دانشگاه صنعتی شریف
دپارتمان مهندسی صنایع

فاز دوم پروژه‌ی پایانی درس

مدلسازی و تصمیم‌گیری داده‌محور

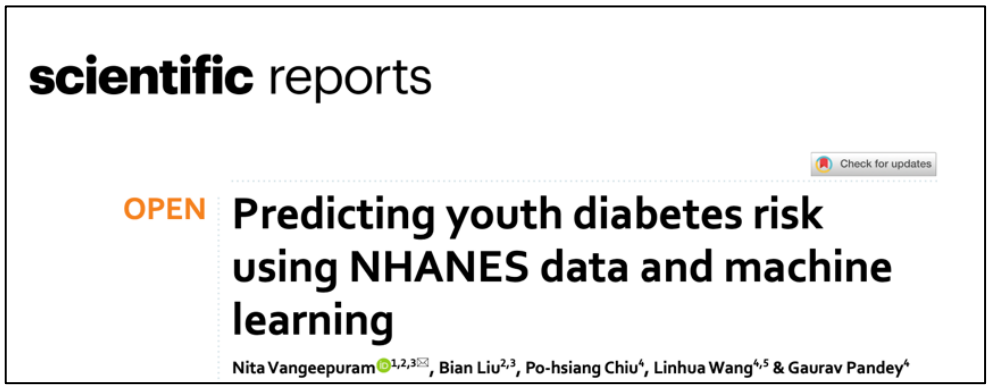
استاد درس | سرکار خانم دکتر صدقی

گردآوری

محمدحسین محمودی ۴۰۱۲۱۱۹۷۴
اشکان فرهودی زارع ۴۰۱۲۱۱۹۱۷

دی ماه ۱۴۰۲

مقدمه


scientific reports

Check for updates

OPEN

Predicting youth diabetes risk using NHANES data and machine learning

Nita Vangeepuram[1,2,3]✉, Bian Liu[2,3], Po-hsiang Chiu[4], Linhua Wang[4,5] & Gaurav Pandey[4]

در مطالعه‌ی فوق که پیشتر در پروپوزال پیشنهادی پروژه به آن پرداخته شده، تلاش شده است تا مدل‌های یادگیری ماشین را مورد ارزیابی قرار دهد و به نتیجه‌ای بهتر از عملکرد شاخص‌های غربالگری دست یابد. در این فاز از پروژه به شرح مسئله به طور کامل پرداخته شده است و تماما فرآیند جمع‌اوری دیتاست به صورت کامل تشریح شده است. همچنین در این گام از پروژه‌ی پایانی این درس مدل‌های یادگیری ماشین تدریس‌شده در کلاس درس به کار گرفته شده تا بتوان به حداقل نتیجه‌ی مطلوب دست یافت. هدف از این گزارش شرحی خلاصه و روان از آنچه انجام شده است می‌باشد.

فهرست مطالب

فصل اول

تعریف مسئله
و گزارشی بر مجموعه‌داده‌ی جمع‌آوری شده

شرح مسئله و ضرورت پژوهش

گسترش پیش‌دیابت و دیابت قندی در بین جوانان، همراه با ضعف در تشخیص آن‌ها، نیاز بحرانی به ابزارهای قوی و قابل دسترس برای غربال‌گری دارد. راهنمایی‌های موجود برای غربال‌گری بالینی کودکان و نوجوانان، از جمله آن‌هایی که توسط انجمن دیابت آمریکا و انجمن پزشکان کودکان آمریکا تأیید شده‌اند، محدودیت‌هایی در تشخیص دقیق جوانان دارای دیابت و پیش‌دیابت بر اساس معیارهای زیست‌نشانگر 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 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 کند و آن‌ها را به سوی مداخلات پیشگیرانه پیشنهادی هدایت کند.

پیش‌دیابت و دیابت در نوجوانان

تفاوت دیابت و پیش‌دیابت قندی

پیش‌دیابت یک شرایط سلامتی است که 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 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 نرسیده‌است. 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 کنند. نظارت منظم و نظارت پزشکی برای افراد مبتلا به پیش‌دیابت برای مدیریت بهتر سلامت آن‌ها ضروری است. [۱]

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 [۲] به پرداخته شده است. به طور خلاصه می‌توان اهمیت این مهم را در موارد زیر خلاصه نمود:

- بیماری مزمن جدی با عوارض

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.

- 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.

- شیوع بالا در میان جوانان

این پژوهش طبق مطالعات پیشین برجسته می‌کند که هر دو نوع دیابت و پیش‌دیابت به طور نگران‌کننده‌ای در میان جوانان شیوع پیدا کرده‌است و تعداد قابل توجهی از جوانان سالانه به دیابت نوع دوم تشخیص داده می‌شوند. پرداختن به این وضعیت‌ها در ابتدای مراحل پیشرفت بیماری بسیار حیاتی است به دلیل افزایش شیوع آن‌ها لازم است تا بیشتر به آن پرداخته شود.

