بنام خدا

اپیدمیولوژی دیجیتال؛ کاربرد داده‌های بزرگ و هوش مصنوعی در پایش سلامت

مولفان:

عرفان شهیررودی

معصومه جمالی زاده

ساناز سلیمانی

زهرا پاشاپور

ناظر علمی:

فائزه تهرودی

انتشارات ارسطو

(سازمان چاپ و نشر ایران – ۱۴۰۵)

نسخه الکترونیکی این اثر در سایت سازمان چاپ و نشر ایران و اپلیکیشن کتاب رسان موجود می باشد

Chaponashr.ir

سرشناسه : شهیررودی ، عرفان ، ۱۳۷۵
عنوان و نام پدیدآورندگان : اپیدمیولوژی دیجیتال؛ کاربرد داده‌های بزرگ و هوش مصنوعی در پایش سلامت/ مولفان : عرفان شهیررودی ، معصومه جمالی زاده ، ساناز سلیمانی ، زهرا پاشاپور
مشخصات نشر : انتشارات ارسطو (سازمان چاپ و نشر ایران)، ۱۴۰۵.
مشخصات ظاهری : ۹۶ ص.
شابک : ۴-۴۹۵-۱۲۹-۶۲۲-۹۷۸
شناسه افزوده: جمالی زاده ، معصومه ، ۱۳۷۷
شناسه افزوده: سلیمانی ، ساناز ، ۱۳۷۷
شناسه افزوده: پاشاپور ، زهرا ، ۱۳۷۶
وضعیت فهرست نویسی : فیپا
یادداشت : کتابنامه.
موضوع : اپیدمیولوژی دیجیتال- کاربرد داده‌های بزرگ - هوش مصنوعی
رده بندی کنگره : TP ۹۸۳
رده بندی دیویی : ۶۶۸/۵۵
شماره کتابشناسی ملی :۸۵۶۳۲۱
اطلاعات رکورد کتابشناسی : فیپا

نام کتاب : اپیدمیولوژی دیجیتال؛ کاربرد داده‌های بزرگ و هوش مصنوعی در پایش سلامت
مولفان : عرفان شهیررودی- معصومه جمالی زاده- ساناز سلیمانی -زهرا پاشاپور
ناشر : انتشارات ارسطو (سازمان چاپ و نشر ایران)
صفحه آرایی و تنظیم:الهام غفاری
طرح جلد : پریا فرجام
ناظر علمی: فائزه تهرودی
تیراژ : ۱۰۰۰ جلد
نوبت چاپ : اول ــ ۱۴۰۵
چاپ : زبرجد
قیمت : ۱۶۰۰۰۰ تومان
فروش نسخه الکترونیکی - کتاب رسان :
https//:chaponashr.ir/ketabresan
شابک : ۴-۴۹۵-۱۲۹-۶۲۲-۹۷۸
تلفن مرکز پخش : ۰۹۱۲۰۲۳۹۲۵۵
www.chaponashr.ir

فهرست

فهرست

فصل اول:
سوگیری انتخاب، پوشش و خطای اندازه‌گیری در منابع داده‌های دیجیتال

مقدمه:

نخستین گام در هر تحلیل اپیدمیولوژیک، ارزیابی کیفیت و ماهیت داده‌های خام است. در حوزه اپیدمیولوژی دیجیتال، این گام اهمیت بیشتری پیدا می‌کند، چرا که داده‌ها نه در قالب پرسشنامه‌های استاندارد و با اهداف پژوهشی مشخص، بلکه به‌عنوان محصول جانبی تعاملات انسانی با فناوری تولید می‌شوند. پیش از آنکه بتوان از این داده‌ها برای پایش طغیان‌ها یا برآورد بار بیماری بهره برد، باید پرسید که «چه کسانی» در این داده‌های دیجیتال دیده می‌شوند و «چه چیزی» واقعاً اندازه‌گیری شده است. این فصل به سه منبع اصلی خطاهای سیستماتیک در این زمینه جدید می‌پردازد: سوگیری انتخاب[۱]، خطای اندازه‌گیری[۲] و خطا در فنوتایپ‌های الگوریتمی. کاربران رسانه‌های اجتماعی یا جستجوگران وب، نمونه‌ای تصادفی از جمعیت عمومی نیستند؛ در نتیجه، سوگیری پوشش[۳] و سوگیری برخوردگر ناشی از گزینش نمونه بر اساس یک متغیر مشترک، تهدیدی جدی برای تعمیم‌پذیری یافته‌هاست. همچنین، تبدیل رفتار دیجیتال (مانند یک جستجوی آنلاین) به یک متغیر اپیدمیولوژیک معنادار (مانند «مواجهه با ویروس») فرآیندی همراه با خطاست که می‌تواند از نوع طبقه‌بندی نادرست افتراقی[۴] یا غیرافتراقی[۵] باشد. هدف این فصل، آشنا کردن خواننده با ابزارهای مفهومی برای تشخیص و برآورد این سوگیری‌ها است.

---

۱ Selection Bias
۲ Measurement Error
۳ Coverage Bias
۴ Differential Misclassification
۵ Non-differential Misclassification

شکل ۱. نمای مفهومی منابع داده و پایش سلامت در اپیدمیولوژی دیجیتال

درس ۱: روایی و پایایی فنوتایپ‌های الگوریتمی در فقدان استاندارد طلایی بالینی

مهم‌ترین چالش در استفاده از منابع داده‌های دیجیتال برای اهداف اپیدمیولوژیک، تعریف و اعتبارسنجی «فنوتایپ‌های الگوریتمی»[۱] است. یک فنوتایپ الگوریتمی، تعریفی عملیاتی و محاسباتی از یک وضعیت سلامت، بیماری یا مشخصه بالینی است که با اعمال مجموعه‌ای از قواعد، کدها یا مدل‌های یادگیری ماشین بر روی داده‌های خام ساخت‌یافته یا غیرساخت‌یافته استخراج می‌شود. این فرآیند که در عمل یک عمل «اندازه‌گیری» در زمینه دیجیتال محسوب می‌شود، نیازمند ارزیابی دقیق ویژگی‌های عملکردی خود، یعنی روایی[۲] و پایایی[۳] است. بر اساس چارچوب‌های کلاسیک اپیدمیولوژی که توسط راثمن[۴] و گرینلند[۵] در کتاب «اپیدمیولوژی مدرن»[۶] تشریح شده، روایی یک ابزار اندازه‌گیری به معنای میزان تطابق مقدار اندازه‌گیری‌شده با مقدار واقعی کمیت مدنظر است. در زمینه فنوتایپ‌های الگوریتمی، روایی به توانایی الگوریتم در تمایز صحیح میان افرادی که واقعاً واجد وضعیت بالینی هدف هستند (موردهای واقعی) و افرادی که فاقد آن هستند (غیرموردهای واقعی) گفته می‌شود. پایایی نیز به درجه ثبات و تکرارپذیری نتایج هنگام اعمال مکرر الگوریتم بر روی داده‌های یکسان، توسط ارزیاب‌های مختلف، یا در نقاط زمانی متفاوت اشاره دارد.

---

[۱] Algorithmic Phenotypes
[۲] Validity
[۳] Reliability
[۴] Rothman
[۵] Greenland
[۶] Modern Epidemiology

در پژوهش‌های بالینی سنتی، ارزیابی روایی معمولاً با مقایسه نتایج ابزار در برابر یک «استاندارد طلایی»[۱] انجام می‌شود که بهترین روش در دسترس برای تعیین وضعیت واقعی بیماری در نظر گرفته می‌شود. برای مثال، برای ارزیابی روایی یک پرسشنامه غربالگری افسردگی، نتایج آن با مصاحبه بالینی ساخت‌یافته توسط روان‌پزشک (به‌عنوان استاندارد طلایی) مقایسه می‌شود. از این مقایسه، شاخص‌های عملکردی کلیدی شامل حساسیت[۲] یعنی نسبت موارد واقعی که به درستی توسط الگوریتم شناسایی شده‌اند، ویژگی[۳] یعنی نسبت غیرموردهای واقعی که به درستی رد شده‌اند، ارزش اخباری مثبت[۴] یعنی احتمال آنکه یک فرد با نتیجه مثبت در الگوریتم، واقعاً بیمار باشد، و ارزش اخباری منفی[۵] محاسبه می‌شوند. این معیارها، که در متون مرجع اپیدمیولوژی مانند کتاب «اپیدمیولوژی گوردیس»[۶] با جزئیات شرح داده شده‌اند، برای فهم میزان خطای طبقه‌بندی و پیامدهای آن بر تخمین‌های اپیدمیولوژیک ضروری هستند.

