برای متغیرهای گروهی کاربرد دارد، استفاده شد.خطای دسته‌بندی که تابعی از اندازه درخت تصمیم‌گیری است، با روش 10-fold cross-validation اندازه‌گیری شد[۱۳]. مجموعه داده به صورت تصادفی به ده مجموعه داده کوچکتر تقسیم‌بندی شد. برای انتخاب بهینه تعداد گره‌ها از درخت اصلی، از روش وجین وارونه (backward pruning) استفاده شد. بعد از ایجاد ده درخت تصمیم‌گیری، میزان خطای دسته‌بندی آنها محاسبه شده و درخت تصمیم گیری که کمترین میزان خطا را داشت به عنوان درخت بهینه انتخاب شد.

برای اجرای الگوریتم درخت‌های رگرسیون و دسته بندی از IBM SPSS Statistics ویرایش ۲۱ استفاده شد که این الگوریتم بر پنج مجموعه داده کاملی که با روش جایگذاری متعدد ایجاد شده بودند، اعمال گشت. مدلی که بالاترین امتیاز را

از لحاظ معیارهای ارزیابی (صحت، حساسیت و ویژگی) کسب کرده بود، مدل منتخب شد.

ارزیابی کارکرد مدل به وسیله سه معیار حساسیت، ویژگی و صحت بررسی شد. فرمول‌های این معیارها در ذیل آورده شده است[1]:

$$\text{صحت} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{\text{صحیح های منفی + صحیح های مثبت}}{\text{اشتباه های منفی + اشتباه های مثبت + صحیح های منفی + صحیح های مثبت}}$$

$$\text{حساسیت} = \frac{\text{TP}}{\text{TP+FN}} = \frac{\text{صحیح های مثبت}}{\text{اشتباه های منفی + صحیح های مثبت}}$$

$$\text{ویژگی} = \frac{\text{TN}}{\text{FP+TN}} = \frac{\text{صحیح های منفی}}{\text{صحیح های منفی + اشتباه های مثبت}}$$

نتایج:

الگوریتم درخت‌های رگرسیون و دسته‌بندی برای استخراج الگوی پنهان مجموعه داده سرطان پستان مورد استفاده قرار گرفت. قبل از اجرای الگوریتم، مجموعه داده به داده‌های آزمایشی و آموزشی تقسیم شد. معمولاً نسبت تشکیل داده‌های آموزشی و آزمایشی در مطالعات مربوط به داده کاوی ۷۰ به ۳۰ یا ۸۰ به ۲۰ است[۱۲،۱]. داده‌های آموزشی، ۷۰٪ داده‌های مجموعه داده این مطالعه را در بر گرفته بودند که الگوریتم برای یادگیری الگوی موجود در مجموعه داده از آنها استفاده می‌کرد. همچنین ۳۰٪ داده‌های مجموعه داده را داده‌های آزمایشی تشکیل دادند که درخت تصمیم‌گیری از داده‌های آزمایشی ایجاد شد و برای پیش‌بینی برچسب دسته (مقادیر متغیر بقا) روی داده‌های آزمایشی اعمال شد. شکل دو درخت تصمیم‌گیری ایجاد شده از الگوریتم درخت‌های رگرسیون و دسته‌بندی را نشان می دهد.

درخت تصمیم‌گیری ایجاد شده، به طور کلی هفده گره و نُه گره برگ دارد که هر گره برگ مرتبط با مجموعه قوانینی است. جدول سه مجموعه قوانین منتج از درخت تصمیم‌گیری را نشان می‌دهد.

برای نیل به دسته بندی صحیح بقا، همه قوانین استخراجی توسط متخصصین سرطان پستان، ارزیابی و تایید شدند. ترتیب قوانین جدول چهار، بر اساس مقدار معیار حساسیت است. با توجه به شکل دو، مشخص است که متغیر مرحله سرطان در بالاترین گره درخت تصمیم‌گیری می‌باشد. همچنین قوانین استخراج شده نیز این امر را تایید می‌کنند. لذا، این متغیر به عنوان مهم‌ترین متغیر پیش‌بینی کننده بقای سرطان پستان تعیین شده است.

