تخمین میزان بقاء پیوند کلیه با استفاده از داده کاوی

لیلا شاهمرادی[۱](Ph.D)، مصطفی لنگری‌زاده[۲](Ph.D)، غلامرضا پورمند[۳](Ph.D)، زیبا اقصایی‌فرد[۳](Ph.D)، علیرضا برهانی[*۱](M.Sc)

۱- گروه مدیریت اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران

۲- گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع‌رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران

۳- مرکز تحقیقات اورولوژی بیمارستان سینا، دانشگاه علوم پزشکی تهران، تهران، ایران

چکیده

هدف: نارسایی کلیه از مشکلات پرهزینه جوامع انسانی به‌شمار می‌رود و استفاده از درمان‌های جایگزین در حوزه کلیه در جهان و ایران رو به افزایش می‌باشد. بقاء یکی از حوزه‌های پیش‌آگهی پزشکی است و داده کاوی فرایند کشف روابط و الگوهای مناسب در داده‌هاست که به عنوان روشی کارامد برای تحلیل بقاء شناخته می‌شود. هدف مطالعه حاضر، پیش‌بینی بقاء پیوند کلیه بیمار بر اساس متغیرهای پیش از پیوند کلیه می‌باشد.

مواد و روش‌ها: به منظور شناسایی عوامل موثر در پیش‌بینی بقاء پیوند از طریق پرسش‌نامه‌ای محقق ساخته، نیازسنجی اطلاعاتی از متخصصان به‌عمل آمد سپس با استفاده از اطلاعات حاصل از تجزیه و تحلیل پرسش‌نامه‌ها، چک‌لیستی تهیه و داده‌های ۵۱۳ پرونده بیمار کلیوی از مرکز تحقیقات اورولوژی سینا، استخراج شد. در مرحله بعد با پیروی از متدلوژی CRISP به منظور تحلیل و داده کاوی از نرم‌افزار IBM SPSS Modeler 14.2 و الگوریتم C5.0 استفاده شد.

یافته‌ها: در این پژوهش متغیرهای شاخص توده بدنی، بیماری مرحله نهایی کلیه و مدت زمان دیالیز بیمار به عنوان موثرترین فاکتورهای دخیل در بقاء پیوند ارزیابی شدند و قوانین استخراج شده از مدل می‌توانند به عنوان الگویی برای پیش‌بینی بقاء پیوند کلیه پیش از عمل جراحی استفاده شوند. صحت مدل ایجاد شده، ٪۹۶،۷۷ تخمین زده شد.

نتیجه‌گیری: میزان بالای صحت مدل C5.0 نشان از قدرت پیش‌بینی بقاء آن دارد. در این مطالعه موثرترین فاکتورهای بقاء پیوند کلیه شناسایی شدند و با توجه به قوانین ایجاد شده برای یک نمونه جدید با ویژگی‌های مشخص، می‌توان به پیش‌بینی دوام پیوند بیمار بر اساس سال پرداخت.

واژه‌های کلیدی: داده کاوی، تحلیل بقاء، پیوند کلیه، پیش‌بینی، متدلوژی CRISP

مقدمه

پیوند عضو به عنوان بهترین درمان جایگزین نارسایی عضو به‌منظور کمک به بازتوانی بیمار انجام می‌شود. به گفته WHO در سال ۲۰۰۸ بر روی ۱۰۴ کشور که نماینده‌ی ۹۰٪ از پیوندهای صورت گرفته در جهان هستند، در حدود ۱۰۰۸۰۰ پیوند عضو انجام گرفته است که از این بین ۶۹۴۰۰ پیوند، مربوط به پیوند کلیه (۴۶٪ از اهداکنندگان زنده) می‌باشد که بیش از ۶۸٪ پیوندهای انجام شده در جهان را به خود اختصاص می‌دهد [۱]. بیماری مرحله نهایی کلیه ESRD (End Stage Renal Disease) در ایران ۳۶۰ مورد در هر یک میلیون نفر تخمین زده شده است و مطالعات، حاکی از روند رو به رشد آمار مبتلایان به این عارضه می‌باشد [۳،۲].

* نویسنده مسئول، تلفن: ۰۹۱۲۵۳۱۰۶۶۵ Email: a-borhani@razi.tums.ac.ir

تاریخ دریافت: ۱۳۹۵/۱/۳۱ تاریخ پذیرش: ۱۳۹۴/۹/۲۱

شیوع و بروز ESRD به ترتیب ۳۵۷ و ۵۷ مورد در هر یک میلیون نفر در سال گزارش شده است [۴]. در ایران تا پایان سال ۱۳۹۲ جمعیت بیماران مزمن کلیوی با درجه نارسایی پیشرفته کلیه که تحت درمان با یکی از روش‌های جایگزین کلیه بوده‌اند به ۵۰۰۰۰ نفر رسید [۵]. عمده‌ترین دلایل ESRD که در ممالک پیشرفته زمینه‌ساز دیالیز و یا پیوند کلیه می‌باشند عبارتند از دیابت، نفرواسکلروز ناشی از فشار خون، گلومرولونفریت مزمن و بیماری کلیه پلی‌کیستیک نوع بالغین Adult Polycystic Kidney Disease (APKD) [۶].

انتخاب دهنده کلیه بر اساس پیش‌بینی میزان موفقیت پیوند صورت می‌گیرد. اما علی‌رغم همه توجهاتی که در پیوند کلیه انجام می‌شود، عوارضی چون واکنش‌های دفع پیوند، نکروز حاد توبولر Acute tubular necrosis (ATN)، عوارض ناشی از جراحی، بیماری‌های عفونی و مسمومیت کلیوی داروهای دریافت شده، شانس بقاء پیوند را تهدید می‌کند [۷]. یکی از انواع داده‌ها که مورد توجه پزشکان است، فاصله زمانی تا وقوع حوادثی مانند مرگ و میر است. یعنی توجه به گروهی از افراد به‌طوری که پس از مدتی برای هر کدام از آن‌ها یک نقطه زمانی به نام شکست یا وقوع حادثه تعریف می‌گردد. از آن‌جایی که این روش‌ها در ابتدا غالباً برای مطالعات مرگ و میر به کار برده می‌شد و به این منظور طراحی گردیده بود، به همین جهت نام " تحلیل زمان بقاء" بر آن نهاده شده است [۸].

