مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال پنجم، شماره چهارم، زمستان ۱۳۹۱

پیش بینی عود مجدد سرطان پستان به کمک سه تکنیک داده کاوی

عباس طلوعی اشلقی: استاد دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران

علی پورابراهیمی: رئیس واحد الکترونیکی دانشگاه آزاد اسلامی

ماندانا ابراهیمی: استادیار پژوهش، مرکز تحقیقات سرطان پستان جهاد دانشگاهی

لیلا قاسم احمد [1]: دانشجوی کارشناسی ارشد مدیریت فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران

چکیده

مقدمه: تعداد و اندازه پایگاه داده‌های پزشکی به سرعت در حال افزایش است و مدل‌های توسعه یافته تکنیک داده‌کاوی می‌توانند برای پزشکان جهت کمک در تصمیم گیری موثر و کاربردی باشند. هدف اصلی از این مقاله، گزارش یک پروژه تحقیقاتی به منظور مقایسه الگوریتم‌های مختلف داده کاوی از طریق مقایسه حساسیت، ویژگی و دقت بین آنها، جهت انتخاب دقیق‌ترین مدل برای پیش بینی عود مجدد سرطان پستان در زنان مبتلا بوده است. در حقیقت بیان کاربرد عملی داده کاوی در حوزه سرطان پستان با استفاده از داده‌های ثبت شده در پایگاه داده است که به فراهم کردن اطلاعات ضروری و دانش مورد نیاز پزشکان در تصمیم گیری بهتر کمک می‌کند.

مواد و روش‌ها: این تحقیق در خصوص بیماران مبتلا به سرطان پستان که حداقل هرکدام به مدت دو سال تحت پیگیری بوده‌اند، انجام شد. اطلاعات این بیماران در مرکز تحقیقات سرطان پستان جهاد دانشگاهی برای پیگیری اقدامات درمانی ثبت و بیماران حداقل به مدت دو سال پس از تشخیص، تحت نظر این مرکز بوده و پیگیری‌های بعدی برای آنها انجام شده است. به منظور توسعه مدل‌های پیش بینی جهت پیش بینی عود سرطان پستان، از درختان تصمیم گیری(C5.0)، ماشین بردار پشتیبان(SVM: Support Vector Machines) و تکنیک‌های شبکه‌های عصبی مصنوعی(ANNs: Artificial Neural Networks) با بهره‌گیری از پایگاه داده مذکور استفاده شده است.

نتایج: بررسی‌های صورت گرفته نشان می‌دهد که دقت در سه الگوریتم داده کاوی، یعنی درخت تصمیم گیری، ANN و SVM به ترتیب ۰/۹۳۶، ۰/۹۴۷ و ۰/۹۵۷ بوده است.

بحث و نتیجه گیری: مدل طبقه بندی SVM در پیش بینی عود مجدد سرطان پستان، حداقل میزان خطا و بیشترین دقت را داشت که بالاتر از درخت تصمیم گیری و مدل ANN بود و دقت پیش بینی در مدل درخت تصمیم گیری(C5.0) نیز پایین‌ترین میزان در بین سه مدل پیش بینی را نشان داد. نتایج به دست آمده حاکی از افزایش درصد صحت نتایج، با بهره گیری از روش‌های تقویت و هرس کردن بوده است.

واژه‌های کلیدی: عود سرطان پستان، درخت تصمیم گیری، ماشین بردار پشتیبان، شبکه‌های عصبی مصنوعی، طبقه بندی.

[1] نشانی نویسنده پاسخگو: دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران، لیلا قاسم احمد. پست الکترونیک: lga_77@yahoo.com

[DOR: 20.1001.1.17359406.1391.5.4.3.3]

مقدمه

سرطان پستان یکی از شایع‌ترین انواع سرطان در زنان است که حدود ۱۰٪ از آنان را در مراحل مختلف زندگی خود، تحت تاثیر قرار می‌دهد(۴). این سرطان شایع‌ترین بدخیمی در میان زنان ایرانی و کانون اصلی توجهات در کشور ایران است. در سال‌های اخیر، میزان شیوع بیماری روند رو به رشدی داشته و داده‌ها نشان می‌دهد که میزان بقای بیماران تا پنج سال پس از تشخیص، ۸۸٪ و ۱۰ سال پس از تشخیص ۸۰٪ بوده است. در واقع، تمام تومورها سرطانی نبوده و ممکن است خوش خیم یا بدخیم باشند. تومورهای خوش خیم رشد غیر طبیعی دارند ولی به ندرت مرگ‌آور هستند. با این حال، تعدادی از توده‌های خوش خیم پستان نیز می‌توانند خطر ابتلا به سرطان پستان را افزایش دهند. همچنین در برخی از زنان با سابقه بیوپسی از توده‌های خوش خیم پستان نیز، خطر سرطان پستان افزایش یافته است. از طرف دیگر، تومورهای بدخیم جدی‌تر بوده و سرطان محسوب می‌شوند ولی تشخیص زود هنگام این نوع از سرطان‌ها شانس درمان موفقیت آمیز را بالا برده است(۴). پیش بینی عود مجدد سرطان پستان یکی از پرطرفدارترین اقدامات انجام شده برای توسعه رویکردهای داده کاوی است. بنابراین، به منظور توسعه بهتر روش‌های شناسایی سرطان پستان ضروری به نظر می‌رسد. روش‌های داده کاوی می‌توانند به کاهش تعداد پاسخ‌ها و نتایج مثبت کاذب و منفی کاذب در تصمیم گیری پزشکان کمک کنند(۵ و ۶). در نتیجه، رویکردهای جدید مانند کشف دانش از پایگاه داده (KDD)، که شامل تکنیک‌های داده کاوی هستند، روز به روز محبوبیت بیشتری یافته و تبدیل به یک ابزار تحقیقاتی مطلوب برای پژوهشگران علوم پزشکی شده‌اند. به کمک آنها پژوهشگران می‌توانند الگوها و روابط بین تعداد زیادی از متغیرها را شناسایی کرده و پیش بینی نتایج حاصل از یک بیماری با استفاده از ذخایر اطلاعاتی موجود در پایگاه‌های داده برای آنها امکان پذیر گشته است(۷) بررسی‌ها و مطالعات گوناگونی در زمینه مشکلات ناشی از پیش بینی بقای بیماران مبتلا به سرطان پستان، با استفاده از روش‌های آماری و شبکه‌های عصبی مصنوعی صورت گرفته، اما فقط تعداد کمی از مطالعات حوزه پزشکی در زمینه عود سرطان با استفاده از روش‌های داده کاوی انجام شده است(۸).

