مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره دوم شماره سوم. ص ۱۵۹-۱۴۹، ۱۳۹۴

بررسی ویژگی‌های بیماران مبتلا به سل با استفاده از روش خوشه‌بندی K-Means

فرزاد فیروزی جهانتیغ[۱]، حکیمه عامری[۲]*

• **دریافت مقاله:** ۹۴/۷/۲۲ • **پذیرش مقاله:** ۹۴/۸/۲۵

مقدمه: به گزارش سازمان سلامت جهانی، بیماری سل بیشترین عامل مرگ و میر در بیماری‌های عفونی است. با توجه به بالا بودن درصد افراد مبتلا به سل و تعداد زیاد مرگ و میر در بین این بیماران، این تحقیق با هدف دسته بندی و پیدا کردن ارتباط بین ویژگی‌های بالینی و دموگرافیک بیماران مختلف انجام شده است.

روش: این پژوهش مطالعه‌ای توصیفی، تحلیلی بوده که به روی ۶۰۰ بیمار مرکز تحقیقات سل بیمارستان مسیح دانشوری انجام شده است. برای انجام دسته‌بندی و تعیین شاخص‌های مشترک بین بیماران از الگوریتم‌های داده کاوی خوشه‌بندی K-Means و قوانین باهم آیی Apriori به کمک نرم افزار SPSS Clementine نسخه ۱۴ استفاده شده است.

نتایج: به کمک شاخص دان، تعداد ۳خوشه به عنوان خوشه بهینه انتخاب شده‌اند. عوامل مشترک بین خوشه‌ها به تفصیل در بخش نتایج آورده شده است. با توجه به ویژگی‌های هر خوشه، می‌توان بیماران را بر اساس میزان تأثیرگذاری عوامل مختلف بر روی آن‌ها دسته بندی کرد.

نتیجه‌گیری: با توجه به نتایج حاصل از این مطالعه، مهم‌ترین عوامل شناسایی شده با استفاده از خوشه‌بندی عبارت بودند از: هموگلوبین، سن، جنسیت، مصرف سیگار، مصرف الکل و کراتینین. همچنین با توجه به قوانین با هم‌آیی، بیشترین ارتباط بین سرفه، کاهش وزن و سرعت رسوب گلبول‌های قرمز یافت شده است.

کلید واژه‌ها: سل، خوشه بندی، قوانین باهم آیی، داده کاوی

• **ارجاع:** فیروزی تیغ جهان فرزاد، عامری حکیمه. بررسی ویژگی‌های بیماران مبتلا به سل با استفاده از روش خوشه‌بندی K-Means. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۴؛ ۲(۳): ۱۵۹-۱۴۹.

۱. دکتری مهندسی صنایع، استادیار، گروه مهندسی صنایع، دانشگاه سیستان و بلوچستان، زاهدان، ایران.

۲. کارشناس ارشد فناوری اطلاعات، تجارت الکترونیک، دانشکده مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی تهران، تهران، ایران.

* **نویسنده مسؤول:** تهران، دانشگاه خواجه نصیرالدین طوسی تهران، دانشکده مهندسی صنایع

• **شماره تماس:** ۰۹۲۱۴۷۰۶۹۹۳ • **Email:** ha.amery@gmail.com

مقدمه

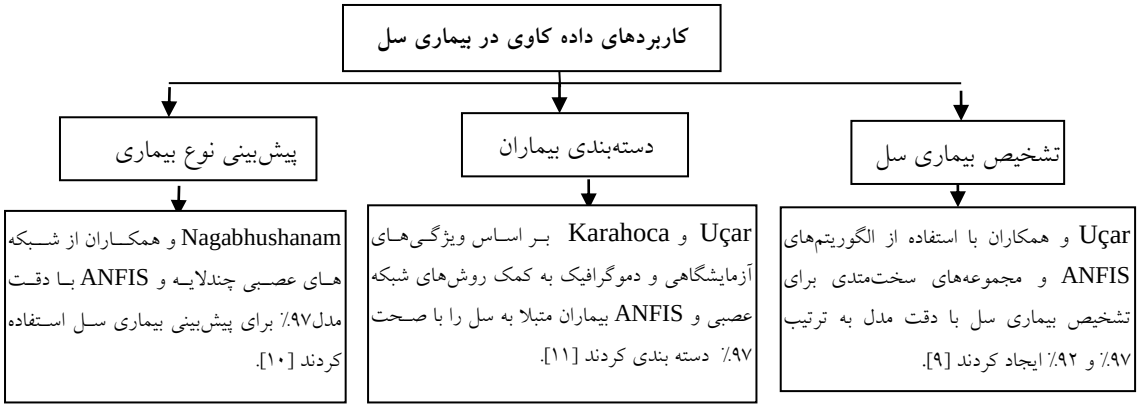
سل یکی از قدیمی‌ترین بیماری‌های انسان است که با بالاترین درجه مرگ و میر در میان بیماری‌های عفونی، همچنان توجه دنیا را به خود جلب کرده است. میزان مرگ و میر ناشی از سل از سال ۱۹۹۰ کاهش ۴۱٪ داشته و هدف رسیدن به مقدار جهانی کاهش ۵۰٪ تا سال ۲۰۱۵ می‌باشد. با این وجود بار جهانی سل همچنان بزرگ است. در سال ۲۰۱۱ به طور تخمینی ۸/۷٪ میلیون مورد جدید سل (۱۳٪ عفونتی همراه HIV) وجود داشت و ۱/۴ میلیون نفر بر اثر بیماری سل جان خود را از دست دادند [۱]. تقریباً یک سوم جمعیت جهان (حدود ۲ میلیارد نفر) آلوده به میکروب سل هستند و در معرض ابتلا به این بیماری مهلک قرار دارند. به گزارش سازمان بهداشت جهانی سالانه ۹ میلیون نفر به سل فعال مبتلا شده و حدود ۱/۵ تا ۲ میلیون نفر بر اثر این بیماری جان خود را از دست می‌دهند. [۲]. بیماری سل که بزرگترین علت مرگ ناشی از بیماری‌های عفونی تک عاملی است (حتی بیشتر از ایدز، مالاریا و سرخک)، دارای مرتبه دهم در بار جهانی بیماری‌ها است و پیش‌بینی می‌شود تا سال ۲۰۲۰ همچنان جایگاه کنونی خود را حفظ کند (و یا تا رتبه هفتم بالا رود) [۲].

