

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

زهرا صداقت طالبی

ارائه دهنده

جناب آقای دکتر مظاهری

استاد راهنما

اردیبهشت  
۱۴۰۵

# Comparison of Feature Selection Methods in Machine Learning Models of Cancer Information Seeking Among United States Adults: Cross-Sectional Study

مقایسه روش های انتخاب ویژگی در مدل های  
یادگیری ماشین جستجوی اطلاعات سرطان در  
میان بزرگسالان ایالات متحده: مطالعه مقطعی

JMIR MEDICAL INFORMATICS

Name

Pubmed, Scopus, DOAJ

Indexing

5.0

Impact factor

Q1

Category

# مشخصات ژورنال

## فهرست اختصارات

| Abbreviation | full                                                   | ترجمه                                    |
|--------------|--------------------------------------------------------|------------------------------------------|
| ML           | Machine Learning                                       | یادگیری ماشین                            |
| LASSO        | Least Absolute Shrinkage and Selection Operator        | عملگر کمترین قدرمطلق انقباض و انتخاب     |
| PCA          | Principal Component Analysis                           | تحلیل مولفه های اصلی                     |
| HINTS        | Health Information National Trends Survey              | پیمایش ملی روندها در اطلاعات سلامت       |
| SVM          | Support Vector Machine                                 | ماشین بردار پشتیبان                      |
| XGBoost      | Extreme Gradient Boosting                              | تقویت گرادیان تصاعدی                     |
| KNN          | K-Nearest Neighbor                                     | نزدیک ترین همسایه                        |
| AUC          | Area Under the receiver operating Characteristic curve | مساحت زیر منحنی مشخصه بهره برداری گیرنده |

# فهرست مطالب



- ۰۱ مقدمه
- ۰۲ هدف مطالعه
- ۰۳ روش اجرا
- ۰۴ یافته ها
- ۰۵ نتیجه گیری
- ۰۶ دیدگاه من



# مقدمه

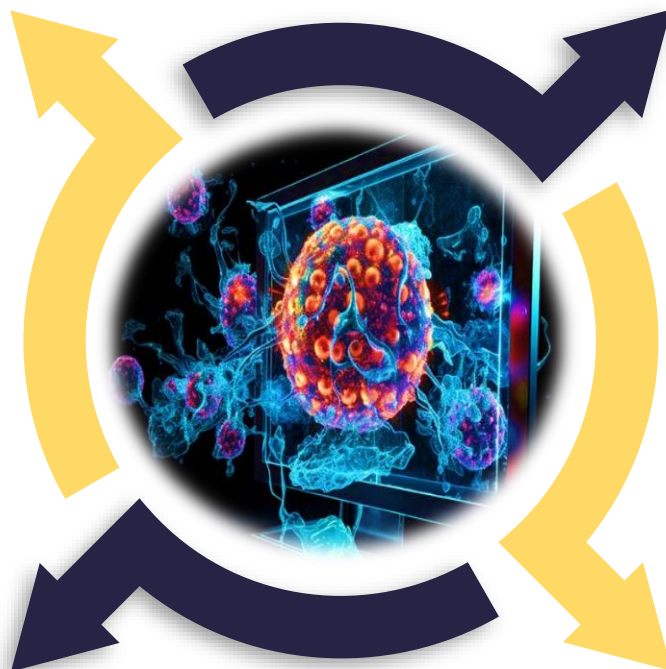
۶/۳۰

## اهمیت موضوع

سرطان یکی از  
بزرگترین چالش های  
سلامت عمومی در  
جهان است

## نقش ML

روش های یادگیری ماشین  
میتوانند داده های با ابعاد  
بالا را مدیریت کرده و  
همچنین داده هایی مفید  
تر در اختیارمان قرار دهند



## نقش فناوری

با پیشرفت علم و  
فناوری، افراد برای اهداف  
مختلف از منابع گوناگون به  
دنبال اطلاعات سلامت  
میگردند

## نقش اطلاعات

بیماران اغلب به طور فعال  
به دنبال اطلاعات مربوط به  
سرطان هستند تا  
تشخیص، گزینه های درمانی  
و مراقبت های خانگی خود  
را بهتر درک کنند

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

# عوامل تاثیر گذار:

| وضعیت سلامت                                                                            | سبک زندگی                                                                        | عوامل اجتماعی-جمعیتی                                                                                                                                                                                    |
|----------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>• بیماریهای مزمن، مانند سرطان و اضطراب</li></ul> | <ul style="list-style-type: none"><li>• مصرف الکل</li><li>• مصرف سیگار</li></ul> | <ul style="list-style-type: none"><li>• جنسیت: زن</li><li>• سن: ۵۵-۶۴</li><li>• وضعیت تأهل: متأهل</li><li>• سطح تحصیلات:</li><li>تحصیلات بالاتر</li><li>• هویت نژادی:</li><li>افراد سیاه پوست</li></ul> |

