TRABAJO PRÁCTICO BIG DATA - EQUIPO 8

1. Selección del Dataset asignado: revisá según tu equipo que dataset tenes que utilizar.

Link dataset: **Healthcare dataset**

2. Definición del Problema: deberán definir un problema específico de negocio o investigación que quieran resolver. Esta definición debe incluir una clara formulación de qué quieren predecir o clasificar y por qué es importante o relevante.

El objetivo de este proyecto es predecir la variable de **"Test Results"** de los pacientes, que están clasificados como **"Normal"** (paciente sano), **"Abnormal"** (paciente enfermo) o **"Inconclusive"** (testeo defectuoso).

La idea es construir un modelo que, a partir de datos como la edad, condición médica, medicación y otras variables, pueda anticipar el resultado más probable del test médico de un paciente.

Esta predicción puede ser útil para: Detectar casos de riesgo antes de obtener los resultados reales del laboratorio, priorizar recursos y atención médica en pacientes con mayor probabilidad de resultados anormales y optimizar los procesos hospitalarios, ayudando a los médicos a tomar decisiones más rápidas.

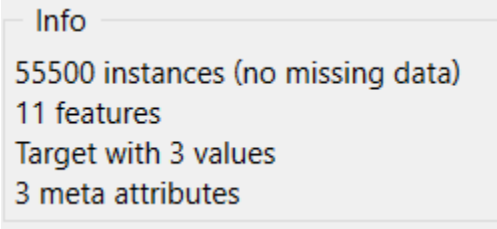
- Variable a predecir: **Test Results**

3. Análisis Exploratorio de Datos:

- Realicen un análisis exploratorio inicial para familiarizarse con los datos.

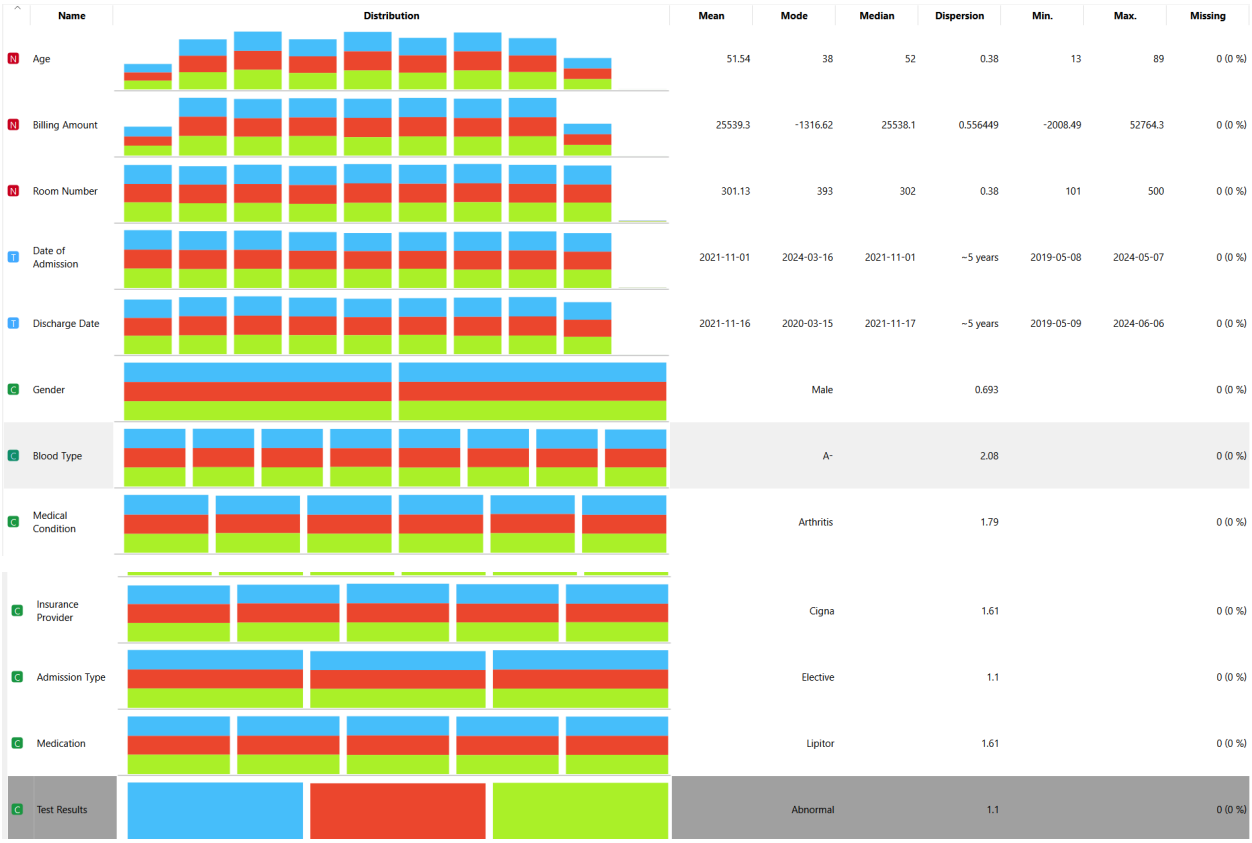
¿Cuántos registros y columnas tiene el dataset?

El dataset está compuesto por 55.000 registros y 15 columnas.



¿Hay valores nulos o repetidos?

El dataset no tiene valores nulos en ninguna de sus variables. Podemos verlo en la columna "missing" donde todas las variables tienen un 0%.



Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
Age		51.54	38	52	0.38	13	89	0 (0 %)
Billing Amount		25539.3	-1316.62	25538.1	0.556449	-2008.49	52764.3	0 (0 %)
Room Number		301.13	393	302	0.38	101	500	0 (0 %)
Date of Admission		2021-11-01	2024-03-16	2021-11-01	~5 years	2019-05-08	2024-05-07	0 (0 %)
Discharge Date		2021-11-16	2020-03-15	2021-11-17	~5 years	2019-05-09	2024-06-06	0 (0 %)
Gender			Male		0.693			0 (0 %)
Blood Type			A-		2.08			0 (0 %)
Medical Condition			Arthritis		1.79			0 (0 %)
Insurance Provider			Cigna		1.61			0 (0 %)
Admission Type			Elective		1.1			0 (0 %)
Medication			Lipitor		1.61			0 (0 %)
Test Results			Abnormal		1.1			0 (0 %)

¿Qué tipo de variables hay?

El dataset está compuesto por las siguientes variables:

- Numeric: Age, Billing Amount, y Room Number
- Categoric: Gender, Blood Type, Medical Condition, Insurance Provider, Admission Type, Medication y Test Results
- Datetime: Date of Admission y Discharge Date
- Text: Name, Doctor y Hospital

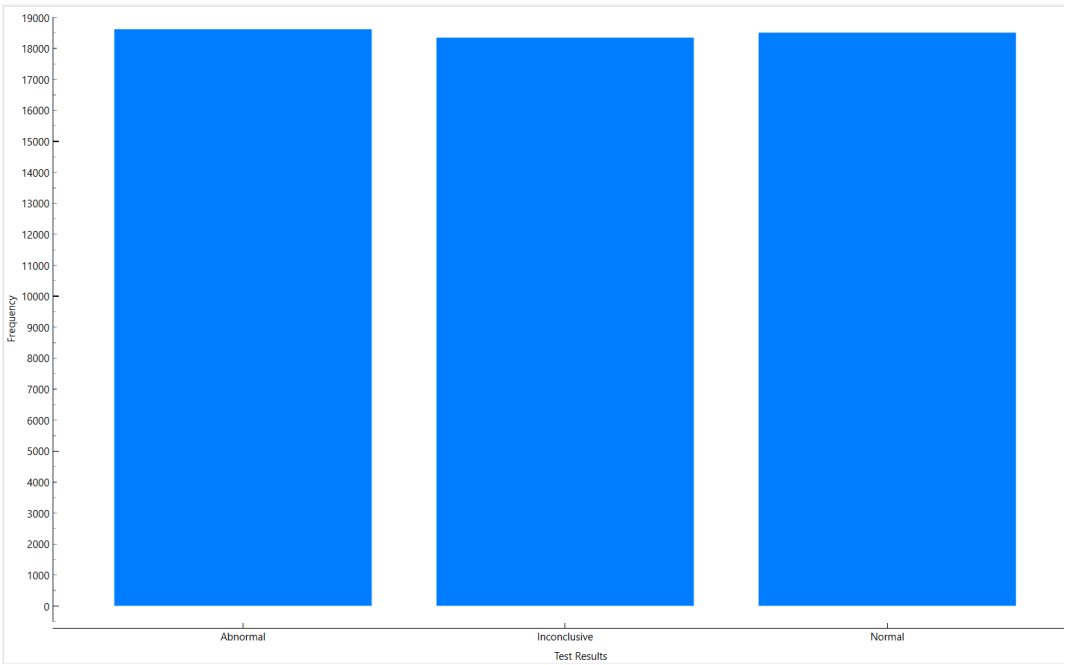
	Name	Type	Role	Values
1	Age	numeric	feature	
2	Gender	categorical	feature	Female, Male
3	Blood Type	categorical	feature	A+, A-, AB+, AB-, B+, B-, O+, O-
4	Medical ...	categorical	feature	Arthritis, Asthma, Cancer, Diabetes, Hypertension, Obesity
5	Date of ...	datetime	feature	
6	Insurance ...	categorical	feature	Aetna, Blue Cross, Cigna, Medicare, UnitedHealthcare
7	Billing Amount	numeric	feature	
8	Room Number	numeric	feature	
9	Admission Type	categorical	feature	Elective, Emergency, Urgent
10	Discharge Date	datetime	feature	
11	Medication	categorical	feature	Aspirin, Ibuprofen, Lipitor, Paracetamol, Penicillin
12	**Test Results**	**categorical**	**target**	**Abnormal, Inconclusive, Normal**
13	Name	text	meta	
14	Doctor	text	meta	
15	Hospital	text	meta	

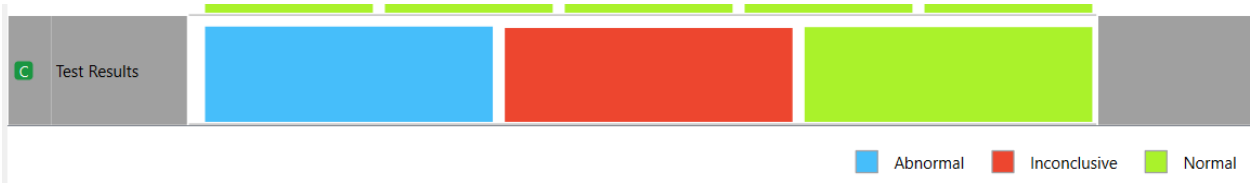
¿Qué representa cada variable?