مطالعات پیشینی با استفاده از داده کاوی و با رویکرد پیش بینی در علوم پزشکی وجود دارند. به عنوان مثال، دلن و همکاران از شبکه‌های عصبی مصنوعی، درخت تصمیم گیری و رگرسیون لجستیک برای توسعه مدل‌های پیش بینی سرطان پستان با تجزیه و تحلیل پایگاه‌های بزرگ داده که از پایگاه داده مشهور ویسکانسین[1] (The Surveillance, Epidemiology, and End Results) گردآوری شده بود، بهره جستند. نتایج تحقیقات آنها نشان داد که الگوریتم درخت تصمیم برای استخراج دانش از داده‌های موجود مقدم بر سایر روش‌ها بود و نتایج به دست آمده از تحقیق، نزدیک به واقعیت بود(۹). همچنین یی و فویانگ از ماشین بردار پشتیبان به تنهایی برای کشف الگوهای تشخیص سرطان پستان، از داده‌های موجود در بیمارستان ویسکانسین استفاده کردند. نتایج به دست آمده نشان داد که SVM برای تشخیص الگوهای سرطان پستان روش مناسبی بود و نتایج به دست آمده با شواهد موجود و واقعی مطابقت داشت(۱۰). لاندین(Lundin M) و همکارانش از مدل شبکه‌های عصبی مصنوعی و رگرسیون لجستیک برای پیش بینی ۵، ۱۰ و ۱۵ ساله بقای بیماران مبتلا به سرطان پستان استفاده کردند. آنها ۹۵۱ بیمار مبتلا به سرطان پستان را مورد مطالعه قرار داده و اندازه تومور، وضعیت گره های لنفی، نوع بافت، تشکیل توبول، نکروز تومور و سن را به عنوان متغیرهای ورودی محسوب کردند. سپس به این نتیجه رسیدند که درختان طبقه بندی و همچنین رگرسیون لجستیک برای تفسیر بالینی بسیار آسان‌تر است(۱۱). پندهارکر و همکاران(Pendharkar PC) از چندین روش داده کاوی برای بررسی الگوهای موجود در سرطان پستان استفاده نمودند. در این مطالعه، آنها نشان دادند که داده کاوی می‌تواند به عنوان یک ابزار ارزشمند در شناسایی شباهت‌ها(الگوها) در مورد سرطان پستان با هدف تشخیص، پیش آگهی و درمان به کار رود(۱۲). این مطالعات مثال‌هایی از کاربرد داده کاوی در علوم پزشکی برای پیش بینی بیماری‌ها هستند.

در تحقیق فعلی، پایگاه داده بیماران مبتلا به سرطان پستان در "مرکز تحقیقات سرطان پستان جهاد دانشگاهی" به کار رفت. از پیش پردازش داده‌ها در نهایت ۵۴۷ نمونه باقی ماند که به دو گروه منتهی به عود یا عدم

[1] SEER

[DOR: 20.1001.1.17359406.1391.5.4.3.3]

عود ختم گردیدند. از این تعداد ۱۱۷ بیمار دچار عود بیماری و ۴۳۰ بیمار نیز فاقد عود مجدد بیماری بودند. نتایج پیش بینی عود مجدد بیماری در هر سه مدل، بالای۹۰٪ بود. در این مقاله، مدل‌هایی که جهت پیش بینی عود در بیماران مبتلا به سرطان پستان به کار می‌روند، توسط تکنیک‌های داده کاوی توسعه یافته و اثربخشی مدل‌ها با مقایسه سه الگوریتم بررسی شدند. جهت تعیین درجه اعتبار مدل، از روش اعتبارسنجی متقابل تا ۱۰ لایه برای ارزیابی دقت طبقه بندی این مدل استفاده گردید. در بخش بعدی این مقاله به بررسی تحقیقات مرتبط صورت گرفته در این زمینه، تشریح تاریخچه این مطالعات، اهداف آن و همچنین ارزیابی سه مدل طبقه‌بندی یعنی درخت تصمیم، ماشین بردار پشتیبان و تکنیک شبکه‌های عصبی مصنوعی برای تجزیه و تحلیل پیش بینی عود مجدد بیماری پرداخته شده است. در بخش بعد نیز نتایج تجربی، نتیجه گیری تحقیق و پیشنهادهایی برای تحقیق‌های آتی ارائه شده‌اند.

به طور کلی سرطان پستان، به تومور بدخیمی اطلاق می‌شود که با رشد غیرقابل کنترل سلول‌های بافت پستان، مرگ و تقسیم سلول‌های طبیعی همراه گردد که شایع‌ترین سرطان در میان زنان است(۶) درمان‌های سرطان پستان به دو دسته اصلی موضعی و سیستماتیک تقسیم می‌شوند. جراحی و پرتودرمانی، نمونه‌هایی از درمان‌های موضعی و شیمی درمانی و هورمون درمانی، نمونه‌هایی از روش‌های درمانی سیستماتیک هستند. معمولاً برای دستیابی به نتایج بهتر، از هر دو نوع درمان با هم استفاده می‌شود(۷).

در این پژوهش، داده‌ها از مرکز نامبرده در بالا جمع آوری شدند. پایگاه داده شامل ۱۱۸۹ نمونه با ۶۴۲ نمونه دارای متغیرهای از دست رفته و ۵۴۷ نمونه دارای کل متغیرهای مورد نیاز بود. از میان ۱۱۸۹ رکورد انتخاب شده بیماران، تعداد بسیار زیادی فاقد اطلاعات کامل بودند. حتی در مواردی اطلاعات یک بیمار به ۳-۲ متغیر محدود می‌شد. به همین دلیل و با توجه به بی‌اثر بودن وجود رکورد موجود، ناچار تصمیم به حذف آن گرفته می‌شد.

