1. CONJUNTO DE DATOS

Nuestro conjunto de datos seleccionados cuenta con 50,000 pacientes y 11 características o variables, este dataset esta enfocado al estudio de factores de riesgo asociados al cáncer de pulmón.

Entre las características encontramos:

- **Edad:** a mayor edad, aumenta la probabilidad de acumulación de daños celulares y riesgo de cáncer.
- **Género:** históricamente los hombres han tenido mayor incidencia, aunque en mujeres fumadoras la brecha se acorta.
- **Años de tabaquismo acumulado:** es el factor de riesgo más fuerte y directo, mientras más alto, mayor riesgo.
- **Exposición al radón:** gas radioactivo que daña el ADN de las células pulmonares, reconocido como segunda causa principal tras el tabaco.
- **Exposición al asbesto:** fibras tóxicas que se alojan en los pulmones y potencian la probabilidad de cáncer.
- **Exposición al humo de segunda mano:** incluso sin fumar, la inhalación pasiva de humo incrementa el riesgo notablemente.
- **Diagnóstico de EPOC (Enfermedad Pulmonar Obstructiva Crónica):** la inflamación y el daño pulmonar crónico predisponen al desarrollo de cáncer.
- **Consumo de alcohol:** actúa como factor indirecto, potenciando el riesgo cuando se combina con tabaco.
- **Antecedentes familiares:** la predisposición genética puede aumentar la vulnerabilidad frente a otros factores.

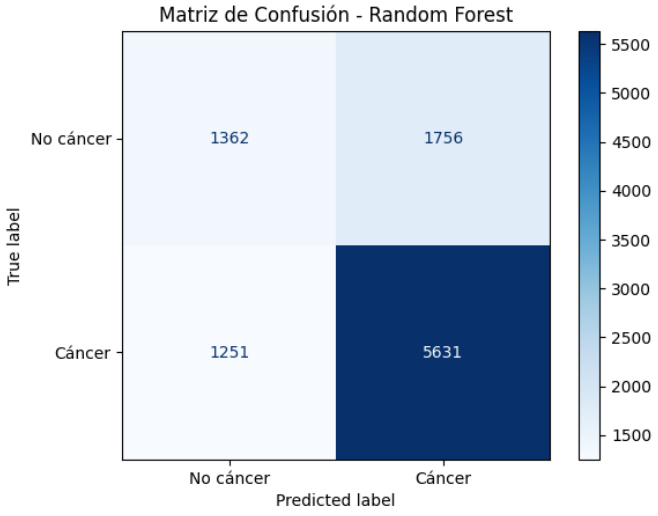
Nuestra columna objetivo o variable dependiente es lung_cancer, que indica si el paciente ha sido diagnosticado o no con la enfermedad. El dataset mezcla variables cuantitativas y cualitativas, por lo que fue necesario realizar un preprocesamiento de los datos con el fin de aplicar los modelos seleccionados. Además, el modelo no cuenta con valores nulos en ninguna columna.

2. MODELOS SELECCIONADOS

2.1 RANDOM FOREST

El Random Forest es un modelo de clasificación (y también de regresión) que funciona como un “bosque” de árboles de decisión. Es decir, en lugar de usar un solo árbol, construye muchos árboles diferentes a partir de muestras aleatorias de los datos y características. Cada árbol da su predicción, y el modelo final decide la clase más votada entre todos ellos.

Resultados obtenidos para Random Forest:



Verdaderos Positivos (TP)	5631
Verdaderos Negativos (TN)	1362
Falsos Positivos (FP)	1756
Falsos Negativos (FN)	1251
Total de datos	10000

Métrica	**Valor**	**Interpretación**
Exactitud (ACC)	0,6993	El modelo acierta en casi el 70% de los diagnósticos totales.
Sensibilidad (SEN), Recall o TPR	0,818221447	El 81,8% de los pacientes con cáncer fueron correctamente identificados.
Especificidad (SPE) o TNR	0,436818473	Solo el 43.7% de los pacientes sin cáncer fueron correctamente identificados.
Precisión o Valor Predictivo Positivo	0,762285095	El 76,2% de los pacientes clasificados como con cáncer realmente lo tenían.
Valor Predictivo Negativo (NPV)	0,521239954	El 52,1% de los pacientes clasificados como sin cáncer realmente estaban sanos.
Tasa de descubrimiento falso (FDR)	0,237714905	El 23,7% de los pacientes diagnosticados como con cáncer no lo tenían.
Tasa de falsos negativos (FNR)	0,181778553	El 18,1% de los pacientes con cáncer fueron clasificados como sanos.
Tasa de falsos positivos (FPR)	0,563181527	El 56,3% de los pacientes sanos fueron clasificados como con cáncer.
Índice de elevación (Lift)	1,107650531	El modelo es ligeramente mejor que adivinar al azar.
F1-score	0,789263438	Valor aceptable entre precisión y recall. Sin embargo, deja mucho que desear para estándares médicos