به عنوان مثال، قانون مربوط به گره برگ شانزده بدین مضمون است: «اگر مقدار متغیر مرحله سرطان، کمتر یا مساوی با ۳C بود و مقدار متغیر تعداد گره های محلی مثبت، کمتر یا مساوی با نُه بود و مقدار متغیر اندازه تومور، کمتر یا مساوی با هفت بود و مقدار متغیر Her2 مساوی با بله بود و مقدار متغیر سن بیشتر از ۵۴ سال بود، آنگاه مقدار پیش‌بینی شده متغیر وضعیت بقا به دسته بله تعلق دارد».

مقادیر معیارهای ارزیابی شامل حساسیت، ویژگی و صحت مدل مطالعه حاضر، به ترتیب عبارت از ۹۳/۵٪، ۵۳٪ و ۸۰/۳٪ است.

جدول (۱):توزیع داده های مفقوده

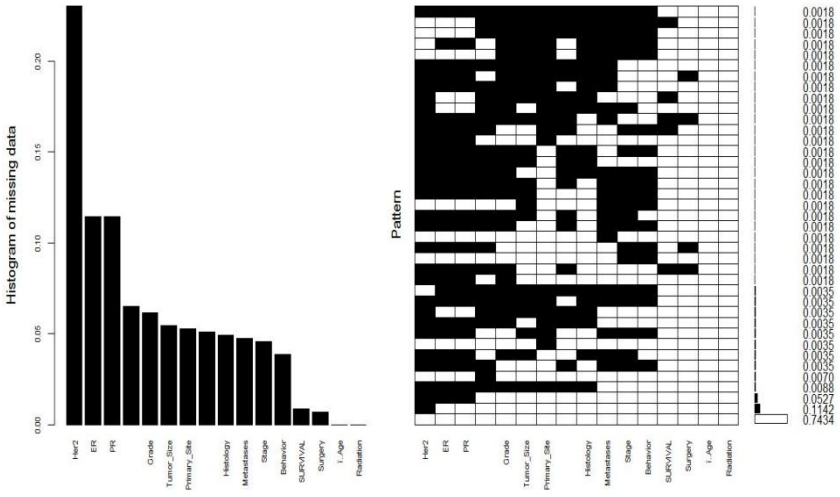
نام متغیر	تعداد مقادیر مفقوده	٪ مقادیر مفقوده
مکان اولیه تومور (Primary Site)	۳۰	۵/۳
بافت شناسی (Histology)	۲۸	۴/۹
اندازه تومور (Tumor Size)	۳۱	۵/۴
متاستاز (Metastases)	۲۷	۴/۷
مرحله تومور (Stage)	۲۶	۴/۶
رفتار تومور (Behavior)	۲۲	۳/۹
درجه تومور (Grade)	۳۵	۶/۲
گره های محلی مثبت (Positive Regional Node)	۲۹	۵/۱
گره های محلی برداشته شده (Removed Regional Node)	۳۷	۶/۵
جراحی (Surgery)	۴	۰/۷
گیرنده Her2	۱۳۱	۲۳
گیرنده ER	۶۵	۱۱/۴
گیرنده PR	۶۵	۱۱/۴
وضعیت بقا (Survival Status)	۵	۰/۹

جدول (۲):الگوی عددی داده های مفقوده

*	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	**
۴۲۳	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۰
۲	۱	۱	۱	۱	۱	۱	۱	۱	۱	۰	۱	۱	۱	۱	۱	۱	۱
۱	۱	۱	۱	۱	۱	۱	۰	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱
۴	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۰	۱	۱	۱	۱
۶۵	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۰	۱
۱	۱	۱	۱	۱	۰	۰	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۲
۳۰	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۰	۰	۰	۳
۱	۱	۱	۱	۱	۰	۰	۰	۱	۱	۱	۰	۱	۱	۱	۱	۱	۴
۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۰	۱	۱	۱	۰	۰	۰	۴
۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۰	۱	۰	۰	۰	۴
۱	۱	۱	۱	۱	۰	۰	۰	۰	۱	۰	۰	۰	۱	۰	۱	۱	۷
۲	۱	۱	۱	۱	۱	۱	۱	۰	۰	۰	۰	۰	۰	۱	۱	۰	۷
۱	۱	۱	۰	۱	۰	۰	۱	۱	۱	۱	۱	۱	۰	۰	۰	۰	۷
۱	۱	۱	۱	۱	۱	۰	۰	۰	۰	۰	۱	۰	۰	۱	۱	۰	۸
۲	۱	۱	۱	۱	۰	۰	۰	۱	۱	۰	۰	۱	۱	۰	۰	۰	۸
۲	۱	۱	۱	۱	۱	۰	۰	۰	۱	۱	۰	۰	۱	۰	۰	۰	۸
۲	۱	۱	۱	۱	۰	۰	۰	۱	۰	۱	۱	۱	۰	۰	۰	۰	۸
۲	۱	۱	۱	۱	۱	۱	۱	۰	۰	۰	۱	۰	۰	۰	۰	۰	۸
۱	۱	۱	۱	۱	۱	۱	۱	۰	۰	۱	۰	۰	۰	۰	۰	۰	۸
۱	۱	۱	۱	۰	۱	۱	۱	۰	۰	۰	۰	۰	۰	۱	۱	۰	۸

