

## تشخیص ابتلا به سرطان پستان با استفاده از روش‌های درخت تصمیم، شبکه عصبی و بیز ساده به منظور ارائه مدل بومی ویژه استان فارس

آزیتا یزدانی<sup>۱</sup>، علی اصغر صفایی<sup>۲</sup>، رضا صفدری<sup>۳</sup>، مریم زحمت کشان<sup>۴</sup>

### چکیده

زمینه و هدف: سرطان پستان شایع‌ترین سرطان و اصلی‌ترین علت مرگ ناشی از سرطان در زنان سراسر جهان به‌شمار می‌رود. تکنولوژی‌هایی مثل داده کاوی، به متخصصان این حوزه، امکان بهبود تصمیم‌گیری را در زمینه‌ی تشخیص زودهنگام فراهم آورده‌اند. هدف از این پژوهش توسعه‌ی مدل تشخیص خودکار سرطان پستان با به‌کارگیری روش‌های داده کاوی و انتخاب مدل بومی ویژه بیماران استان فارس با بالاترین دقت تشخیص می‌باشد.

روش بررسی: در این مطالعه، تعداد ۶۵۴ پرونده در دسترس از بیماران کلینیک تخصصی سرطان پستان مطهری شیراز به‌عنوان نمونه مورد استفاده قرار گرفت که بعد از عملیات پیش پردازش این تعداد به ۶۲۱ پرونده کاهش یافت. برای هر کدام از نمونه‌ها دارای ۲۲ ویژگی در پرونده پزشکی ثبت شده بود که در نهایت ۱۰ ویژگی تاثیرگذار در ساخت مدل استفاده شد. از سه روش درخت تصمیم، بیز ساده و شبکه عصبی مصنوعی به‌منظور تشخیص ابتلا به سرطان پستان و روش 10-fold cross-validation برای ساخت و ارزیابی مدل بر روی مجموعه داده‌ی جمع‌آوری شده بهره گرفته شد.

یافته‌ها: نتایج به‌دست آمده از سه تکنیک ذکر شده نشان داد که هر سه مدل، نتایج امیدبخشی در تشخیص این سرطان دارند. در نهایت، شبکه عصبی مصنوعی، بالاترین دقت ۹۴/۴۹٪ (حساسیت ۹۶/۱۹٪، ویژگی ۸۶/۳۶٪)، در تشخیص ابتلا به سرطان پستان به خود اختصاص داد.

نتیجه‌گیری: بر طبق نتایج حاصل از درخت تصمیم ایجاد شده، ریسک فاکتورهایی چون سن، وزن، سن شروع قاعدگی، یائسگی، مدت زمان مصرف OCP و سن اولین بارداری از جمله عوامل موثر در ابتلای زنان به سرطان پستان در استان فارس شناخته شدند.

واژه‌های کلیدی: سرطان پستان، مدل تشخیص، درخت تصمیم، بیز ساده، شبکه عصبی، عوامل خطرزا

دریافت مقاله: دی ۱۳۹۷

پذیرش مقاله: اردیبهشت ۱۳۹۸

\* نویسنده مسئول:

علی اصغر صفایی؛

دانشکده علوم پزشکی دانشگاه تربیت

مدرس

Email :  
aa.safaei@modares.ac.ir

۱ استادیار گروه فناوری اطلاعات سلامت، دانشکده مدیریت و اطلاع‌رسانی پزشکی، دانشگاه علوم پزشکی شیراز، شیراز، ایران

۲ استادیار گروه انفورماتیک پزشکی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران

۳ استاد گروه مدیریت اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران

۴ دکتری تخصصی مدیریت اطلاعات سلامت، مرکز تحقیقات بیماری‌های غیر واگیر، دانشگاه علوم پزشکی فسا، فسا، ایران

## مقدمه

طبق گزارش سازمان بهداشت جهانی، سرطان پستان شایع‌ترین سرطان در میان زنان است که سالانه ۱۰۵ میلیون زن به آن مبتلا می‌شوند و همچنین علت اصلی مرگ و میر ناشی از سرطان در زنان سراسر جهان به شمار می‌رود (۱). اگرچه دانشمندان علت دقیق سرطان پستان را نمی‌دانند، اما برخی از عوامل خطرزا که موجب افزایش احتمال ابتلای زنان به این سرطان می‌شوند، شناسایی شده‌اند. برخی از مهم‌ترین این ریسک فاکتورها شامل سن، سابقه خانوادگی و خطر ژنتیکی و غیره می‌باشند که در تشخیص‌های اولیه استفاده می‌شوند. در تشخیص‌های زود هنگام این سرطان، ۹۷٪ از زنان برای ۵ سال یا بیشتر زنده مانده‌اند (۲). با توجه به نرخ رشد و وقوع سرطان پستان در سطح جهان، تشخیص زودهنگام سرطان پستان در کاهش تلفات حیاتی امری ضروری به نظر می‌رسد (۳). تشخیص کارآمد، دقیق و زود هنگام سرطان پستان یک مشکل مهم پزشکی در دنیای واقعی است. به‌تازگی روش‌های یادگیری ماشین و داده کاوی به‌طور گسترده در پیش‌بینی‌ها، به‌ویژه در تشخیص پزشکی استفاده می‌شوند. تشخیص پزشکی یکی از مشکلات عمده در برنامه‌های کاربردی پزشکی است (۴). با گسترش استفاده از رایانه‌ها به همراه ابزارهای خودکار، حجم زیادی از اطلاعات پزشکی جمع‌آوری و در اختیار محققان علوم پزشکی قرار گرفته است که با به‌کارگیری تکنیک‌های داده کاوی روابط و الگوهای پنهان در این حجم زیاد داده‌های جمع‌آوری شده کشف می‌شوند (۵). داده کاوی شامل روش‌های مختلف با اهداف گوناگون است. طبقه‌بندی و خوشه‌بندی دو روش رایج در داده کاوی هستند که در علوم پزشکی نیز استفاده می‌شوند. مشکلات تشخیص سرطان پستان اساساً در محدوده‌ی مشکلات طبقه‌بندی مورد بحث قرار می‌گیرند (۶). تاکنون تحقیقات گسترده‌ای در خصوص پیش‌بینی احتمال مبتلا شدن به سرطان پستان (۷)، تشخیص مبتلا بودن به سرطان پستان (۸)، پیش‌بینی مدت زمان بقای بیمار مبتلا به سرطان پستان (۹) و در نهایت احتمال عود مجدد سرطان پستان (۱۰) صورت گرفته است. Umer Khan و همکاران (۱۱) یک مدل ترکیبی بر اساس درخت تصمیم‌گیری فازی بر روی بانک اطلاعاتی "SEER" را بررسی کردند، آنها با استفاده از ترکیبات مختلف، تعدادی از قوانین درخت تصمیم‌گیری، انواع توابع عضویت فازی و تکنیک‌های استنتاج را آزمایش و بررسی نمودند. همچنین عملکرد هر یک از پیش‌آگهی‌های سرطان را مقایسه کردند و طبقه‌بندی درخت تصمیم‌گیری ترکیبی فازی با دقت ۸۵٪ را قوی‌تر و متعادل‌تر از طبقه‌بندی‌های مستقل کریسپ ارزیابی نمودند. Abdelaal و همکاران (۱۲) توانایی طبقه‌بندی SVM با Tree Boost و Tree Forest را در آزمایش داده‌های بانک اطلاعاتی

