Análisis y Modelado de Datos con Orange Data Mining

Presentación del Trabajo Práctico de Big Data del grupo Data Mavericks, enfocado en el análisis y modelado de datos de salud utilizando Orange Data Mining.

Definición del Problema: Predicción de Resultados de Pruebas

Nuestro proyecto aborda la identificación proactiva de pacientes con "Resultados de Pruebas" Anormales o Inconclusos en el Healthcare Dataset. Nuestro objetivo es predecir resultados de pruebas médicas para una intervención temprana.

Buscamos clasificar los resultados en "Normal", "Anormal" o "Inconclusos", priorizando la correcta identificación de los dos últimos para minimizar falsos negativos, que conllevan el mayor costo en salud.

Problema Crítico

Identificar pacientes con resultados de pruebas anormales o inconclusos.

Clasificación Crucial

Predecir "Anormal" o "Inconclusos" para intervención temprana.

Minimizar Falsos Negativos

Evitar la omisión de condiciones médicas serias.

Análisis Exploratorio de Datos e Hipótesis

Realizamos un análisis exploratorio inicial del Healthcare Dataset. Nuestra hipótesis principal es que el "Tipo de Admisión" está directamente relacionado con los "Resultados de Pruebas".

Los ingresos por "Emergencia" o "Urgencia" sugieren un estado de salud más crítico, lo que podría correlacionarse con una mayor probabilidad de resultados "Anormales" o "Inconclusos" en comparación con admisiones "Electivas".

Variables Clave

- Edad
- Género
- Tipo de Sangre
- Condición Médica
- Tipo de Admisión
- Resultados de Pruebas (Objetivo)

Hipótesis Principal

El "Tipo de Admisión" influye en los "Resultados de Pruebas". Admisiones de "Emergencia" o "Urgencia" se asocian con resultados "Anormales" o "Inconclusivos".

Metodología del Proyecto

Aplicaremos técnicas de machine learning, utilizando al menos tres modelos diferentes: Random Forest, Naive Bayes y Regresión Logística, para identificar patrones en los datos de los pacientes.

Mejorar Comprensión

Entender factores que influyen en resultados críticos.

Identificar Proactivamente

Permitir intervención médica temprana.

