# طراحی مدل پیشبینی خطر سرطان کولورکتال مبتنی بر تکنیک داده‌کاوی رگرسیون‌لجستیک

رئوف نوپور[1] هادی کاظمی‌آرپناهی[2] مصطفی شنبه‌زاده[3*]

۱. کارشناسی ارشد، فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران. ORCID: 0000-0003-3370-2375
۲. گروه فناوری اطلاعات سلامت، دانشکده علوم پزشکی آبادان، ایران.
۳. گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی ایلام، ایران.

مجله اطلاع‌رسانی پزشکی نوین؛ دوره ششم؛ شماره چهارم؛ زمستان ۱۳۹۹؛ صفحات ۱۰-۱.

## چکیده

***هدف:*** *استفاده از یادگیری ماشین جهت تشخیص زودرس سرطان کولورکتال نقش مهمی در بهبود شاخص‌های بیماری دارد؛ هدف مطالعه حاضر طراحی مدل پیشبینی بیماری براساس تکنیک‌های داده‌کاوی می‌باشد.*

***روش‌ها:*** *مطالعه حاضر از نوع توصیفی کاربردی بود که در سال ۱۳۹۹ انجام گردید. جامعه پژوهش تمام افرادی (۸۰۰ نفر) بود که جهت بررسی‌های تشخیصی به بیمارستان طالقانی شهرستان آبادان مراجعه کرده بودند. داده‌ها از پرونده الکترونیک بیمار طی سال‌های ۱۳۸۸-۱۳۹۸ استخراج شد. از نرم‌افزار SPSS برای تحلیل اطلاعات استفاده گردید. از روش همبستگی اسپیرمن برای شناسایی فاکتورهای مؤثر در تعیین خطر ابتلا به سرطان کولورکتال در سطح آماری P-Value≤۰/۰۵ استفاده شد. سپس با استفاده از تحلیل رگرسیون‌لجستیک دودویی و روش Enter فاکتورهای مؤثر در تعیین خطر ابتلا به سرطان کولورکتال شناسایی شدند و نهایتاً مدل رگرسیون پیشبینی خطر ابتلا به سرطان کولورکتال طراحی گردید.*

***نتایج:*** *۱۱ متغیر با استفاده از ضریب‌همبستگی اسپیرمن همبستگی معناداری را با کلاس خروجی (ابتلا و عدم ابتلا به سرطان کولورکتال) را نشان دادند. نتایج حاصل از تحلیل رگرسیون‌لجستیک با استفاده از Enter 7 متغیر شانس بالاتری نسبت به سایر متغیرها به دست آوردند. نتایج حاصل از طبقه‌بندی نمونه‌های پژوهش با استفاده از روش Forward LR نشان داد که با این مدل با داشتن میزان صحت، دقت و حساسیت به ترتیب ۹۱ درصد، ۹۳/۵ درصد و ۹۴/۵ درصد عملکرد بالایی داشته است.*

***نتیجه‌گیری:*** *مدل پیشبینی خطر مبتنی بر روش رگرسیون‌لجستیک می‌تواند در ارتقاء صحت و دقت تشخیص بیماری و پیشبینی موثر گروه‌های پرخطر به متخصصین گوارشی کمک‌کننده باشد.*

***کلیدواژه‌ها:*** *سرطان کولورکتال، داده‌کاوی، یادگیری ماشین، رگرسیون‌لجستیک، ماتریس آشفتگی.*

**نوع مقاله:** پژوهشی

**دریافت مقاله:** ۱۳۹۹/۹/۱۶ **اصلاح نهایی:** ۹۹/۱۱/۱ **پذیرش مقاله:** ۹۹/۱۲/۱۲

**ارجاع:** نوپور رئوف، کاظمی‌آرپناهی هادی، شنبه‌زاده مصطفی. طراحی یک مدل پیشبینی و ارزیابی خطر سرطان کولورکتال از طریق تکنیک داده‌کاوی مبتنی بر مدل رگرسیون‌لجستیک. مجله اطلاع‌رسانی پزشکی نوین. ۱۳۹۹؛ (۴)۶: ۱۰-۱.

**مقدمه:**

سرطان کولورکتال (سرطان کولون یا سرطان روده بزرگ) به رشد سلول‌های سرطانی در کولون، رکتوم و یا زائده آپاندیس اطلاق می‌شود که منشأ ایجاد آن ترکیبی از عوامل محیطی و ژنتیکی می‌باشند که در نهایت سلول‌های موکوسی کولونی تبدیل به سلول سرطانی می‌شوند (۱،۲). عوامل متعددی از قبیل عوامل تغذیه‌ای مانند مصرف کم غذاهای پرفیبر و میوه و سبزی، مصرف بیش از حد گوشت قرمز، مصرف الکل، سابقه دیابت ملیتوس، بیماری التهابی روده، سابقه رادیوگرافی لگن، مصرف سیگار، فعالیت کم، چاقی و سایر عوامل در بروز این بیماری مؤثرند (۳،۴).

نویسنده مسئول:
مصطفی شنبه‌زاده
گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی ایلام، ایران.
تلفن: ۹۳۰۰۸۳۳۶۹۱ ۹۸+ پست الکترونیکی: mostafa.shanbezadeh@gmail.com ORCID: 0000-0002-3419-1947



[ DOI: 10.29252/jmis.6.4.1 ]

