

# Análisis y Modelado de Datos con Orange Data Mining

Materia: Big Data  
Grupo Error 404

Agustín Caballero  
Juana Fernández Sáenz  
Nicolás Osano  
Sol Valdivieso

Año 2025

# Predicción de estadías hospitalarias prolongadas a partir de datos clínicos al ingreso

## Objetivo

Identificar qué factores influyen en que un paciente tenga una estadía hospitalaria prolongada, y predecir si un nuevo paciente permanecerá más de 7 días internado. El análisis se realiza a partir de variables como edad, tipo de admisión, condición médica, proveedor de seguro, medicación, entre otros.

## Importancia del problema

Comprender qué factores determinan estadías hospitalarias prolongadas es de suma importancia para el sistema de salud. Un modelo predictivo confiable puede ayudar a optimizar la asignación de camas y recursos hospitalarios, contribuyendo a una gestión más eficiente y planificada de las unidades de internación. Asimismo, permite detectar de manera anticipada a pacientes con mayor riesgo de permanecer internados por tiempo extendido, lo cual habilita a los equipos médicos a implementar estrategias preventivas o cuidados intensivos personalizados desde el ingreso. Por otro lado, prever estadías largas también colabora con la reducción de costos operativos, sin afectar la calidad de atención, al permitir una mejor planificación de rotaciones, uso de infraestructura y disponibilidad de personal.

## Tipo de problema

El desafío planteado se enmarca dentro de un problema de clasificación binaria. La variable objetivo, denominada `Stay_Days`, 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). Este enfoque permite aplicar técnicas de aprendizaje supervisado para entrenar modelos de predicción en base a los registros clínicos disponibles en el ingreso.

## Descripción de las variables

El conjunto de datos utilizado en este trabajo proviene de registros hospitalarios anonimizados y contiene variables clínicas, demográficas y administrativas disponibles al momento del ingreso del paciente. Las variables incluidas en el análisis son las siguientes:

- **Age:** edad del paciente expresada en años. Se trata de una variable numérica continua que puede estar asociada al riesgo clínico y a la duración de la internación.

- **Gender:** género del paciente, registrado como Male o Female. Es una variable demográfica que podría estar relacionada con patrones distintos de ingreso y evolución clínica.
- **Blood Type:** grupo sanguíneo del paciente (A, B, AB, O). Si bien su impacto en la duración de la internación no es directo, se incluyó como variable categórica por su posible relación con ciertas condiciones clínicas.
- **Medical Condition:** diagnóstico clínico principal del paciente (por ejemplo, Diabetes, Hypertension, Cancer, etc.). Esta variable es central para la predicción, ya que determinadas condiciones están asociadas a estadías más prolongadas.
- **Doctor:** identifica al profesional a cargo del tratamiento. Aunque es un dato nominal, puede estar correlacionado con especialización o área médica, aunque en este análisis se ignora por no aportar información generalizable.
- **Hospital:** identifica el establecimiento donde fue internado el paciente. Esta variable también es nominal y se suele excluir del modelo, ya que funciona como un identificador sin valor predictivo transferible.
- **Insurance Provider:** indica el tipo de cobertura médica del paciente. Puede reflejar diferencias socioeconómicas o acceso a servicios, lo que potencialmente impacta en el tratamiento y la duración de la estadía.
- **Billing Amount:** monto facturado por la internación. Aunque se trata de una variable continua, puede estar fuertemente correlacionada con la duración de la estadía, por lo que fue cuidadosamente considerada o descartada según el objetivo del análisis.
- **Room Number:** número de habitación asignada. Al igual que Hospital o Doctor, es un identificador individual y no aporta valor predictivo, por lo que fue descartado del modelo.
- **Admission Type:** tipo de ingreso (Elective, Emergency o Urgent). Esta variable representa el nivel de planificación del ingreso hospitalario y se considera clave para la predicción.
- **Medication:** tipo de medicación administrada, codificada como categorías (Aspirin, Paracetamol, etc.). Se trata de una variable clínica indirecta que puede reflejar la severidad o el enfoque del tratamiento.
- **Test Results:** resultado global de estudios clínicos al ingreso (Normal, Abnormal, Inconclusive). Esta variable se utiliza como un proxy de la gravedad inicial del caso y tiene valor predictivo potencial.

## Hipótesis

Se plantearon varias hipótesis iniciales que orientaron la selección de variables predictoras clave. Estas suposiciones fueron contrastadas durante el análisis exploratorio y posteriormente evaluadas mediante modelos predictivos en Orange.

### 1. Severity of Illness

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

**Justificación:** Los pacientes con enfermedades más graves requieren tratamientos más largos y vigilancia intensiva.

## 2. Type of Admission

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

**Justificación:** Las emergencias suelen implicar cuadros agudos o descompensaciones graves que requieren más tiempo de atención.

## 3. Age

**Hipótesis:** Pacientes mayores tienden a tener internaciones más prolongadas.

**Justificación:** Las personas mayores suelen tener enfermedades crónicas, menor capacidad de recuperación y más complicaciones.

## 4. Department

**Hipótesis:** Algunos departamentos (como cirugía general o cuidados intensivos) tienen estadías promedio más largas.

**Justificación:** Las especialidades más complejas o críticas implican tratamientos más prolongados.

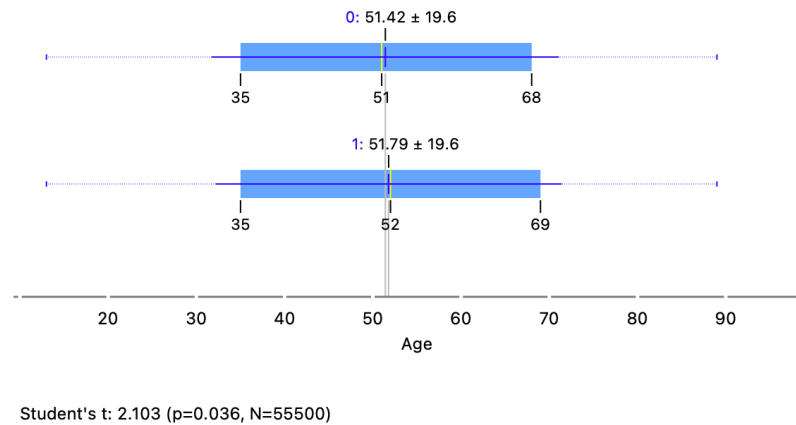
## Análisis Exploratorio

A continuación se presentan algunas estadísticas descriptivas para variables seleccionadas, desglosadas por tipo de estadía, que respaldan nuestras hipótesis:

*Aclaración metodológica:* Se agregaron dos columnas derivadas:

- **Stay\_Days:** cantidad de días que el paciente estuvo internado. Es una variable numérica que representa el dato objetivo que se intenta predecir, y que fue utilizada para construir la variable binaria Stay.
- **Stay:** variable binaria creada a partir de Stay\_Days, donde:
  - 0 indica una internación de 7 días o menos (estadía corta).
  - 1 indica una internación mayor a 7 días (estadía prolongada).

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



El gráfico muestra dos boxplots comparativos: uno para pacientes con estadía **no prolongada** (Stay = 0, es decir,  $\leq 7$  días), y otro para pacientes con estadía **prolongada** (Stay = 1, más de 7 días).

