Análisis Exploratorio e Ingeniería de Características

Centro de Bioinformática y Biología computacional BIOS

EPIDEMIÓLOGOS CONSULTORES S.A.S

BIOS

Índice general

I Introducción

1. Perfilamiento y curado de datos

La elaboración de perfiles de datos incorpora una colección de algoritmos de análisis y evaluación que, cuando se aplican en el contexto adecuado, proporcionan una visión empírica de los problemas potenciales que existen dentro de un conjunto de datos. Se adjunta a este documento, los resultados de los algoritmos sobre cada uno de los conjuntos relevantes de dato entregados.

El objetivo del análisis de anomalías es revisar empíricamente todos los elementos de datos, examinar su distribución de frecuencia y explorar las relaciones entre columnas para revelar posibles valores de datos defectuosos y así establecer su impacto dentro del contexto de producción. En general se busca encontrar las siguientes anomalías:

- Escasez: Identificar las columnas que tienen pocos datos
- Columnas sin usar: indicado ya sea por estar en gran parte despoblado o poblado con el mismo valor en todos los registros
- Análisis de nulidad: que se utiliza para determinar el porcentaje de valores ausentes e identificar representaciones nulas abstractas (por ejemplo, "N / A.º "999-99-9999")
- Atributos sobrecargados, que determinan cuándo se utilizan columnas para almacenar más de un elemento de datos conceptual
- El análisis de frecuencia esperada: la revisión de aquellas columnas cuyos valores se espera que reflejen ciertos patrones de distribución de frecuencia, valida el cumplimiento de los patrones esperados.
- Revisión de valores atípicos: un proceso para aquellas columnas cuyos valores no reflejan la distribución de frecuencia esperada; el objetivo es identificar y explorar aquellos valores cuyas frecuencias son mucho mayores de lo esperado o mucho menores de lo esperado.
- Análisis de rango, utilizado para determinar si los valores de una columna se encuentran dentro de uno (o más) rangos de valores restringidos. Esto puede implicar un análisis de rango único o un agrupamiento de valores más complejo.
- Análisis de formato y / o patrón, que implica inspeccionar patrones alfanuméricos representativos y formatos de valores y revisar la frecuencia.
- Cumplimiento del dominio de valor: en el que las columnas cuyos valores se espera que

cumplan con los dominios de datos conocidos se revisan para identificar los valores que no cumplen.

- Análisis de valor compuesto, que es un proceso que busca conjuntos de valores que parecen estar compuestos por dos o más valores separables. Como ejemplo, considere un código de producto que se compone de un valor único adjunto a un código que representa el sitio de fabricación, como NY-123, donde "NYrepresenta el sitio de fabricación.

1.1 Epilepsia

A continuación, se muestran las características iniciales de los conjuntos suministrados, uno llamado **Control Final** que contiene muestras sin *Epilepsia* que contiene todas las observaciones con epilepsia diagnosticada y *Control* que son observaciones sin epilepsia. De esta manera, se busca describir las características sociodemográficas de los pacientes con epilepsia y controles sanos participantes del estudio. A continuación, se muestran las características iniciales de ambos conjuntos.

Pacientes sanos

CARACTERÍSTICA	VALOR
Numero de variables	20
Numero de observaciones	79
Celdas faltantes	316
Celdas faltantes (%)	20%
Filas duplicadas	0
Tamaño total en memoria	66.7 KiB
Variables Categóricas	13
Variables Numéricas	4
Variables con formato incorrecto	3

Pacientes con Epilepsia

CARACTERÍSTICA	VALOR
Numero de variables	23
Numero de observaciones	56
Celdas faltantes	210
Celdas faltantes (%)	16.3%
Filas duplicadas	0
Tamaño total en memoria	65.9 KiB
Variables Categóricas	18
Variables Numéricas	4
Variables con formato incorrecto	1

Se observa que es considerable la cantidad de datos nulos en ambos conjuntos de datos, mientras que por otro lado, hay columnas que presentan valores contantes para cada uno de los conjuntos individualmente pero muestran la caracterización de los pacientes. A continuación, se muestra el diagrama de barras de los valores nulos por

