استفاده از درخت تصمیم برای پیش‌بینی سطح فریتین سرم در زنان مبتلا به کم‌خونی

پریسا صفایی[۱]
رسول نورالسنا[۲*]
کامران حیدری[۳]
پریا سلیمانی[٤]

۱- گروه مهندسی صنایع، مدیریت سیستم و بهره‌وری، دانشکده مهندسی صنایع، دانشگاه آزاد اسلامی، واحد تهران جنوب، تهران، ایران.
۲- گروه تولید صنعتی، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران، تهران، ایران.
۳- گروه طب اورژانس، بیمارستان لقمان حکیم، دانشکده پزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران.
٤- گروه مهندسی صنایع، سیستم‌های اقتصادی و اجتماعی، دانشکده مهندسی صنایع، دانشگاه آزاد اسلامی، واحد تهران جنوب، تهران، ایران.

[*] نویسنده مسئول: تهران، میدان رسالت، خیابان هنگام، خیابان دانشگاه، دانشگاه علم و صنعت ایران، دانشکده مهندسی صنایع، کدپستی: ۱۳۱۱٤- ۱٦۸٤٦
تلفن: ۰۲۱-۷۳۲۲۵۰۱۷
E-mail: rassoul@iust.ac.ir

چکیده

دریافت: ۱۳۹٤/۰٤/۲۰ ویرایش: ۱۳۹۵/۰۲/۲۲ پذیرش: ۱۳۹۵/۰۲/۲۹ آنلاین: ۱۳۹۵/۰۲/۳۰

زمینه و هدف: داده‌کاوی به‌عنوان فرایند شناسایی و تجزیه و تحلیل مقدار زیادی داده به‌منظور یافتن رفتارها و قوانین معنادار است. داده‌کاوی در بهداشت و درمان فرصت‌های بی‌شماری را برای بررسی الگوهای پنهان از یک مجموعه داده فراهم می‌کند. این الگوها را می‌توان توسط پزشکان برای تشخیص، پیش‌آگاهی و درمان بیماران استفاده کرد. هدف اصلی در این مطالعه استفاده از تکنیک‌های داده‌کاوی برای پیش‌بینی سطح فریتین سرم در زنان مبتلا به کم‌خونی و شناسایی متغیرهای اساسی در تشخیص این بیماری می‌باشد.

روش بررسی: در این پژوهش کاربردی، تعداد ٦۹۰ بیمار و ۲۲ متغیر در جمعیت زنان مبتلا به بیماری کم‌خونی بررسی شده‌اند. داده‌ها مربوط به بیمارانی بود که از اردیبهشت ۱۳۹۲ تا اردیبهشت ۱۳۹۳ به آزمایشگاه بیمارستان‌های امام حسین (ع) و شهدای هفتم‌تیر مراجعه کرده‌اند. از تکنیک درخت تصمیم برای ساخت مدل استفاده شد.

یافته‌ها: دقت مدل دسته‌بند درخت تصمیم با تمام متغیرها ۷۵٪ بود. ترکیب‌های متفاوت از متغیرها جهت یافتن بهترین مدل برای پیش‌بینی بررسی شدند. با توجه به مدل بهینه درخت تصمیم به‌دست آمده، متغیرهای RBC، MCH، MCHC، زخم‌های معده-روده و سرطان معده-روده، به‌عنوان مهم‌ترین عوامل پیش‌بینی‌کننده شناخته شدند. نتایج نشان داد که اگر مقدار متغیرهای Mean corpuscular volume (MCV)، Mean corpuscular hemoglobin concentration (MCHC) و Mean corpuscular hemoglobin (MCH) نرمال و متغیر RBC کمتر از حد نرمال باشند، فرد با احتمال ۹۰٪ به کم‌خونی فقر آهن مبتلا است.

نتیجه‌گیری: با توجه به سادگی و هزینه پایین آزمایش شمارش کامل خون، مدل درخت تصمیم به‌منظور تشخیص بیماری کم‌خونی فقر آهن ایجاد شد. همچنین در این پژوهش تاثیر عوامل جدیدی مانند جراحی‌ها و بیماری‌های مختلف در نظر گرفته شد. قوانین به‌دست‌آمده از مدل درخت تصمیم می‌تواند فرایند تشخیص و درمان بیماران مبتلا به کم‌خونی فقر آهن را بهبود بخشد.

کلمات کلیدی: کم‌خونی، داده‌کاوی، درخت تصمیم، فریتین.