از علائم و نشانه‌های پیش‌آگهی‌دهنده این بیماری می‌توان به مواردی همچون خونریزی از رکتال، وجود درد شکمی، تغییرات عادات روده‌ای، کم‌خونی، خونریزی از مقعد اشاره کرد (۵). این بیماری سومین عامل مرگ از میان سرطان‌ها در بین زنان و مردان در آمریکا محسوب می‌شود (۶). هر ساله ۱/۸۵ میلیون مورد جدید ابتلا به این بیماری و معادل ۱۰/۲ درصد از کل سرطان‌های بدخیم به تعداد این بیماران افزوده می‌شود (۷). در مطالعه‌ای تخمین زده شده است که در سال ۲۰۲۰ حدود ۱۴۷۹۵۰ فرد تشخیص مثبت ابتلا به سرطان کولورکتال خواهند شد که تعداد مرگ در بین آن‌ها حدود ۵۳۲۰۰ نفر خواهد بود که حدود ۱۷۹۳۰ مورد مبتلا به بیماری و ۳۶۴۰ مورد مرگ آن در افراد کم‌تر ۵۰ سال خواهد بود. طبق برآوردهای حاصل از ارزیابی پایگاه داده بین‌المللی GLOBOCAN در سال ۲۰۱۸ حدود ۲ میلیون موارد مبتلا و ۱ میلیون مرگ جدید ناشی از این بیماری در سراسر جهان رخ خواهد داد و همچنین برآورد شده است که به‌طور میانگین نرخ ابتلا به این بیماری و مرگ ناشی از آن به‌طورکلی یک‌روند افزایشی را خواهد داشت (۴). تشخیص سریع و به‌موقع سرطان کولورکتال در مراحل اولیه آن نقش بسیار حیاتی در افزایش میزان بقای ناشی از آن دارد (۸). پژوهش Bosetti و همکاران در مورد بقای سرطان کولورکتال نشان داد که تشخیص به‌موقع سرطان کولورکتال نقشی بسیار مهم در افزایش میزان بقای پنج ساله ناشی از این بیماری و کاهش میزان مرگ‌ومیر آن در کشورهای اروپایی داشته است (۹). همچنین میزان بقای پنج ساله ناشی از سرطان کولورکتال در کشور آمریکا حدود ۶۴ درصد گزارش شده بود که با تشخیص به‌موقع این بیماری در مراحل اولیه میزان بقا در این کشور به ۹۰ درصد افزایش خواهد یافت (۱۰). روش‌های غربال‌گری گوناگونی از جمله ارزیابی خون پنهان در مدفوع (FOB: Fecal Occult Blood)، ارزیابی ایمونوشیمیایی مدفوع (FIT: Fecal Immunochemical Tests) و سیگموئیدسکوپی برای تشخیص سریع این بیماری در مراحل اولیه آن وجود دارد که هر یک از آن‌ها دارای معایب و مزایایی خاص خود می‌باشند (۱۱)؛ مثلاً روش FOB نسبت به سایر روش‌های غربال‌گری ارزان‌تر می‌باشد، اما نسبت به آن‌ها از حساسیت پایین‌تری برخوردار می‌باشد (۸). بنابراین، استفاده از روش مناسب جهت تشخیص سریع این بیماری از ابعاد گوناگون بسیار مهم تلقی می‌شود. فرآیند داده‌کاوی به فرآیند کشف الگوهای ساختاریافته و اطلاعات از داده‌ها به‌منظور حل مشکلات اطلاق می‌شود که از روش‌های آماری گوناگون جهت استخراج بهترین الگوها با توجه به مجموعه داده موجود در پایگاه‌های داده استفاده می‌کنند (۱۲). الگوریتم‌های داده‌کاوی کاربرد بسیاری در حیطه پزشکی از جمله تشخیص و درمان بیماری‌های گوناگون داشته است. این تکنیک‌های آماری تاکنون در تشخیص زودرس بیماری‌ها مختلف نقشی بسیار مفید و حیاتی داشته است (۱۳-۱۵). از طرفی با توجه به تشخیص زودرس این بیماری شانس موفقیت درمان‌های گوناگون در این بیماران افزایش می‌یابد (۸). بنابراین، نقش استفاده از تکنیک‌های آماری برای داده‌کاوی حجم عظیم داده‌های مراقبتی و بهداشتی در تشخیص زودرس این بیماری و جلوگیری از پیشرفت سریع بیماری و درمان مفید به نظر می‌رسد (۱۶،۱۷).

بنابراین با توجه به میزان شیوع بالای سرطان کولورکتال و روند رو به افزایش آن در سطح جهانی از یک طرف و همچنین اهمیت بالای تشخیص زودرس این بیماری در مراحل اولیه در افزایش میزان بقای آن و همچنین معایب ذکر شده روش‌های غربال‌گری، به‌کارگیری راهکاری مناسب جهت تشخیص سریع این بیماری مفید به نظر می‌رسد، بنابراین، هدف از انجام این پژوهش، ایجاد مدل پیشبینی مناسب مبتنی بر تکنیک‌های آماری براساس مجموعه داده‌های گردآوری شده از افراد مبتلا به سرطان کولورکتال و افراد فاقد این بیماری می‌باشد تا بتوان از یک‌سو میزان بقای ناشی از این بیماری را در سطح جامعه افزایش و میزان مرگ‌ومیر آن را کاهش داد.

## مواد و روش‌ها:

این مطالعه توصیفی کاربردی در سال ۱۳۹۹ انجام گرفت. جامعه پژوهش در این مطالعه افرادی بودند که جهت بررسی‌های تشخیصی از لحاظ ابتلا و یا عدم ابتلا به بیماری سرطان کولورکتال به بیمارستان طالقانی شهرستان آبادان بین سال های ۱۳۸۸ تا ۱۳۹۸ مراجعه کرده بودند. در مجموع تعداد ۸۰۰ نمونه از افراد مبتلا و مشکوک به سرطان کولورکتال همراه با ۴۰ عامل خطر، در داخل سیستم پرونده الکترونیک بیمار در آن مراکز ذخیره شده بود. در این میان اطلاعات ۴۰ نمونه که دارای اطلاعات ناقص بودند (دارای حداقل ۷۰ درصد اطلاعات ناقص و یا فاقد اطلاعات) از فرآیند انجام پژوهش حذف گردیدند و باقیمانده آن (۷۶۰ نمونه) در این مطالعه مورد استفاده قرار گرفتند.

به‌منظور بررسی عوامل تأثیرگذار در ایجاد سرطان کولورکتال در افراد با خطر متوسط، افرادی که دارای خطر بسیار بالا در ایجاد سرطان کولورکتال بودند طبق راهنماهای بالینی از جمله برک کشف شدند (۱۸،۱۹). عواملی مانند سوابق شخصی و یا خانوادگی پولیپ و یا سرطان




[ DOI: 10.29252/jmis.6.4.1 ]

کولورکتال، سابقه شخصی بیماری التهابی روده، سابقه خانوادگی بیماری‌های ژنتیکی پولیپ و یا سرطان غیرپولیپی، سابقه خانوادگی سرطان کولورکتال در ۲ نفر از بستگان درجه ۱ یا یک نفر از بستگان درجه ۱ خانواده کمتر ۶۰ سال به‌عنوان عوامل با خطر بالا در نظر گرفته شدند و بنابراین نمونه‌های پژوهش که دارای این عوامل خطر بودند از روند انجام پژوهش حذف گردیدند و اطلاعات ۴۶۸ نفر در انجام تحلیل آماری مورد استفاده قرار گرفت.