ماموگرافی "DDSM" برای استخراج ویژگی‌های توده ماموگرافی بررسی کردند. در این پژوهش، تکنیک SVM نتایج امیدبخشی را برای افزایش دقت تشخیصی طبقه‌بندی داده‌های آموزشی نشان داد. سروستانی و همکاران مقایسه‌ای بین توانایی‌های شبکه‌های عصبی مختلف نظیر شبکه‌های عصبی چند لایه پرسپترون، نقشه خود سازماندهی (SOM)، تابع پایه شعاعی (RBF) و شبکه عصبی مصنوعی احتمالی ارائه داده‌اند. در این پژوهش از مجموعه داده‌های "WBC و NHBCD" برای طبقه‌بندی داده‌ها استفاده شده است. در نهایت PNN با میانگین زمانی ۱/۴۸۷ و کارایی ۱۰۰ درصد بهترین نتیجه را برای طبقه‌بندی در بین الگوریتم‌های شبکه عصبی مصنوعی ارائه داده است. هدف این پژوهش بررسی عملکرد ساختار شبکه‌های عصبی برای تشخیص سرطان پستان بود که نتایج آن نشان داد که شبکه‌های عصبی مصنوعی را می‌توان به‌طور موثری برای تشخیص سرطان پستان مورد استفاده قرار داد (۱۳). مطالعه‌ای که توسط Chao و همکاران (۱۴) انجام شد از ماشین بردار پشتیبان، رگرسیون لجستیک و درخت تصمیم C.5 به منظور طبقه‌بندی نرخ بقای بیماران مبتلا به سرطان پستان و از روش 10-fold cross-validation برای ساخت مدل استفاده شد. نتایج این مطالعه نشان می‌دهد که ایجاد ابزار طبقه‌بندی برای طبقه‌بندی مدل‌ها، به‌طور متوسط به دقت ۹۰ درصد برای هر سه مدل دست یافت. در این مطالعه SVM با دقت ۹۵/۱۵٪ بهترین تکنیک برای ساخت سه طبقه‌بندی در سیستم طبقه‌بندی بقای بیماران بود و رگرسیون لجستیک و درخت تصمیم به ترتیب به دقت ۹۵/۱٪ و ۹۳/۹۵٪ دست یافتند.

در مطالعه‌ی Cakir و Demirel (۱۵)، روش‌های درمان سرطان پستان با استفاده از داده کاوی بررسی گردید که در آن از الگوریتم‌های مختلفی همچون IBL، پرسپترون چند لایه و درخت تصمیم بهره گرفته شده است و سیستم نرم افزاری برای کمک به پزشک انکولوژی برای پیشنهاد استفاده از روش‌های درمان در مورد بیماران مبتلا به سرطان پستان توسعه داده شده است. نتایج این مطالعه، IBL با دقت ۹/۶۳٪ برای هورمون تراپی، پرسپترون چندلایه با دقت ۹۲٪ برای تاموکسیفن، درخت تصمیم با دقت ۹۷/۷۸٪ برای شیمی درمانی و پرسپترون چندلایه با دقت ۹۵/۲۸٪ برای رادیوتراپی را به عنوان بهترین الگوریتم‌ها ارائه داده است.

در مطالعه‌ی Aličković و Subasi (۱۶) روش‌های مختلف داده کاوی برای تشخیص سرطان پستان بررسی گردیده است. در این مطالعه، از مجموعه داده‌های سرطان پستان بانک اطلاعاتی "Wisconsin" استفاده شده است. نتایج به‌دست آمده با مدل Rotation Forest با ۱۴ ویژگی مبتنی بر GA نشان می‌دهد که بالاترین دقت طبقه‌بندی (۹۴/۴۸٪) بوده است. هدف از این پژوهش، توسعه‌ی مدل تشخیص سرطان پستان

اطلاعات مربوط به مراجعات بیمار می‌باشد. از میان این داده‌ها، با توجه به نظر پزشک و بر اساس فرم‌های نظرسنجی فیلهایی که در تشخیص سرطان پستان بسیار موثر و حایز اهمیت هستند انتخاب گردیدند. این فیله‌ها در علم پزشکی با عنوان فاکتور شناخته می‌شوند. هر رکورد شامل ۲۲ ریسک فاکتور قابل پیشگیری و غیرقابل پیشگیری و یک ویژگی کلاس که دارای یکی از دو مقدار سالم و یا مبتلا به سرطان پستان می‌باشد. در این مطالعه تعداد ۶۵۴ پرونده‌ی در دسترس از مراجعه‌کنندگان به کلینیک تخصصی سرطان پستان شهیدمطهری شیراز که دارای دو کلاس افراد سالم و افراد مبتلا به سرطان پستان بودند، جمع‌آوری گردید. تعداد نمونه‌های این مجموعه داده بعد از انجام پیش پردازش و حذف داده‌های پرت به تعداد ۶۲۱ نمونه کاهش یافت که شامل ۴۲۱ نمونه از افراد سالم و ۲۰۰ نمونه از افراد بیمار می‌باشد. جدول ۱ دربردارنده ویژگی‌های مجموعه داده مورد استفاده در این پژوهش به همراه نوع و بازه مقادیر آنهاست.