3

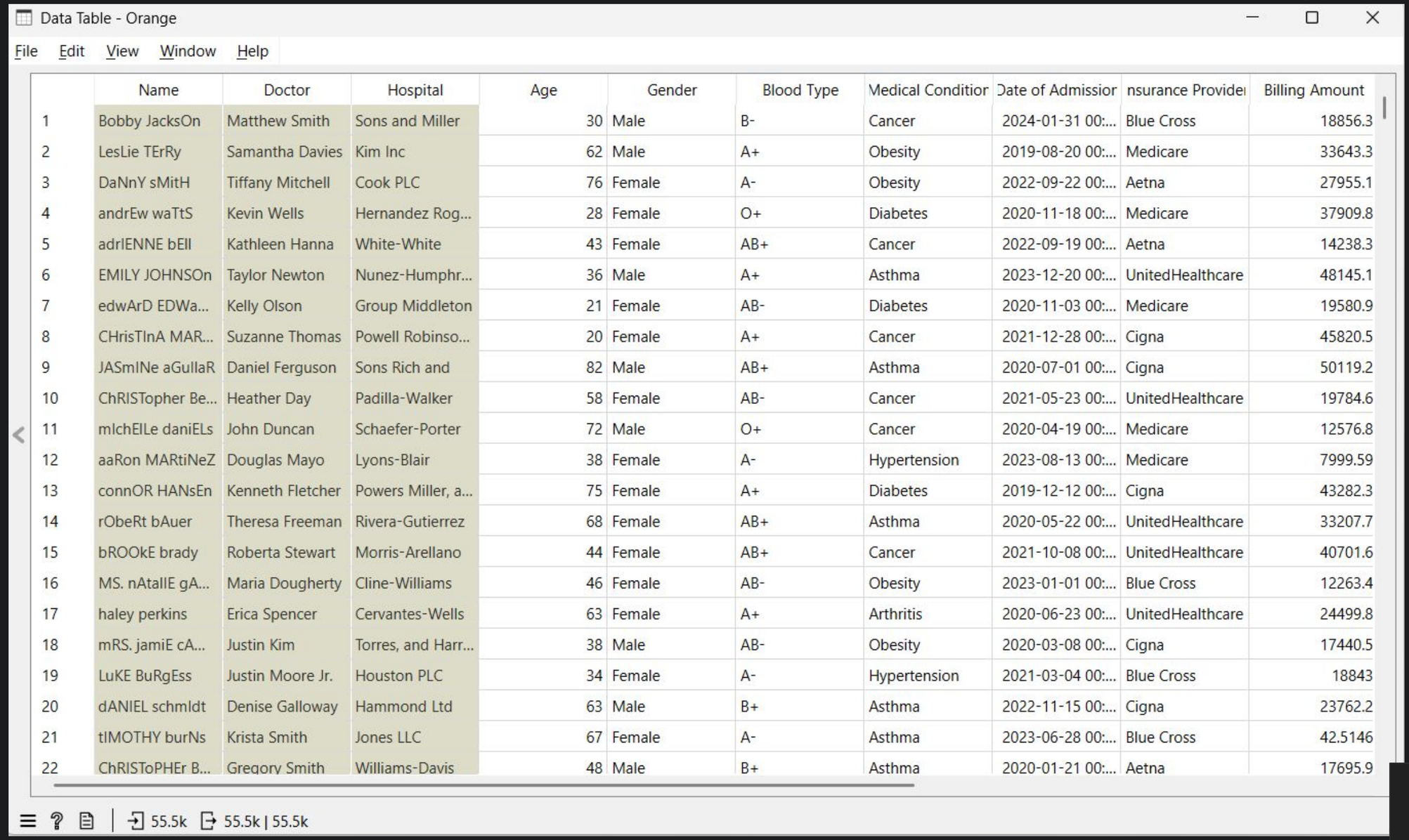
Optimizar Recursos

Enfocar atención en pacientes de mayor riesgo.

Reducir Riesgos

Minimizar la omisión de condiciones graves.

