مجله دانشکده پزشکی، دانشگاه علوم پزشکی تهران، شهریور ۱۴۰۰، دوره ۷۹، شماره ٦، صفحه‌های ٤٥٩ تا ٤٦٧ TUMJ

مقاله اصیل

پیش‌بینی ابتلا به بیماری مزمن کلیوی در شهر اصفهان با استخراج قواعد انجمنی توسط تکنیک‌های داده‌کاوی

فیروزه معین‌زاده[1]، محمد حسین روحانی[2]، مژگان مرتضوی[1]، محمد ستاری[3*]

۱- مرکز تحقیقات بیماری‌های کلیوی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.

۲- گروه تغذیه جامعه، مرکز تحقیقات امنیت غذایی، دانشکده تغذیه و علوم غذایی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.

۳- مرکز تحقیقات فناوری اطلاعات در امور سلامت، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.

[*] نویسنده مسئول: اصفهان، خیابان هزار جریب، دانشگاه علوم پزشکی اصفهان، دانشکده مدیریت و اطلاع‌رسانی پزشکی، طبقه دوم.

تلفن: ۰۳۱-۳۷۹۲۵۱۵۲

E-mail: msattarimng.mui@gmail.com

چکیده

دریافت: ۱٤۰۰/۰۲/۰٤ ویرایش: ۱٤۰۰/۰۲/۱۱ پذیرش: ۱٤۰۰/۰۵/۲۳ آنلاین: ۱٤۰۰/۰٦/۰۱

زمینه و هدف: سالانه میلیون‌ها مرگ به‌دلیل دسترسی نداشتن به درمان مناسب بیماری کلیوی در جهان اتفاق می‌افتد. پژوهشگران قصد دارند از طریق ترکیب بهینه تکنیک‌های به‌کار رفته در مراحل مختلف داده‌کاوی، بیماری مزمن کلیوی را شناسایی نمایند.

روش بررسی: این پژوهش از بهمن ۹۹ تا اردیبهشت ۱٤۰۰ به‌صورت مقطعی انجام شده است. مجموعه داده مورد استفاده شامل ٤۱٤۵ نمونه (بیمار) و ۳۲ صفت (دموگرافیک و بالینی) در نظر گرفته شده است. معیارهای واجد شرایط بودن برای آزمایش شامل بزرگسالان ۱۸ سال به بالا، ساکن اصفهان، مایل به شرکت در مطالعه، نبود تب و سرماخوردگی در زمان انجام آزمایشات آزمایشگاهی، بدون انجام تمرینات سنگین ٤۸ ساعت پیش از آزمایش‌های آزمایشگاهی و ناشتا بودن است. متغیر هدف، بیماری کلیوی است که مقادیرش بیمار و سالم است. در این مطالعه از چهار تکنیک ماشین‌بردار پشتیبان، جنگل تصادفی، شبکه عصبی و CHAID استفاده شده‌اند.

یافته‌ها: براساس معیار Accuracy، ماشین‌بردار پشتیبان نسبت به سایر تکنیک‌ها عملکرد بهتری داشته است. مطمئن‌ترین قانون بیان می‌نماید که اگر فرد از نمک در غذا استفاده نماید و سن وی بین ۵۰ تا ٦۹ باشد و بیماری دیابت داشته باشد به احتمال ۸۲٪ دچار بیماری مزمن کلیوی خواهد شد.

نتیجه‌گیری: همچنین قواعد مستخرج نشان داد که استفاده از نمک در کنار بیماری دیابت می‌تواند منجر به بیماری مزمن کلیوی شود و حتی داشتن بیماری دیابت می‌تواند خطر مرگ‌ومیر بیماران کلیوی را هم افزایش دهد که خود این موضوع قابل تامل است. همچنین افراد مسن هم باید بیشتر مراقب سلامتی خود باشند تا کمتر در معرض بیماری مزمن کلیوی قرار گیرند.

کلمات کلیدی: بیماری مزمن کلیوی، داده‌کاوی، پیش‌بینی.

مقدمه

۱۰٪ از جمعیت جهان دچار بیماری مزمن کلیه هستند و سالانه میلیون‌ها مرگ به‌دلیل دسترسی نداشتن به درمان مناسب در جهان اتفاق می‌افتد. بیماری مزمن کلیه اغلب در مراحل اولیه، یک بیماری بدون علامت است. در واقع وقتی عملکرد کلیه به‌تدریج کاهش می‌یابد، علایم و نشانه‌های مختلف همچون خستگی و نارسایی رشد آشکار می‌شوند.[1] بهترین شاخص موجود برای عملکرد کلی کلیه، میزان فیلتراسیون گلومرولی (GFR) است که برای مشخص‌کردن سطح خون در حال عبور از کلیه به‌کار برده می‌شود. تعریف و طبقه‌بندی بیماری مزمن کلیه با گذشت زمان تکامل یافته است.[2] کلیه‌ها از طریق سه مکانیسم فیلتراسیون، بازجذب و ترشح در دفع

مواد زاید بدن نقش اساسی دارند. بیماری مزمن کلیه به زمانی اطلاق می‌شود که عملکرد کلیه‌ها در مدت بیش از سه ماه کاهش غیرقابل برگشت پیدا کرده باشد و میزان فیلتراسین گلومرولی به کمتر از ٦٠ cc/min/١/٧٣ m^2 رسیده باشد.[٣] این بیماری به‌عنوان یکی از ٢٠ علت مرگ‌ومیر در دنیا به‌شمار می‌رود.[٤] نکته دیگر در مورد این بیماری این است که بیماران مزمن کلیوی شدیدا در معرض بیماری کووید ١٩ هستند و باید بسیار مراقب خود باشند که دچار این بیماری نشوند.[٥]

