Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2023; 10(2): 125-140
10.34172/jhbmi.2023.19

A Pipeline-based Framework for Early Prediction of Diabetes

Abnoosian Karlo[1], Farnoosh Rahman[2]*, Behzadi Mohammad Hassan[3]

• **Received:** 31 Jan 2023 • **Accepted:** 10 Sep 2023

Introduction: Diabetes is a chronic disease worldwide, with an increasing annual death rate. Many health professionals seek innovative ways to detect and treat it early. Rapid advances in machine learning have improved disease diagnosis. However, because of the small amount of labeled data, the frequency of null and missing values, and the imbalance of databases, creating an optimal predictor for disease diagnosis has become a great challenge. This study aimed to present a pipeline-based classification framework for predicting diabetes on two datasets of Indian diabetic patients with two classes (patient and healthy groups) and Iraqi with three classes (patient, healthy, and prediabetes groups).

Method: An important part of this framework is preprocessing. Different ML models based on the One-Vs-One approach for the three-class mode are implemented in the framework. Because of the imbalance of the data set, besides the accuracy evaluation criterion, the area under the receiver operating characteristic (ROC) curve is also used. To increase the level of these two criteria, the Hyper-parameters of each model are optimized using optimization methods to build a powerful model with less training and testing time through various feature selection methods.

Results: The proposed framework was assessed for diabetes prediction on two datasets of Indian and Iraqi diabetic patients. It was revealed that using AdaBoost for the Indian dataset (ACC=89.98, AUC=94.11) and random forest for the Iraqi dataset (ACC=98.66, AUC= 98.62), good accuracy and performance were obtained.

Conclusion: Regarding ACC parameters, precision, accuracy, recall, and F1-Score, the pipeline-based framework has an optimal performance in predicting diabetes, therefore, it can be used in clinical decision support systems.

Keywords: Diabetes Prediction, Machine Learning, Classification, Pipeline, Feature Selection, The Area Under the Receiver Operating Characteristic Curve (AUC)

• **Citation:** Abnoosian K, Farnoosh R, Behzadi MH. A Pipeline-based Framework for Early Prediction of Diabetes. Journal of Health and Biomedical Informatics 2023; 10(2): 125-40. [In Persian] doi:10.34172/jhbmi.2023.19

1. Department of Statistics, School of Convergent Sciences and Technoligies, Science and Research Branch, Islamic Azad University, Tehran, Iran
2. School of Mathematics, Iran University of Science and Technology, Tehran, Iran
3. Department of Statistics, School of Convergent Sciences and Technoligies, Science and Research Branch, Islamic Azad University, Tehran, Iran
***Corresponding Author:** Rahman Farnoosh
Address: Department of Applied Mathematics, Iran University of Science and Technology, University St., Hengam St., Resalat Tehran, Iran
• **Tel:** 021-73225427 • **Email:** rfarnoosh@iust.ac.ir

[DOI: 10.34172/jhbmi.2023.19]

مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره دهم شماره دوم. ص ۱۴۰-۱۲۵، ۱۴۰۲
10.34172/jhbmi.2023.19

چارچوبی مبتنی بر خط لوله برای پیش‌بینی زود هنگام بیماری دیابت

کارلو آبنوسیان[۱]، رحمان فرنوش[۲]*، محمدحسن بهزادی[۳]

• دریافت مقاله: ۱۴۰۱/۱۱/۱۱ • پذیرش مقاله: ۱۴۰۲/۶/۱۹

مقدمه: دیابت یک بیماری مزمن است و میزان مرگ و میر آن در حال افزایش است. متخصصان سلامت به دنبال راهکارهای نوآورانه برای تشخیص و درمان زودهنگام آن هستند. پیشرفت‌های یادگیری ماشینی تشخیص بیماری را بهبود داده است. با این حال، به دلیل کمبود داده‌های برچسب‌گذاری شده، مقادیر ناقص و نامتعادل بودن داده‌ها، ایجاد یک پیش‌بین بهینه برای تشخیص بیماری به یک چالش بزرگ تبدیل شده است. هدف این مطالعه ارائه یک چارچوب طبقه‌بندی مبتنی بر خط لوله برای تشخصیص دیابت در دو مجموعه داده هندی (دو کلاس: بیمار و سالم) و عراقی (سه کلاس: بیمار، سالم و در شرف ابتلا به دیابت) است.

روش: بخش مهم این چارچوب پیش‌پردازش است. مدل‌های مختلف یادگیری ماشین مبتنی بر رویکرد One-Vs-One برای حالت سه‌کلاسه، در چارچوب پیشنهادی پیاده‌سازی شده‌اند. به دلیل نامتعادل بودن مجموعه داده، علاوه بر معیار ارزیابی دقت طبقه‌بندی، مساحت زیر منحنی مشخصه عملکرد گیرنده نیز استفاده می‌شود. با هدف افزایش مساحت زیر منحنی مشخصه عملکرد گیرنده و دقت طبقه‌بند، فراپارامترهای هریک از مدل‌ها با روش‌های بهینه‌سازی جستجوی شبکه‌ای و بیزین بهینه‌سازی می‌شوند برای ساختن مدلی قدرتمند با زمان کم آموزش و آزمایش از روش‌های مختلف انتخاب ویژگی استفاده می‌شود.

نتایج: از طریق شبیه‌سازی، چارچوب پیشنهادی برای تشخیص بیماری دیابت در دو مجموعه داده هندی و عراقی مورد آزمایش قرار گرفت. نتایج نشان داد که با استفاده از AdaBoost در مجموعه داده هندی (AUC=۹۴/۱۱ ، ACC= ۸۹/۹۸) و با استفاده از جنگل تصادفی در مجموعه داده عراقی (AUC=۹۸/۶۲ ، ACC= ۹۸/۶۶)، دقت و عملکرد مطلوبی به دست آمد.

نتیجه‌گیری: از نظر معیارهای ACC، دقت، صحت، یادآور و F1-Score، چارچوب پیشنهادی مبتنی بر خط لوله عملکرد بهینه‌ای دارد و می‌تواند در سامانه‌های پزشکی به عنوان یک برنامه کاربردی مورد استفاده قرار گیرد.

کلیدواژه‌ها: پیش‌بینی بیماری دیابت، یادگیری ماشین، طبقه‌بندی، خط لوله، انتخاب ویژگی، مساحت زیر منحنی مشخصه عملکرد

•ارجاع: آبنوسیان کارلو، فرنوش رحمان، بهزادی محمدحسن. چارچوبی مبتنی بر خط لوله برای پیش‌بینی زود هنگام بیماری دیابت. مجله انفورماتیک سلامت و زیست پزشکی ۱۴۰۲؛ ۱۰(۲): ۴۰-۱۲۵. doi:10.34172/jhbmi.2023.19

۱. گروه آمار، دانشکده علوم و فناوری‌های همگرا، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
۲. دانشکده ریاضی، دانشگاه علم و صنعت ایران، تهران، ایران
۳. گروه آمار، دانشکده علوم و فناوری‌های همگرا، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
* **نویسنده مسئول:** رحمان فرنوش
آدرس: تهران، رسالت، خیابان هنگام، خیابان دانشگاه، دانشکده ریاضی دانشگاه علم و صنعت ایران، گروه ریاضی کاربردی
• **شماره تماس:** ۷۳۲۲۵۴۲۷ - ۰۲۱ • **Email:** rfarnoosh@iust.ac.ir