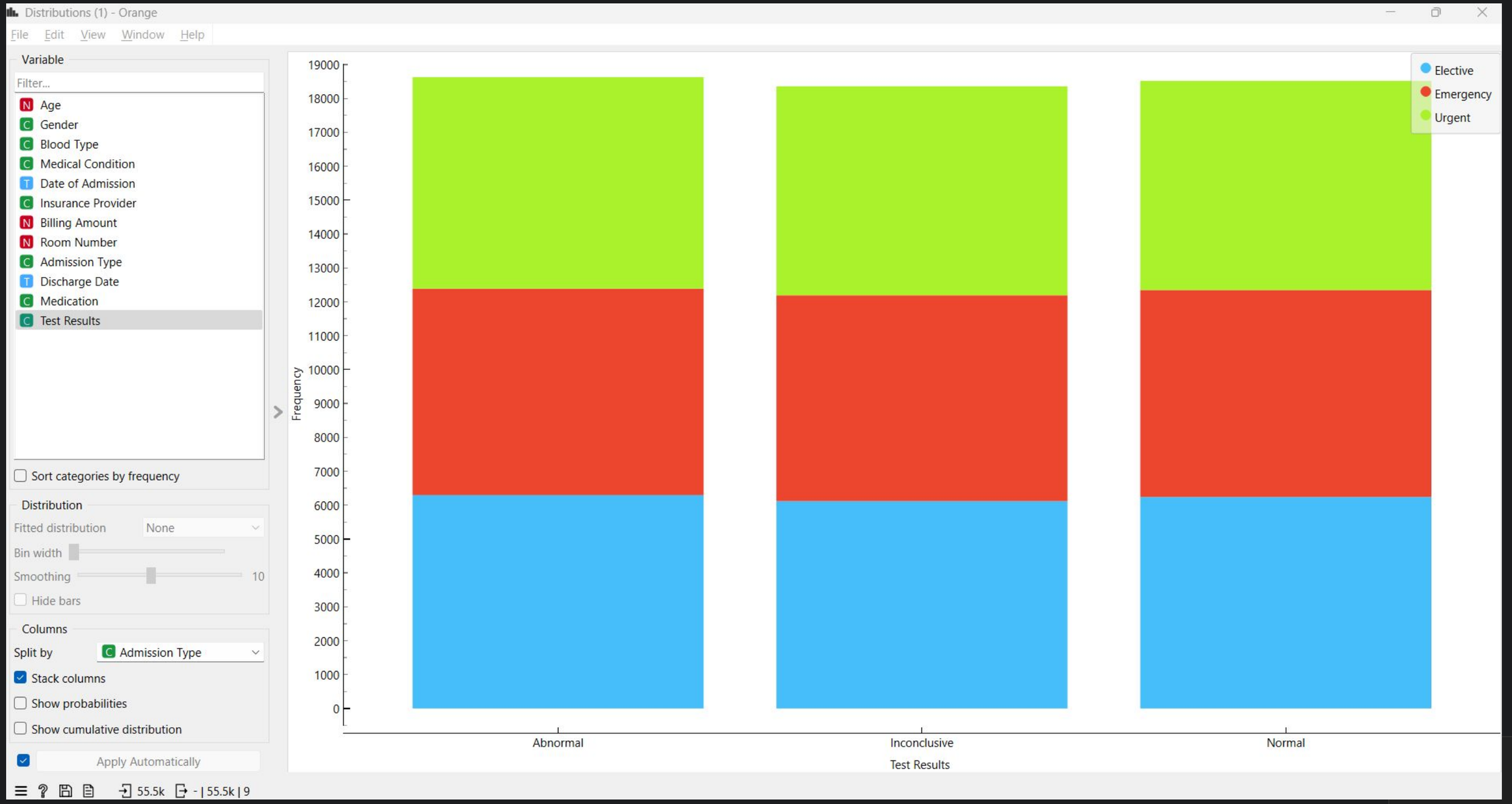
Data Table



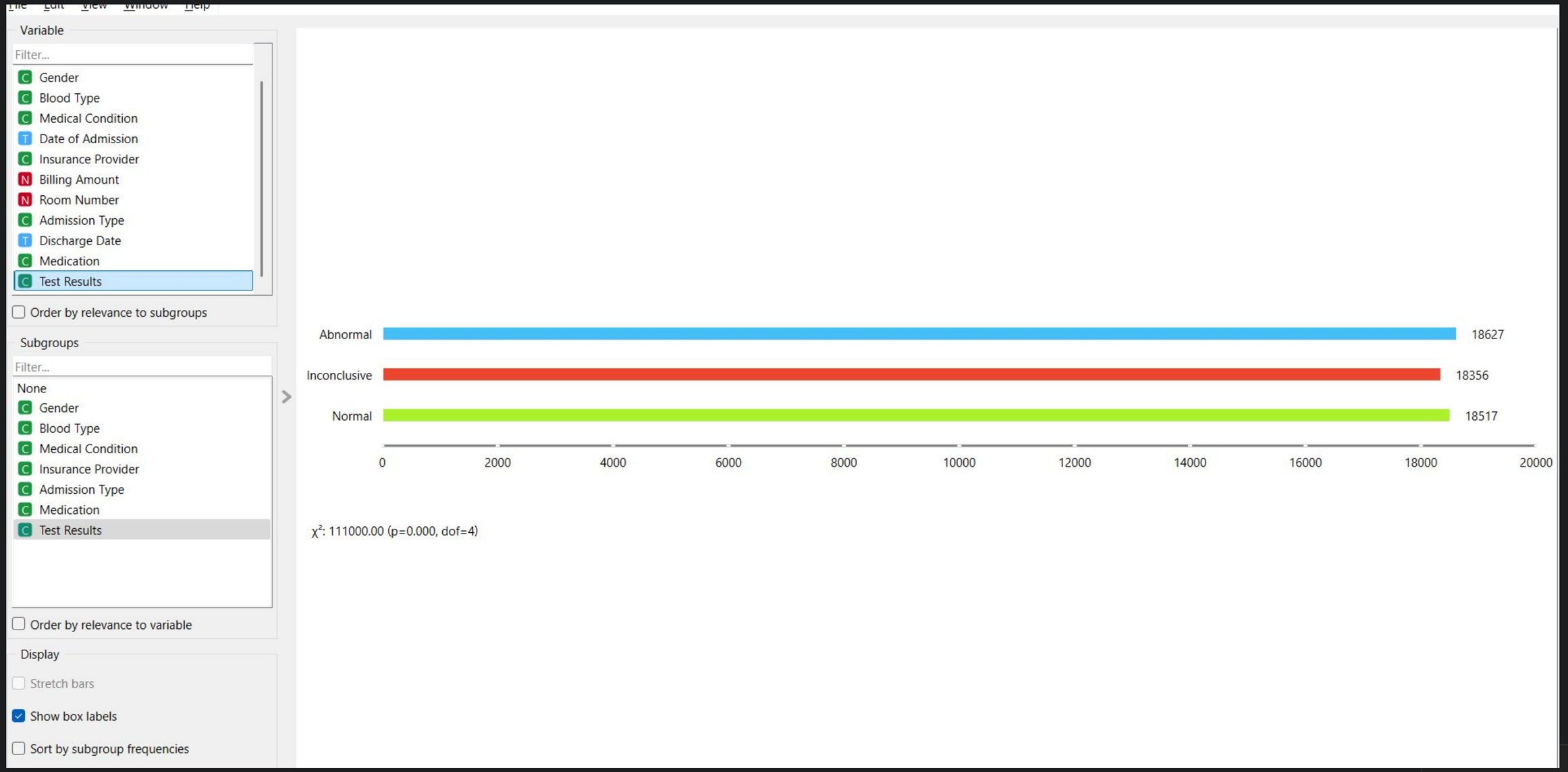
	Name	Doctor	Hospital	Age	Gender	Blood Type	Medical Conditior	Date of Admissior	nsurance Provide	Billing Amount
1	Bobby JacksOn	Matthew Smith	Sons and Miller	30	Male	B-	Cancer	2024-01-31 00:...	Blue Cross	18856.3
2	LesLie TErRy	Samantha Davies	Kim Inc	62	Male	A+	Obesity	2019-08-20 00:...	Medicare	33643.3
3	DaNnY sMitH	Tiffany Mitchell	Cook PLC	76	Female	A-	Obesity	2022-09-22 00:...	Aetna	27955.1
4	andrEw waTtS	Kevin Wells	Hernandez Rog...	28	Female	O+	Diabetes	2020-11-18 00:...	Medicare	37909.8
5	adrIENNE bEll	Kathleen Hanna	White-White	43	Female	AB+	Cancer	2022-09-19 00:...	Aetna	14238.3
6	EMILY JOHNSOn	Taylor Newton	Nunez-Humphr...	36	Male	A+	Asthma	2023-12-20 00:...	UnitedHealthcare	48145.1
7	edwArD EDWa...	Kelly Olson	Group Middleton	21	Female	AB-	Diabetes	2020-11-03 00:...	Medicare	19580.9
8	CHrisTInA MAR...	Suzanne Thomas	Powell Robinso...	20	Female	A+	Cancer	2021-12-28 00:...	Cigna	45820.5
9	JASmINe aGullaR	Daniel Ferguson	Sons Rich and	82	Male	AB+	Asthma	2020-07-01 00:...	Cigna	50119.2
10	ChRISTopher Be...	Heather Day	Padilla-Walker	58	Female	AB-	Cancer	2021-05-23 00:...	UnitedHealthcare	19784.6
