مجله دیابت و متابولیسم ایران. مرداد و شهریور ۱۴۰۰؛ دوره ۲۱ (شماره ۳)۱۶۴-۱۵۱

مقاله پژوهشی

ارایه‌ی یک مدل پیش‌بینی جهت شناسایی افراد مبتلا به دیابت با استفاده از درخت تصمیم

ابوالفضل کاظمی[۱]، حمید بهادر[۲*]

چکیده

مقدمه: امروزه در اکثر بیمارستان‌های ایران بانک اطلاعاتی وسیعی از ویژگی‌های بیماران موجود است که حجم بالایی از اطلاعات مربوط به سوابق بیماری، خانوادگی و پزشکی را شامل می‌شود. پیدا کردن الگوی دانش این اطلاعات می‌تواند در جهت پیش‌بینی عملکرد نظام پزشکی و بهبود فرآیندهای آموزشی کمک شایانی کند.

روش‌ها: تکنیک‌های داده‌کاوی ابزار تحلیلی هستند که برای استخراج دانش معنادار از مجموعه داده‌های بزرگ مورد استفاده قرار می‌گیرند. در این تحقیق از اطلاعات ۵۰۰ نفر از مراجعه‌کنندگان به مرکز بهداشت شهید بلندیان قزوین استفاده شده است. در این تحقیق با استفاده از روش‌های داده‌کاوی درخت تصمیم و شبکه عصبی و شبکه‌ی بیزین یک مدل پیش‌بینی شده انجام شده است.

یافته‌ها: مدل درخت تصمیم بیش‌ترین دقت و شبکه‌ی بیزین کم‌ترین دقت را در تشخیص بیماران دیابت دارد و به تبع آن درخت تصمیم کم‌ترین خطا و شبکه‌ی بیزین بیشترین خطا را دارا هست. مدل درخت تصمیم با ۹۵/۶۸ درصد بیشترین دقت را در پیش‌بینی داشته است.

نتیجه‌گیری: چربی بیشترین تأثیر را در پیش‌بینی بیماری دیابت و جنسیت کمترین تأثیر را در پیش‌بینی بیماری دیابت دارا هست. بر اساس تحلیل درخت تصمیم قوانین به‌دست آمده در بین ویژگی‌های بیان شده متغیرهای سن و میزان قند بیشترین تأثیر را در پیش‌بینی وقوع بیماری دیابت (طبق تحلیل نرم‌افزار) را دارا هستند و با ایجاد رژیم غذایی مناسب می‌توان از ابتلا به این بیماری جلوگیری کرد.

واژگان کلیدی: داده‌کاوی، دیابت، درخت تصمیم، پیش‌بینی، شبکه‌ی عصبی

۱- دانشکده‌ی صنایع، دانشگاه آزاد اسلامی واحد قزوین، قزوین، ایران

۲- گروه کامپیوتر، دانشگاه فنی و حرفه‌ای و پردیس مطهری فرهنگیان استان آذربایجان غربی، ارومیه، ایران

***نشانی:** آذربایجان عربی، ارومیه، خیابان امام، تقاطع خیام شمالی، اداره کل آموزش و پرورش، اداره فناوری اطلاعات و ارتباطات : تلفن: ۰۴۴۳۱۹۳۲۳۲۷، نمابر : ۳۲۲۲۲۰۴۶ ۰۴۴ ، کدپستی ۱۵۷۶۵- ۵۷۱۳۷ ، پست الکترونیک:hamidbahador52@gmail.com

تاریخ دریافت: ۱۴۰۰/۰۴/۱۶ **تاریخ درخواست اصلاح: ۱۴۰۰/۰۵/۰۵** **تاریخ پذیرش: ۱۴۰۰/۰۶/۰۷**

مقدمه

دیابت یک بیماری مزمن است که روی چگونگی تولید و استفاده انسولین در بدن اثر می‌گذارد. اگر سلول‌های بدن در مقابل انسولین مقاومت نشان دهند و یا اگر بدن به اندازه کافی انسولین نسازد در این صورت بدن درست کار نخواهد کرد. اشخاصی که به بیماری دیابت مبتلا هستند معمولاً آثار فیزیکی مشاهده نخواهند کرد مگر اینکه قند خونشان به بیش از دو برابر میزان نرمال برسد. در این حالت حتی برای کسانی که دیابت ندارند علایم بیماری دیابت کم و بیش ظاهر می‌شود [۱].

مسئله‌ی چشم‌گیری که در حوزه‌ی پزشکی وجود دارد آن است که هنگامی که تعداد پارامتر مورد بررسی جهت تشخیص یک بیماری زیاد باشد، پُر واضح است که تشخیص این بیماران حتی برای یک مفرد متخصص پزشکی هم مشکل می‌شود. از طرفی به‌طور معمول آشکار نمودن این که کدام‌یک از پارامترهای تأثیرگذار با یکدیگر ارتباط دارند و بر هم تأثیر می‌گذارند نیز امری ناممکن است. بنابر این از اهدافی که ما در این تحقیق به‌دنبال آن هستیم کشف ارتباط میان فاکتورهای تأثیرگذار بر یک بیماری است [۲].

در این پژوهش سعی شده است که بیماری دیابت مورد نقد، بررسی و پیش‌بینی قرار گیرد. ذکر این نکته ضروری است که ٤ نوع دیابت وجود دارد. دیابت نوع یک که ۱۰ تا ۱۵ درصد کل موارد دیابت را تشکیل می‌دهد که اغلب در سنین زیر ۳۰ سال به‌وجود می‌آید و از این‌رو به آن دیابت جوانی نیز می‌گویند. دیابت نوع دو بیش‌تر در بالغین بالای ۳۰ سال و چاق دیده می‌شود که ۸۵ تا ۹۰ درصد کل موارد دیابت را شامل می‌گردد به این نوع دیابت، دیابت غیر وابسته به انسولین یا دیابت بزرگسالان نیز می‌گویند. نوع سوم دیابت حاملگی است که برای اولین بار در طول حاملگی تشخیص داده می‌شود. این نوع دیابت معمولاً گذراست و نوع ٤ دیابت از علل متفرقه مانند جراحی، داروها، سوءتغذیه و عفونت شامل می‌شود.

