دوفصلنامه محاسبات و سامانه های توزیع شده
سال چهارم، شماره اول، شماره پیاپی ۷، صفحه ۷۸ تا ۸۹، سال ۱۴۰۰

تشخیص بیماری دیابت بر اساس روش کاهش ویژگی هوشمند و یادگیری ترکیبی ماشینی

محمد هادی‌زاده[۱]، دکتر آرش خسروی[*۲]
[۱]کارشناسی ارشد، مهندسی کامپیوتر، گرایش هوش مصنوعی و رباتیک، دانشگاه شهاب دانش.
[۲]استادیار گروه مهندسی کامپیوتر، دانشکده مهندسی، مرکز آموزش عالی محلات، محلات، ایران.

چکیده

استخراج وآنالیز از میان انبوهی از داده های مرتبط به سوابق بیماری و پرونده های پزشکی افراد با استفاده از فرایند داده کاوی می تواند منجر به شناسایی قوانین حاکم در تشخیص بیماری ها شود و اطلاعات ارزشمندی را جهت افزایش دقت درتشخیص بیماری ، پیش بینی و تشخیص بیماری با توجه به عوامل محیطی حاکم در اختیار متخصصان حوزه سلامت قرار دهد. هدف این پژوهش، تشخیص بیماری دیابت با استفاده از ترکیب آنالیز جداکننده خطی و الگوریتم گرگ خاکستری می باشد که بر روی دیتاست PIDD و به زبان پایتون انجام شده است. در این تحقیق با استفاده از این ترکیب و کاهش ویژگی، دقت بالاتری ارائه شده است که در مقایسه با کارهای پیشین به بهبود ۶ درصد رسیدیم.

کلمات کلیدی: تشخیص دیابت-داده کاوی-کاهش ویژگی ها- الگوریتم طبقه بندی- الگوریتم گرگ خاکستری.

Provide Diagnosis of Diabetes Based on Intelligent Feature Reduction and Machine Learning

Mohamad Hadizade[1] , Arash Khosravi[*2]
[1]M.Sc., Computer Engineering, Artificial Intelligence and Robotics, Shahab Danesh University, Iran.
[2]Assistant Professor, Faculty of Engineering, Mahallat Institute of Higher Education, Mahallat, Iran.

Abstract

Extraction and analysis from a large number of data related to disease records and medical records of individuals using the data mining process can lead to the identification of the rules governing the accurate diagnosis of diseases and provide valuable information for accuracy in the disease, forecast and Diagnosis of the disease to be provided to health professionals according to the prevailing environmental factors. The aim of this study was to diagnose diabetes using a combination of linear separator analysis and gray wolf algorithm based on PIDD database and Python language. We were able to provide higher accuracy by using this combination and by reducing the feature, so we achieved a 6% improvement.

Keywords: Diabetes Diagnosis, Data Mining, Diagnosis Accuracy, Classification Algorithm Gray Wolf Algorithm.

تاریخچه مقاله:

تاریخ ارسال: ۱۴۰۰/۰۲/۲۵
تاریخ اصلاحات: ۱۴۰۰/۰۴/۱۹
تاریخ پذیرش: ۱۴۰۰/۰۶/۰۲
تاریخ انتشار: ۱۴۰۰/۰۶/۳۱

Keywords:

Diabetes Diagnosis,
Data Mining,
Diagnosis Accuracy,
Classification Algorithm,
Gray Wolf Algorithm

*ایمیل نویسنده مسئول :
Khosravi.280@gmail.com

۱ – مقدمه

دیابت یکی از بیماری‌هایی است که زندگی افراد زیادی را با چالش روبه رو کرده است. با پیشرفت تسهیلات کامپیوتری ارائه‌شده توسط فناوری اطلاعات، هم‌اکنون پیش‌بینی بسیاری از عوارض و دردها به صورت دقیق‌تر، امکان‌پذیر شده است. نخستین مزیت چشمگیر فناوری اطلاعات آن است که ذخیره سازی عظیم داده های مربوط به اسناد و سوابق قبلی بیمار، حفظ شده و توسط بیمارستان ها به طور مستمر، مورد نظارت و کنترل قرار می گیرد. داده های پزشکی ذخیره شده می توانند برای پزشکان جهت بررسی الگوها در مجموعه داده، مفید باشند و الگوهای یافت شده در مجموعه های داده می توانند برای تشخیص بیماری به کار روند.

داده کاوی فرایندی است که می تواند به تفسیر اطلاعات و دانش بپردازد و به سرعت در حال گسترش و پیشرفت است. مطالعات زیادی از داده کاوی بهره برده اند تا از طریق آن مدل های پیش بینی کننده و فاکتورهای ناشناخته ای را بررسی کنند که در زمینه پزشکی ساخته شده اند [۱]. امروزه با پیشرفت ابزارهای دانش و کامپیوتری، راه های تشخیص بیماری تسهیل تر شده است و دقت آنها افزایش یافته است. بنابراین با کمک سیستم های کامپیوتری می توان یک روشی ارائه داد که به پیش بینی بیماری دیابت بپردازد و از طرف دیگری تاثیر افراد مُفسر آزمایشات را کمتر نماید تا بتوان از خطاهای انسانی و تصمیمات نادرست جلوگیری نمود [۲].

یک سیستم تشخیص کامپیوتری، سیستمی است که بتواند بدون کمک و دخالت انسان به تشخیص بیماری بپردازد و در این میان فقط از ابزار کامپیوتری کمک گرفته شده. لذا این سیستم بر مبنای اطلاعات و دانش کامپیوتری و نرم افزاری می باشد که از چند بلوک مهم تشکیل می شود. این بلوک ها در هر مساله ای می تواند متفاوت باشد اما در اکثر مقالاتی که استفاده شده اند شامل ۱. جمع آوری داده، ۲. پیش پردازش و آماده سازی داده، ۳. انتخاب ویژگی و کاهش ابعاد داده ها، ۴. دسته بندی داده ها و ۵. ارزیابی داده ها هستند که هر کدام از بلوک ها دارای روش ها و تکنیک های خاص خود می باشند که می تواند روی عملکرد سیستم تشخیص و سرعت آن تاثیرگذار باشد.

