

Evaluating the Predictive Performance of Metabolic Biomarkers for 20-Year Incidence Diabetes in the Yazd Healthy Heart Cohort: A Machine Learning Analysis

Akram Ghadiri-Anari¹, Seyedeh Mahdieh Namayandeh², Mohammadreza Ahi^{2*}

1. Diabetes Research Center, Shahid Sadoughi University of Medical Science, Yazd, Iran

2. Clinical Research Development Center, Afshar Hospital, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

Abstract

Background: Type 2 diabetes is a major global health challenge, necessitating the development of accurate prediction models for early intervention. This study aimed to assess the performance of machine learning models in predicting the long-term incidence of diabetes and to identify predictive metabolic biomarkers based on data from the Yazd Healthy Heart Cohort.

Methods: This retrospective study was conducted on 906 non-diabetic individuals from the Yazd Healthy Heart Cohort with a 20-year follow-up. Five machine learning models (Logistic Regression, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest) were implemented based on 17 demographic, clinical, and biochemical variables. The data were randomly split into 70% for training and 30% for testing. Model performance was evaluated using 10-fold cross-validation and metrics including accuracy, sensitivity, specificity, F1-Score, and Area Under the Curve- Receiver Operator Characteristic (AUC-ROC).

Results: During the study, 340 participants (37.5%) developed diabetes. None of the models achieved satisfactory performance (AUC > 0.8). The best performance was observed for the LDA model in the 10-year prediction, with an accuracy of 73% and an AUC of 0.70, although its sensitivity was low (47%). The composite indices triglyceride-glucose (TyG) and atherogenic plasma index (AIP) were identified as the strongest predictors across most models.

Conclusion: Although composite biomarkers such as TyG show significant predictive potential, conventional machine learning models using baseline data are insufficient for predicting diabetes incidence over long-term horizons. The development of dynamic models that incorporate temporal changes in variables and complex pathophysiological interactions is essential to achieve acceptable accuracy in long-term prediction.

Keywords: Type 2 diabetes, Machine learning, Metabolic biomarkers, Cohort, Long-term prediction

Received: September 20, 2025; Accepted: January 10, 2026

Please cite this article as:

Ghadiri-Anari A, Namayandeh SM, Ahi M. Evaluating the Predictive Performance of Metabolic Biomarkers for 20-Year Incidence Diabetes in the Yazd Healthy Heart Cohort: A Machine Learning Analysis. *ijdl*. 2026; 26(3): 283-291.

DOI: [10.18502/ijdl.v26i3.22507](https://doi.org/10.18502/ijdl.v26i3.22507)

*Corresponding Author: Mohammadreza Ahi; Email: ahi.dvm@gmail.com, Add: Afshar Hospital, Jomhuri Blvd, Yazd, Iran, Tel: +983531602125

بررسی قدرت پیش‌گویی بیومارکرهای متابولیک برای تخمین بروز ۲۰ ساله دیابت در کوهورت قلب سالم: یک آنالیز یادگیری ماشین

اکرم قدیری-اناری^۱، سیده مهدیه نماینده^۲، محمدرضا آهی^{۳*}

۱- مرکز تحقیقات دیابت، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران

۲- واحد توسعه تحقیقات بالینی، بیمارستان افشار، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران

چکیده

مقدمه: دیابت نوع دو به‌عنوان یک چالش عمده سلامت جهانی، نیازمند توسعه مدل‌های پیش‌بینی دقیق برای مداخله زودهنگام است. این مطالعه با هدف ارزیابی عملکرد مدل‌های یادگیری ماشین در پیش‌بینی بروز بلندمدت دیابت و شناسایی بیومارکرهای متابولیک پیش‌بین بر پایه داده‌های کوهورت قلب سالم یزد انجام شد.

روش‌ها: این مطالعه گذشته‌نگر بر روی ۹۰۶ فرد بدون دیابت از کوهورت قلب سالم یزد با پیگیری ۲۰ ساله انجام شد. پنج مدل یادگیری ماشین (رگرسیون لجستیک، Linear Discriminant Analysis (LDA)، Support Vector Machine (SVM)، K-Nearest Neighbors (KNN) و جنگل تصادفی) براساس ۱۷ متغیر دموگرافیک، بالینی و بیوشیمیایی پیاده‌سازی شدند. داده‌ها به‌صورت تصادفی به ۷۰٪ برای آموزش و ۳۰٪ برای آزمون تقسیم شدند. عملکرد مدل‌ها با اعتبارسنجی متقابل ۱۰-تایی و معیارهای دقت، حساسیت، ویژگی، F1-Score و Area (AUC-ROC) Under the Curve- Receiver Operator Characteristic ارزیابی شد.

یافته‌ها: در طول مطالعه، ۳۴۰ شرکت‌کننده (۳۷/۵٪) به دیابت مبتلا شدند. هیچ‌یک از مدل‌ها به عملکرد مطلوب یعنی AUC بیش از ۰/۸۰ دست نیافتند. بهترین عملکرد مربوط به مدل LDA در پیش‌بینی ۱۰ ساله با دقت ۷۳٪ و AUC برابر ۰/۷۰ بود، هرچند حساسیت آن نیز پایین (۴۷٪) بود. نمایه‌های ترکیبی تری‌گلیسیرید-گلوکز (TyG) و نمایه آتروژنیک پلاسما (AIP) در اکثر مدل‌ها به‌عنوان قوی‌ترین پیش‌بین‌ها شناسایی شدند. **نتیجه‌گیری:** اگرچه بیومارکرهای ترکیبی مانند TyG پتانسیل پیش‌بینی قابل توجهی دارند، اما مدل‌های متعارف یادگیری ماشین با داده‌های پایه اولیه برای پیش‌بینی بروز دیابت در افق‌های زمانی طولانی مدت ناکافی هستند. توسعه مدل‌های پویا که تغییرات زمانی متغیرها و تعاملات پیچیده پاتوفیزیولوژیک را در نظر می‌گیرند، برای دستیابی به دقت قابل قبول در پیش‌بینی بلندمدت ضروری به نظر می‌رسد.

