مروری بر روش های مبتنی بر درخت تصمیم[1] جهت پیش بینی بیماری دیابت

سعید محمدی [*]، محمد حسین ندیمی

گروه مهندسی کامپیوتر ، دانشکده کامپیوتر ، دانشگاه آزاد اسلامی ، نجف آباد ، اصفهان - ایران

خلاصه

افزایش تعداد بیماران دیابتی ، عواقب این بیماری، هزینه ها و آمار مرگ و میر مرتبط با آن تلاش برای توسعه و بهبود سیستمهای کامپیوتری تشخیص و پیش بینی بر پایه داده کاوی را معنا دار می کند. داده کاوی پزشکی می تواند کمک بزرگی به جامعه پزشکی در کشف الگوهای پنهان ، روابط معنا دار ومهم از بین حجم انبوهی از داده ها واطلاعات موجود با توجه رشد جمعیت بیماران باشد و در نهایت به افزایش صحت پیش بینی بیماری، سادگی روند درمان و کاهش هزینه های مربوط به آن کمک کند. هدف این مطالعه مقایسه تعدادی از کارهای انجام شده اخیر در رابطه با پیش بینی و تشخیص بیماری دیابت بوسیله درخت تصمیم است. در این مقاله روش های ساخت درختان تصمیم بر پایه نوع شکاف داده و انتخاب ویژگی و کشف رابطه آن با صحت نهایی مدل پیش بینی با توجه به تعداد رکوردهای مجموعه داده ، متغیر های ورودی خروجی و ویژگی های انتخاب شده توسط شخص پزشک مورد بررسی قرار می گیرد. در این مطالعه روش های مختلف شکاف داده با مجموعه داده های مختلف از بیماران دیابتی و غیر دیابتی و فاکتور های ریسک[2] متفاوت به همراه صحت نهایی هر کدام از مدل های پیش بینی بررسی و ارزیابی گردید. نتایج تحلیلی نشان داد که تعداد نمونه های مجموعه داده و روش های شکاف داده در میزان صحت طبقه بند ، با هم در ارتباطند و همچنین انتخاب تعداد زیادی از فاکتور های ریسک بخاطر عدم ارتباط برخی از آنها با تشخیص بیماری در افزایش صحت پیش بینی بی تاثیر است.

واژه های کلیدی: داده کاوی پزشکی،پیش بینی، درخت تصمیم ، بیماری دیابت

۱. مقدمه

امروزه بهمراه رشد روز افزون جمعیت ، کاهش فعالیت بدنی افراد به علت گسترش زندگی ماشینی ، مصرف غذاهایی که برای سلامتی مضر هستند وهمچنین روش های متفاوت زندگی ، علاوه بر این که سن ابتلا به دیابت کاهش یافته است ، نرخ ابتلا به آن افزایش یافته است . از طرفی بیماری دیابت می تواند به خطراتی همچون نابینایی و سکته قلبی منجر شود. هزینه جهانی مصرف شده در رابطه با این بیماری در سال۲۰۱۳ برابر ۵۴۸ میلیون دلار می باشد . تعداد بیمارن دیابتی در سال ۲۰۱۴

---

[1] Decision Tree

[*] Corresponding author: Email: saeed-mohamadi@sco.iaun.ac.ir

[2] Risk Factors

آکادمی داده
dataacademy.ir

برابر ۳۸۷ میلیون نفربوده است وپیش بینی شده است تا سال ۲۰۴۰ حدود 642 میلیون نفر در سراسر جهان مبتلا به بیماری دیابت باشند[۱].

در دهه های اخیرحجم اطلاعات و داده های بیماران بسیار گسترده شده است و شخص پزشک نمی تواند روی تمامی فاکتورهای موثر آن تحلیل و ملاحضات مربوطه را داشته باشد ؛ همچنین فاکتورهای ریسک متفاوتی در ارتباط با بیماری دیابت وجود دارد که هر کدام تـاثیرات متفاوتی در فرایند تشخیص و پیش بینی این بیماری دارند . بنابراین استفاده از سیستم های تحلیلی کامپیوتری با توجه به آمار افزایشی ابتلا به این بیماری می تواند تاثیر زیادی در کشف الگوهای مربوط به تشخیص و پیش بینی دیابت و در نهایت کاهش هزینه های درمان آن داشته باشد. داده کاوی یک جستجوی خودکار روی منابع بسیار بزرگ از داده ها برای یافتن الگوها و رفتارهایی که منجر به کشف دانش می شوند است[۲]. یکی از تکنیک های مشهور ساده و قدرتمند داده کاوی که در زمینه های مختلف از جمله پزشکی استفاده می شود ، طبقه بند درخت تصمیم است. شهرت این تکنیک به خاطر پیاده سازی ساده آن برای حجم بالایی از داده در کنار صحت بالا و راحتی درک و تفسیر آن توسط عموم مردم است[۳]. همین موضوع دلیل ما برای انتخاب طبقه بند درخت تصمیم در کار مقایسه ای فعلی جهت پیش بینی بیماری دیابت می باشد.

تا کنون مطالعات زیادی در مورد پیش بینی بیماری دیابت بوسیله درختان تصمیم انجام شده است. برای مثال در سال ۲۰۱۳ آقای Xue-Hui.m و همکارانش بوسیله درخت تصمیم C5.0 ویک مجموعه داده با ۱۸۴۷ نمونه از افراد دیابتی و غیر دیابتی یک مدل پیش بینی را تولید کرده اند و به صحت 77.87% که بیشترین مقدار در مقایسه با دو طبقه بند دیگر در همان مطالعه بود دستیافتند[۴]. همچنین در سال ۲۰۱۴ آقای O.Purnik و همکارانش بوسیله درخت[3] CART و مجموعه داده با ۶۶۴۷ نفراز افراد غیر دیابتی توانستند با صحت 90.50% یک مدل تشخیص و پیش بینی دیابت را بسازند که در۱۱٪ از افراد این بیماری پیش بینی شد. در مطالعات بیان شده و کارهای انجام شده اخیر به مقایسه درخت های تصمیم مختلف با توجه به تعداد رکورد و ویژگی های مجموعه داده اشاره نشده است. اکثر مطالعات از چندین تکنیک طبقه بندی مختلف یا طبقه بند مجموعه ترکیبی[4] جهت بکارگیری چند تکنیک طبقه بندی در جهت ساخت مدل پیش بینی استفاده کرده اند.

درختان تصمیم را می توان با توجه به روش های شکاف داده و انتخاب و یژگی به دسته های زیر تقسیم کرد:

- شکاف بر پایه مفهوم Entropy و Information Theory که خود به دونوع Information Gain و Gain Ratio تقسیم می شود[۵].
- شکاف بر پایه روش Gini Index
- شکاف بر پایه روش آماری Chi-Squared Test

[3] Classification And Regression Tree
[4] Ensemble Classifier

آکادمی داده
dataacademy.ir

شکاف بر پایه آزمایش Chi-Squared دردرخت تصمیم CHAID[5] استفاده شده است که اغلب برای کاربردهای آماری و بازاریابی استفاده می شود و در حوزه پزشکی و کارهای مقایسه شده در این مطالعه کاربردی نداشته است. هدف این مطالعه تحلیل ۷ مورد از کارهای انجام شده اخیر در زمینه پیش بینی بیماری دیابت بوسیله درختان تصمیم با رویکرد مقایسه آنها از نظر الگوریتم های شکاف داده و انتخاب ویژگی و ارتباط آن با تعداد رکوردهای مجموعه داده ، ویژگی ها و فاکتورهای ریسک می باشد. مقایسه تحلیلی کارهای انجام شده با توجه به مقایسه تکنیک های شکاف داده می تواند کمکی به کشف رابطه آن با تعداد رکوردهای مجموعه داده و فاکتور های موثر برای تشخیص و پیش بینی این بیماری شود و درنهایت افزایش صحت طبقه بندی وکاهش هزینه های زمانی ساخت مدل های پیش بینی را در ارتباط با حذف فاکتورهای ریسک کم ارزش یا نامرتبط داشته باشیم. نتایج مقایسه ای نشان داد که با تعداد تعداد رکورد های کم و یکسان، روش Information Gain صحت بیشتری نسبت به Gain Ratio بدست می آورد و همچنین افزایش تعداد نمونه های مجموعه داده در روش Information Gain ارتباط مستقیمی با صحت پیش بینی دارد. با مقایسه روش Gini Index و Information Gain به این نتیجه رسیدیم که افزایش تعداد ویژگی های فاکتورهای ریسک بیماری در روند ساخت مدل پیش بینی به خاطر تفاوت میزان تاثیر آنها با این بیماری نمی تواند ارتباط مستقیمی با افزایش صحت مدل داشته باشد . در ادامه این مقاله به تعاریف و آمارهای مربوط به بیماری دیابت در بخش ۲ اشاره می شود؛ درقسمت ۳ به مفاهیم مربوط به داده کاوی پزشکی و انواع درختان تصمیم می پردازیم بعد از آن در قسمت ۴ توضیح مختصری درمورد تکنیک ها ی طبقه بندی استفاده شده در حوزه داده کاوی و سپس روش ارزیابی طبقه بندی ها داده می شود و در قسمتهای ۵ و ۶ به ترتیب ، مقایسه تحلیلی کارهای اخیر انجام شده در زمینه پیش بینی بیماری دیابت بوسیله درخت تصمیم و نتیجه گیری بیان می شود.

