آشکارسازی و تشخیص بیماری دیابت با استفاده از خوشه بندی و تکنیک‌های داده‌کاوی

فاطمه کردی [*۱]، ابوالفضل اسفندی [۲]، فرزاد همتی [۳]

۱- گروه مهندسی کامپیوتر، واحد بروجرد، دانشگاه آزاد اسلامی، بروجرد، ایران
۲- دانشگاه آزاد اسلامی واحد بروجرد، باشگاه پژوهشگران جوان و نخبگان بروجرد، ایران
۳- گروه مهندسی کامپیوتر، واحد بروجرد، دانشگاه آزاد اسلامی، بروجرد، ایران

چکیده

در این پژوهش سعی داریم به بررسی و تشخیص بیماری دیابت با استفاده از تکنیک‌های داده‌کاوی و خوشه‌بندی بپردازیم. به‌طورکلی مشکل عمده‌ای که در رابطه با بیماری دیابت وجود دارد عدم‌تشخیص به‌موقع و یا به‌طورکلی ضعف در تشخیص این بیماری است. داده‌کاوی یکی از بهترین و دقیق‌ترین راه‌ها در پردازش داده‌ها جهت پیش‌بینی می‌باشد. برای تجزیه‌وتحلیل اطلاعات در این پروژه از نرم‌افزار (MATLAB\R2016a) و ماشین‌بردارپشتیبانی(SVM) به‌منظور تشخیص بیماری استفاده‌شده است. تعداد داده‌های مورداستفاده ۷۶۸ داده است. این داده‌ها شامل ۹ ویژگی است که هریک بیان‌کننده مشخصه‌ای پزشکی مربوط به وضعیت فرد سالم و بیمار مبتلابه دیابت می‌باشد. با استفاده از الگوریتم خوشه‌بندی k-means بدون نظارت، داده‌های نویز دار دارای شرایط خاص را از دیتاست حذف خواهیم کرد تا دقت کلاس‌بندی بالا برود. نتایج، نشان‌دهنده صحت بالاتر روش پیشنهادی با دقت ۹۴/۲ درصد در این پژوهش در مقایسه با سایر روش‌های هوشمند مورداستفاده تاکنون جهت تشخیص بیماری دیابت می‌باشد.

کلمات کلیدی: الگوریتم k-means، الگوریتم ماشین بردار پشتیبان، بیماری دیابت، خوشه‌بندی، داده‌کاوی

۱. مقدمه

دیابت یا بیماری قند یکی از شناخته‌شده‌ترین بیماری‌ها در بین عموم مردم به شمار می‌آید، به‌طورکلی بیماری دیابت رایج‌ترین بیماری غددی و چهارمین علت مرگ در کشورهای پیشرفته می‌باشد و پیشگیری از آن بدون شک موضوع

[*] نویسنده اول : گروه مهندسی کامپیوتر ، واحد بروجرد ، دانشگاه آزاد اسلامی ، بروجرد ، ایران **Email: f.kordi89@gmail.com**
نویسنده دوم : دانشگاه آزاد اسلامی واحد بروجرد ، باشگاه پژوهشگران جوان و نخبگان بروجرد، ایران
Email: abolfazlesfandi@gmail.com
نویسنده سوم : گروه مهندسی کامپیوتر ، واحد بروجرد ، دانشگاه آزاد اسلامی ، بروجرد ، ایران **Email: Hemati.f@gmail.com**

حیاتی، درمانی و اقتصادی در قرن بیست و یکم است. دیابت می‌تواند منجر به مشکلات حاد فیزیکی در افراد و تأثیر اقتصادی بزرگی بر سیستم بهداشت و درمان ملی گردد[۱] . مخارج درمانی بیماران دیابتی در سال ۲۰۱۰ برابر با ۱۱/۶ درصد مخارج کل جهان بوده است. در این راستا شناسایی و تشخیص دقیق و به‌موقع این بیماری از اهمیت بسیار بالایی برخوردار است[۲و۳].

در حالت ساده می‌توان گفت اهمیت پیش‌بینی دیابت از این لحاظ است که بیمار پس‌ازاین آگاهی می‌تواند روش زندگی خود را تغییر داده و از این طریق از وقوع بیماری پیشگیری کند. در طی سال‌های اخیر استفاده از تکنولوژی‌های مدرن در علم پزشکی گسترش‌یافته است، در این پژوهش نیز با در نظر گرفتن موضوع سعی داریم به بررسی و تشخیص بیماری دیابت با استفاده از تکنیک‌های داده‌کاوی و خوشه‌بندی بپردازیم. بیماری دیابت در حال حاضر یکی از شناخته‌شده‌ترین بیماری‌ها به‌حساب می‌آید، این بیماری به‌طورکلی یکی از مخرب‌ترین بیماری‌ها در جهان به‌حساب می‌آید که عمدتاً در کشورهای توسعه‌یافته و درحال‌توسعه در حال گسترش می‌باشد. طبق گزارشی در ایران از هر ۲۰ نفر یک نفر به این بیماری مبتلا می‌باشد وعده‌ی کثیری از این افراد از بیماری خود اطلاع ندارند [۱].

به‌طورکلی مشکل عمده‌ای که در رابطه با بیماری دیابت وجود دارد عدم‌تشخیص به‌موقع و یا به‌طورکلی ضعف در تشخیص این بیماری است که این ضعف نیز به دلیل عدم انتخاب الگوی مناسب توسط پزشک و یا عدم استفاده مناسب از الگوهای استاندارد است؛ بنابراین پیاده‌سازی روشی که بتواند هر فرد را در تشخیص صحیح ابتلا یا عدم ابتلا به این بیماری یاری رساند می‌تواند گام مهمی در جهت پیشگیری و کنترل این بیماری به‌خصوص در مراحل ابتدایی آن باشد. دیابت همانند سایر بیماری‌ها عوارض بسیاری را به دنبال دارد، گرفتگی قلبی و عروقی و در نوع پیشرفته آن نابینایی، قطع اعضای بدن و اختلالات فکری ازجمله عوارض این بیماری می‌باشد، مشکل عمده‌ای که در حال حاضر در رابطه با این بیماری وجود دارد عدم‌تشخیص به‌موقع و یا به‌طورکلی ضعف در تشخیص این بیماری می‌باشد [۴].

