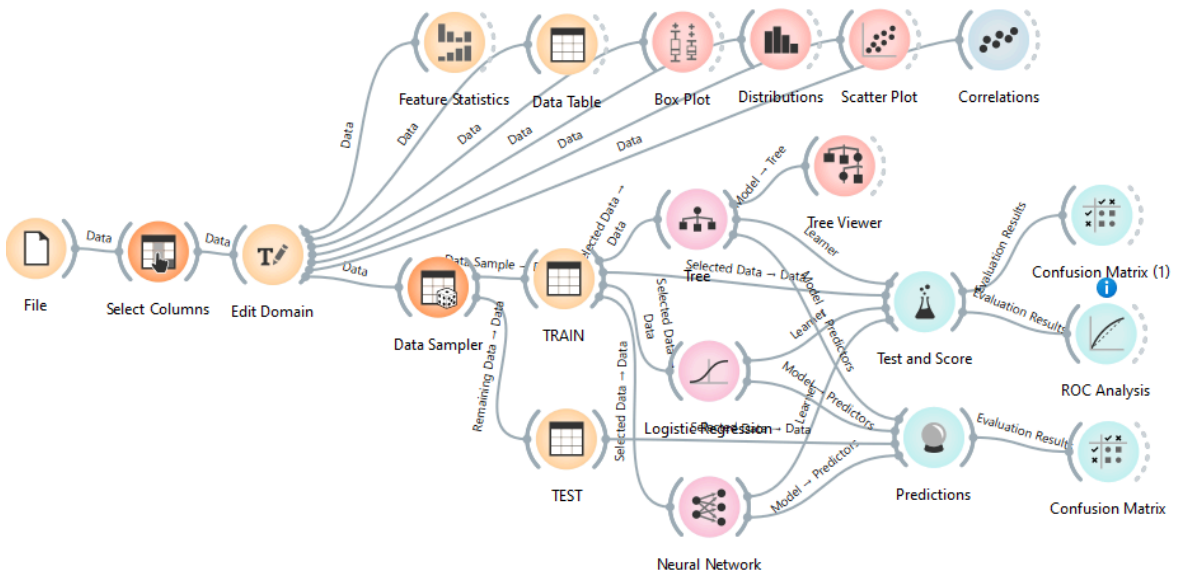
TP GRUPO YARARÁ BIG DATA



DEFINICIÓN DEL PROBLEMA:
En el contexto del sistema de salud, los resultados de pruebas médicas (test results) suelen clasificarse como "Normal", "Abnormal" o "Inconclusive". Considerando que los resultados “abnormal” e “inconclusive” representan una zona de incertidumbre que no aporta información diagnóstica útil, hemos decidido juntar esas clasificaciones bajo el nombre de “abnormal”. Esto lo hacemos con el fin de identificar cuáles son las variables que llevan a que un resultado de un test no normal, siendo que cuando da normal el médico sabe cómo proceder y el resultado contrario trae inconvenientes a la hora de continuar el tratamiento con el paciente.

PROCEDIMIENTO:
El objetivo de este trabajo es analizar las variables de estudio frente al contexto de los “**Test Results**”, comenzando por el análisis exploratorio de datos (**EDA**), viendo si hay valores **atípicos**, **nulos** y la **correlación** entre las variables. Para realizar esto decidimos primero editar los valores (**Edit Domain**) de las variables categóricas a números para poder realizar un análisis más sofisticado. Además, utilizando la herramienta “**Select Columns**” ponemos el target que son, como hemos mencionado, los test results e ignoramos la columna “Room Number” ya que creemos que no aporta valor.

Comenzamos conectando nuestro file (luego de haber editado las columnas y el domain) a un **Data Sampler** para dividir nuestro conjunto de datos en dos subconjuntos: uno de entrenamiento **(Train)** y otro de prueba **(Test)**.

A partir del conjunto de entrenamiento, entrenamos tres modelos distintos: un árbol de decisión **(Tree)**, una regresión logística (**Logistical Regression**) y una red neuronal **(Neural Network)**. Estos tres modelos, junto con el conjunto de Train, se conectan al widget **Test and Score**, que evalúa el rendimiento de los modelos sobre el conjunto de prueba. Desde este widget, generamos dos salidas: una hacia una **Confusion Matrix**, que nos permite

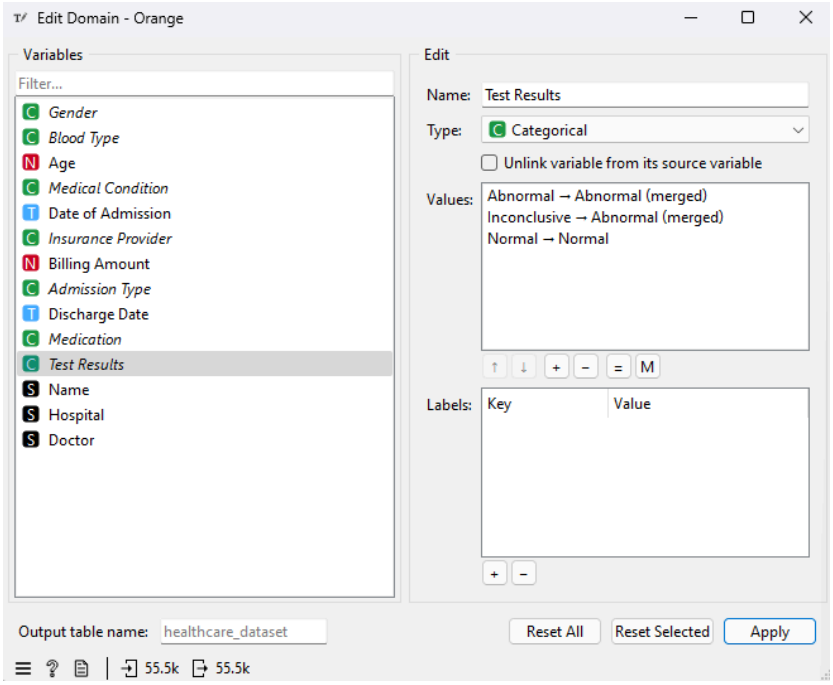
observar los falsos positivos y falsos negativos de cada modelo, y otra hacia un **ROC Analysis**, que visualiza la capacidad de los modelos para discriminar entre clases.