۲. بیماری دیابت

دیابت یک بیماری مزمن یا یک وضعیت متابولیکی (سوخت و ساز)در بدن است که با بالارفتن سطح قند خون در بدن رخ می دهد. بالا رفتن قند خون از اینجا ناشی می شود که لوزالمعده هیچ انسولینی تولید نکند، انسولین تولید شده ناکافی باشد و یا انسولین بخوبی کار خود را انجام ندهد [6]. هورمون انسولین توسط لوزالمعده تولید وباعث شده تا قند خون وارد سلولهای بدن شود وسوخت و ساز صورت گیرد. در بدن انسان از قند خون به عنوان سوخت جهت تولید انرژی استفاده می شود[۷]. دیابت می تواند منجر به تخریب سلول های عصبی و رگ های خون، نابینایی، نارسایی و از کار افتادگی کلیه ، قطع عضو و افزایش ریسک سکته و بیماریهای قلبی شود[۸, ۹]. بر طبق گفته ADA[6] سطح قند خون ناشتا بین ۱۰۰ و ۱۲۵ میلی گرم در دسی لیتر به عنوان پیش دیابت درنظرگرفته می شود[۷].

1.2. انواع دیابت

بیماری دیابت انواع مختلفی دارد که به ترتیب در مورد آنها توضیحاتی خواهیم داد.

[5] Chi-squared Automatic Interaction Detector
[6] American Diabetes Association

آکادمی داده
dataacademy.ir

1.1.2. دیابت نوع ۱

اگر سلول های تولید انسولین در لوزالمعده(که سلول های بتا[7] نام دارند) از بین رفته باشند ، هیچ انسولینی تولید نشده و میزان قند خون افزایش می یابد. این نوع دیابت درهرسنی می تواند رخ دهد اما معمولا دردوران کودکی و یا تا قبل از ۴۰ سالگی پدیدار می شود. علائم و نشانه های آن معمولا بسیارسریع به طورمعمول بعد ازچند هفته آشکار می شوند[۷, ۱۰]. دیابت نوع یک ۵ تا۱۰٪ از تعداد مبتلایان به این بیماری را شامل می شود[۹]. چندین برنامه پیش بینی بصورت آزمایشی در کلینیک های مختلف جهان با مطالعات اضافی برنامه ریزی شده در حال انجام است ولی تا کنون هیچ راه شناخته شده ای برای جلو گیری از دیابت نوع ۱ وجود نداشته است[۱۰].

۲.۱.۲. دیابت نوع۲

اگر لوزالمعده بتواند مقداری انسولین تولید کند اما نه به اندازه کافی و یا انسولین تولید شده به درستی کارنکند ، این نوع دیابت رخ می دهد. اغلب ازسن۲۵ سالگی شروع شده و درافراد بالای ۴۰ سال نیز مشاهده می شود. همچنین دیابت نوع ۲ بین کودکان، نوجوانان وجوانان به صورت فزاینده ای در حال افزایش است. علائم ونشانه هاممکن است به راحتی واضح و آشکار نباشد زیرا شرایط آن به آرامی درطی یک دوره یکساله گسترش می یابد وممکن است تنها دریک آزمایش پزشکی روتین مشخص شود[۷].در این نوع از دیابت ، سلول های بتا در لوز المعده توانایی خود در تولید انسولین کافی و موثر را رفته رفته از دست می دهند[۱۰]. دیابتی های نوع ۲ نیاز به داروی انسولین برای زنده ماندن ندارند ولی۲۰٪ از آنها برای کنترل سطوح قند خون با انسولین درمان می شوند. دیابت نوع دو حدود ۹۰٪ از تعداد مبتلایان به این بیماری را شامل می شود و تشخیص آن در مراحل اولیه کاملا یک کار چالش برانگیز به دنبال وابستگی پیچیده فاکتور های مختلف است [۱۱].

dataacademy.ir

۳.۱.۲. پیش دیابت

پیش دیابت زمانی رخ می دهد که سطح قند خون فرد بالاترازحد طبیعی است، اما برای تشخیص دیابت به اندازه کافی نیست. حدود ۷۰٪ افرادی که پیش دیابت دارند استعداد ابتلا به دیابت نوع ۲ وحمله قلبی را نیز دارا هستند. مطالعات زیادی نشان داده است که کاهش وزن ، فعالیت فیزیکی وهمچنین اصلاح سبک زندگی در بعضی موارد سطح قند خون را به وضعیت معمولی باز گردانده است[۱۰].

۴.۱.۲. دیابت بارداری[8]

اگربدن نتواند انسولین کافی برای پاسخگویی به نیازهای اضافی بدن در زمان بارداری را تولید کند این نوع از دیابت رخ می دهد. ممکن است درطول سه ماهه اول بارداری وجود داشته باشد. این نوع دیابت در ۲ تا ۵ ٪ بارداری ها رخ می دهد و معمولا بعد از دوران بارداری رفع می شود[۷]. مطالعات نشان داده ۴۰٪ از زنانی که دیابت بارداری را تجربه کرده اند در آینده مبتلا به دیابت نوع ۲ شده اند و همچنین فرزند متولد شده مستعد چاقی و ابتلا به دیابت است[۹,۱۰].

[7] Beta Cells
[8] Gestational Diabetes

آکادمی داده
dataacademy.ir

۲.۲. علائم ، تشخیص و درمان

ابتلا به بیماری دیابت منجر به بروز علائمی در بدن فرد بیمار می شود که در ادامه به تعدادی از آنها اشاره کرده و در مورد تشخیص و درمان دیابت مواردی را مطرح می کنیم.

۱.۲.۲ . علائم بیماری دیابت

علائم متداول بروز دیابت شامل، تکرر ادرار، تشنگی زیاد، احساس گرسنگی و اشتهای بالا، عفونت زخم، ترمیم طولانی مدت زخم، اضطراب، خستگی، کاهش وزن، بیحالی شدید و تاری دید می باشد. برخی ازعوامل خطرابتلا به دیابت مانند سابقه خانوادگی، اضافه وزن ، چاقی ، فشارخون بالا ، اعتیاد به مصرف دخانیات و مصرف غذاهای چرب و شیرین می باشد. دیابت می تواند در۱ تا ۲٪ مواقع از عمل جراحی ، بیماری ژنتیکی خاص، دارو، سوء تغزیه و یا عفونت ناشی شود [۷, ۹].

۲.۲.۲ . تشخیص و درمان دیابت

عوامل متعددی مثل معاینه پزشکی، وجود علائم، سابقه پزشکی، آزمایش خون وادرار براساس مقدارقند خون و پروتئین موجود درادراربه تشخیص دیابت کمک می کند. از جمله عواملی که در درمان دیابت موثر است کنترل رژیم غذایی مناسب، انجام فعالیت فیزیکی ، ورزش وهمچنین تزریق انسولین (در ۴۰٪ از مبتلایان به دیابت نوع ۲) می باشد. مطالعات نشان داده است که فعالیت فیزیکی به طور چشمگیری می تواند ریسک ابتلا به دیابت نوع ۲ را کاهش دهد[۷, ۹].

dataacademy.ir

۳.۲. آمار ها

بیماری های مسری از جمله دیابت اصلی ترین علت مرگ ومیر در جهان و به ویژه درکشور های آسیای شرقی محسوب می شوند. قاره آسیا یک سوم جمعیت جهان را در بر دارد. بیشتر از ۶۰٪ جمعیت دیابتی جهان در کشورهای آسیای شرقی زندگی می کنند . همچنین سبک زندگی و نوع تغزیه مردم در سرتا سر دنیا با هم متفاوت است و همین موضوع علت مشاهده آمار های گوناگون مربوط به دیابت در سرتاسر جهان است. طبق گزارشات IDF[9] تعداد بیماران دیابتی در سال ۲۰۱۴ برابر ۳۸۷ میلیون نفر بوده که حدود ۲۰۰ میلیون نفر آن مربوط به کشور های آسیای شرقی بوده است. این درحالی است که نیمی از جمعیت دیابتی جنوب شرق آسیا هنوز مورد تشخیص قرار گرفته نشده اند. پیش بینی شده است تا سال ۲۰۴۰ حدود 642 میلیون نفر در سراسر جهان مبتلا به بیماری دیابت باشند که این مقدار بیشتر از دو برابر جمعیت مبتلا به دیابت در سال ۲۰۰۸ است. سن ابتلا در حال کاهش است. تعداد زنان دیابتی بیشتر ازتعداد مردان دیابتی است و اغلب افراد دارای اضافه وزن می باشند. تا آخر سال ۲۰۱۳ دیابت باعث مرگ 1.5 میلیون نفر درجهان شده و همچنین هزینه های پرداخت شده مربوط به این بیماری ۵۴۸ میلیون دلار بوده است[۱, ۶, ۸]. کشور آمریکا در رابطه با افزایش نرخ شیوع دیابت و هزینه های مصرفی آن رشد بسیار بالایی در دهه اخیر داشته است. در سال ۲۰۱۰ دیابت به عنوان هفتمین علت مرگ ومیر در آمریکا ثبت شده است. در سال ۲۰۱۱ دیابت اصلی ترین دلیل از کار افتادگی کلیه در ۴۴٪ از موارد رخ داده شده آمریکا بوده است. دیابت به رگ های ریز خونی در شبکیه آسیب می رساند که باعث تاری دید می شود ؛ بین سالهای ۲۰۰۵ تا ۲۰۰۸ حدود ۵.۲۸٪ از افراد ۴۰ سال

[9] International Diabetes Federation

آکادمی داده
dataacademy.ir

به بالا، مبتلا به تاری دید چشم ناشی از دیابت بوده اند. در سال ۲۰۱۰ تعداد ۷۳۰۰۰ مورد قطع عضو در افراد ۲۰ سال به بالای مبتلا به دیابت انجام شده است. هزینه های انجام شده مربوط به دیابت اعم از مستقیم و غیر مستقیم در سال ۲۰۱۰ برابر ۲۴۵ میلیون دلار بوده است[۱۰].