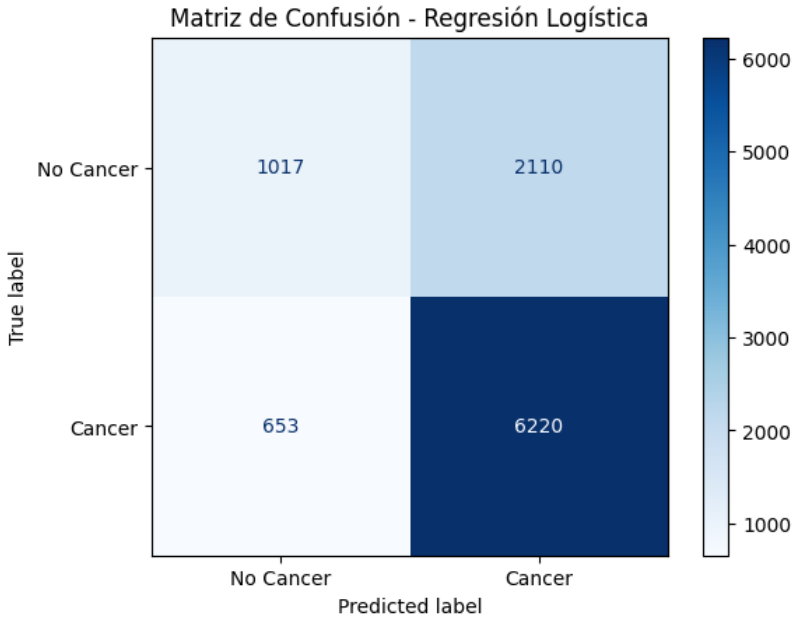
Desempeño del modelo:

el Random Forest ofrece un desempeño moderado, con una buena capacidad para identificar a los pacientes con cáncer, con una sensibilidad del 81,8% y una precisión del 76,2%. Sin embargo, cuenta con una baja capacidad para identificar correctamente a los pacientes sanos (especificidad de 43.7%) y un nivel de falsos negativos demasiado alto para un contexto clínico, con una tasa de falsos negativos de 18,1%.

2.2 REGRESIÓN LOGÍSTICA:

La regresión logística es un modelo matemático y estadístico que se utiliza principalmente para problemas de clasificación binaria (dos categorías). Su objetivo es estimar la probabilidad de que una observación pertenezca a una clase u otra, a partir de un conjunto de variables independientes. Básicamente transforma cualquier valor real en un número entre 0 y 1. De esta manera, el modelo produce probabilidades interpretables. Por ejemplo, si el resultado es 0.8, se entiende como una probabilidad del 80% de pertenecer a la clase positiva.

Resultados para el modelo de regresión logística:



Verdaderos Positivos (TP)	6220
Verdaderos Negativos (TN)	1017
Falsos Positivos (FP)	2110
Falsos Negativos (FN)	653
Total de datos	10000

Métrica	Valor	Interpretación
Exactitud (ACC)	0,7237	El modelo acierta en el 72.37% de los diagnósticos totales.
Sensibilidad (SEN), Recall o TPR	0,904990543	El 90,4% de los pacientes con cáncer fueron correctamente identificados.
Especificidad (SPE) o TNR	0,325231852	Solo el 32.5% de los pacientes sin cáncer fueron correctamente identificados.
Precisión o Valor Predictivo Positivo	0,746698679	El 74.6% de los pacientes clasificados como con cáncer realmente lo tenían.
Valor Predictivo Negativo (NPV)	0,608982036	El 60.9% de los pacientes clasificados como sin cáncer realmente estaban sanos.
Tasa de descubrimiento falso (FDR)	0,253301321	El 25.3% de los pacientes diagnosticados como con cáncer no lo tenían.
Tasa de falsos negativos (FNR)	0,095009457	El 9.5% de los pacientes con cáncer fueron clasificados como sanos.
Tasa de falsos positivos (FPR)	0,674768148	El 67.5% de los pacientes sanos fueron clasificados como con cáncer.
Índice de elevación (Lift)	1,086423221	El modelo es solo ligeramente mejor que el azar.
F1-score	0,818259554	Muestra un mejor balance que Random Forest, aunque sigue siendo bajo para uso clínico.

Desempeño:

El modelo de regresión logística se destaca por su alta sensibilidad de un 90,4%, lo que la convierte en una herramienta más segura para la detección de cáncer, aun cuando sacrifica especificidad (solo un 32.5%) y genera más falsos positivos (tasa de falsos positivos de 67.5%). Sin embargo, consideramos que en un contexto clínico, esto es aceptable, porque es preferible sobre diagnosticar que dejar pasar un caso real. Además, cabe recalcar que este modelo cuenta con una tasa de falsos negativos solo un 9.5%.

3. MEJOR MODELO

Antes de definir que un modelo como el mejor, es importante recalcar que nuestro dataset presenta un desbalance en sus clases, teniendo un 69% de pacientes con cáncer y un 31% de pacientes sin cáncer, por lo que la alta sensibilidad de los modelos pueda deberse a que estos estuvieran sesgados y tuvieran cierta “preferencia” hacia la característica positiva. Ahora bien, en base a las métricas, consideramos que el modelo de regresión logística es la mejor alternativa para este dataset, ya que prioriza la detección temprana de cáncer con una sensibilidad del 90,4% y reduce los falsos negativos, contando con una tasa de solo un 9.5%, aunque aumente el número de falsos positivos lo cual, en medicina, suele ser preferible a dejar pasar un caso real.

✓ lung_cancer

true
34.4k 69%

false
15.6k 31%