Además, conectamos los tres modelos y el Train al widget **Predictions**, lo que nos permite comparar las predicciones concretas de cada modelo.

Al contrastar las métricas obtenidas en la **Confusion Matrix** con las del widget **Predictions**, evaluamos si hay **overfitting o underfitting**, observando especialmente las diferencias en **Accuracy**.

Finalmente, según el tipo de problema que enfrentamos y los costos asociados a errores de clasificación, analizamos métricas como **Precision**, **Recall** o **F1-score** para determinar cuál resulta más relevante para la toma de decisiones.

PRIMER PASO: EDIT DOMAIN



Pasamos las variables categóricas a numéricas para poder hacer un mejor análisis. Test Result juntamos inconclusive con abnormal. para comparar si dio normal o no.

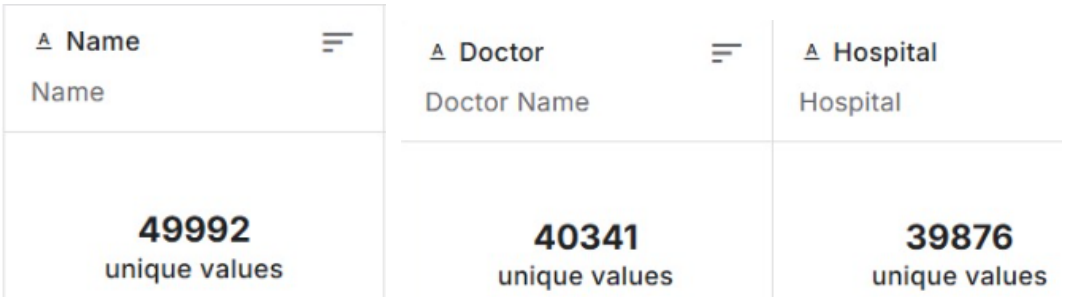
EDA ANÁLISIS EXPLORATORIO DE DATOS:

Billing Amount: que son menores a cero. Creemos que puede haber “Billings” negativos porque hay saldos a favor del paciente.

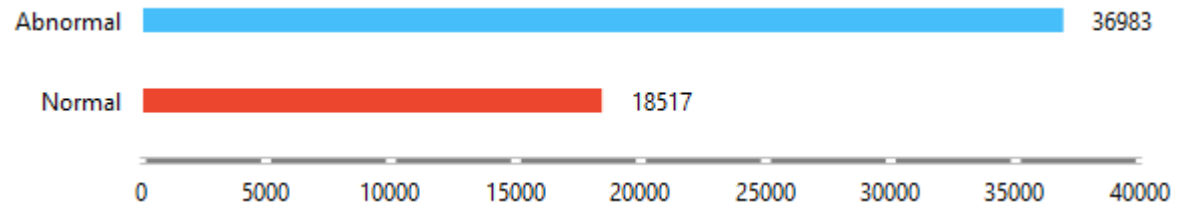
Name y Room Number: no aportan valor

Hospitales: 39.876 valores únicos
Doctores: 40.341 valores únicos

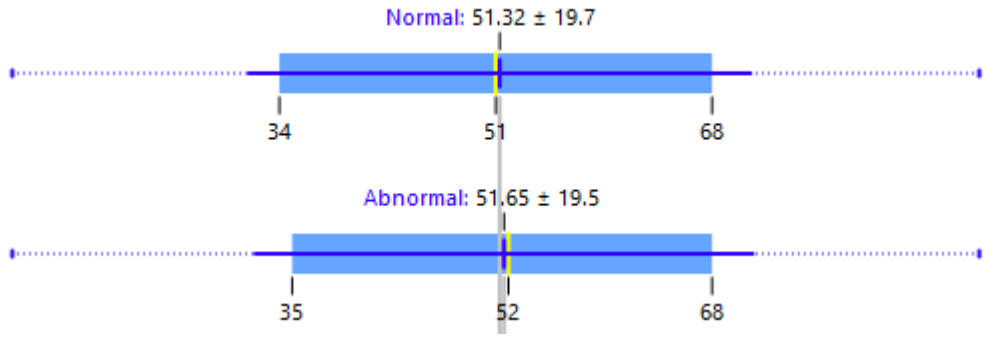
Nombres de personas: 49.992 valores únicos



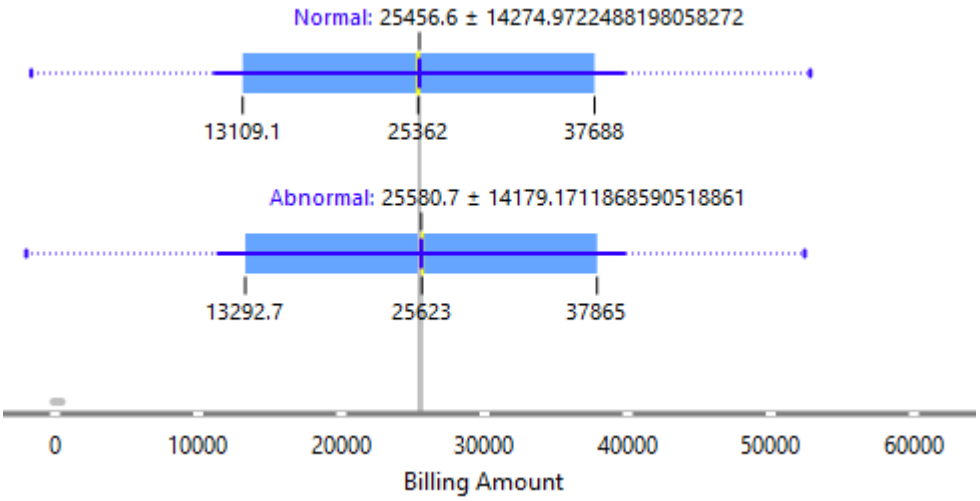
El doble de los test results de los casos resultan en abnormal:



La edad de los pacientes sea el test result que den no varía y no posee outliers:



Lo mismo con los billing amounts:



Notamos que hay billing negativos pero creemos que pueden existir saldos a favor del paciente y eso produce que se vean menores a cero. Hay 108 datos para ser precisos que son negativos