11	mIchElLe daniELs	John Duncan	Schaefer-Porter	72	Male	O+	Cancer	2020-04-19 00:...	Medicare	12576.8
12	aaRon MARtiNeZ	Douglas Mayo	Lyons-Blair	38	Female	A-	Hypertension	2023-08-13 00:...	Medicare	7999.59
13	connOR HANsEn	Kenneth Fletcher	Powers Miller, a...	75	Female	A+	Diabetes	2019-12-12 00:...	Cigna	43282.3
14	rObeRt bAuer	Theresa Freeman	Rivera-Gutierrez	68	Female	AB+	Asthma	2020-05-22 00:...	UnitedHealthcare	33207.7
15	bROOkE brady	Roberta Stewart	Morris-Arellano	44	Female	AB+	Cancer	2021-10-08 00:...	UnitedHealthcare	40701.6
16	MS. nAtalIE gA...	Maria Dougherty	Cline-Williams	46	Female	AB-	Obesity	2023-01-01 00:...	Blue Cross	12263.4
17	haley perkins	Erica Spencer	Cervantes-Wells	63	Female	A+	Arthritis	2020-06-23 00:...	UnitedHealthcare	24499.8
18	mRS. jamiE cA...	Justin Kim	Torres, and Harr...	38	Male	AB-	Obesity	2020-03-08 00:...	Cigna	17440.5
19	LuKE BuRgEss	Justin Moore Jr.	Houston PLC	34	Female	A-	Hypertension	2021-03-04 00:...	Blue Cross	18843
20	dANIEL schmIdt	Denise Galloway	Hammond Ltd	63	Male	B+	Asthma	2022-11-15 00:...	Cigna	23762.2
21	tIMOTHY burNs	Krista Smith	Jones LLC	67	Female	A-	Asthma	2023-06-28 00:...	Blue Cross	42.5146
22	ChRISToPHEr B...	Gregory Smith	Williams-Davis	48	Male	B+	Asthma	2020-01-21 00:...	Aetna	17695.9