امروزه داده‌ها عمده‌ترین دارایی سازمان‌های سلامت بوده و موفقیت سازمان‌های سلامت در گروی جمع‌آوری، ذخیره و تحلیل آن‌هاست [۹]. با رشد سریع در اندازه و تعداد پایگاه داده‌ها، کاوش دانش، قواعد یا اطلاعات سطح بالا از داده‌ها به منظور پشتیبانی از تصمیم‌گیری‌ها و پیش‌بینی رفتارهای آتی، ضروری به نظر می‌رسد [۱۰]. از این رو برای تبدیل این ارزش بالقوه به اطلاعات استراتژیک، بسیاری از سازمان‌ها به داده کاوی روی آورده‌اند [۹] داده کاوی، جستجوی خودکار منابع داده‌ای بزرگ، جهت یافتن الگوهاییست که تحلیل‌های ساده آماری قادر به انجام آن نیستند [۱۱].

"دسته‌بندی" (Classification) یکی از روش‌های پیشگویانه برای تخمین میزان رخداد یک حادثه می‌باشد. درخت تصمیم بر اساس قواعد تصمیم‌گیری، برای پیش‌بینی و دسته‌بندی مورد استفاده قرار می‌گیرد. این روش دارای مزایای متعددی دارد. از جمله این‌که بعد از ساخت درخت می‌توان علت استنتاج قواعد به‌دست آمده را مشاهده نمود و هم‌چنین به شناخت بهتر فیلدهای با اهمیت پرداخت. الگوریتم‌های مختلفی برای آنالیز دسته‌بندی به روش درخت تصمیم همچون C5.0، C&RTree، CHAID و QUEST موجود هستند که اساساً یک فرآیند شبیه به هم را پیاده‌سازی می‌کنند [۱۲].

الگوریتم C5.0 که برای ساخت درخت تصمیم یا مجموعه قوانین استفاده می‌شود، روش خاصی برای بهبود نرخ دقت پیش‌بینی تحت عنوان ترقی دادن (Boosting) دارد. این روش با ساخت مدل‌های چندگانه به صورت متوالی کار می‌کند که می‌تواند دقت را به مقدار قابل توجهی در مدل C5.0 بهبود دهد، ولی این امر نیازمند آموزش طولانی‌تری است.

پیش‌آگهی پزشکی (Medical Prognosis) رشته‌ای در علم پزشکی و شامل برآورد عوارض و عود بیماری و هم‌چنین پیش‌بینی بقاء بیماران می‌باشد. تحلیل بقاء یکی از حوزه‌های پیش‌آگهی پزشکی است که شامل روش‌های مختلف برای برآورد بقاء بیمار مبتلا از زمان شروع بیماری تا یک دوره خاص است [۱۳]. تا کنون مطالعات گسترده‌ای در زمینه پیش‌بینی بقاء بیماران مبتلا به سرطان، بیماری‌های قلبی، بیماری‌های مزمن کلیه، سوختگی و غیره با استفاده از مدل‌های داده کاوی صورت گرفته است. شبکه‌های عصبی مصنوعی، تحلیل رگرسیون، درخت تصمیم و ماشین‌بردار پایه، الگوریتم‌هایی هستند که به وفور برای پیش‌بینی بقاء مورد استفاده قرار گرفته‌اند [۱۴].

هدف از این پژوهش کمک به پیش‌بینی مدت زمان بقاء پیوند کلیه بیمارانی است که در معرض نارسایی کلیه قرار دارند تا تخمینی از میزان بقاء پیوند در اختیار ایشان قرار دهد.

مواد و روش‌ها

در این بخش با نگاهی به استاندارد شش مرحله‌ای CRISP (Cross Industry Process for Data Mining) به کشف دانش در میان داده‌های بیماران پیوند کلیه پرداخته می‌شود. در مطالعه حاضر هر یک از این فازها که خود شامل زیربخش‌هایی می‌شوند، و روند کشف دانش در میان داده‌ها را بیان می‌کنند در شکل ۱ نشان داده شده‌اند. در ادامه به ذکر این مراحل به صورت جزئی پرداخته می‌شود.

تعریف مسئله به نوعی ارزیابی شرایط فعلی، تعریف اهداف داده کاوی و ایجاد یک برنامه زمان‌بندی پروژه داده کاوی می‌باشد.

گام شناخت داده‌ها، نیازمندی‌های داده‌ای را مورد مطالعه قرار می‌دهد. جامعه آماری مطالعه حاضر را گیرندگان و دهندگان پیوند کلیه مرکز تحقیقات اورولوژی بیمارستان سینا تهران تشکیل می‌دادند و مرحله نخست به منظور پیش‌بینی مدت زمان بقاء، یافتن فاکتورهای تاثیرگذار در پیوند کلیه بود.

ابتدا پارامترهای تاثیرگذار در پیش‌بینی بقاء پیوند کلیه بر اساس مطالعه کتب و متون تخصصی تعیین شد و پرسش‌نامه‌ای محقق ساخته جهت نیازسنجی اطلاعاتی از پزشکان متخصص اورولوژی و نفرولوژی در رابطه با اقلام داده‌ای پیش‌بینی بقاء پیوند کلیه طراحی گردید. این پرسش‌نامه از دو قسمت (در کل ۴۶ سؤال) تشکیل شده بود و دربرگیرنده مشخصات فردی (۷ سؤال) و اقلام داده‌ای مورد نیاز به منظور پیش‌بینی بقاء پیوند کلیه (۳۹ سؤال) بود. در کنار سؤالات بسته پرسش‌نامه، یک سؤال باز نیز در هر بخش مطرح گردید تا پژوهشگر سایر نظرات پاسخ‌دهندگان را جویا شود.