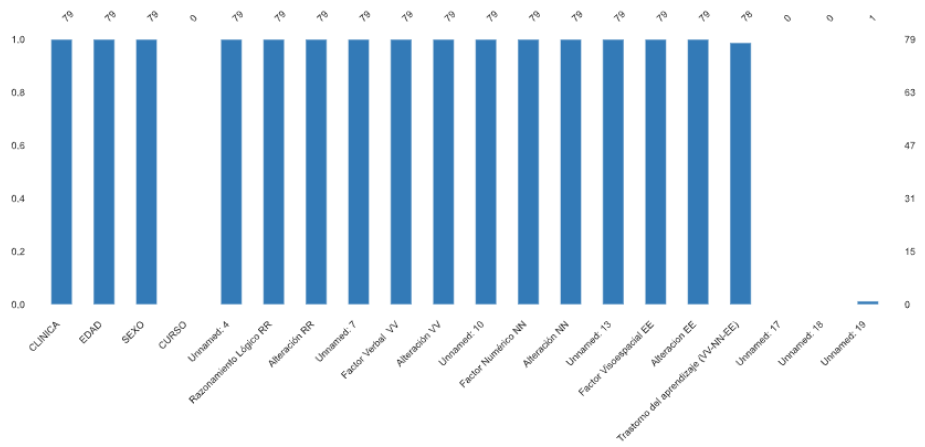

Figura 1.1.1: *Diagrama de barras de valores nulos Conjunto Control*

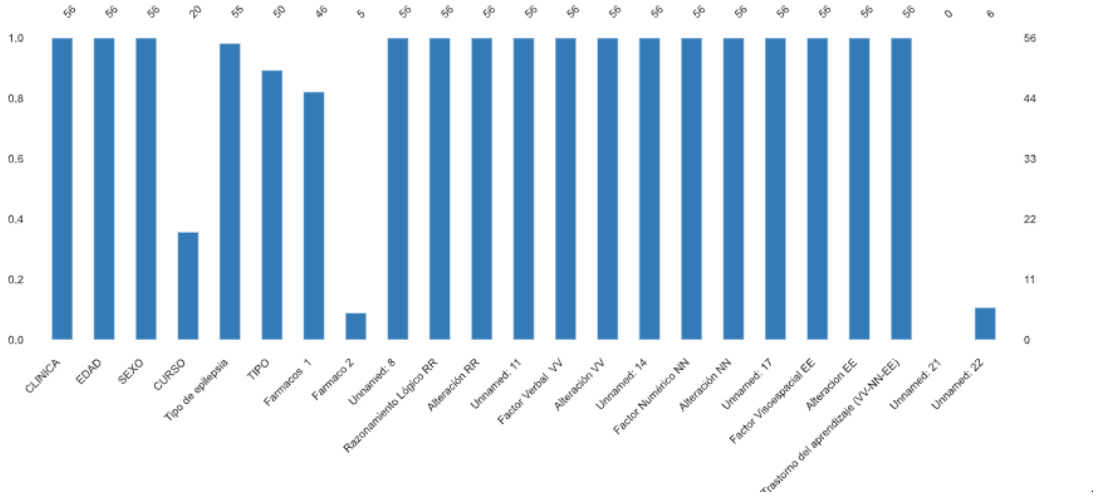

Figura 1.1.2: *Diagrama de barras de valores nulos Conjunto Epilepsia*

Cabe destacar que el valor numérico de cada una de las pruebas no tiene encabezado. Debido que el análisis individual de cada conjunto no tendría sentido, se decide continuar con un nuevo conjunto nacido por la unión de las dos, renombrando las columnas sin nombre y añadiendo clases **"NO APLICA"** si es el caso. De esta manera, se puede caracterizar mejor las variables y el comportamiento en general.

Se muestran las principales características de las variables numéricas del nuevo conjunto creado.