۱	۱	۱	۰	۰	۱	۱	۱	۱	۰	۱	۱	۱	۱	۰	۰	۰	۸
۱	۱	۱	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱	۱	۰	۹
۱	۱	۱	۱	۰	۰	۰	۰	۱	۰	۰	۰	۰	۱	۱	۱	۰	۹
۱	۱	۱	۱	۱	۰	۰	۰	۱	۰	۱	۱	۱	۱	۱	۰	۰	۹
۱	۱	۱	۱	۰	۰	۰	۱	۱	۱	۰	۰	۰	۱	۱	۰	۰	۹
۵	۱	۱	۱	۱	۱	۱	۰	۰	۰	۱	۱	۱	۱	۱	۰	۰	۹
۱	۱	۱	۱	۱	۰	۰	۱	۰	۱	۰	۰	۱	۱	۱	۰	۰	۹
۱	۱	۱	۱	۱	۰	۰	۱	۰	۱	۰	۰	۰	۱	۱	۰	۰	۹
۱	۱	۱	۰	۱	۱	۰	۰	۱	۰	۱	۰	۰	۱	۱	۰	۰	۹
۱	۱	۱	۱	۱	۱	۱	۱	۰	۱	۱	۱	۱	۱	۱	۱	۱	۱۰
۱	۱	۱	۱	۱	۰	۰	۰	۱	۰	۰	۰	۱	۱	۱	۰	۰	۱۰
۱	۱	۱	۱	۱	۰	۰	۰	۱	۰	۰	۰	۰	۱	۱	۰	۰	۱۰
۱	۱	۱	۱	۱	۰	۰	۰	۱	۰	۰	۰	۰	۱	۱	۰	۰	۱۰
۱	۱	۱	۰	۱	۰	۰	۰	۰	۰	۱	۱	۱	۰	۱	۰	۰	۱۰
۱	۱	۱	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱	۱	۱	۱۰
۲	۱	۱	۱	۱	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱	۰	۱	۱۱
۲	۱	۱	۱	۱	۰	۰	۰	۱	۰	۰	۰	۰	۱	۱	۰	۰	۱۱
۱	۱	۱	۰	۰	۱	۰	۱	۰	۱	۰	۱	۱	۰	۰	۰	۰	۱۱
۱	۱	۱	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱	۱	۰	۰	۱۲
	۰	۰	۴	۵	۲۲	۲۶	۲۷	۲۸	۲۹	۳۰	۳۱	۳۵	۳۷	۶۵	۶۵	۱۳۱	۵۳۵

گره های =IX ,بافت شناسی =VIII ,متاستاز =VII ,مرحله تومور =VI ,رفتار تومور =V ,وضعیت بقا =IV ,جراحی =III ,اشعه درمانی =II ,سن =I ,گیرنده PR =XV ,گیرندهER =XIV ,گره های محلی برداشته شده =XIII ,درجه تومور =XII ,اندازه تومور =XI ,مکان اولیه تومور =X ,محلی مثبت تعداد متغیرهای دربرگیرنده مقادیر مفقوده =** ,تعداد مقادیر مفقوده = * ,گیرنده Her2 =XVI