لذا تلاش گردیده با استفاده از شاخص‌ها و مشخصات فردی انسان‌ها مانند سن، جنسیت و همچنین مشخصات بیولوژیکی مانند وزن سطح تری‌گلیسیرید[1]، HDL، BMI ، قند[2]، هر فرد با استفاده از الگوریتم‌های داده‌کاوی مانند الگوریتم درخت تصمیم به ارایه‌ی یک مدل و الگو برای شناسایی افراد مبتلا به بیماری دیابت (از نوع یک و دو) پرداخته شود در این تحقیق سعی شده است با استفاده از روش طبقه‌بندی الگو ایجاد گردد به‌طوری‌که در ابتدا با استفاده از شاخص جینی ریشه و سایر اعضا مشخص می‌شود و درخت ترسیم می‌گردد. برای بررسی این درخت و میزان درستی و صحت آن با ۱۵۰دیتاست آزمایش قرار داده شده است تا مورد داوری قرار بگیرد. و پیاده‌سازی و تحلیل تمام دیتاهای توسط نرم‌افزار کلمنتاین مورد آنالیز جامع قرار گرفته است.

در سال ۱۹۸۹ و ۱۹۹۱ کارگاه‌های کشف دانش و معرفت از پایگاه داده‌ها توسط Fayyad و همکاران برگزار شد. در واقع داده‌کاوی فرآیندی است که در آغاز دهه‌ی ۹۰ پا به عرصه‌ی ظهور گذاشته و با نگرشی نو، به مسأله‌ی استخراج اطلاعات از پایگاه داده‌ها پرداخت. در واقع پژوهش جدی روی موضوع داده‌کاوی از اوایل دهه‌ی ۹۰ شروع شد. پژوهش‌ها و مطالعه‌های زیادی در این زمینه صورت گرفت، همچنین سمینارها، دوره‌های آموزشی و کنفرانس‌هایی نیز برگزار شد و پایه‌های نظری داده‌کاوی در تعدادی از مقاله‌های پژوهشی آورده شد. در فواصل سال‌های ۱۹۹۱ تا ۱۹۹٤ نیز کارگاه‌های کشف دانش و معرفت از پایگاه داده‌ها مجدد توسط Fayyad و دیگران برگزار شد. از سال ۱۹۹۵ داده‌کاوی به‌صورت جدی وارد مباحث آمار شد [۳]. کشف دانش به‌طور رسمی اولین بار توسط Fayyad در اولین کنفرانس بین‌المللی داده‌کاوی و کشف دانش که در سال ۱۹۹۵ در مونترال برگزار شده بود، معرفی شد که به بیان ارتباط تکنیک‌های آنالیز در چندین مرحله با هدف استخراج دانش‌های ناشناخته‌ی قبلی از داده‌های در دسترس می‌پرداخت[۳].

جمعی از اندیشمندان سال ۱۹۹۵ با استفاده از داده‌کاوی، انباره‌های داده‌ی بانک‌های آمریکا را بررسی کرده و بیان کردند که چگونه این سیستم‌ها برای بانک‌های آمریکا قدرت رقابت بیشتری ایجاد می‌کنند. در این سال انجمن داده‌کاوی هم‌زمان با

[1] Triglyceride level

[2] Suger

اولین کنفرانس بین‌المللی «کشف دانش و داده‌کاوی» شروع به کار و یک سازمان علمی به نام ACM-SIGKDD را تأسیس کرد. در سال ۱۹۹۶ اولین شماره‌ی مجله‌ی «کشف دانش از پایگاه داده‌ها» منتشر شد. در همان سال دیدگاهی از داده‌کاوی به‌عنوان «پرس و جو کننده از پایگاه‌های استنتاجی» پیشنهاد شد و Fayyad و Piatetsky پیشرفت‌های کشف دانش و داده‌کاوی را اعلام کردند [۳].

در تحقیقی که توسط عامری و همکاران در سال ۱۳۹۲ با عنوان "استخراج دانش از داده‌های بیماران دیابتی با استفاده از روش درخت تصمیم C5" انجام شد آنها در این تحقیق، برای اولین بار احتمال بروز عوارض میکروواسکولار، ماکروواسکولار و یا هر دو نوع عارضه را در بیماران دیابتی و ویژگی‌های تأثیرگذار بر آنها را مورد بررسی قرار دادند. متغیرهای فشارخون بالا، سن و سابقه‌ی خانوادگی در عوارض مشاهده شده بیشترین تأثیر را داشته‌اند. به کمک درخت تصمیم ایجاد شده، قوانینی استخراج شده‌اند که می‌تواند به‌عنوان الگویی برای پیش‌بینی وضعیت بیماران و احتمال بروز عوارض در آنها استفاده شود. میزان دقت تشخیص بیماران دیابتی با استفاده از درخت تصمیم C5 عدد ۸۹/۷٤ درصد به دست آمده است. آنها در این مقاله نتیجه گرفتند که معایب روش درخت تصمیم نوع هرس کردن است که درخت هزینه‌ی بالایی دارد و در مواردی با تعداد دسته‌های زیاد و نمونه‌های آموزشی کم احتمال خطا بالاست [٤].

Song و همکاران (۲۰۱۷) با استفاده از الگوریتم‌های داده‌کاوی پیش‌بینی از طبقه‌بندی نزدیکترین همسایگی و قواین انجمنی استفاده کردند. در این پژوهش دو الگوریتم ارائه گردید که با استفاده از الگوریتم‌های طبقه‌بندی برای استخراج قوانین طبقه‌بندی انجمنی مورد استفاده قرار گرفت. مدل طبقه‌بندی به‌دست آمده ترکیبی از مزایای طبقه‌بندی انجمنی و درخت تصمیم‌گیری است [۵]. الگوریتم درخت تصمیم در میان الگوریتم‌های موجود از دقت بی‌نظیری برخوردار است. این الگوریتم می‌تواند به‌طور مؤثر بر روی پایگاه داده‌های بزرگ اجرا شود و می‌تواند با هزاران متغیر ورودی بدون حذف متغیر سروکار داشته باشد [۵]. Breault و همکاران [٦] با استفاده از سیستم CART طبقه‌بندی و تجزیه و تحلیل رگرسیون را در سال ۲۰۰۲ انجام داده و وابستگی بین یک سری از ویژگی‌های آنها را استنباط کرده‌اند. میزان دقت طبقه‌بندی ۵۹/۹ درصد بوده است. همچنین Miyaki و همکاران [۷] از روش کارت برای قضاوت عوامل مؤثر بر بروز عوارض دیابت در سال ۲۰۰۲ استفاده کرده‌اند.

