دانشگاه صنعتی شریف
دپارتمان مهندسی صنایع

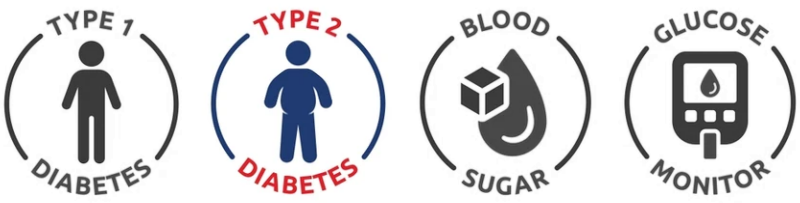
پروژه‌ی پایانی درس

مدلسازی و تصمیم‌گیری داده‌محور

استاد درس | سرکار خانم دکتر صدقی

گردآوری

محمدحسین محمودی ۴۰۱۲۱۱۹۷۴



بهمن ماه ۱۴۰۲

nature
SCIENTIFIC REPORTS

Predicting Youth Diabetes Risk Using NHANES Data and Machine Learning

Nita Vangeepuram, Bian Liu, Po-hsiang Chiu, Linhua Wang, Gaurav Pandey

Abstract

Prediabetes and diabetes mellitus (preDM/DM) have become alarmingly prevalent among youth in recent years. However, simple questionnaire-based screening tools to reliably assess diabetes risk are only available for adults, not youth. As a first step in developing such a tool, we used a large-scale dataset from the National Health and Nutritional Examination Survey (NHANES) to examine the performance of a published pediatric clinical screening guideline in identifying youth with preDM/DM based on American Diabetes Association diagnostic biomarkers. We assessed the agreement between the clinical guideline and biomarker criteria using established evaluation measures (sensitivity, specificity, positive/negative predictive value, F-measure for the positive/negative preDM/DM classes, and Kappa). We also compared the performance of the guideline to those of machine learning (ML) based preDM/DM classifiers derived from the NHANES dataset. Approximately 29% of the 2858 youth in our study population had preDM/DM based on biomarker criteria. The clinical guideline had a sensitivity of 43.1% and specificity of 67.6%, positive/negative predictive values of 35.2%/74.5%, positive/negative F-measures of 38.8%/70.9%, and Kappa of 0.1 (95%CI: 0.06–0.14). The performance of the guideline varied across demographic subgroups. Some ML-based classifiers performed comparably to or better than the screening guideline, especially in identifying preDM/DM youth ($p = 5.23 \times 10^{-5}$). We demonstrated that a recommended pediatric clinical screening guideline did not perform well in identifying preDM/DM status among youth. Additional work is needed to develop a simple yet accurate screener for youth diabetes risk, potentially by using advanced ML methods and a wider range of clinical and behavioral health data.

فهرست مطالب

فصل اول

تعریف مسئله
و گزارشی بر مجموعه‌داده‌ی جمع‌آوری شده

شرح مسئله و ضرورت پژوهش

گسترش پیش‌دیابت[1] و دیابت قندی[2] در بین جوانان، همراه با ضعف در تشخیص آن‌ها، هشداری است به یک نیاز بحرانی به ابزارهای قوی و قابل دسترس برای غربال‌گری[3]. راهنمایی‌های موجود برای غربال‌گری بالینی کودکان و نوجوانان، از جمله آن‌هایی که توسط انجمن دیابت آمریکا[4] و انجمن پزشکان کودکان آمریکا[5] تأیید شده‌اند، محدودیت‌هایی در تشخیص دقیق جوانان دارای دیابت و پیش‌دیابت بر اساس معیارهای زیست‌نشانگر[6] نشان می‌دهد. اختلافات در عملکرد راهنماها در گروه‌های زیرجمعیتی مختلف مانند سن، جنسیت و نژاد یا قومیت نگرانی را در عدم تشخیص صحیح و به موقع این بیماری‌ها افزایش می‌دهد. این مطالعه با آگاهی از این محدودیت‌ها، بر ایجاد یک ابزار غربال‌گری پیشرفته و داده‌محور با استفاده از روش‌های یادگیری ماشین[7] تأکید دارد که شامل طیف گسترده‌ای از داده‌های بهداشت بالینی و رفتاری است . هدف این پژوهش آن است که چالش‌های مطرح‌شده‌ی کنونی را برطرف نماید تا بتواند با دقت بالاتری جوانان مشکوک به این عوارض را شناسایی کند. هدف کلی در این مطالعه ایجاد یک ابزار جامع و دقیق است تا به صورت گسترده اجرا شود و به شناسایی جوانان در خطر کمک و آن‌ها را به سوی مداخلات پیشگیرانه پیشنهادی هدایت کند [۱].

پیش‌دیابت و دیابت در نوجوانان

تفاوت دیابت و پیش‌دیابت قندی

پیش‌دیابت یک شرایط سلامتی است که در آن بیمار با افزایش سطح قند خونی مواجه می‌شود که بالاتر از حد معمول است، اما هنوز به اندازه‌ی کافی بالا نیست تا به عنوان دیابت نوع دوم تشخیص داده شود. این عارضه به عنوان یک مرحله‌ی میانی بین سطوح عادی قند خون و توسعه‌ی دیابت شناخته می‌شود. افراد مبتلا به پیش‌دیابت با خطر بالاتری برای پیشرفت به دیابت نوع دوم مواجه هستند، همچنین خطر ابتلا به بیماری قلبی و سکته‌ی قلبی برای این دست از

[1] Prediabetes
[2] Diabetes Mellitus
[3] Screening Guideline
[4] American Diabetes Association (ADA)
[5] The American Academy of Pediatrics (AAP)
[6] Biomarker Guidelines
[7] Machine Learning

افراد نیز افزایش می‌یابد. تفاوت اصلی بین پیش‌دیابت و دیابت در سطوح قند خون است که در پیش‌دیابت، سطوح قند خون بیش از حد نرمال افزایش یافته‌است، اما هنوز به آستانه‌ی تشخیص دیابت نرسیده است. زمانی که سطوح قند خون به طور مداوم بالا باقی بماند، منجر به تشخیص دیابت نوع دوم می‌شود. لازم به ذکر است که تغییرات در سبک زندگی، مانند کاهش وزن، تغذیه‌ی سالم و فعالیت‌های فیزیکی منظم، معمولاً به افراد مبتلا به پیش‌دیابت کمک می‌کند تا از ظهور دیابت نوع دوم جلوگیری نمایند. نظارت منظم و نظارت پزشکی برای افراد مبتلا به پیش‌دیابت برای مدیریت بهتر سلامت آن‌ها ضروری است [۲].

اهمیت تشخیص زودهنگام دیابت در نوجوانان

تشخیص دیابت و همچنین پیش‌دیابت میان جوانان و نوجوانان از اهمیت بسیار بالایی برخوردار است که در مقاله‌ی مورد مطالعه به تفسیر به آن پرداخته شده است [۱]. به طور خلاصه می‌توان اهم این موارد را به شرح زیر خلاصه نمود:

- دیایت قندی، یک بیماری مزمن جدی با عوارض بلندمدت

دیابت قندی به عنوان یک بیماری مزمن جدی با عوارض طولانی‌مدت زیاد شناخته می‌شود. شناسایی و پرداختن به این بیماری در ابتدای مراحل پیشرفت آن می‌تواند کمک کند تا از عوارض آن جلوگیری کرد یا بیماری به تأخیر انداخته شود. از این رو برای نوجوانان و جوانان که در ابتدای دوره‌ی زندگی خود هستند، پیشگیری و یا جلوگیری از پیشرفت این بیماری دوچندان اهمیت خواهد داشت.

- قابل برگشت بودن پیش‌دیابت با تغییرات سبک زندگی

پیش‌دیابت در متون علمی به عنوان یک شرایط پیش‌گیرنده معرفی شده است که با تغییرات در سبک زندگی و کاهش وزن قابل بازگشت است. شناسایی زودهنگام این وضعیت امکان مداخله‌ی به موقع از طریق تغییرات در سبک زندگی را فراهم می‌کند و احتمالاً جلوی پیشرفت این بیماری به دیابت را می‌گیرد. از این رو ذکر می‌شود که پیش‌دیابت به عنوان یک چراغ خطر در نوجوانان می‌تواند تحت درمان و بررسی دوره‌ای قرار گیرد تا با تغییر در سبک زندگی مانع از پیشرفت آن به دیابت شد.

- شیوع بالا در میان نوجوانان و جوانان

این پژوهش طبق مطالعات پیشین برجسته می‌کند که هر دو نوع دیابت و پیش‌دیابت به طور نگران‌کننده‌ای در میان نوجوانان و جوانان شیوع پیدا کرده‌است و تعداد قابل توجهی از آن‌ها سالانه به دیابت نوع دوم تشخیص داده می‌شوند.

پرداختن به این وضعیت در ابتدای مراحل پیشرفت بیماری بسیار حیاتی است و به دلیل افزایش شیوع آن لازم است تا بیشتر به آن پرداخته شود.

- تفاوت‌ها در شیوع در گروه‌های جمعی

با توجه به آمار و ارقام به‌دست آمده تفاوت‌های بسیاری در شیوع پیش‌دیابت و دیابت در میان گروه‌های جمعی مختلف وجود دارد، از جمله نرخ‌های بالاتر در مردان، آمریکاییان سیاه‌پوست، لاتین‌تبارها و جوانان دارای اضافه‌وزن. شناسایی زودهنگام بیماری جهت پرداختن به این تفاوت‌ها و ارائه‌ی مداخلات هدفمند ضروری است تا بتوان تشخیص‌ها و بعد از آن راهکارهای شخصی‌سازی‌شده برای گروه‌های جمعی مختلف ارائه نمود.

- مشکلات در درمان دیابت در جوانان

دیابت در جوانان به عنوان مجموعه‌ای از مشکلات پیچیده‌تر در مقایسه با افراد میان‌سال توصیف شده است که این چالش‌ها با کاهش سریع‌تر عملکرد سلول‌های بتا که وظیفه‌ی آن‌ها ترشح انسولین است همراه است. شناسایی زودهنگام بیماری امکان مدیریت و پیشگیری مؤثرتر از عوارض این بیماری را فراهم می‌کند که در میان افراد با سنین پایینتر اهمیت بسیار بالایی دارد.

- تأثیر بیشتر بیماری جوانان و نوجوانان از منظر ابعاد اقتصادی

تأثیر اقتصادی بالقوه‌ی بیماری دیابت برای نوجوانان و جوانان نسبت به افراد دارای سنین بالاتر بسیار بیشتر اعلام شده است که با در نظر گرفتن تعداد بیشتر سال‌های زندگی با این بیماری و زمان موجود برای توسعه‌ی عوارض طولانی مدت آن قابل توجیه است. از این رو تشخیص هر چه سریع‌تر این بیماری در میان این زیرگروه از اهمیت بسیار بالایی برخوردار است و پیشگیری از پیشرفت آن سبب کاهش گستره‌ی اثرگذاری آن خواهد شد.