- تفاوت‌ها در شیوع در گروه‌های جمعی

تفاوت‌هایی در شیوع پیش‌دیابت 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.

- 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 و شروع زودهنگام عوارض آن همراه است. شناسایی زودهنگام بیماری امکان مدیریت و پیشگیری مؤثرتر از عوارض آن را فراهم می‌کند.

- تأثیر بیشتر از منظر ابعاد اقتصادی برای جوانان

تأثیر اقتصادی بالقوه‌ی بیماری دیابت برای جوانان نسبت به افراد دارای سن 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.

- تشخیص ضعیف‌تر و عدم آگاهی میان جامعه

با وجود گستره‌ای از راهنماها که توسط سازمان‌هایی مانند انجمن دیابت آمریکا ارائه شده‌است، پیش‌دیابت در میان جوانان اغلب کمتر تشخیص داده می‌شود و بسیاری از جوانان ممکن است مراجعه‌ی سالانه‌ی پیشنهادی توسط سین سازمان‌ها را نداشته باشند و خدمات پیشگیری را دریافت نکنند که این امر منجر به عدم آگاهی از وضعیت خود می‌شود.

- نیاز به ابزارهای اسکرینینگ

همچنین در این پژوهش به نیاز به ابزارهای غربال‌گری موثر تأکید می‌کند چرا که بسیاری از جوانان از پیش‌دیابت یا دیابت خود آگاه نیستند و یک ابزار ساده، غیرتهاجمی 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.

پس به طور کلی می‌توان بیان نمود که تشخیص زودهنگام دیابت و پیش‌دیابت در میان جوانان برای جلوگیری از عوارض، پرداخت به تفاوت‌ها و اجرای مداخلات مؤثر، از جمله تغییرات در سبک زندگی و تدابیر پیشگیری، بسیار حیاتی است.

روش‌های تشخیص بیماری

انجمن دیابت آمریکا

انجمن دیابت آمریکا[1] که در سال ۱۹۳۹ توسط شش پزشک دورنگر 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 پیش می‌برد. این دست از پشتیبانی‌ها انواع دیابت من‌جمله دیابت نوع ۱ و نوع ۲ را تا دیابت بارداری و پیش‌دیابت 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 سبک‌زندگی سالم می‌شود که اقلیت‌ها را برای مقابله با پیچیدگی‌های دیابت حمایت می‌کند. طبق داده‌های موجود از عملکرد آن، سرمایه‌گذاری‌های این انجمن تأثیرگذار بوده و هر دلار سرمایه‌گذاری در تحقیقات دیابت منجر به ۱۲.۴۷ دلار سرمایه‌گذاری اضافی شده است که این یک مهر تایید به 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 یا وضعیت یک بیماری یا شرایط خاص می‌باشد. اصطلاح معیارهای زیست‌نشانگر به مجموعه شرایط یا پارامترها اطلاق می‌شود که باید برآورده شود تا یک ماده‌ی 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، تشخیص، مانیتورینگ، پیش‌آگاهانه، پیش‌بینی‌کننده، پاسخی، و ایمنی می‌شود. یک زیست‌نشانگر می‌تواند یک ویژگی تکی یا یک پنل از ویژگی‌ها باشد. به عنوان مثال، در

[1] American Diabetes Association
[2] Biomarker Criteria

زمینه‌ی درمان سرطان، آزمون‌های زیست‌نشانگر می‌تواند تغییرات ژنتیک مرتبط با سرطان را شناسایی کنند. برخی از آزمون‌ها به تمام ژن‌های سرطان شما نگاه می‌کنند، برخی به تمام ساختار سرطان نگاه می‌کنند و برخی دیگر به تعداد تغییرات ژنتیک در سرطان نگاه می‌کنند.

دستورالعمل غربال‌گری[3]

هرچند که اصطلاح دستورالعمل غربالگری در منابع ارائه شده 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 جمعیت یا در شرایط خاصی 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 خاص سرطان، بر اساس سن، تاریخچه‌ی خانوادگی یا عوامل دیگر، آزمایش‌های معینی برای زیست‌نشانگرهای خاصی به صورت روزانه انجام شود.