داده‌کاوی به‌عنوان تکنیکی برای شناسایی و تشخیص بیماری‌ها ودسته‌بندی بیماران در مدیریت بیماری و پیدا کردن الگوهایی برای تشخیص سریع‌تر بیماران و جلوگیری از بروز عوارض در آنها می‌تواند کمک بسیار بزرگی داشته باشد. افزایش دقت تشخیص، کاهش هزینه‌ها و کاهش منابع انسانی به‌عنوان مزایای معرفی داده‌کاوی در تجزیه و تحلیل پزشکی توسط Al Jarullah در سال ۲۰۰۲ ثابت شده است [۸].

Rohlfing و همکاران [۹] از روش تجزیه و تحلیل رگرسیون خطی برای بررسی ارتباط بین قند خون در دیابت نوع یک ودر سال ۲۰۰۲ استفاده کرده‌اند [۹].

Silverstein و همکاران [۱۰] آزمایش‌هایی را بر روی سه پایگاه داده‌ی پزشکی در زمینه‌ی دیابت انجام داده‌اند و قوانینی برای تشخیص دیابت تولید کرده‌اند و سپس این قوانین را با قوانین از پیش تعیین شده مقایسه کرده‌اند.

Quentin-Trautvetter و همکاران [۱۱] از روش قواعد انجمنی و درخت تصمیم برای استخراج دانش از پایگاه داده پزشکی استفاده کرده‌اند. Juan و همکاران[۱۲] در سال ۲۰۰۷ با استفاده از ترکیب الگوریتم‌های C4.5 و EM (حداکثرانتظار) سیستم پردازش داده‌های دیابت نوع دو را ایجاد کرده‌اند. Huang و همکاران [۱۳] در سال ۲۰۰۷ تحقیقی بر روی شناسایی عوامل عمده‌ی تأثیرگذار بر کنترل دیابت، با به‌کار بستن انتخاب ویژگی ها در سیستم (Feature Selection) مدیریت بیمار انجام دادند.

Tang و همکاران (۲۰۲۰) شناسایی ویژگی‌های بیوشیمیایی و عوامل مؤثر بر این بیماری، ممکن است فرصت‌هایی را جهت مداخله به موقع فراهم آورد. یکی از اصلی‌ترین مشخصه‌های بیوشیمیایی دیابت، هایپرگلایسمی و اختلال در متابولیسم است که در اثر نقص در ترشح انسولین و یا مقاومت نسبت به انسولین به‌وجود می‌آید [١٤].

مشکوتی و همکاران (۱۳۹٤) در مقاله‌ی "تشخیص بیماری دیابت با استفاده از ماشین بردار پشتیبان"به این نتیجه رسیدند که در حالت طبیعی، غذا در معده تبدیل به گلوکز یا قندخون می شود. قند از طریق معده وارد جریان خون شده، سپس لوزالمعده (پانکراس) هورمون انسولین را ترشح می‌کند و این هورمون باعث می‌شود قند از جریان خون وارد سلول‌های بدن گردد، در نتیجه قندخون در حد نرمال و متعادل باقی می‌ماند [۱۵].

بیماری دیابت شامل ٤ نوع است، با بررسی منابع و مقالات منتشر شده مشخص گردید الگوهای مطرح شده مربوط به دیابت نوع دو است، حال در این تحقیق سعی شده علاوه بر نوع ۲ به نوع ۱ هم نگاه ویژه‌ای گردد و با نرم‌افزار کلمنتاین بر روی داده‌ها تحلیل جامعی صورت گیرد.

روش‌ها

داده‌های به‌کار رفته در این مقاله از بیماران و مراجعه کنندگان به مرکز بهداشت شهید بلندیان قزوین به‌دست آمده است که این داده‌ها از اطلاعات ۵۰۰ بیمار و مراجعه کننده به این مرکز حاصل شده است که از این میان ۳۵۰ داده اصلی و ۱۵۰ داده برای تست مدل لحاظ شده است شایان ذکر است کلیه‌ی این داده‌ها متعلق به دوره‌ی زمانی ٦ ماه اول سال ۹۷ است. در این مدل پیشنهادی مراحل مختلف یک فرآیند داده‌کاوی از جمله جمع‌آوری داده‌ها، آماده‌سازی و پیش پردازش روی کل پایگاه داده‌ی ذکر شده انجام گردیده است، با استفاده از داده‌های بیماران و تعریف مشخصه‌های مربوط به بیماران مدل فوق ساخته می‌شود و با تحلیل نتایج حاصل از آن می‌توان پیش‌بینی کرد که افراد مختلف به چه میزان به این بیماری دچار شده‌اند و یا در مرحله‌ی پیش ابتلا و یا احیاناً سالم هستند. داده‌های پایگاه اطلاعاتی در داخل یک فایل جمع آوری شده و برخی از آیتم‌ها کدگذاری شده است که برخی به خصیصه‌های عددی و برخی به خصیصه‌های اسمی معادل گشته‌اند.

معرفی ویژگی‌های مدل

اسامی فیلدها و نوع ویژگی آنها در جدول ۱ آمده است.

جدول ۱- اسامی و ویژگی‌های بیماران

نام اختصاری	نام فیلد	نوع ویژگی
age	سن	عددی
Weight	وزن	عددی
Sex	جنسیت	کیفی- اسمی
Family rec	سابقه خانوادگی	کیفی- اسمی
Married	تاهل	کیفی- اسمی
Sugar	قند خون	عددی
triglyceride	چربی	عددی
HDL	کلسترول مفید	عددی
BMI	نسبت وزن به توان ۲ قد	عددی

اما همان‌طور که ذکر شد برای عملکرد بهتر الگوریتم‌های داده کاوی پس از تجمیع داده‌ها داخل یک فایل باید خصوصیات عددی به خصوصیات گروهی معادل تبدیل شوند. به‌عنوان مثال جنسیت دانشجویان که شامل دو گروه زن و مرد است، کدگذاری شده و به هرگروه یک کد اختصاص داده شده است که کدهای مربوطه در جدول ۲ شرح داده شده است.

جدول ۲- کدگذاری ویژگی‌های بیماران

نام فیلد	کدگذاری	نوع ویژگی
جنسیت	مرد=۲ زن=۱	کیفی- اسمی
تاهل	طلاق=۱ متاهل=۲ مجرد=۳	کیفی – اسمی
سابقه خانوادگی	دارد=۱ ندارد=۲	کیفی- اسمی

پیاده‌سازی با درخت تصمیم

درخت تصمیم یکی از ابزارهای قوی و متداول برای دسته‌بندی و پیش‌بینی است. در ساختار درخت تصمیم پیش‌بینی به‌دست آمده از درخت در قالب یک سری قواعد توضیح داده می‌شود. ساختار درخت تصمیم یک ساختار درختی، شبیه فلوچارت است. در برخی موارد تنها صحت دسته‌بندی و پیش‌بینی مهم است.

