ORIGINAL ARTICLE

Providing a Model for Predicting the Risk of Osteoporosis Using Decision Tree Algorithms

Mona Sharifkhani[2],
Somayeh Alizadeh[1],
Mahnaz Abbasi[3],
Hakimeh Ameri[2]

[1] MSc in Information Technology, Faculty of Industrial Engineering, Khajeh Nasir Toosi University of Technology, Tehran, Iran
[2] Assistant Professor, Department of Information Technology, Faculty of Industrial Engineering, Khajeh Nasir Toosi University of Technology, Tehran, Iran
[3] Associate Professor, Metabolism Diseases Research Centre, Qazvin University of Medical Sciences, Qazvin, Iran

(Received Jun 11, 2014; Accepted August 25, 2014)

Abstract

Background and purpose: Some diseases such as osteoporosis may have no symptom but suddenly cause fractures in different parts of body such as spine, chest, hands and legs, thereby resulting in very painful death in old people. According to a report by Iran's ministry of health 4.6% of people aged 20 to 70 years in Iran are affected by osteoporosis in the spine. This study aimed at determining the factors influencing the incidence of osteoporosis and also providing a predictive model to speed up the detection and reduce diagnostic costs.

Material and Methods: Data was collected by interviewing 670 patients in an orthopedic clinic. The information included demographic information, lifestyle and diseases, and the results of DEXA scan. In this paper, a new model based on the standard methodology CRISP is presented. In modeling, three known data mining methods, the CHAID, C5.0 decision tree, and neural network were used. For data analysis Celementine V.12.0 was used..

Results: In this study, for the first time in Iran, the characteristics affecting osteoporosis in patients has been studied. Using data mining techniques, influencing characteristics of the disease have been identified. According to the created decision tree, some rules are derived that can be used as a model for the prediction of patient's status. Accuracy of built models using the algorithms C.5.0, CHAID, and neural networks were compared. Each algorithm was observed to act better in predicting osteoporosis in a specified group of people.

Conclusion: The accuracy of artificial neural networks algorithm is higher than that of the decision tree algorithm. In this study the most affected factors on osteoporosis were detected. According to the created rules for a new instant with specified features, we can predict whether a patient will probably suffer from osteoporosis or not.

Keywords: Osteoporosis, data mining, decision tree, artificial neural network

J Mazandaran Univ Med Sci 2014; 24(116): 110-118 (Persian).

پژوهشی

مجله دانشگاه علوم پزشکی مازندران
دوره بیست و چهارم شماره ۱۱۶ شهریور سال ۱۳۹۳ (۱۱۸-۱۱۰)

ارائه مدلی برای پیش‌بینی احتمال ابتلا به بیماری پوکی استخوان با استفاده از الگوریتم های درخت تصمیم

مونا شریف خانی[۱]
سمیه علیزاده[۲]
مهناز عباسی[۳]
حکیمه عامری[۲]

چکیده

سابقه و هدف: امروزه بیماری‌هایی مانند پوکی استخوان که بدون هیچ علامتی ناگهان فرد مبتلا را به شکستگی‌های زیادی در نواحی مختلف بدن از جمله ستون فقرات، قفسه سینه، دست‌ها و پاها دچار می‌سازد و در نهایت باعث مرگی دردناک می‌شود در اغلب سالخوردگان مشاهده می‌شود. به گزارش وزارت بهداشت ۴/۶ درصد افراد ۲۰ تا ۷۰ سال در ایران به پوکی استخوان در ستون فقرات مبتلا هستند. هدف این مقاله تعیین عوامل تأثیرگذار در بروز پوکی استخوان و نیز ارائه مدل پیش‌بینی‌کننده‌ای برای تشخیص این بیماری به منظور افزایش سرعت تشخیص و نیز کاهش هزینه‌های تشخیصی می‌باشد.

مواد و روش ها: اطلاعات مربوط به ۶۷۰ بیمار در چهار بخش اطلاعات فردی، سبک زندگی و اطلاعات بیماری‌ها و نتایج دستگاه DEXA بررسی شده است. در این مقاله مدل جدیدی بر اساس متدولوژی استاندارد CRISP ارائه شده است. در بخش مدل سازی از سه روش شناخته شده در داده کاوی به نام های درخت تصمیم CHAID و C5.0 و شبکه عصبی مصنوعی استفاده شده است. برای تحلیل داده ها از نرم افزار Celementine 12.0 استفاده شده است.

یافته ها: در این تحقیق برای اولین بار در ایران ویژگی‌های تاثیرگذار بر پوکی استخوان در بیماران مورد بررسی قرار گرفته است. با استفاده از داده کاوی و روش‌های آن ویژگی‌های تاثیرگذار بر این بیماری شناسایی شده‌اند. به کمک درخت تصمیم ایجاد شده، قوانینی استخراج شده اند که می تواند به عنوان الگویی برای پیش بینی وضعیت بیماران از آن‌ها استفاده کرد. دقت مدل‌های ساخته شده با استفاده از الگوریتم‌های C.5.0، CHAID و شبکه عصبی مصنوعی با یکدیگر مقایسه شده‌اند. نتایج این مقایسه نشان می‌دهد هریک از این الگوریتم‌ها در پیش‌بینی گروهی از افراد بهتر عمل می‌کند.