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 [۴] که این مقاله بر اساس آن پیش رفته است معیارها و دستورالعمل به صورت زیر است:

- دستورالعمل زیست‌نشانگر انجمن دیابت آمریکا برای تشخیص دیابت به معیارهای تست ناشتای گلوکز خون، تست قند دوساعته و تست هموگلوبین گلیکوزیله‌شده رجوع می‌کند و بر اساس آن، دیابت می‌تواند تشخیص داده شود.
- دستورالعمل‌های غربالگری باید در جوانانی که اضافه وزن دارند و یک یا چند عامل خطرناک دیگر نظیر تاریخچه‌ی مادری ابتلا به دیابت، تاریخچه‌ی خانوادگی دیابت نوع ۲، نژاد یا قومیت‌های متسعد به دیابت و نشانه‌های مقاومت به انسولین یا شرایط مرتبط با مقاومت به انسولین دارند، مدنظر قرار گیرد. طبق این

[3] Screening Guideline

آزمایش‌های غربالگری باید پس از شروع دوران نوجوانی یا پس از ده سالگی انجام شود، هر کدام که زودتر رخ دهد.

معرفی مجموعه‌داده

مطالعه‌ی ملی سلامت و تغذیه[4]

درباره‌ی دستور کار مطالعه‌ی ملی سلامت و تغذیه

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[5]، بخشی از مراکز کنترل و پیشگیری از بیماری‌ها[6] 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.

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 ویژگی‌های مجموعه‌داده‌ی نهایی طبق چارت ۱ استخراج شده است و متغیرهای ورودی به آن به شرح زیر است:

مجموعه‌داده‌ی جمعیت‌شناسی[7]

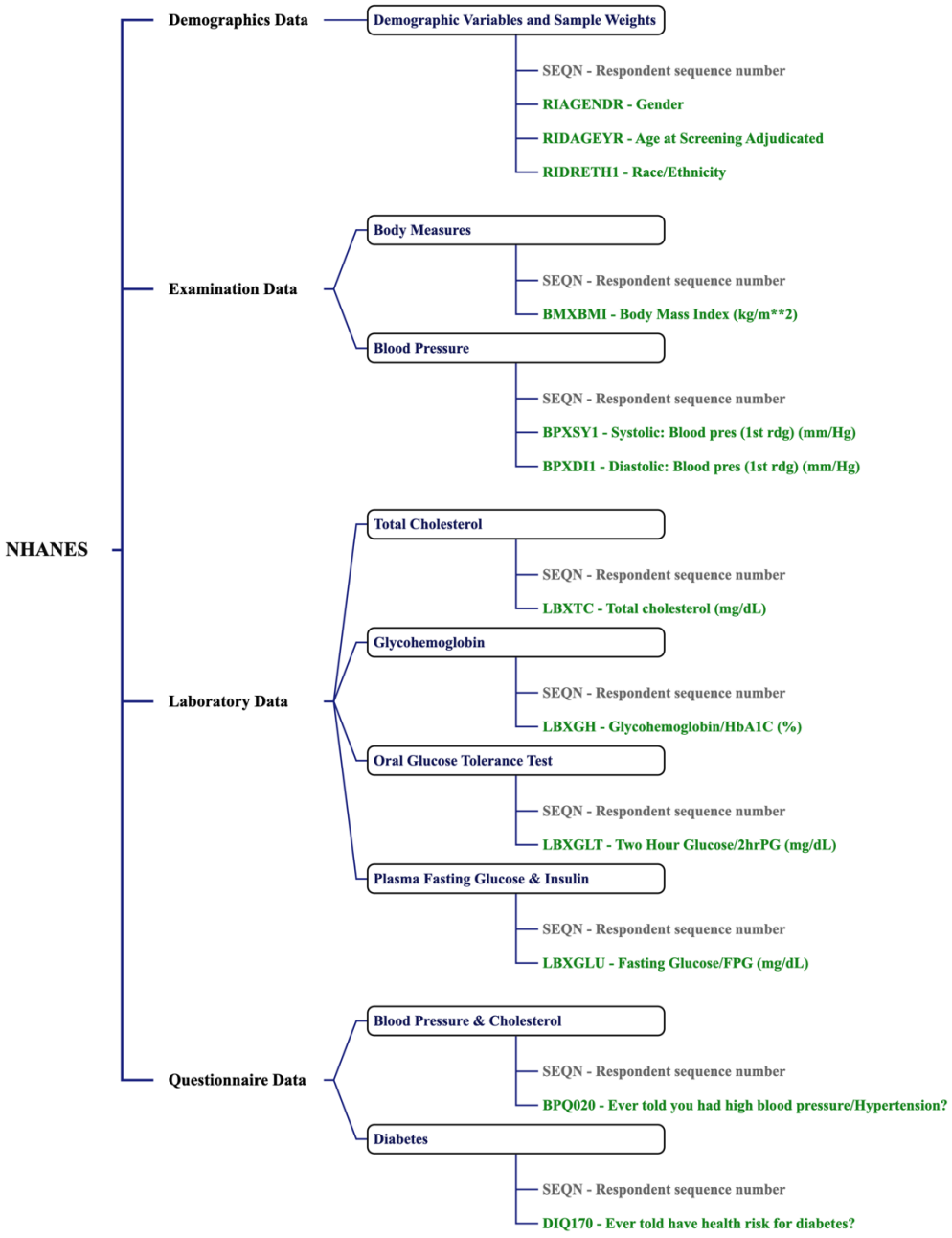
از مجموعه‌داده‌ی جمعیت شناختی که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها)، جنسیت، سن و نژاد استخراج شده است.