حال با توجه به گسترش این بیماری استفاده از روش‌های جدید پزشکی در راستای بهره‌گیری در تشخیص بسیار موردتوجه قرارگرفته است که استفاده از تکنیک‌های داده‌کاوی و خوشه‌بندی در این پژوهش در جهت تشخیص این بیماری در دستور کار ما قرار دارد، به‌طورکلی داده‌کاوی می‌تواند ارتباطات و وابستگی‌های جدید و بدیعی را کشف کند که برای پزشکان مفید هستند و با استفاده از نتایج کار می‌توان نتایج خوبی را در راستای کمک به علم پزشکی گرفت [۵].

۲. خوشه‌بندی

خوشه‌بندی، وظیفه تقسیم یک گروه ناهم‌جنس را در چندین زیرگروه بر عهده دارد. این فرآیند یک تفاوت اساسی با طبقه‌بندی دارد؛ زیرا در این مدل هیچ‌گونه الگوی آموزشی نداریم. خوشه‌بندی به‌طور خودکار ویژگی‌های متمایزکننده زیرگروه‌ها را تعریف می‌کند و زیرگروه‌ها را سازماندهی می‌نماید و به‌عنوان نوعی قابلیت داده‌کاوی غیرمستقیم مطرح است. اغلب از خوشه‌بندی به‌عنوان اولین گام فرآیندهای داده‌کاوی یاد می‌شوند که قبل از سایر فرآیندها برای شناسایی گروهی از رکوردهای مرتبط باهم که بعداً بتوانند نقطه آغاز تحلیل‌ها باشند بر روی رکوردها اعمال می‌شود[۶].

۳. تجزیه‌وتحلیل خوشه‌بندی

تعدادی از تکنیک‌های خوشه‌بندی نظیر الگوریتم‌های خوشه‌بندی داده‌کاوی محور نظیر:

- K-MEANS
- Y-MEANS
- FUZZY C-MEANS

وجود دارند که با توجه به محدودیت‌هایمان در ارتباط با موضوع و تدوین پژوهش به بررسی از الگوریتم K-MEANS می‌پردازیم. همچنین با تلفیق این روش خوشه‌بندی و تکنیک‌های داده‌کاوی به پردازش داده‌های متناسب با پژوهش و تشخیص بیماری دیابت می‌پردازیم[۷].

۳. ۱. خوشه‌بندی K-MEANS

می‌توان گفت این الگوریتم به‌سختی قسمت‌بندی شده و در بسیاری جاها به‌سادگی و با سرعت به کار می‌رود. این الگوریتم فاصله اقلیدسی را با اندازه‌های مشابه استفاده می‌کند. مفهوم خوشه‌بندی سخت این است که یک آیتم در مجموعه داده می‌تواند متعلق به یکی و فقط یک خوشه در یک‌زمان باشد. این الگوریتم یک تحلیل خوشه‌بندی است به‌طوری‌که آیتم‌های گروه‌ها بر اساس مقادیر مشخص در داخل K به خوشه‌هایی که آیتم‌ها در همان خوشه ویژگی‌های مشابه دارند ملحق نمی‌شود. چون آن‌ها در خوشه‌های متفاوت مشخصه‌های متفاوتی دارند. عملکرد اقلیدسی با محاسبه فاصله بین دو آیتم به‌صورت زیر استفاده می‌شود.

$$d\,(p.q) = \sqrt{(p1-q2)^2 + (p2-q2)^2} \qquad (۱)$$

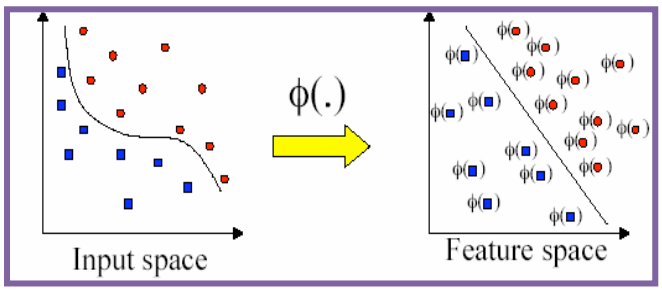
در اینجا $p = (p1.\ p2.....pm)$ و $q = (q1.\ q2....qm)$ دو بردار ورودی با مشخصه‌های m کمیت هستند. الگوریتم اجرا می‌شود تا مجموعه‌های آموزش داده که ممکن است شامل ترافیک عادی و یا غیرعادی شود بدون اینکه قبلاً برچسب خورده باشد. ایده اصلی این دیدگاه بر اساس فرضیه است که ترافیک عادی و غیرعادی از خوشه‌ها می‌باشد. همچنین داده ممکن است شامل بخش‌های مجزایی شود که کدام آیتم‌های داده از آیتم‌های دیگر در خوشه خیلی متفاوت هستند و کدام متعلق به هیچ خوشه‌ای نیست. بخش مجزا با مقایسه شعاع آیتم‌های داده درک می‌شود به‌طوری‌که اگر شعاع یک آیتم داده بزرگ‌تر از یک آستانه ارائه‌شده باشد.پس به‌صورت یک بخش مجزا دیده می‌شود؛ اما این موضوع یعنی پردازش خوشه‌بندی k-means را که تعداد بخش مجزا و کوچک است را بر هم نمی‌زند[۷].

