تشخیص بیماری دیابت با استفاده از مدل رای‌گیری نرم

سکینه اسدی‌امیری*، استادیار، هانا یوسف‌پور[۲]، دانشجوی کارشناسی ارشد، سعیده محمدپور[۳]، دانشجوی کارشناسی ارشد

[۱] گروه مهندسی کامپیوتر- دانشکده مهندسی و فناوری ـ دانشگاه مازندران ـ بابلسر ـ ایران ـ s.asadi@umz.ac.ir

[۲] گروه مهندسی کامپیوتر - دانشکده مهندسی و فناوری ـ دانشگاه مازندران ـ بابلسر ـ ایران ـ h.yousefpour14@umail.umz.ac.ir

[۳] گروه مهندسی کامپیوتر- دانشکده مهندسی و فناوری ـ دانشگاه مازندران ـ بابلسر ـ ایران

s.mohammadpour11@umail.umz.ac.it

چکیده: دیابت یکی از عوامل مهم مرگ و میر در سراسر جهان است و تأثیرات آن بر بیماری‌های کلیوی و قلبی و از دست دادن بینایی قابل توجه است. پیش‌بینی دیابت یک حوزه تحقیقاتی مهم است که می‌تواند به بهبود درمان بیماری کمک کند. در این مقاله، روش جدیدی برای تشخیص بیماری دیابت پیشنهاد شده است. روش پیشنهادی روی مجموعه‌داده دیابت اعمال شده است، ابتدا در مرحله پیش‌پردازش، شناسایی داده‌های پرت و حذف آن‌ها، جایگزین نمودن مقادیر گمشده و نرمال‌سازی داده‌ها انجام می‌شود. پس از پیش‌پردازش داده‌ها با استفاده از الگوریتم لاسو، ویژگی‌های مهم انتخاب می‌شوند. سپس با استفاده از سه طبقه‌بند K-نزدیک‌ترین همسایه، تقویت گرادیان شدید و کت‌بوست، نمونه‌ها به دو کلاس بیماران دیابتی و سالم طبقه‌بندی می‌شوند. در پایان برای بهبود روش پیشنهادی از الگوریتم رای‌گیری نرم برای ادغام سه طبقه‌بند استفاده شده است. مدل پیشنهادی در این پژوهش با استفاده از معیارهای ارزیابی دقت، صحت و پوشش مورد ارزیابی قرار گرفت. این مدل به دقت ۹۴,۴٪، صحت ۹۶,۵٪ و پوشش ۹۲,۷٪ دست یافت. نتایج حاکی از آن هستند که مدل پیشنهادی با افزایش دقت در تشخیص بیماری دیابت نسبت به سایر، عملکرد بهتری داشته است. بنابراین، با استفاده از این مدل، می‌توان افرادی که در معرض خطر ابتلا به دیابت هستند را با دقت بیشتری شناسایی کرد و اقدامات پیشگیرانه‌ای را برای کنترل بیماری دیابت انجام داد.

واژه‌های کلیدی: یادگیری ماشین، انتخاب ویژگی، داده‌کاوی، دیابت، پیش‌بینی

* **نویسنده مسئول، ایمیل نویسنده**

Diabetes Diagnosis Using Soft Voting Model

Sekine Asadi Amiri[1*], Assistant Professor, Hannah Yousefpour[2], Master's Student, Saeide Mohammadpour[3], Master's Student

[1] Dept. Computer Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran, s.asadi@umz.ac.ir

[2] Dept. Computer Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran, h.yousefpour14@umail.umz.ac.ir

[3] Dept. Computer Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran, s.mohammadpour11@umail.umz.ac.it

Abstract: Diabetes is one of the most significant factors leading to death, which can significantly result in kidney diseases, heart diseases, and sight loss. The application of data mining could be helpful for the diagnosis and treatment of this disease. Predicting diabetes is an important area of research that can help improve the treatment process. It is necessary to prevent, monitor, and raise awareness about this disease. In this article, a new method to diagnose diabetes is proposed. The proposed method includes a pre-processing stage in which outlier data is removed. Eventually, by using the K-Nearest Neighbor classifier, Extreme Gradient Boosting and CatBoost, samples have been classified into two classes: diabetic and non-diabetic. In the end, to improve the proposed method, a soft voting algorithm has been used to merge the three classifiers. The proposed method has been applied to the Pima diabetes dataset, which includes information on age, gender, blood pressure, glucose, and other factors related to diabetes. The proposed method in this research was evaluated using evaluation metrics such as accuracy, precision, and recall. This model achieved 94.4. % accuracy, 96.5% precision, and 92.7% recall. The results indicate that the proposed model has performed better than other references by increasing the accuracy in diagnosing diabetes. Therefore, by using this model, it will be possible to identify potential diabetic patients more accurately and ultimately prevent them from becoming diabetic.

Keywords: *Machine Learning, Feature Selection, Data Mining, Diabetes, Prediction*

*** Corresponding author**

۱. مقدمه

دیابت، یک بیماری مزمن است که در آن بدن توانایی کاهش قند خون را از دست می‌دهد. قند خون انرژی لازم برای فعالیت‌های بدن را فراهم می‌کند و اگر سطح آن افزایش یابد، بدن باید برای کاهش آن اقدام کند. کاهش قند خون توسط یک هورمون به نام انسولین انجام می‌شود که در پانکراس تولید می‌شود. این بیماری به دو نوع تقسیم می‌شود. نوع اول، یا دیابت نوع ۱، به دلیل عدم ترشح انسولین رخ می‌دهد و معمولاً در سنین کودکی و نوجوانی بروز می‌کند. نوع دوم، یا دیابت نوع ۲، به دلیل مقاومت بدن به انسولین که ممکن است به دلیل چاقی، رخ دهد. توده بدنی بالا یکی از عوامل مهمی است که می‌تواند بر سلامت انسان تأثیر بگذارد و ممکن است منجر به مشکلاتی مانند دیابت شود. بارداری یک دوره حساس و مهم در زندگی زنان است و نیازمند مراقبت ویژه است. افزایش وزن در طول بارداری طبیعی است و معمولاً بین ۱۱ تا ۱۶ کیلوگرم است. اما اگر افزایش وزن بیش از حد باشد، ممکن است منجر به مشکلاتی مانند دیابت شود. اگر دیابت در طول بارداری رخ دهد، آن را دیابت بارداری می‌نامند که می‌تواند برای مادر و جنین او خطرناک باشد [۱]. دیابت، بیماری مزمنی است که در آن سطح بالای قند خون می‌تواند خطرناک باشد. این بیماری با علائم کوتاه مدتی مانند تشنگی، خستگی، تحریک‌پذیری، گرسنگی شدید و تکرر ادرار همراه است. در بلند مدت، می‌تواند منجر به امراض دیگری مانند آسیب تدریجی به کلیه‌ها، چشم‌ها و قلب شود. این تهدیدات می‌توانند تاثیر مهمی بر سطح بهداشت و زندگی افراد مبتلا به این بیماری داشته باشند.