واژگان کلیدی: دیابت نوع دو، یادگیری ماشین، بیومارکرهای متابولیک، کوهورت، پیش‌بینی بلندمدت

تاریخ دریافت: ۱۴۰۴/۰۶/۲۹

تاریخ پذیرش: ۱۴۰۴/۱۰/۲۰

به این مقاله، به صورت زیر استناد کنید:

Ghadiri-Anari A, Namayandeh SM, Ahi M. Evaluating the Predictive Performance of Metabolic Biomarkers for 20-Year Incidence Diabetes in the Yazd Healthy Heart Cohort: A Machine Learning Analysis. *ijddl*. 2026; 26(3): 283-291.

*نویسنده مسئول: محمدرضا آهی، آدرس: یزد، بلوار جمهوری، بیمارستان افشار، تلفن: ۰۳۵۳۱۶۰۲۱۲۵، پست الکترونیک: ahi.dvm@gmail.com

مقدمه

دیابت ملیتوس نوع دو^۱ (T2DM) به‌عنوان یک اختلال متابولیک چندعاملی، یکی از عمده‌ترین چالش‌های سیستم‌های سلامت در سراسر جهان محسوب می‌شود [۱]. این بیماری با عوارض میکروواسکولار و ماکروواسکولار متعددی از قبیل نروپاتی، رتینوپاتی، نوروپاتی و بیماری‌های قلبی-عروقی همراه است که بار اقتصادی و اجتماعی سنگینی را به جوامع تحمیل می‌کند [۲]. پیش‌بینی به‌موقع افراد در معرض خطر ابتلا، سنگ بنای هر راهبرد پیشگیرانه مؤثر است. در این راستا، شناسایی و به‌کارگیری بیومارکرهای قابل اعتماد و در دسترس از اهمیت فوق‌العاده‌ای برخوردار است.

در سال‌های اخیر، تمرکز تحقیقات از بیومارکرهای سنتی مانند گلوکز ناشتا و HbA1c^۲ به سمت نمایه‌های ترکیبی نوینی که منعکس‌کننده تعاملات پیچیده‌تر متابولیکی هستند، معطوف شده است [۳]. از جمله این نمایه‌ها می‌توان به نمایه تری‌گلیسیرید-گلوکز^۳ (TyG) اشاره کرد که به‌عنوان یک جایگزین غیرتهاجمی برای ارزیابی مقاومت به انسولین پیشنهاد شده است [۴]. مطالعات اپیدمیولوژیک متعددی ارتباط قوی بین سطوح بالای نمایه TyG و افزایش خطر ابتلا به T2DM را نشان داده‌اند [۶]. بیومارکر ترکیبی دیگر، نمایه آتروژنیک پلاسما^۴ (AIP) است که از لگاریتم نسبت تری‌گلیسیرید به کلسترول لیپوپروتئین با چگالی بالا محاسبه می‌شود و نه تنها یک نمایه خطر قلبی-عروقی قوی است، بلکه در شناسایی افراد پیش‌دیابتی در معرض خطر نیز نویدبخش است [۷].

به موازات توسعه این بیومارکرها، تکنیک‌های هوش مصنوعی و یادگیری ماشین^۵ (ML) به‌عنوان ابزارهای قدرتمندی برای تجزیه و تحلیل مجموعه داده‌های پیچیده و چندبعدی در حوزه سلامت ظهور کرده‌اند [۸]. الگوریتم‌های ML از قبیل رگرسیون لجستیک^۶ (LR)، ماشین بردار پشتیبان^۷ (SVM) و جنگل تصادفی^۸ (RF) توانایی خود را در مدل‌سازی روابط غیرخطی و پیش‌بینی پیامدهای سلامت در مطالعات متعدد نشان داده‌اند [۹، ۱۰]. با این وجود، کاربرد این مدل‌ها در پیش‌بینی بروز دیابت با استفاده از بیومارکرهای متابولیک در مطالعات کوهورت با پیگیری بسیار طولانی مدت (۲۰ ساله) هنوز به‌طور گسترده مورد بررسی قرار نگرفته است.

این شکاف دانشی، ما را بر آن داشت تا با استفاده از داده‌های با ارزش و منحصر به فرد کوهورت قلب سالم یزد، که یکی از طولانی‌ترین دوره‌های پیگیری در منطقه را داراست، به بررسی این سؤال پردازیم که: «دقت مدل‌های رایج یادگیری ماشین در پیش‌بینی بروز ۲۰ ساله دیابت چقدر است و کدام یک از بیومارکرهای متابولیک از قدرت پیش‌بینی‌کنندگی بالاتری برخوردارند؟»

روش‌ها

طرح مطالعه و جامعه مورد بررسی

این پژوهش یک مطالعه تحلیلی از نوع کوهورت گذشته‌نگر است که بر روی داده‌های جمع‌آوری شده در قالب «کوهورت قلب سالم یزد» انجام گرفت. جامعه مطالعه را ۲۰۰۰ نفر از شهروندان ساکن شهر یزد تشکیل می‌دادند که در فاز پایه (سال ۱۳۸۲) برای مطالعه کوهورت قلب سالم یزد انتخاب شده بودند. برای ورود به آنالیز حاضر، شرکت‌کنندگان می‌بایست در درجه اول فاقد تشخیص دیابت در ابتدای مطالعه باشند، در درجه دوم در فاز پایه در محدوده سنی ۳۰ تا ۷۰ سال قرار داشته باشند و در نهایت اطلاعات کامل در مورد متغیرهای مورد بررسی در پرونده آنها موجود باشد. بر این اساس، از بین ۲۰۰۰ شرکت‌کننده اولیه، در نهایت ۹۰۶ نفر که فاقد هر گونه داده گمشده در متغیرهای پیش‌بین و پیامد بودند، وارد آنالیز نهایی شدند. در این مطالعه از هیچ روش جایگزینی (Imputation) برای داده‌های گمشده استفاده نشد.