۱.۱.۳. روش کار الگوریتم خوشه‌بندی k-means

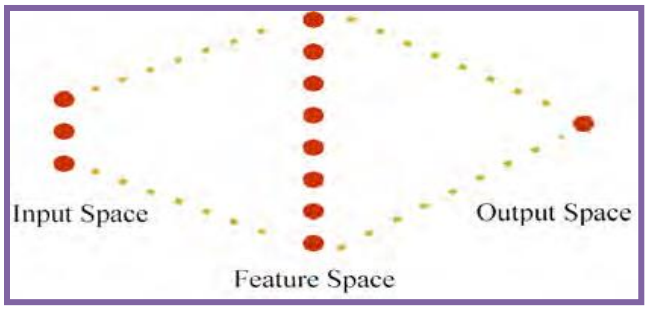
- تعریف تعداد خوشه‌های k برای مثال اگر k=2 باشد.ما در آموزش دو خوشه متفاوت، ترافیک عادی و غیرعادی را فرض می‌کنیم.
- مقداردهی اولیه مراکز ثقل خوشه k، این کار با انتخاب اتفاقی آیتم‌های داده k از مجموعه داده انجام می‌شود.
- محاسبه فاصله هر آیتم تا مراکز ثقل همه‌ی خوشه‌ها با به‌کارگیری فاصله اقلیدسی متریک که برای یافتن شباهت بین آیتم‌ها در مجموعه داده باکار می‌رود.
- اختصاص هر آیتم با نزدیک‌ترین مرکز ثقل خوشه،در این روش همه‌ی آیتم‌ها به خوشه‌های مختلف اختصاص خواهند یافت، به‌طوری‌که هر خوشه آیتم‌هایی با ویژگی‌های مشابه خواهد داشت.
- پس از اختصاص یافتن همه‌ی آیتم‌ها به خوشه‌های مختلف، میانگین خوشه‌های تغییریافته مجدداً محاسبه می‌شود. میانگین اخیراً محاسبه‌شده به‌عنوان مرکز ثقل جدید اختصاص می‌یابد.
- تکرار مرحله ۳ تا زمانی که مرکز ثقل خوشه تغییر نکند.
- برچسب زدن خوشه‌ها به‌صورت عادی و غیرعادی،به تعداد آیتم‌های داده هر خوشه بستگی دارد.
- می‌توان گفت یک مشکل اساسی در این روش خوشه‌بندی وجود دارد،که این مشکل تعیین بخش اولیه و تعداد مناسب خوشه‌های K می‌باشد.همچنین گاهی اوقات منجر به هم ترازی می‌شود،که میانگین کدام پردازش خوشه بندی ممکن است با چند خوشه خالی به پایان برسد [۷].

۴. ماشین بردار پشتیبان

در این پژوهش از ماشین بردار پشتیبانی (SupportVectorMachines)SVMبه‌منظور تشخیص بیماری استفاده‌شده است. SVMیک الگوریتم یادگیری است،که می‌تواند در کاربردهایی نظیر جداسازی و طبقه‌بندی داده‌ها مورداستفاده قرار گیرد. این الگوریتم پس از طی دو مرحله آموزش می‌تواند داده‌های ورودی را به دسته تقسیم کند و به همین جهت از سادگی منحصربه‌فردی نسبت به سیستم‌های هوشمندی هم چون شبکه‌های عصبی برخوردار است. البته به علت اینکه SVM در جداسازی داده به بیش از دو گروه ناتوان است، کمتر در مواردی که تعداد گروه‌های خروجی بیش از دو گروه باشد استفاده می‌شود.



شکل ۱: نگاشت داده‌ها از فضای ورودی به فضای ویژگی



شکل ۲: نمونه‌ای از نوع تبدیل SVM

درصورتی‌که داده در فضای داده‌شده توسط یک خط قابل جداسازی نباشد، الگوریتم با نگاشت ویژگی‌ها به‌وسیله تابع به ابعاد بالاتر سعی می‌کند توسط یک فرا صفحه با دقت بالاتری داده‌ها را تفکیک نماید. حداکثر حاشیه ایده اولیه SVM خطی بوده است که در آن مطابق شکل (۱) و (۲) طبقه‌بندی داده‌های توسط کرنل خطی صورت می‌گیرد.

روش طبقه‌بندی ماشین بردار پشتیبان به این صورت می‌باشد،که با فرض این‌که مجموعه داده‌ها را به‌صورت مجموعه زیر نمایش دهیم:

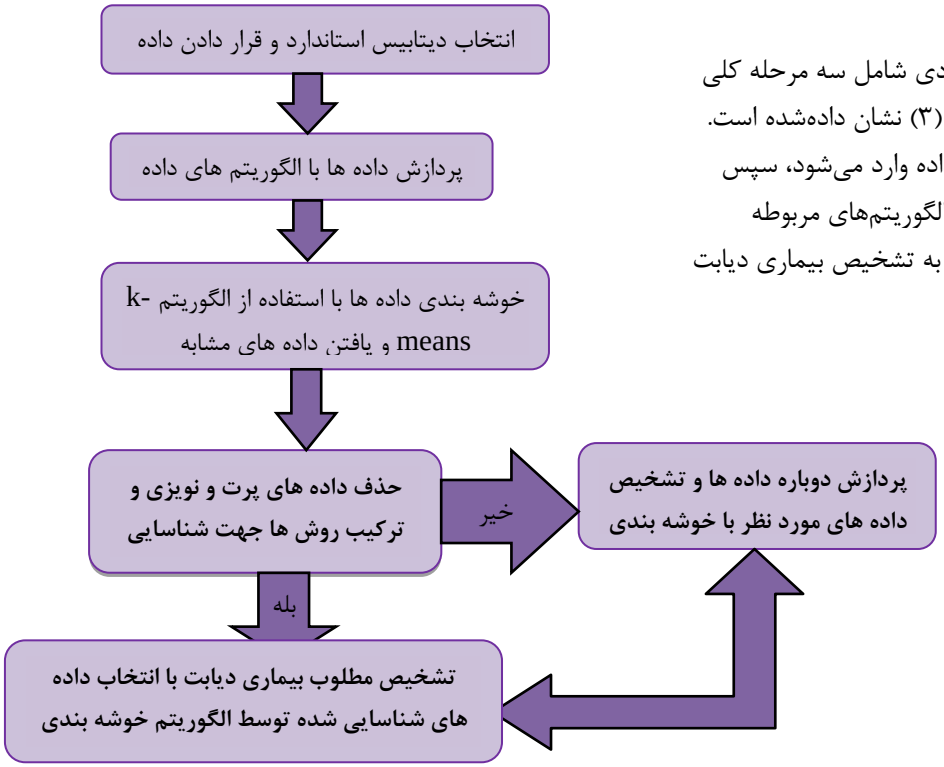
$$[(X1,C1),(X2,C2),...(Xn,Cn)] \quad (۵)$$