روایی پرسش‌نامه با استفاده از روش روایی صوری و محتوا و کسب نظرات متخصصین نفرولوژی و اورولوژی (۷ نفر از اعضای هیئت علمی و حداقل دارای سه سال سابقه کار) تأئید گردید. پایایی پرسش‌نامه نیز از با روش آزمون-بازآزمون سنجیده شد و ضریب همبستگی آن ۹۲٪ به‌دست آمد.

سپس داده‌های به دست آمده از پرسش‌نامه با استفاده از آمار توصیفی و گزارش توزیع فراوانی تحلیل شدند. حاصل این پرسش‌نامه چک‌لیستی بود که از این طریق اقلام داده‌ای مورد نیاز از پرونده بیماران استخراج گردید. چک‌لیست مذکور از سه بخش اصلی شامل متغیرهای ورودی گیرنده کلیه، متغیرهای ورودی دهنده کلیه و متغیر خروجی مورد نیاز تشکیل شده بود. در این مرحله نمونه‌گیری انجام نگرفت و تمامی پرونده‌های بیماران پیوند شده از ابتدای فروردین سال ۱۳۸۶ تا پایان شهریور ۱۳۹۲ توسط مسئولین مرکز اورولوژی بیمارستان سینا در اختیار پژوهشگر قرار گرفت که پس از حذف پرونده‌های ناقص، فرایند داده کاوی بر روی ۵۱۳ فقره پرونده انجام پذیرفت.

داده‌های خام اغلب به ندرت برای داده کاوی قابل استفاده‌اند بنابراین قبل از آنالیز نهایی توسط الگوریتم‌های داده کاوی، باید آن‌ها را پردازش کرد. آماده‌سازی داده یکی از مهم‌ترین و اغلب زمان برترین جوانب پروژه‌های داده کاوی است. هنگامی که منابع در دسترس داده مشخص شدند، بایستی داده‌ها را از آن‌ها انتخاب (Selecting) و پاک‌سازی (Cleaning) کرد و در قالب مورد نظر قرار داد [۱۱].

پس از تجزیه و تحلیل‌های متعدد یک سری از رکوردهای داده به دلیل نقص در اطلاعات و یا غلط بودن اطلاعات حذف گردید. در این بین داده‌های پرت (Outlier) نیز مشخص گردیدند تا نتایج نهایی دچار کم‌ترین میزان خطا باشند. هم‌چنین برخی از رکوردهایی که مقدار سن اهداکننده در آن مشخص نبود نیز به روش میانگین‌گیری مقدارگذاری شدند. از ترکیب دو فیلد قد و وزن با استفاده از فرمول شاخص توده بدنی میزان BMI= Weight(kg) / (Height(m)) ^{2} BMI برای هر گیرنده کلیه پیوندی مشخص گردید تا در مدل‌سازی مورد استفاده قرار گیرد.

گره PCA/Factor روشی قدرتمند برای کاهش داده‌ها از طریق کاهش ابعاد آن‌ها دارد. این روش ترکیبات خطی از فیلدهای ورودی را ایجاد می‌کند و این ترکیبات خطی بهترین تعریف از کل فیلدهاست. تحلیل فاکتور تلاش می‌کند مفاهیم یا فاکتورهایی را تعیین کند که الگوی همبستگی درون مجموعه‌ای از فلیدها را توضیح می‌دهد. در این مرحله از

پیش‌پردازش اطلاعات، سوال ابتلاء بیمار پیوندی به بیماری‌هایی چون سنگ کلیه، سرطان کبد یا غیره به عنوان فاکتور کم‌اهمیت شناخته و حذف شد و علت آن نیز به دلیل نقص در ثبت این موارد در پرونده بیماران به نظر می‌رسد.

شکل ۱. گام‌های روش استاندارد CRISP و مدل پیشنهادی

گره دیگری که به منظور کاهش بعد و انتخاب ویژگی‌ها در پیش‌پردازش در این زمینه مورد استفاده قرار گرفت، Feature Selection بود. در غربالگری رکوردها و پیش‌بینی‌کننده‌ها این گره، مانند گره PCA/Factor، ابتلاء بیمار به بیماری‌های سنگ کلیه، سرطان کبد یا غیره را زاید تشخیص داد. که در نتیجه تصمیم بر آن شد که این مورد از ستون‌های ورودی به عنوان پیش‌بینی‌کننده حذف گردد. از فیلدهای خروجی در چک‌لیست، موارد "سلامت بیمار پیوندی در زمان تماس" و "وضعیت کلیه بیمار پیوندی" به دلایل عدم پاسخ مناسب علمی از سوی بیماران حذف گردیدند و تنها فاکتور خروجی مورد بحث، مدت زمان بقاء کلیه پس از عمل پیوند می‌باشد.

در نتیجه پس از پالایش داده‌ها و وزن‌دهی به فاکتورهای موثر از طریق نرم‌افزار IBM SPSS Modeler 14.2 (جدول ۱)، تعداد ۵۱۳ پرونده بیمار گیرنده کلیه (و به همین تعداد دهنده کلیه) به عنوان نمونه پژوهش جهت ارزیابی انتخاب شدند که توسط الگوریتم درخت تصمیم C5.0 مدل‌سازی و قوانین آن استخراج شد (پیوست ۱).

به منظور ارزیابی مدل‌های طبقه‌بندی، از نمودار Gains استفاده می‌شود به این ترتیب که از پاسخ واقعی و پیش‌بینی مدل، جدولی ساخته و نمودار آن رسم می‌شود که محور عمودی پاسخ واقعی و محور افقی پیش‌بینی مدل است در ادامه نتایج ارزیابی مدل با داده‌های آموزش و داده‌های آزمون به صورت بصری نمایش داده می‌شود.