متغیرهای مطالعه

متغیر پاسخ در این مطالعه، «بروز دیابت نوع دو» در طول دوره پیگیری ۲۰ ساله تعریف شد. تشخیص دیابت براساس آخرین آزمایش ثبت شده در بانک اطلاعاتی کوهورت و مطابق با معیارهای انجمن دیابت آمریکا صورت پذیرفت: سطح گلوکز پلاسمای ناشتا بیشتر یا مساوی ۱۲۶ میلی‌گرم بر دسی‌لیتر و یا مصرف منظم داروهای کاهنده قند خون (خوراکی یا تزریقی). در این مطالعه هفده متغیر پیش‌بین در سه دسته کلی مورد بررسی قرار گرفتند: (۱) متغیرهای دموگرافیک و سابقه (سن، جنس، سابقه خانوادگی دیابت در بستگان درجه اول)، (۲)

^۵ Machine Learning^۶ Logistic Regression^۷Support Vector Machine^۸Random Forest^۱Diabetes Mellitus Type-2^۲Glycated Hemoglobin^۳Triglyceride-Glucose Index^۴Atherogenic Plasma Index

متغیرهای بالینی و آنتروپومتریک (نمایه توده بدنی، وضعیت مصرف سیگار، فشار خون سیستولیک و دیاستولیک) و (۳) بیومارکرهای بیوشیمیایی و متابولیک [HbA1C، کلسترول تام، تری‌گلیسیرید، کلسترول لیپوپروتئین با چگالی بالا و پایین، نیتروژن اوره خون، نمایه تری‌گلیسیرید-گلوکز (TyG)، نمایه تری‌گلیسیرید-گلوکز-نمایه توده بدنی^۱ (TyG-BMI)، نسبت تری‌گلیسیرید به کلسترول لیپوپروتئین با چگالی بالا (TG/HDL-C)^۲ و نمایه آتروژنیک پلاسما (AIP)].

سطوح زمانی پیش‌بینی

به‌منظور بررسی قدرت پیش‌بینی مدل‌ها در افق‌های زمانی مختلف، پیش‌بینی بروز دیابت در سه سطح تعریف و اجرا شد: (۱) سطح اول (پیش‌بینی ۱۰ ساله): پیش‌بینی بروز دیابت از شروع مطالعه تا پایان سال دهم، با استفاده از داده‌های پایه ثبت‌شده در ابتدای مطالعه.

(۲) سطح دوم (پیش‌بینی ۱۰ تا ۲۰ سال): پیش‌بینی بروز دیابت در فاصله سال‌های ۱۰ تا ۲۰، تنها در بین افرادی که تا پایان سال دهم هنوز به دیابت مبتلا نشده بودند.

(۳) سطح سوم (پیش‌بینی تجمعی ۲۰ ساله): پیش‌بینی تجمعی بروز دیابت از شروع مطالعه تا پایان سال بیستم، با استفاده از داده‌های پایه.

لازم به تأکید است که برای هر یک از متغیرهای پیش‌بین، اندازه‌گیری در زمان شروع مطالعه (سال پایه برای سطح زمانی اول و سوم)، سال دهم (سال پایه برای سطح زمانی دوم و سال پایان برای سطح زمانی اول) و همچنین سال بیستم (سال پایان برای سطح زمانی دوم و سوم) موجود بود. در طول دوره پیگیری ۲۰ ساله، هیچ اندازه‌گیری مکرری از این متغیرها انجام نشد. بنابراین، مدل‌های ارائه‌شده صرفاً براساس یک «تصویر لحظه‌ای» از وضعیت متابولیک افراد، اقدام به پیش‌بینی خطر در آینده دور می‌کنند.

پیاده‌سازی مدل‌های یادگیری ماشین و ارزیابی عملکرد

پنج الگوریتم استاندارد یادگیری ماشین انتخاب و پیاده‌سازی شدند. این الگوریتم‌ها شامل رگرسیون لجستیک (LR)، تحلیل تشخیص خطی (LDA)^۳، ماشین بردار پشتیبان (SVM) با کرنل خطی، K-نزدیک‌ترین همسایه^۴ (K-NN) با $k=5$ و جنگل تصادفی (F)

مشکل از ۱۰۰ درخت تصمیم بودند. انتخاب این پنج الگوریتم براساس سابقه گسترده استفاده از این مدل‌ها در مسائل طبقه‌بندی پزشکی، سرعت اجرا، قابلیت تفسیرپذیری نسبی (به‌ویژه در مدل‌های خطی) و توانایی آنها در مدل‌سازی روابط خطی و غیرخطی بین متغیرها بود. این انتخاب امکان مقایسه طیفی از روش‌های ساده تا پیچیده را فراهم می‌کند.