بهینه سازی یکی از ابزارهای ریاضی است که به منظور یافتن نقطه بهینه به کار می رود. نقطه بهینه می تواند نقطه ماکزیمم یا مینیمم باشد. به عبارت دیگر نقطه بهینه، مکانی است که تابع هزینه یا تابع هدف دارای کمترین مقدار یا بیشترین مقدار خود می باشد. از این رو در کارهای طبقه بندی و هوش مصنوعی نیز کاربرد فراوانی دارد. بدین منظور که مجموعه ویژگی هایی انتخاب می گردند که منجر به بالاترین دقت تشخیص شود. این روش انتخاب ویژگی وابسته به طبقه بندی دودویی است و اگر مکان مربوط به ویژگی یک شود یعنی آن ویژگی انتخاب شده است و اگر صفر شود بدین معنی است که ویژگی موردنظر از مجموعه ویژگی ها حذف گردیده است.

الگوریتم گرگ خاکستری یکی از الگوریتم های فرابتکاری است که از رفتارهای موجود در طبیعت الهام می گیرد و این الگوریتم نیز همان گونه از نام آن پیداست، از زندگی گرگ ها الهام گرفته است در زمانی که به دنبال شکار و طعمه می باشند. در تقسیم بندی زیر که برای گرگ ها آمده است، هر کدام از آنها نقش متفاوتی در گروه گله دارند که به تشریح آنها پرداخته خواهد شد [۳, ۴]. گرگ های خاکستری در راس زنجیره غذایی می باشند و دارای زندگی گروهی و اجتماعی می باشند. به طور معمول تعداد متوسط گرگ های هر گله بین پنج تا دوازده عدد است و در هر گله چهار نوع رتبه اصلی وجود دارد.

- **گرگ های الفا:** این گرگ ها که رهبر گروه و گله می باشند و می توانند از نوع مذکر یا مونث باشند. این گرگ ها بر گله تلسط دارند و مواردی از قبیل تعیین محل استراحت یا نحوه شکار را مدیریت می کنند اما علاوه بر رفتار مسلط گرگ های آلفا، نوعی ساختار دموکراتیک هم در گروه وجود دارد.
- **گرگ های بتا:** نقش این گرگ ها، کمک به گرگ های آلفا در فرایند تصمیم گیری و رهبری می باشد و همچنین مستعد انتخاب شدن به جای آنها هستند.
- **گرگ های دلتا:** رتبه بندی این نوع گرگ ها پایین تر از گرگ های بتا هستند و شامل گرگ های پیر، شکارچی ها و گرگ های مراقبت کننده از نوزادان می باشند.
- **گرگ های امگا:** پایین ترین رتبه در هرم سلسله مراتبی گرگ ها هستند که کمترین حق و حقوق را نسبت به بقیه اعضای گروه دارند. آنها همیشه بعد از سایر گرگ ها غذا می خورند و در فرایند تصمیم گیری هیچ گونه مشارکتی ندارند.

مطابق بررسی های Muro و همکارانش، مراحل اصلی شکار گرگ های خاکستری شامل ۳ مرحله اصلی است؛ الف) مشاهده شکار، ردیابی و تعقیب آن، ب) نزدیک شدن، احاطه کردن (حلقه زدن) به دور شکار و گمراه کردن آن تا زمانی که از حرکت باز بماند، ج) حمله به شکار.

روال بهینه سازی با استفاده از گرگ های آلفا، بتا، دلتا و امگا انجام می گیرد. یک گرگ به عنوان آلفا هدایت کننده اصلی الگوریتم فرض می شود و یک گرگ بتا و دلتا نیز مشارکت دارند و بقیه گرگ

ها نیز به عنوان دنبال کننده آنها به شمار می آیند. در شکل زیر این فرایند نمایش داده شده است [۵].

نویسندگان مقاله [۶] نیز از ماشین بردار پشتیبان برای تشخیص بیماری دیابت استفاده کردند اما آنها کاهش ابعاد ویژگی نیز انجام دادند. آنها از الگوریتم بهینه سازی ژنتیک برای انتخاب مجموعه بهینه ویژگی ها استفاده کردند تا بتوانند فقط ویژگی های مهم و تاثیرگذار را به کار ببرند همچنین با استفاده از الگوریتم خوشه بندی K-Means توانستند نویز موجود در داده ها را کاهش دهند و داده ها را پیش پردازش نمودند (۲۶۸ نمونه مربوط به افراد دیابتی بود و ۵۰۰ نمونه مربوط به افراد سالم بود که ۳۷۶ نمونه آنها به پیش پردازش نیاز داشتند). با اجرای روش پیشنهادی خود روی دیتاست با ویژگی های کاهش یافته Pima به دقت تشخیص ۹۸/۷۹ رسیدند. همچنین بیان کردند که دقت روش ماشین بردار پشتیبان بدون استفاده از الگوریتم خوشه بندی K-Means برابر ۹۸/۷۹ بوده است که این اختلاف دقت نقش کارایی الگوریتم خوشه بندی K-Means را نشان می دهد. در مقاله [۷] نویسنده ها از یک سیستم هوشمند ترکیبی برای تشخیص بیماری دیابت اسم بردند که شبکه های عصبی و الگوریتم ژنتیک با هم ترکیب شده اند. آنها ابتدا از الگوریتم بهینه سازی ژنتیک جهت انتخاب ویژگی های دیتاست Pima بهره بردند، الگوریتم ژنتیک از میان ۸ ویژگی موجود، ۴ ویژگی را به عنوان مجموعه بهینه ویژگی ها معرفی کرد. سپس از یک شبکه عصبی پرسپترون چندلایه به عنوان طبقه بند جهت تشخیص استفاده کردند. به این ترتیب نه تنها با استفاده از الگوریتم ژنتیک زمان و هزینه را کاهش می دهند بلکه دقت طبقه بند شبکه عصبی را نیز افزایش دادند و به دقت ۷۹/۱۳۰۴ رسیدند. مقاله [۸] تشخیص زودهنگام دیابت نوع دو را با استفاده از سیستم طبقه بند چندتایی بیان می کند. آنها در این مقاله بیان کرده اند که یه طبقه بند به تنهایی نمی تواند تشخیص های کاملا درست ارائه دهد به همین دلیل جهت بهبود دقت تشخیص بیماری از سیستم های طبقه بندی چندتایی بهره بردند و با استفاده از الگوریتم وزن دهی دینامیکی به هر کدام از طبقه بندها وزنی را اختصاص دادند. با اجرای روش پیشنهادی خود روی دیتاست T2DM با ۴۳۵ رکورد و ۱۱ ویژگی نشان دادند که این روش نسبت به روش های تکی دارای دقت بیشتری می باشد. نویسندگان مقاله [۹] از الگوریتم بهینه سازی ژنتیک برای انتخاب ویژگی استفاده کردند. آنها در این کار مجموعه ویژگی های دیتاست pima مربوط به دیابت را از ۸ به ۴ کاهش داده اند و به دقت ۸۳ درصد رسیدند. در این کار از طبقه بند فازی استفاده شده است و با استفاده از ۷۰ درصد داده ها، آن را آموزش داده اند و با ۳۰ درصد مابقی داده ها، از مدل تست گرفته اند و آن را مورد ارزیابی قرار دادند. همچنین در مقاله [۱۰] نیز انتخاب ویژگی از دیتاست pima انجام شده است. آنها در این کار سه الگوریتم ژنتیک، PSO و الگوریتم جست و جوی هارمونی را با هم ترکیب کرده اند و در نهایت با کمک خوشه بندی Kmeansترکیب کرده اند تا بتوانند مجموعه درست ویژگی ها را انتخاب نمایند. در نهایت با طبقه بندی کننده نزدیک ترین همسایه به تشخیص بیماری دیابت پرداختند و به مقدار صحت ۹۱/۶۵ درصد رسیدند. در مقاله [۱۱] نیز انتخاب ویژگی دیتاست pima با استفاده از الگوریتم جنگل تصادفی انجام شده است و به کمک svm طبقه بندی انجام گرفته است. نتایج صحت به دست آمده برای روش پیشنهادی فوق حدود ۸۱ درصد می باشد در حالی اگر از انتخاب ویژگی استفاده نمی کردند و تنها از svmاستفاده می کردند نتیجه صحت حدود ۷۶ درصد می شد. در مقاله [۱۲] مرجع نویسندگان از ترکیب الگوریتم svm و الگوریتم بهینه سازی ازدحام ذرات pso برای تشخیص بیماری دیابت استفاده کردند بدین صورت که با استفاده از svm با کرنل های مختلف به تشخیص بیماری و با استفاده از الگوریتم pso به انتخاب ویژگی در میان دیتاست pima پرداختند. نتایج شبیه سازی آنها نشان داده است که svm با کرنل خطی دارای بالاترین دقت بوده است. همچنین در مقاله دیگری از شبکه عصبی برای دسته بندی استفاده شده است که دارای دقتی حدود ۹۶ درصد روی ۷۰ درصد داده های تست دیتاست pima بوده است[۱۳]. در مقاله [۱۴]، روش پیشنهادی روی دیتاست pima اجرا شده است و از دو روش دسته بندی شبکه عصبی و XGBoost برای دسته بندی استفاده کرند. برخلاف مقاله قبلی، نشان دادند که XGBoost بهتر از روش شبکه عصبی عمل نموده است و دارای دقت ۷۸ درصد می باشد در حالی که دقت شبکه عصبی کمتر از این مقدار بوده است. البته لازم به ذکر است که انتخاب ویژگی این مقاله، قواعد انجمنی است که با الگوریتم های فراابتکاری متفاوت می باشد.