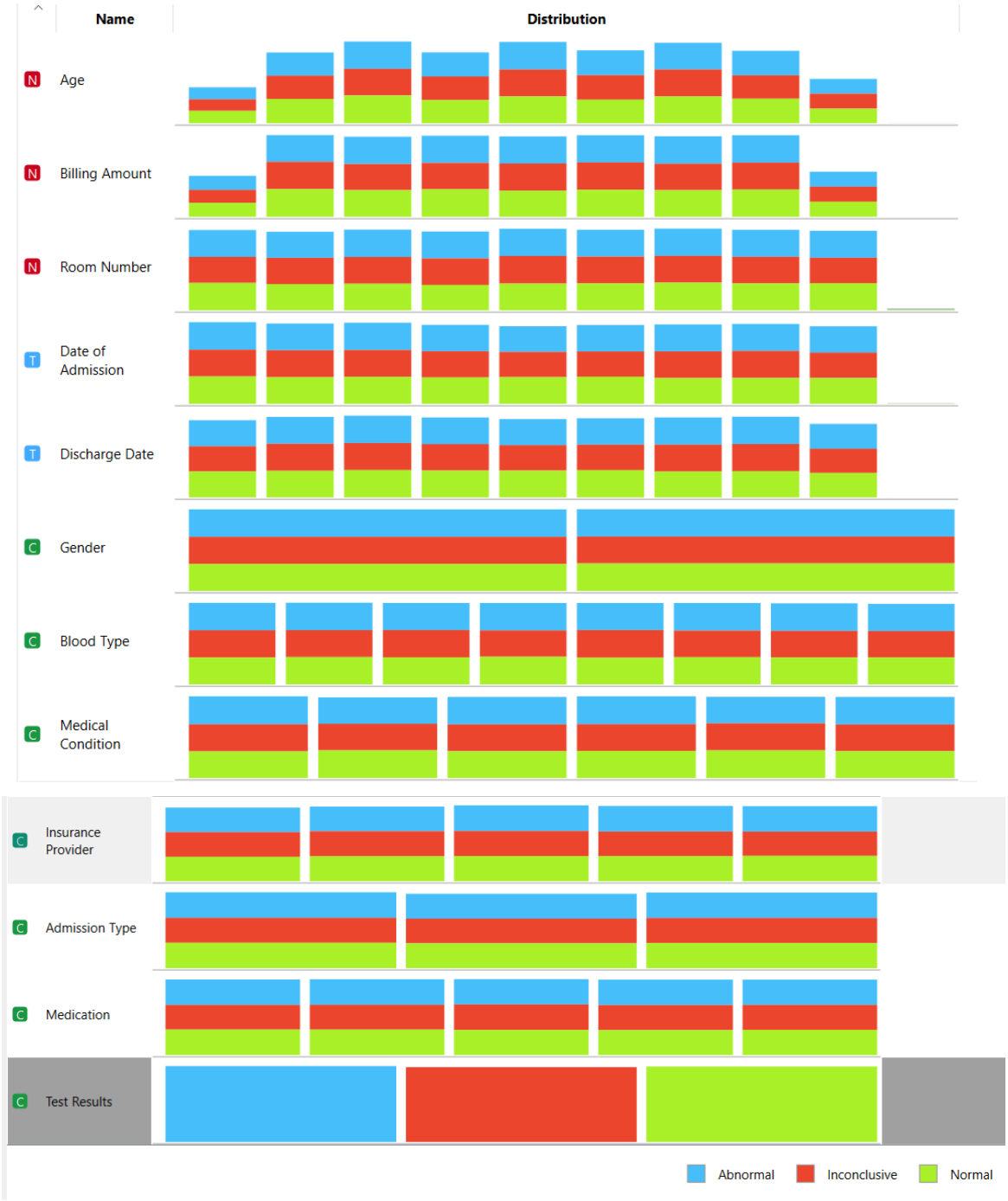
Las variables del dataset representan distintos tipos de información. En primer lugar, las variables numéricas registran valores enteros o decimales, como por ejemplo la edad de los pacientes. Las variables categóricas se dividen en clases o categorías, como sucede con la variable Gender, que puede tomar los valores Male o Female. Las variables de tipo fecha (datetime) contienen información temporal, como la Date of Admission o Discharge Date. Por último, las variables de tipo texto almacenan cadenas de caracteres que pueden variar libremente, como los nombres de los pacientes, médicos u hospitales.

¿Cómo están distribuidos los valores?

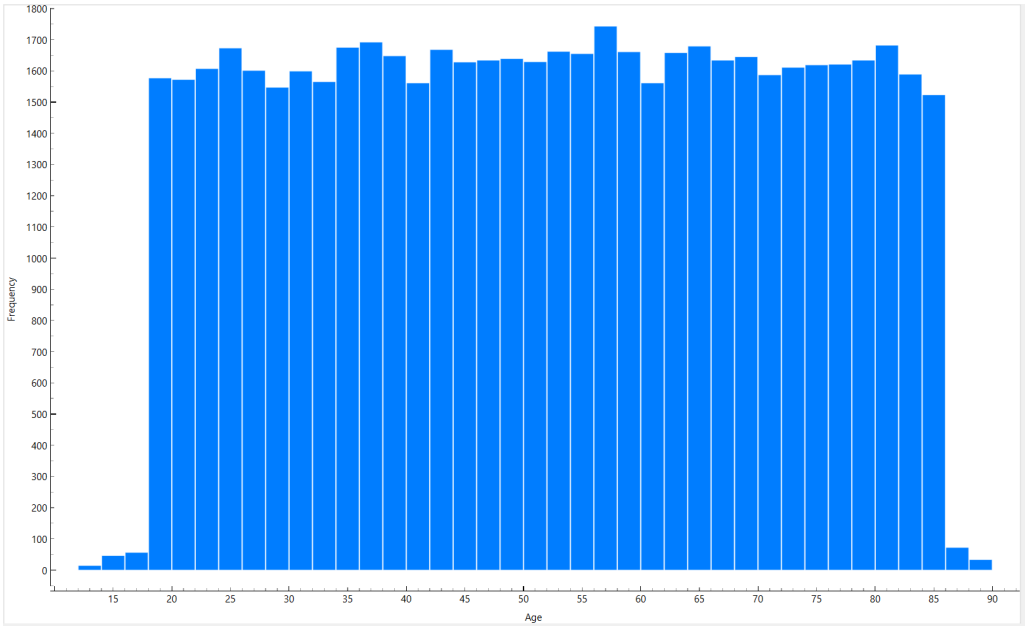
La distribución variable target, Test Results, presenta una distribución equilibrada entre sus 3 clases: Abnormal, Inconclusive y Normal.