شکل ۱: الگوی داده های مفقوده

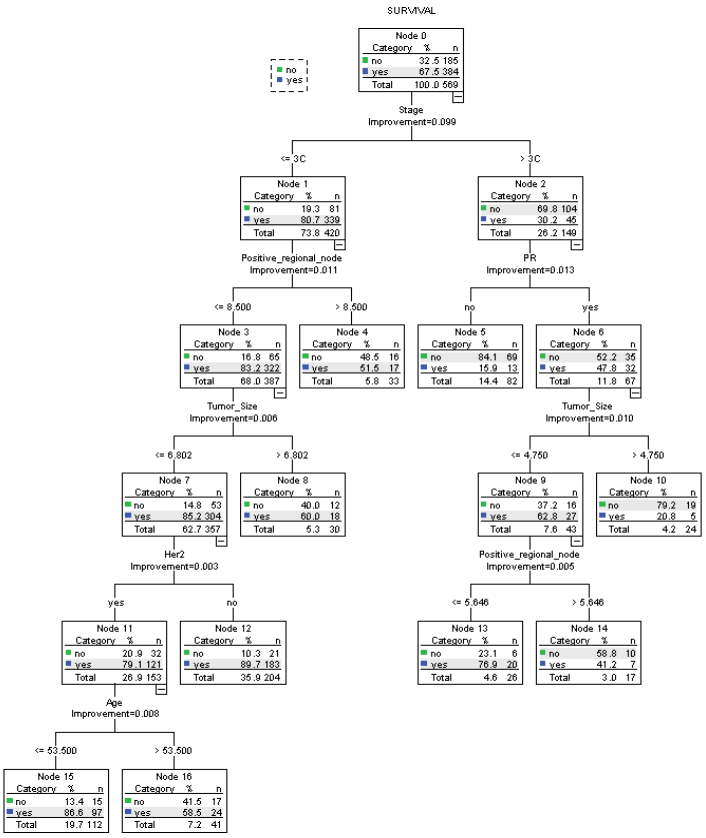
جدول ۳ متغیرهای استفاده شده در توسعه مدل و همچنین آمار توصیفی آنها را بعد از اجرای جایگذاری متعدد نشان می دهد. در مطالعه حاضر، ۱۵ متغیر پیش بینی کننده و یک متغیر پاسخ (وضعیت بقا) وجود دارد.

جدول ۳:آمار توصیفی متغیرهای مطالعه حاضر

متغیرهای گروهی	مقادیر	تعداد	٪
مکان اولیه تومور	Central	۱۰۴	۱۸/۳
	UIQ[1]	۲۱۷	۳۸/۱
	LIQ[2]	۸۳	۱۴/۶
	UOQ[3]	۱۱۲	۱۹/۷
	LOQ[4]	۵۳	۹/۳
متاستاز تومور	Yes	۱۵۸	۲۷/۸
	No	۴۱۱	۷۲/۲
رفتار تومور	In situ	۲۰۸	۳۶/۶
	Malignant	۳۶۱	۶۳/۴
	I	۱۸۷	۳۲/۹
درجه تومور	II	۳۰۲	۵۳/۱
بافت شناسی تومور	III	۸۰	۱۴
	Adenoid	۶	۱/۱
	DCI[5]	۱۷	۳
	Epithelial	۳	۰/۵
	IDC[6]	۴۸۶	۸۵/۴
	Paget disease	۷	۱/۲
	Comedo	۶	۱/۱
	ILC[7]	۱۵	۲/۶
	Inflammatory	۲	۰/۴
	Mucinous	۵	۰/۹
	Papillary	۵	۰/۹
مرحله تومور	Micropapillary	۵	۰/۹
	Medullary	۱۰	۱/۸
	Phyllodes	۲	۰/۴
	۱	۱۸	۳/۲
	۲A	۱۲۳	۲۱/۶
	۲B	۱۰۲	۱۷/۹
	۳A	۱۵۱	۲۶/۵
	۳C	۲۶	۴/۶
	۴	۱۴۹	۲۶/۲
	No	۱۰۹	۱۹/۲
جراحی	Lumpectomy	۲	۰/۴
	Quadrantectomy	۷	۱/۲
	MRM8	۴۴۷	۷۸/۶
	Radical mastectomy	۴	۰/۷