با توجه به اهمیت و میزان مرگ‌ومیر این بیماری، پیش‌بینی زودهنگام و با هزینه کمتر آن از اهمیت بسیاری برخوردار است. علم داده‌کاوی در جهت ارایه الگوی مناسب جهت اخذ تصمیمات بهتر در سال‌های اخیر مطرح شده است.[٦و٧] باتوجه به اینکه حجم گسترده‌ای از داده‌ها در مورد بیماران کلیوی وجود دارد و اینکه اطلاعات ارزشمندی در این داده‌ها نهفته است، داده‌کاوی می‌تواند در اتخاذ تصمیمات مناسب در این زمینه موثر باشد. Subasi از تکنیک جنگل تصادفی، شبکه عصبی، درخت تصمیم و k– نزدیکترین همسایه برای پیش‌بینی ابتلا به بیماری مزمن کلیوی استفاده نمود. مجموعه داده مورد استفاده شامل ٤٠٠ بیمار کلیوی با ٢٤ صفت است. نتایج نشان‌دهنده عملکرد مناسب تکنیک جنگل تصادفی بود.[٨] Gharibdousti و همکاران داده‌های ٢٤ صفت از ٤٠٠ بیمار را برای تشخیص بیماری مزمن کلیوی استفاده نمودند. تکنیک‌های مختلف استفاده شده شامل درخت تصمیم، رگرسیون خطی، ماشین‌بردار پشتیبان و بیزین ساده (Naïve Bayes) بود. این تکنیک‌ها هشت و ٩ صفت را به‌عنوان صفات تاثیرگذار استخراج نمودند.[٩] در واقع هدف این بررسی، این است که نشان دهد استخراج فاکتورهای تاثیرگذار می‌تواند از بررسی فاکتورهایی که تاثیری بر خروجی نهایی ندارند، جلوگیری نماید. Akben از روش خوشه‌بندی K-Means به‌عنوان روش پیش‌پردازش استفاده نمود. سپس، روش‌های طبقه‌بندی (KNN، SVM و Naïve Bayes) برای تشخیص بیماری مزمن کلیوی در داده‌های از پیش‌پردازش شده استفاده شد. نتایج نشان‌دهنده این موضوع بود که استفاده از آزمایش ادرار نتایج بهتری نسبت به استفاده از آزمایش خون دارد.[١٠] Almansour در مطالعه خود از دو تکنیک شبکه عصبی و ماشین‌بردار پشتیبان برای پیش‌بینی ابتلا به بیماری مزمن کلیوی در یک مجموعه داده‌ای ٤٠٠ بیمار با ٢٤ صفت استفاده نمود و عملکرد این دو تکنیک را مورد مقایسه قرار داد. نتایج نشان داد که فاکتور سن، فشارخون، هموگلوبین و کراتین خون را به‌عنوان فاکتورهای تاثیرگذار بر نارسایی کلیوی می‌توان در نظر گرفت.[١١]

با توجه به اینکه نمی‌توان گفت که یک تکنیک برای همه مجموعه‌های داده‌ای مناسب است و با در نظر گرفتن این نکته که تاکنون داده‌های مرتبط با بیماران مزمن کلیوی در سطح شهر اصفهان مورد بررسی قرار نگرفته است، پژوهشگران قصد دارند از طریق ترکیب بهینه تکنیک‌های به‌کار رفته در مراحل مختلف داده‌کاوی، ابتلا به بیماری مزمن کلیوی را براساس ریسک فاکتورهای افراد شناسایی نمایند.

روش بررسی

جمع‌آوری اطلاعات: برای محاسبه اندازه نمونه، با خطای برآورد ٠/٠٥٪، فرضیه شیوع CKD ٦٪ براساس تخمین‌های یک مطالعه اخیر در ایران و با فرض خطای حداقل برابر با P=٠/٠٠٨ به‌دلیل نمونه‌گیری از شهری با جمعیت محدود شهری خواهد بود.[١٢]

در این مطالعه از نمونه‌گیری تصادفی خوشه‌ای چند مرحله‌ای استفاده شد. هر منطقه از شهر با توجه به سیستم بهداشتی درمانی که لایه را پوشش می‌دهد، به‌عنوان یک طبقه جداگانه در نظر گرفته شد. در مرحله اول نمونه‌گیری خوشه‌ای (مناطق انتخاب)، از هر منطقه وجودی، خانه‌ها و محله‌های تحت پوشش یک مرکز بهداشتی درمانی براساس جمعیت محاسبه شده براساس آخرین آمار شهرداری‌ها انتخاب شدند. مرحله دوم به روش طبقه‌ای (انتخاب براساس جنسیت و گروه‌های سنی) و سپس به‌صورت تصادفی ساده (افراد مراجعه‌کننده به مراکز) انجام شد.

معیارهای واجد شرایط بودن شامل بزرگسالان ١٨ سال به بالا، ساکن اصفهان، مایل به شرکت در مطالعه، فقدان تب و سرماخوردگی در زمان انجام آزمایشات آزمایشگاهی، بدون انجام تمرینات سنگین ٤٨ ساعت پیش از آزمایش‌های آزمایشگاهی و ناشتا بودن است. افرادی که پرسشنامه ناقص داشتند یا تمایلی به انجام دقیق آزمایشات نداشتند، از مطالعه خارج شدند. همچنین، زنان باردار یا کسانی که در دوره قاعدگی بودند، از مطالعه خارج شدند.

