ORIGINAL ARTICLE

Data Mining Approach in Prediction of Erythropoietin Dosage in Hemodialysis Patients

Akram Tavousi[1],
Mohammad Mehdi Sepehri[2,3],
Tahereh Malakoutian[4],
Toktam Khatibi[5]

[1] MSc Student in Medical Informatics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[2] Associate Professor, Department of Healthcare Systems Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
[3] Hospital Management Research Center, Iran University of Medical Sciences, Tehran, Iran
[4] Assistant Professor, Department of Internal Medicine, Iran University of Medical Sciences, Tehran, Iran
[5] Assistant Professor, Department of Industrial Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran

(Received April 22, 2015 Accepted August 5, 2015)

Abstract

Background and purpose: Kidney failure reduces the kidney function and in long term it leads to chronic kidney disease. One of the main complications of this disease is irreversible damage to the kidneys (end-stage kidney disease) and hemodialysis is the main method used to treat advanced kidney failure. The main problem associated with hemodialysis is treating anemia caused by lack of erythropoietin secretion in kidney which is usually treated by synthetic erythropoietin. On the other hand, choosing the right dosage of erythropoietin is important because it is expensive and could have some complications. This research aimed at predicting the dosage of erythropoietin and identifying affecting factors.

Materials and methods: Data was collected from a dialysis center in Tehran and data mining methods were used. The input variables were measured in the past 6 months of treating patients with erythropoietin. The sequential data was then converted to the bag of features (BOF) format. Then support vector machines and random forest were applied on the BOF to predict the erythropoietin dosage.

Results: The amount of medication in previous months was found to be an important factor in determining the appropriate dosage of erythropoietin for the next month. In optimal condition, random forest and SVM could predict the erythropoietin dosage with an average accuracy of 90% and 79%, respectively.

Conclusion: This study identified the factors influencing the treatment and control of anemia in hemodialysis patients. These results could be of great benefit in prescribing the proper dosage of erythropoietin, and reducing the treatment cost and duration. Moreover, it helps to prevent the complications caused by excessive use of erythropoietin such as increase in hemoglobin level.

Keywords: Data mining, prediction, erythropoietin, hemodialysis patients

J Mazandaran Univ Med Sci 2015; 25(129): 26-35 (Persian).

مجله دانشگاه علوم پزشکی مازندران
دوره بیست و پنجم شماره ۱۲۹ مهر سال ۱۳۹۴ (۳۵-۲۶)

پیش‌بینی مقدار تجویز داروی اریتروپویتین در بیماران همودیالیزی- رویکرد داده‌کاوی

اکرم طاوسی [۱]
محمد مهدی سپهری [۲،۳]
طاهره ملکوتیان [۴]
توکتم خطیبی [۵]

چکیده

سابقه و هدف: نارسایی کلیه منجر به کاهش عملکرد کلیه‌ها می‌شود و این امر در درازمدت منجر به‌بیماری مزمن کلیه می‌گردد. یکی ازعوارض بیماری مزمن کلیه تخریب برگشت‌ناپذیر کلیه‌ها (رسیدن به مرحله پایانی بیماری کلیه) است. یکی از شایع‌ترین راه‌های درمان بیماران دچار نارسایی کلیوی، همودیالیز است. به‌علاوه یکی از مسائل اصلی در همودیالیز، کمخونی ناشی از کمبود ترشح اریتروپویتین از کلیه‌ها است که معمولا با داروی اریتروپویتین صناعی، درمان می‌شود. از سوی دیگر انتخاب دوز مناسب داروی اریتروپویتین جهت مقابله با کم خونی بیماران همودیالیزی، و با توجه به قیمت بالا و عوارض این دارو، از اهمیت بالایی برخوردار است. لذا این پژوهش به منظور پیش‌بینی دوز داروی اریتروپویتین و شناسایی عوامل اثرگذار بر انتخاب دوز مناسب این دارو از رویکردهای داده کاوی بهره برده و آن‌ها را بر روی داده جمع‌آوری‌شده از بیماران همودیالیزی اعمال می‌کند.

مواد و روش‌ها: داده‌های پژوهش از مرکز دیالیزی در تهران جمع‌آوری‌شده است. فرض می‌شود ورودی مسئله، مشخصه های شش ماهه متوالی از بیمار به همراه مقدار داروی اریتروپویتین مورد استفاده تاکنون است. جهت اعمال رویکردهای داده‌کاوی بر این داده‌ها، آن را تبدیل به یک بردار ویژگی مشخصه نموده و از روش‌های ماشین بردار پشتیبان و جنگل تصادفی برای پیش‌بینی مقدار تجویز دارو استفاده می‌شود.

یافته‌ها: نتایج حاصل از این تحقیق نشان می‌دهد مقادیر داروی تجویز شده در ماه‌های قبل بر مقدار دارو در ماه بعدی تاثیرگذار است. الگوریتم جنگل تصادفی با متوسط صحت ۹۰ درصد و ماشین بردار پشتیبان با متوسط صحت ۷۹ درصد در بهترین حالت، قادر به پیش‌بینی دوز داروی تجویزی هستند.

استنتاج: این تحقیق با شناسایی ویژگی‌های مؤثر بر درمان بیماران همودیالیزی وکنترل کم‌خونی، سبب صرفه‌جویی در هزینه و زمان شده و از عوارض ناشی از تجویز بیش از حد دارو و افزایش هموگلوبین بیمار، خواهد کاست.

