مجله دیابت و متابولیسم ایران. فروردین و اردیبهشت ۱۴۰۲؛ دوره ۲۳ (شماره ۱) ۶۷-۵۳

مقاله پژوهشی

طراحی الگوریتم مبتنی بر داده‌کاوی به‌منظور پیش‌بینی دیابت

نوید رفیعی[1*]

چکیده

مقدمه: دیابت سالانه باعث مرگ‌ومیر فراوانی می‌شود و تعداد افراد زیادی که به این بیماری مبتلا هستند به اندازه‌ی کافی وضعیت سلامت خود را درک نمی‌کنند. این مطالعه یک مدل مبتنی بر داده‌کاوی به‌منظور تشخیص و پیش‌بینی زودهنگام دیابت پیشنهاد می‌کند.

روش‌ها: با وجود اینکه تکنیک کا-میانه ساده است و می‌توان آن را برای طیف گسترده‌ای از انواع داده‌ها استفاده کرد، اما نسبت به موقعیت های اولیه مراکز خوشه که نتیجه‌ی نهایی خوشه را تعیین می‌کنند بسیار حساس است، به‌طوری‌که یا یک مجموعه داده‌ی خوشه‌بندی شده مناسب و کارا را برای مدل رگرسیون لجستیک فراهم می‌کند و یا مقدار کمتری داده را در نتیجه‌ی خوشه‌بندی ناصحیح مجموعه داده‌ی اصلی ارائه می‌دهد. از این‌رو، عملکرد مدل رگرسیون لجستیک را محدود می‌کند. هدف اصلی این مقاله تعیین راه‌های بهبود خوشه‌بندی کا-میانه و نتیجه‌ی دقت رگرسیون لجستیک است. از این‌رو، الگوریتم پیشنهادی شامل تکنیک‌های تحلیل مؤلفه‌های اصلی، کا-میانه و مدل رگرسیون لجستیک است.

یافته‌ها: نتایج به‌دست‌آمده از این مطالعه نشان می‌دهد که توانایی به‌دست آوردن نتیجه دقت خوشه‌بندی کا-میانه بسیار بالاتر از آن چیزی است که سایر محققان در مطالعات مشابه به‌دست آورده‌اند. همچنین در مقایسه با نتایج به‌دست‌آمده از سایر الگوریتم‌ها، مدل رگرسیون لجستیک در سطح بهبود یافته‌ای در پیش‌بینی شروع دیابت اجرا شد. مزیت واقعی دیگر این است که الگوریتم پیشنهادی توانست با موفقیت یک مجموعه داده‌ی جدید را مدل کند.

نتیجه‌گیری: به‌طور کلی، رویکرد پیشنهادی می‌تواند به شکل تأثیرگذاری در پیش‌بینی و تشخیص زودهنگام دیابت استفاده شود.

واژگان کلیدی: دیابت، پیش‌بینی، تحلیل مؤلفه‌های اصلی، کا-میانه، رگرسیون لجستیک

۱- گروه مهندسی صنایع، واحد بندرعباس، دانشگاه آزاد اسلامی، بندرعباس، ایران

***نشانی:** بندرعباس، بلوار دانشگاه، دانشگاه آزاد اسلامی واحد بندرعباس، کدپستی: ۷۹۱۵۸۹۳۴۵۷، تلفن: ۰۹۳۹۹۱۸۲۰۱۰، پست الکترونیک: N.rafiei@iau-tnb.ac.ir

تاریخ دریافت: ۱۴۰۱/۰۹/۲۸ **تاریخ درخواست اصلاح:** ۱۴۰۱/۱۱/۰۲ **تاریخ پذیرش:** ۱۴۰۱/۱۲/۰۳

مقدمه

دیابت از جمله بیماری‌های متابولیک و یک اختلال چندعاملی است که با افزایش مزمن قند خون یا هیپرگلیسمی مشخص می‌شود و ناشی از اختلال ترشح یا عمل انسولین و یا هر دوی آنها است. دیابت یکی از چالش‌های بهداشتی دهه‌های اخیر است که بارِ اقتصادی فراوانی را به جامعه تحمیل می‌کند [۱]. دیابت در میان ۱۰ عامل اصلی مرگ‌ومیر در سال ۲۰۱٦ قرار دارد. در سال ۲۰۱٦، دیابت باعث مرگ ۱/٦ میلیون نفر شد که این رقم در سال ۲۰۰۰ کمتر از ۱ میلیون نفر بود. با این رقم دیابت به‌عنوان هفتمین عامل مرگ‌ومیر شد. تعداد افراد مبتلا به دیابت از ۱۰۸ میلیون نفر در سال ۱۹۸۰ به ٤۲۲ میلیون نفر در سال ۲۰۱٤ رسیده و با شیوع جهانی دیابت در میان بزرگسالان بالای ۱۸ سال، از ٤/۷ درصد در سال ۱۹۸۰ به ۸/۵ درصد در سال ۲۰۱٤ افزایش یافته است. انتظار می‌رود تا سال ۲۰٤۰، ٦٤۲ میلیون بزرگسال (از هر ۱۰ بزرگسال ۱ نفر) به دیابت مبتلا شوند. همچنین، ٤٦/۵ درصد از افرادی که به دیابت مبتلا هستند شناسایی نشده‌اند [۲]. برای کاهش تعداد مرگ‌ومیرهای ناشی از دیابت ضروری است روش‌ها و تکنیک‌هایی ابداع شود که به تشخیص زودهنگام دیابت کمک کند، زیرا تعداد زیادی از مرگ‌ومیرها در بیماران دیابتی به‌دلیل تشخیص دیرهنگام هستند.

به‌منظور دستیابی به تکنیک‌های پیشرفته برای تشخیص زودهنگام دیابت، نیاز است که از فناوری اطلاعات پیشرفته استفاده شود و داده‌کاوی زمینه‌ی مناسبی برای این منظور است. داده‌کاوی توانایی استخراج و کشف الگوهای ناشناخته، پنهان، اما جالب از یک مجموعه داده بزرگ را ارائه می‌دهد. این الگوها می‌توانند به تشخیص پزشکی و تصمیم‌گیری کمک کنند [۳].

تکنیک‌ها و الگوریتم‌های مختلفی طراحی شده است که کاربرد آنها استخراج دانش و اطلاعات در زمینه‌ی تشخیص و درمان بیماران دیابتی از مجموعه داده‌های پزشکی است.

تحلیل مؤلفه‌های اصلی (PCA)[1] یک روش ساده و ناپارامتریک برای استخراج اطلاعات وابسته از مجموعه داده‌های درهم‌ریخته است [٤]. هنگامی که یک مجموعه داده ی بزرگ قرار است توسط کاربر به تعداد خوشه‌های (k) که توسط مراکزشان نمایش داده می‌شوند خوشه‌بندی گردد، تکنیک کا-میانه[2] داده‌ها را با به حداقل رساندن تابع مربعات خطا خوشه‌بندی خواهد کرد و اغلب برخی داده‌ها را به‌دلیل داده‌های پرت اشتباه طبقه‌بندی می‌کند و همچنین، پیچیدگی زمانی بیشتر خواهد بود. برای غلبه بر این مسائل، تکنیک تحلیل مؤلفه‌های اصلی به‌منظور کاهش مجموعه داده‌ها به ابعاد پایین‌تر استفاده می‌شود، درحالی‌که این روش از دست رفتن کمترین اطلاعات را تضمین می‌کند و نقطه‌ی مرکزی بهتری را برای خوشه‌بندی ارائه می‌دهد. تکنیک خوشه‌بندی کا-میانه یک مجموعه داده را به گروه‌های مختلفی از اشیاء مشابه تقسیم می‌کند. خوشه‌هایی که بسیار متفاوت از خوشه‌های دیگر هستند به‌عنوان داده‌های پرت در نظر گرفته شده و کنار گذاشته می‌شوند [۵]. رگرسیون لجستیک یک الگوریتم تحلیلی پیش‌بینی بوده و کاربرد آن زمانی مناسب و کاراست که متغیر وابسته‌ی یک مجموعه داده، دودوئی (باینری) باشد. رگرسیون لجستیک در توصیف و تحلیل داده‌ها، به‌منظور تشریح رابطه بین یک متغیر دودوئی وابسته و یک یا چند متغیر مستقل استفاده می‌شود [٦].

