

# ERROR 404

**Healthcare Dataset**



# OBJETIVO DEL PROYECTO

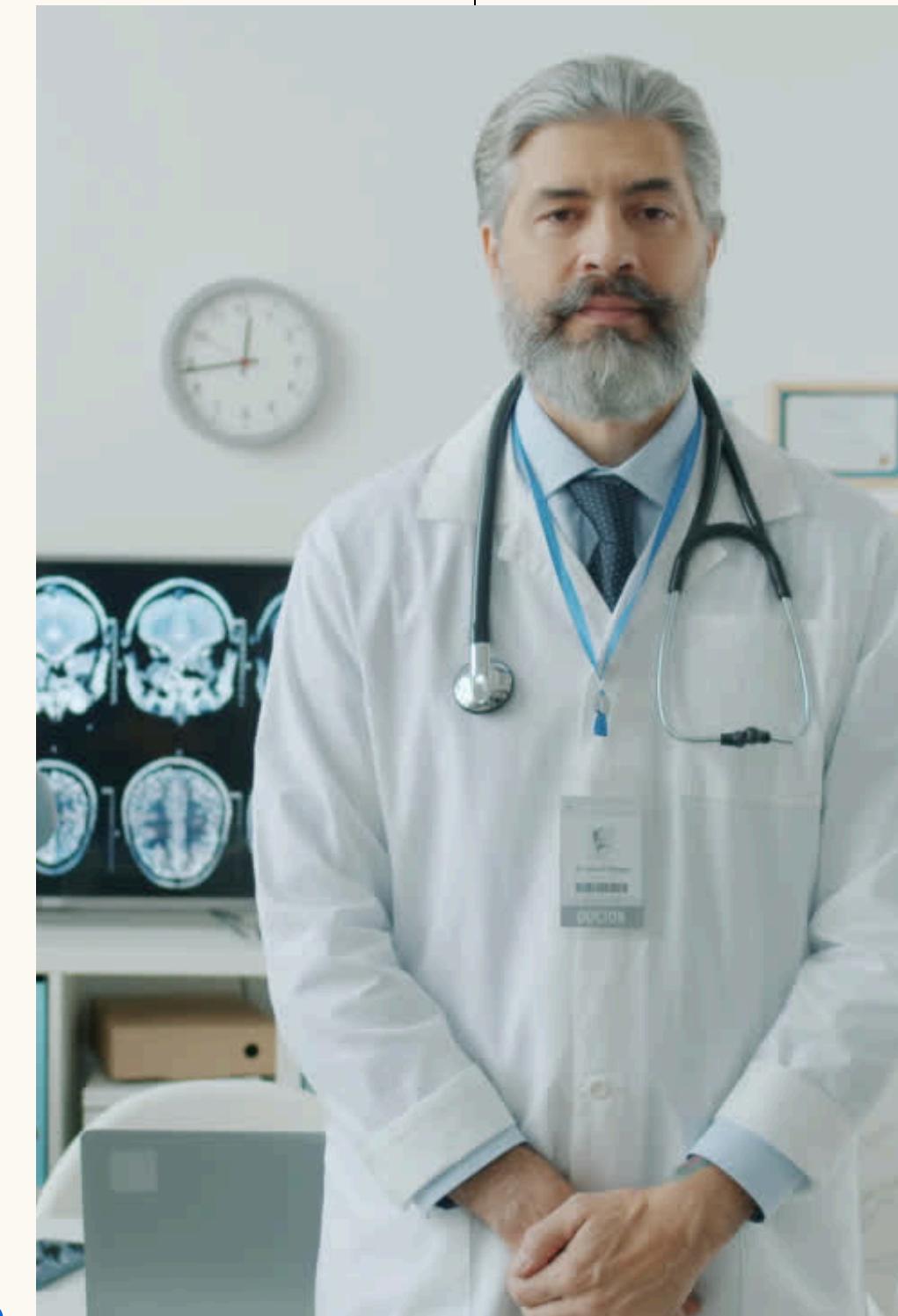
Identificar qué factores influyen en que un paciente tenga una estadía hospitalaria prolongada >7 días.

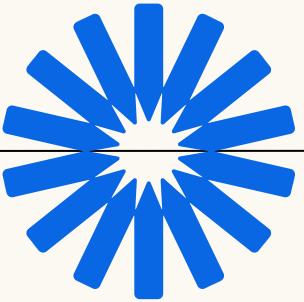
Comprender qué factores determinan las estadías hospitalarias prolongadas es de suma importancia. Un modelo predictivo confiable puede ayudar a optimizar la asignación de camas y recursos hospitalarios.



# TIPO DE PROBLEMA

Creamos una variable nueva binaria, esta es la variable objetivo, denominada Stay, indica si el paciente tuvo una estadía hospitalaria prolongada o no. Se considera como prolongada toda internación que haya superado los siete días (valor 1), y como corta toda internación igual o menor a siete días (valor 0).





# DESCRIPCION DE ALGUNAS VARIABLES

## Edad y genero

-18 o más

-Hombre - Mujer

## Medical condition

- Asma
- Cancer
- Diabetes
- Artritis

ABO (A, B, AB, O) y el sistema Rh (positivo o negativo)