- ضعف در تشخیص میان نوجوانان و عدم آگاهی از این بیماری میان جامعه

با وجود گستره‌ای از راهنماها که توسط سازمان‌هایی مانند انجمن دیابت آمریکا ارائه شده است، پیش‌دیابت در میان جوانان اغلب با دقت کمتری نسبت به بزرگسالان تشخیص داده می‌شود و بسیاری از جوانان ممکن است مراجعه‌ی سالانه‌ی پیشنهادی توسط این سازمان‌ها را نداشته باشند و خدمات پیشگیری را دریافت نکنند که این امر منجر به عدم آگاهی از وضعیت بیمار می‌شود. در نتیجه‌ی این معضل، کسب دانش نسبت به این بیماری در این زیرگروه بسیار بیشتر

از زیرگوه‌های سنی دیگر احساس می‌شود تا بتوان علاوه بر افزایش دقت به افزایش آگاهی نسبت به این بیماری نیز کمک نمود.

- نیاز به ابزارهای اسکرینینگ

همچنین در این پژوهش به نیاز به ابزارهای غربال‌گری مؤثر تأکید می‌کند چرا که بسیاری از جوانان از پیش‌دیابت یا دیابت خود آگاه نیستند و از این رو یک ابزار ساده، غیرتهاجمی[1] و مبتنی بر پرسشنامه به عنوان یک استراتژی مؤثر اولیه برای شناسایی افراد در معرض خطر، قبل از مواجهه با آزمون‌های قطعی و برنامه‌های پیشگیری پیشنهاد می‌شود.

با توجه به موارد فوق به طور کلی می‌توان بیان نمود که تشخیص زودهنگام دیابت و پیش‌دیابت در میان نوجوانان و جوانان برای جلوگیری از عوارض، پرداخت به تفاوت‌ها و اجرای مداخلات مؤثر، از جمله تغییرات در سبک زندگی و تدابیر پیشگیری، بسیار حیاتی است و این پژوهش نیز تمرکز بر همین امر دارد.

[1] Non-invasive

روش‌های تشخیص بیماری

انجمن دیابت آمریکا

انجمن دیابت آمریکا[1] که در سال ۱۹۳۹ توسط شش پزشک با دورنمای آینده‌نگرانه تأسیس شد، به عنوان یک نیروی حیاتی در جنگ با دیابت ظاهر شده است. این سازمان با شبکه‌ای متشکل از ۵۶۵۰۰۰ داوطلب، شامل ۲۰۰۰۰ داوطلب حرفه‌ای در حوزه‌ی سلامت، مأموریت‌هایی شامل آموزش، تحقیقات و حمایت از شرایط متنوع دیابت را پیش می‌برد. فعالیت‌های این انجمن مطالعه روی انواع دیابت من‌جمله دیابت نوع ۱ و نوع ۲ تا دیابت بارداری و پیش‌دیابت را شامل می‌شود. جلسات علمی سالانه‌ی این انجمن به همراه اعضای قابل توجه آن که حدود ۲۰۰۰۰ عضو حرفه‌ای را شامل می‌شود، نشان از تعهد این انجمن به پیشبرد درمان دیابت دارد.

این سازمان در پیشبرد مأموریت‌های خود، پروژه‌های تحقیقاتی بسیاری را تأمین می‌کند که حتی از محدوده‌ی آزمایشگاهی نیز فراتر می‌رود و سبب ترویج سبک زندگی سالم می‌شود که اقلیت‌ها را برای مقابله با پیچیدگی‌های انواع دسته‌ها از بیماری دیابت حمایت می‌کند. طبق داده‌های موجود از عملکرد این انجمن، سرمایه‌گذاری‌های آن تأثیرگذار بوده و هر دلار سرمایه‌گذاری در تحقیقات دیابت منجر به ۱۲.۴۷ دلار سرمایه‌گذاری بیشتر شده است که این موضوع خود یک مهر تاییدی بر تأثیر پایدار این سازمان در شکل‌دهی به یک آینده‌ای فارغ از دیابت است [۳].

دستورالعمل زیست‌نشانگر[2]

زیست‌نشانگر یک ویژگی قابل اندازه‌گیری مولکولی، سلولی یا فیزیولوژیک است که نشانگر حضور و وضعیت یک بیماری یا شرایط خاص است. اصطلاح معیارهای زیست‌نشانگر به مجموعه شرایط یا پارامترهایی اطلاق می‌شود که باید برآورده گردد تا یک مورد خاص به عنوان یک زیست‌نشانگر معتبر برای یک بیماری یا شرایط خاص در نظر گرفته شود. به طور کلی چندین دسته از زیست‌نشانگرها وجود دارد که شامل حساسیت[3]، تشخیص[4]، نظارتی[5]، پیش‌آگاهانه[6]، پیش‌بینی‌کننده[7]،

[1] American Diabetes Association
[2] Biomarker Criteria
[3] Sensitivity
[4] Diagnosis
[5] Monitoring
[6] Proactive
[7] Predictive

پاسخی[1] و ایمنی[2] می‌شوند. یک زیست‌نشانگر می‌تواند یک ویژگی تکی یا یک پنل از ویژگی‌ها باشد؛ به عنوان مثال، در زمینه‌ی درمان سرطان، آزمون‌های زیست‌نشانگر می‌تواند تغییرات ژنتیک مرتبط با سرطان را شناسایی کنند. بعضی از آزمون‌ها به تمام ژن‌های سرطان، برخی به تمام ساختار سرطان و شماری دیگر به تعداد تغییرات ژنتیک در سرطان نگاه می‌کنند.

دستورالعمل غربال‌گری[3]

هرچند که اصطلاح دستورالعمل غربالگری در منابع مختلف به صورت صریح تعریف نشده است، در اصطلاح عام پزشکی، دستورالعمل غربالگری به مجموعه‌ای از شیوه‌ها یا روش‌های توصیه‌شده برای شناسایی افراد با خطر بالای ابتلا به یک شرایط خاص اشاره دارد. این دستورالعمل‌ها معمولاً توسط سازمان‌های بهداشتی یا انجمن‌های حرفه‌ای تدوین می‌شود تا پزشکان و بیماران را در استفاده از بهترین رویکردها برای شناسایی و مدیریت خطرهای بهداشتی هدایت کند.

در زمینه‌ی زیست‌نشانگرها، راهنمایی‌های غربالگری ممکن است مشخص کند که برای یک جمعیت یا در شرایط خاص، کدام زیست‌نشانگرها باید آزمایش شود. به عنوان مثال، یک راهنمایی غربالگری ممکن است توصیه کند که در افراد با خطر بالای ابتلا به یک نوع سرطان، بر اساس سن، تاریخچه‌ی خانوادگی یا عوامل دیگر، آزمایش‌های معینی برای زیست‌نشانگرهای خاصی به صورت روزانه انجام شود.

[1] Responsive
[2] Safety
[3] Screening Guideline

در نتیجه، معیارهای زیست‌نشانگر و دستورالعمل‌های غربالگری ابزارهای حیاتی در پزشکی مدرن هستند که به طور قابل توجهی در تشخیص بیماری‌ها، هدایت طرح‌های درمانی و نظارت بر پیشرفت بیماران نقش دارند. در پژوهشی که این مقاله بر اساس آن پیش رفته است، معیارها و دستورالعمل به صورت زیر است [۴]:

- دستورالعمل زیست‌نشانگر انجمن دیابت آمریکا توصیه‌هایی برای تشخیص دیابت ارائه می‌دهد. این توصیه‌ها به معیارهای تست ناشتای گلوکز خون[۱]، تست قند دوساعته[۲]، و تست هموگلوبین گلیکوزیله‌شده[۳] اشاره دارد. با اعتماد به این دستورالعمل، امکان تشخیص دقیق دیابت و نظارت بر سطح گلوکز در خون فراهم می‌شود.
- دستورالعمل‌های غربالگری برای جوانانی که اضافه وزن دارند و دارای یک یا چند عامل خطرناک دیگر هستند، توصیه می‌شود. این عوامل شامل تاریخچه‌ی مادری ابتلا به دیابت، تاریخچه‌ی خانوادگی دیابت نوع ۲، نژاد یا قومیت‌های متسعد به دیابت و نشانه‌های مقاومت به انسولین[۴] یا شرایط مرتبط با مقاومت به انسولین می‌شوند. این دستورالعمل‌ها باید پس از شروع دوران نوجوانی یا پس از ده سالگی انجام شوند که این اقدامات غربالگری امکان تشخیص زودهنگام دیابت را افزایش می‌دهد.

[1] Fasting Plasma Glucose Test
[2] Two Hour Plasma Glucose Test
[3] Hemoglobin A1c Test
[4] Insulin Resistance

معرفی مجموعه‌داده

مطالعه‌ی ملی سلامت و تغذیه[1]

مطالعه‌ی ملی بررسی سلامت و تغذیه یک برنامه‌ی جامع است که در اوایل دهه‌ی ۱۹۶۰ شروع به کار کرده است و تحت نظر مرکز ملی آمار سلامت[2] که بخشی از مراکز کنترل و پیشگیری از بیماری‌ها[3] است، اداره می‌شود. این برنامه هدف اساسی ارزیابی وضعیت سلامت و تغذیه‌ی همزمان بزرگسالان و کودکان در ایالات متحده را عهده‌دار است. همچنین این مطالعات شامل ترکیب مصاحبه‌ها، معاینات مختلف و آزمایش‌های متمایز است که به صورت مداوم تا کنون ادامه داشته است و برخوردار از یک نمونه‌ی ملی شامل سالانه حدود ۵۰۰۰ نفر از افراد مقیم در شهرستان‌های مختلف در سراسر کشور آمریکا است. این مطالعه هر ساله داده‌هایی شامل اطلاعات بیماران مورد مطالعه در سال گذشته را منتشر می‌کند که در این پژوهش هم منبع داده‌ها وب‌سایت این انجمن است.