۳. داده کاوی پزشکی ، درخت تصمیم

۱.۳. داده کاوی پزشکی

فرایند انتخاب ، کاوش ومدلسازی حجم انبوهی از داده جهت کشف الگوهای ناشناخته یا روابطی که نتایج شفاف و مفیدی را فراهم می کند را داده کاوی گویند[۴]. Jianchao Han در مقاله خود در سال ۲۰۰۹ داده کاوی را اینگونه تعریف کرده است:" به کل فرایند بکارگیری متدولوژی کامپیوتری برای گسترش دانش از طریق داده ها ، داده کاوی می گویند"[۱۲].در دنیای امروز داده های مربوط به بیماران و مریضی های مختلف بسیار گسترده اند، تحلیل و ملاحضات لازم روی تمامی فاکتور های آنها بوسیله شخص پزشک غیر ممکن است. بنابراین به سیستم های هوشمند ی برای ملاحظه فاکتورهای مختلف و شناسایی یک مدل مناسب بین پارامتر های مختلف مشاهده شده نیاز است. از طرفی پزشکان و محققان باید قادر به آزمایش و جستجوی اطلاعاتی باشند که به آنها اجازه می دهد تشخیص دقیق تری از بیماری ها داشته باشند. توسعه سیستم های پشتیبانی تصمیم گیری تشخیص پزشکی می تواند به جامعه پزشکی در روند تشخیص و پیش بینی بیماری ها کمک کند. استفاده از سیستم های هوشمند برای تشخیص و درمان بیماری ها، باعث کاهش جدی خطای سیستم پزشکی و در نتیجه کاهش هزینه ها و از دست رفتن جان انسان ها می شود[۱۱, ۱۳].

dataacademy.ir

داده کاوی در حوزه سلامت می تواند جهت فراهم کردن تحلیل داده های مراکز پزشکی برای ارائه منابع بهتر ، شناسایی و پیش بینی سریعتر از بیماری ها استفاده شود، که صرفه جویی در هزینه های ناخواسته و گران آزمایشات مربوط به حوزه سلامت و پزشکی را در بر دارد[۱۴]. داده کاوی در حوزه پزشکی تفاوت هایی با حوزه های غیر پزشکی دارد وبا چالش هایی رو به رو است برای مثال کیفیت داده های پزشکی به خاطر مقادیرمفقود[۱۰] شده ناشی از بیماریهای مشابه و مریض های مشابه پایین است. سیستم اطلاعات بیمارستان برای اهداف اقتصادی طراحی شده است نه تحقیقاتی. همچنین محقق باید در مورد داده های بیمار، حفظ حریم خصوصی و محرمانگی آن طبق ضوابط تعیین شده مطمئن شود. اشتباه در ورود داده درمرحله پیش پردازش و ارزیابی منجر به کشف الگوهای مشکوک و در نتیجه شکایات مردم می شود[۱۵]. در حوزه پزشکی تکنیکهای مختلف داده کاوی اعم از طبقه بندی ، خوشه بندی وقوانین انجمنی در مطالعات سال های اخیر استفاده شده اند.

تکنیک های طبقه بندی از دوگام پیروی می کنند. در گام اول (یادگیری یا آموزش[۱۱]) :یک مدل طبقه بندی متشکل از قوانین طبقه بندی، باتحلیل داده های یادگیری حاوی برچسب های کلاس ایجاد می شود. در گام دوم (آزمایش[۱۲]) : طبقه بند (بوسیله داده های آزمایش) برای میزان صحت و تواناییش جهت طبقه بندی رکورد های ناشناخته جهت پیش بینی ، آزمایش

[10] Missing Values
[11] Train
[12] Test

آکادمی داده
dataacademy.ir

می شود. مرحله آزمایش در مقایسه با یادگیری بسیار ساده ، با پیچیدگی پائین و نیازمند محاسبات کمتر است. فرایند یادگیری در طبقه بندی به صورت نظارتی و درخوشه بندی به صورت غیر نظارتی انجام می شود. در روش نظارتی برچسب های کلاس پیش بینی بر خلاف روش های غیر نظارتی ، از قبل مشخص است . طبقه بندی جهت دسته بندی داده در برچسب های کلاس از پیش تعریف شده استفاده می شود . کلاس یک خصیصه در مجموعه داده است که در آمار به عنوان یک متغیر وابسته در نظر گرفته می شود.در حوزه پزشکی تکنیک های طبقه بندی می توانند برای کمک به تشخیص و پیش بینی بیماری های متفاوت از جمله دیابت استفاده شوند. الگوریتم های طبقه بندی مختلفی نظیر طبقه بند بیزین ساده ، شبکه عصبی ، SVM[13]، رگرسیون، درخت تصمیم و طبقه بند دسته جمعی برای تشخیص و پیش بینی انواع بیماری ها در مطالعات متعددی مورد بررسی قرار گرفته اند. در بین تکنیک ها و الگوریتم های مختلف داده کاوی در این مقاله قصد داریم روی طبقه بند درخت تصمیم تمرکز کنیم.

۲.۳ . طبقه بند درخت تصمیم

درخت تصمیم یک تکنیک طبقه بندی برپایه یادگیری نظارتی و فرمی جهت بیان یک نگاشت است[8]. توانایی درخت تصمیم در مدلسازی روابط غیرخطی می باشد[۱۶]. در ساختار درخت تصمیم،یک مدل پیش بینی از مجموعه قوانین استخراج شده طبقه بندی بدست می آید[13]. این پیش بینی می تواند در حوزه های مختلف از جمله پزشکی صورت گیرد.

۱.۲.۳. یادگیری

dataacademy.ir

درخت تصمیم با شکاف مجموعه داده منبع به زیر مجموعه هایی بر اساس یک آزمایش مقدار ویژگی توسط عمل بازگشتی تقسیم و غلبه برای انتخاب ویژگی ها بوسیله یک تعداد پویش پارامتری روی مجموعه داده، از بالا تا پایین درخت یادگیری می شود. این فرایند روی هر زیر مجموعه مشتق شده به صورت بازگشتی تکرار می شود. عمل بازگشتی وقتی کامل می شود که یا شکاف امکان پذیر نیست یا یک طبقه بند یکتایی بتواند روی هر عنصر از زیر مجموعه مشتق شده به کارگرفته شود[۱۲, ۱۳].

2.2.۳. القای درخت تصمیم

درخت شامل نودهای ریشه، میانی(فرزند) و برگ است و متد انتخاب ویژگی کلید فرآیند ساخت درخت است. خصیصه های انتخاب شده نود های درخت تصمیم هستند. در خت تصمیم فرض می کند که نمونه ها به کلاس های مختلفی تعلق دارند که مقادیر مختلفی حداقل در یکی از خصیصه ها یشان وجود دارد. القای درخت گام یادگیری طبقه بندی است. بهترین ویژگی که داده های یادگیری را تقسیم می کند، بعنوان ریشه درخت در نظر گرفته می شود. نمونه ها با شروع از نود ریشه بر اساس مقادیر ویژگیشان مرتب شده و طبقه بندی می شوند. هر نود در درخت تصمیم نمایانگر یک ویژگی در یک نمونه جهت طبقه بندی شدن است. و بر اساس آن ویژگی و تمام پاسخ های ممکن ، یک سوال دارد . در رابطه با تصمیم گیری برای یک مورد از ریشه شروع می کند و از یک مسیر بر اساس سوالات و پاسخ های نودهای میانی تا زمانی که بلاخره به یک برگ برسد

[13] Support Vector Machine

آکادمی داده
dataacademy.ir

پیروی می کند و برچسب برگ برای کلاس آن مورد استفاده می شود. در نهایت هر برگ درخت با استفاده از یکی از کلاس ها برچسب می خورد که نمایانگر نتیجه پیش بینی می باشد[۳, ۴, ۱۳, ۱۵]. مسئله القاء طرح ریزی و ایجاد کردن درخت های تصمیم بهینه دودویی یک مسئله NP کامل است و تئوریسین ها راه های اکتشافی موثر را برای نزدیک شدن به ساخت درختان تصمیم بهینه جستجو می کنند[۳].

3.2.3 . الگوریتم های شکاف داده و انتخاب ویژگی

برای پیدا کردن بهترین ویژگی و شکاف داده های مجموعه داده متد های متفاوتی وجود دارد[۳]. هدف الگوریتم شکاف دهنده پیدا کردن متغیر و آستانه ای است که ترتیب همگن نتایج دو زیر گروه یا بیشتر از نمونه ها را بیشینه می کند . این قدم در نود برگ تا کامل شدن ساختار درخت تکرار می شود[۱۷]. بهترین ویژگی برای شکاف توسط متد انتخاب ویژگی انتخاب می شود که بهترین جدا سازی رکورد ها را برای هر یک از برچسب های کلاس انجام می دهد. [15]معمول ترین الگوریتم های ریاضی شکاف مجموعه داده به ۳ دسته تقسیم می شوند که به ترتیب شرح داده می شوند.