مواد و روش‌ها

در این مقاله، سه الگوریتم داده کاوی یعنی درخت تصمیم گیری، ماشین بردار پشتیبان و شبکه‌های عصبی مصنوعی برای پیش بینی عود مجدد سرطان پستان تجزیه و تحلیل شد و بهترین راهکار در پیش بینی عود شناسایی گردید. داده‌هایی که از مرکز تحقیقات سرطان پستان جهاد دانشگاهی جمع آوری گردید مربوط به سال‌های ۱۳۷۶ تا ۱۳۸۷ بود. پایگاه داده قبل از پردازش شامل ۱۱۸۹ نمونه و ۲۶ ویژگی اولیه بود که پس از پردازش اولیه، تمیز کردن داده‌ها و حذف نمونه‌های محتوی متغیرها با مقادیر مفقود شده، ۵۴۷ نمونه با اطلاعات و متغیرهای کامل و ۲۲ ویژگی‌های نهایی باقی ماند.

متغیرهای ذیل قبل از بکارگیری پیش پردازش شدند که شامل: عود مجدد بیماری[۲]، سن تشخیص بیماری[۳]، سن شروع قاعدگی[۴]، سن شروع یائسگی[۵]، سابقه ناباروری[۶]، سابقه فامیلی ابتلا به سرطان پستان[۷]، سابقه ابتلا به سایر انواع سرطان[۸]، محل قرارگیری توده سرطانی[۹]، سمت قرارگیری تومور[۱۰]، اندازه تومور[۱۱]، میزان درگیری غددلنفاوی[۱۲]، تعداد غدد لنفاوی خارج شده پس از جراحی[۱۳]، متاستاز[۱۴]، مثبت بودن درگیری غدد لنفاوی[۱۵]، نتیجه بیوپسی(آسیب شناسی پس از نمونه برداری)، جراحی، درجه تومور، میزان آزاد یا گرفتار بودن حاشیه تومور[۱۶]، گیرنده استروژن(ER)[۱۷]، گیرنده پروژسترون[۱۸]، عامل Her2[۱۹]، شیمی درمانی شدن یا نشدن بیمار، نوع شیمی درمانی، پرتو درمانی پس از برداشتن پستان[۲۰]، هورمون درمانی[۲۱]، مرگ[۲۲] بوده‌اند.

در این تحقیق از نرم افزار کلمنتاین نسخه ۱۲[۲۳] و الگوریتم‌های داده کاوی استفاده شد. نرم افزار کلمنتاین

[2] Local Recurrence
[3] Age diagnosis
[4] Age Menarche
[5] Age Menopause
[6] Infertility
[7] Family History of breast cancer
[8] Other Cancer (CA)
[9] Location
[10] Side
[11] Tumor Size (T)
[12] LN Involvement
[13] LN/Nexion
[14] Metastasis
[15] NPositive
[16] Margin
[17] Estrogen Receptor (ER)
[18] Progesterone Receptor
[19] Human epidermal growth factor receptor 2
[20] Radiotherapy (Rt)
[21] Hormonetherapy
[22] Death
[23] SPSS Clementine 12

[DOR: 20.1.1.17359406.1391.5.4.3.3]

مشتمل بر مجموعه‌ای از ابزارها برای طبقه بندی داده‌ها، رگرسیون، خوشه بندی و قوانین انجمنی است(۱۳) برای پیش‌بینی دو ساله میزان عود مجدد در سرطان پستان، به کمک پایگاه داده نامبرده، ویژگی‌های جمعیت آماری حاوی ۲۲ متغیر از قبیل سن تشخیص، سن یائسگی، ناباروری، سابقه خانوادگی سرطان پستان و سایر شاخص‌ها استخراج گردید. این داده‌ها از بررسی پرونده ۱۱۸۹ بیماری سرطان پستان گردآوری و ثبت شدند ولی به دلیل مفقود بودن برخی از متغیرها در مورد بیماران، تنها از ۵۴۷ مورد استفاده شد. در این مطالعه، تنها بیمارانی که حداقل به مدت دو سال تحت پیگیری قرار داشتند، در تحقیق وارد شدند. برای اعتبارسنجی و ارزیابی مدل، داده‌های جدید به طور تصادفی به کمک اعتبارسنجی متقابل(۱۰ لایه)، به دو دسته آموزشی و آزمایشی مستقل تقسیم گردیدند.

درختان تصمیم‌گیری(مدل C5.0) ، ماشین بردار پشتیبان و تکنیک های شبکه های عصبی مصنوعی:

در این مطالعه، سه روش داده کاوی یعنی درخت تصمیم گیری، شبکه‌های عصبی مصنوعی و ماشین‌های بردار پشتیبان بکار رفتند.

درختان تصمیم‌گیری:

درخت تصمیم‌گیری در بین الگوریتم‌های طبقه‌بندی روش قدرتمندی است که محبوبیت آن با رشد داده کاوی به طور فزاینده‌ای در حال افزایش است. درخت‌های"و، یا" نام دیگر درختان تصمیم است که نمونه‌ها را با مرتب کردن آنها در درخت از گره ریشه به سمت گره‌های برگ دسته‌بندی می‌کنند(۱). نمونه‌ها به نحوی دسته‌بندی می‌شوند که از ریشه به سمت پایین رشد می‌کنند و در نهایت به گره‌های برگ می‌رسد. هر گره داخلی یا غیر برگ با یک ویژگی مشخص می‌شود. این ویژگی سوالی را در رابطه با مثال ورودی مطرح می‌کند. در هر گره داخلی به تعداد جواب‌های ممکن با این سوال شاخه وجود دارد که هر یک با مقدار آن جواب مشخص می‌شوند. برگ‌های این درخت با یک کلاس و یا یک دسته از جواب‌ها مشخص می‌شوند. علت نامگذاری آن با درخت تصمیم این است که این درخت فرایند تصمیم‌گیری برای تعیین دسته یک مثال ورودی را نشان می‌دهد. درخت تصمیم در مسائلی کاربرد دارد که بتوان آنها را به صورتی مطرح نمود که پاسخ واحدی به صورت نام یک دسته یا کلاس ارائه دهند.