روش اصلی برای تشخیص سل تست توبرکولین (tuberculin) و میکروسکوپی sputum - smear و رادیولوژی از سینه می‌باشد. متأسفانه این روش‌ها زمان بر و با کارآیی پایین هستند. مهم‌ترین دلیل مرگ و میر ناشی از سل، تشخیص دیرهنگام و اشتباه در روش درمانی بیماران است. از مهم‌ترین دلایل تشخیص اشتباه می‌توان به تکیه بیش از حد بر رادیولوژی در تشخیص و عدم استفاده از آزمایش میکروسکوپی خلط، تجویز رژیم‌های درمانی اشتباه و فاقد مقبولیت علمی، نقصان در پایش بیماران در طی درمان دارویی، نقصان در پیگیری و بررسی افراد در تماس با بیماران، شناسایی شده است [۱]. استفاده از روش‌های داده کاوی می‌تواند گامی در جهت پیش‌بینی بهتر و دسته بندی بیماران بر اساس ویژگی‌های تأثیر گذار باشد [۳].

داده کاوی از رشته‌های جدیدی است که با به کارگیری و استفاده از داده‌های آماری، به استخراج اطلاعات و الگوهای مفید می‌پردازد. داده کاوی نشان دهنده یک پیشرفت قابل توجه در انواع ابزار تحلیلی در دسترس است و به عنوان یک روش معتبر، حساس و قابل اعتماد برای کشف الگوها و روابط بین آن‌ها، در نظر گرفته می‌شود [۴]. یکی از زمینه‌هایی که می‌توان این دانش را به نحو مؤثری استفاده کرد و نتایج قابل توجهی به دست آورد، داده‌های پزشکی است. افزایش دقت تشخیص، کاهش هزینه‌ها و کاهش منابع انسانی به عنوان مزایای معرفی داده کاوی در تجزیه و تحلیل پزشکی توسط خواجویی و اعتمادی و جایالاکشی ثابت شده است [۶، ۵]. برخی کاربردهای داده کاوی در پزشکی عبارت است از: بررسی میزان تأثیر دارو بر بیماری، شناسایی عوارض جانبی داروها، تعیین نوع درمان، تجزیه و تحلیل داده‌های موجود در پرونده الکترونیک سلامت (Electronic Health Records) EHR، تشخیص و پیش‌بینی انواع بیماری‌ها مانند سرطان، تحلیل عکس‌های پزشکی مانند ماموگرافی، التراسونیک، اشعه X و (Magnetic Resonance Imaging) MRI، ارائه مدل‌های توصیفی بروی داده‌های پزشکی، کنترل عفونت بیمارستان، بهره‌برداری از خدمات سلامت [۷]. به عنوان مثال شریف خانی و همکاران با استفاده از الگوریتم‌های C.5.0، CHAID و شبکه عصبی مصنوعی بیشترین عوامل تأثیرگذار بر پوکی استخوان را شناسایی و معرفی کرده‌اند. با استفاده از داده کاوی و روش‌های آن ویژگی‌های تأثیرگذار بر این بیماری شناسایی شده‌اند. به کمک درخت تصمیم ایجاد شده، قوانینی استخراج شده‌اند که می‌تواند به عنوان الگویی برای پیش‌بینی وضعیت بیماران از آن‌ها استفاده کرد. دقت مدل‌های ساخته شده با استفاده از الگوریتم‌های C.5.0، CHAID و شبکه عصبی مصنوعی با یکدیگر مقایسه شده‌اند. نتایج این مقایسه نشان می‌دهد هریک از این الگوریتم‌ها در پیش‌بینی گروهی از افراد بهتر عمل می‌کند [۸].

مطالعات زیادی در زمینه بیماری‌های ریوی و به خصوص سل با استفاده از تکنیک‌های داده کاوی انجام شده است. مطالعات انجام شده را می‌توان به ۳ گروه کلی که در شکل ۱ نمایش داده شده است، تقسیم کرد [۱۱-۹].

از مهم‌ترین علل شکست جهانی در کنترل بیماری سل تشخیص دیر هنگام و اشتباه در روش درمانی بیماران است. هدف از انجام این تحقیق بررسی ویژگی‌های بیماران مبتلا به سل برای دست‌یابی به دانشی جدید در حوزه تشخیص و شناسایی این افراد می‌باشد. با امید به این که با داشتن الگوهای مناسب، تشخیص بیماری سل دقیق‌تر و سریع‌تر انجام بگیرد و در نتیجه آن تعداد بیماران با سیل‌های سل مقاوم به چند دارو (MDR-TB) کاهش پیدا کند.



شکل ۱: کاربردهای داده کاوی در سل

روش

این پژوهش مطالعه‌ای توصیفی، تحلیلی بوده که به صورت مقطعی در سال۱۳۹۴ انجام گرفت. جامعه پژوهش بیماران مبتلا به سل در مرکز بیماری‌های سل بیمارستان مسیح دانشوری تهران می‌باشد. پرونده مربوط به ۶۰۰ بیمار این مرکز مورد بررسی قرار گرفت. ویژگی‌های مورد بررسی در دو دسته ویژگی های دموگرافیک و بالینی انجام شده‌اند. ویژگی‌های دموگرافیک عبارت بودند از: سن، جنسیت، شغل، سابقه استفاده از سیگار، تعداد سال‌های استفاده از سیگار، سابقه استفاده از الکل و ویژگی‌های بالینی عبارت بودند از: سرفه مزمن، خلط آغشته به خون، کاهش وزن، تعریق شبانه، تب، بیمار HIV+، تعداد آزمایش‌های اسمیر خلط مثبت بیمار، تعداد گلبول‌های سفید در یک میلی‌لیتر خون، مقدار هموگلوبین موجود در خون، تعداد پلاکت‌ها در هر میلی‌لیتر مکعب خون، سرعت رسوب گلبول‌های قرمز خون، قند خون ناشتا، کراتینین، آلبومین، سابقه تماس با بیمار مبتلا به سل و سایر بیماری‌های بیمار. میانگین سن بیماران ۵۳ سال و ۵۰ درصد آن‌ها مرد و بقیه زن هستند.۸۳ درصد بیماران دارای سابقه تماس با بیماران مبتلا به سل بوده‌اند.