به منظور بررسی صحت مدل، داده‌ها به دو بخش آموزش (۷۰٪) و آزمون (۳۰٪) تقسیم شدند. به وسیله داده‌های بخش آموزش، مدل نهایی ایجاد گردید و داده‌های بخش آزمون، مدل به‌دست آمده در مرحله اول را آزمودند. شاخص‌های مختلفی برای ارزیابی صحت روش‌های دسته‌بندی وجود دارد که می‌توان از این دست حساسیت (Sensitivity)، ویژگی (Specificity) و صحت (Accuracy) را نام برد. میزان صحت یک روش دسته‌بندی بر روی مجموعه داده‌های آموزشی، درصد مشاهداتی از مجموعه آموزش است که به درستی توسط روش مورد استفاده، دسته‌بندی شده است. برای محاسبه این شاخص داده‌های آزمون استفاده شدند.

در این مطالعه به منظور محاسبه میزان صحت مدل از ماتریس اغتشاش (Confusion Matrix) استفاده شد. این ماتریس ابزار مفیدی برای تحلیل چگونگی عملکرد روش دسته‌بندی در تشخیص داده‌ها یا مشاهدات دسته‌های مختلف

است. اگر داده‌ها در M دسته قرار گرفته باشند، یک ماتریس دسته‌بندی جدولی با حداقل اندازه M*M است. حالت ایده‌آل این است که بیش‌تر داده‌های مرتبط به مشاهدات روی قطر اصلی ماتریس قرار گرفته باشند و مابقی مقادیر ماتریس صفر یا نزدیک به صفر باشند [۱۵–۱۷].

منظور از برچسب مثبت، مطابق با جدول ۲، یکی از برچسب دسته‌های (کم‌تر از ۱ سال، کم‌تر از ۲ سال، کم‌تر از ۳ سال، کم‌تر از ۴ سال، کم‌تر از ۵ سال، کم‌تر از ۶ سال و بیش‌تر از ۶ سال) است و برچسب منفی، کل مجموعه داده‌ها به‌جز برچسب دسته مثبت می‌باشد.

$$\text{حساسیت} = \frac{\text{تعداد داده‌های برچسب مثبتی که درست دسته‌بندی شده‌اند}}{\text{کل تعداد داده‌های مثبت}}$$

$$\text{ویژگی} = \frac{\text{تعداد داده‌های برچسب منفی که درست دسته‌بندی شده‌اند}}{\text{کل تعداد داده‌های منفی}}$$

$$\text{صحت} = \frac{\text{تعداد داده‌های منفی}}{\text{کل تعداد داده‌ها}}\text{ویژگی} + \frac{\text{تعداد داده‌های مثبت}}{\text{کل تعداد داده‌ها}}\text{حساسیت}$$

جدول ۱. داده‌های ورودی و نوع آن‌ها پس از پاکسازی

ردیف	مشخصه	توضیحات	اسمی	بازه عددی
۱	ESRD	علت نارسایی کلیه	دیابت گلومرونفریت فشارخون پلی کیستیک سنگ کلیه ناشناخته دیگر عوامل	
۲	R-AGE	سن گیرنده کلیه (سال)		[۹–۶۸]
۳	BMI	شاخص توده بدنی گیرنده کلیه		[۱۶–۳۸]
۴	D-Time	مدت زمان دیالیز (ماه)		[۰–۹۸]
۵	R-SEX	جنسیت گیرنده کلیه	مرد زن	
۶	DIALYSIS-TYPE	نوع دیالیز	بدون دیالیز دیالیز خونی دیالیز صفاقی	
۷	PANEL-TEST	تست پانل		[۰–۱۰۰]
۸	FIRST-TRANSPLANT	سابقه پیوند	بله خیر	
۹	RELATIONSHIP	نوع ارتباط گیرنده و دهنده کلیه (خویشاوند، غیرفامیل، جسد)	غیرخویشاوند خویشاوند پیوند از جسد	
۱۰	D-AGE	سن دهنده کلیه (سال)		[۱۷–۵۸]
۱۱	D-SEX	جنسیت دهنده کلیه	مرد زن	

جدول ۲. برچسب دسته خروجی مدل

کمتر از یک سال	بقاء کلیه پیوندی بین ۰ تا ۱۲ ماه
کمتر از دو سال	بقاء کلیه پیوندی بین ۱۲ تا ۲۴ ماه
کمتر از سه سال	بقاء کلیه پیوندی بین ۲۴ تا ۲۶ ماه
کمتر از چهار سال	بقاء کلیه پیوندی بین ۳۶ تا ۴۸ ماه
کمتر از پنج سال	بقاء کلیه پیوندی بین ۴۸ تا ۶۰ ماه
کمتر از شش سال	بقاء کلیه پیوندی بین ۶۰ تا ۷۲ ماه
بیشتر از شش سال	بقاء کلیه پیوندی بیشتر از ۷۲ ماه

نتایج

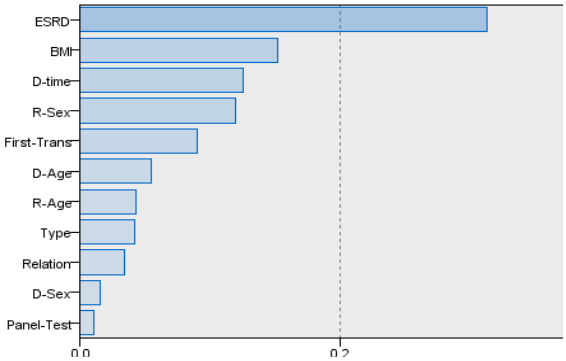
داده کاوی یکی دیگر از مراحل کلیدی در فرایند کشف دانش است [۱۸]. در این پژوهش مدل‌سازی با استفاده نرم‌افزار IBM SPSS Modeler 14.2 صورت یافت و درخت تصمیم C5.0 با ورودی‌های مختلف مورد آزمون قرار گرفت. فیلدهای ورودی، مقادیر به دست آمده از داده‌های بیماران و خروجی، تعداد سال‌های بقاء پیوند کلیه می‌باشد. برچسب دسته مدت زمان بقاء پیوند کلیه در مدل ایجاد شده در جدول ۲ بیان شده است.