واژه های کلیدی: داده‌کاوی، پیش‌بینی، اریتروپویتین، بیماران همودیالیز

مقدمه

مرحله انتهایی بیماری کلیوی یک اختلال پیش‌رونده و غیرقابل برگشت است که در آن توانایی کلیه در دفع مواد زائد متابولیک و حفظ تعادل مایع و الکترولیت‌ها از بین می‌رود(۱). مرکز کنترل و پیشگیری، (Center for disease control and prevention) بیماری‌های کلیه را به عنوان نهمین عامل مرگ در آمریکا دانسته و در

مولف مسئول: محمدمهدی سپهری- تهران: دانشگاه تربیت مدرس، دانشکده پزشکی، بیمارستان هاشمی نژاد E-mail: mehdi.sepehri@modares.ac.ir
۱. دانشجوی کارشناسی ارشد انفورماتیک پزشکی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران
۲. دانشیار، گروه مهندسی سیستم‌های سلامت، دانشکده مهندسی صنایع و سیستم‌ها، دانشگاه تربیت مدرس، تهران، ایران
۳. مرکز تحقیقات مدیریت بیمارستانی، دانشگاه علوم پزشکی ایران، تهران، ایران
۴. استادیار، گروه داخلی، دانشکده پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران
۵. استادیار، گروه مهندسی سیستم‌های سلامت، دانشکده مهندسی صنایع و سیستم‌ها، دانشگاه تربیت مدرس، تهران، ایران
تاریخ دریافت : ۱۳۹۴/۲/۲ تاریخ ارجاع جهت اصلاحات : ۱۳۹۴/۲/۲۱ تاریخ تصویب : ۱۳۹۴/۵/۱۵

ایران نیز طبق آمار، میزان شیوع مرحله انتهایی بیماری کلیه از ۲۳۸ مورد در سال ۲۰۰۰ به ۳۵۷ مورد به ازای هر یک میلیون نفر در سال ۲۰۰۶ افزایش‌یافته است(۲). نارسایی حاد کلیه می‌تواند ناشی از بیماری پرفشاری خون، دیابت، کلیه‌های پلی‌کیستیک، گلومرولونفریت‌ها، التهاب بافت بینابینی، انسداد سیستم ادراری و عوارض حاصل از داروها باشد. درمان در این حالت شامل نظارت بر رژیم غذایی و انتخاب یکی از راه‌های درمان جایگزینی کلیه شامل بستری در بیمارستان و به ویژه پیوند، دیالیز صفاقی و همودیالیز است. مبتلایان به مرحله پایانی بیماری کلیه با GFR کم‌تر از ۱۵ml/min-۱۰ تحت دیالیز قرارگرفته و به علت کمبود ترشح اریتروپویتین از کلیه‌ها دچار کم‌خونی می‌شوند(۳). کم‌خونی یکی از مهم‌ترین عوارض شایع نارسایی مزمن کلیه به‌خصوص همودیالیزی است که با مرگ و میر بالای بیماران همراه بوده و منجر به بسیاری از اختلالات پاتوفیزیولوژیک و کاهش امید به زندگی در این بیماران با غلظت هموگلوبین کم‌تر از ۱۰ گرم در دسی لیتر می‌شود(۴). طبق تعریف سازمان بهداشت جهانی در سال ۲۰۰۱، کم‌خونی به کاهش غلظت هموگلوبین Hb<12g/dl در زنان و Hb<13 در مردان و سطح Hb<11g/dl در زنان یائسه و کودکان ۶ ماهه تا ۵ ساله اطلاق می‌شود(۵). البته هدف درمانی در بیماران دیالیزی مبتلا به کم‌خونی، رساندن هموگلوبین به 11g/dl می‌باشد. تولید ناکافی اریتروپویتین توسط کلیه علت اصلی کم‌خونی است که موجب کاهش تولید گلبول‌قرمز در مغز استخوان در بیماران دیالیزی می‌شود. استفاده از اریتروپویتین صناعی برای جبران این کمبود که از سال ۱۹۹۰ جهت درمان آنمی ایجادشده، یک روش مؤثر و رایج است(۶).

گران‌قیمت بودن داروی اریتروپویتین و مشکلات بیماران در تهیه دارو، از سوی دیگر نادیده گرفتن یک سری از عوامل و شرایط متناسب با بیمار جهت تجویز مقدار دارو، باعث عدم پاسخ درمانی مناسب به درمان با اریتروپویتین می‌شود(۷). ایمنی دارویی یکی از شاخص‌های مهم ایمنی بیمار است. داروی اشتباه یا نامناسب می‌تواند منجر به عوارض جانبی شدید در بیماران و در نتیجه قصور پزشکی شود(۸). با توجه به درصد بالای وقوع کم‌خونی در بیماران همودیالیزی و نیاز آن‌ها به داروی اریتروپویتین جهت درمان، این پژوهش سعی دارد با استفاده از روش داده‌کاوی بر روی اطلاعات جمع‌آوری‌شده از بیماران همودیالیزی در بازه‌های زمانی مختلف، اقدام به پیش‌بینی مقدار داروی اریتروپویتین مرتبط با کم‌خونی در این دسته از بیماران نماید. هدف دیگر شناسایی عواملی است که درکنار مقدار دارو بر روی میزان هموگلوبین تأثیر می‌گذارد. توجه به این نکته ضروری است که استفاده صحیح از این دارو می‌تواند میزان نیاز فرد دیالیزی به تجویز خون را کاهش دهد و تجویز بالای مقدار داور، سبب بالا رفتن هموگلوبین فرد از یک حد مجاز شده و به دنبال آن سبب ایجاد سکته قلبی و مغزی شود. در جدول شماره ۱ به‌طور خلاصه به مطالعاتی که در زمینه‌ی پیش‌بینی مقدار داروی وارفارین دربیماران قلبی با استفاده از روش داده کاوی در مقاله‌ای صورت گرفته، اشاره شده است(۹).