Data Table (1) - Orange

Info
108 instances (no missing data)
11 features
Target with 3 values
3 meta attributes

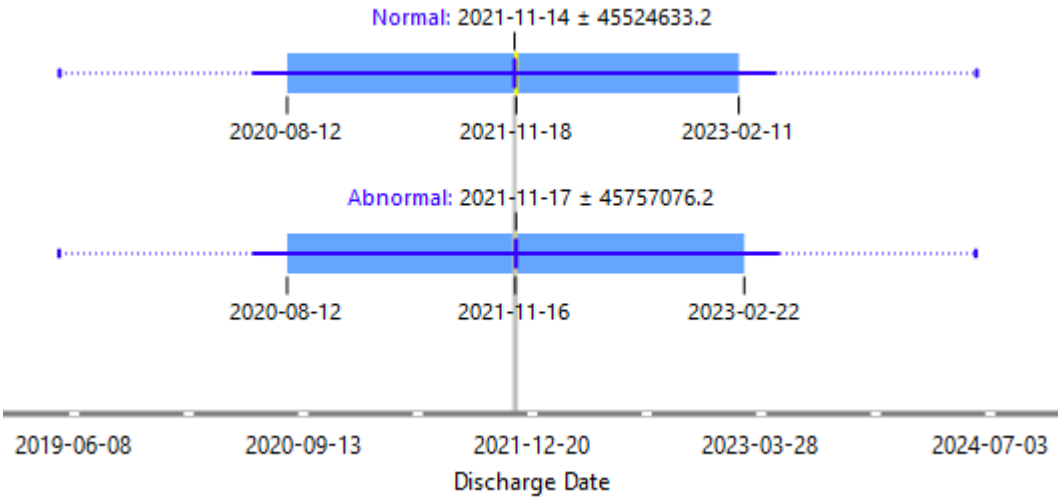
Variables
☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection
☐ Select full rows

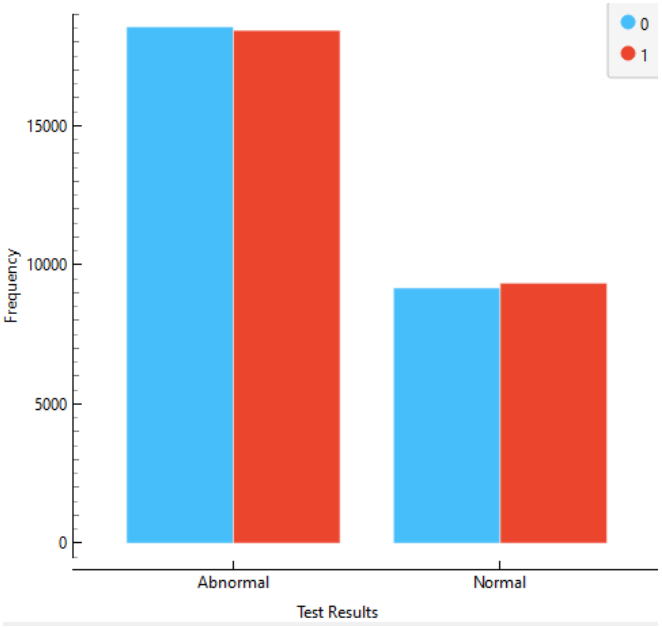
Restore Original Order

	Gender	Blood Type	Medical Conditior	Date of Admissior	nsurance Provide	Billing Amount
1	Female	AB-	Cancer	2019-11-05	Aetna	-502.508
2	Female	AB-	Asthma	2023-02-16	Aetna	-1018.25
3	Male	A+	Hypertension	2021-12-21	Aetna	-306.365
4	Female	O+	Asthma	2021-01-20	Blue Cross	-109.097
5	Female	B-	Diabetes	2021-03-21	Blue Cross	-576.728
6	Male	B+	Diabetes	2023-04-12	Medicare	-135.986
7	Female	B-	Hypertension	2020-04-03	Blue Cross	-370.984
8	Male	AB+	Obesity	2022-06-03	Blue Cross	-1310.27
9	Male	AB+	Arthritis	2022-07-14	Aetna	-692.409
10	Male	AB-	Diabetes	2019-10-13	Blue Cross	-353.865
11	Female	A+	Asthma	2022-06-22	Medicare	-378.961
12	Female	AB+	Diabetes	2019-06-30	Cigna	-367.204
13	Male	AB-	Cancer	2023-03-23	Cigna	-198.284
14	Female	B-	Diabetes	2019-08-21	Cigna	-43.0985
15	Female	A-	Asthma	2019-10-23	Medicare	-857.126
16	Male	B+	Hypertension	2020-01-29	Aetna	-155.082
17	Female	A-	Obesity	2022-04-05	UnitedHealthcare	-211.724
18	Female	B+	Asthma	2020-05-07	Cigna	-656.153
19	Female	B-	Diabetes	2021-02-05	Medicare	-227.995
20	Male	O-	Cancer	2021-04-13	Cigna	-147.072
21	Female	A+	Hypertension	2021-06-09	Blue Cross	-124.756
22	Female	O-	Arthritis	2022-09-15	UnitedHealthcare	-75.8195
23	Female	B-	Arthritis	2022-12-29	Cigna	-26.1125
24	Male	AB+	Cancer	2019-05-30	Blue Cross	-135.719

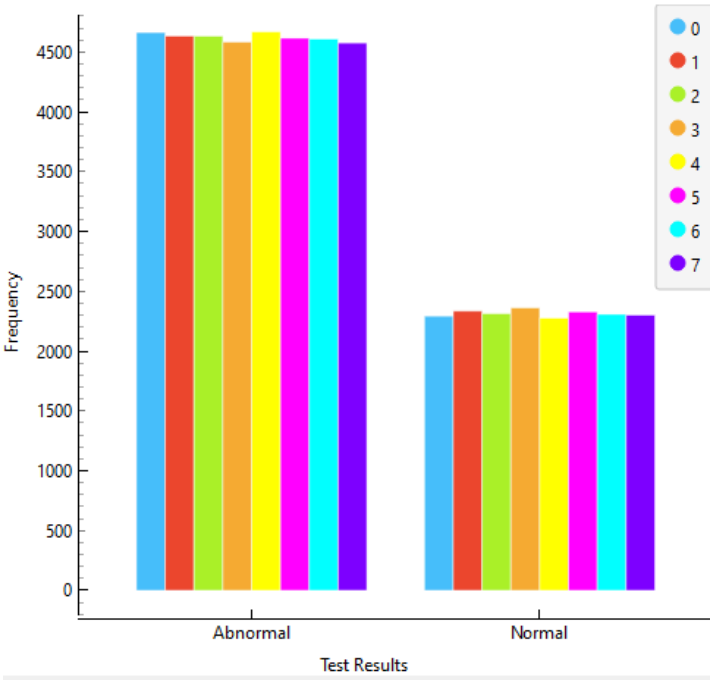
La fecha de admisión y salida del paciente es parecida siendo uno u otro resultado:



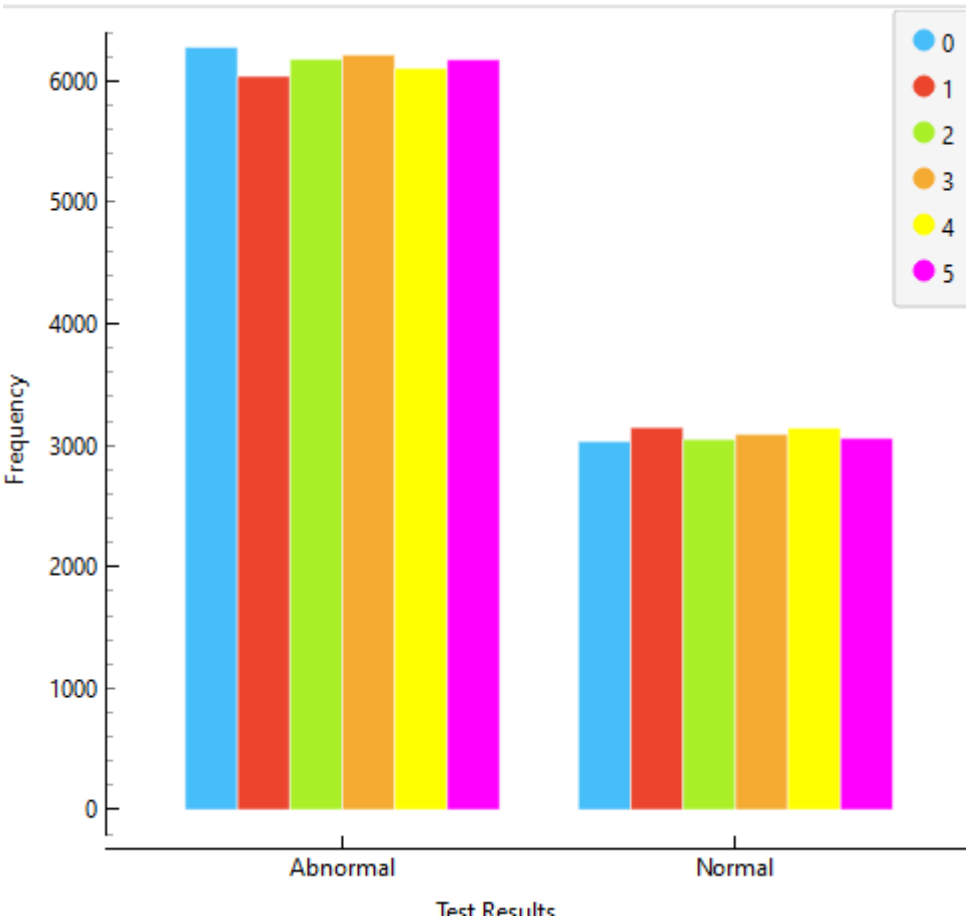
La cantidad de pacientes por género, visto desde cada test result, es similar:



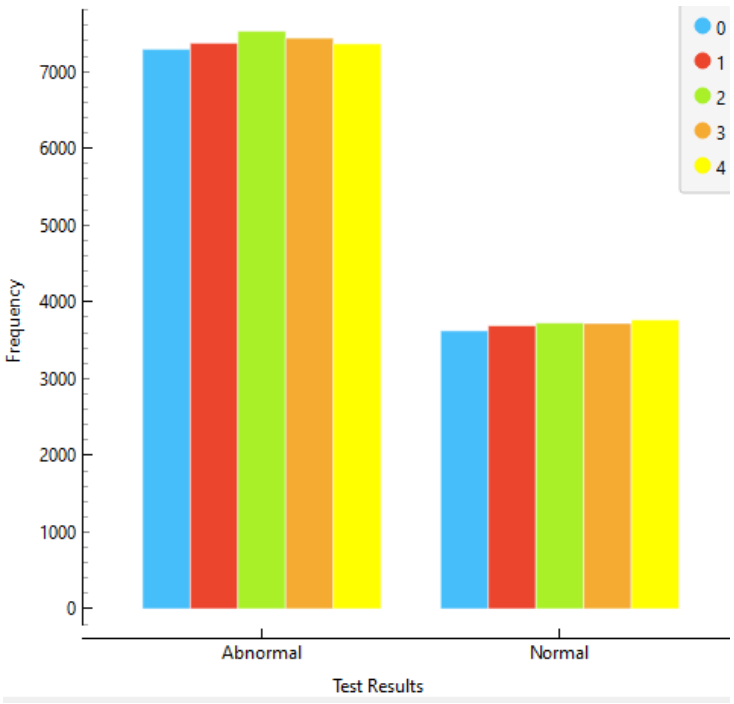
Al igual que los tipos de sangre:



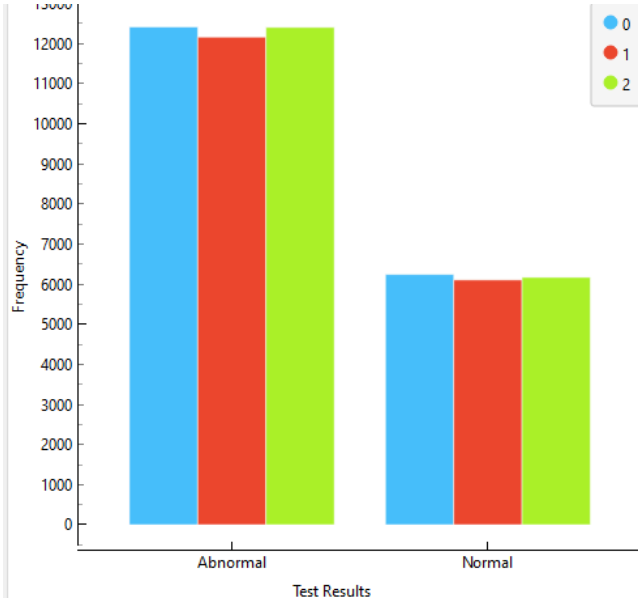
De la misma manera las condiciones médicas de los pacientes se ven en similar porcentaje dentro de cada test result:



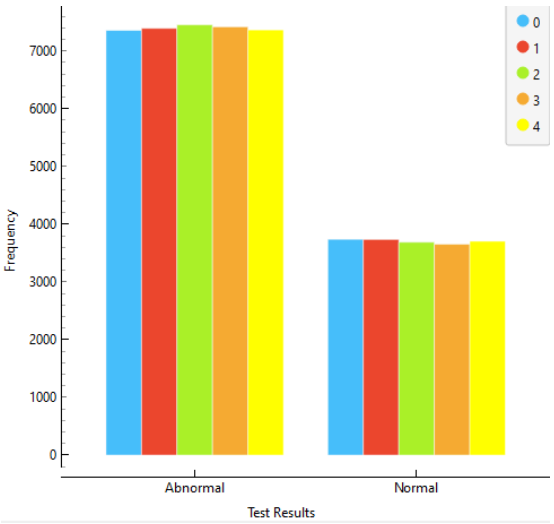
Tambien los seguros médicos:



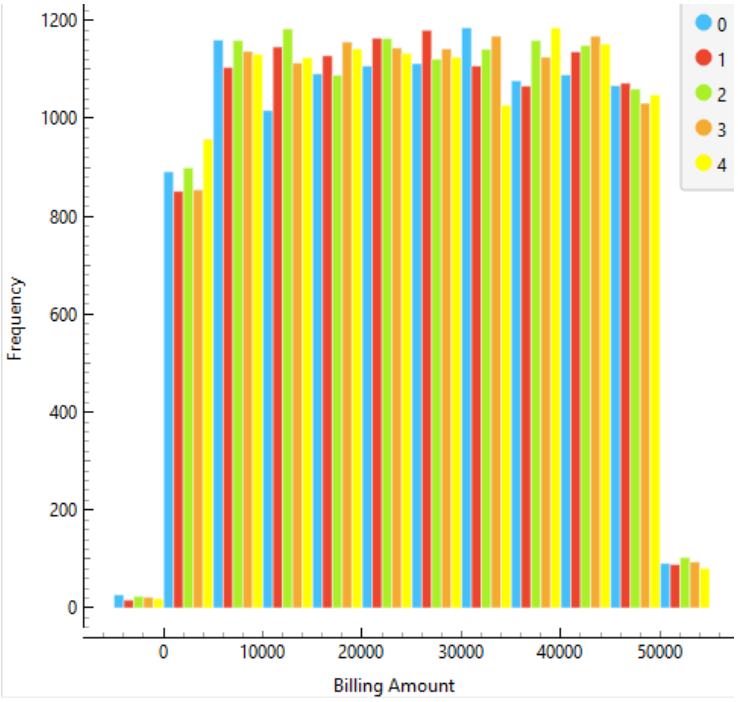
Y el tipo de admisión (emergency, elective o urgent) tampoco varía si resulta normal o anormal el test:



Finalmente, los pacientes que resultan en ambos normal o anormal han utilizado las distintas medicaciones y tenemos la misma cantidad de casos en proporción:



Luego observamos que todas las insurance poseen billings bajos y altos. no hay seguros que parezcan no lograr cubrir precios altos o viceversa:



LOGRAMOS ENTONCES CONCLUIR QUE NO HAY VALORES ATÍPICOS EN LOS BOX PLOTS Y LAS DISTRIBUCIONES DE LAS VARIABLES SEGMENTADAS POR TEST RESULT RESULTAN EN TODOS LOS CASOS SIMILARES.

LUEGO ANALIZAMOS SI EXISTEN VALORES NULOS:

	Name	istributio	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
C	Gender			1		0.693			0 (0 %)
C	Blood Type			1		2.08			0 (0 %)
N	Age		51.54	38	52	0.38	13	89	0 (0 %)
C	Medical Condition			0		1.79			0 (0 %)
T	Date of Admission		2021-11-01	2024-03-16	2021-11-01	~5 years	2019-05-08	2024-05-07	0 (0 %)
C	Insurance Provider			2		1.61			0 (0 %)
N	Billing Amount		25539.3	-1316.62	25538.1	0.556449	-2008.49	52764.3	0 (0 %)
C	Admission Type			0		1.1			0 (0 %)
T	Discharge Date		2021-11-16	2020-03-15	2021-11-17	~5 years	2019-05-09	2024-06-06	0 (0 %)
C	Medication			2		1.61			0 (0 %)
C	Test Results			Abnormal		0.637			0 (0 %)

Se observa claramente en la columna “missing” que NO HAY DATOS FALTANTES para ninguna de las variables.

FINALMENTE ANALIZAMOS LA CORRELACIÓN ENTRE LAS VARIABLES:

La edad no tiene correlación con las variables de Billing ni las fechas de entrada y salida.

1	-0.004	Age	:	Billing Amount
2	-0.000	Age	:	Date of Admission
3	-0.000	Age	:	Discharge Date

El date of admission y el discharge date tienen una correlación positiva perfecta, lo cual tiene sentido, porque son las fechas en las que un paciente entra y se va:

1	+1.000	Date of Admission	:	Discharge Date
2	-0.001	Billing Amount	:	Date of Admission
3	-0.000	Age	:	Date of Admission

El billing no tiene correlación con las fechas de ingreso o egreso:

2	-0.001	Billing Amount	:	Discharge Date
3	-0.001	Billing Amount	:	Date of Admission

Entrenamiento de modelos predictivos:

Como segundo paso, se aplicó un Data Sampler con el objetivo de dividir el conjunto de datos original en dos subconjuntos: el 80% de los datos se destinó al entrenamiento (train) y el 20% restante a la prueba (test). Esta partición permite evaluar el rendimiento de los modelos sobre datos no vistos durante el entrenamiento, favoreciendo así una validación más objetiva de su capacidad de generalización.