هم چنین برای مسائلی مناسب است که مثال‌های آموزشی به صورت زوج(مقدار- ویژگی) مشخص شده باشند. تابع هدف دارای خروجی با مقادیر گسسته باشد. مثلا هر مثال با بله و خیر تعیین شود و یا نیاز به توصیف‌گر فصلی باشد. درخت تصمیم شامل تعدادی از الگوریتم‌ها مانند ID3، C4.5، C5 و طبقه‌بندی است که در این تحقیق از روش C5 استفاده شد(۱۴).

درخت تصمیم برای تقریب توابع گسسته به کار می‌رود.

نسبت به نویز داده های ورودی مقاوم است.

برای داده‌های با حجم بالا کاراست از این رو در داده کاوی استفاده می‌شود.

می‌توان درخت را به صورت قوانین if-then نمایش داد که برای استفاده قابل فهم است.

امکان ترکیب عطفی(AND) و فصلی(OR) فرضیه‌ها را می‌دهد.

در مواردی که مثال‌های آموزشی فاقد همه ویژگی‌ها هستند نیز قابل استفاده است(۲) و (۱۵).

ماشین‌های بردار پشتیبان:

درحالی که روش‌هایی مانند درخت تصمیم‌گیری را نمی‌توان به راحتی در مسائل مختلف به کار برد. این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی از جمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاری دسته‌بندی کنندهٔ SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌شود خطی انتخاب گردد که حاشیه اطمینان بیشتری داشته باشد. ماشین بردار پشتیبان حداکثر حاشیه الگوریتم طبقه‌بندی ریشه در تئوری یادگیری آماری است. این روش برای طبقه‌بندی داده های هر دو خطی و غیر خطی است. در واقع از یک نگاشت غیر خطی در ابعاد جدید برای تبدیل داده‌های آموزشی اصلی به یک بعد بالاتر استفاده می‌کند. با استفاده از نگاشت غیر خطی مناسب داده‌های دو کلاس توسط یک ابرصفحه جدا شده‌اند و خطاهای طبقه بندی به حداقل می‌رسد(۱۶).

شبکه عصبی:

یک شبکه عصبی مصنوعی روشی برای پردازش اطلاعات است که از سیستم‌های عصبی زیستی الهام گرفته شده و مانند مغز پردازش اطلاعات انجام می‌گیرد. عنصرکلیدی

[DOR: 20.1001.1.17359406.1391.5.4.3.3]

این ایده، ساختار جدید سیستم پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی فوق العاده به هم پیوسته تشکیل شده که برای حل یک مسأله با هم هماهنگ عمل می‌کند. شبکه‌های عصبی، نظیر انسان‌ها، با مثال یاد می‌گیرند. یک شبکه عصبی مصنوعی برای انجام وظیفه‌ای مشخص، مانند شناسایی الگوها و دسته بندی اطلاعات، در طول یک پروسه یادگیری، تنظیم می‌شود. در سیستم‌های زیستی یادگیری با تنظیماتی در اتصالات سیناپسی که بین اعصاب قرار دارد همراه است. این روش شبکه عصبی مصنوعی هم می‌باشد(۱۷). این شبکه‌ها قادر به مدل سازی توابع غیر خطی است. شبکه‌های عصبی مصنوعی تکنیک‌های تحلیلی هستند که قادر به پیش بینی مشاهدات جدید(متغیرهای یکسان و یا سایر) پس از اجرای یک فرایند یادگیری به اصطلاح از داده‌های موجود هستند(۱۸).

از الگوریتم شبکه‌های عصبی پرسپترون چندلایه (Perceptrons Multi-Layered) [۲۴] برای مشکلات پیش‌بینی و طبقه‌بندی استفاده می‌شود. اساساً MLP مجموعه‌ای از سلول‌های عصبی غیرخطی سازماندهی شده و متصل به یکدیگر در یک ساختار چند لایه جلورونده[۲۵] است(۱۹). در روش فوق، آن دسته از بیماران که کمتر از دو سال تحت پیگیری بودند نادیده گرفته شدند. فیلدهای بی اثر از تحقیق کنار گذاشته شده و توسط بخش فیلتر، محدودیت آنها جهت عدم ورود به مدل‌سازی تعیین گردیدند. در قسمت Input نیز متغیرهایی همچون Age.Menarc، متاستاز، HER2 و تعداد مثبت درگیری غدد لنفاوی(Npositive) فیلتر شدند. به صورتی که، فیلد کد(به دلیل بی اثر بودن بر تحقیق)، ناحیه قرارگیری تومور (به دلیل مشابهت با فیلد محل قرارگیری تومور)، Her2 و سن شروع قاعدگی(به دلیل تعداد بالای اطلاعات نامعلوم)، Npositive(به دلیل تکرار در فیلد LN/Nexion که خود حاصل تقسیم تعداد غدد لنفاوی درگیر به تعداد غدد لنفاوی خارج شده طی جراحی بود) و متاستاز نیز به دلیل همپوشانی با عود مجدد بیماری حذف گردیدند.

نتایج

در این مطالعه، سه روش داده کاوی مبتنی بر دقت مدل‌ها مقایسه شدند و هدف نهایی دست یابی به مدلی با بالاترین میزان دقت بود. برای پیش پردازش پایگاه داده به منظور انتخاب متغیرها از الگوریتم خاصی استفاده گردید و پس از به کارگیری این الگوریتم بیمارانی که داده‌های آنها فاقد متغیرهای کافی و دارای کاستی های زیادی بودند شناسایی و از پایگاه داده حذف گردیدند. در خصوص حذف متغیرها، متغیرهایی حذف گردیدند که یا با نتایج تحقیق همپوشانی داشتند، و یا اینکه از میان ۱۱۸۹ رکورد موجود، اطلاعات بیماران از این متغیرها بسیار محدود بود. پس از حذف متغیرها مطابق با دو حالت ذکر شده(وجود همپوشانی با سایر متغیرها یا نبودن اطلاعات حداکثری رکوردها از این متغیر) کار بر روی رکوردها آغاز گردید. در مورد بیمارانی که اطلاعاتی از متغیرها در مورد آنها وجود نداشته و یا محدود به ۳-۲ مورد می‌شد، در این خصوص چاره‌ای جز حذف رکورد وجود نداشت. رکوردهای باقیمانده که تعداد کمتری متغیرهای گمشده داشتند، از طریق ماکزیمم سازی مقدار مورد انتظار[۲۶] جایگزین گردیدند.