استنتاج: به طور کلی دقت الگوریتم‌های شبکه عصبی مصنوعی از الگوریتم‌های درخت تصمیم بیش‌تر است. بیش‌ترین عوامل تأثیر گذار بر پوکی استخوان شناسایی شدند. با توجه به قوانین ایجاد شده برای یک نمونه جدید با ویژگی‌های مشخص، می‌توان پیش‌بینی کرد بیمار احتمالا دچار پوکی استخوان خواهد شد یا خیر.

واژه های کلیدی: پوکی استخوان، داده کاوی، درخت تصمیم، شبکه عصبی مصنوعی

مقدمه

به گزارش سازمان جهانی بهداشت سالانه بیش از ۸/۹ میلیون شکستگی در سراسر جهان به دلیل پوکی استخوان اتفاق می‌افتد. تخمین زده شده است که پوکی استخوان ۲۰۰ میلیون زن را در سراسر جهان- حدود

مولف مسئول: حکیمه عامری- تهران- دانشگاه خواجه نصیرالدین طوسی، سید خندان. خیابان دبستان. دانشگاه خواجه نصیرالدین طوسی. دانشکده مهندسی صنایع. طبقه ۴
E-mail: hameri@mail.kntu.ac.ir

۱. کارشناس ارشد مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران ، ایران
۲ استادیار دانشگاه صنعتی خواجه نصیرالدین طوسی، سید خندان. خیابان دبستان، دانشگاه خواجه نصیرالدین طوسی، دانشکده مهندسی صنایع
۳. دانشیار ، مرکز تحقیقات بیماری های متابولیسم، دانشگاه علوم پزشکی قزوین، قزوین، ایران
تاریخ دریافت : ۱۳۹۳/۳/۲۱ تاریخ ارجاع جهت اصلاحات : ۱۳۹۳/۴/۲۳ تاریخ تصویب : ۱۳۹۳/۶/۳

یک دهم از زنان ۶۰ ساله، یک پنجم زنان ۷۰ ساله دو پنجم زنان ۸۰ ساله و دو سوم از زنان ۹۰ ساله را تحت تأثیر قرار داده است(۱، ۲). مطالعه‌ای که چندی پیش توسط وزارت بهداشت ایران انجام شده نشان می‌دهد ۴۷ درصد زنان و ۴۴ درصد مردان بالای ۵۰ سال در ایران دچار کمبود تراکم استخوان هستند و ۴/۶ درصد افراد ۲۰ تا ۷۰ سال در ایران به پوکی استخوان در ستون فقرات مبتلا هستند. هم‌چنین از هر ۴ زن ایرانی بالای ۵۰ سال، یک نفر به پوکی استخوان مبتلاست. این در حالی است که آمارهای جهانی، فاصله معنی‌داری با آمار کشور ما دارد. در سرتاسر دنیا از هر ۳ زن، یک زن و از هر ۵ مرد یک مرد بالای ۵۰ سال دچار پوکی استخوان هستند. مطالعه دیگری در ایران نشان داده است نیمی از زنان بالای ۴۵ سال و ۹۰درصد زنان بالای ۷۵ سال در ایران به پوکی استخوان مبتلا هستند(۳).

پوکی استخوان (Osteoporosis) یا بیماری خاموش به معنی پوک شدن یا نازک شدن استخوان‌ها به هر علتی می‌باشد. این بیماری علایم زیادی نداشته و تنها زمانی بروز می‌کند که استخوان‌ها به راحتی می‌شکنند. فرد مبتلا زمانی به بیماری خود پی می‌برد که در اثر یک زمین خوردن ساده، یک عطسه یا یک سرفه تعدادی از استخوان‌ها دچار شکستگی می‌شود. بر طبق آمار از هر سه زن یک نفر به این بیماری مبتلا می‌شود(۴). در بسیاری از افراد کاهش توده استخوانی که در اثر افزایش سن ایجاد می‌شود برای بروز پوکی استخوان کافی می‌باشد. علت کاهش توده استخوانی با بالا رفتن سن هنوز به درستی مشخص نشده است اما کمبود هورمون استروژن به عنوان یک عامل اصلی در خانم‌ها در هنگام یائسگی مطرح می‌باشد(۵). مشکلات دستگاه گوارش، کبد و کلیه، دیابت، بیماری‌های ریوی و قلبی و عروقی، کمبود استروژن، یائسگی زودرس (زیر ۴۵ سال)، فقدان پریودهای ماهانه (آمنوره)، اختلالات تیروئید، سرطان‌ها، سابقه پوکی استخوان در افراد درجه یک خانواده، سابقه شکستگی قبلی، مصرف داروهایی مانند کورتیکواستروئیدها، داروهای ضد انعقاد، ضد صرع، سیکلوسپورین و درمان با داروهای استروئیدی (کورتون دار) به مدت طولانی، استخوان‌بندی ظریف و کوچک از دیگر عوامل مستعدکننده ابتلای فرد به پوکی استخوان هستند. کمردرد شدید در یک نقطه خاص، درد ناگهانی در ناحیه ستون فقرات و شکستگی‌های بدون دلیل و خودبه‌خودی استخوان از دیگر نشانه‌های بیماری پوکی استخوان به شمار می‌روند(۶، ۷). بسیاری از جنبه های زندگی روزانه شامل رژیم غذایی با کمبود ویتامین D و کلسیم، مصرف الکل و سیگار، مصرف زیاد پروتئین ها، کافئین و نمک و نیز عدم فعالیت بدنی از عوامل تاثیرگذار می باشند. گرچه اثر این عوامل بر روی استخوان و ایجاد شکستگی به طور کلی کمتر از عوامل خطری هستند که قبلاً شرح داده شد، اما از این نظر که می توان این عوامل را اصلاح نمود و خطر پوکی استخوان را کاهش داد اهمیت دارند(۸). از آن‌جایی که بیماری پوکی استخوان قابل پیشگیری است، اهمیت زیادی دارد که هر چه سریع‌تر آن را در افراد تشخیص دهیم. امروزه روش‌هایی وجود دارد که بوسیله آن ها می‌توانیم این بیماری را تشخیص دهیم. روش جذب دوگانه اشعه ایکس (Dual-Energy X-ray Absorptiometry (DEXA)) یکی از بهترین روش‌ها است. بر اساس نتایج این دستگاه، افراد به سه گروه سالم، در مرحله اول فرسایش (استئوپنیا) و مبتلا به پوکی استخوان تقسیم می‌شوند.