مقدمه

به‌دلیل تغییر شیوه زندگی و پیشرفت صنعتی شدن، بیماری‌های مزمن و غیرواگیر به تهدیدی جدی برای سلامتی تبدیل شده است که روز به روز در حال افزایش است [۱،۲]، در این بین، دیابت یکی از مهم‌ترین تهدیدکننده‌های حیات انسان محسوب شده و عوارض و ناتوانی ناشی از آن بار سنگینی به خانواده‌ها، جوامع و دولت‌ها تحمیل می‌کند و از اولویت‌های اصلی برنامه‌های پیشگیری و درمانی محسوب می‌شوند [۳،۴]. بیماری دیابت، ناشی از اختلال‌های متابولیک است. بر اساس آمار، دیابت در سال ۲۰۱۷ بر ۴۵۲ میلیون نفر در سراسر جهان اثر گذاشته است و پیش‌بینی می‌شود این رقم تا سال ۲۰۴۵ به ۶۹۴ میلیون نفر برسد. برخی از مطالعات علمی دیگر نشان داده است که نیم میلیارد نفر در سراسر جهان به بیماری دیابت مبتلا بودند، پیش‌بینی می‌شود در سال‌های ۲۰۳۰ و ۲۰۴۵ به ترتیب به ۲۵٪ و ۵۱٪ افزایش یابد [۸-۵]. این بیماری به یکی از چالش‌های جدی پیش‌رو در کشورهای توسعه یافته و در حال توسعه تبدیل شده است. این بیماری به دلیل عدم ترشح کافی انسولین و یاعدم توانایی بدن در استفاده از انسولین و یا مقاومت بدن به انسولین به وجود می‌آید [۱]. دیابت قندی (Diabetes mellitus) به دو نوع اولیه و ثانویه تقسیم می‌شود، دیابت اولیه خود به سه نوع دیابت نوع ۱، دیابت نوع ۲ و دیابت بارداری تقسیم می‌شود [۸،۹]. دیابت نوع ۱ (وابسته به انسولین) بیشتر در کودکان و جوانان دیده می‌شود [۱]؛ اما ممکن است در هر سنی دیده شود و درصد کمی از افراد مبتلا به دیابت به این نوع مبتلا هستند. علت آن فقدان تولید انسولین می‌باشد و شیوع این نوع دیابت در جوامع مختلف حدوداً ۵٪ است. افراد برای درمان این نوع بیماری باید انسولین تزریق کنند [۱۰]. بیش از ۹۰٪ موارد دیابت قندی دیابت نوع ۲ (غیر وابسته به انسولین)، هستند که در سنین بالا و به صورت آهسته و تدریجی عارض می‌شود [۱۱،۱۲]. علت آن کمبود تولید انسولین یا مقاومت بدن به انسولین می‌باشد. شیوع این نوع دیابت حدوداً ۹۵- ۹۰ درصد است. در ایجاد این بیماری، عوامل ژنتیک و محیطی دخالت دارند. ۱۰ تا ۲۰ سال قبل از تشخیص دیابت نوع ۲، کاهش تحمل گلوکز همراه با افزایش جبرانی انسولین وجود دارد [۲]. عامل مهم دیگر در بروز دیابت نوع ۲، افزایش وزن بدن و چاقی است. در ابتدای شروع بیماری کاهش وزن سبب می‌شود که تحمل به گلوکز بهبود یابد [۳]. دیابت فقط یک بیماری محسوب نمی‌شود بلکه شبکه درهم پیچیده‌ای از عوامل خطرزای محیطی و ژنتیک با پاتوفیزیولوژی‌های مختلف است که به‌دلیل ویژگی‌هایی نظیر زمینه‌سازی و همراهی با سایر بیماری‌ها (قلبی- عروقی، کلیوی، چشمی و معلولیت) به شدت هزینه‌بر و ناتوان کننده است [۵]. با توسعه استانداردهای زندگی، دیابت به طور فزاینده‌ای در زندگی روزمره مردم رایج است و هیچ درمان طولانی‌مدت برای آن وجود ندارد و فقط باید در سطح مراقب‌های بهداشتی، در مراحل اولیه، عوامل زمینه‌ساز این بیماری را تشخیص داد و تشخیص پیش‌دیابت می‌تواند بسیار مهم تلقی شود و تا اینکه از این بیماری جلوگیری کرد و یا در مراحل بعدی این بیماری، کنترل و درمان آسانتر شود و هرچه زودتر تشخیص داده شود، می‌تواند باعث کاهش عواقب جبران ناپذیر و هزینه‌بر این بیماری شود؛ بنابراین چگونگی تشخیص و تجزیه و تحلیل سریع و دقیق دیابت موضوعی است بحث برانگیز، که ارزش مطالعه دارد.

در سال‌های اخیر با رشد سریع یادگیری ماشین، تلاش‌های قابل توجهی برای افزایش کاربردهای تشخیص به کمک رایانه شده است [۲]. یادگیری ماشین می‌تواند با توجه به داده‌های معاینه فیزیکی بیماران طی دوره‌های مختلف زمانی به تشخیص بیماری بپردازد و به‌عنوان یک مرجع برای پزشکان تلقی شود [۶]. تاکنون مدل‌های مختلفی برای پیش‌بینی دیابت ارائه شده است که در جدول ۱ خلاصه برخی از آن‌ها بیان شده است.

[DOI: 10.34172/jhbmi.2023.19]

جدول ۱: خلاصه‌ای از پژوهش‌های گذشته برای پیش‌بین دیابت

نویسندگان	نتیجه
Polat و Güneş [۷]	جدا سازی افراد سالم از دیابتی بر اساس مدلی مبتنی بر تحلیل مؤلفه‌های اصلی(PCA) و استنتاج عصبی، فازی
Yue و همکاران [۸]	ارائه مدلی برای پیش‌بینی دیابت براساس ترکیب الگوریتم بهینه‌سازی ازدحام زرات (QPSO) و SVM با حداقل مربعات وزن
Çalişir و Doğantekin [۹]	پیشنهاد سیستمی برای پیش‌بینی دیابت و لستفاده از تحلیل تشخیص خطی برای کاهش ابعاد و استخراج ویژگی برروی مجموعه داده‌هایی با ابعاد بالا
Beloufa و Chikh [۱۰]	ارایه مدلی ترکیبی برای پیش‌بینی دیابت با استفاده از ترکیب الگوریتم‌های کلونی زنبور عسل و سیستم فازی
Razavian و همکاران [۱۱]	پیشنهاد مدلی رگرسیون خطی با شروع‌های مختلف برای پیش‌بینی بیماری دیابت
Hassan و همکاران [۱۲]	ارائه مدلی ترکیبی برای پیش‌بینی دیابت با استفاده از ترکیب الگوریتم‌های کلونی K-NN وID3
Shirali و همکاران [۱۳]	ارائه مدلی ترکیبی برای پیش‌بینی دیابت با استفاده از ترکیب سیستم استنتاج فازی و الگوریتم کرم شبتاب
Maniruzzaman و همکاران [۱۴]	پیشنهاد یک چارچوب یادگیری ماشین مبتنی بر تحلیل خطی برای طبقه‌بندی دقیق خطر ابتلا به دیابت
Abaker و Saeed [۱۵]	ارائه سیستمی با هدف پیش‌بینی وضعیت سلامتی مراجعین به بیمارستان، استفاده از انتخاب ویژگی متوالی برای پیدا کردن حداقل ویژگی‌ها در الگوریتم‌های یادگیری ماشین

هدف اصلی این پژوهش، ایجاد یک چارچوب مبتنی بر خط لوله برای پیش‌بینی بیماری دیابت برروی دو مجموعه داده، بیماران دیابت هندی (Pima Indians Diabetes Dataset)PID و عراقی (Iraqi Patient Dataset for Diabetes) IPDD انتخاب شد، در مجموعه داده PID بیماران در ۲ کلاس بیماران دیابتی و سالم و در مجموعه داده IPDD بیماران در ۳ کلاس بیماران دیابتی، سالم و در شرف ابتلا به دیابت قرار دارند. پیش‌پردازش بخش مهمی از چارچوب پیشنهادی، جهت رسیدن به یک نتیجه با کیفیت بالا می‌باشد که شامل حذف افزونگی، تبدیل ویژگی، آزمایش نرمال بودن ویژگی‌ها، رد داده پرت، پرکردن مقادیر تهی یا غلط، استانداردسازی، نرمال‌سازی، انتخاب ویژگی می‌باشد. با مشورت پزشک با استفاده از الگوریتم k-NN Imputation، تعداد کمی از مقادیر ویژگی‌های نادرست مجموعه داده IPDD تصحیح شدند. برای مجموعه داده PID، از میانه برای پر کردن مقادیر غلط و تهی استفاده شد. مدل‌های یادگیری ماشین بیز ساده گواسی (Gaussian Naive Bayes)، K–نزدیک‌ترین همسایه (K-Nearest Neighbors)، درخت تصمیم (Decision tree)، جنگل تصادفی (random forest)، PID و ماشین بردار پشتیبان (support vector machine)، در حالت طبقه‌بند دودویی برای مجموعه داده PID و برای مجموعه داده IPDD براساس رویکرد یک‌به‌یک OVO(One-Vs-One) در چارچوب پیشنهادی استفاده شدند. از روش‌های جستجوی شبکه‌ای (Grid Search) و بهینه‌سازی بیزی (Bayesian Optimization) در اعتبارسنجی متقابل برای یافتن ابرپارامترهای بهینه در هر مدل استفاده شد. از طرفی به‌دلیل تفاوت معنادار از نظر تعداد نمونه‌های طبقه‌ها و نامتعادل بودن مجموعه داده‌ها، علاوه بر معیار دقت طبقه‌بندی از AUC برای ارزیابی و مقایسه درست مدل‌ها استفاده شد. تحت شرایط آزمایشی و مجموعه داده یکسان، آزمایش‌های متعددی با ترکیب‌های پیش‌پردازش و مدل‌های مختلف انجام شد تا معیارهای دقت و AUC را به حداکثر برساند. سپس مدل بهینه به عنوان مدل پایه برای چارچوب پیشنهادی، برای پیش‌بینی بهینه استفاده شد. انتخاب ویژگی یکی از مهم‌ترین مراحل در تحقیقات تشخیص بیماری است و اهمیت آن در ساخت مدلی با تعداد ویژگی‌های کم و ساده که نیازمند زمان کم برای آموزش و آزمایش و مخصوصاً قدرتمند در پیش‌بینی بیماری می‌باشد. از روش‌های کاهش ابعاد مانند تحلیل مؤلفه‌های اصلی PCA(Principal Component Analysis) تحلیل مؤلفه‌های مستقل (Independent Component Analysis)ICA و انتخاب ویژگی براساس اهمیت جایگشت ویژگی (Feature Selection Based Feature Importance)FSBFI با استفاده از هریک از مدل‌ها، استفاده شد.