پس از حذف رکوردهای حاوی مقادیر مفقود شده، ۲۹۶ رکورد باقی ماندند. به منظور یافتن بهترین عملکرد پیش بینی، تکنیک‌های طبقه‌بندی داده کاوی شامل درخت تصمیم، شبکه عصبی مصنوعی و ماشین بردار پشتیبان، به کار گرفته شده و پارامترهای متعددی به صورت تصادفی انتخاب و بررسی شدند. جدول یک، شاخصه های پیش بینی را به منظور مدل سازی عود مجدد سرطان پستان نشان می‌دهد(جدول ۱).

[24] MLPs
[25] Feed forward
[26] Expectation-maximization

[DOR: 20.1.1.17359406.1391.5.4.3.3]

جدول ۱: متغیرهای پیش بینی برای مدل سازی عود مجدد سرطان پستان

۱	عود مجدد (Local Recurrence)	۰= عود بیماری ۱= عدم عود
۲	سن تشخیص بیماری (Age Diagnosis)	0 = "<35" 1 = "35-44" 2 = "45-55" 3 = ">55"
۳	سن قاعدگی (Age Menarche)	۰= شروع قاعدگی بیشتر از ۱۲ سال ۱= شروع قاعدگی ۱۲ ≤ (خطرساز)
۴	سن یائسگی (Age Menopause)	۰= سن شروع یائسگی کمتر از ۵۰ سال ۱= سن شروع یائسگی بیشتر از یا مساوی ۵۰ سال (خطرساز) ۲= فرد هنوز یائسه نشده
۵	سابقه نازایی (Infertility)	۰= بیمار سابقه نازایی ندارد (No) ۱= بیمار سابقه نازایی دارد (Yes)
۶	سابقه فامیلی ابتلا به سرطان پستان (Family History)	۰= سابقه ای وجود ندارد (No) ۱= سابقه بیماری در خانواده وجود دارد (Yes)
۷	سابقه فامیلی ابتلا به سایر انواع سرطان (Other Cancer)	0 = "no" 3 = "colon cancer" 1 = "mail breast cancer" 4 = "ovarian cancer" 2 = "prostate cancer" 5 = "utrus cancer"
۸	محل قرارگیری توده سرطانی (Location)	1 = "uoq" 12 = "upper half" 2 = "uiq" 13 = "latral half" 3 = "loq" 14 = "uoq and liq" 4 = "liq" 24 = "medial half" 5 = "central(nipple areole)" 30 = "three quadrant" 6 = "axilla" 34 = "lower half" 50 = "diffuse"
۹	سمت قرار گیری تومور (Side)	۱= تومور در پستان سمت راست ۲= تومور در پستان سمت چپ ۳= تومور دوطرفه
۱۰	اندازه تومور (Tumor Size)	1 = "<2" 2 = "2-5" 3 = ">5" 4 = "chest wall or skin"
۱۱	میزان درگیری غدد لنفاوی (LN involvement)	0 = "no" 1 = "1-3" 2 = "4-9" 3 = ".>9"
۱۲	تعداد کل غدد لنفاوی خارج شده پس از جراحی (LN involvement)	
۱۳	نسبت تعداد غدد لنفاوی درگیر به تعداد غدد لنفاوی خارج شده پس از جراحی (LN involvement/ Nexion)	
۱۴	درگیری غددلنفاوی (NPositive)	

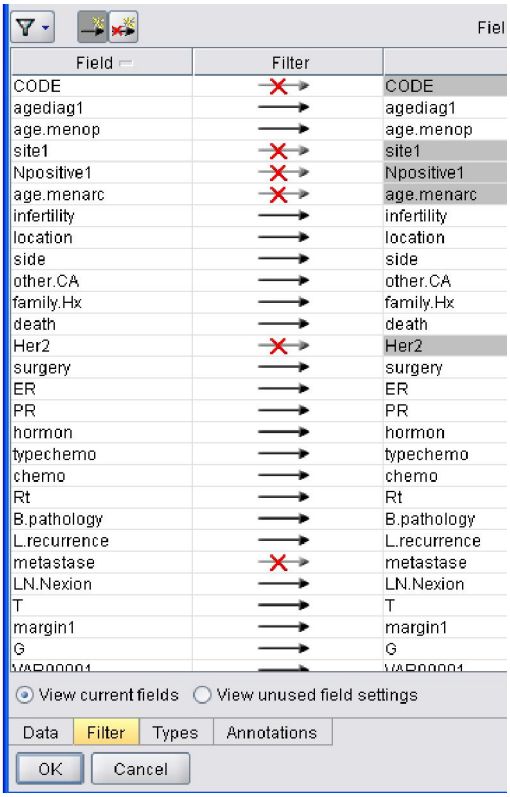
[DOR: 20.1001.1.17359406.1391.5.4.3.3]

