مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره دوم شماره دوم. ص۷۵-۵۹، ۱۳۹٤

توسعه یک مدل رده‌بندی حوزه محور جهت کشف زود هنگام افراد در خطر ابتلا به سرطان روده بزرگ

ایمان برازنده[1]*، محمدرضا غلامیان[2]، عبدالحسن طلائی زاده[3]، محمدامین پورحسین قلی[4]

• **دریافت مقاله:** ۹۴/۵/۲۵ • **پذیرش مقاله:** ۹۴/۶/۱۸

مقدمه: در این تحقیق نشان داده می‌شود که می‌توان با استفاده از تکنیک‌های داده‌کاوی مدل‌هایی برای تشخیص سبک زندگی افراد از لحاظ پرخطر یا کم‌خطر بودن برای ابتلا به سرطان روده بزرگ توسعه داد.

روش: در این بررسی گذشته‌نگر، مجموعه داده‌ای شامل ۸۴ فرد بیمار و ۲۲۵ فرد سالم، شامل ۲۵ خصیصه جمع آوری شد. این اطلاعات شامل بیمارانی است که تشخیص آن‌ها مربوط به سال‌های ۱۳۸۵ تا سه ماهه اول ۱۳۹۳ می‌باشد. از پرکاربردترین تکنیک‌ها در ادبیات انفورماتیک پزشکی شامل ماشین بردار پشتیبان، بیزین ساده، درخت تصمیم و نزدیکترین همسایگی برای توسعه مدل‌ها استفاده شد. سنجه جدید غیرتکنیکی توسعه داده شد که کارایی مدل‌ها برای حوزه پزشکی را مشخص می‌کند. از دیدگاه داده کاوی حوزه محور برای تعیین مدل قابل اجرا استفاده شد.

نتایج: مدل‌های توسعه داده شده با کارایی قابل قبولی قادر به تشخیص سبک زندگی افراد هستند. سنجه غیرتکنیکی توسعه داده شده به خوبی می‌تواند ارزش واقعی تک تک پیش‌بینی‌ها، چه درست و چه نادرست را با هزینه‌های واقعی مشخص کند و یک میزان واقعی از هزینه‌های صرفه‌جویی شده در نظام سلامت توسط هر مدل را نشان دهد. از میان مدل‌های توسعه داده شده تنها دو مدل توانست معیارهای تعیین شده جهت استفاده در دنیای واقعی را ارضا کند.

نتیجه‌گیری: مدل‌های توسعه داده شده نه تنها باید از لحاظ تکنیکی ارزیابی شوند، بلکه باید از لحاظ سنجه‌های مورد پذیرش برای حوزه پزشکی و همچنین قابلیت اجرا برای حل واقعی مسأله نیز بررسی گردند.

کلید واژه‌ها: داده‌کاوی حوزه محور، رده‌بندی، سرطان‌های روده بزرگ، کشف زودهنگام سرطان

•**ارجاع:** برازنده ایمان، غلامیان محمدرضا، طلائی زاده عبدالحسن، پورحسین قلی محمدامین. توسعه یک مدل رده بندی حوزه محور جهت کشف زود هنگام افراد در خطر ابتلا به سرطان روده بزرگ. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۴؛ ۲(۲): ۷۵ -۵۹.

۱. کارشناس ارشد مهندسی فناوری اطلاعات، مربی، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده برق و کامپیوتر، دانشگاه آزاد اسلامی واحد ماهشهر، ماهشهر، ایران
۲. دکترای مهندسی صنایع، استادیار، گروه مهندسی صنایع، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران
۳. فلوشیپ جراحی سرطان، دانشیار، مرکز تحقیقات سرطان و آلاینده‌های محیطی و نفتی، گروه جراحی سرطان، دانشکده پزشکی، دانشگاه علوم پزشکی جندی شاپور، اهواز، ایران
۴. دکترای آمار زیستی، استادیار، مرکز تحقیقات بیماری های گوارش و کبد، پژوهشکده بیماری‌های گوارش و کبد، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

* **نویسنده مسؤول:** خوزستان، بندرماهشهر، خیابان دانشگاه، دانشگاه آزاد اسلامی واحد ماهشهر

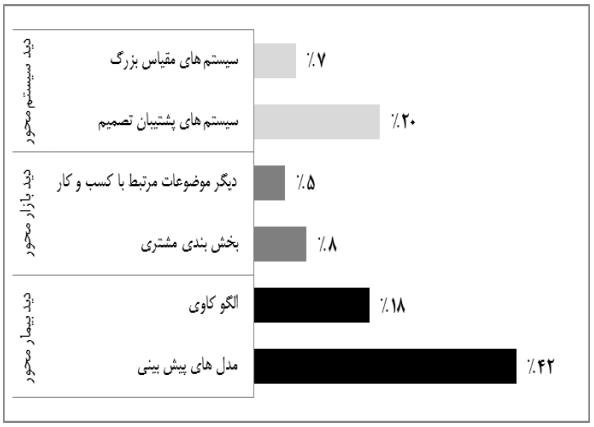
• **شماره تماس:** ۰۶۱۵۲۳۲۷۰۷۰ • **Email:** barazandeh_i@ind.iust.ac.ir

مقدمه

هنگامی که سرطان به موقع تشخیص داده شود، با احتمال بسیار بیشتری درمان‌پذیراست و به درمان‌های مؤثر پاسخ می‌دهد. یکی از این سرطان‌های درمان‌پذیر، سرطان روده بزرگ است. پیش‌بینی، یک مسأله همیشگی در سیر بیماری سرطان از غربالگری تا درمان‌های تسکینی است. در واقع، انکولوژی (Oncology) به طور ذاتی یک مسأله پیش‌بینی است. بسیاری از سرطان‌ها در مراحل اولیه هیچ علامتی ندارند و درمان تنها بر اساس این پیش‌بینی که پیشرفت تومور سرانجام کیفیت زندگی یا بقای بیمار را تحت تأثیر قرار می‌دهد، توصیه می‌شود. البته چه بهتر که از مدل‌های دقیق به جای مدل‌های ذهنی و تخمینی استفاده کرد [۱]. کشف دانش و داده‌کاوی مجموعه‌ای از تکنیک‌های مبتنی بر یادگیری ماشین است که می‌تواند به توسعه مدل‌های پیش‌بینی دقیق کمک کند. Tan و همکاران [۲] داده‌کاوی را استخراج خودکار دانش جدید و مفید از منابع داده‌ای حجیم موجود داده‌کاوی تعریف کرده‌اند. از نظر این تحقیق دانش تولیدی از داده‌های حوزه سلامت وقتی مفید است که به تصمیم‌گیری درست در این حوزه کمک کند.

Wells و همکاران [۳] معتقدند استراتژی‌های پیش‌بینی، خطر سرطان روده بزرگ با این که در کاهش مرگ از این بیماری مؤثرند اما مستلزم صرف هزینه بسیار بالا، خطر بروز عوارض نادر اما جدی مثل پارگی روده و سازگاری ضعیف می‌باشند. به همین دلیل تخمین دقیق خطر می‌تواند بازدهی غربالگری را با هدف قرار دادن بیماران پرخطر برای غربالگری زودهنگام و یا متناوب‌تر و همچنین به تأخیر انداختن غربالگری در افراد با خطر پایین، افزایش دهد. با توجه به رخداد بالای سرطان روده بزرگ، هزینه تحمیلی قابل توجه آن بر جامعه، میزان در دسترس بودن آزمایش‌های غربالگری، وجود مدل‌هایی که احتمال ابتلای افراد به سرطان روده بزرگ را بر اساس اطلاعات فاکتورهای خطر تخمین بزند می‌تواند به پزشکان و بیماران در تعیین رژیم‌های غربالگری مؤثر باشد [۴]. بنابراین هدف این تحقیق تشخیص سبک زندگی افراد با استفاده از مدل‌های پیش‌بینی در داده کاوی به منظور تسریع غربالگری افراد پرخطر و تعویق غربالگری افراد کم خطر برای کاهش بار مالی آن از افراد و نظام سلامت و همچنین کاهش آسیب‌های ناخواسته می‌باشد.

برازنده و غلامیان [۵] نشان داده‌اند که کاربردهای کشف دانش و داده‌کاوی در صنعت سلامت را می‌توان به سه دیدگاه اصلی یعنی دید بیمار محور، دید بازار محور و دید سیستم محور تقسیم کرد. شکل۱ این دیدگاه‌ها را نشان می‌دهد. دید بیمار محور شامل تحقیقاتی است که به داده‌کاوی محض روی داده‌های صنعت سلامت پرداخته‌اند. دید بازار محور شامل تحقیقاتی است که به بیماران به دید مشتری نگاه کرده‌اند. دید سیستم محور که شامل تحقیقاتی است که به توسعه سیستم-های پشتیبان تصمیم پرداخته‌اند.