با این حال، کاربرد این چارچوب در اپیدمیولوژی دیجیتال با یک مانع اساسی روبه‌روست: در انبوه داده‌های تولیدشده توسط کاربران در رسانه‌های اجتماعی، جستجوهای وب، یا حتی بخش‌هایی از سوابق پزشکی الکترونیک[۷]، یک استاندارد طلایی معتبر، در دسترس، یا حتی قابل تعریف برای بسیاری از فنوتایپ‌های مدنظر وجود ندارد. برای مثال، اگر هدف، شناسایی موارد «اضطراب مرتبط با پاندمی» از میان پست‌های توییتر باشد، هیچ روش قطعی و مرجع بی‌خطایی

---

[۱] Gold Standard
[۲] Sensitivity
[۳] Specificity
[۴] Positive Predictive Value - PPV
[۵] Negative Predictive Value - NPV
[۶] Gordis Epidemiology
[۷] Electronic Health Records - EHRs

برای تأیید اینکه یک توییت خاص واقعاً بازتاب‌دهنده اضطراب بالینی است و نه یک ابراز هیجانی گذرا، در دست نیست. به‌طور مشابه، شناسایی یک «فنوتایپ شبه-آنفلوانزایی»[۱] از روی جستجوهای گوگل، فاقد تأیید آزمایشگاهی یا تشخیص پزشک به‌عنوان یک استاندارد طلایی خارجی است. این وضعیت، همان‌طور که در کارهای پاول[۲] و همکاران در مجله «اپیدمیولوژی بالینی»[۳] مطرح شده، به «مسئله فقدان استاندارد طلایی»[۴] منجر می‌شود. در غیاب یک مرجع قطعی، برآوردهای ساده‌انگارانه حساسیت و ویژگی می‌توانند به‌شدت گمراه‌کننده باشند؛ چرا که اگر استاندارد مرجع خود ناقص باشد، عملکرد ابزار تحت ارزیابی به اشتباه، پایین‌تر یا بالاتر از مقدار واقعی آن برآورد خواهد شد.

برای برخورد با این چالش، روش‌شناسان اپیدمیولوژی دیجیتال راهبردهای متعددی را توسعه داده‌اند. یکی از رایج‌ترین رویکردها، ایجاد یک «استاندارد مرجع تقریبی»[۵] از طریق حاشیه‌نویسی[۶] دستی توسط متخصصان انسانی است. در این فرآیند، نمونه‌ای تصادفی یا هدفمند از واحدهای دیجیتال (مانند توییت‌ها، پست‌ها، یا قطعات متن بالینی) انتخاب شده و توسط دو یا چند ارزیاب آموزش‌دیده بر اساس یک دستورالعمل کدگذاری[۷] استاندارد، برچسب‌گذاری می‌شود. پایایی بین

---

۱ Influenza-like Illness - ILI Phenotype
۲ Pavlou
۳ Clinical Epidemiology
۴ Imperfect Gold Standard Problem
۵ Approximate Reference Standard
۶ Annotation
۷ Codebook

ارزیاب‌ها[۱] که معمولاً با آماره کاپای کوهن[۲] یا کاپای فلیس[۳] سنجیده می‌شود، خود به‌عنوان پیش‌نیازی برای روایی عمل می‌کند؛ چرا که پایایی پایین نشان‌دهنده آن است که فنوتایپ هدف به قدر کافی واضح تعریف نشده است. با این وجود، این رویکرد پرزحمت، زمان‌بر و گران است و مقیاس‌پذیری محدودی دارد. علاوه بر این، خود حاشیه‌نویسی انسانی نیز بی‌خطا نیست و ممکن است تحت تأثیر خطاهای سیستماتیک خاص خود، مانند خستگی یا تفسیرهای شخصی، قرار گیرد.

