مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال نهم، شماره اول، بهار ۱۳۹۵

ارائه مدلی جهت تعیین احتمال ابتلا به سرطان سینه با بکارگیری الگوریتم EM در عامل‌های خطر

مهدی حمصیان اتفاق: دانشکده مهندسی کامپیوتر، واحد نجف آباد، دانشگاه آزاد اسلامی، نجف آباد، اصفهان، ایران
محمدحسین ندیمی شهرکی: دانشکده مهندسی کامپیوتر، واحد نجف آباد، دانشگاه آزاد اسلامی، نجف آباد، اصفهان، ایران

چکیده

مقدمه: هر عاملی که باعث افزایش احتمال ابتلا به سرطان پستان در فرد شود عامل خطر محسوب می‌گردد. آگاهی از این عوامل خطر در تعیین زنانی که خطر بالای ابتلا به سرطان پستان دارند کمک کننده است و همچنین این اجازه را می‌دهد تا با مداخله در برخی از عوامل خطر فردی و اجتماعی، خطر بروز سرطان را در فرد و جامعه تعدیل کرد. هدف از این تحقیق ارایه مدلی ریاضی جهت تعیین میزان ریسک ابتلا به بیماری سرطان پستان در افراد مراجعه کننده به مراکز غربالگری می‌باشد.

روش بررسی: داده‌هایی که در این مقاله جهت ارزیابی مورد استفاده قرار گرفته‌اند از کنسرسیوم نظارت بر سرطان پستان (BCSC) وابسته به موسسه ملی سرطان آمریکا اقتباس و جهت انجام تحلیل داده‌های این مقاله از نرم‌افزار Weka استفاده شده است. این مطالعه بر روی ۶۳۱۸۶۳۸ نفر مراجعه‌کننده به مراکز غربالگری در کشور آمریکا از سال ۲۰۰۰ تا ۲۰۰۹ صورت گرفته است. جهت خوشه‌بندی موثر داده‌ها از الگوریتم EM نیز بهره برده شده است.

یافته‌ها: همه عوامل خطر به‌طور یکسان خطر سرطان پستان را افزایش نمی‌دهند، بعضی از آنها، عوامل خطرناک‌تری بوده و خطر ابتلا را بیشتر بالا می‌برند. لذا با تعیین ضریب تاثیرگذاری در جامعه آماری نمونه می‌توان به مدلی دست یافت که بر اساس آن امکان پیش‌بینی احتمال ابتلا به سرطان سینه وجود خواهد داشت.

نتیجه‌گیری: نتایج این مطالعه نشان داد که با استفاده از الگوریتم EM و خوشه‌بندی موثر می‌توان میزان تاثیرگذاری هر یک از عوامل خطر را در یک مجموعه داده مورد بررسی قرار داده و برای هر یک ضریبی را به عنوان ضریب موثر بدست آورده و در نهایت با جمع این ضرایب در مقدار عامل خطر میزان ریسک ابتلا به خطر سرطان سینه را پیش‌بینی نمود.

واژه‌های کلیدی: سرطان سینه، ریسک فاکتور، داده‌کاوی، غربالگری، الگوریتم EM.

* نشانی نویسنده پاسخگو: اصفهان، دانشگاه آزاد اسلامی واحد نجف آباد، دانشکده مهندسی کامپیوتر، مهدی حمصیان اتفاق.
نشانی الکترونیک: Mahdi_hemmasian@yahoo.com

مقدمه

سرطان پستان یکی از شایع‌ترین بیماری‌های زنان است. شناسایی و تشخیص زودهنگام این بیماری می‌تواند در درمان آن بسیار موثر باشد. سرطان پستان دومین سرطان شایع در دنیا و اولین سرطان شایع در زنان است. در مجموع، پنجمین علت مرگ ناشی از سرطان و اولین علت مرگ ناشی از سرطان در زنان در کشورهای توسعه نیافته سرطان سینه می‌باشد. سرطان پستان با تغییرات سلولی ایجاد می‌شود که بالاخره روزی به سرطان مبتلا می‌شوند. تئوری دوم یا تئوری جهش ژنی یا موتاسیون[1] است که در حقیقت همان سلول‌های طبیعی هر عضوی از جمله پستان، تحت تاثیر عوامل محیطی خود تغییر ماهیت داده و سرطانی می‌شوند. با وجودی که در دو دهه اخیر پیشرفت‌های مهمی در زمینه تشخیص زودرس و درمان به موقع و کاهش مرگ‌ومیر برای سرطان پستان در زنان ایجاد شده است اما هنوز این سرطان جزء شایع‌ترین بیماری‌های بدخیم زنان می‌باشد (۱).