مقدمه

داده‌کاوی (Data mining) یک حوزه میان‌رشته‌ای و با رشد سریع است که حوزه‌های مختلفی همچون پایگاه داده، آمار، یادگیری ماشین و سایر زمینه‌های مرتبط را با هم تلفیق کرده است تا داده‌ها و دانش ارزشمند نهفته در حجم بزرگی از داده‌ها را استخراج نماید.[۱]

داده‌کاوی در بهداشت و درمان فرصت‌های بی‌شماری را برای بررسی الگوهای پنهان از یک مجموعه داده فراهم می‌کند. این الگوها را می‌توان برای تشخیص، پیش‌آگاهی و درمان بیماران استفاده کرد.[۲] کم‌خونی (Anemia) از قدیمی‌ترین، رایج‌ترین و گسترده‌ترین اختلالات خونی و بر اساس تعریف سازمان بهداشت جهانی (World Health Organization) وضعیتی است که در آن تعداد و حجم

سلول‌های قرمز خون یا ظرفیت حمل اکسیژن ناکافی باشد.[۳] براساس آمار سازمان بهداشت جهانی در سال ۲۰۱۱، ۲۹٪ از زنان غیر باردار و ۳۸٪ از زنان باردار در سنین ۱۵ تا ۴۹ سال در جهان مبتلا به بیماری کم‌خونی هستند.[۴] فقر آهن به‌عنوان علت اصلی کم‌خونی شناخته شده است[۵] و تنها اختلال تغذیه‌ای می‌باشد که به‌طور قابل‌توجهی در کشورهای صنعتی نیز شایع است.[۶]

کم‌خونی فقر آهن (Iron deficiency anemia) با کاهش کیفیت زندگی، عملکرد جسمی و ذهنی و نتایج نامطلوب بالینی همراه است.[۷] تشخیص کم‌خونی فقر آهن در دو مرحله انجام می‌گیرد. در مرحله اول باید مولفه‌های آزمایش شمارش کامل خون (Complete blood count) بررسی شوند. شمارش کامل خون یک آزمایش روتین با هزینه کم می‌باشد که برای شناسایی کم‌خونی و دیگر اختلالات خونی استفاده شده است.[۸]

بر اساس تعریف سازمان بهداشت جهانی سطح هموگلوبین کمتر از ۱۳ gr/dl برای مردان و کمتر از ۱۲ gr/dl برای زنان به‌معنای وجود کم‌خونی است. در مرحله دوم پس از تایید کم‌خونی، وجود فقر آهن باید مورد ارزیابی قرار گیرد.[۹] آزمایش‌های اندازه‌گیری آهن خون پر هزینه‌اند و برای هر بیمارستانی قابل دسترس نیستند. در این پژوهش متغیر فریتین برای بررسی وجود فقر آهن در بیماران انتخاب شده است، زیرا غلظت این متغیر متناسب با کل ذخایر آهن بدن می‌باشد[۱۰] و کاهش سطح فریتین سرم نشان‌دهنده بیماری کم‌خونی فقر آهن است.[۱۱]

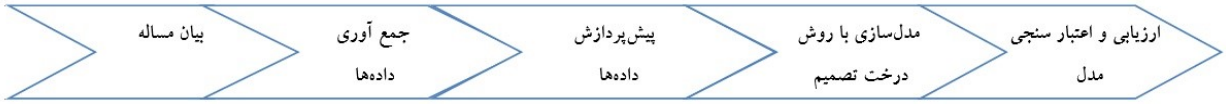
تاکنون پژوهش‌های بسیاری پیرامون کاربرد روش‌های داده‌کاوی در پزشکی انجام شده است.[۱۲-۱۴] با این وجود در زمینه کم‌خونی فقر آهن مطالعات محدودی انجام گرفته که اکثر آنها نیز اپیدمیولوژیک می‌باشند.[۱۵-۱۷] با توجه به سادگی و هزینه پایین آزمایش شمارش کامل خون هدف اصلی در این پژوهش استفاده از مولفه‌های این آزمایش و به‌کارگیری روش‌های داده‌کاوی برای یافتن رویکردی جهت پیش‌بینی سطح فریتین سرم طی یک مرحله و همچنین تعیین متغیرهای مهم و تاثیرگذار در این پیش‌بینی است.

