مجله علوم پزشکی رازی دوره ۲۵، شماره ۹، آذر ۱۳۹۷ http://rjms.iums.ac.ir

مقاله پژوهشی

آنالیز داده‌های مربوط به بیماران مبتلا به سرطان روده بزرگ با استفاده از تکنیک‌های داده کاوی
مطالعه موردی: بیماران مرکز تحقیقات کولورکتال بیمارستان شهید فقیهی شیراز

الهام پروین نیا: گروه مهندسی کامپیوتر، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران (*نویسنده مسئول) parvinnia@iaushiraz.ac.ir
میثم محمدی: گروه مهندسی کامپیوتر، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران
علی محمد بنان زاده: مرکز تحقیقات کولورکتال دانشگاه علوم پزشکی شیراز، شیراز، ایران
سیده پرستو خیامی: بخش مهندسی کامپیوتر و فن آوری اطلاعات، دانشگاه شیراز، شیراز، ایران

چکیده

زمینه و هدف: روند رو به رشد سرطان روده بزرگ در سال‌های اخیر، لزوم اتکا به شیوه‌های مطمئن و جدید را برای شناسایی و کنترل این بیماری بیشتر آشکار می‌کند. داده کاوی یکی از این روش‌هاست که از مهم‌ترین کاربردهای آن، کشف الگوهای پنهان مابین داده‌های بیماران در پایگاه داده‌های بزرگ است. در این مطالعه، به بررسی و کشف الگوهای ناشناخته در یک مجموعه داده واقعی سرطان روده بزرگ پرداخته می‌شود.

روش کار: در این تحقیق مجموعه داده‌های مربوط به ۴۰۰ بیمار سرطان کولورکتال شامل ۴۲ ویژگی مورد بررسی قرار گرفته است. این اطلاعات از طریق مرکز تحقیقات کولورکتال دانشگاه علوم پزشکی شیراز بین سال‌های ۱۳۸۷ تا ۱۳۹۵ جمع‌آوری شده است. پس از انجام مراحل پیش پردازش، از طریق الگوریتم Fp-Growth روابط پنهان بین ویژگی‌های این داده‌ها کشف شده است.

یافته‌ها: با استفاده از الگوریتم فوق الذکر و کشف ارتباط میان بعضی از ویژگی‌ها، قوانینی حاصل شد که به پیشنهاد پزشک متخصص و اهمیت ویژگی‌ها، این قوانین در هفت گروه مورد بررسی قرار گرفته‌اند.

نتیجه‌گیری: نتایج حاصل از بررسی قوانین نشان می‌دهد که مرحله پاتالوژیک و سن بیمار دارای اثر معنادار در نرخ بقاء بیماران داشته‌اند. همچنین درصد ابتلای زنان و مردان به سرطان رکتال بیش از کلون می‌باشد و جنسیت در بقای بیمار تأثیری ندارد. از نتایج دیگر حاصل از بررسی این دیتاست می‌توان به عدم وجود رابطه معنادار بین مرحله پاتالوژیک بیمار و اطلاعات دموگرافیک اشاره کرد.

تعارض منافع: گزارش نشده است.
منبع حمایت کننده: گزارش نشده است.

کلیدواژه‌ها
داده کاوی،
قوانین انجمنی،
سرطان کلون،
سرطان رکتال،
کشف روابط پنهان

تاریخ دریافت: ۹۷/۴/۲۳
تاریخ پذیرش: ۹۷/۶/۱۴

شیوه استناد به این مقاله:
Parvinnia E, Mohammadi M, Bananzadeh AM, Khayami SP. Analysis of data on patients with colon cancer using the data mining techniques- Case study: Patients at colorectal research center of Shaheed Faghihi hospital in Shiraz. Razi J Med Sci.2018;25(9):47-56.

Razi Journal of Medical Sciences.2018;25(9):47-57. http://rjms.iums.ac.ir

Original Article

Analysis of data on patients with colon cancer using the data mining techniques
Case study: Patients at colorectal research center of Shaheed Faghihi hospital in Shiraz

Elham Parvinnia, Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran (*Corresponding author) parvinnia@iaushiraz.ac.ir
Meysam Mohammadi, Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran
Ali Mohammad Bananzadeh, Colorectal Research Center, Shiraz University of Medical Science, Shiraz, Iran
Seyedeh Parastoo Khayami, Department of Computer Engineering and Information Technology, Shiraz University, Shiraz, Iran

Abstract

Background: In recent years the growing trend of colon cancer has revealed that we need some safe and new methods to detect and control this disease. Data mining is one of these methods, one of its most important applications is the discovery of hidden patterns between data in a large database. In this study, we explore and discover unknown patterns in a real colon cancer data set.
Methods: In this study, the information of 400 colorectal cancer patients, with 42 cfeature has been studied.This information was collected through the Colorectal Research Center, Shiraz University of Medical Sciences, between 2008 and 2016. After performing the data set preprocessing, the hidden relationships between the features of this data are discovered through the Fp-Growth algorithm.
Results: Ater using this algorithm and discovering the relationship between some of the features, some rules have been developed. Based on the suggestion of the specialist and the importance of the features, the rules have been studied in seven groups.
Conclusion: The results of the review of the laws indicate that the pathologic stage and the age of the patient had a significant effect on the survival rate of the patients.
Also, the percentage of men and women with rectal cancer is greater than that of the clone, and the sex does not affect the survival of the patient.
Other findings from the review of this data can be the lack of a meaningful relationship between the patient's pathologic stage and the demographic information.