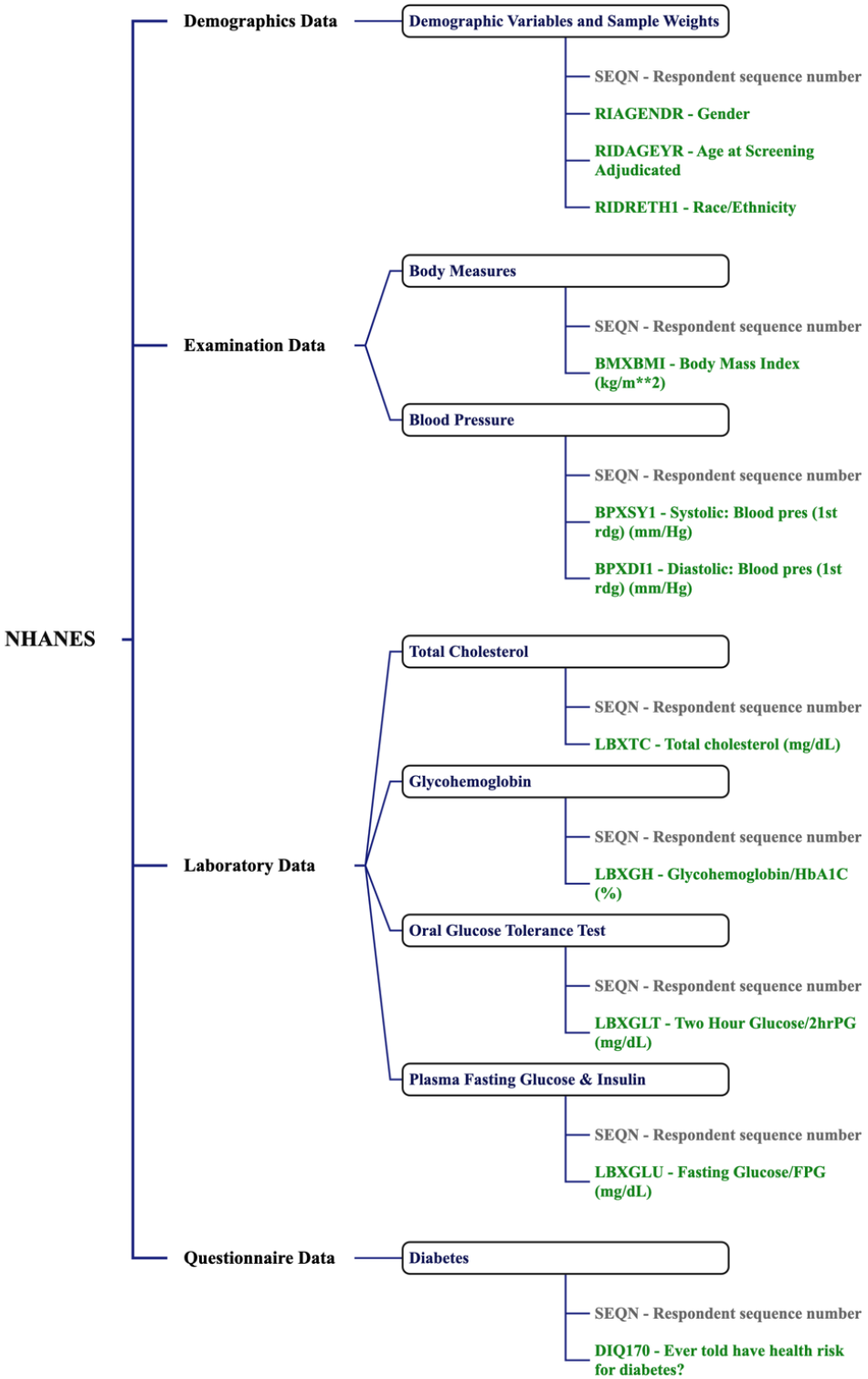
مجموعه‌داده‌های مورد استفاده

همانطور که پیشتر ذکر شد در وب‌سایت مطالعه‌ی ملی سلامت و تغذیه داده‌های بسیاری موجود است که به بخش‌های مختلفی تقسیم‌بندی می‌شود. به طور خاص در این پژوهش متغیرهای مجموعه‌داده‌ی نهایی طبق چارت ۱ استخراج شده است و متغیرهای ورودی از هر مجموعه‌داده به مدل‌های مورد استفاده به شرح زیر است:

مجموعه‌داده‌ی جمعیت‌شناسی[4]

از مجموعه‌داده‌ی جمعیت شناختی که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها)، جنسیت، سن و نژاد استخراج شده است.

[1] National Health and Nutrition Examination Survey (NHANES)
[2]. National Center for Health Statistics (NCHS)
[3]. Centers for Disease Control and Prevention (CDC)
[4] Demographic Data



نمودار ۱- سلسله‌مراتب متغیرهای استخراج‌شده از مجموعه‌داده‌ها در بخش‌های مختلف

مجموعه‌داده‌ی شاخص‌های فیزیکی[1]

از مجموعه‌داده‌ی شاخص‌های فیزیکی که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و شاخص توده‌ی بدنی[2] استخراج شده است. لازم به ذکر است که در مطالعات نهایی و همچنین برچسب‌گذاری بیماران از آمار صدک شاخص توده‌ی بدنی استفاده شده است که در مرحله‌ی جمع‌آوری مجموعه‌داده این متغیر بر حسب شاخص توده‌ی بدنی بیماران ایجاد شده است.

مجموعه‌داده‌ی فشار خون[3]

از مجموعه‌داده‌ی فشار خون که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و فشار خون سیستولیک[4] و فشار خون دیاستولیک[5] استخراج شده است. در این مطالعه از این متغیرها که به ترتیب نشان دهنده‌ی فشار خون در برابر دیواره های شریانی هنگام انقباض (تپش) قلب و فشار خون زمانی که قلب بین ضربان ها استراحت می کند هستند، برای برچسب‌گذاری بیماران و تشخیص وضعیت فشار خون آن‌ها استفاده شده است.

مجموعه‌داده‌ی کلسترول کل[6]

از مجموعه‌داده‌ی کلسترول کل که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و کلسترول خون استخراج شده است. از این متغیر در تشخیص ابتلا یا عدم ابتلا به فشار خون در میان بیمارن استفاده شده است.

مجموعه‌داده‌ی گلیکوهموگلوبین[7]

از مجموعه‌داده‌ی گلیکوهموگلوبین که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و هموگلوبین گلیکوزیله[8] استخراج شده است که به طور مستقیم یکی از معیارهای

[1] Body Measures Data
[2]. Body Mass Index (BMI)
[3] Blood Pressure Data
[4]. Systolic Blood Pressure
[5]. Diastolic Blood Pressure
[6] Total Cholestrol Data
[7] Glycohemoglobin Data
[8]. HbA1c (Hemoglobin A1c)

زیست‌نشانگر جهت تشخیص وضعیت ابتلا به دیابت در میان بیماران است. این متغیر در میان بیماران نشان‌دهنده‌ی میانگین سطح گلوکز (قند) خون بیمار در ۲ یا ۳ ماه گذشته است.

مجموعه‌داده‌ی تست تحمل گلوکز خوراکی[1]

از مجموعه‌داده‌ی تست تحمل گلوکز خوراکی که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و قند خون دو ساعته[2] استخراج شده است. این متغیر همانند متغیر هموگلوبین گلیکوزیله یکی از شاخص‌های زیست‌نشانگر جهت تشخیص بیماری دیابت است.

مجموعه‌داده‌ی گلوکز ناشتا و انسولین پلاسما[3]

از مجموعه‌داده‌ی گلوکز ناشتا و انسولین پلاسما که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و آزمایش قند خون ناشتا[4] که همانند دو متغیر قبل به طور مستقیم در برچسب‌گذاری بیماران به کار گرفته‌شده، استخراج شده است.

مجموعه‌داده‌ی دیابت[5]

از مجموعه‌داده‌ی دیابت که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و سابقه‌ی شخص و یا خانوادگی بیماری دیابت استخراج شده است که مشخص می‌کند آیا بیمار یا یکی از افراد خانواده‌ی نزدیک وی مشکوک به ابتلا به این بیماری بوده‌اند یا خیر.

[1] Oral Glucose Tolerance Test Data
[2] 2hrPG (Two Hour Plasma Glucose)
[3] Plasma Fasting Glucose and Insulin Data
[4]. FPG (Fasting Plasma Glucose)
[5] Diabetes Data

فصل دوم

جمع‌آوری داده

فرآیند جمع‌آوری مجموعه‌داده‌ی نهایی

با توجه به آن که شروع هر پروژه‌ی مبتنی بر داده با استخراج دانش موجود و سپس استخراج داده‌ی مرتبط با آن همراه است، در این پژوهش هم که از این قاعده مستثنی نیست این فرآیند تکرار شده است و بعد از تعیین متغیرهای مورد نیاز برای طراحی پژوهش به استخراج مجموعه‌داده‌های مرتبط با حوزه پرداخته شده است. این فرآیند که نیازمند تسلط بر دانش کلی حوزه است در گام‌های زیر انجام شده است:

۱. استخراج مجموعه‌داده‌های شامل متغیرهای مورد نیاز برای تحلیل و مدلسازی از پایگاه داده‌ی انهنس
۲. ادغام مجموعه‌داده‌های هر سال از ۲۰۰۵ تا ۲۰۱۶ با یکدیگر
۳. حذف بیماران خارج از رده‌ی سنی ۱۲ تا ۱۹ سال از مجموعه‌داده‌ی ادغام‌شده
۴. تشکیل متغیرهای مورد نیاز برای تحلیل‌ها و مقایسات بیشتر مثل متغیر فشار خون
۵. برچسب‌گذاری بیماران بر اساس شاخص‌های زیست‌نشانگر انجمن دیابت آمریکا

با توجه به آن که گام‌های فوق خارج از محدوده‌ی فعالیت‌های پروژه در این درس است، تمامی گام‌های استخراج و ادغام مجموعه‌داده‌ها مبتنی بر زبان برنامه‌نویسی پایتون انجام شده است که فایل‌های مرتبط با آن در انبار گیت‌هاب[1] این پروژه موجود است.

استخراج و ادغام مجموعه‌داده‌ها

همانطور که پیشتر ذکر شد، مجموعه‌داده‌ی مورد استفاده به صورت آماده در اختیار نبوده است؛ بنابراین برای طراحی پژوهش جمع‌آوری دستی و سپس ادغام آن، امری ضروری بود. در نتیجه‌ی این امر و پس از مطالعات بسیار و بررسی مقاله‌ی مورد مطالعه، اطلاعات لازم و متغیرهای دقیق مسئله مشخص شد و مجموعه‌داده‌ی نهایی به وسیله‌ی داده‌های استخراج‌شده ساخته شد. این داده‌ها همانطور که در بخش قبل ذکر شد از چهار بخش داده‌های دموگرافیک، معاینه‌ای، آزمایشگاهی و پرسشنامه‌ای وب‌سایت انهنس استخراج شد و به علت نوع فایل این مجموعه‌داده‌ها (.XPT) عملیات استخراج و ترکیب مجموعه‌داده‌ها بر پایه‌ی زبان پایتون انجام شده است.

[1] https://github.com/mamishere/Predicting-Youth-Diabetes-Risk-Using-NHANES-Data-and-Machine-Learning

برچسب‌گذاری بیماران در مجموعه‌داده

با توجه به آنچه در مقاله‌ی پایه ارائه شده است، برچسب‌گذاری غربال‌گری بر اساس شاخص‌های زیر انجام می‌شود [۱].