این سرطان، اولین سرطان در بین زنان ایرانی است که منجر به از دست رفتن ۰/۰۱ سال به ازای هر یک هزار نفر جمعیت می‌شود و بار ناشی از سرطان پستان، ۶٪ سال‌های عمر از دست رفته به خاطر سرطان‌ها را در کشور به خود اختصاص می‌دهد (۲). با وجودی که در دو دهه اخیر پیشرفت‌های مهمی در زمینه تشخیص زودرس و درمان به موقع و کاهش مرگ‌ومیر برای سرطان پستان در زنان ایجاد شده است اما هنوز این سرطان جزء شایع‌ترین بیماری‌های بدخیم زنان می‌باشد. سرطان پستان می‌تواند همه خصوصیات دیگر سرطان‌ها را داشته باشد. تومور پستان معمولاً کیسه‌های شیری را گرفتار می‌کند اما در مواردی ساختارهای پشتیبانی کننده پستان نیز منشا تومورهای بدخیم می‌شود. خطر ایجاد سرطان پستان در طول عمر زنان ۱۲/۵٪ یعنی یک مورد از هشت مورد و خطر مرگ ناشی از سرطان پستان ۳/۶٪ یعنی حدود یک مورد از بیست و هشت مورد می‌باشد. در حدود ۵٪ سرطان‌های پستان ارثی و ۸۰ تا۹۰ درصد اسپورادیک هستند و اگرچه در سنین بالای ۵۰ سالگی شایع‌تر است ولی در هر سنی ممکن است رخ دهد. بر اساس پژوهش‌های انجام شده سرطان سینه فراوان‌ترین نوع سرطان در میان زنان ایرانی به شمار می‌رود. چنانچه این بیماری در مراحل اولیه شناسایی[2] شود می‌توان میزان بقا[3] بیمار را تا ۸۵٪ افزایش داد. با توجه به اینکه این بیماری، یک بیماری چند عاملی محسوب می‌شود، در صورتی‌که عامل‌های خطر به خوبی برای افراد مراجعه کننده به مراکز غربالگری شناسایی شوند و طبق عوامل محیطی که فرد با آن در ارتباط است، با تغییر سبک زندگی فرد مراجعه‌کننده و با بکارگیری شیوه‌های تغذیه درمانی[4] و ورزش درمانی[5] درصد ابتلا در فرد مراجعه‌کننده را می‌توان به شکل چشمگیری کاهش داد. لذا تعیین و بکارگیری مدلی که بتواند میزان درصد ابتلا را بر اساس معیارهای خطر پیش‌بینی کند بسیار حایز اهمیت می باشد.

عوامل خطرزا:

علت دقیق سرطان پستان تاکنون شناسایی نشده است و این بیماری به عنوان یک بیماری چند عاملی[6] محسوب می‌شود. از عواملی که در بیماری سرطان پستان دخیل هستند به عنوان عوامل خطرزا (Risk Factors) یاد می‌شود. ریسک فاکتورها به دو گروه ریسک فاکتورهای قطعی و نسبی تقسیم می‌شوند: ریسک فاکتورهای قطعی در تمام جمعیت و ریسک فاکتورهای نسبی در گروه مشخصی از جمعیت مورد بررسی قرار می‌گیرد. داشتن یک ریسک فاکتور حتی چند مورد به معنی ابتلای شخص به بیماری نخواهد بود. برخی از زنانی که سرطان پستان دارند هیچ ریسک ابتلا بعضاً نداشته‌اند و این در حالی است که همه زنان در ریسک ابتلا به سرطان پستان قرار دارند (۲). به اختصار در ذیل بعضی از عوامل خطرزای این بیماری ذکر شده است:

- سن (۳).
- سابقه فردی ابتلا به سرطان پستان (۴).
- سابقه خانوادگی (۵،۶).
- تغییرات ویژه در پستان (۷).
- تغییرات در ژن (۸-۱۰).
- سابقه باروری[7] و یائسگی[8] (۶،۱۱و۱۲).

[1] Mutation

[2] Early Stage
[3] Survivability
[4] Nutritional therapy
[5] Exercise therapy
[6] Multi Factorial
[7] Pregnancy
[8] Menopause

زنانی که اولین بار قبل از دوازده سالگی قاعده شده باشند برای ابتلا به سرطان پستان در معرض خطر بیشتری هستند.

◆در زنانی که بالاتر از سن ۵۵ سالگی یائسه می‌شوند خطر ابتلا به سرطان پستان بیشتر است.

◆بنابر مطالعات گسترده، هیچ ارتباطی بین سقط جنین یا حاملگی ناموفق با سرطان پستان وجود ندارد (۱۳).

- نژاد[9] (۱۴).
- پرتودرمانی قفسه سینه (۱۵،۱۶).
- تراکم پستان (۱۷).
- اضافه وزن یا چاقی[10] مفرط پس از یائسگی (۱۹و۱۴،۱۸).
- کمبود فعالیت بدنی (فیزیکی) (۲۰-۲۲).
- مصرف الکل (۲۳-۲۵).

داده‌کاوی: داده‌کاوی به معنای کشف دانش نهفته در محیط تحقیقاتی است (۲۶). داده‌کاوی از روش‌هایی است که در تشخیص یا پیش‌بینی انواع سرطان‌ها از جمله سرطان سینه بسیار استفاده می‌شود. رویکردهای داده‌کاوی می توانند با کاهش تعداد نتایج مثبت کاذب و منفی کاذب در تصمیم‌گیری پزشکان برای شناسایی بهتر سرطان پستان کمک کنند. مطالعات مکرری با استفاده از تکنیک‌های داده‌کاوی و با رویکرد پزشکی وجود دارند. به عنوان مثال دلن و همکاران از شبکه‌های عصبی مصنوعی، درخت تصمیم‌گیری و رگرسیون لجستیک برای توسعه مدل‌های پیش‌بینی سرطان پستان با تجزیه و تحلیل پایگاه‌داده‌های بزرگ داده بهره جستند. در مطالعات بسیاری از الگوریتم SVM برای دسته‌بندی تومورها استفاده شده است و در دسته‌ای دیگر از مطالعات از الگوریتم K-means برای بخش‌بندی تصاویر ماموگرافی بهره برده شده است، استفاده از تکنیک‌های داده‌کاوی روزبه‌روز افزایش یافته و دامنه کاربرد این علوم در حوزه پزشکی بسیار فراتر از گذشته شده است.