با توجه به اینکه هدف این تحقیق تحلیل و پیش‌بینی بیماری دیابت با استفاده از بیماران مراجعه کننده به مرکز بهداشت شهید بلندیان است، پس قند خون متغیر هدف و یا همان برچسب انتخاب می‌شود. در گام آخر باید الگوریتم و تکنیک مورد نظر جهت به‌دست آوردن نتایج و پیش‌بینی انتخاب می‌گردد. مهم‌ترین و اساسی‌ترین الگوهایی که در درخت تصمیم استفاده می‌شود ID3 و C4.5(J48) که هر دو متعلق به ساختمان داده‌ای درخت هستند.

ID3 یک درخت تصمیم هرس نشده می‌سازد که فقط داده nominal قبول می‌کند و J48 یک درخت هرس نشده از C4.5 است. البته باید به این نکته اشاره نمود که الگوریتم‌های دیگر ایجاد درخت تصمیم مانند CART و CHAID نیز برای دسته بندی ساختار تقریباً مشابهی دارند و هدف همگی آنها به‌دست آوردن درختی با کیفیت بالا و نرخ خطای کم در دسته‌بندی داده‌ها است و بیشتر تفاوت‌ها در شیوه‌ی شاخه زدن و هرس شاخه‌هاست. همچنین الگوریتم CHAID داده‌های عددی را قبول نمی‌کند و داده‌ها باید به‌صورت nominal باشند.

در این تحقیق با توجه به اینکه برخی از داده‌ها nominal هستند و برخی Numeric از الگوریتم C4.5(J48) استفاده شده است. الگوریتم C4.5(J48) یکی از اساسی‌ترین و مهم‌ترین الگوریتم درخت تصمیم است.

استخراج و انتخاب قواعد

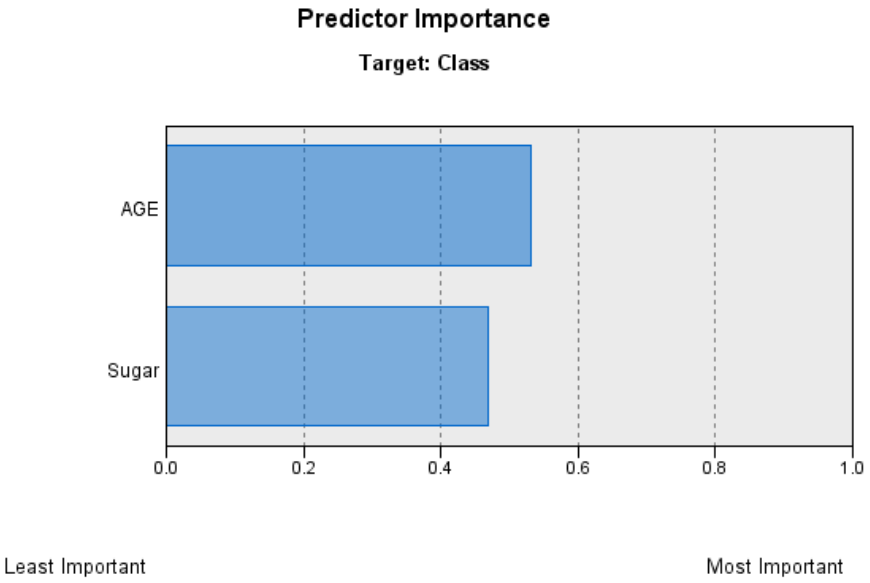
مشخص است که نتایج حاصل از درخت بسیار زیاد و گسترده است. لذا کلیه‌ی قواعدی که از درخت استخراج شده‌اند آنهایی مورد قبول و تحلیل قرار می‌گیرند که از احتمال بالایی نسبت به سایرین برخوردار باشند. باید به این نکته اشاره کرد که گاهی ممکن است قواعدی استخراج و انتخاب گردند که نباید آنها را منطقی دانست و ممکن است در تضاد با هم باشند که این امر بدیهی است. به‌طور کلی می‌توان گفت که تک تک درخت‌ها و قاعده‌ها را می‌توان بررسی کرد و این درخت‌ها را از زبان ریاضی به زبان قابل فهم تبدیل کرد. هر قاعده دارای احتمال وقوع خواهد بود. بر اساس تعداد رکورد منطبق با قاعده تعریف شده از بین رکوردهای آن شاخه به‌دست می‌آید.

در معادله‌ی ۱، p احتمال وقوع هر شاخه در بین سایر رکوردها، n_c تعداد رکوردهای منطبق با قاعده و n تعداد کل رکوردهای بررسی شده در شاخه است.

معادله‌ی (۱): $$P=\frac{n_c}{n}$$

از میان ۹ متغیر پیش‌بین شامل: سن، جنسیت، وضعیت تأهل، سابقه‌ی خانوادگی، شغل، وزن، چربی، قند و شاخص‌های BMI و HDL در اجرای مدل درخت تصمیم متغیرهای سن و میزان قند همان‌طور که در شکل ۱ مشاهده می‌کنید بیشترین تأثیر را در پیش‌بینی وقوع بیماری دیابت دارند.

میزان تأثیر هر کدام در جدول ۳ مشخص شده است.



شکل ۱- میزان تأثیرات قند خون و سن بر روی پیش‌بینی مدل

جدول ۳- تأثیرات قند خون و سن بر روی پیش‌بینی مدل

Nodes	Importance
Sugar	0.4684
AGE	0.5316

بحث

در اجرای مدل درخت تصمیم همان‌طور که مشاهده می‌شود متغیرهای سن و میزان قند به‌عنوان متغیرهای پیش‌بین انتخاب شده‌اند. همچنین طبق آنالیز، دقت پیش‌بینی در جدول ٤ برای داده‌های تمرین ۹۹/۷۱ درصد و برای داده‌های آزمون ۱۰۰ درصد پیش‌بینی درست است.