ADA/AAP preDM/DM risk for children (at-risk if overweight plus one or more additional risk factors)	NHANES variables used (at-risk if overweight plus one or more additional risk factors)
Overweight (BMI > 85th percentile for age and sex, weight for height > 85th percentile, or weight > 120% of ideal for height) **A**[a]	BMI ≥ 85th percentile[b]
Additional risk factors:	Additional risk factors:
Maternal history of gestational diabetes during the child's gestation **A**[a]	Not available:
Family history of type 2 diabetes in first- or second-degree relative **A**[a]	Ever been told by a doctor or other health professional that you have health conditions or a medical or family history that increases your risk for diabetes?
Race/Ethnicity (Native American, African American, Latino, Asian American, Pacific Islander) **A**[a]	Non-White race/ethnicity (non-Hispanic Black, Hispanic, other)
Signs of insulin resistance or conditions associated with insulin resistance (hypertension, dyslipidemia, acanthosis nigricans, polycystic ovary syndrome, or small-for-gestational-age birth weight). **B**[a]	Hypertension[c]: Blood pressure ≥ 90th percentile or ≥ 120/80 mm Hg for children ≥ 13 years; Dyslipidemia[d]: total cholesterol ≥ 170 mg/dL

تصویر ۱- شاخص‌های غربال‌گری تشخیص پیش‌دیابت و دیابت

بر اساس تعاریفی که پیشتر بر شاخص‌های زیست نشانگر ارائه شده‌است، قواعد فوق در تشخیص غربال‌گری بیماران کاربرد دارد که در مطالعه به قیاس آن با مدل‌های یادگیری ماشین پرداخته شده است و در این فاز از تمرین از این قیاس چشم‌پوشی شده است و تمرکز این بخش از مطالعه بر متغیرهای زیست‌نشانگر که ورودی مدل‌های یادگیری ماشین می‌باشند، است.

در این مطالعه که وضعیت پیش‌دیابت یا دیابت بر اساس معیارهای زیست‌نشانگر انجمن دیابت آمریکا تعیین شده است سطوح بالای هر یک از سه زیست‌نشانگر به صورت زیر تعریف شده است:

- گلوکز ناشتای بزرگتر یا مساوی ۱۰۰ میلی‌گرم بر دسی‌لیتر
- گلوکز دو ساعته‌ی پس از وعده‌ی غذایی بزرگتر یا مساوی ۱۴۰ میلی‌گرم بر دسی‌لیتر
- هموگلوبین گلیکوزیله‌ی بزرگتر یا مساوی ۵.۷ درصد

از آنجا که تعداد کمی از جوانان بر اساس معیارهای تشخیصی زیست‌نشانگر مبتلا به دیابت نوع دوم بوده‌اند، گردآورندگان این مطالعه جوانان دارای پیش‌دیابت و دیابت را در یک دسته ترکیب کردند. این معیارهای تشخیصی بر اساس یک مطالعه‌ی معتبر در این حوزه است که این مطالعه نیز به آن ارجاع داده‌است و در جدول دوم می‌توان روش‌های ارزیابی بیماران را با جزئیات بیشتری مشاهده کرد [۴].

Prediabetes
A1C 5.7% to <6.5% (39 to <48 mmol/mol). The test should be performed in a laboratory using a method that is NGSP certified and standardized to the DCCT assay. IFG: fasting glucose ≥100 but <126 mg/dL (≥5.6 but <7.0 mmol/L). IGT: 2-h plasma glucose ≥140 but <200 mg/dL (≥7.8 but <11.1 mmol/L) during an OGTT. The test should be performed as described by the World Health Organization, using a glucose load containing the equivalent of 1.75 mg/kg (max 75 g) anhydrous glucose dissolved in water.*
Diabetes
A1C ≥6.5% (≥48 mmol/mol). The test should be performed in a laboratory using a method that is NGSP certified and standardized to the DCCT assay.* OR FPG ≥126 mg/dL (7.0 mmol/L). Fasting is defined as no caloric intake for at least 8 h.* OR 2-h plasma glucose ≥200 mg/dL (11.1 mmol/L) during an OGTT. The test should be performed as described by the World Health Organization, using a glucose load containing the equivalent of 1.75 mg/kg (max 75 g) anhydrous glucose dissolved in water* OR In a patient with classic symptoms of hyperglycemia or hyperglycemic crisis, a random plasma glucose >200 mg/dL (11.1 mmol/L).
FPG, fasting plasma glucose; IFG, impaired fasting glucose; IGT, impaired glucose tolerance; max, maximum. *In the absence of unequivocal hyperglycemia, result should be confirmed by repeat testing.

تصویر ۲- شاخص‌های زیست‌نشانگر تشخیص پیش‌دیابت و دیابت

در این فاز از پروژه تمرکز بر روی بخش یادگیری ماشین مقاله بوده است و تلاش شده تا بر اساس شاخص‌های زیست‌نشانگر تشخیص پیش‌دیابت و یا دیابت به پیش‌بینی وضعیت بیماران پرداخته شود. مقایسه‌ی فرآیند غربال‌گری با نتیجه‌ی پیش‌بینی ماشینی در فاز بعدی قابل ارائه است و در این گام از آن گذر شده است.

فصل سوم

تحلیلی کاوش‌گرانه بر مجموعه‌داده

شرحی بر مجموعه‌داده

مجموعه‌داده‌ی مورد استفاده شامل ۲۷۳۰ رکورد و ۱۶ ویژگی است که از این ۱۶ ویژگی ۱۴ عدد آن‌ها (تمام ویژگی‌ها به جز شناسه‌ی بیماران و برچسب وضعیت غربالگری آنها) در طول تشریح مسئله مورد استفاده قرار می‌گیرد. در ادامه می‌توان گزارشی از ویژگی‌های مجموعه‌داده‌ی مورد تحلیل را مشاهده نمود.

متغیرهای گسسته

در ابتدای بررسی مجموعه‌داده متغیرهای گسسته و پیوسته از یگدیگر جدا شدند که می‌توان در جدول زیر به طور خلاصه به درکی از ویژگی‌های متغیرهای گسسته رسید.

جدول ۱- نسبت‌های پراکندگی بیماران برای ویژگی‌های مختلف

متغیر	مقادیر خاص	درصد نسبت به کل
جنسیت	مرد	۵۳
	زن	۴۷
نژاد	سفید غیر لاتین	۲۸
	سیاه غیر لاتین	۲۶
	لاتین	۳۷
	نژادهای دیگر	۹
ریسک بیماری دیابت	دارای ریسک	۱۱
	فاقد ریسک	۸۹
فشار خون بالا	دارای فشار خون بالا	۹۹
	دارای فشار خون نرمال	۱
برچسب شاخص زیست‌نشانگر دیابت	مبتلا به پیش‌دیابت یا دیابت	۷۱
	عدم ابتلا به پیش‌دیابت یا دیابت	۲۹

این متغیرها که از جنس کیفی[1] محسوب می‌شوند که در مدل هم به همین صورت مورد استفاده قرار گرفته‌اند، دارای ویژگی‌های به شرح زیر هستند:

جنسیت و نژاد

این متغیرها که از مجموعه‌داده‌ی دموگرافیک استخراج شده‌اند شامل مرد و زن در جنسیت و سفید غیرلاتین، سیاه غیر لاتین، لاتین و غیره در نژاد هستند و در جدول۱ می‌توان نسبت آن‌ها را به طور کلی در مجموعه‌داده‌ی نهایی مشاهده نمود.

ریسک بیماری دیابت

این ویژگی که از مجموعه‌داده‌ی پرسشنامه استخراج شده است، در صورتی مقدار مثبت را به خود اختیار می‌کند که بیمار و یا خانواده‌ی درجه‌ی یک وی تا کنون پس از معاینه توسط پزشک محتمل به ابتلا به دیابت در آینده خطاب شده باشد که در جدول۱ می‌توان نسبت آن‌ها را به طور کلی در مجموعه‌داده‌ی نهایی مشاهده نمود.

فشار خون بالا

این ویژگی که هنگام جمع‌اوری مجموعه‌داده ایجاد شده است، زمانی مقدار یک را اختیار می‌کند که بیمار فشار خون بزرگتر مساوی ۱۲۰/۸۰ میلی‌متر جیوه یا کلسترول کل خون بالای ۱۷۰ میلی‌گرم بر دسی‌لیتر داشته باشد که در جدول۱ می‌توان نسبت آن‌ها را به طور کلی در مجموعه‌داده‌ی نهایی مشاهده نمود.

برچسب شاخص زیست‌نشانگر دیابت

و در نهایت برچسب مورد استفاده در مدل‌های یادگیری ماشین طبق آنچه در مقاله ذکر شده هنگام جمع‌آوری مجموعه‌داده محاسبه شده است و برای مثبت‌شدن آن تنها یکی از سه شرط ارائه‌شده در ۱۹ این گزارش کفایت می‌کند.

[1] Categorical Features

متغیرهای پیوسته

همانطور که پیش‌تر ذکر شد در ابتدای بررسی مجموعه‌داده متغیرهای گسسته و پیوسته از یگدیگر جدا شدند که می‌توان در جدول زیر به طور خلاصه به بینشی از ویژگی‌های متغیرهای پیوسته رسید.

جدول ۲- آماره‌های متغیرهای پیوسته

ردیف	متغیر	کمینه	چارک اول	میانگین	میانه	چارک سوم	بیشینه
۱	سن	۱۲	۱۴	۱۵.۴۹	۱۶	۱۷	۱۹
۲	شاخص توده‌ی بدنی	۱۳.۳	۱۹.۹	۲۴.۲۷	۲۲.۶۹	۲۷.۲	۶۸.۶
۳	فشار خون سیستولیک	۷۶	۱۰۴	۱۱۰.۶	۱۱۰	۱۱۸	۱۵۴
۴	فشار خون دیاستولیک	۱۲	۵۲	۵۹.۷۱	۶۰	۶۸	۹۶
۵	گلوکز خون ناشتا	۶۱	۸۹	۹۴.۲۷	۹۴	۹۹	۲۵۴
۶	گلوکز خون دو ساعته	۳۱	۳۲	۹۷.۸۲	۹۶	۱۱۰	۲۲۲
۷	هموگلوبین گلیکوزیله	۴	۵	۵.۲۱	۵.۲	۵.۴	۹.۵
۸	کلسترول کل خون	۶۶	۱۳۶	۱۵۶.۷۴	۱۵۴	۱۷۴	۳۲۰

این متغیرها که از جنس عددی[1] محسوب می‌شوند که در مدل هم به همین صورت مورد استفاده قرار گرفته‌اند، دارای ویژگی‌های به شرح زیر هستند:

سن و شاخص توده‌ی بدنی

این دو متغیر از دو مجموعه‌داده‌ی دموگرافیک و شاخص‌های فیزیکی استخراج شده است و همانطور که در در جدول۲ قابل مشاهده است افراد مورد مطالعه در این پژوهش دارای کمینه، بیشیته و میانه‌ی شاخص توده‌ی بدنی به ترتیب ۱۳.۳، ۶۸.۶ و ۲۲.۶۹ هستند. با توجه به مقدار چارک سوم که برابر با ۲۷.۲ است، توزیع این متغیر چوله به راست خواهد بود.