مجموعه داده مورد استفاده در پژوهش دارای ۴۰ عامل خطر در تعیین میزان خطر ابتلا به سرطان کولورکتال در افراد مراجعه‌کننده به مرکز درمانی بودند. این عوامل به‌عنوان متغیرهای ورودی در تجزیه و تحلیل آماری مورد استفاده قرار گرفتند. کلاس خروجی مجموعه داده استفاده شده دارای دو حالت بودند که مربوط به افراد مبتلا به سرطان کولورکتال و فاقد سرطان بودند که در افراد فاقد این بیماری با عدد صفر و افراد مبتلا به این بیماری با عدد ۱ مشخص شده بودند.

به‌منظور پردازش اطلاعات موجود در سیستم پرونده الکترونیک بیمار از نرم‌افزار SPSSنسخه ۲۵ استفاده گردید. از روش همبستگی اسپیرمن برای تعیین روابط بین هر یک عوامل خطر ابتلا به سرطان کولورکتال و کلاس خروجی استفاده شد. در این روش تعیین رابط بین دو متغیر که به ضریب‌همبستگی دومتغیره (دودویی) معروف است میزان همبستگی بین دو متغیر در سطح P-Value≤۰/۰۵ بررسی می‌شود و در صورتی که میزان P-Value≤۰/۰۵ باشد نشان‌دهنده همبستگی و روابط معنادار بین دو متغیر گوناگون می‌باشد. در این پژوهش عوامل خطری که میزان همبستگی آن‌ها با کلاس خروجی پژوهش در سطح P-Value≤۰/۰۵ بود به عنوان فاکتورهای خطر مؤثر در تعیین سرطان کولورکتال در افراد عادی در نظر گرفته شدند.

برای ایجاد مدل پیشبینی خطر ابتلا به سرطان کولورکتال با توجه به تعداد حالات کلاس خروجی (حالت صفر برای افراد فاقد سرطان کولورکتال و حالت ۱ افراد مبتلا به سرطان کولورکتال)، از روش رگرسیون‌لجستیک دودویی استفاده گردید. ابتدا تمامی متغیرهای تعیین خطر ابتلا به سرطان کولورکتال که با استفاده از روش اسپیرمن در سطح P-Value≤۰/۰۵ از نظر آماری معنادار شناخته شده بودند، وارد مدل رگرسیون‌لجستیک دودویی Enter شده بودند و عوامل خطری که میزان شانس بسیار پایین داشتند و یا حد پایین اطمینان ۹۵ درصد آن صفر و یا نزدیک به صفر بودند، از مدل رگرسیون حذف گردیدند. متغیرهای با شانس بالاتر به‌عنوان عوامل خطر نهایی در تعیین خطر سرطان کولورکتال شناخته شدند. سرانجام مدل رگرسیون نهایی تعیین خطر سرطان کولورکتال به روش Forward LR براساس مهم‌ترین عوامل خطر سرطان کولورکتال و بالاترین شانس آن‌ها ایجاد گردید. در این روش تحلیل رگرسیون‌لجستیک، متغیرها به ترتیب وارد مدل رگرسیون‌لجستیک شده و با اضافه شدن هر متغیر به مدل رگرسیون میزان کارایی آن با استفاده از معیارهای مختلف ارزیابی سنجیده می‌شود. مثلا در گام اول یک متغیر وارد مدل رگرسیون شده و کارایی مدل سنجیده می‌شود و در گام هفتم، هفت متغیر وارد مدل رگرسیون شده و میزان پیشگویی مدل براساس این هفت متغیر و معیارهای مختلف ارزیابی با گام های قبلی مقایسه شده و ارزیابی می‌شود.

جهت ارزیابی عملکرد مدل رگرسیون پیشبینی خطر ابتلا به سرطان کولورکتال از ماتریس آشفتگی استفاده گردید و نهایتاً میزان دقت، صحت و حساسیت مدل سنجیده شد (+ نشان‌دهنده نمونه‌های بیمار و - نشان‌دهنده موارد غیربیمار می‌باشد) ( جدول ۱).

**جدول ۱- ماتریس آشفتگی**

| موارد مورد انتظار | موارد واقعی | |
|---|---|---|
| | + | - |
| + | True Positive (TP) | False Positive (FP) |
| - | False Negative (FN) | True Negative (TN) |

در این پژوهش معیارهای TP و TN تعداد موارد بیمار و سالم می‌باشد که به درستی توسط مدل طبقه‌بندی‌شده بودند، FP تعداد افراد سالم می‌باشد که به‌اشتباه توسط مدل بیمار تشخیص داده شده بودند و FN تعداد افراد بیماری بوده است که توسط مدل به‌عنوان سالم در نظر گرفته شده بودند.

رابطه ۱: $Precision= TP/ (TP+FP)$

رابطه ۲: $Sensitivity= TP/ (TP+FN)$

رابطه ۳: $Accuracy= (TP+TN)/ (TP+TN+FN+FP)$

## یافته‌ها:

پس از جداسازی نمونه‌های دارای عوامل پرخطر از افراد مبتلا و غیرمبتلا به سرطان کولورکتال ۴۶۸ نمونه باقی ماندند که دارای عوامل پرخطر نبودند، بنابراین به عنوان نمونه‌های متوسط‌خطر و یا کم‌خطر ابتلا به سرطان کولورکتال در نظر گرفته شدند که از این تعداد ۲۷۴ نمونه (۵۸/۵ درصد) مرتبط با افراد فاقد بیماری و ۱۹۴ نمونه (۴۱/۵ درصد) متعلق به افراد مبتلا به سرطان کولورکتال بودند. نتایج حاصل تحلیل





[ DOI: 10.29252/jmis.6.4.1 ]

همبستگی بین هر یک از عوامل خطر ابتلا به سرطان کولورکتال با کلاس خروجی (ابتلا و یا عدم ابتلا به سرطان کولورکتال) با استفاده از ضریب‌همبستگی اسپیرمن در سطح آماری مقدار احتمال (P-Value)، در جدول ۲ نشان داده شده است.