روش

مجموعه داده‌ها

در این پژوهش از دو مجموعه داده PID و IPDD استفاده شد (لینک **۱** و **۲**)، مجموعه داده PID از ۷۶۸ نمونه خانم مبتلا

[DOI: 10.34172/jhbmi.2023.19]

به دیابت نوع ۱ و سالم، از جمعیت هندی ساکن در نزدیکی شهر فینیکس مرکز ایالت آریزونا به‌دست آمد که ۲۶۸ مورد آن بیمار دیابتی (مثبت) و ۵۰۰ مورد آن فرد سالم (منفی)، با هشت ویژگی متفاوت است (جدول ۲).

مجموعه داده

1. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
2. https://data.mendeley.com/datasets/wj9rwkp9c2/1

جدول ۲: خلاصه‌ای از مجموعه داده بیماران هندی

عنوان انگلیسی ویژگی	توضیحات		Mean ± Std
Pregnant	Number of times pregnant	تعداد دفعات وضع حمل (برای زنان)	۶۹/۲۸۰±۶۲/۲۱۶
Glucose	Plasma glucose concentration a 2 Hours in an Oral Glucose Tolerance Test	غلظت گلوکز پلاسمای خون	۶۹/۲۸۰±۶۲/۲۱۶
Pressure	Diastolic blood pressure (mm Hg)	فشار خون دیاستولیک	۶۹/۲۸۰±۶۲/۲۱۶
Triceps	Triceps skin fold thickness (mm)	ضخامت پوست ماهیچه سه سر بازویی	۶۹/۲۸۰±۶۲/۲۱۶
Insulin	2-Hour serum insulin (µU/ml)	انسولین سرم دوساعته	۶۹/۲۸۰±۶۲/۲۱۶
BMI	Body Mass Index (Weight in kg/(Height in inches)^2)	نمایه توده بدنی	۶۹/۲۸۰±۶۲/۲۱۶
Pedigree	Diabetes Pedigree Function	داشتن سابقه دیابت	۶۹/۲۸۰±۶۲/۲۱۶
Age	Age in years	سن	۶۹/۲۸۰±۶۲/۲۱۶

مجموعه داده IPDD از ۱۰۰۰ نمونه زن و مرد مبتلا به دیابت نوع ۲ و سالم، ۲۰ الی ۷۹ ساله که از معاینات فیزیکی مرکز تخصصی غدد درون ریز و متابولیسم بیمارستان الکیندی در عراق به‌دست آمد که ۸۳۷ مورد آن بیمار دیابتی، ۱۰۳ مورد آن فرد سالم و ۵۳ مورد آن فرد در شرف ابتلا به دیابت، با ۱۱ ویژگی متفاوت است(جدول ۳).

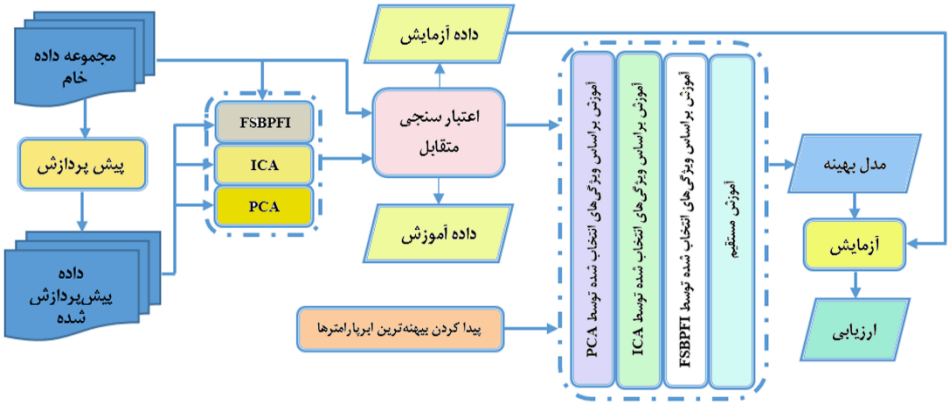
جدول ۳: خلاصه‌ای از مجموعه داده بیماران عراقی

عنوان انگلیسی ویژگی	توضیحات		Mean ± Std
Gender	0 for females and 1 for male	۰ مؤنث و ۱ مذکر	-
Age	Age in years	سن	۸/۸۵۵۷±۵۳/۷۳۹
FBS	The result of a blood sugar taken after a patient fasted for at least eight hours (mmol/l)	نتیجه قند خون پس از حداقل ۸ ساعت ناشتا	۱۰/۱۴۴۳±۵/۰۸۴۴
BUN	BUN is the amount of urea nitrogen that's in your blood (mmol/l)	مقدار نیتروژن اوره موجود در خون	۵/۱۸۰۸±۳/۳۴۸۶
Cr	The Blood levels of chromium (mmol/l)	سطح کروم خون	۶۹/۲۸±۶۲/۲۷۶۴
Chol	The Fast Cholesterol levels (mmol/l)	سطح کلسترول خون	۴/۹۰۹۲±۲/۰۰۴
TG	The Tri Glycoside Levels (mmol/l)	سطح تری گلیسیرید موجود در خون	۲/۳۵۰۶±۱/۳۹۸۸
LDL	The Low-Density Lipoprotein (mmol/l)	سطح لیپوپروتئین با چگالی کم در خون	۲/۶۱۴۵±۱/۱۱۷۵
HDL	The High-Density Lipoprotein (mmol/l)	سطح لیپوپروتئین با چگالی بالا در خون	۱/۲۰۶۷±۰/۶۵۹۴
BMI	The Body Mass Index (Weight in kg / (Height in m)^2)	نمایه توده بدنی	۲۹/۴۲۵۵±۴/۸۵۵۳
HbA1c	The For the previous two to three months, average blood glucose (sugar) levels (mmol/l)	میانگین گلوکز پلاسما را طی ۸ تا ۱۲ هفته گذشته	۸/۲۶۲۳±۲/۵۳۷۰

[DOI: 10.34172/jhbmi.2023.19]

چارچوب پیشنهادی

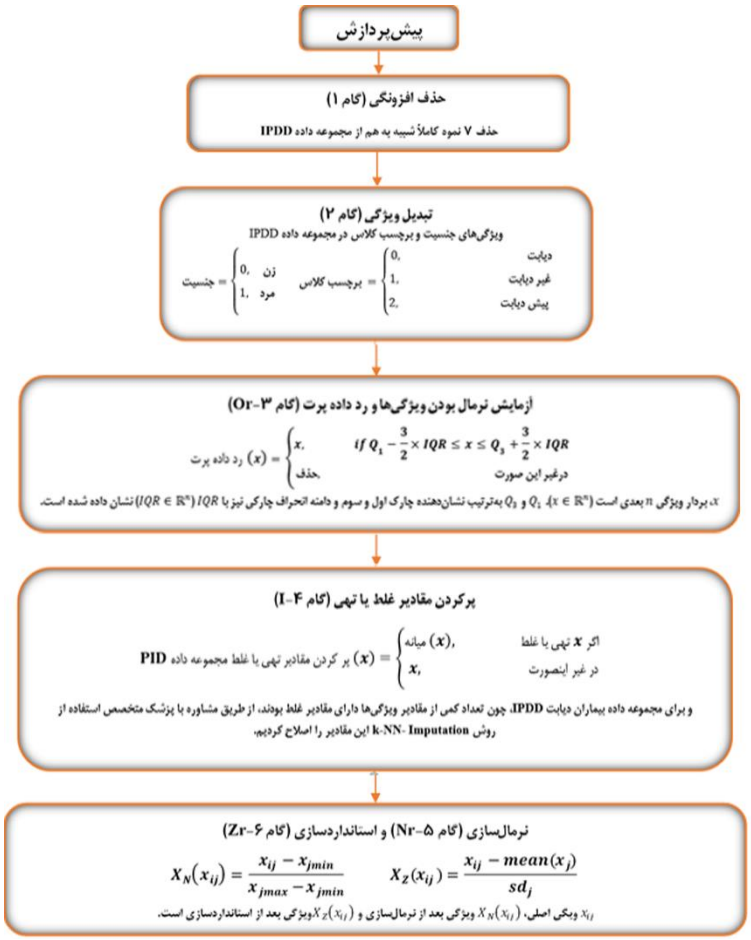
چارچوب پیشنهادی در این مطالعه در شکل ۱ نشان داده شده است.



شکل ۱: چارچوب پیشنهادی برای پیش‌بینی بهینه بیماری دیابت

پیش‌پردازش

پیش‌پردازش داده‌ها اولین و مهم‌ترین مرحله در چارچوب پیشنهادی است، زیرا پیش‌پردازش می‌تواند به طور قابل توجهی کیفیت داده‌ها را بهبود بخشد و داده‌های باکیفیت تأثیر مستقیم بر عملکرد مدل‌های یادگیری ماشین دارد [۱۶]. مراحل پیش‌پردازش در شکل ۲ نمایش داده شده و در ادامه برخی از گام‌های مهم شرح داده شده است.