[1] Numeric Features

فشار خون سیستولیک و دیاستولیک

این دو متغیر از مجموعه‌داده‌ی فشار خون استخراج شده است که ویژگی‌های آن به طور مشخص در جدول۲ قابل مشاهده است. در ادامه و فصل بعد بیان می‌شود که به علت مقدار نامتعارف فشار خون دیاستولیک ۳۰ بیمار، رکوردهای آن‌ها از مجموعه‌داده‌ی نهایی حذف شد.

گلوکز خون ناشتا، دوساعته و هموگلوبین گلیکوزه

این متغیرها از مجموعه‌داده‌ی شاخص‌های انسولین و گلوکز استخراج شده است که ویژگی‌های آن به طور مشخص در جدول۲ قابل مشاهده است و از آن‌جا که این مفاهیم خارج از حیطه‌ی دانش گردآوری است، صحت و دقت آن‌ها با خبره به بحث گذاشته شد و پیش از مدل‌سازی تایید ایشان اخذ شد.

کلسترول کل خون

این متغیر نیز از مجموعه‌داده‌ی کسترول خون استخراج شده است که ویژگی‌های آن به طور مشخص در جدول۲ قابل مشاهده است و صحت مقادیر آن مورد تایید خبره قرار گرفت.

گزارشی از تحلیل روابط بین متغیرهای تاثیرگذار

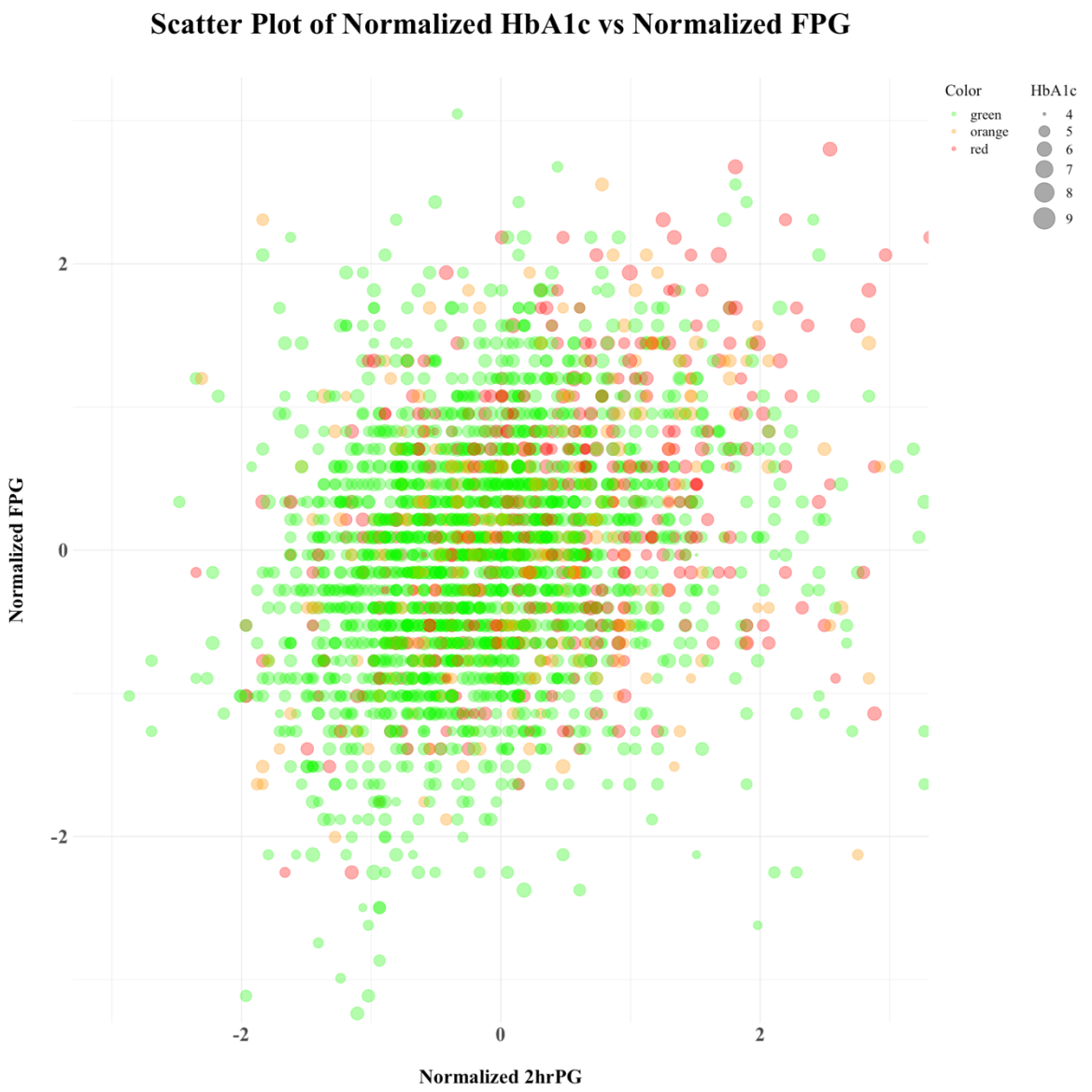
هدف مطالعه‌ی نیتا ونگیپورام[1] و همکاران در مطالعه‌ی مذکور صرفاً تلاش بر طراحی یک مدل پیش‌بینی ابتلا به پیش‌دیابت و یا دیابت میان نوجوانان ۱۲ تا ۱۹ سال است و متغیرهای ورودی به مدل‌های به کار گرفته شده در این مطالعه متغیرهای آشنا میان متخصصان این حوزه است و ارتباط بین این متغیرها موردی نیست که در دانش این حوزه امری جدید باشد. ضمن این امر در مقاله هم به ارتباط این متغیرها با یکدیگر پرداخته نشده است و همچنین هیچ نموداری جهت شناخت جامعه طراحی نشده است.

علیرغم موارد فوق، بعد از همفکری با دکتر امید کهندل گرگری[2] و دریافت نظرهای خبره جهت تصمیم بر راهکارهای ارائه‌ی گزارشی تصویری از مجموعه‌داده، عزم بر آن شد تا تست‌های گلوکز بیماران و همچنین میزان کلسترول خون آن‌ها مورد بررسی قرار گیرد تا بتوان نگاهی دقیق‌تر نسبت به نمونه‌ی مورد آزمون داشت.

[1] Nita Vangeepuram
[2] Omid Kohandel Gargari (Google Scholar)

مقایسه‌ای بر آزمون‌های مختلف گلوکز خون در بیماران

همانطور که مبرهن است گلوکز خون یکی از مهم‌ترین پارامترهای موثر در ابتلا به بیماری دیابت است و با توجه به نظر خبره تلاش شد تا در این بخش به وسیله‌ی نموداری از رکوردها رابطه‌ی بین گلوگز ناشتا و دوساعته را در کنار شاخص توده‌ی بدنی و هموگلوبین گلیکوزه مورد بررسی قرار داد.



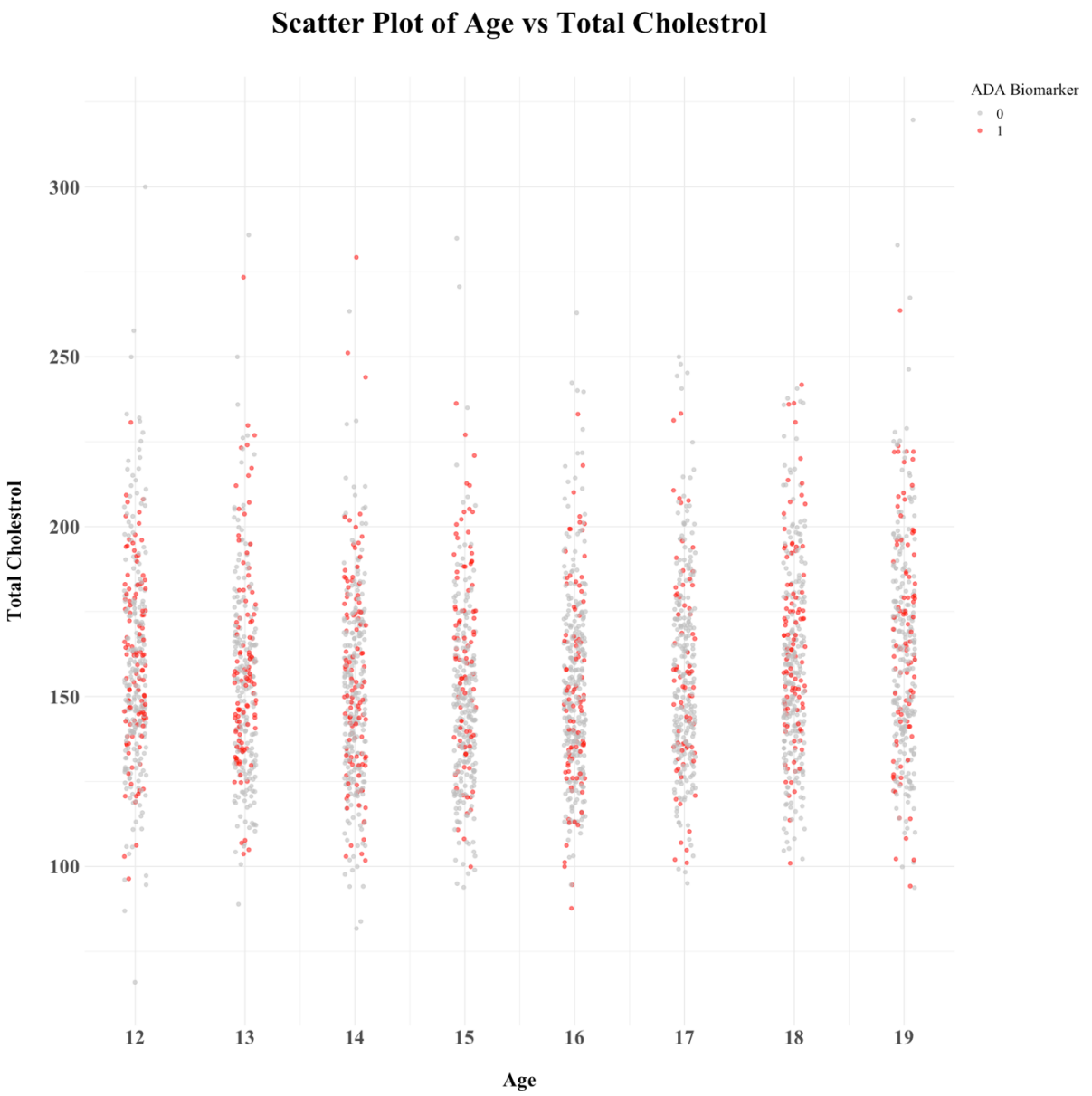
نمودار ۱- مقایسه‌ی انواع گلوکز خون در بیماران

به وضوح روشن است که با حرکت به سمت افزایش هر یک از شاخص‌های گلوکز قند خون، هموگلوبین گلیکوزه و همچنین شاخص توده‌ی بدنی افزایش می‌یابد که این امر گواهی است بر ادعای آزمون‌های غربال‌گری آمریکا بر اهمیت

شاخص توده‌ی بدنی و همچنین آزمون زیست‌نشانگر استفاده شده در این مجموعه‌داده که بر اهمیت گلوکز در انواع مختلف تاکید داشت.