امروزه در دنیایی زندگی می‌کنیم که حجم عظیمی از داده‌های خام ما را احاطه کرده است. در حالی که نیازمند دانش زیادی هستیم. دانشی که بتواند ما را در جهت بهبود زندگی یاری دهد. داده‌کاوی علمی است که ما را در رسیدن به دانش نهفته در داده‌ها یاری می‌کند. یکی از حوزه‌هایی که نیازمند این دانش است، حوزه پزشکی و سلامت است. ابزارهای داده کاوی به دو گروه کلی دسته بندی و پیش بینی تقسیم می شوند.

دسته‌بندی شامل روش‌ها و الگوریتم‌های بسیاری است که درخت تصمیم و شبکه عصبی مصنوعی پرکاربردترین آن ها هستند. در این تحقیق برای روش درخت تصمیم دو الگوریتم C5.0 و CHAID مورد استفاده قرار گرفته است. از الگوریتم های پرکاربرد درخت تصمیم الگوریتم C5.0 است. این الگوریتم می‌تواند برای بیان دسته بندی به صورت درخت تصمیم و یا مجموعه قوانین به کار برده شود. گروه بندی در ساخت مدل C5.0 باعث ترکیب ویژگی هایی می شود که با درنظر گرفتن فیلد خروجی، الگوهای مشابهی دارند. انتخاب Boosting در C5.0 باعث بهبود نرخ دقت پیش بینی می شود. این روش با ساخت مدل های چندگانه به صورت متوالی کار می کند(۹، ۱۰). الگوریتم CHAID می‌تواند درخت‌های غیر دودوئی بسازد، یعنی هر گسست می تواند بیش از دو شعبه داشته باشد. Exhaustive CHAID یک متد از الگوریتم CHAID است که تمرکز و دقت بیشتری برای پیدا کردن کلیه گسست‌های ممکن دارد. در این متد پس از پیدا کردن کلیه گسست‌ها، ادغام دسته ها ادامه می‌یابد تا زمانی که تنها دو دسته برای هر پیش بینی کننده باقی بمانند. پس از آن به گام انتخاب گسست متغیر رفته و از بین پیش‌بینی‌کننده‌ها، آن موردی را که بیش‌ترین گسست مشهود را دارد، انتخاب می کند. برای داده های زیاد و هم‌چنین داده های دارای متغیرهای پیش بینی پیوسته، این الگوریتم به زمان محاسبه بیشتری نیاز دارد(۱۰، ۱۱). روش شبکه عصبی مصنوعی نیز مانند درخت CHAID از متدهای مختلفی برای ساخت مدل استفاده می کند. سریع‌ترین متد Quick است ولی نسبت به سایر متدها دقت کم‌تری دارد. این متد از ویژگی‌های داده برای انتخاب یک توپولوژی مناسب برای شبکه عصبی مصنوعی استفاده می کند. متد Multiple از ویژگی‌های داده، چندین توپولوژی را به صورت موازی تولید می کند و سپس توپولوژی را که کم‌ترین خطا را دارد به عنوان توپولوژی نهایی انتخاب می‌کند. متد Exhaustive prune کم‌ترین سرعت را دارد ولی نسبت به سایر متدها از دقت بیش‌تری برخوردار است. این متد شبکه بزرگی را ایجاد می‌کند سپس در زمان آموزش واحدهای ضعیف تر در لایه ورودی و لایه های پنهان را به منظور یافتن بهترین توپولوژی هرس می کند(۱۲،۱۳). در جدول شماره ۱ برخی از تحقیقات انجام شده در حوزه داده کاوی و بیماری پوکی استخوان به صورت خلاصه مورد بررسی قرار گرفته است(۱۷- ۱۴).

