به نام خدا

فرم پیشنهاد تحقیق:

بررسی ریسک ابتلا به سرطان ریه از طریق تجزیه و تحلیل عوامل خطر

اعضای گروه : مبینا شهبازی، مهدیس سپهوند

استاد: دکتر ابوالفضل ولی‌زاده

پاییز ۱۴۰۳

سوال اولیه:

میزان تاثیر عوامل خطر مختلف بر روی ریسک ابتلا به سرطان ریه چگونه است؟

اهمیت موضوع:

سرطان ریه خطرناک ترین و کشنده ترین نوع سرطان است. طبق گزارشی که در مقاله[2] آمده است هر ساله 85 نفر از 100 نفر به دلیل سرطان ریه جان خود را از دست می دهند. سرطان ریه یک مسئله مهم بهداشت جهانی است که عوامل خطر مختلفی بر توسعه آن تأثیر می گذارد. درک این عوامل برای راهبردهای پیشگیری و تشخیص زودهنگام بسیار مهم است. اگر سرطان ریه در مراحل اولیه تشخیص داده شود، نرخ بقای بیماران را به طور قابل توجهی افزایش می‌دهد و این امر غربالگری را بسیار مهم می کند. [4]

در پژوهش‌های مرتبط با این موضوع همگی بر اهمیت سیگار کشیدن و مصرف انواع دخانیات تأکید کرده‌اند و آن را یکی از مهم‌ترین عوامل خطر در ابتلا به سرطان ریه می‌دانند. حساسیت ژنتیکی و استعداد خانوادگی نیز به عنوان عوامل مستعدکننده شناخته شده‌اند که می‌توانند با عوامل محیطی ترکیب شده و ریسک ابتلا را افزایش دهند. علاوه بر این، مقالات به نقش آلودگی هوا و قرار گرفتن در معرض خطرات شغلی پرداخته‌اند. رژیم غذایی نامتعادل نیز از دیگر عوامل خطری است که در مقاله سوم به آن‌ها پرداخته شده است و در کنار سایر عوامل به افزایش ریسک سرطان ریه کمک می‌کنند.[5][4][3]

امروزه افزایش تشخیص سرطان ریه در بین افرادی که هرگز سیگار نمی‌کشند نشان می‌دهد که عواملی فراتر از سیگار کشیدن در ایجاد این بیماری نقش دارند.[4]در این پروژه قصد داریم به بررسی اثرگذاری و نقش عوامل مختلف ذکر شده در تعیین ریسک بپردازیم.

ویژگی‌های مورد بررسی:

- سن
- جنسیت
- قرارگرفتن در معرض آلودگی هوا
- میزان مصرف الکل
- میزان آلرژی به گرد و غبار
- خطرات شغلی
- ریسک ژنیتیکی
- سابقه بیماری مزمن ریوی
- رژیم غذایی متعادل
- چاقی
- میزان مصرف دخانیات کنونی
- میزان مصرف دخانیات گذشته
- درد ناحیه قفسه سینه
- میزان سرفه‌های خونی
- میزان خستگی
- میزان کاهش وزن
- تنگی نفس
- خس خس سینه
- مشکل در بلع
- فشردگی ناخن انگشتان
- سرفه خشک
- خروپف کردن

سوالات اصلی:

1. آیا رابطه معناداری بین سن افراد و ابتلا به سرطان ریه وجود دارد یا خیر؟ آیا افزایش سن باعث افزایش ریسک ابتلا میشود؟ توزیع سنی درجه ریسک‌های مختلف چگونه است؟
2. آیا رابطه معناداری بین جنسیت افراد و ابتلا به سرطان ریه وجود دارد یا خیر؟
3. آیا رابطه معناداری بین آلودگی هوا و ابتلا به سرطان ریه وجود دارد یا خیر؟
4. آیا رابطه معناداری بین میزان مصرف الکل و ابتلا به سرطان ریه وجود دارد یا خیر؟
5. آیا رابطه معناداری بین حساسیت به گرد و غبار و ابتلا به سرطان ریه وجود دارد یا خیر؟