دیابت به مرور زمان اتفاق می‌افتد و علائمی از خود نشان می‌دهد. شناسایی این علائم می‌تواند به تشخیص دیابت در مراحل اولیه کمک کند. با تشخیص این علائم و افزایش فعالیت بدنی و تغذیه سالم، افراد می‌توانند سلامت خود را حفظ کنند. با این حال، با وجود اهمیت فراوان تشخیص دیابت در مراحل اولیه، پزشکان هنوز نتوانسته‌اند روش موثقی برای تشخیص این علائم ارائه دهند. بنابراین، شناسایی مراحل اولیه دیابت یکی از چالش‌های مهم دنیای امروز است [۲]. با ورود الگوریتم‌های تشخیصی به عرصه پزشکی، انقلابی در فرآیند تشخیص زودهنگام بیماری‌ها رخ داد. این الگوریتم‌ها با استفاده از داده‌های جمع‌آوری شده از مبتلایان به بیماری مورد نظر، برای شناسایی عوامل موثر در بیماری به‌کار رفته‌اند. این روش‌ها می‌توانند الگوهای پنهان و اطلاعات مفیدی را از داده‌های حجیم استخراج کنند و به پزشکان در بهبود تشخیص دیابت کمک کنند [۳]. استفاده از داده‌های سلامت الکترونیکی، داده‌های سنسوری و داده‌های بالینی می‌تواند به پزشکان در کشف الگوهای جدید در تشخیص و پیش‌بینی دیابت کمک کند. همچنین، این روش‌ها می‌توانند در پیش‌بینی عوارض احتمالی دیابت و کمک به بیماران در مدیریت بیماری مفید باشند. به طور کلی، استفاده از استخراج داده و الگوریتم‌های یادگیری می‌تواند به بهبود قابل توجه در تشخیص دیابت منجر شود [۴].

دقت در تشخیص دیابت یکی از چالش‌های علم پزشکی می‌باشد. در این پژوهش، با توجه به اهمیت این مسئله، سعی شده است تا دقت در شناسایی افراد در معرض خطر بیماری دیابت افزایش داده شود. در روش پیشنهادی ابتدا داده‌های پرت شناسایی و حذف می‌شوند. در مرحله بعد از الگوریتم لاسو برای انتخاب ویژگی‌های مهم و حذف ویژگی‌های غیرضروری استفاده شده است. پس از آن، از سه طبقه‌بند K-نزدیکترین همسایه[1] (KNN)، تقویت گرادیان شدید[2] (XGBoost) و کت بوست (CatBoost) برای طبقه‌بندی نمونه‌ها به دو کلاس دیابتی و غیردیابتی استفاده شده است. در انتها برای بهبود روش پیشنهادی از مدل رای‌گیری

[1] K-Nearest Neighbors

[2] Extreme Gradient Boosting

نرم استفاده شده است. رای‌گیری نرم[3] یک تکنیک قدرتمند برای ادغام پیش‌بینی‌های کسب شده توسط الگوریتم‌های یادگیری ماشین است. این پژوهش شامل پنج بخش است. بخش اول، مقدمه‌ای است که در آن تعریف دیابت، سیر بیماری و اهمیت آن بیان شده است. بخش دوم به بررسی پیشینه ادبیات و کارهای انجام شده در زمینه تشخیص و بهبود دقت در تشخیص دیابت می‌پردازد. بخش سوم به تعریف روش کار در این پژوهش می‌پردازد که شامل پیش‌پردازش، انتخاب ویژگی و مدل پیشنهادی است. بخش چهارم نتایج حاصل از الگوریتم پیشنهادی این پژوهش را بررسی می‌کند. در نهایت، بخش پنجم به نتیجه‌گیری می‌پردازد.

۲. پیشینه ادبیات

با توجه به اهمیت تشخیص زودهنگام بیماری دیابت، تحقیقات گسترده‌ای با استفاده از مدل‌های یادگیری ماشین برای تشخیص و پیشگیری از دیابت انجام شده است. در مرجع [۵]، رویکردی نوین با استفاده از تکنیک‌های داده‌کاوی برای افزایش دقت پیش‌بینی و قدرت تعمیم بالای پیش‌بینی‌های مدل ارائه شده است. این پژوهش پس از پیش‌پردازش مجموعه‌داده با استفاده از الگوریتم K-means بهبود یافته و رگرسیون لجستیک، به قدرت تشخیص خوبی دست یافته است. در مرجع [۶]، برای ایجاد یک مدل تشخیصی با قدرت تعمیم بالا، از مجموعه‌داده‌ای از بنگلادش استفاده شده و الگوریتم‌های بیز ساده، رگرسیون لجستیک و جنگل تصادفی مورد بررسی قرار گرفته است. همچنین در مرجع [۷]، مدلی با استفاده از روش‌های داده‌کاوی برای پیش‌بینی موارد دیابت ارائه شده است. در نهایت، در این پژوهش با استفاده از چندین روش، از جمله J48[4]، بیز ساده و ماشین بردار پشتیبان، پیش‌بینی برای تشخیص دیابت انجام شده است.

در مرجع [۸]، برای پیش‌بینی وقوع دیابت در اندونزی، از مجموعه‌داده‌ای استفاده شده است که شامل هفت عامل خطر دیابت، از جمله سن، جنسیت، شاخص توده بدنی، سابقه خانوادگی دیابت، فشار خون، مدت زمان دیابتی و سطح گلوکز خون به عنوان عوامل کلیدی در شناسایی دیابت است. در نهایت، از روش خوشه‌بندی K-means و روش طبقه‌بندی بیز ساده، برای تشخیص بیماری دیابت استفاده شده است. در مرجع [۹]، با استفاده از اطلاعات بیماران بیمارستان نیروی دفاع بحرین و بررسی داده‌ها با استفاده از روش‌های داده‌کاوی، تلاش شده است تا شروع بیماری دیابت را پیش‌بینی کنند. مرجع [۱۰] یک تجزیه و تحلیل جامع را در راستای پیش‌بینی دیابت با استفاده از یادگیری ماشین، روش‌های داده‌کاوی و ابزارهای مرتبط انجام داده است. این پژوهش با استفاده از سه روش مختلف داده‌کاوی، از جمله بیز ساده، ماشین بردار پشتیبان و درخت تصمیم، به بررسی موضوع پرداخته است. مرجع [۱۱] در راستای پیش‌بینی احتمال بروز بیماری قلبی برای بیماران دیابتی، با استفاده از روش‌های پیشرفته احتمالی و بر اساس دقت پیش‌بینی‌شان، ارزیابی و تحلیل انجام داده است.

مرجع [۱۲] برای شناسایی بهترین طبقه‌بند برای ارزیابی نتایج بالینی، از یک مجموعه استاندارد معیارها و شبکه‌های بیز ساده و توابع پایه شعاعی استفاده کرده است. در مرجع [۱۳]، با مقایسه روش‌های مختلف یادگیری ماشین، مدلی برای بهبود عملکرد پیشنهاد شده است. روش‌های پیش‌بینی مانند K–نزدیک‌ترین همسایه و رگرسیون لجستیک، از جمله الگوریتم‌های مورد استفاده در این پژوهش بوده‌اند. مرجع [۱۴] از سه الگوریتم یادگیری K–نزدیک‌ترین همسایه، آدابوست و LightGBM برای پیش‌بینی دیابت استفاده کردند. سپس با استفاده ار رای‌گیری نرم بین این سه الگوریتم، دسته مناسب هر نمونه را تعیین کردند. مرجع [۱۵]

[3] Soft Voting

[4] Java implementation of the C4. 5 algorithm

برای پیش‌بینی بیماری دیابت از رای‌گیری نرم بین سه الگوریتم پرسپترون چند لایه، ماشین بردار پشتیبان و K–نزدیک‌ترین همسایه استفاده کردند. مرجع [۱۶] از الگوریتم‌های رگرسیون لجستیک، ماشین بردار پشتیبان، K –نزدیک‌ترین همسایه، تقویت گرادیان، بیز ساده، و جنگل تصادفی برای پیش‌بینی بیماری دیابت استفاده کردند. سپس از رای‌گیری سخت برای ترکیب سه الگوریتم بیز ساده، جنگل تصادفی و ماشین بردار پشتیبان استفاده کردند. با توجه به اینکه الگوریتم K–نزدیک‌ترین همسایه در اغلب پژوهش‌های انجام شده، الگوریتمی امیدبخش در تشخیص بیماری دیابت بوده است، می‌توان از این الگوریتم و دیگر الگوریتم‌های تجمعی استفاده کرد که قدرت تشخیص خوب خود را در داده‌های پیچیده ثابت کرده‌اند. بنابراین، استفاده از این الگوریتم‌ها امکان ساخت مدل تشخیصی با دقت بالاتری را فراهم می‌آورد.