## Blood type

- Anormal
- Normal
- Inconclusa

**Test result  
(variable target sugerida)**

## Admission type

- Por elección
- Urgencia
- Emergencia

# ALGUNAS HIPOTESIS



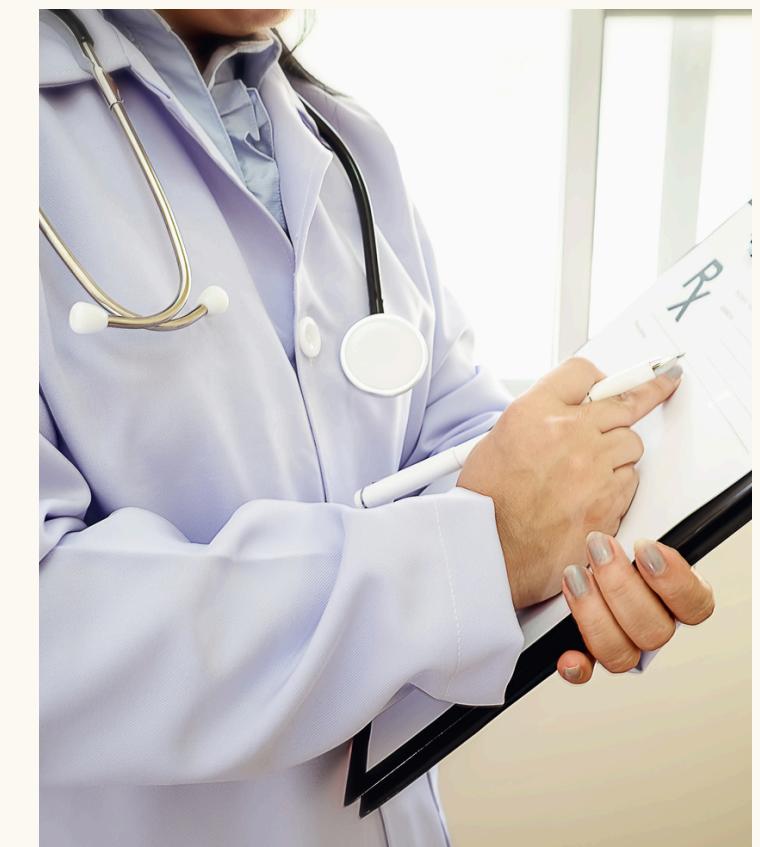
## Severity of Illness

A mayor severidad del cuadro clínico, mayor será la duración de la internación.



## Age

Pacientes mayores tienden a tener internaciones más prolongadas.

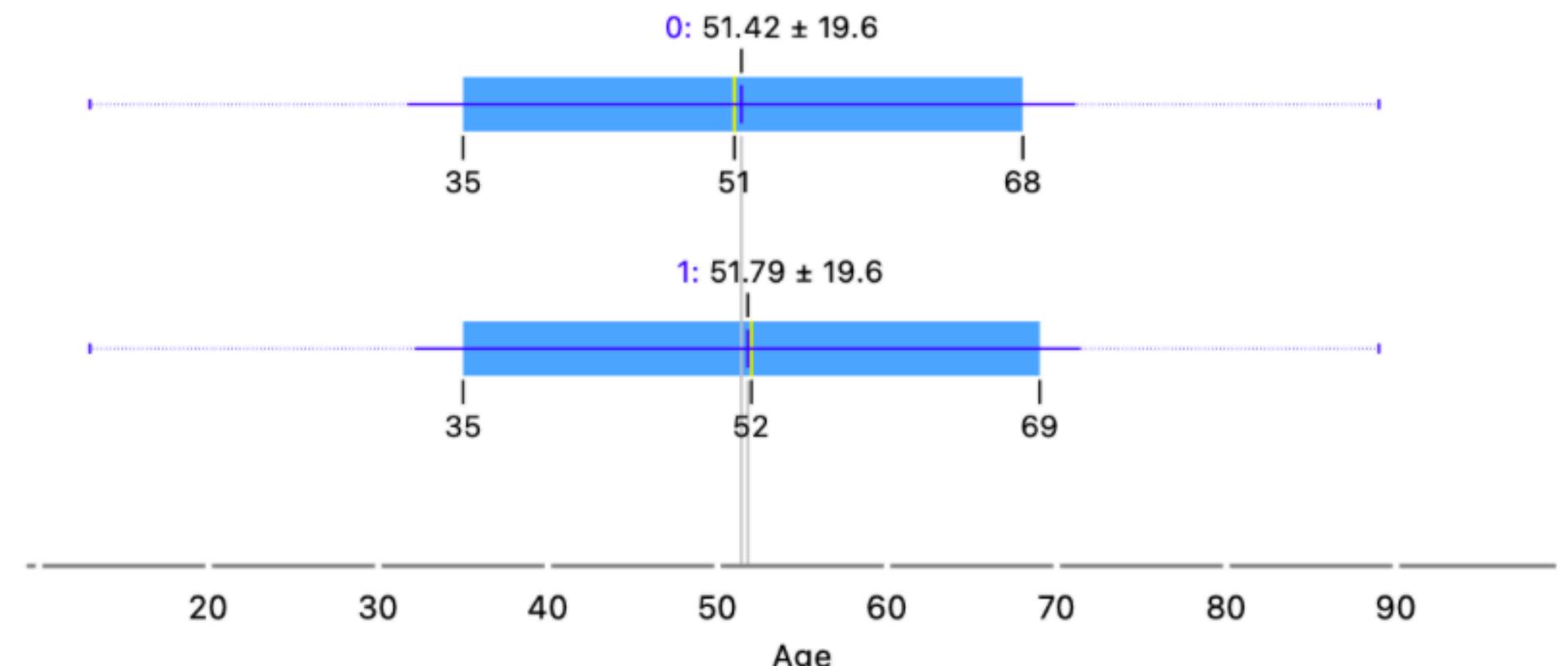


## Type of Admission

Las admisiones de emergencia están más asociadas con estadías prolongadas que las admisiones rutinarias.