Mosaic display

Distributions



Box Plot


File Edit View Window Help
Variable
Filter...
Gender
Blood Type
Medical Condition
Date of Admission
Insurance Provider
Billing Amount
Room Number
Admission Type
Discharge Date
Medication
Test Results
Order by relevance to subgroups
Subgroups
Filter...
None
Gender
Blood Type
Medical Condition
Insurance Provider
Admission Type
Medication
Test Results
Order by relevance to variable
Display
Stretch bars
Show box labels
Sort by subgroup frequencies
Abnormal
Inconclusive
Normal
18627
18356
18517
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
χ²: 111000.00 (p=0.000, dof=4)

Preprocesamiento y Selección de Variables

El preprocesamiento de datos es fundamental para preparar el conjunto de datos para el modelado. Hemos seleccionado variables relevantes para la predicción de los "Resultados de Pruebas", excluyendo identificadores únicos o logísticos.

Variables como "Nombre", "Fecha de Admisión" y "Número de Habitación" fueron consideradas menos relevantes. Las variables seleccionadas incluyen "Edad", "Género", "Tipo de Sangre", "Condición Médica", "Tipo de Admisión" y "Medicamentos".

Variables Relevantes

- Edad
- Género
- Tipo de Sangre
- Condición Médica
- Tipo de Admisión
- Medicamentos