شکل۱: داده کاوی در صنعت سلامت

برازنده و غلامیان [۵] همچنین نشان می‌دهند که علاقه محققین این حوزه به توسعه مدل‌های پیش‌بینی بیش از هر حوزه دیگری است. در تحقیق اسفندیاری و همکاران [۶] نیز نتایج مؤید این مسأله است. ساخت مدل‌های پیش‌بینی بسته به نوع داده‌ها و تکنیک مورد استفاده می‌تواند ساده یا بسیار پیچیده باشد. انواع مختلف درخت‌های تصمیم، یکی از ساده-ترین تکنیک‌ها، هم از نظر پیاده‌سازی و هم از نظر درک برای کاربران نهایی می‌باشد. Moon و همکاران [۷] از درخت تصمیم برای مدل‌سازی و یافتن رابطه بین میانگین تعداد سیگار در روز استفاده کرده‌اند. Parhizai و همکاران [۸] از درخت تصمیم برای شناسایی روابط بین فاکتورهای روانشناسانه و ابعاد خستگی در پرستاران استفاده کرده‌اند. Kumar و همکاران [۹] معتقدند که اگرچه سادگی فهم و توضیح درخت تصمیم برای خبرگان حوزه راحت‌تر است، اما استفاده از درخت تصمیم موجب فداشدن دقت به خاطر سادگی می‌شود. شبکه-های عصبی مصنوعی جعبه‌های سیاه مبتنی بر هوش مصنوعی هستند که با وجود بیش تناسب از دیگر الگوریتم‌های پرطرفدارتر هستند. با این حال دقت و کارایی یک تکنیک به میزان زیادی به حوزه مورد کاوش، انتخاب خصیصه‌ها، نوع داده‌ها و نرم افزار مورد استفاده برای داده‌کاوی بستگی دارد [۱۰]. Marcano و همکاران [۱۱] از سه تکنیک درخت

تصمیم، شبکه عصبی چند لایه پرسپترون و شبکه عصبی رگرسیون عمومی، برای پیش‌بینی نتایج توان بخشی شناختی در بیماران با آسیب‌های مغزی استفاده کرده‌اند که از لحاظ دقت به دست آمده نیز به ترتیب ذکر شده بوده‌اند. از دیگر تکنیک‌های مورد استفاده ماشین بردار پشتیبان است که اگرچه تفسیر رده‌بندهایی که از این دیدگاه استفاده می‌کنند مشکل است اما کارایی رده‌بندی می‌تواند بسیار بالا باشد [۱۰]. Chi و همکاران [۱۲] یک الگوریتم سیستم خبره مبتنی ماشین بردار پشتیبان پیشنهاد می‌کند که سرعت، هزینه و دقت در تشخیص بیماری را بهینه کرده و تصمیمات تشخیصی را با یک پیکربندی متوالی بیان می‌کند. این الگوریتم به صورت پویا آزمایش‌هایی را تعیین می‌کند که با حداقل هزینه و زمان، احتمالاً دقت تشخیص را افزایش می‌دهد و با تشخیص بیماری متوقف می‌شود. رده‌بندهای احتمالی مثل بیزین ساده، شبکه-های بیزین و نزدیکترین همسایگی، اگرچه کمتر اما در ادبیات انفورماتیک پزشکی دیده می‌شوند [۱۰]. در تحقیق Huang و همکاران [۱۳] تشخیص بیماران با وضعیت بد کنترل گلوکوز خون از بیماران با وضعیت خوب کنترل گلوکوز، بر اساس داده-های روانشناسانه و تجربی بررسی شده است. سه روش رده بندی با ویژگی‌های مختلف در این تحقیق شامل یک رده‌بند احتمالی مثل بیزین ساده، درخت تصمیم و نزدیکترین همسایگی استفاد شده است. دقت رده بندی برای تکنیک‌های مختلف و برای تعداد مختلف خصیصه‌ها مقایسه شده است.

با وجود مهلک بودن بیماری سرطان روده بزرگ، تحقیقات اندکی روی این بیماری در حوزه کشف دانش و داده‌کاوی انجام شده است. از میان این تحقیقات اندک، غربالگری این بیماری جهت کشف زودهنگام با پایین‌ترین اقبال و اهتمام محققین روبه رو بوده است. جدول ۱ جزئیات برخی از این تحقیقات را نشان می‌دهد. Ong و همکاران [۱۴] با توجه به این که بیش از ۵۰ درصد بیماران سرطان روده بزرگ با عود مجدد بیماری بعد از جراحی مواجه هستند، یک سیستم پشتیبان تصمیم برای پیش‌بینی زودهنگام عود مجدد سرطان با استفاده از استدلال مبتنی بر مورد توسعه داده‌اند. در تحقیق Anand و همکاران [۱۵] مدلی برای پیش‌بینی مدت بقای بیماران مبتلا به سرطان روده بزرگ پس از درمان توسعه داده شده است. در این تحقیق نشان داده شده که می‌توان از تکنیک‌های مبتنی بر هوش مصنوعی بدون نگرانی از کاهش دقت به جای تکنیک‌های آماری در تحلیل بقاء استفاده کرد. در تحقیق Grumett و همکاران [۱۶] دو مدل برای پیش‌بینی عود مجدد بیماری سرطان روده بزرگ پس از جراحی توسعه داده شده است. نتایج این تحقیق نشان می‌دهد که شبکه عصبی بهتر عمل کرده است. Shi و همکاران [۱۷] با در نظر گرفتن یک مجموعه از نقاط مرزی برای هر تومورمارکر که نه به صورت تصادفی، بلکه از طریق یک مبنای تئوریک معرفی شده در تحقیق محاسبه شده‌اند، در قالب کرومزوم‌های طراحی شده برای الگوریتم ژنتیک و در نهایت تحلیل داده‌ها با استفاده از استدلال مورد محور، یک الگوریتم پشتیبان تصمیم هوشمند برای تشخیص سرطان روده بزرگ توسعه داده است.

Roman و همکاران [۱۸] به بررسی امکان غربالگری سرطان روده بزرگ با استفاده میزان پروتئین‌های سلنیوم در خون پرداخته‌اند. در این تحقیق نشان داده شده است که مرحله اولیه سرطان روده بزرگ می‌تواند به تغلیظ پروتئین‌های سلنیوم در خون، به خصوص سلنوآلبومین وابسته باشد. Al-Bahrani و همکاران [۱۹] از چندین نوع درخت تصمیم و همچنین تکنیک رگرسیون لجستیک برای توسعه رده‌بندهای ساده و تجمیعی مختلف برای پیش‌بینی بقای بیماران سرطان روده بزرگ استفاده کرده و در نهایت رده‌بندها با هم مقایسه شده‌اند. رده‌بند تجمیعی مبتنی بر رأی اکثریت، بهترین نتیجه را داشته است. Stojadinovic و همکاران [۲۰] نیز با استفاده از شبکه باور بیزین سیستم پشتیبان تصمیم برای پیش‌بینی بقاء توسعه داده-اند.

جدول ۱: تحقیقات در حوزه داده کاوی بر بیماری سرطان روده بزرگ

تحقیق	خصیصه‌ها	تکنیک‌ها	تعداد بیماران	مرحله
Ong و همکاران [۱۴]	دموگرافیک، گرید جراح، CEA، داده‌های رادیولوژی، داده‌های شیمی درمانی، داده‌های پاتولوژی،	k- نزدیکترین همسایه، CART و درخت تصمیم C4.5	۶۷۴	پیش‌آگهی
Anand و همکاران [۱۵]	دموگرافیک، داده‌های پاتولوژی	شبکه عصبی مصنوعی، k- نزدیکترین همسایه و رگرسیون	۲۱۶	پیش‌آگهی
Grummet و همکاران [۱۶]	دموگرافیک، داده‌های پاتولوژی	شبکه عصبی مصنوعی و رگرسیون	۴۰۳	پیش‌آگهی
Shi و همکارن [۱۷]	پنج تومورمارکر	استدلال مورد محور به همراه الگوریتم ژنتیک	۵۷۸	تشخیص
Roman و همکاران [۱۸]	پروتئین‌های سلنیوم	رگرسیون لجستیک، درخت تصمیم و شبکه عصبی مصنوعی	۶۲	غربالگری
Al-Bahrani و همکاران [۱۹]	اکثرا داده‌های پاتولوژیک	رده بندهای ساده و تجمیعی مبتنی بر انواع درخت تصمیم و همچنین رگرسیون لجستیک	۱۰۵۱۳۳	پیش‌آگهی
Stojadinovic و همکاران [۲۰]	اکثرا داده‌های پاتولوژیک	شبکه باور بیزین	۱۴۶۲۴۸	پیش‌آگهی

از طرف دیگر در این تحقیقات به ندرت سعی شده سنجه‌هایی در نظر گرفته شود که ارزش مدل ارائه شده را برای خبرگان حوزه سلامت مشخص کند [۵]. از موارد اصلی مورد تأکید داده‌کاوی حوزه محور طراحی سنجه‌های ارزیابی است که برای کسب و کار و کاربران نهایی جذاب و قابل فهم باشد. برازنده و غلامیان [۵] بیان می‌کنند که از میان تحقیقات این حوزه تعداد بسیار اندکی به این مهم پرداخته‌اند و بسیاری از تحقیقات به ارائه مدل و ارزیابی تکنیکی آن بسنده کرده و در نتیجه دیدگاهی غیرکاربردی دارند. بیش از ۹۰ درصد تحقیقات این حوزه از سنجه دقت برای ارزیابی مدل نهایی استفاده می‌کنند [۶]. با توجه به این که در مجموعه‌های داده این حوزه معمولاً توزیع کلاس‌ها متوازن نیست، در چنین موقعیتی سنجه دقت نامناسب‌ترین سنجه ارزیابی خواهد بود. در چنین مواردی در نظر گرفتن هزینه کلاس‌ها، برای کاهش خطاهای نوع اول و دوم می‌تواند راه حل مناسبی باشد. اما کاهش این نوع خطاها پایین‌ترین توجه را در بین محققین داشته است [۶].

هدف این تحقیق کمک به بهبود غربالگری سرطان روده بزرگ با تلاش برای کنترل سبک زندگی فرد برای کاهش احتمال ابتلای وی به این بیماری و کشف بیماری در مراحل اولیه است. این هدف با توسعه مدلی که بتواند با دقت قابل قبولی الگوی سبک زندگی افراد را از نظر خطر ابتلا به سرطان روده بزرگ پیش‌بینی کند، تحقق یافته است.