# ANALISIS EXPLORATORIO

**Gráfico 1:** Distribución de Edad según tipo de estadía hospitalaria

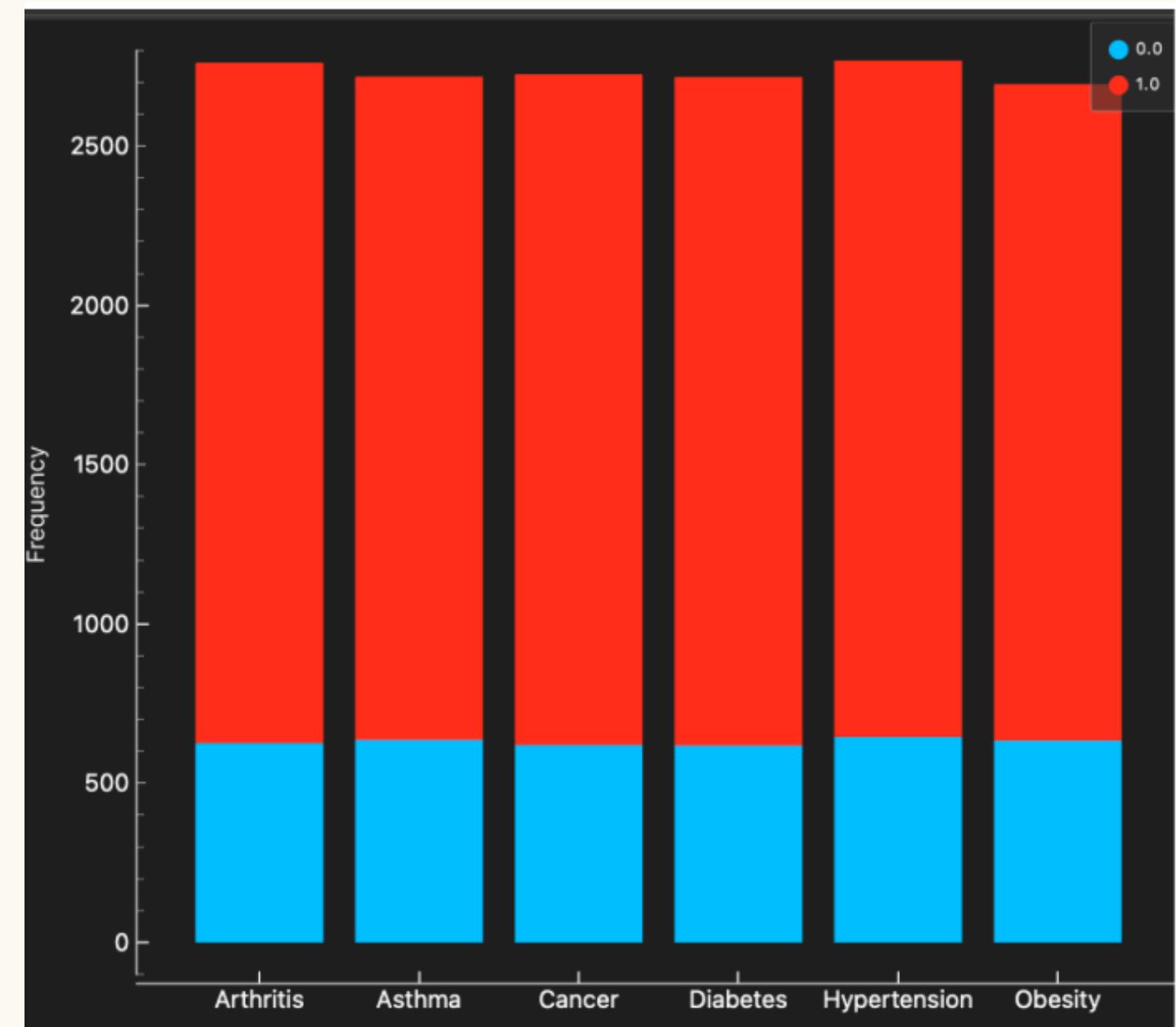


Student's t: 2.103 (p=0.036, N=55500)

## Gráfico 2: Condición médica y su relación con la duración de la internación

- En todos los grupos de condición médica, la proporción de pacientes con estadía prolongada es mayoritaria.
- No se observa una diferencia significativa entre condiciones: todas tienen proporciones visualmente similares.
- Incluso enfermedades más críticas como cáncer o diabetes no muestran una diferencia clara frente a enfermedades más manejables como asma o artritis.

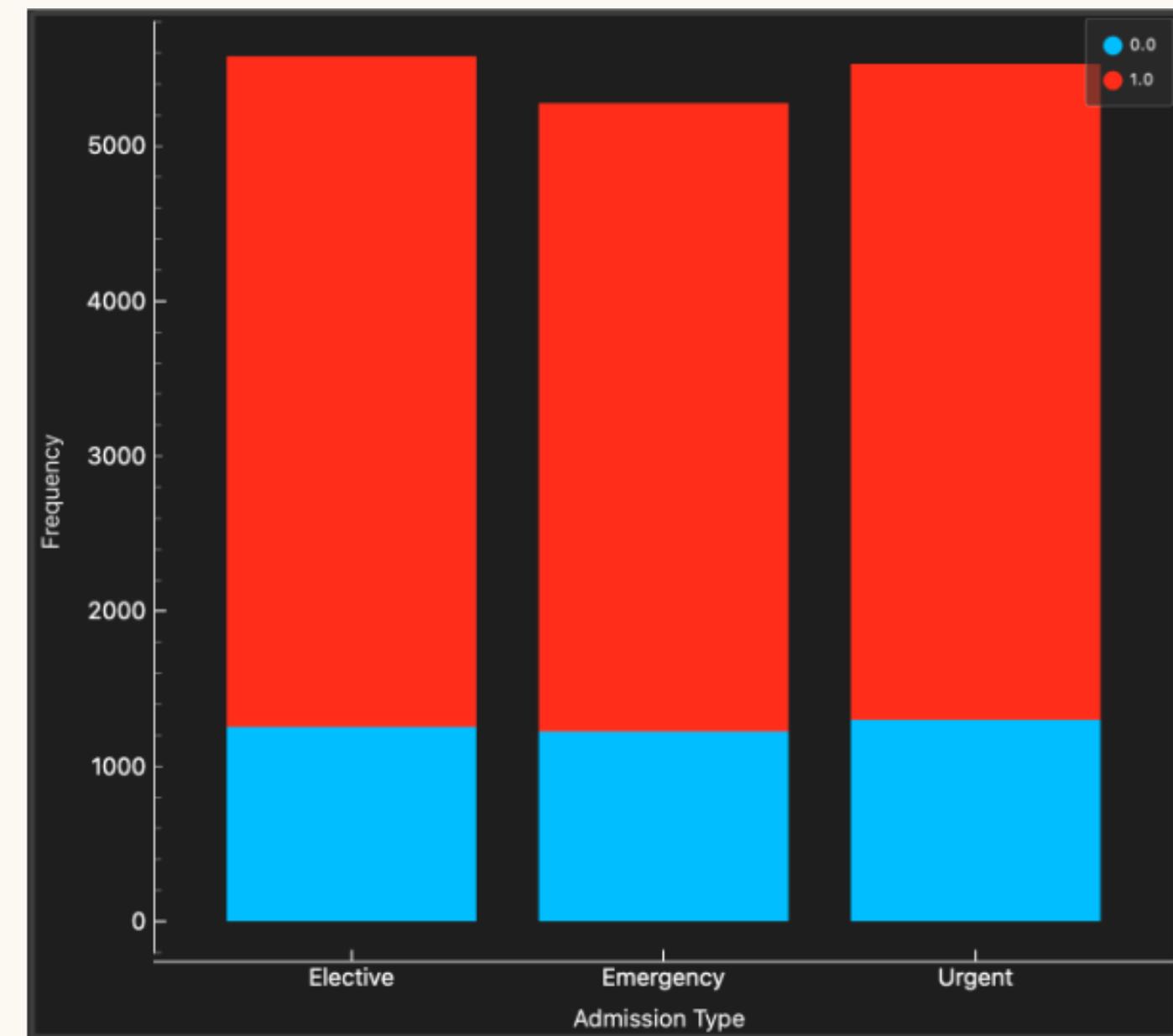
● estadía prologada    ● estadía corta-normal