Conflicts of interest: None
Funding: None.

Keywords
Data mining, Association rules,
Colon cancer,
Rectal cancer,
Discover hidden relationship

Received: 14/06/2018
Accepted: 06/10/2018

Cite this article as:
Parvinnia E, Mohammadi M, Bananzadeh AM, Khayami SP. Analysis of data on patients with colon cancer using the data mining techniques- Case study: Patients at colorectal research center of Shaheed Faghihi hospital in Shiraz. Razi J Med Sci.2018;25(9):47-56.

مقدمه

در جهان امروز با توجه به پیشرفت‌های به دست آمده در جمع‌آوری و قابلیت ذخیره‌سازی داده‌ها در بسیاری از علوم، محققان با حجم انبوهی از اطلاعات روبه‌رو شده‌اند. در حال حاضر داده کاوی مهم‌ترین روش برای بهره‌وری مؤثر و سریع از داده‌های حجیم است. از این رو می‌توان با استفاده از آن جهت شناسایی الگوها، ارتباط عناصر مختلف در پایگاه داده‌ها، جهت کشف دانش نهفته در داده‌ها و تبدیل آن‌ها به اطلاعات مبادرت ورزید (۱).

یکی از زمینه‌هایی که نیازمند استفاده از این ابزارها جهت تحلیل داده‌های وسیع و مدل‌سازی پیشگویانه با روش‌های محاسباتی جدید است، علم پزشکی می‌باشد (۲). محققین تلاش می‌کنند تا از علم داده کاوی برای به دست آوردن اطلاعات و روابط مفید بین عوامل خطرزا در بیماری‌ها استفاده کنند (۳).

یکی از بیماری‌هایی که امروزه گریبان گیر بسیاری شده و یکی از علل اصلی مرگ ومیر در سراسر جهان به شمار می‌آید، سرطان است. در اغلب کشورها به‌ویژه کشورهای توسعه‌یافته، سرطان دومین علت مرگ‌ومیر بعد از بیماری‌های قلبی است (۴ و ۵).

بر اساس آمار منتشر شده توسط سازمان بهداشت جهانی ۸,۲ میلیون مرگ‌ومیر در سال ۲۰۱۲ مربوط به این بیماری بوده است که از این میان سرطان روده بزرگ (Colorectal) با ۶۹۴۰۰۰ مرگ‌ومیر در رده چهارم از این نوع بیماری قرار داشته است (۶).

عواملی ازجمله سن، جنس، تأهل، تفاوت‌های قومی و نژادی می‌توانند زمینه‌ساز سرطان کولورکتال باشند (۴ و ۷). عوامل خطر برای سرطان کولورکتال به عنوان یک سرطان واحد، ممکن است با عوامل خطر اختصاصی برای سرطان‌های کولون و رکتال به صورت جداگانه، متفاوت باشند (۴ و ۸).

از طرف دیگر سرطان‌های روده بزرگ در ناحیه رکتوم دارای ویژگی‌های بیولوژیک، روش‌های درمانی، الگوی عود و میزان بقای متفاوتی از سرطان‌های ناحیه کولون می‌باشند (۹ و ۱۰).

همچنین عوامل خطر دیگری همچون مکان جغرافیایی و سبک و عوامل بالینی و پاتولوژیکی نظیر شاخص توده‌ی بدنی، درجه و اندازه‌ی تومور، میزان وسعت نفوذ دیواره‌ی روده، متاستاز دور، متاستاز غدد لنفی و مرحله‌ی پاتولوژیک تومور نیز بر میزان بقای بیماران مبتلا به سرطان کولورکتال نقش دارند (۱۱ و ۱۲).

در مطالعه انجام شده توسط برازنده و همکاران (۵) به این موضوع پرداخته‌اند که می‌توان با کمک تکنیک‌های داده کاوی، مدل‌هایی را برای تشخیص شیوه زندگی افراد از لحاظ میزان خطر ابتلا به سرطان روده بزرگ ارائه داد. احمدی و همکاران (۱۲) مطالعه‌ای با هدف تعیین بقا، خطر نسبی مرگ و عوامل مؤثر بر آن در بیماران مبتلا به سرطان کولورکتال انجام داده‌اند. در این مطالعه نابرابری نسبت خطر ناشی از سرطان کولورکتال در بین قومیت‌های ایرانی مورد بررسی قرار داده شده است. در مقایسه بین اقوام ایرانی، قوم کرد کم‌ترین و قوم لر بیشترین زمان بقا را در سرطان کولورکتال داشته‌اند.

در مطالعه دیگری روشنایی و همکاران (۴) کوشیده‌اند با تعیین عوامل مؤثر بر بیماران سرطان کولورکتال به بررسی مدت بقای بیماران بپردازند. در این مطالعه آن‌ها تلاش کرده‌اند تا با استفاده از مدل پارامتری مناسب عوامل مؤثر بر پیش‌آگهی و بقای بیماران مبتلا به سرطان کولورکتال را تعیین کنند.

در مطالعه جعفر آبادی و همکاران (۱۱) عوامل تشخیصی مؤثر در بقای سرطان کولورکتال، به عنوان یکی از مهم‌ترین مشکلات و تهدیدهای بهداشت عمومی در کشور مطرح شده است. هدف از این مطالعه تحلیل بقا برای تعیین عوامل خطر بروز سرطان‌های کولون و رکتال، در بیماران متاستازی بوده است.