6. آیا رابطه معناداری بین خطرات شغلی و ابتلا به سرطان ریه وجود دارد یا خیر؟
7. آیا رابطه معناداری بین سابقه خانوادگی و ابتلا به سرطان ریه وجود دارد یا خیر؟
8. آیا رابطه معناداری بین سابقه بیماری مزمن ریوی و ابتلا به سرطان ریه وجود دارد یا خیر؟
9. آیا رابطه معناداری بین رژیم غذایی متعادل و ابتلا به سرطان ریه وجود دارد یا خیر؟
10. آیا رابطه معناداری بین چاقی و ابتلا به سرطان ریه وجود دارد یا خیر؟
11. آیا رابطه معناداری بین میزان مصرف دخانیات کنونی و ابتلا به سرطان ریه وجود دارد یا خیر؟
12. آیا رابطه معناداری بین میزان مصرف دخانیات گذشته و ابتلا به سرطان ریه وجود دارد یا خیر؟
13. آیا رابطه معناداری بین درد ناحیه قفسه سینه و ابتلا به سرطان ریه وجود دارد یا خیر؟
14. آیا رابطه معناداری بین میزان سرفه‌های خونی و ابتلا به سرطان ریه وجود دارد یا خیر؟
15. آیا رابطه معناداری بین میزان خستگی و ابتلا به سرطان ریه وجود دارد یا خیر؟
16. آیا رابطه معناداری بین میزان کاهش وزن و ابتلا به سرطان ریه وجود دارد یا خیر؟
17. آیا رابطه معناداری بین تنگی نفس و ابتلا به سرطان ریه وجود دارد یا خیر؟
18. آیا رابطه معناداری بین خس خس سینه و ابتلا به سرطان ریه وجود دارد یا خیر؟
19. آیا رابطه معناداری بین مشکل در بلع و ابتلا به سرطان ریه وجود دارد یا خیر؟
20. آیا رابطه معناداری بین فشردگی ناخن انگشتان و ابتلا به سرطان ریه وجود دارد یا خیر؟
21. آیا رابطه معناداری بین سرفه خشک و ابتلا به سرطان ریه وجود دارد یا خیر؟
22. آیا رابطه معناداری بین خروپف کردن و ابتلا به سرطان ریه وجود دارد یا خیر؟

سوالات با ترکیب فیچرها:

- آیا میانگین ریسک ابتلا به سرطان ریه برای زنان سیگاری و مردان سیگاری متفاوت است؟
- آیا تفاوت معناداری بین ریسک ابتلا برای افراد چاقی که رژیم غذایی مناسب دارند یا ندارند وجود دارد؟
- آیا افرادی که میزان بالای خستگی و سرفه‌های خونی دارند، نسبت به افرادی با همین سن و جنسیت ولی بدون این علائم، ریسک ابتلای بالاتری دارند؟
- آیا تفاوت معناداری بین ریسک ابتلا به سرطان ریه برای افرادی که سابقه بیماری مزمن ریوی دارند و افرادی که این سابقه را ندارند، وجود دارد؟
- آیا تفاوت معناداری در ریسک ابتلا به سرطان ریه بین افرادی که میزان بالایی از دخانیات مصرف کرده‌اند و در حال حاضر ترک کرده‌اند با افرادی که همچنان به مصرف دخانیات ادامه می‌دهند، وجود دارد؟
- آیا تفاوت معناداری در ریسک ابتلا به سرطان ریه بین افرادی که دارای خطرات شغلی بالایی هستند و رژیم غذایی متعادل دارند در مقایسه با افرادی که خطرات شغلی کمتری دارند و رژیم غذایی مناسبی ندارند، وجود دارد؟
- آیا افراد با ریسک ژنتیکی بالا که سیگار نمی‌کشند، ریسک ابتلای کمتری نسبت به افراد با ریسک ژنتیکی پایین که سیگار می‌کشند، دارند؟

- آیا تفاوت معناداری در ریسک ابتلا به سرطان ریه برای افرادی که سرفه خشک دارند و همچنین در معرض آلودگی هوا هستند، بیشتر از افرادی است که فقط در معرض آلودگی هوا قرار دارند؟
- آیا تفاوت معناداری بین ریسک ابتلا به سرطان ریه برای افرادی که چاقی و مصرف دخانیات بالا دارند و افرادی که فقط یکی از این دو عامل را دارند، وجود دارد؟
- آیا میانگین ریسک ابتلا به سرطان ریه برای افراد بالای 60 سال با سابقه بیماری مزمن ریوی متفاوت از افراد بالای 60 سال بدون این سابقه است؟
- آیا تفاوت معناداری در ریسک ابتلا به سرطان ریه برای افرادی که فشردگی ناخن انگشتان را تجربه می‌کنند و دخانیات مصرف می‌کنند، نسبت به افرادی که فشردگی ناخن ندارند ولی همچنان سیگاری هستند، وجود دارد؟

محدودیت‌ها:

کنترل گروه (گروه‌های کنترل و مقایسه):

برای بررسی تأثیر عوامل خطر، لازم است که افراد با شرایط متفاوت (مثلاً افرادی که دخانیات مصرف می‌کنند و افرادی که مصرف نمی‌کنند) در گروه‌های کنترل و مقایسه قرار گیرند. اما گاهی اوقات انتخاب گروه‌های مناسب مشکل است. ممکن است عواملی مانند دسترسی به اطلاعات دقیق، هزینه، یا محدودیت‌های زمانی، مانع از تشکیل گروه‌های کنترل بهینه شود. همچنین اگر گروه‌های کنترل به درستی انتخاب نشوند، نتایج نهایی دقت کمتری خواهند داشت.