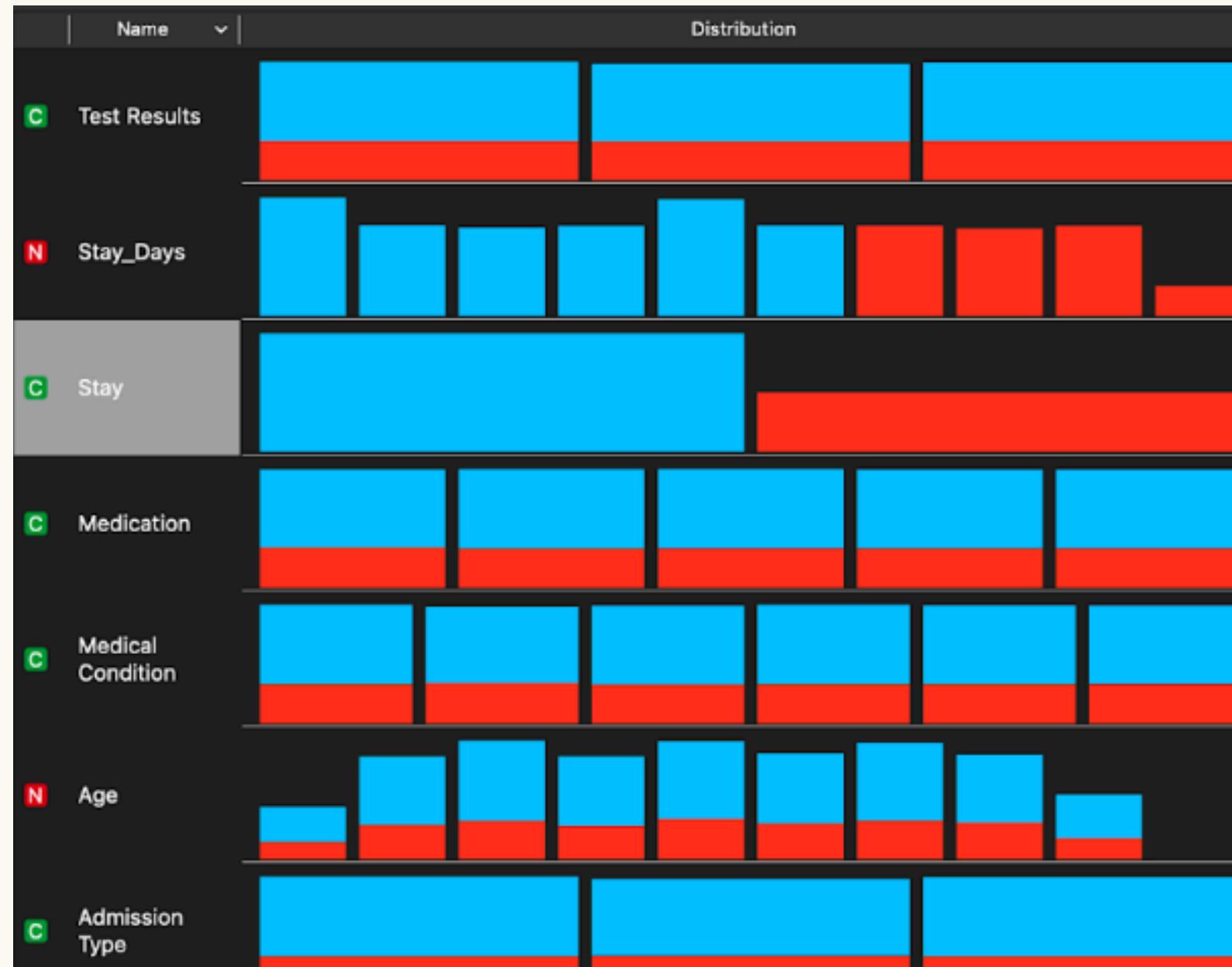
### Gráfico 3: Tipo de admisión y su relación con estadías prolongadas

- En los tres tipos de admisión, la mayoría de los pacientes permanecen más de 7 días internados.
- Sin embargo, se observa que las admisiones electivas presentan ligeramente más casos de estadías cortas en comparación con urgencias o emergencias.

● estadía prologada ● estadía corta-normal



# ESTADÍSTICAS DESCRIPTIVAS DE VARIABLES NUMÉRICAS



● estadía prolongada

● estadía corto-normal

- 1. Test Results:** Los resultados normales o anormales no muestran diferencias en la duración de internación, por lo que no aportan valor predictivo claro.
- 2. Stay:** Existe un leve desbalance hacia estadías prolongadas (63,7%), pero es manejable para modelos de clasificación.
- 3. Medication:** Las proporciones de estadías largas y cortas son similares en todos los medicamentos, indicando bajo poder predictivo individual.
- 4. Medical Condition:** Ninguna condición médica muestra diferencias claras en la duración de internación, posiblemente por una codificación simplificada.
- 5. Age:** La edad no parece influir significativamente en la duración de internación, con proporciones equilibradas en todos los rangos.
- 6. Admission Type:** Es la única variable con diferencia clara: las admisiones electivas se asocian a estadías cortas, y las de urgencia a prolongadas.