Xi یک نمونه از داده‌ها را نشان می‌دهد.Ci کلاس یا طبقه مربوط به نمونه Xi را مشخص می‌کند. به‌عنوان‌مثال Ci می‌تواند نشان‌دهنده افرادی که دیابت دارند یا دیابت ندارند،باشد.و آن‌ها را به دوطبقه $c_i = \{-1,1\}$ تفکیک کنیم. هر Xi یک بردار p بعدی از اعداد حقیقی است.

روش‌های طبقه‌بندی خطی، سعی دارند که با ساختن یک ابر سطح (که عبارت است از یک معادله خطی)، داده‌ها را از هم تفکیک کنند. روش طبقه‌بندی ماشین بردار پشتیبان که یکی از روش‌های طبقه‌بندی خطی است،بهترین ابر سطحی را پیدا می‌کند،که با حداکثر فاصله (maximum margin)، داده‌های مربوط به دوطبقه را از هم تفکیک کند[۱۰ و ۱۱].

۵. الگوریتم پیشنهادی

الگوریتم پیشنهادی شامل سه مرحله کلی می‌باشد که در شکل (۳) نشان داده‌شده است. در مرحله اول پایگاه داده وارد می‌شود، سپس به اجرای تکنیک‌ها و الگوریتم‌های مربوطه می‌پردازیم و درنهایت به تشخیص بیماری دیابت دست می‌یابیم.



شکل ۳: الگوریتم پیشنهادی

۶. پایگاه داده

هدف این پژوهش تشخیص بیماری دیابت می‌باشد که برای محقق شدن این هدف از پایگاه داده uci استفاده کرده‌ایم. این دیتاست شامل اطلاعات زنان با ۹ ویژگی و ۷۶۸ سطر می‌باشد که توسط سازمان بهداشت جهانی معیارهای تشخیص دیابت گردآوری‌شده و شرح داده‌ها به‌صورت زیراست:

۱- تعداد دفعات بارداری[۱]
۲- غلظت گلوکز پلاسما ۲ ساعت در آزمایش تحمل گلوکز خوراکی خون[۲]
۳- فشارخون دیاستولیک (mmHg)
۴- ضخامت پوست ماهیچه[۳] (mm)
۵- انسولین سرم ۲ ساعته[۴] (UU / میلی‌لیتر)
۶- شاخص توده (جرم) بدن[۵] (وزن kg / ((ارتفاع در متر) ^ ۲))
۷- داشتن سابقه دیابت[۶]
۸- سن[۷]
۹- نتیجه آزمایش[۸]

۷. ابزارهای پیاده‌سازی

در این پژوهش جهت تشخیص دیابت از روش خوشه‌بندی استفاده‌شده است و سعی خواهیم کرد تا عمل خوشه‌بندی را با فیلدهای ورودی انجام بدهیم و داده‌های نویز دار را حذف کنیم. در ادامه کار با استفاده از الگوریتم بردار پشتیبان، عمل کلاس‌بندی را نیز انجام خواهیم داد. حال برای خوشه‌بندی تمام ورودی‌ها را با ویژگی‌های ۱ تا ۸ در نظرخواهیم گرفت. بعد از خوشه‌بندی دیتاست به هشت خوشه، درصورتی‌که در خوشه موردنظر داده بر چسب‌دار هرکدام از داده‌ها کمتر باشد آن داده بر چسب‌دار را از خوشه موردنظر حذف خواهیم کرد، بعد از حذف داده‌های نویز دار، دقت کلاس‌بندی داده‌های بدون نویز افزایش قابل‌توجهی خواهد داشت.حال برای تست و آموزش داده‌ها با بردار پشتیبان، ¼ دیتاست را برای تست در نظر گرفتیم و از بقیه دیتاست برای آموزش شبکه بردار پشتیبان استفاده می‌کنیم که برای پیاده‌سازی الگوریتم پیشنهادی از شبیه‌ساز متلب (MATLAB\R2016a) استفاده‌شده است.

[1] Numberpregnant
[2] Plasma glucose
[3] skin fold thickness
[4] serum insulin
[5] Body mass
[6] Diabetes pedigree
[7] Age
[8] Class variable

۸. ارزیابی الگوریتم‌های دسته‌بندی

در این بخش معیارهای ارزیابی برای دسته‌بندی موردبررسی قرار خواهند گرفت.ارزیابی یک مدل دسته‌بندی می‌تواند براساس نمونه‌های آموزشی و آزمایشی صورت گیرد. برای ارزیابی باید برچسبی که مدل دسته‌بندی به نمونه‌ها نسبت داده‌شده است با برچسبی که نمونه‌ها متعلق به آن است مقایسه شود. وقوع حالات مختلف برای دسته‌ها با توجه به مجموعه داده‌های ورودی برای دسته‌بندی با مقادیر TP,FP,TN, FN برای دودسته مثبت و منفی در جدول (۱) نشان داده‌شده است.

جدول ۱: حالات هزینه‌ی دسته‌ها

	تخمین زده‌شده	
	مثبت	منفی
مثبت	TP	FN
منفی	FP	TN

- ✓ **TP**: مخفف True Positive است و بیانگر تعداد نمونه‌هایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز آن‌ها را به‌درستی به دسته مثبت نسبت داده است.
- ✓ **FP**: مخفف False Positive است و بیانگر تعداد نمونه‌هایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی آن‌ها را به‌اشتباه به دسته مثبت نسبت داده است.
- ✓ **TN**: مخفف True Negative است و بیانگر تعداد نمونه‌هایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز آن‌ها را به‌درستی به دسته منفی نسبت داده است.
- ✓ **FN**: مخفف False Negative است و بیانگر تعداد نمونه‌هایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته بندی آن‌ها را به‌اشتباه به دسته منفی نسبت داده است.