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من



# انتخاب ویژگی

چندین روش انتخاب ویژگی در  
یادگیری ماشین استفاده شده  
است، مانند روش LASSO و  
الگوریتم Boruta

انتخاب ویژگی یک گام حیاتی در  
یادگیری ماشین است، نه تنها  
برای کاهش ابعاد داده هاست، بلکه  
برای آشکار کردن مرتبط ترین  
ویژگی ها بدون از دست دادن  
اطلاعات زیاد

PCA هم برای کاهش ابعاد داده  
استفاده میشود، اما چون همه  
متغیرها در هر عامل باقی میماند،  
یک روش انتخاب ویژگی نیست

۰۱ مقدمه

۰۲ هدف مطالعه

۰۳ روش اجرا

۰۴ یافته ها

۰۵ نتیجه گیری

۰۶ دیدگاه من



# هدف مطالعه

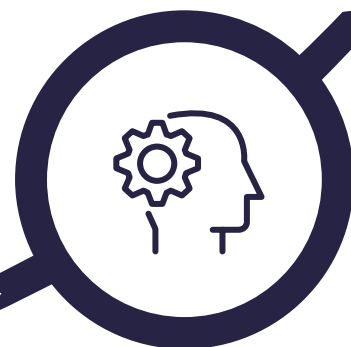
10/30

## اهداف مطالعه

مقایسه روش های انتخاب  
ویژگی - Boruta.LASSO  
ترکیبی از Boruta و LASSO  
و روش های مبتنی بر PCA



توسعه ابزارهای یادگیری  
ماشین برای پیش بینی  
جستجوی اطلاعات سرطان در  
میان بزرگسالان آمریکایی



شناسایی و پیش بینی  
عوامل موثر بر رفتار  
جستجوی اطلاعات  
سرطان



۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من



مرکز آموزش عالی  
علوم پزشکی  
وارستگی



# روش اجرا

۱۲/۳۰

# معرفی داده ها و جامعه آماری

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

✓ منبع داده: (HINTS6)

✓ دوره جمع آوری داده:

۷ مارس ۲۰۲۲ تا ۸ نوامبر ۲۰۲۲

✓ جامعه هدف: افراد ۱۸ سال به بالا در

ایالات متحده

✓ حجم اولیه نمونه: ۶۲۵۲ نفر

✓ حجم نمونه نهایی پس از پردازش:

۵۵۰۵ نفر

✓ نرخ پاسخ خانوار: ۲۸.۱٪

# متغیرها و پردازش داد ها

- ✓ **متغیر پیامد:** دسته بندی افراد به "جستجوگر اطلاعات سرطان" (پاسخ بله به سوال آیا تا کنون دنبال اطلاعات سرطان بوده اید؟) و "غیر جستجوگر" (پاسخ خیر)
- ✓ **تعداد متغیر های اولیه:** ۸۷ (شامل: عوامل جمعیتی، مصرف الکل و دخانیات، سابقه بیماری ها، باورها، شبکه های اجتماعی و...)
- ✓ **حذف متغیر های با بیش از ۱۰ درصد داده ناموجود**
- ✓ **تعداد متغیر های نهایی:** ۷۴
- ✓ **تقسیم داده:** ۷۰٪ برای آموزش و ۳۰٪ برای آزمون

مقدمه

هدف مطالعه

روش اجرا

یافته ها

نتیجه گیری

دیدگاه من

**Table 1.** Prevalence of cancer information seeking across demographic factors.

| Variable                      | Total, n | Seeking, n | Prevalence, % <sup>a</sup> | P value |
|-------------------------------|----------|------------|----------------------------|---------|
| <b>Gender</b>                 |          |            |                            |         |
| Male                          | 2189     | 949        | 43.4                       | <.001   |
| Female                        | 3316     | 1681       | 50.7                       |         |
| <b>Age group</b>              |          |            |                            |         |
| 18-49 years                   | 1935     | 895        | 46.3                       | .01     |
| 50-64 years                   | 1607     | 791        | 49.2                       |         |
| 65-74 years                   | 1234     | 623        | 50.5                       |         |
| >75 years                     | 729      | 321        | 44.0                       |         |
| <b>Race</b>                   |          |            |                            |         |
| Non-Hispanic White            | 3175     | 1738       | 54.7                       | <.001   |
| Non-Hispanic African American | 878      | 314        | 35.8                       |         |
| Hispanic                      | 984      | 371        | 37.7                       |         |
| Non-Hispanic Asian            | 286      | 120        | 42.0                       |         |
| Other                         | 182      | 87         | 47.8                       |         |
| <b>Education<sup>c</sup></b>  |          |            |                            |         |
| Less than high school         | 1300     | 375        | 28.8                       | <.001   |
| Some college                  | 1574     | 721        | 45.8                       |         |
| Bachelor's degree             | 1550     | 843        | 54.4                       |         |
| Postbaccalaureate degree      | 1070     | 687        | 64.2                       |         |
| <b>Income (US \$)</b>         |          |            |                            |         |
| <19,999                       | 904      | 271        | 30.0                       | <.001   |
| 20,000-49,999                 | 1431     | 569        | 39.8                       |         |
| 50,000-74,999                 | 946      | 488        | 51.6                       |         |
| >75,000                       | 2224     | 1302       | 58.5                       |         |
| Overall                       | 5505     | 2630       | 47.8                       |         |

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

# روش های انتخاب ویژگی

✓ LASSO:

کاهش ابعاد و انتخاب ویژگی با جریمه.