با استفاده از نرم‌افزار IBM SPSS Modeler 14.2 فاکتورهای ورودی جدول ۱، به ترتیب شکل ۲ اولویت‌بندی شده‌اند.

همان‌طور که از شکل ۲ قابل استنباط است، فاکتورهای تاثیرگذار بر روی مدت زمان بقاء پیوند توسط مدل بهینه شده، به ترتیب اهمیت عبارتند از: بیماری مرحله نهایی کلیه، شاخص توده بدنی، مدت زمان دیالیز بیمار پیوندی، جنسیت گیرنده کلیه، سابقه قبلی پیوند کلیه، سن دهنده کلیه، سن گیرنده کلیه، نوع دیالیز، رابطه دهنده و گیرنده کلیه، جنسیت دهنده کلیه، تست پانل.

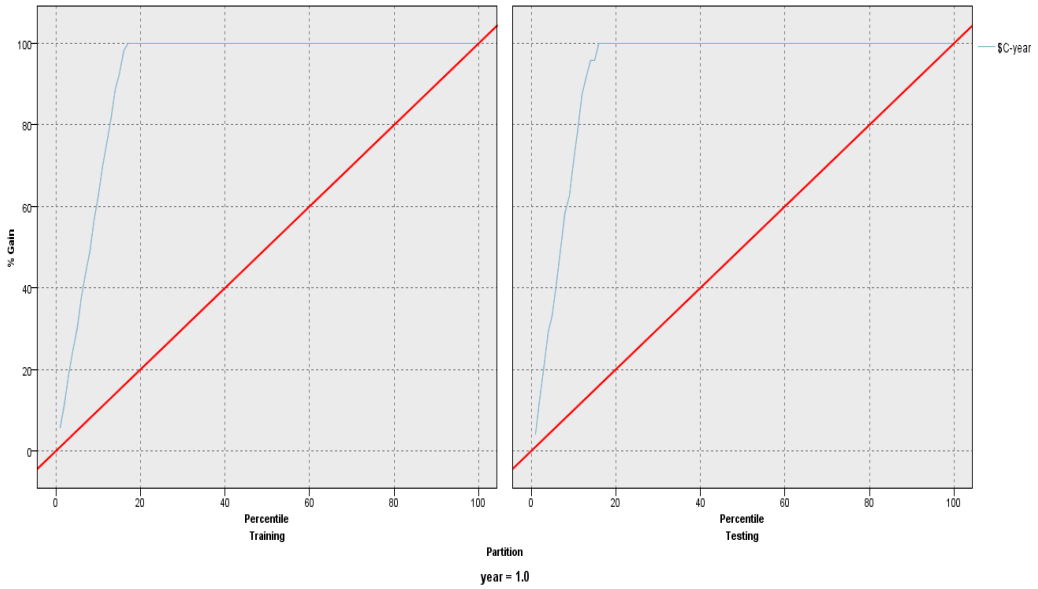
نمودار سمت راست، نمودار Gain حاصل از داده‌های آزمون و نمودار سمت چت، داده‌های آموزش درخت C5.0 را به تصویر کشیده است که محور عمودی پاسخ واقعی و محور افقی پیش‌بینی مدل را (ارزیابی میزان پیش‌بینی بقاء کم‌تر از یک سال) نشان می‌دهد.

ماتریس اغتشاش حاصل از مدل C5.0 در جدول ۳ قابل مشاهده می‌باشد. با استفاده از این ماتریس و روابطی که در بخش ارزیابی بدان اشاره شد، می‌توان صحت، ویژگی و حساسیت مدل به‌دست آمده را با استفاده از داده‌های آزمون سنجید.



شکل۲. فاکتورهای موثر بر بقاء پیوند کلیه به ترتیب اولویت از مدل

جدول ۴، شاخص‌های مذکور برای هر کدام از برچسب دسته‌ها را با استفاده از ماتریس اغتشاش به صورت جداگانه نمایش می‌دهد. همان‌گونه که مشاهده می‌کنید میزان حساسیت، ویژگی و صحت داده‌های آزمون از طریق مدل C5.0 ارائه شده به ترتیب ۹۰/۸۵٪، ۵۲٪ و ۸۷/۲۱٪ محاسبه گردیده است.



شکل ۳. نمودار Gain حاصل از داده‌های آزمون و آزمون درخت C5.0 بقاء کمتر از یک سال

جدول۳. ماتریس اغتشاش حاصل از داده‌های آزمون و آزمون درخت C5.0

مقایسه خروجی بقاء پیوند کلیه بین مقادیر واقعی و پیش بینی شده

	آموزش		آزمایش	
صحیح	304	90.21%	160	90.91%
اشتباه	33	9.79%	16	9.09%
مجموع	337		176	

ماتریس اغتشاش (سطرها مقادیر واقعی را نشان می دهند)

'Partition' = آموزش	1.000000	2.000000	3.000000	4.000000	5.000000	6.000000	7.000000
1.000000	53	0	0	0	0	0	0
2.000000	1	2	1	0	0	0	0
3.000000	1	0	26	1	0	0	0
4.000000	0	0	2	41	3	0	0
5.000000	1	0	0	0	44	2	0
6.000000	2	0	0	2	2	65	7
7.000000	0	0	0	0	0	8	73
'Partition' = آزمایش	1.000000	2.000000	3.000000	4.000000	5.000000	6.000000	7.000000
1.000000	23	0	1	0	0	0	0
2.000000	0	2	0	0	0	0	0
3.000000	0	0	12	1	0	0	0
4.000000	0	0	1	13	3	0	0
5.000000	0	0	0	2	26	2	1
6.000000	1	0	0	0	3	43	0
7.000000	0	0	0	0	0	1	41

بحث و نتیجه‌گیری

در این پژوهش محققین با استفاده از درخت الگوریتم C5.0، به پیش‌بینی احتمال بقاء پیوند بیمار کلیوی را پیش از مبادرت ورزیدن به عمل پیوند، پرداختند. میزان اهمیت هر یک از عوامل موثر بر بقاء پیوند کلیه و روابط بین داده‌ها کشف شد که در شکل ۲، قابل مشاهده می‌باشد. ارزیابی به عمل آمده حاکی از صحت بالای مدل ایجاد شده، داشت.