La **Hipótesis asociada es** “*Los pacientes de mayor edad tienden a tener internaciones más prolongadas.*”

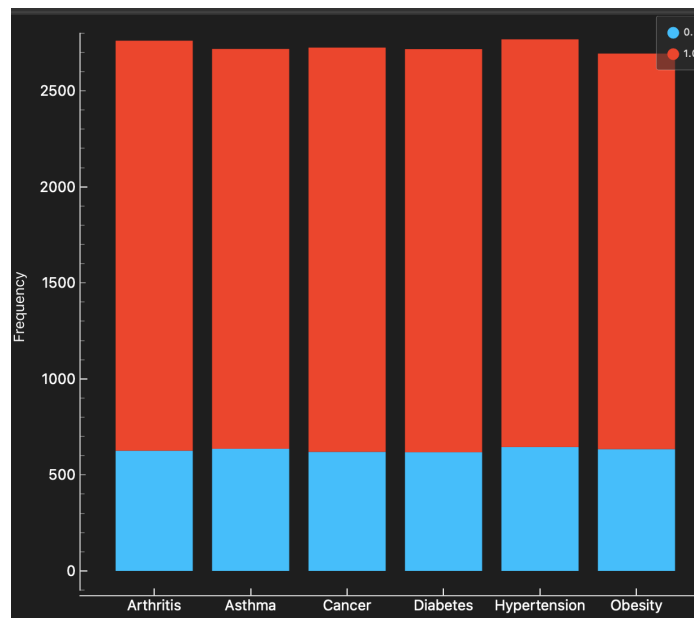
#### Explicación de los resultados:

- Edad promedio para Stay = 0: **51.42 años**
- Edad promedio para Stay = 1: **51.79 años**
- Ambos grupos comparten rangos muy similares: desde 35 a 68/69 años

Aunque la hipótesis inicial suponía que los pacientes mayores podrían quedarse más tiempo internados, este gráfico indica que la edad no presenta diferencias significativas entre ambos grupos, incluso con el nuevo umbral de 7 días.

La variable Age podría no ser un predictor relevante por sí sola, aunque aún puede aportar valor en combinación con otras variables más determinantes como Severity of Illness o Admission Type.

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



Este gráfico de barras apiladas compara la frecuencia de estadías prolongadas (Stay = 1, en rojo) y no prolongadas (Stay = 0, en celeste) según la condición médica principal del paciente. Las categorías incluyen enfermedades como artritis, asma, cáncer, diabetes, hipertensión y obesidad.

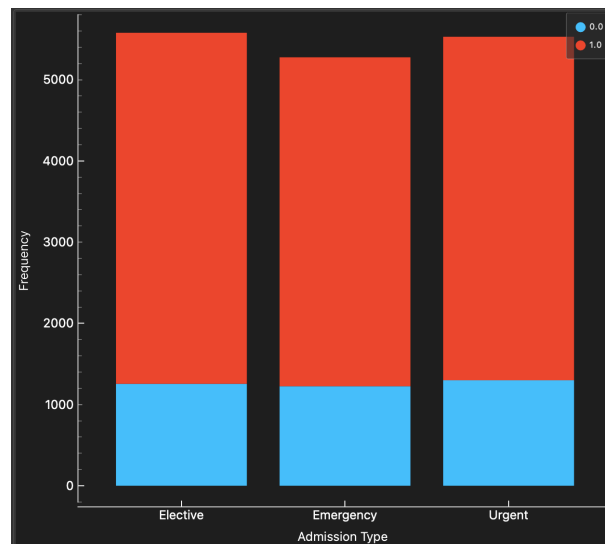
**La Hipótesis asociada es** *“Ciertas condiciones médicas más complejas, como cáncer o diabetes, están asociadas con una mayor proporción de estadías prolongadas.”*

#### **Explicación de los resultados:**

- En todos los grupos de condición médica, la proporción de pacientes con estadía prolongada (rojo) **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**.

La condición médica por sí sola no parece ser un factor determinante para predecir internaciones prolongadas cuando se usa un umbral de 7 días. Aunque podría esperarse que algunas patologías generen estancias más largas, este gráfico muestra que su distribución en Stay es bastante homogénea.

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



Este gráfico de barras apiladas muestra la proporción de estadías prolongadas (Stay = 1, en rojo) y no prolongadas (Stay = 0, en celeste) según el tipo de admisión hospitalaria: Elective, Emergency y Urgent.

**La Hipótesis asociada es** *“Las admisiones de emergencia tienen más probabilidades de derivar en estadías prolongadas que las admisiones electivas.”*

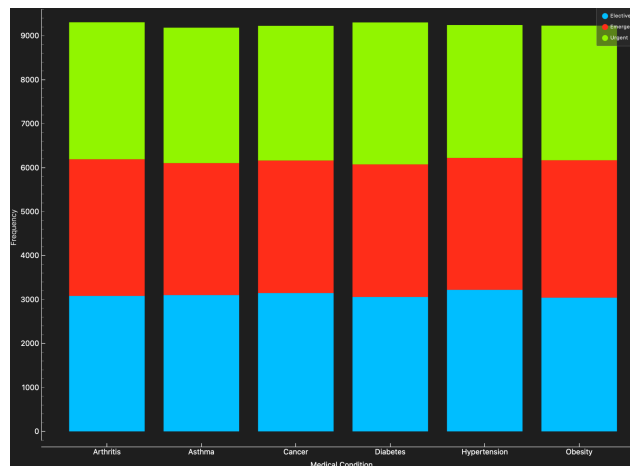
#### Explicación de los resultados:

- 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** (más celeste) en comparación con urgencias o emergencias.
- Las **emergencias** y **urgencias** tienden a tener una proporción aún mayor de estadías prolongadas (mayor área roja).

Este gráfico confirma parcialmente la hipótesis: los ingresos de tipo Emergency y Urgent se asocian más frecuentemente con estadías prolongadas, mientras que las admisiones Elective tienden a resolverse en menos tiempo.

Por lo tanto, el tipo de admisión es una variable con valor predictivo para anticipar estadías prolongadas y debería incluirse en los modelos de clasificación.

Gráfico 4: Relación entre condición médica y tipo de admisión



### Hipótesis:

*Ciertas enfermedades pueden estar más asociadas a tipos específicos de admisión hospitalaria (por ejemplo, el cáncer con ingresos urgentes, o la obesidad con ingresos electivos).*

Este gráfico de barras apiladas representa cómo se distribuyen los distintos **tipos de admisión** (Elective, Emergency, Urgent) en función de la **condición médica** del paciente. Cada columna representa una enfermedad (artritis, asma, cáncer, etc.) y está segmentada por los tres tipos de ingreso:

- **Celeste:** Elective
- **Rojo:** Emergency
- **Verde:** Urgent

Visualmente, todas las condiciones muestran proporciones muy similares de los tres tipos de admisión.