[H]	count	mean	std	min	25%	50%	75%	max
EDAD	135.0	9.400000	1.173081	7.0	8.5	9.0	10.5	11.0
Razonamiento Lógico RR	135.0	61.051852	29.103090	1.0	44.0	63.0	88.0	99.0
Factor Verbal VV	135.0	60.496296	29.724073	1.0	43.0	67.0	82.0	99.0
Factor Numérico NN	135.0	52.251852	35.407167	1.0	23.0	45.0	93.0	99.0
Factor Visoespacial EE	135.0	44.829630	31.440667	1.0	13.0	42.0	71.0	99.0

Se observa una alta desviación estandar entre los datos, lo que implica que por lo general no hay tendencia en el resultado numérico de los exámenes. Se muestra la tabla de valores nulos del nuevo

conjunto creado:

[H] Índice	Columnas	Porcentaje de datos nulos %
0	CLINICA	0.0
1	EDAD	0.0
2	SEXO	0.0
3	Razonamiento Lógico RR	0.0
4	Razonamiento Lógico RR.1	0.0
5	Alteración RR	0.0
6	Factor Verbal VV	0.0
7	Factor Verbal VV.1	0.0
8	Alteración VV	0.0
9	Factor Numérico NN	0.0
10	Factor Numérico NN.1	0.0
11	Alteración NN	0.0
12	Factor Visoespacial EE	0.0
13	Factor Visoespacial EE.1	0.0
14	Alteracion EE	0.0
15	Trastorno del aprendizaje (VV-NN-EE)	0.74
16	Tipo de epilepsia	0.0
17	TIPO	0.0
18	Farmacos 1	0.0
19	Farmaco 2	0.0

Con base en estos resultados, se definen los siguientes mecanismos de limpieza.

- Estandarizar los nombres de las columnas
- Codificar las variables de entrada de tipo categóricas.
- Llenar con media aritmética las columnas de tipo continua cuyos valores falten o tengan formato incorrecto
- Llenar con moda las columnas de tipo nominal cuyos valores falten o tengan formato incorrecto
- Realizar mapeo de columnas incorrectas i.e, la columna sexo tienes valores diferentes haciendo referencia a la misma información (MASCULINO se escribe como Hombre, Masculino, MASCULINO,etc)

Ahora, se muestran las distribuciones de las principales características:

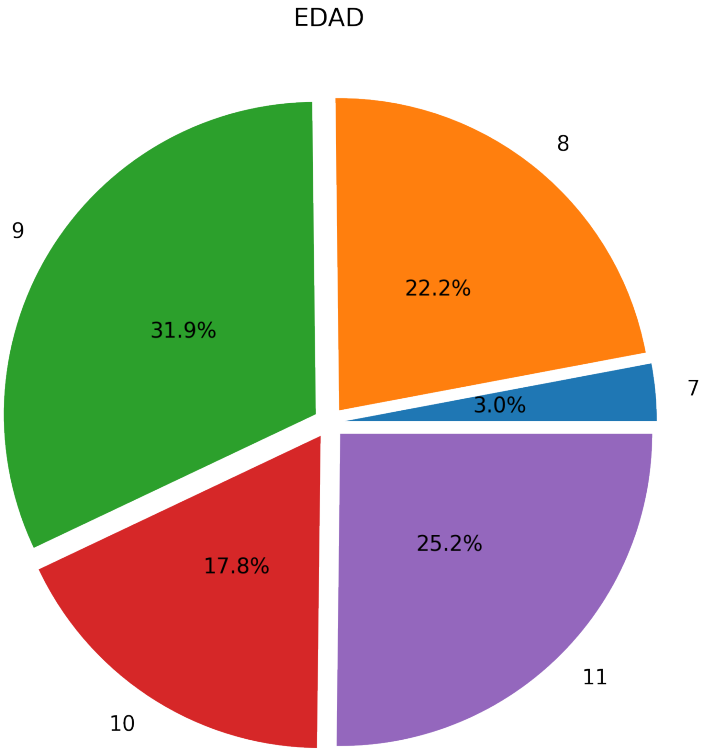

Figura 1.1.3: *Diagrama de pastel edad*

Se observa como el rango de la edad se encuentra entre 7 y 11 años pocas muestras para el conjunto de 7 años.