مشخص‌کردن صفات منتخب: شامل اطلاعات ٤١٤٥ بیمار و فرد سالم است. اطلاعات بیماران از طریق یک مطالعه مقطعی با در نظر

گرفتن عموم جامعه جمع‌آوری شده است. بیماران شامل بیماران نارسایی کلیه و مزمن کلیوی بودند. از این تعداد، ۵۳۲ بیمار دچار بیماری مزمن کلیوی و ٦۵ بیمار دچار نارسایی و بقیه سالم بودند. در واقع ۵۳۲ نفر در سطح یک یا دو بیماری کلیوی بودند و ٦۵ نفر وضعیت نامناسبی از لحاظ شدت بیماری نسبت به گروه ۵۳۲ نفری داشتند و دچار نارسایی کلیوی بودند. مجموعا ۳۲ خصوصیت به‌عنوان خصوصیاتی که می‌توانند بر وضعیت بیماران کلیوی تاثیرگذار باشند، در نظر گرفته شدند. خصوصیات منتخب شامل سن، جنسیت، وضعیت تاهل، شغل، تحصیلات، وزن، قد، BMI، استفاده از سیگار، استفاده از قلیان، استفاده از الکل، فشارخون بالا، بیماری قلبی، دیابت، مصرف نمک، صدمه مغزی و عروقی، کم‌خونی، بارداری، سابقه استفاده از داروی گیاهی، سابقه استفاده از داروی شیمیایی، نسبت دور کمر به دور باسن، GFR، نوع داروی مصرفی گیاهی، نوع داروی مصرفی شیمیایی، لیپوپروتیین با چگالی بالا، لیپوپروتیین با چگالی پایین، کلسترول، تری‌گلیسرید، قند خون، کراتینین خون، تعداد گلبول خون و نسبت آلبومین به کراتینین بودند.

تبدیلات داده‌ای: جهت پردازش بهتر داده‌ها مخصوصا داده‌های عددی، تبدیلاتی انجام شد بدین صورت که سن افراد تبدیل به پنج بازه عددی شد. قد، وزن و BMI هم تبدیل به پنج یا شش بازه عددی شدند. برخی داده‌های عددی مانند کلسترول، لیپوپروتیین با چگالی بالا و لیپوپروتیین با چگالی پایین که می‌توان براساس مقادیرشان سه قسمت، نرمال، پرخطر و نزدیک به پرخطر را در نظر گرفت تقسیم شدند. در مورد داده‌های عددی که حالت دودویی دارند مانند استفاده از سیگار، استفاده از الکل و استفاده از قلیان، عدد یک معادل با بله و عدد صفر معادل با خیر در نظر گرفته شد. در مورد برخی داده‌ها مانند وضعیت تاهل و جنسیت نیاز به تبدیل خاصی نبود. با توجه به اینکه تکنیک شبکه عصبی و ماشین‌بردار پشتیبان برای داده‌های عددی استفاده می‌شوند. از تابع تبدیل داده‌های غیرعددی پیش از اعمال این تکنیک‌ها استفاده شد.

انتخاب هدف: صفت بیماری به‌عنوان کلاس هدف در نظر گرفته شد. کلیه بیماری‌ها مانند بیماری مزمن کلیوی و نارسایی کلیه به عنوان مقدار کلاس هدف در نظر گرفته شدند. بیماری مزمن و نارسایی به‌عنوان یک مقدار بیماری کلیوی در نظر گرفته شدند. در واقع با توجه به اینکه بیماری نارسایی کلیوی نوعی شدید از بیماری مزمن کلیوی است بنابراین می‌توانند در هم ادغام شده و به‌عنوان یک مقدار به نام بیماری کلیوی در نظر گرفته شوند. از این‌رو یک کلاس هدف دو مقداره با مقادیر بیمار و سالم در نظر گرفته شد. البته یکسری از افراد مشکوک به بیماری مزمن کلیوی بودند، از آنجایی که این افراد یک بار آزمایش غیرطبیعی داشتند، عملا باید آنها را جز بیماری مزمن کلیوی در نظر گرفت.

مدلسازی: تکنیک‌های مورد استفاده در این قسمت شامل جنگل تصادفی، شبکه عصبی و ماشین‌بردار پشتیبان بودند. در تکنیک جنگل تصادفی، تعدادی درخت تصمیم به‌صورت تصادفی ایجاد می‌شود. سپس این درخت‌ها با هم ادغام می‌شوند تا بتوان به یک تصمیم‌گیری رسید.[۱۳]

در روش ماشین‌بردار پشتیبان، هر نمونه به عنوان یک نقطه در فضای n بعدی در نظر گرفته می‌شود.[۱٤] متغیر n متناظر با تعداد صفات خواهد بود. بنابراین در مطالعه n برابر با ۳۲ در نظر گرفته شد. هدف یافتن یک خط راست برای دسته‌بندی نقاط مختلف به دو گروه سالم و بیمار است.

شبکه‌های عصبی با الهام‌گرفتن از مغز انسان، عملیات مختلف را انجام می‌دهند. مغز انسان از واحدهای کوچک پردازشی به نام نورون ساخته شده است.[۱۵] شبکه عصبی شامل سه واحد ورودی، نهان و خروجی است. هر کدام از واحدها متناظر با صفات مختلف است و واحد خروجی تعیین کننده‌ی وضعیت نهایی نمونه است. تعداد تکرار روش شبکه عصبی برابر با ۵۰۰ در نظر گرفته شد. الگوریتم CHAID برخلاف روش جنگل تصادفی، یک روش قطعی است و از درخت برای نمایش اطلاعات استفاده می نماید.