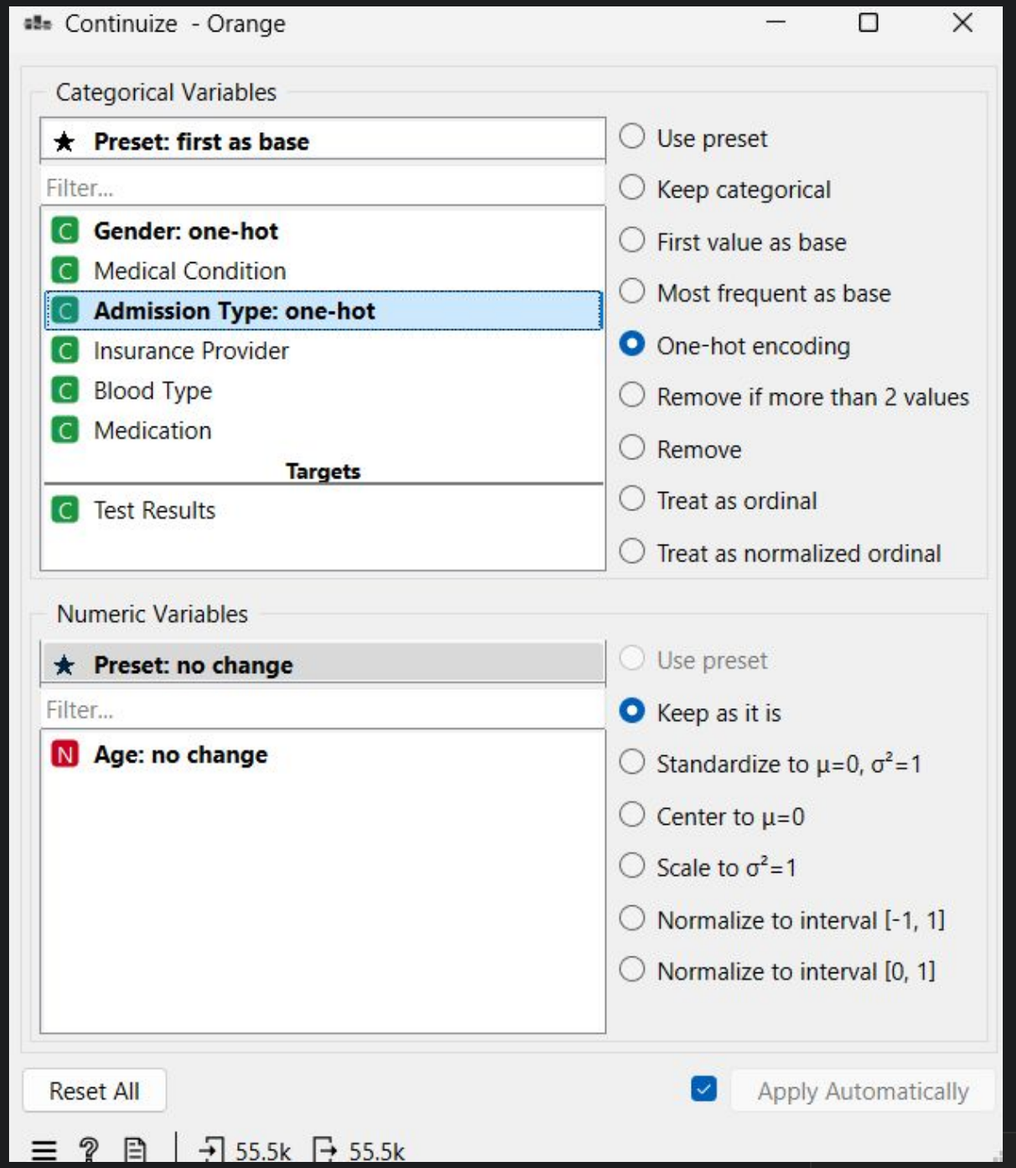
Variables Menos Relevantes

- Nombre
- Fecha de Admisión
- Fecha de Alta
- Médico
- Hospital
- Número de Habitación

Preprocesamiento y Selección de Variables

Continuize Widget

Se aplicó la transformación **one-hot encoding** a variables categóricas como **Gender y Admission Type** para que los modelos pudieran procesarlas adecuadamente. Las variables numéricas, como Age, se mantuvieron en su formato original



Modelado: Selección y Entrenamiento de Modelos

Hemos probado tres modelos de aprendizaje automático en Orange Data Mining para comparar su rendimiento en la predicción de "Resultados de Pruebas": Random Forest, Naive Bayes y tree

Cada modelo fue entrenado con el conjunto de datos de entrenamiento y evaluado con el conjunto de datos de prueba, utilizando los widgets "Test and Score" y "Predictions" para analizar su desempeño.

Random Forest

Algoritmo de conjunto robusto para clasificación.

Naive Bayes

Clasificador probabilístico, simple y rápido.

tree

Random Forest Confussion Matrix

Predicted

Actual	Abnormal	Inconclusive	Normal	∑
Abnormal	11966	516	503	**12985**
Inconclusive	529	11870	496	**12895**
Normal	533	489	11948	**12970**
∑	**13028**	**12875**	**12947**	**38850**

Show perfomance scores Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall
Random Forest	0.985	0.921	0.921	0.921	0.921

38.9k | M 38.9k | 38.9k | 1×38850

tree Confussion Matrix

Predicted

Actual	Abnormal	Inconclusive	Normal	Σ
Abnormal	12165	399	421	**12985**
Inconclusive	1378	11065	452	**12895**
Normal	1349	1302	10319	**12970**
Σ	**14892**	**12766**	**11192**	**38850**