در مطالعه محمد آخوند و همکاران (۷) نشان داده شده است که از جمله پارامترهایی که بقای بیماران را تحت تأثیر قرار می‌دهند می‌توان به وسعت تهاجم تومور به دیواره‌ی روده، متاستاز به گره‌های لنفاوی مجاور و متاستاز تومور به ارگان‌های دیگر اشاره کرد. در این تحقیق به بررسی تأثیر برخی عوامل بالینی و پاتولوژیک بر بقای بیماران مبتلا به سرطان کولون و رکتوم در ایران و عوامل مؤثر پیش‌آگهی دهنده در بیماران مبتلا به سرطان کولورکتال پرداخته شده است.

در تحقیق مقیمی و همکاران (۱۳) تجزیه و تحلیل توصیفی داده‌ها روی اطلاعات دموگرافیک و بالینی

صورت گرفته و منحنی بقای با استفاده از روش کاپلان – میر محاسبه شده است. بر اساس تجزیه و تحلیل تک متغیره متغیرهای شاخص توده بدنی، وضعیت تأهل، مرحله پاتولوژیک تومور (Stage)، وضعیت نفوذ به دیواره روده، متاستاز تومور و مرحله پاتالوژیک بر بقا بیمار مؤثر شناخته شده‌اند.

ویلز و همکاران (۱۴) معتقدند استراتژی‌های غربالگری جهت پیش‌بینی سرطان روده بزرگ با اینکه در کاهش مرگ‌ومیر ناشی از این بیماری مؤثر است اما مستلزم صرف هزینه بسیار بالا و در مواردی به بروز عوارض جدی مثل پارگی روده منجر می‌شود. به همین دلیل ایجاد الگوهایی جهت شناسایی بیماران پرخطر برای غربالگری می‌تواند به کاهش این هزینه‌ها منجر شود. هدف این تحقیق ارائه مدل‌های پیش‌بینی با توجه به سبک زندگی افراد با کمک علم داده کاوی به منظور تسریع غربالگری افراد پرخطر برای کاهش بار مالی و همچنین کاهش آسیب‌های ناخواسته می‌باشد.

در مطالعه انجام گرفته بر روی بیماری، کیانی و همکاران (۱۵) سعی کرده‌اند به کشف الگوهای پنهان با استفاده از قوانین انجمنی و الگوریتم Apriori بپردازند. در مطالعه دیگری افضلی و همکاران به معرفی بهترین مدل هوشمند مبتنی بر داده کاوی برای پیش‌بینی و تشخیص سرطان کبد در مراحل اولیه پرداخته‌اند. در این مقاله 6 مدل ماشین یادگیری از نظر دقت، حساسیت، ویژگی و سطح زیر نمودار با یکدیگر مقایسه شده و مدل دسته‌بندی VFI (Voting Feature Interval) بیشترین میزان دقت دسته‌بندی ویژگی و سطح زیر منحنی را به خود اختصاص داده است.

محمودی و همکاران (۱۶) به بررسی علل و عوامل تأثیرگذار در بروز بیماری سرطان معده با استفاده از پیاده‌سازی الگوریتم Apriori پرداخته‌اند. کیانی و آتشی (۱۷) در مطالعه دیگر درصدد برآمده‌اند یک مدل پیش‌بینی مبتنی بر دسته‌بندی داده‌ها برای پیش‌بینی عود مجدد سرطان سینه با استفاده از الگوریتم EM (Expectation-Maximization) ارائه دهند و در پایان یک مدل پیش‌آگهی عود مجدد سرطان سینه با به‌کارگیری درخت j48 ارائه شده است.

در یک تحقیق (۱۸) عوامل ریسک برای سه نوع بیماری شامل بیماری قلبی، سرطان سینه و هپاتیت به وسیله قوانین انجمنی نادر بررسی شده است. قوانین نادر با در نظر گرفتن کل داده‌ها به صورت یکجا تولید می‌شوند.

در تحقیق دیگری (۱۹) که در مرکز سرطان شاه حسین در شهر عمان کشور اردن انجام شده است با استفاده از قوانین انجمنی یک طبقه بند پیشرفته برای تشخیص سرطان سینه طراحی شده است. قوانین ایجاد شده ویژگی‌های مهم در تشخیص بیماری را نیز شناسایی می‌نمایند.

در مقاله کیم و همکاران (۲۰) با استفاده از قوانین انجمنی ژن‌های مربوط به بیماری شناسایی شده‌اند.

ژن‌های بیماری به وسیله آنالیز ترم‌های MeSH و استخراج اثرگذاری ژن‌ها بر یکدیگر به وسیله قوانین انجمنی تشخیص داده می‌شوند.

روش کار

قوانین انجمنی (Association Rules): قوانین انجمنی یکی از تکنیک‌های اصلی داده کاوی است و تقریباً مهم‌ترین شکل از کشف و استخراج الگوها در سیستم‌های یادگیری غیر هدایت شده می‌باشد. این روش روابط و وابستگی‌های متقابل بین مجموعه بزرگی از اقلام داده‌ای را نشان می‌دهند. الگوریتم‌های زیادی برای کشف قوانین انجمنی تاکنون ارائه شده‌اند که معروف‌ترین الگوریتم‌ها برای کشف قوانین انجمنی الگوریتم Appriori، Fp-Growth هستند.

این روش‌ها دارای دو مرحله اصلی هستند.