پاپایونیون و همکاران در مقاله‌ای که به تجزیه و تحلیل ۱۸ پارامتر آزمایشگاهی همچون کلسیم، کلسترول، آهن، گلوکز، کراتینین و غیره بین ۱۰۸ بیمار مبتلا به نارسایی کلیوی

جدول۴. محاسبه حساسیت، ویژگی و صحت برای داده‌های آزمون

C5.0 Model	حساسیت (درصد)	ویژگی (درصد)	صحت (درصد)
کمتر از یک سال	۹۶	۵۰	۹۱,۵
کمتر از دو سال	۱۰۰	۱۰۰	۱۰۰
کمتر از سه سال	۹۲	۳۳	۸۷
کمتر از چهار سال	۷۶	۵۰	۷۳
کمتر از پنج سال	۸۴	۳۳	۷۹
کمتر از شش سال	۹۱	۵۰	۸۷
بیشتر از شش سال	۹۷	۵۰	۹۳
جمع کل	۹۰,۸۵	۵۲	۸۷,۲۱

مرحله نهایی (ESRF) پرداخته شده با استفاده از تجزیه و تحلیل خوشه‌ای که یکی دیگر از الگوریتم‌های داده کاوی می‌باشد، الگو تشابه در میان همه این پارامترها را به‌دست آورده است [۲۴]. از دیگر پژوهش‌های مرتبط با کاوش بین داده‌های بالینی می‌توان به مقاله‌ای تحت عنوان "مدل آنالیز و پیش‌بینی اثربخش از بیماری قلبی با استفاده از روش‌های دسته‌بندی" اشاره کرد که با استفاده از الگوریتم‌های دسته‌بندی داده کاوی درخت تصمیم، ناوی بیزین (Naive Bayes) و شبکه‌های عصبی به پیش‌بینی وقوع سکته با استفاده از متغیرهای مرتبط پرداخته است و برای کاهش ابعاد این عارضه، تعیین می‌کند که چه متغیرهایی نقش بیش‌تری در افزایش سکته‌های قبلی ایفا می‌کنند [۲۵].

پژوهش پاپایونیون تنها از پارامترهای آزمایشگاهی به عنوان ورودی مدل استفاده گرده است، که در مطالعه‌ی پیش رو، بدین سبب که تناسب پارامترهای آزمایشگاهی دهنده و گیرنده‌ی کلیه، پیش شرطِ مبادرت به پیوند می‌باشد، و در صورت عدم این تناسبات، پیوند کلیه انجام نمی‌گردد، این موارد لحاظ نگردیده است.

در مقاله‌ای تحت عنوان پیش‌بینی بیماری کلیوی آلوگرافت مزمن با استفاده از درخت تصمیم که لوفارو و همکاران انجام دادند، از الگوریتم C4.8 و فاکتورهای آزمایشگاهی بیماران پیوندی استفاده گردید که میزان صحت این مدل کم‌تر از ۸۳٪ به دست آمد [۷] در همان سال گریکو و همکاران در پژوهشی با استفاده از درختی دودوئی در ۴ سطح، بقاء یا رد پیوند بیماران کلیوی را پیش‌بینی نمودند. بدین صورت که دو حالت شکست (رد پیوند) یا عدم شکست را در نظر گرفتند و به عنوان فاکتورهای موثر، در صورت عدم رد حاد در ریشه‌ی درخت، در سطح اول رد پیوند مزمن، در سطح دوم عامل تاخیر در کارکرد پیوند، در سطح سوم شاخص توده‌بندی و در سطح آخر (برگ درخت) نتیجه شکست یا عدم شکست را بررسی می‌شد. حساسیت درخت گریکو ۸۸/۲٪ و ویژگی آن ۷۳/۸٪ تخمین زده شد.

در مطالعه حاضر نیز از الگوریتم C5.0 که بهینه‌سازی شده الگوریتم C4.8 می‌باشد استفاده گردید. که میزان بقاء پیوند مدل در هر مورد جدید پیوندی با صحتی برابر با ۹۶,۷۷٪، تخمین زده شد. از سوی دیگر در این مدل می‌بایست فیلد هدف از نوع طبقه‌ای باشد تفاوت مدل مطالعه حاضر با مدل لوفارو و مدل گریکو در خروجی درخت می‌باشد که خروجی درخت دو مطالعه مذکور تنها شکست یا عدم شکست در پیوند را نشان می‌داد اما خروجی درخت مطالعه حاضر دودوئی نبود و در شش حالت مختلف، مدت زمان بقاء پیوند را بیان می‌کرد [۱۹].

مدل‌های یادگیری ماشین برای پیش‌بینی تشخیص بیماری کبد، نام مطالعه‌ای است که منتظری و همکاران در آن، به منظور تشخیص هوشمند بیماری کبد، الگوریتم‌های دسته‌بندی مختلفی چون ناوی بیزین، Trees Random Forest 1NN، AdaBoost و SVM را مورد مقایسه قرار دادند که در نهایت مدل "درختان جنگل تصادفی" دارای بالاترین میزان دقت با ۷۲٪ به عنوان مدل برتر شناخته شد [۲۰].