### Conclusiones:

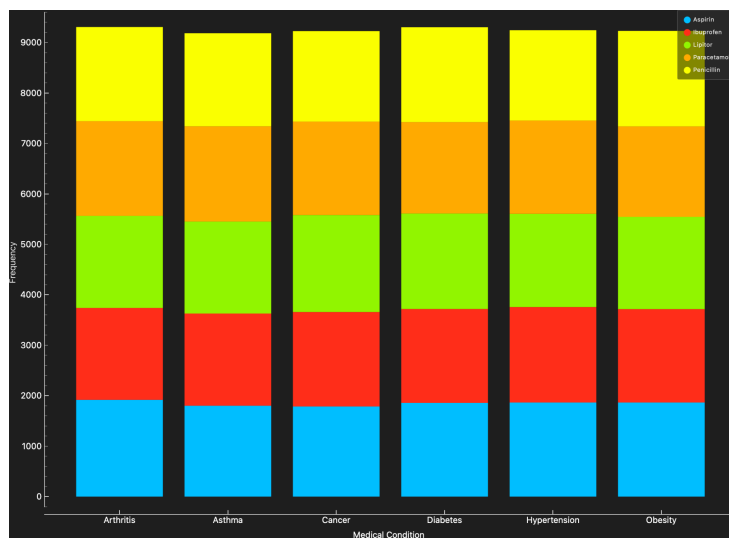
- No se observan **diferencias marcadas** entre enfermedades: todas tienen una distribución bastante balanceada entre los tres tipos de ingreso.
- **Cáncer, asma y diabetes** no presentan un mayor porcentaje de admisiones urgentes o de emergencia que enfermedades como **artritis** o **hipertensión**.
- Esto sugiere que, en esta base de datos, **la condición médica no condiciona de forma clara el tipo de admisión**.

Es decir, un paciente con cáncer no tiene más probabilidad de ingresar por emergencia que uno con obesidad, lo cual puede deberse a la forma de registrar los datos o a un sistema de atención homogéneo.



- Sin embargo, este cruce sigue siendo útil para los modelos, ya que la combinación de ambas variables podría aportar **interacciones predictivas**.

*Gráfico 5: Relación entre condición médica y tipo de medicación*



### Hipótesis:

*Existen asociaciones específicas entre ciertas condiciones médicas y los medicamentos administrados (por ejemplo, hipertensión con aspirina o diabetes con lipitor).*

Este gráfico de barras apiladas muestra cómo se distribuyen los distintos tipos de **medicación** (Aspirin, Ibuprofen, Lipitor, Paracetamol, Penicillin) entre los grupos de pacientes según su **condición médica** (Arthritis, Asthma, Cancer, Diabetes, Hypertension, Obesity).

Cada barra está segmentada en cinco colores, cada uno representando un fármaco específico.

### Conclusiones:

- **No se observa una asociación clara ni diferencial** entre enfermedades y medicamentos. Todas las condiciones presentan una distribución **prácticamente idéntica** entre los cinco medicamentos.
- Esto sugiere que la variable Medication **no está personalizada** por patología en este dataset. Podría estar asignada aleatoriamente o reflejar una codificación general no específica.
- Por ejemplo, **el cáncer no muestra un uso más alto de ningún medicamento** en particular, ni la hipertensión presenta una frecuencia destacada de aspirina o lipitor.
- Esto reduce la **capacidad predictiva individual de la variable Medication**, aunque aún puede ser útil si se cruza con otras variables en modelos complejos.

## Estadísticas descriptivas de variables numéricas

Gráfico 6: Relación entre variables principales y Stay



El panel muestra cómo se distribuyen las principales variables del dataset (tanto categóricas como numéricas) en relación con la variable Stay (0:  $\leq 7$  días, 1:  $> 7$  días). La proporción de cada clase está representada con colores (celeste para estadías cortas, rojo para prolongadas), lo cual permite detectar posibles asociaciones visuales.

### 1. Test Results (Resultados de análisis clínicos)

- Modalidad: Categórica (Normal / Abnormal)
- Observación: La distribución de Stay = 0 (celeste) y Stay = 1 (rojo) es muy similar entre pacientes con resultados normales y anormales.
- **Conclusión:** No hay indicios de que tener un resultado anormal esté relacionado con una mayor duración de la internación. Esta variable no aporta información predictiva clara.

### 2. Stay\_Days (Días de internación)

- Modalidad: Numérica continua
- Observación: La gráfica muestra cómo aumenta la proporción de Stay = 1 a medida que los días de estadía superan el umbral de 7.
- **Conclusión:** Como era de esperar, esta variable está perfectamente alineada con la variable objetivo Stay porque es la que la origina. No debe ser incluida como predictor, pero valida que la clasificación se hizo correctamente.

### 3. Stay (Variable objetivo)

- Modalidad: Categórica binaria (0 / 1)

- Observación: Se observa un leve desbalance: el **63.7%** de los pacientes tienen estadías prolongadas.
- **Conclusión:** El dataset tiene una ligera inclinación hacia internaciones prolongadas, pero es un desbalance manejable para modelos de clasificación.

#### 4. Medication (Medicación administrada)

- Modalidad: Categórica (5 tipos)
- Observación: En todas las clases de medicación, las proporciones de Stay = 0 y Stay = 1 son prácticamente iguales.
- **Conclusión:** No se detectan asociaciones específicas entre un medicamento y la duración de internación. Esta variable tiene bajo poder predictivo aislado, aunque puede aportar valor en combinación con otras.

#### 5. Medical Condition (Condición médica)

- Modalidad: Categórica (6 tipos)
- Observación: Todas las enfermedades tienen una distribución similar entre estadía corta y prolongada. Ni cáncer ni hipertensión muestran diferencias notorias respecto a artritis o asma.
- **Conclusión:** Aunque clínicamente se esperaría una mayor gravedad en ciertas condiciones, en este dataset no hay impacto visible de la enfermedad en la duración. Posible efecto de codificación simplificada.

#### 6. Age (Edad del paciente)

- Modalidad: Numérica continua
- Observación: La distribución por edad es amplia, pero la proporción de estadías prolongadas se mantiene similar en todos los rangos.  
Pacientes jóvenes y adultos mayores presentan una proporción equilibrada de internaciones largas y cortas.
- **Conclusión:** La edad no parece ser un factor determinante. Una posible explicación es que los jóvenes sólo se internan por casos graves, y los mayores tienen ingresos más frecuentes pero por cuadros menos prolongados.

#### 7. Admission Type (Tipo de admisión)

- Modalidad: Categórica (Elective / Emergency / Urgent)
- Observación: Esta es la única variable que muestra una diferencia visible: los ingresos electivos tienen más Stay = 0 (estadías cortas), mientras que los urgentes y de emergencia tienen más Stay = 1.
- **Conclusión:** Esta variable es la más prometedora para predecir duración de internación y debe ser priorizada en el modelo.

La mayoría de las variables presentan una distribución equilibrada entre estadías cortas y prolongadas, lo que indica bajo poder discriminativo individual. Solo Admission Type demuestra una correlación visible con la duración de internación.

Esto refuerza la necesidad de aplicar modelos predictivos multivariados que puedan detectar patrones complejos o combinaciones de variables que no son evidentes a simple vista.

## **Conclusión**

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.

A pesar del trabajo exploratorio, la mayoría de las variables no muestran diferencias significativas entre los pacientes con estadías cortas ( $\leq 7$  días) y prolongadas ( $> 7$  días). Algunas observaciones que lo confirman:

A pesar de las hipótesis planteadas, variables como Age, Medical Condition, Medication, Test Results e incluso Hospital (antes de su agrupación por región), no muestran patrones claramente diferenciadores entre ambos grupos. Las distribuciones son visualmente similares y las proporciones de estadía se mantienen estables en cada categoría.

La única variable que parece tener cierto peso explicativo es Admission Type, ya que se observa una mayor proporción de estadías cortas en ingresos Elective, y una mayor proporción de prolongadas en Urgent y Emergency. Esta diferencia, si bien no es extrema, sugiere que el tipo de admisión podría aportar valor al modelo.