این کار تحقیقاتی تحلیل مؤلفه‌های اصلی را برای کاهش ابعاد پیشنهاد می‌کند تا زمانی که الگوریتم کا-میانه به‌کار گرفته می‌شود، به تعریف مراکز اولیه‌ی مناسب برای مجموعه داده‌های ما کمک کند. سپس، کا-میانه برای یافتن داده‌های پرت و خوشه‌بندی داده‌ها به گروه‌های مشابه، همراه با رگرسیون لجستیک به‌عنوان یک طبقه‌بندی کننده برای مجموعه داده‌ها استفاده می‌شود. در این مقاله، بخش ۲ مروری بر کارهای مرتبط انجام شده توسط سایر پژوهشگران در زمینه‌ی پیش‌بینی و تشخیص دیابت را ارائه می‌دهد. بخش ۳ جزئیات روش‌های تجربی را نشان می‌دهد. بخش ٤

[1] Principal Component Analysis
[2] K-means

نتیجه‌ی تجربی را توصیف کرده، درحالی‌که بخش ۵ کار این مقاله را به پایان می‌رساند.

روش‌ها

همان‌طور که گفته شد دیابت یکی از شناخته شده‌ترین بیماری‌های غیر واگیر در جهان است و براساس برآوردها هفتمین علت مرگ‌ومیر است و پیش‌بینی می‌شود که تا سال ۲۰۴۰ میزان دیابت در بزرگسالان در سراسر جهان به ٦٤٢ میلیون نفر برسد. تشخیص زودهنگام دیابت همواره هدف اصلی پژوهشگران و متخصصان پزشکی بوده است. با دسترسی به نوآوری‌های تکنولوژیکی گسترده، مطالعات مشترک نشان داده‌اند که با به‌کارگیری الگوریتم‌ها و مهارت‌های کامپیوتری مانند داده‌کاوی می‌توان به‌منظور تشخیص دیابت به تکنیک‌های کارا، مقرون به‌صرفه و سریع دست یافت.

پژوهشگران بسیاری مدل‌های پیش‌بینی مختلفی را با استفاده از داده‌کاوی برای پیش‌بینی و تشخیص دیابت توسعه داده‌اند. Iyer و همکاران [۷] در مطالعه‌ی خود به‌منظور پیش‌بینی شروع دیابت استفاده از الگوریتم دسته‌بندی بیز ساده[1] را پیشنهاد کردند که نتیجه دقت را ۷۹/٥٦ درصد نشان داد. Jhaldiyal و Mishra [۸] به‌منظور طبقه‌بندی بیماران دیابتی از تحلیل مؤلفه های اصلی و یک ماشین بردار پشتیبان (SVM)[2] استفاده کرد و نتایج تجربی حاصل از این مطالعه نشان داد که با توجه به دقت طبقه‌بندی ۹۳/٦٦ درصد، سطح قبلی را می‌توان بهبود داد. Kadhm و همکاران [۹] با به‌کارگیری الگوریتم کا-نزدیک‌ترین همسایگی (KNN)[3]، استفاده از درخت تصمیم را به‌منظور تخصیص هر نمونه داده به کلاس مناسب خود پیشنهاد کرد که منجر به حذف داده‌های نامطلوب می‌شد. Wu و همکاران [۱۰] برای تشخیص دیابت مدلی را براساس الگوریتم کا-میانه و الگوریتم رگرسیون لجستیک طراحی کردند که مدل به دقت ۹٥/٤۲ درصد دست یافت. Karegowda و همکاران [۱۱] در شناسایی و حذف داده‌های پرت از خوشه‌بندی کا-میانه، در استخراج ویژگی‌های وابسته از الگوریتم ژنتیک و انتخاب ویژگی مبتنی بر همبستگی (CFS)[4]، و در نهایت در طبقه‌بندی بیماران دیابتی از کا-نزدیک‌ترین همسایگی استفاده کردند. Patil و همکاران [۱۲] یک مدل پیش‌بینی ترکیبی را پیشنهاد کرد که خوشه‌بندی کا-میانه را برای مجموعه داده‌های اصلی به‌کار می‌گیرد و الگوریتم طبقه‌بندی درخت تصمیم را در ایجاد مدل طبقه‌بندی کننده استفاده می‌کند که نتیجه دقت طبقه‌بندی ۹۲/۳۸ درصد بود. Farajollahi و همکاران [۱۳] یک رویکرد مبتنی بر درخت تصمیم و ماشین بردار پشتیبان را به‌منظور کاش ابعاد ویژگی‌های استخراج شده پیشنهاد داد که در آن الگوریتم‌های رگرسیون لجستیک و *XGBoost* به‌عنوان طبقه‌بندی کننده به‌کار گرفته شدند و نتیجه دقت ۸۳ درصد بود.

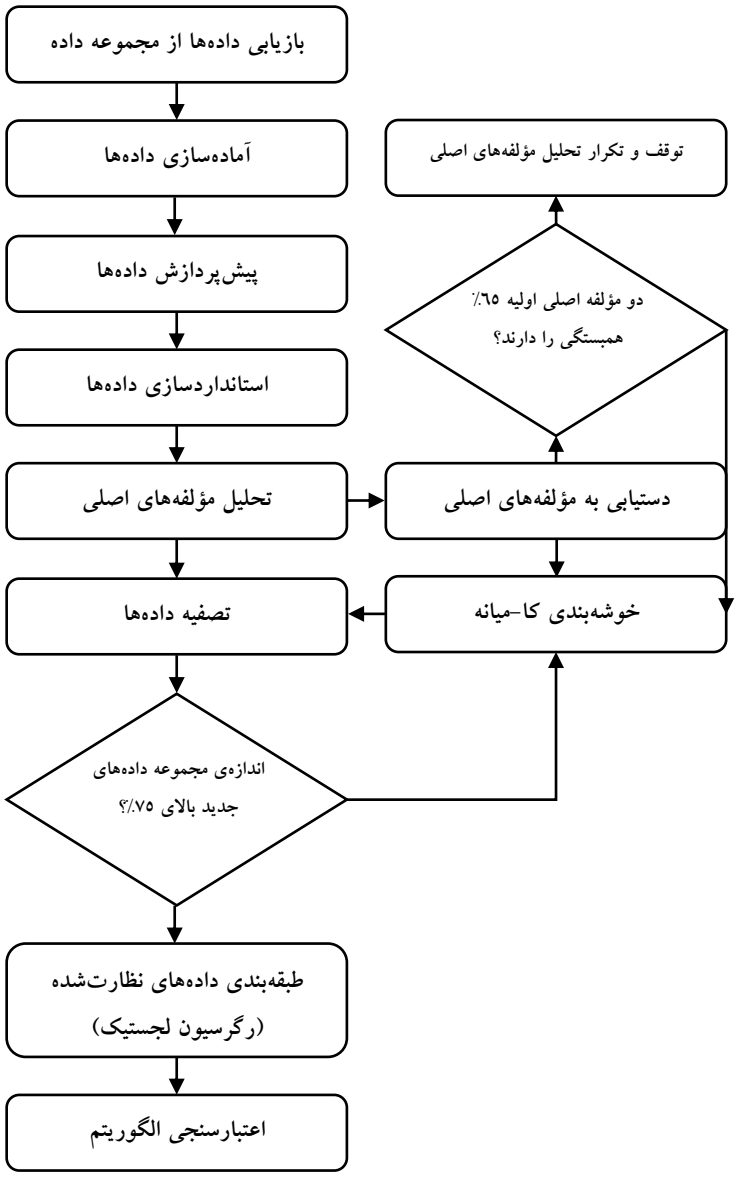
با توجه به بررسی مطالعات پیشین و در نظر گرفتن نیاز به یک الگوریتم پیش‌بینی مؤثر، بهبود الگوریتم پیش‌بینی موجود یکی از وظایف اصلی این تحقیق خواهد بود. درحالی‌که نتایج بزرگی توسط پژوهشگران مختلف به‌دست آمد، مرحله‌ی پیش پردازش داده‌های آنها، مقدار داده‌های در دسترس را به‌منظور پیش‌بینی و طبقه‌بندی نهایی آنها محدود می‌کند. بنابراین، به پیشنهاد یک مدل برای پیش‌پردازش داده‌ها نیاز است که حجم زیادی از داده‌های قابل استفاده را تولید کند و همچنین دقت الگوریتم طبقه‌بندی را افزایش دهد.