مجموعه داده به‌صورت تصادفی و با حفظ نسبت بروز دیابت (Stratified Split) به دو بخش آموزش (۷۰٪) و آزمون (۳۰٪) تقسیم شد. برای اطمینان از استحکام و قدرت تعمیم‌پذیری تخمین‌های عملکرد مدل‌ها، یک راهبرد اعتبارسنجی متقابل (cross-validation) ۱۰-تایی (10-fold) در مرحله آموزش و ارزیابی مدل‌ها به‌کار گرفته شد. به‌دلیل متفاوت بودن مقیاس متغیرها، تمامی ویژگی‌ها با استفاده از StandardScaler استاندارد شدند. به‌منظور کاهش ابعاد داده و حذف هم‌خطی (Multicollinearity) بین متغیرها، از روش تحلیل مؤلفه‌های اصلی^۵ (PCA) استفاده شد. تعداد مؤلفه‌های اصلی به گونه‌ای انتخاب شد که بیش از ۹۵٪ از واریانس داده‌ها را حفظ کنند. لازم به توضیح است که در این مطالعه، فرآیند کاهش ابعاد با استفاده از تحلیل مؤلفه‌های اصلی (PCA) به‌صورت مجزا و متناسب با ویژگی‌های هر زیرمجموعه از داده‌ها انجام شد. از آنجاکه عملکرد مدل‌های مختلف یادگیری ماشین و نیز قدرت پیش‌بینی متغیرها در سطوح زمانی متفاوت ممکن است متغیر باشد، تصمیم بر این شد که PCA به‌صورت مستقل برای هر مدل در هر سطح زمانی اجرا گردد. بنابراین، برای ۵ مدل یادگیری ماشین (LDA، رگرسیون لجستیک، KNN، SVM و جنگل تصادفی) در ۳ سطح زمانی (۱۰ ساله، ۱۰-۲۰ ساله و ۲۰ ساله)، در مجموع ۱۵ بار فرآیند PCA به‌طور جداگانه اعمال شد.

در هر یک از این ۱۵ اجرا، معیار توقف انتخاب مؤلفه‌ها، تبیین بیش از ۹۵ درصد از واریانس کل داده‌های مربوط به آن مدل و سطح زمانی خاص بود. با توجه به تفاوت‌های طبیعی در ساختار داده‌ها و همبستگی بین متغیرها در هر یک از این ۱۵ سناریو، تعداد مؤلفه‌های اصلی انتخاب‌شده اندکی متفاوت بود. بر این اساس، در هر اجرا بین ۸ تا ۱۰ مؤلفه اصلی انتخاب شدند که حد آستانه ۹۵٪ واریانس را تأمین می‌کردند. مؤلفه‌های اصلی انتخاب شده در پیش‌بینی نهایی برای هر مدل، در تحلیل اهمیت متغیرها لحاظ شده و نتایج آن در جدول ۳ منعکس گردیده است.

⁴ K-Nearest Neighbors

⁵Principal Component Analysis

¹ Triglyceride-Glucose-Body Mass Index

²Triglyceride/High Density Lipoprotein-Cholesterol

³ Linear Discriminant Analysis

یافته‌ها

ویژگی‌های پایه شرکت‌کنندگان

در مجموع، ۹۰۶ شرکت‌کننده با میانگین سنی ۱۱/۴۴ (انحراف معیار: ۱۳/۴۱) سال در آنالیز نهایی قرار گرفتند. ۵۳/۶٪ از شرکت‌کنندگان را مردان (۴۸۶ نفر) و ۴۶/۴٪ را زنان (۴۲۰ نفر) تشکیل می‌دادند. در طول ۲۰ سال پیگیری، ۳۴۰ نفر (۳۷/۵٪) به دیابت مبتلا شدند. از این تعداد، ۲۰۷ مورد (۲۲/۸٪ از کل نمونه) در دهه اول و ۱۳۳ مورد (۱۴/۷٪ از کل نمونه) در فاصله سال‌های ۱۰ تا ۲۰ مطالعه به این عارضه مبتلا گردیدند.

عملکرد مدل‌های یادگیری ماشین در سطوح زمانی مختلف

پیش‌بینی بروز دیابت در سه سطح زمانی مجزا مورد بررسی قرار گرفت. در سطح اول پیش‌بینی بروز دیابت از شروع مطالعه تا پایان سال دهم، با استفاده از داده‌های پایه انجام شد. در سطح دوم پیش‌بینی بروز دیابت در فاصله سال‌های ۱۰ تا ۲۰، در بین افرادی که تا پایان سال دهم هنوز به دیابت مبتلا نشده بودند صورت گرفت. در سطح سوم هم پیش‌بینی تجمعی بروز دیابت از شروع مطالعه تا پایان سال بیستم انجام شد. جدول ۱ عملکرد کلی مدل‌ها را در این سه سطح براساس معیارهای دقت، امتیاز F1 و AUC نشان می‌دهد. به‌منظور اطمینان از پایایی و قابلیت تعمیم نتایج، ارزیابی عملکرد تمامی مدل‌ها با به‌کارگیری روش اعتبارسنجی متقابل ۱۰-تایی انجام پذیرفت.

این رویکرد امکان حفظ حداکثر اطلاعات مؤثر و در عین حال حذف هم‌خطی را به‌صورت خاص برای هر شرایط ارزیابی فراهم آورد. برای شناسایی مؤثرترین متغیرهای پیش‌بین در هر مدل، از روش‌های مناسب با ساختار هر الگوریتم استفاده شد. برای مدل‌های خطی (LDA و رگرسیون لجستیک)، ضرایب استاندارد شده به‌عنوان معیار اهمیت در نظر گرفته شد. برای مدل جنگل تصادفی، معیار کاهش ناخالصی (Gini Importance) محاسبه گردید. برای مدل‌های SVM و K-NN، از روش اهمیت مبتنی بر جایگشت (Permutation Importance) استفاده شد. عملکرد هر مدل بر روی مجموعه آزمون با استفاده از معیارهای دقت (Accuracy)، Recall یا حساسیت (Sensitivity)، ویژگی (Specificity)، ارزش اخباری مثبت (Positive Predictive Value- PPV)، ارزش اخباری منفی (Negative Predictive Value- NPV)، امتیاز F1 (F1-Score) و مساحت زیرمنحنی مشخصه عملکرد گیرنده^۱ (ROC AUC) ارزیابی شد.