جدول ۴- تأثیرات حضور ویژگی قند خون بر روی مدل

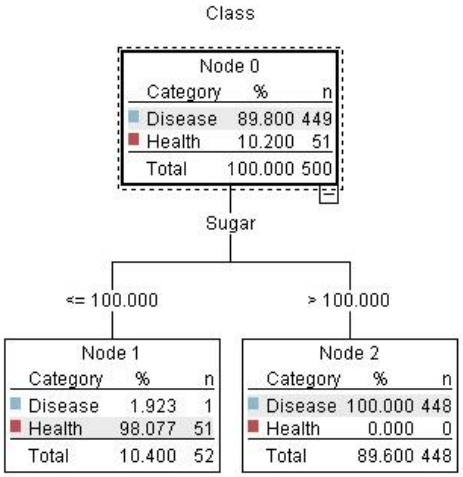
Results for output field Class

Comparing $C-Class with Class

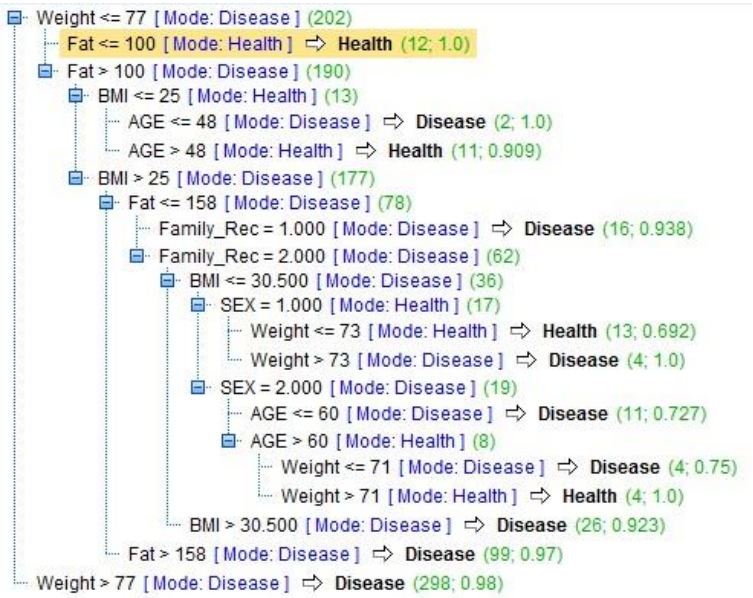
'Partition'	1_Training		2_Testing	
Correct	346	99.71%	153	100%
Wrong	1	0.29%	0	0%
Total	347		153	

خروجی آنالیز درخت نرم‌افزار کلمنتاین نیز در شکل ۲ آمده است. با توجه به شکل ۲ از آنجا که با بالاتر بودن قند از ۱۰۰ بیماری با ۱۰۰ درصد اتفاق می‌افتد در تحلیل بعدی متغیر قند از مدل حذف می‌شود. چرا که وجود این متغیر باعث نادیده گرفته شدن تأثیر سایر متغیرهای پیش‌بین شده است. با حذف قند به شکل ۳ برای پیش‌بین می‌رسیم، که چربی بیشترین تأثیر و جنسیت کمترین تأثیر را در مدل پیش‌بینی داشته‌اند. متغیرهای پیش‌بین در مدل به‌ترتیب چربی، BMI، وزن، سابقه‌ی خانوادگی، سن و جنسیت است.

درخت تصمیم نیز در شکل ۳ به ازای داده‌های بیماران مشاهده می‌شود.



شکل ۲- تأثیرات قند خون بر روی مدل درخت تصمیم



شکل ۳- مدل درخت تصمیم

همان‌طور که در جدول ۵ مشاهده می‌فرمائید دقت مدل فوق حدوداً ۹۶٪ و خطای مدل فوق حدوداً ٤٪ است. و همچنین چربی بیشترین تأثیر را در پیش‌بینی بیماری دیابت دارا است و جنسیت کمترین تأثیر را در پیش‌بینی بیماری دیابت دارا هست.

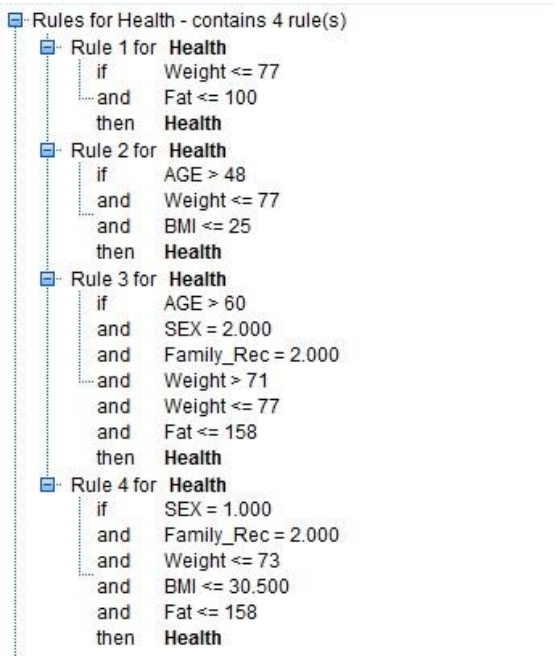
بر اساس تحلیل درخت تصمیم قوانین زیر برای مدل ایجاد می‌شود، برای پیش‌بینی سالم بودن شخص ٤ قانون به‌صورت شکل ٤ بیان می‌شود.

جدول ۵- دقت پیش‌بینی مدل درخت تصمیم

Results for output field Class

Comparing $C-Class with Class

'Partition'	1_Training		2_Testing	
Correct	332	95.68%	147	96.08%
Wrong	15	4.32%	6	3.92%
Total	347		153	

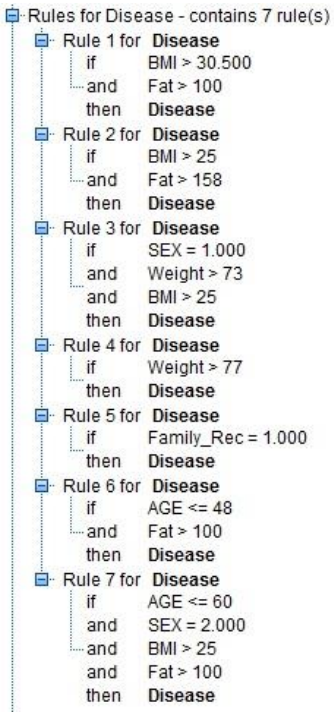


شکل ۴- رول‌های افراد سالم تحقیق

برای پیش‌بینی بیمار بودن شخص، ۷ قانون به‌صورت شکل ۵ بیان می‌شود.