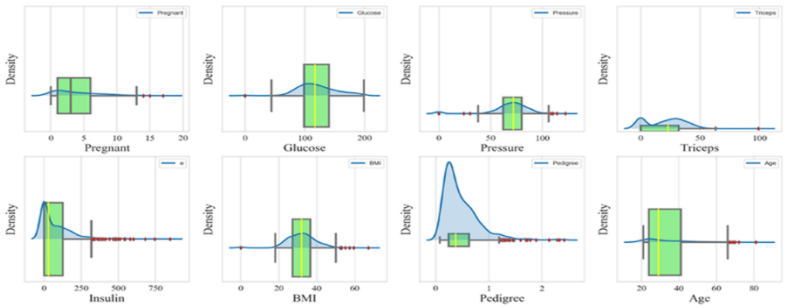


شکل ۲: گام‌های پیش‌پردازش

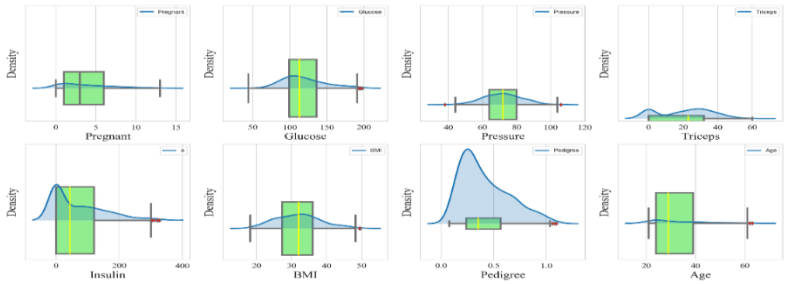
[DOI: 10.34172/jhbmi.2023.19]

• آزمایش نرمال بودن ویژگی‌ها و رد داده پرت (گام Or-۳)

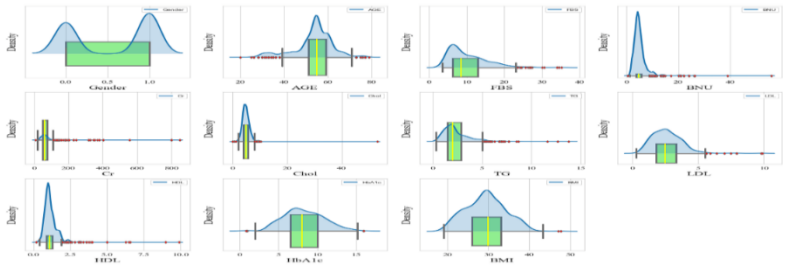
داده پرت، داده‌ای است که با دیگر داده‌های هم‌گروه فاصله چشمگیری داشته باشد [۱۷]. و باید از توزیع داده‌ها حذف شوند؛ زیرا طبقه‌بندها به محدوده داده و توزیع ویژگی حساس هستند.



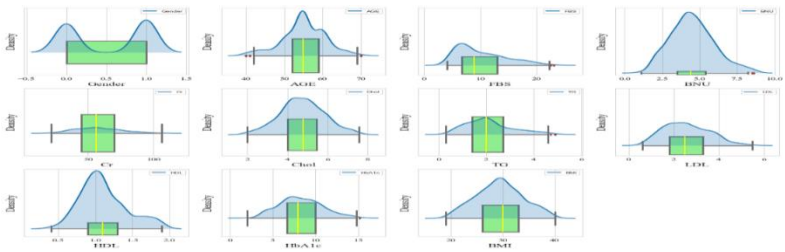
الف. مجموعه PID در حضور داده پرت



ب. مجموعه داده PID بعد از حذف داده پرت



ج. مجموعه داده IPDD در حضور داده پرت



د. مجموعه داده IPDD بعد از حذف داده پرت

شکل ۳: نمودار جعبه‌ای توزیع ویژگی‌های مجموعه داده‌ها قبل و بعد از حذف داده پرت

[DOI: 10.34172/jhbmi.2023.19]

با توجه به قسمت‌های الف و ج از شکل ۳ مشاهده شد که، وجود داده پرت باعث شده است که، اکثر ویژگی‌ها در هر دو مجموعه داده دارای چولگی شوند. وجود چولگی باعث برآورد مقدار پیش‌بینی شده کم و یا زیاد می‌شود. حذف داده پرت توزیع چوله را به سمت توزیع با میانگین صفر انتقال می‌دهد (شکل ۳، قسمت‌های ب و د)، که نشان‌دهنده نزدیکی میانگین و میانه در توزیع است. گام سوم (شکل ۲) در پیش‌پردازش فرآیند حذف داده پرت را نشان می‌دهد.

- **پرکردن مقادیر غلط یا تهی (گام I-٤)**

مقادیر گم‌شده (غلط) یا تهی [۱۸] مقادیری هستند که ممکن است منجر به پیش‌بینی یا استنتاج نادرست برای هر کلاس در طبقه‌بندی شوند با مشورت پزشک با استفاده از الگوریتم k-NN Imputation، با مقدار K= ۴ تعداد کمی از مقادیر ویژگی‌های نادرست مجموعه داده IPDD تصحیح شد. برای مجموعه داده PID، از میانه برای پر کردن مقادیر غلط و تهی استفاده شد که این مراحل در گام چهارم (شکل ۲) در پیش‌پردازش نشان داده شده است.

- **نرمال‌سازی (گام Nr-٥) و استانداردسازی (گام Zr-٦)**

از آن جا که برخی از ویژگی‌های پیوسته دارای برد گسترده‌ای هستند، نرمال‌سازی و استانداردسازی می‌تواند تأثیر بسزایی بر عملکرد طبقه‌بند داشته باشد. برای تبدیل برد ویژگی‌ها به بازه [0.1] از نرمال‌سازی min-max [۱۹] و برای استانداردسازی ویژگی‌ها با مقادیر عددی پیوسته به ویژگی‌هایی با میانگین صفر و انحراف معیار یک، از Z-score استفاده شد (گام‌های ۵ و ۶ پیش‌پردازش شکل ۲) [۲۰].

اعتبارسنجی متقابل

اعتبارسنجی متقابل یک روش آماری برای ارزیابی و مقایسه کارایی مدل‌ها است که داده‌ها را به دو بخش تقسیم می‌کند: یک بخش برای آموزش و بخش دیگر برای اعتبارسنجی یا آزمایش است.

شکل ٤: نمودار تقسیم‌بندی مجموعه داده‌ها، برای تنظیم ابرپارامترها و ارزیابی

در اعتبارسنجی متقابل، داده‌ها به k بخش مساوی (یا تقریباً مساوی) تقسیم می‌شوند. سپس، k بار آموزش و اعتبارسنجی تکرار می‌شود، در هر بار تکرار یک بخش از داده‌ها برای اعتبارسنجی و K-1 بخش باقی مانده برای آموزش استفاده می‌شود [۲۱]. در این پژوهش، در حلقه داخلی الگوریتم‌های بهینه‌سازی فراپارامتر (بهینه‌سازی بیزی و جستجوی شبکه‌ای) اعمال می‌شود [۲۲،۲۳]، ابرپارامترها با استفاده از ۴ بخش دقیق آموزش و تنظیم شدند. داده‌های آزمایش برای ارزیابی مدل در حلقه بیرونی با استفاده از فراپارامترهای بهینه، پنج بار تکرار می‌شوند (شکل ۴).

انتخاب ویژگی

استراتژی کاهش بعد و انتخاب ویژگی می‌تواند به کاهش ابعاد و اجتناب از استفاده از ویژگی‌های اضافی کمک کند. در این پژوهش، از PCA [۲۴]، ICA [۲۵]، برای کاهش ابعاد و از FSBPFI [۲۶،۲۷] برای پیداکردن ویژگی‌های اساسی در هریک از مدل‌های یادگیری ماشین استفاده شد.

[DOI: 10.34172/jhbmi.2023.19]

مدل‌های یادگیری ماشین

مدل‌های مختلف طبقه‌بندی، در حالت دو کلاسه برای مجموعه داده PID و چندکلاسه برای مجموعه داده IPDD، K-نزدیک‌ترین همسایه (K-Nearest Neighbors) [۲۸]، ماشین بردار پشتیبان (support vector machine) دوکلاسه [۲۹] و چندکلاسه [۳۰]، مدل دوکلاسه [۳۱] و چندکلاسه AdaBoost [۳۲، ۳۳]، درخت تصمیم (Decision tree) [۳۴]، جنگل تصادفی (random forest) [۳۵]، ماشین بیز ساده گواسی (Gaussian Naive Bayes) [۳۶]، برای آموزش و آزمایش در چارچوب پیشنهادی استفاده شد. برای گسترش الگوریتم‌های طبقه‌بندی باینری به چند کلاسه از رویکرد یک‌به‌یک (One-Vs-One) OVO استفاده شد [۳۷] و ابرپارامترهای هر مدل با استفاده از روش‌های بهینه‌سازی Grid Search و Bayesian Optimization شد.