۱.۳.۲.۳ . شکاف داده بر پایه Information Gain

مفهوم Information Gain توسط آقای Hunt و همکارانش در سال ۱۹۶۶ ارائه شد. آقای Quinlan در سال ۱۹۷۹ درخت تصمیم ID3 را ارائه داد. این درخت تصمیم برای انتخاب ویژگی جهت شکاف داده ها و القای درخت تصمیم از Information Gain برپایه Entropy استفاده می کند. الگوریتم ID3 یک رویکرد حریصانه است که در آن درخت تصمیم گیری به صورت بازگشتی ازبالا به پایین بروش تقسیم وغلبه ساخته می شود. درخت با مجموعه یادگیری از تاپل ها و برچسب های کلاس مربوط به آنها ساخته می شود. مجموعه یادگیری با فرایند ساخت درخت تصمیم به صورت بازگشتی به زیر مجموعه های کوچکتر تقسیم می شود . از عمل Entropy برای تقسیم نمونه ها به دو زیر کلاس استفاده شده ومیزان همگنی و یکدست بودن مجموعه داده محاسبه می شود ؛ اگر مجموعه داده کاملا همگن بود Entropy "۰" می شود. در غیر اینصورت به صورت مساوی تقسیم می شود ، تا Entropy مقدار "۱" داشته باشد. Information Gain بستگی به کاهش Entropy دارد. یک ویژگی با بیشترین Entropy بیشترین مقدار Information Gain را برمی گرداند. بنابراین یک ویژگی با بیشترین Information Gain به عنوان ویژگی شکاف در درخت تصمیم به ریشه منتقل می شود. سپس نودی که بیشترین Information Gain را دارد ، والد نسل بعدی می شود. این فرایند تکرار می شود تا زمانی که به نود برگ برسد و درخت تصمیم کامل شود. درخت تصمیم بر اساس پیدا کردن بیشترن شاخه های همگن ساخته می شود. رابطه (۱) نحوه محاسبه Information Gain و رابطه (۲) نحوه محاسبه Entropy را نشان می دهد.

$$\text{Info Gain(Parent, Child)} = \text{Entropy(parent)} - [p(c_1) * \text{Entropy}(c_1) + p(c_2) * \text{Entropy}(c_2) \ldots] \quad (۱)$$

$$\text{Entropy} = \sum_{i=1}^{m} - p(c_i) * Log_2 * p(c_i) \quad (2)$$

آکادمی داده
dataacademy.ir

در رابطه (۱) Parent و Child نمایانگر والد و فرزند می باشند . بر طبق رابطه (۲) احتمال نود فرزند i برابر با $p(c_i)$ است و m تعداد کلاس ها برای ویژگی هدف است[3, 14]. بعد از درخت ID3 آقای Quinlan در سال ۱۹۹۳ الگوریتم C4.5 را ارائه داد که توسعه ای از الگوریتم ID3 است. درختC4.5 از روش Gain Ratio جهت القا استفاده می کند. همانطور که از رابطه (۳) مشخص است، Gain ratio از تقسیم Information Gain بروی Splitting Information بدست می آید. برای کاهش تاثیرات تبعیضانه ای که ممکن است به خاطر تعداد زیادی از مقادیر برای یک ویژگی رخ دهد از این الگوریتم شکاف استفاده می شود . یک درخت تصمیم بر اساس Gain Ratio از لحاظ اندازه گیری صحت و مدیریت مسائل بزرگ ، بهتر از Information Gain می باشد. درخت تصمیم با در نظر گرفتن تعداد و اندازه شاخه ها برای یک ویژگی مورد نظرساخته می شود. Gain Ratio تاثیرات تبعیضانه Information Gain را کاهش می دهد و به صورت یکنواخت مقادیر وسیعی را به یک ویژگی خاص اعطا می کند. یک ویژگی با بیشترین مقدار Gain Ratio بعنوان ویژگی شکاف دهنده برای ساخت درخت تصمیم انتخاب می شود[۱۴]. برای بدست آوردن Splitting Information از رابطه (۴) استفاده می کنیم ؛ که در آن مجموعه داده D به تعداد v پارتیشن مطابق با v خروجی از آزمایش روی صفت خاصهٔ A تجزیه می شود[۲]. درخت های تصمیم دیگری مانند j48,C5.0 نیز از Gain Ratio استفاده می کنند.

$$\text{Gain Ratio}=\frac{\text{Information Gain}}{\text{Splitting Info}} \qquad (۳)$$

$$\text{Split Info}_A(D)=\sum_{j=1}^{v}\frac{|D_j|}{|D|}*Log_2[\frac{|D_j|}{|D|}] \qquad (4)$$

dataacademy.ir

۲.۳.۲.۳. شکاف داده برپایهٔ Gini Index

Gini Index در سال ۱۹۸۴ توسط آقای Breiman و همکارانش ارائه شد . درخت CART از Gini Index استفاده می کند. این درخت برای مسائل طبقه بندی و رگرسیون استفاده می شود. Gini Index سطح ناخالصی داده های دریافت شده را اندازه گیری کرده ، یک شکاف باینری برای هر متغیردر نظر گرفته و یک درخت دودویی می سازد؛ که خروجی هرنود میانی درخت دقیقا دو کلاس برای ویژگی دریافت شده است. Gini Index برای هر ویژگی محاسبه می شود و سپس بصورت بازگشتی ویژگی با کمترین مقدار بعنوان ویژگی شکاف دهنده داده انتخاب شده و درخت ساخته می شود[۱۴]. برای فهم ساده، درخت تصمیم به مجموعه ای از قوانین IF-Then از ریشه تا برگ تجزیه می شود. قسمت IF بعنوان پیش گویی کننده و قسمت Then بعنوان برچسب کلاس نود برگ (نتیجه مورد انتظار) در نظر گرفته می شود[۱۶]. اگر احتمال i اُمین کلاس p_i باشد، برای k کلاس هدف از ویژگی مورد نظر Gini Index به صورت رابطه (۵) محاسبه می شود.

$$\text{Gini Index}=1-\sum_{i=1}^{k}p_i^2 \qquad (۵)$$

۳.۳.۲.۳ . شکاف داده بر پایه Chi-Squared

Chi-Squared test در سال ۱۹۰۰ توسط آقای Pearson ارائه شد[۱۸]. درخت تصمیم CHAID از تکنیک Chi-Squared استفاده می کند. در این درخت از معیار انتخاب صفت بر اساس آزمایش X^2 استفاده می شود. به طور معمول در

آکادمی داده
dataacademy.ir

زمینه بازاریابی مستقیم برای انتخاب گروه های مشتریان وپیش بینی چگونگی پاسخ آنها به بعضی از متغیر هایی که تاثیر روی بقیه متغیر ها می گذارند ، استفاده می شود. CHAID تعامل بین متغیرها را در مجموعه داده با شناسایی گروه های گسسته ای از پاسخدهندگان کشف کرده ، ودرپی کشف و پیش بینی تاثیرات برمتغیرهای وابسته با درنظرگرفتن پاسخدهیشان روی متغیرهای توضیحی می باشد. ازآنجاییکه CHAID به داده آماری نیاز دارد ، نیازی به گسسته سازی متغیر های عددی ندارد[۱۲]. درخت CHAID در مطالعات مربوط به تشخیص و پیش بینی بیماری توسط داده کاوی بسیار کمتر از بقیه درخت های تصمیم مورد استفاده قرار گرفته است.

۴.۲.۳. بیش برازش[۱۴] در درخت تصمیم

یک درخت تصمیم یا هر فرضیه یادگیری از h بیش برازش داده های یادگیری گفته می شود اگر یک فرضیه دیگری از h (h_2) وجود داشته باشد که دارای خطای بزرگتری از h است وقتی روی داده های یادگیری ، آزمایش شده باشد؛ و اما دارای یک خطای کوچکتر از h باشد وقتی روی کل مجموعه داده آزمایش شده است. این اتفاق زمانی رخ می دهد که میانگین Entropy کمتر از آستانه تعیین شده است ، یا تعداد داده در یک شاخه درخت کمتر از تعداد داده تعیین شده در یک دایرکتوری است[۱۳]. برای جلو گیری از بیش برازش ۲ روش مطرح می شود که به اختصار شرح داده می شود.

- توقف الگوریتم یادگیری قبل از اینکه به نقطه کاملا برازنده (یا اندازه بودن) از داده های یادگیری برسد.
- هرس کردن کامل درخت تصمیم. هرس مشکل بیش برازش مجموعه داده را با متد های آماری حل می کند . اگر شاخه های مهم بصورت مناسب هرس شوند فهم ساختار داده کلی برای افراد ساده می شود .

dataacademy.ir

اگردرخت در هر دو روش دارای صحت پیش بینی یکسان بود، درخت با تعداد برگ کمتر ترجیح داده می شود.
دو نوع هرس وجود دارد. پیش هرس در هر مرحله القا درخت تصمیم همراه با فاز ساختن درخت انجام می شود. پس هرس شاخه ها را از یک درخت کامل رشد کرده حذف می کند. برای هر نود غیر برگ در درخت الگوریتم هرس نرخ خطای مورد انتظاری تخمین می زند که اگر زیر درخت در آن نود هرس شد رخ می دهد. بنابراین نرخ خطای مورد انتظار زمانی رخ می دهد که نود هرس نشده با استفاده از نرخ های خطای هر شاخه ای که ترکیب شده بوسیله وزن دهی مطابق نسبت مشاهدات در امتداد هر شاخه تخمین زده شود. اگر هرس کردن نود به نرخ خطای بیشتری از آنچه انتظار میرفت منجر شود، زیردرخت نگه داشته می شود در غیر اینصورت هرس می شود[۴].

۵.۲.۳. مزایا و معایب درختان تصمیم

مزایا:

- یکی از خصیصه های مفید درخت های تصمیم قابلیت درک ، تفسیر ، تجسم تمرکز کلاس داده و فهم ساده آن توسط اشخاصی است که آن را مشاهده می کنند . همچنین افراد به راحتی متوجه می شوند که چرا یک نمونه در یک کلاس خاص طبقه بندی شده است و می توانند با این دید که کدام خصیصه تاثیر بیشتری روی یک کلاس داده گذاشته است ، کل ساختار داده را تحلیل کنند[۳, ۱۵]

[14] Over Fitting

آکادمی داده
dataacademy.ir

- درخت تصمیم برای ویژگی های گسسته جهت ساخت مدل طبقه بندی بسیار مناسب است و سرعت قابل توجهی در انجام کار دارد.
- پیاده سازی درخت تصمیم ساده است ، برای حجم بالایی از داده به خوبی کار می کند و نتایج با دقت و صحیحی از طبقه بندی را به دست می آورد[۱۳].