A continuación, se procedió a entrenar tres modelos distintos utilizando exclusivamente el conjunto de entrenamiento: Árbol de Decisión (Tree), Regresión Logística (Logistic Regression) y Red Neuronal (Neural Network). Cada uno de estos algoritmos representa una aproximación diferente al problema de clasificación, lo que permite comparar sus resultados y seleccionar el modelo más adecuado según métricas de rendimiento específicas.

Luego de entrenar los modelos con el conjunto de datos correspondiente al 80% de los registros (train), se evaluó su rendimiento utilizando varias métricas: AUC, CA (Accuracy), F1 Score, Precisión y Recall.

AUC (Area Under the ROC Curve): Mide la capacidad general del modelo para distinguir entre las clases positiva y negativa. Un valor más alto indica mejor poder de discriminación.

CA (Classification Accuracy): Porcentaje total de predicciones correctas (aciertos totales / total de casos). Simple pero puede ser engañoso en datos desbalanceados.

F1-Score: Media armónica (un balance) entre Precision y Recall. Útil cuando necesitas equilibrar ambas métricas, especialmente con datos desbalanceados.

Precision (Precisión): De las veces que el modelo predijo la clase positiva, ¿cuántas veces acertó? (Importante cuando los Falsos Positivos son costosos).

Recall (Sensibilidad): De todos los casos que realmente eran positivos, ¿cuántos logró encontrar el modelo? (Importante cuando los Falsos Negativos son costosos).

Los resultados del Test and Score fueron los siguientes:

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.542	0.595	0.591	0.587	0.595	0.072
Logistic Regression	0.503	0.666	0.533	0.444	0.666	0.000
Neural Network	0.500	0.500	0.514	0.555	0.500	0.000

Tree: mostró un desempeño moderado, con un AUC de 0.542 y una Accuracy de 0.595. La métrica F1 alcanzó 0.591, mientras que la Precisión fue de 0.587 y el Recall de 0.595.

Logistic Regression: presentó una AUC de 0.503, con una Accuracy algo superior (0.666). No obstante, su F1 Score (0.533) y Precisión (0.444) fueron inferiores al modelo anterior, a pesar de un Recall más alto (0.666).

Neural Network: obtuvo resultados inferiores, con AUC y Accuracy de 0.500, un F1 Score de 0.514, Precisión de 0.555 y Recall de 0.500.

En base a estos resultados, el modelo de Árbol de Decisión parece haber sido el que mejor equilibrio mostró entre precisión y recall, al observar el F1, aunque las diferencias no son demasiado pronunciadas. Aun así, el valor bajo de Area Under de ROC Curve (AUC) en todos los modelos indican que los algoritmos no lograron una separación efectiva entre las clases, lo que nos indica que con estos datos no se puede crear un modelo efectivo.

Para el caso que estamos evaluando, los test results, un **falso positivo sería el más costoso** porque seria **que se clasifique como un test normal uno que fue anormal**. Esto provocaría que ese paciente no obtenga el procedimiento que debe dársele por tener resultados anormales en su test y se lo dejaría ir pensando que esta todo bien. Por eso la métrica con mayor relevancia para el caso que analizamos es la **Precision.** Esta métrica posee mayor desempeño en el tree, seguida por la neural network y le sigue la logistic regresion. Sin embargo, los tres modelos rondan por el porcentaje 50 de veces que predijo que era positivo (normal) cuántas acertó. Es decir, **la mitad de las veces**. Esto lleva a que no sea **para nada un modelo confiable** para predecir los resultados de los tests porque traería consecuencias muy graves **en el contexto de salud**.

Análisis de la Matriz de Confusión

Además de las métricas globales, se analizó el comportamiento específico de cada modelo mediante su **matriz de confusión**.

Tree: Este modelo logró clasificar correctamente un buen porcentaje de casos Abnormal, aunque presenta errores importantes, tanto en falsos positivos (61,6%) como en falsos negativos (31,1%).

		Predicted		
		Abnormal	**Normal**	**∑**
Actual	**Abnormal**	68.9 %	61.6 %	**100610**
	Normal	31.1 %	38.4 %	**50350**
	∑	**104582**	**46378**	**150960**

Logistic Regression: En este caso, el modelo nunca predijo la clase "Normal". Es decir, todos los ejemplos fueron clasificados como Abnormal, independientemente de su clase real. Todos los casos Normales fueron mal clasificados como Abnormal. Este comportamiento sugiere un fuerte desequilibrio en la predicción.

		Predicted		
		Abnormal	**Normal**	**∑**
Actual	**Abnormal**	66.6 %	NA	**100610**
	Normal	33.4 %	NA	**50350**
	∑	**150960**	**0**	**150960**

Neural Network: El modelo neuronal tuvo un rendimiento mixto, con una alta tasa de falsos positivos (66,6%). Aunque predijo ambas clases, sus resultados no fueron satisfactorios.

		Predicted		
		Abnormal	**Normal**	**∑**
Actual	**Abnormal**	66.6 %	66.6 %	**100610**
	Normal	33.4 %	33.4 %	**50350**
	∑	**75480**	**75480**	**150960**

Conclusión del análisis de confusión:

El Árbol de Decisión mostró un desempeño más equilibrado que los otros modelos, aunque aún con errores notables. La Regresión Logística presentó un comportamiento problemático al no predecir nunca la clase *Normal*, lo que lo hace inviable para tareas que requieren distinguir ambas clases. La Red Neuronal tuvo un rendimiento más simétrico en cuanto a predicciones, pero igualmente insuficiente.