✓ PCA:

کاهش ابعاد با تبدیل متغیر های هم بسته به مولفه های مستقل.

✓ Boruta:

الگوریتم Boruta با استفاده از مدل

random forest ویژگی های مهم را

به صورت خودکار انتخاب میکند.

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من



# الگوریتم های یادگیری ماشین مورد استفاده



۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

# تحلیل آماری نهایی

- ✓ استفاده از آزمون کای دو برای بررسی ارتباط متغیرها
- ✓ فاصله اطمینان ۹۵٪ برای عوامل مرتبط با جستجوی اطلاعات سرطان
- ✓ نرم افزار SPSS
- ✓ برای ارزیابی و مقایسه عملکرد مدل ها از معیار AUC و آزمون DeLong استفاده شد
- ✓ برای تخمین نسبت شانس از رگرسیون لجستیک مرحله ای استفاده شد

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من



# یافته ها

19/۳۰

# آمار توصیفی

➤ از بین ۵۵۰۵ پاسخ دهنده، ۸،۴۷٪ اعلام کردند  
➤ تفاوت های معنادار بین جستجوگران و  
غیر جستجوگران:

- سن
- جنسیت
- تحصیلات
- درآمد
- متغیرهای مرتبط با سلامت



۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

# نتایج انتخاب ویژگی

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

در مجموع، ۲۷ متغیر  
بین این دو روش  
مشترک بودند

روش LASSO  
نیز ۴۲ متغیر را  
انتخاب کرد

الگوریتم Boruta  
تعداد ۴۳ متغیر مهم را  
انتخاب کرد

در روش PCA، مولفه هایی نگه داشته شدند که  
بیشترین واریانس داده ها را توضیح میدادند.

# عملکرد مدل ها

- مدل Random Forest با استفاده از ویژگی های انتخاب شده توسط LASSO، بالاترین مقدار AUC و همچنین دومین دقت بالا را به دست آورد.
- بیشترین دقت توسط مدل های SVM و رگرسیون لجستیک بدست آمد.
- مدل Random Forest در روش های مختلف انتخاب ویژگی، عملکرد پایداری نشان داد.

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

# مهم ترین متغیرها

رگرسیون لجستیک مرحله ای، مهم ترین متغیرهای تاثیرگذار را شناسایی کرد که عبارت اند از:

- ❖ آگاهی از تست های ژنتیکی
- ❖ مشاهده ویدیوهای مرتبط با سلامت در شبکه های اجتماعی
- ❖ علاقه به غربالگری سرطان
- ❖ دسترسی به پرونده پزشکی آنلاین

- ❖ سابقه شخصی ابتلا به سرطان
- ❖ نگرانی درباره ابتلا به سرطان
- ❖ سابقه خانوادگی سرطان
- ❖ درآمد بالاتر
- ❖ آگاهی از ویروس HPV

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من



# نتیجه گیری

۲۴/۳۰



# بحث و تحلیل

مقدمه

۰۱

هدف مطالعه

۰۲

روش اجرا

۰۳

یافته ها

۰۴

نتیجه گیری

۰۵

دیدگاه من

۰۶

مدل Random Forest در روش های  
مختلف انتخاب ویژگی عملکرد بهتری نشان  
داد.



نتایج نشان داد که روش های Boruta و  
LASSO متغیرهای مرتبط را به درستی  
شناسایی کردند و نسبت به PCA عملکرد  
بهتری داشتند.



## چه افرادی بیشتر به دنبال اطلاعات سرطان هستند؟؟؟



۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من

# محدودیت ها

LIMITATION



داده ها  
مقطعی بودند

داده ها  
براساس  
خوداظهاری  
افراد بودند

نرخ پاسخ  
دهی نسبتاً  
پایین بود

۰۱

مقدمه

۰۲

هدف مطالعه

۰۳

روش اجرا

۰۴

یافته ها

۰۵

نتیجه گیری

۰۶

دیدگاه من



# دیدگاه من

۲۸/۳۰



افزایش نرخ پاسخ دهی با استفاده از مدل های  
Gamification

استفاده از مدل های Boruta و LASSO بصورت  
ترکیبی

انجام مطالعه در چند کشور با ویژگی های متفاوت

مقدمه

۰۱

هدف مطالعه

۰۲

روش اجرا

۰۳

یافته ها

۰۴

نتیجه گیری

۰۵

دیدگاه من

۰۶

سپاس از توجه شما