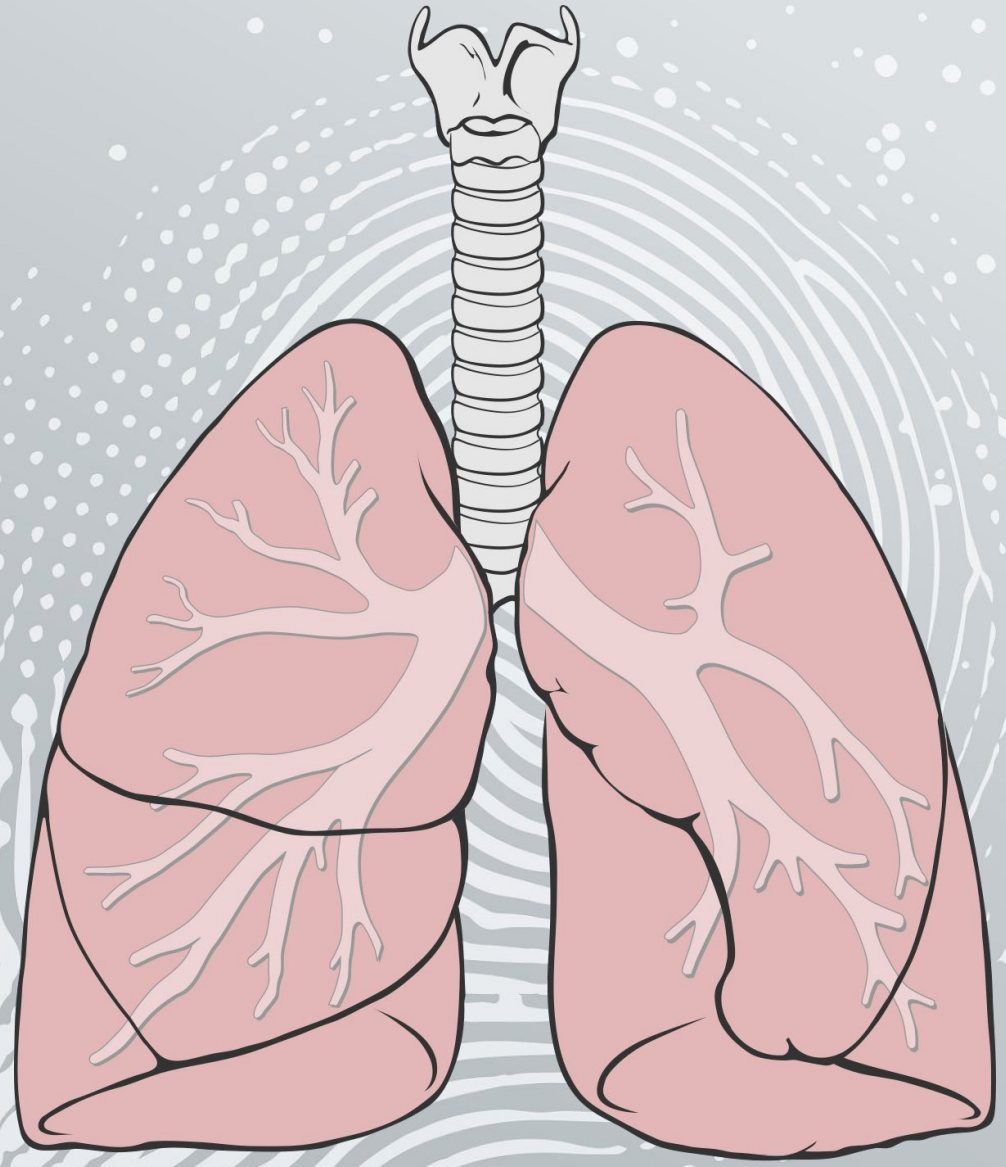


Diagnóstico cáncer de pulmón a partir de consumo de cigarrillo y contaminación del ambiente.