	Yes	۵۰۶	۸۸/۹
اشعه درمانی	No	۶۳	۱۱/۱
	Yes	۲۵۵	۴۴/۸
گیرنده Her2	No	۳۱۴	۵۵/۲
	Yes	۲۶۱	۴۵/۹
گیرنده ER	No	۳۰۸	۵۴/۱
	Yes	۲۸۸	۵۰/۶
گیرنده PR	No	۲۸۱	۴۹/۴
وضعیت بقا	Yes	۳۸۴	۶۷/۵
	No	۱۸۵	۳۲/۵
متغیرهای پیوسته	میانگین	انحراف معیار	بازه
سن	۴۸/۶	۱۰/۶	۲۵–۸۲
اندازه تومور	۴	۱/۹	۰/۵–۱۶/۸
تعداد گره های محلی مثبت	۳/۶	۴/۳	۰–۲۶
تعداد گره های محلی برداشته	۸/۴	۵/۷	۱–۳۸

۱:Upper inner quadrant ۲:Lower inner quadrant ۳:Upper outer quadrant۴:Lower outer quadrant۵:Ductal carcinoma in situ۶:Invasive ductal carcinoma۷:Invasive lobular carcinoma۸:Modified radical mastectomy



جدول ٤: قوانین استخراج شده برای هر یک از گروه‌های متغیر بقا

وضعیت بقا	گره برگ	شماره قانون	محتوی قانون	حساسیت (٪)
yes	۱۲	۱	اگر مرحله سرطان C<=۳ و تعداد گره های محلی مثبت <= ۹ و اندازه تومور <= ۷ و Her2= خیر	۸۹/۷
	۱۵	۲	اگر مرحله سرطان C<=۳ و تعداد گره های محلی مثبت <= ۹ و اندازه تومور <= ۷ و Her2= بله و سن <= ۵۴	۸۶/۶
	۱۳	۳	اگر مرحله سرطان C>۳ و PR= بله و اندازه تومور <= ۵ و تعداد گره های محلی مثبت <= ۶	۷۶/۹
	۸	۴	اگر مرحله سرطان C<=۳ و تعداد گره های محلی مثبت <= ۹ و اندازه تومور > ۷	۶۰
	۱۶	۵	اگر مرحله سرطان C<=۳ و تعداد گره های محلی مثبت <= ۹ و اندازه تومور <= ۷ و Her2= بله و سن > ۵۴	۵۸/۵
	۴	۶	اگر مرحله سرطان C<=۳ و تعداد گره های محلی مثبت > ۹	۵۱/۵
no	۵	۱	اگر مرحله سرطان C>۳ و PR= خیر	۸۴/۱
	۱۰	۲	اگر مرحله سرطان C>۳ و PR= بله و اندازه تومور > ۵	۷۹/۲
	۱۴	۳	اگر مرحله سرطان C>۳ و PR= بله و اندازه تومور <= ۵ و تعداد گره های محلی مثبت > ۶	۵۸/۸

بحث:

در مطالعه حاضر، مهم‌ترین متغیر برای پیش‌بینی وضعیت بقای سرطان پستان، مرحله سرطان شناخته شد. به هرحال، این نتیجه، در مطالعات قبلی حوزه داده کاوی سرطان پستان، به دست نیامده است. در مطالعهkyng-cheng و همکاران[۸]، متغیر تعداد گره‌های محلی مثبت، مهم‌ترین متغیر پیش‌بینی کننده وضعیت بقای سرطان پستان تعیین شد. در مطالعه حاضر، متغیر تعداد گره‌های محلی مثبت، بعد از متغیر مرحله سرطان، مهم‌ترین متغیر پیش‌بینی کننده تعیین شد. این عدم هماهنگی ممکن است ناشی از ماهیت الگوریتم‌های به کار رفته شده یا مجموعه داده استفاده شده باشد. گرچه، عدم تناسب بین مهم ترین متغیرهای به دست آمده در مطالعه حاضر و مطالعه کونگ جنگ ، دیده می‌شود، اما، تناسب‌هایی نیز در برخی از متغیرهای تشکیل دهنده قوانین استخراجی این دو مطالعه وجود دارد. متغیرهای: مرحله سرطان، تعداد گره‌های محلی مثبت، اندازه تومور و Her2(human epidermal growth factor receptor 2) اولین قانون مطالعه حاضر را بابالاترین مقدار حساسیت (۸۹/۷٪) تشکیل دادند. سه متغیر از میان متغیرهای فوق (مرحله سرطان، تعداد گره های محلی مثبت و اندازه تومور) در اولین قانون استخراجی مطالعه kyng-chengنیز دیده می‌شد. در مطالعه حاضر، Her2 یکی از مهم‌ترین متغیرها برای پیش‌بینی وضعیت بقا مشخص شد. این متغیر در حقیقت ژنی است که نقش مهمی در گسترش سرطان پستان دارد[۱۷]. چون متغیر Her2 در مجموعه داده SEER وجود ندارد، لذا در مطالعه kyng-chengنیز هیچ گزارشی از آن ارائه نشده است.