معایب:

- اغلب درختان تصمیم اگر مسئله نیاز به پارتیشن بندی قطری یا اُریب[۱۵] داشته باشد خوب کار نمی کنند.
- درخت تصمیم می تواند برای بعضی از مفاهیم به خاطر وجود تکرار[۱۶] در یک مسئله و همچنین تعداد ویژگی های زیاد در مجموعه داده به شدت نمایش پیچیده ای داشته باشد. برای رفع این مشکل باید عمل هرس انجام شود یا استفاده از الگوریتمی که ویژگی های پیچیده در یک نود را با دوری از تکرار و کپی کردن پیاده سازی کند [۱۳].

۴. طبقه بندی در حوزه سلامت و ارزیابی طبقه بندی ها

1.۴. الگوریتم های طبقه بندی در حوزه سلامت

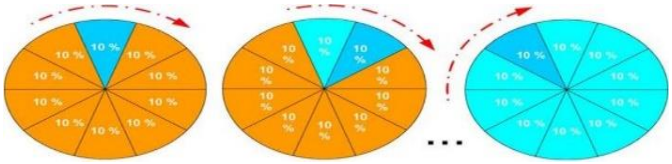
الگوریتم های طبقه بندی زیادی در حوزه سلامت وجود دارند که با توجه به معایب و محدودیت هایی که دارا هستند ما را دراستفاده درخت تصمیم برای پیش بینی بیماری دیابت مصمم نمود. در حوزه داده کاوی پزشکی از شبکه عصبی نیز استفاده می شود. شبکه عصبی از پس داده های نویز دار به خوبی بر می آید ولی فرایند یادگیری آن بسیار طولانی است؛ (حدود $\frac{10}{31}$ کند تر از درخت CART). در شبکه عصبی فقدان قدرت توصیف و شفاف سازی مدل بر خلاف درخت تصمیم، بدلیل تعداد زیاد نود های وزن دهی شده را شاهد هستیم. وزن دهی شبکه بسیار مهم و حساس است و روی نتیجه طبقه بندی تاثیر می گذارد. از آنجاییکه بصورت جعبه سیاه عمل می کند نتایج شبکه عصبی برخلاف درختان تصمیم بوسیله پزشک قابل تفسیر نیست. طبقه بند بیزین ساده الگوریتم طبقه بندی ساده ای است و سرعت بالایی دارد. یک مشکل آن فرضیه بنیادیش مبنی مستقل بودن خصیصه ها از همدیگر است. مثلا در فیلد پزشکی بسیاری از بیماری ها و شرایط سلامتی به شدت به هم وابستگی دارند . مثل فشار خون و BMI[۱۷] که در تشخیص بیماری دیابت به هم وابستگی دارند و در نتیجه طبقه بندی می تواند خطا ایجاد کند. در طبقه بند SVM نیاز به انتخاب یک تابع کرنل مناسب داریم که خود یک چالش محسوب می شود؛ زیرا این طبقه بند توابع کرنل زیادی فراهم می کند.SVM برای مسائل طبقه بندی دو کلاسه طراحی شده ، مرحله یادگیری آن بسیار کند است و منابع محاسباتی گسترده ای نیاز دارد[۱۵].

[15] Diagonal Partitioning
[16] Replication
[17] Body Mass Index

آکادمی داده
dataacademy.ir

۲.۴. ارزیابی طبقه بند ها

قبل از ساخت و ارزیابی مدل پیش بینی باید عمل پیش پردازش داده که شامل تمیز کردن داده ، حذف مقادیر نویز دار و نرمالیزه کردن داده است انجام شود. پس از آن عمل K-Fold Cross Validation انجام می شود. بصورتی که کل مجموعه داده تصادفی به K زیر مجموعه ویژه متقابل (از D_1 تا D_k) با اندازه های حدودا یکسان تقسیم می شود. تعداد زیر مجموعه ها معمولا ۱۰ در نظر گرفته می شود. مدل طبقه بندی K دفعه آموزش و آزمایش می شود. هر دفعه مدل طبقه بندی روی همه داده ها به جز یک قسمت آموزش می بیند و سپس روی تک قسمت باقیمانده آزمایش می شود. این فرایند در شکل (۱) نشان داده شده است. این عمل باعث می شود در مدل پیش بینی بیش برازش رخ ندهد[۸, ۱۷].



شکل(۱):عمل 10-fold Cross validation [17]

بعد از ساخته شدن مدل پیش بینی و آموزش و آزمایش آن ، ماتریس اغتشاش[18] تولید می شود. مقادیر TN,TP,FN,FP به ترتیب نمایانگر False Positive (به اشتباه بیمار پیش بینی شده)، False Negative(به اشتباه سالم پیش بینی شده)، True Positive(بدرستی بیمار پیش بینی شده) و True Negative(بدرستی سالم پیش بینی شده) می باشند. مدلی که کمترین مقدارFN را داشته باشد مناسب تر از بقیه طبقه بند ها است و تاثیر زیادی در نتیجه ارزیابی دارد. ماتریس اغتشاش در جدول (۱) نمایش داده شده است.

جدول(۱):ماتریس اغتشاش

		کلاس های پیش بینی شده	
		Yes	No
کلاس های واقعی	Yes	TP	FN
	No	FP	TN

برای ارزیابی و محاسبه کارایی مدل پیش بینی معمولا از ۳ معیار ارزیابی Accuracy(صحت) , Sensitivity (حساسیت) و Specificity (ویژگی) استفاده می شود که نحوه محاسبهٔ آنها به ترتیب در رابطه های (۶),(۷) و (۸) آمده است.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (۶)$$

[18] Confusion Matrix

آکادمی داده
dataacademy.ir

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (۷)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (۸)$$

۵ . تحلیل مقایسه ای مدل های پیش بینی

دراین قسمت از مقاله ۷ کار انجام شده اخیر با تکنیک های مختلف درخت تصمیم در جهت تشخیص و پیش بینی بیماری دیابت با یکدیگر مقایسه می گردد.

۱.۵. مطالعه اول

در سال ۲۰۰۹ آقای Jianchao. H و همکارانش یک مدل پیش بینی بر پایه طبقه بند درخت تصمیم جهت پیش بینی بیماری دیابت ساختند. از مجموعه داده PIMA که شامل اطلاعات افراد دیابتی و غیر دیابتی است استفاده شد. مجموعه داده شامل ۷۶۸ نمونه با ۸ متغیر است که تمام نمونه ها زنان هندی با حد اقل سن ۲۱ سال هستند. ویژگی های استفاده شده در این مطالعه برای پیش بینی بیماری دیابت برابر با تعداد دفعات بارداری بیمار، مقدار آزمایش قند خون ۲ ساعته ، فشار خون انبساطی[۱۹]، ضخامت پوست عضله بازو، میزان انسولین ۲ ساعته ، BMI ، سابقه خانوادگی دیابت، سن و کلاس پیش بینی است فاز پیش پردازش داده به تنهایی حدود۶۰٪ زمان کل پروژه را صرف می کند و تلاش گسترده ای در کل فرایند داده کاوی دارد. بعد از فاز پیش پردازش داده و مدیریت داده های مفقودی ویژگی های ضخامت پوست عضله بازو، میزان انسولین ۲ ساعته از لیست متغیر ها حذف شدند و بقیه متغیر ها گسسته سازی شده و به دسته های مختلفی تقسیم بندی شدند . با این کار تمرکز روی زیر مجموعه ها به جای تک تک متغیر ها افزایش پیدا می کند و پیچیدگی تحلیل مدل پیش بینی را بدون از دست دادن صحت کاهش می دهد. در نهایت ۷۲۳ رکورد برای ساخت مدل پیش بینی انتخاب شد. در این تحقیق برای اینکه مقیاس تاثیر هر متغیر روی نتایج بخاطر تفاوت در رنج آن ها استاندارد سازی شود از نرمالسازی Min-Max استفاده شد؛ برای مثال BMI از رنج ۱۸. ۲ تا۶۷. ۱ و قند خون از ۴۴ تا ۱۹۹ است. درخت تصمیم ID3 انتخاب و به وسیله ابزار Rapid miner مدل طبقه بندی ساخته شد. تحلیل داده توسط نرم افزار Rapid miner نشان داد که بیمار با قند خون بالا بیشترین احتمال برای ابتلا به دیابت را داراست . با این دلیل مقدار قند خون ناشتا[۲۰] به عنوان نود ریشه درخت در نظر گرفته شد. در مجموعه داده PIMA ۲۴۸ بیمار دیابتی و ۴۷۵ نفر بیمار نبودند. درخت تصمیم ۲۳۱ نفر بیمار دیابتی و ۴۹۲ نفرغیر دیابتی را پیش بینی کرد. از ۲۳۱ نفر ۳۳ نفر اشتباه پیش بینی شد که پیش بینی صحیح به ۱۹۸ نفر کاهش یافت . پس از بدست آوردن ماتریس اغتشاش مقادیر Accuracy وSensitivity و Specificity به ترتیب برابر 88.50% 83. 79٪ و ۰۵. 93٪ بدست آمد.

[19] Diastolic Blood Pressure
[20] Fasting Blood Sugar

آکادمی داده
dataacademy.ir

۲.۵ . مطالعه دوم

در سال ۲۰۱۱ آقای Asma. A و همکارانش از طبقه بند درخت تصمیم برای پیش بینی دیابت نوع ۲ استفاده کردند. از مجموعه داده PIMA استفاده شده و همهٔ ۷۶۸ نفر از بیماران انتخاب شده منحصراً زنان هندی با حداقل ۲۱ سال سن هستند. مجموعه داده شامل ۹ ویژگی تعداد دفعات بارداری بیمار، مقدار تست قند خون ۲ ساعته، فشار خون انبساطی، ضخامت پوست عضله بازو، میزان انسولین ۲ ساعته، BMI ، سابقه خانوادگی دیابت، سن و کلاس پیش بینی است . پس از پیش پردازش داده و مدیریت داده های مفقودی تعداد۷۲۴ نمونه با ۶ ویژگی قند خون ۲ ساعته ,فشار خون انبساطی ,BMI , سابقه خانوادگی دیابت و سن باقی ماندند.