جدول شماره ۱: مروری بر مطالعات پیشین

اسم نویسنده	پایگاه داده	تعداد نمونه	الگوریتم مورد استفاده
(۹) Byrne et al	Irish	۱۰۳	ANN
(۱۰) Ya-Han Hu et al	Taiwan	۵۸۷	kNN,SVR,MLP

در تحقیقات مشابه روش K-NN و رگرسیون جز روش‌های پرکاربرد است. حال ما می‌خواهیم با روش جنگل تصادفی (Random Forest) و ماشین‌های بردار پشتیبان (Support Vector Machine) که جزء روش‌های پرکاربرد داده‌کاوی درمطالعات به‌جهت صحت پیش‌بینی بالاتر است، بر روی داده‌های متواتر زمانی استفاده کنیم. در ادامه پس از آشنایی با اریتروپویتین و الگوریتم‌های داده‌کاوی مورد استفاده، با گام‌های رسیدن به مقدار صحت بالاتر و خصیصه‌های تاثیرگذار در پژوهش آشنا خواهید شد.

اریتروپویتین

اکثر بیماران مبتلا به نارسایی کلیه که به مدت طولانی دچار اختلال عملکرد کلیه هستند، مبتلا به کم‌خونی می‌شوند. کم‌خونی به دلیل کاهش ترشح هورمون اریتروپویتین است که ۹۰ درصد آن در غده فوق کلیه ساخته و ترشح می‌شود و با تأثیر بر مغز استخوان باعث افزایش تولید گلبول‌های قرمز خون می‌گردد. این دارو جهت جبران و اصلاح کم‌خونی ایجاد شده مورد استفاده قرار می‌گیرد. دارو به شکل مایع بی‌رنگی است که به صورت آمپول‌های ۲۰۰۰، ۴۰۰۰، ۶۰۰۰ و ۱۰۰۰۰ واحد عرضه می‌شود. بنابراین جهت درمان کم‌خونی نیاز به تجویز داروی اریتروپویتین مطابق با نیاز هر بیمار برای رسیدن به حد هموگلوبین هدف داریم. در این راستا اریتروپویتین صناعی به عنوان یک داروی اصلی جهت درمان کم‌خونی، بلوغ اریتروبلاست‌ها را تحریک کرده و آنمی را که یکی از علل عمده مرگ و میر در بین بیماران همودیالیزی است را تا حدودی رفع می‌نماید. طبق آمار بیش از ۹۰ درصد بیماران آمریکایی که تحت همودیالیزی قرار دارند، اریتروپویتین صناعی دریافت می‌کنند(۱۱). پاسخ به دوز با افزایش میزان هموگلوبین در فرد سنجیده می‌شود. گران‌قیمت بودن داروی اریتروپویتین و مشکلات بیماران در تهیه دارو، از سوی دیگر نادیده گرفتن یکسری از عوامل و شرایط متناسب با بیمار جهت تجویز مقدار دارو، باعث عدم پاسخ درمانی مناسب به درمان با اریتروپویتین می‌شود(۷). استفاده صحیح از این دارو می‌تواند میزان نیاز فرد دیالیزی به خون را کاهش دهد و تجویز بالای دوز دارو سبب بالارفتن هموگلوبین فرد از یک حد مجاز شده و به دنبال آن سبب ایجاد سکته قلبی و مغزی می‌شود.

مواد و روش ها

روش شناسی تحقیق جاری مبتنی بر رویکردهای داده کاوی برای پیش‌بینی دوز داروی اریتروپویتین و انتخاب مشخصه‌های تمایزگذار برتر است. لذا لازم است ابتدا به مرور مختصر الگوریتم‌های داده‌کاوی بپردازیم که در تحقیقات پیشین برای حل مسایل پیش‌بینی و انتخاب ویژگی معرفی شده‌اند و در این تحقیق برای این منظور مورد استفاده قرار خواهند گرفت. سپس مراحل روش شناسی تحقیق را بیان کنیم.

رویکردهای داده‌کاوی

داده‌کاوی فرایند کشف دانش مطلوب از مقدار بزرگی از داده است که در پایگاه داده ذخیره‌شده است. در داده‌کاوی معمولاً به کشف الگوهای مفید از میان داده‌ها اشاره می‌شود. منظور از الگوی مفید، مدلی در داده‌ها است که ارتباط میان یک زیرمجموعه از داده‌ها را توصیف می‌کند و معتبر، ساده، قابل‌فهم و جدید است. امروزه استفاده از داده‌کاوی در حوزه پزشکی داده‌کاوی در حال رشد است. بهره‌گیری از داده‌کاوی در حوزه بهداشت و درمان برای همه افراد درگیر در این حوزه مفید است. داده‌کاوی یکی از ابزارهایی است که به پزشکان جهت تصمیم‌گیری و سازمان‌دهی داده‌ها کمک می‌کند. امروزه درمان‌ها پس از شناسایی عوامل خاص هر بیمارصورت می‌گیرد. مقدار دارو یکی از عواملی است که برای هر فرد به‌صورت خاص در پزشکی تجویز می‌شود. الگوریتم‌های داده‌کاوی به دو دسته کلی نظارتی و غیرنظارتی و یا پیش‌بینی و توصیفی تقسیم‌شده است. در الگوریتم‌های پیش‌بینی، هدف پیش‌بینی یک ویژگی خاص بر مبنای ویژگی‌های دیگر است. ویژگی پیش‌بینی شونده متغیر وابسته و بقیه متغیرها مستقل نامیده می‌شود. اما در الگوریتم‌های توصیفی هدف استخراج الگو از داده‌هاست که نیاز به تحلیل نتایج دارد(۱۲).