**جدول ۲- فاکتورهای مؤثر در تعیین خطر سرطان کولورکتال در سطح آماری**

| نام متغیر | نوع متغیر | P.Value | ویژگی | میزان همبستگی |
|---|---|---|---|---|
| مصرف میوه و سبزیجات (تعداد سرو شده در روز) | چندحالتی | ۰/۰۰۱> | ۲><br>۳-۲<br>٤-۳<br>٤< | -۰/۲۸۷ |
| مصرف چربی حیوانی (تعداد سرو شده در روز) (هر وعده معادل ۵۰ گرم میوه و سبزیجات مصرف‌شده در روز می‌باشد) | چندحالتی | ۰/۰۰۱> | ۱><br>۲-۱<br>۳-۲<br>۳< | ۰/۲۲۳ |
| مصرف گوشت قرمز و فرآوری شده (تعداد سرو شده در روز) (هر وعده معادل ۵۰ گرم میوه و سبزیجات مصرف‌شده در روز می‌باشد) | چندحالتی | ۰/۰۰۱> | ۱><br>۲-۱<br>۳-۲<br>۳< | ۰/۲٤٥ |
| ورزش (برحسب ساعت) | چندحالتی | ۰/۰۰۱> | ۱><br>۲-۱<br>۲< | -۰/۲۸۱ |
| سن (بر حسب سال) | چندحالتی | ۰/۰۰۵ | ٤٥><br>٦٥-٤٥<br>٦٥< | ۰/۰۹۵ |
| شاخص توده بدنی (نسبت وزن (بر حسب کیلوگرم) به قد((بر حسب متر)$^2$) | چندحالتی | ۰/۰۰۱ | ۱۸/۵><br>۲٤/۹-۱۸<br>۲۹/۹-۲۵<br>۳٤/۹-۳۰<br>۳۹/۹-۳۵<br>۵۰-٤۰<br>۵۰< | ۰/۱۲۱ |
| مصرف قرص آسپرین (نیم واحد قرص در روز) | چندحالتی | >۰/۰۰۱ | ۲><br>۳-۲<br>٤-۳<br>٤< | -۰/۲۰٦ |
| مصرف قرص آسپرین (برحسب سال) | چندحالتی | ۰/۰۰۱> | ۱><br>۲-۱<br>۳-۲<br>۳< | -۰/۱۹۵ |
| مصرف سیگار (تعداد پاکت در روز) | چندحالتی | ۰/۰۰۱> | ۱><br>۲-۱<br>۳-۲<br>۳< | ۰/۳۰۱ |
| مصرف سیگار (برحسب سال) | چندحالتی | ۰/۰۰۱> | ۱><br>۵-۱<br>۱۰-۵<br>۱۰< | ۰/۲۷۸ |
| سوابق خانوادگی | چندحالتی | ۰/۰۰۷ | ندارد<br>درجه ۳ | ۰/۱۲۹ |





[ DOI: 10.29252/jmis.6.4.1 ]

| | |
|---|---|
| درجه ۲ | |
| درجه ۱ | |

نتایج حاصل از تعیین شانس فاکتورهای مؤثر در سرطان کولورکتال با استفاده از روش Enter رگرسیون‌لجستیک دودویی در جدول ۳ نشان داده شده است. میزان شانس هفت متغیر مصرف میوه و سبزیجات، مصرف گوشت قرمز، مصرف چربی حیوانی، فعالیت فیزیکی، میزان مصرف سیگار برحسب پاکت در روز و در سال و سابقه خویشاوندی دریکی از بستگان بالای ۶۰ سال در افراد در سطح معناداری P-Value≤۰/۰۵ بود، بنابراین این هفت فاکتور نسبت به سایر متغیرها از شانس بالاتری در تعیین خطر ابتلا به سرطان کولورکتال در سطح آماری برخوردار بودند.

**جدول ۳- نسبت شانس شاخص‌های مهم تعیین‌کننده CRC**

| متغیر | P-value | نسبت شانس | نسبت شانس یا اطمینان ۹۵٪ |
|---|---|---|---|
| مصرف میوه و سبزیجات | ۰۰۰/۱ | ۱/۲۳۳ | ۰/۸۵-٤/٦٦٥ |
| مصرف گوشت قرمز | <۰/۰۰۱ | ۱/۱۱۷ | ۰/۸۸۹- ۳/۲٤۷ |
| مصرف چربی حیوانی | <۰/۰۰۱ | ۱/۱۹٦ | ۰/۹۱۲-٤/۲٥٦ |
| فعالیت بدنی (ساعت) | ۰/۰۱ | ۰/۹۹٥ | ۰/٥۸٥-۱/٤٤۱ |
| شاخص توده بدنی | ۰/۱٥ | ۰/۲٤ | ۰/۱۲-۰/٥۸۷ |
| سن | ۰/۰۸ | ۱/۰٦٦ | ۰/٥٤۱-۱/٥۷٤ |
| مصرف آسپرین (قرص در روز/۲) | ۰/۱۳ | ۰/٥۲ | ۰/۲۱٤-۰/۷٤۸ |
| آسپرین (سالیانه) | ۰/۲۲ | ۰/۲۱ | ۰/۲۷۹-۰/۸٥۷ |
| مصرف دخانیات (بسته در روز) | <۰/۰۰۱ | ۱/۷۸ | ۱/۲۲۳-٥/۲۷٤ |
| مصرف دخانیات (در سال) | <۰/۰۰۱ | ۱/٥٤٤ | ۱/۱۱٥-٤/۰۱ |
| سابقه ژنتیکی | ۰/۰۲ | ۱/٤٦ | ۰/۸۷۱-۲/۲۲۱ |