معیارهای ارزیابی

در مرحله آزمایش برای ارزیابی عملکرد مدل‌ها، از معیارهای ارزیابی برای حالت دو کلاسه و چندکلاسه با میانگین میکرو استفاده شد (جدول ۴) [۳۸، ۳۹].

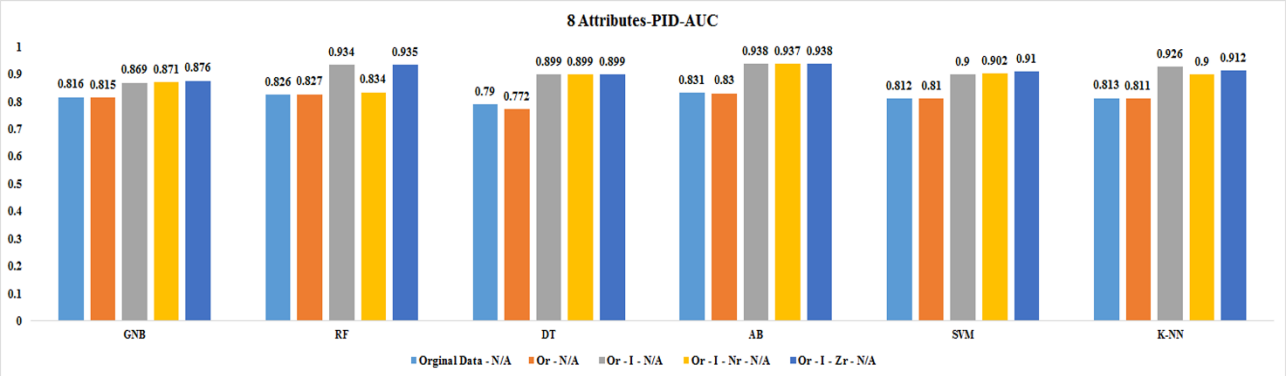
جدول ٤: خلاصه‌ای از معیارهای ارزیابی عملکرد مدل‌های پیش‌بینی بیماری دیابت

معیار ارزیابی	تعریف		دوکلاسه	چندکلاسه
دقت طبقه‌بند (دوکلاسه) و میانگین دقت طبقه‌بند (چند کلاسه) (Average Accuracy)	اندازه پیش‌بینی درست مدل را ارزیابی می‌کند.	(7)	$\frac{tp+tn}{tp+tn+fp+fn}$	$\frac{\sum_{i=1}^{k}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}}{k}$
صحت (Precision)	تعداد اسناد مثبت صحیح به کل اسناد مثبت صحیح و ناصحیح بازگردانده شده توسط مدل.	(8)	$\frac{tp}{tp+fp}$	$\frac{\sum_{i=1}^{k} tp_i}{\sum_{i=1}^{k} tp_i+fp_i}$
یادآوری (Recall)	تعداد اسناد مثبت صحیح به کل اسناد مثبت صحیح و منفی ناصحیح بازگردانده شده توسط مدل.	(9)	$\frac{tp}{tp+fn}$	$\frac{\sum_{i=1}^{k} tp_i}{\sum_{i=1}^{k}(tp_i+fn_i)}$
امتیاز F1	میانگین هارمونیک از دقت و یادآوری.		$\frac{2\times Precision\times Recall}{Precision+Recall}$	
AUC	معیار AUC-ROC، درباره قابلیت مدل در زمینه تشخیص کلاس‌ها به ما اطلاعات می‌دهد. هرچه AUC بالاتر باشد، مدل بهتر است.			

که در این جدول TN (True Negative)، TP (Positive True) و FP (False Positive) False) FN(Negative به‌ترتیب تعداد منفی‌های واقعی، مثبت‌های واقعی، مثبت‌های کاذب، منفی‌های کاذب و k تعداد کل کلاس‌ها (در اینجا برابر با ۳) را نشان می‌دهند.

نتایج

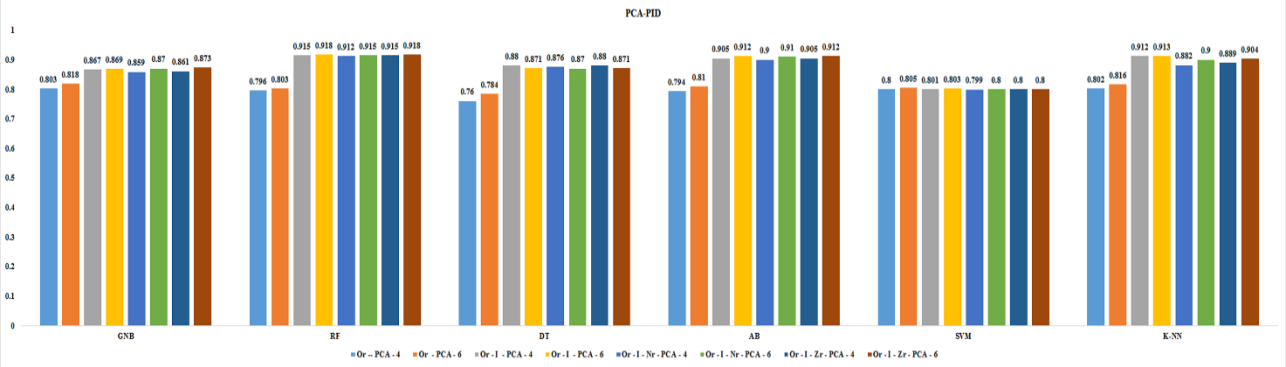
این بخش، نتایج کمی را برای انتخاب بهینه‌ترین پیش‌پردازش، کاهش ابعاد، انتخاب مؤثرترین ویژگی‌ها و مدل‌ها در خط لوله پیشنهادی نشان می‌دهد.



شکل ۵: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف با استفاده از همه ویژگی‌ها از نظر AUCبرای داده‌های بیماران دیابت هندی

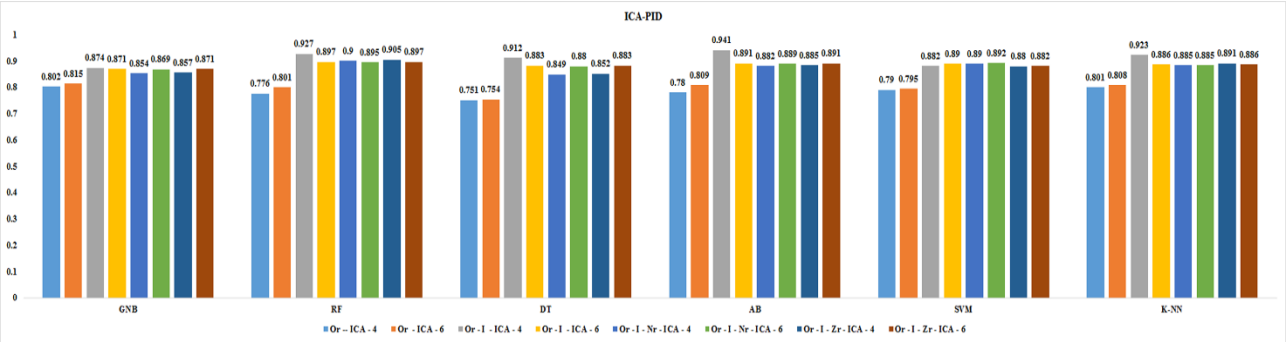
[DOI: 10.34172/jhbmi.2023.19]

برای مجموعه داده PID، زمانی که از همه ویژگی‌ها (۸ ویژگی) استفاده شد، مدل AB در دوحالت با استفاده از پیش‌پردازش‌های (گام‌های ۱ الی ۵) و استانداردسازی (گام‌های ۱ الی ۴ و ۶) بهینه‌ترین عملکرد را از نظر AUC برابر با ۰/۹۳۸، نسبت به سایر مدل‌ها کسب کند (شکل ۵).



شکل ٦: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف از نظر AUC با استفاده از PCA برای کاهش ابعاد برای داده‌های بیماران هندی

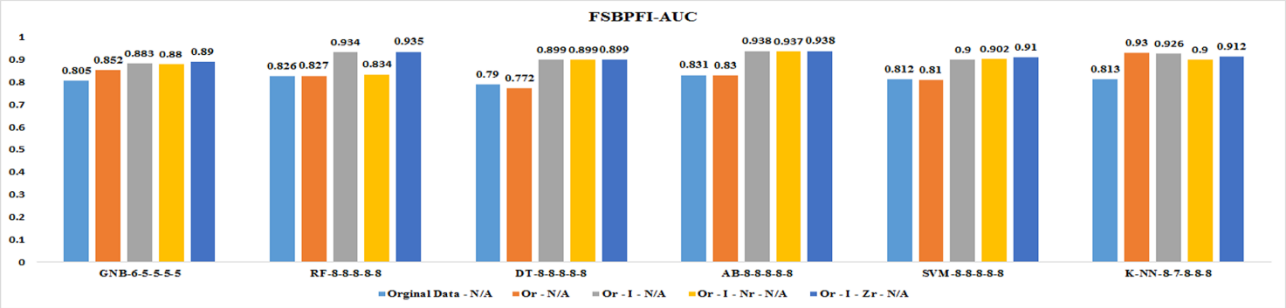
در این پژوهش از PCA با تحلیل واریانس در دوحالت واریانس ۹۵٪ و ۹۸٪ داده‌ها استفاده شد که برای مجموعه داده PID، ۸ ویژگی به ۴ و ۶ ویژگی به ترتیب کاهش داده شد و مدل RF در دوحالت پیش‌پردازش (گام‌های ۱ الی ۵) و PCA با ۶ ویژگی و (گام‌های ۱ الی ۴ و ۶) و PCA با ۶ ویژگی، توانست بهینه‌ترین عملکرد را از نظر AUC نسبت به سایر مدل‌ها کسب کند (AUC= ۰/۹۱۸) (شکل ۶).