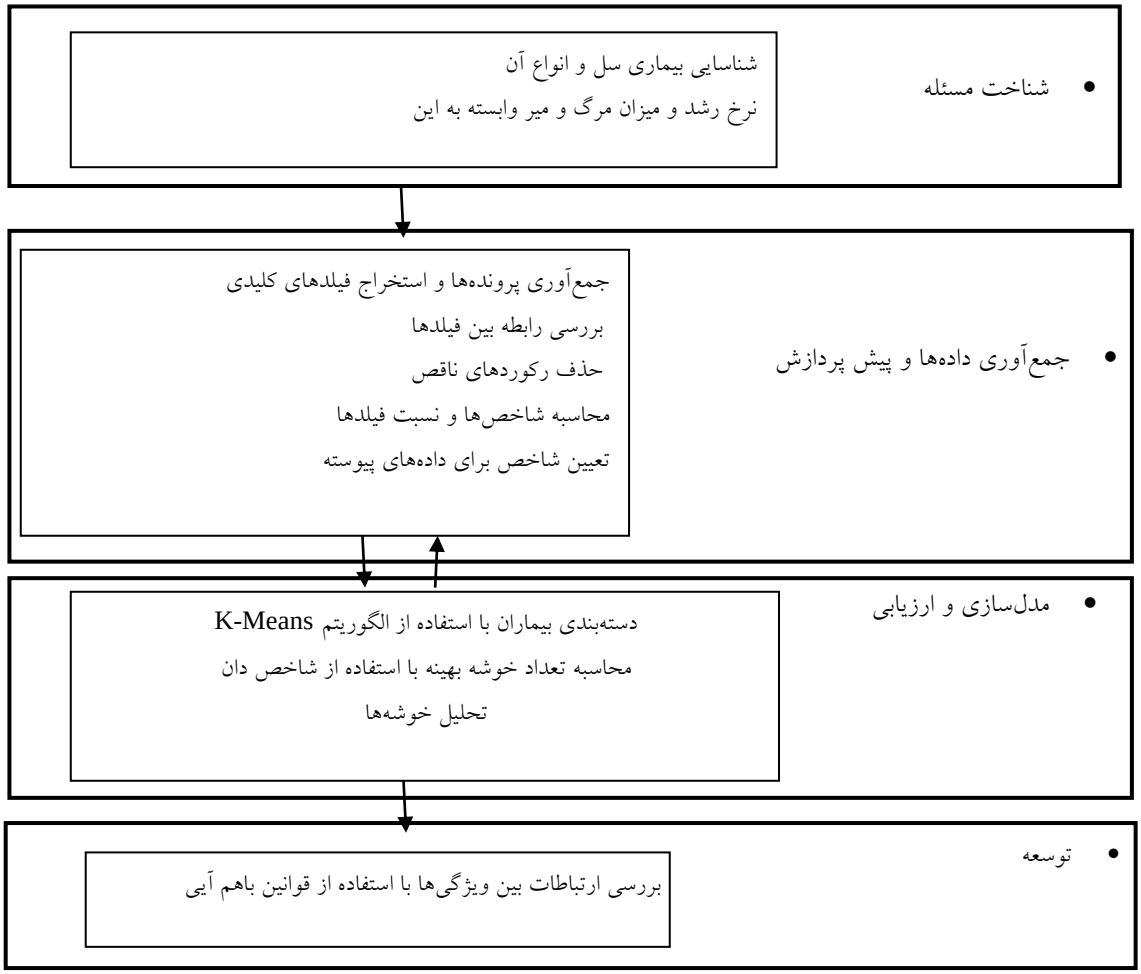
روش مورد استفاده برای دسته‌بندی بیماران، روش خوشه‌بندی K-Means از تکنیک‌های داده‌کاوی می‌باشد. خوشه‌بندی، یک روش یادگیری بدون نظارت است که روی دسته‌های از قبل تعریف شده و یا ویژگی خاصی به عنوان هدف تکیه ندارد و نمونه‌های مشابه با هم در یک حجم داده را گروه‌بندی می‌کند. داده‌ها برای انجام عمل خوشه‌بندی وارد مدل K-Means می‌شوند. در واقع خوشه بندی شکلی از یادگیری به وسیله مشاهدات است [۱۲]. با استفاده از شاخص دان از بین خوشه های متفاوت که به عنوان ورودی به مدل داده‌ایم، خوشه‌بندی بهینه را به دست می‌آوریم. این شاخص برای داشتن خوشه‌های متمرکز با مرزهای مشخص مورد استفاده قرار می‌گیرد.

برای پیدا کردن ارتباط مؤثر بین ویژگی‌های مختلف بیماران از الگوریتم Apriori، از مجموعه الگوریتم‌های پرکاربرد قوانین باهم آیی استفاده شده است. برای این منظور از نرم افزار SPSS Clementine نسخه ۱۴ استفاده شده است.

برای داشتن یک داده‌کاوی مؤثر علاوه بر نیاز به داده‌های مرتبط، باید از یک فرآیند و روش داده‌کاوی مناسب نیز بهره‌مند شویم. روشی که کلیه مراحل داده‌کاوی اعم از جمع آوری داده، آماده‌سازی داده، مدل‌سازی و ارزیابی را در برگیرد [۱۲]. بدین منظور بر مبنای روش CRISP (Cross Industry Process For Data Mining)، فرآیند انجام کار صورت گرفته است. مراحل این فرآیند در شکل ۲ قابل مشاهده است.