با تمرکز بر سه الگوریتم طبقه‌بندی درخت تصمیم، بیز ساده و شبکه عصبی مصنوعی و انتخاب مدل با بالاترین دقت تشخیص به منظور بهبود تصمیم‌گیری پزشکان در امر تشخیص زودهنگام این سرطان می‌باشد. در ادامه این مقاله، در ابتدا مجموعه داده‌ی جمع‌آوری شده از کلینیک تخصصی سرطان پستان مطهری شیراز و روش پژوهش تشریح خواهند شد. سپس، مراحل آماده سازی مجموعه داده‌ها و روش‌های ایجاد مدل بررسی می‌گردند. در نهایت نتایج به دست آمده از روش‌های مختلف مقایسه و ارزیابی خواهند شد.

## روش بررسی

### • مجموعه داده

داده‌های جمع‌آوری شده در این پژوهش از میان داده‌های موجود در بانک اطلاعاتی کلینیک تخصصی سرطان پستان شهیدمطهری شیراز به دست آمده است. این داده‌ها شامل دو جدول مشخصات بیمار و

جدول ۱: ویژگی‌های مجموعه داده سرطان پستان

نام فاکتور	شرح	نوع داده	بازه مقادیر داده ها
سن	تاریخ تولد	تاریخ شمسی	۲۴-۹۲
وضعیت تاهل	وضعیت تاهل	رشته‌ای	مجرد-متاهل-مطلقه
وزن	وزن بیمار در زمان ابتلا	عددی صحیح	۳۸-۱۳۲
قد	قد بیمار مبتلا	عددی صحیح	۱۴۰-۱۸۳
سن قاعدگی	شروع قاعدگی بیمار	عددی صحیح	۱۳-۲۲
سن حاملگی	اولین حاملگی بیمار	عددی صحیح	۱۳-۴۶
تعداد زایمان	تعداد زایمان بیمار	عددی صحیح	۰-۱۴
تعداد سقط	تعداد سقط جنین در بیمار	عددی صحیح	۰-۶
شیردهی	تعداد ماه‌های شیردهی	عددی صحیح	۰-۱۰۰۰
وضعیت یائسگی	آیا بیمار یائسه شده یا خیز	رشته‌ای	Negative-positive
سن یائسگی	سن یائسه شدن بیمار	عددی صحیح	۳۰-۶۱
OCP	استفاده بیمار از قرص ضدبارداری	رشته‌ای	Negative-positive
مدت OCP	تعداد ماه‌هایی که از OCP استفاده کرده	عددی صحیح	۰-۳۶۰
HRT	سابقه هورمون درمانی	رشته‌ای	Negative-positive
سابقه بیماری پستان	سابقه بیماری خوش خیم پستان از قبل (کیست، توده چربی و ...)	رشته‌ای	Negative-positive
جراحی پستان	آیا بیمار سابقه جراحی پستان داشته است؟	رشته‌ای	Negative-positive
پرتو درمانی	آیا بیمار سابقه پرتو درمانی دارد؟	رشته‌ای	Negative-positive
سابقه خانوادگی	آیا بیمار سابقه فامیلی سرطان پستان دارد؟	رشته‌ای	Negative-positive
سیگار	آیا بیمار سابقه مصرف سیگار دارد؟	رشته‌ای	Negative-positive
قلیان	آیا بیمار سابقه استعمال قلیان دارد؟	رشته‌ای	Negative-positive
الکل	آیا بیمار سابقه مصرف الکل دارد؟	رشته‌ای	Negative-positive
ورزش	آیا بیمار فعالیت ورزشی منظم دارد؟	رشته‌ای	Negative-positive

شناسایی و بنابر استاندارد، ۵ درصد آنها، حذف گردیدند. نتیجه این فرایند حذف ۳۳ رکورد از مجموع داده‌های جمع‌آوری شده، بود.

#### • گسسته سازی

هدف از این روش، این است که داده‌ها را بر حسب قواعدی در دسته‌بندی‌هایی قرار دهیم و دسته‌ای را که تعداد داده‌های موجود در آن بسیار کم باشد، کنار گذاریم. در موارد بسیار، پزشکان بر اساس دسته‌بندی‌هایی که بر اساس وضعیت‌های مختلف وجود دارد نسبت به تشخیص، اقدام می‌نمایند. به‌عنوان مثال فشار خون همسان در دو فرد که‌نسان و نوجوان نشان از وضعیت‌های متفاوتی دارد؛ بنابراین در این مثال عملیات گسسته سازی بر اساس سن افراد صورت گرفته است. برای این منظور از "Discretize" استفاده شد. در این مرحله ویژگی‌هایی همچون سن، وزن و غیره که دارای مقادیر پیوسته عددی بودند، با استفاده از تکنیک گسسته سازی به‌چندین دسته تقسیم‌بندی شدند (جدول ۲).