پژوهش پیش رو از این نظر دارای اهمیت است که می تواند به کمک داده کاوی به پیش بینی بیماری دیابت بپردازد. در چند سال گذشته موجی از علاقه در استفاده از آنالیز تشخیص خطی نشان داده شده است. روش طبقه بندی آنالیز تشخیص خطی یکی از روش های نظارت شونده است که از آن برای طبقه بندی و دسته بندی استفاده می شود. تجربه نشان داده است که ماشین های بردار پشتیبان عملکرد خوبی بر روی طیف وسیعی از مشکلات داشته اند. در این تحقیق روش تشخیص از دو تکنیک آنالیز تشخیص خطی جهت تشخیص دیابت و تکنیک الگوریتم بهینه سازی گرگ خاکستری جهت انتخاب ویژگی های مهم تشکیل شده است که یک روش نوین

و جدید در این زمینه می باشد. هدف از این پژوهش، تشخیص بیماری دیابت با استفاده از ترکیب آنالیز تشخیص خطی و الگوریتم گرگ خاکستری می باشد که روی دیتاست PIDD و به زبان پایتون انجام شده است.

۲ – متدولوژی و ارزیابی روش

انتخاب ویژگی[1] یا کاهش داده[2]، یکی دیگر از بلوک های اصلی روش پیشنهادی می باشد که در این پژوهش مورد توجه قرار گرفته است. همان گونه که در مرحله قبل توضیح داده شد، استخراج ویژگی و انتخاب ویژگی دو فرایندی هستند که با مشخصه های سیستم سروکار دارند. در استخراج ویژگی بایستی سعی شود که تصویر یا نمونه دیگر به تعداد مشخصه تبدیل شود طوری که مشخصه های نماینده آن تصویر باشند. اما در انتخاب ویژگی بایستی از بین مجموعه ویزگی ها، تعدادی انتخاب گردند بدین ترتیب بعد ماتریس ویژگی کاهش می یابد برای همین است که به انتخاب ویژگی، کاهش ابعاد نیز گفته می شود[۱۵]. تشخیص در داده کاوی براساس طبقه بندی انجام می گیرد. طبقه بندی یعنی اینکه افراد مورد آزمایش را با توجه به ویژگی هایی که دارند، به دو دسته بیمار و سالم دسته بندی نماییم . داده کاوی براساس طبقه بندی انجام می گیرد. در این کار دو روش وجود دارد که عبارتند از یادگیری نظارت شونده و یادگیری بدون ناظر. در روش یادگیری با ناظر، علاوه به مجموعه ویژگی ها، به بردار لیبل یا تارگت نیز نیاز است تا کلاس هر فرد مشخیص شود. کلاس هر فرد، لیبلی است که به آن فرد داده شده تا نشان دهد که بیمار است یا سالم است. در روش یادگیری بدون ناظر نیازی به لیبل و تارگت برای افراد نیست و آنها براساس ویژگی هایشان خوشه بندی می گردند. لذا معیاری وجود ندارد که سنجیده شود آیا تشخیص درست بوده است یا خیر. به همین دلیل است که برای تشخیص از روش های یادگیری با ناظر استفاده می شود تا بتوان در نهایت، دقت سیستم تشخیص را محاسبه نمود و به بررسی کارایی آن پرداخت. فلوچارت الگوریتم به کار گرفته شده در این پژوهش در شکل ۱ نشان داده شده است.

۳-ارزیابی مدل تشخیص

برای ارزیابی روش پیشنهادی از معیارهای صحت و دقت استفاده می شود اما ابتدا لازم است تعاریف زیر بیان گردد:

- TP (تعداد مثبت‌های درست[3]): تعداد مواردی است که روش آنها را بیمار تشخیص داده است و در واقع بیمار هم بوده اند.
- TN (تعداد منفی‌های درست[4]): تعداد مواردی است که روش آنها را سالم تشخیص داده است و در واقع سالم هم بوده اند.
- FP (تعداد مثبت‌های غلط[5]): تعداد مواردی است که روش آنها را بیمار تشخیص داده است و و در واقع سالم بوده اند.
- FN (تعداد منفی‌های غلط[6]): تعداد مواردی است که روش آنها را سالم تشخیص داده است و ولی در واقع بیمار بوده اند.