با توجه به پارامترهای مطرح‌شده در جدول (۱) معیارهای ارزیابی مختلفی ارائه‌شده است که ازجمله مهم‌ترین آن‌ها می‌توان به معیار درستی، دقت، فراخوانی اشاره کرد[۸و۹].

۹. معیارهای ارزیابی

- ✓ میزان صحت دسته‌بندی (accuracy)
- ✓ دقت تشخیص کل دیابتی مثبت (recall)
- ✓ دقت تشخیص دیابتی مثبت (precision) [۸و۹].

۹. ۱. معیار Accuracy (دقت، صحت)

مهم‌ترین معیار برای تعیین کارایی یک الگوریتم دسته‌بندی می‌باشد. این معیار، دقت کل یک دسته‌بندی را محاسبه می‌کند. این معیار نشان‌دهنده این موضوع است که چند درصد از کل مجموعه داده‌ها به‌درستی دسته‌بندی‌شده است.

$$Accuracy=\frac{TP+TN}{TP+FP+FN+TN}\times 100\% \quad (۲)$$

دو مقدار TP و TN مهم‌ترین مقادیری هستند که باید بیشینه شوند تا کارایی دسته‌بندی به حداکثر برسد[۸و۹].

۹. ۲. معیار Precision

درصدی از نمونه‌ها را نشان می‌دهد که از میان تمامی نمونه‌ها که توسط دسته‌بندی به آن دسته نسبت داده‌شده‌اند، درست دسته‌بندی‌شده‌اند.به عبارتی دقت دسته‌بندی دسته i را با توجه به‌کل مواردی نشان می‌دهد که برچسب i برای نمونه موردبررسی توسط دسته‌بندی پیشنهادشده است. نحوه‌ی محاسبه‌ی این معیار نشان داده‌شده است. اندیس i در این پارامتر به این مفهوم است که پارامترها باید برای هر دسته i محاسبه شوند.

$$Precision = \frac{TP_i}{TP_i+FP_i} \quad (۳)$$

۹. ۳. معیار Recall

برای یک دسته‌بندی، درصد نمونه‌هایی را نشان می‌دهد که از میان تمام نمونه‌ها متعلق به آن دسته، به‌درستی، دسته‌بندی‌شده است. به عبارتی دقت دسته‌بندی دسته i را با توجه به‌کل نمونه‌های با برچسب i نشان می‌دهد.

$$Recall_i = \frac{TP_i}{TP_i+FN_i} \quad (۴)$$

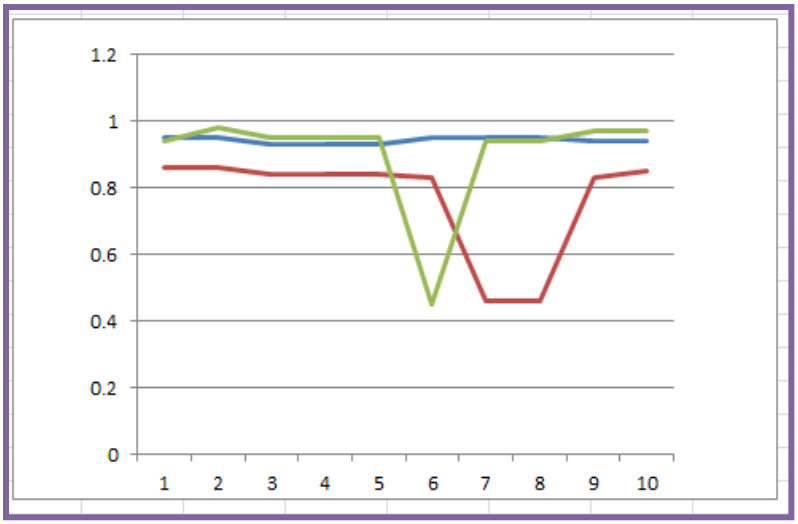
نکته قابل‌توجه این است که معیار Recall کارآیی دسته‌بندی را با توجه به تعداد رخداد دسته i نشان می‌دهد درحالی‌که معیارPrecision اساساً مبتنی بر دقت پیش‌بینی دسته می‌باشد و بیانگر آن است که به چه میزان می‌توانیم به خروجی دسته‌بندی اعتماد کنیم[۸و۹].

مقدار دقت حساسیت تشخیص با تغییر تکرار الگوریتم را در پژوهش انجام‌شده به‌صورت زیر در جدول (۲) می‌باشد.

جدول ۲: دفعات اجرای نمودار به همراه مقادیر سه معیار موردبررسی

معیار ارزیابی	رنگ قابل‌نمایش	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰
Accuracy	آبی	۰/۹٤	۰/۹٥	۰/۹٥	۰/۹٥	۰/۹۳	۰/۹۳	۰/۹۳	۰/۹٥	۰/۹٥	۰/۹٤
Recall	قرمز	۰/۸۳	۰/٤٦	۰/٤٦	۰/۸۳	۰/۸٤	۰/۸٤	۰/۸٤	۰/۸٦	۰/۸٦	۰/۸٥
Precision	سبز	۰/۹۷	۰/۹٤	۰/۹٤	۰/٤٥	۰/۹٥	۰/۹٥	۰/۹٥	۰/۹۸	۰/۹٤	۰/۹۷