#### • انتخاب ویژگی

از آنجایی که تعداد ویژگی‌های موجود در پرونده بیماران زیاد است، باید از بین این ویژگی‌ها، ویژگی‌های تاثیرگذار را به‌منظور ساخت مدل انتخاب کنیم. برای این منظور از ضریب جینی استفاده شد. در این مرحله از "Weight by Gini Index" به‌منظور وزن‌دهی به ویژگی‌ها بهره گرفته شد. ویژگی‌ها بعد از دریافت وزن می‌توانند به کمک "Select by Weights" انتخاب شوند. ۱۰ ویژگی که دارای بیش‌ترین وزن بودند، انتخاب شدند و بعد از تایید توسط پزشکان متخصص در فرایند تولید مدل استفاده شدند (جدول ۲).

در جدول ۱ تمامی ریسک فاکتورهای که در پرونده بیماران ثبت شده‌اند با جزئیات نمایش داده شده است. از این ۲۲ ریسک فاکتور به‌عنوان ویژگی در فرایند مدل سازی بهره گرفته خواهد شد.

#### • پیش پردازش

اهمیت آماده سازی داده‌ها به‌دلیل این واقعیت است که فقدان داده با کیفیت برابر با فقدان کیفیت در نتایج کاوش است و ورودی بد، خروجی بد به‌دنبال دارد (۱۷). وظیفه اصلی پیش پردازش داده‌ها، سازماندهی داده‌ها در شکل‌های استاندارد برای داده‌کاوی و یا سایر عملیات مبتنی بر کامپیوتر است که در ادامه بدان اشاره شده است.

#### • ویژگی‌های فاقد مقدار

در پرونده‌های جمع‌آوری شده برخی از ویژگی‌ها فاقد مقدار بودند که از روش "impute missing value" با استفاده از الگوریتم نزدیک‌ترین همسایه با اندازه  $k=5$  و معیار فاصله اقلیدسی به‌منظور مقداردهی به آن‌ها استفاده شد.

#### • تشخیص مقادیر ناهنجار (پرت)

در ادامه عملیات پیش پردازش، نوبت به داده‌هایی می‌رسد که اصطلاحاً پرت هستند. مقادیر این دسته از داده‌ها نرمال نیستند. به‌عبارتی برخی از نمونه‌ها دارای ویژگی‌هایی با مقادیر غیراستاندارد هستند به‌عنوان مثال اندازه قد فرد رکورد ۱۷۵، "۰" ثبت شده است که این اندازه با اندازه سایر نمونه‌ها فاصله زیادی دارد. برای حذف داده‌های پرت از "Detect Outlier" با پارامتر فاصله اقلیدسی بهره گرفته شد. به این ترتیب، رکوردهایی که دارای فاصله بیشتری نسبت به مابقی نمونه‌ها هستند

جدول ۲: ویژگی‌های برگزیده به همراه رده و ضریب جینی

ویژگی	رده ویژگی	ضریب جینی
سن	کمتر از ۳۰	۰/۰۳
	۳۰-۴۰	۱
	۴۰-۵۰	۲
	بالای ۵۰	۳
وزن	کمتر از ۵۰	۰/۰۷
	۵۰-۶۰	۱
	۶۰-۷۰	۲
	بالاتر از ۷۰	۳
سن شروع قاعدگی	زیر ۱۲ سالگی	۰/۱۳
	۱۲-۱۶	۱
	بالای ۱۶ سالگی	۲
سن یائسگی	زیر ۴۰ سال	۰/۲۶
	۴۰-۵۰	۱
		۲

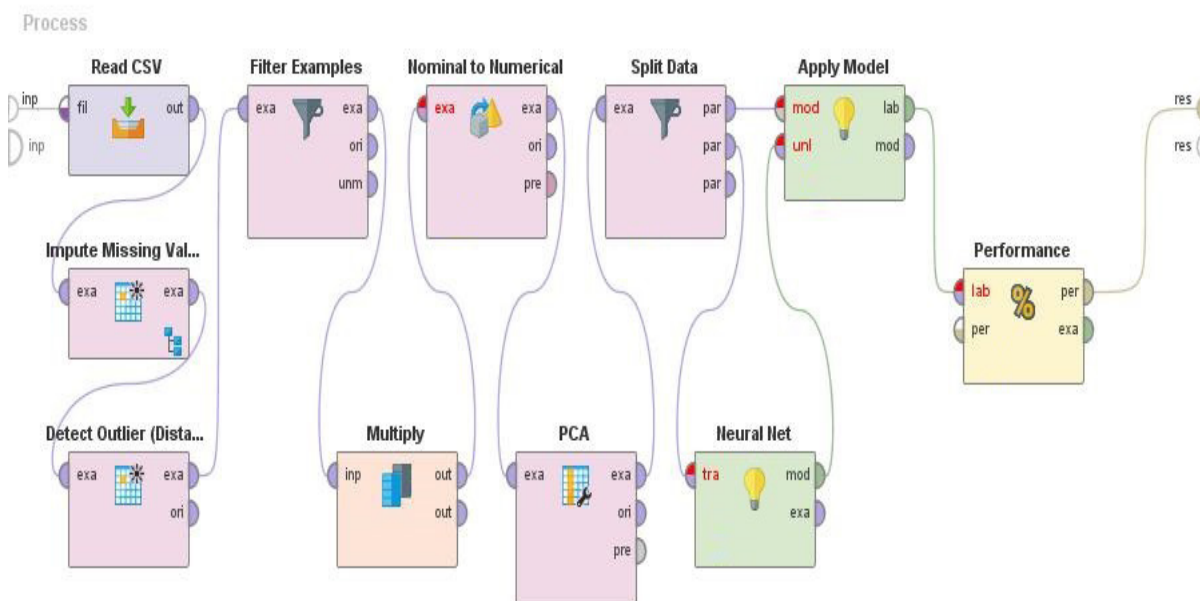
	۳	بالای ۵۰ سال	
مدت زمان مصرف OCP	۱	زیر ۲۰ ماه	
	۲	۲۰-۳۰	
	۳	بالای ۵۰ ماه	
سن اولین بارداری	۱	زیر ۲۰ سال	
	۲	۲۰-۳۰	
	۳	۳۰-۴۰	
سابقه خانوادگی	۴	بالای ۴۰ سال	
	۱	دارد	
	۰	ندارد	
سابقه بیماری پستان	۱	دارد	
	۰	ندارد	
	۱	دارد	
فعالیت ورزشی	۱	دارد	
	۰	ندارد	
	۱	کمتر از ۲۴ ماه	
تعداد ماه‌های شیردهی	۲	۲۴-۴۸	
	۳	بالای ۴۸	