الگوریتم EM: یکی از انواع الگوریتم‌های کارا در خوشه‌بندی، در حالتی که تعداد خوشه‌های انتخاب شده تصادفی باشد الگوریتم EM می‌باشد. الگوریتم EM در اواخر سال‌های ۱۹۷۰ توسط روبین، دمپستر ولارد معرفی و گسترش داده شد که روش محاسباتی برای برآورد داده به خصوص داده‌های گمشده می‌باشد (۲۷-۲۹).

الگوریتم EM به صورت گسترده روشی قابل استفاده برای محاسبه برآوردهای ماکزیمم درستنمایی در محاسبات مکرر است. انتخاب نام EM به این علت است که در هر تکرار الگوریتم یک مرحله امید ریاضی‌گیری و بعد از آن یک ماکسیمم‌سازی انجام می‌شود. مدل احتمال به کار رفته در این الگوریتم، توزیع نرمال است. زیرا فرض می‌کند که مجموعه داده‌ها، می‌توانند به عنوان یک ترکیب خطی از توزیع نرمال چند متغیره در آیند.

مواد و روش‌ها

این مطالعه از نوع کمی و گذشته‌نگر[11] است. در این مطالعه عملیات مربوط به داده‌کاوی و کشف روابط پنهان بین ریسک فاکتورها و همچنین میزان تاثیر آنها از مجوعه داده BCSC استفاده شده است. مجموعه داده فوق، حاصل اطلاعات جمع‌آوری شده ۶۳۱۸۶۳۸ نفر از ژانویه ۲۰۰۰ تا دسامبر سال ۲۰۰۹ در کشور آمریکا می‌باشد که شامل اطلاعات مفیدی از قبیل سن، نژاد/ قومیت، سابقه خانوادگی سرطان پستان، سن شروع قاعدگی، سن در هنگام تولد اولین نوزاد، تراکم پستان، استفاده از درمان جایگزینی هورمون، وضعیت یائسگی، BMI، سابقه بیوپسی و سابقه سرطان پستان می‌باشد. علت انتخاب مجموعه داده فوق جامعه آماری بسیار بالا بوده که امکان کشف روابط همبستگی و وابستگی بین ریسک فاکتورها را به شکل خوبی فراهم می‌نماید. این تحقیق از نوع گذشته نگر و کمی می‌باشد.

متغیرهای این مجموعه داده که در حقیقت همان ریسک فاکتورها می‌باشند طبق جدول شماره ۱ تقسیم‌بندی شده‌اند. لازم بذکر است به علت محاسبات بسیار سنگین مجموعه داده فوق، نتایج تحقیق این مقاله در بازه زمانی ۲۰۰۵ تا ۲۰۰۹ ارایه خواهد شد. در نمودارهای زیر میزان پراکندگی هر کدام از ریسک فاکتورها به نمایش در آمده است.

[9] Ethnicity
[10] Obesity
[11] Retrospective

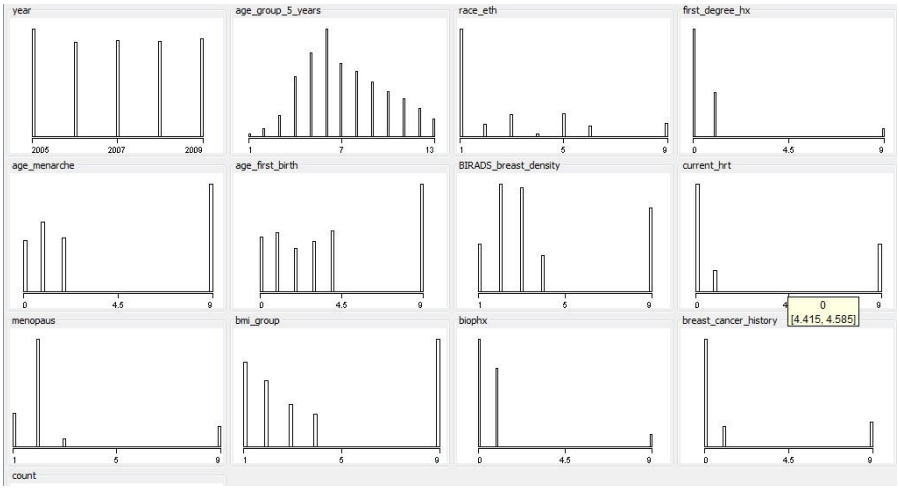
جدول ۱: خروجی حاصل از اجرای الگوریتم EM در محیط Weka (سال ۲۰۰۹)