رابطه‌ی بین سن و کلسترول خون

مورد دیگری که توسط خبره مورد توجه قرار گرفت تاثیر سن بر میزان کلسترول خون بود که میتوان در این نمودار به طور حدودی این ادعا را رد نمود و بر خلاف تفکر عمومی این میزان کلسترول در سنین بالاتر اختلاف خود را نشان می‌دهد.



نمودار ۲- کلسترول کل خون و سن بیمارن

همچنین تاثیر بسزایی از کلسترول خون روی شاخص زیست‌نشانگر پیش‌دیابت و دیابت نمی‌توان مشاهده کرد و این مجددا تاییدی است بر رابطه‌ی محاسبه‌ی وضعیت بیمار مبتلا به دیابت توسط شاخص‌های زیست‌نشانگر ارائه‌شده توسط انجمن دیابت آمریکا.

فصل چهارم

گزارشی بر فرآیند مدلسازی و نتایج آن

آماده‌سازی مجموعه‌داده

با توجه به آن‌که مجموعه‌داده‌ی مورد استفاده فاقد هر گونه مقدار اشتباه و یا گم‌شده بود، در فرآیند پاک‌سازی داده‌ها صرفاً به بررسی ساختمان‌داده‌ی موجود و صحت مقادیر آن پرداخته شد که با توجه به نظر خبره تصمیم بر حذف رکوردهای دارای فشار خون دیساتولیک کمتر از ۴ گرفته شد. به جز این مورد که شامل ۳۰ رکورد از ۲۷۳۰ بیمار موجود بود، هیچ نیازی به تغییر در مجموعه‌داده احساس نشد.

ضمن این امر لازم به ذکر است که در مجموعه‌داده‌های استخراج‌شده‌ی ابتدایی شاخص توده‌ی بدنی به صورت استاندارد ارائه نشده بود که در هنگام جمع‌آوری داده‌ها به اصلاح این امر جهت همگامی فرآیند کلی پروژه با آن چه در مقاله انجام شده است، پرداخته شد.

مورد دیگری که مورد توجه است برچسب‌گذاری داده‌ها می‌باشد که این عملیات هم در فاز جمع‌آوری مجموعه داده به انجام رسید و می‌توان ادعا نمود که در این فاز از پروژه بعد از جمع‌آوری و برچسب‌گذاری داده‌ها با یک داده‌ی تمیز و آماده روبه‌رو می‌شویم.

شاخص‌های ارزیابی عملکرد مدل‌های یادگیری

همانند تمامی پروژه‌ها و مطالعات مبتنی بر یادگیری ماشین، در این مطالعه نیز جهت بررسی عملکرد مدل‌های یادگیری از شاخص‌های ارزیابی مختص مدل‌های دسته‌بندی[1] استفاده شده است. شاخص‌های استفاده‌شده در این مطالعه طیف گسترده‌ای از موارد را پوشش می‌دهد که در این فاز تنها به مواردی که در کلاس درس مورد بحث قرار گرفته پرداخته می‌شود و در فاز بعد عملکرد مدل‌های خارج از محدوده‌ی تدریس و همچنین شاخص‌های ارزیابی دیگر بررسی می‌شود. در ادامه تعریف شاخص‌های مذکور و همچنین روابط محاسبه‌ی آن‌ها ارائه می‌شود.

صحت[2]

شاخص صحت یک معیار اساسی استفاده‌شده در ارزیابی مدل‌های طبقه‌بندی می‌باشد که صحت کلی پیش‌بینی‌ها را با مقایسه‌ی تعداد نمونه‌های صحیح پیش‌بینی‌شده و تعداد کل نمونه‌ها اندازه‌گیری می‌کند.

[1] Classification
[2] Accuracy

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

مقدار بالاتر این شاخص نشان‌دهنده‌ی عملکرد بهتر مدل است اما با این حال، صحت محدودیت‌های خاص خود را دارد، به ویژه در شرایطی که توزیع کلاس‌ها نامتوازن باشد. در چنین مواردی، یک مدل ممکن است با پیش‌بینی کلاس اکثریت، صحت بالایی را حاصل کند در حالی که از کلاس اقلیت غافل می‌شود [۵].

حساسیت یا فراخوانی[1]

حساسیت یا فراخوانی یک معیار مهم در ارزیابی مدل‌های طبقه‌بندی است که نشان‌دهنده‌ی توانایی مدل در شناسایی صحیح نمونه‌های مثبت به ازای تمام نمونه‌های مثبت واقعی می‌باشد.

$$Sensitivity\ or\ Recall = \frac{TP}{TP + FN}$$

این معیار زمانی اهمیت خود را نشان می‌دهد که هدف آن باشد مدل بیشترین تعداد مثبت را شناسایی کند. شاخص حساسیت به ویژه در مواقعی که تشخیص درست نمونه‌های مثبت حیاتی است، مانند زمینه‌های پزشکی یا تشخیص بیماری‌ها از اهمیت بالایی برخوردار است [۵].

خاصیت[2]

خصوصیت یک معیار حیاتی در ارزیابی مدل‌های طبقه‌بندی است، به ویژه در مواقعی که تشخیص صحیح نمونه‌های منفی از اهمیت زیادی برخوردار است. این شاخص توانایی یک مدل را در تشخیص درست منفی‌ها از بین تمام نمونه‌های منفی واقعی اندازه‌گیری می‌کند.

$$Specifity = \frac{TN}{TN + FP}$$

[1] Sensitivity or Recall
[2] Specifity

این شاخص به ویژه در شرایطی کاربردی است که پیامدهای اشتباه مثبت (شناسایی نادرست یک نمونه‌ی منفی به عنوان مثبت) زیاد باشد و شناسایی دقیق نمونه‌های واقعی منفی امری حیاتی باشد. به عنوان مثال، در تشخیص پزشکی یا برنامه‌های امنیتی که یک اشتباه مثبت ممکن است منجر به درمان‌های ناپایدار یا هشدارهای غیرضروری شود، خصوصیت به عنوان یک معیار حیاتی در ارزیابی مدل برجسته می‌شود [۵].

دقت[1]

شاخص دقت یکی از معیارهای اصلی در ارزیابی مدل‌های طبقه‌بندی است که تعداد نمونه‌های مثبتی را که به درستی تشخیص داده شده‌اند، نسبت به کل تعداد نمونه‌های مثبت تعیین می‌کند. دقت به ویژه در شرایطی کاربرد دارد که هزینه یا پیامد اشتباه مثبت (تشخیص یک نمونه منفی به عنوان مثبت) زیاد باشد.

$$Precision = \frac{TP}{TP\ +\ FP}$$

در واقع، دقت نشان‌دهنده‌ی این است که از نمونه‌هایی که مدل به عنوان مثبت تشخیص داده است، چه مقداری از آنها به درستی مثبت هستند. مقدار بالاتر دقت به معنای دقت بیشتر مدل در تشخیص نمونه‌های مثبت و اجتناب از اشتباه مثبت است [۵].

اف‌یک[2]

شاخص اف‌یک معیاری مهم در ارزیابی مدل‌های طبقه‌بندی است که دقت و حساسیت را به یک مقدار واحد ترکیب می‌کند. این معیار به ویژه در شرایطی که توازن بین دقت و حساسیت مهم است، بسیار مفید می‌باشد.

$$F1\ Score = 2\ \frac{Precision\ *\ Recall}{Precision\ +\ Recall}$$

مقدار بالاتر این شاخص نشان‌دهنده‌ی عملکرد بهتر مدل در ترکیب صحیح بین دقت و حساسیت است. این معیار به ویژه در زمینه‌هایی مانند پزشکی یا امنیت که هر دو نوع اشتباه ممکن است پیامدهای جدی داشته باشند، به کار می‌رود [۵].

[1] Precision
[2] F1 Score

مدلسازی مسئله

پیش از مدل‌سازی، همانند هر مسئله‌ی مبتنی بر یادگیری ماشین دیگری، مجموعه‌داده با نسبت ۸۰ به ۲۰ به دسته‌های یادگیری و آزمون تقسیم شدند و سپس با استفاده از مدل‌های تدریس‌شده در کلاس درس مورد ارزیابی قرار گرفتند که گزارش عملکرد آن‌ها در ادامه‌ی این گزارش ارائه شده است.

لازم به ذکر است که نتایج مدل‌های ارائه شده، خروجی پیش از بهبود الگوریتم‌هاست که امر بهبود و بهینه‌سازی الگوریتم در فاز بعد از این پژوهش مورد توجه و بررسی قرار خواهد گرفت.

مدل رگرسیون لجستیک[1]

اولین و شاید ساده‌ترین مدلی که برای دسته‌بندی یک مجموعه‌داده به ذهن می‌رسد رگرسیون لجستیک است که در کنار سرعت عمل بالا توانایی متوسطی از خود نشان می‌دهد. در جدول۳ می‌توان نتایج مدل‌سازی را مورد بررسی قرار داد.

جدول ۳- نتایج یادگیری و آزمون مدل یادگیری رگرسیون لجستیک بر حسب درصد

	نتایج یادگیری (بر حسب درصد)	**نتایج آزمون** (بر حسب درصد)
Accuracy	۵۵.۴	۵۵.۷
Sensitivity	۵۴.۷	۴۹.۷
Specifity	۵۵.۶	۵۸.۲
Precision	۳۳.۵	۳۲.۷
F1 Score	۴۱.۵	۳۹.۴

مدل رگرسیون لجستیک برای پیش‌بینی دیابت و پیش‌دیابت یک صحت حدود ۵۵.۴ درصدی در مجموعه‌ی یادگیری و یک دقت حدود ۵۵.۷ درصدی در مجموعه‌ی آزمون ارائه داده است. به طور قابل توجهی، شاخص حساسیت که

[1] Logistic Regression

نشان‌دهنده‌ی توانایی شناسایی صحیح افراد مبتلا به دیابت است، به ترتیب ۵۴.۷ درصد در مجموعه‌ی آموزش و ۴۹.۷ درصد در مجموعه‌ی آزمون بود و این امر نشان می‌دهد که مدل رگرسیون لجستیک توانایی متوسطی در تشخیص موارد مثبت واقعی دارد. شاخص صحت که نشان‌دهنده‌ی توانایی شناسایی صحیح افراد غیر دیابتی است، به ترتیب نتیجه‌ی ۵۵.۶ درصدی در آموزش مدل و ۵۸.۲ درصدی در آزمون مدل از خود ارائه کرد که این مورد نیز نشان می‌دهد عملکرد متوسط مشابهی در تشخیص داده‌های منفی واقعی وجود دارد. سایر شاخص‌ها نیز در جدول قابل مشاهده هست.