در مطالعه‌ای که توسط اشرفی و همکاران بر روی ۳۱۶ بیمار پیوند کلیه صورت پذیرفت مشخصات دموگرافیک دهنده و گیرنده پیوند، نوع پیوند، محل پیوند، شاخص توده بدنی گیرنده پیوند و وضعیت دیابت گیرنده پیوند از پرونده‌های بیماران استخراج گردید و مرگ بیمار و یا انتقال بیمار به دیالیز به عنوان نقطه پایان در تحلیل بقاء در نظر گرفته شد [۲۱]. حسن‌زاده و همکاران به منظور تحلیل بقاء ۱۰ ساله پیوند کلیه و تعیین عوامل موثر بر آن، علاوه بر متغیرهای فوق، فاصله زمانی بعد از پیوند تا تولید اولین ادرار (Cold ischemic time)، نسبت دهنده (فامیل و غریبه)، سمت کلیه اهدایی، مدت زمان دیالیز قبل از عمل، مقدار کراتینین زمان ترخیص و مدت زمان بستری در بیمارستان را نیز در مطالعه خویش به‌کار بردند [۲۲]. در ادامه پژوهش حسن‌زاده، این‌بار حشیانی با هدفی دیگر، در مطالعه‌ای گذشته‌نگر به بررسی میزان بقاء پیوند کلیه پرداخت که متغیرهای تحت بررسی در این مطالعه سن و جنس دهنده و گیرنده کلیه بودند

[۲۳]. نقطه قوت مطالعه حاضر در مقایسه با پژوهش‌های انجام شده‌ی فوق روش تعیین داده‌های موثر در پیش‌بینی بقاء است که این بررسی به صورت کاملاً علمی و توسط پرسش‌نامه‌های توزیع شده بین متخصصین نفرولوژیست و اورولوژیست صورت پذیرفت. از جمله مواردی که در تعیین عوامل موثر در پیش‌بینی بقاء می‌توان بدان اشاره نمود، بررسی مجموعه عوامل نارسایی کلیه در مطالعه حاضر می‌باشد که در پژوهش اشرفی تنها یکی از این عوامل (وضعیت دیابت گیرنده) ملحوظ نظر قرار گرفته است.

در دیگر پژوهش که توسط صالح‌نسب و همکاران صورت گرفت [۲۳]، همچون مطالعه حاضر چک‌لیستی تهیه شد و اطلاعات پرونده بیماران کلیوی استخراج شد و مدل‌سازی بر روی آن‌ها صورت گرفت. گرچه صالح‌نسب نیز، در نظر داشته است با استفاده از داده کاوی به استخراج الگوهای موثر در پیش‌بینی بقاء پیوند بپردازد اما به دلیل تفاوت نگرش هدف مطالعه پیش رو در مقابل پژوهش وی و سایر مطالعات فوق، برخی فاکتورهای مورد استفاده چون نوع رژیم دارویی سرکوبگر ایمنی، فاصله زمانی بعد از پیوند تا تولید اولین ادرار، مقدار کراتینین زمان ترخیص و مدت زمان بستری مورد استفاده قرار نگرفتند زیرا هدف از پژوهش حاضر، پیش‌بینی بقاء پیوند، پیش از صورت گرفتن عمل پیوند می‌باشد اما فاکتورهای یاد شده در مطالعات مذکور مربوط به متغیرهای بعد از پیوند کلیه می‌باشند لذا در مطالعه پیش رو، کاربردی ندارند.

با توجه به سنجش میزان اهمیت متغیرهای پیش بین بقاء پیوند در مطالعه حاضر، بیماری مرحله نهایی کلیه، شاخص توده بدنی و مدت زمان دیالیز پیش از پیوند به عنوان موثرترین فاکتورها تعیین شدند که منطبق بر یافته‌های پژوهش‌های گذشته است. با مقایسه تحقیق‌های قبلی در حوزه داده کاوی و بقاء پیوند کلیه، مشخص است که مدل ارائه شده در این مقاله بالاترین صحت را داراست و از دیگر نقاط قوت مطالعه حاضر، اجرای تمام مراحل کشف دانش طبق استاندارد CRISP بود که در سایر پژوهش‌ها بدان اشاره‌ای نشده بود. از سوی دیگر به منظور سهولت استفاده از مدل درختی ایجاد شده، محققین به طراحی و پیاده‌سازی برنامه‌ای کاربردی تلفن همراه تحت دو پلتفرم اندروید و iOS پرداختند که کاربر می‌تواند با وارد کردن فیلدهای ورودی مطابق جدول ۱، میزان بقاء پیوند کلیه پیش‌بینی شده را در قالب یکی از سطرهای جدول ۲ مشاهده نماید.

از کاستی‌های این مطالعه می‌توان به این موارد اشاره کرد که نتایج این تحقیق وابسته به داده‌های تنها یک بیمارستان می‌باشد، پیشنهاد می‌شود برای بررسی بیش‌تر در این زمینه، در مطالعات بعدی از داده‌های مراکز تحقیقاتی دیگر نیز استفاده و نتایج با هم مقایسه شود. از سوی دیگر ناحیه جغرافیایی که موسسه تحقیقاتی در آن واقع شده است، سطح رفاه، طبقه اجتماعی و شغل دهنده و گیرنده کلیه، می‌تواند در میزان بقاء پیوند کلیه موثر باشد که در این پژوهش به علت فقدان این گونه اطلاعات در مدارک پزشکی بیماران، لحاظ نشده است.

تشکر و قدردانی

محقق بر خود لازم می‌داند از همکاری پرسنل محترم مرکز اورولوژی بیمارستان سینا که در این تحقیق مساعدت لازم را داشتند، کمال تشکر و قدردانی را به عمل آورد. هم‌چنین از زحمات و راهنمایی‌های سرکار خانم دکتر دهقانی کمال تشکر را دارم.

منابع

[1] WHO. Transplantation. Documentation center: http://www.who.int/transplantation/publications/en; 2008 [cited 2013 February]; Department of Essential Health Technologies (HSS/EHT/CPR)]. Available from: http://www.who.int/transplantation/gkt/statistics/en./

[2] Mahdavi-Mazdeh M. Why do we need chronic kidney disease screening and which way to go. Iran J Kidney Dis 2010; 4: 275-281. (Persian).

[3] Nedjat S, Montazeri A, Holakouie K, Mohammad K, Majdzadeh R. Psychometric properties of the Iranian interview-administered version of the world health organization's quality of life questionnaire (WHOQOL-BREF): a population-based study. BMC Health Serv Res 2008; 8: 1. (Persian).

