FACTORES ASOCIADOS AL CÁNCER DE PULMÓN

PAULA BELAZA HERNAIZ
2º BIOMEDICINA

INTRODUCCIÓN

El cáncer de pulmón es una enfermedad multifactorial influida tanto por síntomas respiratorios como por hábitos de vida. Este análisis explora cómo el consumo de alcohol y la presencia de tos se relacionan con el diagnóstico de cáncer utilizando el dataset Survey Lung Cancer.

HIPÓTESIS

"Los pacientes que consumen alcohol y presentan tos frecuente tienen mayor probabilidad (mayor ODDs) de ser diagnosticados con cáncer de pulmón".

OBJETIVO

Evaluar si la tos y el consumo de alcohol se asocian con un mayor diagnóstico de cáncer de pulmón mediante un análisis exploratorio y un modelo predictivo básico

DATASET

- 309 pacientes
- 16 variables (1 numérica, resto binarias/categóricas)
- Sin valores faltantes
- Clase objetivo desbalanceada (87% "YES")
- Fuente: Kaggle

```
> str(datos)
'data.frame':   309 obs. of  16 variables:
 $ GENDER                : chr  "M" "M" "F" "M" ...
 $ AGE                   : int  69 74 59 63 63 75 52 51
 $ SMOKING               : int  1 2 1 2 1 1 2 2 2 2 ...
 $ YELLOW_FINGERS        : int  2 1 1 2 2 2 1 2 1 2 ...
 $ ANXIETY               : int  2 1 1 2 1 1 1 2 2 2 ...
 $ PEER_PRESSURE         : int  1 1 2 1 1 1 1 2 1 2 ...
 $ CHRONIC.DISEASE       : int  1 2 1 1 1 2 1 1 1 2 ...
 $ FATIGUE               : int  2 2 2 1 1 2 2 2 2 1 ...
 $ ALLERGY               : int  1 2 1 1 1 2 1 2 1 2 ...
 $ WHEEZING              : int  2 1 2 1 2 2 2 1 1 1 ...
 $ ALCOHOL.CONSUMING     : int  2 1 1 2 1 1 2 1 1 2 ...
 $ COUGHING              : int  2 1 2 1 2 2 2 1 1 1 ...
 $ SHORTNESS.OF.BREATH   : int  2 2 2 1 2 2 2 2 1 1 ...
 $ SWALLOWING.DIFFICULTY: int  2 2 1 2 1 1 1 2 1 2 ...
 $ CHEST.PAIN            : int  2 2 2 2 1 1 2 1 1 2 ...
 $ LUNG_CANCER           : chr  "YES" "YES" "NO" "NO" ..
```

DICCIONARIO DE VARIABLES:

- AGE: Edad del participante.
- GENDER: Género (M/F).
- ALCOHOL.CONSUMING: Consumo de alcohol (1=No, 2=Sí).
- COUGHING: Tos frecuente (1=No, 2=Sí).
- LUNG_CANCER: Diagnóstico final de cáncer (YES/NO).
- FATIGUE, WHEEZING, SHORTNESS.OF.BREATH, etc.: Síntomas binarios (1=No, 2=Sí).

METODOLOGÍA

- Limpieza y recodificación
- EDA univariante y bivariante
- SMOTE para balanceo
- Matriz Cramer's V
- Modelo logístico

Dataset crudo → Limpieza → EDA univariante → Balanceo → Gráficos → Modelo logístico → Conclusiones

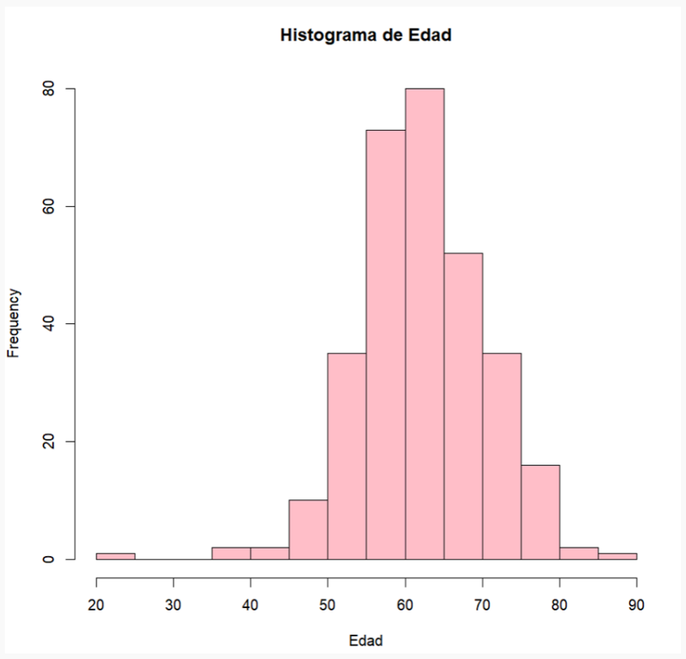


Figura 1. Histograma de la distribución de edad

EXPLORATORY DATA ANALYSIS (EDA)