روش بررسی

در این پژوهش کاربردی تنها جمعیت زنان مبتلا به بیماری کم‌خونی مورد مطالعه قرار گرفت زیرا گروهی پرخطر برای ابتلا به این بیماری محسوب می‌شوند.[۱۸] داده‌های به‌دست‌آمده از بیماران مبتلا به کم‌خونی در جمعیت زنان جمع‌آوری شد که از اردیبهشت ۱۳۹۲ تا اردیبهشت ۱۳۹۳ به آزمایشگاه بیمارستان‌های امام حسین (ع) و شهدای هفتم‌تیر واقع در شهر تهران مراجعه کردند. این داده‌ها مطابق جدول‌های ۱ و ۲، شامل متغیرهای آزمایشگاهی و بالینی بیماران مبتلا به کم‌خونی بودند که از پرونده‌ها و نتایج آزمایش آنها گردآوری شد. مجموعه داده‌ها در این مطالعه شامل ۶۹۰ رکورد و ۲۲ فیلد (ویژگی‌های بیماران) بود. برای انجام تمام مراحل پژوهش از R Statistical Computing Project, Version 3.2.0 (http://www.r-project.org) استفاده شد. این مراحل در شکل ۱ نشان داده شد.

در مرحله سوم به آماده‌سازی و پیش‌پردازش (Preprocess) داده‌ها پرداخته شد. داده‌های به‌دست‌آمده از دنیای واقعی اغلب شامل مقادیر گم‌شده (Missing data)، داده‌های نویز (Noise) و پرت (Outlier) می‌باشند. بنابراین به‌منظور حصول نتیجه بهتر و دستیابی به بهبود کیفیت داده‌های واقعی برای داده‌کاوی، پیش‌پردازش داده‌ها لازم است.

در این پژوهش از روش K-Means برای شناسایی نقاط دورافتاده یا پرت استفاده شد. الگوریتم خوشه‌بندی K-Means از جمله مشهورترین الگوریتم‌های یادگیری بدون نظارت است که در آن مجموعه داده‌ها به تعداد خوشه‌های از پیش تعیین‌شده تقسیم می‌شوند. تعداد پنج رکورد از ۶۹۰ داده توسط این روش به‌عنوان نقاط پرت شناسایی شدند و پس از بررسی این نمونه‌ها توسط پزشک متخصص، حذف گردیدند. در مرحله چهارم، از روش درخت تصمیم



شکل ۱: مراحل انجام پژوهش

جدول ۱: متغیرهای آزمایشگاهی پژوهش

نام متغیر	نوع متغیر	حدود نرمال	نام متغیر در مدل
شمارش گلبول‌های قرمز	عددی	٤/٢–٥/٤	RBC
هموگلوبین	عددی	١٢–١٦	HB
هماتوکریت	عددی	٣٦–٤٨	HCT
حجم متوسط هموگلوبین	عددی	٨٠–١٠٠	MCV
وزن متوسط هموگلوبین	عددی	٢٧–٣٠	MCH
غلظت متوسط هموگلوبین	عددی	٣٠–٣٥	MCHC
پهنای گلبول قرمز	عددی	١٠/٥–١٥/٥	RDW
آهن	عددی	١٢–١٥٠	IRON
تری‌گلیسیرید	عددی	<١٦٠	TG
کلسترول	عددی	<٢٠٠	CHOL
فریتین	عددی	١٨–١٠٠	Ferritin

جدول ۲: متغیرهای بالینی تحقیق

نام متغیر	نوع متغیر	نام متغیر در مدل	سطوح متغیر
سن	عددی	Age	
شاخص توده بدنی	عددی	BMI	
زخم‌های روده- معده	طبقه‌ای	Gastrointestinal ulcer	زخم معده، زخم دوازدهه
بیماری‌های روده- معده	طبقه‌ای	Gastrointestinal diseases	پولیپ روده بزرگ، واریس مری، بواسیر، واریس هموروییید
بیماری‌های التهابی روده	طبقه‌ای	Inflammatory bowel diseases	بیماری التهابی روده، آرتریت روماتوییید، سلیاک
سرطان‌های روده- معده	طبقه‌ای	Gastrointestinal cancer	مری، معده، روده، روده بزرگ، کبد
عفونت‌های روده- معده	طبقه‌ای	Gastrointestinal infection	کرم روده، سایر عفونت‌ها
جراحی‌های روده- معده	طبقه‌ای	Gastrointestinal surgery	مری، معده، روده، روده بزرگ
عادت‌های روزمره	طبقه‌ای	Habit	سیگار، مصرف الکل، پیکا
داروهای مصرفی	طبقه‌ای	Drug	استروییید، سایر داروها
سایر علل کم‌خونی	طبقه‌ای	Other anemia cause	دارد، ندارد

(Decision tree) برای مدل‌سازی استفاده شد. درخت تصمیم یکی از روش‌های داده‌کاوی و از ابزارهای قوی و متداول برای دسته‌بندی (Classification) و پیش‌بینی است. ساختار درخت تصمیم یک ساختار درختی، شبیه فلوچارت است. بالاترین گره در درخت، گره ریشه است و گره‌های برگ، دسته‌ها یا توزیع دسته‌ها را نشان می‌دهند.