[4] Abbaszadeh S, Nourbala M, Taheri S, Ashraf A, Einollahi B. Renal transplantation from deceased donors in Iran. Saudi J Kidney Dis Transplant 2008; 664-668. (Persian).

[5] Mortazavi N. Bright outlook in dialysis technology in Iran. MED LAB Engine Magazine 2014; 157: 75-77. (Persian).

[6] Faraj Zade A. Principles of Urology: TUMS; 1382. (Persian.)

[7] Lofaro D, Maestripieri S, Greco R, Papalia T, Mancuco D, Conforti D, Bonofiglio R. Prediction of chronic allograft nephropathy using classification trees. Transplant Proc 2010; 42: 1130-1133.

[8] ET L. Statistical methods for survival data analysis. 2nd ed. New York: John Wiley Sons Inc; 1992.

[9] Moghadasi H, Hoseini A, Asadi F, Jahanbakhsh M. Data mining and its application in health. Health Inform Manag 2012. (Persian).

[10] Pang-Ning T, Steinbach M, Vipin K. Introduction to Data Mining Addison Wesley 2005.

[11] Hassanzadeh M, Razavi Ebrahimi A. Comparison classificaion of data mining algorithms in medical sciences. Iranian J Med Inform 2012; 2. (Persian).

[12] Alizadeh S, Malek Mohamadi S. Data mining & knowledge discovery step by step with clementine. Tehran Iran Khajeh Nasir Univ 2014. (Persian).

[13] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining method. Artif Intell Med 2004; 27: 1-15.

[14] Liu Zea HI, Media B. Efficient support vector machine method for survival prediction with SEER data in Arabnia advances in computational biology. New York: Springer Science; 2010.

[15] Alizadeh S, Ghazanfari M, B T. Data mining and knowledge discovery. 2nd ed. Tehran Iran: Iran University of Science and Technology; 2011. (Persian).

[16] Han J, M K. Data Mining: Concepts and Techniques. 2 Kaufmann; 2006.

[17] Ameri H, Alizadeh S, Barzegari A. Knowledge extraction of diabetics' data by decision tree method. J Health Administrat 2013; 16: 58-72. (Persian).

[18] Pal NR, L J. Advanced techniques in knowledge discovery and data mining. 1nd ed. New York: Springer Science+Business Media; 2004.

[19] Greco R, Papalia T, Lofaro D, Maestripieri S, Msncuso D, Bonofiglio R. Decisional trees in renal transplant follow-up. Transplant Proc 2010; 42: 1134-1136.

[20] Montazeri M, Montazeri M. Machine learning models for predicting the diagnosis of liver disease. Koomesh 2014; 16: 53-59. (Persian).

[21] Ashrafi M, Hamidi Beheshti M, Shahidi Sh, Ashrafi F. Application of artificial neural network to predict graft survival after kidney transplantation: Reports of 22 years follow up of 316 patients in Isfahan. Tehran Univ Med J 2009; 67: 353-359. (Persian).

[22] Hasan zadeh J, Salahi H, Rajaei far A, Zeighami B, Hashyani A. 10-year survival analysis of its influencing factors in patients with renal transplantation and transplantation from a living donor transplant center Namazi Hospital 2011; 28-39. (Persian).

[23] Almasi Hashiani A, Rajaeefard A, Hassanzade J, Salahi H. Survival analysis of renal Transplantation and its relationship with age and sex. Koomesh 2010; 11: 302-306. (Persian).

[24] Papaioannou A, Karamanis G, Rigas L, Spanos T, Z R. Determination and modelling of clinical laboratory data of healthy individuals and patients with end-stage renal failure. Central Eur J Med 2009; 4: 12.

[25] Sudha A, Gayathri P, Jaisankar N. Effective analysis and predictive model of stroke disease using classification methods. Int J Computer Appl 2012; 43: 26-31.

Estimating survival rate of kidney transplants by using data mining

Leila Shahmoradi (Ph.D)[1], Mostafa Langarizadeh (Ph.D)[2], Gholamreza Pourmand (Ph.D)[3], Ziba Aghsaei fard (Ph.D)[3], Alireza Borhani (M.Sc)[*1]

[1]- Dept. of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran

2- Dept. of Health Information Management, School of Health Management and Information Science, Iran University of Medical Sciences, Tehran, Iran

3- Urology Research Center, Tehran University of Medical Sciences, Tehran, Iran

(Received: 19 Apr 2016; Accepted: 11 Dec 2016)

Introduction*:* todays, kidney failure is one of the costly problems of human society and use of renal replacement therapy is increasing in the world and Iran. Survival analysis is one of the fields in medical prognosis and data mining is a process of discovering unknown relationship and is a useful pattern from data and is known as a highly efficient method in survival analysis. Conclusively, the purpose of this study is predicting the survival of the kidney transplant patient's according to variables before kidney transplant.

Materials and Methods*:* In order to identify important factors for predicting survival in kidney transplant, informative requirements assessment was done by using self-designed questionnaire. Then, obtained information from the analysis of questionnaire was reviewed and data from 513 medical record of kidney patient in Sina Urology Research Center was extracted. Ultimately, by applying CRISP methodology, data mining was done by IBM SPSS Modeler 14.2 and C.5 algorithm.

Results*:* In this study, BMI, ESRD and dialysis time were evaluated as the most effective factors in survival kidney transplant and extracted rules from the model can be used for predicting the survival of the transplanted kidney before the surgery. Accuracy rate of this model was estimated at 96.77%.

Conclusion*:* The high accuracy rate of C5.0 model shows the power of it in survival prediction. Furthermore, the most effective kidney transplant survival factors were identified and kidney transplanted survival of a new patient with distinctive features, can be predicted.

Keywords*:* Data Mining, Survival Analysis, Kidney Transplantation, Prediction, CRISP Methodology

* Corresponding author. Tel: +98 9125310665

a-borhani@razi.tums.ac.ir