هدف این مقاله ساخت مدلی برای پیش‌بینی بیماری پوکی استخوان است. به کمک این روش‌ها می‌توان بدون نیاز به روش‌های تشخیصی، احتمال ابتلا به این بیماری را تشخیص داد و نیز عوامل موثر در این بیماری را شناسایی کرد.

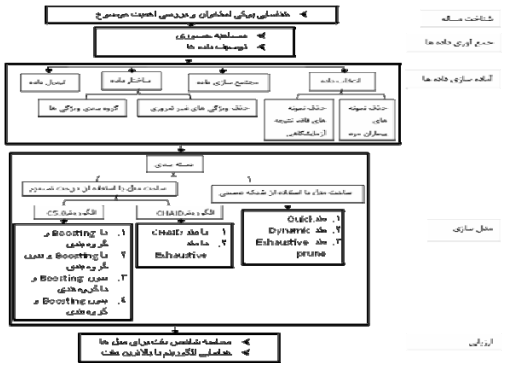
مواد و روش ها

برای داشتن یک داده‌کاوی موثر علاوه بر نیاز به داده های مرتبط، باید از یک فرآیند و روش داده‌کاوی مناسب نیز بهره‌مند شویم. روشی که کلیه مراحل داده‌کاوی اعم از جمع آوری داده، آماده سازی داده، مدل سازی و ارزیابی را دربرگیرد(۱۷). بدین

جدول شماره ۱: بررسی مطالعات انجام شده به کمک داده کاوی در حوزه پوکی استخوان

نویسنده و سال	تعداد نمونه ها	روش به کار رفته	توضیحات
ونگ و دیگران (۲۰۰۵)	۲۹۳۴ زن	ترکیب درخت تصمیم و شبکه عصبی	پیش بینی خطر ابتلا به پوکی استخوان در زنان
گالو و دیگران(۲۰۱۰)	۱۴ فرد سالم و ۱۴ بیمار مبتلا به پوکی استخوان	درخت تصمیم (الگوریتم C4.5)	تشخیص پوکی استخوان
مودانی و دیگران(۲۰۱۱)	۲۸۴۵	جنگل تصادفی درخت تصمیم Random Forest (RF)	ارائه سیستم هوشمند مبتنی بر وب به منظور پیش بینی احتمال ابتلا به بیماری پوکی استخوان
دربهر و دیگران(۲۰۰۹)	۷۹۳۶ مورد مبتلا به پسوریازیس و ۱۴۸۳۵ گروه شاهد	رگرسیون لجستیک	بررسی ارتباط بیماری پسوریازیس (صدف) و پوکی استخوان جهت شناسایی عوامل تاثیرگذار

منظـور بـر مبنـای متـدولوژی (Cross-Industry Standard Process for Data Mining Methodology CRISP(CRISP-DM Methodology))، فرآیند انجـام کار صورت گرفته است. مراحل ایـن فرآینـد در تـصویر شماره ۱ قابل مشاهده است.



درک مسئله

در این مرحله از متدولوژی به شناخت سیستم و بیان اهداف مسئله می پردازیم. به دلیل بالا رفتن متوسط سنی جامعه، تعداد مبتلایان به بیماری پوکی استخوان هر روزه در حال افزایش است. هدف، ارائه مدل پیش‌بینی کننـده ای برای احتمال بروز بیماری پوکی اسـتخوان بـه منظـور افـزایش سـرعت تـشخیص و نیـز کـاهش هزینـه‌هـای تشخیصی می‌باشد.

جمع آوری و توصیف داده

به منظور جمع‌آوری داده در ایـن پـژوهش از روش مصاحبه استفاده شده اسـت. بـدین منظـور بـا تـک تـک مراجعه کنندگان به یک مطـب متخـصص روماتولوژی به صورت حضوری و تک تک مـصاحبه انجـام شـده و پرسشنامه‌هـا تکمیـل گردیـد. ایـن پرسـشنامه‌هـا بـه طـور مستمر در طول ۹ ماه در سال ۱۳۹۱ برای ۶۷۰ زن و مـرد تکمیل گردید. تعداد ویژگی های درنظر گرفته شـده در پرسشنامه ۶۰ مـورد شـامل سـه بخـش اطلاعـات فـردی، سبک زندگی و اطلاعات بیماری ها است. بخش چهارم پرسـشنامه مربـوط بـه نتـایج دسـتگاه DEXA (دسـتگاه تشخیص پوکی استخوان) است که توسط پزشک معالج تکمیل گردیده است.

آماده سازی داده

این مرحله شامل چهـار زیـر مرحلـه، انتخـاب داده، سـاختار داده، مجتمـع سـازی داده و تبـدیل فرمـت داده می‌باشد. با توجه بـه ایـن کـه حـدود ۹۴درصـد مـصاحبه شوندگان زن بودند پرسشنامه‌هـای مربـوط بـه مـردان بـه دلیل کم بودن تعداد از بررسی در ایـن پـژوهش حـذف گردیدند. هم‌چنین سه پرسشنامه نیز به دلیل ناقص بـودن نتـایج تـست DEXA حـذف گردیدنـد. برخـی از ویژگی‌های موجود در پرسشنامه مثـل شـماره پرونـده بـا توجه به اهداف مسئله مورد نیاز نبوده و حذف گردیدنـد و برخی مثل قـد، وزن، میـزان ورزش بـا توجـه بـه دامنـه گسترده‌ای که داشتند گروه‌بنـدی شـده‌انـد و یـا بـه زیـر ویژگی‌هایی تقسیم شدند.