روش کار درخت به این صورت است که یک رکورد جدید در گره ریشه وارد می‌شود و در این گره یک آزمون صورت می‌گیرد تا معلوم شود که این رکورد به کدام یک از گره‌های فرزند (شاخه‌های پایین‌تر) تعلق دارد. تمام رکوردهایی که به یک برگ از درخت می‌رسند، در یک دسته قرار می‌گیرند.

در این پژوهش برای ساخت درخت تصمیم از الگوریتم کارت (CART) استفاده شد. روش کارت شاخه‌های خود را به‌صورت دوتایی و تنها بر اساس یک فیلد (متغیر مستقل) انشعاب می‌زند. معیار انتخاب مشخصه برای تفکیک در هر گره، شاخص جینی (Gini index) است. در این پژوهش متغیر هدف، فریتین سرم می‌باشد که دارای مقادیر پیوسته است.

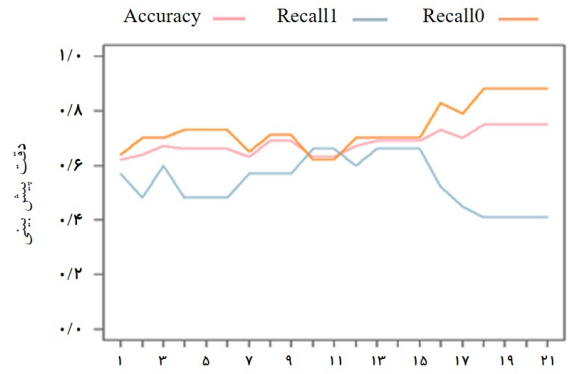
برای استفاده از مدل‌های دسته‌بندی به‌ویژه درخت تصمیم بهتر است که متغیر هدف گسسته باشد. بنابراین مقادیر فریتین پایین‌تر از ۱۸ ng/ml به‌عنوان دسته مبتلا به بیماری کم‌خونی فقر آهن و بالاتر از ۱۸ به‌عنوان دسته افراد نرمال (عدم وجود کم‌خونی فقر آهن) در نظر گرفته شد. این مقدار بر اساس مطالعات پیشین در زمینه تشخیص کم‌خونی فقر آهن و همچنین با توجه به کیت‌های آزمایشگاهی استفاده شده برای نمونه‌گیری از خون بیماران در آزمایشگاه‌های بیان شده تعیین شده است.[۱۹و۲۰]

برای ساخت مدل درخت، کل مشاهدات یا نمونه‌ها با نسبت ۷۰٪ به ۳۰٪ به‌ترتیب به دو بخش آموزش (Train) و آزمایش (Test) تقسیم شدند. در مرحله پنجم اثربخشی درخت ساخته شده اندازه‌گیری شد. برای این کار از داده‌های مجموعه آزمایش استفاده می‌شود که در ساخت مدل استفاده نشدند. پس از اجرای مدل، معیارهای ارزیابی Accuracy، Recall1 و Recall0 با توجه به ماتریس اغتشاش (Confusion matrix) به‌دست آمدند. Recall1 دقت پیش‌بینی‌های درست برای دسته بیماران مبتلا به بیماری کم‌خونی فقر آهن و Recall0 دقت پیش‌بینی‌های درست برای دسته غیر بیمار می‌باشد. بدیهی است که هر چه مقدار Recall1 بالاتر باشد، ارزش پیش‌بینی انجام گرفته نیز بالاتر می‌رود.