Show perfomance scores Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall
Tree	0.979	0.864	0.863	0.869	0.864

38.9k | M 38.9k | 38.9k | 1×38850

Logistic Regression Confussion Matrix

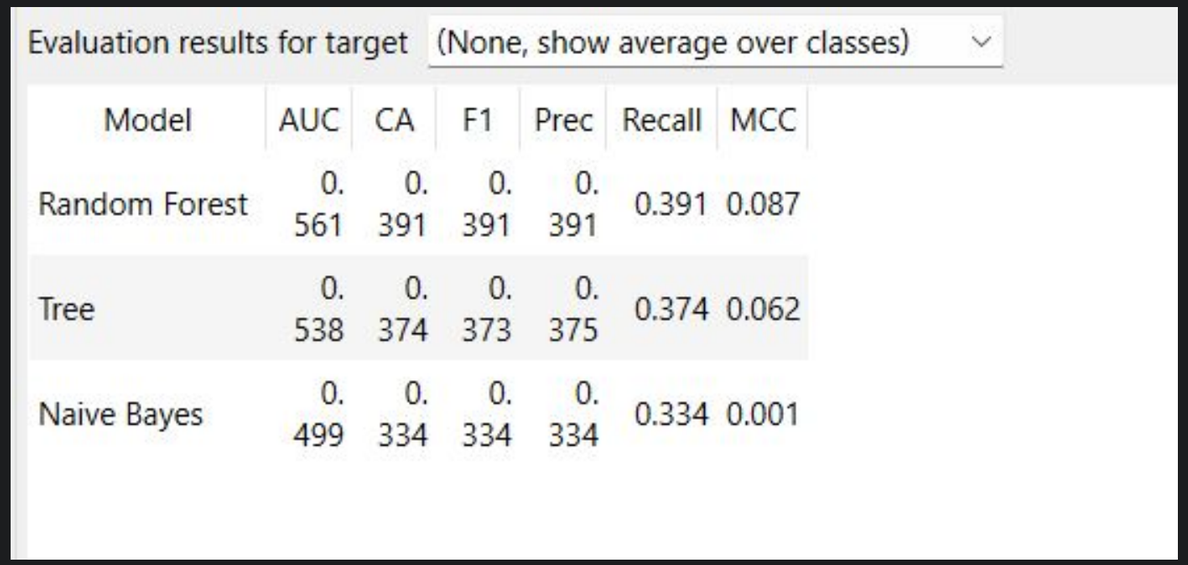
Predicted

Actual	Abnormal	Inconclusive	Normal	Σ
Abnormal	3604	3769	5612	**12985**
Inconclusive	3583	3801	5511	**12895**
Normal	3605	3703	5662	**12970**
Σ	**10792**	**11273**	**16785**	**38850**

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.502	0.336	0.333	0.336	0.336	0.004

Evaluación del Modelo: Priorizando el Recall

La evaluación es crítica, y en salud, el costo de un falso negativo es significativamente mayor. Por ello, nuestra métrica principal es el Recall (Sensibilidad) para las clases "Anormal" e "Inconclusivo", que mide la proporción de casos positivos reales correctamente identificados.
La Matriz de Confusión es fundamental para una evaluación granular, permitiéndonos observar los Verdaderos Positivos y Falsos Negativos para cada clase y seleccionar el modelo con la menor cantidad de Falsos Negativos en las clases críticas.



Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.561	0.391	0.391	0.391	0.391	0.087
Tree	0.538	0.374	0.373	0.375	0.374	0.062
Naive Bayes	0.499	0.334	0.334	0.334	0.334	0.001

Interpretación de Predicciones y Conclusiones

Estos tres modelos analizados nos dan como la variable más importante la edad, lo cual tiene sentido dado a que en los temas de salud la edad suele ser de los factores más relevantes en torno a la medicina y la gravedad del diagnóstico.

Los Admission Types aparecen tanto en el Modelo Tree como en el Random Forest, por lo que respalda nuestra hipótesis de que es una variable clave para predecir los resultados médicos. Sin embargo, los resultados nos muestran que el resultado del diagnóstico es multicausal.