---

[4] National Health and Nutrition Examination Survey (NHANES)
[5] National Center for Health Statistics (NCHS)
[6] Centers for Disease Control and Prevention (CDC)
[7] Demographic Data

مجموعه‌داده‌ی شاخص‌های فیزیکی[8]

از مجموعه‌داده‌ی شاخص‌های فیزیکی که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و شاخص توده‌ی بدنی[9] استخراج شده است.



نمودار ۱- چارت متغیرهای استخراج‌شده از دیتاست‌های موجود

[8] Body Measures Data
[9] Body Mass Index (BMI)

مجموعه‌داده‌ی فشار خون[10]

از مجموعه‌داده‌ی فشار خون که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و فشار خون سیستولیک[11] و فشار خون دیاستولیک[12] استخراج شده است.

مجموعه‌داده‌ی کلسترول کل[13]

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.

مجموعه‌داده‌ی گلیکوهموگلوبین[14]

از مجموعه‌داده‌ی گلیکوهموگلوبین که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و هموگلوبین گلیکوزیله[15] استخراج شده است.

مجموعه‌داده‌ی تست تحمل گلوکز خوراکی[16]

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 استخراج شده است.

مجموعه‌داده‌ی گلوکز ناشتا و انسولین پلاسما[17]

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[18] استخراج شده است.

---

[10] Blood Pressure Data
[11] Systolic Blood Pressure
[12] Diastolic Blood Pressure
[13] Total Cholestrol Data
[14] Glycohemoglobin Data
[15] HbA1C
[16] Oral Glucose Tolerance Test Data
[17] Plasma Fasting Glucose and Insulin Data
[18] FPG

مجموعه‌داده‌ی فشار خون و کلسترول[19]

از مجموعه‌داده‌ی تست فشار خون و کلسترول که در وب‌سایت مطالعه‌ی سلامت و تغذیه موجود است متغیرهای شماره‌ی بیمار (جهت یکپارچه‌سازی مجموعه‌داده‌ها) و سابقه‌ی شخص و یا خانوادگی بیماری دیابت آن‌ها استخراج شده است.

مجموعه‌داده‌ی دیابت[20]

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.

[19] Blood Pressure and Cholestrol Data
[20] Diabetes Data

فصل دوم

فرآیند جمع‌آوری داده

فرآیند جمع‌آوری مجموعه‌داده‌ی نهایی

شرح مراحل جمع‌اوری و برچسب‌گزاری مجموعه‌داده

استخراج و ادغام مجموعه‌داده‌ها

همانطور که پیشتر ذکر شد، مجموعه‌داده‌ی مورد استفاده به صورت آماده در اختیار نبود و برای طراحی پژوهش نیاز به جمع‌اوری دستی آن بود. پس از مطالعات بسیار و بررسی مقاله‌ی مورد مطالعه، نتیجه‌گیری بابت اطلاعات لازم به استخراج گرفته 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 استخراج شد و به علت نوع فایل این مجموعه‌داده‌ها (XPT.) عملیات استخراج و ترکیب مجموعه‌داده‌ها بر پایه‌ی زبان پایتون انجام شده است.

برچسب‌گذاری بیماران در مجموعه‌داده

با توجه به آنچه در مقاله‌ی پایه [۲] ارائه شده است، برچسب‌گذاری غربال‌گری بر اساس شاخص‌های زیر انجام می‌شود.

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

جدول ۱- شاخص‌های غربال‌گری تشخیص پیش‌دیابت و دیابت

بر اساس تعاریفی که پیشتر بر شاخص‌های زیست نشانگر ارائه شده‌است، قواعد فوق کاربرد در تشخیص غربال‌گری بیماران 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.

همچنین جهت برچسب‌گذاری شاخص زیست‌نشانگر تشخیص پیش‌دیابت و دیابت از مطالعه‌ای [۴] که این مقاله به آن ارجاع داده است استفاده شده که در روبرو قاعده‌ی محاسبه‌ی این برچسب‌ها ارائه شده است.

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.

جدول ۲- شاخص‌های زیست‌نشانگر تشخیص پیش‌دیابت و دیابت

در این فاز از پروژه که تمرکز بر روی بخش یادگیری ماشین مقاله بوده است، تلاش شده تا بر اساس شاخص‌های زیست‌نشانگر تشخیص پیش‌دیابت و یا دیابت به پیش‌بینی وضعیت بیماران پرداخته شود و مقایسه‌ی فرآیند غربال‌گری با نتیجه‌ی پیش‌بینی ماشینی در فاز بعدی قابل ارائه 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 ویژگی‌های به جز آیدی بیماران و برچسب وضعیت غربالگری آنها) در طول تشریح مسئله مورد استفاده قرار می‌گیرد. در ادامه می‌توان گزارشی از ویژگی‌های مجموعه‌داده 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.