روش

مجموعه داده: مجموعه داده جمع‌آوری شده برای تحقیق یکی از محصولات اصلی تحقیق به شمار می‌رود. خصیصه‌های این داده‌ها از طریق مطالعه ادبیات تخصصی سرطان روده بزرگ [۲۳-۲۱،۳،۴] و همچنین مشاوره خبرگان حوزه تعیین شده است. این مجموعه داده شامل اطلاعات ۳۰۹ فرد است. اطلاعات این مجموعه داده در طول هشت ماه از طریق تماس تلفنی، مصاحبه حضوری و بررسی پرونده‌های بیماران جمع آوری شده است. حدود ۸۴ نفر از این بیماران مبتلا به سرطان روده بودند که بین سال‌های ۱۳۸۵ تا سه ماهه اول سال ۱۳۹۳ تشخیص داده شده‌اند و از مراجعه کنندگان یکی از کلینیک‌های پزشکی اهواز بوده‌اند. اطلاعات حدود ۲۲۵ فرد سالم نیز که در ۱۰ سال اخیر هیچ نشانه‌ای از این بیماری را نداشته‌اند در این مجموعه قرار داده شده است. این اطلاعات شامل ۲۵ خصیصه است. خصیصه‌ها در سه گروه دسته‌بندی شده‌اند: خصیصه‌های دموگرافیک، سبک زندگی و درمانی. جدول ۲، ۳، ۴ این خصیصه‌ها را به همراه ویژگی آن‌ها نشان می‌دهند. ویژگی کلاس در این مجموعه سبک زندگی نام دارد که دارای دو مقدار پرخطر و کم خطر است. در این مجموعه داده، دوره زمانی مورد نظر به طور متوسط در ۱۰ سال گذشته بوده است. این دوره زمانی برای بیماران پیش از تشخیص بوده است. به عنوان مثال در مورد مصرف هفتگی ماهی، میزان مصرف ماهی در ۱۰ سال گذشته جمع‌آوری شده است.

جدول۲: خصیصه‌های سبک زندگی

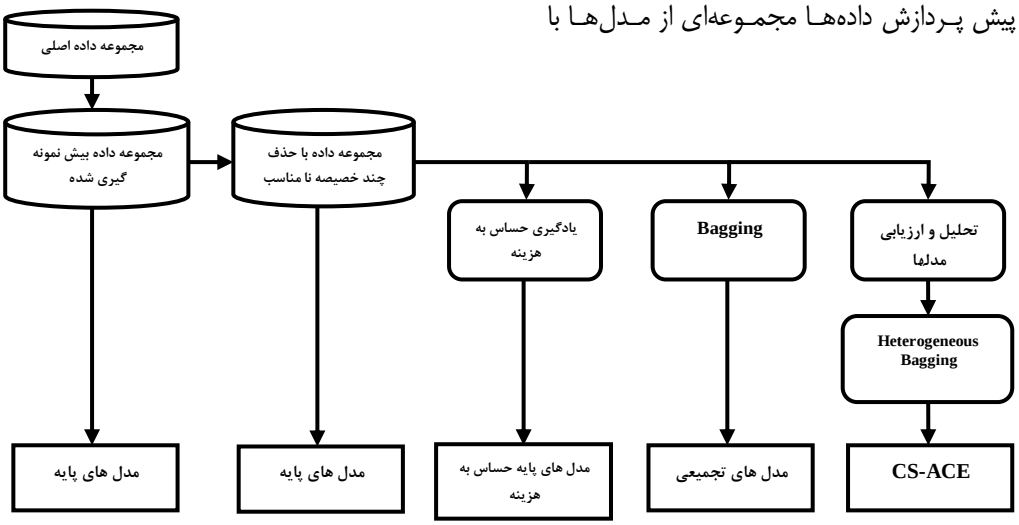
نام متغیر	مقدار متغیر	شرح متغیر
استرس	۰ تا ۳	خیلی زیاد / زیاد / کم / خیلی کم
نوبت کاری شبانه	Yes/No	نوبت کاری شبانه بیش از ۵ سال در طول زندگی
سیگار	۰ تا ۵	خیر / کمتر از ۲۰ سال / ۲۰ تا ۳۰ سال / ۳۱ تا ۴۰ سال / ۴۱ تا ۵۰ سال
الکل	۰ تا ۵	خیر/ کمتر از ۲ لیوان / ۲ تا ۴ لیوان / ۵ تا ۶ لیوان / بیش از ۶ لیوان
فعالیت فیزیکی	۰ تا ۴	هیچ/ یک ساعت یا کتر / ۱ تا ۲ ساعت / ۲ تا ۳ ساعت / ۳ تا ۴ ساعت
مصرف گوشت قرمز	۰ تا ۵	هیچ/ ۱ تا ۲ وعده/ ۳ تا ۴ وعده / ۵ تا ۶ وعده / ۷ تا ۱۰ وعده / بیش از ۱۰ وعده
مصرف ماهی	۰ تا ۵	هیچ/ ۱ تا ۲ وعده/ ۳ تا ۴ وعده / ۵ تا ۶ وعده / ۷ تا ۱۰ وعده / بیش از ۱۰ وعده
مصرف سبزیجات	۰ تا ۵	هیچ/ ۱ تا ۲ وعده/ ۳ تا ۴ وعده / ۵ تا ۶ وعده / ۷ تا ۱۰ وعده / بیش از ۱۰ وعده
مصرف گوشت فرآری شده	۰ تا ۵	هیچ/ ۱ تا ۲ وعده/ ۳ تا ۴ وعده / ۵ تا ۶ وعده / ۷ تا ۱۰ وعده / بیش از ۱۰ وعده

جدول ۳: خصیصه‌های درمانی

نام متغیر	مقدار متغیر	شرح متغیر
سابقه خانوادگی سرطان روده	Yes/No	شامل پدر/مادر/خواهر/برادر/فرزند
سابقه مصرف قرص‌های شامل آسپرین	Yes/No/ Not Now	شامل بافرین، ب-یر، اکزدرین
سابقه مصرف قرص‌های ضد التهابی	Yes/No/ Not Now	شامل ایبوبروفن، دیکلوفناک، ناپروکسین
سابقه مصرف مولتی ویتامین	Yes/No	
سابقه مصرف ویتامین D	Yes/No	
سابقه دیابت	Yes/No	
سابقه تشخیص پولیپ روده بزرگ	Yes/No/Don't Know	تشخیص از طریق کولونوسکوپی و سیگموئیدوسکوپی
مصرف هورمون‌های زنانه مثل استروژن	Yes/No	شامل استروژن، پروجستین و غیره
عادت ماهیانه	Yes/No	

در این تحقیق برای حل مشکل نامتوازن بودن داده‌ها از روش نمونه‌گیری و به طور خاص بیش نمونه‌گیری از کلاس اقلیت به صورت ترکیبی SMOTE (Synthetic Minority Oversampling Technique) استفاده شده است. این روش توسط Chawla و همکاران [۲۴] ارائه شده است.

مدل‌سازی: شکل ۲ ساختار مدل‌سازی جامع صورت گرفته در تحقیق را نشان می‌دهد. همان طور که در شکل نشان داده شده پس از پیش پردازش داده‌ها مجموعه‌ای از مدل‌ها با ویژگی‌های مختلف شامل مدل‌های پایه غیر حساس به هزینه، مدل‌های پایه حساس به هزینه و مدل‌های تجمیعی همگن و غیر همگن توسعه داده شده‌اند. پس از ارزیابی، از میان مدل-های پایه، برترین‌ها در جنبه‌های مختلف انتخاب شده تا در کنار هم مدل تجمیعی دقیق و حساس به هزینه CS-ACE (Cost-Sensitive Accurate Ensemble) را تشکیل دهند.



شکل۲: ساختار مدل‌سازی

تکنیک‌های مورد استفاده در این تحقیق از برترین تکنیک‌های رده‌بندی مورد استفاده در ادبیات انفورماتیک پزشکی می‌باشند [۵،۶،۲۵]. این تحقیق از گروه رده‌بندهای مشتاق، از رده‌بندهای ماشین بردار پشتیبان (Support Vector Machine) SVM با هسته تابع با پایه رادیال (Radial Basis Function)، بیزین ساده (Naïve Bayes) و درخت تصمیم استفاده کرده است. از گروه رده‌بندهای تأخیری، رده‌بند نزدیک-ترین همسایگی (Nearest Neighbor) مورد استفاده قرار گرفته است. در این تحقیق برای اجرای الگوریتم‌های داده‌کاوی از نرم افزار Weka استفاده شد. برای محاسبات آماری، رسم نمودارها و جداول از نرم افزار Excel استفاده شد.

هسته RBF دو پارامتر دارد : C و γ. مقادیر مناسب برای زوج (C,γ) موجب افزایش دقت رده‌بند می‌شود. برای یافتن بهترین (C,γ)، اجرای جستجوی روی مشبکه (Grid Search) با استفاده از تقابل سنجی توصیه شده است [۲۶،۲۷]. در این تحقیق این جستجو به کار گرفته شد.

سنجه‌های ارزیابی تکنیکی: در این بخش سنجه‌های ارزیابی عملکرد الگوریتم‌های رده‌بندی استفاده شده در تحقیق معرفی می‌شوند. این سنجه‌ها مستخرج از مفهوم ماتریس درهم‌ریختگی [۲] هستند. این سنجه‌ها مهمترین سنجه‌های ارزیابی کارایی تکنیکی در حوزه داده‌کاوی در صنعت سلامت بوده و در اکثر تحقیقات این حوزه استفاده شده‌اند [۲۸].

از جمله سنجه‌های ارزیابی که بر اساس ماتریس درهم‌ریختگی تعریف می‌شود می‌توان به دقت (Accuracy)، حساسیت (Sensitivity) (یا فراخوانی (Recall)، ویژگی(Specificity)، صحت (Precision) اشاره کرد که در ادبیات انفورماتیک پزشکی استفاده می‌شوند [۳۰-۲۸]. از دیگر سنجه‌های مورد استفاده یکی از این سنجه‌ها، ترکیبی از صحت و فراخوانی است که F-measure نام دارد و دیگری که ترکیبی از حساسیت و ویژگی است که g-means نام دارد.

$$F = 2.\frac{percision.recall}{percision + recall}$$

$$g-means = \sqrt{acc^{+}.acc^{-}}$$

در این رابطه acc^{+} نشان دهنده حساسیت و acc^{-} نشان دهنده ویژگی است. این سنجه در بسیاری از تحقیقات برای ارزیابی رده‌بندهایی ساخته شده روی داده‌های نامتوازن استفاده شده است [۳۱].

سنجه‌های ارزیابی غیرتکنیکی: رده‌بندی حساس به هزینه: یادگیری حساس به هزینه یکی از موضوعات داغ تحقیق در یادگیری ماشین است [۳۲]. در اغلب مدل‌های سنتی رده‌بندی از دقت به عنوان معیاری برای سنجش دقت رده‌بندی استفاده می‌شود، اما این سنجه فقط تعداد مواردی را که به طور صحیح رده‌بندی شده در نظر می‌گیرد و هزینه خطای رده‌بندی منتج از خطاهای نوع I (مثبت کاذب) و خطای نوع II (منفی کاذب) را برابر فرض می‌کند. اما این خطاها در دنیای واقعی برابر نیستند. به عنوان مثال هزینه پیش‌بینی سرطان برای بیمار غیر سرطانی متفاوت از پیش‌بینی عکس این مسأله است. در نتیجه دقت سنجه مناسبی برای نمایه سازی کارایی مدل‌های رده‌بندی نیست [۳۳].

طراحی جدول هزینه: در این بخش جدول هزینه‌ای که در مدل پیشنهادی مورد استفاده قرار می‌گیرد طراحی می‌شود. این جدول میزان حساسیت مدل را مشخص می‌کند. جدول ۴ هزینه درمان این بیماری را در مراحل مختلف در سال ابتدایی تشخیص، به ازای هر سال بقاء پس از سال اول و سال پایانی نشان می‌دهد [۳۴].

جدول ۴: هزینه درمان سرطان روده بزرگ در مراحل مختلف در سال اول تشخیص

مرحله	سرطان کولن	سرطان رکتوم	میانگین
0	$ ۱۸۰۵۲	$ ۱۳۹۵۴	$ ۱۶۷۶۲
I	$ ۲۷۷۸۳	$ ۲۵۶۵۹	$ ۲۷۰۹۹
II	$ ۳۵۰۵۵	$ ۴۰۲۱۷	$ ۳۶۰۹۲
III	$ ۴۱۲۲۲	$ ۴۳۵۱۸	$ ۴۱۷۹۶
IV	$ ۴۲۴۰۱	$ ۳۹۴۳۶	$ ۴۱۵۶۲

اگر واحد محاسبه هزینه را هزار دلار در نظر بگیریم و هزینه تست کولونوسکوپی را با colCost و هزینه درمان سرطان را با (care) نشان دهیم آنگاه جدول ۵ فرمول های توسعه جدول هزینه را نشان می دهد.

جدول۵: فرمول‌های توسعه جدول هزینه

	منفی پیش بینی شده	مثبت پیش بینی شده
منفی واقعی	-colCost	β*colCost
مثبت واقعی	α*care	-care/γ

در تعیین سنجه هزینه می‌توان از جداول هزینه مختلفی استفاده کرد تا علاوه بر تعیین هزینه هر مدل، مدلی را انتخاب کرد که از نظر سنجه‌های تکنیکی دیگر نیز خوب عمل کرده باشد [۳۵]. به همین علت برای این که بین سنجه هزینه و سنجه-های تکنیکی با اهمیت مصالحه انجام شود تا بهبود یکی به قیمت از دست دادن دیگری نباشد از ضرایب α، β و γ ($\alpha, \beta, \gamma > 0$) برای افزایش یا کاهش تأثیر هر کدام از عناصر جدول هزینه و همچنین بهبود سنجه‌های ارزیابی تکنیکی استفاده می‌کنیم. مقدار متغیر care برابر با میانگین مجموع اختلاف هزینه‌های سال ابتدایی مراحل (0,I) از مراحل بالای بیماری یعنی (II,III,IV) در نظر گرفته شده است.

$$care = \frac{\sum_{i=II}^{IV}(costS_i - costS_0) + \sum_{i=II}^{IV}(costS_i - costS_I)}{6}$$

سنجه مقرون به صرفه‌گی مالی: هزینه‌های سربار سرطان روده بزرگ بسیار بالا است و به میزان قابل توجهی بستگی به فاز درمان، بخش درگیر سرطان (بدنه روده یا رکتوم) و همچنین مراحل بیماری دارد [۳۴]. با توجه به حوزه کاری تحقیق پیش رو هر پیش‌بینی، هزینه یا سود مختص به خود را دارد و صحیح نیست که به عنوان مثال هزینه تمامی منفی‌های کاذب یکسان در نظر گرفته شود چرا که شخصی که بیماری‌اش در دنیای واقعی در مرحله II کشف شده با شخصی که بیماری‌اش در مرحله IV کشف شده هزینه‌های متفاوتی دارند. به همین دلیل احتیاج به سنجه‌ای داریم که این تمایز را بین پیش‌بینی‌های مختلف قائل باشد. از این رو دیگر سنجه‌ای که می‌تواند مورد علاقه حوزه سلامت باشد میزان تأثیری است که مدل پیشنهادی در کاهش هزینه‌های بیمار و حتی نظام سلامت خواهد داشت. در این تحقیق، چنین سنجه‌ای را توسعه و سنجه مقرون به صرفه‌گی مالی مدل MCE (Model Cost Effectiveness) نام‌گذاری می-کنیم.

در این سنجه از میان داده‌های تست در هر فولد تقابل سنجی، رکوردهایی که سبک زندگی آن‌ها به درستی پرخطر تشخیص داده شده، یعنی TPها، را جدا کرده و با توجه به جدول ۵ و رابطه زیر میانگین هزینه‌ای که با کشف زودهنگام بیماری در مراحل 0 یا I صرفه جویی می‌شود را محاسبه می‌کنیم. مقدار این رابطه TP_{save} می‌نامیم.

$$TP_{save} = \sum_{i=II}^{IV} numS_i * (costS_i - baseCost) + numS_I * baseCost$$

$$baseCost = \frac{costS_0 + costS_I}{2}$$

در این رابطه $numS_i$ تعداد افرادی هستند که سبک زندگی آن‌ها پرخطر تشخیص داده شده و در دنیای واقعی بیماری آن-ها در مرحله i کشف شده است. $costS_i$ هزینه درمان در مرحله i بوده و $costS_0$ و $costS_I$ به ترتیب هزینه درمان در مراحل 0 و I هستند و baseCost میانگین هزینه این دو مرحله است.

مقدار زیان مالی ناشی از پیش‌بینی نادرست منفی‌های کاذب باید از میزان صرفه‌جویی شده در قسمت قبل کسر شود. این مقدار را کمی سختگیرانه‌تر و از طریق رابطه زیر محاسبه می-کنیم. مقدار این رابطه را FN_{loss} می‌نامیم.

$$FN_{loss} = \sum_{i=I}^{IV} numS_i * costS_i$$

در این رابطه $numS_i$ تعداد افرادی هستند که سبک زندگی آن‌ها به اشتباه کم‌خطر تشخیص داده شده و در دنیای واقعی بیماری آن‌ها در مرحله i کشف شده است.

زیانی که مدل به خاطر پیش‌بینی نادرست مثبت‌های کاذب به بار می‌آورد را FP_{loss} نامیده و از رابطه زیر محاسبه می‌کنیم. در این تحقیق فرض می‌کنیم که فرد یک تست کولونوسکوپی در صورت این نوع پیش‌بینی‌ها انجام می‌دهد.

$$FP_{loss} = num_{FP} * colCost$$

در این رابطه num_{FP} تعداد مثبت‌های کاذب و colCost هزینه تست کولونوسکوپی به علاوه هزینه‌های جانبی می‌باشد.

با توجه به سه رابطه بالا میزان مقرون به صرفه‌گی مدل (MCE) برای داده‌های تست از رابطه زیر محاسبه می‌شود.

$$MCE = \frac{\sum_{fold=1}^{n}(TP_{save} - (FN_{loss} + FP_{loss}))_{fold}}{n}$$

نتایج

همان طور که در ساختار مدل سازی یعنی شکل ۲ مشخص است، مدل‌ها ابتدا روی مجموعه داده با تمام خصیصه‌ها اجرا شدند اما نتایج پایین‌تر از حد انتظار بود. برای بهبود نتایج، تصمیم به حذف خصیصه‌های نامناسب و ایجاد زیر مجموعه بهتری از خصیصه‌ها گرفته شد. این حذف خصیصه‌ها از طریق تکنیک‌های بصری‌سازی داده‌ها از جمله درخت تصمیم انجام شد.

جدول ۷ نتایج اجرای رده‌بندها با حذف ۴ خصیصه مصرف هورمون، عادت ماهیانه، مصرف الکل و همچنین گوشت قرمز را نشان می‌دهد. با حذف این خصیصه‌ها تمامی سنجه‌های ارزیابی در تمامی رده‌بندها بهبود داشت. این موضوع نشان می‌دهد که انتخاب یک زیرمجموعه مناسب از ویژگی‌های مؤثر، دقت را افزایش می دهد، زیرا ویژگی‌های نامرتبط کارایی الگوریتم را کم می‌کنند.

نتایج نشان داد که رده‌بند ماشین بردار پشتیبان ساده و ماشین بردار پشتیبان تجمیعی با دقت ۸۳/۷ ٪ در مجموع بهتر از دیگر الگوریتم‌ها عمل کرده‌اند. پس از این دو رده‌بند، K-NN با دقت ۸۳/۲٪ قرار داشت، که اختلاف کمی با یکدیگر دارند اما ماشین بردار پشتیبان در تشخیص افراد با سبک زندگی پرخطر بسیار بهتر از K-NN عمل کرده بود. در مجموع رده‌بندهای مبتنی بر ماشین بردار پشتیبان از لحاظ سنجه دقت و سنجه‌های بسیار مهم F-measure و g-means بهتر از دیگر رده‌بندها بودند.

نکته دیگر جالب توجه این است که در تمامی الگوریتم‌ها به جزء Naïve Bayes، ویژگی از حساسیت بیشتر بود. به این معنی که رده‌بندها به طور کلی در تشخیص افراد با سبک زندگی سالم بهتر از افراد با سبک زندگی پرخطر عمل کرده بودند. نکته جالب توجه دیگر این است که از نظر سنجه حساسیت، Naïve Bayes بهتر از دیگر الگوریتم‌ها عمل کرده بود. به این معنی که این رده‌بند بیشتر تمایل دارد که افراد با سبک پرخطر را تشخیص دهد.