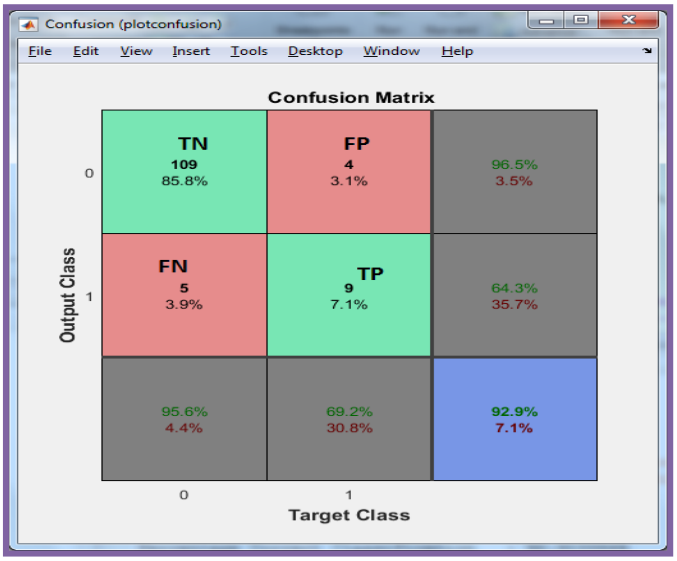
در شکل (۴) مشاهده می‌کنید که دقت الگوریتم پیشنهادی نرخ ثابتی را دارد اما دو معیار دیگر با افت شاخص مواجه شده است، البته با میانگین‌گیری نرخ ثابتی دارند.



شکل ۴: مقایسه سه معیار با دفعات تکرار ۱۰ بار

۱۰. ماتریس آشوب

فصل مشترک تعاریفی که برای مفهوم آشوب ارائه‌شده است، تأکید بر این نکته است که آشوب دانش بررسی رفتار سیستم‌هایی است که اگرچه ورودی آن‌ها قابل‌تعیین و اندازه‌گیری است،اما خروجی این سیستم‌ها ظاهری تصادفی دارد.شاید به همین دلیل بود که استوارت ریاضیدان برجسته این موضوع را مفهومی احتمالاتی می‌دانست، اما چیزی نگذشت که وی تعریف خود را اصلاح کرد و به تعریفی رسید که تقریباً مورد تأیید عمومی قرار دارد. بر اساس این تعریف، آشوب به توانایی یک الگو و مدل ساده گفته می‌شود که اگرچه خود این الگو نشانی از پدیده‌های تصادفی در خود ندارد،اما می‌تواند منجر به ظهور رفتارهای بسیار در محیط شود. شکل (۵) ماتریس آشوب را نمایش می‌دهد.



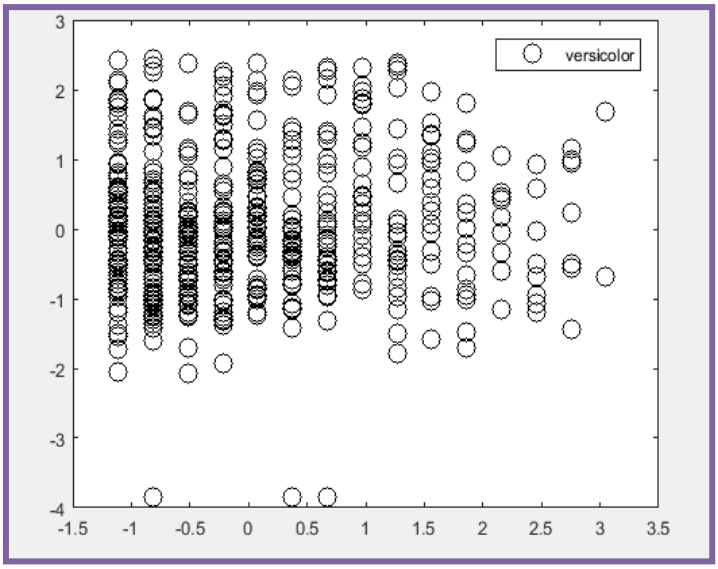
شکل ۵: عملکرد ماتریس آشوب بر روی پروژه

۱۱. روش‌های پیاده‌سازی در این پژوهش

- پیاده‌سازی بدون الگوریتم k-means
- پیاده‌سازی با الگوریتم k-means
- کاهش ویژگی‌ها

۱۱. ۱. پیاده‌سازی بدون الگوریتم k-means

داده‌های مورد آزمایش و تست را تعیین می‌کنیم. نهایتاً دقت صحت کلاس‌بندی را اندازه‌گیری می‌کنیم. بردارهای پشتیبان را نیز در تصویر زیر مشاهده می‌کنید. همچنین با بررسی ماتریس آشوب می‌توان تعداد کلاس‌بندی اشتباه را مشاهده کرد. ۰/۴۱۱۵، دقت کلاس‌بندی است که بر اساس تعداد درستی برنامه تشخیص داده‌شده است.این از روش ¼ داده‌ها۳۴ مورد است یعنی صحت تشخیص درست مثبت ۷۲/۹۱ درصد می‌باشد. در شکل (۶) خروجی بردار پشتیبان نمایش داده‌شده است.



در خروجی بردارپشتیبان، محور x فضای چندبعدی است که شامل چندین ویژگی خواهد بود و محور y خروجی با برچسب داده ها خواهد بود. درواقع به تعداد ویژگی‌ها برچسب تولید می‌کند،دایره‌های خالی متعلق به خوشه‌بندی ۱، دایره‌های توپر هم برای خوشه‌بندی ۰ است.