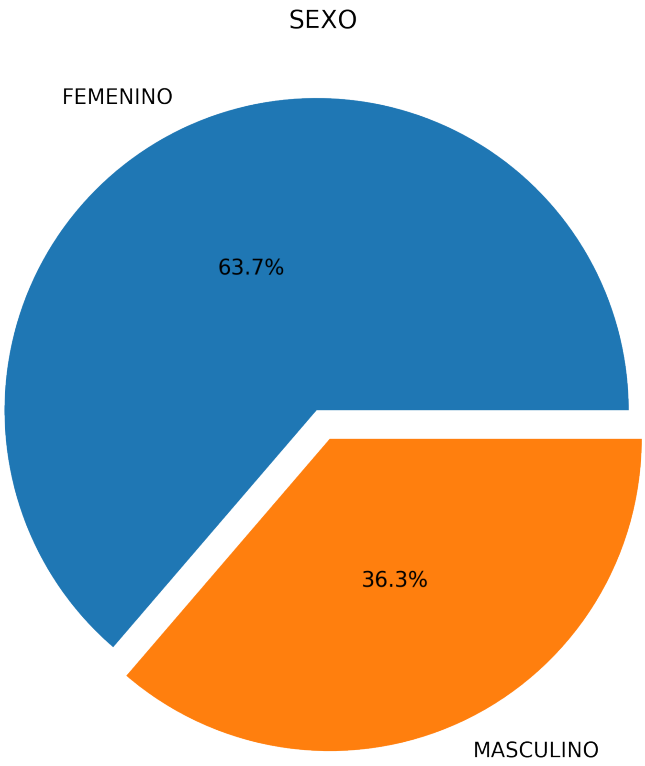

Figura 1.1.4: *Diagrama de pastel sexo*

La mayoría de pacientes son mujeres

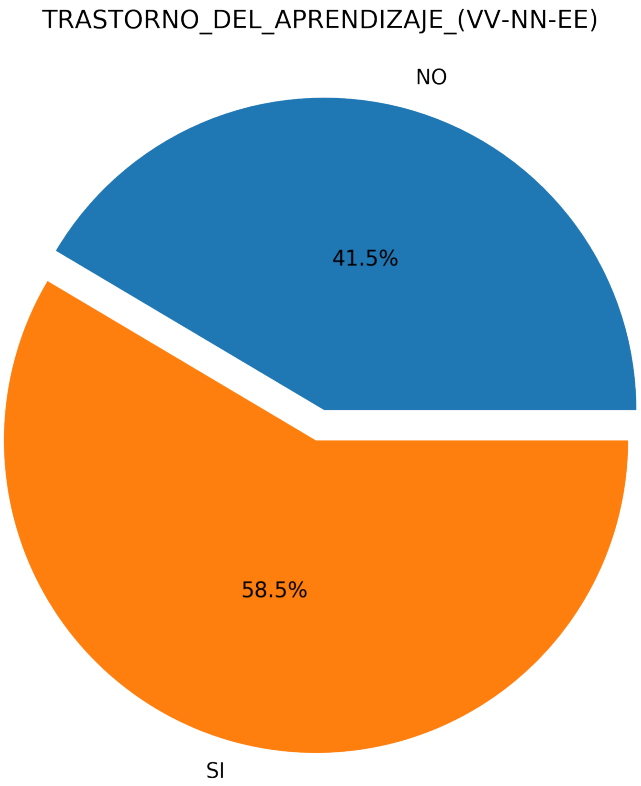

Figura 1.1.5: *Diagrama de pastel tipo de trastorno*

La mayoría de pacientes tiene algún trastorno de aprendizaje independientemente de su condición clínica.