قواعد انتخاب شده از درخت تصمیم

حال پس از آنکه مدل درخت تصمیم شکل گرفت و به دست آمد نمونه‌ای از رول‌های خروجی از مدل فوق به شرح جدول ٦ است.



شکل ۵- رول‌های افراد بیمار تحقیق

جدول ۶- رول‌های خروجی از مدل درخت تصمیم

Rule 1
If weight<=77 and Fat<=100 then health
اگر وزن شخص کوچکتر یا مساوی ۷۷ و چربی آن کوجکتر یا مساوی ۱۰۰ باشد آنگاه شخص سالم می‌باشد.

Rule 2
48 and Weight<=77 and BMI<=25 then health>If AGE
اگر سن بزرگتر از ٤٨ و وزن کوجکتر یا مساوی ۷۷ و BMI کوجکتر یا مساوی ۲۵ باشد آنگاه شخص سالم است.

Rule 3
If AGe>60 and SEX=2 and family- rec=2 and Weight>71 and weight<=77 and Fat<=158 then Health
اگر سن بزرگتر از ٦٠ باشد و مرد باشد و سابقه خانوادگی نداشته باشد و وزن بزرگتر از ۷۱ و کوچکتر یا مساوی ۷۷ باشد و چربی کوجکتر مساوی ۱۵۸ داشته باشد آنگاه سالم است.

Rule 4
If SEX=1 and Family-Rec=2 and Weight<=73 and BMI<=30.500 and Fat<=158 Health
اگر زن باشد و سابقه خانوادگی نداشته باشد و وزن کوچکتر مساوی ۷۳ باشد و BMI کوچکتر مساوی ۳۰،۵۰۰ باشد و چربی کوپکتر مساوی ۱۵۸ باشد آنگاه شخص سالم است.

Rule 5
If BMI>30.500 and Fat>100 then Disease
اگر BMI بزرگتر از ۳۰،۵۰۰ باشد و چربی بزرگتر از ۱۰۰ باشد انگاه شخص مریض است.

Rule6
If BMI>25 and Fat>158 then Disease
اگر BMI بزرگتر از ۲۵ باشد و چربی بزرگتر از ۱۵۸ باشد آنگاه شخص مریض است.

Rule7
If Sex=1 and Weight>73 and BMI>25 then Disease
اگر وزن بزرگتر از ۷۳ و BMI بزرگتر از ۲۵ و زن باشد آنگاه شخص مریض است.

برای مقایسه‌ی میزان دقت دیتاهای پایگاه داده بیماران با دو مدل دیگر از مدل‌های داده‌کاوی (شبکه‌ی بیزین و عصبی) نیز انجام می‌شود با مقایسه‌ی این سه مدل می‌توان با استفاده از درصد خطاهای اعلام شده از طرف نرم‌افزار کلمنتاین فهمید کدام یک از این سه الگوریتم با توجه به درصد خطای اعلام شده دقیق‌تر پیش‌بینی می‌کند.

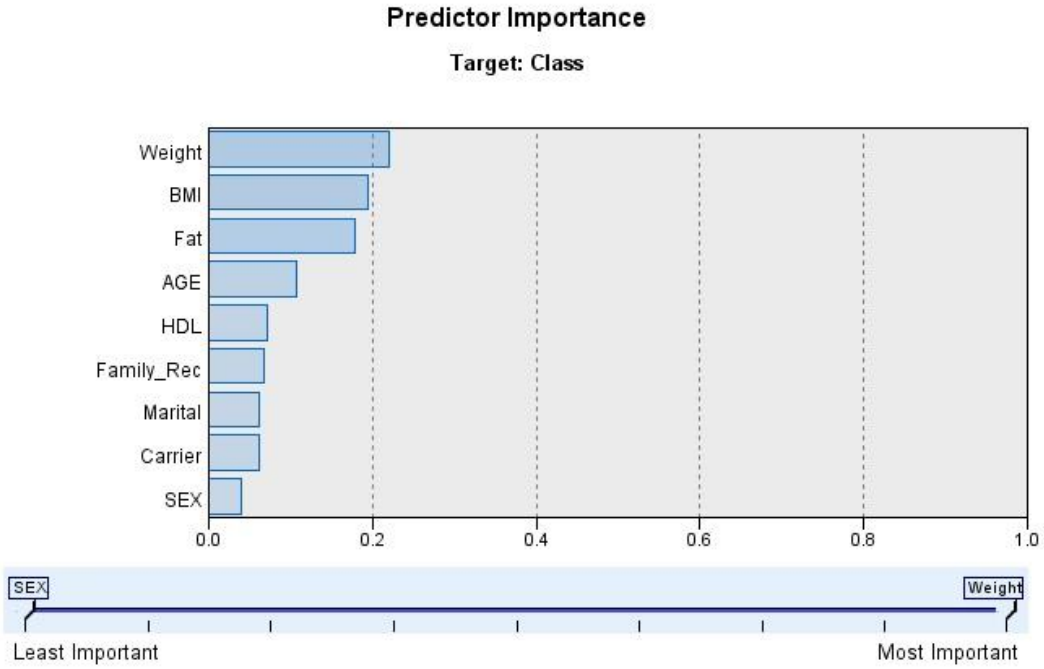
پیاده‌سازی با شبکه‌ی عصبی

با توجه به جدول ۷ مشخص گردید که میزان دقت این مدل حدوداً ۹۵٪ و میزان خطای آن حدوداً ۵٪ است. همان‌طور که در شکل ٦ آمده است وزن اشخاص بیشترین تأثیر در پیش‌بینی بیماری فوق را دارد و جنسیت کمترین تأثیر را در پیش‌بینی دارد.

جدول ۷- دقت پیش‌بینی مدل شبکه عصبی

Results for output field Class
Comparing $N-Class with Class

'Partition'	1_Training		2_Testing	
Correct	330	95.1%	138	90.2%
Wrong	17	4.9%	15	9.8%
Total	347		153	



شکل ۶- تأثیرات ویژگی‌های بیماران در مدل شبکه عصبی

شبکه‌ی بیزین

با توجه به جدول ۸ همان‌طور که ملاحظه می‌شود دقت این مدل حدوداً ۹۳٪ و خطای آن حدوداً ۷٪ است.