الف) شناسایی الگوهای پرتکرار

ب) استخراج قوانین با استفاده از الگوهای پرتکرار

روش Appriori در شناسایی الگوهای پرتکرار در مقایسه با Fp-Growth با اسکن‌های متوالی مجموعه داده، هزینه اضافی به سیستم تحمیل می‌کند. روش Fp-Growth مشکل یافتن الگوهای مکرر طولانی را به جستجوی بازگشتی الگوهای کوتاه‌تر در یک پایگاه داده کوچک‌تر تبدیل می‌کند. این روش به صورت قابل‌ملاحظه‌ای هزینه‌های جستجو را کاهش می‌دهد.

پارامترهای ارزیابی قوانین انجمنی: فرض کنید قانونی به نام R داریم که به شکل R:**A=>B** می‌باشد که در آن A و B زیرمجموعه‌ای از اشیاء می‌باشند. پارامترهای زیر جهت ارزیابی قوانین استخراج شده از الگوهای پرتکرار

به کار می‌روند.

پشتیبان: از نسبت تعداد تراکنش‌هایی که در آن اشیاء A و B هر دو حضور دارند، به کل تعداد رکوردها، پشتیبان (Support) به دست می‌آید که بر اساس رابطه ۱ تعریف می‌شود:

$$\text{Support (A)}=\frac{A \cup B \subseteq |T|}{|T|}$$

پشتیبان دارای مقداری عددی بین صفر و یک می‌باشد و هر چه این میزان بیشتر باشد، نشان می‌دهد که این دو شیء بیشتر با هم در ارتباط هستند. کاربر می‌تواند با مشخص کردن یک آستانه برای این معیار، مجموعه مکرر را به دست آورد که پشتیبان آن‌ها بیشتر از مقدار آستانه باشد.

اطمینان: برای تعیین ارزش قانون از معیار اطمینان (Confidence) و بر اساس رابطه زیر استفاده می‌شود.

$$Conf\ (A \rightarrow B) = \frac{SUP(A \cup B)}{SUP(A)}$$

معیار اطمینان نیز مقداری عددی بین صفر و یک می‌باشد که هر چه این عدد بزرگ‌تر باشد بر کیفیت قانون افزوده خواهد شد.

Lift: از معیارهای دیگر قوانین انجمنی می‌توان به معیار Lift اشاره کرد که این معیار میزان استقلال میان اشیاء A و B را نشان می‌دهد و بر اساس رابطه زیر محاسبه می‌شود.

$$Lift\ (A \rightarrow B) = \frac{Conf(A \rightarrow B)}{SUP(B)}$$

مقدار lift می‌تواند مقدار عددی بین صفر تا بی نهایت باشد. چنانچه این معیار از عدد یک کمتر باشد، نشان‌دهنده این است که A و B با یکدیگر رابطه منفی دارند. هر چه مقدار این معیار بیشتر از عدد یک باشد، نشان‌دهنده این است که A اطلاعات بیشتری درباره B فراهم می‌کند که در این حالت جذابیت قانون A=>B بالاتر ارزیابی می‌شود. ترکیب این معیار به همراه Support و Confidence جزء بهترین روش‌های کاوش قوانین انجمنی است.

Conviction: مشکل معیار lift حساس بودن به تعداد نمونه‌های مجموعه داده، به‌ویژه برای مجموعه تراکنش‌های کوچک می‌باشد. از این رو معیار Conviction برای جبران این نقص معرفی شده است. محدوده قابل تعریف برای این معیار در حوزه ۰/۵ تا بی نهایت قرار می‌گیرد که هر چه این مقدار بیشتر باشد، نشان‌دهنده این است که آن قانون جذاب‌تر می‌باشد. مقدار این معیار برای دلالت‌های منطقی یعنی در جایی که Confidence قانون یک می‌باشد برابر با بی نهایت است و چنانچه A و B مستقل از هم باشند، برابر با عدد یک خواهد بود.

$$Conv\ (A \rightarrow B) = \frac{1 - SUP(B)}{1 - Conf(A \rightarrow B)}$$

جمع‌اوری داده: در این پژوهش داده‌های مربوط به ۴۰۰ بیمار سرطان کولورکتال مراجعه‌کننده بین سال‌های ۱۳۸۷ تا ۱۳۹۵ به بخش انکولوژی بیمارستان شهید فقیهی شیراز جمع‌آوری و هر ساله وضعیت بیمار پس از درمان از طریق تماس تلفنی پیگیری و ثبت شده است. از پرونده بیماران ۲۳ ویژگی جهت یافتن الگوهای موجود در بین آن‌ها استخراج شد.

از ۴۰۰ بیمار مراجعه کنند ۲۴۱ نفر مبتلا به سرطان رکتال ۱۵۱ نفر مبتلا به سرطان کلون و تعداد ۸ نفر نیز به کولورکتال مبتلا بوده‌اند. ۲۳ ویژگی مربوط به بیماران در سه مرحله مجزا جمع‌آوری شده است.

۱. مرحله تشخیص بیماری

۲. مرحله درمان و عمل جراحی و پیگیری‌های بعد از درمان

۳. مرحله پاتولوژی

هر یک از مراحل جداگانه بررسی می‌شود.

ویژگی‌های مربوط به مرحله تشخیص: در بدو ورود بیمار به مرکز تحقیقات، اطلاعات اولیه از وضعیت عمومی بیمار ثبت می‌گردد. به این ویژگی‌ها اطلاعات دموگرافیک گفته می‌شود. نمونه این ویژگی‌ها در جدول ۱ آورده شده و از قبیل نوع سرطان، سن، جنسیت، شاخص توده بدنی(Body Mass Index-BMI)، وضعیت تأهل ازدواج خانوادگی، حاملگی، سقط جنین، وجود سرطان در دیگر اعضای خانواده و وضعیت سیگار، قلیان و... است.