تحلیل‌های آماری و ملاحظات اخلاقی

تمامی محاسبات با استفاده از زبان برنامه‌نویسی پایتون (نسخه ۳.۱۱.۵) و کتابخانه scikit-learn (نسخه ۱.۲.۲) انجام گرفت. این مطالعه پس از اخذ تأییدیه از کمیته اخلاق دانشگاه علوم پزشکی شهید صدوقی یزد با کد اخلاق IR.SSU.SRH.REC.1403.032 انجام شد و محرمانگی داده‌ها در تمام مراحل رعایت گردید.

جدول ۱- مقایسه عملکرد مدل‌های یادگیری ماشین در پیش‌بینی بروز دیابت در سطوح زمانی مختلف بر روی مجموعه آزمون (۲۷۲ نفر) از کوهورت قلب سالم یزد

مدل‌ها	سطح اول (۱۰ سال)			سطح دوم (۱۰-۲۰ سال)			سطح سوم (۲۰ سال)		
	دقت	AUC	F1-Score	دقت	AUC	F1-Score	دقت	AUC	F1-Score
رگرسیون لجستیک	۶۷٪	۰/۶۵	۰/۵۵	۵۵٪	۰/۵۲	۰/۴۵	۶۸٪	۰/۶۶	۰/۵۶
LDA	۷۳٪	۰/۷	۰/۵۷	۵۶٪	۰/۵۴	۰/۴۷	۶۹٪	۰/۶۷	۰/۵۸
SVM	۷۱٪	۰/۶۸	۰/۵۴	۴۷٪	۰/۴۸	۰/۳۹	۶۶٪	۰/۶۴	۰/۵۳
KNN	۵۵٪	۰/۵۲	۰/۴۲	۵۱٪	۰/۵	۰/۴۱	۵۹٪	۰/۵۶	۰/۴۶
جنگل تصادفی	۶۲٪	۰/۶	۰/۴۹	۵۵٪	۰/۵۲	۰/۴۴	۶۵٪	۰/۶۳	۰/۵۲

AUC: Area Under the Curve; LDA: Linear Discriminant Analysis; SVM: Support Vector Machine; KNN: K-Nearest Neighbors

کلی در سطح اول (پیش‌بینی ۱۰ ساله) و توسط مدل LDA با دقت ۷۳٪ و AUC برابر ۰/۷۰ مشاهده شد. عملکرد تمامی

همان‌طور که در جدول ۱ مشهود است، هیچ‌یک از مدل‌ها به عملکرد مطلوب (AUC > ۰/۸) دست نیافتند. بهترین عملکرد

^۱Area Under the Curve- Receiver Operator Characteristic

مدل‌ها در سطح دوم (پیش‌بینی وقوع دیابت در دهه دوم) به‌طور محسوسی ضعیف‌تر بود.

عملکرد تفکیک‌شده مدل LDA در سطح زمانی اول

با توجه به عملکرد نسبی بهتر مدل LDA در سطح اول، ماتریس درهم‌آمیختگی (Confusion Matrix) و سایر معیارهای عملکردی آن بر روی مجموعه آزمون در این سطح به تفکیک

محاسبه و در جدول ۲ ارائه شده است. لازم به توضیح است که ماتریس درهم‌آمیختگی ارائه شده در جدول ۲، بر روی مجموعه آزمون محاسبه شده است. مطابق پروتکل تقسیم‌بندی داده‌ها، ۷۰٪ از داده‌ها (۶۳۴ نفر) برای آموزش و ۳۰٪ (۲۷۲ نفر) برای آزمون مدل به‌صورت تصادفی و با حفظ نسبت بروز دیابت انتخاب شدند. بنابراین، ارزیابی عملکرد نهایی مدل‌ها (از جمله ماتریس درهم‌آمیختگی) تنها بر روی این ۲۷۲ نفر (مجموعه آزمون) انجام شده است.

جدول ۲- عملکرد تفکیک‌شده مدل LDA در پیش‌بینی بروز دیابت در سطح اول (۱۰ ساله) بر روی مجموعه آزمون (۲۷۲ نفر)

معیار	مقدار	تفسیر	ماتریس درهم‌آمیختگی			
حساسیت (Recall)	۰/۴۷	نسبت مبتلایان واقعی که به درستی شناسایی شدند	ماتریس درهم‌آمیختگی			
ویژگی (Specificity)	۰/۹۵	نسبت افراد سالم که به درستی شناسایی شدند				
ارزش اخباری مثبت (PPV)	۰/۷۳	نسبت موارد پیش‌بینی‌شده مثبت که واقعاً مبتلا بودند	واقعی:	۱۹۹ (TN)	پیش‌بینی مثبت	۲۱۰
ارزش اخباری منفی (NPV)	۰/۸۶	نسبت موارد پیش‌بینی‌شده منفی که واقعاً سالم بودند	منفی:	۳۳ (FN)	پیش‌بینی منفی	۶۲
امتیاز (F1)	۰/۵۷	میانگین توازن‌یافته حساسیت و دقت	مجموع	۲۳۲	مجموع	۲۷۲

PPV: Positive Predictive Value; NPV: Negative Predictive Value; FP: False Positive; TN: True Negative; TP: True Positive; FN: False Negative

نتایج جدول ۲ نشان می‌دهد که اگرچه مدل LDA از ویژگی بسیار بالایی (۰/۹۵) برخوردار است و می‌تواند افراد سالم را به خوبی شناسایی کند، اما حساسیت آن بسیار پایین (۰/۴۷) است. این بدان معناست که این مدل تقریباً نیمی از افرادی که در آینده به دیابت مبتلا خواهند شد را شناسایی نمی‌کند (مثبت کاذب بالا). امتیاز F1 پایین (۰/۵۷) نیز این ضعف را تأیید می‌کند. بنابراین حتی بهترین مدل در بهترین شرایط زمانی این مطالعه، از نظر بالینی برای غربالگری قابل اتکا نیست. بنابراین، به‌منظور پرهیز از تفصیل غیرضروری و تأکید بر این پیام کلیدی، از ارائه جداول مشابه برای سطوح زمانی دیگر یا مدل‌های با عملکرد ضعیف‌تر خودداری شد.