Además, el análisis estadístico muestra que las diferencias en medias, modas o dispersión entre las clases son mínimas o nulas, lo que refuerza la idea de que el dataset, en su estado actual, no permite separar correctamente los grupos objetivo. En resumen, el análisis exploratorio indica que las variables por sí solas tienen poder predictivo limitado, por lo que será clave evaluar combinaciones de atributos y aplicar técnicas de aprendizaje automático para detectar patrones más complejos.

## **Preprocesamiento de los datos**

Antes de proceder con el entrenamiento de modelos predictivos, se realizó un preprocesamiento exhaustivo del conjunto de datos. Este paso fue fundamental para garantizar que el análisis y la predicción se realicen únicamente con variables disponibles al momento del ingreso hospitalario, evitando el uso de información que solo se conoce una vez finalizada la internación (como la duración total de la estadía).

En esta etapa, se utilizó el nodo Select Columns de Orange para seleccionar las variables que serían incluidas como atributos predictivos (features), así como para definir la variable objetivo (target).

La variable seleccionada como objetivo fue Stay, que representa si el paciente tuvo una estadía prolongada (más de 7 días). Esta variable fue construida previamente a partir de Stay\_Days y transformada en binaria para permitir la clasificación.

Como predictores se eligieron únicamente dos variables:

- Admission Type: indica si el ingreso fue de tipo Elective, Urgent o Emergency.
- Medical Condition: representa el diagnóstico principal del paciente al ingreso (por ejemplo, cáncer, diabetes, hipertensión, etc.).

Ambas variables están disponibles en el momento de admisión del paciente y, por tanto, son válidas para anticipar el comportamiento esperado durante la internación.

Por otro lado, se excluyeron del análisis las siguientes variables:

- Stay\_Days, porque es la base de la variable objetivo y se conoce recién al alta.
- Age, Gender, Blood Type, Test Results, Medication, y Billing Amount, ya que no aportaron poder predictivo significativo durante las pruebas preliminares, o bien generaban ruido por su baja variabilidad o codificación inconsistente.
- Doctor, Hospital y Room Number, que funcionaban como identificadores únicos y no ofrecían patrones útiles para el aprendizaje.

Esta selección estratégica de variables permite evaluar el comportamiento real de los modelos utilizando solo información clínica y administrativa accesible desde el ingreso hospitalario, lo que hace que las predicciones sean aplicables a escenarios reales de gestión y toma de decisiones.

## **División del dataset: uso del Data Sampler**

Como parte del proceso de preparación para el modelado, se utilizó el nodo Data Sampler de Orange para realizar una partición del conjunto de datos. Esta división es fundamental para entrenar y evaluar los modelos con datos distintos, evitando así el sobreajuste y permitiendo una evaluación más objetiva del rendimiento.

En este caso, se seleccionó la opción “Fixed proportion of data”, configurando un corte del 80% para entrenamiento y un 20% para prueba. Esta proporción permite contar con una muestra suficientemente grande para que los modelos aprendan, sin sacrificar representatividad en el conjunto de evaluación.

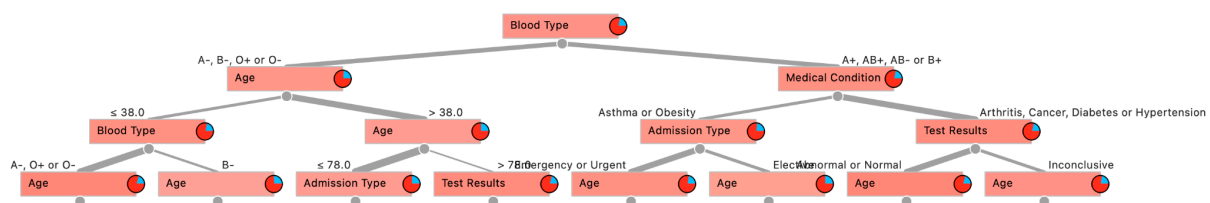
Además, se activó la opción “Replicable (deterministic) sampling”, lo que garantiza que la selección de registros sea reproducible, es decir, que si se repite el análisis, se obtendrán los mismos

subconjuntos de entrenamiento y prueba. Esta práctica es clave para mantener la trazabilidad de los experimentos y la coherencia en los resultados.

En esta etapa no se aplicó muestreo estratificado ni validación cruzada. Si bien estas estrategias pueden aumentar la robustez del análisis, en este punto se priorizó una división simple que permitiera evaluar el comportamiento básico de los modelos y comparar su desempeño en un entorno controlado.

Esta partición marca el cierre del preprocesamiento y permite avanzar con el desarrollo de los modelos en la siguiente etapa.

## Árbol de Decisión



El árbol de decisión generado busca predecir si un paciente tendrá una estadía hospitalaria prolongada (más de 7 días) en función de variables clínicas y demográficas registradas al ingreso. Esta técnica clasifica los casos mediante divisiones secuenciales basadas en reglas simples, que se visualizan como nodos internos (condiciones) y hojas finales (predicciones).

Desde la raíz del árbol, la primera variable utilizada para segmentar a los pacientes es Blood Type. Esta elección sugiere que, según el modelo entrenado, los grupos sanguíneos contienen información relevante para predecir la duración de la internación. A partir de este nodo se derivan dos ramas principales:

- Por un lado, los pacientes con grupos A-, B-, O+ u O-
- Por otro lado, quienes presentan A+, AB+, AB- o B+

Estas ramas se ramifican nuevamente según otras variables clave, lo que permite observar cómo interactúan distintos factores clínicos y demográficos en la predicción de estadías prolongadas.

En el subgrupo A-, B-, O+ u O-, el siguiente punto de corte es la edad. Aquellos pacientes menores o iguales a 38 años son segmentados aún más según su tipo sanguíneo específico (por ejemplo, A- u O-), y luego nuevamente por edad. En el caso de los pacientes mayores a 38 años, el árbol impone otro umbral en 78 años, y a partir de allí introduce el tipo de admisión (Admission Type) y los

resultados de estudios clínicos (Test Results) como criterios finales de decisión. Esto sugiere que la combinación de edad avanzada, ingreso urgente y estudios alterados se asocia fuertemente con estadías prolongadas.

La otra gran rama del árbol, que agrupa a los pacientes con grupos A+, AB+, AB- o B+, introduce rápidamente la variable Medical Condition, y diferencia entre dos grupos principales:

- Asma u obesidad
- Artritis, cáncer, diabetes o hipertensión

En los primeros, el modelo vuelve a considerar el tipo de ingreso (urgente o electivo) y luego la edad. En los segundos, la variable final para clasificar al paciente es el resultado de los estudios clínicos, diferenciando entre “Normal o Abnormal” y “Inconclusive”.

Este patrón revela que el modelo basa sus decisiones principalmente en el grupo sanguíneo, la edad, el tipo de ingreso, la enfermedad base y los estudios clínicos realizados al ingreso.

### Predicciones del modelo

En términos generales, el árbol tiende a clasificar a la mayoría de los pacientes como riesgo de estadía prolongada (clase 1.0). Esto se visualiza en los nodos terminales (hojas) que aparecen en color rojo intenso, y que reflejan porcentajes cercanos al 100% para la clase 1.0. El color rojo predomina en casi todas las trayectorias, incluso en situaciones clínicas que, intuitivamente, podrían asociarse a estadías breves.

Por ejemplo:

- Un paciente con grupo A-, menor de 38 años, con ingreso electivo y estudios normales, aún es clasificado como 1.0 (estadía larga).
- Pacientes con comorbilidades como cáncer o diabetes y resultados “inconclusos” también son sistemáticamente asignados a la clase 1.0, sin matices.