تشخیص سرطان پستان مورد استفاده قرار گرفته اند. سه روش بیز ساده، درخت تصمیم و شبکه عصبی مصنوعی در این مطالعه استفاده شده‌اند. تمامی مراحل پیاده سازی مدل، در نرم افزار RapidMiner صورت گرفته است. در شکل ۱ مدل شبکه عصبی بر روی داده‌های مطالعه نشان داده شده است:

با بهره‌گیری از ضریب جینی ۱۰ ویژگی که تاثیر بیشتری در فرایند مدل سازی خواهند داشت انتخاب شدند. این ۱۰ ویژگی به-ترتیب اهمیت در جدول ۲ نشان داده شده‌اند.

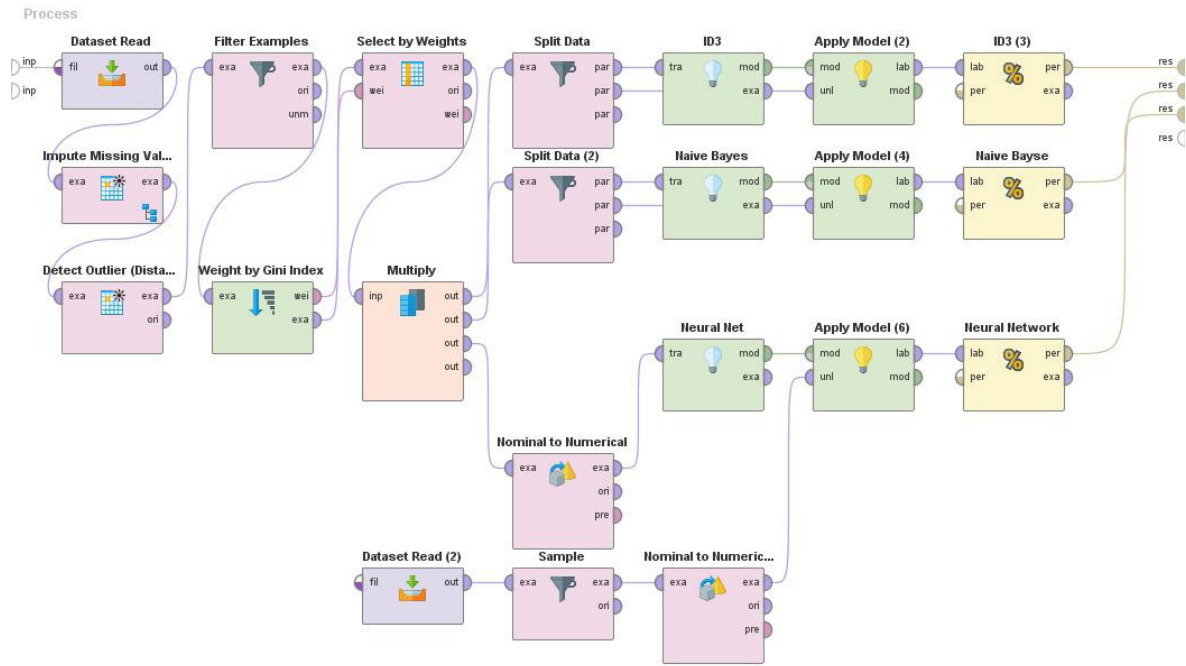
● اعمال روش‌های طبقه‌بندی روی داده‌ها

به‌طورکل روش‌های طبقه‌بندی کننده مختلفی تاکنون به‌منظور



شکل ۱: مدل شبکه عصبی بر روی داده‌ها در نرم‌افزار RapidMiner

شکل شماره ۲، تصویر کامل مدل‌های درخت تصمیم، بیز ساده و شبکه عصبی در نرم‌افزار RapidMiner را نشان داده‌است.



شکل ۲: شکل کامل پیاده سازی سه مدل در نرم افزار RapidMiner

## یافته‌ها

تصمیم، شبکه عصبی مصنوعی و بیز ساده با استفاده از چهار معیار دقت (Accuracy)، صحت (Precision)، حساسیت (Sensitivity) و ویژگی (Specificity) و معیار اف (F-measure) مورد ارزیابی قرار گرفتند. در ارزیابی‌های صورت گرفته، هر سه مدل نتایج امیدبخشی را در تشخیص بیماری سرطان پستان نشان دادند. جدول ۳ نتایج ارزیابی سه مدل مذکور را نشان می‌دهند.

در این مطالعه مجموعه داده مراجعه کنندگان به کلینیک تخصصی سرطان پستان شهیدمطهری شیراز جمع‌آوری و عملیات پیش پردازش بر روی نمونه‌ها انجام گرفت. در ادامه سه الگوریتم طبقه‌بندی با روش "10-fold cross validation" برای ایجاد و ارزیابی مدل بر روی مجموعه داده‌ها انتخاب گردیدند. نتایج به‌دست آمده از سه تکنیک درخت

جدول ۳: ارزیابی مدل‌های تشخیص سرطان پستان

معیار اف	ویژگی	حساسیت	صحت	دقت	
٪۹۶/۰۷	٪۸۵/۰	٪۹۵/۱۴	٪۹۷/۰۲	٪۹۳/۵۰	درخت تصمیم
٪۹۶/۶۵	٪۸۶/۳۶	٪۹۶/۱۹	٪۹۷/۱۲	٪۹۴/۴۹	شبکه عصبی
٪۹۴/۲۸	٪۸۱/۸۲	٪۹۲/۵۲	٪۹۶/۱۲	٪۹۰/۶۷	بیز ساده

مطابق با این ماتریس، مدل ایجاد شده توسط الگوریتم‌های شبکه عصبی مصنوعی، با میزان ٪۹۶/۱۹ بالاترین حساسیت در تشخیص صحیح افراد بیمار و با مقدار ٪۸۶/۳۶ بالاترین ویژگی در تشخیص درست افراد سالم را دارد (جدول ۵).