ارزیابی: در مرحله ارزیابی، ٤۱٤۵ نمونه و ۳۲ صفات در نظر گرفته شدند. برای تقسیم‌بندی مجموعه داده‌ای به مجموعه آموزشی و تست، از روش 10-fold cross validation استفاده شد. بر این اساس ٤۱٤۵ نمونه موجود به ۱۰ گروه تقسیم شدند. طی ۱۰ بار تکرار آزمایش هربار ۹ گروه که ۹۰٪ مجموعه داده‌ای اصلی را تشکیل می‌دادند به‌عنوان مجموعه داده آموزشی و یک گروه باقیمانده، که ۱۰٪ مجموعه داده‌ای اصلی را تشکیل می‌داد به‌عنوان مجموعه داده‌ای تست در نظر گرفته شدند. از مجموعه داده‌های آموزشی برای تولید الگوهای ورودی جهت ساخت مدل طبقه‌بند و از مجموعه داده‌ای تست برای ارزیابی صحت آن استفاده شد.

یافته‌ها

معیارهای ارزیابی: یکی از معیارهای ارزیابی، Accuracy است که هرچه دقت این معیار نزدیکتر باشد، نتیجه بهتری خواهد داشت. این معیار براساس فرمول زیر محاسبه می‌شود:[16]

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}.$$

معیارهای دیگر شامل False positive rate است که هر چه به یک نزدیکتر باشد نشان‌دهنده عملکرد نامناسب‌تر روش است.[17] معیار بعدی Confidence است که هر چه به یک نزدیکتر باشد عملکرد بهتری دارد.[18] همچنین از معیار Support نیز استفاده شده است. این معیار تعداد تکرار قانون را در مجموعه داده‌ای نشان می‌دهد.[19]

در این طرح از مجموعه داده‌ای مربوط به افراد در معرض نارسایی کلیوی استفاده شده است. مشخصات این مجموعه داده‌ای، مقادیر و تعداد استفاده از این مقادیر در جداول ۱ و ۲ آمده است. سن افراد به پنج بازه ۱۰ تا ۲۹، ۳۰ تا ٤۹، ۵۰ تا ٦۹ و ۷۰ تا ۸۹ و ۹۰ تا ۱۰۰ تقسیم شده است که بیشترین تعداد مربوط به افراد بین ۳۰ تا ٤۹ ساله است. همچنین تعداد زن‌ها بیشتر از مردان است. افراد متاهل تقریبا پنج برابر افراد مجرد هستند. شغل بیشتر بیماران خانه‌دار است. همچنین غالب آنها تحصیلات زیر دیپلم دارند. وزن، قد و BMI به ترتیب به پنج، پنج و شش بازه تقسیم شدند. تعداد افرادی که سیگار مصرف می‌کنند تقریبا سه برابر افرادی است که قلیان مصرف می‌کنند (جدول ۱).

۵۲۱ نفر دچار کم‌خونی، ۱۳۰ نفر بیماری قلبی و ۵۰۱ نفر دیابت دارند. ۲٤۵۲ نفر تری‌گلیسرید بالا دارند (جدول ۲) که رقم قابل‌توجهی است. تقریبا نیمی از مجموعه داده‌ای دارای تری‌گلیسرید بالا هستند. همچنین بیش از نیمی از افراد یعنی ۲٦٤۲ نفر نمک مصرف می‌کنند. داروی ایبوبروفن در بین داروهای شیمیایی و آویشن در بین داروهای گیاهی بیشترین مصرف را دارند.

در این مجموعه داده‌ای، یک کلاس هدف دو مقداری تعریف شده است. دو مقدار برابر است با سالم و بیمار. از بین کل اعضا، ۵۹۹ نفر بیمار هستند در مقابل ۳٤۹۱ نفر سالم هستند که تقریبا چهار برابر تعداد افراد بیمار هستند (جدول ۳). تکنیک‌های مختلف بر روی

جدول ۱: خصوصیات، مقادیر و تعداد مربوط به اطلاعات دموگرافیک بیماران کلیوی

خصوصیات	مقادیر(تعداد)
سن*	۲۹-۱۰(٦٦۱) و ٤۹-۳۰(۱٦۹۰) و ٦۹-۵۰(۱٤۸٦) و ۸۹-۷۰(۳۰٦) و ۱۰۰-۹۰(۱)
جنسیت**	مرد (۱۵۵٤) و زن (۲۲۳۹)
وضعیت تاهل**	مجرد (۵۹۵) و متاهل (۳۲۵۰)
شغل**	بازنشسته (۳۸۵) و خانه‌دار (۱۷۲۹) و آزاد (۷۱۱) و کارمند (۳۳۲) و فرهنگی (٤٦) و شاغل در بیمارستان (٤۵) و کارگر (۸۰) و پزشک (۱۷)
تحصیلات*	بی‌سواد (۷۵۳) و زیردیپلم (۱۳۳۱) و دیپلم (۱۱۹۸) و لیسانس (٦۰۸) و بالاتر از لیسانس (٦۰۸)
وزن*	۵۰-۳۱(٤۹۸) و ۷۰-۵۱(۱٦۸۳) و ۹۰-۷۱(۱۵۷۳) و ۱۱۰-۹۱(۳٤٤) و ۱۱۰ و بیشتر (٤۵)
قد*	۱۳۰-۱۱۰(۳) و ۱۵۰-۱۳۱(۲۵۰) و ۱۷۰-۱۵۱(۲۵٤۷) و ۱۹۰-۱۷۱(۹۵۸) و ۲۱۰-۱۹۱(۱۳)
BMI*	کمتر یا مساوی ۲۰ (۱۸۵) و کمتر یا مساوی ۲٤ و بیشتر از ۲۰ (۷۹۱) و کمتر یا مساوی ۲۸ و بیشتر از ۲٤ (۱۳۸۵) و کمتر یا مساوی ۳۲ و بیشتر از ۲۸ (۸۷٦) و کمتر یا مساوی ۳٦ و بیشتر از ۳۲ (۳٤٤) و کمتر یا مساوی ٤۰ و بیشتر از ۳٦ (۹۸)
استفاده از سیگار**	خیر (۳۱۸٦) و بله (۸۸۲)
استفاده از قلیان**	خیر (۳۷۸۸) و بله (۲٦۲)
استفاده از الکل**	خیر (۳٦۵۰) و بله (۳۹۰)
نسبت دور کمر به دور باسن**	کمتر از یک (۳٦۱۹) و بیشتر یا مساوی یک (۵۲۵)