جنگل تصادفی

الگوریتم جنگل تصادفی یک روش یادگیری مبتنی بر دسته‌ای از درخت‌های تصمیم است. مدل پیش‌بینی جنگل‌های تصادفی، بر اساس میانگین‌گیری از نتایج حاصل از تمامی درخت‌های تصمیم مربوطه استوار است.

جنگل‌های تصادفی متشکل از مجموعه‌ای از درخت‌های تصمیم‌گیری است. در این الگوریتم درختان تصادفی بردار ورودی را می‌گیرند، آن را با هر درخت در جنگل کلاس‌بندی می‌کنند. خروجی، برچسب‌های کلاسی هستند که از اکثریت آراء دریافت شده است. در حال حاضر یکی ازبهترین الگوریتم‌های یادگیری است و برای بسیاری از مجموعه داده‌ها، دسته‌بندی با صحت بالایی انجام می‌دهد و ویژگی مثبت دیگر این دسته‌بندی‌کننده این است که روی مجموعه داده‌های بزرگ بسیار خوب عمل می‌کند. در الگوریتم RF، برای تشکیل هر درخت، دسته متفاوتی از الگوهای موجود، با در نظر گرفتن جایگزینی دوباره هر الگوی انتخاب‌شده، صورت می‌گیرد. بر اساس الگوریتم RF در مرحله رشد هر درخت، در هر گره، دسته‌ای از ویژگی‌ها به صورت تصادفی انتخاب می‌شوند و بهترین انشعاب در میان دسته، ویژگی انتخاب شده برای تشکیل گره‌های جدید بعدی در نظر گرفته می‌شود(۱۳).

روش ماشین های بردار پشتیبان

الگوریتم ماشین بردار پشتیبان از روش‌های یادگیری با ناظر است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. هدف (Support vector Machine) SVM پیدا کردن بهترین تابع برای طبقه‌بندی است به نحوی‌که بتوان اعضای دو کلاس را در مجموعه داده‌ها از هم تشخیص داد. SVM با استفاده از یک خط به نام مرز تصمیم، نمونه‌ی کلاس‌های مختلف را از هم جدا می‌کند، این مرز تصمیم بردار پشتیبان نامیده می‌شود. هدف دسته‌بندی SVM پیدا کردن یک مرز تصمیم با حاشیه دسته‌کننده حداکثر است. مرز تصمیم می‌تواند خطی و یا غیرخطی باشد. این الگوریتم برای داده‌هایی با ابعاد بالا خوب عمل می‌کند(۱۴). جمع‌آوری داده‌ها ابتدا مبانی نظری از مقالات منتشر شده در پایگاه‌های اطلاعاتی گردآوری شده و پس از مشاوره با پزشک در رابطه با ویژگی‌های بیماران همودیالیزی، داده‌ها نهایی شد. اطلاعات موردنیاز این پژوهش، پس از حذف پرونده‌های ناقص شامل ۱۷۱ بیمار مرکز دیالیزی در تهران است که به صورت دستی ازپرونده کاغذی بیماران جمع‌آوری گردید. فرمت داده‌ها به صورت متنی و عددی است. این داده‌ها برای هر بیمار در شش ماه متوالی اندازه‌گیری شده است. برخی داده‌ها مانند اطلاعات دموگرافیک ثابت بوده و برخی از ویژگی‌ها در شش ماه متوالی و برخی در فواصل بین چند ماه اندازه‌گیری شده است. پس از جمع‌آوری داده‌ها، برای سهولت کار، داده‌ها وارد محیط اکسل شد.