در ادامه چند معیار معروف و شناخته‌شده در حوزه تشخیص بیماری معرفی می‌گردند:

۱. دقت[7]:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{(معادله-۱)}$$

۲. حساسیت[8]:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{(معادله-۲)}$$

۳. خصوصیت[9]:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \text{(معادله-۳)}$$

۴. دقت کل[10] یا صحت:

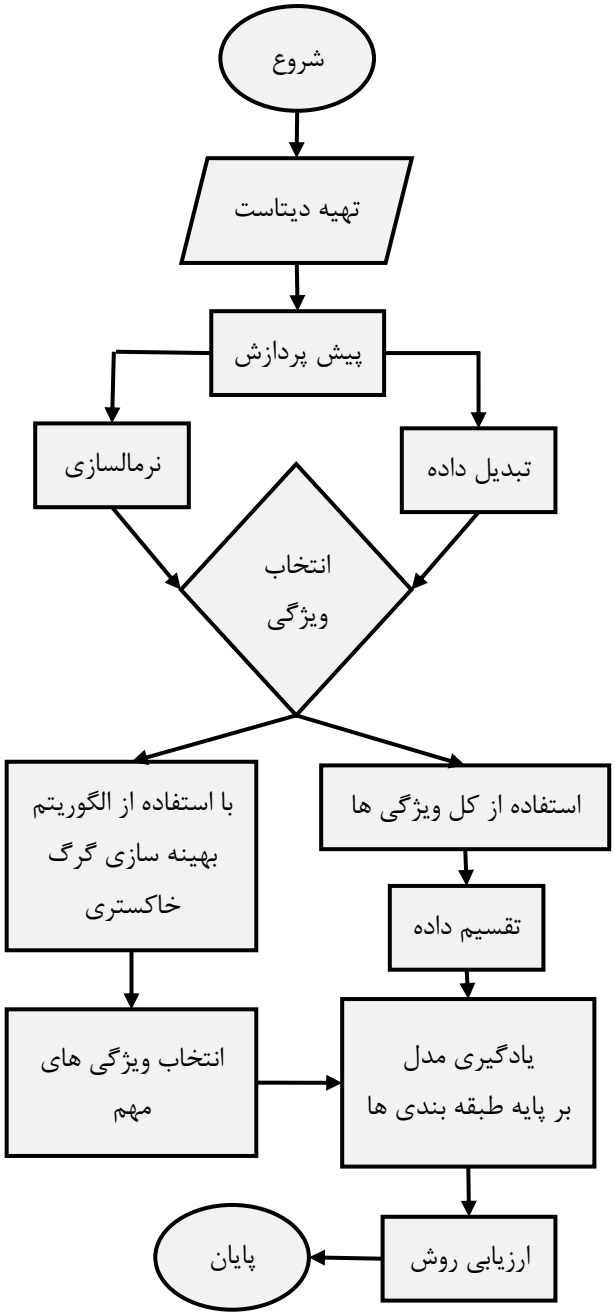
$$(\text{ACC}) = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(معادله ۴)}$$

۵. سطح زیر منحنی دقت[11]:

$$\text{AUC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad \text{(معادله-۵)}$$

[1]Feature Selection
[2]Data Reduction

[3]True positive
[4]True negative
[5]False positive
[6]False negative
[7]Precision
[8]Sensitivity
[9]Specificity
[10]Overall accuracy(ACC)
[11]Area under curve(AUC)



(شکل-۱): فلوچارت الگوریتم به کار گرفته شده در این پژوهش

۴ – بانک داده و پیش‌پردازش داده‌ها

مجموعه داده ای که در این پژوهش به کار رفته است، دیتاست معروف در زمینه بیماری دیابت نوع ۲ است که به PIDD[۱۲] شناخته می شود[۱۶]. مقالات زیادی از این دیتاست استفاده کرده اند که می توان به مراجع[۱۷-۲۲] رجوع نمود. این دیتاست در سال ۲۰۱۱ از انستیتو دیابت جهانی گرفته شده است و توسط Peter Turney گردآوری و برچسب گذاری شده است طوری که به راحتی می توانست برای کارهای طبقه بندی باینری مورد استفاده قرار گیرد. این دیتاست از مرجع مربوطه قابل دانلود کردن می باشد. در شکل ۲ نمایی از رکوردهای اول آن نشان داده شده است.

data.csv

```
1   Pregnancies,Glucose,BloodPressure
2   6,148,72,35,0,34,1,50,1
3   1,85,66,29,0,27,0,31,0
4   8,183,64,0,0,23,1,32,1
5   1,89,66,23,94,28,0,21,0
6   0,137,40,35,168,43,2,33,1
7   5,116,74,0,0,26,0,30,0
8   3,78,50,32,88,31,0,26,1
9   10,115,0,0,0,35,0,29,0
10  2,197,70,45,543,31,0,53,1
11  8,125,96,0,0,0,0,54,1
12  4,110,92,0,0,38,0,30,0
13  10,168,74,0,0,38,1,34,1
14  10,139,80,0,0,27,1,57,0
15  1,189,60,23,846,30,0,59,1
16  5,166,72,19,175,26,1,51,1
```

(شکل-۲): دیتاست به کار گرفته شده در این پژوهش

در این دیتاست ۷۶۸ نفر تحت آزمایش می باشند و برای آنها ۹ ستون داده قرار گرفته است. ۸ ستون اول ویژگی ها می باشند و ستون آخر نشان می دهد که فرد به بیماری دیابت مبتلا می باشد یا خیر. برای این کار عدد "1" مبتلا بودن به دیابت و عدد "0" نرمال بودن را نشان می دهد. فراخوانی این دیتاست از طریق کتابخانه pandas در پایتون انجام شده است. لازم به ذکر است که همه افراد تحت آزمایش در این دیتاست دارای جنسیت زن هستند و حداقل ۲۱ سال سن دارند. آنها اهل phoenix، Arizona و یا USA. بوده اند[۲۳].