ادامه جدول شماره ۱

۱۵	ناحیه قرارگیری تومور	1 = "local" 2 = "axilla" 3 = "regional" 4 = "Local (post MRM)" 12 = "local & axilla" 13 = "local & regional" 14 = "local & local (post MRM)"
۱۶	متاستاز (Metastasis)	0 = "no" 1 = "bone" 2 = "liver" 3 = "lung" 4 = "brain" 5 = "others" 6 = "not specified"
۱۷	ناحیه مبتلا به سرطان (Site)	
۱۸	تعیین تعداد دقیق غدد لنفاوی درگیر	بین صفر تا حداکثر ۲۷ عدد
۱۹	نتیجه آسیب شناسی پس از نمونه برداری (Biopsy of Pathology)	1 = "lcis" 2 = "dcis" 3 = "idc" 4 = "ilc" 5 = "medullary" 6 = "microinvasion" 7 = "paget's disease" 8 = "others" 9 = "inflamatory carcinoma" 11 = "sarcoma" 12 = "metastatic(un known origin)" 13 = "lymphoma"
۲۰	نوع جراحی صورت گرفته (Surgery)	1 = "MRM" 2 = "Breast preservation" 3 = "bilatral MRM" 4 = "bilatral BCS" 5 = "bilatral MRM & BCS"
۲۱	درجه بندی تومور (Grade)	1 = "1" 2 = "2" 3 = "3"
۲۲	میزان آزاد یا گرفتار بودن حاشیه تومور (Margin)	0 = "free (>=2cm)" 1 = "closed (<=2cm)" 2 = "involve"
۲۳	مقدار گیرنده های استروژن (Estrogen Receptor)	۰= منفی بودن مقدار ۱= مثبت بودن مثبت بودن مقدار، خطر عود مجدد را افزایش می دهد
۲۴	مقدار گیرنده های پروژسترون (Progestron Receptor)	۰= منفی بودن مقدار ۱= مثبت بودن مثبت بودن مقدار، خطر عود مجدد را افزایش می دهد.
۲۵	مقدار Her2	۰= منفی بودن مقدار ۱= مثبت بودن مثبت بودن مقدار، خطر عود مجدد را افزایش می دهد
۲۶	شیمی‌درمانی شدن یا نشدن بیمار (Chemoteraphy)	0 = "no" 1 = "yes" 19 = "chemo+Herceptin"
۲۷	نوع شیمی درمانی (Typechemo)	1 = "neo adjvant" 2 = "adjvant"
۲۸	پرتودرمانی پس از جراحی برداشتن پستان (Radiotherapy)	۰= پرتو درمانی انجام نشده ۱= پرتو درمانی انجام شده وقتی پرتودرمانی انجام نشده است، احتمال بیشتری برای عود مجدد سرطان وجود دارد.
۲۹	هورمون درمانی (Hormone)	0 = "no" 1 = "tamoxifen" 2 = "raloxifen" 3 = "femara or letrozol" 4 = "aromazin (exemstane)" 5 = "megace" 6 = "others"
۳۰	مرگ (Death)	0 = "unrelated" 1 = "related"

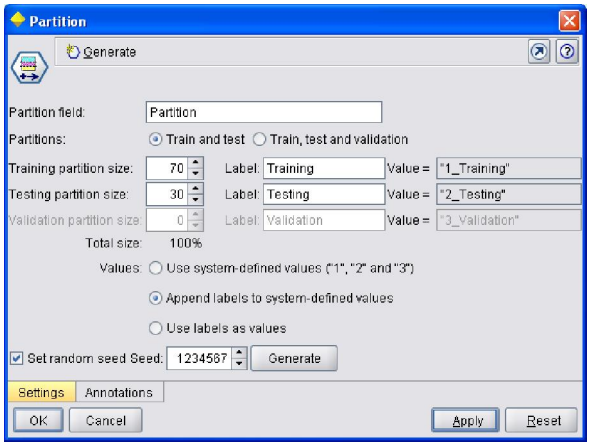
[DOR: 20.1.1.17359406.1391.5.4.3.3]

تقویت کردن[۲۷] و آموزش، دقت نتایج حاصله از درخت تصمیم را افزایش می دهند. هرچند که همه متغیرها برای پیش بینی نیازمند این درجه از دقت نیستند. در خصوص ماشین‌های بردار پشتیبان نیز، این مدل‌ها قبلا الگوهای موفقی در علوم بیوانفورماتیک، تشخیص سرطان و .... ارائه داده‌اند. الگوریتم ماشین‌های بردار پشتیبان اولیه در ۱۹۶۳ توسط وپنیک ابداع شد و در سال ۱۹۹۵ توسط کورتس برای حالت غیرخطی تعمیم داده شد. ماشین بردار پشتیبان می‌تواند، از یک سری داده‌های آموزشی عملکرد رگرسیون و دسته بندی را ایجاد کند(۲۰). در این بررسی برای مقایسه مدل‌ها و طبقه بندی داده‌ها تکنیک‌های داده کاوی بکار رفتند. داده‌ها شامل ۲۲ متغیر ورودی اولیه و ۵۴۷ رکورد اطلاعاتی بودند. مجموعه داده‌های گردآوری شده، مستقیما از پایگاه داده SPSS که شامل منابع داده است، فراخوانی شدند. در میان داده‌های گردآوری شده، برخی از داده‌ها شامل داده‌های غلط، ناپایدار یا مفقود شده بودند. شکل شماره یک متغیرهای فیلتر شده در کلمنتاین را نشان می‌دهد(شکل ۱).



شکل ۱: متغیرهای فیلتر شده در کلمنتاین

جهت آزمون روایی این تحقیق، داده‌ها به دو دسته داده‌های تست و داده‌های آموزشی در تکنیک درخت تصمیم تقسیم شدند که خروجی‌های نهایی بررسی شده و روایی پژوهش تایید شد. به طور کلی مبانی نظری این تحقیق که شامل تجربه‌های مشابه هستند نیز روایی ابزار گردآوری داده‌ها را نیز پشتیبانی می‌کنند. براین اساس در این مدل باید خروجی نهایی اختلاف قابل توجهی(تفاوت معنادار) در میان داده‌های تست و داده‌های آموزشی خود نداشته باشند. داده‌ها با نسبت ۷۰ به ۳۰ به ترتیب برای داده‌های آموزشی و آزمایشی پارتیشن بندی شدند. یعنی تعداد مجموعه‌های آموزش ۷۰ درصد از کل داده‌ها بوده و ۳۰ درصد باقیمانده به عنوان داده‌های تست درنظرگرفته شدند. نمودار یک، دو مدل آموزشی و آزمایشی را نشان می‌دهد. این شکل، نمودار به دست آمده از دو مدل آموزشی و آزمایشی را نشان می‌دهد. بالاتر بودن خطوط، اشاره به مدل‌های به دست آمده بهتری دارد. همچنین درصد داده‌های آموزشی و آزمایشی در شکل دو، نشان داده شده است. برای این کار داده‌های آموزشی را وان به دو قسمت تقسیم کرد: اول آن دسته که مدل بر اساس آنها ساخته می‌شود و دوم گروهی که برای ارزیابی مدل استفاده می‌شود. داده‌های گروه دوم که دسته آنها مشخص است را به مدل داده و خروجی مدل با دسته‌های مشخص توسط مدل مقایسه می‌گردد و سپس دقت کل مدل استخراج می‌شود. در واقع، برچسب شناخته شده از نمونه آزمون با نتایج دسته‌بندی مقایسه می‌شود(شکل ۲).