# PRIMERAS CONCLUSIONES

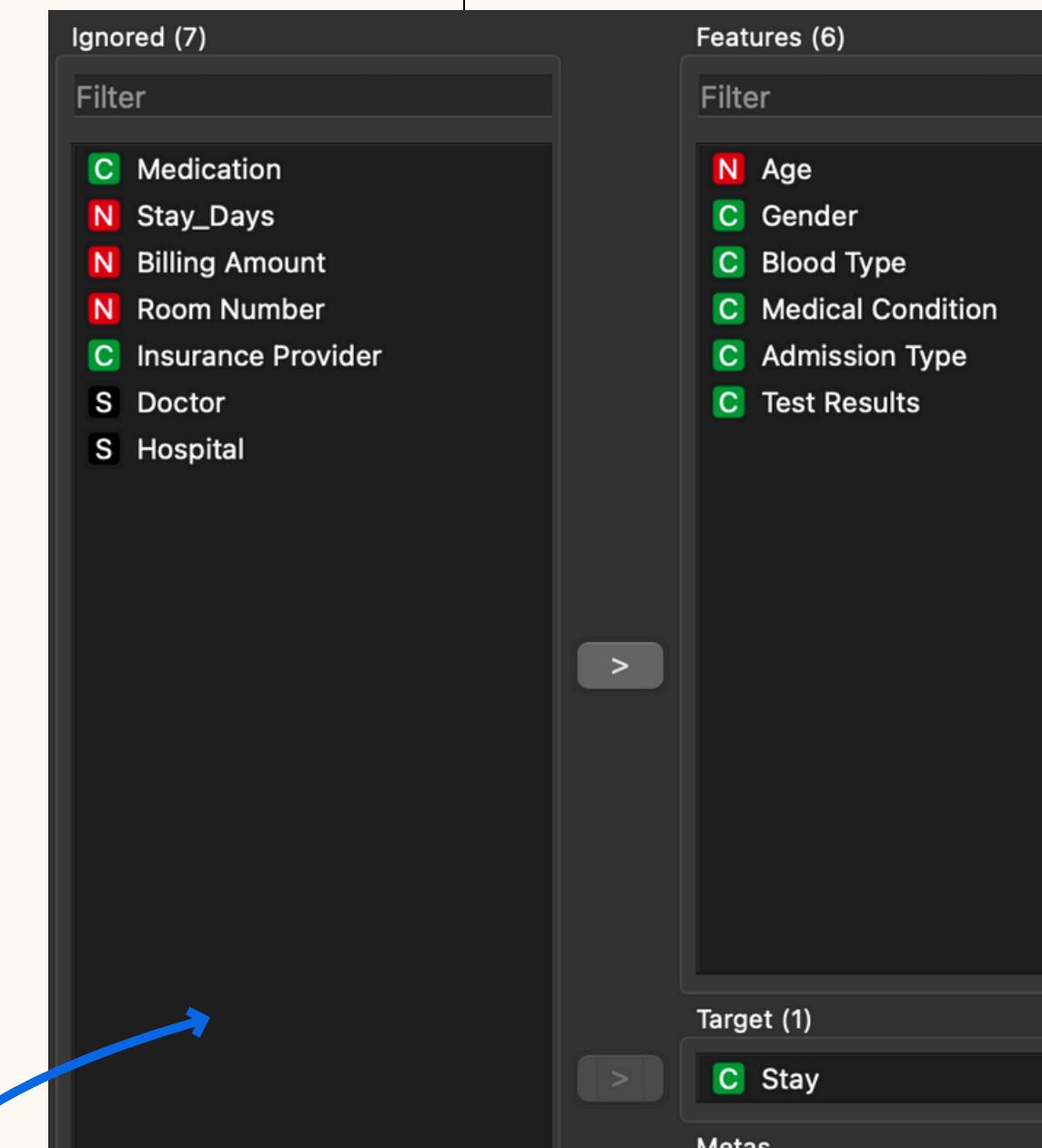
Tras realizar un análisis estadístico y gráfico de las variables incluidas en el dataset, se concluye que **los resultados obtenidos hasta el momento no son suficientemente sólidos como para alimentar un modelo predictivo confiable.** La mayoría de las variables no muestran diferencias significativas



# PREPROCESAMIENTO DE LOS DATOS

**1. Selección inicial de variables:** Se utilizó el nodo select columns de para incluir solo variables que aportaban al modelo y definir la variable objetivo.

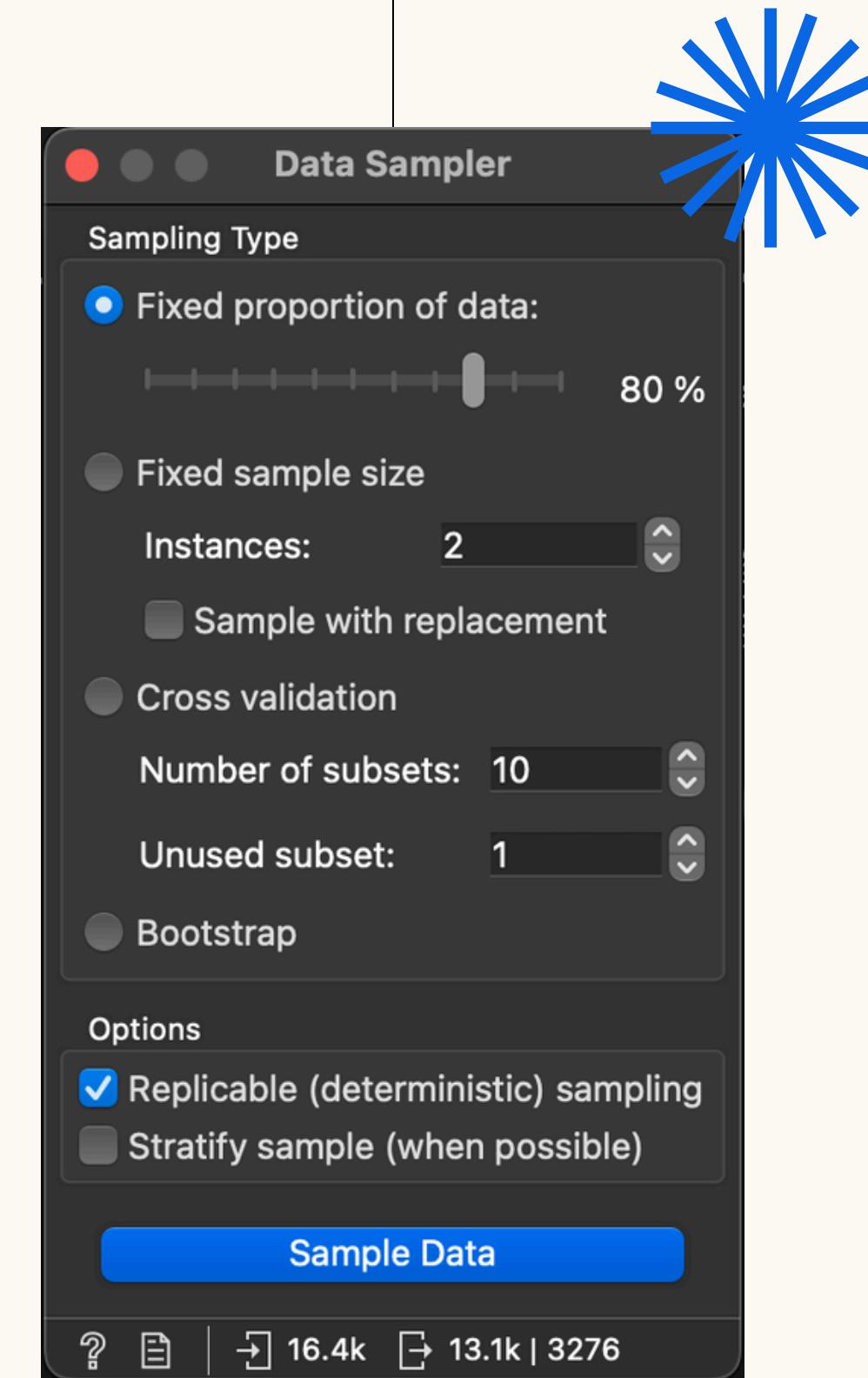
**2. Definir la variable objetivo:** Se eligió Stay como target, construido a partir de Stay\_Days.