19000
18000
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency
Abnormal
Inconclusive
Normal
Test Results

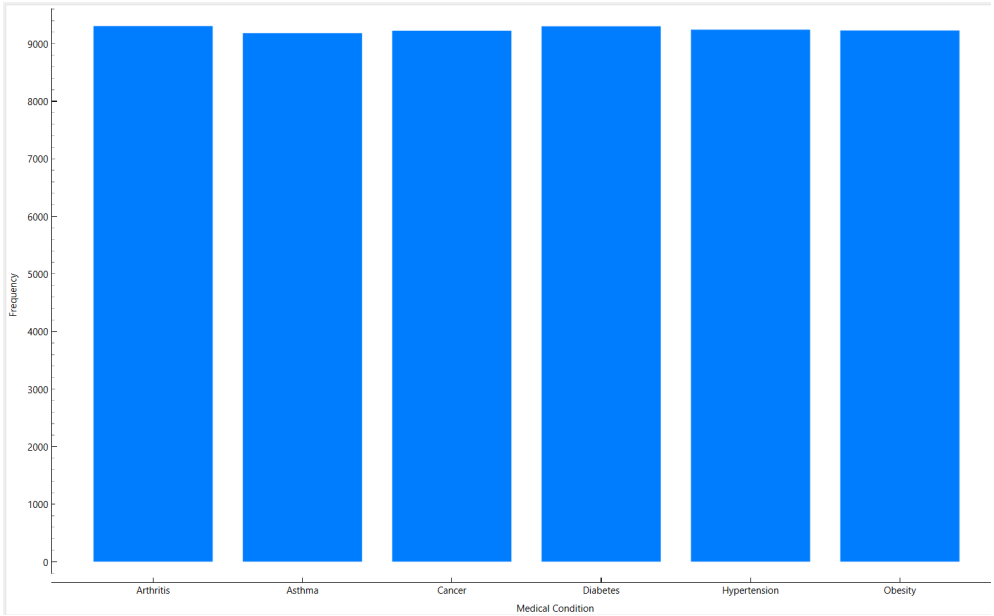

Test Results
Abnormal
Inconclusive
Normal



La variable edad también presenta una distribución equilibrada teniendo la mayor concentración de personas en el rango etario desde los 18 hasta los 86 años.



La distribución de las condiciones médicas también refleja tener una distribución equilibrada entre las mismas.



- Planteen hipótesis sobre qué variables creen que son más relevantes para predecir el resultado de interés.

Las variables que creemos pueden ser relevantes para predecir nuestro resultado de interés son Age, Gender, Medication, Insurance Provider, Blood Type y Medical Condition.

- Utilicen gráficos y estadísticas para apoyar estas hipótesis y documenten sus hallazgos.

Correlaciones:

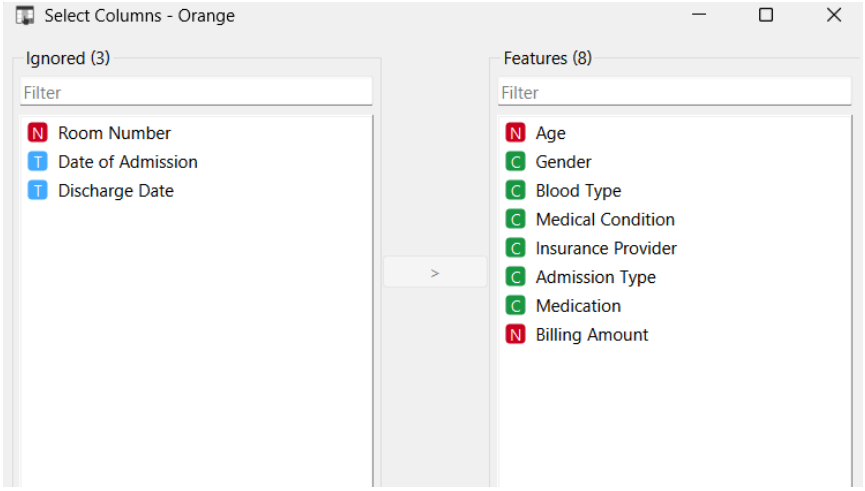
1	-0.503	Test Results=0	:	Test Results=2
2	-0.500	Test Results=0	:	Test Results=1
3	-0.009	Medical Condition=Hypertension	:	Test Results=0
4	-0.008	Medical Condition=Asthma	:	Test Results=0
5	-0.007	Insurance Provider=Blue Cross	:	Test Results=0
6	+0.007	Medical Condition=Arthritis	:	Test Results=0
7	+0.005	Medical Condition=Diabetes	:	Test Results=0
8	+0.004	Insurance Provider=Medicare	:	Test Results=0
9	+0.003	Medical Condition=Obesity	:	Test Results=0
10	-0.003	Admission Type=Emergency	:	Test Results=0
11	-0.003	Blood Type=4	:	Test Results=0
12	+0.003	Insurance Provider=UnitedHealthcare	:	Test Results=0
13	+0.003	Age	:	Test Results=0
14	-0.003	Insurance Provider=Aetna	:	Test Results=0
15	-0.003	Medication=Lipitor	:	Test Results=0
16	+0.003	Insurance Provider=Cigna	:	Test Results=0
17	+0.003	Blood Type=6	:	Test Results=0
18	+0.003	Admission Type=Elective	:	Test Results=0
19	-0.003	Blood Type=2	:	Test Results=0
20	-0.002	Gender=1	:	Test Results=0
21	+0.002	Gender=0	:	Test Results=0