مدل سازی داده

به منظور پیش‌بینی احتمال بروز پـوکی اسـتخوان از مدل پیش‌بینی کننده استفاده شده است. بـدین منظـور از دو روش درخت تصمیم و شبکه عصبی مـصنوعی بـرای ساخت مـدل اسـتفاده شـده اسـت. بـا توجـه بـه ایـن کـه مجموعه هدف شامل سه متغیر سـالم، اسـتئوپنیا و پـوکی استخوان است نیاز به الگوریتمی داریـم کـه بـه توانـد بـا مجموعه داده غیـر دودوئـی کـار کنـد. از ایـن رو بـرای روش درخت تصمیم از الگوریتم‌های C.5.0 و CHAID و بـرای روش شـبکه عـصبی مـصنوعی از سـه متـد Exhaustive prune، Quick و Multiple اسـتفاده شـده است. ورودی تمام این الگـوریتم هـا شـامل ۶۱ ویژگـی آماده شده است. مجموعه داده به دو بخـش آموزشـی و آزمایـشی بـه نـسبت ۸۰ بـه ۲۰ تقـسیم شـده اسـت و بـا اسـتفاده از روش تقـسیم ۱۰ بخـشی (10-fold) مـدل هـا ساخته شده‌اند. روش تقـسیم ۱۰ بخـشی، بـدین صـورت است که ابتدا مجموعه داده بـه صـورت تـصادفی بـه ۱۰

بخش تقسیم می‌شود. سپس در هر بار یکی از این ۱۰ قسمت به عنوان داده آزمایشی انتخاب می‌شود و ۹ قسمت دیگر به عنوان داده آموزشی مورد استفاده قرار می‌گیرند.

ارزیابی

به منظور مقایسه مدل های ساخته شده، از شاخص دقت (Precision) استفاده شده است(۱۸).

$$\text{دقت} = \frac{TP(\text{True Positive})}{TP + FP(\text{False Positive})}$$

عنصر «مثبت درست» به مشاهداتی اشاره دارد که توسط روش دسته‌بندی، درست دسته بندی شده اند. عنصر «مثبت غلط» به.مشاهداتی اشاره دارد که توسط روش دسته‌بندی، غلط دسته بندی شده اند. در جدول شماره ۲ دقت مدل‌های ساخته شده با یکدیگر مقایسه شده است. همان‌گونه که مشاهده می‌شود در میان مدل‌های ساخته شده مدل شبکه عصبی مصنوعی با متد Multiple بیش‌ترین دقت را برای پیش‌بینی افراد مبتلا به پوکی استخوان دارد. برای افراد سالم مدل درخت تصمیم C.5.0 همراه با گروه‌بندی و Boosting بیش‌ترین دقت را دارد و برای افراد استئوپنیا مدل شبکه عصبی مصنوعی با متد Exhaustive prune بیش‌ترین دقت را دارد.

یافته ها

با وجود این که مدل‌های ساخته شده با شبکه عصبی مصنوعی دقت بیش‌تری در پیش‌بینی نسبت به سایر مدل‌های ساخته شده دارند ولی مانند یک جعبه سیاه (Black Box) عمل می‌کنند و نمی‌توان قوانین پیش‌بینی‌کننده را استخراج نمود. برخلاف روش شبکه عصبی، روش درخت تصمیم امکان استخراج قوانین پیش‌بینی‌کننده را دارد. برخی از قوانین استخراج شده از مدل با Boosting و Grouping الگوریتم C.5.0 به شرح زیر می‌باشد.

قانون ۱: اگر فردی یائسه شده باشد بین ۱ تا ۵ لیوان دوغ در هفته مصرف کند و زانو درد نداشته باشد یا زانو درد خفیف باشد آنگاه سالم است ۹۶/۲۶ درصد.

قانون ۲: اگر فردی یائسه نشده باشد و کم‌تر از ۲ فنجان نوشابه در هفته مصرف کند و جراحی تخمدان داشته باشد آن‌گاه مبتلا به استئوپنیا است ۸۷/۵۰ درصد.

قانون ۳: اگر فردی یائسه شده باشد، و اصلاً دوغ مصرف نکند و مبتلا به بیماری اعصاب و روان باشد آنگاه مبتلا به پوکی استخوان است ۹۰ درصد.