خطاهای خودگزارشی:

بسیاری از داده‌ها در مطالعات مربوط به سلامت، بر اساس گزارش‌های خود افراد جمع‌آوری می‌شود (مثلاً میزان مصرف دخانیات، رژیم غذایی، یا سابقه خانوادگی). این اطلاعات ممکن است دقیق نباشند، زیرا افراد ممکن است برخی اطلاعات را به درستی یادآوری نکنند یا به دلایل شخصی اطلاعات را تغییر دهند. این خطاها می‌توانند بر اعتبار نتایج تاثیر بگذارند.

عدم توانایی بررسی علیت:(Causation)

این نوع مطالعات عمدتاً ارتباطات (Correlation) بین عوامل و ریسک ابتلا را بررسی می‌کنند و به ندرت می‌توانند علیت (یعنی اینکه یک عامل خاص مستقیماً باعث ابتلا شود) را ثابت کنند. برای بررسی علیت نیاز به آزمایش‌های کنترل شده تصادفی است که معمولاً امکان‌پذیر نیستند.

پراکندگی اطلاعات و عدم دسترسی به داده‌های کامل:

در برخی موارد، دسترسی به داده‌های کامل و دقیق ممکن نیست. برای مثال، اطلاعات در مورد میزان دقیق مواجهه با آلودگی هوا یا جزئیات دقیق رژیم غذایی ممکن است محدود یا غیرقابل دسترس باشند. این محدودیت می‌تواند دقت تحلیل را تحت تأثیر قرار دهد.

سوگیری انتخاب:(Selection Bias)

این محدودیت زمانی پیش می‌آید که روش انتخاب افراد به نحوی باشد که برخی افراد به‌طور ناخواسته بیشتر یا کمتر در مطالعه

حضور داشته باشند. برای مثال، اگر افراد مبتلا به سرطان تمایل بیشتری به شرکت در این مطالعه داشته باشند، نتایج ممکن است به نفع افراد بیمار سوگیری پیدا کند.

پیچیدگی‌های تحلیلی و محدودیت‌های مدل‌های آماری:

مدل‌های آماری که برای تحلیل این داده‌ها استفاده می‌شوند، ممکن است قادر به درک کامل روابط پیچیده بین متغیرها نباشند. برای مثال، تحلیل رابطه بین چندین عامل همزمان (مانند سیگار کشیدن و آلودگی هوا) ممکن است نیاز به مدل‌های پیچیده‌تر داشته باشد که از نظر محاسباتی و تحلیلی به چالش بکشاند.

تأثیر عوامل محیطی متغیر و غیرقابل کنترل:

برخی عوامل محیطی مانند تغییرات فصلی، سطح آلودگی هوا و وضعیت سلامت عمومی ممکن است بر نتایج تأثیر بگذارند، اما قابل کنترل نیستند. این متغیرها ممکن است به‌طور تصادفی تأثیر متفاوتی در نتایج ایجاد کنند و تحلیل‌های آماری را دشوارتر کنند.

نحوه جمع آوری دیتا:

در این پروژه قصد داریم از یک دیتاست[1] که در سایت kaggle ارائه شده استفاده کنیم که بر مبنای آن پژوهش‌هایی نیز انجام گرفته است. ویژگی‌های مورد بررسی که پیش‌تر تعریف شدند به ترتیب معادل ستونهای دیتاست مورد استفاده هستند.

همچنین با استفاده از زبان R یک خلاصه از دیتاست داریم که در زیر مشاهده می کنید.

```
── Data Summary ────────────────────
                           Values
Name                       df1
Number of rows             1000
Number of columns          26
_____________________
Column type frequency:
  character                2
  numeric                  24

── Variable type: character ──────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 Patient.Id            0             1   2   5     0     1000          0
2 Level                 0             1   3   6     0        3          0
```