1	-0.500	Test Results=0	:	Test Results=1
2	-0.497	Test Results=1	:	Test Results=2
3	-0.008	Insurance Provider=UnitedHealthcare	:	Test Results=1
4	+0.008	Blood Type=4	:	Test Results=1
5	+0.007	Insurance Provider=Blue Cross	:	Test Results=1
6	+0.006	Medication=Lipitor	:	Test Results=1
7	-0.005	Blood Type=3	:	Test Results=1
8	+0.005	Age	:	Test Results=1
9	+0.005	Insurance Provider=Aetna	:	Test Results=1
10	-0.004	Medication=Aspirin	:	Test Results=1
11	-0.004	Admission Type=Elective	:	Test Results=1
12	+0.004	Billing Amount	:	Test Results=1
13	-0.004	Insurance Provider=Medicare	:	Test Results=1
14	-0.004	Gender=1	:	Test Results=1
15	+0.004	Gender=0	:	Test Results=1
16	+0.003	Medical Condition=Hypertension	:	Test Results=1
17	+0.003	Blood Type=0	:	Test Results=1
18	-0.003	Medication=Ibuprofen	:	Test Results=1
19	-0.003	Blood Type=5	:	Test Results=1
20	-0.003	Medical Condition=Diabetes	:	Test Results=1
21	+0.003	Blood Type=2	:	Test Results=1

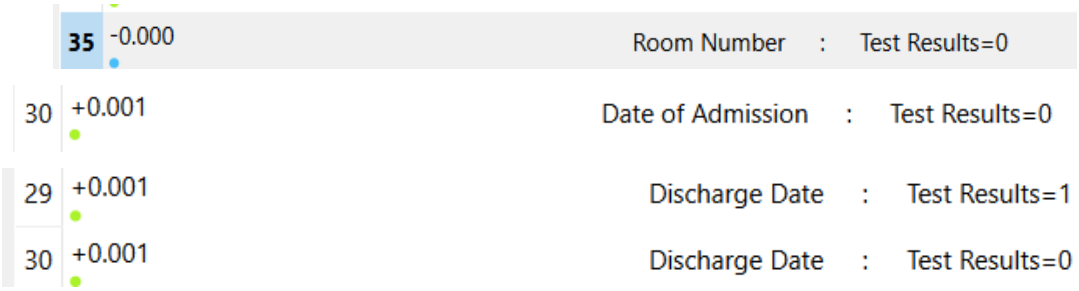
1	-0.503	Test Results=0	:	Test Results=2
2	-0.497	Test Results=1	:	Test Results=2
3	+0.008	Medical Condition=Asthma	:	Test Results=2
4	-0.008	Age	:	Test Results=2
5	-0.008	Medical Condition=Arthritis	:	Test Results=2
6	-0.006	Gender=0	:	Test Results=2
7	+0.006	Gender=1	:	Test Results=2
8	+0.006	Medical Condition=Hypertension	:	Test Results=2
9	+0.005	Blood Type=3	:	Test Results=2
10	+0.005	Insurance Provider=UnitedHealthcare	:	Test Results=2
11	-0.005	Blood Type=4	:	Test Results=2
12	-0.004	Billing Amount	:	Test Results=2
13	-0.004	Medication=Paracetamol	:	Test Results=2
14	+0.003	Medication=Aspirin	:	Test Results=2
15	-0.003	Blood Type=0	:	Test Results=2
16	-0.003	Medical Condition=Cancer	:	Test Results=2
17	-0.003	Medication=Lipitor	:	Test Results=2
18	-0.003	Insurance Provider=Cigna	:	Test Results=2
19	-0.002	Medical Condition=Obesity	:	Test Results=2
20	-0.002	Admission Type=Urgent	:	Test Results=2
21	+0.002	Medication=Ibuprofen	:	Test Results=2

4. Preprocesamiento y Selección de Variables: Limpiar y transformar datos para que el modelo los pueda procesar. Es importante definir qué variables son o no relevantes al problema.

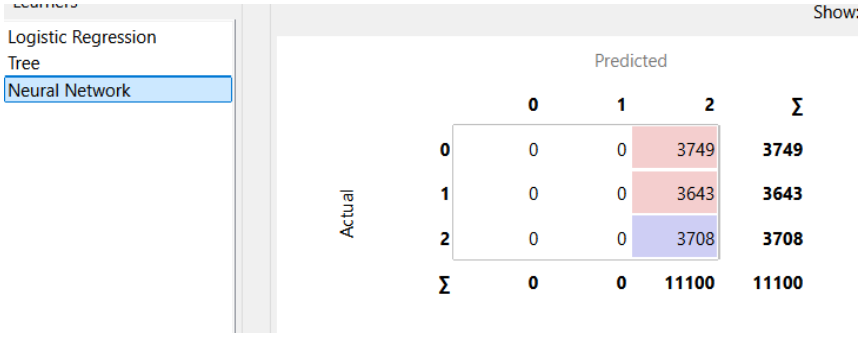
- Decidimos eliminar columnas irrelevantes: Room Number, Date of Admission y Discharge Date.