تمامی این موارد از ماتریس در هم‌ریختگی قابل استخراج است که در ادامه ارائه شده:

جدول ٤- ماتریس درهم‌ریختگی پیش‌بینی مدل رگرسیون لجستیک روی مجموعه‌ی آزمون

		Reference	
		No	**Yes**
Prediction	**No**	۲۲۳	۷۹
	Yes	۱۶۰	۷۸

مدل درخت تصمیم[1]

مزیت مدل درخت تصمیم در قابلیت بالای تفسیر آن است که می‌تواند به نحوی سبب تصمیم‌گیری بهتر افراد فعال در حوزه‌ی مورد مطالعه شود و برخی روابط پنهان را کشف نماید.

[1] Decision Tree

جدول ۵- نتایج یادگیری و آزمون مدل یادگیری درخت تصمیم بر حسب درصد

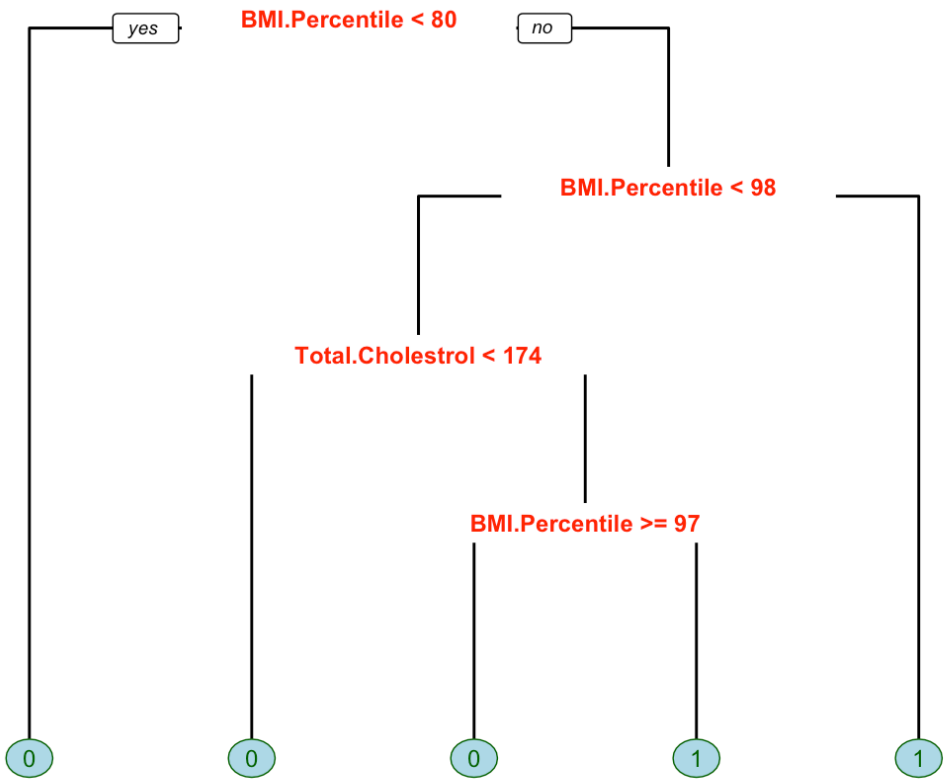
	نتایج یادگیری (بر حسب درصد)	**نتایج آزمون** (بر حسب درصد)
Accuracy	۶۹.۲	۶۶.۱
Sensitivity	۳۲.۵	۲۱.۷
Specifity	۸۴.۲	۸۴.۳
Precision	۴۵.۷	۳۶.۱
F1 Score	۳۷.۹۸	۲۷.۱

در این مسئله مدل درخت تصمیم برای پیش‌بینی دیابت و پیش‌دیابت در مجموعه‌ی آموزش یک صحت حدود ۶۹.۲ درصد و در مجموعه‌ی آزمون صحت حدود ۶۶.۱ درصد را از خود نشان داد. همچنین حساسیت این مدل که نمایانگر توانایی شناسایی صحیح افراد مبتلا به دیابت است، به ترتیب ۳۲.۵ درصد در مجموعه‌ی آموزش و ۲۱.۷ درصد در مجموعه‌داده‌ی آزمون بود. این در حالی است که مدل دقت بالایی در شاخص ویژگی به میزان ۸۴.۲ درصد در آموزش مدل و ۸۴.۳ درصد در آزمون آن از خود نشان داد و سایر شاخص‌ها نیز در جدول قابل مشاهده هست. تمامی این موارد از ماتریس در هم‌ریختگی قابل استخراج می‌باشد که در ادامه ارائه شده است:

جدول ٦- ماتریس درهم‌ریختگی پیش‌بینی مدل درخت تصمیم روی مجموعه‌ی آزمون

		Reference	
		No	**Yes**
Prediction	**No**	۳۲۳	۱۲۳
	Yes	۶۰	۳۴

همچنین نتیجه‌ی مدل‌سازی این الگوریتم روی مجموعه‌داده‌ی یادگیری به صورت زیر است:



نمودار ۳- درخت تصمیم

مدل جنگل تصادفی[1]

علیرغم این که مدل جنگل تصادفی در عمل سرعت پایینی دارد و همچنین قابلیت تفسیرپذیری ندارد، اما قابلیت بالایی در پیش‌بینی به خصوص پیش‌بینی داده‌های نامتعادل دارد که این مورد نیز از این قاعده مستثنی نیست.

[1] Random Forest

جدول ۷- نتایج یادگیری و آزمون مدل یادگیری جنگل تصادفی بر حسب درصد

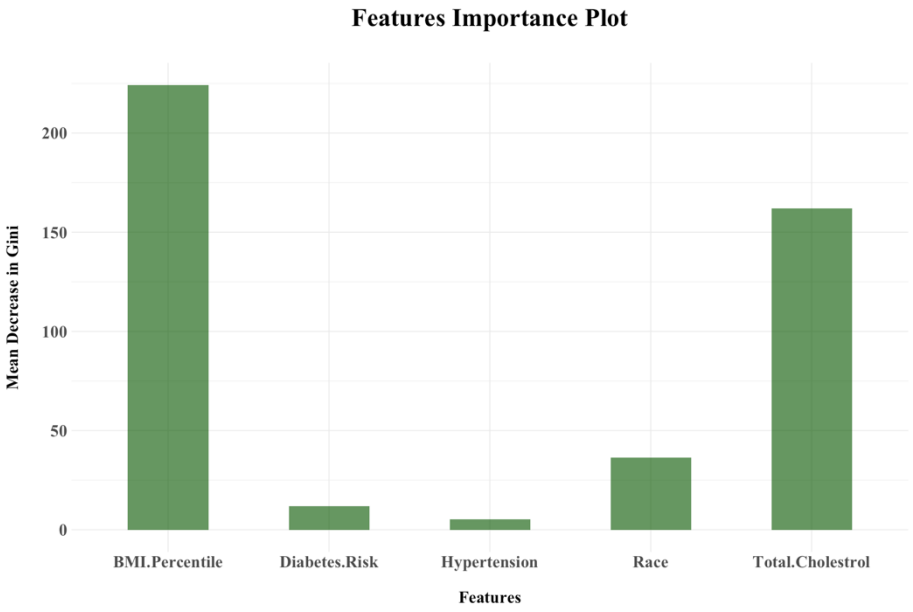
	نتایج یادگیری (بر حسب درصد)	**نتایج آزمون** (بر حسب درصد)
Accuracy	۸۸.۷	۶۶.۵
Sensitivity	۷۰.۰	۲۵.۵
Specifity	۹۶.۳	۸۳.۳
Precision	۸۸.۵	۳۸.۴۶
F1 Score	۷۸.۱۷	۳۰.۶۶

در این مسئله مدل جنگل تصادفی دقت بالایی را از خود نشان داده است و با حدود ۸۸.۷ درصد صحت در مجموعه‌ی آموزش و ۶۶.۵ درصد در مجموعه‌ی آزمون یکی از بهترین نتایج را بین نتایج ارائه‌شده از خود نشان داد. همچنین شاخص ویژگی مدل نیز عملکرد مناسب ۹۶.۳ درصدی در داده‌های یادگیری و ۸۳.۳ درصدی در داده‌های آزمون داشت که نشان‌دهنده‌ی توانایی آن در شناسایی صحیح موارد غیر دیابتی است. با این حال، حساسیت به نسبتاً پایین این مدل با مقدار ۷۰ درصد در مجموعه‌ی آموزش و ۲۵.۵ در مجموعه‌ی آزمون نشان‌دهنده‌ی چالش در تشخیص موارد مثبت واقعی است و سایر شاخص‌ها نیز در جدول قابل مشاهده هست. تمامی این موارد از ماتریس در هم‌ریختگی قابل استخراج می‌باشد که در ادامه ارائه شده است:

جدول ۸- ماتریس درهم‌ریختگی پیش‌بینی مدل جنگل تصادفی روی مجموعه‌ی آزمون

		Reference	
		No	**Yes**
Prediction	**No**	۳۱۹	۱۱۷
	Yes	۶۴	۴۰

همچنین می‌توان در نمودار زیر مقایسه‌ای بر میزان اثرگذای متغیرهای ورودی نشان داد:



نمودار ٤- میزان اثرگذاری متغیرهای ورودی در پیش‌بینی مدل جنگل تصادفی

فصل پنجم

تعریف مسئله
و گزارشی بر مجموعه‌داده‌ی جمع‌آوری شده

سایر شاخص‌های ارزیابی عملکرد مدل‌های یادگیری

علاوه بر شاخص‌های ارزیابی فاز قبل پروژه (فصل چهارم)، در این بخش سه شاخص دیگر نیز استفاده شده است که در ادامه به اختصار به بررسی هر کدام پرداخته می‌شود.

شاخص مقدار پیشبینی‌کننده‌ی مثبت[1]

این شاخص، نسبت تعداد پیش‌بینی‌های صحیح مثبت به کل پیش‌بینی‌های مثبت را محاسبه می‌کند و در شرایطی حیاتی است که هزینه انجام یک پیش‌بینی نادرست مثبت بالا باشد. این متریک به ارزیابی قابلیت اعتبار مدل در پیش‌بینی یک نتیجه‌ی مثبت کمک می‌کند.

$$PPV = \frac{TP}{TP + FP}$$

شاخص مقدار پیشبینی‌کننده‌ی منفی[2]

این شاخص، نسبت تعداد پیش‌بینی‌های صحیح منفی به کل پیش‌بینی‌های منفی را محاسبه می‌کند و در زمانی اهمیت دارد که هزینه‌ی خطاهای منفی بالا باشد، زیرا به دقت پیش‌بینی‌های منفی تمرکز دارد.

$$NPV = \frac{TN}{TN + FN}$$

بیش‌نمونه‌گیری

با توجه به عدم هم‌اندازه بودن رکوردهای دارای تارگت ۱ و ۰، با توجه به آن‌چه در مطالعه‌ی مذکور انجام شده است، بیش‌نمونه‌گیری انجام شد و مجدداً داده‌ها به دو دسته‌ی یادگیری و آزمون با نسبت ۸۰ به ۲۰ تقسیم شد. بعد از این گام تنها الگوریتم‌های یادگیری پیاده‌سازی شد و نتایج هر یک از این مدل‌ها به همراه بررسی نتایج آن‌ها در ادامه ارائه شده است.