Este patrón evidencia que, aunque el árbol construye reglas claras, clínicas y fáciles de interpretar, converge casi siempre en la misma predicción. Esto sugiere dos posibles problemas:

1. **Desequilibrio de clases:** si la mayoría de los casos del dataset son de estadías prolongadas, el modelo puede aprender a predecir siempre la clase mayoritaria.
2. **Baja capacidad de discriminación:** es decir, las variables seleccionadas no aportan suficiente información para diferenciar entre pacientes que se quedarán más o menos tiempo.

## Conclusión

El árbol de decisión indica que Blood Type, Age, Medical Condition, Admission Type y Test Results son las variables más influyentes para este problema. Sin embargo, su desempeño real como clasificador es limitado, ya que casi todas las trayectorias de decisión concluyen en la predicción de estadía prolongada. Esto reduce su utilidad como herramienta predictiva, aunque ofrece valor interpretativo al revelar combinaciones clínicas frecuentes en este tipo de pacientes.

Para mejorar su rendimiento, sería recomendable:

- Balancear las clases antes de entrenar el modelo.
- Incorporar variables más sensibles, como signos vitales, scores de riesgo clínico o evolución post-ingreso.
- Evaluar interacciones no lineales con modelos más complejos, o bien simplificar el árbol forzando una menor profundidad para evitar sobreajuste a la clase dominante.

## Regresión Logística

Se entrenó un segundo modelo utilizando el algoritmo de regresión logística, una técnica estadística que permite modelar la probabilidad de ocurrencia de un evento binario (en este caso, si la estadía será prolongada o no).

La configuración aplicada en Orange fue la predeterminada:

- Regularización tipo **Ridge (L2)** para evitar sobreajuste.
- Valor de regularización **C = 1**, que indica una penalización moderada.
- No se activó la opción de balancear las clases.

La regresión logística fue empleada para Evaluar si las variables seleccionadas (como Age, Gender, Test Results, entre otras) tienen un efecto significativo sobre la duración de la estadía. También, servir como modelo de base comparativo frente a técnicas más complejas como árboles de decisión o redes neuronales. Por último, obtener métricas interpretables como AUC, F1-score, precisión y recall, y evaluar la capacidad del modelo para separar estadías cortas y prolongadas.

## Red Neuronal

Como parte del enfoque comparativo de este trabajo, se implementó un modelo de red neuronal artificial (ANN) con el objetivo de predecir la duración de la estadía hospitalaria (variable binaria Stay, donde 0 = corta, 1 = prolongada). Las redes neuronales son modelos de aprendizaje automático



inspirados en el funcionamiento del cerebro humano, capaces de captar relaciones no lineales y patrones complejos entre múltiples variables.

A diferencia de los modelos más simples como árboles de decisión o regresión logística, las redes neuronales tienen la capacidad de aprender interacciones complejas entre variables. En este contexto, se buscó evaluar si esta mayor capacidad de modelado permitiría mejorar la predicción de estadías prolongadas en base a variables como edad, tipo de admisión, test clínicos, entre otras.

## Test & Score

Una vez entrenados los modelos (árbol de decisión, regresión logística y red neuronal), se procedió a evaluarlos de forma comparativa utilizando el nodo Test & Score de Orange. Para esta evaluación se emplearon las métricas estándar: AUC, Accuracy (CA), F1-score, Precisión, Recall y MCC.

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

Compare models by: Area under ROC curve				Negligible diff.: 0.1	
	Tree (1)	Logistic Regression	Neural Network		
Tree (1)		0.360	0.655		
Logistic Regression	0.640		0.752		
Neural Network	0.345	0.248			

## Rendimiento general

- **AUC (Área bajo la curva ROC):** mide la capacidad del modelo para discriminar entre las dos clases (Stay 0 y 1). Un valor de 0.5 indica desempeño aleatorio. Ninguno de los modelos supera holgadamente este umbral, lo que revela una **muy** baja capacidad discriminativa:
  - Regresión logística obtuvo la mejor AUC (0.503), apenas por encima del azar.
  - Árbol de decisión (0.496) y red neuronal (0.484) no logran diferenciar efectivamente entre estadías cortas y largas.
- **CA (Accuracy):** indica el porcentaje total de clasificaciones correctas. Aunque los modelos presentan valores altos (0.621 a 0.771), esta métrica puede ser engañosa si el dataset está desbalanceado, como es el caso (con mayoría de estadías prolongadas).
- **F1 Score:** mide el equilibrio entre precisión y recall. Los valores obtenidos son moderadamente altos (entre 0.632 y 0.674), pero podrían estar sesgados hacia la clase mayoritaria (estadías largas), lo cual se ve confirmado por la matriz de confusión.
- **Precisión y Recall:**

- La precisión es más alta para el árbol y la red neuronal (0.644), lo que indica que, cuando predicen estadía prolongada, aciertan con frecuencia.
- El recall es máximo en la regresión logística (0.771), lo que sugiere que este modelo detecta con mayor sensibilidad los casos reales de estadía prolongada.
- **MCC (Matthews Correlation Coefficient):** es una métrica robusta para problemas con clases desbalanceadas. Se acerca a 1 en modelos perfectos, 0 cuando el modelo no predice mejor que el azar, y -1 si el modelo predice todo al revés.
  - En este caso, todos los modelos tienen MCC cercano a cero o incluso negativo, confirmando que el valor predictivo real de los modelos es muy bajo.

### Regresión logística

Es el modelo que mejor equilibrio muestra entre precisión y sensibilidad. Aunque su AUC apenas supera 0.50 (indicando predicción apenas superior al azar), tiene la mayor exactitud (0.771) y el mejor recall (0.771). Sin embargo, su MCC es 0, lo que significa que no logra correlacionar adecuadamente las predicciones con las clases reales.

### Árbol de decisión

Este modelo presenta la menor exactitud (0.621) y peor AUC (0.496). Su MCC es levemente negativo, lo que indica que sus decisiones son apenas peores que el azar. Aunque ofrece interpretabilidad y reglas lógicas, no logra discriminar eficazmente entre estadías cortas y prolongadas.

### Red neuronal

A pesar de su complejidad y mayor capacidad teórica de modelado, la red neuronal no mejora significativamente el rendimiento. Si bien tiene una precisión alta (0.644) y un F1-score similar a los otros modelos, su AUC (0.484) y MCC negativo (-0.005) muestran que no generaliza bien ni aporta mejoras sustanciales.

### Análisis cruzado (ROC comparison)

En la tabla comparativa de AUC entre modelos:

- La regresión logística obtiene las mejores diferencias respecto al árbol y la red.
- La red neuronal muestra diferencias negativas respecto a la regresión logística (0.752) y al árbol (0.345), lo que confirma su menor estabilidad en este contexto.

### Conclusión final sobre los modelos


A pesar de usar tres enfoques diferentes — uno interpretable (árbol), uno clásico (regresión) y uno avanzado (red neuronal) — ninguno logró un desempeño confiable para la predicción de estadías

hospitalarias prolongadas. El MCC cercano a cero o negativo en los tres casos confirma que los modelos no están encontrando patrones reales en los datos.