یافته‌ها

این بخش شامل مراحل پیش‌رو است: توصیف داده‌ها، تکنیک پیش‌پردازش و الگوریتم طبقه‌بندی. الگوریتم پیشنهادی با ترکیب مزایای استفاده از تکنیک تحلیل مؤلفه‌های اصلی، تکنیک کا-میانه و مدل رگرسیون لجستیک، طراحی شده و به کار گرفته می‌شود. سپس، یک روش جدید با استفاده از تحلیل مؤلفه‌های اصلی به‌منظور تبدیل مجموعه‌ی اولیه ویژگی‌ها پیشنهاد می‌شود و بدین‌ترتیب، مشکل همبستگی حل خواهد شد که این مشکل یافتن رابطه بین داده‌ها را برای الگوریتم طبقه‌بندی دشوار می‌کند [۲]. کاربرد تحلیل مؤلفه‌های اصلی به

[1] Naïve Bayes
[2] Support Vector Machine
[3] K-nearest neighbor
[4] Correlation based feature selection

فیلتر کردن ویژگی‌های غیرمرتبط کمک می‌کند، در نتیجه زمان آموزش و هزینه را پایین آورده و همچنین عملکرد مدل را افزایش می‌دهد [٤]. پس از اجرای تحلیل مؤلفه‌های اصلی، سپس نتیجه برای خوشه‌بندی بدون نظارت با استفاده از کا-میانه تصویب می‌شود، چرا که کا-میانه توانایی اشاره به داده های پرت را دارد [۵]. نتیجه‌ی خوشه کا-میانه تصفیه شده و رگرسیون لجستیک برای ایجاد طبقه‌بندی نظارت شده‌ی ما برای مجموعه داده‌ها به‌کار گرفته می‌شود. فلوچارت مدل پیشنهادی در شکل ۱ نشان داده شده است.



شکل ۱- فلوچارت الگوریتم پیشنهادی

ابزار داده‌کاوی

آناکوندا یک ابزار رایگان و باز زبان برنامه‌نویسی پایتون است که شامل بیش از ۲۵۰ بسته محبوب برای کاربردهای مرتبط با علم داده و یادگیری ماشینی است. با استفاده از این بسته، می‌توان وظایف داده‌کاوی مرتبط را بر روی مجموعه داده‌ها اجرا نموده و الگوریتم پیشنهادی را طراحی و پیاده‌سازی

کرد. با پیش‌پردازش کارامد مجموعه داده‌های اصلی، اجرای تحلیل مؤلفه‌های اصلی و شبیه‌سازی آزمایشات مشابه با دیگر پژوهشگران، نشان داده می‌شود که دقت تشخیص دیابت با استفاده از تکنیک‌های داده‌کاوی می‌تواند بهبود یابد.

توصیف مجموعه داده‌ها

برای این مطالعه مجموعه داده‌های دیابت به‌دست آمده از بایگانی بیمارستان امام علی شهر کرمانشاه استفاده شده است. با استفاده از رابطه‌ی برآورد حجم نمونه‌ی کوکران، مجموعه داده‌ها شامل ۳۸٤ نمونه‌ی بیماران زن است که از نظر دیابت مورد بررسی قرار گرفتند. مجموعه داده‌ها در مجموع دارای ۸ شاخصه با یک کلاس هدف هستند که شاخصه‌ها نشان‌دهنده‌ی معیارهای تشخیص پزشکی بوده و کلاس هدف وضعیت هر فرد آزمایش شده را نشان می‌دهد. در مجموعه داده‌ها به‌طور کلی ٦۷ نمونه آزمایش شده مثبت و ۱۲۵ نمونه‌ی آزمایش شده منفی وجود دارد. شاخصه‌ها در این مجموعه داده‌ها شامل موارد زیر هستند:

- تعداد دفعات بارداری (Preg)
- غلظت گلوکز پلاسما در دو ساعت در آزمایش تحمل گلوکز خوراکی (Plas)
- فشار خون دیاستولیک (Pres)
- ضخامت چین‌های پوستی سه سر بازو (Skin)
- انسولین سرم دو ساعته (Insu)
- نمایه‌ی توده‌ی بدنی (BMI)
- داشتن سابقه‌ی دیابت (Pedi)
- سن (Age)
- متغیر هدف (Diag)

پیش‌پردازش داده‌ها

پایگاه‌های داده‌های جهان واقعی امروزی به‌دلیل اندازه‌های داده‌های عموماً بزرگ و منشاء احتمالی آنها از منابع چندگانه و ناهمگون، به‌شدت در معرض داده‌های نویزی، گمشده و ناسازگار هستند. در فرایند داده‌کاوی به‌منظور پیش‌بینی و تشخیص دیابت، کیفیت داده‌ها عامل مهمی است زیرا داده های با کیفیت پایین ممکن است منجر به نتایج نامناسب و پیش‌بینی با دقت پایین شود [١٤]. در این مطالعه به‌منظور ایجاد مجموعه‌ی داده اصلی مؤثرتر و کاربردی‌تر برای پیش بینی دیابت، چندین تکنیک پیش‌پردازش مختلف با استفاده از بسته‌های مختلف پیشنهادی در محیط توسعه‌ی یکپارچه آناکوندا به‌کار گرفته شد.

در ابتدا شاخصه‌های مختلف مورد بررسی دقیق‌تر واقع شدند و با مشورت با یک متخصص تغذیه‌ی حرفه‌ای ارتباط پزشکی هر شاخصه با پیش‌بینی و تشخیص دیابت تجزیه‌و تحلیل شد. مشخص شد که تعداد دفعات باردار شدن کمترین اهمیت را برای جهت‌گیری تحقیق فعلی دارد. سپس تصمیم گرفته شد تکنیک مشابه استفاده شده توسط Wu و همکاران [۱۰] با تبدیل این شاخصه‌ی عددی به شاخصه‌ی اسمی با ارزش‌های ۰ و ۱ به‌کار گرفته شود که عدد ۱ نشان دهنده‌ی بارداری بیمار در گذشته بوده و عدد ۰ نشان می‌دهد که بیمار هرگز باردار نبوده است. این کار باعث می‌شود پیچیدگی تجزیه‌وتحلیل مجموعه داده‌ها کاهش یابد. آمارهای مجموعه داده‌های اصلی و پیش‌پردازش شده در جدول ۱ قابل مشاهده است. در ادامه تجزیه‌وتحلیل آماری مجموعه داده‌های ما حضور مقادیر گمشده را پیشنهاد کرد. جدول ۲ نتیجه آماری مجموعه داده‌های ما را نشان می‌دهد.