رویکرد پیشرفته‌تر برای اعتبارسنجی در فقدان استاندارد طلایی، استفاده از مدل‌های آماری است که عدم قطعیت در وضعیت واقعی بیماری را به‌طور مستقیم مدل‌سازی می‌کنند. شاخص‌ترین این روش‌ها، «مدل‌های کلاس پنهان»[۴] است که ریشه در کارهای گودمن[۵] و هابرمن[۶] در دهه ۱۹۷۰ دارد. در این چارچوب، وضعیت واقعی و مشاهده‌نشده (پنهان) بیماری به‌عنوان یک متغیر طبقه‌بندی‌شده در نظر گرفته می‌شود که تنها از طریق چندین شاخص اندازه‌گیری ناقص (مانند نتیجه یک الگوریتم، نظر یک پزشک، و یک نشانگر زیستی غیرقطعی) خود را نشان می‌دهد. مدل کلاس پنهان با فرض استقلال شرطی[۷] خطاهای اندازه‌گیری این شاخص‌ها در هر کلاس پنهان، به‌طور هم‌زمان شیوع کلاس‌های پنهان و میزان حساسیت و ویژگی

---

۱ Inter-rater Reliability
۲ Cohen’s Kappa
۳ Fleiss’ Kappa
۴ Latent Class Models - LCMs
۵ Goodman
۶ Haberman
۷ Conditional Independence

هر یک از شاخص‌ها را برآورد می‌کند. برای نمونه، در مطالعه‌ای که توسط کولین[۱] و همکاران برای اعتبارسنجی فنوتایپ الگوریتمی آرتریت روماتوئید با استفاده از داده‌های سوابق الکترونیک سلامت انجام شد، از یک مدل کلاس پنهان بیزی[۲] استفاده شد که در آن، خروجی الگوریتم، کدهای تشخیصی بین‌المللی بیماری‌ها[۳] و تجویز داروهای خاص به‌عنوان شاخص‌های ناقص در نظر گرفته شدند و توزیع‌های پیشین[۴] آگاهی‌بخش برای پارامترهای نامعلوم از مطالعات قبلی استخراج شد. این رویکرد بیزی که در کتاب «آمار بیزی کاربردی در پزشکی»[۵] شرح داده شده، امکان ترکیب عدم قطعیت و دانش پیشین را فراهم می‌کند و برآوردهایی واقع‌بینانه‌تر از روایی یک فنوتایپ الگوریتمی در غیاب یک استاندارد طلایی کامل به‌دست می‌دهد. به این ترتیب، پژوهشگر می‌تواند به جای یک پاسخ قطعی «بله یا خیر» در مورد روایی، یک توزیع احتمال برای حساسیت و ویژگی به‌دست آورد که بازتاب‌دهنده عدم قطعیت موجود در کل فرآیند اندازه‌گیری دیجیتال است.

---

[۱] Kohlin
[۲] Bayesian Latent Class Model
[۳] ICD codes
[۴] Prior Distributions
[۵] Applied Bayesian Statistics in Medicine

درس ۲: سوگیری برخوردگر و سوگیری بازماندگی در کوهورت‌های دیجیتال مبتنی بر رسانه‌های اجتماعی