۳. مجموعه‌داده

مجموعه‌داده‌ای که در این پژوهش استفاده شده، مربوط به دیابت در بین زنان قوم پیما[5] در هند است. این مجموعه‌داده توسط موسسه ملی دیابت و بیماری‌های گوارشی و کلیوی آماده شده است [۱۷]. هدف از این مجموعه‌داده، پیش‌بینی ابتلا به دیابت بر اساس اندازه‌گیری‌های تشخیصی خاص است که در مجموعه‌داده وجود دارد. مجموعه‌داده شامل ۷۶۸ نمونه، هشت ویژگی پیش‌بینی‌کننده و یک ویژگی که نشان دهنده کلاس (هدف) است. در این پژوهش، ۷۰ درصد از مقادیر مجموعه‌داده برای آموزش، ۱۰ درصد برای اعتبارسنجی و ۲۰ درصد باقی‌مانده برای آزمایش استفاده شده است. جدول (۱) نشان‌دهنده خلاصه آماری مجموعه‌داده دیابت هندی پیما می‌باشد. متغیرهای پیش‌بینی‌کننده شامل تعداد حاملگی‌ها، گلوکز، فشار خون، ضخامت پوست، انسولین، شاخص توده بدنی، فاکتورهای مشخص‌کننده دیابت و سن می‌باشد. این مجموعه‌داده دارای دو کلاس، یعنی ابتلا و عدم ابتلا به دیابت، است. شکل (۱) نشان‌دهنده ماتریس همبستگی ویژگی‌ها با یکدیگر است. ماتریس همبستگی نشان می‌دهد که دو ویژگی چقدر مرتبط هستند. ضریب همبستگی در محدوده صفر و یک است که هر چه مقدار آن به یک نزدیک‌تر باشد، یعنی وابستگی بین دو متغیر بیشتر است.

شکل (۲) نمودار جفت شده[6] ویژگی‌های مجموعه‌داده بر اساس کلاس را نشان می‌دهد که این ویژگی‌ها از بالا به پایین و از چپ به راست به‌ترتیب جدول می‌باشند. این نمودار امکان کشف رابطه بین ویژگی‌های موجود در مجموعه‌داده را فراهم می‌کند. اگر نقاط در نمودار پراکنده باشند، به معنی عدم وجود رابطه آشکاری است. در مقابل، اگر نقاط تقریباً در یک خط مستقیم قرار گیرند، نشان‌دهنده رابطه خطی بین آن‌ها است.

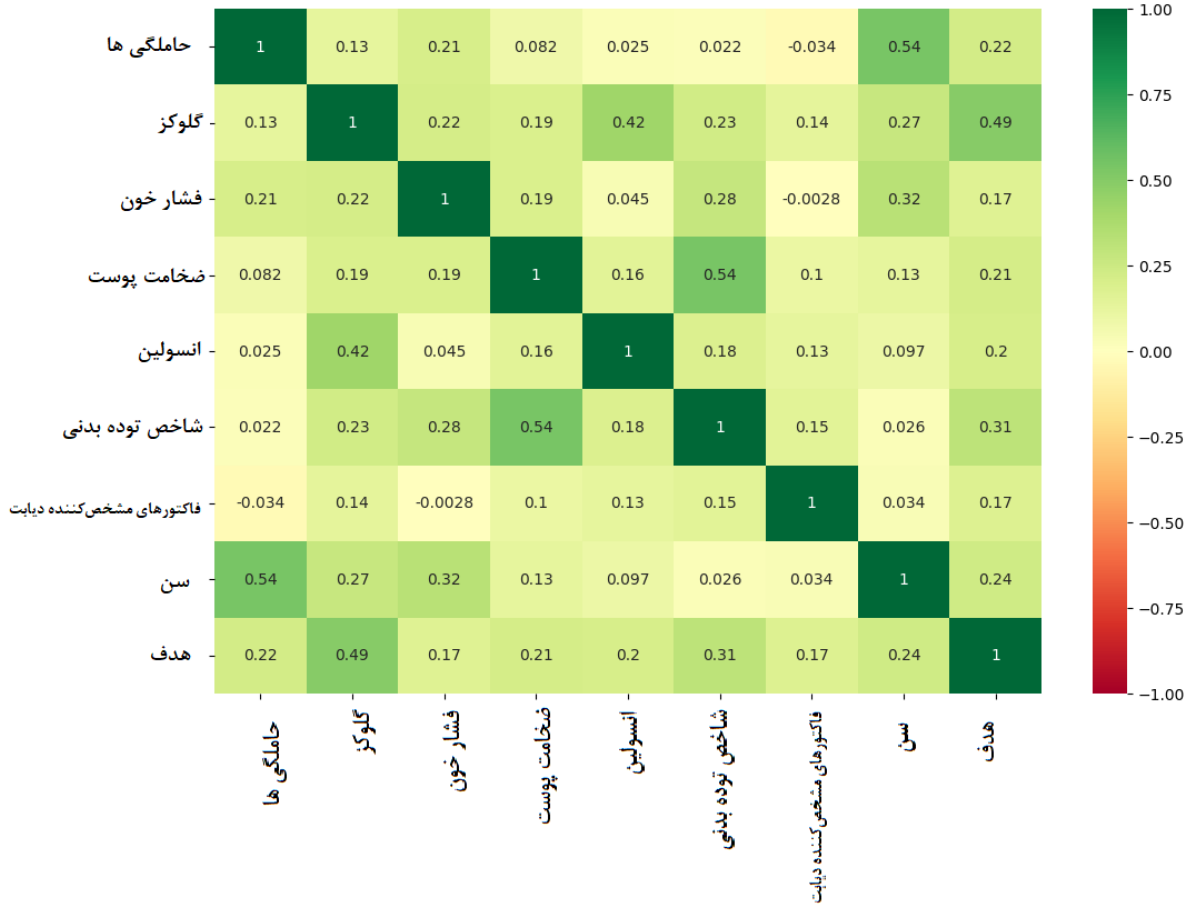
در یک نمودار جفت‌شده، قطرها نمودارهایی هستند که یک ویژگی را در برابر خودش نشان می‌دهند، که این می‌تواند توزیع داده‌ها برای آن ویژگی خاص را نشان دهد. قطر نمودار جفت شده، پراکندگی داده‌ها در مجموعه‌داده را نشان می‌دهد. براساس نمودار پراکندگی که در شکل (۲) نشان داده شده است، داده‌ها به‌صورت جفتی با تجمع شبیه‌ترین خروجی‌ها همراه هستند. بیشترین همبستگی در ویژگی‌ها بین تعداد حاملگی‌ها و سن، ضخامت پوست و شاخص توده بدنی، و گلوکز و انسولین مشاهده می‌شود. این همبستگی بر اساس ارقام مثبت نمودار پراکندگی است. همچنین، با توجه به ماتریس همبستگی که رابطه قوی بین هیچ یک از ویژگی‌ها و خروجی را نشان نمی‌دهد در نتیجه هیچ یک از ویژگی‌ها به تنهایی قادر به پیش‌بینی بیماری دیابت نمی‌باشد.