جدول ۱- آمارهای مجموعه داده‌های اصلی و پیش‌پردازش شده

آمار	مجموعه داده	Preg	Plas	Pres	Skin	Insu	BMI	Pedi	Age
مقدار	اصلی	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰
	پیش‌پردازش	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰	۳۸٤/۰۰
میانگین	اصلی	۰/۷٦	۱۱۰/۸۹	۵۹/۱۱	۱۰/٦٤	٦۹/۸۰	۲۲/۰۰	۰/۳۷	۳۲/۲٤
	پیش‌پردازش	۰/۷٦	۲۱/٦۹	٦۲/٤۱	۱۹/۱۵	۱٤۵/۵۵	۲۲/٤٦	۰/۵۷	۳۲/۲٤
انحراف معیار	اصلی	۰/۲۵	۲۱/۹۷	۱۸/۹۵	۱٤/۹۵	۱۰۵/۲٤	۷/۸۸	۰/۲۳	۱۰/۷٦
	پیش‌پردازش	۰/۵	۲۰/٤٤	۱۱/۱۰	۷/۷۹	۷۵/۰۲	٦/۸۸	۰/۲۳	۱۰/۷٦
حداقل	اصلی	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۷	۱۵/۰۰
	پیش‌پردازش	۰/۰۸	۳٤/۰	۱٤/۰۰	۷/۰۰	۱٤/۰۰	۱۵/۲۰	۰/۰۷	۱۵/۰۰
۲۵٪	اصلی	۱/۰۰	۸۹/۰۰	۵۲/۰۰	۰/۰۰	۰/۰۰	۱۷/۳۰	۰/۱٤	۲۰/۰۰
	پیش‌پردازش	۱/۰۰	۸۹/۷۵	۵٤/۰۰	۱۵/۰۰	۱۱۱/۵۰	۱۷/۵۰	۰/۱٤	۲۰/۰۰
۵۰٪	اصلی	۱/۰۰	۱۰۷/۰۰	٦۲/۰۰	۱۳/۰۰	۲۰/۵۰	۲۲/۰۰	۰/۲۷	۲۵/۰۰
	پیش‌پردازش	۱/۰۰	۱۰۷/۰۰	٦۲/۲۰	۱۹/۲۵	۱٤۵/۵۵	۲۲/٤۰	۰/۲۷	۲۵/۰۰
۷۵٪	اصلی	۱/۰۰	۱۳۰/۲۵	۷۰/۰۰	۲۲/۰۰	۱۱۷/۲۵	۲٦/٦۰	۰/۵۳	۳٦/۰۰
	پیش‌پردازش	۱/۰۰	۱۳۰/۲۵	۷۰/۰۰	۲۲/۰۰	۱٤۵/۵۵	۲٦/٦۰	۰/۵۳	۳٦/۰۰
حداکثر	اصلی	۱/۰۰	۱۸۹/۰۰	۱۱۲/۰۰	۸۹/۰۰	۸۳٦/۰۰	۵۷/۱۰	۲/۳۲	۷۵/۰۰
	پیش‌پردازش	۱/۰۰	۱۸۹/۰۰	۱۱۲/۰۰	۸۹/۰۰	۸۳٦/۰۰	۵۷/۱۰	۲/۳۲	۷۵/۰۰

جدول ۲- ماتریس درهم‌ریختگی

کلاس	۱ (مثبت)	۰ (منفی)
منفی پیش‌بینی شده	۲	۱۹۵
مثبت پیش‌بینی شده	۱۰۳	۷

از نتایج آماری مشاهده می‌شود که غلظت گلوکز پلاسما، فشار خون دیاستولیک، ضخامت چین‌های پوستی، انسولین سرم دو ساعته، نمایه‌ی توده‌ی بدنی حداقل مقدار ۰ دارند. دانش پزشکی بیان می‌دارد که چنین شاخصه‌هایی (نتایج پزشکی) نمی‌توانند ۰ باشند بنابراین مشخص می‌شود که مجموعه داده‌ها شامل یک مقدار گمشده هستند که در صورت عدم رسیدگی به آن کیفیت نتایج و دقت مدل را مختل می‌کند. به‌منظور رسیدگی به مقادیر گمشده در مجموعه داده‌ها روش‌های مختلفی پیشنهاد شده است. در این مطالعه، مقادیر گمشده با میانگین مشخصه جایگزین می‌شود.

به‌عنوان بخشی از پیش‌پردازش داده‌ها، با اجرای بی مقیاس‌سازی مجموعه داده‌ها، مقادیر داده‌های اصلی به گونه‌ای مقیاس‌بندی می‌شوند که در داخل یک محدوده‌ی مشخص کوچک از مقادیر ۰ تا ۱ قرار گیرند. این امر موجب بهبود سرعت و کاهش پیچیدگی زمان اجرا می‌گردد. با استفاده از امتیاز Z، مجموعه داده V بی‌مقایس می‌شود تا یک مجموعه جدید از مقادیر بی‌مقیاس V' با رابطه‌ی (۱) حاصل می‌شود:

$$V' = \frac{V - Y}{Z} \qquad (۱)$$

که در این معادله V' مقدار بی‌مقیاس شده جدید، V مقدار قبلی، Y مقدار میانگین و Z انحراف استاندارد را نشان می دهد.

طراحی الگوریتم پیشنهادی

الگوریتم پیشنهادی از سه مرحله‌ی فرعی ایجاد شده است. در مرحله‌ی اول طراحی، با استفاده از تکنیک تحلیل مؤلفه های اصلی کاهش ابعاد بر روی مجموعه داده‌های از پیش پردازش شده انجام می‌گیرد. سپس، مؤلفه‌ی اصلی انتخاب شده با استفاده از تکنیک کا-میانه خوشه‌بندی می‌شود تا به داده‌های پرت رسیدگی شده و هرگونه طبقه‌بندی ناصحیح حذف شود. در نهایت، داده‌های به درستی خوشه‌بندی و طبقه‌بندی شده، به‌منظور طبقه‌بندی نظارت شده‌ی ما به‌عنوان ورودی به مدل رگرسیون لجستیک مورد استفاده قرار می گیرند.

تحلیل مؤلفه‌های اصلی

در طول تجزیه‌وتحلیل داده‌ها، اغلب یافتن تمامی روابط شاخصه‌ها بسیار دشوار است. تکنیک تحلیل مؤلفه‌های اصلی به حجم بزرگی از اطلاعات محصور شده در داده‌های همبسته‌ی اولیه اجازه می‌دهد تا به مجموعه‌ای از عناصر متعامد جدید تبدیل شوند. بنابراین، کشف روابط پنهان، افزایش بصری‌سازی داده‌ها، تشخیص داده‌های پرت و طبقه بندی در داخل ابعاد جدید تعریف شده امکان‌پذیر می‌شود. هنگامی که نیاز به یادگیری بدون نظارت بر روی چنین مجموعه داده‌هایی باشد، کاربرد تحلیل مؤلفه‌های اصلی بر روی مجموعه‌ی داده می‌تواند کمک بزرگی باشد، زیرا به مقداردهی اولیه‌ی مؤثر مراکز به‌منظور خوشه‌بندی کمک خواهد کرد.

از آنجاکه تحلیل مؤلفه‌های اصلی یک زیرفضای ویژگی ایجاد می‌کند که واریانس را در امتداد محورها به حداکثر می‌رساند، ابتدا مجموعه داده‌ها را در یک مقیاس واحد یعنی میانگین برابر با ۰ و واریانس برابر با ۱، استاندارد می‌کنیم تا نتیجه‌ی تحلیل مؤلفه‌های اصلی بهبود یابد. این اقدام لازمه‌ی عملکرد بهینه‌ی بسیاری از الگوریتم‌های یادگیری ماشینی است [۱۵].

هدف ما در اینجا تبدیل مجموعه‌ی داده X از بعد p به مجموعه‌ی نمونه جدید Y از بعد کوچکتر L $(L < p)$ بوده که در اینجا Y مؤلفه‌ی اصلی X است، یعنی:

$$Y = PC(X) \qquad (۲)$$

از این‌رو، به‌صورت زیر عمل می‌شود [۱۵]:

- سازمان‌دهی مجموعه داده‌ها:

X دارای مجموعه‌ای از بردارهای n $(x_1, x_2, ..., x_n)$ است که هر عنصر x_i یک نمونه‌ای از مجموعه‌ی داده است.