Reinaldo Cárdenas
Juan Herrera
Camilo Carvajal

Motivación

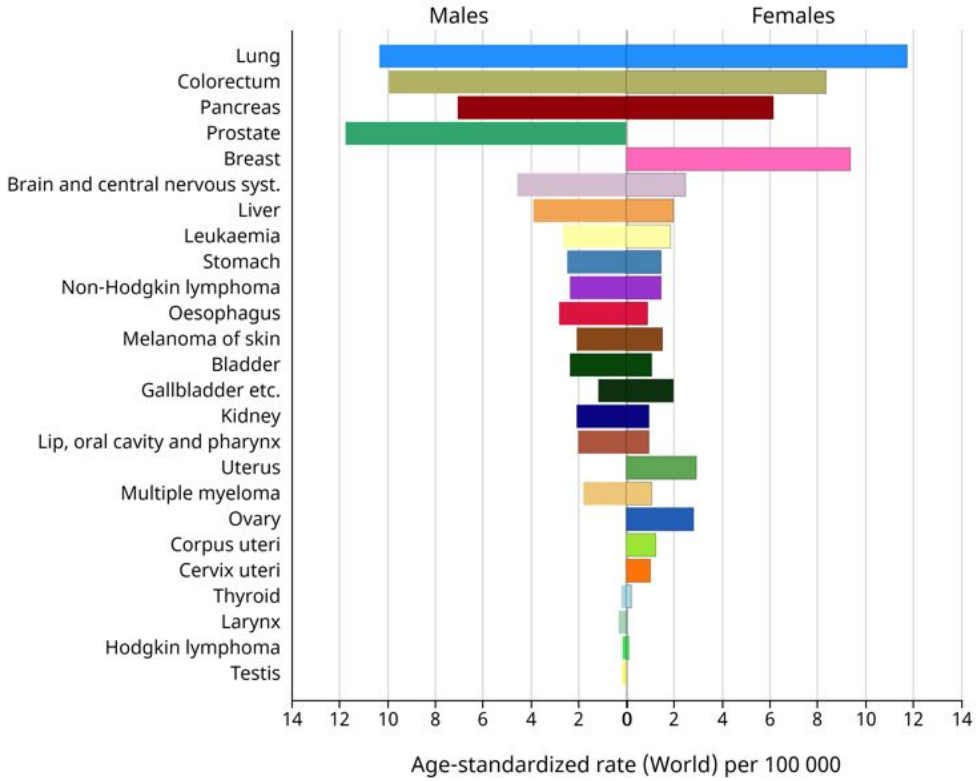


TABLE 1. New cases and deaths for 36 cancers and all cancers combined in 2022.

Cancer site	Incidence			Mortality		
	Rank	New cases	% of all sites	Rank	Deaths	% of all sites
Lung	1	2,480,301	12.4	1	1,817,172	18.7
Female breast	2	2,308,897	11.6	4	665,684	6.9
Colorectum	3	1,926,118	9.6	2	903,859	9.3
Prostate	4	1,466,680	7.3	8	396,792	4.1
Stomach	5	968,350	4.9	5	659,853	6.8
Liver	6	865,269	4.3	3	757,948	7.8
Thyroid	7	821,173	4.1	24	47,485	0.5
Cervix uteri	8	661,021	3.3	9	348,189	3.6
Bladder	9	613,791	3.1	13	220,349	2.3
Non-Hodgkin lymphoma	10	553,010	2.8	11	250,475	2.6

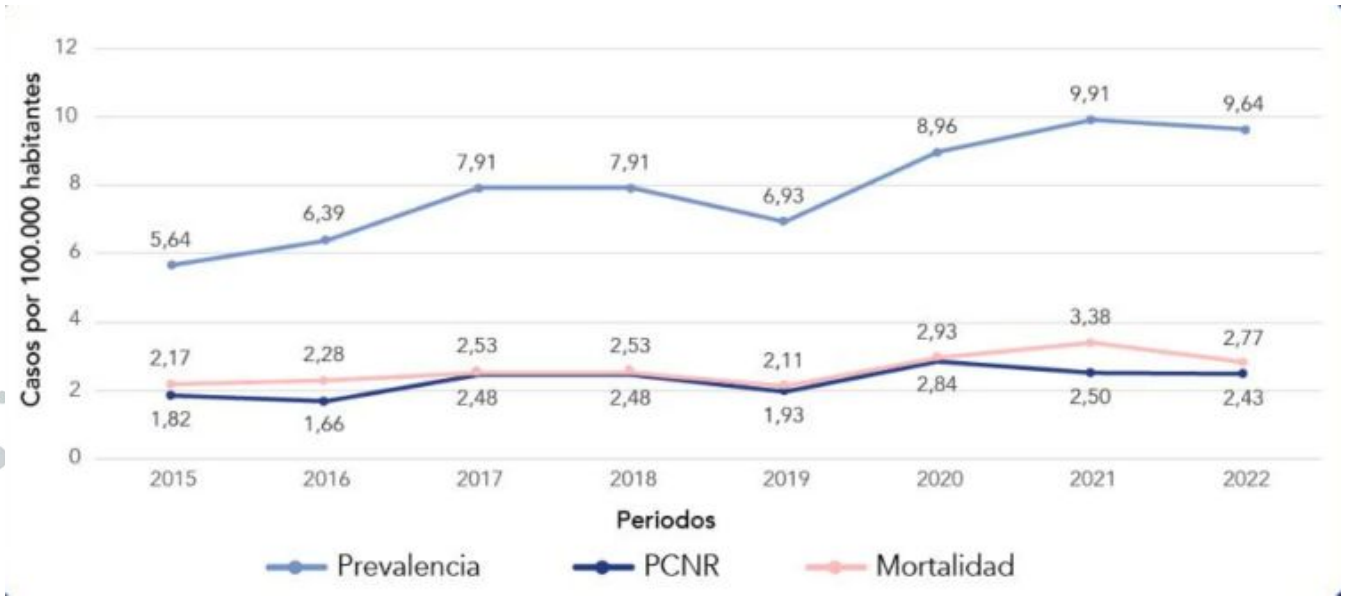
A cancer Journal for clinicians.