در مهم‌ترین گام تحقیق (آماده‌سازی داده‌ها یا پیش پردازش داده‌ها) به بررسی پرونده بیماران پرداخته شده است. در جهان واقعی، داده همیشه کامل نیست و در مورد اطلاعات پزشکی، این موضوع همیشه درست است. برای حذف تعدادی از تناقض‌ها و داده‌های ناقص در ارتباط با داده‌ها از پردازش داده استفاده شده است. بسیاری از تکنیک‌های پردازش داده توسط Chen و Astebro و Han و همکاران [۱۲،۱۳] ارائه شده‌اند. Witten و همکاران در [۱۴] ثابت کردند که حذف عاقلانه،

یک روش کارآمد به جای جایگزین کردن ارزش‌ها با تکنیک‌هایی مانند میانگین، انتساب تصادفی، انتساب رگرسیون و مدل‌های بیزی است. طهماسبی و همکاران نشان دادند که جایگزینی مقادیر مفقود به کمک تکنیک‌های مختلف مثل شبکه‌های عصبی و یا نزدیکترین همسایگی برتری نسبت به هم ندارند. در این تحقیق مواردی را که ارزش صفر برای ویژگی‌های آزمایشگاهی و دموگرافیک داشتند، حذف شده‌اند [۱۵].



شکل ۲: مدل پیشنهادی

نتایج

متغیرهای تعداد گلبول‌های سفید (White Blood Cell) WBC، مقدار هموگلوبین موجود در خون (Hemoglobin) (HB)، تعداد پلاکت‌ها (PLT)، سرعت رسوب گلبول‌های قرمز(Erythrocyte Sedimentation Rate) ESR، میزان قند خون ناشتا (Fasting Blood Sugar) FBS، مقدار کراتینین (Creatinine) و آلبومین (Albumin) که دارای ارزش‌های عددی به صورت دامنه‌ای بودند براساس منابع و سایت‌های معتبر علمی [۱۹-۱۶] و تأیید پزشک متخصص به صورت کدبندی شده استفاده شده‌اند. در نتیجه پس از پالایش داده‌ها به رکوردهایی با مشخصات جدول ۱ رسیدیم.

جدول ۱: داده‌ها و ارزش‌های مربوطه پس از پیش پردازش

ویژگی	علامت استفاده شده	ارزش	نوع
سن بر اساس سال	Age	کمی گسسته	فاصله‌ای
شغل	Job	طبقه‌ای گسسته	اسمی
جنسیت	Sex	زن=۰ مرد =۱	اسمی
سرفه مزمن	cough	بلی=۱ خیر =۰	اسمی
خلط آغشته به خون	Blood-tinged sputum	بلی=۱ خیر =۰	اسمی
کاهش وزن	lose weight	بلی=۱ خیر =۰	اسمی
تعریق شبانه	Sweating at night	بلی=۱ خیر =۰	اسمی
تب	Fever	بلی=۱ خیر =۰	اسمی
سابقه تماس با بیمار مبتلا به سل	Contact history with TB	بلی=۱ خیر =۰	اسمی
سابقه استفاده از سیگار	smoking	بلی=۱ خیر =۰	اسمی
تعداد سال‌های استفاده از سیگار	duration of smoking	کمی گسسته	فاصله‌ای
سابقه استفاده از الکل	alcoholic	بلی=۱ خیر =۰	اسمی
بیمار HIV+	HIV	بلی=۱ خیر =۰	اسمی
تعداد آزمایش‌های اسمیر خلط مثبت بیمار	BK	عددی صحیح بین ۰ تا ۳	عددی فاصله ای
تعداد گلبول‌های سفید در یک میلی لیتر خون	WBC	Leucopenia: 1=<3999; Normal: 2=4000-11000; Leukocytosis: 3=>11001;	عددی فاصله‌ای
مقدار هموگلوبین موجود در خون	HB	Anemia :Male: <14 & female: <12=1; Normal: Male : 14-16 & female: 12-14 =2; Hemochromatosis: Male: >16 & female: >14=3;	عددی فاصله‌ای
تعداد پلاکت‌ها در هر میلی‌لیتر مکعب خون	PLT	Thrombocytopenia: 1= <150000; Normal: 2=150000-450000; Thrombocytosis:3=>450000;	عددی فاصله‌ای
سرعت رسوب گلبول‌های قرمز خون	ESR (Erythrocyte Sedimentation rate)	Normal: Children: ESR 3-13=1; Normal: Male <50: ESR up to 15 & >50 :ESR up to 20=1; Inflammation: Male <50: ESR more than15 & >50 :ESR more than 20=2; Normal: Female <50: ESR up to 20 & >50 :ESR up to 30=1; Inflammation: Female <50: ESR more than20 & >50 :ESR more than 30=2;	عددی فاصله‌ای
قند خون ناشتا	Fbs(Fasting blood sugar in mg/dl)	Normal: 1=<100; Prediabetic: 2=101-126; Diabetic:3= more than>= 127	فاصله‌ای
کراتینین	creatinine	Male: <0.5 &female: <0.4=1; Normal: Male: 0.5-1.2 & female: 0.4-1.1=2; Azotemia: Male: >1.3 & female: >1.2=3; Normal: Children :0-0.7=2;	فاصله‌ای عددی
آلبومین	albumin	1= <3.5; Normal: 2=3.5-5.5; 3=>5.5;	فاصله‌ای عددی
سایر بیماری‌های بیمار	Other diseases	طبقه‌ای گسسته	اسمی

داده‌ها و ارزش‌های مربوطه پس از پیش پردازش و پس از خوشه‌بندی به کمک الگوریتم K-Means، شاخص دان برای پیدا کردن تعداد خوشه بهینه به صورت زیر محاسبه می شود:

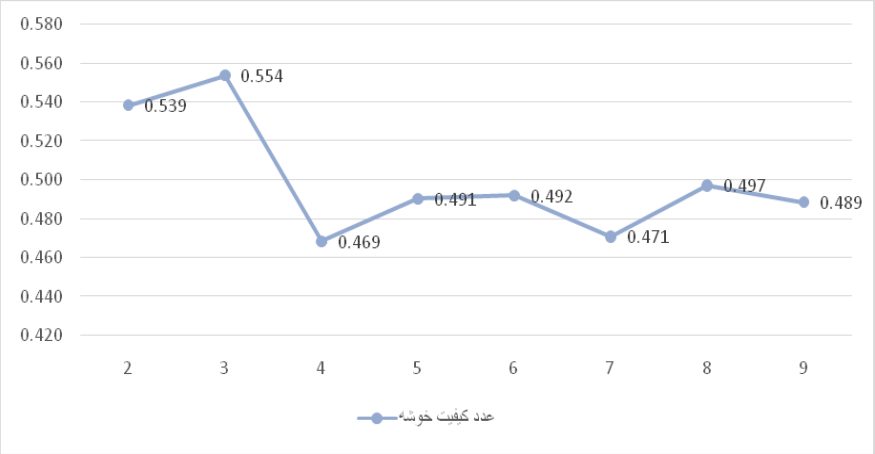
$$D= \min_{i=1..nc}\{ \min_{j=i+1..nc}(\frac{d(ci,cj)}{\max k=1..nc(diam(ck))})\} \quad (۱)$$

که در آن (D(ci,cj) و Diam(ci) به صورت زیر محاسبه می‌شوند.

$$D(c_i, c_j) = \min_{x\in ci,\ y\in cj}\{d(x, y)\} \quad (۲)$$

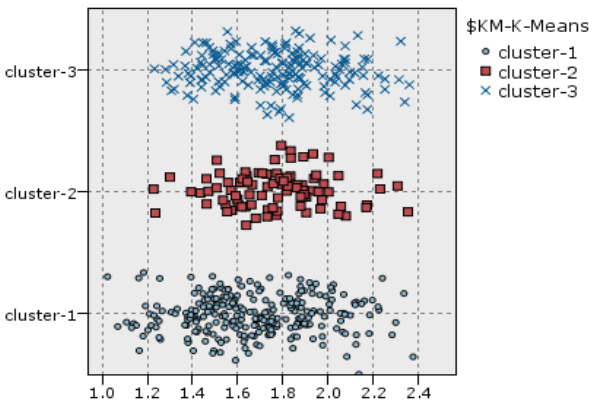
$$Diam(c_i) = \max_{x,\ y\in ci}\{d(x, y)\} \quad (۳)$$

هدف این شاخص ماکزیمم کردن فاصله برون خوشه‌ای در ضمن مینیمم کردن فاصله درون خوشه‌ای است. مقادیر این شاخص هر چه بزرگتر باشد بهتر است. بنابراین تعداد خوشه‌ای که مقدار این شاخص را زیادتر نماید، مقدار بهینه تعداد خوشه‌ها است [۲۰]. تعداد خوشه بهینه به دست آمده به کمک شاخص دان، ۳ خوشه است (شکل ۳). در ارزیابی و تحلیل خوشه‌بندی، بررسی می‌کنیم که در هر خوشه چه رکوردهایی، با چه مقادیری قرار گرفته است.



شکل ۳: تعیین تعداد خوشه بهینه به کمک شاخص دان

مهم‌ترین عامل در خوشه‌بندی، معیار شباهت است. به این معنی که اشیاء داخل یک خوشه به هم شبیه هستند. شباهت هر خوشه نسبت به متوسط اشیاء آن خوشه سنجیده می‌شود. خوشه‌ای به عنوان خوشه بهینه شناخته می‌شود که اشیاء هر خوشه در دسته‌های مجزا قرار داشته باشند و تداخلی با هم نداشته باشند. هر چه خوشه‌ها متمرکزتر باشند، عمل خوشه بندی بهینه‌تر انجام شده است. برای این منظور از پلات برای نمایش گرافیکی خوشه‌ها استفاده شده است که در شکل ۴ نمایش داده شده است. با توجه به این پلات مشخص می‌شود که خوشه‌های مختلف مرزهای کاملاً مجزایی دارند.



شکل ٤: نمایش تراکم خوشه‌ها با استفاده از گراف پلات

فاکتورهایی که با استفاده از خوشه بندی به روشK- Means در تعداد خوشه بهینه ۳ به عنوان فاکتورهای با اهمیت شناخته شده‌اند عبارت بودند از: تعداد دفعه‌های آزمایش خلط، سرعت رسوب خون، هموگلوبین، عرق شبانه، میزان گلبول‌های سفید خون، آلبومین، سن، استفاده از الکل، استفاده از سیگار و طول مدت آن، تب، ایدز، نوع شغل، کاهش شدید وزن و جنسیت.

هر یک از ویژگی‌ها وابسته به میزان و دامنه تغییرات درون خوشه‌ای، به عنوان ویژگی با اهمیت، بدون اهمیت و یا مرزی معرفی می‌شوند. به عنوان مثال در خوشه ۱ با تعداد ۲۶۰ رکورد، ۲۵۸ نفر زن و ۲ نفر مرد و خوشه ۲ یا مجموع ۸۸ رکورد، ۳ نفر زن و ۸۵ نفر مرد، خوشه ۳ با مجموع ۱۷۶ رکورد ۲ نفر زن و ۱۷۴ نفر مرد هستند. بنابراین ویژگی جنسیت در هر ۳ خوشه، فاکتور با اهمیت تعیین شده است. فاکتورهای با اهمیت به تفکیک هر خوشه در جدول ۲ آورده شده‌اند.