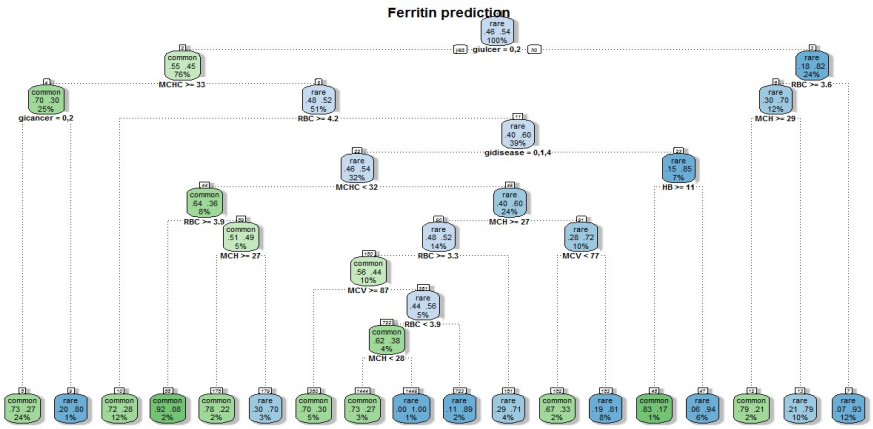
یافته‌ها

ترکیبات مختلف از متغیرهای آزمایشگاهی و بالینی در هر مرحله بررسی و بهترین ترکیب از متغیرها انتخاب شد. همچنین در هر مرحله برای تعیین متغیرها و صحت پیش‌بینی از نظر پزشک متخصص استفاده شده است. معیارهای ارزیابی Accuracy، Recall1 و Recall0 با توجه به ماتریس اغتشاش برای هر مرحله به‌دست آمدند. شکل ۲ این معیارهای ارزیابی را بر اساس ورود متغیرها به‌صورت یک به یک در مدل نشان می‌دهد. این نتایج با منحنی‌های صورتی (Accuracy)، آبی (Recall1) و نارنجی (Recall0) نشان داده شده‌اند.

همان‌طور که در منحنی مشاهده می‌شود، بالاترین دقت در روش درخت تصمیم ۷۵٪ می‌باشد که با استفاده از تمام متغیرها به‌دست آمده است. اما با تحلیل ماتریس اغتشاش در این مرحله در می‌یابیم که دقت پیش‌بینی کلاس مثبت (وجود کم‌خونی فقر آهن) که از اهمیت ویژه‌ای برخوردار است، ۴۱٪ و دقت پیش‌بینی کلاس منفی (عدم وجود کم‌خونی فقر آهن) ۸۹٪ است. بنابراین با توجه به پایین بودن دقت در پیش‌بینی کلاس مثبت، مدل ایجاد شده با تمام متغیرها قابلیت لازم برای پیش‌بینی را ندارد. با تحلیل نتایج هر مرحله و نظر کارشناس، بهترین مدل پیش‌بینی با استفاده از متغیرهای RBC، MCH، MCHC، MCV، HB، giulcer، gidisease، gicancer، giinfection و gisurgery به‌دست آمد. با وجود این متغیرها نتایج معیارهای ارزیابی دقت پیش‌بینی (Accuracy) ۶۸٪، Recall1 و Recall0 به‌ترتیب ۵۵٪ و



شکل ۲: معیارهای ارزیابی بر اساس تعداد متغیرها در مدل درخت تصمیم



شکل ۳: درخت تصمیم برای پیش‌بینی کم‌خونی فقر آهن

۷٤٪ بود. شکل ۳ درخت ایجاد شده با استفاده از متغیرهای مدل بهینه برای تشخیص کم‌خونی ناشی از فقر آهن را نشان می‌دهد.

نتایج درخت تصمیم در قالب قواعد اگر- آنگاه قابل استخراج بود. براساس این قواعد در صورتی‌که متغیر RBC کمتر از حد نرمال و متغیرهای MCH، MCHC و MCV نرمال بودند، فرد با احتمال بسیار زیادی به کم‌خونی فقر آهن مبتلا بود. در بیماران با RBC پایین‌تر از حد نرمال، MCH نرمال و مبتلا به بیماری‌های مری و بواسیر با توجه به متغیر gidisease، احتمال ۹٤٪ برای ابتلا به کم‌خونی فقر آهن وجود داشت.

با توجه به اینکه حدود نرمال برای RBC بین ٤/۲ و ۵/٤ میلیون در میکرولیتر است، حال اگر RBC کمتر از سه (به‌شدت پایین) باشد و بیمار زخم معده نیز داشته باشد، با احتمال بیش از ۹۰٪ به کم‌خونی فقر آهن مبتلا است. در صورتی‌که MCH و RBC نرمال باشند، با احتمال ۲۱٪ کم‌خونی فقر آهن به‌وجود می‌آید. به‌طور کلی نتایج نشان داد که RBC و MCH کمتر از حد نرمال و giulcer (زخم معده، زخم دوازدهه) تاثیر زیادی در بروز کم‌خونی فقر آهن دارند.