توزیع متغیر دسته (متغیر بقا) در مطالعه kyng-cheng نسبت به مطالعه حاضر تعادل کمتری داشت. به عبارت دیگر، یکی از مقادیر دسته نسبت به مقدار دیگر آن، پر تعدادتر بود. در چنین مواقعی، عملکرد بیشتر الگوریتم‌های داده کاوی با تورش همراه استو نتایج واقعی را انعکاس نمی‌دهند[۶]. توزیع متغیر دسته در مطالعه آنها، ۹۰/۶۸٪ (بقا) به ۹/۳۲٪ (عدم بقا) بود. این توزیع در مطالعه حاضر، ۶۷/۵٪ (بقا) به ۳۲/۵٪ (عدم بقا) بود. لازم به ذکر است که در مطالعه kyng-cheng از روش SMOTE (synthetic minority oversampling technique) برای مدیریت این چالش استفاده شد.

در مطالعاتی که در حوزه داده کاوی بقای سرطان پستان در مجموعه داده SEER انجام شده است[۹-۱،۴]، به غیر از مطالعه احمدی[۱۰]، در هیچ یک از آنها، جایگذاری مقادیر مفقوده انجام نگرفته است. چون مجموعه داده SEER تعداد قابل توجهی

رکورد دارد، در مطالعات فوق الذکر ترجیح داده شده است که به جای جایگذاری مقادیر مفقوده، این مقادیر حذف شوند. در مطالعه حاضر، جایگذاری مقادیر مفقوده از ریزش قابل توجه اطلاعات درباره وضعیت بقای سرطان پستان جلوگیری کرده است. دلیل دیگر جایگذاری مقادیر مفقوده، پایین بودن تعداد رکوردهای مجموعه داده می باشد. روش به کار رفته برای جایگذاری مقادیر مفقوده در مطالعه احمدی[10] و مطالعه حاضر جایگذاری متعدد بود.

مقدار معیار حساسیت مدل مطالعه حاضر که برابر با ۹۳/۵ ٪ بود، از مقدار به دست آمده در مطالعات احمدی[10]، kyng[8] و Thongkam[6] بالاتر بود. در حالیکه نسبت به مقدار حساسیت به دست آمده در مطالعات Endo،[1]Delen،[7]Lieou[5] و یکی از مطالعات دیگر Kyng[9] پایین‌تر بود. این معیار در مطالعه Bellachia[4] گزارش نشده بود. در میان مطالعات انجام شده در زمینه بقای سرطان پستان، معیار ویژگی مطالعه حاضر (۵۳٪) بالاتر از مقدار ویژگی مطالعه های Lieou[7] و Endo[5] بود. لازم به ذکر است که در مطالعه Bellachia[4] این مقدار گزارش نشده است. بر اساس نتایج به دست آمده، واضح است که مدل ایجاد شده در مطالعه حاضر، پیش بینی وضعیت بقا را نسبت به وضعیت عدم بقا بهتر انجام داده است. معیار صحت همه مطالعه ها به غیر از مطالعه Kyng[8] از معیار صحت مطالعه حاضر (۸۰/۳٪) بالاتر بودند. با اینکه مقادیر معیارهای ارزیابی: صحت و ویژگی مطالعه حاضر از بیشتر مطالعات پایین تر بودند. لذا، این مساله، محدودیتی برای مطالعه حاضر محسوب نمی شود. چون، معیارهای حساسیت، ویژگی و صحت، معیارهای کامل و بدون نقصی برای قضاوت راجع به عملکرد یک مدل محسوب نمی شوند.