[5] PIMA

[6] Pair Plot

جدول (۱): خلاصه آماری مجموعه‌داده دیابت هندی پیما (منبع: محاسبات محقق).

شاخص‌ها	تعداد حاملگی‌ها	گلوکز	فشار خون	ضخامت پوست	انسولین	شاخص توده بدنی	فاکتورهای مشخص‌کننده دیابت	سن
شمارش	768.0	768.0	768.0	768.0	768.0	768.0	768.0	768.0
میانگین	3.845	120.894	69.105	20.536	79.799	31.992	0.471	33.240
انحراف معیار	3.36	31.97	19.35	15.95	115.2	7.88	0.331	11.760
کمینه	0.0	0.0	0.0	0.0	0.0	0.0	0.078	21.0
٪۲۵	1.0	99.0	62.0	0.0	0.0	27.3	0.24375	24.0
٪۵۰	3.0	117.0	72.0	23.0	30.5	32.0	0.3725	29.0
٪۷۵	6.0	140.2	80.0	32.0	127.25	36.6	0.62625	41.0
بیشینه	17.0	199.0	122.0	99.0	846.0	67.1	2.42	81.0



شکل (۱): ماتریس همبستگی ویژگی‌ها در مجموعه‌داده پیما (منبع: محاسبات محقق).

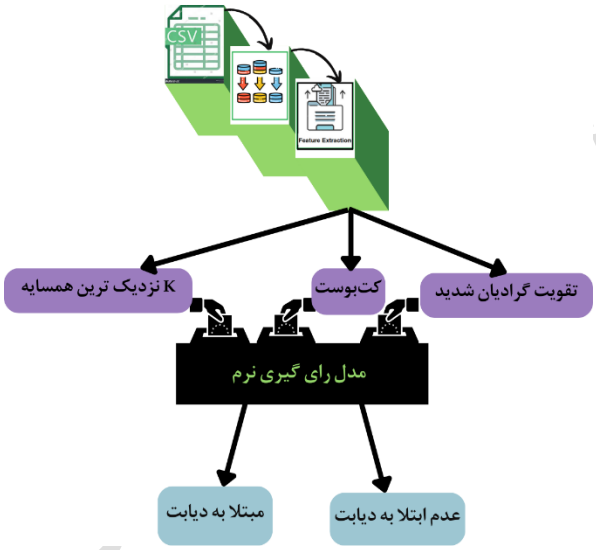
شکل (۲): نمودار جفت‌شده ویژگی‌های مجموعه‌داده (منبع: محاسبات محقق).

۴. روش‌شناسی

در این مقاله، روش جدیدی برای تشخیص بیماری دیابت معرفی شده است. در مرحله پیش‌پردازش، مقادیر گمشده با صفر جایگزین می‌شوند و سپس داده‌ها نرمال‌سازی[۷] می‌شوند. در مرحله بعدی، با استفاده از روش انتخاب ویژگی لاسو، اهمیت هر یک از ویژگی‌ها محاسبه و ویژگی‌های با اهمیت بیشتر انتخاب می‌شوند. در نهایت، یک روش با استفاده از رای‌گیری نرم و سه رای‌دهنده، یعنی K–نزدیک‌ترین همسایه، تقویت گرادیان شدید و

[7] Normalization

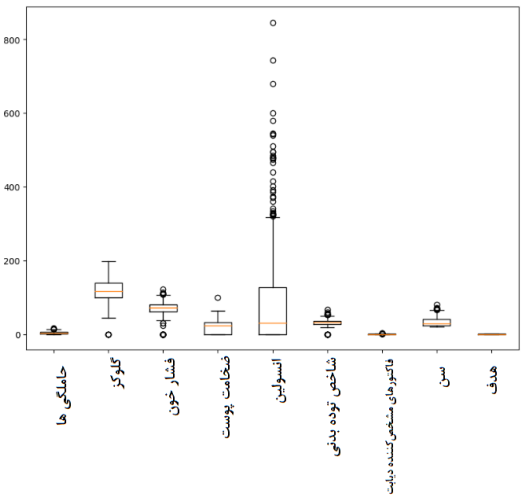
کت‌بوست، برای افزایش دقت در تشخیص بیماری دیابت ارائه شده است. شکل (۳) فلوچارت روش پیشنهادی در تشخیص بیماری دیابت را نشان می‌دهد.



شکل (۳): فلوچارت روش پیشنهادی برای تشخیص بیماری دیابت (منبع: محاسبات محقق).

۱,۴. پیش پردازش داده‌ها

قبل از آموزش هر الگوریتم، پیش‌پردازش داده‌ها ضروری است. پیش‌پردازش، یکی از مهم‌ترین بخش‌های داده‌کاوی است که به بررسی داده‌ها برای یافتن مقادیر پرت و گم‌شده و بررسی ویژگی‌ها برای بهبود عملکرد می‌پردازد. شکل (۴) نشان‌دهنده نمودار جعبه‌ای این مجموعه‌داده می‌باشد. با توجه به شکل‌های (۴) و (۲) داده‌های پرت در این مجموعه‌داده وجود دارد. داده‌های پرت تفاوت قابل توجهی با بقیه اعضای مجموعه دارند. از آنجا که داده‌های پرت می‌توانند فرآیند تشخیص را برای یادگیرنده دشوار کند، در این پژوهش داده‌های پرت حذف شدند. روش مورد استفاده برای شناسایی داده‌های پرت، روش فاصله بین چارکی[۸] (IQR) می‌باشد. هرآنچه که در نمودار جعبه‌ای از چارک دامنه کمینه و بیشینه خارج باشد، داده پرت محسوب می‌شود. نهایتا مقادیر داده‌های پرت با میانگین داده‌ها جایگزین می‌شوند. در این پژوهش، در مرحله پیش‌پردازش، مقادیر گم‌شده با مقدار ثابت NaN، با مقدار صفر جایگزین می‌شوند. سپس، داده‌ها نرمال‌سازی می‌شوند. این عمل، با کاهش دامنه ویژگی‌های مجموعه‌داده، داده‌ها را به یک دامنه مشخص می‌برد. در این مرحله، دامنه اعداد بین ۱ و ۱- قرار داده شده است.



شکل (۴): نمودار جعبه‌ای نمونه‌های موجود در مجموعه‌داده (منبع: محاسبات محقق).