- یافتن میانگین با استفاده از رابطه‌ی (۳):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (۳)$$

- محاسبه‌ی واریانس با استفاده از رابطه‌ی (٤):

$$\overline{X} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{(n-1)} \qquad (٤)$$

- محاسبه‌ی کواریانس با استفاده از رابطه‌ی (۵):

$$X^{n \times n} = (x_{ij}, x_{ij} = \mathrm{cov}(Dim_i, Dim_j)) \qquad (۵)$$

که $X^{n \times n}$ ماتریس داده با n سطر و n ستون است و Dim_i بیانگر i اُمین بعد است.

- محاسبه‌ی مقادیر ویژه و بردارهای ویژه:

هسته‌ی تکنیک تحلیل مؤلفه‌های اصلی، بردار ویژه و مقادیر ویژه ماتریس کواریانس است. بردارهای ویژه جهت فضای ویژگی جدید را تعیین می‌کنند درحالی‌که مقادیر ویژه تعیین کننده اهمیت و بزرگی هستند.

اگر A یک ماتریس مربع n بعدی باشد، سپس یک بردار غیر صفر x عضوی از R^n بردار ویژه A یا عملگر ماتریس T_A نامیده می‌شود، اگر Ax یک مضرب اسکالر x باشد. یعنی:

$$Ax = \lambda x \qquad (٦)$$

اسکالر λ یک مقدار ویژه A نامیده شده و به x بردار ویژه مربوط به λ گفته می‌شود. از آنجاکه بردارهای ویژه مربوط به مقدار ویژه ماتریس A بردارهای غیر صفر هستند رابطه (۷) را برقرار و برآورده می‌کنند.

$$(\lambda I - A)x = 0 \qquad (۷)$$

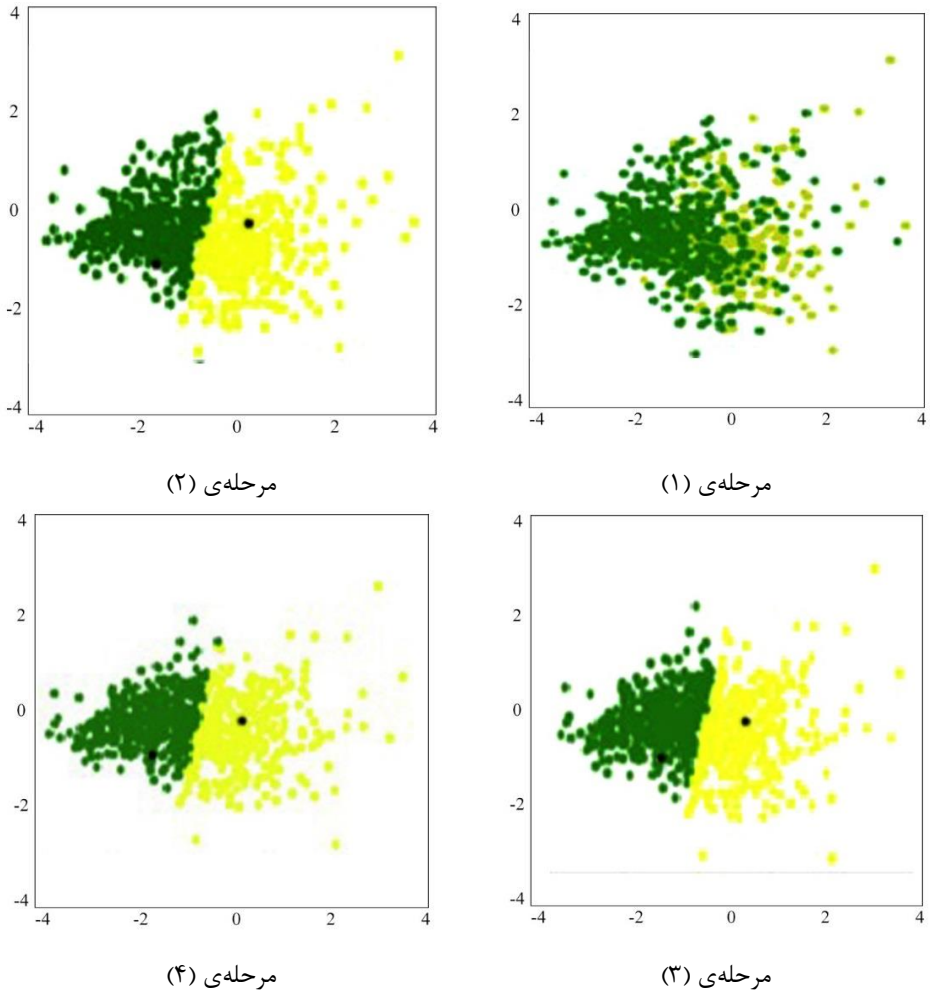
مجموعه E را به‌عنوان فضای ویژه، برای تمامی بردارهای x تعریف می‌کنیم که رابطه‌ی (۷) را برآورده می‌کنند.

$$E = \{x : (\lambda I - A)x\} = 0 \qquad (۸)$$

- هنگامی‌که فضای ویژه از ماتریس کواریانس پیدا شد، گام بعدی این است که بردارهای ویژه را براساس مقدار ویژه از بیشترین به کمترین مقدار مرتب کنیم. این‌کار مؤلفه‌های با اهمیت کمتر را حذف کرده و مؤلفه‌های اصلی باقی می‌مانند که تقریب خوبی را از داده‌ای اصلی فراهم می‌کنند. مؤلفه‌های اصلی جدید در مرحله‌ی بعدی طراحی الگوریتم، به‌عنوان ورودی برای تکنیک خوشه‌بندی کا-میانه استفاده خواهند شد.

خوشه‌بندی کا-میانه

کا-میانه یکی از ساده‌ترین و کاراترین الگوریتم‌های طبقه بندی بدون نظارت و یک تکنیک خوشه‌بندی مبتنی بر تقسیم‌بندی شناخته شده است که تلاش می‌کند تعداد مشخصی از خوشه‌ها که توسط مراکزشان نشان داده شده را پیدا کند. همچنین، این روش یک الگوریتم خوشه‌بندی مبتنی بر فاصله‌ی معمولی است که در آن از فاصله به‌عنوان یک معیار تشابه استفاده می‌شود، یعنی فاصله‌ی کمتر بین اشیاء شباهت بیشتری را نشان می‌دهد. با اعمال مراحل زیر، شکل ۲ روش گرافیکی برای خوشه‌بندی کا-میانه را نشان می‌دهد [۲]:



مرحله‌ی (۱) مرحله‌ی (۲)

مرحله‌ی (۳) مرحله‌ی (۴)

شکل ۲- رویکرد خوشه‌بندی کا-میانه

- مرحله‌ی (۱) در شکل ۲ کل مجموعه‌ی داده را نمایش می‌دهد. k = ۲ مقداردهی اولیه می‌شود چرا که متغیر هدف شامل دو خروجی ممکن مثبت و منفی است.

- در مرحله‌ی (۲) باید برای هر داده‌ی ورودی، مرکز خوشه‌ای که نزدیک‌ترین به داده است تعیین شود که این کار با استفاده از رابطه‌ی (۹) صورت می‌گیرد.'

$$S_t^{(t)} = \{x_p : \| x_p - m_i^{(2)} \|^2 \le \| x_p - m_j^{(2)} \|^2 \ \forall_j, 1 \le j \le k\} \quad (۹)$$

- در مرحله‌ی (۳) با محاسبه‌ی مجدد میانگین هر داده ورودی اختصاص داده شده به خوشه با به‌کارگیری رابطه ی (۱۰)، مراکز خوشه به‌روزرسانی می‌شود.

$$m_i^{(t+1)} = \frac{1}{\left|S_t^{(t)}\right|} \sum_{s_j \in S_t^{(t)}} x_j \quad (۱۰)$$

- مرحله‌ی (٤) نشان می‌دهد که به‌منظور توقف خوشه کا-میانه، در مراحل (۲) و (۳) حلقه‌ای ایجاد می‌شود و تا زمانی که یک هم‌گرایی در مقدار میانگین خوشه‌ها ایجاد شود حلقه به‌کار خود ادامه می‌دهد.