شکل ۷: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف از نظر AUC با استفاده از ICA برای کاهش ابعاد برای داده‌های بیماران هندی

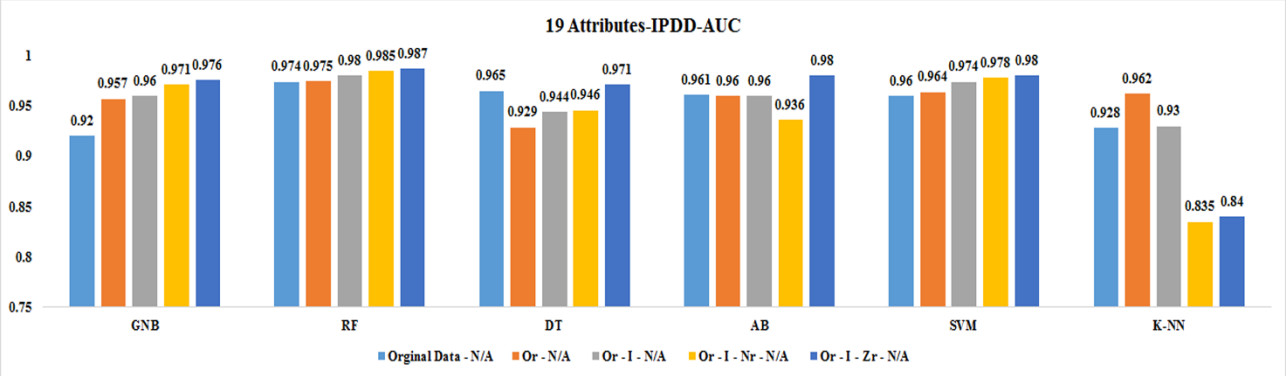
استفاده از الگوریتم ICA (شکل ۷) در دوحالت به ۴ و ۶ تعداد ویژگی انجامید، در این حالت مدل AB توانست با پیش‌پردازش (گام‌های ۱ الی ۴) با تعداد ۶ ویژگی مقدار AUC را نسبت به سایر مدل‌ها و در تمامی آزمایش‌ها مختلف کسب کند (۰/۹۴۱ =AUC).

[DOI: 10.34172/jhbmi.2023.19]



شکل ۸: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف از نظر AUC با استفاده از FSBPFI برای انتخاب ویژگی برای داده‌های بیماران هندی

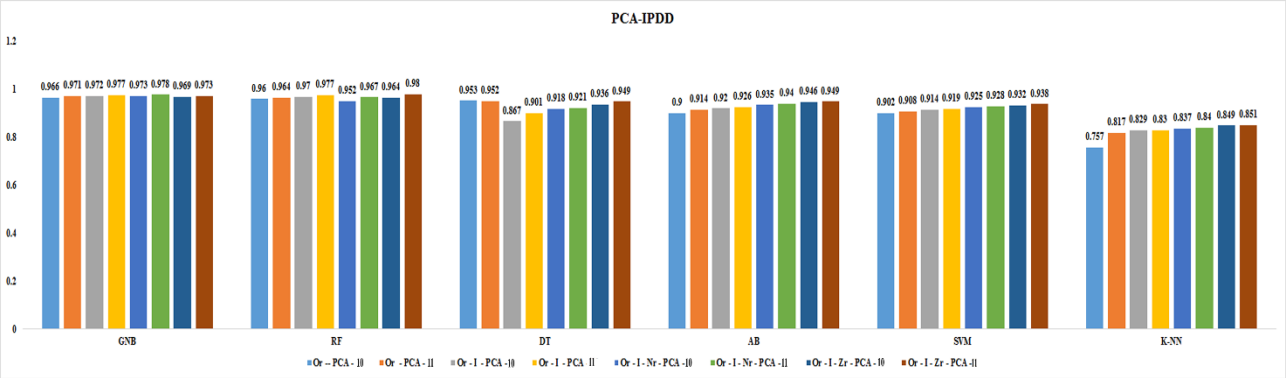
و در حالت استفاده از FSBPFI (شکل ۸) مدل AB توانست همان ۸ ویژگی یعنی تمام ویژگی‌ها را انتخاب کند و از نظر عملکرد بهینه‌ترین مقدار را نسبت به سایر مدل‌ها در حالت‌های استفاده از پیش‌پردازش‌های (گام‌های ۱ الی ۴) و (گام‌های ۱ الی ۴ و ۶) کسب کند (AUC= ۰/۹۳۸). FSBPFI با GNB توانست عملکرد این الگوریتم را نسبت به تمامی آزمایش‌های قبلی خودش بهبود دهد که بهینه‌ترین حالت این الگوریتم زمانی اتفاق می‌افتد که ۵ ویژگی با استفاده از پیش‌پردازش (گام‌های ۱ الی ۴) صورت پذیرد که در این حالت بیش‌ترین مقدار AUC یعنی ۰/۸۸۳ به‌دست می‌آید. الگوریتم K-NN نیز در حالت استفاده از انتخاب ویژگی براساس اهمیت جایگشت ویژگی با ۷ ویژگی و استفاده از پیش‌پردازش (گام‌های ۱ الی ۳) بهترین عملکرد را در سایر حالات استفاده از پیش‌پردازش‌های مختلف و روش‌های مختلف انتخاب ویژگی از خود نشان دهد. مدل‌های SVM، AB، DT، RF در تمام حالت‌های مختلف پیش‌پردازش همان ۸ ویژگی یعنی تمام ویژگی‌ها را انتخاب کردند.



شکل ۹: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف با استفاده از همه ویژگی‌ها از نظر AUCبرای داده‌های بیماران دیابت IPDD

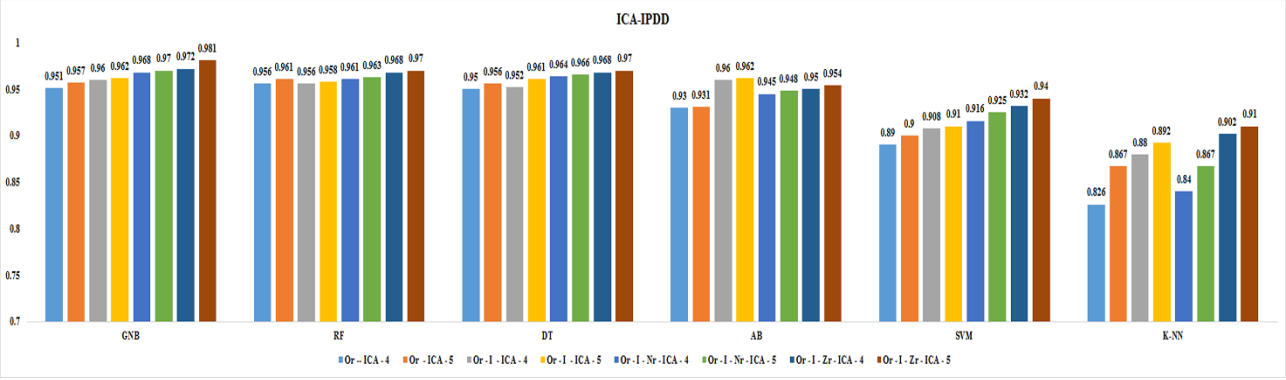
برای مجموعه داده IPDD، زمانی که از همه ویژگی‌ها (۱۱ ویژگی) استفاده شد، مدل RF توانست با استفاده از پیش‌پردازش (گام‌های ۱ الی ۴ و ۶) بهینه‌ترین عملکرد را از نظر AUC (AUC= ۰/۹۸۷)، نسبت به سایر مدل‌ها با تمام روش‌های انتخاب ویژگی و پیش‌پردازش‌ها، کسب کند (شکل ۹).