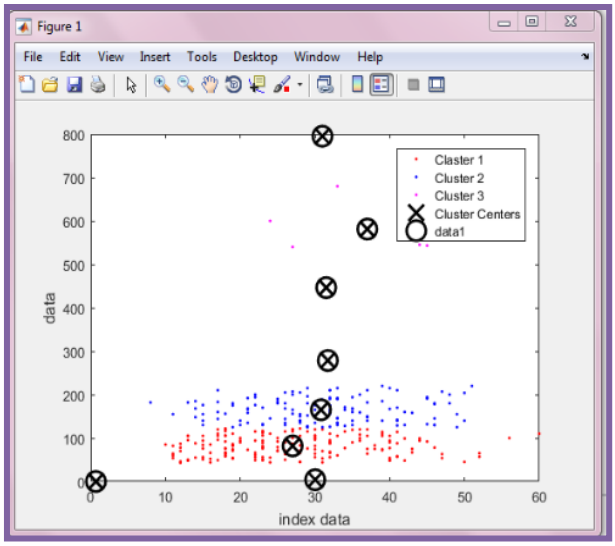
شکل ۶: خروجی بردار پشتیبان بدون الگوریتم k-means

دقت کلاس‌بندی در حالت کلی به‌صورت زیر می‌باشد:

Percentage Correct Classification: 72. 916667%

Percentage Incorrect Classification: 27. 083333%

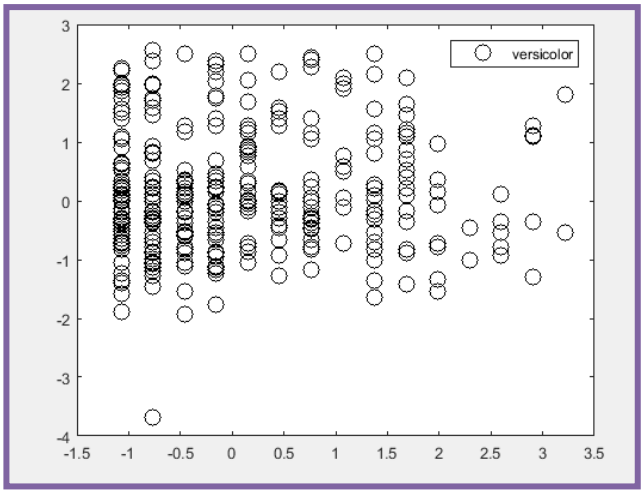
۱۱. ۲. پیاده‌سازی با الگوریتم k-means



در این روش از کار با استفاده از الگوریتم خوشه بندی k-means داده‌ها را به ۸ خوشه تقسیم کرده همه خوشه‌ها را برای حذف داده‌های نویز دار بررسی می‌کنیم.در هر خوشه با بررسی داده‌های بر چسب‌دار متعلق به یک کلاس درصورتی‌که از دیگر داده‌ها کمتر باشد، به‌عنوان نویز شناخته و از دیتاست حذف می‌کنیم.بعد از آماده‌سازی دیتاست داده‌ها، عمل کلاس بندی را با استفاده از بردارهای پشتیبان انجام داده و دقت کلاس‌بندی را اندازه‌گیری می‌کنیم ¼ دیتاست را به‌عنوان تست در نظر می‌گیریم.

شکل ۷: خروجی خوشه‌بندی

بعد از هر بار اجرا حدود۳۰۰سطر از داده نویز دار را حذف کرده،صحت دیتاست را بالا می‌بریم. بعد از کلاس‌بندی میزان خطا به‌صورت قابل‌توجهی کاهش‌یافته است. تابعK-means مجموعه را به k خوشه‌ی کاملاً مجزا تبدیل می‌نماید و اندیس هر خوشه را به داده‌های مجموعه نسبت می‌دهد. نمونه‌ها معمولاً به‌وسیله بردارهایی (نقاطی) در یک فضای چندبعدی نمایش داده می‌شوند که هر بعد یک صفت (متغیر) مشخص از نمونه‌ها را نشان می‌دهد.مانند سن، فشارخون و غیره اگر فرض کنیم.در شکل (۷)و (۸) خروجی خوشه‌بندی و بردار پشتیبان نمایش داده‌شده است.



شکل ۸: خروجی بردار پشتیبان با الگوریتمk-means

اگر n نمونه داشته باشیم که هرکدام دارایd صفت باشند، پس یک ماتریس نمونهn×dرا می‌توانیم تشکیل دهیم که هر سطر این ماتریس یک نمونه و هر ستون یک صفت مربوط به نمونه‌ها می‌باشد.داده‌ها با استفاده از روشmin-max نرمال‌سازی می‌شوند،نرمال‌سازی تغییر مقیاس داده‌ها به‌گونه‌ای است که آن‌ها را به یک‌فاصله کوچک و معین نگاشت می کند. مطابق شکل نه‌تنها مراکز خوشه‌ها(⊗)مشخص‌شده است.بلکه هر خوشه بارنگی خاص نمایش داده‌شده و محورXدر این شکل سن افراد و محورYتعداد داده‌هایی که در این خوشه قرارگرفته‌اند را مشخص می‌کند.

نمایی از مراکز خوشه‌های هشت‌گانه از چهل‌ویک صدم به ۲۱صدم کاهش‌یافته است.۰/۲۶۳۲،دقت کلاس‌بندی است که براساس تعداد درستی برنامه تشخیص داده‌شده است و میزان خطای کلاس‌بندی که کاهش‌یافته است، با بررسی ماتریس آشوب صحت کلاس‌بندی ۹۰درصد است.

Percentage Correct Classification: 93. 181818%

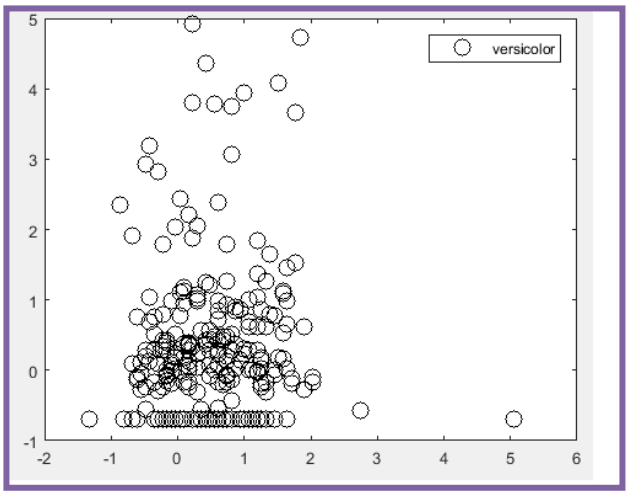
Percentage Incorrect Classification: 6. 818182%