سپس، نتیجه‌ی خوشه کا-میانه با حذف داده‌های خوشه‌بندی شده‌ی ناصحیح تصفیه شده و برای پیدا کردن مجموعه داده ی جدید به‌منظور طبقه‌بندی با استفاده از رابطه‌ی (۱۱)، تصمیم‌گیری می‌شود. اگر اندازه‌ی مجموعه داده‌ی جدید بالای ۷۵ درصد بود، طبقه‌بندی داده‌های نظارت شده ادامه پیدا می‌کند، در غیر این‌صورت مرحله‌ی کا-میانه تا تعیین یک اندازه‌ی مناسب تکرار می‌گردد.

(۱۱) مجموع کل/داده‌های چپ = اندازه‌ی جدید

پس از تصفیه داده های خوشه بندی شده، ۳۱۷ بیمار خوشه بندی شده بطور صحیح بدست آمد که به عنوان ورودی برای آموزش الگوریتم رگرسیون لجستیک استفاده می شود.

مدل رگرسیون لجستیک

در بسیاری از حوزه‌ها مانند علوم زیستی کاربرد مدل رگرسیون لجستیک اهمیت ویژه‌ای پیدا کرده است. رگرسیون لجستیک زمانی استفاده می‌شود که هدف طبقه‌بندی اقلام داده‌ها به دسته‌ها است. معمولاً در رگرسیون لجستکی متغیر هدف باینری است، به این معنی که فقط شامل داده‌هایی است که به‌صورت ۰ یا ۱ طبقه‌بندی شده که در این مطالعه به بیمار مثبت و منفی برای دیابت اشاره دارد. هدف مدل رگرسیون لجستیک در اینجا پیدا کردن بهترین تناسب است به‌طوری‌که از نظر تشخیصی برای توصیف رابطه‌ی بین متغیر هدف و متغیرهای پیش‌بینی کننده‌ی منطقی عمل کند. مدل رگرسیون لجستیک براساس مدل رگرسیون خطی ارائه شده در رابطه‌ی (۱۲) زیر است [۱٦]:

$$y = h_\theta(x) = \theta^T x \quad (۱۲)$$

رابطه‌ی (۱۲) برای پیش‌بینی مقادیر باینری $(y^i \in \{0,1\})$ بسیار نارکارا خواهد بود، بنابراین، به‌منظور پیش‌بینی احتمال اینکه یک بیمار معین با شاخصه‌های مشخص، متعلق به کلاس ۱ یا مثبت باشد، در مقابل پیش‌بینی احتمال اینکه این بیمار متعلق به کلاس ۰ یا منفی باشد، تابعی در رابطه‌ی (۱۳) معرفی می‌شود [٦].

$$P(y=1|x) = h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)} = \sigma(\theta^T x) \quad (۱۳)$$
$$P(y=0|x) = 1 - P(y=1|x) = 1 - h_\theta(x)$$

با استفاده از رابطه‌ی (۱٤) که به تابع سیگماوار معروف است می‌توان مقدار $\theta^T x$ را در محدوده [۰, ۱] نگه داشت. سپس، جستجو برای مقدار θ آغاز می‌شود به‌طوری‌که اگر x متعلق به کلاس ۱ باشد احتمال $P(y=1|x) = h_\theta(x)$ بزرگ است و اگر x متعلق به کلاس ۰ باشد این احتمال کوچک است (یعنی $P(y=0|x)$ بزرگ است).

$$\sigma(t) = \frac{1}{(1+e^{-t})} \quad (۱٤)$$

با پایان یافتن مدل‌سازی و پیاده‌سازی موفقیت‌آمیز مدل رگرسیون لجستیک، در بخش بعدی خروجی‌ها و نتایج مورد بحث قرار می‌گیرد.

یافته‌ها

یک نتیجه‌ی اصلی به‌دست آمده در استفاده از تکنیک تحلیل مؤلفه‌های اصلی این است که این فرایند ایراد و مشکل داشتن ویژگی‌های زائد را که به خوشه‌بندی کمکی نمی‌کنند به حداقل مقدار ممکن می‌رساند. از آنجاکه کاهش تعداد متغیرها در مجموعه داده‌های اصلی به مدیریت داده‌های نویزی و پرت کمک می‌کند، بنابراین در این مطالعه تکنیک تحلیل مؤلفه‌های اصلی نتیجه تکنیک کا-میانه را بهبود می بخشد. مزیت اصلی تحلیل مؤلفه‌های اصلی فشرده‌سازی داده ها در زمان پیدا کردن مؤلفه‌های اصلی داده‌ها است، یعنی با کاهش تعداد ابعاد بدون از دست دادن اطلاعات زیادی، به یک فرایند ضروری برای تعیین تعداد خوشه‌ها و ارائه‌ی یک چارچوب آماری برای مدل‌سازی ساختار خوشه تبدیل می شود.

از طرفی، کارایی و دقت هر مدل پیش‌بینی و تشخیص از اهمیت بالایی برخوردار است و باید قبل از به‌کارگیری چنین مدلی برای پیاده‌سازی اطمینان حاصل شود. در این مطالعه خروجی مدل با استفاده از معیارهای ارزیابی مختلف تجزیه‌و تحلیل و ارزیابی شد و نتیجه در جدول ۳ نشان داده شده است

جدول ۳- خلاصه عملکرد مدل

معیارهای ارزیابی	مقدار
نمونه‌های به درستی طبقه‌بندی شده	۱۵۰
نمونه‌های به اشتباه طبقه‌بندی شده	٤
تعداد کل نمونه‌ها	۱۵٤
مقدار ROC	۰/۹٦۷
میانگین مربعات خطا	۰/۰۲٦
آماره‌ی کاپا	۰/۹٤۲

یکی از روش‌های بررسی و ارزیابی عملکرد دسته‌بندی دودوئی، نمودار مشخصه عملکرد یا به اختصار منحنی ROC[1] است. ROC یک نمودار گرافیکی است که عملکرد یک طبقه‌بندی کننده را نشان می‌دهد و این اجازه را می‌دهد تا عملکرد مدل را در تمامی آستانه‌های ممکن دنبال کنیم [۱۷]. در آزمایش ما، مقدار ROC برابر ۰/۹٦۷ بود. همچنین، آماره‌ی کاپا، الگوریتم طبقه‌بندی ما را با یک الگوریتم طبقه‌بندی تصادفی مقایسه می‌کند و به ما می‌گویند که الگوریتم طقه‌بندی ما، چه میزان از یک الگوریتم تصادفی بهتر عمل کرده است. آماره‌ی کاپا معیاری است که دقت مشاهده شده را با دقت مورد انتظار مقایسه می‌کند و معمولاً مقداری بین ۰ تا ۱ را دارد [۱۷]. در این آزمایش مقدار آماره ی کاپا ۰/۹٤۲ بود. با توجه به مقادیر معیارهای عملکرد، نتایج حاکی از آن است که عمکرد مدل عالی است.