[DOI: 10.34172/jhbmi.2023.19]



شکل ۱۰: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف از نظر AUC با استفاده از PCA برای کاهش ابعاد برای داده‌های بیماران IPDD

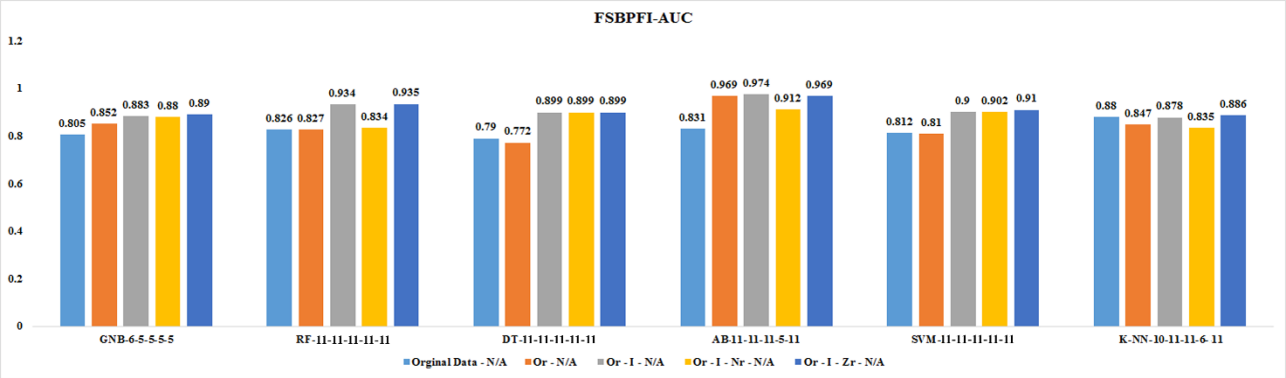
این مجموعه داده که دارای ۱۱ ویژگی است، استفاده از PCA منجر به ۱۰ و ۱۱ بعد برای هر دو واریانس ۹۵٪ و ۹۸٪ داده‌ها می‌شود که در این حالت مدل RF با پیش‌پردازش (گام‌های ۱ الی ۴ و ۶) و با ۱۱ وِیژگی توانست، بهینه‌ترین عملکرد را از نظر AUC نسبت به سایر مدل‌ها کسب نماید (AUC=۰/۹۸)، (شکل ۱۰). استفاده از PCA در با واریانس ۹۸٪ داده‌ها، ابعاد را کاهش نداد و همان ۱۱ ویژگی را برگرداند.



شکل ۱۱: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف از نظر AUC با استفاده از ICA برای کاهش ابعاد برای داده‌های بیماران IPDD

استفاده از ICA (شکل ۱۱) در دوحالت به ۴ و ۵ تعداد ویژگی انجامید، در این حالت مدل GNB توانست با پیش‌پردازش (گام‌های ۱ الی ۴ و ۶) و تعداد ۵ ویژگی بیش‌ترین مقدار AUC را نسبت به سایر مدل‌ها و در تمامی آزمایش‌ها کسب کند (AUC= ۰/۹۸۱).

[DOI: 10.34172/jhbmi.2023.19]



شکل ۱۲: نتایج حاصل از مقایسه مدل‌ها و پیش‌پردازش‌های مختلف از نظر AUC با استفاده از FSBPFI برای انتخاب ویژگی برای داده‌های بیماران IPDD

و در استفاده از FSBPFI (شکل ۱۲)، مدل RF توانست همان ۱۱ ویژگی (تمام ویژگی‌ها) را انتخاب کند و از نظر عملکرد بهینه‌ترین مقدار را با پیش‌پردازش (گام‌های ۱ الی ۴ و ۶) نسبت به سایر مدل‌ها، کسب کند (AUC= ۰/۹۳۵)، مدل‌های SVM، DT، RF در حالت‌های مختلف پیش‌پردازش همان ۱۱ ویژگی را انتخاب کردند. در نهایت جدول ۵، بهینه‌ترین نتایج آزمایش‌های مختلف را از نظر معیارهای ارزیابی ارائه می‌دهد.

جدول ۵: مقایسه مدل‌های بهینه حاصل از نظر معیارهای مختلف ارزیابی

مدل	مجموعه داده	پیش‌پردازش	دقت	صحت	یادآوری	امتیاز *F1*	*AUC*
RF	PID	گام‌های ۱ الی ۴ و ۶ و ۶- *ICA*	۸۹/۹۱	۹۰/۲۳	۹۱/۱۵	۸۹/۲۱	۹۳/۵۲
AB	PID	گام‌های ۱ الی ۴ و ۶- *ICA*	۸۹/۹۹	۹۰/۰۱	۹۰/۰	۸۸/۰۴	۹۴/۱۱
RF	IPDD	گام‌های ۱ الی ۴ و ۶ و ۱۲- *N/A*	۹۸/۶۶	۹۷/۴۴	۹۷/۲۳	۹۵/۲۱	۹۸/۶۲

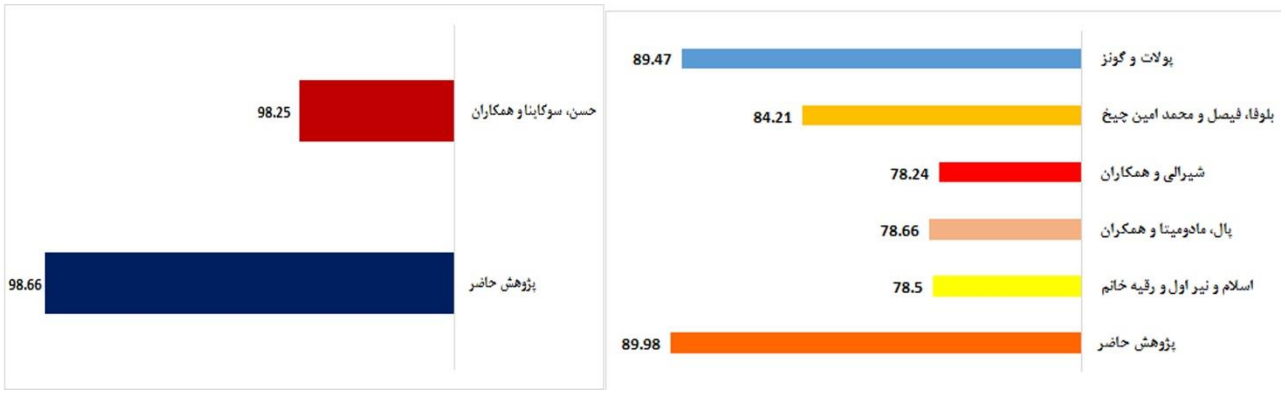
بحث و نتیجه‌گیری

چارچوب پیشنهادی ابتکاری بر اساس خط لوله برای پیش‌بینی دیابت برروی دو مجموعه داده، بیماران دیابت PID و IPDD در این پژوهش توانست موجب افزایش عملکرد مدل‌های دسته‌بندی نسبت به سایر روش‌های قبلی پیاده‌سازی شده در این پژوهش شود. تا کنون روش‌های مختلفی برای تشخیص بیماری دیابت بر اساس مجموعه داده بیماران دیابت PID و یک مورد برای بیماران IPDD پیشنهاد شده است که در جداول ۶ و نمودار ۱۳ به مقایسه این روش‌ها با چارچوب پیشنهادی این پژوهش پرداخته شد.

جدول ٦: مقایسه روش‌های مختلف پیاده‌سازی شده در تشخیص بیماری دیابت

پژوهشگران	پایگاه داده	روش پیاده‌سازی	دقت	سال
Polat و Güneş [۷]	PID	مدلی مبتنی بر PCA و استنتاج عصبی–فازی	٪۸۹/۴۷	۲۰۰۷
Beloufa و Chikh [۱۰]	PID	ترکیب الگوریتم کلونی زنبورعسل و سیستم فازی	٪۸۴/۲۱	۲۰۱۳
Shirali و همکاران [۱۳]	PID	ترکیب سیستم استنتاج فازی سوگنو و الگوریتم کرم شبتاب	٪۷۸/۲۴	۲۰۱۶
Pal و همکاران [۴۰]	PID	مدل یادگیری ماشین مبتین بر الگوریتم‌های تحلیل تشخیص خطی، k-NN، SVM، RF، برای پیش‌بینی زودهنگام بیماری دیابت نوع ۲ پیشنهاد دادند	٪۷۸/۶۶	۲۰۲۱
پژوهش حاضر	PID	چارچوب پیشنهادی مبتنی بر خط لوله براساس پیش‌پردازش‌های مختلف و انتخاب ویژگی	٪۸۹/۹۹	۲۰۲۳
Hassan و همکاران [۱۲]	IPDD	طراحی یک سیستم تشخیص دیابت از ترکیب الگوریتم‌های K-NN و ID3	٪۹۸/۲۵	۲۰۱۴
پژوهش حاضر	IPDD	چارچوب پیشنهادی مبتنی بر خط لوله براساس پیش‌پردازش‌های مختلف و انتخاب ویژگی	٪۹۸/۶۶	۲۰۲۳

[DOI: 10.34172/jhbmi.2023.19]



(الف) PID (ب) IPDD

شکل ۱۳: مقایسه روش‌های مختلف پیاده‌سازی شده برای پیش‌بینی بیماری دیابت از نظر دقت طبقه‌بندی

همان‌طور که در جدول ۶ و نمودار ۱۳ و مشاهده شد در حالت استفاده از مجموعه داده‌های PID چارچوب پیشنهادی توانست با عملکردی مناسب‌تر از نظر دقت (برابر با ۸۹/۹۸٪) دیگر افراد را به دو دسته بیمار و سالم طبقه‌بندی یا در شرف ابتلا به دیابت تقسیم کند و در هنگام استفاده از مجموعه داده IPDD نیز توانست نسبت به کارهای قبلی دقتی برابر با ۹۸/۶۶٪ کسب کند. نتایج شبیه‌سازی نشان داد شاخص دقت به‌دست‌آمده نسبت به سایر روش‌های شبیه‌سازی شده بر این مجموعه داده (جدول ۶ و نمودار ۱۳) بهبود یافته است. پیش‌بینی دیابت و یا پیش‌آگهی از دیابت می‌تواند باعث مراجعه زودهنگام و درمان به‌موقع و جلوگیری از عوارض ناشی از آن شود در واقع، چارچوب پیشنهادی براساس خط لوله می‌تواند برای بهبود تشخیص بیماری دیابت مورد استفاده قرار گیرد. از آنجایی که پیش‌پردازش داده‌ها و روش‌های انتخاب ویژگی تأثیر قابل توجهی بر عملکرد مدل‌ها دارد، توصیه می‌شود برای کارهای آینده از سایر روش‌های پیش‌پردازش داده‌ها و انتخاب ویژگی استفاده شود.

