آموزش مقدماتی علوم داده با پایتون

تابستان ۱۴۰۲

پروژه پایانی (۲)

مدلسازی برای طراحی سامانه

پیش‌بینی خطر ابتلا به دیابت

بیان مسئله

دیابت یکی از شایع‌ترین بیماری‌ها در دنیاست که هر ساله بر زندگی میلیون‌ها نفر 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. این وضعیت در نهایت منجر به کاهش کیفیت زندگی و کاهش امید به زندگی می‌شود.

پس از تجزیه غذاهای مختلف به قندها در طول فرآیند هضم، قند وارد خون می‌شود. این عمل باعث می‌شود که غده پانکراس انسولین آزاد کند تا سلول‌ها بتوانند از قند موجود در خون برای تولید انرژی استفاده کنند. دیابت به عدم تولید کافی انسولین یا ناتوانی بدن در استفاده از انسولین برای تبدیل قند خون به انرژی اتلاق می‌شود.

عوارضی همچون بیماری‌های قلبی، از دست دادن دید، قطع اندام‌های پایین در نتیجه‌ی بریدگی، بیماری‌های کلیوی و موارد دیگر همگی ناشی از سطوح بالای قند خون در افراد دیابتی است. برای دیابت هیچ درمان مشخصی وجود ندارد. با این حال راهکارهایی مانند از دست‌دادن وزن، تغذیه سالم، فعالیت بدنی و دریافت درمان‌های پزشکی در کاهش آسیب‌های بیماری و کنترل آن نقش مهمی ایفا می‌کنند. تشخیص زودهنگام می‌تواند منجر به تغییرات سبک زندگی مبتلایان و درمان‌های موثرتر شود. بنابراین مدل‌های پیش‌بینی خطر دیابت، ابزارهای بسیار مهمی برای ارتقای آگاهی جامعه و مسئولان بهداشت و سلامت عمومی در کشورها است.

بر اساس اعلام مرکز کنترل و پیشگیری بیماری‌های آمریکا (CDC) تا سال ۲۰۱۸، ۲/۳۴ میلیون آمریکایی دیابت دارند و ۸۸ 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.

برای این منظور می‌خواهیم سامانه‌ای ایجاد کنیم که اطلاعاتی درباره وضعیت کاربر را از وی دریافت کند و خطر ابتلا به دیابت را به وی گزارش کند. این سامانه به یک مدل برای پیش‌بینی دیابت نیاز دارد. در این پروژه می‌خواهیم این مدل را با استفاده از داده‌هایی که در اختیار داریم ایجاد کنیم.

داده‌ها

سیستم نظارت بر عوامل مخاطرات رفتاری (BRFSS) یک نظرسنجی تلفنی مربوط به سلامت است که سالانه توسط مرکز کنترل و پیشگیری از بیماری‌ها در آمریکا 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 ۲۰۱۵ در فایلی با فرمت csv در دسترس 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.

جدول زیر داده‌ی تمیز شده شامل ۲۱ ویژگی و ۲۵۳،۸۶۰ ردیف از پاسخ‌های افراد به نظرسنجی BRFSS در سال ۲۰۱۵ است. متغیر هدف (Diabetes_۰۱۲) در این داده از سه کلاس تشکیل شده است:

- عدم ابتلا به دیابت و یا دیابت در دوران بارداری (۰)
- پیش دیابت (۱)
- دیابت (۲)

این داده در لینک زیر در دسترس است:

d-learn.ir/diabetes_۰۱۲_health_indicators_brfss۲۰۱۵/

پرسش‌ها

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

۱. آیا نتایج نظرسنجی BRFSS شانسی برای ارائه یک پیش‌بینی قابل قبول از وضعیت فرد دارد؟ برای پاسخ به این سوال یک بررسی توصیفی کفایت می‌کند، نیازی به مدلسازی برای پاسخ به این سوال نیست.
۲. با توجه به پاسخی که به پرسش نخست دادید آیا تلفیق کلاس پیش‌دیابت و دیابت می‌تواند تصمیم معقولی باشد؟ اگر بله، این دو دسته را برای پاسخ به سوالات بعدی با یکدیگر ادغام کنید.
۳. آیا کلاس‌های متغیر هدف (ابتلا و عدم ابتلا به دیابت) به یکدیگر نزدیک است؟ عدم توازن کلاس‌های متغیر هدف چه تاثیری بر مدلسازی خواهد داشت؟ (ادغام دیابت و پیش‌دیابت می‌تواند به مسئله عدم توازن نیز کمک کند)
٤. برای حل مسئله عدم توازن کلاس‌های متغیر هدف چه راهکاری را در پیش می‌گیرید؟
۵. برای گزارش خطر ابتلا به دیابت یک مدل دسته‌بندی (classification) ایجاد کنید و عملکرد آن را با استفاده از داده‌های train و validation و test گزارش کنید.
٦. مهم‌ترین متغیرها در پیش‌بینی دیابت کدام متغیرها هستند؟

۷. کدام متغیرها در پیش‌بینی دیابت نقشی بازی نمی‌کنند و بهتر از است آن‌ها را از مدل نهایی حذف کنیم تا اطلاعات غیرضروری از کاربر دریافت نکنیم؟
۸. مدل خود را برای خودتان و چند نفر از دوستانتان تست کنید و خطر ابتلا به دیابت را به آن‌ها گزارش کنید. آیا پاسخ‌ها معقول است؟
۹. آیا مدل شما می‌تواند در مدت زمان معقولی (مثلا در کمتر یک ثانیه) پاسخ خود را ارائه کند؟ مدل خود را از نظر حجم پردازش مورد نیاز و در هر پیش‌بینی و پیچیدگی محاسباتی آن تحلیل کنید.

منابع داده و توضیحات بیشتر

Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type ۲Diabetes Using Machine Learning Techniques. Prev Chronic Dis ۲۰۱۹;۱۶:۱۹۰۱۰۹.

http://dx.doi.org/۱۰٫۵۸۸۸/pcd۱۶٫۱۹۰۱۰۹

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

برای دریافت اطلاعات بیشتر درباره متغیرهای پیش‌بینی (ویژگی‌ها) به منابع بالا مراجعه کنید.