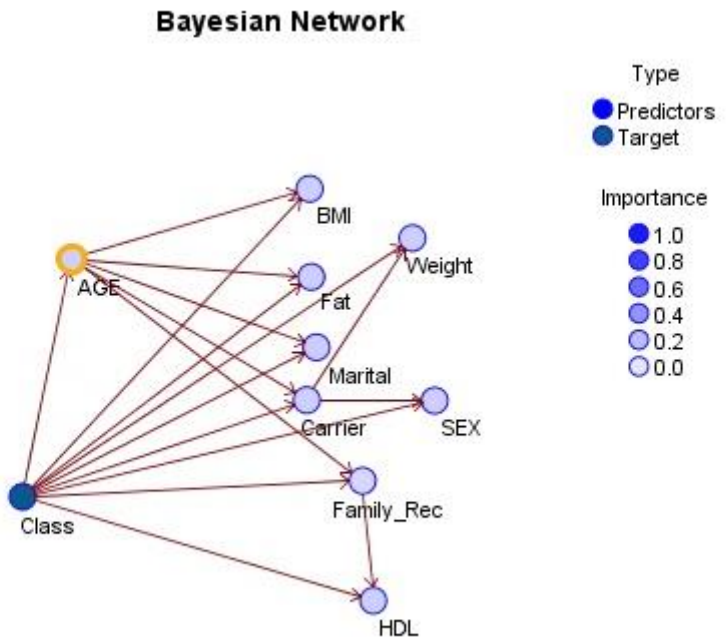
مقایسه‌ی ۳ مدل: با توجه به ۳ مدل ایجاد شده در این تحقیق همان‌گونه که مشاهده می‌شود مدل درخت تصمیم بیش‌ترین دقت و شبکه بیزین کم‌ترین دقت را دارا هست و به تبع آن درخت تصمیم کم‌ترین خطا و شبکه‌ی بیزین بیشترین خطا را دارا هست. همچنین چربی در درخت تصمیم بیش‌ترین تأثیر را در پیش‌بینی بیماری دیابت دارا هست، وزن در شبکه‌ی عصبی و سن در شبکه‌ی بیزین بیش‌ترین تأثیر را دارا هست. شکل ۸ مدل شبکه‌ی بیزین و شکل ۹ مقایسه‌ی میزان دقت سه روش را نشان می‌دهد.

جدول ۸- دقت پیش‌بینی مدل شبکه بیضین

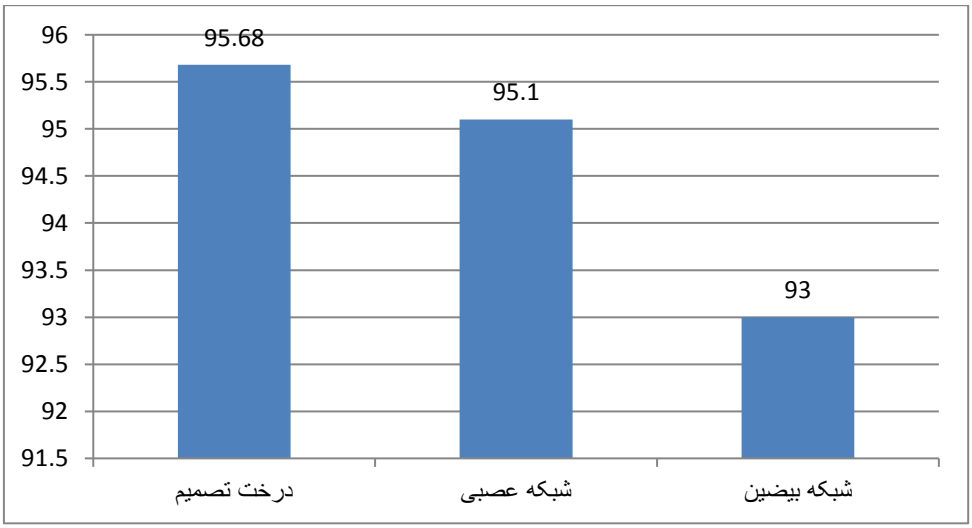
Results for output field Class

Comparing $B-Class with Class

'Partition'	1_Training		2_Testing	
Correct	323	93.08%	147	96.08%
Wrong	24	6.92%	6	3.92%
Total	347		153	



شکل ۸- مدل شبکه بیضین



شکل ۸- نمودار مقایسه دقت پیش‌بینی ۳ مدل

نتیجه‌گیری

امروزه در اکثر بیمارستان‌های ایران بانک اطلاعاتی وسیعی از ویژگی‌های بیماران موجود است که حجم بالایی از اطلاعات مربوط به سوابق پزشکی افراد است. پیدا کردن الگوی دانش این اطلاعات می‌تواند در جهت تحلیل و پیش‌بینی انواع بیماری‌ها و بهبود فرآیند پزشکی کمک شایانی کند. در عمق درون این حجم از داده‌ها الگوها و روابط بسیار جالبی میان پارامترهای مختلف به‌صورت پنهان باقی می‌ماند.

هدف این است که با به‌کارگیری مدل‌های داده‌کاوی و تحلیل داده های جمع آوری شده که از جامعه‌ی اطلاعات بیماران استفاده شده است، به سؤالاتی در زمینه‌ی رفتار بیماران و نظام پزشکی و تحلیل و پیش‌بینی سطح علمی آنها در حال و آینده پاسخ داده شود.

با ارایه‌ی این مدل پیشنهادی می‌توان نتایج و دستاوردهای زیر را در حوزه‌ی تحلیل و پیش‌بینی عملکرد نظام پزشکی با توجه به ویژگی بیماران ارایه نمود:

۱- در بین ویژگی‌های بیان شده متغیرهای سن و میزان قند بیشترین تأثیر را در پیش‌بینی وقوع بیماری دیابت (طبق تحلیل نرم‌افزار) را دارا هستند و با ایجاد رژیم غذایی مناسب می‌توان از ابتلا به این بیماری جلوگیری کرد.

۲- طبق تحلیل نرم‌افزار کلمتاین در مدل درخت تصمیم پس از قند خون به‌ترتیب چربی، BMI، وزن، سابقه‌ی خانوادگی، سن و جنسیت بیش‌ترین تأثیر را در جهت پیش‌بینی این بیماری دارا هستند.

۳- با توجه به الگوریتم‌های داده شده (درخت تصمیم، شبکه‌ی بیزین، شبکه‌ی عصبی) الگوریتم درخت تصمیم کم‌ترین خطای پیش‌بینی را دارا هست که مشخص می‌کند این مدل یک مدل دقیق‌تر نسبت به دو مدل شبکه‌ی عصبی و شبکه‌ی بیزین است.