¿Por qué? Por su baja correlación con test resultados, que al fin y al cabo es lo que buscamos predecir.



Modelo con las variables eliminadas:

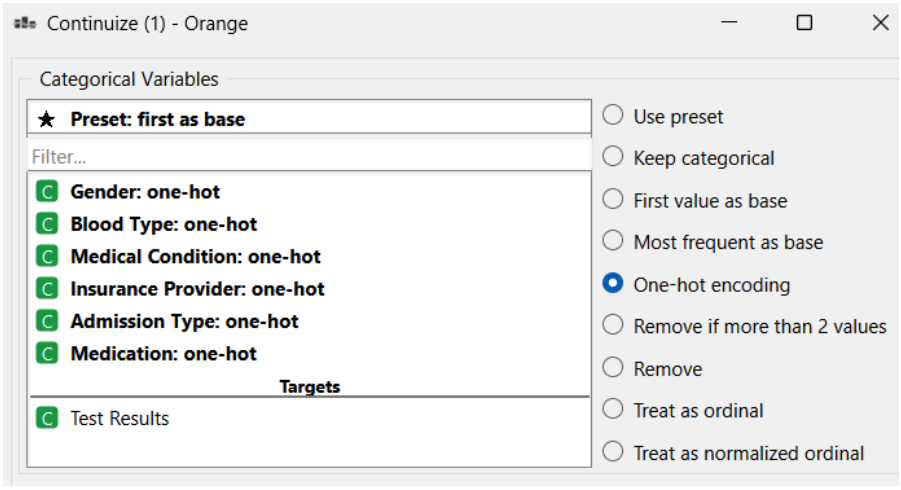


Predicted

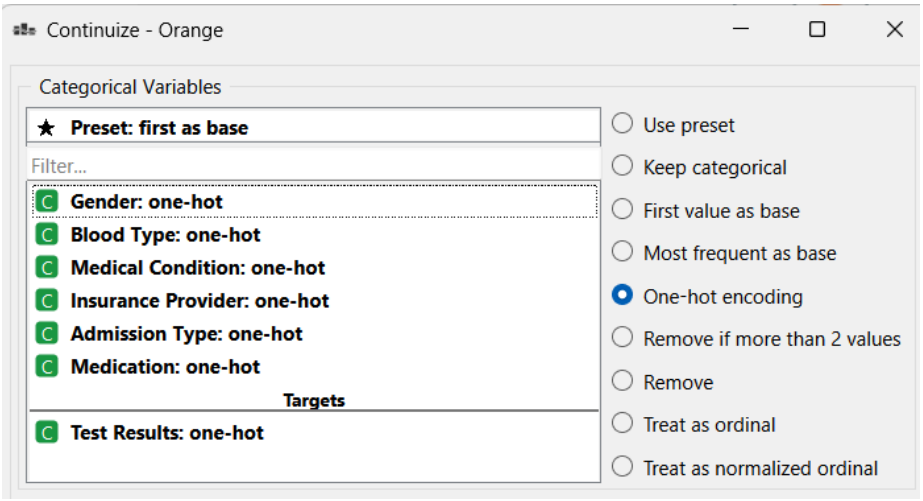
Actual	0	1	2	Σ
0	2638	1111	0	3749
1	2591	1052	0	3643
2	2619	1089	0	3708
Σ	7848	3252	0	11100

- Codificación de variables categóricas: aplicamos one-hot encoding a las variables

Continuize para entrenamiento:



Continuize para analizar las correlaciones:



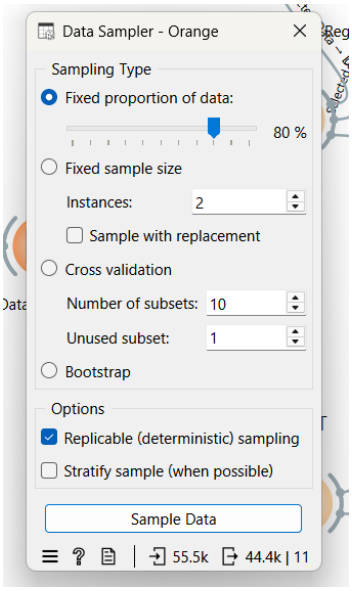
- Esto lo hicimos así porque si dejamos one-hot en la target, nos toma como que hay múltiples variables target para hacer el modelo.

5. Modelado: Prueben varios modelos de aprendizaje automático disponibles en Orange Data Mining. Deben experimentar con al menos tres tipos diferentes

Implementamos tres modelos en Orange Data Mining:

- Tree
- Logistic Regression
- Neural Network

Los tres modelos fueron entrenados y testeados con los mismos datos, donde utilizamos el 80% de los datos para entrenamiento y el 20% para la prueba.



6. Evaluación del Modelo: Evalúen los modelos utilizando las métricas apropiadas. Comparen los modelos entre sí y seleccionen el más adecuado basándose en los resultados de las métricas y los requisitos del problema.