شکل ۲: درصد داده‌های آموزشی و آزمایشی

[27] Boosting

دقت مدل، درصد تعداد دفعاتی است که نمونه‌های آزمایشی با موفقیت بسته‌بندی می‌شوند. اگر دقت مدل قابل قبول باشد می‌توان مدل را برای دسته‌بندی داده‌هایی که دسته آنها مشخص نیستند، به کار برد. برای انجام آموزش و آزمایش داده‌ها به صورت تصادفی به ۱۰ قسمت تقسیم می‌شوند و در ۱۰ آزمایش مختلف تقسیم می‌شوند. سپس این داده‌ها به دو کلاس عود مجدد و عدم عود مجدد تقسیم می‌شوند. مقادیر داده‌های عود، ۱۱۷ و مقادیر داده‌های دچار عدم عود، ۴۳۰ مورد هستند. ارتباط بین کلاس‌های واقعی و کلاس‌های پیش بینی شده ماتریس confusion نامیده می‌شود. این ماتریس برای محاسبه میزان دقت مدل به دست آمده به کار می‌رود. فرمول ذیل برای محاسبه دقت، حساسیت و ویژگی به کار می‌رود.

نتایج تعیین میزان دقت به دست آمده از مدل‌ها، مقادیر عددی حاصل از محاسبات انجام شده توسط نرم افزار کلمنتاین و آنالیز نهایی آن را نشان می‌دهد. در تحقیقات مشابه انجام شده نیز، دقت مدل SVM بالاتر از مدل‌های دیگر بوده است. یعنی این مدل دقت بالاتری نسبت به سایر مدل‌ها داشته است و همچنین در تعیین متغیرهای اثرگذار بر عود نیز، متغیرهای مشخص شده با متغیرهای دنیای واقعی منطبق بودند. یعنی جزو متغیرهای موثر بر پیش بینی عود بودند.

دقت: عبارت است از تعداد نمونه‌هایی که به درستی تشخیص داده می‌شوند، نسبت به کل نمونه‌ها.

حساسیت: احتمال پیش بینی درست عود توسط الگوریتم‌ها(مثبت واقعی تقسیم بر منفی کاذب+مثبت واقعی).

ویژگی: احتمال پیش بینی درست عدم عود توسط الگوریتم‌ها(منفی واقعی تقسیم بر مثبت کاذب+منفی واقعی).

$$Sensitivity= TP/ (TP+ FN)$$
$$Specificity= TN/ (TN+ FP)$$
$$Accuracy= (TP+ TN)/ (TP+ FP+ TN+ FN)$$

TP: تعداد نمونه‌هایی که به درستی مثبت تشخیص داده می‌شوند.

TN: تعداد نمونه‌هایی که به درستی منفی تشخیص داده می‌شوند.

FP: تعداد نمونه‌هایی که به اشتباه مثبت تشخیص داده می‌شوند.

FN: تعداد نمونه‌هایی که به اشتباه منفی تشخیص داده می‌شوند.

جدول شماره دو، میزان دقت، حساسیت و ویژگی را برای سه روش‌های مختلف دسته‌بندی نشان می‌دهد(جدول ۲).

جدول ۲: مقایسه نتایج حاصل از سه مدل داده کاوی

	دقت	حساسیت	ویژگی
درخت تصمیم (C5.0)	۰/۹۳۶	۰/۹۵۸	۰/۹۰۷
شبکه های عصبی مصنوعی (MLP)	۰/۹۴۷	۰/۹۵۶	۰/۹۲۸
ماشین بردار پشتیبان (SVM)	۰/۹۵۷	۰/۹۷۱	۰/۹۴۵

با مقایسه و ارزیابی معیارهای ذکر شده، میزان دقت، حساسیت و سایر ویژگی‌های ذکر شده مشخص می‌گردد. براساس نتایج بدست آمده از مدل داده کاوی و نظرات کارشناسان می‌توان مدلی ارائه نمود که بتواند در تصمیم گیری در مورد احتمال عود یا عدم عود بیماری پزشک یا کارشناس مربوطه را یاری رساند(۲۱ و ۲۲).

در تحقیقات مشابه کاربرد داده کاوی از مقایسه الگوریتم‌های داده کاوی در پیش بینی بقا بیماران یا عود نیز استفاده شده است. بالا بودن حجم داده‌های آنالیز شده، بکارگیری متغیرهای مناسب، ایجاد مدل‌های پیش بینی هدفمند و جایگزینی داده‌های مفقود شده با روش مناسب، موجب دستیابی به نتایج بهتری در داده کاوی خواهد گردید(۲۳). در حالی که داده کاوی اطلاعات مفیدی را فراهم می‌کند و کارکنان تیم درمانی را در شناسایی الگوهای پنهان یاری می‌دهد، محدودیت‌هایی نیز وجود دارند که داده کاوی قادر به انجام آنها نیست. همه الگوهای یافت شده از طریق داده کاوی، جالب نیستند و برای جالب توجه بودن یک الگو، آن الگو باید منطقی و در عمل قابل اجرا باشد. بنابراین داده کاوی نیازمند مداخلات انسانی برای بکارگیری دانش استخراج شده است. برای مثال داده‌کاوی می‌تواند به ایجاد تشخیص یا تجویز درمان کمک کند، ولی هنوز نتوانسته است جایگزین مهارت‌های تجربی پزشکان گردد.

[DOR: 20.1.1.17359406.1391.5.4.3.3]

در خصوص متغیرهای بکار رفته پس از آنالیز انجام شده توسط نرم افزار کلمنتاین، ترتیب اهمیت متغیرها در آنالیز نهایی مشخص گردید. بر اساس نتایج حاصله، متغیرهای اثرگذار بر عود بیماری، مثل میزان درگیری غدد لنفی، اندازه تومور، التهاب غدد لنفی که توسط متخصصین نیز جزو ریسک فاکتورهای عود محسوب می‌شوند، در آنالیز نرم افزار کلمنتاین نیز در صدر عوامل خطرساز عود قرار گرفتند.