متغیر	خوشه	۰	۱	۲	۳	۴	۵	۶
	مقدار	۱۴.۰-	۱۴.۰-	۰۵.۰-	۱۳.۰-	۲۵.۰-	۰۷.۰-	۲۳.۰-
گروه سنی								
	میانگین	۵۹۳۱.۲	۷۰۰۹.۲	۳۵۵۴.۲	۴۴۰۹.۲	۵۸۸۷.۲	۴۹۱۴.۲	۶۰۹۱.۲
توزیع استاندارد	انحراف معیار	۶۲۵۷.۰	۵۵۳۱.۰	۷۵۹۹.۰	۷۱۵۴.۰	۶۴۱۲.۰	۷۰۸۱.۰	۶۰۸۹.۰
قومیت								
	میانگین	۲۰۵۴.۲	۹۳۱۸.۱	۷۱۱۸.۳	۱۶۳۸.۴	۹۳۱۵.۲	۵۶۱۸.۳	۰۲۸۸.۲
توزیع استاندارد	انحراف معیار	۳۹۷۴.۲	۶۹۵۱.۱	۱۱۹۶.۳	۲۰۹۶.۳	۲۰۹۱.۲	۹۸۴۶.۲	۹۴۵۸.۱
سابقه سرطان در فامیل درجه یک								
	میانگین	۸۷۲۴.۰	۵۴۹۹.۱	۸۶۰۸.۰	۲۵۹۱.۰	۸۱۶.۰	۲۱۴۹.۲	۰۸۱۶.۱
توزیع استاندارد	انحراف معیار	۰۹۶۷.۲	۰۴۷۱.۳	۲۳۴.۲	۵۲۶۶.۰	۰۹۴۳.۲	۵۷۹۷.۳	۴۵۷۵.۲
سن منارک								
	میانگین	۱۶۶۲.۱	۳۸۷.۴	۸۶۴۳.۸	۹	۵۱۶۴.۸	۰۸۴۳.۶	۱۳۲.۱
توزیع استاندارد	انحراف معیار	۷۸۱۲.۰	۹۱۳۳.۳	۹۸۸۷.۰	۰۱۷.۰	۸۷۳۵.۱	۸۲۳۴.۳	۷۶۵۵.۰
سن اولین بارداری								
	میانگین	۵۲۴۷.۷	۲۵۷۷.۶	۳۳۷۲.۸	۷۷۷۸.۶	۰۸۰۴.۵	۷۸۹۵.۸	۹۴۲۱.۱
توزیع استاندارد	انحراف معیار	۸۴۹۵.۲	۴۲۶۴.۳	۹۱۵۴.۱	۱۹۵۸.۳	۹۲۰۱.۲	۰۵۱۴.۱	۵۲۱۳.۱
BIRADS تراکم پستان								
	میانگین	۹۲۵۶.۳	۴۲۷۹.۴	۳۶۸۲.۳	۹۵۹۱.۳	۲۷۷.۴	۰۵۴۱.۴	۱۳۹۹.۴
توزیع استاندارد	انحراف معیار	۷۰۵۳.۲	۸۷۹۹.۲	۲۵۹۹.۲	۶۳۶۷.۲	۸۵۵۶.۲	۶۸۱۷.۲	۷۶۰۹.۲
هورمون درمانی								
	میانگین	۰	۶۱۹۲.۶	۰۱۶۹.۲	۸۶۳۳.۰	۰	۹۱۲.۸	۰
توزیع استاندارد	انحراف معیار	۵۶۵۹.۳	۸۶۵۷.۳	۷۲۵۴.۳	۵۰۶.۲	۵۶۵۹.۳	۸۸۵۶.۰	۵۶۵۹.۳
یائسگی								
	میانگین	۸۸۵۹.۱	۵۴۰۲.۱	۲۲۰۸.۱	۳۴۴۲.۱	۵۱۰۹.۱	۸۱۵.۸	۷۳۷۷.۱
توزیع استاندارد	انحراف معیار	۴۴۶۲.۲	۶۴۲۱.۱	۵۹۱۷.۰	۸۰۵۴.۰	۷۵۳۲.۱	۱۹۷۸.۱	۰۷۳.۲
BMI								
	میانگین	۰۷۹۲.۴	۰۶۴۳.۳	۶۱۶۹.۸	۵۲۴۳.۸	۹۰۸۲.۱	۶۱۸۷.۶	۱۲۵۸.۲
توزیع استاندارد	انحراف معیار	۳۱۷۲.۳	۸۴۵۲.۲	۵۸۶۹.۱	۸۴۱۸.۱	۰۳۶۷.۱	۴۱۴۹.۳	۱۲۸۸.۱
سابقه بیوپسی								
	میانگین	۵۴۶۵.۱	۱۸۹۸.۱	۸۷۹۱.۴	۳۲۸۲.۰	۳۱۴.۰	۵۸۰۹.۲	۴۴۳۱.۰
توزیع استاندارد	انحراف معیار	۰۸۷۳.۳	۷۲۹۷.۲	۲۸۴۵.۴	۴۶۹۵.۰	۴۶۴۱.۰	۷۷۰۵.۳	۴۸۹۱.۱
سابقه سرطان پستان در بیمار								
	میانگین	۰۵۶۷.۰	۴۴۷.۲	۹	۱۰۵۸.۰	۱۰۷۵.۰	۷۰۵۲.۴	۰۰۰۹.۰
توزیع استاندارد	انحراف معیار	۲۳۱۳.۰	۹۴۷۸.۳	۹۳۴۶.۲	۳۰۷۶.۰	۳۰۹۷.۰	۴۲۰۵.۴	۰۳۰۸.۰