Análisis ROC

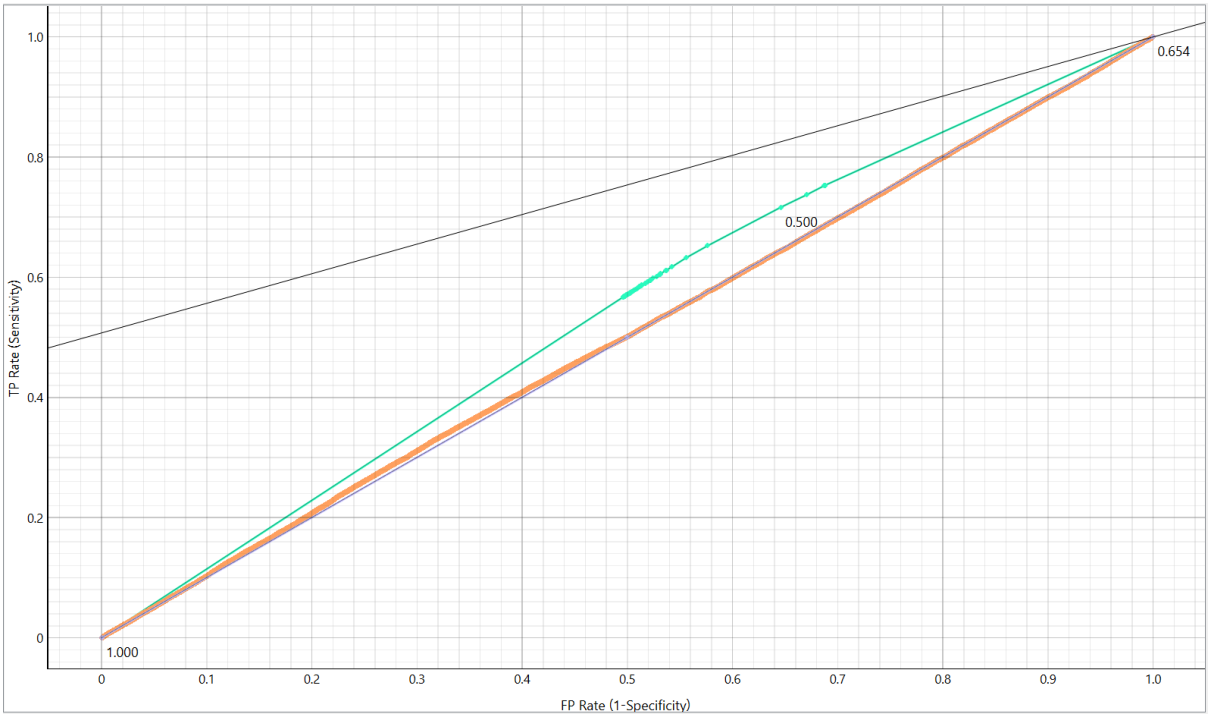
Para evaluar el rendimiento de los modelos desde el punto de vista de su capacidad discriminativa, se utilizaron **curvas ROC**, tanto considerando como clase objetivo *Abnormal* como *Normal*. Estas curvas permiten observar cómo varía la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) al modificar el umbral de decisión.

Target: Abnormal

En la primera gráfica, donde se estableció *Abnormal* como clase positiva, se observa que:

- El **modelo Tree** presenta la mejor curva ROC, superando ligeramente la línea de no discriminación (diagonal), alcanzando un AUC de aproximadamente **0.654**, lo que indica una capacidad de discriminación moderadamente mejor que el azar.

- Tanto **Logistic Regression** como **Neural Network** tienen curvas muy cercanas a la diagonal, con un AUC cercano a **0.500**, lo cual refleja un rendimiento equivalente al azar, sin poder distinguir correctamente entre clases.

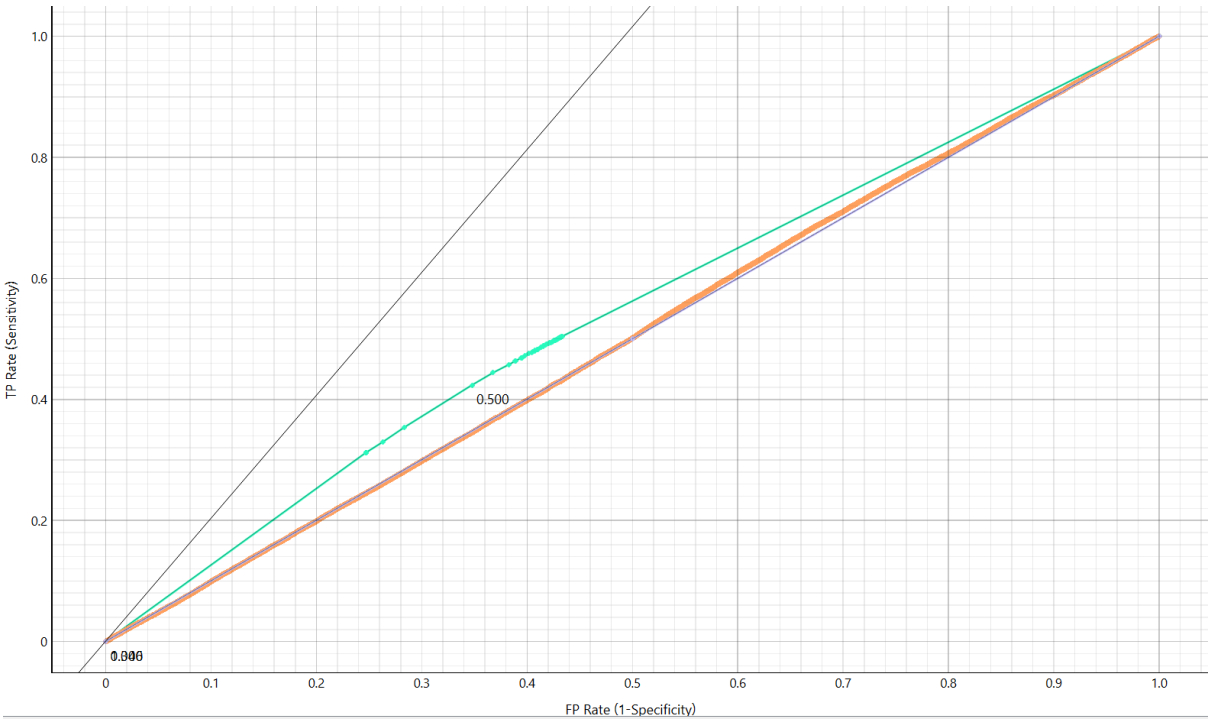


Target: Normal

En la segunda gráfica, donde se invierte el target a *Normal*, el patrón se repite:

- El **modelo Tree** nuevamente muestra un leve mejor rendimiento que el resto, aunque no llega a valores ideales.