با توجه به این که مانند روش درخت تصمیم نمی‌توانیم قوانین روش شبکه عصبی مصنوعی را بیان کنیم در جدول شماره ۳ ویژگی‌های با اهمیت دو مدل

جدول شماره۲: مقایسه دقت مدل های پیش‌بینی کننده

عنوان الگوریتم	عنوان متد	افراد مبتلا به پوکی استخوان	افراد استئوپنیا	افراد سالم	پیش بینی اشتباه افراد مبتلا به پوکی استخوان	پیش بینی اشتباه افراد استئوپنیا	پیش بینی اشتباه افراد سالم
C.5.0	۱- با Boosting بدون Grouping	۹۲/۳۰ درصد	۸۱/۴۸	۹۸	۷/۷	۱۸/۵۲ درصد	۲ درصد
	۲- بدون Boosting بدون Grouping	۶۷/۳۰	۸۵/۱۸	۸۴	۳۲/۷۰	۱۴/۸۲ درصد	۱۶ درصد
	۳- با Boosting با Grouping	۸۶/۵۳	۸۵/۱۸	۹۴	۱۳/۴۷	۱۴/۸۲ درصد	۶ درصد
	۴- بدون Boosting با Grouping	۸۴/۶۱	۷۰/۳۷	۹۸	۱۵/۳۹	۲۹/۶۳ درصد	۲ درصد
CHAID	CHAID	۸۶/۵۰	۸۴/۹۳	۸۷/۷۱	۱۳/۵۰	۱۵/۰۷ درصد	۱۲/۲۹ درصد
	Exhaustive CHAID	۸۴/۰۴	۸۴/۹۳	۸۶/۸۴	۱۵/۹۶	۱۵/۰۷ درصد	۱۳/۱۶ درصد
شبکه عصبی	Exhaustive prune	۹۰/۱۸	۹۵/۳۹	۹۵/۱۷	۹/۸۲	۴/۶۱ درصد	۴/۸۳ درصد
	Quick	۷۸/۵۲	۹۲/۸۸	۹۲/۵۴	۴۸,۲۱	۷/۱۲ درصد	۷/۴۶ درصد
	Multiple	۹۵/۷۰	۹۲/۸۸	۹۵/۶۱	۴/۳۰	۷/۱۲ درصد	۴/۳۹ درصد

شبکه عصبی مصنوعی Multiple و Exhaustive prune ذکر شده است. اهمیت ویژگی‌ها از بالا به پایین در این دو مدل کاسته می‌شود.

جدول شماره ۳: ویژگی های با اهمیت در دو مدل شبکه عصبی

شبکه عصبی مصنوعی با متد Exhaustive prune	شبکه عصبی مصنوعی با متد Multiple
سن یائسگی	میزان قرارگیری روزانه در معرض نور
زانو درد	سن یائسگی
مصرف نوشیدنی دوغ در طول هفتگی	ابتلا به قوز/ گودی کمر
مدت زمان پیاده روی در طول هفته	زانو درد
میزان قرارگیری روزانه در معرض نور	مصرف نوشیدنی دوغ در طول هفتگی
مدت زمان استفاده از مکمل کلسیم	میزان مصرف آب روزانه
مدت زمان استفاده از کورتون	استفاده از دندان مصنوعی
میزان مصرف آب روزانه	مدت زمان استفاده از مکمل کلسیم
ابتلا به قوز/ گودی کمر	کوتاه شدن قد
میزان مصرف چای روزانه	ورزش در گذشته

بحث

در این تحقیق با استفاده از الگوریتم های داده کاوی تلاش کردیم احتمال ابتلا به پوکی استخوان را با استفاده از الگوریتم‌های داده کاوی پیش‌بینی کنیم. مطابق جدول شماره ۳ بر اساس مدل شبکه عصبی، مهم‌ترین عوامل موثر بر پوکی استخوان معرفی شده‌اند. الگوریتم‌های درخت تصمیم C5.0 و CHAID و شبکه عصبی مصنوعی برای مدل سازی استفاده شدند. در میان مدل‌های ساخته شده مدل شبکه عصبی مصنوعی با متد Multiple بیش‌ترین دقت را برای پیش‌بینی افراد مبتلا به پوکی استخوان دارد. برای افراد سالم مدل Boostingو گروه‌بندی درخت تصمیم C.5.0 بیش‌ترین دقت را دارد و برای افراد استئوپنیا مدل شبکه عصبی مصنوعی با متد Exhaustive prune بیش‌ترین دقت را دارد. دقت مدل پیش بینی کننده پوکی استخوان معرفی شده توسط Wang و همکاران (۲۰۰۵) در یکی از بهترین مدل‌های ترکیبی درخت تصمیم و شبکه عصبی مصنوعی برابر ۷۰/۵ درصد است(۱۳). در حالی که دقت مدل پیش‌بینی ارائه شده توسط ما برای افراد مبتلا به پوکی استخوان با شبکه عصبی مصنوعی با متد Multiple و معادل ۹۵/۷۰ درصد به دست آمده است.

Gao و همکاران (۲۰۱۰) در تحقیقی با بررسی ویژگی‌های استخراج شده از طیف فرکتال تصویر میکروسی‌تی به پیش‌بینی احتمال پوکی استخوان در افراد پرداخته اند. از درخت تصمیم C4.5 برای مدل پیش‌بینی‌کننده استفاده کرده‌اند. دقت مدل به‌دست آمده ۹۲/۹ درصد است(۱۴).

Moudani و همکاران (۲۰۱۱) به پیش‌بینی ابتلا به بیماری پوکی استخوان با استفاده از روش جنگل تصادفی (Random Forest) درخت تصمیم پرداخته‌اند. مدل به کار رفته چهار کلاس بدون خطر، کم خطر، خطر متوسط، خطر زیاد و خطر شدید را برای احتمال ابتلا به پوکی استخوان پیش‌بینی می‌کند. ۱۵ ویژگی را مورد تحلیل قرار داده‌اند(۱۵).