همان‌گونه که مشاهده می‌شود بعضی از ویژگی‌ها به صورت دسته‌بندی شده هستند؛ مانند نوع بیماری یا جنسیت. درصورتی‌که برخی از ویژگی‌ها با یک عدد پیوسته بیان شده‌اند مانند سن یا شاخص توده بدن (BMI). در مرحله پیش پردازش هر یک از این موارد نیاز به تغییر شکل خواهند داشت.

ویژگی‌های مربوط به عمل جراحی و پیگیری‌های بعدی: ویژگی‌های مربوط به درمان پیشنهادی و نوع عمل جراحی در جدول ۲ خلاصه شده است. در این جدول ویژگی مربوط به پیگیری بعد از عمل و وضعیت زنده ماندن یا عود مجدد بیماری در سال‌های (یک تا پنج بعد از درمان) می‌باشد.

ویژگی‌های مربوط به پاتولوژی: ویژگی‌های مربوط به این مرحله شامل تاریخ تشخیص بیماری، محل تومور، اندازه تومور، وضعیت تومور و مرحله پاتولوژیک بیماری (I,II,III,IV). این ویژگی‌ها در جدول ۳ آورده شده است.

آماده‌سازی داده‌ها (Preprocessing): شامل حذف داده‌های تکراری و گسسته سازی (Discretization) است. در اولین مرحله پس از جمع‌آوری داده‌ها، جهت آماده‌سازی داده وجود رکوردهای تکراری در دیتابیس چک گردید که نشان داد رکورد تکراری وجود ندارد. با

جدول ۱- ویژگی های دموگرافیک

ویژگی	مقادیر ثبت شده	تعداد (درصد)
	Colon	(٪۳۸) ۱۸۱

جدول ۳- اطلاعات پاتولوژی

ویژگی	مقادیر ثبت شده	تعداد (درصد)
تاریخ شروع بیماری	۱۳۹۴-۱۳۶۴	---
تاریخ تشخیص	۱۳۹۴-۱۳۷۵	---
متاستازهای کبدی	بله	---
	خیر	
اندازه تومور	۱۲-۰	---
مرحله پاتولوژیک	I	۹۰ (٪۲۲/۵)
	II	۱۱۱ (٪۲۸)
	III	۹۵ (٪۲۴)
	IV	۱۰ (٪۲/۵)
	ناشناخته	۹۴ (٪۲۳)

جدول ۴- بازه‌های گسسته‌سازی شده ویژگی سن

حالت	بازه اول	بازه دوم	بازه سوم
۱	Under 40	Between 40 and 65	Over 65
۲	Under 40	Between 40 and 70	Over 70
۳	Under 45	Between 45 and 65	Over 65
۴	Under 45	Between 45 and 70	Over 70
۵	Under 50	Over 50	

جدول ۵- بازه‌های گسسته‌سازی‌شده ویژگی شاخص توده بدنی

حالت	بازه ۱	بازه ۲	بازه ۳	بازه ۴	بازه ۵	بازه ۶	بازه ۷	بازه ۸	بازه ۹	بازه ۱۰
۱	کمتر از ۲۵	بیشتر از ۲۵								
۲	کمتر از ۳۰	بیشتر از ۳۰								
۳	کمتر از ۲۰	۳۰-۲۰	۴۰-۳۰	بیشتر از ۴۰						
۴	کمتر از ۱۸/۵	۲۵-۱۸/۵	۳۰-۲۵	۳۵-۳۰	۴۰-۳۵	بیشتر از ۴۰				
۵	کمتر از ۱۸/۵	۲۲/۹-۱۸/۵	۲۴/۹-۲۳	۲۷/۴-۲۵	۲۹/۹-۲۷/۵	۳۲/۴-۳۰	۳۴/۹-۳۲/۵	۳۷/۴-۳۵	۳۹/۹-۳۷/۵	بیشتر از ۴۰

توجه به وجود ویژگی‌هایی از نوع پیوسته مانند سن و شاخص توده بدنی نیاز به گسسته سازی این ویژگی‌ها جهت بررسی بازه‌های مختلف محسوس است.

ویژگی سن به پیشنهاد پزشک متخصص به سه بازه تقسیم شده است. بازه‌هایی که در نظر گرفته شد به پیشنهاد پزشک متخصص در ۵ حالت مختلف بررسی گردید. جزییات آن در جدول ۴ آورده شده است.

همچنین برای شاخص توده بدن به پیشنهاد پزشک متخصص ۵ حالت مختلف با بازه‌های متفاوت مورد بررسی قرار گرفت. در جدول ۵ این حالت‌ها خلاصه شده است.

با توجه به پراکندگی انواع سرطان در ویژگی سرطان در سایر اعضای خانواده، ویژگی Other Family که نشان‌دهنده وجود یا عدم وجود سرطان در دیگر اعضای خانواده است و با Yes، No مقداردهی می‌شود، ایجاد شد تا در صورت وجود رابطه در بروز سرطان در خویشاوندان و بروز سرطان در شخص بیمار این مورد قابل بررسی باشد.

یکی دیگر از ویژگی‌های ایجاد شده جهت کاهش اثرات مقادیر گم شده در ویژگی وضعیت بیمار در سال‌های پس از عمل، ویژگی the latest status of Patient است که نشان‌دهنده آخرین وضعیت ثبت شده از بیمار در مجموع پنج سال می‌باشد که علاوه بر بررسی وضعیت بیمار در سال‌های یک تا پنج سال این ویژگی نیز مورد بررسی قرار گرفت.