۱۱. ۳. کاهش ویژگی‌ها

حال با تست حالت‌های مختلف،بهترین حالت ورودی را برای کاهش خطا و افزایش دقت انتخاب کردیم،که از ویژگی‌های۴تا۸ به ترتیب استفاده‌شده است:

۴- skin fold thickness (ضخامت پوست ماهیچه (mm))
۵- serum insulin (انسولین سرم ۲ ساعته (UU / میلی‌لیتر)
۶- Body mass (شاخص توده بدن (وزن kg / ((ارتفاع در متر) ^ ۲))
۷-Diabetes pedigree (داشتن سابقه دیابت)
۸- Age (سن)

به‌عنوان ورودی در نظر گرفتیم،که از الگوریتمk-means برای حذف داده‌های نویز دار استفاده می‌کنیم.صحت تشخیص کلاس‌بندی افزایش قابل‌ملاحظه‌ای داشت،که در حالت بدون حذف نویز تأثیر چندانی داشت. درروش اولی ذکرشده صحت تشخیص کلاس‌بندی ۹۳/۷ درصد افزایش یافت. میزان خطا نیز به۱۷درصد رسید.در شکل(۹)نیز تفکیک داده‌ها با بردار پشتیبان قابل‌مشاهده است.



شکل ۹: خروجی بردار پشتیبان با الگوریتم k-means و کاهش ویژگی‌ها

Percentage Correct Classification: 93. 700787%

Percentage Incorrect Classification: 6. 299213%

۱۲. مقایسه روش پیشنهادی با سایر روش‌ها

در این جدول(۳)به‌صورت خلاصه روش‌های بررسی‌شده در خصوص بیماری دیابت با الگوریتم‌های داده‌کاوی و درصد درستی آن‌ها در مقایسه با روش پیشنهادی می‌باشد.

جدول ۳: روش‌های بررسی‌شده بیماری دیابت با داده‌کاوی

محقق	روش	Accuracy
پلیت و گون[1](۲۰۰۸)	LS – SVM (10- fold CV)	۷۸/۲۱
ایسا و مامات[2] (۲۰۱۱)	GDA-LS – SVM (10- fold CV)	۸۲/۰۵
آیبینو و همکاران(۲۰۱۱)	Clustered – HMLP	۸۰/۵۹
پلیت و گونز (۲۰۰۷)	Combining PCA and ANFIS	۸۱/۲۸
قهرمانی و اله وردی(۲۰۰۸)	Hybrid system (ANN and FNN)	۸۴/۲
پتیل (۲۰۱۰)	Hybrid Prediction Model (HPM)	۹۲/۳۸
فاطمه کردی(۲۰۱۷)	svm with kmeans	۹۴/۲

۱ Polat and Gunes
2 Isa&Mamat

۱۳. نتیجه‌گیری

در این پژوهش ما یک روش با خطای کمتر نسبت به سایر روش‌های تشخیص دیابت توسعه دادیم. این دیتاست شامل داده‌های نویز دار و داده‌های تکراری است. در این روش ما با استفاده از الگوریتم خوشه‌بندی k-means برای تشخیص داده‌های نویز دار استفاده کردیم.بعد از پاک‌سازی داده‌ها،داده‌های بدون نویز داردایم.داده‌ها بیشترین تأثیر را در صحت کلاس‌بندی با دقت بالا رادارند.در ادامه کار که داده‌ها را بهوسیلهsvm کلاس‌بندی کردیم،نسبت به سایر روش‌های قبلی کارشده از صحت و دقت بالایی برخوردار است.همچنین با انتخاب ویژگی‌های مهم تأثیر زیادی در صحت کلاس‌بندی داده‌ها دارند.میزان صحت تشخیص داده‌ها به مرز۹۴/۲درصد رسید.

۱۴. منابع و مأخذ

۱. مرجب، سونا. و فاضلی، امین. (۱۳۹۵)، " *استخراج ویژگی و تشخیص هوشمند بیماری رتینوپاتی دیابتی درتصاویر شبکیه* "، سومین کنفرانس ملی توسعه علوم مهندسی، مازندران،تنکابن، موسسه آموزش عالی آیندگان.

2. Bose, I.(2006),*"Data Mining in Diabetes Diagnosis and Detection"*, Encyclopedia of Data Warehousing and Mining: IZ, p. 257.

3. Huang ,Y. P. and Black, N. and Harper, R.(2007),*"Feature selection and classification model construction on type 2 diabetic patients' data",* Artificial intelligence in medicine, vol. 41, pp. 251-262.

4. Sankar, K. Pal. and Pabitra, Mitra.(2009),"*Pattern Recognition Algorithms for Data Mining*", (Book), Chapman & Hall/CRC.

5. Acharya, U. R. and Faust, O. and Sree, S.V. and Ghista,D.N. and Dua,S. and Joseph, P. and Tamura,T.(2013),"*An integrated diabetic index using heart rate variability signal features for diagnosis of diabetes*", Computer methods in biomechanics and biomedical engineering, 16(2), 222-234.

6. Snehal,A. Mulay. And Devale, P. R. and Garje,G.V. (2010), *"Intrusion Detection System using Support Vector Machine and Decision tree*" international journal computer application (0975 –8887) Volume 3 –No.3.

7. Sonukumari, A.(2013) , "*A data mining approach for the Diagnosis of Diabetes Mellitus* ". 7th International Conference on Intelligent Systems and Control,Coimbatore , TamilNadu, India.

8. Domingos,P.(1999), "*MetaCost: a general method for making classifiers cost-sensitive*," in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 155-164.

9. Engelgau, Michael M. and Narayan, K.M. Venkat . and Herman William H.(2000), *"Screening for type 2 diabetes"*, Diabetes care, vol. 23, pp. 1563-1580.

10. Han,J. and Kamber,M. and Pei,J.(2006), " *Data mining: concepts and techniques*" Morgan kaufmann.

11. Xue,H. and Yang,Q. and Chen,S .(1999),*"SVM: Support Vector Machines"*