جدول۶: رده بندهای غیرحساس به هزینه با خصیصه‌های برگزیده

الگوریتم	دقت	حساسیت	ویژگی	F-measure	g-means
SVM	٪۸۳/۷ ±۴/۷۹	٪۸۰/۹۵	٪۸۵/۸۸	٪۸۳/۷۰	٪۸۳/۳۳
Naïve Bayes	٪۸۰/۱ ±۴/۷۱	٪۸۵/۶۰	٪۸۳/۵۶	٪۸۰/۱۰	٪۷۹/۴۸
J48	٪۷۶/۶ ±۵/۵۷	٪۷۱/۴۳	٪۸۰/۴۴	٪۷۶/۶۰	٪۷۵/۸۰
K-NN (K=3)	٪۸۳/۲ ±۴/۵۴	٪۷۵	٪۸۹/۳۳	٪۸۳/۱۰	٪۸۱/۸۵
Bagging – SVM	٪۸۳/۷ ±۴/۸۹	٪۷۹/۱۷	٪۸۷/۱۱	٪۸۳/۷۰	٪۸۳/۰۴
Bagging – Naïve Bayes	٪۸۰/۹ ±۴/۴۹	٪۷۷/۹۸	٪۸۳/۱۱	٪۸۰/۹۰	٪۸۰/۵۰
میانگین (بدون J48)	٪۸۲/۳۲ ±۱/۷۰	٪۷۹/۷۴	٪۸۵/۸۰	٪۸۲/۳۰	٪۸۱/۶۴

ماتریس‌های درهم‌ریختگی رده‌بندها: ماتریس درهم‌ریختگی رده‌بندها نشان می‌دهد که بیشترین تشخیص سبک زندگی پرخطر مربوط به ماشین بردار پشتیبان با ۱۳۶ تشخیص بود. بیشترین تشخیص سبک زندگی سالم مربوط به K-NN با ۲۰۱ تشخیص بود. نکته دیگر این است که کمترین منفی کاذب (FN) را نیز ماشین بردار پشتیبان دارا بود. در حوزه سلامت پایین بودن منفی کاذب یک مزیت بزرگ برای رده‌بندها محسوب می‌شود. نکته دیگر این است که هرچند مثبت‌های کاذب رده‌بند K-NN پایین‌تر از دیگر رده‌بندها بود، در رده‌بند ماشین بردار پشتیبان نسبت منفی‌های کاذب به مثبت‌های کاذب پایین‌تر از دیگر رده‌بندها بود.

رده‌بندی حساس به هزینه: ماتریس هزینه پایه‌ای که بر اساس جدول ۷ (ارائه شده در بخش قبل) طراحی شده در جدول ۸ نمایش داده شده است. در این ماتریس علاوه بر زیان از تشخیص نادرست، سود ناشی از تشخیص درست نیز در نظر گرفته شده است.

جدول ۷: جدول هزینه پایه

	پرخطر پیش بینی شده	کم خطر پیش بینی شده
کم خطر واقعی	۸	-۴
پرخطر واقعی	-۱۰۷	۱۰۷

جدول ۸: ماشین بردار پشتیبان حساس به هزینه

الگوریتم	دقت	↓ دقت	↓منفی کاذب	↑مثبت کاذب	هزینه	F-measure	g-means
CS-SVM-1	٪۸۱/۱۸ ±۳/۹۶	٪۲/۵۱	٪۳۱	٪۶۳	-۲۳۰۲	٪۸۱/۳۰	٪۸۱/۷۴
CS-SVM-2	٪۸۱/۶۶ ±۳/۶۷	٪۲/۰۳	٪۳	٪۲۸	-۲۹۹	٪۸۱/۷۰	٪۸۱/۶۶
CS-SVM-3	٪۷۶/۱۱ ±۵/۶۶	٪۷/۵۸	٪۴۷	٪۱۴۱	-۱۱۰۶	٪۷۶/۱۰	٪۷۶/۸۹
CS-SVM-4	٪۷۹/۱۶ ± ۴/۶۵	٪۴/۵۳	٪۳۱	٪۸۸	-۱۱۶۰	٪۷۹/۲۰	٪۷۹/۸۳
میانگین	٪۷۹/۵۳ ± ۲/۵۲				-۶۳۶	٪۷۹/۵۸	٪۸۰

جدول ۹: نزدیکترین همسایگی حساس به هزینه

الگوریتم	دقت	↓دقت	↓منفی کاذب	↓مثبت کاذب	هزینه	F-measure	g-means
CS-KNN-1	٪۷۵/۳۵ ±۴/۳۰	٪۷/۸۶	٪۶۴	٪۲۱۳	-۳۱۲۲	٪۷۷/۳۰	٪۷۸/۱۷
CS-KNN-2	٪۸۱/۹۲ ± ۶/۳۳	٪۱/۲۹	٪۴۰	٪۹۱	-۹۵۱	٪۸۲	٪۸۲/۲۹
CS-KNN-3	٪۶۸/۱۷ ±۶/۲۹	٪۱۵/۰۴	٪۸۶	٪۳۹۶	-۳۱۳۲	٪۶۶/۹۰	٪۶۷/۴۰
CS-KNN-4	٪۷۴/۰۵ ± ۳/۹۱	٪۹/۱۶	٪۸۱	٪۲۹۲	-۹۰۴	٪۷۳/۶۰	٪۷۴/۴۶
میانگین	٪۷۴/۸۷ ±۵/۶۴				-۲۰۲۷	٪۷۴/۹۵	٪۷۵/۵۸

جدول ۱۰: بیزین ساده حساس به هزینه

الگوریتم	دقت	↓دقت	↓منفی کاذب	↓مثبت کاذب	هزینه	F-measure	g-means
CS-NB-1	٪۷۸/۴۰ ± ۳/۸۳	٪ ۱/۷۹	٪ ۵۱	٪ ۷۶	-۲۴۲۰	٪ ۷۸/۴۰	٪ ۷۹/۱۵
CS-NB-2	٪۷۹/۹۳ ±۴/۶۵	٪ ۰/۲۶	٪ ۳۴	٪ ۴۱	-۴۷۳	٪ ۸۰	٪ ۸۰/۳۳
CS-NB-3	٪ ۷۴/۸۲ ±۲/۹۳	٪ ۵/۳۷	٪ ۶۶	٪ ۱۳۰	-۱۷۰۰	٪ ۷۴/۶۰	٪ ۷۵/۵۲
CS-NB-4	٪۷۷/۳۸ ±۳/۲۶	٪ ۲/۸۱	٪ ۵۴	٪ ۸۹	۸۱۰	٪ ۷۷/۴۰	٪ ۷۸/۱۷
میانگین	٪ ۷۷/۶۳ ± ۲/۱۵				-۹۴۵	٪ ۷۷/۶۰	٪ ۷۸/۲۹

نتایج جداول فوق نشان می‌دهد که در مجموع K-NN با میانگین هزینه (۲۰۲۷ -)، از لحاظ میانگین هزینه‌ها بسیار بهتر از دیگر رده‌بندها بهتر عمل کرده بود. اما اگر مصالحه کلی سنجه‌ها مدنظر باشد همچنان با فاصله بالایی ماشین بردار پشتیبان بهتر عمل کرده بود.

رده‌بند تجمیعی ناهمگن:

نتایج رده‌بندهای بالا نشان داد که هرکدام از رده‌بندها در سنجه خاصی بهتر از دیگر رده‌بندها عمل کرده بودند. به عنوان مثال رده‌بندهای مبتنی بر ماشین بردار پشتیبان از نظر دقت و F-measure بهتر از دیگر رده‌بندها بودند. رده بند Naïve Bayes حساسیت خوبی داشت و همچنین وقتی پای هزینه به میان آمده رده بند K-NN بهتر از دیگر رده‌بندها عمل کرده بود. در این بخش رده‌بندی توسعه می‌دهیم که از مزایای تمامی رده-بندها در کنار هم استفاده کرده و توانسته تمامی سنجه‌ها را در کنار هم به سطح بالایی برساند.

این رده‌بند که رده‌بند تجمعی دقیق و حساس به هزینه CS-ACE (Cost Sensitive – Accurate Ensemble) نام گرفته است از ترکیب رده‌بندهای غیر حساس به هزینه SVM، KNN (K=3) و Naïve Bayes و همچنین رده-بندهای حساس به هزینه CS-SVM-3، CS-KNN-1 و CS-NB-1 تشکیل شده است. رده بند تجمیعی CS-ACE از میانگین احتمال‌ها برای رأی گیری استفاده می‌کند. برای انتخاب رده‌بندهای حساس به هزینه، شکل‌های مختلفی از انتخاب‌ها بررسی شد. به عنوان مثال یک بار رده‌بندهایی انتخاب شدند که F-measure بالایی داشتند، یا رده بندهایی که کمترین هزینه را داشتند. اما بهترین انتخاب مصالحه بین دقت، F-measure و همچنین میزان کاهش منفی‌های کاذب همزمان با عدم افزایش زیاد مثبت‌های کاذب بود که در این حالت بهترین نتیجه به دست آمد. جدول ۱۱ نتایج ارزیابی این رده‌بند را نشان می‌دهد. قابل ذکر است، سطح زیر نمودار ROC برای این رده‌بند برابر ۰/۹۱ بوده است.

جدول ۱۱: رده بند CS-ACE

الگوریتم	دقت	حساسیت	ویژگی	هزینه	F-measure	g-means
CS-ACE	٪ ۸۴/۲۲±۳/۵۲	٪ ۸۴/۲۰	٪ ۸۱/۳۴	-۲۰۷۶	٪ ۸۴/۳۰	٪ ۸۲/۷۶

جدول ۱۱ نشان می‌دهد که رده‌بند توسعه داده شده در اکثر سنجه‌های مهم از میانگین نتایج و حتی از نتایج تک تک رده-بندهای غیرحساس به هزینه و حساس به هزینه بسیار بهتر عمل کرده است. به این ترتیب ما رده‌بندی داریم که هم دقیق، با F-measure بالا و هم حساس به هزینه است. در واقع مصالحه بین این سه سنجه مهم به بهترین وجه ممکن شکل گرفته است.

سنجه مقرون به صرفه‌گی مالی مدل:

در این تحقیق رده‌بندهای مختلف از لحاظ سنجه‌های مختلف تکنیکی بررسی شدند، اما در دنیای واقعی این ارزیابی تکنیکی هدف نهایی نیست. هدف نهایی استفاده از داده‌کاوی بهینه کردن سنجه هزینه است. جدول ۱۲ نتایج رده‌بند CS-ACE را در کنار رده‌بندهای حساس به هزینه‌ای نشان می‌دهد که بالاترین F-measure و دقت را داشته‌اند، که اجراهای حاصل از ماتریس هزینه دوم بود. همان طور که مشاهده می‌شود رده‌بند CS-ACE در تمامی سنجه‌ها از دیگر رده‌بندها پیش است.

جدول ۱۲: مقایسه CS- ACE با اجراهای حساس به هزینه باجدول هزینه دوم

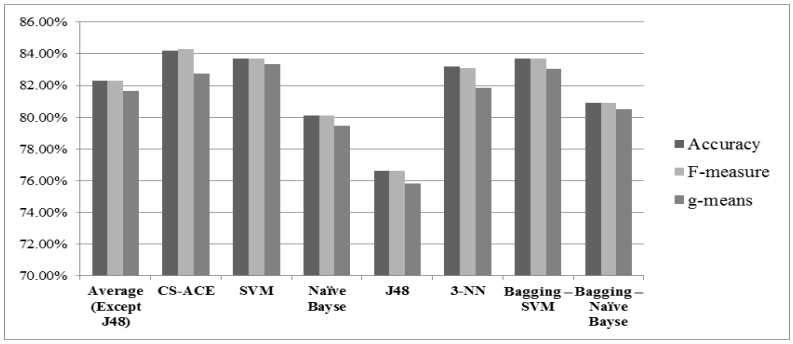
الگوریتم	دقت	AUC	F-measure	g-means	MCE
CS-ACE	٪۸۴/۲۲	۰/۹۱	٪۸۴/۳۰	٪۸۲/۷۶	$ ۱۴۹۵۸۲
CS-SVM-2	٪۸۱/۶۶	۰/۸۹	٪۸۱/۷۰	٪۸۱/۶۶	$ ۸۳۷۹۶
CS-NB-2	٪۷۹/۹۳	۰/۸۸	٪۸۰	٪۸۰/۳۳	$ ۱۰۴۹۹۸
CS-KNN-2	٪۸۱/۹۲	۰/۹۰	٪۸۲	٪۸۲/۲۹	$ ۱۱۲۳۷۸

در جدول ۱۳رده بند CS-ACE به همراه اجراهای حساس به هزینه تکنیک‌ها با استفاده از ماتریس هزینه اول نشان داده شده است. این رده‌بندها نسبت به رده‌بندهای، F-measure پایین-تری داشتند. این کاهش خصوصاً برای رده‌بند نزدیکترین همسایگی قابل توجه بود. در مقابل این کاهش، بهبود قابل توجهی نیز در میزان صرفه‌جویی مالی به وجود آمده بود. با این حال فقط رده‌بند نزدیکترین همسایگی توانسته صرفه‌جویی بالاتری نسبت به رده‌بند CS-ACE داشته باشد. این افزایش اما به قیمت کاهش ۵ درصدی سنجه مهم F-measure نسبت به CS-KNN-2 و اختلاف ۷ درصدی با رده‌بند CS-ACEبود. انتخاب هرکدام از این رده‌بندها بستگی به مصالح حوزه دارد که تصمیم بگیرد کدام سنجه را قربانی دیگری کند.

جدول۱۳: مقایسه CS-ACE با اجراهای حساس به هزینه با جدول هزینه اول

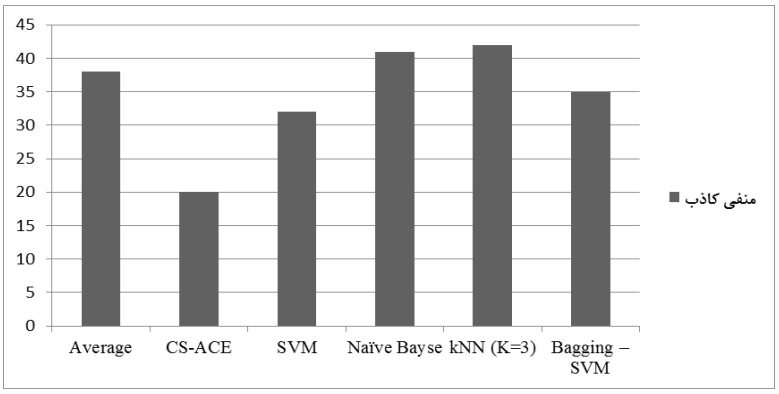
الگوریتم	دقت	AUC	F-measure	g-means	MCE
CS-ACE	٪۸۴/۲۲	۰/۹۱	٪۸۴/۳۰	٪۸۲/۷۶	$ ۱۴۹۵۸۲
CS-SVM-1	٪۸۱/۱۸	۰/۸۸	٪۸۱/۳۰	٪۸۱/۷۴	$ ۱۳۱۰۵۶
CS-NB-1	٪۷۸/۴۰	۰/۸۸	٪۷۸/۹۳	٪۷۹/۲۳	$ ۱۳۴۱۷۴
CS-KNN-1	٪۷۷/۳۷	۰/۹۰	٪۷۷/۳۰	٪۷۸/۱۷	$ ۱۵۷۲۶۴

نمودار شکل ۳ نشان می‌دهد که F-measure و دقت رده بند CS-ACE بالاتر از میانگین و همچنین تک تک رده‌بندهای غیرحساس به هزینه ساده و تجمیعی همگن می‌باشد. این رده‌بند تنها از نظر g-means اندکی پایین‌تر از رده‌بند ساده و تجمیعی ماشین بردار پشتیبان بوده که این مسأله به خاطر پایین-تر بودن ویژگی می‌باشد. اما از نظر حساسیت رده‌بند CS-ACE با اختلاف قابل توجهی بهتر عمل کرده است.



شکل ۳: مقایسه CS-ACE با رده بندهای غیر حساس به هزینه

نکته مهم این است که این رده‌بند در عین F-measure و دقت بالا تعداد منفی‌های کاذب پایینی دارد، به طوری که این مقدار از میانگین منفی‌های کاذب رده‌بندهای غیر حساس به هزینه که حدود ۳۸ است نزدیک به ۵۰ ٪ کمتر است. شکل ۴ رده‌بندها را از این نظر مقایسه می‌کند.



شکل ۴: مقایسه رده بند CS-ACE با دیگر رده بندها از نظر تعداد منفی های کاذب

به این ترتیب ما رده‌بندی داریم که هم دقیق، با F-measure بالا و هم حساس به هزینه است. در واقع مصالحه بین این سه سنجه مهم به بهترین وجه ممکن شکل گرفته است.

گزارش ارزیابی کاربردی: هدف داده‌کاوی حوزه محور شناسایی الگوهای قابل اجرا است. داده‌کاوی حوزه محور معتقد است فقط الگوهایی قابل اجرا و قادر به حل مسائل دنیای واقعی هستند که هم سنجه‌های تکنیکی و هم سنجه‌های جذاب از نظر حوزه را ارضا کنند [۳۶]. با توجه به این هدف اگر آستانه سنجه‌های تکنیکی دقت، g-means و F-measure را ۸۰ درصد در نظر گرفته و آستانه سنجه مقرون به صرفه‌گی را $ ۱۳۰،۰۰۰ در نظر بگیریم آنگاه جدول ۱۵ رده بندها را از نظر قابلیت اجرا مقایسه می‌کند.

جدول ۱۴: مقایسه رده بندها از نظر قابلیت اجرا

مدل	گذر از آستانه				
	دقت	F-measure	G-MEANS	MCE	قابلیت اجرا
CS-ACE	↑	↑	↑	↑	↑
CS-SVM-1	↑	↑	↑	↑	↑
CS-SVM-2	↑	↑	↑	↓	↓
CS-KNN-1	↓	↓	↓	↑	↓
CS-KNN-2	↑	↑	↑	↓	↓
CS-NB-1	↓	↓	↓	↑	↓
CS-NB-2	↓	↑	↑	↓	↓

از میان بهترین رده‌بندهای توسعه داده شده فقط دو مدل توانسته‌اند هم سنجه‌های تکنیکی و هم سنجه‌های غیرتکنیکی را ارضا کنند. نکته قابل توجه این است که هرچند مدل‌های زیادی از لحاظ تکنیکی جذاب هستند، اما در واقع تعداد مدل-هایی که واقعاً قابل اجرا و کاربردی می‌باشند اندک هستند.

بحث و نتیجه گیری

سه شکاف کلیدی مشاهده شده در ادبیات حوزه، انگیزه محققین در این تحقیق می‌باشد. این سه شکاف عبارت از ۱) عدم توجه به فاز غربالگری، ۲) عدم توجه به بعضی بیماری‌ها از جمله سرطان روده بزرگ و ۳) دیدگاه‌های غیرکاربردی تحقیقات این حوزه هستند. امروزه تحقیقات کمی، به خصوص در داخل کشور برای

پل زدن بر شکاف موجود تلاش می‌کنند. این تحقیق نشان داد که برنامه غربالگری سرطان به طور کلی و غربالگری سرطان روده بزرگ به طور خاص، با استفاده از مدل‌های مبتنی بر کشف دانش و داده‌کاوی هم از لحاظ کیفی و هم از لحاظ مسائل مالی قابل بهبود است. همچنین اتخاذ دیدگاه‌های کاربردی، نتایج تحقیقات را برای حوزه سلامت که قرار است از آن‌ها برای حل مشکلات و مسائل خود استفاده کند قابل فهم و قابل اجرا می‌کند.

ارزش اصلی تحقیق، ارائه مدلی است که با کارایی بالایی به غربالگری بیماری مهلک سرطان روده بزرگ و کشف این بیماری در مراحل اولیه و حتی کنترل سبک زندگی فرد برای کاهش ابتلای او به این بیماری کمک می‌کند. این نوع کار در ادبیات حوزه تحقیق چه در جهان و چه در کشور کمتر انجام شده است. از طرف دیگر این کار تعریف یک مسأله جدید در حوزه کشف دانش و داده کاوی با توجه به یکی از مشکلات موجود در کشور و دنیا و سپس تحلیل این مسأله با هدف حل مسأله می‌باشد. با این نگاه مدل‌های پایه غیر حساس به هزینه با استفاده از تکنیک‌های ماشین بردار پشتیبان، بیزین ساده، نزدیکترین همسایگی و درخت تصمیم توسعه داده شده و ارزیابی و مقایسه شدند. نتایج نشان داد که پس از حذف چند خصیصه سنجه‌های ارزیابی در تمامی رده‌بندها بهبود داشته است. این موضوع نشان می‌دهد که انتخاب یک زیرمجموعه مناسب از خصیصه‌های مؤثر، نتایج را بهبود می‌بخشد و ویژگی‌های نامناسب می‌تواند تأثیر منفی در کارایی مدل‌ها داشته باشد. این تأثیر به خصوص در ماشین بردار پشتیبان و بیزین ساده چشمگیر بود. نتایج نشان می‌دهد که حذف چهار خصیصه حداکثر حدود ۵٪ در سنجه دقت تأثیر داشته است. ماشین بردار پشتیبان در این حوزه بهتر از دیگر تکنیک‌ها عمل کرده است. تکنیک بیزین ساده بهترین تکنیک از لحاظ تشخیص سبک زندگی پرخطر بود. بدترین نتایج در میان رده‌بندها را رده‌بند درخت تصمیم به خود اختصاص داده است. این مسأله نشان از پیچیده بودن مرزهای تصمیم و همچنین غیرخطی بودن آن در داده‌ها دارد. چهار ماتریس هزینه بر اساس هزینه‌های واقعی سرطان روده بزرگ طراحی، بر اساس این ماتریس‌ها مدل‌های حساس به هزینه توسعه داده شد. این مدل‌ها با یکدیگر و همچنین با مدل‌های پایه مقایسه شدند. استفاده از هزینه‌های واقعی سرطان روده بزرگ برای طراحی جدول هزینه یکی از ویژگی‌های اصلی این تحقیق است. چنین مسأله‌ای به ندرت در ادبیات حوزه دیده شده بود. چرا که با این کار علاوه بر این که تکنیکی داریم که هزینه تشخیص اشتباه در آن‌ها بالاتر رفته است، هزینه نهایی به دست آمده از مدل نیز یک سنجه واقعی و کاربردی برای ارزیابی و مقایسه مدل‌ها در اختیار محققین قرار می‌دهد. تکنیک نزدیکترین همسایگی حساس به هزینه بهتر از دیگر تکنیک‌ها عمل کرد. در نهایت یک مدل تجمیعی غیرهمگن توسعه داده شد که دقیق، حساس به هزینه، با F-measure بالا و منفی‌های کاذب پایین می‌باشد و از کلیه مدل‌های توسعه داده شده از نظر سنجه‌های تکنیکی و غیر تکنیکی برتر بود. نکته جالب توجه‌ای که در نتایج مشاهده شد این بود که رده‌بندهای تجمیعی همگن برتری خاصی را نسبت به رده‌بندهای ساده از خود نشان ندادند. اما با ترکیب رده‌بندهایی که هر کدام مزیت خاصی داشتند و استفاده از آن‌ها در کنار هم رده‌بند ارزشمندی با نتایج خوب توسعه داده شد. نکته بسیار جالب توجه دیگر در این تحقیق نتایج نسبتاً خوب تکنیک نزدیکترین همسایگی بود که با توجه به این که چنین مسأله‌ای در ادبیات حوزه کمیاب است برای محققین غیرمنتظره بود. اما از آنجایی که دلیل خوبی این نتایج بیشتر به خاطر عملکرد خوب این تکنیک در پیش‌بینی سبک زندگی کم‌خطر است نسبت ماشین بردار پشتیبان و حتی بیزین ساده از اهمیت کمتری برخوردار است.

به طور قطع یکی از ارزش‌های مهم و برتری این تحقیق نسبت به تحقیقات مشابه، مجموعه داده آن است که به صورت هدفمند جمع‌آوری شده و می‌تواند به عنوان پایه تحقیقات بعدی قرار گیرد. خصیصه‌های این مجموعه داده با هدف غربالگری سرطان روده بزرگ انتخاب و داده‌های تحقیق از منابع مختلف و در طول هشت ماه جمع‌آوری گردیده است. چنین مجموعه داده‌ای با خصیصه‌های در نظر گرفته شده در کشور بی‌نظیر و در جهان کم نظیر می‌باشد. در بسیاری از تحقیقات به خصوص در داخل کشور، محققین بر استفاده از مجموعه داده‌های از پیش آماده روی وب و حتی مجموعه داده‌های آماده بیمارستانی اصرار دارند. در حالی که با اتخاذ دیدگاه حل مشکلات می‌توان به طور هدفمند و به منظور حل مشکل مورد نظر، خصیصه‌ها را مشخص کرده، مجموعه داده‌ها را جمع آوری و برای انجام داده‌کاوی آماده کرد.

اکثر مقالات حوزه داده‌کاوی در صنعت سلامت بعد از ارزیابی تکنیکی مدل‌های توسعه داده شده و یا الگوهای کشف شده متوقف می‌شوند. اما ارزش دیگر تحقیق، نگاه جدید به حل مسأله، یعنی نگاه حوزه محور به حل مسأله است. چرا که هدف نهایی خود را از ابتدا قابلیت اجرای مدل پیشنهادی در صنعت سلامت و نه صرفاً ارائه مدل و تحلیل تکنیکی آن قرار داده

بودیم. در این تحقیق سنجه‌ای ارائه شد که این قابلیت را داشت که ارزش تک تک پیش‌بینی مدل را به طور مجزا و با نگاهی واقعی ارزیابی کند. این سنجه برخلاف سنجه‌های معمول ارزیابی مدل‌ها که تمامی پیش‌بینی‌های هم نوع در ماتریس درهم‌ریختگی را یکسان در نظر می‌گرفت، با توجه به هزینه‌های واقعی درمان سرطان روده بزرگ، ارزش واقعی تک تک پیش-بینی‌ها را مورد سنجش قرار داده و یک ارزیابی واقعی و کاربردی از مدل توسعه داده شده ارائه می‌دهد. مکانیزم ارزیابی ارائه شده در این سنجه می‌تواند در حوزه‌های کاربرد داده‌کاوی در صنعت سلامت نیز به کار گرفته شود. در نهایت کارایی مدل‌ها هم از نظر سنجه‌های تکنیکی و هم از نظر سنجه مورد علاقه حوزه سلامت ارزیابی شد. چراکه از نظر داده‌کاوی حوزه محور فقط مدل‌هایی قابلیت اجرا دارند که از هردو نظر تکنیکی و حوزه جذاب و کارا باشند. نکته بسیار جالب توجه از این نظر این بود که از میان بهترین رده‌بندهای توسعه داده شده فقط دو رده‌بند توانسته‌اند هم سنجه‌های تکنیکی و هم سنجه‌های غیرتکنیکی را ارضا کنند. این مسأله ایراد بزرگ بسیاری از تحقیقات یعنی عدم توجه به مسأله کاربردی بودن و همچنین اهمیت اتخاذ دیدگاه کاربردی در تحقیقات را آشکار می‌سازد. انتخاب هرکدام از این رده‌بندها بستگی به مصالح حوزه یا کسب و کار دارد که تصمیم بگیرد کدام سنجه را قربانی دیگری کند. این رده‌بندها دو رده‌بند CS-ACE و همچنین ماشین بردار پشتیبان حساس به هزینه CS-SVM-1 به ترتیب با F-measure های ۸۴/۳۰٪ و ۸۱/۳۰٪ بودند. این مسأله کارایی ماشین بردار پشتیبان در حوزه سلامت را نیز مورد تأیید قرار می‌دهد.

در حال حاضر محققین در حال تکمیل تحقیقات خود برای پیاده-سازی برنامه کاربردی مبتنی بر رده‌بندهای توسعه داده شده برای گوشی‌های هوشمند می‌باشند. این برنامه کاربردی در کنار بررسی سبک زندگی فرد از نظر خطر ابتلا به سرطان روده بزرگ، یکسری پیشنهادهای سفارشی شده نیز در جهت کاهش این خطر و همچنین تشخیص زودهنگام سرطان روده بزرگ ارائه می‌دهد. از این نظر در آموزش افراد برای بهبود سبک زندگی خود نیز مؤثر خواهد بود. به این ترتیب هر دو جزء اصلی در برنامه‌های کشف زود هنگام درباره سرطان یعنی آموزش برای ارتقاء تشخیص زود هنگام و دیگری غربالگری در نظر گرفته شده است. از مسیرهای توسعه این تحقیق می‌توان به توسعه سنجه پیشنهادی تحقیق اشاره کرد. توسعه سنجه‌های غیرتکنیکی جدیدی که علاوه بر هزینه افزایش طول عمر افراد در نتیجه استفاده از مدل‌های پیشنهادی و کشف زود هنگام بیماری را محاسبه و ارزیابی کند. استخراج دانش صریح با استفاده از قوانین انجمنی و درخت تصمیم می‌تواند یکی دیگر از مسیرهای تحقیق باشد. قوانین انجمنی روابط و وابستگی‌های متقابل بین مجموعه بزرگی از اقلام داده‌ای را نشان می‌دهند. استفاده از دیگر تکنیک پرکاربرد در داده‌کاوی برای توسعه مدل رده‌بندی یعنی شبکه عصبی مصنوعی می‌تواند دیگر حوزه کاری برای توسعه بیشتر تحقیق باشد. در نهایت مطمئناً افزایش تعداد بیماران در مجموعه داده‌ها از منابع دیگر می‌تواند بر کارایی و اعتبار مدل بیافزاید.