توصیف داده‌ها

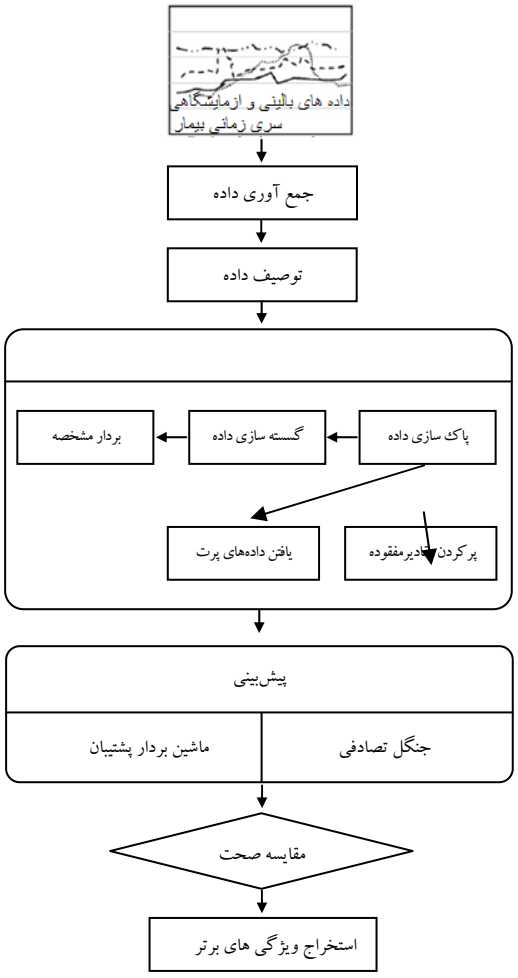
مجموعه داده دارای ۱۷۱ رکورد است. حدود ۴۳ درصد آنان را زنان و حدود ۵۷ درصد را مردان تشکیل می‌دهند. میانگین سنی بیماران ۵۷ سال است. تعداد ویژگی‌های موجود در پایگاه داده ۲۷ ویژگی است که برخی از آن‌ها با توجه به روال بیمارستان و بررسی شش‌ماهه پژوهش، هر ماه اندازه‌گیری شده و دارای شش مقدار هستند و برخی هر سه ماه یک بار اندازه‌گیری می‌شوند. هر رکورد بیماران همودیالیزی را می‌توان به چهار دسته اطلاعات دموگرافیک بیمار، اطلاعات مربوط به بیماری همراه، اطلاعات درمانی و اطلاعات آزمایشگاهی تقسیم کرد. در تصویر شماره ۱ شمای کلی از جریان تحلیل داده‌ها آورده شده است که مراحل آن در ادامه توضیح داده شده است.

انواع داده‌های گردآوری شده

در کل می‌توان داده‌های این پژوهش را به ۴ دسته‌ی زیر دسته‌بندی نمود:

۱. اطلاعات دموگرافیک بیمار شامل جنس، سن، وضعیت تأهل، شغل، سطح تحصیلات، وضعیت سیگاری و وزن است.

۲. اطلاعات مربوط به بیماری همراه شامل سابقه دیابت، هپاتیت، بیماری قلبی، فشارخون، وضعیت پیوند کلیه و مدت‌زمان دیالیز است.



تصویر شماره ۱: جریان تحلیل داده های بیماران همودیالیزی

۳. اطلاعات آزمایشگاهی شامل درصد اشباع ترانسفرین، هموگلوبین، کراتینین، کفایت دیالیز، کلسیم، فسفر، هورمون پاراتیروئید، فریتین، تری‌گلیسیرید، HDL، LDL وآلبومین است.

۴. اطلاعات درمانی شامل میزان مصرف اریتروپویتین و ونوفر در هفته است.

کفایت دیالیز، سطح سرمی پارتیروئید و فریتین، تری گلیسیرید، HDL و LDL و آلبومین جز ویژگی‌هایی است که تعداد اندازه‌گیری متفاوتی دارند.

پیش‌پردازش و آماده سازی داده‌ها

پس از درک موضوع و داده‌ها، مرحله آماده‌سازی داده‌ها که شامل پیش‌پردازش داده‌هاست، آغاز می‌شود که نقش مهمی در داده‌کاوی ایفا می‌کند. در این مرحله، داده‌ها آماده فرایند داده‌کاوی می‌شوند تا به هدف تعیین مقدار تجویز داروی اریتروپویتین در بیماران همودیالیزی جهت مقابله با کم‌خونی دست یابند. پیش‌پردازش شامل مراحل زیر است:

پاک‌سازی داده

داده‌های واقعی همراه با مقادیر مفقودشده و دور افتاده هستند. بنابراین جهت افزایش کیفیت دانش کسب شده باید مرحله پاک‌سازی داده‌ها صورت گیرد. در این مرحله، داده‌های غیر معتبر از مجموعه داده‌های آموزشی خارج می‌شوند. داده‌های پرت و ناقص نمونه‌هایی از داده‌هایی هستند که باید پاک‌سازی در مورد آن‌ها انجام گردد. پاک‌سازی در این تحقیق شامل پر کردن مقادیر مفقوده و شناسایی داده‌های پرت و مقداردهی آن‌ها است(۱۵).

پر کردن مقادیر مفقوده

سیاست رفع مقادیر مفقوده روش Miss Forest است. در مطالعه‌ای که بر روی دقت روش‌های رفع مقادیر مفقوده در داده‌های آزمایشگاهی صورت گرفته است، نشان داد که Miss Forest نسبت به متدهای دیگر در مدل‌های پیش‌بینی بالینی، خطای کم‌تری برای متغیرهای پیوسته و دسته‌بندی‌شده دارد(۱۶). این روش محتمل‌ترین مقدار (most probable value) جهت پرکردن داده فراموش شده است که با ایجاد درخت‌هایی از مقادیر، بهترین مقدار ممکن را پیدا می‌کند. این کار با استفاده از تابع Miss Forest در محیط R به‌روش زیرصورت گرفته است:

missForest (xmis, ntree=100, mtry= floor (sqrt (ncol(xmis))

xmis: ماتریسی از مقادیر ورودی

Maxiter: ماکزیمم تعداد تکرار در هر دفعه

Ntree: تعداد درخت‌های رشد کننده در هر جنگل

Mtry: تعداد متغیرهای انتخابی در هربار به‌صورت تصادفی(۱۷).

شناسایی داده پرت با نمودار هیستوگرام

برای شناسایی داده‌های پرت، از دو روش شناسایی با نمودار هیستوگرام و مقایسه مقدار حداقل، حداکثر و

میانگین هر یک از مقادیر خصیصه‌ها، به دلیل قابل فهم و استفاده در داده‌های پزشکی استفاده شده است(۱۸).

