پیش‌بینی ابتلا به دیابت در مردم آمریکا

دیابت یکی از شایع‌ترین بیماری‌ها در ایالات متحده، وسایر نقاط جهان، است که هر ساله بر زندگی میلیون‌ها آمریکایی تأثیر می‌گذارد و بار مالی قابل توجهی را بر اقتصاد کشورها به دنبال دارد. دیابت یک بیماری مزمن و پیش‌رونده است که در آن افراد قابلیت تنظیم سطح قند خون را به خوبی از دست می‌دهند، که این می‌تواند منجر به کاهش کیفیت زندگی و امید به زندگی شود. پس از تجزیه غذاهای مختلف به قندها در طول فرآیند هضم، قندهای حاصل به خون رها می‌شود. این عمل باعث می‌شود که پانکراس انسولین رها کند. انسولین به سلول‌های بدن کمک می‌کند تا از قندهای موجود در خون برای تولید انرژی استفاده‌کنند. دیابت به عدم تولید کافی انسولین توسط بدن یا ناتوانی در استفاده از انسولین به طور کافی تعریف می‌شود.

عوارضی همچون بیماری‌های قلبی، از دست دادن دید، قطع اندام‌های پایین در نتیجه‌ی بریدگی و بیماری‌های کلیوی، ناشی از سطوح بالای قند خون در افراد دیابتی است. در حالی که برای دیابت هیچ درمان مشخصی وجود ندارد، راهکارهایی همانند از دست‌دادن وزن، تغذیه سالم، فعالیت بدنی و دریافت درمان‌های پزشکی می‌تواند در کاهش آسیب‌های این بیماری در بسیاری از بیماران مفید باشد. تشخیص زودهنگام می‌تواند منجر به تغییرات سبک زندگی و درمان‌های موثرتر شود. بنابراین مدل‌های پیش‌بینی خطر دیابت ابزارهای بسیار مهمی برای جامعه و مسئولان بهداشت و سلامت عمومی است.

شناخت مقیاس این مساله نیز بسیار مهم است. براساس اعلام مرکز کنترل و پیشگیری از بیماری‌های آمریکا (CDC) تا سال ۲۰۱۸، ۳۴/۲ میلیون آمریکایی دیابت دارند و 88 میلیون نفر نیز پیش‌دیابت دارند. علاوه بر این، CDC برآورد می‌کند که ۱ نفر از هر ۵ نفر دیابتی، و تقریبا ۸ نفر از هر ۱۰ نفر پیش‌دیابتی از وضعیت خود اطلاعی ندارند. در حالی که انواع مختلفی از دیابت وجود دارد، دیابت نوع دوم شایع‌ترین نوع آن است و شیوع آن بستگی به سن، تحصیلات، درآمد، محل زندگی، نژاد و سایر عوامل تعیین‌کننده‌ی اجتماعی سلامت دارد. بار بیماری بیشتر بر دوش افراد با وضعیت اقتصادی پایین است. دیابت همچنین بار سنگینی را بر اقتصاد به دنبال دارد. مجموع هزینه‌ی افراد تشخیص‌داده شده به دیابت تقریبا ۳۲۷ میلیارد دلار، و هزینه‌های کل به همراه افراد دیابتی تشخیص داده نشده و و افراد پیش‌دیابتی در حدود سالانه ۴۰۰ میلیارد دلار برآورد می‌شود.

داده‌ها

سیستم نظارت بر عوامل خطر رفتاری (BRFSS) یک نظرسنجی تلفنی مربوط به سلامت است که سالانه توسط مرکز کنترل و پیشگیری از بیماری‌ها جمع‌آوری می‌شود. هر سال، این نظرسنجی پاسخ‌های بیش از ۴۰۰،۰۰۰ آمریکایی را در مورد رفتارهای خطرناک سلامت، بیماری‌های مزمن و استفاده از خدمات پیشگیری جمع‌آوری می‌کند. این نظرسنجی هر ساله از سال ۱۹۸۴ برگزار می‌شود. نتایج این نظرسنجی در سال 2015 در تعدادی فایل فرمت csv در دسترسی است. این مجموعه داده اصلی شامل پاسخ‌های ۴۴۱،۴۵۵ پرسش‌شونده و ۳۳۰ ویژگی است. این ویژگی‌ها یا به‌صورت مستقیم از شرکت‌کنندگان پرسیده شده‌اند و یا متغیرهای محاسبه‌شده بر اساس پاسخ‌های شرکت‌کنندگان فردی هستند.