در ابتدا به‌منظور تعیین عملکرد مدل، از روش اعتبارسنجی متقابل k-fold استفاده شد که به ما این اجازه را می‌دهد مشخص کنیم که مدل با داده‌های جدید چه میزان عملکرد خوبی دارد [۱۷]. انتخاب ما از اعتبارسنجی متقابل 10-fold به این معناست که مجموعه داده‌ها به ۱۰ زیرمجموعه تقسیم می‌شوند. در هر آزمایش، یک زیرمجموعه به‌عنوان مجموعه ی آزمون و نه زیرمجموعه‌ی دیگر مجموعه‌ی آموزشی را تشکیل می‌دهند. سپس خطای میانگین در تمام ۱۰ آزمایش محاسبه شد تا عملکرد کل مدل به‌دست آید. این روش به حل دو موضوع کمک می‌کند. اول اینکه مشکل جهت‌گیری

[1] Receiver Operating Characteristic

را کاهش می‌دهد زیرا تقریباً تمامی داده‌ها برای برازش استفاده می‌شوند و دوم اینکه مشکل واریانس تا حد زیادی کاهش می‌یابد. ماتریس درهم‌ریختگی روشی محبوب برای ارائه‌ی خلاصه‌ای از یافته‌های پیش‌بینی است و نتایج شاخص‌های «مثبت واقعی»، «منفی واقعی»، «مثبت کاذب» و «منفی کاذب» را ارائه می‌دهد. جدول ۲ ماتریس درهم ریختگی مدل ما را نشان می‌دهد.

به‌منظور ارزیابی بیشتر عملکرد مدل، مجموعه داده‌های خود را با در نظر گرفتن تغییراتی با چهار الگوریتم مختلف مدل کردیم. تغییرات شامل «مجموعه داده‌های اصلی»، «داده‌های پردازش‌شده PCA»، «داده‌های پردازش‌شده PCA و K-means» و «داده‌های K-means» بودند. نتایج در جدول ٤ مشخص است.

جدول ۴- مقایسه‌ی مدل با الگوریتم‌های مختلف

الگوریتم	دقت مجموعه داده‌ها			
	اصلی	پردازش شده با PCA	خوشه‌بندی شده با K-means	پردازش‌شده با PCA+K-means
الگوریتم رگرسیون لجستیک	۰/۷۵	۰/٦۹	۰/۸۰	۰/۹۵
الگوریتم کا-نزدیک‌ترین همسایگی	۰/۷۳	۰/٦۷	۰/۹۱	۰/۹٤
الگوریتم *XGBoost*	۰/۷٤	۰/٦٤	۰/۹۳	۰/۹۱
الگوریتم ماشین بردار پشتیبان	۰/۷٤	۰/۷۰	۰/۸۱	۰/۹۰
الگوریتم دسته‌بندی بیز ساده	۰/۷۲	۰/۷۱	۰/۸٤	۰/۸۸

نتایج جدول بالا نشان می‌دهد که تکنیک یکپارچه‌ی تحلیل مؤلفه‌های اصلی و کا-میانه، به‌جزء عملکرد الگوریتم *XGBoost*، دقت عملکرد الگوریتم‌های مختلفی را که مجموعه داده‌ها را با آنها مدل کردیم بهبود می‌دهد. نتایج نشان می‌دهد دقت ۹۳ درصد در زمان یکپارچه‌شدن کا-میانه با الگوریتم *XGBoost*، به دقت ۹۱ درصد در زمان یکپارچه شدن تکنیک تحلیل مؤلفه‌های اصلی و کا-میانه پیشنهادی ما با الگوریتم *XGBoost* کاهش می‌یابد. علاوه بر این، کا-میانه یک روش خوب به‌منظور بهبود دقت هر یک از الگوریتم‌ها است، درحالی‌که استفاده تحلیل مؤلفه‌های اصلی به تنهایی نتیجه دقت را کاهش می‌دهد.

بحث

نتایج تجربی نشان داد که تحلیل مؤلفه‌های اصلی دقت الگوریتم خوشه‌بندی کا-میانه را افزایش می‌دهد و از نتایج جدول ٦ قابل مشاهده است که در مقایسه با دیگر مطالعات، در این مطالعه ۳۱۷ مجموعه داده‌ی خوشه‌بندی شده به‌طور صحیح به‌دست آمد. در میان آنها، نزدیک‌ترین نتیجه به مطالعه ما مربوط به Kadhm و همکاران [۹] است که دقت بالایی را از اندازه نمونه ۵۷۰ به‌دست آمده از خوشه‌بندی کا-میانه را دارد. با اشاره به نتایج تجربی به‌دست آمده، می‌توان به وضوح نشان داد که تکنیک تحلیل مؤلفه‌های اصلی و کا-میانه پیشنهادی، دقت الگوریتم طبقه‌بندی رگرسیون لجستیک برای مجموعه داده‌ها بهبود می‌بخشید.

جدول ۵- مقایسه‌ی نتیجه‌ی خوشه‌بندی کا-میانه

نویسنده (سال)	روش‌شناسی	داده‌های به‌درستی خوشه‌بندی شده	درصد دقت
الگوریتم پیشنهادی این مطالعه	PCA+K-means	۳۱۷	۸۲/۵۵
Kadhm و همکاران [۹]	K-means	۵۷۰	۷٤/۲۱
Patil و همکاران [۱۲]	K-means	٤۳۳	٥٦/۳۸
Karegowda و همکاران [۱۱]	Cascaded K-means	۲۹۹	۳۸/۹۳

یکی از نگرانی‌های اصلی توسعه‌ی الگوریتم‌های یادگیری ماشینی برای کاربردهای پزشکی، قابلیت اطمینان چنین الگوریتمی در صورت اجرای عملی است. به‌منظور ارزیابی عملکرد الگوریتم این مطالعه، از یک مجموعه داده‌ی دیگر با همکاری بیمارستان امام رضا در شهر کرمانشاه اطلاعاتی از پرونده‌ی بیمارانی که به‌منظور دیابت آزمایش شده، استخراج شد. یک مجموعه داده را از ۵۰۰ ثبت تصادفی تشکیل دادیم، درحالی‌که تنها آن دست از ثبت‌هایی در نظر گرفته شد که هیچ مقدار گمشده‌ای برای ویژگی‌های مورد نیاز ما نداشتند. مجموعه داده شامل ۱۱ شاخصه به نام‌های سن، نمایه‌ی توده ی بدنی، سابقه‌ی خانوادگی، تکرر ادرار، خستگی، کاهش وزن، الگوی غذا خوردن، ورزش منظم، فشار خون، قند خون ناشتا و تست تحمل گلوکز خوراکی است. متغیر کلاس دارای توزیعی از ۲۵۳ مورد منفی و ۲٤۷ مورد مثبت است.

مراحل پیش‌پردازشی که برای مجموعه داده‌های قبلی به‌کار گرفته شد بر روی همین مجموعه داده نیز اجرا شد و سپس خروجی برای ارزیابی عملکرد مدل پیشنهادی به‌کار گرفته شد. به‌منظور نمایش کاربرد بیشتر مدل، نتیجه‌ی آن را با استفاده از مجموعه داده‌ی جدید با الگوریتم‌های دیگر مقایسه شد که نتیجه در جدول ٦ قابل مشاهده است. نتایج نشان داد که دقت عملکرد مدل ۸۷ درصد بود که تأکید می کند که رویکرد پیشنهادی حتی زمانی که با مجموعه داده‌های دیگری استفاده می‌شود می‌تواند قابل اعتماد باشد.