Para conocer la estructura del dataset y detectar posibles patrones. Se analizaron:

- Distribuciones (edad)
- Estadísticos descriptivos
- Relación variables ↔ diagnóstico
- Cramer's V
- Selección de alcohol y tos como variables relevantes

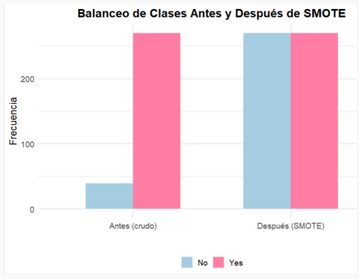
Variable	Categorías y frecuencias
Género	• M: 162 (52.4%) • F: 147 (47,6%)
Alcohol	• Sí: 172 (55,7%) • No: 137 (44,3%)
Tos	• Sí: 179 (57,9%) • No: 130 (42,1%)
Cáncer	• Sí: 270 (87,4%) • No: 39 (12,6%)

Cáncer	Frecuencia
Sí	• 270 (87,4%)
No	• 39 (12,6%)

→ **Desbalance de la variable objetivo (Lung cancer)**

BALANCEO DE DATOS (SMOTE)

Dataset original (87% YES / 13% NO)
↓
Problema: desbalance
↓
Aplicamos SMOTE
↓
Dataset balanceado (~50% / 50%)
↓
Se usa solo para: gráficas descriptivas
(NO se usa para el modelo logístico)



¿Por qué fue necesario balancear?

- La variable objetivo estaba fuertemente desbalanceada (87% de cáncer).
- Este desbalance distorsiona las gráficas de proporciones.
- SMOTE genera casos sintéticos de la clase minoritaria (NO cáncer).
- El resultado se usa únicamente para visualización, NO para la regresión logística.

RESULTADOS

ALCOHOL Y DIAGNÓSTICO DE CANCER

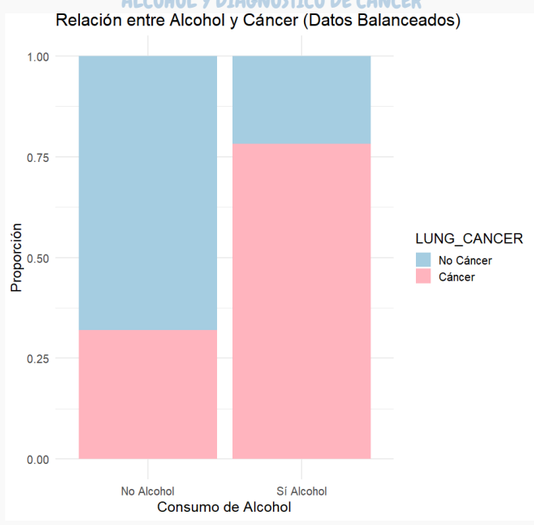


Figura 2. Proporción de cáncer según consumo de alcohol tras el balanceo de clases. Los pacientes que consumen alcohol presentan un mayor porcentaje de cáncer en comparación con los que no consumen.

TOS FRECUENTE Y CÁNCER DE PULMÓN

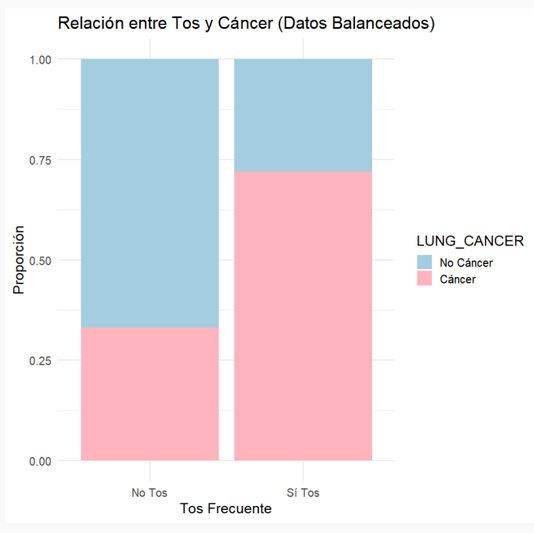


Figura 3. Relación entre tos frecuente y diagnóstico de cáncer en el dataset balanceado. La tos es el factor individual más fuertemente asociado al cáncer.