بحث

۳۰٪ تا ۵۰٪ کم‌خونی در کودکان و زنان به‌دلیل فقر آهن به‌وجود می‌آید. بنابراین می‌توان گفت که فقر آهن شایع‌ترین علت کم‌خونی است. در حال حاضر در فرایند شناسایی کم‌خونی فقر آهن نیاز به نمونه‌گیری دوباره از خون می‌باشد که اغلب با عدم مراجعه بیماران به‌دلیل زمان و هزینه بالا همراه است. با توجه به اهمیت تشخیص و درمان این بیماری در این مطالعه سعی بر آن شده است تا عوامل مهم در بروز کم‌خونی ناشی از فقر آهن با استفاده از روش‌های داده‌کاوی شناسایی شوند. به‌این منظور به پیش‌بینی سطح فریتین خون با استفاده از بهترین ترکیب ممکن از متغیرهای آزمایش شمارش کامل خون و متغیرهای بالینی پرداخته شد. در اغلب پژوهش‌های پیشین در زمینه بیماری کم‌خونی فقر آهن تنها به بررسی آثار نامطلوب این بیماری بر روی جامعه اکتفا شده است و در مطالعات اندکی با استفاده از تکنیک‌های داده‌کاوی به پیش‌بینی و شناخت این بیماری پرداخته‌اند. Casanova و همکارانش به پیش‌بینی بیماری کم‌خونی فقر آهن در زنان باردار با استفاده از متغیرهای HB، RBC، MCV، MCH، MCHC، RDW، HCT، BMI، فریتین، سن، نژاد و همچنین سوابق بیماران پرداختند. به‌این منظور آن‌ها از روش‌های تحلیل رگرسیون لجستیک، منحنی مشخصه عملکرد (ROC) و تحلیل طبقه‌ای استفاده کردند. دو متغیر HB و RDW به‌عنوان مهم‌ترین فاکتورهای پیش‌بینی‌کننده کم‌خونی فقر آهن در دوران بارداری شناخته شدند.[۲۱] Azarkhish و همکاران به پیش‌بینی سطح آهن سرم بر اساس چهار متغیر آزمایشگاهی MCH, MCHC, Hb/RBC و MCV پرداختند. آن‌ها

یک شبکه عصبی هوشمند (Artificial neural network, ANN) و یک سیستم استنتاج عصبی- فازی تطبیقی (Adaptive neuro-fuzzy inference system, ANFIS) را برای تشخیص کم‌خونی فقر آهن گسترش دادند. نتایج پژوهش آنها نشان داد که تحلیل شبکه‌های عصبی نسبت به دو روش دیگر از عملکرد بالاتری در پیش‌بینی برخوردار است.[۲۲] همچنین در پژوهش دیگری از روش شبکه‌های عصبی برای تشخیص کم‌خونی فقر آهن در زنان استفاده شده است. این مطالعه تنها با شش متغیر آزمایش شمارش کامل خون (RBC، HB، HCT، MCH، MCHC، MCV) انجام شد. نتایج این پژوهش دقت بالای این روش را نشان داد.[۲۳]

در پژوهش کنونی بر خلاف مطالعات پیشین علاوه بر داده‌های آزمایش شمارش کامل خون از متغیرهای بالینی استفاده شده است و بررسی تاثیر عوامل جدیدی مانند جراحی، بیماری‌های مختلف و زخم‌های دستگاه گوارش مهم‌ترین وجه تمایز این مطالعه نسبت به سایر بررسی‌ها می‌باشد. همچنین تعیین متغیرهای با اهمیت در تشخیص کم‌خونی فقر آهن که در این پژوهش انجام شده است، می‌تواند به پزشکان برای تشخیص سریع‌تر این بیماری یاری رساند. نتایج این پژوهش نشان داد که زخم‌های معده و دوازدهه، بیماری‌های مری، بواسیر و سرطان معده در بروز کم‌خونی فقر آهن تاثیر به‌سزایی دارند.