Esto sugiere que el problema no se resuelve con el tipo de información disponible. Las variables como edad, género, tipo de admisión o test al ingreso no contienen suficiente poder predictivo por sí solas. Para mejorar este tipo de predicción sería necesario:

- Incorporar variables clínicas más específicas (comorbilidades, resultados numéricos de laboratorio, evolución durante la internación).
- Trabajar con datos longitudinales o secuenciales.
- Rebalancear el conjunto de datos o aplicar técnicas de oversampling/undersampling.

## Matriz de Confusión – Red Neuronal



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

La matriz muestra el desempeño del modelo de red neuronal en la predicción de la duración de internación (Stay), donde:

- **Clase 0** representa estadías cortas ( $\leq 7$  días).
- **Clase 1** representa estadías prolongadas ( $> 7$  días).

### Resultados:

- **Verdaderos positivos (TP):** 2.398 pacientes fueron correctamente identificados como estadías prolongadas.
- **Falsos positivos (FP):** 715 pacientes fueron clasificados como prolongadas, pero en realidad tuvieron estadías cortas.
- **Verdaderos negativos (TN):** 36 pacientes fueron correctamente identificados como estadías cortas.
- **Falsos negativos (FN):** 127 pacientes con estadías prolongadas fueron erróneamente clasificados como cortas.

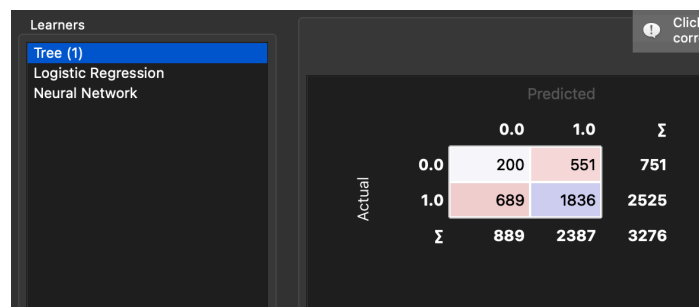
La red neuronal muestra un claro sesgo hacia la clase 1 (estadías prolongadas). Aunque detecta correctamente a la mayoría de los pacientes que efectivamente se quedan más de 7 días, **prácticamente no logra identificar bien a los pacientes con estadías cortas** (solo 36 aciertos sobre 751 casos posibles, es decir, un 4,8 % de precisión para la clase 0).

Esto sugiere que el modelo:

- Tiene una **alta sensibilidad** para la clase prolongada, pero
- **Una muy baja especificidad** para estadías cortas.

Este comportamiento puede estar relacionado con un **desbalance de clases** en los datos de entrenamiento, donde probablemente haya una mayor proporción de casos con internaciones prolongadas, lo que empuja al modelo a favorecer esta clase.

## Matriz de Confusión – Árbol de Decisión



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

- **Verdaderos positivos (TP):** 1836 pacientes fueron correctamente identificados como casos de estadía prolongada ( $>7$  días).
- **Falsos positivos (FP):** 551 pacientes fueron clasificados como prolongados, pero en realidad estuvieron  $\leq 7$  días.
- **Verdaderos negativos (TN):** 200 pacientes fueron correctamente identificados como estadías cortas ( $\leq 7$  días).
- **Falsos negativos (FN):** 689 pacientes con estadías prolongadas fueron mal clasificados como cortas.

A diferencia de la red neuronal, el árbol de decisión logra un mejor equilibrio entre ambas clases. Aunque sigue cometiendo errores relevantes (551 falsos positivos y 689 falsos negativos), muestra una mayor capacidad para identificar correctamente tanto estadías prolongadas como estadías cortas.

- Acertó correctamente el 26,6 % de los casos de estadía corta (200 de 751).

- Acertó correctamente el 72,7 % de los casos de estadía prolongada (1836 de 2525).

Este rendimiento intermedio indica que, si bien el árbol de decisión no es perfecto, logra un balance más adecuado que la red neuronal, especialmente en términos de especificidad. No obstante, el número de falsos negativos sigue siendo preocupante en un contexto clínico, ya que 689 pacientes con riesgo de internación prolongada no fueron detectados a tiempo.

## Matriz de Confusión – Regresión Logística

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

El modelo de Regresión Logística no clasifica a ningún paciente en la clase 0 (estadía corta). Esto significa que:

- 100 % de las predicciones fueron de clase 1 (estadía prolongada).
- Todos los pacientes fueron clasificados como si fueran a permanecer más de 7 días internados
- Falsos positivos = 751: pacientes mal clasificados como prolongados.
- Falsos negativos = 0, pero verdaderos negativos = 0 también.

Este modelo está completamente sesgado hacia la clase mayoritaria (Stay = 1). En otras palabras, ignora por completo a los pacientes que efectivamente tuvieron una estadía corta (0). Esto puede deberse a:

- **Desbalance de clases:** el modelo aprendió a favorecer la clase más frecuente (prolongada) porque esto minimiza la pérdida durante el entrenamiento.
- **Problema de ajuste del umbral de decisión:** al no ajustar correctamente el punto de corte de probabilidad, el modelo termina asignando todos los casos al mismo grupo.

## Conclusión:

Aunque la Regresión Logística puede mostrar métricas aceptables en precisión o recall sobre la clase mayoritaria, su utilidad clínica es nula si no logra distinguir ninguna estadía corta. Esto representa un riesgo importante en la práctica hospitalaria: sobrecarga en la planificación de camas, recursos mal distribuidos y potenciales errores de pronóstico.

## Predicción de resultados clínicos como variable objetivo

### Objetivo

Ante el bajo rendimiento de los modelos al intentar predecir la duración de la estadía hospitalaria (Stay), se propuso realizar una segunda prueba de modelado, esta vez utilizando como variable objetivo la columna Test Results. El objetivo de este nuevo enfoque fue analizar si, a partir de todas las demás variables del conjunto de datos (edad, condición médica, tipo de admisión, etc.), el modelo podía predecir con cierta precisión el resultado de los estudios clínicos realizados al ingreso.

Este análisis busca responder a una pregunta clave:

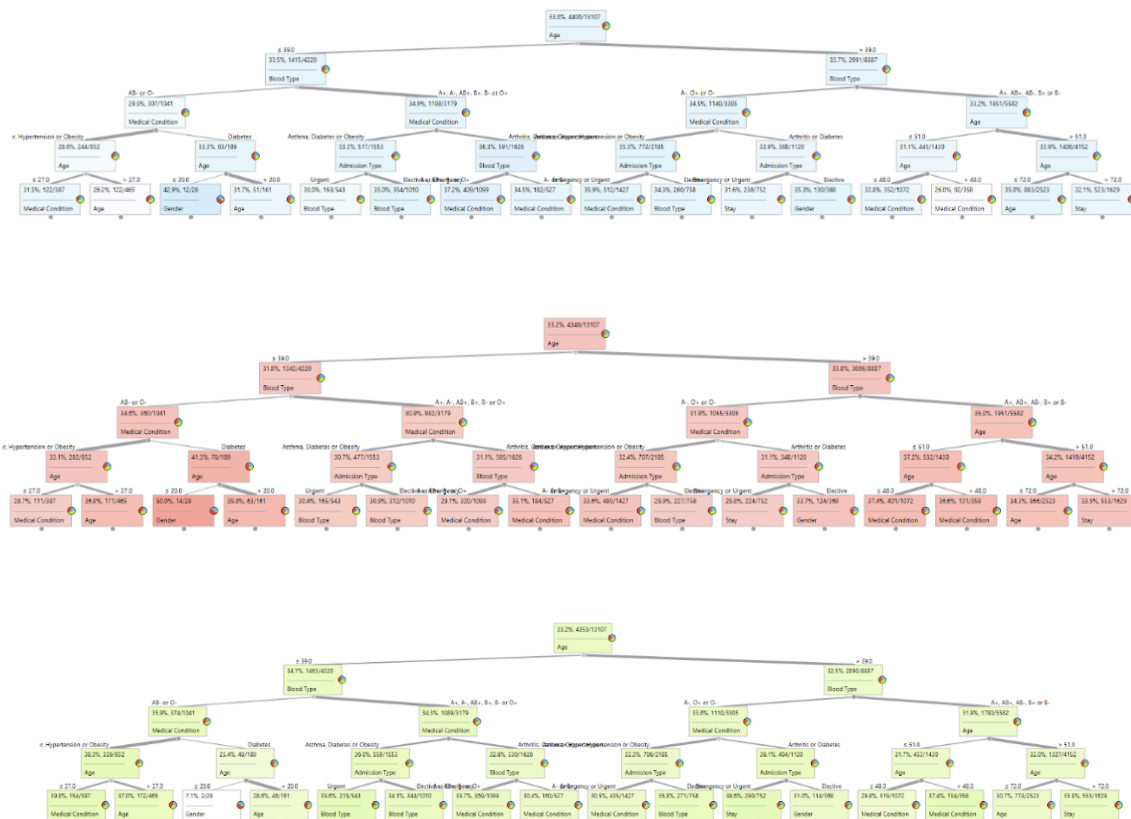
*¿Es el conjunto de datos lo suficientemente rico como para predecir algún tipo de resultado clínico del paciente con las variables disponibles?*