Feature Importance Random Forest:

1. Age
2. Date of Admision
3. Discharge Date
4. Gender
5. Admission Type - Urgent
6. Admission Type - Elective

Feature Importance Tree:

1. Date of Admision
2. Discharge Date
3. Age
4. Gender
5. Admission Type - Elective
6. Admission Type - Emergency

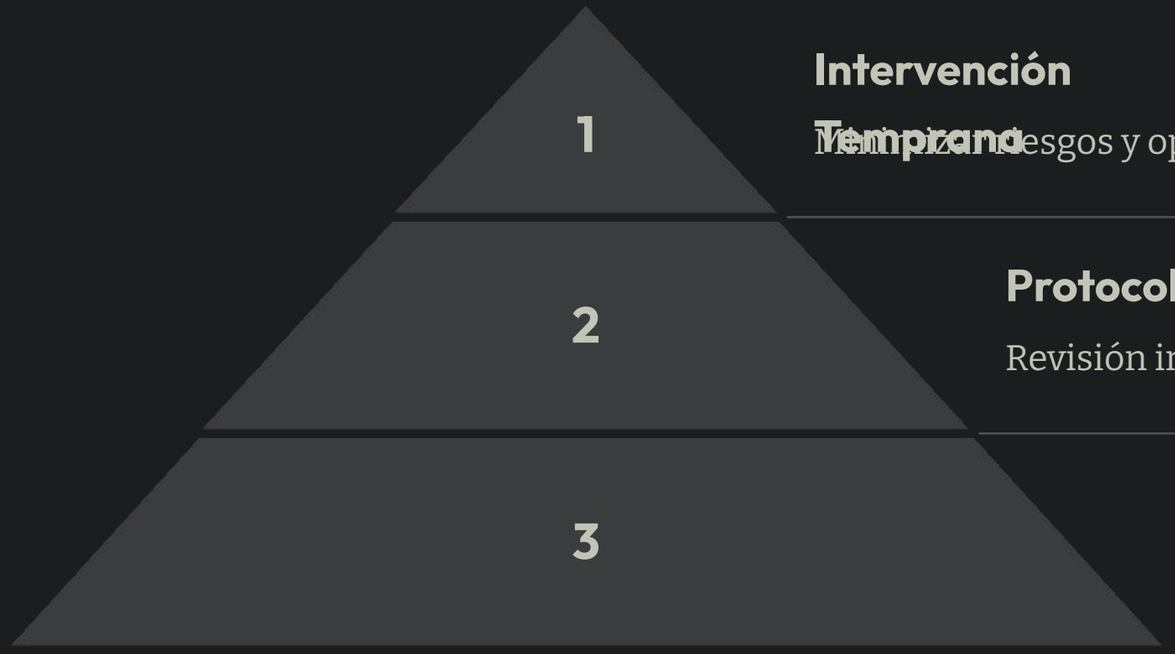
Feature Importance Naive Bayes:

1. Blood Type - AB
2. Medical Condition - Asthma
3. Age
4. Discharge Date
5. Medical Condition - Obesity
6. Medical Condition - Hypertension

Conclusiones

El análisis de "Feature Importance" para Random Forest confirma nuestra hipótesis: el "Tipo de Admisión" es una variable crucial. "Edad", "Género - Femenino", y ciertos "Medicamentos" también son relevantes, mostrando la naturaleza multifactorial de los resultados.

Estamos prediciendo para identificar proactivamente pacientes con alta probabilidad de resultados "Anormales" o "Inconclusivos", resolviendo la detección tardía de condiciones críticas. Si un paciente es predicho como "Anormal" o "Inconclusivo", se activa un protocolo de atención prioritaria, incluyendo revisión inmediata y pruebas adicionales. El Recall es nuestra métrica clave, ya que el costo de un falso negativo en salud es inaceptablemente alto.



Intervención Temprana

Minimizar sesgos y optimizar recursos.

Protocolo Prioritario

Revisión inmediata y pruebas adicionales.

Recall como Métrica Clave

Priorizar la detección de casos críticos.