طبقه‌بندی نمونه‌های واقعی برحسب موارد پیشبینی شده حاصل از تحلیل رگرسیون‌لجستیک دودویی با روش Forward LR در هفت گام با استفاده از هفت متغیر به دست آمده با شانس بالا انجام گردیده بود. با در نظر گرفتن حد آستانه‌ای ۰/۵ در جداسازی نمونه‌هایی سرطانی و غیرسرطانی (بالاتر از ۰/۵ نمونه‌های سرطانی و کوچک‌تر از ۰/۵ نمونه‌های فاقد سرطان)، نتایج حاصل از مدل پیشبینی نشان داد که مدل رگرسیون حاصل اکثر نمونه‌های واقعی سرطان کولورکتال را در حد بالای ۰/۵ و نمونه‌های فاقد سرطانی را در حد پایین ۰/۵ خصوصاً در فراوانی تکرار کمتر از ۱۰ حاصل از مدل طبقه‌بندی کرده بودند. نتایج حاصل از طبقه‌بندی افراد مبتلا و فاقد سرطان براساس ماتریس آشفتگی در گام اول و هفتم مدل رگرسیون‌لجستیک دودویی در جدول ۴ نشان داده شده است. نتایج جدول بیانگر این است که با اضافه شدن هفت متغیر مهم و با شانس بالا به مدل رگرسیون کارایی حاصل از آن رشد چشمگیری داشته است به طوری که میزان TP آن ۶۵ عدد (۳۳/۲ درصد) و میزان TN حاصل از آن ۶۴ عدد (۲۳/۳ درصد) افزایش داشته است.

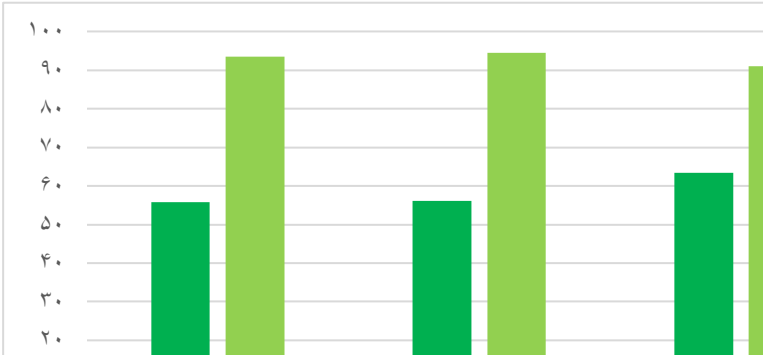
**جدول ۴- ماتریس آشفتگی طبقه‌بندی موارد در گام‌های ۱ و ۷**

| گام | TN | FN | FP | TP |
|---|---|---|---|---|
| ۱ | ۱۸۸ | ۸٥ | ۸٦ | ۱۰۹ |
| ۷ | ۲٥۲ | ۲۰ | ۲۲ | ۱۷٤ |

نتایج حاصل از مقایسه میزان دقت، حساسیت و صحت حاصل از ماتریس آشفتگی جدول ۱ حاصل از گام اول و هفتم مدل رگرسیون در نمودار ۱ نشان داده شده است. نتایج حاصل از ارزیابی عملکرد مدل براساس نمودار ۱ با استفاده از سه معیار دقت، حساسیت و صحت حاصل از ارزیابی گام اول و آخر مدل نشان داد که در گام آخر مدل نسبت با گام اول آن با اضافه شدن هفت متغیر مهم و شانس بالا، عملکرد الگوریتم به‌طور محسوسی افزایش یافت به‌طوری‌که میزان دقت از ۵۵/۸ درصد به ۹۳/۵ درصد، میزان حساسیت از ۵۶/۱ درصد به ۹۴/۵ درصد و میزان صحت مدل از ۶۳/۴ درصد به ۹۱ درصد افزایش یافت.




[ DOI: 10.29252/jmis.6.4.1 ]




**تصویر ۱- شاخص‌های عملکردی طبقه‌بندی موارد در گام‌های ۱ و ۷**

## بحث و نتیجه‌گیری:

تشخیص و پیش‌آگهی سریع و به موقع بیماری سرطان کولورکتال می‌تواند در بهبود کیفیت درمان و افزایش بقاء بیماران تاثیر بسزایی داشته باشد. سرطان کولورکتال بیماری است که شروع بسیار ساکتی داشته و وقتی علائم و نشانه‌های خود را بروز می‌دهند که بیماری در فرد بسیار پیشرفت کرده و میزان امید به زندگی بیماران کاهش می‌یابد. در سطح جهانی و حتی در کشور ایران این بیماری شیوع بسیار بالایی داشته است (۱۷). در طول سالیان بسیار به دلیل پیش‌آگهی کم در بیماران، سرطان کولورکتال را به یکی از مهم‌ترین علل مرگ‌ومیر در جهان تبدیل کرده است. بنابراین استفاده از راهکارهای کمک تشخیصی که بتواند به پزشکان در شناسایی و تشخیص به‌موقع این بیماری کمک‌کننده باشد، بسیار مفید به نظر می‌رسد (۲۰). در بسیاری از کشورها برنامه‌های تشخیص و غربالگری برای سرطان کولورکتال در افراد ارائه شده‌اند تا بتوانند در پیشگیری و به تبع آن افزایش میزان بقای بیماران مبتلا به این سرطان مؤثر واقع شوند؛ با این وجود برنامه‌های تشخیص و غربال‌گری بیماری سرطان بر روی علائم و نشانه‌های بیمار متمرکز شده‌اند، بنابراین غربالگری بیماران با استفاده از علائم و نشانه‌ها می‌تواند امید به زندگی در بیماران را با توجه به ماهیت سرطان کولورکتال تا حد قابل ملاحظه‌ای کاهش دهد (۲۱،۲۲). از این رو استفاده از برنامه‌های تشخیص و غربال‌گری سرطان کولورکتال با استفاده از ریسک فاکتورها بر روی افراد عادی جامعه و نه بیماران باعث می‌شود که میزان امید به زندگی افراد جامعه افزایش یابد و پیشگیری از این بیماری در مرحله قبل از شروع پیدایش علائم و نشانه‌ها اتفاق افتد و به تبع آن میزان مرگ‌ومیر ناشی از آن در سطح جامعه کاهش یابد (۲۳). استفاده از فناوری‌های نوین و پیشرفته از جمله هوش مصنوعی و تکنیک‌های داده‌کاوی از طریق الگوهای آماری و محاسباتی می‌تواند راهکارهای پزشکی و دانش پزشکان را شبیه‌سازی کرده و در غربال‌گری مؤثر سرطان کولورکتال در سطح جامعه مؤثر واقع شود (۲۴). از آنجایی که در حیطه علوم پزشکی در زمینه تشخیص و درمان بیماری‌ها مسائل پیچیده و مبهم به وفور مشاهده می‌شود، استفاده از این روش می‌تواند در زمینه تشخیص و درمان بیماری بسیار کمک‌کننده باشد. امروزه استفاده از فناوری‌های محاسباتی به‌روز و پیشرفته در قالب سیستم‌های یادگیری ماشین، یادگیری عمیق و داده‌کاوی داده‌های بزرگ، قابلیت‌های تحلیلی عمیق، اثربخش و غیرتهاجمی را به منظور حمایت از تصمیم‌گیری پزشکان و سیاست‌گذاران بهداشتی نسبت به روش‌های سنتی آماری، بررسی‌های بالینی و ارزیابی‌های آزمایشگاهی فراهم کرده است (۲۵،۲۶). پژوهش‌هایی در خصوص کاربرد روش رگرسیون‌لجستیک در تشخیص و پیش‌آگهی صحیح، دقیق و به موقع وضعیت های بدخیمی انجام پذیرفته که در زیر به مواردی از آن‌ها اشاره شده است.