۲,۴. انتخاب ویژگی با الگوریتم لاسو

لاسو یا عمل‌گر گزینش و انقباض کمترین قدرمطلق[۹] در سال ۱۹۹۶ توسط رابرت تیب‌شیرانی معرفی شد [۱۸]. این الگوریتم توسیع الگوریتم حداقل مربعات معمولی (OLS) است. OLS روشی برای تخمین پارامترهای یک مدل رگرسیون خطی است که سعی می‌کند مجموع مجذور باقی‌مانده‌ها را به حداقل برساند. لاسو از یک تکنیک کاهش بعد استفاده می‌کند و به جای کمینه‌سازی مجموع مربعات، از یک تابع جریمه بر روی جمع قدرمطلق ضرایب مدل رگرسیونی بهره می‌برد. این تابع جریمه تعداد پارامترهای مدل را کنترل می‌کند، که این موضوع سبب کاهش

[8] Interquartile range

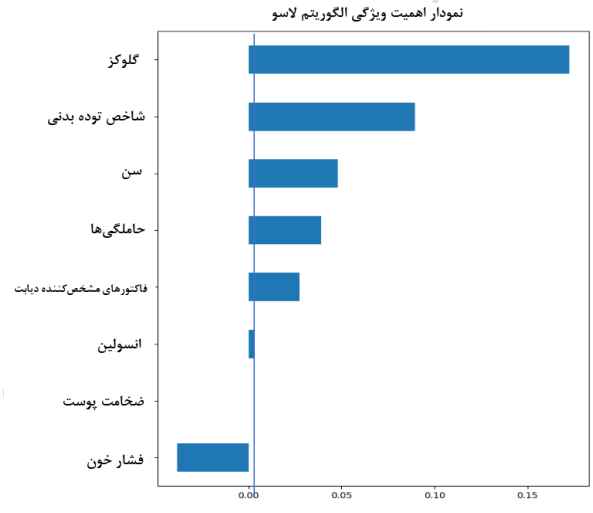
[9] Least Absolute Shrinkage and Selection Operator

پیچیدگی مدل و جلوگیری از بیش‌برازش می‌شود. انتخاب لاسو را می توان به صورت زیر تعریف کرد [۱۸]:

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{lasso}} = \boldsymbol{arg}_{\boldsymbol{\beta}}\, \boldsymbol{min} \left\{ \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \left(\boldsymbol{y_i} - \boldsymbol{\beta_0} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{x_{ij}}\, \boldsymbol{\beta_j} \right)^{\mathbf{2}} + \boldsymbol{\lambda} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} |\boldsymbol{\beta_j}| \right\} \quad (۱)$$

در این رابطه λ پارامتر تنظیم‌کننده است، به این معنی که اگر مقدارش برابر با صفر باشد، مدل به رگرسیون عادی تبدیل شده و همه متغیرها در آن حضور خواهند داشت و اگر مقدار آن افزایش یابد تعداد متغیرهای مستقل در مدل کاهش می‌یابند. لاسو جریمه ای را به مجموع مجذور باقی‌مانده اضافه می کند. این جریمه برابر است با ضرب ضرایب بتا در پارامتر λ، که در کند کردن یا شدت بخشیدن جریمه تاثیر می‌گذارد. برای مثال، اگر λ کمتر از ۱ باشد، جریمه را کند می‌کند و اگر بالای ۱ باشد، جریمه را شدت می‌بخشد. از این رو این روش انتخاب ویژگی، مجموع باقی‌مانده مربع‌ها را کاهش می‌دهد به شرط اینکه مجموع مقدار مطلق ضرایب کمتر از یک مقدار ثابت باشد. لاسو در ابتدا در زمینه حداقل مربعات تعریف شد، اما می توان آن را به مدل های بسیار متنوعی نیز تعمیم داد. لاسو هم دقت پیش‌بینی و هم قابلیت تفسیر مدل را بهبود می‌بخشد. اگر همبستگی بالایی در گروه پیش‌بینی کننده ها وجود داشته باشد، لاسو تنها یکی از آنها را انتخاب می‌کند و بقیه را به صفر می‌رساند. این تغییرپذیری برآوردها را با کوچک کردن برخی از ضرایب دقیقاً به صفر کاهش می‌دهد و مدل های قابل تفسیر تولید می‌کند.

لاسو یک مدل خطی است که با برازش ابر صفحه‌ای، داده‌ها را به دو کلاس جدا می‌کند. این مدل هم می‌تواند عمل انتخاب ویژگی و هم منظم سازی[10] که برای جلوگیری از بیش‌برازش است را انجام دهد. این الگوریتم برای انتخاب ویژگی مفید است، زیرا می‌تواند مهم‌ترین ویژگی‌های در دسترس را شناسایی کند. همچنین در مسئله طبقه‌بندی، لاسو اغلب برای انجام طبقه‌بندی دودویی استفاده می‌شود، جایی که هدف پیش‌بینی یکی از دو برچسب ممکن است. انتخاب ویژگی با انتخاب بهترین ویژگی‌ها و حذف ویژگی‌های نامرتبط، قدرت پیش‌بینی مدل‌ها را افزایش می‌دهد. الگوریتم لاسو، با توجه به عملکرد خوب و معمول در طبقه‌بندی دودویی و عدم بیش‌برازش، می‌تواند اهمیت ویژگی‌ها را به خوبی نشان دهد [۱۸]. شکل (۵) نتایج الگوریتم لاسو را در تشخیص اهمیت ویژگی در هشت ویژگی موجود در مجموعه‌داده را نشان می‌دهد. چهار ویژگی برتر که توسط الگوریتم لاسو انتخاب شده‌اند، به ترتیب شامل گلوکز، شاخص توده بدنی، سن، حاملگی‌ها، فاکتور پیش‌بینی کننده دیابت و انسولین هستند که در این پژوهش استفاده شده‌اند.



شکل (۵): اهمیت ویژگی‌های موجود در مجموعه‌داده با استفاده از الگوریتم لاسو (منبع: محاسبات محقق).

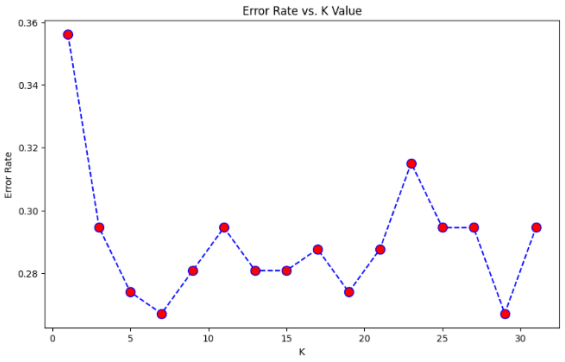
10 Regularization

۳,۴. طبقه‌بند

در این گام، با استفاده از ویژگی‌های انتخاب شده، به آموزش مدل پرداخته می‌شود. برای ساخت یک مدل رای‌گیری، ابتدا باید الگوریتم‌های مشارکت‌کننده در رای‌گیری آموزش داده شوند. در این پژوهش، از الگوریتم‌های رای‌گیری نرم، K–نزدیک‌ترین، تقویت گرادیان شدید و کت‌بوست استفاده شده است.

۱,۳,۴. الگوریتم K–نزدیک‌ترین همسایه

الگوریتم K–نزدیک‌ترین همسایه یکی از الگوریتم‌های یادگیری ماشین تحت نظارت است که برای حل مسائل طبقه‌بندی استفاده می‌شود. این الگوریتم برای مسائل طبقه‌بندی، K–نزدیک‌ترین همسایه را پیدا و با اکثریت آرا نزدیک‌ترین همسایگان کلاس را پیش‌بینی می‌کند. شکل (۶) نشان‌دهنده عملکرد الگوریتم K–نزدیک‌ترین همسایه در بازه فرد K، ۱ الی ۳۱ است. محور عمود در این شکل مقدار خطا و محور افقی مقدار K را نشان می‌دهد. در این پژوهش K برابر با ۲۹ انتخاب شده است، چراکه K بزرگ‌تر به تصمیم‌گیری روان‌تر الگوریتم منجر می‌شود.