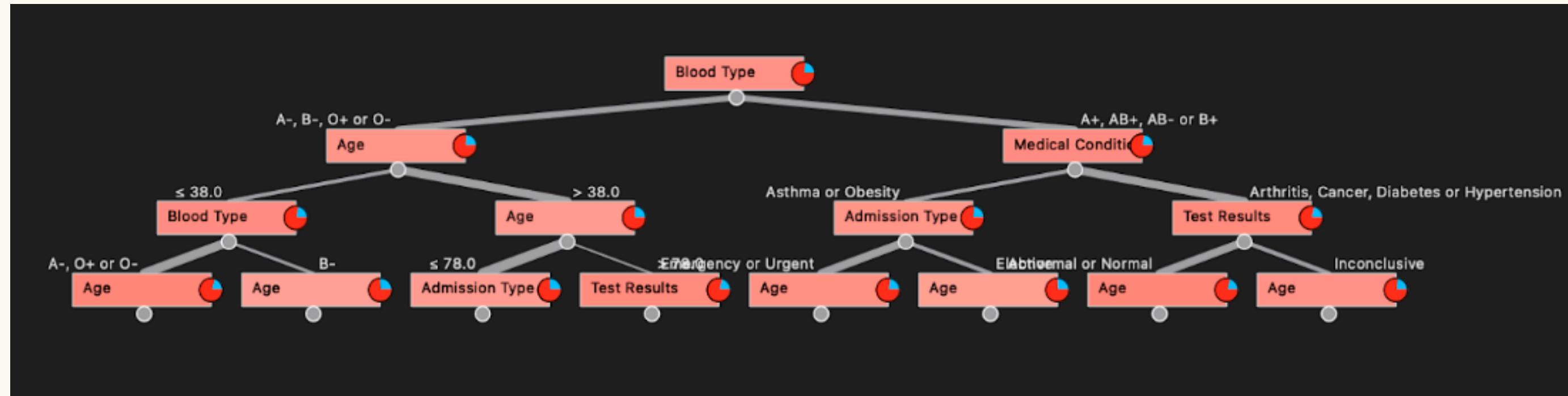
# PREPROCESAMIENTO DE LOS DATOS

**1. División del dataset:** Se utilizó el nodo **Data Sampler** para separar el 80% de los datos para entrenamiento y el 20% para prueba.

**2. Muestra replicable:** Se activó la opción “Replicable sampling” para asegurar consistencia en los resultados al repetir el análisis.



# PRIMER MODELO - ARBOL DE DECISIÓN



Modelo entrenado con las variables definidas anteriormente para predecir el tipo de estadía

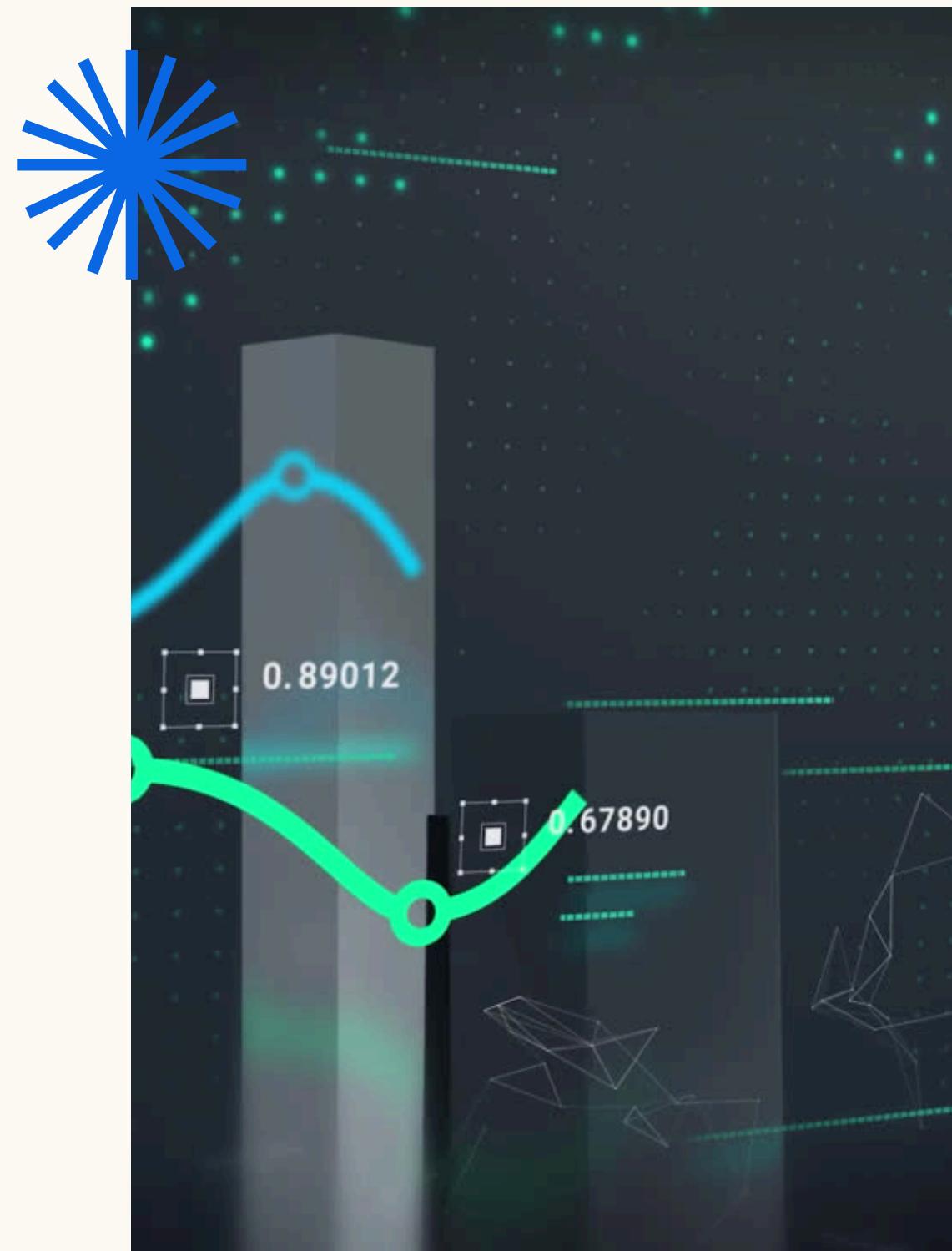
- El árbol clasifica a casi todos los pacientes como de estadía prolongada, lo que sugiere un posible sesgo del dataset o baja capacidad de discriminación.