در پژوهش Aslam و همکارانش از مدل رگرسیون‌لجستیک برای تشخیص سرطان ریه استفاده و در نهایت سیستم با میزان حساسیت و اختصاصیت به ترتیب ۹۵/۸ درصد و ۹۲/۳ درصد برای افراد سیگاری و ۹۶/۲ درصد و ۹۰/۶ درصد برای افراد غیرسیگاری قادر به تشخیص موارد سرطان بود (۲۷). در مطالعه Hsieh و همکارانش از روش رگرسیون‌لجستیک برای تشخیص سرطان پانکراس در بیماران مبتلا به دیابت استفاده و در نهایت سطح زیر نمودار (AUC) برابر با ۰/۷۲۷ قابلیت غربالگری موارد بیمار از سالم داشت (۲۸). نتایج تحقیق Rodrigues با هدف استفاده از روش رگرسیون‌لجستیک برای تشخیص بیماری سرطان پستان با حساسیت ۸۰ درصد استفاده کرد (۲۹). در بررسی Weppler و همکارانش از این مدل برای تشخیص سرطان سر و گردن با سطح زیر نمودار ۸۵ درصد استفاده کرد (۳۰). در پژوهش حاضر پس از جداسازی نمونه‌های پرخطر سرطان کولورکتال ۴۶۸ نمونه از





[ DOI: 10.29252/jmis.6.4.1 ]

موارد خطر متوسط یا خطر پایین سرطان کولورکتال به دست آمد، نتایج حاصل از تحلیل همبستگی اسپیرمن در سطح آماری P نشان داد که ۱۱ متغیر در تعیین خطر سرطان کولورکتال نقش مؤثری دارند که در این بین عوامل تغذیه‌ای مانند مصرف گوشت قرمز و فرآوری‌شده، مصرف میوه و سبزیجات، مصرف چربی‌های حیوانی و عوامل اپیدمیولوژیکی مانند مصرف سیگار با بالاترین میزان همبستگی در سطح P کمتر از ۰/۰۰۰۱ از اهمیت قابل‌توجه‌ای در تعیین خطر ابتلا به سرطان کولورکتال برخوردارند.

نتایج حاصل از تحلیل رگرسیون با روش Enter نشان داد که متغیرهای مصرف گوشت قرمز و فرآوری‌شده و مصرف چربی‌های حیوانی و مصرف سیگار شانس بالاتری نسبت به سایر متغیرها در سطح معنادار آماری برخوردار بودند و به‌عنوان مهم‌ترین متغیرهای ورودی در ایجاد مدل رگرسیون در نظر گرفته شدند. نتایج حاصل از ایجاد مدل رگرسیون‌لجستیک دودویی با استفاده از روش Forward LR و هفت متغیر با شانس بالا و در سطح معنادار آماری در هفت گام با استفاده از ماتریس طبقه‌بندی نشان داد که با اضافه شدن هفت متغیر مؤثر در تعیین خطر ابتلا به سرطان کولورکتال، تعداد موارد درست پیشبینی شده توسط مدل (TP و TN) در گام هفتم نسبت به گام اول افزایش چشمگیری داشته بود و هم‌زمان تعداد موارد به‌اشتباه طبقه‌بندی‌شده توسط مدل (FN و FP) در انتهای گام هفتم کاهش چشمگیری داشته بود، همچنین میزان صحت، دقت و حساسیت مدل در گام هفتم حاصل از تحلیل رگرسیون نسبت به گام اول افزایش چشمگیری داشت که این موضوع نشان‌دهنده قدرت پیشبینی بالای این ریسک فاکتورها در ایجاد مدل پیشبینی خطر ابتلا به سرطان کولورکتال بود. بنابراین نتایج حاصل از تحلیل مدل رگرسیون با استفاده از روش Forward LR در گام هفتم (آخر) نشان داد که مدل حاصل از عملکرد بهینه با میزان دقت ۹۳/۵ درصد، حساسیت ۹۴/۵ درصد و صحت ۹۱ درصد برخوردار بود.

مهم‌ترین محدودیت‌های پژوهش حاضر، محدود بودن تعداد نمونه‌ها در پایگاه داده انتخابی، تک مرکزی بودن پایگاه داده و نیز وجود برخی رکوردهای اطلاعاتی غیریکپارچه، ناقص، دارای خطا و موارد غیرطبیعی در فیلدهای اطلاعاتی بود. از طریق اعمال الگوریتم‌های یادگیری ماشین در پایگاه‌های داده بزرگ برگرفته از چند مرکز و نیز توجه به کمیت و کیفیت مستندسازی می‌توان قابلیت‌های الگوریتم‌ها را در تشخیص صحیح موارد بالا برد. به علاوه در پژوهش حاضر از داده‌های آزمایشگاهی به عنوان ورودی سیستم‌ها استفاده نشد که توجه به این کلاس داده عملکرد تشخیصی و پیشبینی سیستم‌ها را بهبود خواهد بخشید.

نتایج حاصل از ارزیابی سیستم تعیین خطر ابتلا به سرطان کولورکتال با استفاده از روش‌های داده‌کاوی مجهز به تکنیک‌های آماری نشان داد که این مدل پیشبینانه می‌تواند در تشخیص و غربال‌گری سرطان کولورکتال در سطح جامعه مؤثر واقع شود؛ بنابراین می‌تواند به پزشکان در تشخیص زودرس سرطان کولورکتال و کاهش خطاهای تشخیصی کمک کند و به تبع آن میزان امید به زندگی افراد افزایش و میزان مرگ‌ومیر و ناخوشی‌های ناشی از سرطان کولورکتال در سطح جامعه می‌یابد. از طرفی در کاهش هدر رفت زمان، بار اقتصادی و پیامدهای روانی و جسمی ناشی از اقدامات درمانی در بیماران مبتلا سرطان کولورکتال و به تبع آن بهبود شاخص‌های بهداشتی جوامع مؤثر واقع شود.

## تشکر و قدردانی:

نویسندگان بر خود لازم می‌دانند از همکاری مسئولان بیمارستان طالقانی آبادان و همچنین معاونت تحقیقات و فناوری دانشکده علوم پزشکی آبادان تشکر و قدردانی نمایند.

## تأییدیه اخلاقی:

این مطالعه دارای تاییدیه اخلاقی به شماره IR.ABADANUMS.REC.1399.049 از دانشکده علوم پزشکی آبادان است.

## تعارض منافع:

نویسندگان مقاله تعارض منافعی ندارند.

## سهم نویسندگان:

رئوف نوپور (نویسنده اول) تحلیل داده و آماده‌سازی نسخه اولیه مقاله ۴۰ درصد؛ هادی کاظمی‌آرپناهی (نویسنده دوم) گردآوری و تهیه پیش‌نویس مقاله ۲۰ درصد؛ مصطفی شنبه‌زاده (نویسنده سوم و مسئول) نظارت و تائید نسخه نهایی مقاله ۴۰ درصد.

## حمایت مالی:

این مقاله با حمایت مالی معاونت تحقیقات دانشکده علوم پزشکی





[ DOI: 10.29252/jmis.6.4.1 ]

آبادان انجام شده است.

## Reference


1. Binefa G, Rodríguez Moranta F, Teule À, Medina-Hayas M. Colorectal cancer: From prevention to personalized medicine. World J Gastroenterol . 2014; 20(22):6786-808. DOI: 10.3748/wjg.v20.i22.6786
2. Sameer AS. Colorectal cancer: Molecular mutations and polymorphisms. Front Oncol. 2013; 3:114. DOI: 10.3389/fonc.2013.00114
3. Johnson CM, Wei C, Ensor JE, Smolenski DJ, Amos CI, Levin B, et al. Meta-analyses of colorectal cancer risk factors. Cancer Causes Control. 2013; 24(6):1207-22. DOI: 10.1007/s10552-013-0201-5
4. Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. Prz Gastroenterol. 2019; 14(2):89-103. DOI: 10.5114/pg.2018.81072
5. John SKP, George S, Primrose JN, Fozard JBJ. Symptoms and signs in patients with colorectal cancer. Colorectal Dis. 2011; 13(1):17-25. DOI: 10.1111/j.1463-1318.2010.02221.x
6. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin. 2020; 70(3):145-64. DOI: 10.3322/caac.21601
7. Mattiuzzi C, Sanchis Gomar F, Lippi G. Concise update on colorectal cancer epidemiology. Ann Transl Med. 2019; 7(21):609. DOI:10.21037/atm.2019.07.91
8. Bhat SK, East JE. Colorectal cancer: Prevention and early diagnosis. Medicine. 2015; 43(6):295-8. DOI:10.1016/j.mpmed.2015.03.009
9. Bosetti C, Levi F, Rosato V, Bertuccio P, Lucchini F, Negri E, et al. Recent trends in colorectal cancer mortality in Europe. Int J Cancer. 2011; 129(1):180-91. DOI: 10.1002/ijc.25653
10. Moghimi-Dehkordi B, Safaee A. An overview of colorectal cancer survival rates and prognosis in Asia. World J Gastrointest Oncol. 2012; 4(4):71-5. DOI: 10.4251/wjgo.v4.i4.71
11. Hol L, Van Leerdam ME, Van Ballegooijen M, Van Vuuren AJ, Van Dekken H, Reijerink JCIY, et al. Screening for colorectal cancer: Randomised trial comparing guaiac-based and immunochemical faecal occult blood testing and flexible sigmoidoscopy. Gut. 2010; 59(01):62-8. DOI: 10.1136/gut.2009.177089
12. Contreras-Valdes A, Amezquita-Sanchez JP, Granados-Lieberman D, Valtierra-Rodriguez M. Predictive data mining techniques for fault diagnosis of electric equipment. Appl Sci. 2020; 10(3):950. DOI: 10.3390/app10030950
13. Zubi ZS, Saad RA. Using some data mining techniques for early diagnosis of lung cancer. In: Bojkovic Z, Kacprzy J, editors. Proceedings of the 10th WSEAS international conference on Artificial intelligence Knowledge Engineering and Data Bases; 2011 Feb 20. Wisconsin: World Scientific and Engineering Academy and Society; 2011.P. 32-7.
14. Kharya S. Using data mining techniques for diagnosis and prognosis of cancer disease. IJCSEIT. 2012; 2(2):55-66. DOI: 10.5121/ijcseit.2012.2206
15. Sinha A, Sahoo B, Rautaray SS, Pandey M. Predictive model prototype for the diagnosis of breast cancer using big data technology. In: Kolhe M, Tiwari S, Trivedi M, Mishra K, editors. Advances in Data and Information Sciences. Singapore: Springer; 2020. P. 455-64. DOI:10.1007/978-981-15-0694-9_43
16. Sabouri S, Esmaily H, Shahidsales S, Emadi M. Survival prediction in patients with colorectal cancer using artificial neural network and cox regression. Int J Cancer Manag. 2020; 13(1): e81161. DOI: 10.5812/ijcm.81161
17. Xie W, Xie L, Song X. The diagnostic accuracy of circulating free DNA for the detection of KRAS mutation status in colorectal cancer: A meta-analysis. Cancer Med. 2019; 8(3):1218-31. DOI: 10.1002/cam4.1989
18. Mankaney G, Sutton RA, Burke CA. Colorectal cancer screening: Choosing the right test. Cleve Clin J Med. 2019; 86(6):385-92. DOI: 10.3949/ccjm.86a.17125
19. Issa IA, Noureddine M. Colorectal cancer screening: An updated review of the available options. World J Gastroenterol. 2017; 23(28):5086-96. DOI: 10.3748/wjg.v23.i28.5086
20. Rieger AK, Mansmann UR. A bayesian scoring rule on clustered event data for familial risk assessment: An example from colorectal cancer screening. Biom J. 2018; 60(1):115-27. DOI: 10.1002/bimj.201600264
21. Pourhoseingholi MA, Kheirian S, Zali MR. Comparison of basic and ensemble