متغیرهای RBC، MCH، MCHC در میان متغیرهای آزمایش شمارش کامل خون اهمیت بالایی برای تشخیص کم‌خونی فقر آهن دارند. همچنین از متغیر فریتین به‌عنوان متغیر هدف استفاده نشده است و این در صورتی است که مطابق نظر متخصصان و پژوهش‌های انجام گرفته، سطح فریتین خون به‌عنوان بهترین متغیر در شناسایی بیماری کم‌خونی فقر آهن می‌باشد. با وجود نقاط قوت بیان‌شده، این مطالعه با محدودیت‌هایی نیز رو به رو بوده است که می‌توان به تعداد کم نمونه‌های جمع‌آوری شده برای بررسی این بیماری اشاره کرد. بنابراین جمع‌آوری داده‌های بیش‌تر می‌تواند برای پژوهش‌های آتی به‌عنوان عامل مهمی در به‌دست آوردن نتایج تعمیم‌پذیرتر در نظر گرفته شود. همچنین استفاده از متغیرهای بیش‌تر مانند ترانسفرین (Transferrin) جهت توسعه و بهبود در نتایج حاصل شده و به‌کارگیری سایر روش‌های داده‌کاوی برای دست‌یابی به نتایج کامل‌تر پیشنهاد می‌شود. انجام پژوهش کنونی بر روی بیماران بستری شده در بیمارستان و همچنین افراد سالم می‌تواند منجر به نتایج جالب‌تری شود زیرا وجود کم‌خونی فقر آهن در بیماران بستری شده در بیمارستان به‌طور وسیع می‌باشد. بنابراین از مطالعه بر روی این بیماران می‌توان نتایج جالب‌تری به‌دست آورد.

مدل درخت تصمیم نشان داد که مقادیر کمتر از حد نرمال برای متغیرهای RBC و MCH و وجود زخم‌های دستگاه گوارش (زخم معده، زخم دوازدهه) تاثیر زیادی در بروز کم‌خونی فقر آهن دارند. همچنین مطابق ساختار درخت تصمیم به‌دست آمده، متغیرهای RBC، MCH، MCHC و gicancer به‌عنوان مهم‌ترین عوامل در پیش‌بینی بیماری کم‌خونی فقر آهن شناخته شدند. نتایج به‌دست آمده از این مدل می‌تواند به پزشکان در تشخیص و درمان بیماری کم‌خونی فقر آهن کمک کند.

سپاسگزاری: این مقاله بخشی از پایان‌نامه تحت عنوان "استفاده از درخت تصمیم برای پیش‌بینی سطح فریتین سرم در بیماران مبتلا به کم‌خونی" در مقطع کارشناسی ارشد در سال ۱۳۹٤ می‌باشد که با حمایت دانشگاه آزاد اسلامی واحد تهران جنوب اجرا شده است.

References

1. Chakrabarti S, Ester M, Fayyad U, Gehrke J, Han J, Morishita S, et al. Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, 2006.
2. Milovic B, Milovic M. Prediction and decision making in health care using data mining. *Int J Publ Health Sci (IJPHS)* 2012;1(2):69-78.
3. World Health Organization (WHO). Worldwide prevalence of anaemia 1993-2005. WHO Global Database on Anaemia. [Internet] 2008 [cited 2015 Apr 15]; Available from: http://www.who.int/nutrition/publications/micronutrients/anaemia_iron_deficiency/9789241596657/en/
4. World Health Organization (WHO). Global Nutrition Targets 2025: Anaemia policy brief. [Internet] 2014 [cited 2015 Apr 15]; http://www.who.int/nutrition/publications/globaltargets2025_policybrief_anaemia/en/
5. McLean E, Cogswell M, Egli I, Wojdyla D, de Benoist B. Worldwide prevalence of anaemia, WHO Vitamin and Mineral Nutrition Information System, 1993-2005. *Public Health Nutr* 2009;12(4):444-54.
6. World Health Organization (WHO). Iron deficiency anaemia: assessment, prevention and control. A guide for programme managers. [Internet] 2014 [cited 2015 Apr 15];