[12] Pima Indians Diabetes Database

(جدول-۱): ویژگی‌های دیتاست

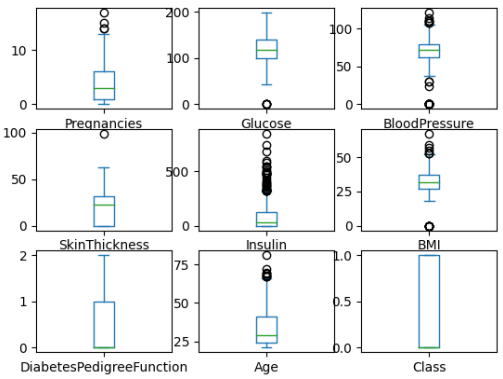
نام ویژگی	نماد به کار رفته در دیتاست	برچسب
تعداد دفعات بارداری[۱۳]	Pregnancies	X1
غلظت گلوکز پلاسمای خون[۱۴] mg/dl	Glucose	X2
فشار دیاستولیک خون[۱۵] $mmHg$	BloodPressure	X3
چربی زیر پوست[۱۶] mm	SkinThickness	X4
میزان قند ناشتا[۱۷] mmU/l	Insulin	X5
شاخص جرم بدن[۱۸] kg/m^2	BMI	X6
سابقه خانوادگی[۱۹]	DiabetesPedigree Function	X7
سن[۲۰] فرد (سال)	Age	X8

پیش پردازش داده ها

توجه داشته باشید که همه داده های جدول ۱ دیتاست از نوع عددی هستند و نیازی به تبدیل داده کیفی به عددی نمی باشد. توصیف داده ها از نظر ویژگی های آماری نظیر تعداد، میانگین، انحراف معیار، مینیمم، چندک اول و دوم و سوم و ماکزیمم توسط دستور describe انجام شده است که در جدول ۲ گزارش آن آمده است. خلاصه جدول ۲ به صورت نمودار جعبه ای برای هر یک از ویژگی ها در شکل ۴-۳ رسم شده است.

(جدول-۲): توصیف آماری داده‌ها

	X1	X2	X3	X4	X5	X6	X7	X8
Count	768	768	768	768	768	768	768	768
Mean	3.84	120.89	69.10	20.53	79.79	32.04	0.37	33.24
Std	3.36	31.97	19.35	15.94	115.24	7.88	0.510	11.76
Min	0	0	0	0	0	0	0	21
25%	1	99	62	0	0	27	0	24
50%	3	117	72	23	30	32	0	29
75%	6	140.25	80	32	127.25	37	1	41
Max	17	199	122	99	846	67	2	81



(شکل-۱-۳): توصیف آماری داده‌ها به کار گرفته شده در این پژوهش

این نوع نمودار توزیع داده را خیلی ساده تر بیان می کند. توزیع داده یعنی پراکندگی آن است، یعنی اینکه مقادیر یک ویژگی خاص چقدر از هم دور هستند یا به هم نزدیک هستند. مشاهده می گردد ۶۵/۱۰ درصد از داده ها مربوط به افراد نرمال(۵۰۰ نفر) و ۳۴/۸۹ درصد مربوط به افراد دیابتی(۲۶۸نفر) می باشد. با توجه به جدول (۳) مشاهده

[13]Number of times pregnant
[14]Plasma glucose concentration
[15]Diastolic blood pressure
[16]Triceps skin fold thickness
[17]Hour serum insulin
[18]Body mass index
[19]Diabetes pedigree function
[20]Age

می گردد که دامنه توزیع از هر ویژگی به ویژگی دیگر متفاوت است لذا نرمال سازی بایستی انجام گیرد تا داده ها در محدوده [۱-۰] قرار گیرند، این امر از روی هیستوگرام ویژگی ها نیز واضح است.

تقسیم داده

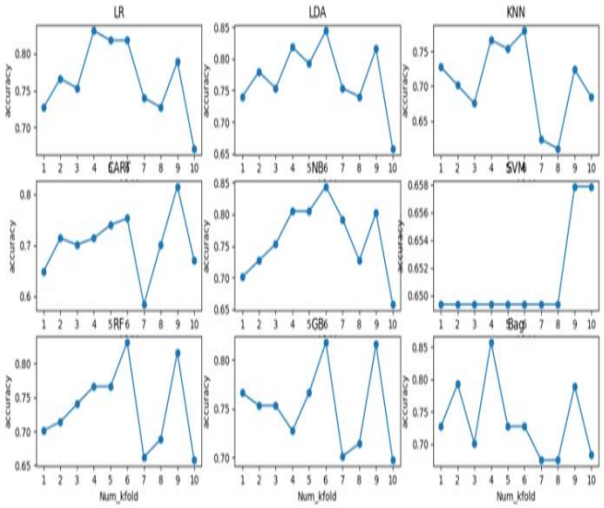
در این مرحله لازم است داده ها به دو دسته آموزش و آزمایش تقسیم شوند.

- داده های آموزش داده هایی هستند که برای آموزش الگوریتم های طبقه بندی به کار می روند. این داده ها باید حجم بیشتری نسبت به داده های آزمایش داشته باشند تا مدل بهتر و دقیق تری به ما ارائه دهد. لذا ۸۰ درصد داده ها به آموزش اختصاص یافته اند که تعداد آنها ۶۱۵ عدد می باشد.
- داده های آزمایش: این داده ها بایستی حین آموزش به کار نرفته باشند و به عبارتی برای سیستم آشنا نباشند تا بتوان قابلیت اطمینان تشخیص سیستم را بالا برد و بتوان به دقت به دست امده اعتماد داشت. از داده های استفاده نشده در فرایند آموزش استفاده می شود که مابقی ۲۰ درصد داده ها هستند و ۱۵۳ عدد می باشند.

لازم به ذکر است که این تقسیم بندی کاملا تصادفی است و در هر بار ممکن است ترتیب داده ها تفاوت داشته باشد. برای این کار از تکنیک cross-validation با تعداد فولد kfold=10 استفاده می شود و با استفاده از انواع طبقه بندی بهترین طبقه بندی را انتخاب کردید از جمله طبق بندی های که استفاده شد