یکی از محدودیت‌هایی که پژوهش حاضر با آن مواجه بود، حجم مجموعه داده استفاده شده بود. بنابراین، داده‌های آموزشی این مجموعه داده نیز بیشتر نبودند. ارتباط مستقیمی بین قدرت یادگیری الگوریتم‌های داده کاوی و تعداد داده‌های آموزشی وجود دارد[18، 12]. محدودیت دیگر در این مطالعه، حجم مقادیر مفقوده متغیر Her2 بود. گرچه، Her2 تنها متغیر در مطالعه حاضر بود که مقادیر مفقوده آن بالاتر از ۲۰ ٪ بود که آنها هم توسط یکی از روش‌های جایگذاری پر شد.

نتیجه‌گیری:

هدف اصلی مطالعه حاضر، استخراج الگو و قوانین با اهمیت درباره بقای سرطان پستان از یک مجموعه داده بومی با استفاده از یک الگوریتم داده کاوی بود. نتیجه اصلی این مطالعه نشان داد که متغیر مرحله سرطان، مهم‌ترین متغیر پیش‌بینی کننده بقای سرطان پستان است. یافته دیگر این مطالعه نشان داد که مدل ایجاد شده، وضعیت بقای بیماران مبتلا به سرطان پستان را که احتمال زنده ماندن آنها بالاتر است، نسبت به بیماران با احتمال بقای پایین تر را بهتر پیش بینی می کند. نتیجه‌گیری کلی که از مطالعه حاضر قابل استنباط است، تاثیر داده های مفقوده و عدم تعادل در تعداد برچسب دسته بر نتایج پژوهش است که در صورت امکان باید از جانب پژوهشگران مدیریت شوند تا از اثرات منفی آنها جلوگیری به عمل آید. این مطالعه اولین پژوهشی است که از الگوریتم درخت‌های رگرسیون و دسته بندی در یک مجموعه داده بومی مربوط به ایران استفاده کرده است. در پژوهش‌های به چاپ رسیده قبلی در زمینه داده کاوی بقای سرطان پستان، از این الگوریتم استفاده نشده است. تحقیق بعدی برای به کارگیری الگوریتم درخت های رگرسیون و دسته بندی در مجموعه داده ای با رکوردهای بیشتر و همچنین مجموعه داده ای که در برگیرنده اطلاعات درباره بیمارانی باشد که به مدت ۱۰ یا ۱۵ سال بعد از تشخیص سرطان پستان مورد پی گیری منظم قرار گرفته اند، ضروری به نظر می رسد.

تشکر و قدردانی:

این مقاله بخشی از طرح پژوهشی تحت عنوان « تحلیل بقاء بیماران سرطان سینه ، دستگاه گوارش و ریه مراجعه کننده به مرکز پژوهشی درمانی امید ارومیه با راهکار داده کاوی» در سال ۱۳۹۲ کد ۹۲-۰۱-۵۲-۱۱۴۰می باشد که با حمایت دانشگاه علوم پزشکی ارومیه اجرا شده است. بر خود لازم می دانیم از مسئولین و پرسنل مرکز تحقیقاتی و درمانی امیدارومیه که ما را در انجام این مطالعه یاری نمودند، قدردانی نماییم.

References:

1. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparision of three data mining methods. Artif Intell Med 2005: 34(2):113-127.

2. Garcia M, Jemal A, Ward E, Center M, Hao Y SR, Thun M. Global Cancer Facts & Figures. Atlanta, GA:American Cancer Society 2007.

3. Jami MS, Tavassoli M, Hemati S. Association of the Length of CA Dinucleotide Repeat in the Epidermal Growth Factor Receptor with Risk and Age of Breast Cancer Onset in Isfahan. J Isfahan Med Sch 2010; 26 (88):22-30

4. Bellaachia A, Guven E. Predicting Breast Cancer Survivability using Data Mining Techniques. In: Proceedings of Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM) 2006 April 22; Bethesda, MD, USA; 2006.

5. Endo A, Shibata T, Tanaka H. Comparison of Seven Algorithms to Predict Breast Cancer Survival. IJBSCHS 2008; 13 (2):11-16

6. Thongkam J, Xu GD, Zhang YC, Huang FC. Toward breast cancer survivability prediction models through improving training space. EXPERT SYST APPL 2009; 36 (10):12200-12209.

7. Ya-Qin L, Cheng W, Lu Z. Decision tree based predictive models for breast cancer survivability on imbalanced data. In: The 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE); 2009. June 11-13; Beijing: China 2009. p. 1-4.