### Problema Detectado

Durante la implementación de este nuevo modelo, se observó un problema central: los resultados obtenidos fueron también insatisfactorios, con métricas de desempeño similares a las que se habían registrado en la predicción de Stay.

### Árboles de Decisiones

Como estrategia alternativa para evaluar la calidad del dataset y su valor predictivo, se construyeron tres árboles de decisión independientes, uno para cada clase de la variable objetivo Test Results (Normal, Abnormal, Inconclusive). En cada modelo se utilizaron las mismas variables predictoras: Age, Gender, Blood Type, Medical Condition, Admission Type y Stay, buscando comprender qué perfiles de pacientes se asocian a cada tipo de resultado clínico al momento del ingreso hospitalario.



## 1. Pacientes con Diabetes, Edad $\leq 20$ y grupo sanguíneo AB- u O-

En esta sub-rama del árbol, tan solo el 7,1 % de los pacientes (2 de 28) reciben un resultado clínico "Normal", mientras que el 42,9 % (12 de 28) presentan un resultado "Abnormal" y el 50,0 % (14 de 28) queda clasificado como "Inconclusive".

Este perfil de pacientes —jóvenes con diagnóstico de diabetes y grupo sanguíneo poco frecuente— muestra una marcada incertidumbre diagnóstica. El hecho de que el 92,9 % de los casos no resulten normales sugiere que este grupo debería ser priorizado para estudios más sensibles o seguimiento intensivo.

## 2. Pacientes menores de 27 años con diagnóstico de Hypertension, Obesity o Other

En esta rama, el 39,8 % (154 de 387) obtuvo un test "Normal", mientras que el 31,5 % (122 de 387) fue "Abnormal", y el 28,7 % (111 de 387) "Inconclusive".

Aunque se observa un leve predominio de resultados normales, más del 60 % de los pacientes aún presentan resultados dudosos o alterados. Esto sugiere que, aun en edades tempranas, condiciones como hipertensión u obesidad pueden tener efectos clínicos significativos que requieren monitoreo riguroso.

### 3. Pacientes con Asthma, Diabetes u Obesity, con admisión Urgent y edad $\leq 39$

Aquí, el 36,0 % arrojó un resultado "Normal", el 33,3 % fue "Abnormal" y el 30,7 % "Inconclusive".

En contextos de ingreso urgente, incluso pacientes jóvenes con enfermedades crónicas como asma, diabetes u obesidad presentan una alta tasa de resultados no concluyentes o anormales (64 % en total). Esto pone de manifiesto la complejidad diagnóstica en estos casos, y refuerza la necesidad de intervención rápida y eficiente.

### 4. Pacientes entre 40 y 51 años con grupo sanguíneo A+, AB+, AB–, B+ o B–

Este segmento arroja un 31,7 % de resultados "Normal", frente a un 31,1 % "Abnormal" y un 37,2 % "Inconclusive".

En este grupo etario intermedio, los tests inconclusos predominan, lo cual sugiere que este perfil presenta ambigüedades clínicas difíciles de resolver en una sola evaluación. Esto podría deberse a condiciones subyacentes no capturadas por las variables registradas o a limitaciones en la especificidad del estudio clínico.

## Conclusión

Los árboles de decisión revelan que incluso al segmentar pacientes con múltiples variables clínicas, una proporción significativa de los casos termina en resultados "Inconclusive" o "Abnormal", lo que evidencia una falta de precisión diagnóstica en varias combinaciones clínicas. Además, las ramas no permiten identificar con alta certeza grupos claramente normales, lo que refuerza la idea de que el dataset presenta una alta heterogeneidad clínica, baja discriminación entre clases y posibles limitaciones en la calidad o especificidad de los tests realizados.

## Test & Score

Para valorar el desempeño del modelo en la predicción de la variable Test Results, se recurrió a métricas estándar como AUC (Área Bajo la Curva ROC), Accuracy (Precisión Global), Precision, Recall, F1-score y MCC (Coeficiente de Correlación de Matthews). Estas medidas ofrecen una visión integral sobre la capacidad del modelo para clasificar correctamente a los pacientes dentro de las categorías clínicas "Normal", "Abnormal" e "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



De los tres algoritmos entrenados (Árbol de Decisión, Regresión Logística y Red Neuronal), el Árbol de Decisión fue el que mostró ligeramente mejor rendimiento, aunque las diferencias con los otros modelos fueron mínimas. Por esta razón, se tomó dicho árbol como referencia principal para el análisis.

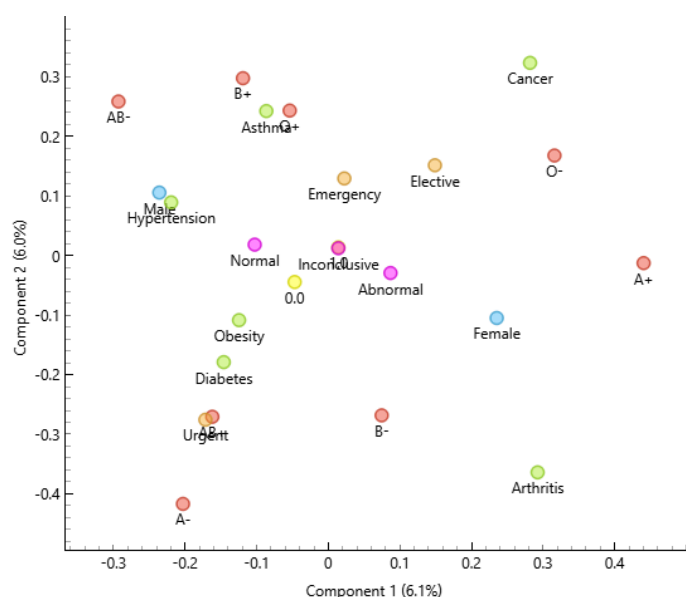
### Área Bajo la Curva ROC (AUC) = 0,510

Esta métrica representa la capacidad del modelo para distinguir entre clases. En este caso, un valor de 0,510 indica que el árbol apenas supera la aleatoriedad (un modelo completamente aleatorio tendría  $AUC = 0,500$ ). Es decir, la habilidad del árbol para separar correctamente los resultados “Normal”, “Abnormal” e “Inconclusive” es prácticamente nula. En términos prácticos, este modelo no aporta valor real en cuanto a discriminación de clases, lo que sugiere que la información disponible en las variables predictoras no es suficiente para predecir con certeza el tipo de resultado clínico del paciente.