با استفاده از تکنیک خوشه‌بندی در داده‌کاوی، اطلاعات مجموعه داده فوق طبق جدول شماره ۲ تقسیم‌بندی شده است تا میزان مشارکت هر یک از ریسک فاکتورها در هر خوشه بررسی گردد. در شکل شماره ۱ نمودارهای هیستوگرام بر اساس ریسک فاکتورهای استفاده شده در مطالعه از سال ۲۰۰۰ تا سال ۲۰۰۹ نمایش داده شده است. پس از مطالعه و بررسی هر یک از خوشه‌ها، میزان همبستگی و عضویت هر یک از ریسک فاکتورها در خوشه‌ها مشاهده می‌شود که این موضوع بیانگر این است که به عنوان مثال ریسک فاکتور سن اولین زایمان با کدام و یا کدامین ریسک فاکتورها در یک دسته قرار گرفته است. در ذیل تفسیر کامل هر یک از این ریسک فاکتورها را که در مجموعه داده BCSC بیان شده است ارایه می‌گردد.

خوشه‌بندی مجموعه داده فوق با استفاده از الگوریتم EM صورت پذیرفته است که یکی از بهترین الگوریتم‌های خوشه‌بندی می‌باشد.

$$l(\theta) = \log p(D \mid \theta) = \log \sum_{H} p(D, H \mid \theta)$$



شکل۱: نمودارهای هیستوگرام بر اساس ریسک فاکتورهای استفاده شده در مطالعه (سال۲۰۰۰- ۲۰۰۹)

جدول ۲: دسته‌بندی متغیرهای مورد مطالعه

تعریف متغیرها و نحوه مقداردهی آنها					
f_d_hx	۱	سابقه سرطان سینه در فامیل درجه یک دارد	c_hrt	۱	سابقه هورمون درمانی دارد
	۰	سابقه سرطان سینه در فامیل درجه یک ندارد		۰	سابقه هورمون درمانی ندارد
a_m	۱	سن یائسگی≤ ۵۰	m	۱	سن منارک ≥ ۱۳
	۰	سن یائسگی> ۵۰		۰	سن منارک < ۱۳
a_f_b	۱	سن اولین بارداری ≤ ۳۰	bmi_group	۱	قرارگیری در طبقه‌ی BMI ≤ ۲۵
	۰	سن اولین بارداری>۳۰		۰	قرارگیری در طبقه‌ی BMI > ۲۵
Biophx	۱	سابقه بیوپسی از پستان دارد	۰ ≤ E ≤ ۱۰۰		
	۰	سابقه بیوپسی از پستان ندارد			

یافته‌ها

همان‌طوری که در ضمیمه ۲ مشاهده می‌گردد با اجرای الگوریتم EM و دسته بندی ریسک فاکتورها بالاترین میزان مشارکت افراد در خوشه ۶ می‌باشد. در حقیقت در خوشه ۶ ریسک فاکتورها به نحوی قرار گرفته‌اند که شامل افرادی است که بالاترین ریسک ابتلا به سرطان سینه در آنها توسط پزشک اندازه‌گیری شده است. به منظور نمایش خروجی الگوریتم EM که در محیط نرم افزار Weka اجرا شده است قسمتی از آن که شامل افراد مورد مطالعه در سال ۲۰۰۹ می‌باشد در جدول ذیل ارایه می‌گردد. با اجرای الگوریتم EM داده‌های موجود در مجموعه داده بر اساس میزان تجمعی بودن و فراوانی ریسک فاکتورها در یک شخص مبتلا دسته‌بندی شدند. هر یک از این دسته‌بندی‌ها افرادی از لحاظ ریسک فاکتور موثری که پزشک برای آنها تعیین نموده بود شمارش شده‌اند. به عنوان مثال در ریسک فاکتور گروه سنی سال ۲۰۰۹ در خوشه ۶ دارای میانگین ۲/۶۰۹۱ می‌باشد که این نشانگر این موضوع است که میانگین گروه سنی این افراد طبق جدول پیوست نزدیک به گروه ۳ یعنی سن بین ۳۵ تا ۳۹ سال می‌باشد. دقت عملکرد این مدل در مطالعه انجام شده ۸۷/۹۹٪ بوده است که حاصل مقایسه خروجی الگوریتم و نظر قطعی پزشک در پایان می‌باشد.

با توجه به تحلیل‌های فوق و آنالیز داده‌ها می‌توانیم به فرمول زیر اشاره نماییم که میزان احتمال ابتلا به ریسک یک نفر را به ما اعلام می‌دارد:

E= (f_d_hx *0.05 + a_m*0.04 + a_f_b *0.1 + c_hrt *0.07 + m *0.1 + bmi_group*0.04 + biophx*0.1)*200

فرمول ارایه شده در این مطالعه نشانگر این است که در نمونه های مورد مطالعه سن منارک و سابقه بیوپسی دارای بالاترین تاثیر و ضریب ۰/۱ در جامعه آشکار گردیدند و ریسک فاکتور سابقه هورمون درمانی و سن اولین بارداری کمترین تاثیر را در ابتلا افراد داشته است.