Dreiher و همکاران (۲۰۰۹) مطالعه‌ای در زمینه ارتباط دو بیماری پوکی استخوان و پسوریازیس با استفاده از رگرسیون لجستیک چند متغیره انجام داده‌اند. داده‌های مورد بررسی شامل سن، جنسیت، نژاد وضعیت اجتماعی و اقتصادی، و تشخیص‌های مزمن از جمله پوکی استخوان و بیماری‌های مشکوک یا شناخته شده به عنوان عوامل خطر برای پوکی استخوان (برای مثال سیگار کشیدن، اختلالات تیروئید، بیماری التهابی روده، هپاتیت مزمن، آرتریت روماتوئید،و بیماری انسدادی مزمن ریوی)، حالات مرتبط با کاهش فعالیت بدنی (کوری و افسردگی) و چاقی (یک عامل محافظتی در برابر پوکی استخوان) هستند(۱۶). نتایج بدست آمده نشان می‌دهد که پسوریازیس (صدف) ارتباط قابل توجهی با پوکی استخوان در مردان دارد ولی در بین زنان این دو بیماری با هم ارتباطی ندارند. همچنین با افزایش سن شیوع بیماری‌ها افزایش می‌یابد، اما ارتباط کاهش می‌یابد. خلاصه مقالات ذکر شده در جدول شماره ذکر شده است.

با مقایسه تحقیق‌های قبلی در حوزه داده کاوی و پوکی استخوان، مشخص است که مدل ارائه شده در

جدول شماره ٤: خلاصه مقالات مرتبط با پوکی استخوان

نویسنده مقاله	تعداد ویژگی	تعداد نمونه	صحت	هدف	نحوه جمع آوری اطلاعات	نحوه انجام کار
Wang و دیگران	۳۳	۲۹۳۴ زن	ذکر نشده است . میانگین دقت برای داده های آزمایشی ۶۰/۳۹درصد می باشد. در بهترین مدل ۷۰/۵درصد است	پیش بینی خطر ابتلا به پوکی استخوان در زنان	اطلاعات مربوط به سوابق پزشکی و سبک زندگی این بیماران به وسیله پرسشنامه‌ای شامل ۶۰ سوالی و نتایج آزمایش سنجش تراکم استخوان جمع آوری شده است.	روش به کار گرفته شده در این مطالعه ترکیبی از دو روش درخت تصمیم و شبکه‌های عصبی است که از آن با عنوان گروه ترکیبی Hybrid Ensembles نام برده شده‌است. این روش بدین صورت ساخته شده که در ابتدا چندین مدل شبکه عصبی و چندین مدل درخت تصمیم برای این داده‌ها ساخته شده است سپس تعداد m عدد از مدل های درخت تصمیم و N-m عدد از مدل‌های شبکه عصبی بگونه‌ای باهم ترکیب گردیده‌اند که نقص‌های یکدیگر را برطرف نماید و باعث افزایش دقت این روش شوند. بدین منظور سه مدل ترکیبی (فقط شامل شبکه عصبی، فقط شامل درخت تصمیم و ترکیبی از هر دو) ساخته شده است که در این مقاله فقط به برخی از این مدل‌ها اشاره شده‌است. این مطالعه نشان می‌دهد که به کارگیری ترکیبی از دو مدل درخت تصمیم و شبکه عصبی کارآیی مدل منفرد را تا حدود ۵٪ افزایش می دهد و به ساخت مدلی برای پیش‌بینی دقیق‌تر خطر ابتلا به پوکی استخوان در زنان کمک می‌کند.
~~گائو~~ Gao و دیگران	۱	۱۴ فرد سالم و ۱۴ بیمار مبتلا به پوکی استخوان	ذکر نشده است. دقت ۹۲/۹درصد می باشد.	کمک به پزشکان در تشخیص پوکی استخوان	مجموعه داده Micro-CT استخوان میله‌ای دانشگاه تکنولوژی ایندهوون (Eindhoven University of Technology)	در این مطالعه طیف چندگانه فرکتال از Micro-CT بیوپسی برداشته شده از استخوان میله‌ای استخراج می‌شود. این کار به منظور ساخت نمودارهایی از تصاویر سه بعدی انجام می‌شود تا بتوان این تصاویر را با تصاویر نمونه‌های نرمال مقایسه کرد. سپس درخت تصمیم و الگوریتم C.4.5 برای دسته بندی و بررسی این روش به کارگرفته می‌شود.
Moudani و دیگران	۱۵	۲۸۴۵	۹۹/۹۲درصد	پیش‌بینی احتمال ابتلا به پوکی استخوان در ۴ کلاس بدون خطر، کم خطر، خطر متوسط، خطر زیاد و خطر شدید	ذکر نشده است	از روش جنگل تصادفی درخت تصمیم برای شناسایی موارد پوکی استخوان استفاده شده‌است.
Dreiher و دیگران	ذکر شده است	۷۹۳۶ مورد مبتلا به پسوریازیس و ۱۴۸۳۵ گروه شاهد	ذکر نشده است	بررسی ارتباط دو بیماری پوکی استخوان و صدف جهت شناسایی عوامل تاثیرگذار مشترک	مجموعه داده افراد ۵۱ تا ۹۰ ساله پایگاه داده سازمان بهداشت و درمان فلسطین اشغالی	این مطالعه با استفاده از یک مدل رگرسیون لجستیک چند متغیره انجام شده است. برای پیاده سازی این مدل از نرم افزار SPSS و Stata استفاده شده است. نتایج بدست آمده نشان می دهد که پسوریازیس ارتباط قابل توجهی با پوکی استخوان در مردان دارد ولی در بین زنان این دو بیماری با هم ارتباطی ندارند. همچنین با افزایش سن شیوع بیماری ها افزایش می‌یابد، اما ارتباط کاهش می یابد.