[1] Positive Predictive Value
[2] Negative Predictive Value

مدل‌سازی مسئله

در این گام از پروژه مدل‌های مورد استفاده در فصل قبل مجددا پیاده‌سازی شد، با این تفاوت که در این بخش مدل‌سازی روی داده‌ها بیش‌نمونه‌گیری شده انجام شد و همچنین جهت اطمینان از حفظ بهترین نتیجه‌گیری مدل‌ها اعتبارسنجی نیز شدند. علاوه بر این، در این بخش جهت اطمینان از دریافت بهترین نتیجه الگوریتم بیز ساده که یک الگوریتم قابل اطمینان و سریع در مسائل دسته‌بندی است نیز مورد استفاده قرار گرفت. همچنین لازم به ذکر است که مقدار مناسب سرحد[1] تعیین برچسب متناسب با آنچه در مقاله انجام شده است، تعیین شده.

مدل رگرسیون لجستیک

در این بخش مدل رگرسیون لجستیک آموزش داده شد و به وسیله‌ی ۵ لایه مورد اعتبارسنجی قرار گرفت و سپس با استفاده از یک ماتریس ابهام با تمرکز بر روی کلاس مثبت، عملکرد مدل ارزیابی شد که در ادامه می‌توان خروجی این نتیجه را مشاهده نمود.

جدول ۹- نتایج یادگیری و آزمون مدل یادگیری رگرسیون لجستیک بر حسب درصد

	نتایج یادگیری (بر حسب درصد)	**نتایج آزمون** (بر حسب درصد)
Accuracy	۵۷.۶	۵۸.۵
Sensitivity	۴۸.۸	۴۵.۹
Specifity	۶۱.۱	۶۳.۷
PPV	۳۳.۹	۳۴.۱
NPV	۷۴.۴	۷۴.۱
F1 Score	۴۰.۰	۳۹.۱

با توجه به آنچه در مطالعه انجام شده بود، در این بخش تصمیم بر تعیین سرحد اطمینان با مقدار ۰.۳ شد.

[1] Threshold

مدل درخت تصمیم

در این بخش مدل درخت تصمیم آموزش داده شد و به وسیله‌ی ۱۰ لایه مورد اعتبارسنجی قرار گرفت و در یکی از گریدها مقدار به بهینه‌ی خود رسید و سپس با استفاده از یک ماتریس ابهام با تمرکز بر روی کلاس مثبت، عملکرد مدل ارزیابی شد که در ادامه می‌توان خروجی این نتیجه را مشاهده نمود.

جدول ۱۰- نتایج یادگیری و آزمون مدل یادگیری درخت تصمیم بر حسب درصد

	نتایج یادگیری (بر حسب درصد)	**نتایج آزمون** (بر حسب درصد)
Accuracy	۶۶.۱	۶۹.۲
Sensitivity	۲۱.۷	۳۲.۵
Specifity	۸۴.۳	۸۴.۲
PPV	۳۶.۱	۴۵.۷
NPV	۷۲.۴	۷۵.۳
F1 Score	۲۷.۱	۳۶.۰

با توجه به آنچه در مطالعه انجام شده بود، در این بخش تصمیم بر تعیین سرحد اطمینان با مقدار ۰.۳ شد.

مدل جنگل تصادفی

در این بخش مدل جنگل تصادفی آموزش داده شد و به وسیله‌ی ۱۰ لایه مورد اعتبارسنجی قرار گرفت و در یکی از گریدها مقدار به بهینه‌ی خود رسید و سپس با استفاده از یک ماتریس ابهام با تمرکز بر روی کلاس مثبت، عملکرد مدل ارزیابی شد که در ادامه می‌توان خروجی این نتیجه را مشاهده نمود.

جدول ۱۱- نتایج یادگیری و آزمون مدل یادگیری جنگل تصادفی بر حسب درصد

	نتایج یادگیری (بر حسب درصد)	**نتایج آزمون** (بر حسب درصد)
Accuracy	۶۳.۴	۶۲.۶
Sensitivity	۴۲.۱	۳۹.۵
Specifity	۷۲.۷	۷۲.۱
PPV	۳۸.۲	۳۶.۷
NPV	۷۵.۳	۷۴.۴
F1 Score	۴۰.۱	۳۸.۰

همچنین لازم به ذکر است که در این بخش مقدار لاپلاس ۱ قرار داده شد تا مانع هر گونه مشاهده‌ی احتمال و ایجاد خلل در محاسبات مدل شود. با توجه به آنچه در مطالعه انجام شده بود، در این بخش تصمیم بر تعیین سرحد اطمینان با مقدار ۰.۳۱ شد.

مدل بیز ساده

در این بخش مدل بیز ساده آموزش داده شد و به وسیله‌ی ۱۰ لایه مورد اعتبارسنجی قرار گرفت سپس با استفاده از یک ماتریس ابهام با تمرکز بر روی کلاس مثبت، عملکرد مدل ارزیابی شد که در ادامه می‌توان خروجی این نتیجه را مشاهده نمود.

جدول ۱۲- نتایج یادگیری و آزمون مدل یادگیری بیز ساده بر حسب درصد

	نتایج یادگیری (بر حسب درصد)	**نتایج آزمون** (بر حسب درصد)
Accuracy	۷۸.۲	۶۷.۶
Sensitivity	۴۲.۳	۱۸.۵
Specifity	۹۲.۷	۸۷.۷
PPV	۷۰.۱	۳۸.۲
NPV	۷۹.۸	۷۲.۴
F1 Score	۵۲.۸	۲۴.۹

با توجه به آنچه در مطالعه انجام شده بود، در این بخش تصمیم بر تعیین سرحد اطمینان با مقدار ۰.۲ شد.

جمع‌بندی و ارائه‌ی پیشنهادات

همانطور که پیشتر ذکر شد، این مطالعه پیرو افزایش روزافزون پیش دیابت و دیابت در میان جوانان انجام شده است و همانطور که بیان شد تشخیص ناکافی این شرایط موردی چالش برانگیز در میان افراد کم‌سن و سال است و از این رو بر لزوم استفاده از ابزارهای دقیق جایگزین تاکید می‌کند. در این مطالعه عملکرد شاخص‌های غربالگری نسبت به مدل‌های یادگیری ماشین مورد ارزیابی قرار گرفت که در ادامه به طور خلاصه می‌توان خروجی حاصل از مدل‌سازی را برای مجموعه‌داده‌ی آزمون مشاهده نمود.

جدول ۱۳- نتایج یادگیری و آزمون ممدل‌های مورد استفاده

	رگرسیون لجستیک	درخت تصمیم	جنگل تصادفی	بیز ساده
Accuracy	۵۸.۵	۶۹.۲	۶۲.۶	۶۷.۶
Sensitivity	۴۵.۹	۳۲.۵	۳۹.۵	۱۸.۵
Specifity	۶۳.۷	۸۴.۲	۷۲.۱	۸۷.۷
PPV	۳۴.۱	۴۵.۷	۳۶.۷	۳۸.۲
NPV	۷۴.۱	۷۵.۳	۷۴.۴	۷۲.۴
F1 Score	۳۹.۱	۳۶.۰	۳۸.۰	۲۴.۹

با توجه به این جدول از مدل‌های مختلف یادگیری ماشین برای پیش‌بینی می‌توان مشاهده کرد که عملکردهای مختلفی در میان مدل‌های مورد بررسی وجود دارد. برای نمونه در حالی که مدل بیز ساده حساسیت پایینی دارد و اشتباهات مثبت را به حداقل می‌رساند، مدل رگرسیون لجستیک در حساسیت از عملکرد بهتری برخوردار است که برای شناسایی موارد مثبت حیاتی است. رگرسیون لجستیک تعادلی بین حساسیت، خاصیت و دقت دارد که این امر آن را به یک گزینه قوی برای رویکرد جامع تبدیل می‌کند.

با توجه به آن چه در مطالعه انجام شده و محدودیت‌های آن می‌توان موارد زیر را برای بهبود و توسعه‌ی آن پیشنهاد داد:

- یکی از مواردی که مورد توجه است، دیتای قدیمی این مطالعه است که می‌توان با افزودن دیتای موجود در وب‌سایت تا سال ۲۰۲۲ به دقت و صحت این مطالعه افزود.
- مورد بعدی که در این مطالعه به چشم می‌رسد، بهره‌گیری نویسندگان آن از داده‌های تشخیص آزمایشگاهی است که می‌توان برای گسترش این مطالعه از داده‌های عوارض بلندمدت بیماری مثل کمای دیابتی استفاده نمود. این بدان معناست که با توجه به شرایط بیماران مبتلا به دیابت که گاها به کمای دیابتی وارد می‌شوند که در آن قند بیمار ناگهان بالا می‌رود و هوشیاری کم می‌شود؛ پیشنهاد می‌شود که در این در بازه‌ی ده‌ساله وضعیت فالوآپ بیماران مورد مطالعه قرار گیرد تا بتوان با اطمینان خاطر از مدل پیش‌بینی دفاع نمود.
- در این مطالعه شرایط متابولیسمی مثل چربی خون بالا و مشکلات هورمونی مورد بررسی قرار نگرفته که می‌تواند منجر به ایجاد عوارض بیشتر این بیماری و افزایش شدت آن شود که پیش‌بینی این مورد هم می‌تواند مورد بررسی قرار گیرد.
- مورد آخر که در این مطالعه قابل ارائه است، بررسی شاخصه‌های رشد جوانان می‌باشد. همانطور که جامعه‌ی مورد مطالعه افراد ۱۲ تا ۱۹ سال را مورد بررسی قرار می‌دهد؛ می‌توان شاخصه‌های بلوغ و رشد آن‌ها را نیز در نظر گرفت.

منابع و مراجع

[1] N. Vangeepuram, B. Liu, P. hsiang Chiu, L. Wang, and G. Pandey, “Predicting youth diabetes risk using NHANES data and machine learning,” *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/S41598-021-90406-0.

[2] “Prediabetes - Your Chance to Prevent Type 2 Diabetes | CDC.” Accessed: Dec. 28, 2023. [Online]. Available: https://www.cdc.gov/diabetes/basics/prediabetes.html

[3] “Diabetes Research, Education, Advocacy | ADA.” Accessed: Dec. 26, 2023. [Online]. Available: https://diabetes.org/

[4] S. Arslanian, F. Bacha, M. Grey, M. D. Marcus, N. H. White, and P. Zeitler, “Evaluation and Management of Youth-Onset Type 2 Diabetes: A Position Statement by the American Diabetes Association,” *Diabetes Care*, vol. 41, no. 12, pp. 2648–2668, Dec. 2018, doi: 10.2337/DCI18-0052.

[5] A. Géron, “Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems,” *O’Reilly Media*, p. ```, 2017.