علاوه براین با گسترش چشمگیر کاربرد یادگیری عمیق در پیش‌بینی زودهنگام بیماری در حوزه سلامت و درمان، پیشنهاد می‌شود.

تشکر و قدردانی

خداوند سبحان را سپاسگزارم که به ما توفیق انجام و اتمام پژوهش حاضر را عنایت فرمود. از تمامی دوستان و اساتید به ویژه خانم دکتر میترا اسماعیلی آزاد متخصص دکترای تخصصی بهداشت مرکز تحقیقات پوست دانشکده علوم پزشکی شهید بهشتی که با راهنمایی‌های مناسب ما را در این پژوهش یاری کردند، کمال تقدیر و تشکر را داریم.

تعارض منافع

نویسندگان این مقاله اعلام می‌کنند که این پژوهش هیچ‌گونه تعارض منافعی ندارد.

References

1. Lonappan A, Bindu G, Thomas V, Jacob J, Rajasekaran C, Mathew KT. Diagnosis of diabetes mellitus using microwaves. Journal of Electromagnetic Waves and Applications 2007;21(10):1393-401. https://doi.org/10.1163/156939307783239429

2. Iancu I, Mota M, Iancu E. Method for the analysing of blood glucose dynamics in diabetes mellitus patients. In2008 IEEE International Conference on Automation, Quality and Testing, Robotics; 2008 May 22-25; Cluj-Napoca, Romania: IEEE; 2008. p. 60-5. doi: 10.1109/AQTR.2008.4588883

3. Robertson G, Lehmann ED, Sandham W, Hamilton D. Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. Journal of Electrical and Computer Engineering 2011;2011. https://doi.org/10.1155/2011/681786

4. Abnoosian K, Farnoosh R, Behzadi MH. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics 2023; 337. https://doi.org/10.1186/s12859-023-05465-z

[DOI: 10.34172/jhbmi.2023.19]

5. Otte C, Gold SM, Penninx BW, Pariante CM, Etkin A, Fava M, et al. Major depressive disorder. Nature reviews. Nat Rev Dis Primers 2016;2:16065. doi: 10.1038/nrdp.2016.65
6. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine 2022;128:102289. https://doi.org/10.1016/j.artmed.2022.102289
7. Polat K, Güneş S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing 2007;17(4):702-10. https://doi.org/10.1016/j.dsp.2006.09.005
8. Yue C, Xin L, Kewen X, Chang S. An intelligent diagnosis to type 2 diabetes based on QPSO algorithm and WLS-SVM. International Symposium on Intelligent Information Technology Application Workshops; 2008 Dec 21-22; Shanghai, China: IEEE; 2008. p. 117-21. doi: 10.1109/IITA.Workshops.2008.36
9. Çalişir D, Doğantekin E. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications 2011;38(7):8311-5. https://doi.org/10.1016/j.eswa.2011.01.017
10. Beloufa F, Chikh MA. Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer Methods and Programs in Biomedicine 2013;112(1):92-103. https://doi.org/10.1016/j.cmpb.2013.07.009
11. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 2015;3(4):277-87. https://doi.org/10.1089/big.2015.0020
12. Hassan S, Karbat AR, Towfik ZS. Propose Hybrid KNN-ID3 for Diabetes Diagnosis System. International Journal of Scientific & Engineering Research 2014;5(9):1087-104.
13. Shirali M, Madmoli Y, Roohafza J, Karimi H, Baboli Bahmaei A, Ertebati S. Improvement diagnosis of diabetes using a combination of sugeno fuzzy inference systems and firefly algorithms. Iranian Journal of Diabetes and Metabolism 2016;15(3):172-6. [In Persian]
14. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, Suri HS, Abedin MM, El-Baz A, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst 2018;42(5):92. doi: 10.1007/s10916-018-0940-7.
15. Abaker AA, Saeed FA. A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications. Informatica 2021;45(1).
doi: https://doi.org/10.31449/inf.v45i1.3111
16. Tegunov D, Cramer P. Real-time cryo-electron microscopy data preprocessing with Warp. Nature Methods 2019;16(11):1146-52.
17. Cousineau D, Chartier S. Outliers detection and treatment: a review. International Journal of Psychological Research 2010;3(1):58-67.
18. Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics 2012;13:1-9.
19. Ali PJ, Faraj RH, Koya E, Ali PJ, Faraj RH. Data normalization and standardization: a technical report. Mach Learn Tech Rep 2014;1(1):1-6.
20. Mohamad, Ismail Bin, and Dauda Usman. "Research Article Standardization and its Effects on K-Means Clustering Algorithm. Research Journal of Applied Sciences, Engineering and Technology 2013;6(17): 3299-303. http://dx.doi.org/10.19026/rjaset.6.3638
21. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The Kin K-fold cross validation. 20th European Symposium on Artificial Neural Networks; 2012 Apr 25-27; Bruges: Computational Intelligence and Machine Learning; p. 441-6.
22. Syarif I, Prugel-Bennett A, Wills G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. TELKOMNIKA (Telecommunication Computing Electronics and Control). 2016;14(4):1502-9. doi: http://doi.org/10.12928/telkomnika.v14i4.3956
23. Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH. Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology 2019;17(1):26-40. https://doi.org/10.11989/JEST.1674-862X.80904120
24. Jackson DA. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 1993;74(8):2204-14. https://doi.org/10.2307/1939574
25. Yuan H, Wu N, Chen X. Mechanical compound fault analysis method based on shift invariant dictionary learning and improved FastICA algorithm. Machines. 2021;9(8):144. https://doi.org/10.3390/machines9080144
26. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26(10):1340-7. https://doi.org/10.1093/bioinformatics/btq134
27. Galkin F, Aliper A, Putin E, Kuznetsov I, Gladyshev VN, Zhavoronkov A. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. BioRxiv. 2018:507780. doi: https://doi.org/10.1101/507780
28. Cunningham P, Delany SJ. k-Nearest neighbour classifiers-A Tutorial. ACM Computing Surveys (CSUR). 2021;54(6):1-25. https://doi.org/10.1145/3459665
29. Noble WS. What is a support vector machine?. Nature Biotechnology 2006;24(12):1565-7.
30. Angulo C, Ruiz FJ, González L, Ortega JA. Multi-classification by using tri-class SVM. Neural Processing Letters 2006;23:89-101.
31. Vezhnevets A, Vezhnevets V. Modest AdaBoost-teaching AdaBoost to generalize better. Graphicon 2005;12(5): 987-97.

[DOI: 10.34172/jhbmi.2023.19]

32. Hastie T, Rosset S, Zhu J, Zou H. Multi-class adaboost. Statistics and its Interface 2009;2(3):349-60. doi: https://dx.doi.org/10.4310/SII.2009.v2.n3.a8
33. Kégl B. The return of AdaBoost. MH: multi-class Hamming trees. arXiv preprint arXiv:1312.6086. 2013.
34. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends 2021;2(01):20-8. https://doi.org/10.38094/jastt20165
35. Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 2016;114:24-31.
https://doi.org/10.1016/j.isprsjprs.2016.01.011
36. Xu S. Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science 2018;44(1):48-59.
https://doi.org/10.1177/0165551516677945
37. Fernández A, López V, Galar M, Del Jesus MJ, Herrera F. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-Based Systems 2013;42:97-110.
https://doi.org/10.1016/j.knosys.2013.01.018
38. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756. 2020.
https://doi.org/10.48550/arXiv.2008.05756
39. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 2015;5(2):1-11.
40. Pal M, Parija S, Panda G. Improved prediction of diabetes mellitus using machine learning based approach. 2nd International Conference on Range Technology (ICORT); 2021 Aug 5-6; Chandipur, Balasore, India: IEEE; 2021. p. 1-6. doi: 10.1109/ICORT52730.2021.9581774

[DOI: 10.34172/jhbmi.2023.19]