مطالعات کوهورت[۱] که در آن‌ها گروهی از افراد در طول زمان دنبال می‌شوند تا بروز پیامدهای سلامت در میان زیرگروه‌های با مواجهه متفاوت مقایسه شود، از بخش‌های اصلی پژوهش‌های اپیدمیولوژیک به حساب می‌آیند. در اپیدمیولوژی دیجیتال، مفهوم «کوهورت دیجیتال» به مجموعه‌ای از کاربران گفته می‌شود که از طریق ردپاهای دیجیتالشان در یک یا چند پلتفرم آنلاین، مانند رسانه‌های اجتماعی، تالارهای گفتگو یا اپلیکیشن‌های سلامت همراه، شناسایی و پیگیری می‌شوند. این کوهورت‌ها، برخلاف کوهورت‌های سنتی که با نمونه‌گیری هدفمند از یک جمعیت مبدأ مشخص شکل می‌گیرند، به‌صورت فرصت‌طلبانه و بر اساس حضور دیجیتالی افراد شکل می‌گیرند. این شیوه شکل‌گیری، زمینه را برای بروز دو نوع خطای سیستماتیک به هم پیوسته فراهم می‌کند: سوگیری انتخاب ناشی از برخوردگر، و سوگیری بازماندگی که به آن پارادوکس برکسن در زمینه دیجیتال نیز گفته می‌شود. درک این مفاهیم برای ارزیابی اعتبار مطالعاتی که از داده‌های شبکه‌های اجتماعی برای استنتاج روابط علی یا حتی توصیفی استفاده می‌کنند، حیاتی است.

سوگیری برخوردگر[۲] زمانی رخ می‌دهد که یک عامل مشترک (متغیر ثالث) که خود معلول دو متغیر دیگر (مواجهه و پیامد، یا علل آن‌ها) است، به نحوی در طراحی یا تحلیل مطالعه شرطی‌سازی یا کنترل شود. این شرطی‌سازی می‌تواند از طریق انتخاب نمونه بر اساس آن عامل،

---

[۱] Cohort Studies

[۲] Collider Bias

طبقه‌بندی در تحلیل، یا تعدیل در مدل رگرسیونی صورت پذیرد. نتیجه این عمل، ایجاد یک ارتباط آماری ساختگی میان دو متغیری است که در جمعیت اصلی هیچ ارتباطی با یکدیگر ندارند، یا تحریف یک ارتباط موجود. در کتاب راهنمای اپیدمیولوژی مدرن، هرنان[۱] و روبینز[۲] با استفاده از نمودارهای علّی غیرچرخه‌ای[۳] این پدیده را به‌طور کامل تشریح کرده‌اند. در یک نمودار علّی، یک برخوردگر گرهی است که دو یا چند مسیر علّی به آن وارد می‌شوند. شرطی‌سازی بر روی آن، مسیری را که به‌طور طبیعی مسدود است، باز می‌کند و اجازه می‌دهد اطلاعات میان علل آن جریان یابد. در زمینه مطالعات دیجیتال، مهم‌ترین و رایج‌ترین برخوردگر، خودِ «حضور در پلتفرم» یا «ورود به مطالعه» است.

برای یک مطالعه کوهورت دیجیتال که شرکت‌کنندگان را از یک رسانه اجتماعی خاص مانند فیسبوک یا توییتر جذب می‌کند، ورود به مطالعه تابعی از دو عامل اصلی است: داشتن دسترسی به اینترنت و سواد دیجیتال از یک سو (که خود با وضعیت اقتصادی-اجتماعی، سن و محل زندگی مرتبط است) و از سوی دیگر، داشتن دلیلی برای فعالیت در آن پلتفرم خاص که می‌تواند شامل وضعیت سلامت فرد (مواجهه یا پیامد مدنظر) یا علاقه به موضوعات سلامت باشد. این وضعیت یک برخوردگر کلاسیک ایجاد می‌کند: اگر هم مواجهه (مثلاً یک وضعیت شغلی پراسترس) و هم پیامد (مثلاً اضطراب) احتمال حضور فعال و جذب شدن در یک مطالعه درباره سلامت روان در اینستاگرام را افزایش دهند، آنگاه مطالعه تنها بر روی افرادی انجام می‌شود که این برخوردگر (حضور در مطالعه) در آن‌ها رخ داده است. نتیجه این انتخاب، ایجاد یا تقویت کاذب ارتباط منفی

---

[۱] Hernán

[۲] Robins

[۳] Directed Acyclic Graphs - DAGs