از محدودیت‌های تحقیق می‌توان به مشکلات جمع‌آوری داده در حوزه سلامت اشاره کرد. بسیاری از خصیصه‌های مؤثر در پرونده‌های بیماران درج نشده است و نزد خود بیمار در پرونده-های خانگی نگهداری می‌شود. از طرف دیگر تخصصی بودن بعضی از خصیصه‌های مؤثر امکان جمع‌آوری آن‌ها را از طریق تماس با بیمار غیر ممکن می‌سازد. پیشنهاد می‌شود که برای تعریف تحقیقات در این حوزه، با توافقات بلند مدت با مراکز تحقیقاتی در حوزه سلامت در جهت جمع‌آوری مؤثر و هدفمند داده‌ها تلاش شود. محدودیت دیگری که این نوع تحقیقات به طور کلی با آن مواجه هستند این است که حجم جزئیاتی که در کشف دانش قابل اجرا در حوزه سلامت باید به آن‌ها توجه داشت بسیار بالا است. این مسأله ضرورت کار تیمی در تحقیقات این-چنینی و همکاری بسیار نزدیک با مشاوران و خبرگان حوزه سلامت را اجتناب ناپذیر می‌سازد. بنابراین تعیین تیم‌های تحقیقاتی از مجموعه ای از افراد با مسئولیت‌های مشخص در هر دو حوزه فنی و سلامت می‌تواند به تعریف تحقیقات اثربخش و کاربردی کمک شایانی کند.

References

1. Vickers AJ. Prediction models in cancer care. CA Cancer J Clin. 2011;61(5):315-26.

2. Tan PN, Steinbach M, Kumar V. Introduction to data mining. 1th ed. Pearson Addison-Wesley; 2005.

3. Wells BJ, Kattan MW, Cooper GS, Jackson L, Koroukian S. Colorectal cancer predicted risk online (CRC-PRO) calculator using data from the multi-ethnic cohort study. J Am Board Fam Med. 2014;27(1):42-55.

4. Freedman AN, Slattery ML, Ballard-Barbash R, Willis G, Cann BJ, Pee D, et al. Colorectal cancer risk

prediction tool for white men and women without known susceptibility. J Clin Oncol. 2009;27(5):686-93.
5. Barazandeh I, Gholamian MR. Knowledge discovery and data mining applications in the healthcare industry: a comprehensive study. USA: IGI-Global; 2015.
6. Esfandiari N, Babavalian MR, Eftekhari Moghadam AM, Tabar VK. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications. 2014;41(9):4434-63.
7. Moon SS, Kang S-Y, Jitpitaklert W, Kim SB. Decision tree models for characterizing smoking patterns of older adults. Expert Systems with Applications. 2012;39(1):445-51.
8. Parhizi S, Steege LM, Pasupathy KS. Mining the relationships between psychosocial factors and fatigue dimensions among registered nurses. International Journal of Industrial Ergonomics. 2013;43(1):82-90.
9. Kumar M, Ghani R, Mei ZS. Data mining to predict and prevent errors in health insurance claims processing. KDD '10 the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2010 Jul 25– 28; USA: ACM ; 2010. p. 65-74.
10. Bellazzi R, Ferrazzi F, Sacchi L. Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011;1(5):416-30.
11. Marcano-Cedeño A, Chausa P, García A, Cáceres C, Tormos JM, Gómez EJ. Data mining applied to the cognitive rehabilitation of patients with acquired brain injury. Expert Systems with Applications. 2013;40(4):1054-60.
12. Chi CL, Street WN, Katz DA. A decision support system for cost-effective diagnosis. Artif Intell Med. 2010;50(3):149-61.
13. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med. 2007;41(3):251-62.
14. Ong LS, Shepherd B, Tong LC, Seow-Choen F, Ho YH, Tang CL, et al. The Colorectal Cancer Recurrence Support (CARES) System. Artif Intell Med. 1997;11(3):175-88.
15. Anand SS, Smith AE, Hamilton PW, Anand JS, Hughes JG, Bartels PH. An evaluation of intelligent prognostic systems for colorectal cancer. Artif Intell Med. 1999;15(2):193-214.
16. Grumett S, Snow P, Kerr D. Neural networks in the prediction of survival in patients with colorectal cancer. Clin Colorectal Cancer. 2003;2(4):239-44.
17. Shi J, Su Q, Zhang C, Huang G, Zhu Y. An intelligent decision support algorithm for diagnosis of colorectal cancer through serum tumor markers. Comput Methods Programs Biomed. 2010;100(2):97-107.
18. Roman M, Jitaru P, Agostini M, Cozzi G, Pucciarelli S, Nitti D, et al. Serum seleno-proteins status for colorectal cancer screening explored by data mining techniques - a multidisciplinary pilot study. Microchemical Journal. 2012;105:124-32.
19. Al-Bahrani R, Agrawal A, Choudhary A. Colon cancer survival prediction using ensemble data mining on SEER data. Big Data, 2013 IEEE International Conference on; 2013 Oct 6-9. Silicon Valley, CA: IEEE; 2013.
20. Stojadinovic A, Bilchik A, Smith D, Eberhardt JS, Ward EB, Nissan A, et al. Clinical decision support and individualized prediction of survival in colon cancer: bayesian belief network model. Ann Surg Oncol. 2013;20(1):161-74.
21. Colorectal Cancer American Cancer Society; 2013 [cited 2014 Mar 17]. Available from: http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/index.
22. Bowel cancer risk factors: cancer research uk; 2014 [cited 2014 Mar 17]. Available from: http://www.cancerresearchuk.org/cancer-info/cancerstats/types/bowel/riskfactors/bowel-cancer-risk-factors#page.
23. Norat T, Bingham S, Ferrari P, Slimani N, Jenab M, Mazuir M, et al. Meat, fish, and colorectal cancer risk: the European Prospective Investigation into cancer and nutrition. J Natl Cancer Inst. 2005;97(12):906-16.
24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16(1):321-57.
25. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81-97.
26. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2(3):1-27.
27. Hsu CW, Chang CC, Lin CJ. A Practical Guide to SupportVector Classication. 2010.
28. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, et al. A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed. 2013;111(1):52-61.
29. Cho I, Park I, Kim E, Lee E, Bates DW. Using EHR data to predict hospital-acquired pressure ulcers: a prospective study of a Bayesian Network model. Int J Med Inform. 2013;82(11):1059-67.
30. Gil D, Johnsson M. Using support vector machines in diagnoses of urological dysfunctions. Expert Systems with Applications. 2010;37(6):4713-8.
31. Akbani R, Kwek S, Japkowicz N. Applying Support Vector Machines to Imbalanced Datasets. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D, editors. Machine Learning: ECML 2004. Lecture Notes in Computer Science. Italy, Pisa: Springer Berlin Heidelberg; 2004. p. 39-50.
32. Qi Z, Tian Y, Shi Y, Yu X. Cost-Sensitive Support Vector Machine for Semi-Supervised Learning. Procedia Computer Science. 2013;18:1684-9.
33. Kim J, Choi K, Kim G, Suh Y. Classification cost: An empirical comparison among traditional classifier, Cost-Sensitive Classifier, and Meta Cost. Expert Systems with Applications. 2012;39(4):4013-9.

34. Lang K, Lines LM, Lee DW, Korn JR, Earle CC, Menzin J. Lifetime and treatment-phase costs associated with colorectal cancer: evidence from SEER-Medicare data. Clin Gastroenterol Hepatol. 2009 Feb;7(2):198-204.
35. Alizadehsani R, Hosseini MJ, Boghrati R, Ghandeharioun A, Khozeimeh F, Alizadeh Sani Z. Exerting Cost-Sensitive and Feature Creation Algorithms for Coronary Artery Disease Diagnosis. International Journal of Knowledge Discovery in Bioinformatics. 2012;3(1):59-79.
36. Longbing C. Domain-Driven Data Mining: Challenges and Prospects. Knowledge and Data Engineering, IEEE Transactions on. 2010;22(6):755-69.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2015; 2(2): 59-75

A Domain-Driven Classification Model to Early Detection of Individuals Having High Risk to Develop Colorectal Cancer

Iman Barazandeh[1]*, Mohammad Reza Gholamian [2], Abdolhasan Talaiezadeh [3], Mohammad Amin Pourhoseingholi [4]

• **Received:** 16 Aug, 2015 • **Accepted:** 9 Sep, 2015

Introduction: The aim of this research is to improve Colorectal Cancer screening trying to control an individual lifestyle to reduce the probability of developing Colorectal Cancer, detect the disease in early stages, and then accelerate the screening of risky individuals and postpone the screening of ones with low risk.

Method: In this retrospective study information of 309 individuals including 84 patients whose diagnosis had been between years 2006 to 2013 and 225 healthy individuals were collected through phone or face to face interviews and exploring patient medical records. Popular techniques to develop classification models in clouding support vector machine, naive bayes, k-nearest neighbor, and decision tree were applied. Finally actionable models were determined according to both types of measures and based on domain-driven data mining approach.

Results: The results show that most of the developed models have acceptable evaluation results in predicting lifestyles. The developed non-technical measure clearly distinguishes the value of every false negative prediction and every true positive prediction itself. And finally, the actionable classifiers have been selected for domain practitioners. Only two of all the developed classifiers could satisfy both technical and non-technical measures.

Conclusion: The results showed developed models must not only be evaluated by technical measures, but also be evaluated by medical domain interestingness, and also their application ability to actual problem solving should be explored.

Keywords: Domain -Riven Data Mining, Classification, Colorectal Neoplasms, Early Detection of Cancer

• **Citation:** Barazandeh I, Gholamian MR, Abdolhasan Talaiezadeh A, Pourhoseingholi MA. A Domain-Driven Classification Model to Early Detection of Individuals Having High Risk to Develop Colorectal Cancer. Journal of Health and Biomedical Informatics 2015; 2(2): 59-75.

1. M.Sc. in Information Technology Engineering, Lecturer, Information Technology & Computer Engineering Dept., School of Electrical and Computer, Islamic Azad University Mahshahr Branch, Mahshahr, Iran.
2. Ph.D. in Industrial Engineering, Assitant Professor, Industrial Engineering Dept., School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
3. M.D., Associate Professor, Cancer Surgery Dept., Petroleum and Environmental Pollutants Research Center, School of Medical, Ahvaz Jundishapur University of Medical Science, Ahvaz, Iran.
4. Ph.D. in Biostatistics, Assitant Professor, Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

***Correspondence**: Islamic Azad University, Mahshahr branch, Daneshgah Avenue, Mahshahr, Iran
• **Tel:** 06152327070 • **Email:** barazandeh_i@ind.iust.ac.ir