

موضوع: پیش‌بینی ترک سیگار در بزرگسالان با استفاده از داده‌های BRFSS: تحلیل نقش عوامل رفتاری، تاریخچه مصرف، و مشخصه‌های فردی-جغرافیایی

اعضا گروه: آرمین خوجوی - حامد حسامی - مهدیه ابراهیمی - صادق اکبریان

سؤال اصلی پژوهش: با چه دقتی می‌توان ترک سیگار در موج بعدی (یا دوره زمانی مشخص پس از خط مبنا) را در میان بزرگسالان با استفاده از داده‌های BRFSS پیش‌بینی کرد؟ سهم نسبی عوامل رفتاری (مانند تعداد سیگار در روز)، تاریخچه مصرف (مانند سن شروع، مدت مصرف)، مشخصه‌های فردی مانند BMI، سلامت روان و مشخصه‌های جغرافیایی (مانند محل سکونت شهری/روستایی) در این پیش‌بینی چقدر است؟

- **فرضیه ۱:** استعمال کمتر سیگار در ۳۰ روز پیش از خط مبنا، سن شروع بالاتر و سابقه کوتاه‌تر مصرف سیگار، شانس ترک موفق در موج بعدی را افزایش می‌دهد.
- **فرضیه ۲:** وضعیت وزنی بهتر (مثلاً BMI در محدوده نرمال یا اضافه‌وزن در مقایسه با کم‌وزنی یا چاقی مفرط) و شاخص‌های بهتر سلامت روان، با احتمال بالاتر ترک موفق سیگار مرتبط هستند
- **فرضیه ۳:** ارتباط منفی بین مشکلات سلامت روان و ترک موفق سیگار، در مناطق روستایی و کم‌جمعیت شدیدتر از مناطق شهری است (بررسی اثر تعاملی بین سلامت روان و جغرافیا).

طرح مطالعه: مطالعه طولی (Longitudinal) با استفاده از داده‌های موجود در BRFSS (با فرض امکان اتصال داده‌های افراد در طول زمان یا استفاده از متغیرهایی که وضعیت ترک را در موج بعدی نشان می‌دهند). یا مطالعه مقطعی (Cross-sectional) در صورتی که متغیر هدف (ترک سیگار) و متغیرهای پیش‌بین هم‌زمان اندازه‌گیری شده باشند و هدف تحلیل ارتباط باشد

جامعه و نمونه آماری: بزرگسالان (۱۸ سال و بالاتر) شرکت‌کننده در پیمایش BRFSS که در موج پایه به‌عنوان "سیگاری فعلی (Current)" - Smoker تعریف دقیق بر اساس متغیر SMOKER3 (یا مشابه) شناسایی شده‌اند و داده‌های مربوط به وضعیت سیگار کشیدن آن‌ها در موج پیگیری (Follow-up) موجود است. معیارهای ورود و خروج نمونه به‌دقت تعریف می‌شود. تحلیل‌ها با استفاده از وزن‌های نمونه‌گیری (BRFSS مانند LLCPTW) انجام خواهد شد تا قابلیت تعمیم نتایج به جمعیت بزرگسال ایالات متحده افزایش یابد.

• متغیرها:

○ **متغیر وابسته (Outcome):** ترک موفق سیگار در موج پیگیری (به‌صورت باینری: بله/خیر). تعریف دقیق بر اساس ترکیبی از متغیرها مانند SMOKER3 (در موج پیگیری) مقدار "Never" یا "Former" و احتمالاً متغیرهایی مانند STOPSMK2 (مدت زمان ترک) یا پرسش مستقیم در مورد ترک کامل (مشابه رویکرد).

○ **متغیرهای مستقل (Predictors):**

▪ **عوامل رفتاری (H1):** تعداد روزهای سیگار کشیدن در ۳۰ روز گذشته SMKDAY2

▪ **تاریخچه مصرف (H1):**

▪ **مشخصه‌های فردی (H2, H3):**

▪ **BMI:** متغیر محاسبه‌شده BMI5_

- سلامت روان: تعداد روزهای با سلامت روان نامناسب در ۳۰ روز گذشته (MENTHLTH)
- دموگرافیک: گروه سنی, (`_AGE_g`) جنسیت, نژاد/قومیت, (`_IMPRACE`) سطح تحصیلات, (`_EDUCAG`) وضعیت تأهل, (`MARITAL`)
- مشخصه‌های جغرافیایی: (`H3`) وضعیت سکونت شهری/روستایی, (`_METSTAT`) ایالت (`_STATE`) در صورت نیاز به تحلیل‌های منطقه‌ای
- متغیرهای کمکی: (`Covariates`) وضعیت اشتغال, (`EMPLOY1`) سطح درآمد, (`_INCOMG`) وضعیت دسترسی به پزشک, (`PERSDOC2`) آخرین معاینه پزشکی, (`CHECKUP1`)

روش‌ها:

- تحلیل آماری و یادگیری ماشین:
 - آمار توصیفی: برای شرح ویژگی‌های نمونه مطالعه.
 - مدل‌سازی پیش‌بینی:
 - استفاده از الگوریتم‌های طبقه‌بندی (Classification Algorithms) مانند:
 - رگرسیون لجستیک (برای بررسی ارتباط و پیش‌بینی اولیه و آزمون فرضیه‌ها).
 - ماشین بردار پشتیبان (SVM)
 - درخت تصمیم (Decision Tree)
 - جنگل تصادفی (Random Forest)
 - گرادیان بوستینگ مانند XGBoost, LightGBM.
 - ارزیابی مدل: محاسبه معیارهای دقت (Accuracy), سطح زیر منحنی ROC (AUC), دقت (Precision), بازیابی (Recall) و امتیاز F1 با استفاده از روش‌هایی مانند اعتبارسنجی متقابل (Cross-Validation).
 - تحلیل سهم نسبی عوامل:
 - استفاده از شاخص‌های اهمیت ویژگی (Feature Importance) در مدل‌های درختی مانند Random Forest, XGBoost
 - مقایسه عملکرد مدل‌ها با و بدون گروه‌های مختلف متغیرها (مثلاً مدل فقط با عوامل رفتاری در مقابل مدل کامل).
 - تحلیل ضرایب در مدل رگرسیون لجستیک
 - آزمون فرضیه‌ها:

▪ **فرضیه ۱ و ۲:** بررسی ضرایب و معناداری آماری متغیرهای مربوطه در مدل رگرسیون لجستیک یا تحلیل اهمیت ویژگی‌ها.

▪ **فرضیه ۳:** وارد کردن جمله تعاملی (Interaction Term) بین شاخص سلامت روان و متغیر محل سکونت (روستایی/شهری) در مدل رگرسیون لجستیک و بررسی معناداری آن.

▪ پردازش داده‌ها :

- نحوه تعریف عملیاتی متغیر وابسته (ترک سیگار).
- نحوه مدیریت داده‌های گمشده (Missing Data) مثلاً حذف لیستی، جایگزینی چندگانه
- نحوه تبدیل متغیرها در صورت نیاز (مثلاً تبدیل متغیرهای پیوسته به طبقه‌ای، ساخت متغیرهای ترکیبی).
- نحوه اتصال داده‌ها بین موج‌های مختلف (در صورت مطالعه طولی).
- در نظر گرفتن وزن‌های نمونه‌گیری (Survey Weights) در تحلیل‌ها برای تعمیم‌پذیری نتایج به جمعیت هدف.

نتایج مورد انتظار:

- انتظار داریم که مدل‌های یادگیری ماشین (به‌خصوص مدل‌های آنسامبل مانند Random Forest یا XGBoost) به دقتی قابل قبول (مثلاً AUC بالاتر از ۰.۷ یا ۰.۷۵) در پیش‌بینی ترک سیگار دست یابند.
- پیش‌بینی می‌شود که مطابق فرضیه ۱، عوامل مرتبط با تاریخچه و شدت مصرف (تعداد کمتر سیگار، سن شروع بالاتر، دوره مصرف کوتاه‌تر) جزو پیش‌بین‌های قوی ترک موفق باشند.
- انتظار می‌رود طبق فرضیه ۲، BMI در محدوده نرمال/اضافه‌وزن (نه کم‌وزنی یا چاقی مفرط) و سلامت روان بهتر، با شانس بالاتر ترک مرتبط باشند.
- پیش‌بینی می‌شود تحلیل تعاملی، فرضیه ۳ را تأیید کند و نشان دهد که اثر منفی سلامت روان ضعیف بر ترک سیگار، در مناطق روستایی بارزتر است.
- انتظار می‌رود تحلیل اهمیت ویژگی‌ها، سهم نسبی هر گروه از عوامل (رفتاری، تاریخچه، فردی، جغرافیایی) را مشخص کند و احتمالاً عوامل رفتاری و تاریخچه مصرف سهم بیشتری نسبت به مشخصه‌های دموگرافیک پایه داشته باشند.
- **اهمیت نتایج:** این یافته‌ها می‌تواند به طراحی مداخلات هدفمندتر برای ترک سیگار کمک کند. شناسایی افراد در معرض خطر بالاتر عدم موفقیت در ترک و درک نقش عوامل زمینه‌ای مانند جغرافیا و سلامت روان، برای برنامه‌ریزی‌های بهداشت عمومی ارزشمند خواهد بود.

اهمیت و نوآوری پژوهش:

- استفاده از یک دیتاست بزرگ و معتبر ملی (BRFSS).

- به‌کارگیری روش‌های پیشرفته یادگیری ماشین برای پیش‌بینی دقیق‌تر ترک سیگار.
- تحلیل جامع سهم نسبی گروه‌های مختلف عوامل (فراتر از بررسی تک‌تک متغیرها).
- بررسی نقش تعدیل‌گر جغرافیا در ارتباط بین سلامت روان و ترک سیگار (آزمون فرضیه ۳).