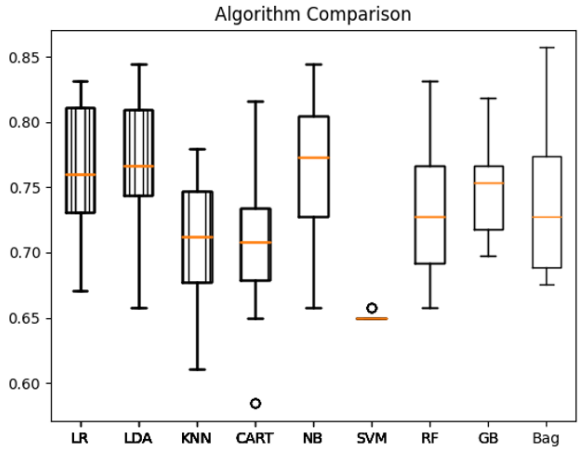
۱- رگرسیون لگاریتمی: Logistic Regression (LR)
۲- آنالیز جداکننده خطی: Linear Discriminant Analysis (LDA)
۳- نزدیک ترین همسایگی: K-Nearest Neighbors (KNN)
۴- درخت رگرسیون و کلاسبندی: Classification and Regression Trees (CART)
۵- بیزین ساده گوسی: Gaussian Naive Bayes (NB)
۶- ماشین بردار پشتیبان: Support Vector Machines (SVM)
۷- جنگل تصادفی: Random Forest
۸- بوستینگ گرادیان: Gradient Boosting
۹- Bagging

اکنون مدل تشخیص می تواند هر یک از موارد فوق باشد. برای انتخاب یکی از آنها بایستی معیار صحت را مورد بررسی قرار دهیم. در زیر معیار صحت به ازای فولدهای مختلف برای هر یک از طبقه بندی کننده های فوق رسم شده است:



(شکل -۲-۳) نمودار صحت به ازای طبقه بندی کننده های مختلف برحسب شماره

برای اینکه بتوان به خوبی تشخیص داد که کدام یک از طبقه بندی کننده ها از بقیه بهتر است از مقادیر صحت های نشان داده شده در نمودار فوق، میانگین گرفته شده است و نمودار 4 رسم شده است مزیتی که نمودار جعبه ای دارد می تواند مینیمم و ماکزیمم را همراه با مقدار میانگین نشان دهد اما با این وجود در جدول 5 مقادیر میانگین صحت هر ۱۰ فولد به همراه انحراف معیار آمده است تا بتوان انتخاب دقیقی را انجام داد.



(شکل-۴) نمودار جعبه ای مقایسه صحت به ازای طبقه بندی کننده های مختلف

در قسمت قبل مشخص شد که مدل انالیز جداکننده خطی دارای بالاترین مقدار صحت می باشد. در این قسمت قصد داریم معیارهای ارزیابی را به ازای داده های تست محاسبه نماییم. جدول (۵)

(جدول-۵): مقایسه مقدار صحت میانگین و انحراف استاندارد طبقه بندی کننده های مختلف

نام طبقه بندی کننده	مقدار میانگین صحت	مقدار انحراف معیار میانگین
Logistic Regression	0.764234	0.048013
Linear Discriminant Analysis	0.769446	0.050389
K Neighbors	0.704426	0.054192
Decision Tree	0.704528	0.059019
Gaussian NB	0.761637	0.054822
SVM	0.651059	0.003418
Random Forest	0.734381	0.057201
Gradient Boosting	0.751316	0.040548
Bagging	0.735680	0.056720

با توجه به جدول فوق اگر بخواهیم یک طبقه بندی کننده انتخاب نماییم مسلما LDA دارای بالاترین مقدار صحت و SVM نیز دارای پایین ترین مقدار صحت می باشد. بنابراین LDA به عنوان طبقه بندی کننده منتخب در نظر گرفته می شود.

با توجه به جدول فوق اگر بخواهیم یک طبقه بندی کننده انتخاب نماییم مسلما LDA دارای بالاترین مقدار صحت و SVM نیز دارای پایین ترین مقدار صحت می باشد. بنابراین LDA به عنوان طبقه بندی کننده منتخب در نظر گرفته می شود.

(جدول-۶): نتایج معیارهای ارزیابی طبقه بندی کننده LDA با هشت ویژگی.

معیار ارزیابی	مقدار
accuracy	0.7857
recall	0.5636
precision	0.775
F1	0.6526
تعداد کلاس 0	99
تعداد کلاس 1	55
تعداد تشخیص های درست TP+TN	90+31
تعداد تشخیص های غلط FN+FP	24+9

۵ – نتایج

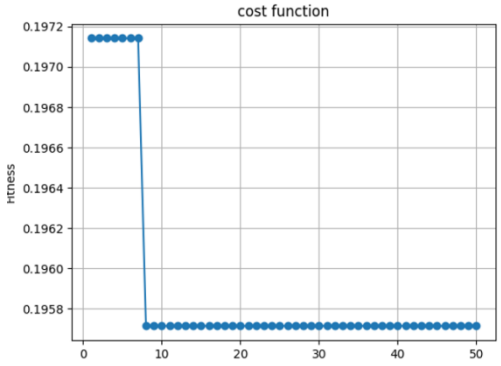
در ادامه هدف این است که به کمک الگوریتم گرگ خاکستری ویژگی های مهم تر را شناسایی کنیم. به عبارت دیگر الگوریتم گرگ خاکستری می تواند با الگوهای مکرر، ویژگی های مهم را مشخص نماید. پارامترهای الگوریتم گرگ خاکستری در جدول ۷ آمده است:

(جدول- ۷) پارامترهای به کار رفته در الگوریتم گرگ خاکستری.

نام	مقدار
تعداد ویژگی ها(مجهولات)	۸
تعداد تکرار	۵۰
کران بالا	۱
کران پایین	۰
مقدار آستانه	۰/۵

انتخاب ویژگی یک فرایند مهم در حوزه داده کاوی و یادگیری ماشینی است چرا که در عین حالی که از پیچیدگی مسئله می کاهد می تواند منجر به افزایش سرعت روش نیز گردد. نتایج تابع هدف برای الگوریتم گرگ خاکستری روی دیتاست مذکور به صورت نمودار ۷ می باشد. روش انتخاب ویژگی مبتنی بر تابع هدف و متغیرهای تصمیم می باشد که در این مقاله، متغیرهای تصمیم همان انتخاب ویژگی ها می باشد یعنی به این صورت که اگر مقدار موقعیت ویژگی بزرگتر از مقدار آستانه باشد یعنی ویژگی انتخاب شود و در غیر این صورت انتخاب نمی شود.

و یا تابع هدف نیز ترکیبی از دو معیار خطای طبقه بندی و نسبت تعداد ویژگی به ماکزیمم ویژگی است که معیار اول با ضریب ۰.۹۹ و معیار دو با ضریب ۱-۰.۹۹ وارد شده است. نتایج تابع هدف الگوریتم گرگ خاکستری روی دیتاست مذکور به صورت زیر می باشد(نمودار ۸)



(شکل-۸): نتیجه تابع هدف الگوریتم گرگ خاکستری.