این مجموعه داده شامل ۳ فایل است:

۱) فایل diabetes_012_health_indicators_BRFSS2015.csv: داده‌ی تمیز شده شامل ۲۵۳،۸۶۰ ردیف از جواب‌های داده شده به نظرسنجی BRFSS2015 که متغیر هدف Diabetes_012 در آن سه کلاس دارد:

- 0: عدم ابتلا به دیابت و یا دیابت در دوران بارداری
- 1: پیش دیابت
- 2: دیابت

کلاس‌های این مجموعه داده متوازن نیستند. این داده شامل ۲۱ متغیر است.

۲) فایل diabetes_binary_5050split_health_indicators_BRFSS2015.csv: داده‌ی تمیز شده شامل ۷۰،۶۹۲ ردیف از جواب‌های داده شده به نظرسنجی BRFSS2015 که متغیر هدف Diabetes_binary در آن دو کلاس دارد:

- 0: عدم ابتلا به دیابت
- 1: پیش دیابت و یا دیابت

این داده متوازن بوده و به صورت ۵۰-۵۰ شامل کلاس‌های ۰ (عدم ابتلا به دیابت)، و ۱ (پیش دیابت و یا دیابت) است . این داده نیز دارای ۲۱ متغیر است.

۳) فایل diabetes_binary_health_indicators_BRFSS2015.csv: داده‌ی تمیز شده شامل ۲۵۳،۸۶۰ ردیف از جواب‌های داده شده به نظرسنجی BRFSS2015 می‌باشد. متغیر هدف Diabetes_binary است که دو کلاس دارد:

- 0: عدم ابتلا به دیابت
- 1: پیش دیابت و یا دیابت

این داده نیز دارای ۲۱ متغیر بوده و در متغیر هدف نامتوازن است.

لینک دانلود داده‌ها:

https://d-learn.ir/diabetes_012_health_indicators_brfss2015/

https://d-learn.ir/diabetes_binary_5050split_health_indicators_brfss2015/

https://d-learn.ir/diabetes_binary_health_indicators_brfss2015/

پرسش‌ها

با استفاده از داده‌های ارائه شده به پرسش‌های زیر پاسخ دهید:

۱. آیا نتایج نظرسنجی BRFSS شانسی برای اینکه پیش‌بینی قابل قبولی از اینکه یک فرد دیابت دارد یا نه، ارائه کند دارد؟ (یک بررسی توصیفی کفایت می‌کند، نیازی به مدلسازی نیست)
۲. کدام عوامل خطر بیشترین قدرت پیش‌بینی ابتلا به دیابت را دارند؟
۳. آیا می‌توانیم از زیرمجموعه‌ای از عوامل خطر برای پیش‌بینی اینکه یک فرد دیابت دارد یا نه، استفاده کنیم؟ به عبارت دیگر آیا می توانیم تنها از تعدادی از متغیرها در مدل استفاده کنیم و از بقیه صرف نظر کنیم؟ با چه روشی مدل نهایی را انتخاب می‌کنید؟
۴. کدام متغیرها در مدل نهایی شما نقش بازی می‌کنند و عملکرد مدل پیش‌بینی با استفاده از آن‌ها چگونه است؟ استراتژی شما برای استفاده از داده train و validation و test در ساخت مدل و گزارش عملکرد آن چیست؟
۵. آیا می‌توانیم برای ایجاد یک سامانه ارائه احتمال ابتلا به دیابت با انتخاب هوشمندانه ویژگی‌ها شکل کوتاه‌تری از پرسش‌ها ایجاد کنیم و تنها پرسیدن چند سوال با دقت قابل قبول پیش‌بینی کنیم که

آیا شخصی ممکن است دیابت داشته باشد یا در خطر بالای دیابت باشد؟ در نظر داشته باشید می‌خواهیم در این سامانه همه محاسبات سمت کاربر (در مرورگر) انجام شود و چیزی سمت سرور فرستاده نشود، چون در این صورت سامانه قابلیت پیاده‌سازی نخواهد داشت. برای پاسخ به این سوال کافیست مدل از منظر خواسته شده تحلیل کنید. نیازی به انجام محاسبه نیست.

مراجع

Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019;16:190109. http://dx.doi.org/10.5888/pcd16.190109

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset