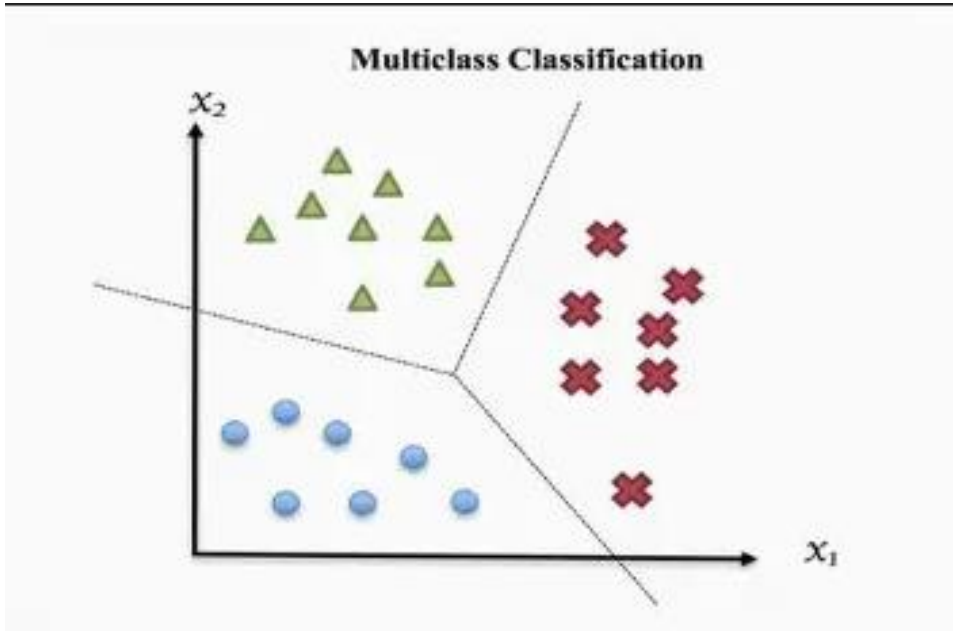
ML_Lung_Cancer

David Torosyan Petrosyan

¿Cuales son tus hábitos?

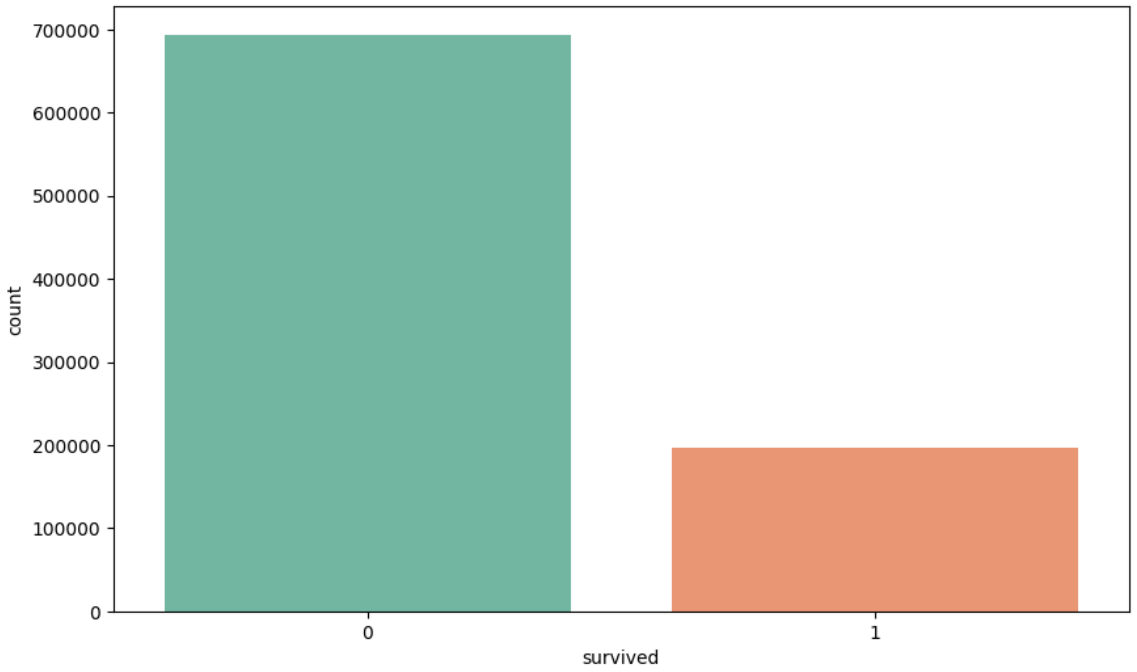
- Teniendo un historial medico podemos predecir hasta si puedes sobrevivir de un cancer de pulmón.
- ¿Fumas,tienes hipertensión,has tenido cancer de pulmón previamente, en que país vives?
- Con unos cuántos factores más podemos predecir si sobrevivirás o no



Base de Datos

- Tenemos una base de datos con 890000 pacientes y 17 características
- Analizaremos la variable que queremos predecir y las características más importantes

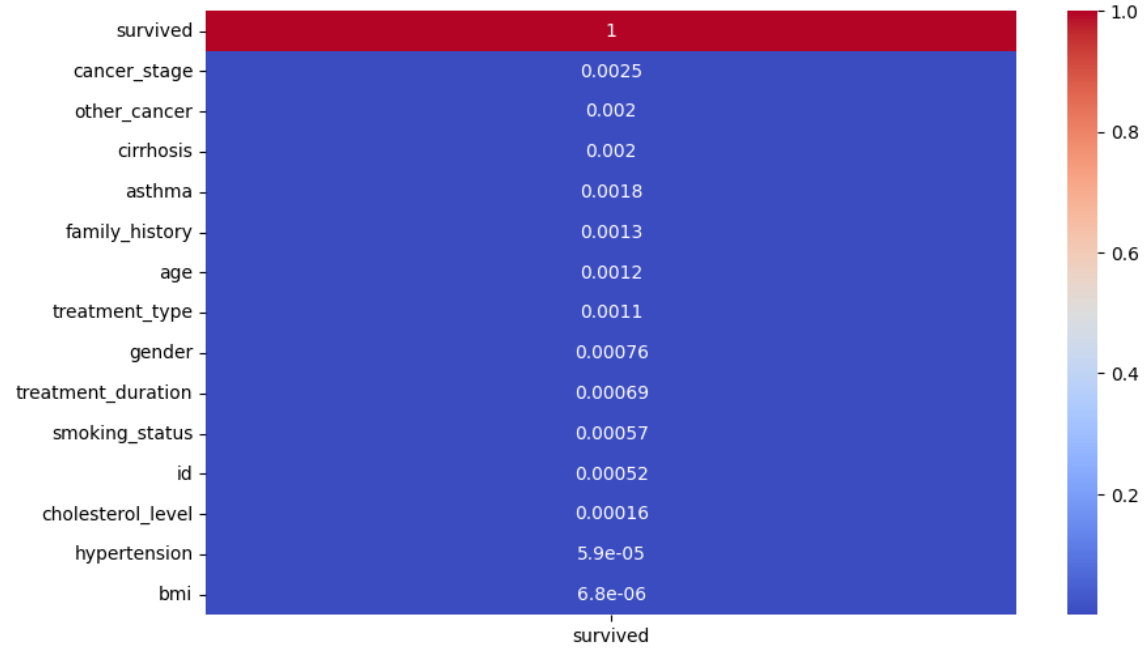
Variable target o de respuesta



Se puede observar que hay un gran desbalanceo

Aumentaremos los casos de la clase minoritaria sintéticamente para darle mayor rigor e importancia

Correlaciones target-variables predictoras



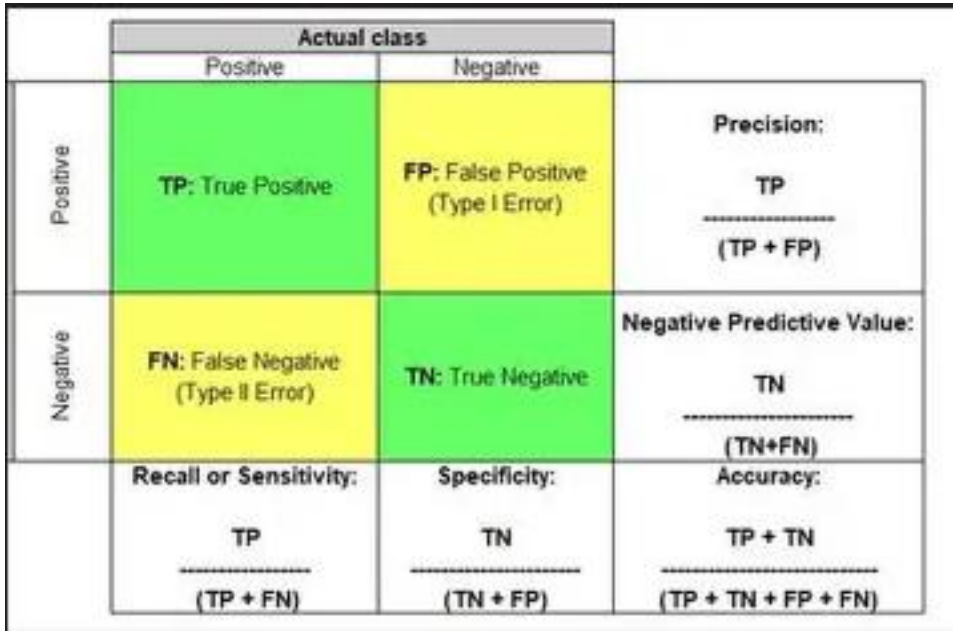
No tenemos grandes rasgos de correlaciones entre las variables lo cual nos dificultará la precisión en las predicciones

Clasificación de las variables importantes

- Las variables como country, diagnosis_date y end_treatment_date las eliminaremos porque no nos aportan mucha información.
- Variables como smoking_status,cancer_stage que nos muestran correlaciones con la target muy bajas y unas distribuciones muy dispersas las incluiré en la lista de variables importantes para poder analizar y completar el modelo.

¿Cómo sabemos que nuestro modelo es el correcto?

- Nos han pedido que nos centremos en la cantidad de aciertos que podemos hacer a la hora de predecir que una persona sobrevivirá.
- Por lo que utilizaremos la métrica de evaluación de precisión.



	Actual class: Positive	Actual class: Negative	
Positive	**TP:** True Positive	**FP:** False Positive (Type I Error)	**Precision:** TP / (TP + FP)
Negative	**FN:** False Negative (Type II Error)	**TN:** True Negative	**Negative Predictive Value:** TN / (TN+FN)
	Recall or Sensitivity: TP / (TP + FN)	**Specificity:** TN / (TN + FP)	**Accuracy:** TP + TN / (TP + TN + FP + FN)

Modelo de entrenamiento

- LGMBClassifier como modelo estrella
- Para la clase mayoritaria (0), el modelo alcanza una precisión del 78% y un recall del 98%, indicando una alta capacidad para identificar no sobrevivientes.
- Sin embargo, para la clase minoritaria (1), el recall es extremadamente bajo (2%), con una precisión del 23% y un F1-score de 0.03, lo que refleja una pobre detección de sobrevivientes.

	precision	recall	f1-score	support
0	0.78	0.98	0.87	138799
1	0.23	0.02	0.03	39201
accuracy			0.77	178000
macro avg	0.50	0.50	0.45	178000
weighted avg	0.66	0.77	0.69	178000

Conclusiones

- Nuestro modelo no ha llegado a ser lo esperado, podríamos pensar en la poca correlación entre las variables predictoras y la target
- Necesitaremos más información y mas casos para poder hacer un análisis exhaustivo y conciso