── Variable type: numeric ──

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	index	0	1	500.	289.	0	250.	500.	749.	999	
2	Age	0	1	37.2	12.0	14	27.8	36	45	73	
3	Gender	0	1	1.40	0.491	1	1	1	2	2	
4	Air.Pollution	0	1	3.84	2.03	1	2	3	6	8	
5	Alcohol.use	0	1	4.56	2.62	1	2	5	7	8	
6	Dust.Allergy	0	1	5.16	1.98	1	4	6	7	8	
7	OccuPational.Hazards	0	1	4.84	2.11	1	3	5	7	8	
8	Genetic.Risk	0	1	4.58	2.13	1	2	5	7	7	
9	chronic.Lung.Disease	0	1	4.38	1.85	1	3	4	6	7	
10	Balanced.Diet	0	1	4.49	2.14	1	2	4	7	7	
11	Obesity	0	1	4.46	2.12	1	3	4	7	7	
12	Smoking	0	1	3.95	2.50	1	2	3	7	8	
13	Passive.Smoker	0	1	4.20	2.31	1	2	4	7	8	
14	Chest.Pain	0	1	4.44	2.28	1	2	4	7	9	
15	Coughing.of.Blood	0	1	4.86	2.43	1	3	4	7	9	
16	Fatigue	0	1	3.86	2.24	1	2	3	5	9	
17	Weight.Loss	0	1	3.86	2.21	1	2	3	6	8	
18	Shortness.of.Breath	0	1	4.24	2.29	1	2	4	6	9	
19	Wheezing	0	1	3.78	2.04	1	2	4	5	8	
20	Swallowing.Difficulty	0	1	3.75	2.27	1	2	4	5	8	
21	Clubbing.of.Finger.Nails	0	1	3.92	2.39	1	2	4	5	9	
22	Frequent.Cold	0	1	3.54	1.83	1	2	3	5	7	
23	Dry.Cough	0	1	3.85	2.04	1	2	4	6	7	
24	Snoring	0	1	2.93	1.47	1	2	3	4	7	

در این دیتاست فیچرها شامل دو نوع numeric و character هستند. تعداد فیچر ها 22 عدد و تعداد سمپل ها 1000 است. یکی از دلایل انتخاب این دیتاست این است که تعداد زیاد نمونه ها موجب می شود که توزیع سمپل ها به توزیع نرمال نزدیک تر شود. همچنین از نکات مثبت این دیتاست این است که missing value نداریم.

با استفاده از این دیتاست نیز یک مقاله نوشته شده است که در واقع یک پژوهش بر طراحی یک مدل جدید برای پیش بینی اولیه سرطان ریه هست. آن ها ابتدا یک پایگاه داده بین المللی سرطان را برای تعیین شایع ترین علائم و عوامل خطر سرطان ریه از دیدگاه پزشکی استاندارد تجزیه و تحلیل کردند. سپس پرسشنامه های پزشکی را بین پزشکان متخصص در زمینه های مختلف مانند داخلی، جراحی قفسه سینه، جراحی عمومی و انکولوژی و بیماران بررسی کردند و در نهایت با ادغام نظرات پزشکی بین المللی از گزارش های منتشر شده فیچر های مناسب را استخراج کرده و دیتاست را آماده و استفاده کرده اند.

عملکرد ابزار پیش بینی سرطان ریه با آزمایش آن با استفاده از موارد پزشکی محلی و مقایسه نتایج با نظر پزشکی محلی مورد ارزیابی قرار گرفت. علاوه بر این، رویکردهای یادگیری ماشین برای تجزیه و تحلیل 1000 پرونده بیمار از یک مجموعه داده بین المللی برای مقایسه نتایج ابزار با نتایج بین المللی استفاده شد. این ابزار

دقت و حساسیت بالایی در پیش بینی خطر سرطان ریه دارد. همچنین در نتایج این مقاله ذکر شده است که accuracy مدل 93.33٪ است . میزان specificity آن نیز به مقدار 90.47٪ است.[2]

مراجع:

[1]https://www.kaggle.com/datasets/thedevastator/cancerpatientsandairpollutionanewlink/data

[2] Ahmad AS, Mayya AM. A new tool to predict lung cancer based on risk factors. Heliyon. 2020 Feb 26;6(2):e03402. doi: 10.1016/j.heliyon.2020.e03402. PMID: 32140577; PMCID: PMC7044659.

[3] Risk factors for lung cancer worldwide. Malhotra, Jyoti and Malvezzi, Matteo and Negri, Eva and La Vecchia, Carlo and Boffetta, Paolo. http://erj.ersjournals.com/content/erj/48/3/889

[4] Schabath MB, Cote ML. Cancer Progress and Priorities: Lung Cancer. Cancer Epidemiol Biomarkers Prev. 2019 Oct;28(10):1563-1579. doi: 10.1158/1055-9965.EPI-19-0221. PMID: 31575553; PMCID: PMC6777859.

[5] Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. Eur Respir J. 2016 Sep;48(3):889-902. doi: 10.1183/13993003.00359-2016. Epub 2016 May 12. PMID: 27174888.