اهمیت نسبی متغیرهای پیش‌بین

به منظور شناسایی مؤثرترین متغیرها در پیش‌بینی بروز دیابت، اهمیت نسبی متغیرهای پیش‌بین در تمامی پنج مدل یادگیری ماشین و در هر سه سطح زمانی مورد تحلیل قرار گرفت. به این

منظور از ضرایب استاندارد شده برای مدل‌های خطی (LDA و رگرسیون لجستیک) و معیار کاهش ناخالصی (Gini Importance) برای مدل‌های مبتنی بر درخت (جنگل تصادفی) استفاده شد. برای مدل‌های SVM و K-NN نیز از روش‌های مبتنی بر جایگشت (Permutation Importance) استفاده شد.

نتایج این تحلیل که در جدول ۳ ارائه شده است، الگوهای جالبی را نشان می‌دهد. مهم‌ترین یافته، ثبات و پایداری رتبه دو نمایه ترکیبی TyG (نمایه تری‌گلیسیرید-گلوکز) و AIP (نمایه آتروژنیک پلاسما) در اکثر مدل‌ها و سطوح زمانی مختلف بود. جالب توجه آنکه نمایه توده بدنی (BMI) به تنهایی در رتبه‌های میانی قرار گرفت، اما زمانی که در قالب نمایه ترکیبی TyG-BMI ظاهر شد، اهمیت بسیار بالاتری را در مدل‌های مبتنی بر درخت مانند جنگل تصادفی نشان داد. این موضوع به وضوح بر برتری نمایه‌های تلفیقی که تعاملات پیچیده‌تر متابولیک را منعکس می‌کنند، در مقایسه با پارامترهای منفرد تأکید دارد.

بحث

این مطالعه با بهره‌گیری از داده‌های یک کوهورت گذشته‌نگر با پیگیری ۲۰ ساله، به ارزیابی قدرت پیش‌گویی مدل‌های رایج یادگیری ماشین و بیومارکرهای متابولیک در پیش‌بینی بروز دیابت نوع دو پرداخت. یافته اصلی ما نشان داد که علی‌رغم استفاده از الگوریتم‌های مختلف و بیومارکرهای نوین، هیچ‌یک از مدل‌ها نتوانستند به سطح

دقت قابل قبولی برای استفاده در محیط بالینی دست یابند. بهترین عملکرد متعلق به مدل LDA با AUC برابر ۰/۷۰ در پیش‌بینی ۱۰ ساله بود که عمدتاً ناشی از ویژگی بالا (۰/۹۵) و در مقابل، حساسیت بسیار پایین (۰/۴۷) آن بود. این الگوی عملکردی حاکی از آن است که مدل مذکور تمایل زیادی به «پیش‌بینی منفی» دارد و در شناسایی واقعی افراد در معرض خطر (که هدف اصلی غربالگری است) ناتوان عمل می‌کند.

جدول ۳- رتبه‌بندی اهمیت متغیرهای پیش‌بین در مدل‌های مختلف (رتبه ۱ بالاترین اهمیت) بر اساس داده‌های پایه ۹۰۶ نفر

متغیرها	سطح اول (۱۰ ساله)					سطح دوم (۱۰-۲۰ سال)					سطح سوم (۲۰ ساله)				
	RF	KNN	SVM	LDA	LR	RF	KNN	SVM	LDA	LR	RF	KNN	SVM	LDA	LR
TyG	1	2	2	2	1	-	1	1	1	1	2	2	2	2	1
AIP	2	1	1	1	2	-	5	5	3	-	1	1	1	1	2
TyG-BMI	-	-	-	-	-	4	3	3	4	7	4	4	5	-	4
TG	9	7	7	-	9	2	2	2	2	4	-	-	-	-	7
TG/HDL-C	6	9	9	-	6	-	-	-	-	2	8	8	6	6	8
BMI	-	-	-	7	-	7	7	7	7	-	6	7	4	4	6
BUN	7	4	4	8	7	-	-	-	-	-	7	7	8	10	5
HDL-C	3	3	3	3	-	-	-	-	-	10	3	3	4	3	-
Age	4	5	5	6	-	-	-	-	-	8	3	5	-	-	1
Family History	-	-	-	-	-	8	8	8	8	6	6	6	9	7	-
Cholesterol	8	-	-	5	9	6	6	-	-	9	-	-	-	9	9
LDL-C	5	6	6	4	-	-	-	-	-	-	-	-	-	8	-
HbA1C	-	-	-	-	-	4	4	5	3	3	-	-	-	-	2
Sex	-	8	8	-	5	-	-	6	5	-	-	-	10	-	-
Smoke	-	-	-	-	-	-	-	-	-	-	5	5	7	5	-
SBP	-	-	-	-	-	3	-	-	-	-	-	-	-	-	3
DBP	-	-	-	-	-	10	-	-	-	-	-	-	-	-	10

توجه: خانه‌های دارای "-" در این جدول به معنای قرار نگرفتن آن متغیر در میان متغیرهای پیش‌بین آن مدل خاص است