آزمون آماری: * Mann Whitney U test، ** Independent paired t-test، P<۰/۰۵ معنادار در نظر گرفته شد. Body Mass Index (BMI)

جدول ۲: خصوصیات، مقادیر و تعداد مربوط به داده‌های آزمایشگاهی مربوط به مجموعه داده‌ای بیماران کلیوی

خصوصیات	مقادیر(تعداد)
فشارخون بالا**	خیر (۳۳۵۰) و بله (۷۵۲)
بیماری قلبی**	خیر (۳۹۰۰) و بله (۱۳۰)
دیابت**	خیر (۳۶۰۰) و بله (۵۰۱)
صدمه مغزی و عروقی**	خیر (۳۳۳۷) و بله (۷۶۳)
کم خونی**	خیر (۳۵۸۰) و بله (۵۲۱)
بارداری**	خیر (۲۳٤۰) و بله (۱۷٤۹)
سابقه استفاده از دارو شیمیایی**	خیر (۲۵٤۷) و بله (۱۵۰۲)
سابقه استفاده از دارو گیاهی**	خیر (۳۱۷٤) و بله (۸۷۲)
GFR**	غیر نرمال (۱٤۱) و نرمال (۳۲۱۷)
لیپوپروتیین با چگالی بالا (HDL) **	بدون‌خطر (۳۳۵۱) و پرخطر (۱۸)
لیپوپروتیین با چگالی پایین*	بدون‌خطر (۲۹۰۱) و نزدیک به پر خطر(۳۸۰) و پرخطر (۸۶)
کلسترول*	نرمال (۲٤۸۱) و نزدیک به پرخطر (۵۷۶) و پرخطر (۳۱۱)
تری‌گلیسرید*	نرمال (۱٤۶۹) و نزدیک به پرخطر (۲۱۸) و پرخطر (۲۵٤۲)
نسبت آلبومین به کراتینین**	غیرنرمال (۲٤۳۳) و نرمال (۱۳۵۳)
قندخون*	نرمال (۱۰۳۰) و پیش‌دیابت (۱۰۵) و دیابت (۲۲۳۳)
کراتینین خون*	پایین (۳۹۹) و نرمال (۲۸۸۰) و بالا (۹۰)
تعداد گلبول خون*	پایین (۳۱۵۵) و نرمال (۷۷) و بالا (۱۱۶)
نوع داروی مصرفی شیمیایی**	استامینوفن (۷) و ایبوبروفن (۷۶۱) و ایندومتاسین (۱۵) و دیکلوفناک (۲۱۸) و سلکسیب (۲۰) و کتورولاک (٤) و مفنامیک اسید (۱۳) و ناپرکسن (۵) و نوافن (٤)
نوع داروی مصرفی گیاهی**	آویشن (۳۹) و اسطوقودوس (۶) و بابونه (۸) و انواع چای غیر از سیاه (۳۶) و گل گاوزبان (۳۷) و انواع عرقیجات (۱٤) و زنجبیل (۵) و رازیانه (٤) و شوید (۵)

آزمون آماری: * Mann Whitney U test، ** Independent Paired t-test ، P<۰/۰۵ معنادار در نظر گرفته شد.

جدول ۳: کلاس‌ها، مقادیر مختلف و فراوانی آنها

کلاس	مقدار	فراوانی
افراد	سالم	۳٤۹۱
	بیمار	۵۹۹

جدول ٤: میزان Accuracy و خطای تکنیک‌های مختلف داده‌کاوی

تکنیک‌ها	جنگل تصادفی	ماشین بردار پشتیبان	شبکه عصبی	CHAID
صحت	٪۸٤/۳۵	٪۸۶/۸۱	٪۸۶/۰۵	٪۶۸/۷۸
False positive rate	۰/۰۷	۰/۰٤	۰/۰۵	۰/۱۶

جدول ۵: قوانین مستخرج از مجموعه داده‌ای کلیه

قوانین	وضعیت	نتیجه	اطمینان
قانون ۱	اگر شغل آزاد است و وزن بین ۷۱ تا ۹۰ باشد و وضعیت تری‌گلیسرید پرخطر باشد	فرد دچار بیماری مزمن کلیوی خواهد شد	۰/۷۵
قانون ۲	اگر از نمک در غذا استفاده نماید و سن بین ۵۰ تا ٦۹ باشد و بیماری دیابت داشته باشد	فرد دچار بیماری مزمن کلیوی خواهد شد	۰/۸۲
قانون ۳	اگر کراتینین خون نرمال باشد و BMI بین ۲۸ تا ۳۲ باشد و از قلیان و سیگار استفاده نشود و سن بین ۳۰ تا ٤۹ باشد	فرد دچار بیماری مزمن کلیوی نخواهد شد	۰/۷۲
قانون ٤	اگر از نمک در غذا استفاده شود و لیپوپروتیین با چگالی پایین نزدیک به پرخطر باشد و کراتینین خون پایین باشد	فرد دچار بیماری مزمن کلیوی خواهد شد	۰/٦۸
قانون ۵	اگر GFR نرمال باشد و BMI بین ۲٤ تا ۲۸ باشد و کراتینین خون نرمال باشد	فرد دچار بیماری مزمن کلیوی نخواهد شد	۰/۵٦
قانون ٦	اگر لیپوپروتیین با چگالی پایین نرمال باشد فرد دیابت نداشته باشد و GFR نرمال باشد و BMI بین ۲٤ تا ۲۸ باشد	فرد دچار بیماری مزمن کلیوی نخواهد شد	۰/٤۵
قانون ۷	اگر فشارخون بالا باشد و از سیگار استفاده شود	فرد دچار بیماری مزمن کلیوی خواهد شد	۰/٦۹
قانون ۸	اگر فرد بیماری دیابت داشته باشد و صدمه مغزی و عروقی دیده باشد	فرد دچار بیماری مزمن کلیوی خواهد شد	۰/۵۸