جدول ۲: ویژگی‌های با اهمیت به تفکیک خوشه‌ها

خوشه ۱	خوشه ۲	خوشه ۳
HB	BK	HB
سن	ESR	تعریق شبانه
مصرف الکل	تعریق شبانه	Albumin
ایدز	WBC	مصرف الکل
جنسیت	سن	مدت مصرف سیگار
مصرف سیگار	مدت مصرف سیگار	ایدز
	تب	تب
	کاهش وزن	کاهش وزن
	جنسیت	جنسیت
	مصرف سیگار	مصرف سیگار

بیشترین جمعیت مربوط به خوشه یک می‌باشد که ۹۹٪ آن را زنان تشکیل می‌دهند. ۳۶٪ این افراد آزمایش اسمیر خلط مثبت را انجام نداده‌اند، ۸۳٪ خلط آغشته به خون نداشته‌اند و سرعت رسوب گلبول‌های قرمز خون ۸۶٪ بیماران در محدوده غیرنرمال قرار دارد. ۵۳٪ قند خون ناشتا در بازه نرمال دارند. ۵۱٪ هموگلوبین در بازه کم خونی و ۹٪ در محدوده همو کروماتوزیس (Hemochromatosis) هستند. ۷۰٪ عرق شبانه ندارند. ۷۰٪ پلاکت، ۷۵٪ آلبومین و ۷۵٪ تعداد گلبول سفید نرمال دارند. ۹۵٪ سرفه مزمن دارند. ۱۴٪ کراتینین در بازه آزومتی دارند. ۷۱٪ تب و ۷۸٪ کاهش وزن داشته‌اند. ۰/۸٪ افراد این گروه مبتلا به HIV+ هستند.

خوشه ۲، ۱۶/۵٪ از جمعیت آماری در این تحقیق را تشکیل می‌دهند که ۹۷٪ آن‌ها مرد هستند، ویژگی‌هایی به شرح زیر دارند: ۴۲٪ آزمایش اسمیر خلط مثبت را انجام نداده‌اند و ۲۲٪ تنها یک بار این آزمایش را انجام داده‌اند. ۷۶٪ خلط آغشته به خون نداشتند. ۹۰٪ تماسی با بیمار مبتلا به سل نداشته‌اند. سرعت رسوب گلبول‌های قرمز خون ۷۴٪ بیماران در محدوده غیرنرمال قرار دارد. ۲۵٪ این بیماران قند خون ناشتا در محدوده پیش دیابت دارند. ۶۵٪ هموگلوبین در بازه کم خونی دارند. ۸۴٪ عرق شبانه ندارند. ۹٪ این بیماران پلاکت در بازه ترومبوسیتوپنی (Thrombocytopenia) دارند در حالی که پلاکت ۸۸٪ بیماران نرمال است. ۱۰٪ گلبول سفید در بازه لکوسیتوسیز (Leukocytosis) دارند و ۸۷/۵ ٪ گلبول سفید نرمال دارند. متوسط سن این بیماران ۶۵ سال است. ۸۱٪ بیماران آلبومین در بازه نرمال دارند. ۷٪ بیماران مصرف الکل دارند. ۹۴٪ سرفه مزمن دارند. ۱۷٪ کراتینین در بازه آزومتی دارند. ۵۵٪ بیماران این گروه سیگار مصرف کرده‌اند. ۹۴٪ تب نداشته‌اند. ۶۲/۵٪ کاهش وزن داشته‌اند.

خوشه ۳ که ۳۳/۵٪ از کل افراد این مطالعه را تشکیل می‌دهند که ۹۹٪ آن‌ها مرد هستند ویژگی‌های مشترکی به شرح ذیل دارند: ۳۶٪ این بیماران ۳ بار آزمایش اسمیر خلط مثبت را انجام داده اند و ۳۵٪ اصلاً این آزمایش را انجام نداده اند. ۷۵٪ این بیماران خلط آغشته به خون نداشتند. ۷۹/۵ ٪ سابقه تماس با بیمار مبتلا به سل ندارند. سرعت رسوب گلبول‌های قرمز خون ۸۷٪ بیماران در محدوده غیرنرمال (التهاب) بوده است. ۵۶٪ بیماران قند خون ناشتای نرمال دارند. ۸۰٪ بیماران هموگلوبین

در محدوده آنمی هستند و ۳٪ هموگلوبین در محدوده هموکروماتوزیس(Hemochromatosis) دارند. ۹۳٪ عرق شبانه دارند. پلاکت ۶۳٪ در بازه نرمال و ۱۲٪ در بازه ترومبوسیتوسیز (Thrombocytosis) قرار دارند. ۲۳٪ بیماران گلبول سفید در بازه لکوسیتوسیز (Leukocytosis) دارند در حالی که ۷۰٪ گلبول سفید در محدوده نرمال دارند. ۶۱٪ این بیماران آلبومین در بازه نرمال دارند. ۷٪ این بیماران سابقه مصرف الکل دارند. ۹۶٪ بیماران سرفه مزمن دارند. ۱۴٪ کراتینین در محدوده آزوتومیا(azotemia) و ۸۶٪ کراتینین نرمال دارند. ۵۰٪ این بیماران سابقه مصرف سیگار دارند که ۹۱٪ آن‌ها بیشتر از ۱۵ سال سیگار مصرف داشته‌اند. متوسط سن این بیماران ۴۵ سال است. ۶۱٪ این بیماران آلبومین در بازه نرمال دارند. ۷٪ این بیماران سابقه مصرف الکل دارند. ۹۶٪ سرفه مزمن دارند. ۱۴٪ کراتینین در بازه آزومتی دارند. ۹۵٪ تب داشته‌اند. ۹۳/۷۵٪ کاهش وزن داشته‌اند. ۷٪ افراد این گروه مبتلا به HIV+هستند.

برای بررسی بهتر و دقیق‌تر ارتباط بین ویژگی‌های مختلف از قوانین با هم آیی (Association Rules) استفاده کرده‌ایم. برای این منظور خروجی حاصل از خوشه‌بندی را به مجموعه رکوردهای موجود خود اضافه کرده‌ایم. سپس برای داشتن قوانین باهم‌آیی مطمئن‌تر معیارهای اطمینان (Confidence) و پشتیبان (Support) را به ترتیب ۹۵ و ۷۰ درصد تعیین کردیم. استخراج قواعد با هم آیی نوعی عملیات داده کاوی است که با هدف یافتن ارتباط بین ویژگی‌ها در مجموعه داده‌ها انجام می‌شود [۲۰]. جدول ۳ قوانین استخراج شده از گره Apriori را نمایش می‌دهد.