شبکه عصبی مصنوعی با دقت تشخیص ٪۹۴/۴۹، یک مدل با دقت قابل قبول به‌منظور استفاده در فرایند تشخیص ابتلا به سرطان پستان نسبت به دو روش درخت تصمیم و بیز ساده ارائه داده است. به‌منظور بررسی کارایی الگوریتم‌ها از ماتریس درهم ریختگی استفاده می‌شود.

جدول ۴: ماتریس درهم ریختگی درخت تصمیم

صحت	کلاس واقعی سالم	کلاس واقعی ابتلا به سرطان	
٪۹۷/۰۲	۳	۹۸	کلاس پیش بینی ابتلا به سرطان
٪۷۷/۲۷	۱۷	۵	کلاس پیش بینی سالم
	٪۸۵	٪۹۵/۱۴	حساسیت

جدول ۵: ماتریس درهم ریختگی شبکه عصبی مصنوعی

صحت	کلاس واقعی سالم	کلاس واقعی ابتلا به سرطان	
%۹۷/۱۲	۳	۱۰۱	کلاس پیش بینی ابتلا به سرطان
%۸۲/۶۰	۱۹	۴	کلاس پیش بینی سالم
	%۸۶/۳۶	%۹۶/۱۹	حساسیت

جدول ۶: ماتریس درهم ریختگی بیز ساده

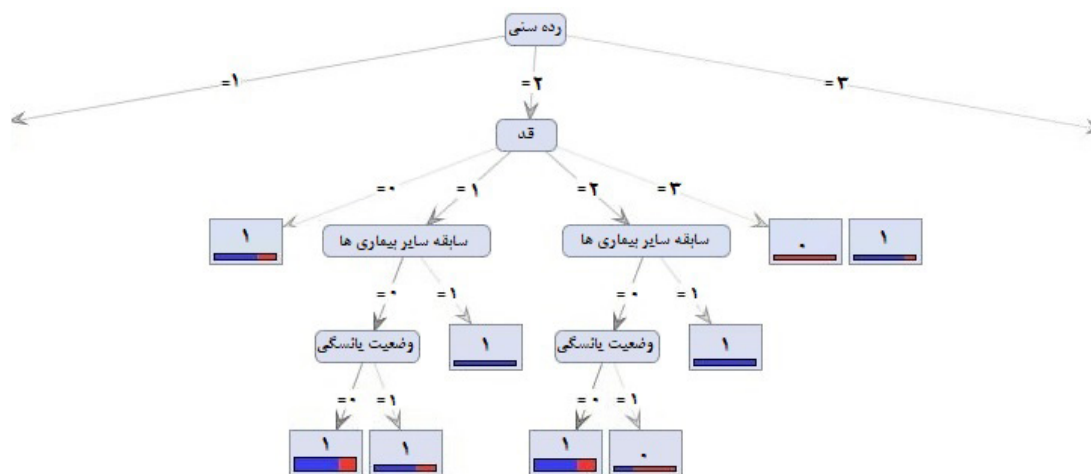
صحت	کلاس واقعی سالم	کلاس واقعی ابتلا به سرطان	
%۹۶/۱۲	۴	۹۹	کلاس پیش بینی ابتلا به سرطان
%۶۹/۲۳	۱۸	۸	کلاس پیش بینی سالم
	%۸۱/۸۲	%۹۲/۵۲	حساسیت

تفسیر قانون اول بیان می‌کند که زنان بین ۳۰ تا ۴۰ سال که وزن آنها بالای ۷۰ کیلوگرم است و هیچ‌گونه فعالیت ورزشی ندارند و اولین بارداری آنها بعد از سن چهل سالگی می‌باشد، بیش از سایر زنان در معرض ابتلا به سرطان پستان قرار دارند. تفسیر قانون دوم نشان می‌دهد که زنان بالای ۵۰ سال که مدت زمان مصرف قرص OCP در آنها بالاتر از ۵۰ ماه می‌باشد و سن شروع قاعدگی و یائسگی آنها به ترتیب کمتر از ۱۲ سالگی و کمتر از ۴۰ سالگی است در معرض ابتلا به سرطان پستان قرار دارند.

ساختار درخت تصمیم مجموعه‌ای از قوانین اگرآنگاه را در اختیار ما قرار می‌دهد (شکل ۳). این قوانین از ترکیبات عطفی و فصلی بین ویژگی‌های درخت تولید می‌شوند. صحت قوانین با نظر پزشک متخصص بررسی و تحلیل گردید. در ادامه به دو قانون برگرفته از درخت تصمیم تولید شده اشاره می‌کنیم:

سرطان=class If (سن فرد=۳) And (وزن=۴) And (فعالیت ورزشی=۰) And (مدت زمان بارداری=۴) Then class=سرطان

سرطان=class If (سن یائسگی=۱) And (سن شروع قاعدگی=۱) And (مدت زمان استفاده از OCP=۳) And (سن=۴) Then class=سرطان



شکل ۳: بخشی از درخت تصمیم تولید شده

که بیش‌ترین استفاده را در کارهای مشابه داشتند با روش ten-fold cross validation برای ایجاد مدل بر روی نمونه داده‌های بومی انتخاب گردیدند. در ارزیابی‌های صورت گرفته هر سه مدل نتایج امیدبخشی را در تشخیص بیماری سرطان پستان نشان دادند. درخت تصمیم با دقت %۹۳/۵۰ (حساسیت %۹۵/۱۴، ویژگی %۸۵/۰) بیز ساده دقت %۹۰/۶۷

## بحث

در این مطالعه با توجه به اهمیت مدل بومی پیش بینی تشخیص بیماری سرطان پستان، مجموعه داده کاملاً بومی استان فارس از کلینیک تخصصی سرطان پستان شهید مطهری شیراز جمع‌آوری و عملیات پیش پردازش بر روی نمونه‌ها انجام گرفت. در ادامه سه الگوریتم طبقه‌بندی

قبول به منظور استفاده در تشخیص ابتلا به سرطان پستان در بیماران ایرانی خصوصاً استان فارسی را ارایه داد.