http://www.who.int/nutrition/publications/micronutrients/anaemia_iron_deficiency/WHO_NHD_01.3/en/
7. Shander A, Javidroozi M, Ozawa S, Hare GM. What is really dangerous: anaemia or transfusion? *Br J Anaesth* 2011;107 Suppl 1:i41-59.
8. Yeh JS, Cheng CH. Using hierarchical soft computing method to discriminate microcyte anemia. *Expert Sys Appl* 2005;29(3):515-24.
9. Shander A1, Goodnough LT2, Javidroozi M3, Auerbach M4, Carson J5, Ershler WB, et al. Iron deficiency anemia: bridging the knowledge and practice gap. *Transfus Med Rev* 2014;28(3):156-66.
10. Wang W, Knovich MA, Coffman LG, Torti FM, Torti SV. Serum ferritin: Past, present and future. *Biochim Biophys Acta* 2010;1800(8):760-9.
11. Goddard AF1, McIntyre AS, Scott BB. Guidelines for the management of iron deficiency anaemia. British Society of Gastroenterology. *Gut* 2000;46 Suppl 3-4:IV1-IV5.
12. Marcano-Cedeño A, Chausa P, García A, Cáceres C, Tormos JM, Gómez EJ. Data mining applied to the cognitive rehabilitation of patients with acquired brain injury. *Expert Sys Appl* 2013;40(4):1054-60.
13. Chao CM1, Yu YW, Cheng BW, Kuo YL. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. *J Med Syst* 2014;38(10):106.
14. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Sys Appl* 2014;41(4):1476-82.
15. Cook JD. Iron-deficiency anaemia. *Baillieres Clin Haematol* 1994;7(4):787-804.
16. Polin V, Coriat R, Perkins G, Dhooge M, Abitbol V, Leblanc S, et al. Iron deficiency: from diagnosis to treatment. *Dig Liver Dis* 2013;45(10):803-9.
17. Powers JM1, Buchanan GR2. Diagnosis and management of iron deficiency anemia. *Hematol Oncol Clin North Am* 2014;28(4):729-45, vi-vii.
18. Friedman AJ1, Chen Z, Ford P, Johnson CA, Lopez AM, Shander A, et al. Iron deficiency anemia in women across the life span. *J Womens Health (Larchmt)* 2012;21(12):1282-9.
19. Guyatt GH1, Patterson C, Ali M, Singer J, Levine M, Turpie I, et al. Diagnosis of iron-deficiency anemia in the elderly. *Am J Med* 1990;88(3):205-9.
20. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med* 1992;7(2):145-53.
21. Casanova BF, Sammel MD, Macones GA. Development of a clinical prediction rule for iron deficiency anemia in pregnancy. *Am J Obstet Gynecol* 2005;193(2):460-6.
22. Azarkhish I, Raoufy MR, Gharibzadeh S. Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. *J Med Syst* 2012;36(3):2057-61.
23. Yılmaz Z, Bozkurt MR. Determination of women iron deficiency anemia using neural networks. *J Med Syst* 2012;36(5):2941-5.

TUMJ monthly Tehran University Medical Journal

Tehran University Medical Journal, April 2016; Vol. 74, No. 1: 50-57 *Original Article*

Using decision tree to predict serum ferritin level in women with anemia

Parisa Safaee M.Sc.[1]
Rassoul Noorossana Ph.D.[2*]
Kamran Heidari M.D.[3]
Parya Soleimani Ph.D.[4]

1- Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.
2- Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
3- Department of Emergency Medicine, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
4- Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.

[*] Corresponding author: Department of Industrial Engineering, Iran University of Science and Technology, University St., Hengam Ave., Resalat Sq., Tehran, Iran, 16846-13114
Tel: +98 21 73225017
E-mail: rassoul@iust.ac.ir

Abstract

Received: 11 Jul. 2015 Revised: 11 May 2016 Accepted: 18 May 2016 Available online: 19 May 2016

Background: Data mining is known as a process of discovering and analysing large amounts of data in order to find meaningful rules and trends. In healthcare, data mining offers numerous opportunities to study the unknown patterns in a data set. These patterns can be used to diagnosis, prognosis and treatment of patients by physicians. The main objective of this study was to predict the level of serum ferritin in women with anemia and to specify the basic predictive factors of iron deficiency anemia using data mining techniques.

Methods: In this research 690 patients and 22 variables have been studied in women population with anemia. These data include 11 laboratories and 11 clinical variables of patients related to the patients who have referred to the laboratory of Imam Hossein and Shohada-E- Haft Tir hospitals from April 2013 to April 2014. Decision tree technique has been used to build the model.

Results: The accuracy of the decision tree with all the variables is 75%. Different combinations of variables were examined in order to determine the best model to predict. Regarding the optimum obtained model of the decision tree, the RBC, MCH, MCHC, gastrointestinal cancer and gastrointestinal ulcer were identified as the most important predictive factors. The results indicate if the values of MCV, MCHC and MCH variables are normal and the value of RBC variable is lower than normal limitation, it is diagnosed that the patient is likely 90% iron deficiency anemia.

Conclusion: Regarding the simplicity and the low cost of the complete blood count examination, the model of decision tree was taken into consideration to diagnose iron deficiency anemia in patients. Also the impact of new factors such as gastrointestinal hemorrhoids, gastrointestinal surgeries, different gastrointestinal diseases and gastrointestinal ulcers are considered in this paper while the previous studies have been limited only to assess laboratory variables. The rules of the decision tree model can improve the process of diagnosing and treatment of the patients with iron deficiency anemia and reduce their costs.

Keywords: anemia, data mining, decision trees, ferritins.