جدول۳: قواعد باهم آیی استخراج شده از مجموعه داده‌ها

تالی	مقدم	تعداد نمونه	Support %	Confidence %
سرفه=۱	از دست دادن وزن=۱	۴۲۳	۸۰/۷۲۵	۹۶/۴۵۴
سرفه=۱	ESR = 2 and creatinine = 2	۳۷۸	۷۲/۱۳۷	۹۶/۲۹۶
سرفه=۱	ESR = 2	۴۴۲	۸۴/۳۵۱	۹۵/۹۲۸
سرفه=۱	creatinine = 2	۴۴۷	۸۵/۳۰۵	۹۵/۵۲۶
سرفه=۱	تست خلط اسمیر=۰	۴۱۶	۷۹/۳۸۹	۹۵/۱۹۲
سرفه=۱	PLT = 2 and creatinine = 2	۳۷۳	۷۱/۱۸۳	۹۵/۱۷۴

قوانین استخراج شده را می‌توان به صورت زیر تشریح کرد:

- در ۴۴۲ نمونه وقتی سرعت رسوب گلبول‌های قرمز خون در بازه التهابی قرار داشته، سرفه مزمن گزارش شده است. از این تعداد ۳۷۸ نمونه کراتینین در بازه نرمال داشته‌اند.
- ۳۳۷ نمونه با داشتن پلاکت و کراتینین نرمال، سرفه شدید داشته‌اند.
- ۴۱۶ نمونه خلط آغشته به خون نداشته‌اند، ولی سرفه مزمن داشته‌اند.
- ۴۲۳ نمونه کاهش وزن همراه با سرفه مزمن داشته‌اند.

بحث و نتیجه‌گیری

در این تحقیق با استفاده از الگوریتم‌های داده کاوی تلاش کردیم ارتباطات بین ویژگی‌های مختلف بیماران مبتلا به سل دموگرافیک بیماران مبتلا به سل با استفاده از تکنیک‌های داده کاوی گزارش کاملی در دسترس ما نبوده است. در ادامه تلاش را استخراج کنیم. برای این کار از الگوریتم غیرنظارتی خوشه بندی K-Means و قوانین باهم‌آیی Apriori استفاده کرده‌ایم. مهم‌ترین عوامل شناسایی شده با استفاده از خوشه‌بندی عبارت بودند از: هموگلوبین، سن، جنسیت، مصرف سیگار، مصرف الکل و کراتینین. با توجه به قوانین با هم‌آیی، بیشترین ارتباط بین سرفه، کاهش وزن و سرعت رسوب گلبول‌های قرمز یافت شده است.

پژوهش‌های گذشته در حوزه داده کاوی بیماری سل، مربوط به پیش‌بینی و دسته‌بندی بیماران می‌باشند. گزارش‌های منتشر شده از این پژوهش‌ها در رابطه با دقت و صحت الگوریتم‌های پیش‌بینی می‌باشند. در رابطه با فراوانی و ارتباطات ویژگی‌های بیماران سل، گزارش‌های در دسترس ما به کمک نرم افزارهای آماری انجام شده بودند. در حوزه بررسی ویژگی‌های بالینی و می‌کنیم چند نمونه از کارهایی که تا حدی نزدیک به کار ما هستند را بررسی کنیم.

Asha و همکاران از ۷۰۰ داده واقعی جمع‌آوری شده از یک بیمارستان شهری برای تشخیص بیماری سل به کمک تکنیک-های خوشه‌بندی و دسته‌بندی استفاده کرده‌اند. داده‌های مورد استفاده از این تحقیق شامل سن، سرفه، از دست دادن وزن، تب، عرق شبانه، خلط آغشته به خون، درد سینه، ایدز، یافته‌های رادیوگرافی، خلط سینه، خس‌خس و نوع سل بوده است [۲۱]. Bakar و Febriyani به بررسی بیماری سل ۲۳۳ رکورد از بیماران پرداخته‌اند. ویژگی‌های مورد استفاده عبارت بودند از: سن، جنسیت، وزن، سرفه بیش از ۳ هفته، عرق شبانه، تب، خلط سینه و خلط آغشته به خون. روش استفاده در این تحقیق گسسته‌سازی به کمک رگرسیون بوده است [۲۲]. Abdallah و Ali بررسی فاکتورهای همه‌گیری مربوط به سل غیر ریوی در شرق سودان پرداخته‌اند. براساس یافته‌های این پژوهش محل سکونت و شغل رابطه مستقیمی و معناداری با سل نداشتند، پایین بودن سطح آموزش، جنسیت مؤنث، عدم تزریق واکسن و آن‌هایی که دخانیات مصرف می‌کنند، بیشترین میزان سل غیر ریوی را داشتند [۲۳].

با توجه به تحقیقات انجام شده، بیشترین فاکتورهایی که مورد بررسی قرار گرفته‌اند عبارت بودند از: سن، سرفه، خلط سینه، تب، عرق شبانه، از دست دادن وزن و ایدز. پیشنهاد ما برای کارهای بعدی، بررسی ارتباطات این ویژگی‌ها و بیماری‌های همراه مبتلایان به سل می‌باشد. با این هدف که با کنترل عوامل مؤثر، کمک به کاهش بروز این بیماری‌ها در افراد مبتلا به سل داشته باشیم.

References

1. Rusdah U, Winarko E. Review on data mining methods for tuberculosis diagnosis. Information Systems International Conference (ISICO); 2013 Dec 2–4; Surabaya, Indonesia: Institut Teknologi Sepuluh Nopember (ITS); 2013.

2. Nasehi M, Mirhaghani L. National Guidelines TB. 2th ed. Tehran: Andishmand; 2008. Persian.

3. Ameri H, Alizadeh S, Barzegari A. Knowledge Extraction of diabetics' data by decision tree method. J Health Adm. 2013; 16(53):58-72. Persian.