(حساسیت ۹۲/۵۲٪، ویژگی ۸۱/۸۲٪) و در نهایت شبکه عصبی دقت ۹۴/۴۹٪ (حساسیت ۹۶/۱۹٪، ویژگی ۸۶/۳۶٪). شبکه عصبی با بالاترین دقت پیش‌بینی بر روی مجموعه داده‌های بومی، یک مدل با دقت قابل

جدول ۷: مقایسه‌ی دقت مدل پیشنهادی با کارهای پیشین

کارهای پیشین	درخت تصمیم	بیزساده	شبکه عصبی
لطیف و همکاران (۱۸)	٪۹۲	٪۹۱	-
محمودی و همکاران (۱۹)	٪۹۳/۹۴	٪۹۵/۹	٪۹۵/۱
دهقان و همکاران (۲۰)	-	٪۹۸/۳	٪۹۷/۵
طلوعی و همکاران (۲۱)	٪۹۳/۶	-	٪۹۴/۷
مدل ارائه شده	٪۹۳/۵۰	٪۹۰/۶۷	٪۹۴/۴۹

تولید شده از روش‌های مختلف داده‌کاوی، همچون شبکه عصبی مصنوعی که دقت قابل قبولی در این پژوهش ارایه داده است، در جهت پشتیبانی از فرایند تصمیم‌گیری پزشکان موثر واقع گردند. بر اساس ارزیابی صورت گرفته در این مقاله، سه مدل ایجاد شده خصوصاً درخت تصمیم، بر روی مجموعه داده‌های جمع‌آوری شده می‌توان مهم‌ترین عوامل موثر در تشخیص سرطان پستان را عواملی چون سن، اضافه وزن، سن شروع قاعدگی و یائسگی، مدت زمان مصرف OCP و سن اولین بارداری دانست. نتایج بررسی این پژوهش بر روی جامعه مورد مطالعه، نشان می‌دهد که زنان بالای ۵۰ سال که یائسه هستند و از مشکل چاقی رنج می‌برند نسبت به سایر گروه‌های سنی، بیشتر در معرض خطر ابتلا به سرطان پستان قرار گرفته‌اند. همچنین از دیگر عوامل خطرزا، در این پژوهش حاملگی دیررس می‌باشد که با همراه شدن با عواملی چون مصرف بالای OCP، احتمال ابتلا به این سرطان را افزایش می‌دهد. بنابراین پیشنهاد می‌گردد که در برنامه‌ریزی‌های پیشگیری از ابتلا به این سرطان تغییر سبک زندگی افراد، افزایش فعالیت‌های ورزشی مورد توجه ویژه قرار گیرند تا از شیوع بیش از پیش این سرطان جلوگیری گردد.

بر اساس جدول ۷، نتیجه‌ی مطالعه‌ی ارایه شده با کارهای مشابه ارزیابی گردید. بر اساس این ارزیابی، دقت‌های به‌دست آمده بسیار نزدیک به کارهای پیشین بوده و در بعضی موارد نیز دقت بهتری در برداشته است. با توجه به این‌که به‌منظور ساخت مدل در این مطالعه از داده‌های استان فارس بهره گرفته شده است، لذا به‌منظور تشخیص ابتلا به سرطان پستان در بیماران بومی این استان، استفاده از این مدل نسبت به سایر مدل‌های ساخته شده بر روی داده‌های غیربومی، می‌تواند نتایج بهتری را در برداشته باشد.

با توجه به اینکه این مدل بر روی داده‌های استان فارس ایجاد گردیده است، می‌توان نقش دیتاست‌های بومی را در جهت بهبود فرایند تشخیص بیماری‌های بومی ارزیابی نمود، اما با توجه به در دسترس نبودن مدل‌های دیگر، از این موضوع می‌توان به‌عنوان محدودیت مطالعه صورت گرفته یاد کرد.

## نتیجه‌گیری

با توجه به اهمیت تشخیص زودهنگام سرطان پستان در مراحل اولیه، انتظار می‌رود به‌کارگیری سیستم‌های تشخیصی مبتنی بر مدل‌های

## منابع

- Hübner-Bloder G & Ammenwerth E. Key performance indicators to benchmark hospital information systems—A Delphi study. *Methods of Information in Medicine* 2009; 48(6): 508-18.
- Beckmann KR, Lynch JW, Hiller JE, Farshid G, Houssami N, Duffy SW, et al. A novel case-control design to estimate the extent of over-diagnosis of Breast cancer due to organised population-based mammography screening. *International Journal of Cancer* 2015; 136(6): 1411-21.