Aunque ofrece reglas claras, no diferencia bien entre estadías largas y cortas.

# SEGUNDO MODELO- REGRESIÓN LOGÍSTICA

Se entrenó un modelo con configuración predeterminada en Orange.

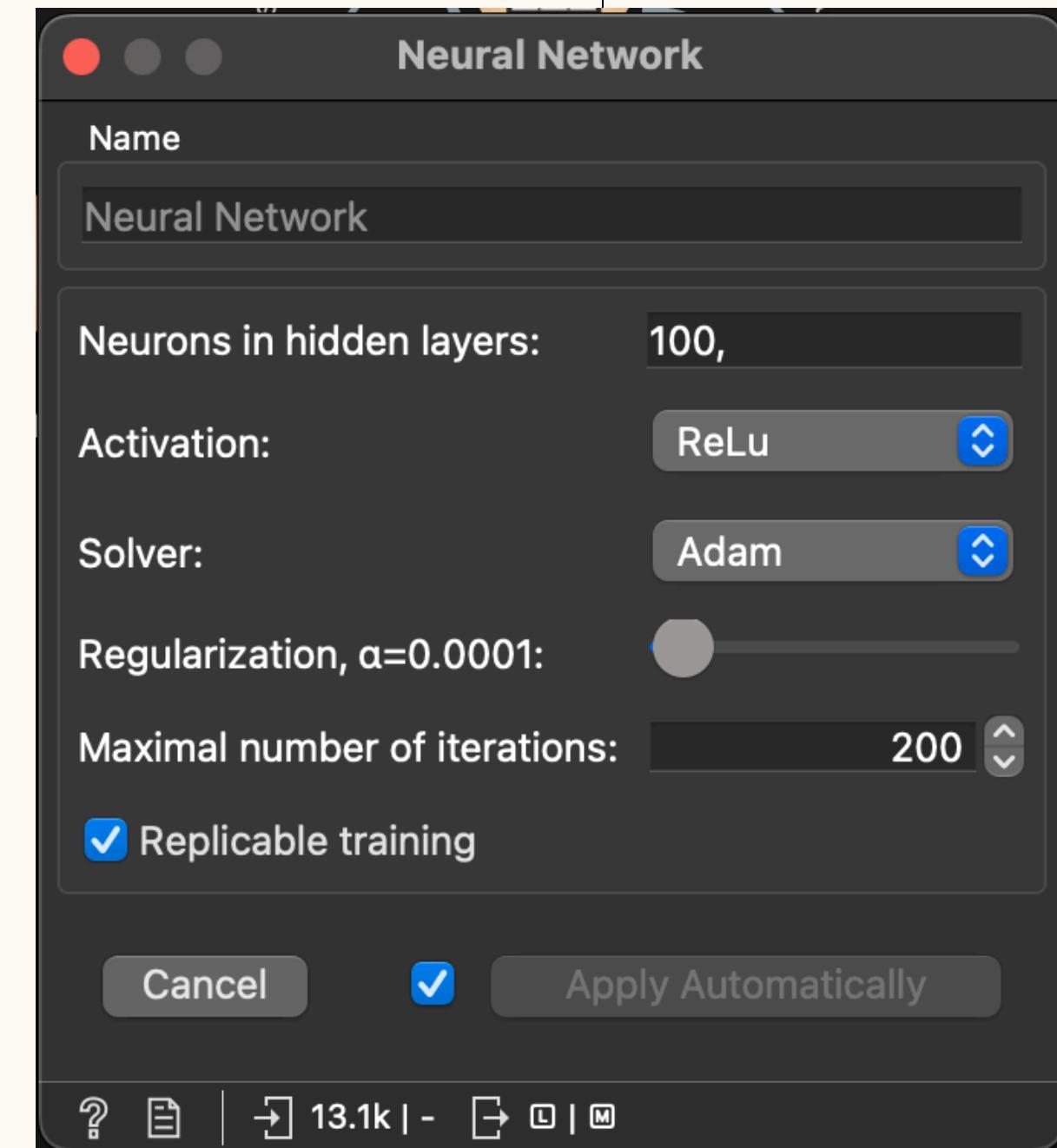
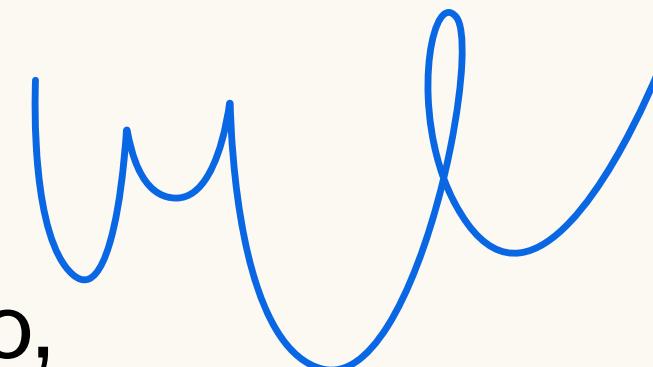
Las variables no mostraron una relación lineal clara con el target, por lo que el modelo no alcanzó un desempeño destacado.