4. Al Jarullah AA. Decision tree discovery for the diagnosis of type II diabetes. Innovations in Information Technology (IIT), 2011 International Conference on; 2011 Apr 25-27; Abu Dhabi: IEEE; 2011.p. 303-7.

5. Khajehei M, Etemady F. Data mining and medical research studies. Computational Intelligence, Modelling and Simulation (CIMSiM). 2th International Conference on; 2010 Sep 28-30; Bali: IEEE; 2010. p. 119 –22.

6. Jayalakshmi T, Santhakumaran A. A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. Data Storage and Data Engineering (DSDE), 2010 International Conference on; 2010 Feb 9-10; Bangalore: IEEE; 2010. p. 159-63.

7. Ameri H, Alizadeh S, Barzegari A. Identification of influencing factors for heart attack in diabetic patients using C & R algorithm. Daneshvar Medicine. 2014; 21(112):1-13. Persian.

8. Sharifkhani M, Alizadeh S, Abbasi M, Ameri H. Providing a model for predicting the risk of osteoporosis using decision tree algorithms. J Mazandaran Univ Med Sci. 2014; 24(116):110-18. Persian.

9. Uçar T, Karahoca A, Karahoca D. Tuberculosis disease diagnosis by using adaptive neuro fuzzy inference system and rough sets. Neural Computing and Applications. 2013; 23(2): 471-83.

10. Nagabhushanam D, Naresh N, Raghunath A, Praveen Kumar K. Prediction of Tuberculosis Using Data Mining Techniques on Indian Patient's Data National Conference on Research Issues and Recent Trends in Computer Science & Information Technology (NCRRCSIT 2013), At Sir C R Reddy College of Engineering; 2013 Oct – Dec 4; Eluru, AP, India; 2013.

11. Uçar T, Karahoca A. Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches. Procedia Computer Science. 2011; 3:1404-11.

12. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 2th ed. San Francisco: Morgan Kaufmann; 2006.

13. Chen G, Astebro TB. How to deal with missing categorical data: test of a simple Bayesian method. Organizational Research Methods. 2003; 6(3):309-27.

14. Witten IH, Frank E, Hall MA. Data Mining: Practical machine learning tools and techniques. 3th ed. USA: Morgan Kaufmann; 2011.

15. Tahmasbi H, Amoozgar M, Adine H. Replacement of missing values and its effect on the classification accuracy in medical data mining. Journal of Health and Biomedical Informatics. 2015; 2(1):24-32. Persian.

16. Burris CA, Ashwood ER, Burns DE. Tietz textbook of clinical chemistry and molecular diagnostics. 4th ed. Louis: Elsevier Saunders; 2006.

17. Mc Pherson RA, Mattew R, Princus MR. Henry's clinical diagnosis and management by laboratory methods. 22th ed. Philadelphia: Elsevier Saunders; 2011.

18. Lujambio I, Sottolano M, Luzardo L, Robaina S, Krul N, Thijs L, et al. Estimation of glomerular filtration rate based on serum cystatin C versus creatinine in a uruguayan population. Int J Nephrol. 2014; 2014:837106.

19. Alizadeh S, Ghazanfari M, Teimorpour B. Data Mining and Knowledge Discovery. 2th ed. Tehran: Iran University of Science and Technology; 2011. Persian.

20. Alizadeh S, Malekmahmodi S. Data mining & knowledge discovery step by step clementine. Tehran:

Khaje Nasir Toosi University of Technology; 2011. Persian.
21. Asha T, Natarajan S, Murthy KN. A data mining approach to the diagnosis of tuberculosis by cascading clustering and classification. Journal of Computing. 2011; 3(4):1-8.
22. Bakar AA, Febriyani F. Rough neural network model for tuberculosis patient categorization. Proceedings of the International Conference on Electrical Engineering and Informatics; 2007 Jun 17-19; Indonesia: Institut Teknologi Bandung; 2007.p. 765–8.
23. Abdallah TM, Ali AA. Epidemiology of tuberculosis in Eastern Sudan. Asian Pac J Trop Biomed. 2012;2(12):999-1001.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2015; 2(3): 149-159

The Investigation of TB Patients Features with K-Means Clustering

Farzad Firuzi Jahantigh[1], Hakimeh Ameri[2]*

Introduction: According to the World Health Organization, TB is the largest cause of death among infectious diseases. Due to the high percentage of tuberculosis infection and the high number of death among these patients, this study was carried out to categorized and find the relationship between different clinical and demographical characteristics.
Method: This descriptive analytical study was done on 600 patients from Masih Daneshvari hospital tuberculosis research center. K-means clustering, Apriori association rules, and data mining algorithms (SPSS Clementine software) were used for clustering and determining the common characteristics among patients.
Results: Based on DUNN index, 3 clusters were chosen as optimal cluster. The common factors between clusters have been described in details in findings section. According to the characteristics of each cluster, patients can be classified based on the effectiveness of various factors
Conclusion: According to the results of this study, the most important identified factors by the use of clustering are Hemoglobin, age, sex, smoking, alcohol and Creatinine. Based on the association rules the highest rate of relationship is found between cough, weight loss, and ESR.

Key words: Tuberculosis, Clustering, Association rules, Data mining

• **Received:** 14 Oct, 2015 • **Accepted:** 16 Nov, 2015

• **Citation:** Firuzi Jahantigh F, Ameri H. The investigation of TB patients features with K-Means clustering. Journal of Health and Biomedical Informatics 2015; 2(3): 149-159.

1. Ph.D. Industrial Engineering, Assistant Professor of Industrial Engineering Dept., University of Sistan and Baluchestan, Zahedan, Iran
2. M.Sc. in Information Technology, Industrial Engineering Dept., Khaje Nasir Toosi University of Technology, Tehran, Iran
***Correspondence**: Industrial Engineering Dept., Khaje Nasir Toosi University of Technology, Tehran, Iran
• **Tel:** 09214706993 •**Email:** ha.amery@gmail.com