گسسته سازی داده

با توجه به این که مقادیر آزمایشگاهی به‌صورت اعداد اعشاری بوده و تنوع مقدار در آن بسیار زیاد است، جهت افزایش دقت و کارایی الگوریتم‌ها پس از مشورت با پزشک مشاور، داده‌ها به بازه‌هایی طبق نظر پزشک تبدیل شدند.

تبدیل به بردار مشخصه

داده‌ها پس از پر شدن مقادیر مفقوده، اصلاح داده‌های پرت و گسسته کردن داده‌های پیوسته برای این که در الگوریتم بردار پشتیبان بتوان فاصله نقاط از هم را سنجید، به‌صورت باینری تبدیل می‌شوند. مثلاً اگر یک ویژگی به ۴ دسته تقسیم شود ما برای نمایش ورودی‌ها به۳ بیت نیاز داریم (جدول شماره ۲).

جدول شماره ۲: تبدیل متغیرهای طبقه‌بندی‌شده به باینری

۰	۰۰۰
۱	۰۰۱
۲	۰۱۰
۳	۱۰۰

مرحله آزمایش

در این تحقیق از نرم افزار R استفاده شده است. جهت پیش‌بینی ۷۰ درصد داده‌های به عنوان داده‌های آموزش استفاده شدند. در روش‌های دسته بندی، یک فیلد به عنوان فیلد خروجی در نظر گرفته می‌شود که در این‌جا داروی تجویزی در هر ماه به عنوان خروجی و مشخصه‌های اندازه‌گیری شده در هر ماه به عنوان ورودی در نظر گرفته شدند. جهت پیش‌بینی در الگوریتم svm، کرنل، نوع خط مرز تصمیم را نشان می‌دهد که در این تحقیق از کرنل RBF استفاده شده است. ابتدا با تابع tune.svm بهترین گامای تابع کرنل را جهت روش SVM پیدا نموده و برای هر ماه به صورت جداگانه بر روی داده‌های آزمایش، پیش‌بینی صورت گرفت. هم‌چنین بر روی داده‌های آموزش که در روش SVM استفاده شده بود، الگوریتم RF اجرا گردید.

یافته‌ها

در روش پیشنهادی جهت بررسی سری زمانی برای پیش‌بینی طبقه‌بندی مقدار اریتروپویتین، ابتدا فرض می‌شود ورودی مسئله یک سری زمانی از فاکتورهای بیمار به همراه دوز داروی مورد استفاده شده تاکنون است. سپس این مسئله به‌عنوان یک مسئله دسته‌بندی توالی با تبدیل آن به بردار مشخصه در نظر گرفته می‌شود. از روش RF (Random Forest) و SVM جهت پیش‌بینی در محیط R استفاده‌شده است. جهت پیش‌بینی مقدار دارو، ۳ فرضیه به‌صورت زیر در نظر گرفته می‌شود:

۱. خروجی ماه‌های قبل به‌عنوان ورودی ماه بعدی وارد شود.

۲. خروجی فقط یک ماه قبل به عنوان ورودی ماه بعدی وارد شود.

۳. خروجی ماه‌های قبل به عنوان ورودی وارد نشود.

جهت پیش‌بینی با طبقه‌بندی جنگل تصادفی و بردار پشتیبان تصمیم، به‌صورت سری زمانی از بردار مشخصه استفاده می‌شود. فرض می‌شود ورودی هر ماه یک سری زمانی از ورودی‌هاست. هر ویژگی به بردار مشخصه تبدیل می‌شوند. در جداول زیر مقدار صحت برای هر سه فرضیه آمده است. در ابتدا فرض شده است که خروجی اریتروپویتین ماه‌های قبل به عنوان ورودی در ماه بعدی وارد شده است. نتایج حاصل از این فرضیه در جدول شماره ۳ آمده است:

جدول شماره ۳: صحت الگوریتم با تاثیر خروجی ماه‌های قبل به عنوان ورودی

صحت دسته بند	ماه اول	ماه دوم	ماه سوم	ماه چهارم	ماه پنجم	ماه ششم
SVM	۷۳/۷	۷۸/۱	۷۵/۶	۷۵	۷۸/۹	۷۶/۳
RF	۷۶	۸۹	۸۸	۸۶	۸۹	۹۰

در فرضیه بعدی خروجی اریتروپویتین یک ماه قبل به عنوان ورودی در ماه بعدی وارد شده است. نتایج حاصل ازصحت پیش‌بینی این‌فرضیه درجدول شماره۴ آمده است.

فرضیه آخر مربوط به صحت پیش‌بینی خروجی اریتروپویتین بدون تاثیر خروجی‌های ماه قبل است (جدول شماره ۵).