مجموعه داده‌های بیماری مزمن کلیوی پیاده‌سازی شدند. در بین تکنیک‌های مورد بررسی، تکنیک ماشین‌بردار پشتیبان میزان Accuracy بیشتری نسبت به دو تکنیک دیگر داشته است به گونه‌ای که توانسته است نزدیک به ۸۷٪ Accuracy داشته باشد. پس از این تکنیک، شبکه عصبی توانسته است دومین عملکرد را به نام خود ثبت نماید و سپس جنگل تصادفی سومین تکنیک می‌باشد. البته عملکرد این تکنیک از لحاظ نرخ خطای False positive rate نزدیک به دو تکنیک دیگر است. بدترین عملکرد مربوط به Chaid با ۰/۱٦ است (جدول ٤). آستانه قواعد برای میزان Confidence ۵۰٪ هست. بنابراین قواعدی که میزان اطمینان آنها بیشتر از ۵۰٪ است انتخاب شدند. همچنین قواعد انتخابی باید حداقل برای ۲۰ رکورد برقرار باشند، در واقع support آنها حداقل ۲۰ رکورد باشد. مجموعا ۱۷ قاعده استخراج شد که ۱۰ قاعده برتر در جدول ٤ موجود است. همانطور که در جدول ۵ مشاهده می‌نمایید، در قاعده اول بیان شده است که اگر شغل آزاد باشد، وزن بین ۷۱ تا ۹۰ باشد و وضعیت تری‌گلیسرید پرخطر باشد، به احتمال ۷۵٪ فرد دچار بیماری مزمن کلیوی خواهد شد. همچنین قاعده ۲ بیان می‌نماید اگر از نمک در غذا استفاده نماید و سن بین ۵۰ تا ٦۹ باشد و بیماری دیابت داشته باشد به احتمال ۸۲٪ فرد دچار بیماری مزمن کلیوی خواهد شد. اما در بین قوانینی که برای افراد سالم وجود دارد (افرادی که بیماری مزمن کلیوی نخواهند داشت) می‌توان گفت که قانون ۳ بیان می‌کند اگر کراتینین خون نرمال باشد و BMI بین ۲۸ تا ۳۲ باشد، از قلیان و سیگار استفاده نشود و سن بین ۳۰ تا ٤۹ باشد آنگاه به احتمال ۷۲٪ فرد دچار بیماری مزمن کلیوی نخواهد شد (جدول ۵).

بحث

در این پژوهش محدودیت خاصی وجود نداشت و سعی بر این بود که ارتباط بین ریسک فاکتورهای مختلف بر بیماری مزمن کلیوی بررسی شود. چهار تکنیک ماشین‌بردار پشتیبان، شبکه عصبی، CHAID و جنگل تصادفی به همین منظور مورد استفاده قرار گرفتند. تکنیک ماشین‌بردار پشتیبان عملکرد بهتری نسبت به سایر تکنیک‌ها داشت و توانست در بیش از ۸۵٪ مواقع تشخیص صحیحی داشته باشند که عملکرد مناسبی است. با توجه به اینکه ۳۲ صفت غیرعددی استفاده شده است، می‌توان گفت در حالتی که تعداد صفات زیاد است و نوع صفات غیرعددی است، عملکرد تکنیک ماشین‌بردار