عدد حاصل از رابطه فوق بین ۰ تا ۱۰۰ خواهد بود که میزان احتمال ریسک ابتلا به یک فرد را با استفاده از ریسک فاکتورهای سابقه سرطان پستان در فامیل درجه۱(دارد، ندارد)، سن یائسگی (**≥۵۰ و <۵۰**)، سابقه بیوپسی، سابقه هورمون درمانی، سابقه menopause زودرس، سن اولین زایمان و قرارگیری در گروه‌بندی با BMI خاص نشان می‌دهد.

ضمیمه۱: قوانین مستخرج از الگوریتم EM برای تعیین میزان تاثیر و مشارکت عوامل خطر

قوانین مستخرج
1. current_hrt=0 biophx=0 2593 ==> breast_cancer_history=0 2475 conf:(0.95)
2. race_eth=1 current_hrt=0 biophx=0 1424 ==> breast_cancer_history=0 1359 conf:(0.95)
3. first_degree_hx=0 current_hrt=0 biophx=0 1592 ==> breast_cancer_history=0 1519 conf:(0.95)
4. current_hrt=0 menopaus=1 biophx=0 2217 ==> breast_cancer_history=0 2106 conf:(0.95)
5. first_degree_hx=0 current_hrt=0 menopaus=1 biophx=0 1319 ==> breast_cancer_history=0 1250 conf:(0.95)
6. age_group_5_years=3 current_hrt=0 biophx=0 1646 ==> breast_cancer_history=0 1558 conf:(0.95)
7. age_group_5_years=3 current_hrt=0 menopaus=1 biophx=0 1371 ==> breast_cancer_history=0 1288 conf:(0.94)
8. first_degree_hx=0 menopaus = 1 biophx=0 1529 ==> breast_cancer_history=0 1434 conf:(0.94)
9. first_degree_hx=0 biophx=0 1977 ==> breast_cancer_history=0 1852 conf:(0.94)
10. race_eth=1 menopaus=1 biophx=0 1413 ==> breast_cancer_history=0 1323 conf:(0.94)

ضمیمه ۲: دسته‌بندی ریسک فاکتورها و افراد مورد مطالعه بر اساس میزان ریسک ابتلا به سرطان سینه

نام متغیر	توضیحات		کدینگ
سال	**سال مشاهده بیمار**		**۲۰۰۰- ۲۰۰۹**
گروه سنی	سن بیماران به ۱۳ گروه تقسیم شده است	۱	۲۹ ≤ سن ≤ ۱۸
		۲	۳۴ ≤ سن ≤ ۳۰
		۳	۳۹ ≤ سن ≤ ۳۵
		۴	۴۴ ≤ سن ≤ ۴۰
		۵	۴۹ ≤ سن ≤ ۴۵
		۶	۵۴ ≤ سن ≤ ۵۰
		۷	۵۹ ≤ سن ≤ ۵۵
		۸	۶۴ ≤ سن ≤ ۶۰
		۹	۶۹ ≤ سن ≤ ۶۵
		۱۰	۷۴ ≤ سن ≤ ۷۰
		۱۱	۷۹ ≤ سن ≤ ۷۵
		۱۲	۸۴ ≤ سن ≤ ۸۰
		۱۳	۸۵ ≤ سن
قومیت	طبقه بندی بیماران بر اساس قومیت	۱	غیر اسپانیائی سفید پوست
		۲	غیر اسپانیائی سیاه پوست
		۳	جزایز اقیانوس آرام
		۴	بومی آمریکائی
		۵	اسپانیائی
		۶	متفرقه
		۹	نامشخص
سابقه سرطان پستان	سابقه سرطان پستان در فامیل درجه یک	۰	خیر
		۱	بلی
		۹	نامشخص
سن منارک	طبقه‌بندی سن منارک بیمار	۰	سن ≤ ۱۴
		۱	۱۳ ≤ سن ≤ ۱۲
		۲	۱۲ ≤ سن
		۹	نامشخص
سن اولین بارداری	سن بیمار به هنگام اولین زایمان	۰	۲۰ ≤ سن
		۱	۲۴ ≤ سن ≤ ۲۰
		۲	۲۹ ≤ سن ≤ ۲۵
		۳	سن ≤ ۳۰
		۴	نخست‌زا
		۹	نامشخص
BIRADS تراکم پستان	طبقه‌بندی تراکم بافت پستان	۱	تقریباً بطور کامل متراکم
		۲	غده‌های متراکم و پراکنده
		۳	متراکم ناهمگن
		۴	بسیار متراکم
		۹	نامشخص
هورمون	هورمون درمانی	۰	خیر
		۱	بلی

ادامه ضمیمه ۲

	۹	نامشخص
یائسگی	۱	یائسگی زودرس
وضعیت یائسگی بیمار	۲	یائسگی دیررس
	۳	یائسگی ناشی از جراحی
	۹	نامشخص
گروه‌بندی BMI	۱	$۱۰ \leq BMI \leq ۲۴,۹۹$
	۲	$۲۵ \leq BMI \leq ۲۹,۹۹$
BMI بیماران	۳	$۳۰ \leq BMI \leq ۳۴,۹۹$
	۴	$۳۵ \leq BMI$
	۹	نامشخص
بیوپسی	۰	خیر
انجام بیوپسی در گذشته	۱	بلی
	۹	نامشخص
سابقه سرطان پستان در بیمار	۰	خیر
مشخص کننده اینکه بیمار قبلاً مبتلا بوده یا خیر؟	۱	بلی
	۹	نامشخص