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

جدول ٤- نسبت‌های پراکندگی بیماران برای ویژگی‌های مختلف

شرح این متغیرها به صورت زیر است:

جنسیت و نژاد

این متغیرها که از مجموعه‌داده‌ی دموگرافیک استخراج شده‌اند شامل مرد و زن در جنسیت و صفید غیرلاتین، سیاه غیر لاتین، لاتین و غیره در نژاد هستند.

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 وی تا کنون پس از معیانه توسط پزشک محتمل به ابتلا به دیابت در آینده خطاب شده باشد.

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.

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 یک شدن آن تنها یکی از سه شرط زیر کفایت می‌کند.

- تست گلوکز خون ناشتای بزرگتر مساوی ۱۰۰ میلی‌گرم بر دسی‌لیتر
- تست گلوکز خون دو ساعته‌ی بزرگتر مساوی ۱۴۰ میلی‌گرم بر بر دسی‌لیتر
- تست هموگلوبین گلیکوزیله بیشتر از ۶ درصد

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.

	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

جدول ۵- آماره‌های متغیرهای پیوسته

سن و شاخص توده‌ی بدنی

این متغیرها از دو مجموعه‌داده‌ی دموگرافیک و شاخص‌های فیزیکی استخراج شده است که ویژگی‌های آن به طور مشخص در جدول فوق قابل مشاهده است.

فشار خون سیستولیک و دیاستولیک

این دو متغیر از مجموعه‌داده‌ی فشار خون استخراج شده است که ویژگی‌های آن به طور مشخص در جدول فوق قابل مشاهده است.

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 جدول فوق قابل مشاهده است.

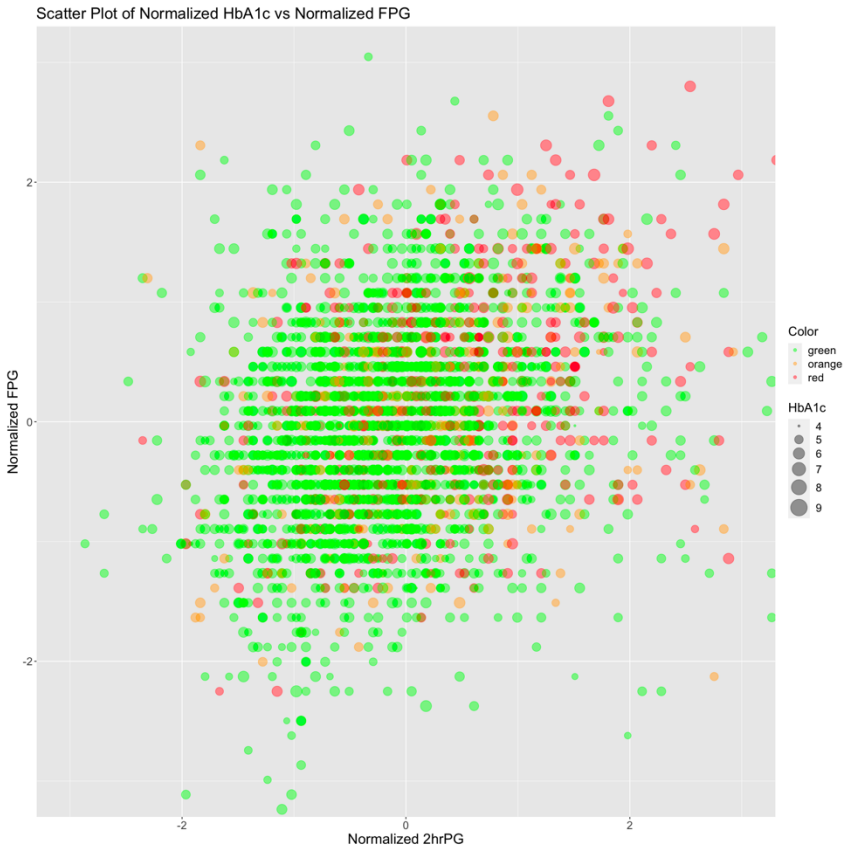
کلسترول کل خون

این متغیر نیز از مجموعه‌داده‌ی کسترول خون استخراج شده است که ویژگی‌های آن به طور مشخص در جدول فوق قابل مشاهده است.