INTERACCIÓN ALCOHOL Y TOS EN EL CÁNCER DE PULMÓN

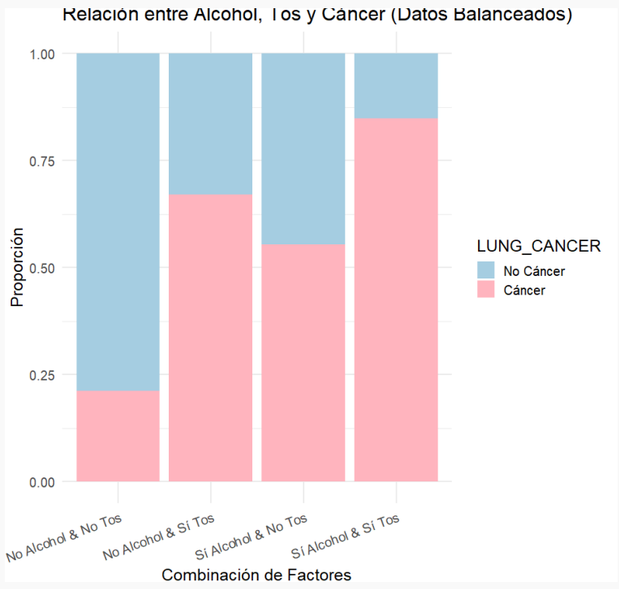


Figura 4. Combinación de factores y cáncer en el dataset balanceado. El grupo con consumo de alcohol y tos frecuente ('Sí Alcohol & Sí Tos') presenta la mayor proporción de cáncer, apoyando la hipótesis planteada

HEATMAP CRAMER'S (MATRIZ DE ASOCIACIONES)

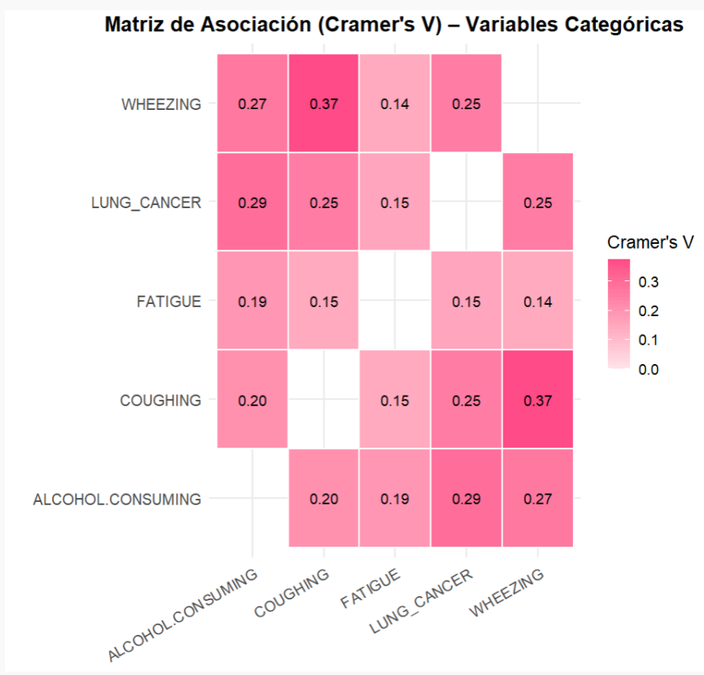


Figura 5. Matriz de asociación entre variables categóricas mediante Cramer's V. La tos y el consumo de alcohol presentan las asociaciones más altas con el cáncer, justificando su selección como variables principales del estudio

Interpretación

La tos (COUGHING) muestra una asociación más alta con el cáncer de pulmón (Cramer's V = 0.25), seguida por el consumo de alcohol (0.29).
Los síntomas respiratorios (tos y sibilancias) también presentan asociaciones entre sí (0.37).
El resto de variables muestran asociaciones débiles

MODELO LOGÍSTICO BINARIO

Variables incluidas: Alcohol, Tos, Interacción

Predictor	Coeficiente (β)	OR (e^β)	p-value	IC-95%
Intercepto	0,6819	1,98	0,004	[1,25-3,19]
Alcohol	1,7786	5,92	6,02e-05	[2.62–15.24]
Tos	1,3398	3,82	0,000815	[1.79–8.73]

INTERPRETACIÓN

- El modelo logístico muestra que tanto el consumo de alcohol (OR = 5.92; IC95% = 2.62–15.24) como la tos frecuente (OR = 3.82; IC95% = 1.79–8.73) aumentan significativamente la probabilidad de cáncer (p < 0.001). Ambos predictores presentan intervalos de confianza por encima de 1, confirmando su relevancia estadística y apoyando la hipótesis del estudio.