بحث و نتیجه گیری

داده‌کاوی قادر به کشف و استخراج دانش جدید از داده‌های گذشته نگر است. نحوه پیش پردازش داده‌ها و هم چنین متغیرهای منتخب، تاثیر قابل توجهی در کشف دانش دارد. تکنیک‌های داده‌های مختلفی وجود دارند که به منظور پیش بینی عود مجدد سرطان پستان به کار می‌روند. در این مقاله از ۲۲ متغیر با مقایسه سه روش طبقه‌بندی در داده کاوی استفاده شد که این مدل‌ها شامل شبکه‌های عصبی مصنوعی، درخت تصمیم C5.0 و ماشین بردار پشتیبان بودند. نتایج تجربی کارایی و موثر بودن هر سه روش را نشان می‌دهد که بر اساس حساسیت، ویژگی و دقت مقایسه شدند. برای پیدا کردن ساختار مطلوب و افزایش دقت و صحت نتایج، از هرس کردن و روش تقویت استفاده گردید. پایگاه داده موجود مورد بررسی قرار گرفته و داده‌ها به دو قسمت آموزش و آزمایشی تقسیم شدند که به ترتیب نسبت نتایج حاصل از آنها ۷۰ به ۳۰ است.

نتایج حاصله نشان داد که ماشین بردار پشتیبان با دقت ۰/۹۵۷، بهترین پیش بینی کننده در طبقه‌بندی بوده و دقت آزمون داده‌ها به کمک دو روش شبکه‌های عصبی مصنوعی و درختان تصمیم نیز به ترتیب ۰/۹۴۷ و ۰/۹۳۶ است. مدل شبکه‌های عصبی مصنوعی نیز با استفاده از آنالیز حساسیت به منظور شناسایی اهمیت نسبی فاکتورهای پیش بینی کننده تجزیه و تحلیل گردیدند. در تجزیه و تحلیل صورت گرفته، برخی از رکوردها شامل متغیرها و ویژگی‌های مفقود شده بودند.

مفقود بودن برخی از متغیرهای مهم مانند HER2، S–فاز و شاخص DNA می‌تواند بطور قابل توجهی صحت و درستی اطلاعات را کاهش دهد. به دلیل ناکافی بودن متغیرهای مورد نظر در این تحقیق، برخی از رکوردها به ناچار حذف گردیدند که این امر تعداد رکوردها را از ۱۱۸۹ به ۵۴۷ مورد کاهش داد که مسلماً این کاهش تعداد در نتایج حاصله اثرات چشم‌گیری داشت. نتایج تحقیق حاکی از آن است که در مقایسه سه مدل طبقه‌بندی SVM، درخت تصمیم گیری و مدل ANN، حداقل میزان خطا و بیشترین دقت مربوط به SVM بود. دقت پیش بینی در مدل درخت تصمیم‌گیری(C5.0) نیز پایین‌ترین میزان در بین سه مدل پیش بینی را نشان داد. نتایج به دست آمده نشان داد که SVM مناسب‌ترین مدل در میان سه مدل جهت پیش بینی عود مجدد سرطان پستان بود. همچنین متغیرهایی که توسط نرم افزار کلمنتاین جزو متغیرهای اثرگذار بر عود مجدد شناخته شدند، متغیرهایی بودند که از نظر پزشکان متخصص سرطان پستان نیز به عنوان متغیرهای پیش بینی عود تعیین گردیده‌اند. داده کاوی می‌تواند راهنمای پزشکان در پیش بینی عود سرطان پستان باشد و دقت نتایج حاصل از مدل‌ها نیز کاملا به واقعیت نزدیک است. در تحقیقات آتی می‌توان با ادغام چند پایگاه داده، تعداد رکوررها را افزایش داده و از متغیرهای کاربردی‌تر نیز استفاده کرد. همچنین می‌توان پیگیری ۵ ساله بیماران را نیز به جای دو سال لحاظ نمود.

[DOR: 20.1001.1.17359406.1391.5.4.3.3]

References

۱- غضنفری مهدی، علیزاده سمیه، تیمورپور بابک. داده کاوی وکشف دانش. انتشارات دانشگاه علم و صنعت ایران ۱۳۸۷.

۲- پورحسن مژگان، اتابکی گلناز، جودکی مجید، مینایی بیدگلی بهروز. رده بندی با استفاده از ترکیب قوانین انجمنی و درخت تصمیم. دومین کنفرانس داده کاوی ایران.

۳- دچرنی آلن. بیماری‌های زنان کارنت، ترجمه: قطبی نادر، نیک روش نادر، سلیمانی محمدرضا. چاپ دوم تهران، انتشارات تیمورزاده، ۱۳۷۹.

4- American Cancer Society (2006): URL http://www.cancer.org

5- Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

6- Calle J. Breast cancer facts and figures 2003-2004. American Cancer Society 2004. http://www.cancer.org/. (last accessed: Jan. 2010).

7- Breast cancer Q & A/ facts and statistics (http://www.komen.org/bci/bhealth/QA/q_and_a.asp).

8- Karabatak M, Cevdet M. An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications 2009; 36: 3465–9.

9- Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. J. Artificial Intelligence in Medicine 2010; 34: 113-27.

10- Yi W, Fuyong W. Breast cancer diagnosis via support vector machines. in Proc. the Twenty.

11- Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999; 57(4): 281-6.

12- Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M. Associations statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 1999; 17: 223–32.

13- SPSS Clementine 12.0, 2007. Data mining workbench software. Product of SPSS inc. http://www.cad100.net/247_data-mining-workbench-SPSS-Clementine-12.html

14- Jerez-Aragone´s JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 2003; 27: 45-63.

15- Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M. Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Syst Applic 1999; 17: 223-32.

16- Zhou ZH, Jiang Y. Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. IEEE Trans Inf Technol Biomed 2003; 7(1): 37-42.

17- Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine 2005; 34(2): 113-27.

18- Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco: Morgan Kaufmann; 2005.

19- Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network. Neural Netw 1990; 2:359-66.

20- Levine MS. Canonical Analysis and Factor Comparison: Sage Publications Inc; 1977.

21- Kovalerchuc B, Triantaphyllou E, Ruiz JF, Clayton J. Fuzzy logic in computer-aided breast- cancer diagnosis: Analysis of lobulation. Artificial Intelligence in Medicine 1997; 11: 75–85.

[DOR: 20.1.1.17359406.1391.5.4.3.3]

22- Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999; 16: 3-23.

23- Abdelghani Bellaachia, Erhan Guven. Predicting Breast Cancer Survivability 2008; 3:11-16

[DOR: 20.1001.1.17359406.1391.5.4.3.3]