گزارشی از تحلیل روابط بین متغیرهای تاثیرگذار

مقایسه‌ای بر گلوکز خون در بیماران

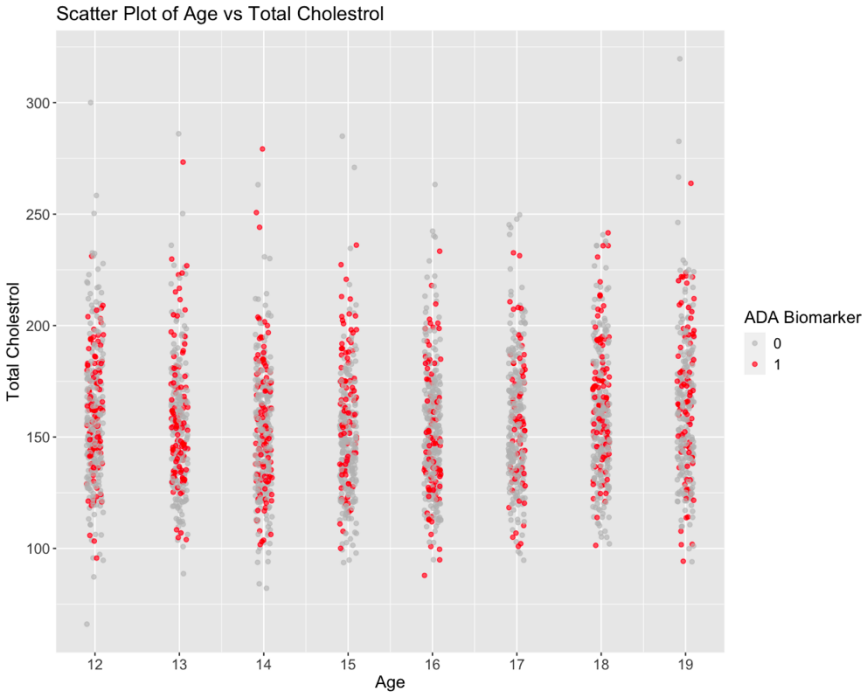
همانطور که مبرهن است گلوکز خون یکی از مهم‌ترین پارامترهای موثر در ابتلا به بیماری دیابت است و با توجه به نظر خبره تلاش شد در این بخش به وسیله‌ی نموداری از رکوردها رابطه‌ی بین گلوگز ناشتا و دوساعته را در کنار شاخص توده‌ی بدنی و هموگلوبین گلیکوزه مورد بررسی قرار داد. به وضوح روشن است که با حرکت به سمت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



نمودار ۱- مقایسه‌ی انواع گلوکز خون در بیماران

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



نمودار ۲- کلسترول کل خون و سن بیمارن

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 ساختمان داده‌ی موجود و صحت مقادیر آن پرداخته شد که با توجه به نظر خبره تصمیم به 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 پیوسته ارائه نشده بود که در هنگام جمع‌اوری داده‌ها به اصلاح این امر جهت همگامی با مقاله پرداخته شد.

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

مدلسازی مسئله

پیش از مدل‌سازی، همانند هر مسئله‌ی مبتنی بر یادگیری ماشین دیگری، مجموعه داده 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، مرتبط با خروجی پیش از بهبود الگوریتم‌های مورد استفاده است که این امر در فاز بعد از این پژوهش مورد توجه قرار خواهد گرفت.

رگرسیون لجستیک[۲۱]

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 جدول صفحه‌ی بعد می‌توان نتایج مدل‌سازی را مورد بررسی قرار داد.

[21] Logistic Regression

مدل رگرسیون لجستیک برای پیش‌بینی دیابت و پیش‌دیابت یک دقت 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 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 ویژگی 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. با این حال، مقدار پیش‌بینی مثبت و مقدار پیش‌بینی منفی دارای مقادیر نسبتاً کم به ترتیب ۳۳.۵ درصد و ۳۲.۸ درصد است که نشان می‌دهد باید دقت بیشتری در تفسیر احتمالات پیش‌بینی شده داشته باشیم چرا که توانایی مدل در پیش‌بینی دقیق موارد مثبت و منفی محدود است.

Metric	Training Results	Testing Results
Accuracy	0.5537	0.5574
95% CI	(0.5324, 0.5748)	(0.5144, 0.5998)
Kappa	0.0872	0.0686
Sensitivity	0.5470	0.4968
Specificity	0.5564	0.5822
Pos Pred Value	0.3353	0.3277
Neg Pred Value	0.7502	0.7384
Prevalence	0.2903	0.2907
Detection Rate	0.1588	0.1444
Detection Prevalence	0.4736	0.4407
Balanced Accuracy	0.5517	0.5395

جدول ٦- جدول شاخص‌های ارزیابی مدل رگرسیون لجستیک

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

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

جدول ۷- ماتریس درهم‌ریختگی پیش‌بینی مدل رگرسیون لجستیک روی مجموعه‌ی آزمون

که بر همین اساس نیز می‌توان شاخص F را نیز محاسبه نمود که برابر است با ۵۳.۶۲ درصد در داده‌های آزمون.