شکل (۶): خطای الگوریتم K–نزدیک‌ترین همسایه با مقادیر متفاوت k (منبع: محاسبات محقق).

۲,۳,۴. الگوریتم تقویت گرادیان شدید

الگوریتم تقویت گرادیان شدید، یک تکنیک یادگیری تحت نظارت است که به خانواده الگوریتم‌های یادگیری ماشین مربوط به درخت‌های تصمیم تقویت‌شده با گرادیان[11] تعلق دارد. این الگوریتم، گروهی از یادگیرندگان ضعیف را با استفاده از فرآیند ادغام گروهی تقویت می‌کند و آن‌ها را به یک یادگیرنده قدرتمند تبدیل می‌کند. در هر مرحله، این الگوریتم سعی می‌کند تا تعداد خطاهای آموزشی انجام شده توسط یادگیرنده ضعیف را کاهش دهد [۱۹]. الگوریتم تقویت گرادیان شدید، با استفاده از یک تابع خطا، سعی می‌کند تا خطای پیش‌بینی را کاهش دهد. در هر مرحله، یک یادگیرنده ضعیف به مدل اضافه می‌شود که سعی می‌کند خطای باقی‌مانده از مراحل قبلی را کاهش دهد. این فرآیند به‌صورت تکرار شونده ادامه می‌یابد تا اینکه خطای مدل به یک حد مقبول برسد یا تعداد مراحل مشخص شده به پایان برسد. در نهایت، پیش‌بینی مدل نهایی، مجموع پیش‌بینی‌های تمام یادگیرندگان ضعیف خواهد بود.

۳,۳,۴. کت‌بوست

کت‌بوست یک الگوریتم تجمعی یادگیری ماشین است یعنی با استفاده از مجموعه‌ای از یادگیرنده‌های ضعیف مدلی قوی‌تر می‌سازد. کاهش خطا نیز در این الگوریتم با تقویت گرادیان درخت‌های تصمیم در هر بار آموزش انجام می‌پذیرد. این الگوریتم همانند دیگر الگوریتم‌های تجمعی برآن است تا با ساخت مجموعه‌ای از پیش‌بینی کننده‌های ضعیف و کاهش خطای آن‌ها به یک مدل قوی دست یابد. کت‌بوست برخلاف دسته‌بندهای سنتی که نیاز به رمزگذاری یک‌طرفه دارند، ویژگی‌ها را پردازش کرده و دقت دسته‌بندی آن بالا است. همچنین

11 Gradient-Boosted Decision Trees

رمزگذاری مرتب رابطه بین ویژگی‌ها و کلاس را در نظر می‌گیرد که به مراتب سبب بهبود عملکرد در مقایسه با رمزگذاری یک‌طرفه می‌شود.

۴,۳,۴. رای‌گیری نرم

مدل یادگیری رای‌گیری، مدلی است که در آن از چندین الگوریتم برای ساخت پیش‌بینی نهایی استفاده می‌شود. در این پژوهش، از الگوریتم رای‌گیری نرم استفاده شده است، زیرا نسبت به الگوریتم رای‌گیری سخت، مزایای بیشتری دارد و نیازهای این پژوهش را برآورده می‌کند. الگوریتم رای‌گیری، با توجه به مزایایی مانند جلوگیری از بیش‌برازش و تفسیرپذیری پیشرفته، باعث افزایش دقت و تعمیم‌پذیری بیشتر الگوریتم‌های مورد استفاده شده است [۲۰]. در مدل رای‌گیری نرم، مدل‌ها احتمالات برای هر کلاس ارائه می‌دهند و این احتمالات را برای ایجاد پیش‌بینی نهایی، وزن‌دهی و میانگین‌گیری می‌کند. در مقابل، در مدل رای‌گیری سخت، نتیجه نهایی بر اساس بیشترین رای به کلاس تعیین می‌شود.

گرچه مدل یادگیری نرم ممکن است بار محاسباتی بیشتری نسبت به رای‌گیری سخت داشته باشد، اما نیاز به مدل‌ها برای تولید احتمالات یا امتیازات، اطمینان از نتیجه نهایی را فراهم می‌کند. ترکیب الگوریتم‌های یادگیری‌ماشین یا الگوریتم‌های رای‌گیری نرم و سخت مدلی پیشرفته برای پیش‌بینی‌تولید می‌کند، که پیش‌بینی‌ها را ادغام می‌کند تا نتیجه قوی‌تر و دقیق‌تری تولید شود. با استفاده از نقاط قوت مدل‌های مختلف، رای‌گیری می‌تواند نقاط ضعف آن‌ها را کاهش دهد و بعضاً منجر به عملکرد بهتر از هر مدل تکی شود. در این پژوهش مدل رای‌گیری نرم با سه رای دهنده K–نزدیک‌ترین همسایه ، تقویت گرادیان شدید و کت‌بوست به عنوان طبقه بند با هدف افزایش دقت در تشخیص دیابت مورد استفاده قرار گرفت. برای الگوریتم کت‌بوست وزنی معادل با ۰,۸، تقویت گرادیان شدید وزنی برابر با ۰,۱ و برای K–نزدیک‌ترین همسایه وزنی برابر با ۰,۱ انتخاب شده‌است. وزن‌ها به‌صورت آزمایشی تعیین شده و سپس بالاترین دقت حاصل از وزن ۰,۸–۰,۱–۰,۱ مورد استفاده قرار گرفت.

۵. تحلیل و نتایج

در این بخش، ابتدا معیارهای ارزیابی مورد استفاده در پژوهش تعریف می‌شوند. سپس، نتایج حاصل از روش پیشنهادی بر روی مجموعه‌داده پیما ارائه می‌شود. در نهایت، روش پیشنهادی با کارهای مرتبط دیگر مقایسه و تحلیل می‌شود.

۱,۵. معیارهای ارزیابی

در این بخش، معیارهای ارزیابی مورد بررسی قرار خواهد گرفت. از سه معیار ارزیابی دقت، صحت و پوشش در این پژوهش استفاده شده است. معیارهای دقت، صحت و پوشش طبق روابط زیر تعریف می‌شوند [۲۱]:

دقت[12] نشان‌دهنده پیش‌بینی درست خروجی توسط مدل است و بصورت زیر محاسبه می‌شود.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (۲)$$

صحت[13] نشان‌دهنده آن است که پیش‌بینی مثبت مدل تا چه اندازه صحیح است و بصورت زیر محاسبه می‌شود.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (۳)$$

12 Accuracy

13 Precision

پوشش[14] نشان دهنده این است که پیش‌بینی های منفی مدل تا چه اندزاه صحیح است و بصورت زیر محاسبه می‌شود.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (۴)$$

در روابط بالا، TP [15] نشان دهنده مثبت صحیح است، یعنی کلاس مثبت به درستی به آن نسب داده شده است. FP[16] نشان دهنده مثبت کاذب است، یعنی کلاس مثبت به اشتباه به آن نسبت داده شده است. FN [17] نشان دهنده منفی کاذب است، یعنی کلاس منفی به اشتباه به آن نسبت داده شده است. TN[18] نشان دهنده منفی صحیح است، یعنی کلاس منفی به درستی به آن نسب داده شده است.