پشتیبان نسبت به سایر تکنیک‌ها بهتر است. البته با توجه به اینکه صفات غیرعددی غالبا ترتیبی هستند و می‌توان آنها را تبدیل به انواع عددی نمود بنابراین عملکرد ماشین بردار پشتیبان قابل توجیه است. البته تکنیک شبکه عصبی عملکرد نزدیکی نسبت به تکنیک ماشین‌بردار پشتیبان از لحاظ میزان Accuracy داشت و زیر یک صدم با تکنیک ماشین‌بردار پشتیبان اختلاف داشت. برای افراد بیمار، پنج قانون و برای افراد سالم، سه قانون استخراج شده است. افراد بیمار معادل با افراد دچار بیماری مزمن کلیوی کلیه و افراد سالم، افرادی هستند که دچار بیماری مزمن کلیوی کلیه نشده‌اند. مطمئن‌ترین قاعده با میزان اطمینان ۸۲٪ بیان می‌کند که اگر از نمک در غذا استفاده نماید و سن بین ۵۰ تا ٦٩ باشد و بیماری دیابت داشته باشد، فرد دچار بیماری مزمن کلیوی خواهد شد. مصرف نمک به دلیل سدیم بالای آن می‌تواند بر عملکرد کلیه تاثیر منفی بگذارد. همچنین این قانون بیان می‌نماید که افراد مسن بیشتر در معرض ابتلا به بیماری‌های مزمن کلیوی هستند. همچنین بررسی‌های دیگر نشان داده‌اند که سن به عنوان یک فاکتور موثر در بیماری مزمن کلیوی مطرح شده است و افراد مسن نسبت به افراد جوان بیشتر در معرض بیماری‌های کلیوی هستند.[۲۰] از طرف دیگر این قانون به نوعی تاییدکننده ارتباط بین بیماری دیابت و بیماری مزمن کلیوی است. همچنین این پژوهش نشان داده است که در افراد با درجات بالای بیماری مزمن کلیوی که بیماری دیابت هم دارند، افسردگی و مرگ‌ومیر بیشتر خواهد بود.[۲۱] در واقع تاثیر همزمان این دو بیماری بر سلامت افراد قابل توجه است. دومین قانون از لحاظ اطمینان با میزان اطمینان ۷۵٪ بیان می‌کند که که اگر شغل آزاد باشد، وزن بین ۷۱ تا ۹۰ باشد و وضعیت تری‌گلیسیرید پرخطر باشد، به احتمال ۷۵٪ فرد دچار بیماری مزمن کلیوی خواهد شد. این قانون به نوعی ارتباط بین تری‌گلیسرید و بیماری مزمن کلیوی در افراد با وزن زیاد را بیان می‌نماید. Soohoo در پژوهش خود به‌این نتیجه رسید که در صورتی که تری‌گلیسرید در بیماران کلیوی افزایش یابد، میزان پرخطر بودن این بیماری افزایش خواهد یافت و وضعیت بیماری فرد بدتر خواهد شد.[۲۲] سومین قانون از لحاظ اطمینان با میزان اطمینان ۷۲٪ بیان می‌کند که اگر کراتینین خون نرمال باشد، BMI بین ۲۸ تا ۳۲ باشد، از قلیان و سیگار استفاده نشود و سن بین ۳۰ تا ٤٩ باشد آنگاه به احتمال ۷۲٪ فرد دچار بیماری مزمن کلیوی نخواهد شد. در واقع تاثیر عدم استفاده از سیگار در جوانان در کاهش ریسک ابتلا به بیماری‌های کلیوی را بیان می‌نماید. Yacoub در پژوهش خود مشخص نمود که افراد سیگاری نسبت به غیرسیگاری بیشتر در معرض بیماری‌های کلیوی هستند.[۲۳] بنابراین عدم مصرف سیگار می‌تواند به‌عنوان یک عامل موثر در پیشگیری از بیماری‌های کلیوی باشد. همچنین طبق سومین قانون از لحاظ میزان اطمینان، ریسک ابتلا افراد با شغل آزاد به بیماری‌های کلیوی بالا است. در واقع کارگران شاغل در بخش‌های مختلف، رانندگان تاکسی و اتوبوس، رانندگان وسایل نقلیه سنگین، نانوایان، قنادها و آشپزها نیز باتوجه به‌کار در محیط گرم بیشتر در معرض ابتلا به ناراحتی‌های کلیوی در اثر کم آبی بدن هستند. نتایج پیش‌بینی بیماری کلیوی با استفاده از تکنیک‌های داده‌کاوی نشان داد که تکنیک ماشین‌بردار پشتیبان در حالتی که صفات غیرعددی ولی ترتیبی هستند عملکرد بهتری نسبت به سه تکنیک شبکه عصبی، جنگل تصادفی و CHAID داشته است. البته تکنیک شبکه عصبی عملکرد نزدیکی از لحاظ میزان Accuracy نسبت به تکنیک ماشین‌بردار پشتیبان داشته است و تقریبا کمتر از یک درصد اختلاف دارند. همچنین قواعد مستخرج نشان داد که استفاده از نمک در کنار بیماری دیابت می‌تواند منجر به بیماری مزمن کلیوی شود. بیماری زمینه‌ای دیابت حتی می‌تواند خطر مرگ‌ومیر بیماران کلیوی را هم افزایش دهد که خود این موضوع قابل تامل است. افزون‌براین، افراد مسن هم باید بیشتر مراقب سلامتی خود باشند تا کمتر در معرض بیماری مزمن کلیوی قرار گیرند.

سپاسگزاری: این مقاله حاصل از طرح تحقیقاتی تحت عنوان "پیش‌بینی ابتلا به بیماری مزمن کلیوی در شهر اصفهان" با استفاده از داده‌کاوی مصوب دانشگاه علوم پزشکی اصفهان در سال ۱٤۰۰ به کد اخلاق IR.MUI.NUREMA.REC.1400.079 می‌باشد که با حمایت دانشگاه علوم پزشکی اصفهان اجرا شده است.

References

1. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. *The lancet* 2017;389(10075):1238-52.
2. Levey AS, Becker C, Inker LA. Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: a systematic review. *Jama* 2015;313(8):837-46.

3. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. *The lancet* 2017;389(10075):1238-52.
4. Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management: a review. *Jama* 2019;322(13):1294-304.
5. Henry BM, Lippi G. Chronic kidney disease is associated with severe coronavirus disease 2019 (COVID-19) infection. *Int Urol Nephrol* 2020(6):1139-4.
6. Nasser IM, Abu-Naser SS. Lung Cancer Detection Using Artificial Neural Network. *Int J Eng Inform Sys (IJEAIS)*. 2019;3(3):17-23.
7. Bagherian H, Haghjooy Javanmard S, Sharifi M, Sattari M. Using data mining techniques for predicting the survival rate of breast cancer patients: a review article. *Tehran Univ Med J* 2021;79(3):176-86
8. Subasi A, Alickovic E, Kevric J. Diagnosis of chronic kidney disease by using random forest. CMBEBIH 2017: Springer; 2017. p. 589-94.
9. Gharibdousti MS, Azimi K, Hathikal S, Won DH, editors. Prediction of chronic kidney disease using data mining techniques. IIE Annual Conference Proceedings; 2017: Institute of Industrial and Systems Engineers (IISE).
10. Akben S. Early stage chronic kidney disease diagnosis by applying data mining methods to urinalysis, blood analysis and disease history. *IRBM* 2018;39(5):353-8.
11. Almansour NA, Syed HF, Khayat NR, Altheeb RK, Juri RE, Alhiyafi J, et al. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Comput Biol Med* 2019;109:101-11.
12. Vidal EJ, Alvarez D, Martinez-Velarde D, Vidal-Damas L, Yuncar-Rojas KA, Julca-Malca A, et al. Perceived stress and high fat intake: A study in a sample of undergraduate students. *PLoS One* 2018;13(3):e0192827.
13. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst Appl* 2019;134:93-101.
14. Pisner DA, Schnyer DM. Support vector machine2020:101-21.
15. Albawi S, Mohammed TA, Al-Zawi S , editors. Understanding of a convolutional neural network. In2017 International Conference on Engineering and Technology (ICET); 2017:IEEE.
16. Yin M, Wortman Vaughan J, Wallach H. Understanding the effect of accuracy on trust in machine learning models. InProceedings of the 2019 chi conference on human factors in computing systems; 2019:1-12.
17. Al Jallad K, Aljnidi M, Desouki MS. Anomaly detection optimization using big data and deep learning to reduce false-positive. *J Big Data* 2020;7(1):1-2.
18. Cheng F, Zhang H, Yuan D, Sun M. Leveraging semantic segmentation with learning-based confidence measure. *Neurocomputing* 2019;329:21-31.
19. Hikmawati E, Surendro K, editors. How to Determine Minimum Support in Association Rule. InProceedings of the 2020 9th International Conference on Software and Computer Applications 2020:6-10.
20. Raman M, Green D, Middleton RJ, Kalra PA. Comparing the impact of older age on outcome in chronic kidney disease of different etiologies: a prospective cohort study. *J Nephrol* 2018;31(6):931-9.
21. Young BA, Von Korff M, Heckbert SR, Ludman EJ, Rutter C, Lin EH, et al. Association of major depression and mortality in Stage 5 diabetic chronic kidney disease. *Gen Hosp Psychiatry* 2010;32(2):119-24.
22. Soohoo M, Moradi H, Obi Y, Kovesdy CP, Kalantar-Zadeh K, Streja E. Serum triglycerides and mortality risk across stages of chronic kidney disease in 2 million US veterans. *J Cin lipidol* 2019;13(5):744-53.e15.
23. Yacoub R, Habib H, Lahdo A, Al Ali R, Varjabedian L, Atalla G, et al. Association between smoking and chronic kidney disease: a case control study. *BMC Public Health* 2010;10(1):1-6.

TUMJ Tehran University Medical Journal, September 2021; Vol. 79, No. 6: 459-467 Original Article

Prediction of chronic kidney disease in Isfahan with extracting association rules using data mining techniques

Firouzeh Moeinzadeh Ph.D.[1]
Mohammad Hossein Rouhani Ph.D.[2]
Mojgan Mortazavi Ph.D.[1]
Mohammad Sattari Ph.D.[3*]

1- Isfahan Kidney Diseases Research Center, Isfahan University of Medical Sciences, Isfahan, Iran.
2- Department of Community Nutrition, School of Nutrition and Food Science, Isfahan University of Medical Sciences, Isfahan, Iran.
3- Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran.

[*]Corresponding author: Floor two, School of Management and Medical Informatics, Isfahan University of Medical Sciences, Hezar Jarib St., Isfahan, Iran.
Tel: +98-31-37925152
E-mail: msattarimng.mui@gmail.com

Abstract

Received: 24 Apr. 2021 Revised: 01 May 2021 Accepted: 14 Aug. 2021 Available online: 23 Aug. 2021

Background: Millions of deaths occur around the world each year due to lack of access to appropriate treatment for chronic kidney disease patients. Given the importance and mortality rate of this disease, early and low-cost prediction is very important. The researchers intend to identify chronic kidney disease through the optimal combination of techniques used in different stages of data mining.

Methods: This cross-sectional research was conducted from February 1999 to May 2014. The used data set included 4145 samples and 32 attributes, where Each sample corresponded to a patient and each attribute corresponded to the demographic and clinical traits. There were several eligibility criteria for the patients for clinical testing. These criteria for the clinical testing included having 18 years of age and older, living in Isfahan city, willing to participate in the study, lack of fever and cold during laboratory tests, no strenuous exercise 48 hours before laboratory tests, and fasting. Individuals who had an incomplete questionnaire or were unwilling to perform accurate tests were excluded from the study. The target variable is kidney disease, the values of which include sick and healthy. Four data mining techniques have been used in the dataset. These techniques are support vector machine (SVM), random forest (RF), artificial neural network (ANN) and Chi-square automatic interaction detection (CHAID).

Results: Accuracy is the evaluation criteria for comparing available data mining methods. Based on the accuracy criterion, the support vector machine performed better than other techniques (random forest, neural network and CHAID). The best rule is that if the patients consume salt in their diet, their age is between 50 and 69, and they have diabetes. they are 82% more likely to develop chronic kidney disease.

Conclusion: The derived rules also showed that if we use salt and we have diabetes, we are at the risk of developing chronic kidney disease. Moreover, having diabetes can increase the risk of mortality in chronic kidney patients. Aged people should also be more careful about getting chronic kidney disease. Because, they are more prone to develop chronic kidney disease.

Keywords: chronic kidney disease, data mining, prediction.