نقش الگوریتم بهینه سازی گرگ خاکستری در انتخاب ویژگیهای مهم است که در این جا از بین ۸ ویژگی موجود در دیتاست، ۴ ویژگی را انتخاب کرده است. که عبارتند از ویژگی های X1-X3-X5-X6 لازم به ذکر است که این نتیجه طی چندین بار اجرا گرفته شده است. چون برای مقداردهی موقعیت اولیه xها از تابع رندوم استفاده شده است، لازم بود که چندین بار اجرا شود تا اطمینان خاطر حاصل شود. نتایج طبقه بندی با الگوریم گرگ خاکستری و آنالیز جداکننده خطی به ازای ۱۰ درصد داده های تست آمده است که نتایج در جدول ۹ و شکل ۸ قابل رویت می باشد.

(جدول-۹): نتیجه داده ها به همراه کلاس واقعی آنها با استفاده از الگوریتم گرگ خاکستری و آنالیز جداکننده خطی

معیار ارزیابی	مقدار
accuracy	۰.۸۳۱۱
recall	۰.۶۵۵۱
precision	۰.۸۶۳۶
F1	۰.۷۴۵۰
تعداد کلاس0	۴۸
تعداد کلاس1	۲۹
تعداد تشخیص های درست(TP+TN)	۱۹+۴۵
تعداد تشخیص های غلط(FN+FP)	۳+۱۰

	precision	recall	f1-score	support
0	0.82	0.94	0.87	48
1	0.86	0.66	0.75	29
accuracy			0.83	77
macro avg	0.84	0.80	0.81	77
weighted avg	0.84	0.83	0.83	77

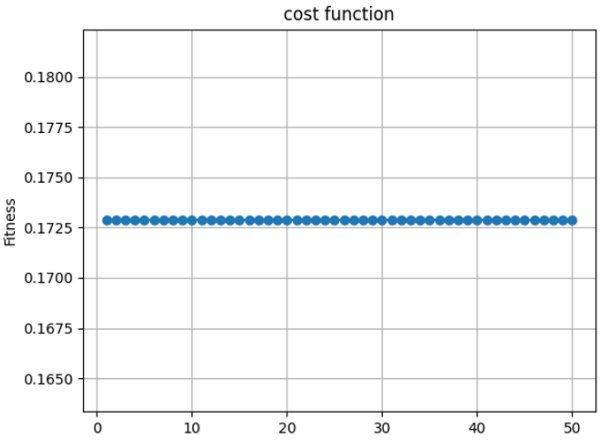
Feature Size: 4

(شکل-۸): نمایش نتایج گرفته شده از ویژوال استودیو.

برای بررسی انتخاب ویژگی با الگوریتم های فراابتکاری دیگری نیز انجام دادیم و برای این کار از الگوریتم ژنتیک و pso استفاده شد این دو الگوریتم نیز مجموعه ویژگی ها را به صورت زیر انتخاب کردند:

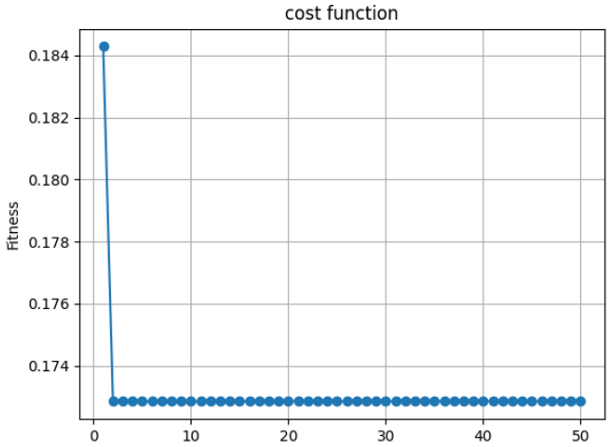
Feat=x1-x3-x5-x6

اما مقدار تابع هزینه آنها حدود دو درصد با هم تفاوت داشت و همچنین زمان اجرای آنها نیز با هم فرق می کرد که در نمودار ۱۰ نتایج آنها آمده است:



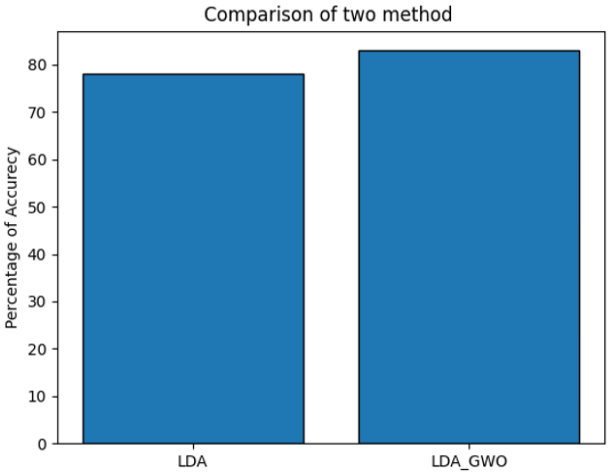
(نمودار-۱۰): نمودار تابع هزینه الگوریتم pso

نمودار فوق نشان می دهد که تابع هزینه به صورت میانگین حدود ۰.۱۷۲۵ می باشد در حالی که تابع هزینه الگوریتم ژنتیک از ۰.۱۸ شروع می شود و به همان حدود می رسد (نمودار ۱۱) که این تفاوت نیز به دلیل تصادفی بودن مقدار اولیه موقعیت ها می باشد که رخ داده است.



(نمودار-۱۱) نمودار تابع هزینه الگوریتم ژنتیک

هدف این قسمت این است که بین دو روش انجام شده قسمت پیشین، مقایسه ای انجام گیرد تا اهمیت انتخاب ویژگی و نقش الگوریتم گرگ خاکستری مشخص شود. در شکل ۱۲ نمودار میله ای درصد دقت کل را به ازای داده های آزمایشی برای روش آنالیز جداکننده خطی و روش پیشنهادی را نشان میدهد.



(نمودار-۱۲): نمودار مقایسه دقت کل داده های آزمایشی به روش LDA و روش پیشنهادی

۶- نتیجه گیری

هدف این مقاله تشخیص بیماری دیابت بوده است که برای این کار از ۷۶۸ رکورد داده های موجود در دیتاست PIMA استفاده گردید طوری که ۸۰ درصد داده ها به آموزش اختصاص یافت و ۲۰ درصد مابقی به آزمایش مدل اختصاص یافت.+ مجموعه ۶۱۵ داده صرف آموزش و ۱۵۳ داده صرف آزمایش مدل گردید. روش طبقه بندی در این کار با استفاده آنالیز جداکننده خطی بوده است که زمانی که به تنهایی مورد استفاده قرار گرفت دارای دقت ۷۸ درصد شد. اما با ترکیب این طبقه بندی کننده با الگوریتم گرگ خاکستری توانستیم یک روش ترکیبی با دقت بالاتر ارائه دهیم و دقت تشخیص از ۷۸ درصد به ۸۴ درصد رسید در حالی که به جای ۸ ویژگی از ۴ ویژگی استفاده گردید. به عنوان نتیجه گیری آخر می توان گفت انتخاب ویژگی به روش الگوریتم گرگ خاکستری منجر به بهبود نتایج مدل آنالیز جداکننده خطی شد.