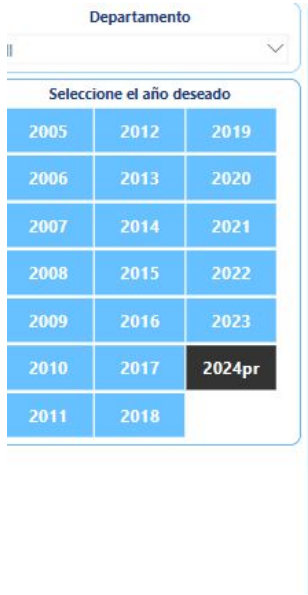
El cáncer de pulmón es el cáncer de mayor incidencia y mortalidad en el mundo
https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834
https://gco.iarc.fr/en

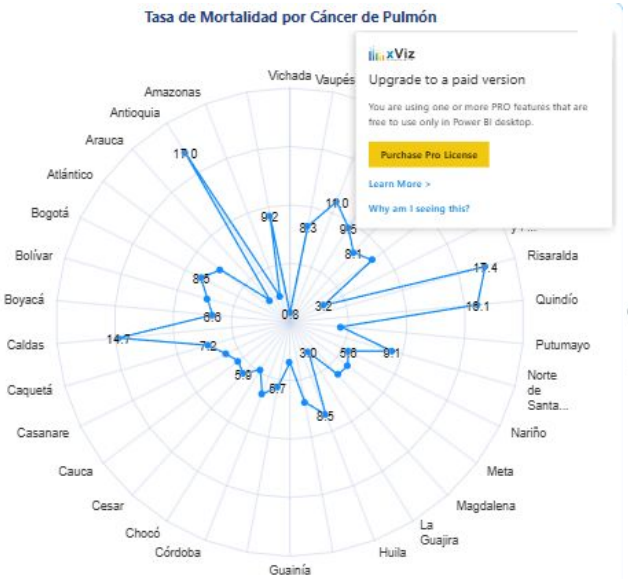
- https://www.who.int/news-room/fact-sheets/detail/lung-cancer
- https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834

#LaUISqueQueremos

Motivación





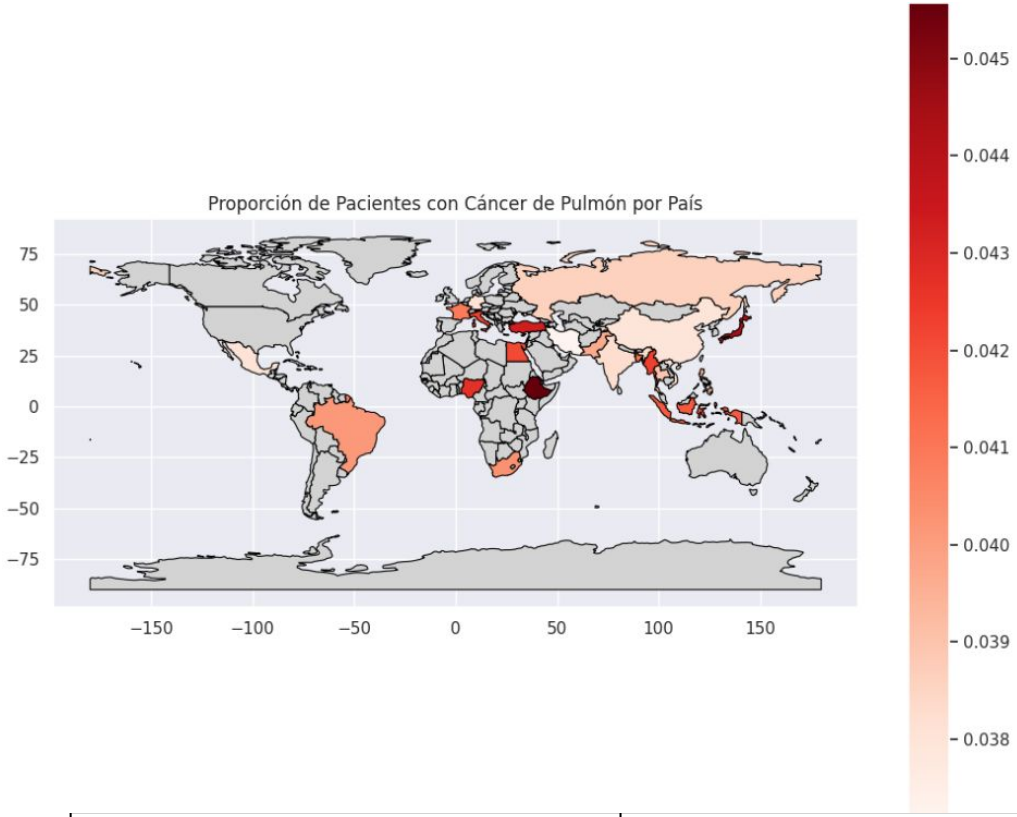


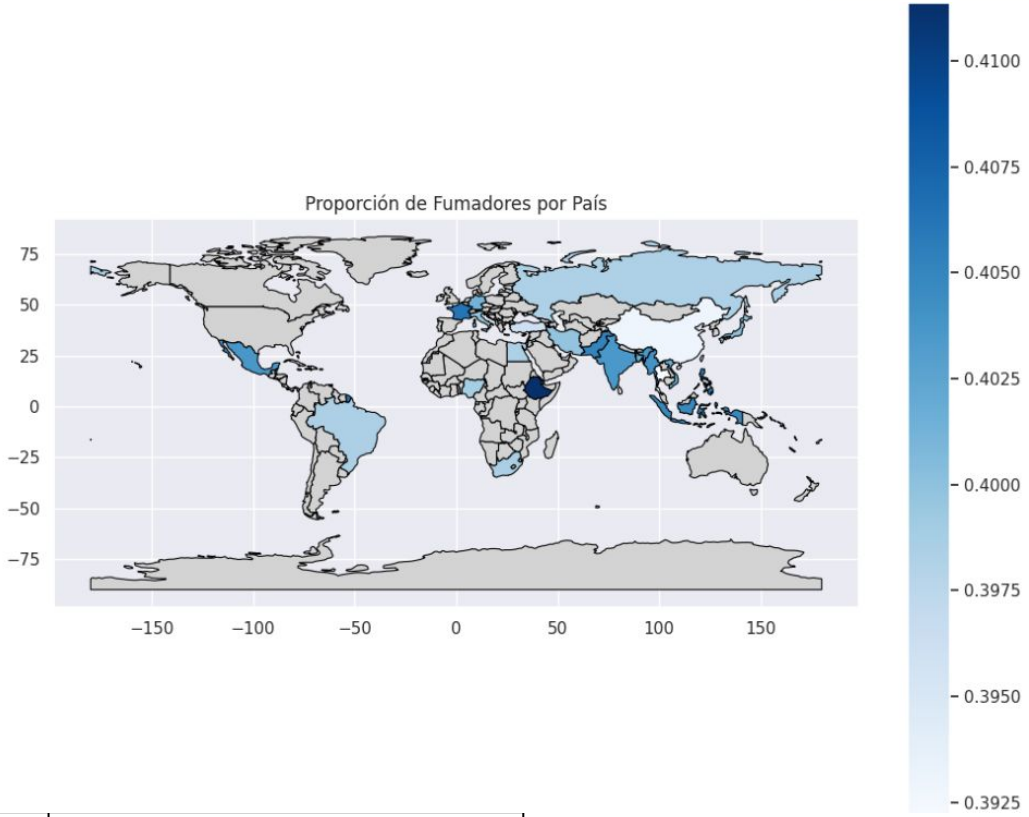
Casos cáncer pulmón por 100.000 habitantes en Colombia.

En Colombia el cáncer de pulmón es el segundo cáncer de mayor incidencia, detrás del cáncer de mama, pero el primero en mortalidad.

- https://www.asivamosensalud.org/indicadores/enfermedades-cronicas-no-transmisibles/tasa-de-mortalidad-por-cancer-de-pulmon
- https://pmc.ncbi.nlm.nih.gov/articles/PMC9512683/
- https://cuentadealtocosto.org/cancer/dia-mundial-del-pulmon-2023/

Dataset





Lung_Cancer_Diagnosis	No	Yes
Count	211.671	8.961
%	95.94%	4.06%

https://www.kaggle.com/datasets/aizahzeeshan/lung-cancer-risk-in-25-countries

Dataset

	Age	Gender	Smoker	Years_of_Smoking	Cigarettes_per_Day	Passive_Smoker	Lung_Cancer_Diagnosis	Air_Pollution_Exposure	Occupational_Exposure	Indoor_Pollution
0	80	Male	Yes	30	29	No	No	Low	Yes	No
1	53	Male	No	0	0	Yes	No	Low	Yes	No
2	47	Male	Yes	12	6	Yes	No	Medium	No	No
3	39	Female	No	0	0	No	No	Low	No	No
4	44	Female	No	0	0	Yes	No	Medium	Yes	No

Lung_Cancer_Diagnosis	No	Yes
Count	211.671	8.961
%	95.94%	4.06%

https://www.kaggle.com/datasets/aizahzeeshan/lung-cancer-risk-in-25-countries

#LaUISqueQueremos

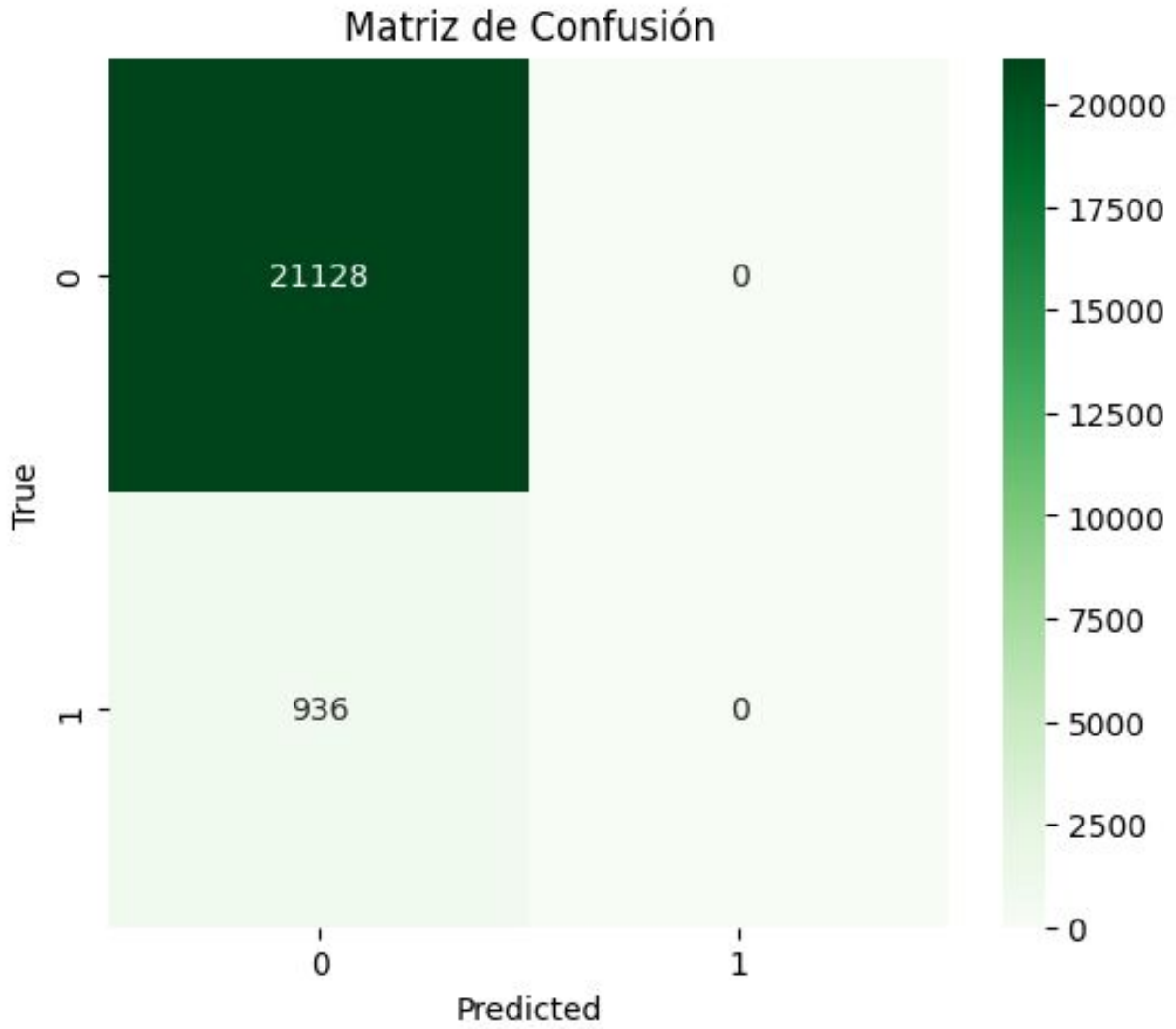
Red neuronal

Train - Test: 90-10

Resultados:

- 4 épocas
- Train accuracy: 96%
- Test accuracy: 95.75%
- Loss 0.1571

El modelo no clasifica ningún paciente con cáncer como paciente con cáncer



Alternativas

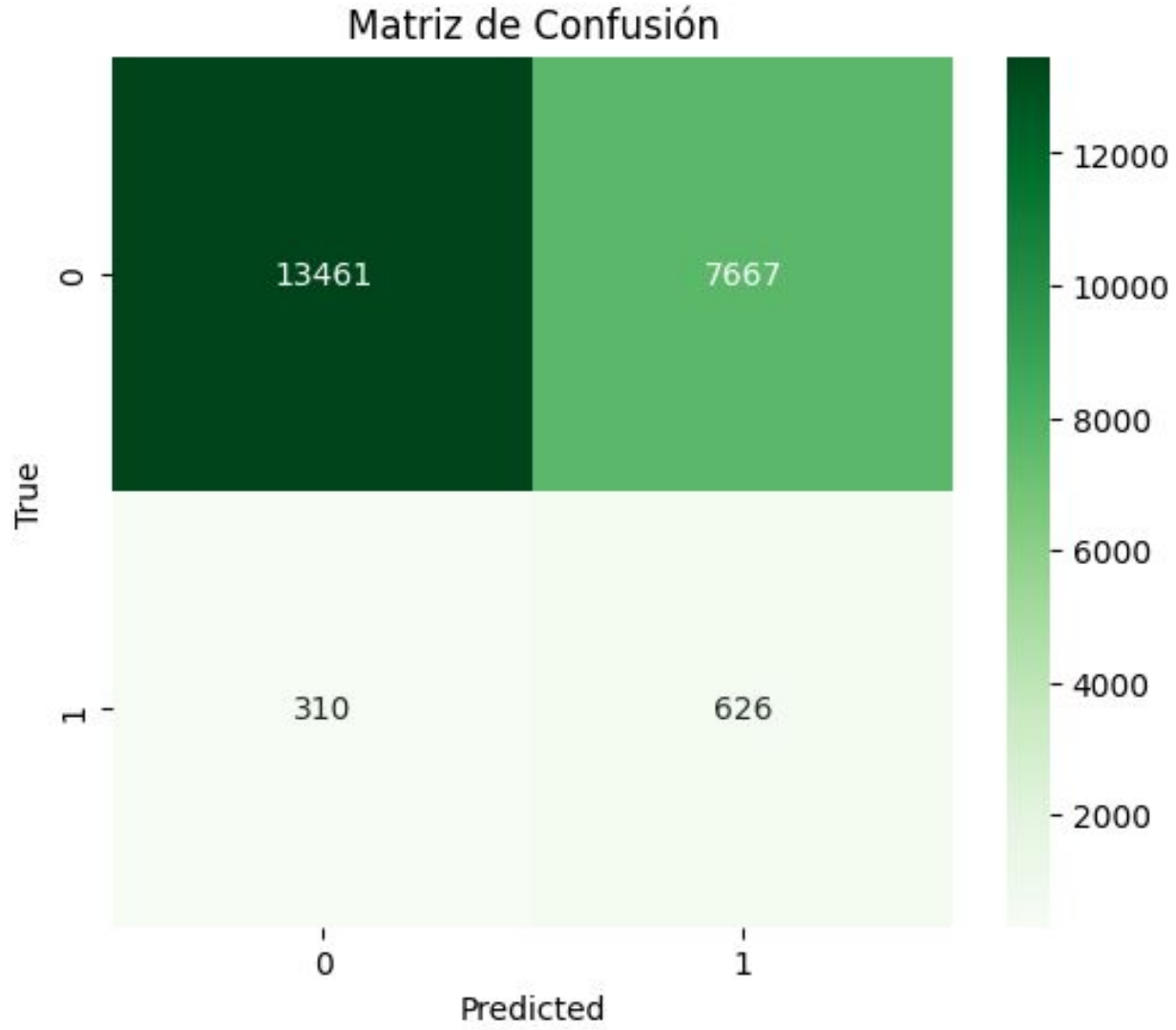
1. Cost-Sensitive Learning

Class weights:

- Clase 0 -> 1
- Clase 1 -> 15

Accuracy 64%

El modelo ya no está sesgado, ahora es capaz de clasificar la clase minoritaria, pero el costo es el aumento en los falsos positivos

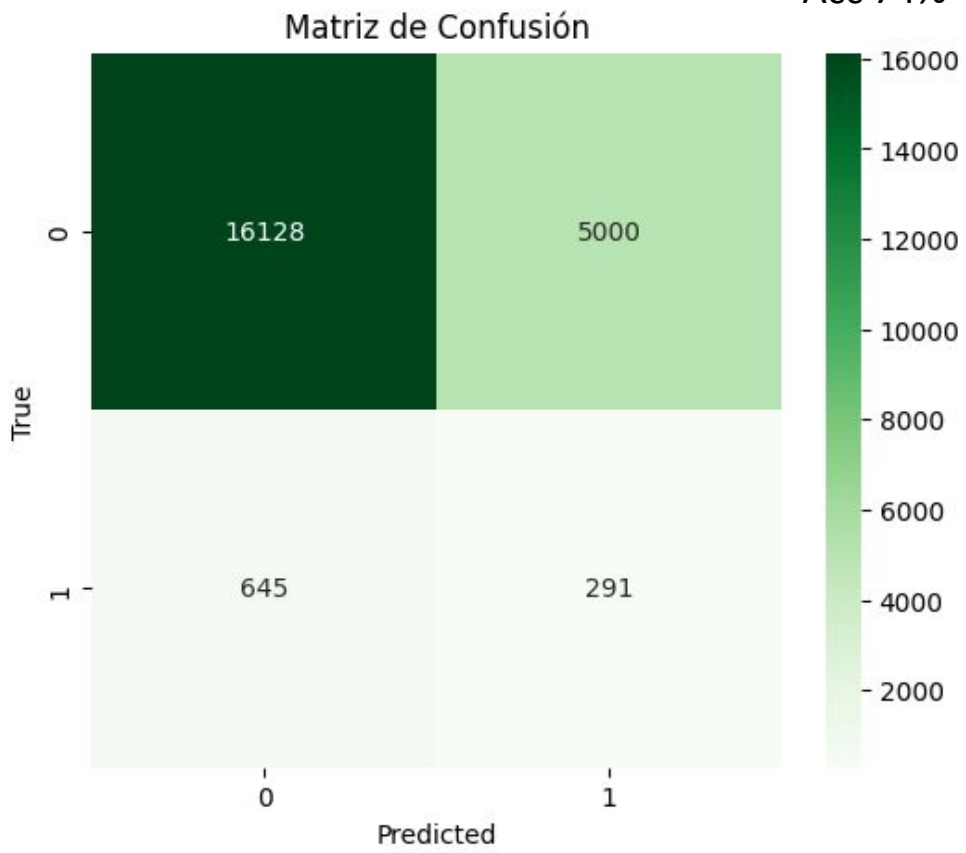


Alternativas

2. Oversampling - SMOTE

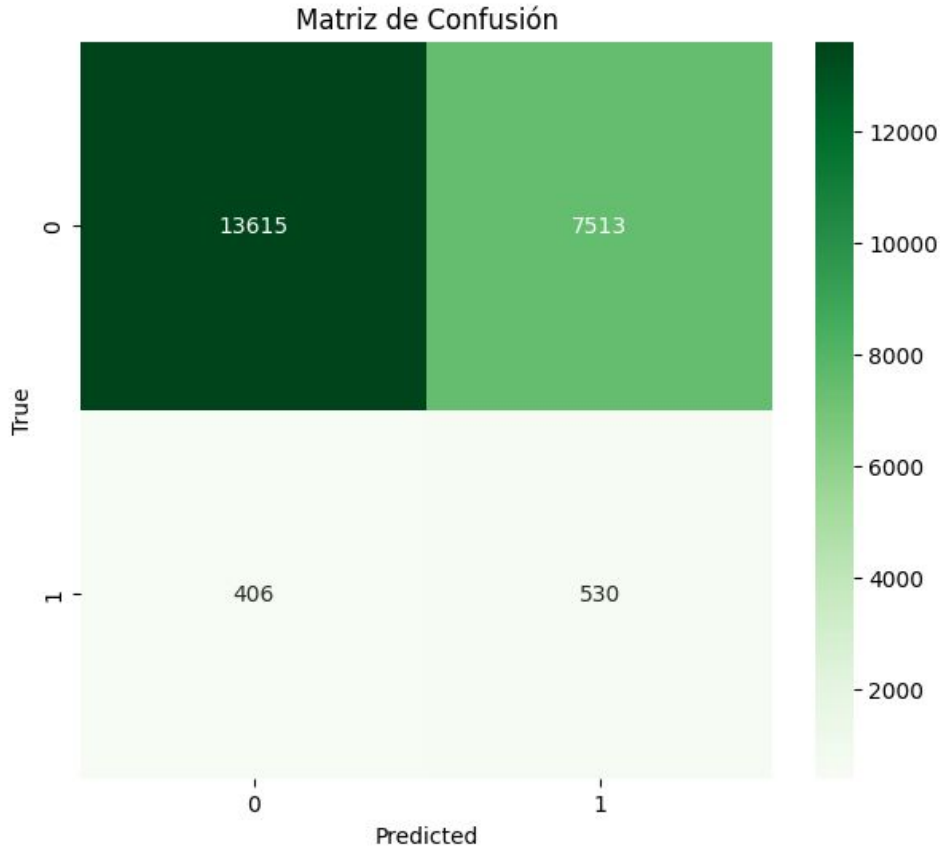
13D

Acc 74%



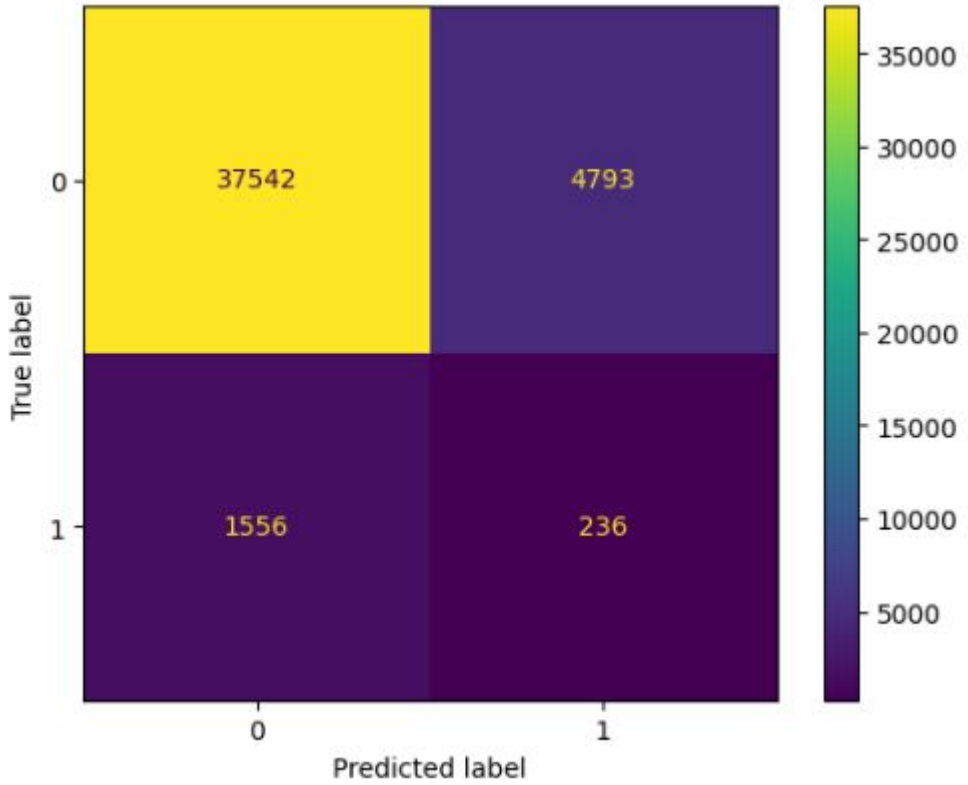
8D (PCA)

Acc 64%



Métodos usados/Modelos

1. Random Forest



Random Forest con class_weight='balanced'

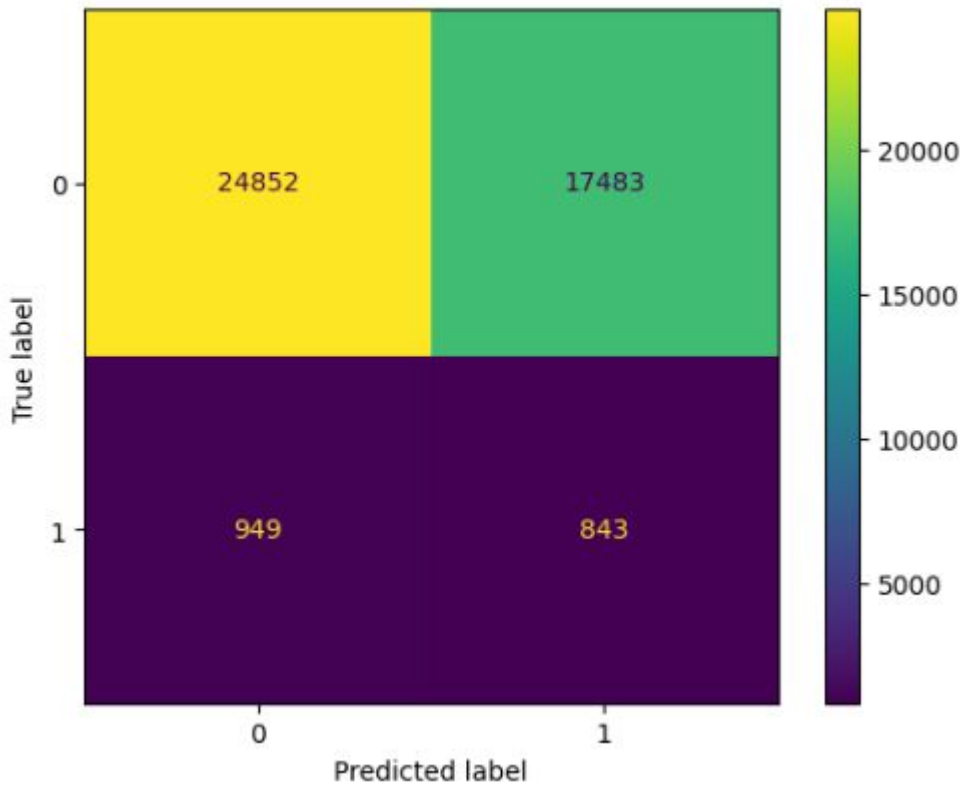
```
[[37542  4793]
 [ 1556   236]]
              precision    recall  f1-score   support

           0       0.96      0.89      0.92     42335
           1       0.05      0.13      0.07      1792

    accuracy                           0.86     44127
   macro avg       0.50      0.51      0.50     44127
weighted avg       0.92      0.86      0.89     44127
```

Métodos usados/Modelos

2. Decision Tree



Decision Tree con SMOTE

```
[[24852 17483]
 [  949   843]]
              precision    recall  f1-score   support

           0       0.96      0.59      0.73     42335
           1       0.05      0.47      0.08      1792

    accuracy                           0.58     44127
   macro avg       0.50      0.53      0.41     44127
weighted avg       0.93      0.58      0.70     44127

AUC: 0.5848
```

#LaUISqueQueremos

Conclusiones

- Las estrategias para el desbalance (Cost-Sensitive Learning, SMOTE) lograron un recall de la clase minoritaria, pero con un sacrificio en la precisión o el recall de la clase mayoritaria.
- Las características disponibles en el dataset no contienen suficiente información discriminatoria para distinguir de manera confiable entre la presencia o ausencia de cáncer de pulmón
- Superposición de clases, no hay distinción entre pacientes con cáncer o sin cáncer de pulmón.
- El diagnóstico de cáncer en la práctica clínica va mucho más allá de los antecedentes, la exposición ambiental o los hábitos. Requiere una combinación de pruebas físicas, análisis de laboratorio y estudios de imagen, lo que subraya la complejidad de clasificar a un paciente basándose únicamente en datos de historial.

Universidad Industrial de Santander

UIS 1948 2025

¡Gracias!