3. Chaurasia V & Pal S. A novel approach for Breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering* 2014; 2(1): 1-17.
4. Liou DM & Chang WP. Applying data mining for the analysis of Breast cancer data. *Data Mining in Clinical Medicine* 2015; 1246(1): 175-89.
5. Js S, Shenoy PD, KR V & Patnaik LM. Cancer prognosis prediction model using data mining techniques. *Data Mining and Knowledge Engineering* 2014; 6(1): 21-9.
6. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. Classification of Breast cancer histology images using convolutional neural networks. *PloS One* 2017; 12(6): e0177544.
7. Chaurasia V, Pal S & Tiwari BB. Prediction of benign and malignant Breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology* 2018; 12(2): 119-26.
8. Lu J, Hales A, Rew DA, Keech M, Fröhlingsdorf C, Mills-Mullett A, et al. Data mining techniques in health informatics: A case study from Breast cancer research, Switzerland: *International Conference on Information Technology in Bio-and Medical Informatics*, 2015.
9. Ruiz A, Sebah M, Wicherts DA, Castro-Benitez C, van Hillegersberg R, Paule B, et al. Long-term survival and cure model following liver resection for Breast cancer metastases. *Breast Cancer Research And Treatment* 2018; 170(1): 89-100.
10. Ojha U & Goel S. A study on prediction of Breast cancer recurrence using data mining techniques, Noida: 7<sup>th</sup> International Conference on Cloud Computing, Data Science & Engineering-Confluence, 2017.
11. Umer Khan M, Choi JP, Shin H & Kim M. Predicting Breast cancer survivability using fuzzy decision trees for personalized healthcare, Vancouver, British Columbia, Canada: 30<sup>th</sup> Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008.
12. Abdelaal MMA, Sena HA, Farouq MW & Salem ABM. Using data mining for assessing diagnosis of Breast cancer. Available at: <https://annals-csis.org/proceedings/2010/pliks/158.pdf>. 2010.
13. Sarvestani AS, Safavi AA, Parandeh NM & Salehi M. Predicting Breast cancer survivability using data mining techniques, USA: 2<sup>nd</sup> International Conference on Software Technology and Engineering, 2010.
14. Chao CM, Yu YW, Cheng BW & Kuo YL. Construction the model on the Breast cancer survival analysis use support vector machine, logistic regression and decision tree. *Journal of Medical Systems* 2014; 38(10): 106.
15. Cakır A & Demirel B. A software tool for determination of Breast cancer treatment methods using data mining approach. *Journal of Medical Systems* 2011; 35(6): 1503-11.
16. Aličković E & Subasi A. Breast cancer diagnosis using GA feature selection and rotation forest. *Neural Computing and Applications* 2017; 28(4): 753-63.
17. Blake R & Mangiameli P. The effects and interactions of data quality and problem complexity on classification. *Journal of Data and Information Quality* 2011; 2(2): 8.
18. Latif AM, Momeny M, Sarram R, Agha Sarram M, Pour Ahmadi A & Haj Ebrahimi Z. Using data mining and genetic algorithm for diagnosis of Breast cancer. *Iranian Quarterly Journal of Breast Disease* 2016; 9(1): 45-56[Article in Persian].
19. Mahmoodi MS, Mahmoodi SA, Haghghi F & Mahmoodi SM. Determining the stage of Breast cancer by data mining algorithms. *Iranian Quarterly Journal of Breast Diseases* 2014; 7(2): 36-44[Article in Persian].
20. Dehghan P, Mogharabi M, Zabbah I, Layeghi K & Maroosi A. Modeling Breast cancer using data mining methods. *Journal of Health and Biomedical Informatics* 2018; 4(4): 266-78[Article in Persian].
21. Toloiee-Ashlaghi A, Pourebrahimi A, Ebrahimi M & Ghasem-ahmad L. Using data mining techniques for prediction Breast cancer recurrence. *Iranian Journal of Breast Diseases* 2013; 5(4): 23-34[Article in Persian].

# Diagnosis of Breast Cancer Using Decision Tree, Artificial Neural Network and Naive Bayes to Provide a Native Model for Fars Province

Azita Yazdani<sup>1</sup> (Ph.D) - Ali Asghar Safaei<sup>2</sup> (Ph.D.) - Reza Safdari<sup>3</sup> (Ph.D.) -  
Maryam Zahmatkeshan<sup>4</sup> (Ph.D.)

1 Assistant Professor, Department of Health Information Technology, School of Management and Medical Informatics, Shiraz University of Medical Sciences, Shiraz, Iran

2 Assistant Professor, Department of Medical Informatics, School of Medical Sciences, Tarbiat Modarres University, Tehran, Iran

3 Professor, Department of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran

4 Ph.D. in Health Information Management, Noncommunicable Diseases Research Center, Fasa University of Medical Sciences, Fasa, Iran

## Abstract

Received: Dec 2018

Accepted: Apr 2019

**Background and Aim:** Breast cancer is the most common type of cancer and the main cause of death from cancer in women worldwide. Technologies such as data mining, have enabled experts in this area to improve decision making in the early diagnosis of the disease. Therefore, the purpose of this research is to develop an automatic diagnostic model for Breast cancer by employing data mining methods and selecting the model with the highest accuracy of diagnosis for Fars province.

**Materials and Methods:** In this study, 654 available patient records of Motahari Breast cancer Clinic in Shiraz" were used as the sample. The number of records was reduced to 621 after the pre-processing operation. These samples had 22 features that ultimately ten were used as effective features in the design of the model. Three types of Decision tree, Naive Bayes and Artificial neural network were used for diagnosis of Breast cancer and 10-fold cross-validation method for constructing and evaluating the model on the collected data set.

**Results:** The results of the three techniques mentioned all three models showed promising results in detecting Breast cancer. Finally, the artificial neural network accounted for the highest accuracy of 94/49%(sensitivity 96/19%, specificity 86/36%) in the diagnosis of Breast cancer.

**Conclusion:** Based on the results of the decision tree, the risk factors such as age, weight, age of menstruation, menopause, duration of OCP usage, and the age of the first pregnancy were among the factors affecting the incidence of Breast cancer in women of Fars province.

**Keywords:** Breast Cancer, Diagnostic Model, Decision Tree, Naive Bayes, Artificial Neural Network, Risk Factors

\* Corresponding Author:

Safaei AA

Email:

aa.safaei@modares.ac.ir