جدول شماره ۴: صحت الگوریتم با تاثیر خروجی ماه قبل به عنوان ورودی

صحت دسته بند	ماه اول	ماه دوم	ماه سوم	ماه چهارم	ماه پنجم	ماه ششم
SVM	۷۳/۷	۷۸/۱	۷۱/۱	۷۳/۱	۷۵/۵	۷۵/۶
RF	۷۶	۸۹	۸۸	۸۶	۸۳	۸۳

جدول شماره ۵: صحت الگوریتم بدون تاثیر خروجی ماه‌های قبل

صحت دسته بند	ماه اول	ماه دوم	ماه سوم	ماه چهارم	ماه پنجم	ماه ششم
SVM	۷۳/۷	۷۱/۴	۷۱	۷۱/۸	۷۴/۵	۷۱/۴
RF	۷۶	۶۷	۶۹	۶۵	۶۶	۶۷

لازم به توضیح است که صحت، یکی از معیارهای ارزیابی مدل‌های دسته‌بندی است که مقدار آن برابر درصد مشاهدات مجموعه آموزشی است که توسط روش مورد استفاده، به‌درستی دسته‌بندی‌شده است(۱۹). هر یک از جداول، مربوط به یکی از فرضیه‌ها است که نشان دهنده تاثیرگذاری یا عدم تاثیر اریتروپویتین تجویزی قبلی بر داروی تجویزی بعدی است. در هر فرضیه هر دو الگوریتم SVM و RF مورد بررسی قرار گرفتند. از نتایج صحت کسب شده می‌توان به این نتیجه رسید که الگوریتم RF به دلیل ایجاد درخت‌هایی از نتایج، جهت پیش‌بینی صحت بالاتری دارد. هم‌چنین در ماه دوم و پنجم به دلیل اندازه‌گیری برخی خصیصه‌ها و تاثیرگذاری آن‌ها، صحت نسبت به ماه قبلی افزایش یافته است.

بحث

گران‌قیمت بودن داروی اریتروپویتین و مشکلات بیماران در تهیه دارو از یک سو و از سوی دیگر نادیده گرفتن یک سری از عوامل و شرایط متناسب با بیمار، جهت تجویز مقدار دارو، باعث عدم پاسخ درمانی مناسب به درمان با اریتروپویتین می‌شود. در تعیین مقدار اریتروپویتین تجویزی با استفاده از ویژگی‌ها، الگوریتم SVM با استفاده از کرنل radial در بهترین حالت پیش‌بینی ۷۹ درصدصحت و الگوریتم جنگل تصادفی با نزدیک ۹۰ درصد به دست آمد. هر دو الگوریتم در فرضیه‌ی تاثیر مقادیر داروی تجویزی ماه‌های قبل به عنوان ورودی، در هر دو مدل SVM و RF بهترین حالت را داشتند. با مقایسه مقدار جینی در الگوریتم جنگل تصادفی که بالاترین صحت را داشت، متغیرهای تاثیر گذار در هر ماه به شرح جدول شماره ۶ شناسایی شدند.

جدول شماره ۵: انتخاب ویژگی های برتر

ماه	متغیرهای تأثیرگذار
۱	هموگلوبین، وزن، سن، مدت دیالیز، ونوفر، جنسیت، شغل، سطح سواد
۲	اریتروپویتین، هموگلوبین، ترانسفرین، ونوفر، سن، مدت دیالیز، شغل، سطح سواد، کلسیم
۳	اریتروپویتین، شغل، سن، وزن، هموگلوبین، مدت دیالیز، بیماری قلبی، کراتینین، سطح سواد، سیگاری
۴	اریتروپویتین، وزن، سن، سطح سواد، فسفر، هموگلوبین، وضعیت ازدواج، پیوند کلیه، شغل
۵	اریتروپویتین، هموگلوبین، پاراتیروئید، وزن، ترانسفرین، سن، مدت دیالیز، سطح سواد
۶	اریتروپویتین، سن، هموگلوبین، شغل، ونوفر، فسفر، مدت دیالیز، وزن، هپاتیت

در ابتدا لازم است اشاره شود که امروزه، هموگلوبین جز ویژگی‌هایی است که پزشکان با توجه به محدوده استاندارد این ویژگی اقدام به تجویز اریتروپویتین می‌کنند. در این تحقیق از مقایسه ویژگی‌های مهم در هر ماه به این نتیجه می رسیم که مقدار داروی تجویزی در هر ماه بر مقدار دارو در ماه بعدی تاثیر مستقیم دارد. سطح سواد جز مشخصه‌هایی به دست آمد که در مطالعات به آن توجه زیادی نشده است، اما جز عوامل تاثیرگذار بر موضوع مطالعه شناسایی شده است. می‌توان رابطه سطح سواد با مقدار دارو را به این صورت تحلیل کرد که فردی که سطح سواد بالاتری دارد، نسبت به زمان دریافت دارو حساس‌تر است و داروهای خود را در زمان تعیین شده مصرف می‌کند. علاوه بر این فرد با تحصیلات بالاتر نسبت به تغذیه و مواد غذایی مصرفی دقیق تر بوده و سعی در مصرف مواد غذایی با خواص بالاتر جهت تامین منابع بدنی دارد. با توجه به اندازه‌گیری هورمون پاراتیروئید در دو مرتبه اندازه‌گیری، نقش پاراتیروئید به عنوان یکی از عوامل تاثیرگذار در درمان کم‌خونی در بیماران همودیالیز شناسایی شده است. پاراتیروئید به عنوان یکی از علل تشدیدکننده کم‌خونی و هم‌چنین مقاومت به درمان با اریتروپویتین در بیماران