برای کارهای آینده پیشنهاد می شود روش های بیشتری را نیز بررسی نماییم و استفاده از دیتاست‌های بزرگتر به منظور بررسی بیشتر داده‌ها تا بتوان سیستمی با قابلیت اطمینان بالاتر ارئه داد و همچنین استفاده از دیتا ست بزرگتر با ویژگی های بیشترو اهداف دیگری مانند بالا بردن سرعت یا کاهش نرخ خطا را نیز اضافه نمود.

۷ – مراجع

[1] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, no. 01, p. 1, 2017.

[2] D. J. Hand, "Principles of data mining," Drug safety, vol. 30, no. 7, pp. 621-622, 2007.

[3] D. M. Nathan et al., "Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes," Diabetes care, vol. 32, no. 1, pp. 193-203, 2009.

[4] B. M. Frier and M. Fisher, Hypoglycaemia in clinical diabetes. John Wiley & Sons, 2007.

[5] S. E. Inzucchi et al., "Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)," Diabetologia, vol. 55, no. 6, pp. 1577-1596, 2012.

[15] M. Jafari-Eskandari, N. Moghaddam-shebeilo, and E. Khodabakhshi, "The importance of intellectual capital in knowledge-based organizations for obtaining competitiveness superiority."

[16] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proceedings of the Annual Symposium on Computer Application in Medical Care, 1988: American Medical Informatics Association, p. 261.

[17] J. Eggermont, J. N. Kok, and W. A. Kosters, "Genetic programming for data classification: Partitioning the search space," in Proceedings of the 2004 ACM symposium on Applied computing, 2004: ACM, pp. 1001-1005.

[18] M. L. Raymer, T. E. Doom, L. A. Kuhn, and W. F. Punch, "Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 33, no. 5, pp. 802-813, 2003.

[19] T. Jiang and A. B. Owen, "Quasi-regression for visualization and interpretation of black box functions," ed: Stanford University, Stanford, 2002.

[20] P. Sykacek and S. J. Roberts, "Adaptive classification by variational Kalman filtering," in Advances in Neural Information Processing Systems, 2003, pp. 753-760.

[21] M. Skurichina, L. I. Kuncheva, and R. P. Duin, "Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy," in International Workshop on Multiple Classifier Systems, 2002: Springer, pp. 62-71.

[22] J. Garcke, M. Griebel, and M. Thess, "Data mining with sparse grids," Computing, vol. 67, no. 3, pp. 225-253, 2001.

[23] P. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study," ARPN Journal of Engineering and Applied Science, vol. 10, no. 1, pp. 8-13, 2015.

[6] A. Iyer, S. Jeyalatha, and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," arXiv preprint arXiv:1502.03774, 2015.

[7] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in engineering software, vol. 69, pp. 46-61, 2014.

[8] J. Zhu, Q. Xie, and K. Zheng, "An improved early detection method of type-2 diabetes mellitus using multiple classifier system," Information Sciences, vol. 292, pp. 1-14, 2015.

[9] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in 2017 international conference on computing networking and informatics (ICCNI), 2017: IEEE, pp. 1-5.

[10] X. Li, J. Zhang, and F. Safara, "Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm," Neural Processing Letters, pp. 1-17, 2021.

[11] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, vol. 1: IEEE, pp. 141-146.

[12] D. K. Choubey, S. Tripathi, P. Kumar, V. Shukla, and V. K. Dhandhania, "Classification of Diabetes by Kernel based SVM with PSO," Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), vol. 14, no. 4, pp. 1242-1255, 2021.

[13] E. A. Frimpong, A. Oluwasanmi, E. Y. Baagyere, and Q. Zhiguang, "A feedforward artificial neural network model for classification and detection of type 2 diabetes," in Journal of Physics: Conference Series, 2021, vol. 1734, no. 1: IOP Publishing, p. 012026.

[14] P. Tiwari and V. Singh, "Diabetes disease prediction using significant attribute selection and classification approach," in Journal of Physics: Conference Series, 2021, vol. 1714, no. 1: IOP Publishing, p. 012013.

آرش خسروی مدرک کارشناسی خود را در رشته مهندسی نرم افزاردر سال ۱۳۸۲ از دانشگاه صنعتی اصفهان، مدرک ارشناسی ارشدخود را در رشته مهندسی فناوری اطلاعات در سال ۹۲ و مدرک دکتری خود را در رشته مهندسی فناوری اطلاعات، گرایش سیستمهای اطلاعاتی در سال ۹۶ از دانشگاه صنعتی مالزی اخذ کرده است. ایشان در حال حاضر به عنوان هیات علمی مرکز آموزش عالی محلات مشغول به کار هستند. زمینه های پژوهشی مورد علاقه ایشان عبارتند از: هوش تجاری، سیستم های پیشنهاد دهنده، مدیریت دانش مشتری، داده کاوی، متن کاوی و فناوری اطلاعات درپزشکی.

Khosravi.280@gmail.com

محمد هادی زاده
مدرک کارشناسی را از دانشگاه علمی کاربردی و کارشناسی ارشد را در رشته هوش مصنوعی از دانشگاه شهاب دانش قم اخذ کرده است و زمینه پژوهش و علایق ؛هوش مصنوعی و رایانش و سیستم های توزیع شده می باشد.
آدرس رایانه : mohdhadi1399@gmail.com

روش ارجاع به مقاله :م. هادی زاده، آ. خسروی. تشخیص بیماری دیابت بر اساس روش کاهش ویژگی هوشمند و یادگیری ترکیبی ماشینی. دوفصلنامه محاسبات و سامانه های توزیع شده سال چهارم، شماره اول، شماره پیاپی ۷، صفحه ۷۸ تا ۸۹، سال ۱۴۰۰.

How to cite: Mhamad hadizade , Arash. khosravi. Provide diagnosis of diabetes based on intelligent feature reduction and machine learning. Journal of Distributed Computing and Systems(JDCS), Vol 4, Issue 1, Page 78-89, 2021.