# RED NEURONAL - TERCER MODELO

Se entrenó una red neuronal mult capa con los siguientes ajustes:

La red mostró buen desempeño, pero su complejidad no aporto valor diferencial.



# TEST AND SCORE

Model	AUC	CA	F1	Prec	Recall	MCC
Tree (1)	0.496	0.621	0.632	0.644	0.621	-0.006
Logistic Regression	0.503	0.771	0.671	0.594	0.771	0.000
Neural Network	0.484	0.743	0.674	0.644	0.743	-0.005

	Tree (1)	Logistic Regression	Neural Network
Tree (1)		0.360	0.655
Logistic Regression	0.640		0.752
Neural Network	0.345	0.248	

- **AUC:** Ningún modelo supera claramente el azar (0.5). La regresión logística tuvo la mejor AUC (0.503), pero sigue siendo muy baja.

- **Accuracy (CA):** Resultados entre 0.621 y 0.771, pero poco confiables por el desbalance de clases.

- **F1 Score:** Moderadamente alto en todos (0.63–0.67), reflejando sesgo hacia estadías prolongadas.

- **Precisión:** Más alta en el árbol y red neuronal (0.644).

- **Recall:** Máximo en regresión logística (0.771).

- **MCC:** Cercano a cero en todos los casos → baja capacidad real de predicción.

Los modelos no logran discriminar bien entre estadías cortas y prolongadas, probablemente por un dataset desbalanceado o poca información predictiva.

# MATRIZ DE CONFUSION

**Red Neuronal:** El modelo muestra un alto desempeño para detectar estadías prolongadas, pero falla casi por completo al identificar estadías cortas. Esto evidencia un fuerte sesgo hacia la clase mayoritaria, con una muy baja especificidad. Es probable que este comportamiento esté influido por el desbalance de clases en los datos.

**Árbol de Decisión:** Este modelo logra un mayor equilibrio entre clases. Identifica correctamente el 72,7% de las estadías prolongadas y el 26,6% de las cortas. Sin embargo, aún presenta errores importantes, especialmente en los falsos negativos, lo cual puede ser crítico en un contexto hospitalario.

**Regresión Logística:** La regresión clasifica todos los casos como estadías prolongadas, ignorando por completo la clase corta. No identifica ningún verdadero negativo ni falso negativo. Esto puede deberse tanto al desbalance de clases como a una mala calibración del umbral de decisión.

**Conclusión:** Ninguno de los tres modelos logra una predicción adecuada de la variable Stay. Todos tienden a favorecer la clase mayoritaria, lo que limita seriamente su utilidad práctica. En especial, la falta de identificación de estadías cortas puede generar problemas en la gestión hospitalaria. Por lo tanto, aplicar estos modelos en su forma actual no resulta recomendable, ya que no ofrecen un valor predictivo confiable.

		Predicted		
		0.0	1.0	$\Sigma$
Actual	0.0	36	715	751
	1.0	127	2398	2525
		$\Sigma$	163	3113
				3276

		Predicted		
		0.0	1.0	$\Sigma$
Actual	0.0	200	551	751
	1.0	689	1836	2525
		$\Sigma$	889	2387
				3276

		Predicted		
		0.0	1.0	$\Sigma$
Actual	0.0	0	751	751
	1.0	0	2525	2525
		$\Sigma$	0	3276
				3276

# NUEVO ENFOQUE DEL PROYECTO

Ante el bajo rendimiento al predecir la estadía hospitalaria, se probó un nuevo enfoque: predecir los resultados clínicos **Test\_Results** usando las demás variables del dataset (edad, diagnóstico, tipo de admisión, etc.).

**Problema Detectado:** Los resultados también fueron insatisfactorios, con métricas similares al primer intento. Esto sugiere que el conjunto de datos no tiene suficiente información para realizar predicciones clínicas confiables.



# CONCLUSIONES A PARTIR DE TEST AND SCORE

**Test & Score:** Se evaluó el modelo con métricas estándar: AUC, Accuracy, Precision, Recall, F1-score y MCC.

- AUC = 0,510 → Apenas mejor que el azar.
- Accuracy ≈ 34 %,
- F1-score ≈ 34 %,
- MCC = -0,006 → El modelo no generaliza bien ni encuentra patrones claros.

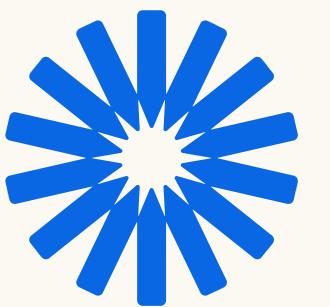
## Matriz de Confusión:

- Sobrepredicción de "Abnormal": El modelo tiende a etiquetar demasiados casos como "Abnormal", aunque no lo sean.
- Poca identificación de pacientes sanos: Solo 266 de 1.061 pacientes realmente "Normales" fueron correctamente clasificados.
- Alta confusión entre "Inconclusive" y "Abnormal": El modelo confunde muchos casos "Inconclusive" como "Abnormal", mostrando poca precisión entre estas clases.

Esto indica que las variables del dataset no permiten predecir con precisión el tipo de resultado clínico (Normal, Abnormal, Inconclusive).

Model	AUC	CA	F1	Prec	Recall	MCC
Neural Network	0.491	0.333	0.331	0.333	0.333	-0.003
Tree	0.510	0.343	0.341	0.342	0.343	0.013
Logistic Regression	0.486	0.326	0.325	0.327	0.326	-0.010

		Predicted			
		Abnormal	Inconclusive	Normal	S
Actual	Abnormal	478	321	316	1115
	Inconclusive	430	380	290	1100
	Normal	441	354	266	1061
Σ		1349	1055	872	3276



**GRACIAS!**