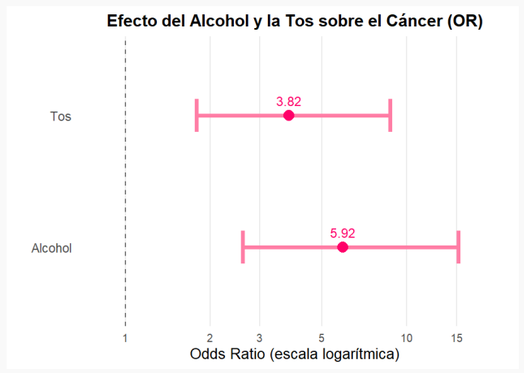


Figura 6. Forest plot del modelo logístico mostrando los odds ratios e intervalos de confianza (95%). Tanto el consumo de alcohol como la tos incrementan significativamente la probabilidad de cáncer

El modelo logístico indica que tanto el consumo de alcohol (OR=5.92) como la tos (OR=3.82) incrementan significativamente la probabilidad de cáncer. Ambos efectos presentan intervalos de confianza por encima de 1, lo que confirma su relevancia estadística

CURVA ROC

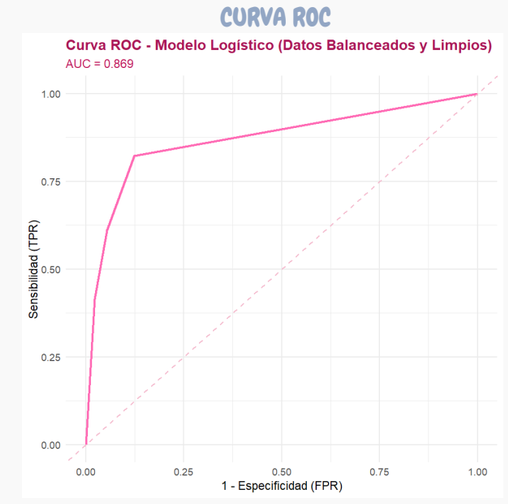


Figura 7. Curva ROC con dataset balanceado y limpiado (SMOTE). Modelo predictivo.

AUC = 0.8688 indica una excelente capacidad discriminativa del modelo, mostrando que puede diferenciar entre cáncer y no cáncer con alta precisión.

DISCUSIÓN

Los resultados del modelo logístico son coherentes con los patrones observados en el análisis exploratorio. La tos muestra un efecto significativo y consistente en todas las etapas del análisis, mientras que el alcohol presenta un efecto moderado pero también significativo. La combinación de ambos factores es el grupo con mayor proporción de cáncer en los datos balanceados, apoyando la hipótesis inicial.

LIMITACIONES DEL ESTUDIO

- El dataset original es desbalanceado y requiere técnicas de balanceo para visualizaciones.
- Las variables son binarias (Sí/No)-->limitada profundidad analítica.
- El origen del dataset es de encuesta y puede no representar datos clínicos reales.
- No se pueden establecer relaciones causales, solo asociativas

CONCLUSIONES

El consumo de alcohol y la tos frecuente aumentan la probabilidad de cáncer en este dataset, con resultados consistentes entre el análisis exploratorio y el modelo logístico. Estos hallazgos confirman la hipótesis planteada.

MENSAJE CLAVE

- La tos es el predictor más fuerte.
- El alcohol también incrementa el riesgo.
- La combinación de ambos factores muestra la mayor proporción de cáncer.
- El modelo logístico confirma el EDA

BIBLIOGRAFÍA

- World Health Organization. (2023). Cancer: Key facts. https://www.who.int/news-room/fact-sheets/detail/cancer
- National Cancer Institute. (2022). Lung cancer: Overview. https://www.cancer.gov/types/lung
- Cancer Research UK. (2023). Lung cancer symptoms. https://www.cancerresearchuk.org/about-cancer/lung-cancer/symptoms
- International Agency for Research on Cancer. (2021). Alcohol consumption and carcinogenesis. https://publications.iarc.fr
- Bagnardi, V., et al. (2015). Alcohol consumption and site-specific cancer risk: A comprehensive review and meta-analysis. Annals of Oncology, 26(8), 1523–1535. https://doi.org/10.1093/annonc/mdv164
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387
- UCLA Institute for Digital Research and Education. (n.d.). Cramer's V in categorical association. https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-is-cramers-v/
- Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, 28(5). https://www.jstatsoft.org/article/view/v028i05
- Kaggle. (2022). Survey Lung Cancer Dataset. La tos es el predictor más fuerte.
- El alcohol también incrementa el riesgo.
- La combinación de ambos factores muestra la mayor proporción de cáncer.
- El modelo logístico confirma el EDA