LR: Logistic Regression; LDA: Linear Discriminant Analysis; SVM: Support Vector Machine; KNN K-Nearest Neighbors, RF: Random Forest; TyG: Triglyceride-Glucose; AIP: Atherogenic Index of Plasma; TyG-BMI: Triglyceride-Glucose-Body Mass Index; TG: Triglyceride; TG/HDL-C: Triglyceride/High Density Lipoprotein-Cholesterol; BMI: Body Mass Index; BUN: Blood Urea Nitrogen; HDL-C: High Density Lipoprotein-Cholesterol; LDL-C: Low Density Lipoprotein-Cholesterol; HbA1C: Glycated Hemoglobin; SBP: Systolic Blood Pressure; DBP: Diastolic Blood Pressure

یافته‌های ما از نظر محدوده دقت و AUC به‌دست آمده، با نتایج برخی مطالعات مشابه هم‌خوانی دارد. برای مثال، مطالعه Haneef و همکاران (۲۰۲۱) که با استفاده از داده‌های ثبت سلامت فرانسه انجام شد، به دقت ۶۷٪ در پیش‌بینی بروز دیابت دست یافت [۱۲]. به‌طور مشابه، مطالعه Leihner و همکاران (۲۰۲۴) که بر روی داده‌های متابولومیک کار کرد، دقت ۷۷٪ را برای پیش‌بینی دیابت گزارش نمود [۱۳]. این هم‌خوانی نشان می‌دهد که پیش‌بینی دیابت در سطح جمعیت و با افق زمانی طولانی، حتی با روش‌های پیشرفته نیز با چالش مواجه است. در مقابل، مطالعاتی مانند Thabit و همکاران (۲۰۲۲)، Ditsas و Trigka (۲۰۲۲) و همچنین Liu و همکاران (۲۰۲۴) به دقت‌های بسیار بالایی (بیش از ۹۸٪) دست یافته‌اند [۱۶-۱۴]. این اختلاف فاحش می‌تواند ناشی از عوامل روش‌شناختی متعددی مانند حجم و نوع مجموعه داده (مثلاً استفاده از داده‌های بسیار گزینش‌شده یا کوچک‌تر)، پروتکل‌های متفاوت پیش‌پردازش و مهندسی ویژگی، نوع الگوریتم‌های به‌کار رفته و تفاوت در معیارهای ارزیابی باشد. بنابراین، مقایسه مستقیم اعداد بدون در نظر

گرفتن زمینه روش‌شناسی هر مطالعه می‌تواند گمراه‌کننده باشد. اگرچه مدل‌های کلاسیک آماری مانند رگرسیون لجستیک به‌دلیل سادگی و تفسیرپذیری بالا همچنان پرکاربرد هستند، مدل‌های یادگیری ماشین (مانند SVM و جنگل تصادفی) این پتانسیل را دارند که روابط غیرخطی و تعاملات پیچیده‌تری را بین متغیرها شناسایی کنند. با این حال، نتایج این مطالعه نشان داد که حتی این مدل‌های پیشرفته‌تر نیز نتوانستند به دقت قابل قبولی در افق زمانی بسیار طولانی دست یابند. این امر نشان می‌دهد که ماهیت مسئله (پیش‌بینی بلندمدت با داده‌های ایستا) ممکن است چالش بزرگ‌تری نسبت به نوع الگوریتم باشد.

یکی از یافته‌های مهم و ثابت در بین مدل‌های مختلف در این مطالعه، اهمیت بالای نمایه‌های ترکیبی، به‌ویژه TyG و AIP بود (جدول ۳). این نتایج، یافته‌های مطالعات قبلی را تأیید می‌کند که TyG را به‌عنوان یک نمایه قوی برای مقاومت به انسولین و پیش‌بینی دیابت معرفی کرده‌اند [۵، ۶]. به‌نظر می‌رسد که این نمایه‌ها نسبت به نمایه‌های منفرد، توانایی بهتری در انعکاس تعاملات پیچیده متابولیکی

نتیجه‌گیری

جمع‌بندی نهایی این مطالعه حاکی از آن است که اگرچه بیومارکرهای ترکیبی متابولیک (به‌ویژه TyG و AIP) به‌عنوان پیش‌بین‌های قدرتمندی شناسایی شدند، اما مدل‌های متعارف یادگیری ماشین با داده‌های پایه‌اولیه، برای پیش‌بینی دقیق بروز دیابت در یک افق طولانی مدت کافی نیستند. چالش اصلی، پویایی ذاتی بیماری دیابت و تغییرات عوامل خطر در طول زمان است. بنابراین، حرکت به سوی توسعه مدل‌های پیش‌بینی پویا و یکپارچه‌گرا که قادر به شبیه‌سازی روند تکاملی بیماری باشند، یک گام ضروری و اجتناب‌ناپذیر در مسیر پیش‌بینی واقعی و به موقع این اپیدمی جهانی است. تأکید بر ناکافی بودن مدل‌های ایستا و لزوم حرکت به سمت مدل‌های پویا، مهم‌ترین پیام این پژوهش است.

تضاد منافع

نویسندگان هیچ‌گونه تعارض منافع از نظر مادی و معنوی ندارند.

سپاسگزاری

نویسندگان از مرکز توسعه تحقیقات بالینی (RDC) بیمارستان افشار دانشگاه علوم پزشکی شهید صدوقی یزد، به‌خاطر حمایت، همکاری و مساعدت‌هایشان در طول دوره مطالعه، صمیمانه سپاسگزارند.

پاتورنز دیابت دارند. اهمیت بالای شاخص TyG در مطالعه ما، با پژوهش‌های اخیر که نقش این شاخص را در حوزه‌های مختلف متابولیک بررسی کرده‌اند نیز همسو بوده و بر جایگاه ویژه این شاخص ترکیبی به‌عنوان یک بیومارکر فراگیر متابولیک تأکید می‌کند [۱۷].