[ DOI: 10.29252/jmis.6.4.1 ]

data mining methods in predicting 5-year survival of colorectal cancer patients. Acta Inform Med. 2017; 25(4):254-8. DOI: 10.5455/aim.2017.25.254-258

22. Roberts PO, de Souza TG, Grant BM, Wanliss MG, Leake P-AE, Johnson AR, et al. Pathologic factors affecting colorectal cancer survival in a Jamaican population: The UHWI experience. J Racial Ethn Health Disparities. 2020; 7:413-20. DOI: 10.1007/s40615-019-00669-7
23. Arunkumar C, Ramakrishnan S. Prediction of cancer using customised fuzzy rough machine learning approaches. Healthc Technol Lett. 2019; 6(1):13-8. DOI: 10.1049/htl.2018.5055
24. Sha S, Du W, Parkinson A, Glasgow N. Relative importance of clinical and sociodemographic factors in association with post- operative in- hospital deaths in colorectal cancer patients in New South Wales: An artificial neural network approach. J Eval Clin Pract. 2020; 26(5):1389-98. DOI: 10.1111/jep.13318
25. Celesti A, Ruggeri A, Fazio M, Galletta A, Villari M, Romano A. Blockchain- based healthcare workflow for tele-medical laboratory in Federated Hospital IoT Clouds. Sensors. 2020; 20(9):2590. DOI: 10.3390/s20092590
26. Chamola V, Hassija V, Gupta V, Guizani M. A comprehensive review of the COVID-19 pandemic and the role of IoT, Drones, AI, Blockchain, and 5G in managing its impact. IEEE Access. 2020; 8:90225-65. DOI: 10.1109/ACCESS.2020.2992341
27. Tirzite M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer with electronic nose and logistic regression analysis. J. Breath Res. 2019; 13(1):016006.
28. Hsieh MH, Sun LM, Lin CL, Hsieh MJ, Hsu CY, Kao CH. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Cancer Manag Res. 2018; 10:6317-24. DOI: 10.2147/CMAR.S180791
29. Morais Rodrigues F, Silv́erio Machado R, Kato RB, Rodrigues DLN, Valdez Baez J, Fonseca V, et al. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. Gene. 2020; 726:144168. DOI: 10.1016/j.gene.2019.144168
30. Weppler S, Schinkel C, Kirkby C, Smith W. Lasso logistic regression to derive workflow-specific algorithm performance requirements as demonstrated for head and neck cancer deformable image registration in adaptive radiation therapy. Phys Med Biol. 2020; 65(19):195013.



[ DOI: 10.29252/jmis.6.4.1 ]

# Designing a model for predicting colorectal cancer risk based on regression-logistic data mining technique

Raoof Nopour[1] Hadi Kazemi Arpanahi[2] Mostafa Shanbehzadeh[3]*

1. MSC, Health Information Technology, Department of Health Information Technology and Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran. ORCID: 0000-0003-3770-2375
2. Department of Health Information Technology, Abadan Faculty of Medical Sciences, Abadan, Iran.
3. Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran.

(Received 6 Dec, 2020 Accepted 2 Mar, 2021)

## Original Article

## Abstract

**Aim:** Using machine learning for the early detection of this disease has an important role in improving disease indicators. Therefore, this study aims to design a disease prediction model based on data mining techniques to help in early diagnosis and provide evidence-based services.

**Methods:** This is an applied descriptive studyconducted in 2020. The study population was all patients (800 people) referred to Taleghani Hospital in Abadan for diagnostic tests. The data were derived from the electronic records of during 2009-2010. The Spearman correlation method was used to identify the effective factors in determining the risk of CRC. Then, Binary Logistic Regression (BLR) analysis and Enter method, effective factors in determining the risk of CRC were identified. Finally, the regression prediction model for CRC was developed. SPSS 17 was used to analyze statistical data. P-value $\geq 0.05$ was considered significant.

**Results:** Eleven variables using the Spearman correlation coefficient showed a significant correlation with the output class (with and without colorectal cancer). The results of regression-logistic analysis using Enter 7 variables obtained a higher chance than other variables. The results of classifying the research samples using the Forward LR method showed that with this model, accuracy, precision, and sensitivity (91%, 93.5%, and 94.5%, respectively) had high performance.

**Conclusion:** Designing a risk prediction model based on logistic regression plays an important role in rapid, accurate, and timely screening of patients in improving the quality of care and increasing the life expectancy of patients. The proposed model in the present study can help gastroenterologist to improve the diagnosis accuracy, precision, and effective prediction of high-risk groups.

**Key Words:** Colorectal Cancer, Data Mining, Machine Learning, Logistic Regression, Confusion Matrix.

**Citation:** Nopour R, Kazemi Arpanahi H, Shanbehzadeh M. Designing a predicting and evaluation model for colorectal cancer by data mining techniques based on the logistic regression model. J Mod Med Info Sci. 2021; 6(4):1-10.

*Correspondence:*
*Mostafa Shanbehzadeh*
*Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran.*
*Tel: + 98 9300833691 Email: mostafa.shanbezadeh@gmail.com ORCID: 0000-0002-3419-1947*





[ DOI: 10.29252/jmis.6.4.1 ]