Este bajo AUC, combinado con valores moderados de Accuracy ( $\sim 0,62$ ), F1-score ( $\sim 0,63$ ) y un MCC negativo ( $-0,006$ ), evidencia que el modelo no logra generalizar ni identificar patrones sólidos para esta tarea. Si bien puede acertar en algunos casos, lo hace sin una lógica consistente detrás, lo cual puede explicarse por ruido, datos insuficientes o variables poco relevantes.

### Correspondence Analysis

El Correspondence Analysis confirma de forma visual que cada clase (“Normal”, “Abnormal” e “Inconclusive”) se comportan casi idénticamente en función de las variables predictoras utilizadas. En conjunto, estas variables no aportan información suficiente para distinguir con claridad a los pacientes “Normal” de los “Abnormal” o “Inconclusive”, lo que explica por qué el modelo rinde prácticamente como lanzar una moneda.



**CA = 0,343**

Este valor representa el porcentaje de clasificaciones correctas. Con tres clases relativamente balanceadas, un modelo que asigna etiquetas al azar alcanzaría un 0,33 (33,3 %) de aciertos. En cambio, nuestro árbol de decisión acierta únicamente al 34,3 % de los pacientes; de cada 100, clasifica correctamente a unos 34. Es decir, el modelo apenas supera al azar por un 1 % y, por lo tanto, su desempeño global sigue siendo muy bajo.

**Precisión = 0,342**

La precisión mide, de todas las veces que el modelo asigna la etiqueta X (ya sea “Normal”, “Abnormal” o “Inconclusive”), qué proporción de esas asignaciones es realmente correcta. En este caso, cuando el modelo predice que un paciente pertenece a la clase X, solo acierta en el 34,2 % de los casos. Dicho de otro modo, de cada 100 predicciones que hace para esa clase, alrededor de 66 son falsos positivos (etiquetó al paciente como X, pero en realidad no lo es), lo que evidencia un exceso de falsos positivos.

**Recall = 0,343**

Recall mide qué proporción de los casos positivos reales el modelo identifica correctamente. En este modelo, un Recall de 0,343 indica que, de cada 100 pacientes que verdaderamente pertenecen a una clase (por ejemplo, “Normal”), sólo 34 fueron identificados correctamente y los otros 66 se perdieron como falsos negativos.

En salud, esto significa que la mayoría de los pacientes que requieren seguimiento o tratamiento no serían detectados, lo cual puede derivar en retrasos en el diagnóstico y empeoramiento de su condición. Por ello, en medicina se suele exigir un Recall mucho más alto (idealmente por encima de 0,80 – 0,90) para minimizar los falsos negativos y garantizar la seguridad del paciente.

**F1 - Score = 0,341**

La media entre precisión y recall es de 0,341, muy cercana a 0,33 (azar puro). Esto indica un desempeño deficiente:

- De cada 100 pacientes a los que el modelo asigna “Normal”, sólo 34 son realmente “Normal” (precisión baja).
- De cada 100 pacientes que en verdad son “Normal”, sólo 34 son detectados; los otros 66 se pierden como falsos negativos (recall bajo).

Un F1 del 34,1 % implica que el modelo etiqueta incorrectamente a muchos pacientes (falsos positivos) o no reconoce a otros en su clase real (falsos negativos). En la práctica:

- **Falsos negativos** (pacientes no detectados) pueden sufrir demoras en el tratamiento, empeorar su condición y enfrentar complicaciones serias.
- **Falsos positivos** (pacientes sanos etiquetados erróneamente) acarrear exámenes innecesarios, costos adicionales, ansiedad en el paciente y sobrecarga del sistema de salud.

Por todo esto, un F1 de 0,34 sugiere que el árbol de decisión no es apto para decisiones clínicas: Es prácticamente un resultado al azar.

## Confusion Matrix

Logistic Regression					
Tree					
Neural Network					
		Predicted			
		Abnormal	Inconclusive	Normal	$\Sigma$
Actual	Abnormal	478	321	316	1115
	Inconclusive	430	380	290	1100
	Normal	441	354	266	1061
$\Sigma$		1349	1055	872	3276

Clase	Explicación
<b>Abnormal</b>	De los 1.115 pacientes que en verdad son "Abnormal", el modelo solo identificó correctamente a 478 pacientes. El resto (321+316) son falsos negativos: El modelo no detecta esos "Abnormal" y los asigna a otras dos clases. Son falsos negativos porque "se pierden" dado que el modelo no los reconoce como "Abnormal" cuando de hecho, lo son...
<b>Inconclusive</b>	De los 1.100 pacientes que en verdad son "Inconclusive", el modelo identificó correctamente a 380 pacientes. El resto (430 + 290 = 720) son falsos negativos: El modelo no detecta esos "Inconclusive" y los etiqueta como "Abnormal" (430 casos) o como "Normal" (290 casos). Esos 720 se "pierden" porque, aunque son inconclusos en realidad, el modelo los clasifica en otra clase.
<b>Normal</b>	Normal:

	De los 1.061 pacientes que en verdad son "Normal", el modelo identificó correctamente a 266 pacientes. El resto ( $441 + 354 = 795$ ) son falsos negativos: El modelo no detecta esos "Normal" y los asigna a "Abnormal" (441 casos) o a "Inconclusive" (354 casos). Esos 795 se "pierden" porque, siendo sanos, el modelo los clasifica erróneamente como "Abnormal" o sin diagnóstico claro ("Inconclusive")
--	--

## Conclusiones de la matriz de confusión

### Tendencia a sobre Etiquetar "Abnormal"

Podemos inferir que el modelo tiende a sobre etiquetar la clase "Abnormal". La columna "Predicted Abnormal" suma 1 349 predicciones, mientras que "Predicted Normal" solo alcanza 872 y "Predicted Inconclusive" 1 055. Esto muestra que el árbol clasifica muchos casos como "Abnormal" aun cuando no lo sean.

### Baja identificación de pacientes sanos ("Normal")

A simple vista, la celda "Actual Normal – Predicted Normal" es la más pequeña de su fila (266 pacientes), en contraste con "Predicted Abnormal" (441) y "Predicted Inconclusive" (354). En otras palabras, de los 1 061 pacientes que realmente son "Normal", solo 266 se predicen correctamente; los demás se etiquetan erróneamente como "Abnormal" o "Inconclusive".

### Alta confusión entre "Inconclusive" y "Abnormal"

Al observar la fila "Actual Inconclusive", el cuadro "Predicted Abnormal" (430 pacientes) es más grande que el "Predicted Inconclusive" (380). Esto indica que, de los 1 100 casos verdaderamente "Inconclusive", más se clasifican como "Abnormal" (430) que los que el modelo acierta como "Inconclusive" (380). En otras palabras, el árbol confunde con mayor frecuencia los casos "Inconclusive" con "Abnormal" cuando debería identificarlos correctamente como "Inconclusive".