- **Logistic Regression** y **Neural Network** vuelven a mostrar curvas muy próximas a la diagonal, confirmando que **no son capaces de discriminar eficazmente entre Normal y Abnormal**.



Evaluación sobre el conjunto de test

Una vez entrenados los modelos con el conjunto de entrenamiento, se aplicaron al 20% restante de los datos (conjunto de test) para evaluar su capacidad de generalización. Los resultados obtenidos son los siguientes:

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.567	0.616	0.612	0.608	0.616	0.119
Logistic Regression	0.500	0.666	0.532	0.443	0.666	0.000
Neural Network	0.500	0.666	0.532	0.443	0.666	0.000

Análisis:

- El **Árbol de Decisión** es el único modelo que muestra cierta capacidad discriminativa (AUC = 0.567). El porcentaje total de predicciones correctas (aciertos totales / total de casos), accuracy, posee una métrica medianamente aceptable, al igual que la precision y recall; teniendo valores cercanos al 60% en los tres casos.

- Tanto la **Regresión Logística** como la **Red Neuronal** presentan **AUC = 0.500**, lo que indica que su capacidad de discriminación es equivalente a lanzar una moneda. Pese a que su **accuracy** es un poco más elevada (0.666), con una **precisión baja**

(0.443) y reflejan un comportamiento desequilibrado y poco confiable.

Esto confirma que una elevada exactitud (accuracy) por sí sola no garantiza un buen modelo, especialmente si las clases están desbalanceadas o el modelo aprende un solo patrón dominante.

TRAIN vs TEST

Train:

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.542	0.595	0.591	0.587	0.595	0.072
Logistic Regression	0.503	0.666	0.533	0.444	0.666	0.000
Neural Network	0.500	0.500	0.514	0.555	0.500	0.000

Test:

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.567	0.616	0.612	0.608	0.616	0.119
Logistic Regression	0.500	0.666	0.532	0.443	0.666	0.000
Neural Network	0.500	0.666	0.532	0.443	0.666	0.000

Analisis:

Árbol de Decisión (Tree)

- Accuracy: pasa de 0.595 en entrenamiento a 0.616 en test, una ligera mejora.

- Interpretación: este comportamiento no indica overfitting, ya que el rendimiento no cae al pasar al conjunto de test. De hecho, mejora un poco, lo cual podría deberse a una mejor adecuación a patrones generales del conjunto de datos o a una leve aleatoriedad.
- Conclusion: **no hay overfitting ni underfitting** porque las metricas no son alevosamente distintas entre el train y test

Regresión Logística

- Accuracy: se mantiene constante en 0.666 tanto en entrenamiento como en test.

- Interpretación: este comportamiento sugiere que no hay overfitting, pero tampoco hay evidencia de que el modelo esté capturando relaciones complejas.

- Conclusión: este modelo parece estar ajustado a un nivel básico, sin aprender demasiado del patrón de datos. **No hay señales claras de underfitting ni overfitting.**

Red Neuronal

- Accuracy: mejora fuertemente de 0.500 en entrenamiento a 0.666 en test.
- Precision: disminuye de 0.55 a 0.44
- Recall: aumenta de 0.5 a 0.66

- El comportamiento del modelo sugiere un **posible caso de underfitting**. Aunque la accuracy mejora notablemente de 0.500 en entrenamiento a 0.666 en test, esta diferencia no implica un verdadero aprendizaje, ya que durante el entrenamiento el modelo muestra una performance muy baja. Además, mientras que el recall aumenta (de 0.50 a 0.66), indica ndo que detecta más casos positivos en test, la precisión disminuye (de 0.55 a 0.44), lo que implica un mayor número de falsos positivos. **Esta combinación refleja que el modelo no está captando adecuadamente los patrones durante el entrenamiento y tiene un rendimiento inestable, característico de un modelo con capacidad insuficiente o mal ajustado.**

Conclusión Final

.A lo largo de este trabajo se desarrolló un sistema de clasificación orientado a predecir si un resultado de test médico es Normal o Abnormal, considerando como “abnormal” tanto los resultados originalmente clasificados como Abnormal e Inconclusive, debido a que ambos representan situaciones clínicas que no ofrecen una guía clara de actuación médica. En este contexto, los falsos positivos, es decir, predecir Normal cuando en realidad el resultado es Abnormal, constituyen el error más costoso, ya que pueden llevar a una decisión médica inadecuada con consecuencias potencialmente graves para el paciente.

Tras analizar los resultados obtenidos por los tres modelos entrenados (Árbol de Decisión, Regresión Logística y Red Neuronal), se identificó que, aunque en ciertos casos la accuracy fue razonablemente alta (hasta 0.666 en test), ello no se tradujo necesariamente en un buen rendimiento clínico, ya que modelos como la Regresión Logística y la Red Neuronal presentaron una precisión baja (0.44) y un número elevado de falsos positivos. Esto compromete su utilidad, pues implicaría clasificar erróneamente como Normales casos que requieren atención médica adicional.

El Árbol de Decisión, aunque con métricas moderadas en general (recall de 0.616, precisión de 0.608 en test), fue el modelo que mostró un rendimiento más equilibrado y con

menor tasa de falsos positivos, lo cual lo hace más adecuado para este tipo de problema donde el costo del error no es simétrico. Además, su comportamiento estable entre entrenamiento y test indica que no sufre ni de overfitting ni de underfitting significativo, y por tanto, generaliza de manera razonable.

No obstante, es importante remarcar que, aun siendo el mejor de los tres modelos evaluados, el rendimiento del Árbol de Decisión no alcanza un nivel aceptable para su aplicación real en el ámbito de la salud. Las métricas obtenidas, si bien relativamente mejores, siguen siendo bajas en términos absolutos. En un contexto clínico, donde las decisiones basadas en modelos predictivos pueden afectar directamente la atención de los pacientes, no se puede tolerar un modelo con tasas de error tan elevadas, especialmente en lo referente a falsos positivos. Esto subraya la necesidad de continuar trabajando en la mejora del modelo