8. Wang K-J, Makond B, Chen K-H, Wang K-M. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Applied Soft Computing 2014; 20:15-24

9. Wang K-J, Makond B, Wang K-M.An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. BMC Med Inform Decis Mak 2013; 13 (1):124

10. Afshar HL, Ahmadi M, Roudbari M, Sadoughi F. Prediction of Breast Cancer Survival Through Knowledge Discovery in Databases. Glob J Health Sci 2015; 7 (4):392-399

11. Carol D, Rebecca S, Priti B, Ahmedin J. Breast cancer statistics. CA 2011 Nov-Dec)cited 2012 Jun 12(; Available from:URL:http:// onlinelibrary. wiley.com/doi/10.3322/caac.20134/full

12. Han J, Kamber M, Pei J, editors. (2011) Data Mining: Concepts and Techniques. 3rded.Burlington:Morgan Kaufmann Publishers Inc 2011.

13. Breiman L, Friedman J, Stone CJ, OlshenRA.Classification and regression trees.Boca Raton, FL: CRC Press 1984.

14. Magnani M. Techniques for dealing with missing data in knowledge discovery tasks.Computer Science 2004; 1 (1):1-10.

15 .Vesin A, Azoulay E, Ruckly S, Vignoud L, Rusinovà K, Benoit D, Soares M, Azeivedo-Maia P, Abroug F, Benbenishty J. Reporting and handling missing values in clinical studies in intensive care units. Intensive Care Med 2013; 39 (8):1396-1404

16. Scheffer JA. Dealing with Missing Data. Research Letters in the Information and Mathematical Sciences 2002; 3:153-160

17. Bellenir K. Breast Cancer Sourcebook: Basic Consumer Health Information about Breast Health and Breast Cancer. Detroit: Omnigraphics 2009.

18. Witten I, Frank E, Hall M. Data mining: practical machine learning tools and techniques. Burlington:MorganKaufmannPublishersInc 2011.

JCRPS Journal of Clinical Research in Paramedical Sciences, Vol 6, No 2, Summer 2017 Original Article

Presentation of hidden knowledge from a localbreast cancer dataset by the classification and regression trees

Hadi LotfnezhadAfshar [1] , Lili Rahmatnejad [2] , Bahlol Rahimi [1], Hamid Reza Khalkhali [*3],

1. Department of Health Information Management, Health Information Technology Department, School of Paramedicine, Urmia University of Medical Sciences, Urmia, Iran

2. Department of Midwifery, School of Nursing & Midwifery, Urmia University of Medical Sciences, Urmia, Iran

3. Department of Biostatistics, Patient Safety Research Center, School of Medicine, Urmia University of Medical Sciences, Urmia, Iran

[*]**Corresponding Author:**
Iran, Urmia, Urmia University of Medical Sciences, School of Medicine, Patient Safety Research Center, Department of Biostatistics

Email: khalkhali@umsu.ac.ir

Abstract

Introduction: The using of standard knowledge discovery methods such as decision trees, in context ofthe breast cancer has been studied. Presentation of undiscovered relationship among data in formats such as: visualization and formulating are the reasons of decision trees popularity. An algorithm from this group that has not been used in the previous published papers, applied in current study.

Methods: A dataset included data about 569 patients' records between the years 2007 and 2010 was used. The missing data handling method was multiple imputation (MI). IBM statistics 21 was the used software for running MI and developing the model. The developed model was evaluated against the criteria such as: accuracy, sensitivity and specificity.

Results: A decision tree with seventeen nodes produced by the model. A set of clinically meaningful if-then rules were produced from nine nodes. It was clear from these rules that the variable that showed the stage of cancer was the most important variable to predict living probability of breast cancer. The performance of produced model for criteria (sensitivity, specificity and accuracy) was: 93.5, 53 and 80.3 percentage respectively.

Conclusion: The model created in current study as the first model in living probability of breast cancer revealed practical undiscovered rules from a not large dataset.

Key words: Breast neoplasms, survival, machine learning, Regression Analysis

How to cite this article
LotfnezhadAfshar H, Rahmatnejad L, Rahimi B, Khalkhali HR. Presentation of hidden knowledge from a localbreast cancer dataset by the classification and regression trees . J Clin Res Paramed Sci 2017; 6(2):123-134.