گسسته سازی متغیر های عددی جهت کاهش پیچیدگی مسئله انجام شد. برای مثال تعداد دفعات بارداری ۰ یا ۱ برابر کم ، بین ۲ تا ۵ برابر متوسط و بیشتر از ۶ بار ، زیاد در نظر گرفته شد. برای القای درخت تصمیم از نرم افزار WEKA استفاده شد. مجموعه داده با انتخاب الگوریتم J48 که یک یادگیرنده و پیاده سازی از درخت تصمیم C4. 5 می باشد طبقه بندی شد. در نهایت از ویژگی های انتخاب شده و ۷۲۴ نمونه مجموعه داده 566 نمونه یا ۱۸. ۷۸٪ بدرستی طبقه بندی شد که Accuracy مدل محسوب می شود و ۱۵۸ نمونه یا ۸۲. ۲۱٪ از کل نمونه ها اشتباه طبقه بندی شد. Specificity و Sensitivity مدل پیش بینی به ترتیب برابر80.50 ٪و ۷۲.۳۳٪ بدست آمد.

dataacademy.ir

۳.۵. مطالعه سوم

در سال ۲۰۱۳ Xue-Hui Meng و همکارانش طبقه بند های رگرسیون ، شبکه عصبی و درخت تصمیم را برای پیش بینی دیابت و پیش دیابت بوسیله فاکتور های ریسک معمول دراین بیماری به کار گرفتند. نمونه های مورد مطالعه از مردم کشور چین می باشند. مجموعاً ۱۴۸۷ فرد ۲۰ سال به بالا در این مطالعه شرکت داده شده که اطلاعات آنها از سال ۲۰۰۷ تا ۲۰۰۸ جمع آوری شده است. ۷۳۵ نفر دیابتی (با قند خون ناشتا بالا) و پیش دیابتی و۷۵۲ نمونه معمولی (بوسیله چک آپ فیزیکی در ۲ سال گذشته) تائید شدند . مصاحبه استاندارد برای بدست آوردن تاریخچه خانوادگی دیابت ، اندازه گیری های مربوط به بدن و فاکتورهای ریسک سبک زندگی انجام شد. سپس مدل های پیش بینی با ۱۲ متغیر ورودی و ۱ متغیر خروجی بوسیله اطلاعات استاندارد دریافت شده از مصاحبه ایجاد شد. رضایت کتبی از شرکت کننده ها قبل از جمع آوری داده بدست آمد. داده های یادگیری شامل ۷۰٪ شرکت کنندگان(۱۰۳۱مورد) و داده های آزمایش ۳۰٪ شرکت کنندگان (۴۵۶مورد) را تشکیل دادند. مدل پیش بینی با ابزار spss modeler ver14. 1 طرح ریزی شد. ویژگی های مهم در مورد نمونه ها در این تحقیق از جمله سن، جنسیت ، وضعیت تاهل ، سطح سواد، تاریخچه خانوادگی دیابت، مصرف قهوه، میزان خواب، BMI، فعالیت فیزیکی، استرس کاری، مصرف غذاهای شور و مصرف ماهی می باشند. مقدار BMI با وزن به کیلوگرم تقسیم بر مربع قد به متر (kg/m^2) محاسبه شد؛ و BMI بیشتر مساوی ۲۵ به عنوان اضافه وزن در نظر گرفته شد. متغیر مستقل(خروجی) یک دسته بندی باینری دارد (۰ یعنی معمولی و ۱ یعنی دیابتی). در این مطالعه فاکتور های ریسک سبک زندگی شامل متغیر های مختلفی می باشد که استرس کاری یکی از آنها محسوب می شود؛ بنابراین کل تعداد متغیر های

آکادمی داده
dataacademy.ir

ورودی ۱۹ مورد شد. ازدرخت C5. 0 بر پایه Information Gain جهت ساخت مدل پیش بینی استفاده شده و درخت تصمیم بهترین نتیجه را دربین دو طبقه بند دیگر با Sensitivity،Accuracy وSpecificity بترتیب ۷۷.۸۷٪، ۸۰.۶۸٪ و ۷۵.۱۳ ٪ بدست آورد. سن بیشتر ، تاریخچه خانوادگی دیابت ، BMI بالا ، مصرف غذای شور، فعالیت بدنی کم و میزان استرس مربوط به کار با این بیماری ارتباط مثبت داشت در حالی که سطح تحصیل و نوشیدن قهوه ارتباط منفی با این بیماری داشت. بقیه ویژگی ها ازجمله مصرف سیگار،مصرف الکل ، خوردن سبزیجات و میوه ارتباط مستقیم با این بیماری داشتند.

۴.۵ . مطالعه چهارم

در سال ۲۰۱۴ آقای Saba . b و همکارانش درخت های تصمیم C4. 5 , Id3, CART را به صورت یک مجموعه طبقه بند ترکیبی برای پیش بینی بیماری دیابت بوسیله دو مجموعه داده Uci,Biostate به کار گرفته اند. طبقه بندهای دسته جمعی استفاده شده در این مطالعه از تکنیک های Majority Voting, Adaboost, Bayesian Boosting, Stacking and Bagging استفاده می کنند. نتایج تجربی نشان داد که تکنیک Bagging بهتر از بقیه عمل می کند. هدف اصلی این مطالعه شناسایی بهترین چارچوب طبقه بندی دسته جمعی برای درختان تصمیم است که کمک به شناسایی موثر بیماران دیابتی و مهمتراز آن صحت بالای بدست آمده دارد. درخت های تصمیم مختلف با معیارهای شکاف مختلف از جمله Information Gain, Gain Ratio , Gini Index به عنوان طبقه بند های پایه استفاده شدند. هر طبقه بند بر اساس داده های آموزش مرحله یادگیری را اجرا می کند. ایده اساسی طبقه بند های دسته جمعی، وزن دهی چندین طبقه بند مجزا و ترکیب آنها برای بدست آوردن نتیجه ایست که بهتر و شایسته تر از هر یک از طبقه بند ها می باشد . در این مطالعه از ابزار Rapid Miner5 استفاده شد. بوسیله 10 Fold cross validation از ۹۰٪ داده ها بعنوان یادگیری و ۱۰٪ داده ها برای آزمایش استفاده شد. نتایج بدست آمده نشان داد که بهترین مجموعه طبقه بند ترکیبی Bagging می باشد. برای ارزیابی کلی طبقه بند مقادیر Accuracy وSensitivity و Specificity به ترتیب برابر با ۹۱.۵۶٪ ، ۹۵.۶۳٪ و ۶۸.۳۳٪ بدست آمد.

۵.۵. مطالعه پنجم

در سال ۲۰۱۴توسط O. Purnik و همکارانش برای شناسایی افرادی که ریسک پایینی در ابتلا به دیابت دارند بوسیله پایگاه داده TLGS که مربوط به مرکز قندخون و چربی شهر تهران است یه مدل پیش بینی کننده ساختند. اطلاعات افراد طی ۴ فاز در سالهای مختلف جمع آوری شد: فاز ۱:1999-2001, فاز ۲:2002-2005 , فاز ۳ :2005-2008 و فاز ۴: 2009-2012. تعداد ۱۵۰۰۰نفر با سن بیشتر مساوی ۳ سال از۱۳ منطقه مختلف شهر تهران با متد نمونه برداری خوشه تصادفی انتخاب شدند. افرادی که از قبل دیابت داشتند حذف شد و ۱۰۳۱۰ نفر بدون دیابت باقی ماندند. اطلاعات افرادی که دارای مقدیر مفقودی درمورد مقدار قند ناشتا و قند دوساعته بودند از مجموعه داده حذف شدند. در نهایت۶۶۴۷ نفر غیر دیابتی برای ساختن مدل پیش بینی انتخاب شدند. بوسیله پرسشنامه و مصاحبه اطلاعاتی از قبیل :سن ، جنسیت ، وضعیت تاهل ، میزان تحصیلات، مصرف دخانیات ، فعالیت های بدنی تاریخچه پزشکی و دارویی جمع اوری شد. اندازه قد، وزن، دور کمر، ران ، مچ و BMI بر طبق استاندارد بدست آمد. فشار خون از بازوی راست با فشار سنج جیوه ای بدست آمد و قند خون ناشتا بوسیله آزمایش خون اندازه گیری شد. سپس نمونه های خون جهت تحلیل در آزمایشگاه مرکز تحقیقات ۳۰ تا ۴۵ دقیقه

آکادمی داده
dataacademy.ir

در سانتریفیوژ نگهداری و خُنَک شد . برای کلی سازی، داده سطح ابتدایی داده با مفاهیم سطح بالا جایگزین شد؛ برای مثال تحصیلات به ۳ سطح کمتر از ۵ سال ، ۵-۱۲ سال و بیشتر از ۱۲ سال تقسیم بندی شد. مجموعه داده نهایی به ۶۰ متغیر ورودی (پیش بینی کننده ها) و یک متغیر نتیجه که به ۲ دسته (رخداد دیابت نوع ۲ و عدم رخداد دیابت نوع ۲) تعریف شده است، تقسیم شد. برای ساخت مدل پیش بینی از ابزار Konstanz Information Miner استفاده شد. ٪۷۰ از افراد برای یادگیری ٪۳۰ از نمونه ها و برای آزمایش انتخاب شدند. از درخت CART و Gini Index برای معیار انتخاب ویژگی استفاده شد. درخت تصمیم با مجموعه یادگیری شامل ۴۶۵۲ رکورد ساخته شد و با ۱۹۹۵ رکورد، آزمایش شد. برای اولین متغیر جدا سازی، سطح قند خون ناشتا انتخاب شد. نتایج ارزیابی مدل پیش بینی برابر با Accuracy ٪۹۰.۵۰، sensitivity 31. 1%و specificity 97. 9% بدست آمد. تمرکز این مطالعه روی درختان تصمیم جهت پیش بینی احتمالات رخ ندادن دیابت نوع ۲ روی افراد غیر دیابتی است. درخت تصمیم در این مطالعه صحت پایینی در پیش بینی افراد دیابتی داشت. تنها در۷۲۹ نفر(٪۱۱) بیماری دیابت ، شناسایی شد.