۲,۵. نتایج

در این بخش، مجموعه‌داده‌های سوابق بیماران دیابت هندی پیما به منظور پیش‌بینی دیابت مورد بررسی قرار گرفته است. با توجه به اهمیت تشخیص دیابت، در این مقاله مدلی کارآمد با استفاده از رای‌گیری نرم ایجاد شده است. روش پیشنهادی، که از رای‌گیری نرم بین سه الگوریتم K-نزدیک‌ترین همسایه، تقویت گرادیان شدید و کت‌بوست استفاده می‌کند، با ادغام نتایج این سه الگوریتم، به دقت بالاتری نسبت به هر یک از الگوریتم‌ها به تنهایی رسیده است. دقت اولیه الگوریتم‌های K-نزدیک‌ترین همسایه برابر با ۸۱,۸۱٪، تقویت گرادیان شدید برابر با ۸۲,۴۶٪ و کت‌بوست برابر با ۹۲٪ بوده است. نتایج روش پیشنهادی در دو حالت با حذف داده‌های پرت و بدون حذف داده‌های پرت در جدول (۲) نشان داده شده است. همان‌طور که از نتایج مشخص است، با حذف داده‌های پرت، عملکرد مدل پیشنهادی به دقت ۹۴,۴٪، صحت ۹۶,۵٪ و پوشش ۹۲,۷٪ رسیده است. جدول (۲) نشان می‌دهد آنچه مدل پیش‌بینی کرده است تا ۹۶٪ صحت دارد. همچنین مدل پیشنهادی پوشش مورد قبولی را نیز از خود نشان داده است. در این جدول نتایج حاصل از روش پیشنهادی بدون حذف مقادیر پرت نیز نشان داده شده است. این حالت، به دقت ۹۳,۶٪، صحت ۹۵,۴٪ و پوشش ۹۲,۵٪ رسیده است.

جدول (۲): نتایج روش پیشنهادی برای تشخیص بیماری دیابت در ۲ حالت با حذف داده‌های پرت و بدون حذف داده‌های پرت.

روش	روش رای‌گیری نرم	روش رای‌گیری نرم
حذف داده پرت	بله	خیر
جداسازی داده	80-20	80-20
دقت	94.4٪	93.6٪
صحت	96.5٪	95.4٪
پوشش	92.7٪	92.5٪

ماتریس درهم‌ریختگی مدل پیشنهادی این پژوهش با حذف داده‌های پرت در شکل (۷) نشان داده شده است. ماتریس درهم‌ریختگی با نشان دادن نمونه‌هایی که به درستی و یا به غلط در هر دسته قرار گرفتند، اطلاعات مهمی در راستای کارایی مدل ارائه می‌کند [۲۲]. همان‌طور که از ماتریس درهم‌ریختگی مشخص است روش پیشنهادی به خوبی قادر است نمونه‌های هر دو کلاس را با دقت خوبی شناسایی کند.

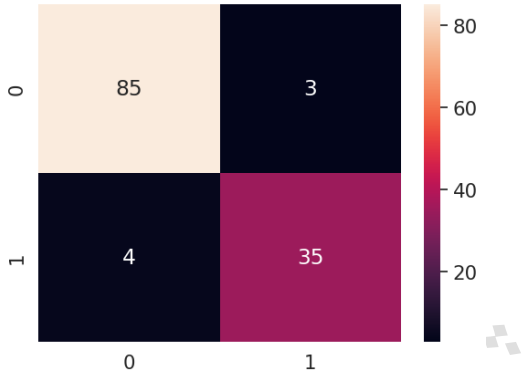
14 Recall

15 True Positive

16 False Positive

17 False Negative

18 True Negative



شکل (۷): ماتریس درهم‌ریختگی روش پیشنهادی برای تشخیص دیابت با حذف داده‌های پرت.

در جدول (۳) نتایج این پژوهش با دیگر کارهای مرتبط مقایسه شده است. این جدول دقت چندین مدل مختلف برای تشخیص بیماری دیابت را نشان می‌دهد. روش پیشنهادی این پژوهش، که از روش رای‌گیری نرم با استفاده از سه رای‌دهنده، K–نزدیک‌ترین همسایه، تقویت گرادیان شدید و کت‌بوست استفاده می‌کند، در مقایسه با مدل‌های موجود [۱۴–۱۶، ۲۳–۲۸] روش‌های ترکیبی، رای‌گیری نرم و شبکه عصبی دقت بیشتری داشته‌اند. از این‌رو، با توجه به مقایسه‌های انجام شده در این جدول، می‌توان گفت که مدل پیشنهادی این پژوهش، مدل کارآمدتری برای تشخیص بیماری دیابت است. با توجه به آن که کیفیت زندگی یک متغیر مهم است، از دیدگاه تحقیقات بالینی و فناوری، طراحی یک مدل یادگیری ماشین بر اساس داده‌های بیوشیمیایی که وضعیت سلامت انسان را ثبت می‌کند به یک نیاز ضروری تبدیل شده است. مدل پیشنهادی این پژوهش، که از سه الگوریتم K–نزدیک‌ترین همسایه، تقویت گرادیان شدید و کت‌بوست برای شرکت در مدل رای‌گیری نرم استفاده کرده است، الگوریتمی امیدبخش برای آینده تشخیص دیابت خواهد بود.

جدول (۳): مقایسه عملکرد روش پیشنهادی با دیگر پژوهش ها روی مجموعه‌داده پیما.

#	مرجع	نوع جداسازی داده	روش	دقت
۱	روش پیشنهادی	۲۰-۸۰	رای‌گیری نرم (با حذف داده‌های پرت)	94.4%
۲	روش پیشنهادی	۲۰-۸۰	رای‌گیری نرم (بدون حذف داده‌های پرت)	93.6%
۳	روش پیشنهادی	۲۰-۸۰	کت‌بوست	92%
۴	[۲]	K-Fold	ماشین بردار پشتیبان	77.3%
۵	[۱۴]	K-Fold	LightGBM+K-NN+Adaboost	90.7%
۶	[۱۵]	K-Fold	رای‌گیری نرم	85.7%
۷	[۱۶]	K-Fold	رای‌گیری سخت	77.%
۸	[۲۳]	K-Fold	رگرسیون لجستیک	77%
۹	[۲۴]	۳۰-۷۰	بیز اصلاح شده با الگوریتم ازدحام ذرات	78.6%
۱۰	[۲۵]	۳۰-۷۰	جنگل تصادفی	79.57%
۱۱	[۲۶]	۳۰-۷۰	درخت تصمیم	75.6%
۱۲	[۲۷]	۲۰-۸۰	رای‌گیری نرم (جنگل تصادفی، بیز، رگرسیون لجستیک)	79.08%
۱۳	[۲۸]	K-Fold	شبکه عصبی عمیق	89%

۶. نتیجه‌گیری

دیابت یک بیماری خود ایمنی است که یکی از عوامل مهم مرگ و میر در سراسر جهان می‌باشد و تأثیرات قابل توجهی در بیماری‌های کلیوی، بیماری‌های قلبی و از دست دادن بینایی دارد. این بیماری می‌توانند سبب امراض متعددی شود و کیفیت زندگی و سلامت انسان‌ها را به خطر بیاندازد. با به‌کارگیری الگوریتم‌های یادگیری ماشین در پزشکی می‌توان الگوهای پنهان بیماری، عوامل موثر در بیماری و تشخیص آن را شناسایی کرد. از این‌رو، تشخیص زودهنگام و دقیق بیماری با استفاده از این الگوریتم‌ها می‌تواند کمک شایانی به جامعه و پزشکان کند. بر این اساس، در