این مقاله بالاترین دقت را دارد. هم‌چنین می‌تواند سه حالت مبتلا به پوکی استخوان، مبتلا به استئوپنیا و یا سالم را پیش‌بینی کند. با توجه به ویژگی‌ها مؤثر شناخته شده واضح است که کمبود ویتامین D، سن یائسگی، فعالیت بدنی و زانو درد بیش‌ترین اثر را برروی احتمال بروز پوکی استخوان دارند که منطبق بر یافته‌های پژوهش‌های گذشته است. با توجه به مشکلات فراوانی که در مسیر جمع‌آوری داده وجود دارد، به نظر می رسد ارائه یک استاندارد ملی به منظور ثبت اطلاعات بیماران می‌تواند تا حد زیادی از مشکلات جمع آوری داده به کاهد و مسیری را برای انجام تحقیقات گسترده‌تر فراهم سازد. به کارگیری سایر الگوریتم‌های پیش‌بینی کننده و نیز مشخص ساختن عوامل موثر در بروز پوکی استخوان می‌تواند زمینه‌های دیگری از تحقیقات را فراهم سازد. هم‌چنین ساخت برنامه پیش‌بینی کننده‌ای بر اساس این مدل‌ها پیشنهاد می‌گردد.

References

1. Johnell O, Kanis JA .An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int. 2006; 17(12):1726-1733.
2. Kanis JA. Assessment of osteoporosis at the primary health-care level. WHO scientefic Group Technical Report, University of Sheffield, UK: 66,2007; Available online at : http:// www. iofbonehealth. org/facts-statistics# category-14
3. Rohollahi F. Take a look at prevalence of osteoporosis in the world and Iran. Hamshahri newspaper. Health part. 2007; 4150. Available online at: http:// www. isaarsci. ir/ general
4. Cooper C, Melton LJ. Epidemiology of osteoporosis. Trends Endocrinol Metab. 1992; 3(6): 224-229. PMID: 18407104.
5. Compston J. The British Medical Association Family Doctor Guide to Osteoporosis. 11th ed. London: Dorling Kindersley; 2002.
6. Poole KE,Compston JE. Osteoporosis and Its Management. BMJ. 2006; 333(7581): 1251-1256.
7. Ebeling PR. Clinical practice, Osteoporosis in Men. N Engl J Med. 2008; 358(1): 1474-1482.
8. Lindsay, R. and Cosman, F. Harrison's Principles of Internal Medicine: Osteoporosis. 18th ed. The McGraw Hill; 2012: 3131–3136.
9. Alizadeh S, Ghazanfari M, Teimorpour B .Data Mining and Knowledge Discovery. 2nd ed. Tehran: Publication of Iran University of Science and Technology; 2011. (Persian).
10. Han J, Kamber M. Data Mining: Concepts and Techniques.2nd ed. San Francisco: Morgan Kaufman Publisher; 2006.
11. Alizadeh S, malek mahmodi S. data mining and knowlade discovery,step by step. Tehran: KN Toosi University; 2011.
12. IBM Knowledge Center.IBM SPSS Modeler V15.0.0 documentation (software) IBM Corporation 1994, 2013. Available online at : http://pic.dhe.ibm.com/infocenter/spssmodl/v15 r0m0/index.jsp? topic=% 2Fcom. ibm.spss. modeler. help%2 Fneuralnet_ modeltab. htm.
13. Wang W, Richards G, Rea S. Hybrid Data Mining Ensemble for Predicting Osteoporosis Risk. Conf Proc IEEE Eng Bio Soc. 2005. 1: 886-889.
14. Gao Z, Hong W, Xu Y, Zhang T, Song Z, Liu J. Osteoporosis Diagnosis Based on the Multifractal Spectrum Features of Micro-CT Images and C4.5 Decision Tree. 1st International Conference on Pervasive Computing, Signal Processing and Applications. 2010; sep, 17-19. Harbin, China.
15. Moudani W, Shahin A, Chakik F, Rajab D. Intelligent Predictive Osteoporosis System. INT J COMPUT APPL. 2011; 32(5):28-37.
16. Dreiher J, Weitzman D, Cohen A. Psoriasis and Osteoporosis: A Sex-Specific Association? J Invest Dermatol. 2009; 129 (1): 1643-1649.
17. Shearer C. The CRISP-DM Model: The New Blueprint for DataMining. Journal of Data Warehousing, 2000; 5(4): 13-22.