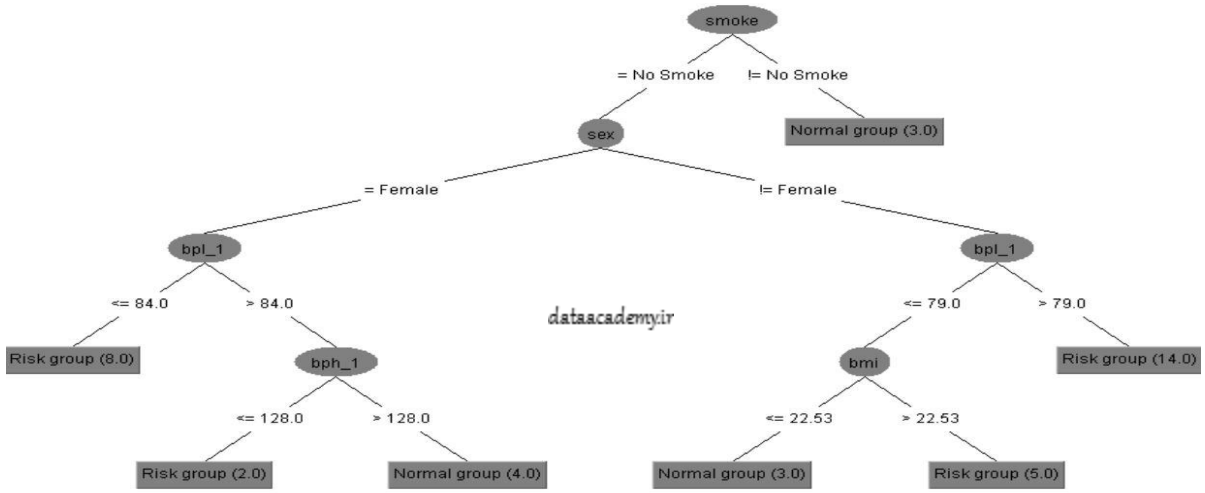
۶.۵. مطالعه ششم

در سال ۲۰۱۵ آقای Nongyao.n و همکارانش از طبقه بند های رگرسیون، درخت تصمیم، بیزین ساده و شبکه عصبی جهت طبقه بندی ریسک بیماری دیابت استفاده کردند و همچنین از تکنیک های Bagging and Boosting برای توانمندی طبقه بندها استفاده شد. هدف این مطالعه پیش بینی بیماری دیابت برای هر فرد بدون نیاز به آزمایش خون و رفتن به بیمارستان بوده است. مجموعه داده اولیه از ۲۶ واحد مراقبت اصلی (PCU) دربیمارستان SPR تایلند در بین سالهای ۲۰۱۲- ۲۰۱۳ جمع آوری شده است. هر فرد فرم گزینشی را پر می کند که برای شناسایی گروه ریسک دیابت در این مطالعه استفاده شده است. متغیر های ورودی خروجی به وسیله مشاوره با شخص پزشک تعیین می شوند و سپس از اطلاعات عمومی در فرم گزینش انتخاب می شوند. دراین مطالعه ۱۱متغیر ورودی و یک متغیر خروجی دو بخشی در نظر گرفته شده است. متغیر های ورودی به ترتیب جنسیت، BMI، قد، دورکمر، وزن، فشارخون انقباضی[21]، فشار خون انبساطی ، سابقه فشار خون در خانواده ، سابقه دیابت در خانواده، مصرف الکل، سیگار کشیدن و متغیر خروجی متغیر قند خون ناشتا است که تقسیم می شود به دو گروه معمولی و ریسکی. گروه معمولی افرادی هستند که مقدار قند خون ناشتای آنها کمتر از ۱۰۰ میلی گرم در دسی لیتر است و گروه ریسکی بین ۱۰۰ تا ۱۲۵ می باشد. در این تحقیق افراد بالای ۱۲۵ لحاظ نشده اند چرا که آنها در گروه دیابت طبقه بندی شده و تحت درمان می باشند. پس از حذف رکوردهای دارای مقادیر مفقودی از مجموعه داده ، مجموعه داده نهایی شامل ۳۰۱۲۲ نفر شد که به دو گروه معمولی ۱۹۱۴۵ نفر و ریسکی ۱۰۹۷۷ نفر تقسیم شد. باترکیب الگوریتم های Bagging ,Boosting, Random Forest و ۴ طبقه بند ذکر شده ۱۳ مدل طبقه بندی برای پیش بینی ساخته شد. الگوریتم Bagging با تولید تصادفی یک مجموعه داده یادگیری شروع می کند. سپس جایی که یک کلاس پیش بینی ایجاد شده است را مدل می کند. این فرایند چندین بار تکرار می شود تا هر دفعه عمل جابه جایی داده ها صورت گیرد. پیش بینی خروجی نهایی به وسیله رای اکثریت مدل های پیش بینی شده ارائه می شود. الگوریتم Boosting با به کارگیری یک وزن به تمام مشاهدات مجموعه داده در حال یادگیری شروع می شود. سپس طبقه بندش (مثلا درخت تصمیم) مدلسازی می شود.

[21] Systolic Blood Pressure

آکادمی داده
dataacademy.ir

سپس این گام ها چندین بار تکرار می شود و کلاس های پیش بینیشان ترکیب می شود. از اینرو آخرین خروجی پیش بینی از رای اکثریت ترکیبات می باشد. الگوریتم Random Forest از الگوریتم Bagging مدلسازی شده است. این روش بصورت بالقوه ای می تواند صحت طبقه بندی را بهبود بخشد و همچنین روی مجموعه داده هایی با تعداد بسیار زیادی از متغیر های ورودی بخوبی کار می کند. الگوریتم با ساختن یک ترکیبی از درختانی که هر کدام برای یک کلاس رای داده اند شروع می شود. در این مطالعه درخت تصمیم بر پایه Information Gain ساخته شد. سپس مدل با 10-Fold cross validation آزمایش شد. بیشترین Accuracy به ترتیب برای مدل های Random Forest و درخت تصمیم با مقدار 85.55٪ و85.09% و کمترین مقدار مربوط به طبقه بند ترکیبی Bagging و بیزین ساده با 80.96٪ بوده است. در شکل(۲) قسمتی از درخت تصمیم ساخته شده در این مطالعه نمایش داده شده است. عدد داخل پرانتز در برگ های درخت احتمال تعلق گرفتن نمونه به کلاس پیش بینی مورد نظر است



شکل(۲): مثالی از درخت تصمیم [8]

۷.۵. مطالعه هفتم

درسال ۲۰۱۵ آقای Mahmoud . h و همکارانش از طبقه بند های درخت تصمیم، بیزین ساده، SVM،Knn[۲۲] و شبکه عصبی برای تشخیص بیماری دیابت نوع ۲ در شهرتبریز استفاده کردند. هدف این تحقیق کمینه کردن هزینه اقتصادی سربار از خطای طبقه بندی دیابت بوسیله تکنیک ها و الگوریتم های مختلف در سیستم سلامت ، با در نظر گرفتن هزینه های غلط طبقه بندی افراد سالم به عنوان بیمار و بالعکس بوسیله پارامتر های نرخ TP وFP بوده است. مجموعه داده دارای ۲۵۳۶ رکورد که برای این بیماری آزمایش شده اند ، می باشد. داده این بیماران بوسیله دانشگاه علوم پزشکی تبریز طی ۳ ماه اول سال ۲۰۱۰ جمع آوری شد. تعداد دیابتی ها 2305 نفر و غیر دیابتی ها ۲۳۱ است. بنابراین مجموعه داده نامتعادل است. ۱۳ متغیر یا ویژگی در مجموعه داده با مشورت پزشکان و متخصصان انتخاب شد که عبارتند ازجنسیت، سن، وزن، قد، BMI، سابقه خانوادگی دیابت، سابقه بارداری، سابقه دیابت بارداری، سابقه سقط جنین، سابقه فشار خون بالا، سابقه مصرف دارو برای

[22] K-Nearest Neighbor

آکادمی داده
dataacademy.ir

فشار خون بالا، میزان فشار خون انبساطی و میزان فشار خون انقباضی. از Information Gain برای ساخت درخت تصمیم و برای پیاده سازی و آزمایش تکنیک های طبقه بندی از نرم افزار WEKA ورژن ۳. ۶ استفاده شد. برای تحلیل کارایی هر طبقه بند از10-fold cross validation استفاده شد. در کاربردهای پزشکی میانگین هزینه طبقه بندی اشتباه FN بیشتر از FP است و مدل طبقه بندی مناسب تر است که کمترین مقدار FN را داشته باشد. زیرا پیش بینی منفی اشتباه برای فرد بیمار عواقب خطرناکی را برای وی در پی دارد. از آنجائیکه مقدارSensitivity مهم تر از Specificity یا FPR[۲۳] است ما نیاز به کمینه کردن مقدار FPR به کمترین مقدار قابل قبول برای TPR[۲۴] داریم. 5nn و درخت تصمیم به ترتیب بیشترین Sensitivity را با مقدار 99.87٪ و 97.85٪ داشتند. شبکه عصبی و درخت تصمیم کمترین FP را با Specificity به ترتیب 96.8٪ و 66.97٪ دارا بودند و 5nn و درخت تصمیم دارای کمترین FN بودند. بیشترین Accuracy به ترتیب به شبکه عصبی و درخت تصمیم با ٪97.44 و 95.03٪ اختصاص یافت.