محدودیت‌ها

این مطالعه با چندین محدودیت روبرو بود که باید در تفسیر نتایج در نظر گرفته شوند:

(۱) **ماهیت پویای دیابت در مقابل داده‌های ایستا:** مدل‌های ما براساس یک «تصویر» لحظه‌ای (داده‌های پایه) از افراد ساخته شدند. درحالی‌که دیابت یک بیماری پویا است که تحت تأثیر تغییرات تدریجی در سبک زندگی، وزن، رژیم غذایی و وضعیت متابولیک در طول دهه‌ها قرار دارد. عدم در نظرگیری این تغییرات زمانی یک محدودیت عمده است.

(۲) **فرض استقلال مشاهدات:** ما از مدل‌های کلاسیک یادگیری ماشین استفاده کردیم که فرض می‌کنند مشاهدات مستقل از یکدیگر هستند. درحالی‌که در داده‌های طولی، اندازه‌گیری‌های مکرر روی یک فرد وابسته هستند. این امر می‌تواند بر برآورد دقت واقعی مدل تأثیر بگذارد.

(۳) **حجم نمونه:** اگرچه کوهورت قلب سالم از مطالعات طولانی‌مدت ارزشمند است، اما حجم نمونه نهایی (۹۰۶ نفر) برای آموزش مدل‌های بسیار پیچیده ممکن است کافی نباشد.

(۴) **عدم در نظرگیری سایر عوامل:** عوامل دیگری مانند رژیم غذایی دقیق، فعالیت بدنی کمی، استرس و عوامل اجتماعی-اقتصادی که در خطر دیابت نقش دارند، در مدل‌ها لحاظ نشدند.

References

1. Abdul R, Kushnir O. Diabetes: a global challenge and the impact of melatonin. Prirodnichi, matematichni nauki ta osvita v meditsini. *Natural, Mathematical Sciences and Education in Medicine*. 2024; 1(1):9-21.
2. Mishra S, Tiwari P, Yadav R, Patel PS. An extensive analysis of diseases associated with diabetes. *J Pharma Insights Res*. 2024; 2(3):174-87.
3. Lyons TJ, Basu A. Biomarkers in diabetes: hemoglobin A1c, vascular and tissue markers. *Transl Res*. 2012; 159(4):303-12.
4. Liberty IA, Putri NK. Association between triglyceride-glucose index (Tyg index) and type 2 diabetes mellitus: A systematic review. *Epidemiol Biostat Public Health*. 2020; 17(1).
5. Lee DY, Lee ES, Kim JH, Park SE, Park CY, Oh KW, et al. Predictive value of triglyceride glucose index for the risk of incident diabetes: a 4-year retrospective longitudinal study. *PLoS One*. 2016; 11(9):e0163465.
6. Low S, Khoo KCJ, Irwan B, Sum CF, Subramaniam T, Lim SC, et al. The role of triglyceride glucose index in development of Type 2 diabetes mellitus. *Diabetes Res Clin Pract*. 2018; 143:43-9.
7. Şıvgın H. Predicting Cardiovascular Risk in Prediabetes Status: The Role of Atherogenic Plasma Index in Individuals with Impaired Fasting Glucose. *Osmangazi Med J*. 2023;45(5):796-804.
8. Katoch A, Singh N, Kumar P, Sharma R, Sharma R, Arya S, et al. A Machine Learning-based Approach to Diabetes Prediction. *Int J Sci Res Sci Technol*. 2023; 10(3):131-41.
9. Ahmed N, Hassan SI, Hussain MZ. Applications of Machine Learning for Diabetes Prediction. *Ann Comput Sci Inf Syst*. 2023; 38:1-6.
10. Rani KJ. Diabetes prediction using machine learning. *Int J Sci Res Comput Sci Eng Inf Technol*. 2020; 6(4):294-305.
11. Kaushik A, Sehgal A, Vora S, Palan V, Patil S. Presaging the Signs of Diabetes Using Machine Learning Algorithms. In: *2021 12th International Conference on Computing Communication and*

- Networking Technologies (ICCCNT)*; 2021: p. 1-5.
12. Haneef R, Kab S, Hrzic R, Fuentes S, Fosse-Edorh S, Cosson E, et al. Use of artificial intelligence for public health surveillance: a case study to develop a machine Learning-algorithm to estimate the incidence of diabetes mellitus in France. *Arch Public Health*. 2021; 79(1):168.
 13. Leiharer A, Muendlein A, Saely CH, Plattner T, Larcher B, Mader A, et al. 1321-P: A Machine-Learning Approach on Metabolomic Data to Predict Type 2 Diabetes Mellitus Incidence. *Diabetes*. 2024; 73(Supplement_1):1321-P.
 14. Thabit QQ, Fahad TO, Dawood AI. Detecting diabetes using machine learning algorithms. In: *2022 Iraqi International Conference on Communication and Information Technologies (IICCIT)*; 2022. p. 131-6.
 15. Dritsas E, Trigka M. Data-driven machine-learning methods for diabetes risk prediction. *Sensors (Basel)*. 2022; 22(14):5304.
 16. Liu YQ, Chang TW, Lee LC, Chen CY, Hsu PS, Tsan YT, et al. Use of Machine Learning to Predict the Incidence of Type 2 Diabetes Among Relatively Healthy Adults: A 10-Year Longitudinal Study in Taiwan. *Diagnostics (Basel)*. 2024; 15(1):72.
 17. Mohseni Sh, Pejman Sani M. Association between triglyceride-glucose related indices and bone health: A review study. *ijdl*. 2025; 25(4):299-307.