همودیالیزی محسوب می‌شود. یکی دیگر از شایع‌ترین علل عدم پاسخ به اریتروپویتین، کمبود آهن است. تجویز اریتروپویتین به بیماران نارسایی حادکلیه، موجب مصرف بیش‌تر آهن و بروز کمبود آن خواهد داشت. بنابراین برای جبران آهن از آهن وریدی ونوفر استفاده می‌شود. بنابراین ونوفر در کنار داروی اریتروپویتین از تاثیر مهمی برخوردار است و درنهایت این‌که اریتروپویتین تجویزی در ماه‌های پیشین بر روی مقدار داروی تجویزی در ماه بعدی تاثیر گذار است. با توجه به این که در ماه دوم و پنجم، درصد صحت بالاتری یافت شده، می‌توان به این نتیجه رسید که خصیصه‌هایی که در این ماه‌ها اندازه‌گیری می‌شود، در پیش‌بینی تاثیرگذار بوده است. علی‌رغم این که نتایج حاصل از پیش‌بینی انجام شده در این تحقیق می‌تواند در تصمیم‌گیری به پزشکان یاری دهد، اما هنوز نمی‌توان آن را جایگزین پزشک کرد. از آن‌جایی‌که نتایج این تحقیق وابسته به داده‌های یک بیمارستان می‌باشد، پیشنهاد می‌شود برای بررسی بیش‌تر در این زمینه، در مطالعات بعدی از داده‌های مراکز درمانی دیگر و الگوریتم های دیگر استفاده کرده و نتایج را با هم مقایسه نمود. بررسی تاثیرگذاری زمان و نوع صافی، دور پمپ خون هم می‌تواند به عنوان عناصر تاثیرگذار بر درمان باشد که در مطالعات بعدی بهتر است این عوامل در نظر گرفته شود.

سپاسگزاری

از زحمات پرسنل محترم واحد دیالیز بیمارستان هاشمی‌نژاد تهران که در مراحل انجام این تحقیق همکاری لازم را داشتند، کمال تشکر و قدردانی را دارم.

References

1. Surrena H. Handbook for Brunner and Suddarth's textbook of medical-surgical nursing. Lippincott: Williams & Wilkins; 2009.
2. Abbaszadeh S, Nourbala MH, Taheri S, Ashraf A, Einollahi B. Renal transplantation from deceased donor in Iran. Saudi J Kidney Dis Transpl 2008; 19(4): 664-668
3. Roche A, Jenkins K, Johnson C. Chronic kidney disease anaemia: diagnosis and screening. Nursing Times 2008; 104(8): 26-27.
4. Regidor DL, Kopple JD, Kovesdy CP, Kilpatrick RD, McAllister CJ, Aronovitz J, et al. Associations between changes in hemoglobin and administered erythropoiesis-stimulating agent and survival in hemodialysis patients. J AM Soc Nephrol 2006 17(4): 1181-1191.
5. Schmidt RJ, Dalton CL. Treating anemia of chronic kidney disease in the primary care setting: cardiovascular outcomes and management recommendations. Osteopath Med Primary Care 2007; 1: 14.
6. Costa E, Belo L, Quintaniha A, Santo Silva A. Resistance to recombinant human erythropoietin therapy in haemodialysis patienrs-focus on inflamatatory cytokines, leukocyte activation, iron status and erythrocyte damage. J Nephrol Ren Transplant 2008; 2: 66-83.
7. Macdougall IC, Provenzano R, Sharma A, Spinowitz BS, Schmidt RJ, Pergola PE, et al. Peginesatide for anemia in patients with chronic kidney disease not receiving dialysis. N Engl J Med 2013; 368(4): 320-332.
8. Shirley J, PJ, Sarah G, Benator JD. JCAHO initiative seeks to improve patient safety. Drug Benefit Trends 2003; 15(1): 23-24.
9. Byrne S, Cunningham P, Barry A, Graham I, Delaney T, Corrigan OI. Using neural nets for decision support in prescription and outcome prediction in anticoagulation drug

therapy. The fifthin ternational workshop on intelligent data analysis in medicine and pharmacology: a workshop at the 14th European conference on Art. Germany, Berlin; 2000.

10. Hu YH, Wu F, Lo Cl, Tai CT. Predicting warfarin dosage from clinical data: A supervised learning approach. Artif Intell Med 2012; 56(1): 27-34.
11. Fadrowski JJ, Frankenfield D, Amaral S, Brady T, Gorman GH, Warady B, et al. Children on long-term dialysis in the United States: findings from the 2005 ESRD clinical performance measures project. Am J Kidney Dis 2007; 50(6): 958-966.
12. Bellazzi R, Sacchi L, Caffi E, de Vincenzi A, Nai M, Manicone F, et al. Implementation of an automated system for monitoring adherence to hemodialysis treatment: A report of seven years of experience. Int J Med Inform 2012; 81(5): 320-331.
13. Liaw A, Wiener M. Classification and Regression by randomForest. R news 2002; 2(3): 18-22.
14. Teymourpour B, Alizadeh S, Ghazanfari M. Data mining and knowledge discovery. 4th ed. Tehran: Publication of university of science and technology tehran; 2014. (Persian).
15. Natarajan K, Li J, Koronios A. Data mining techniques for data cleaning. London: Springer-Verlag; 2010. p. 796-804.
16. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ OPEN 2013; 3(8): e002847.
17. Stekhoven DJ, Buhlmann P. MissForest-nonparametric missing value imputation for mixed-type data. Bioinformatics 2012; 28(1): 112-118.
18. Laurikkala J, Juhola M, Kentala E, Lavrac N, Miksch S, Kavsek B. Informal identification of outliers in medical data. Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology. Germany, Berlin: Citeseer; 2000.
19. Asghari M, Lizadeh S, Abolmasum Faranak M. Utilizing Data Mining Techniques for Investigating Factors Influencing the Failure of Intrauterine Insemination Infertility Treatment. Journal of Tehran University of Medical Science 2013; 16(54): 46-55 (Persian).