8.5. جدول مقایسه

در جدول(۲) مقایسه تحلیلی ۷ مقاله بررسی شده از نظر تکنیک شکاف داده، نوع درخت تصمیم ، تعداد نمونه و ویژگی ها و همچنین ارزیابی آنها نمایش داده شده است.

جدول(۲): مقایسه تحلیلی طبقه بندی های درخت تصمیم در مقالات مطرح شده

مطالعه	سال	تکنیک شکاف داده	درخت تصمیم	تعداد نمونه	تعداد متغیر (ویژگی ها)	ابزار داده کاوی	Specificity	Sensitivity	Accuracy
۱	۲۰۰۹	Information Gain	ID3	723	Input:6 Output:1	Rapid Miner	93. 05%	79. 83%	88. 50%
۲	2011	Gain Ratio	C4. 5:J48	724	Input:6 Output:1	WEKA	80. 50%	72. 33%	78. 18%
۳	2013	Information Gain	C5. 0	1487	Input:19 Output:1	Spss modeler ver14. 1	75. 13%	80. 68%	77. 87%
۴	2014	Information Gain Ratio, Gini Index	Ensemble Bagging: CART C4. 5 , ID3	_	_	Rapid Miner ver. 5	68. 33%	95. 63%	91. 56%
۵	2014	Gini Index	CART	6647	Input:60 Output:1	KNIME Ver.2. 6	97. 90%	31. 10%	90. 50%
۶	2015	Information Gain	_	۳۰۱۲۲	Input:11 Output:1	_	_	_	85. 09%
۷	2015	Information Gain	_	۲۵۳۶	Input:13 Output:1	WEKA ver3. 6	66. 97%	97. 08%	95. 03%

[23] False Positive Rate
[24] True Positive Rate

آکادمی داده
dataacademy.ir

۶. نتیجه گیری

در این مقاله ابتدا به تعاریف مربوط بیماری دیابت ، انواع آن و آمار ها اخیر پرداخته شد. سپس در مورد داده کاوی در حوزه پزشکی ، انواع درختان تصمیم ، روشهای القای آنها و معایب و مزایای آنها توضیحاتی داده شد. به مقایسه ۷ مورد از کارهای اخیر در زمینه پیش بینی دیابت بوسیله طبقه بند درخت تصمیم پرداخته شد. در مطالعه اول تعداد رکوردها و ویژگی های کمی بررسی شده است ولی بخاطر پیش پردازش مناسبی که روی داده ها انجام شده، صحت مناسبی از مدل را شاهد بودیم. در مطالعه دوم تعداد نمونه ها یکی بیشتر از مطالعه اول بوده و تعداد ویژگی ها برابر مطالعه اول بوده است ولی با تغییر تکنیک شکاف داده حدود ۱۰٪ کاهش صحت طبقه بند را شاهد بودیم.یکی از دلایل آن می تواند خصوصیت Gain Ratio باشد که در مجموعه داده های کوچک صحت کمتری نسبت به مجموعه داده ها و مسائل بزرگتر دارد. در مطالعه سوم به همراه افزایش تعداد نمونه ها نسبت به دو مطالعه قبلی تعداد فاکتور های ریسک نیز افزایش پیدا کردند. بعضی از این فاکتور ها مثل سطح تحصیل و نوشیدن قهوه ارتباطی با بیماری دیابت نداشتند . یکی از دلایل کاهش صحت طبقه بندی نسبت به ۲ کار قبلی می تواند افزایش تعداد فاکتور های نامربوط به بیماری در مدل پیش گیری باشد. در مطالعه چهارم از طبقه بند مجموعه ترکیب استفاده شده است که یک طبقه بند کار آمد به حساب می آید و صحت خوبی را بدست آورده است. در این مطالعه ازتعداد نمونه ها و ویژگی ها صحبتی نشده است و از این جنبه نمی توانیم اظهار نظری داشته باشیم. در مطالعه پنجم نمونه های مجموعه داده افراد غیر دیابتی بوده اند و هدف پیش بینی دیابت برای افرادی بوده است که ریسک پایینی در ابتلا به این بیماری دارند. بنابراین تعداد افراد غیر دیابتی بالا باعث افزایش مقدار TN و کاهش مقدار TP می شود. مقدار پایین TPنسبت به FN تاثیر مستقیمی روی کاهش Sensitivity می گذارد و میزان آن با مقدار ۳۱.۱٪ ثبت می شود. افزایش تعداد متغیرها نسبت به مقالات قبلی تأثیری در افزایش صحت نداشته است ولی در مورد افزایش تعداد نمونه ها نمی توانیم این ادعا را داشته باشیم. در مطالعه ششم ۱۳ طبقه بند مختلف مورد آزمایش قرار گرفت که درخت تصمیم با اختلاف بسیار ناچیز دومین صحت بالای این مطالعه را بدست آورد. از مطالعه ششم نتیجه می گیریم که طبقه بند های ترکیبی تاثیر مثبتی در بهبود نتایج طبقه بندی و پیش گیری الگوریتم های طبقه بندی مجزا دارند. مطالعه هفتم با اختلاف ۲٪ و میزان ۹۵.۰۳٪ دومین صحت بالای طبقه بندی را درمقایسه با شبکه عصبی در مطالعه مربوطه بدست آورد. این عدد بیشترین مقدار در مقایسه تحلیلی ما در این مقاله می باشد. تفاوت ویژگی ها و فاکتور های ریسک در مقاله هفتم با بقیه کارهای بررسی شده استفاده از ویژگی هایی مربوط به دیابت بارداری می باشد. در نهایت نتیجه می گیریم که افزایش تعداد ویژگی های مورد بررسی تأثیر زیادی در افزایش صحت نداشته، افزایش تعداد نمونه ها ی مورد بررسی درکل در بهبود صحت پیش گیری موثر است و بهترین روش شکاف داده با توجه به ارزیابی کلی Specificity Sensitivity, Accuracy, در این مطالعه به Information Gain اختصاص می یابد.

dataacademy.ir

آکادمی داده
dataacademy.ir

مراجع:

1. D. A. G. Idf. (2015),"Update of mortality attributable to diabetes for the IDF Diabetes Atlas: Estimates for the year 2013," *Diabetes research and clinical practice,* vol. 109, p. 461.
2. M H. Nadimi ,M.Taki and F.Habibollahi.(2014), *Data Mining Concepts And Applications* vol. 1.
3. T. N. Phyu, "Survey of classification techniques in data mining.(2009)," in *Proceedings of the International MultiConference of Engineers and Computer Scientists,* pp. 18-20.
4. X. -H. Meng, Y. -X. Huang, D. -P. Rao, Q. Zhang, and Q. Liu.(2013),"Comparison of 3 data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung journal of medical sciences,* vol. 29, pp. 93-99.
5. R. M. Goodman and P. Smyth.(1990),"Decision tree design using information theory," *Knowledge Acquisition,* vol. 2, pp. 1-19.
6. F. Aguiree, A. Brown, N. H. Cho, G. Dahlquist, S. Dodd, T. Dunning, *et al.* .(2013),"IDF diabetes atlas,".
7. R. W. Grant and M. S. Kirkman.(2015),"Trends in the Evidence Level for the American Diabetes Association's “Standards of Medical Care in Diabetes” From 2005 to 2014," *Diabetes care,* vol. 38, pp. 6-8.
8. N. Nai-arun and R. Moungmai.(2015),"Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Computer Science,* vol. 69, pp. 132-14.
9. C. f. D. Control, Prevention, C. f. D. Control, and Prevention.(2011),"National diabetes fact sheet on diabetes and prediabetes in the United States," *Atlanta, GA: US, Centers for Disease Control and Prevention,* vol. 201.
10. A. D. Association, "National Diabetes Statistics Report.(2014)," *Estimates of diabetes and its burden in the epidemiologic estimation methods. Natl Diabetes Stat Rep,* pp. 2009-2012.
11. A. A. Al Jarullah, (2011)"Decision tree discovery for the diagnosis of type II diabetes. " (IIT), International Conference on , pp. 303-307.
12. J. Han, J. C. Rodriguez, and M. Beheshti, "Discovering decision tree based diabetes prediction model.(2009)" in *Advances in Software Engineering,* ed: Springer, pp. 99-109.
13. M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia,(2015)"Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran," *International Journal of Diabetes in Developing Countries,* pp. 1-7.
14. S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "An Efficient Rule-Based Classification of Diabetes Using ID3, C4. 5, & CART Ensembles.(2014)," *(FIT),12th International Conference on,* pp. 226-231.
15. I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. -F. Chang, *et al.* , "Data mining in healthcare and biomedicine: a survey of the literature.(2012)" *Journal of medical systems,* vol. 36, pp. 2431-2448.
16. A. Ramezankhani, O. Pournik, J. Shahrabi, D. Khalili, F. Azizi, and F. Hadaegh, "Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study.(2014)" *Diabetes research and clinical practice,* vol. 105, pp. 391-398.
17. D. Delen, G. Walker, and A. Kadam,(2005)"Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial intelligence in medicine,* vol. 34, pp. 113-127.
18. R. L. Plackett, "Karl Pearson and the chi-squared test,(1983)" *International Statistical Review/Revue Internationale de Statistique,* pp. 59-72.

آکادمی داده
dataacademy.ir