بحث

با استفاده از تکنیک‌های داده‌کاوی به راحتی می‌توانیم به کشف روابط پنهان بین داده‌های یک مجموعه پرداخته و حتی به مدل‌سازی آن‌ها بپردازیم. الگوریتم EM ما را به سمت انتخاب بهترین خوشه هدایت می‌کند. این الگوریتم همگرایی قابل اعتمادی دارد. یعنی با شروع از هر تعداد خوشه، همگرایی تقریباً همیشه به یک نقطه ماکسیمم موضعی حتمی است. مطالعه فعلی در یک جامعه آماری با حجم بالا و با هدف کشف روابط پنهان و میزان تاثیرگذاری هر یک از ریسک فاکتورهای مورد مطالعه در سرطان پستان صورت پذیرفت. جامعه آماری با حجم بالا افزایش دقت پیش بینی را به همراه خواهد داشت. دقت پیش‌بینی در این مطالعه ۸۷/۹۹٪ می‌باشد. برای انجام پژوهش، نیمی از داده‌ها با رعایت اصول حفظ نمونه آماری برای کشف الگو انتخاب گردید و نیمی دیگر از داده‌ها برای تست و ارزیابی الگوی کشف شده مورد استفاده قرار گرفت.

نتیجه‌گیری

در این پژوهش با استفاده از الگوریتم EM که یکی از الگوریتم‌های داده‌کاوی می‌باشد اقدام به دسته‌بندی ریسک فاکتورها گردید. در پژوهش‌های مشابه از الگوریتم‌های مختلفی همچون SVM و شبکه‌های عصبی استفاده شده است. منتها یکی از تفاوت‌های بارز و نوآورانه این مطالعه با پژوهش‌های قبلی این است که در این مطالعه برای هر دسته‌بندی ضریب تاثیرگذاری اختصاص یافته است. در این تحقیق با فرض اینکه تمامی ریسک فاکتورهای شناخته شده و موثر در سرطان سینه به یک میزان فرد را دچار ابتلا نمی‌کنند و هر کدام میزان اثرگذاری خاصی دارند اقدام به دسته‌بندی این ریسک فاکتورها و میزان اثرگذاری آن‌ها در مجموعه داده شد. سپس از اعمال الگوریتم بر روی مجموعه داده، بر اساس تعداد معیارهای خطر مورد مطالعه، این مجموعه به ۷ خوشه تقسیم گردید و در نتیجه ۷ معیار پرخطر که در این مجموعه اثرگذارتر از دیگر معیارها بودند مشخص گردید. در انتها نسبت به تقسیم و تسهیم این معیارها در تحقیق اقدام شد که هر یک به تنهایی چه درصدی از کل مجموعه را تحت تاثیر قرار داده بودند و در انتهای کار برای هر یک از معیارهای خطر بر اساس کسری از مجموعه که موثر بودند ضریبی به آن اختصاص داده شد. حاصل مدل ریاضی ارایه شده میزان ریسک ابتلا به سرطان سینه در افراد مراجعه‌کننده به مراکز غربالگری سرطان سینه را پیش‌بینی خواهد نمود.

References

۱. اکبری محمداسماعیل، سرطان در ایران. انتشارات دارالفکر قم؛ ۱۳۸۷.

۲.گیتی معصومه. اصول تشخیص و درمانی بیماری‌های پستان. انتشارات رضوان پرتو؛ ۱۳۸۱.

3. Arvold ND, Taghian AG, Niemierko A, Raad RFA, Sreedhara M, Nguyen PL, Bellon JR, Wong JS, Smith BL, Harris JR. Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. Journal of Clinical Oncology 2011; 29(29):3885-91.

4. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M. The life history of 21 breast cancers. Cell 2012; 149(5): 994-1007.

5. Phipps AI, Buist DS, Malone KE, Barlow WE, Porter PL, Kerlikowske K, Li CI. Family history of breast cancer in first-degree relatives and triple-negative breast cancer risk. Breast cancer research and treatment 2011; 126(3):671-8.

6. Augustinsson A, Ellberg C, Kristoffersson U, Olsson H. Increasing age at first full-time pregnancy correlates to use of oral contraceptives before age 20 in women with a family history of breast cancer. Cancer Research 2015; 75(15):2747.

7. Rasmussen CB, Kjær SK, Ejlertsen B, Andersson M, Jensen MB, Christensen J, Langballe R, Mellemkjær L. Incidence of metachronous contralateral breast cancer in Denmark 1978–2009. International journal of epidemiology 2014; 202.

8. Norquist B, Harrell M, Walsh T, Mandell J, Bernards S, Agnew K, Lee M, Pennington K, King M, Swisher E. Abstract AS09: Germline mutations in cancer susceptibility genes in BRCA1 and BRCA2 negative families with ovarian and breast cancer. Clinical Cancer Research 2015; 21:(16): AS09.

9. Greenup R, Buchanan A, Lorizio W, Rhoads K, Chan S, Leedom T, King R, McLennan J, Crawford B, Marcom PK. Prevalence of BRCA mutations among women with triple-negative breast cancer (TNBC) in a genetic counseling cohort. Annals of surgical oncology 2013; 20(10): 3254-8.

10. Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkäs K, Roberts J, Lee A, Subramanian D, De Leeneer K, Fostira F. Breast-cancer risk in families with mutations in PALB2. New England Journal of Medicine 2014; 371(6): 497-506.

11. Azim HA, Santoro L, Russell-Edu W, Pentheroudakis G, Pavlidis N, Peccatori FA. Prognosis of pregnancy-associated breast cancer: a meta-analysis of 30 studies. Cancer treatment reviews 2012; 38(7):834-42.

12. Azim HA, Kroman N, Paesmans M, Gelber S, Rotmensz N, Ameye L, De Mattos-Arruda L, Pistilli B, Pinto A, Jensen MB. Prognostic impact of pregnancy after breast cancer according to estrogen receptor status: a multicenter retrospective study. Journal of clinical oncology 2013; 31(1):73-9.

13. Cancer C.G.o.H.F.i.B. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. The lancet oncology 2012; 13(11):1141-51.

14. Amadou A, Ferrari P, Muwonge R, Moskal A, Biessy C, Romieu I, Hainaut P. Overweight, obesity and risk of premenopausal breast cancer according to ethnicity: a systematic review and dose-response meta-analysis. Obesity Reviews 2013; 14(8): 665-78.

15. Mamounas EP, Bandos H, White JR, Julian TB, Khan AJ, Shaitelman SF, Torres MA, McCloskey SA, Vicini FA, and Ganz PA. Abstract OT1-3-02: Will chest wall and regional nodal radiotherapy post mastectomy or the addition of regional nodal radiotherapy to breast radiotherapy post lumpectomy reduce the rate of invasive cancer events in patients with positive axillary nodes who convert to ypN0 af. Cancer Research 2015; 75(9): OT1-3-02-OT01-03-02.

16. Group E.B.C.T.C. Effect of Radiotherapy After Breast-Conserving Surgery on 10-Year Recurrence and 15-Year Breast Cancer Death: Meta-Analysis of Individual Patient Data for 10,801 Women in 17 Randomized Trials. Obstetrical & Gynecological Survey 2012; 67(2):92-4.

17. Gierach GL, Ichikawa L, Kerlikowske K, Brinton LA, Farhat GN, Vacek PM, Weaver DL, Schairer C, Taplin SH, Sherman ME. Relationship between mammographic density and breast cancer death in the Breast Cancer Surveillance Consortium. Journal of the National Cancer Institute 2012.

18. Ewertz M, Jensen MB. Gunnarsdóttir KÁ, Højris I, Jakobsen EH, Nielsen D, Stenbygaard LE, Tange UB, Cold S. Effect of obesity on prognosis after early-stage breast cancer. Journal of Clinical Oncology 2011; 29(1):25-31.

19. Anderson GL, Neuhouser ML. Obesity and the risk for premenopausal and postmenopausal breast cancer. Cancer prevention research 2012; 5(4):515-21.

20. Steindorf K, Ritte R, Eomois PP, Lukanova A, Tjonneland A, Johnsen NF, Overvad K, Østergaard JN, Clavel-Chapelon F, Fournier A. Physical activity and risk of breast cancer overall and by hormone receptor status: the European prospective investigation into cancer and nutrition. International Journal of Cancer 2013; 132(7):1667-78.

21. Friedenreich CM. Physical activity and breast cancer: review of the epidemiologic evidence and biologic mechanisms. Clinical Cancer Prevention 2011;125-39.

22. Beasley JM, Kwan ML, Chen WY, Weltzien EK, Kroenke CH, Lu W, Nechuta SJ, Cadmus- Bertram L, Patterson RE, Sternfeld B. Meeting the physical activity guidelines and survival after breast cancer: findings from the after breast cancer pooling project. Breast cancer research and treatment 2012; 131(2):637-43.

23. Newcomb PA, Kampman E, Trentham-Dietz A, Egan KM, Titus LJ, Baron JA, Hampton JM, Passarelli MN, Willett WC. Alcohol consumption before and after breast cancer diagnosis: associations with survival from breast cancer, cardiovascular disease, and other causes. Journal of clinical oncology 2013; JCO. 2012. 2046. 5765.

24. Liu Y, Colditz GA, Rosner B, Berkey CS, Collins LC, Schnitt SJ, Connolly JL, Chen WY, Willett WC, Tamimi RM. Alcohol intake between menarche and first pregnancy: a prospective study of breast cancer risk. Journal of the National Cancer Institute 2013; djt213.

25. Chen WY, Rosner B, Hankinson SE, Colditz GA, Willett WC. Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk. Jama 2011; 306(17):1884-90.

26. Han J, Kamber M, Pei J. Data mining: concepts and techniques: concepts and techniques. Elsevier 2011.

27. Mustapha N, Jalali M, Bozorgniya A, Jalali M. Navigation Patterns Mining Approach based on Expectation Maximization Algorithm. World Academy of Science, Engineering and Technology, 2009; 50:855-9.

28. Gupta MR, Chen Y. Theory and use of the EM algorithm. Now Publishers Inc 2011.

29. Hierarchical E. Clustering Dynamic Textures with the Hierarchical EM Algorithm for Modeling Video. Ieee transactions on pattern analysis and machine intelligence 2013; 35(7).