درخت تصمیم[۲۲]

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.

در این مسئله مدل درخت تصمیم برای پیش‌بینی دیابت و پیش‌دیابت در مجموعه‌ی آموزش یک دقت حدود ۶۹.۲ درصد و در مجموعه‌ی آزمون دقت 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.

Metric	Training Results	Testing Results
Accuracy	0.6921	0.6611
95% CI	(0.6722, 0.7116)	(0.6195, 0.701)
Kappa	0.1831	0.0679
Sensitivity	0.32536	0.21656
Specificity	0.84214	0.84334
Pos Pred Value	0.45740	0.36170
Neg Pred Value	0.75321	0.72422
Prevalence	0.29028	0.29074
Detection Rate	0.09444	0.06296
Detection Prevalence	0.20648	0.17407
Balanced Accuracy	0.58375	0.52995

جدول ۸- جدول شاخص‌های ارزیابی مدل درخت تصمیم

این امر نشان‌دهنده‌ی توانایی مدل در شناسایی صحیح بیماران غیر دیابتی است و حساسیت پایین آن که در این مسئله یک ضعف بزرگ محسوب می‌شود، نشان‌دهنده‌ی چالش‌های ممکن در تشخیص موارد مثبت واقعی است. مقدار پیش‌بینی مثبت و مقدار پیش‌بینی منفی همچنین به ترتیب با مقادیر متوسط ۴۵.۷ درصد و ۳۶.۲ درصد ارائه شده‌است، که نشان‌دهنده‌ی عملکرد متوازن مدل درخت تصمیم در پیش‌بینی موارد مثبت و منفی در این مجموعه‌داده است.

[22] Decision Tree

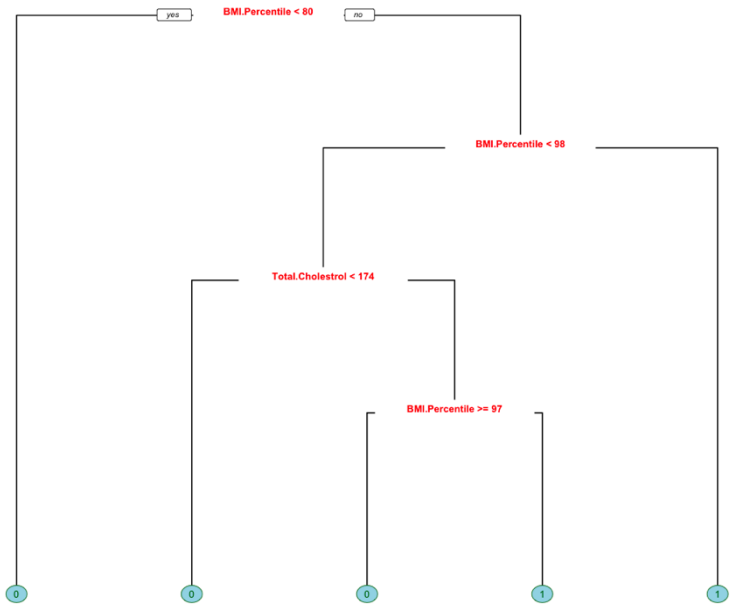
تمامی این موارد از ماتریس در هم‌ریختگی قابل استخراج می‌باشد که در ادامه ارائه شده است:

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

جدول ۹- ماتریس درهم‌ریختگی پیش‌بینی مدل درخت تصمیم روی مجموعه‌ی آزمون

که بر همین اساس نیز می‌توان شاخص F را نیز محاسبه نمود که برابر است با ۳۴.۵۲ درصد در داده‌های آزمون.

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



نمودار ۳- مدل درخت تصمیم نهایی

جنگل تصادفی[۲۳]

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.

در این مسئله مدل جنگل تصادفی دقت بالایی را از خود نشان داده است و با حدود ۸۸.۷ درصد دقت 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. مقدار پیش‌بینی مثبت به نسبت بالا و برابر با ۸۸.۵ درصد است، اما مقدار پیش‌بینی منفی با مقدار ۷۳.۲ درصد از آن کمتر است. این امر نیز نشان‌دهنده‌ی آن است که در حالی که مدل در پیش‌بینی موارد مثبت مؤثر است، امکان بهبود در توانایی آن در پیش‌بینی صحیح موارد منفی وجود دارد.