این پژوهش تلاش شده است تا مدلی جهت افزایش دقت در تشخیص بیماری دیابت ارائه شود. برای ساخت این مدل از مجموعه‌داده پیما استفاده شده است. در گام نخست پیش‌پردازش داده‌ها انجام می‌شود، چرا که با بررسی نمونه‌های موجود در مجموعه‌داده می‌توان از ایجاد یک مدل دقیق اطمینان حاصل کرد. از این‌رو، در این پژوهش به حذف داده‌های پرت، جایگزین نمودن مقادیر از دست رفته با میانگین پرداخته شده است. سپس برای کاهش ابعاد و پیدا کردن ویژگی‌های مهم و تاثیرگذار، از الگوریتم لاسو استفاده شده است. در نهایت برای ساخت یک مدل با دقت بالاتر، از الگوریتم رای‌گیری نرم استفاده شده است. این روش، با استفاده از سه الگوریتم رای‌دهنده، K–نزدیک‌ترین همسایه، تقویت گرادیان شدید و کت‌بوست، دقت تشخیص بیماری دیابت را بهبود داده است.

مراجع

[1] N. Arora, A. Singh, M. Z. N. Al-Dabagh, and S. K. Maitra, “A Novel Architecture for Diabetes Patients’ Prediction Using K-Means Clustering and SVM,” Mathematical Problems in Engineering, vol. 2022, pp. 1–9, Aug. 2022, doi: https://doi.org/10.1155/2022/4815521.

[2] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” Procedia Computer Science, vol. 132, pp. 1578–1585, 2018, doi: https://doi.org/10.1016/j.procs.2018.05.122.

[3] Z. Salih Ageed et al., “Comprehensive Survey of Big Data Mining Approaches in Cloud Systems,” Qubahan Academic Journal, vol. 1, no. 2, pp. 29–38, Apr. 2021, doi: https://doi.org/10.48161/qaj.v1n2a46.

[4] W. Haoxiang and S. S, “Big Data Analysis and Perturbation using Data Mining Algorithm,” March 2021, vol. 3, no. 1, pp. 19–28, Apr. 2021, doi: https://doi.org/10.36548/jscp.2021.1.003.

[5] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining,” Informatics in Medicine Unlocked, vol. 10, pp. 100–107, 2018, doi: https://doi.org/10.1016/j.imu.2017.12.006.

[6] Islam, R. Ferdousi, S. Rahman, and H. Bushra, “Likelihood prediction of diabetes at early stage using data mining techniques,” 2020, pp. 113–125. doi: https://doi.org/10.1007/978-981-13-8798-2_12.

[7] F. G. Woldemichael and S. Menaria, “Prediction of Diabetes Using Data Mining Techniques,” 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), May 2018, doi: https://doi.org/10.1109/icoei.2018.8553959.

[8] C. Fiarni, E. M. Sipayung, and S. Maemunah, “Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm,” Procedia Computer Science, vol. 161, pp. 449–457, 2019, doi: https://doi.org/10.1016/j.procs.2019.11.144.

[9] A. Aldallal and A. A. A. Al-Moosa, “Using Data Mining Techniques to Predict Diabetes and Heart Diseases,” IEEE Xplore, Sep. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8552051

[10] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017, doi: https://doi.org/10.1016/j.csbj.2016.12.005.

[11] A. Kumar, P. Kumar, A. Srivastava, A. Kumar, K. Vengatesan, and A. Singhal, “Comparative analysis of data mining techniques to predict heart disease for diabetic patients,” M. Singh, G. P. K, V. Tyagi, J. Flusser, T. Ören, and G. Valentino, Eds., Springer Singapore, 2020, pp. 507–518.

[12] T. R. Mahesh et al., “Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease,” Computational Intelligence and Neuroscience, vol. 2022, p. e4451792, Jul. 2022, doi: https://doi.org/10.1155/2022/4451792.

[13] A. Oza and A. Bokhare, “Diabetes prediction using logistic regression and k-nearest neighbor,” M. Saraswat, H. Sharma, K. Balachandran, J. H. Kim, and J. C. Bansal, Eds., Springer Nature Singapore, 2022, pp. 407–418.

[14] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, “An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes,” International Journal of Computational Intelligence Systems, vol. 16, no. 1, Feb. 2023, doi: https://doi.org/10.1007/s44196-023-00184-y.

[15] A. Mahabub, “A robust voting approach for diabetes prediction using traditional machine learning techniques,” SN Applied Sciences, vol. 1, no. 12, Nov. 2019, doi: https://doi.org/10.1007/s42452-019-1759-7.

[16] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, “Voting Classification-

Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," Mobile Information Systems, vol. 2022, pp. 1–16, Mar. 2022, doi: https://doi.org/10.1155/2022/6521532.
[17] UCI Machine Learning. Pima Indians Diabetes Database. 2016.
[18] Muthukrishnan R, Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. In: 2016 IEEE International Conference on Advances in Computer Applications (ICACA). IEEE; 2016.
[19] H. Veisi, G. H. Reza, and M. Ebrahimi, "Improving the performance of machine learning algorithms for heart disease diagnosis by optimizing data and features," Soft Computing Journal, vol. 8, Art. no. 1, 2021, doi: https://doi.org/10.22052/8.1.70. Available: https://scj.kashanu.ac.ir/article_111441.html
[20] F. Leon, S.-A. Floria, and C. Bădică, "Evaluating the effect of voting methods on ensemble-based classification," 2017, pp. 1–6. doi: https://doi.org/10.1109/INISTA.2017.8001122.
[21] R. Taimourei-Yansary, M. Mirzarezaee, M. Sadeghi, and N. A. Babak, "Predicting invasive disease-free survival time in breast cancer patients using semi-supervised graph-based machine learning techniques," Soft Computing Journal, vol. 10, Art. no. 1, 2022, doi: https://doi.org/10.22052/scj.2022.243330.1039.
[22] R. Akhoondi and R. H, "A novel fuzzy-genetic differential evolutionary algorithm for optimization of a fuzzy expert systems applied to heart disease prediction," Soft Computing Journal, vol. 6, Art. no. 2, 2021, Available: https://scj.kashanu.ac.ir/article_111420.html
[23] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," Measurement: Sensors, p. 100605, Dec. 2022, doi: https://doi.org/10.1016/j.measen.2022.100605.
[24] G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, "Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods," Machines, vol. 7, no. 4, p. 74, Dec. 2019, doi: https://doi.org/10.3390/machines7040074.
[25] D. Choubey, P. Kumar, S. Tripathi, and S. Kumar, "Performance evaluation of classification methods with PCA and PSO for diabetes," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 9, Dec. 2020, doi: https://doi.org/10.1007/s13721-019-0210-8.
[26] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Computing and Applications, Mar. 2022, doi: https://doi.org/10.1007/s00521-022-07049-z.
[27] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," Jan. 2021.
[28] P. Hounguè and A. G. Bigirimana, "Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network," Journal of Computer and Communications, vol. 10, no. 11, pp. 15–28, 2022, doi: https://doi.org/10.4236/jcc.2022.1011002.