جدول ۶- الگوریتم‌های مختلف به‌منظور انجام مقایسات با مجموعه داده‌ی جدید

الگوریتم	دقت مجموعه داده‌ها			
	اصلی	پردازش‌شده با PCA	خوشه‌بندی شده با K-means	پردازش‌شده با PCA + K-means
الگوریتم رگرسیون لجستیک	۰/٤٥	۰/٤٦	۰/۷۳	۰/۸۷
الگوریتم کا-نزدیک‌ترین همسایگی	۰/٥۱	۰/٤٦	۰/٦٥	۰/۷٦
الگوریتم *XGBoost*	۰/٤۹	۰/٤۷	۰/۷۲	۰/۸۳
الگوریتم ماشین بردار پشتیبان	۰/٤۸	۰/٤٥	۰/٤۳	۰/٥٦
الگوریتم دسته‌بندی بیز ساده	۰/٥۲	۰/٥۰	۰/٦۲	۰/۸۰

نتیجه‌گیری

هدف از این مقاله طراحی یک مدل کارا بود که دیابت را پیش‌بینی کند. پس از بررسی دقیق سایر کارهای منتشر شده، یک مدل جدید پیشنهاد شد که شامل استفاده از تکنیک تحلیل مؤلفه‌های اصلی برای کاهش ابعاد، تکنیک کا-میانه برای خوشه‌بندی و مدل رگرسیون لجستیک برای طبقه‌بندی بود. با قصد بهبود نتایج تکنیک کا-میانه استفاده شده توسط سایر پژوهشگران، در ابتدا تکنیک تحلیل مؤلفه‌های اصلی در مجموعه داده‌ها به‌کار گرفته شد. اگرچه تحلیل مؤلفه‌های

اصلی یک تکنیک شناخته‌شده است، اما به کارایی آن در بهبود خوشه‌بندی کا-میانه و به لحاظ مدل طبقه‌بندی رگرسیون لجستیک توجه کافی نشده است. با توجه به آزمایش ما نشان داده شد که طراحی یک مدل رگرسیون لجستیک بهبودیافته برای پیش‌بینی دیابت با استفاده از یکپارچه‌سازی تکنیک‌های تحلیل مؤلفه‌های اصلی و کا-میانه امکان‌پذیر است. دستاوردهای به‌دست آمده از این مطالعه نشان می‌دهد توانایی به‌دست آوردن نتیجه دقت خوشه‌بندی کا-میانه، بسیار بالاتر از آنچه است که سایر پژوهشگران در مطالعات مشابه به‌دست آورده‌اند. همچنین، در مقایسه با نتایج به‌دست آمده از الگوریتم‌های دیگر، مدل رگرسیون لجستیک در سطح بهبود یافته‌ای در پیش‌بینی شروع دیابت اجرا شد. مزیت واقعی دیگر این است که مدل ما توانایی مدل‌سازی موفقیت‌آمیز یک مجموعه داده‌ی جدید را دارد. در کل، رویکرد پیشنهادی در پیش‌بینی و تشخیص زودهنگام دیابت می‌تواند به‌طور مؤثر مورد استفاده قرار گیرد.

سپاسگزاری

نویسنده مقاله مراتب تقدیر و تشکر خود را از حمایت‌های حوزه تحقیقات و فناوری دانشگاه علوم پزشکی و خدمات درمانی کرمانشاه اعلام می‌دارد.

مآخذ

۱. ایزدی ندا، رحیمی مهرعلی، رضوان مدنی فاطمه، شتابی حمیدرضا، دربندی میترا. بررسی اپیدمیولوژی بیمار دیابت نوع II در مراجعه‌کنندگان به کلینیک دیابت استان کرمانشاه در سال ۹۳-۱۳۹۲: یک گزارش کوتاه. *مجله دانشگاه علوم پزشکی رفسنجان* ۱۳۹۶؛ ۱۶(۱): ۹۰-۸۳.

2. Zhu C, Idemudia CU, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Inform Med Unlocked* 2019; 17: 10079.
3. Jothi N, Abdul Rashid N, Husain W. Data Mining in Healthcare – A Review. *Procedia Comput Sci* 2015; 72: 306-313.
4. Pérez-Montalvo E, Zapata-Velásquez M-E, Benítez-Vázquez LM, Cermeño-González J-M, Alejandro-Miranda J, Martínez-Cabero M-Á, et al. Model of monthly electricity consumption of healthcare buildings based on climatological variables using PCA and linear regression. *Energy Rep* 2022; 8(9): 250-258.

۵. یلوه الهام، نوروزی یعقوب، خطیر اشکان. مروری نظام‌مند بر پژوهش‌های بهبود الگوریتم کا-میانه برای خوشه‌بندی‌بندی داده‌ها. *پژوهش‌نامه پردازش و مدیریت اطلاعات* ۱۴۰۰؛ ۳۷(۲): ۵۵۶-۵۲۷.

6. Chang Y-C, Chang K-H, Lin Y-X. Establishment of Business Loan Default Prediction Model by Integrating Survival Analysis with Logistic Regression. *Sci. Iran* 2022.
7. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of Diabetes Using Classification Mining Techniques. *Int J Data Min Knowl Manag Process* 2015; 5(1): 1-14.
8. Jhaldiyal T, Mishra, PK. Analysis and prediction of diabetes mellitus using PCA, REP and SVM. *Int J Eng Tech Res* 2014; 2(8): 164-166.
9. Kadhm MS, Ghindawi IW, Mhawi DE. An accurate diabetes prediction system based on K-means clustering and proposed classification approach. *Int J Appl Eng Res* 2018; 13(6): 4038-4041.
10. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. *Inf Med Unlocked* 2018; 10: 100-107.
11. Karegowda AG, Jayaram MA, Manjunath AS. Cascading Kmeans Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. *Int J Eng Adv Res Tech* 2012; 1(3): 147-151.
12. Patil BM, Joshi RC, Toshnival D. Hybrid prediction model for type-2 diabetic patients. *Expert Syst Appl* 2010; 37(12): 8102-8108.
13. Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F. and Sayadi, M.J. Diabetes diagnosis using machine learning. *Front Health Inf* 2021; 10(65): 1-5.
14. Moradi M, Modarres M, Sepehri MM. Detecting factors associated with polypharmacy in general

practitioners' prescriptions: A data mining approach. *Sci Iran* 2020.

15. Kong X, Hu C, Duan Z. *Principal component analysis networks and algorithms*. Singapore: Springer; 2017.

۱۶. اسدزاده شروین، رفیعی نوید، نیاکی سید تقی اخوان. طراحی اقتصادی-آماری یک نمودار کنترل به‌منظور پایش مدت‌زمان بقای بیماران. *مهندسی صنایع و مدیریت شریف* ۱۳۹۹؛ ۱-۳۶(۲/۲): ۹۷-۸۷.

17. Marcot BG. Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecol Modell* 2012; 230: 50-62.

Design an Algorithm based on Data Mining to Predict Diabetes

Navid Rafiei[1*]

1. Department of Industrial Engineering, Bandar Abbas Branch, Islamic Azad University, Bandar Abbas, Iran

ABSTRACT

Background: Diabetes entails a great quantity of deaths each year and a great quantity of people living with the disease do not find out their health status early sufficient. In this paper, we advance a data mining-based model for prematurely diagnosis and prediction of diabetes.

Methods: Although K-means is simple and can be utilized for a vast diversity of data kinds, it is wholly sensitive to initial locations of cluster centers which specify the final cluster result, which either enables an efficiently and adequate clustered dataset for the logistic regression model, or presents a lesser amount of data as a result of wrong clustering of the main dataset, thereby restricting the proficiency of the logistic regression model. The main purpose of this study is was to specify procedures of ameliorating the k-means clustering and logistic regression accuracy consequence. Therefore, our algorithm comprises of principal component analysis technique, k-means technique and logistic regression model.

Results: The results obtained from this study show that the ability to obtain the result of K-means clustering accuracy is much higher than what other researchers have obtained in similar studies. Also, compared to the results obtained from other algorithms, the logistic regression model was implemented at an improved level in predicting the onset of diabetes. Another real advantage is that the proposed algorithm was able to successfully model a new dataset.

Conclusion: In general, the proposed approach can be effectively used in predicting and early diagnosis of diabetes.

Keywords: Diabetes, Prediction, Principal component analysis, K-means, Logistic regression

* Bandar Abbas, University Boulevard, Islamic Azad University, Bandar Abbas branch, Postal Code: 7915893457, Tel: +989399182010, Email: N.rafei@iau-tnb.ac.ir