یافته‌ها

این پژوهش با هدف یافتن روابط معنی‌دار بین ویژگی‌های مختلف موجود در دیتاست مرکز تحقیقات کولورکتال دانشگاه علوم پزشکی شیراز و بر اساس الگوریتم FP-Growth انجام گردید. با استفاده از این الگوریتم قوانینی به صورت زیر تعریف می‌شوند.

If Permission => Conclusion

که در آن قسمت permission از ترکیب عطفی شروط روی مقادیر ویژگی‌ها و قسمت Conclusion نتیجه قانون بر اساس مقادیر ویژگی‌هاست. این قوانین رابطه بین ویژگی‌ها و همبستگی آن‌ها را بیان می‌کند. برای همه قوانین، معیارهای Support Confidence، Lift، Conviction اندازه‌گیری شد. قوانینی که Lift، Conviction قابل قبولی (بالای عدد یک) داشتند در

جدول ۶- قوانین استخراج شده

۱. تاثیر BMI بر شانس بقای بیمار						
	Permission	Conclusion	Support	Confidence	Lift	Conviction
۱	BMI = 18.5-25	the status of Patient = Alive with disease	۰.۳	۰.۶۹	۱.۲۳	۱.۴۱
۲	BMI = 32.5 - 34.9	the status of Patient = Alive with disease	۰.۱۲	۰.۶۳	۱.۱۱	۱.۱۶
۲. تاثیر جنسیت بر نوع سرطان						
	Permission	Conclusion	Support	Confidence	Lift	Conviction
۳	Gender = Male	Disease = Rectal cancer	۰.۳۶	۰.۶۵	۱.۰۸	۱.۱۴
۴	Gender = Male	Disease = Colon cancer	۰.۱۹	۰.۳۴	۱.۰۱	۱.۰۰
۵	Gender = Female	Disease = Rectal cancer	۰.۲۵	۰.۵۴	۱.۱۰	۱.۰۶
۶	Gender = Female	Disease = Colon cancer	۰.۱۸	۰.۳۹	۱.۰۳	۱.۰۲
۳. تاثیر نوع بیماری بر جنسیت						
	Permission	Conclusion	Support	Confidence	Lift	Conviction
۷	Disease = Rectal cancer	Gender = Male	۰.۳۶	۰.۵۹	۱.۰۸	۱.۱۰
۸	Disease = Rectal cancer	Gender = Female	۰.۲۵	۰.۴۱	۰.۹۱	۱.۰۳
۹	Disease = Colon cancer	Gender = Male	۰.۱۸	۰.۴۹	۱.۱۰	۱.۰۹
۱۰	Disease = Colon cancer	Gender = Female	۰.۱۹	۰.۵۱	۱.۱۳	۱.۱۲

جدول ۶- ادامه

۴. تاثیر جنسیت در بقا از سرطان رکتال و کلون

	Permission	Conclusion	Support	Confidence	Lift	Conviction
۱۱	Gender = Male	the status of Patient = Alive with disease	۰.۳۱	۰.۵۷	۱.۰۰	۱.۰۰
۱۲	Gender = Female	the status of Patient = Alive with disease	۰.۲۵	۰.۵۷	۱.۰۰	۱.۰۰

۵. تاثیر نوع بیماری بر سن افراد

	Permission	Conclusion	Support	Confidence	Lift	Conviction
۱۳	Disease = Rectal cancer	Dis Age = Old	۰.۳۳	۰.۵۵	۰.۹۸	۰.۹۸
۱۴	Disease = Rectal cancer	Dis Age = Middle-Aged	۰.۱۸	۰.۲۹	۱.۷۹	۱.۰۵
۱۵	Disease = Rectal cancer	Dis Age = Young	۰.۰۹	۰.۱۶	۰.۵۷	۱.۰۰
۱۶	Disease =Colon cancer	Dis Age = Old	۰.۲۲	۰.۵۸	۱.۰۴	۱.۰۵
۱۷	Disease =Colon cancer	Dis Age = Middle-Aged	۰.۱۰	۰.۲۶	۰.۹۳	۰.۹۷
۱۸	Disease =Colon cancer	Dis Age = Young	۰.۰۶	۰.۱۶	۱.۰۰	۱.۰۰

۶. تاثیر مرحله پاتالوژیک بیماری در بقای بیمار

	Permission	Conclusion	Support	Confidence	Lift	Conviction
۱۹	Staging = I	the status of Patient = Alive without disease	۰.۱۵	۰.۶۶	۱.۱۷	۱.۲۹
۲۰	Staging = II	the status of Patient = Alive without disease	۰.۱۶	۰.۵۹	۱.۰۵	۱.۰۶
۲۱	Staging = III	the status of Patient = Alive without disease	۰.۱۴	۰.۵۵	۰.۹۸	۰.۹۸
۲۲	Staging = IV	the status of Patient = Alive without disease	۰.۰۰	۰.۰۰	۰.۰۰	۰.۴۴