Metric	Training Results	Testing Results
Accuracy	0.8866	0.6648
95% CI	(0.8724, 0.8996)	(0.6233, 0.7046)
Kappa	0.7066	0.0974
Sensitivity	0.7002	0.25478
Specificity	0.9628	0.83290
Pos Pred Value	0.8851	0.38462
Neg Pred Value	0.8870	0.73165
Prevalence	0.2903	0.29074
Detection Rate	0.2032	0.07407
Detection Prevalence	0.2296	0.19259
Balanced Accuracy	0.8315	0.54384

جدول ۱۰- جدول شاخص‌های ارزیابی مدل جنگل تصادفی

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

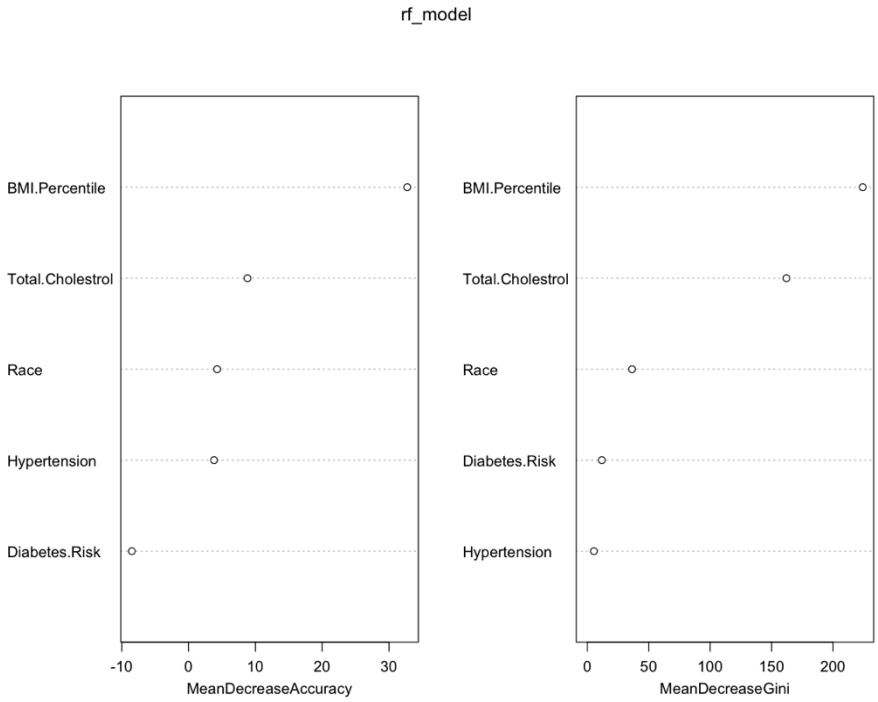
		Reference	
		No	**Yes**
Prediction	**No**	۳۱۹	۱۱۷
	Yes	۶۴	۴۰

جدول ۱۱- ماتریس درهم‌ریختگی پیش‌بینی مدل جنگل تصادفی روی مجموعه‌ی آزمون

[23] Random Forest

که بر همین اساس نیز می‌توان شاخص F را نیز محاسبه نمود که برابر است با ۳۹.۰۵ درصد در داده‌های آزمون.

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



نمودار ۲- میزان اثرگذاری متغیرهای ورودی در پیش‌بینی مدل جنگل تصادفی

جمعبندی

در این فاز از پروژه مدل‌های رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی برای پیش‌بینی دیابت ارزیابی شدند و هرکدام ویژگی‌های عملکرد متفاوتی را نشان دادند. به طور خلاصه، مدل جنگل تصادفی دقت کلی بالاتری را نشان داد، اما در حساسیت در مجموعه‌ی آزمون عملکرد چندان مطلوبی نداشت. مدل درخت تصمیم عملکرد متوازنی داشت، در حالی که مدل رگرسیون لجستیک محدودیت‌هایی از خود نشان داد، به ویژه در حساسیت نسبت به مقادیر مورد پیش‌بینی. به طور کلی برای بررسی دقیق‌تر مدل‌های به‌کار گرفته‌شده نیاز است تا در فاز بعد عملیات بیشتری روی مدل‌ها جهت بهبود آن‌ها انجام شود.

اقلام تحویلی فاز بعد

همانطور که در طول تشریح و تحلیل مسئله قابل مشاهده است، تمرکز در این فاز از پروژه روی دریافت درک درستی از مجموعه‌داده و پرسش مطرح‌شده است. علاوه بر این در این فاز از پروژه با استفاده از مدل‌های تدریس‌شده در کلاس در به پیش‌بینی پرداخته شده‌است که برای دریافت بهترین نتیجه کافی نیست. در نتیجه با توجه به محدودیت‌های این فاز از پروژه در فاز بعدی تلاش میشود تا موارد زیر نیز در نظر گرفته شود:

- بهبود و بهینه‌سازی مدل‌های رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی
- ارائه‌ی راهکار مناسب برای حل مسئله‌ی عدم تعادل در داده‌ها
- بررسی و آزمون مدل‌های خارج از چارچوب کلاس و ارائه‌ی قیاسی بر مدل‌های بررسی‌شده

منابع و مراجع

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

[2] 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.

[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.