Antes del preprocesamiento y selección de variables:

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.564	0.410	0.409	0.410	0.410	0.115
Neural Network	0.500	0.328	0.162	0.108	0.328	0.000
Logistic Regression	0.495	0.333	0.254	0.220	0.333	-0.005

Después del preprocesamiento y selección de variables:

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.494	0.332	0.330	0.331	0.332	-0.004
Tree	0.543	0.379	0.378	0.379	0.379	0.068
Neural Network	0.518	0.349	0.348	0.349	0.349	0.023

Antes de realizar la comparación de métricas, planteamos que está buscando predecir y que representa cada parte de la matriz de confusión. Tomando como Abnormal (el resultado es atípico y el paciente está enfermo), Normal (el resultado es bueno y el paciente no está enfermo) e Inconclusive (el resultado es defectuoso y no se puede determinar si el paciente está enfermo). La matriz del "tree model" previa al preprocesamiento es la siguiente:

Actual	Abnormal	Inconclusive	Normal
Abnormal	1693	1075	981
Inconclusive	1185	1499	959
Normal	1263	1085	1360
∑	**4141**	**3659**	**3300**

Posterior al preprocesamiento

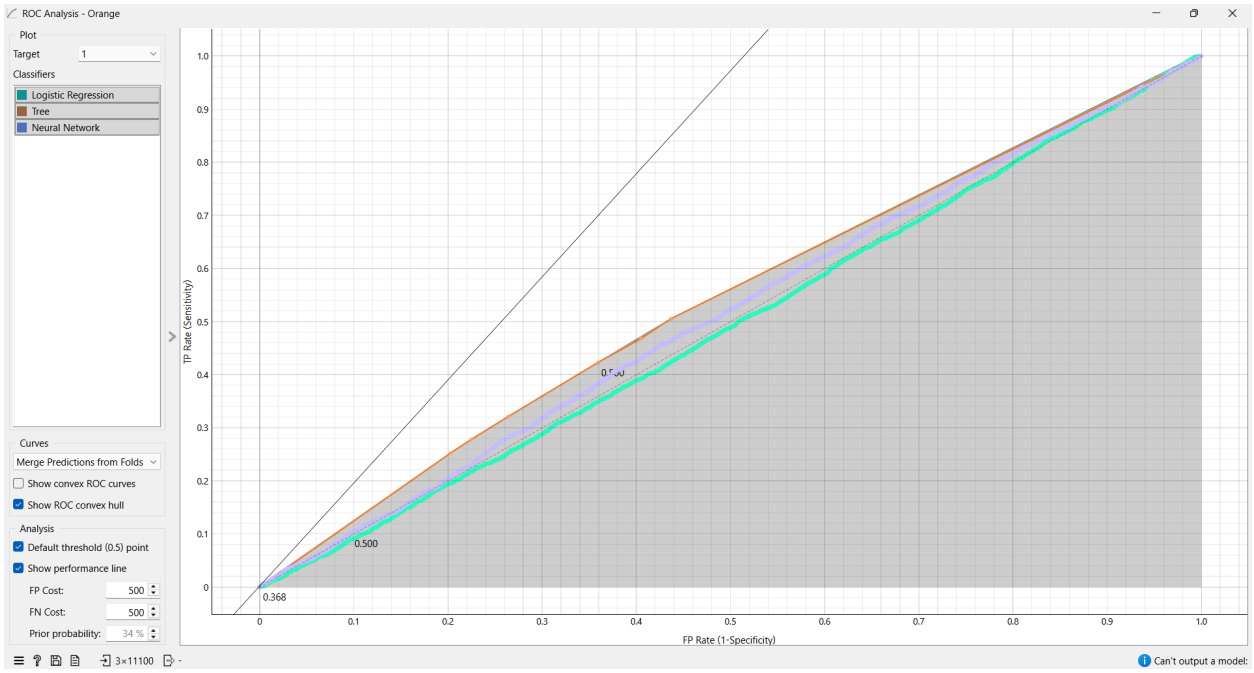
		Predicted			
		Abnormal	**Inconclusive**	**Normal**	**∑**
Actual	**Abnormal**	1613	1101	1035	**3749**
	Inconclusive	1328	1368	947	**3643**
	Normal	1369	1114	1225	**3708**
	∑	**4310**	**3583**	**3207**	**11100**

Basándonos en la clase "ABNORMAL":

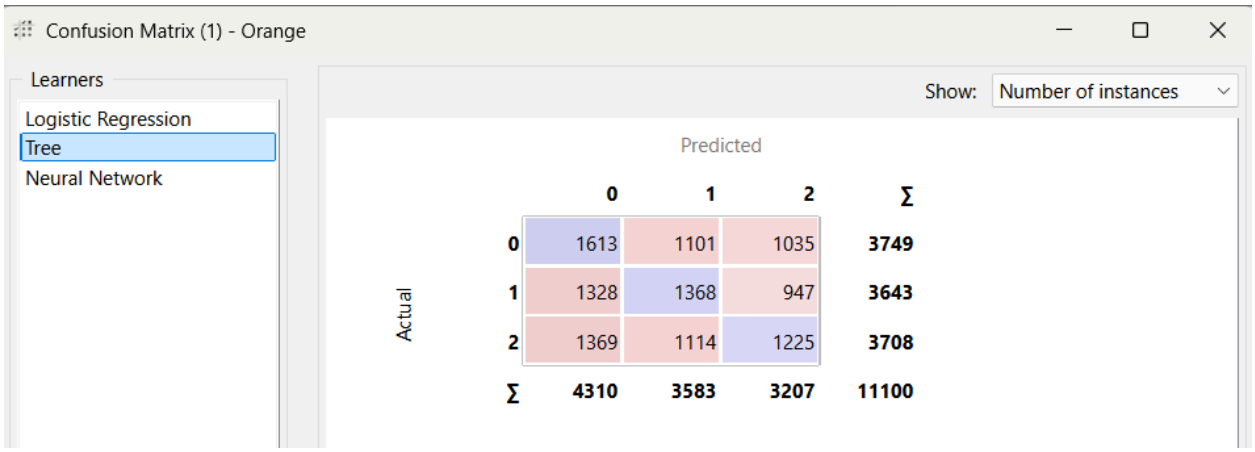
- VP (verdadero positivo): el modelo predice Abnormal y el valor real es Abnormal.
- FP (falso positivo): el modelo predice Abnormal pero el valor real es Normal o Inconclusive.
- FN (falso negativo): el modelo predice Normal o Inconclusive pero el valor real es Abnormal.
- VN (verdadero negativo): el modelo predice NO-Abnormal y el valor real es Normal o Inconclusive .

Como primer vistazo comparamos accuracy ("AUC"), donde Tree resultó el más preciso, mientras que Logistic Regression fue el menos preciso. Dado que el costo

clínico más alto es dar por sano a alguien que no lo está (un falso negativo), priorizamos la métrica de Recall, tomando este criterio, Tree también lidera como modelo. Sin embargo, ningún modelo es preciso y podemos notarlo con el análisis del ROC.



Análisis de la Matriz de Confusión:



El modelo predijo correctamente 1.613 casos anormales, son aquellos en los que tanto la predicción como el valor real coincidieron en la categoría Abnormal. Estos se clasifican como verdaderos positivos, el modelo predijo un paciente enfermo y el paciente estaba enfermo.

Sin embargo, también clasificó erróneamente 2.697 casos como “Abnormal” cuando en realidad pertenecían a las categorías Inconclusive (1.328) o Normal (1.369), es decir, falsos positivos.

Por otro lado, 2.136 casos que realmente eran anormales fueron clasificados como no anormales, es decir como Inconclusive (1.1001) o Normal (1035), con lo cual estaríamos hablando de falsos negativos. Este tipo de error es el más grave, ya que implica que el modelo no logra detectar una condición potencialmente patológica.

Finalmente, el modelo acertó en 2.593 casos no anormales, clasificándolos correctamente como Inconclusive (1.368) o Normal (1225), es decir verdaderos negativos.

Debido a que el costo de un falso negativo es muy alto en este tipo de problema —porque podría significar pasar por alto un caso realmente anormal—, la métrica más importante a considerar es la sensibilidad (recall). Esta métrica refleja la capacidad del modelo para detectar correctamente los casos anormales y, en este caso, su valor del 38% indica que el modelo aún necesita mejoras para reducir el riesgo de diagnósticos no detectados.

En el modelo de árbol de decisión, se registraron 2.136 falsos negativos, mientras que la regresión logística tuvo 2.261 y la red neuronal 2.234. Esto significa que el árbol de decisión es el que menos casos anormales deja sin detectar, un aspecto clave en este tipo de problema, donde no identificar un caso realmente anormal puede tener consecuencias graves.

Aunque las diferencias no son enormes, el modelo de árbol de decisión es el más adecuado, ya que su menor cantidad de falsos negativos y su mayor sensibilidad lo convierten en la opción preferible para evitar que los casos anormales pasen desapercibidos.

7. Interpretación de las Predicciones: Analicen qué variables son consideradas más importantes por el modelo y cómo se relacionan estas con sus hipótesis iniciales.

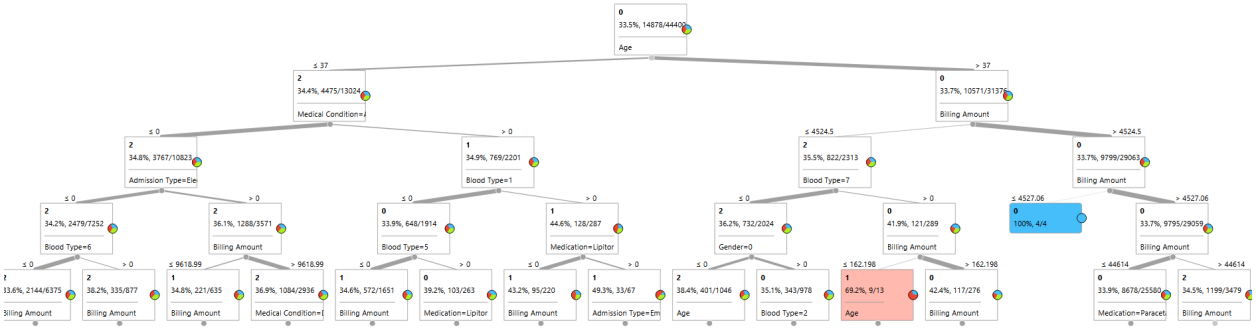
El análisis de importancia de variables en el modelo de Árbol de Decisión mostró que las más influyentes fueron:

- Age
- Medical Condition
- Billing Amount

Nuestra hipótesis inicial fue: “Las variables que creemos pueden ser relevantes para predecir nuestro resultado de interés son Age, Medication y Medical Condition.”

El modelo confirmó la importancia de la variable “Age” ya que aparece en la raíz (lo cual indica que es la que más influencia tiene sobre el test) y que se distribuye en mayores de 37 y menores de 37. Las ramas del árbol comienzan a distribuirse en “Billing Amount” y “Medical Condition” mostrando importancia en estas dos

variables. Prosiguiendo con “Blood Type” y “Admission Type” que presentan una menor importancia.



Como conclusiones podemos determinar que:

1. En jóvenes pacientes, menores de 37 años, pesan más la condición clínica registrada y el tipo de admisión (electivo/urgencia/emergencia) que el costo. El Billing Amount aporta menos al principio.
2. En los pacientes mayores de 37 años, el costo tiene mayor relevancia, debido a que puede actuar como un indicador indirecto en base a la complejidad del caso, ya que se requieren más estudios dando mayor Billing Amount.