۷. تاثیر نوع بیماری بر مرحله پاتولوژیک بیماری

	Permission	Conclusion	Support	Confidence	Lift	Conviction
۲۳	Disease = Rectal cancer	Staging = I	۰.۱۷	۰.۲۷	۱.۲۱	۱.۰۶
۲۴	Disease = Rectal cancer	Staging = II	۰.۱۶	۰.۲۶	۰.۹۴	۰.۹۸
۲۵	Disease = Rectal cancer	Staging = III	۰.۱۲	۰.۱۹	۰.۸۰	۰.۹۴
۲۶	Disease = Rectal cancer	Staging = IIII	۰.۰۱	۰.۰۱	۰.۵۰	۰.۹۸
۲۷	Disease = Colon cancer	Staging = I	۰.۰۵	۰.۱۴	۰.۶۲	۰.۹۰
۲۸	Disease = Colon cancer	Staging = II	۰.۱۲	۰.۳۰	۱.۰۸	۱.۰۳
۲۹	Disease = Colon cancer	Staging = III	۰.۱۲	۰.۳۲	۱.۳۵	۱.۱۲
۳۰	Disease = Colon cancer	Staging = IIII	۰.۰۲	۰.۰۴	۱.۶۰	۱.۰۷

جدول زیر آورده شدند. برای آنکه دامنه قوانین گسترده شود کمینه پشتیبانی (Minimum support) برابر با ۱۰ درصد در نظر گرفته شده است. بر همین اساس ارزش قوانین با توجه به مقدار سطح اطمینان (Confidence) بررسی شده است.

پس از بررسی‌های انجام شده و به پیشنهاد پزشک متخصص و اهمیت ویژگی‌ها، قوانین به دست آمده در هفت گروه مورد بررسی قرار گرفته‌اند. این هفت گروه و قوانین مربوطه آن در جدول ۶ آورده شده است.

بر اساس آزمایش‌های انجام گرفته بر روی دیتاست موجود، نتایج نشان می‌دهد که هیچ رابطه معناداری بین داده‌های دموگرافیک و مرحله پاتولوژیک بیماری وجود ندارد. در این بررسی ویژگی‌های دموگرافیک به دو صورت تک به تک و یا مجموعه‌های مختلف در نظر گرفته شده و تأثیر آن بر ویژگی مرحله پاتولوژیک بیماری بررسی گردید؛ اما نتایج حاصل دارای مقادیر قابل قبول متغیرهای ارزیابی نبودند.

بحث و نتیجه‌گیری

قوانین به دست آمده در قسمت یافته‌های تحقیق مورد ارزیابی پزشک متخصص قرار گرفت و نتایج زیر حاصل گردید:

تأثیر BMI بر شانس بقای بیمار: بر اساس قوانین ۱ و ۲ و در بررسی انجام شده بر روی BMI افراد، آن‌هایی که در بازه ۱۸/۵ تا ۲۵ و بازه ۳۲/۵ تا ۳۵ قرار داشته‌اند به ترتیب با ۶۹٪ و ۶۳٪ از شانس بیشتری برای بقا بدون عود مجدد بیماری برخوردار بوده‌اند.

تأثیر جنسیت بر نوع سرطان: بر اساس قوانین ۳ و ۴ از بین مردان مراجعه‌کننده به این مرکز احتمال ابتلای مردان به سرطان رکتال ۶۵٪ می‌باشد درحالی‌که احتمال ابتلای آن‌ها به کلون ۳۴٪ و مابقی به کولورکتال مبتلا هستند.

همچنین قوانین ۵ و ۶ نشان می‌دهد ۵۴٪ زنان به رکتال و ۳۴٪ به کلون و مابقی به کولورکتال مبتلا هستند. در نتیجه ابتلای زنان و مردان در این دیتاست به سرطان رکتال بیش از کلون می‌باشد.

تأثیر نوع بیماری بر جنسیت: بر اساس قوانین ۷ و ۸ از مجموع زنان و مردان مبتلا به رکتال، مردان با ۵۹٪ سهم بیشتری نسبت به زنان با ۴۱٪ در ابتلا به این بیماری دارند.

درحالی‌که بر اساس قوانین ۹ و ۱۰ زنان با ۵۱٪ سهم بیشتری در مقایسه با مردان با ۴۹٪ در ابتلا به کلون دارند.

درصد ابتلای مردان به رکتال بیش زنان است درحالی‌که در سرطان کلون درصد ابتلای زنان بیشتر از مردان می‌باشد.

تأثیر نوع بیماری بر جنسیت: بر اساس قوانین ۱۱ و ۱۲ مردان با ۶۶٪ و زنان با ۶۷٪ از احتمال بقای تقریباً مساوی در ۵ سال اول برخوردارند. در نتیجه می‌توان گفت جنسیت در بقای بیمار تأثیری ندارد.

تأثیر نوع بیماری بر سن افراد: بر اساس قوانین،۱۴ و ۱۵ و بر اساس رده‌های سنی (Old بالای ۶۵ سال، Middle-Aged بین ۴۰ تا ۶۵ سال و Young زیر ۴۰ سال) در نظر گرفته شده، افراد مسن با ۵۵٪ و میان سال با ۲۹٪ و جوان ۱۶٪ به ترتیب در رده‌های مختلف ابتلا به سرطان رکتال هستند.

و همچنین بر اساس قوانین ۱۶، ۱۷ و ۱۸ و بر اساس رده‌های سنی (Old بالای ۶۵ سال، Middle-Aged بین ۴۰ تا ۶۵ سال و Young زیر ۴۰ سال) در نظر گرفته شده افراد مسن با ۵۸٪ و میان سال با ۲۶٪ و جوان ۱۶٪ به ترتیب در رده‌های مختلف ابتلا به سرطان کلون قرار دارند که می‌توان نتیجه گرفت سن افراد در ابتلای آن‌ها به سرطان کلون و رکتال نقش مهمی ایفا می‌کند.