٤- بر طبق خروجی مدل درخت تصمیم و با توجه به داده‌های بیماران مشخص گردید اگر یک خانم با وزن بیشتر از ۷۳ و BMI بیشتر از ۲۵ باشد قطعاً آن خانم مبتلا به بیماری دیابت است و یا اگر یک آقا با وزن کمتر و یا مساوی ٦۰ باشد و چربی بیشتر از ۱۰۰ داشته باشد و BMI بزرگتر از ۲۵ داشته باشد آن آقا مبتلا به این بیماری است.

۵- بر طبق مدل درخت تصمیم و با توجه به داده‌های بیماران مشخص گردید اگر یک خانم با وزن بیشتر از ٦۰ باشد و دارای سابقه‌ی خانوادگی نیز نباشد و وزنی بین ۷۱ تا ۷۷ داشته باشد و چربی کمتر از ۱۵۸ داشته باشد آن خانم سالم است. و یا اگر یک آقا سابقه‌ی خانوادگی نداشته باشد و وزنی کمتر از۷۳ داشته باشد و چربی کمتر از ۱۵۸ داشته باشد آن شخص سالم است.

سپاسگزاری

بدین‌وسیله نویسندگان مراتب تشکر و قدردانی را ازمرکز بهداشت شهید بلندیان قزوین بابت همکاری در جمع‌آوری داده‌ها اعلام می‌دارند.

مآخذ

1. Hosseini SE, Tavakoli F, Karami M. Medicinal Plants in the treatment of Diabetes mellitus. *Clinical Excellence* 2014; 2(2):64-89.

۲. ذباح، ایمان؛ اسکندری، اسما؛ سرداری، زهرا؛ نوقندی، ابوالفضل. تشخیص بیماری دیابت با استفاده از شبکه‌ی عصبی مصنوعی و عصبی–فازی. *مجله‌ی دانشگاه علوم پزشکی تربیت حیدریه*، ۱۳۹۷؛٦(۲):۱۰-۲۰.

3. Fayyad U, Piatetsky-Shapiro G, & Smyth P. From data mining to knowledge discovery in databases. *AI magazine* 1996; 17(3):37.

٤. عامری، حکیمه؛ علیزاده، سمیه؛ برزگری، اکبر. استخراج دانش از داده‌های بیماران دیابتی با استفاده از روش درخت تصمیم C5. *فصلنامه‌ی مدیریت سلامت*، ۱۳۹۲؛۱٦(۵۳):۵۸-۷۲.

5. Song Y, Liang J, Lu J, & Zhao X. An efficient instance selection algorithm for k nearest neighbor regression. *Neuro computing* 2017; 251:26-34.

6. Breault JL. Goodall CR., and Fos PJ.Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine* 2002; 26(1-2):37-54.

7. Miyaki K, Takei I, Watanabe K, Nakashima H, & Omae K. Novel statistical classification model of type 2 diabetes mellitus patrents for tailor-made prevention using data mining algorithm. *Journal of epidemiology/Japan Epidemiological Association* 2002; 12(3):243.

8. Al Jarullah AA. Decision tree discovery for the diagnosis of type II diabetes. In Innovations in Information Technology (IIT). *International Conference* on, 2011; pp. 303-307. IEEE.

9. Rohlfing CL, Wiedmeyer HM, Little RR, England JD, Tennill A, & Goldstein DE. Defining the relationship between plasma glucose and HbA1c analysis of glucose profiles and HbA1c in the Diabetes Control and Complications Trial. *Diabetes Care* 2002; 25(2):275-278.

10. Silverstein C, Brin S, Motwani R, & Ullman J, Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery* 2000; 4(2-3):163-192.

11. Quentin-Trautvetter J, Devos P, Duhamel A, & Beuscart R. Assessing association rules and decision

trees on analysis of diabetes data from the DiabCare program in France. *Studies in health technology and informations* 2002; 90, 557.

12. Juan G, Luo S, Jia H, Zhang T, and Han Y. Type 2 diabetes data processing with EM and C4.5algorithm. In Complex Medical Engineering, 2007. CME 2007. IEEE/ICME *International Conference* on, 2007; pp. 371-377. IEEE.
13. Huang Y, McCullagh P, Black N, & Harper R. Feature selection and classification model construcion on type 2 diabtic patients'data. *Artificial intelligence in medicine* 2007; 41(3):251-262.
14. Tang C, Li L, Shi J, Wu D, Wang M, Wu Y, et al. Curcumin in age-related diseases. Die Pharmazie-An *International Journal of Pharmaceutical Sciences 2020*; 75(11):534-9.

۱۵. مشکوتی، الهام؛ معینی، علی. تشخیص بیماری دیابت با استفاده از ماشین بردار پشتیبان. *کنفرانس بین‌المللی یافته های نوین پژوهشی در مهندسی برق و علوم کامپیوتر*، ۱۳۹٤.

Provide a Predictive Model to Identify People with Diabetes Using the Decision Tree

Abolfazl Kazemi[1], Hamid Bahador[2]*

1. *Department of Industry - Faculty of Industry, Islamic Azad University, Qazvin Branch, Qazvin, Iran*
2. *Department of Computer, Technical and Vocational University, Khoy Branch, Iran*

ABSTRACT

Background: Today, in most hospitals in Iran, there is an extensive database of patient characteristics that includes a large amount of information related to medical, family and medical records. Finding a knowledge model of this information can help to predict the performance of the medical system and improve educational processes.
Methods: Data mining techniques are analytical tools that are used to extract meaningful knowledge from a large data set. In this study, the information of 500 people referred to Shahid Bolandian Health Center in Qazvin has been used. In this research, a predicted model has been performed using decision tree data mining methods and neural network and Bayesian network.
Results: The decision tree model has the highest accuracy and the Bayesian network has the lowest accuracy in diagnosing diabetic patients, and consequently the decision tree has the least error and the Bayesian network has the highest error. The decision tree model with 95.68% had the highest accuracy in prediction.
Conclusion: Fat has the greatest effect in predicting diabetes and gender has the least effect in predicting diabetes. Based on the decision tree analysis, the rules obtained among the stated characteristics of age and sugar variables have the greatest effect in predicting the occurrence of diabetes (according to software analysis) and by creating a proper diet can prevent this disease Prevented.

Keywords: Data Mining, Diabetes, Decision Tree, Prediction, Neural Network

* Information and Communication Technology Department- Imam St. Intersection of Khayyam North- Department of Education, Urmia, West Azerbaijan, Iran. Tel: +984431932327. Fax: +984432222046, Email: hamidbahador52@gmail.com