تأثیر مرحله پاتالوژیک بیماری در بقای بیمار: قوانین ۱۹، ۲۰، ۲۱ و ۲۲ نشان‌دهنده تأثیر مرحله پاتولوژیک در بقای بیمار است. بر اساس این قوانین بیمارانی که در stage یک قرار دارند ۶۶٪، بیماران stage دوم ۵۹٪، بیماران stage سوم ۵۵٪ و بیماران stage چهارم ۰٪ احتمال بقا در پنج سال اول ابتلا به این بیماری را دارند.

تأثیر نوع بیماری بر مرحله پاتولوژیک بیماری: قوانین ۲۳، ۲۴، ۲۵ و ۲۶ نشان می‌دهد از بیماران مبتلا به سرطان رکتال ۲۷٪ در stage یک، ۲۶٪ stage دوم، ۱۹٪ Stage سوم و ۱٪ در stage چهارم قرار دارند و مابقی افراد مبتلا به سرطان رکتال با شیمی‌درمانی در هیچ stage قرار نمی‌گیرند.

قوانین ۲۷، ۲۸، ۲۹ و ۳۰ که از بیماران مبتلا به سرطان کلون ۱۴٪ در stage یک، ۳۰٪ stage دوم، ۳۲٪ Stage سوم و ۴٪ در stage چهارم قرار دارند و مابقی افراد مبتلا به سرطان کلون با شیمی‌درمانی در هیچ stage قرار نمی‌گیرند.

همچنین قوانین ۲۳ تا ۳۰ نشان می‌دهد تشخیص سرطان کلون در stage های بالاتری نسبت به رکتال انجام می‌شود بدین معنی که این بیماری دیرتر خود را بروز می‌دهد.

References

1. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011 Jun 9.
2. Milovic B. Prediction and decision making in health care using data mining. Int j pub health sci; 2012 Dec 1. 1(2):69-78.
3. Hassanzadeh M, Ebrahimi SA. Data Mining Algorithms for Medical Sciences. Iran J Med Inform; 2013 Jun 15. 2(2):23-6. (Persian)
4. Roshanaei G, Komijani A, Sadighi A, Faradmal J. Prediction of survival in patients with colorectal cancer referred to the Hamadan MRI center using of Weibull parameter model and determination of its

5. risk factors during 2005-2013. Arak Med Uni J; 2014. 16(11). (Persian)

6. Barazandeh I, Gholamian MR, Talaiezadeh A, Pourhoseingholi MA. A Domain-Driven Classification Model to Early Detection of Individuals Having High Risk to Develop Colorectal Cancer. J Health Biomed Inform; 2015. 2(2):59-75. (Persian)

7. World Health Organization 2015 Available from:http://www.who.int/cardiovascular_diseases/resources/atlas/en/

8. Akhoond MR, Kazemnejad A, Hajizadeh E, GAnbary Motlagh A, Zali MR. Comparison of influential factors affecting survival of patients with colon and rectum cancer using competing risks model. Koomesh; 2011 Jan 15. 12(2):119-28. (Persian)

9. Liang H, Wang XN, Wang BG, Pan Y, Liu N, Wang DC, et al. Prognostic factors of young patients with colon cancer after surgery. World J Gastroenterol; 2006 Mar 7. 12(9):1458.

10. Roncucci L, Fante R, Losi L, Di Gregorio C, Micheli A, Benatti P, et al. Survival for colon and rectal cancer in a population-based cancer registry. Eur J Cancer; 1996 Feb 1. 32(2):295-302.

11. Liang S, Carlin BP, Gelfand AE. Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. Ann Appl Stat; 2008 Oct 8. 3(3):943.

12. Jafarabadi MA, Mohammadi SM, Hajizadeh E, Fatemi SR. An evaluation of 5-year survival of metastatic colon and rectal cancer patients using cumulative incidence models. Koomesh; 2013. 14(2). (Persian)

13. Ahmadi A, Mobasheri M, Hashemi Nazari SS. Survival time and relative risk of death in patients with colorectal cancer in an Iranian population: a cohort study. J Mazandaran Uni Med Sci; 2014 Apr 15. 24(111):2-8. (Persian)

14. Moghimi B, Safaee A, Zalee M. [Evaluation of survival and prognostic factors in patients with colorectal cancer]. J Med Sci; 2008. 16(1). (Persian)

15. 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 Jan 1. 27(1):42-55.

16. Atashi A, Kiani B, Abbasi E, Nazeri N. Discovery of hidden patterns in breast cancer patients data using data mining to examine the data with a real data set. Koomesh; 2015. 56-64. (Persian)

17. Mahmoodi SA, Mirzaei K, Mahmoodi SM. Using association rules for the detection of risk factors in gastric cancer. J Health Biomed Inform; 2015 Mar 15. 1(2):95-103.

18. Kiani B, Atashi A. A prognostic model based on data mining techniques to predict breast cancer recurrence. J Health Biomed Inform; 2014. 1(1):26-31.

19. Borah A, Nath B. Identifying risk factors for adverse diseases using dynamic rare association rule mining. Expert Sys Appl; 2018. 113:233-63.

20. Alwidian Hammo BH, Obeid N. WCBA: Weighted classification based on association rules algorithm for breast cancer disease. Appl Soft Comput; 2018. 62:536-49.

21. Kim J, Bang C, Hwang H, Kim D, Park C, Park S. IMA: Identifying disease-related genes using MeSH terms and association rules. J Biomed Inform; 2017. 76:110-23.