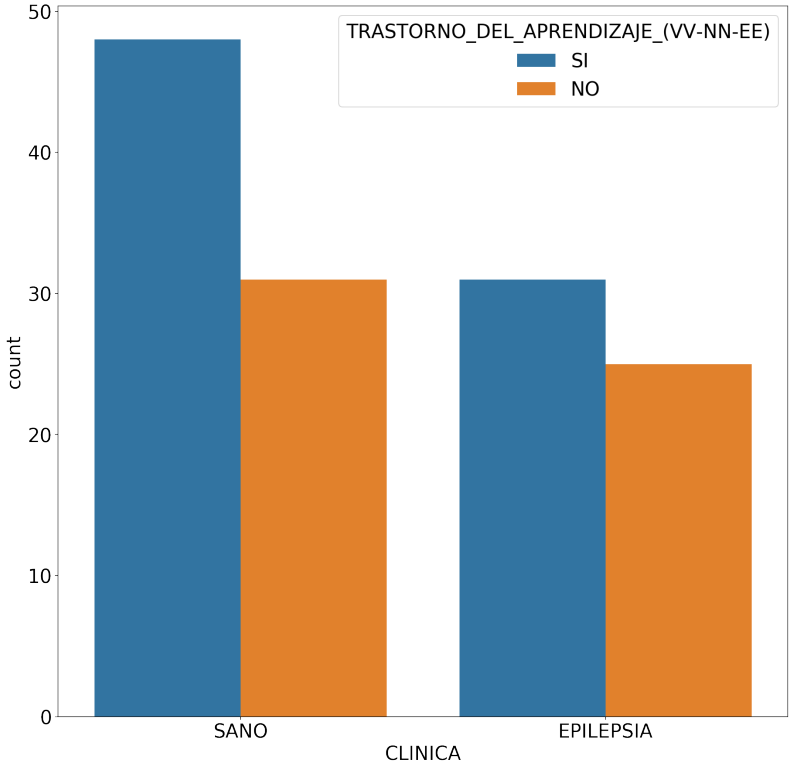

Figura 1.1.6: *Diagrama de barras CLÍNICA por tipo de epilepsia*

Es mucho más frecuente la presencia de trastornos de aprendizaje en pacientes con Epilepsia. Finalmente, se realiza un scatter tridimensional del puntaje numérico de las pruebas VV NN y EE.

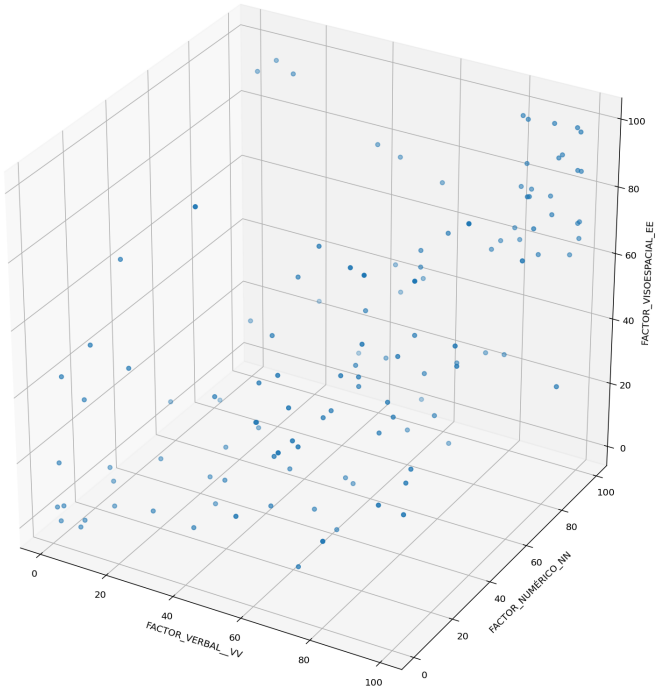

Figura 1.1.7: *Scatter de puntajes*

Como se pudo evidenciar en la matriz de correlación 2.1.8, los puntajes entre las pruebas están altamente correlacionados, tendiendo todos una relación lineal.

II Análisis Exploratorio e Ingeniería de Características

2. Epilepsia

La validación de la selección de variables y el rendimiento predictivo es crucial en la construcción de modelos multivariados robustos que generalizan bien, minimizan el sobreajuste y facilitan la interpretación de los resultados. La selección de variables inapropiada conduce en cambio a un sesgo de selección, lo que aumenta el riesgo de sobreajuste del modelo y descubrimientos de falsos positivos. Aunque existen varios algoritmos para identificar un conjunto mínimo de la mayoría de las variables informativas (es decir, el problema mínimo-óptimo), pocos pueden seleccionar todas las variables relacionadas con la pregunta de investigación (es decir, el problema totalmente relevante), es por esto que para este caso se utiliza el algoritmo MUVR.

2.1 Correlaciones y análisis de relevancia de variables

El objetivo de este paso es comprender cómo las variables del conjunto de datos de entrada varían de la media entre sí, o en otras palabras, para ver si existe alguna relación entre ellas. Porque a veces, las variables están altamente correlacionadas de tal manera que contienen información redundante. Entonces, para identificar estas correlaciones, calculamos la matriz de covarianza.

Se muestra las correlaciones entré las viables; Para esta tarea se utilizó la correlación *Pearson ´s R* para datos continuos-continuos, *Ratio de correlación* para datos categóricos-continuos y finalmente *Cramer's V* para comparaciones categóricas-categóricas:

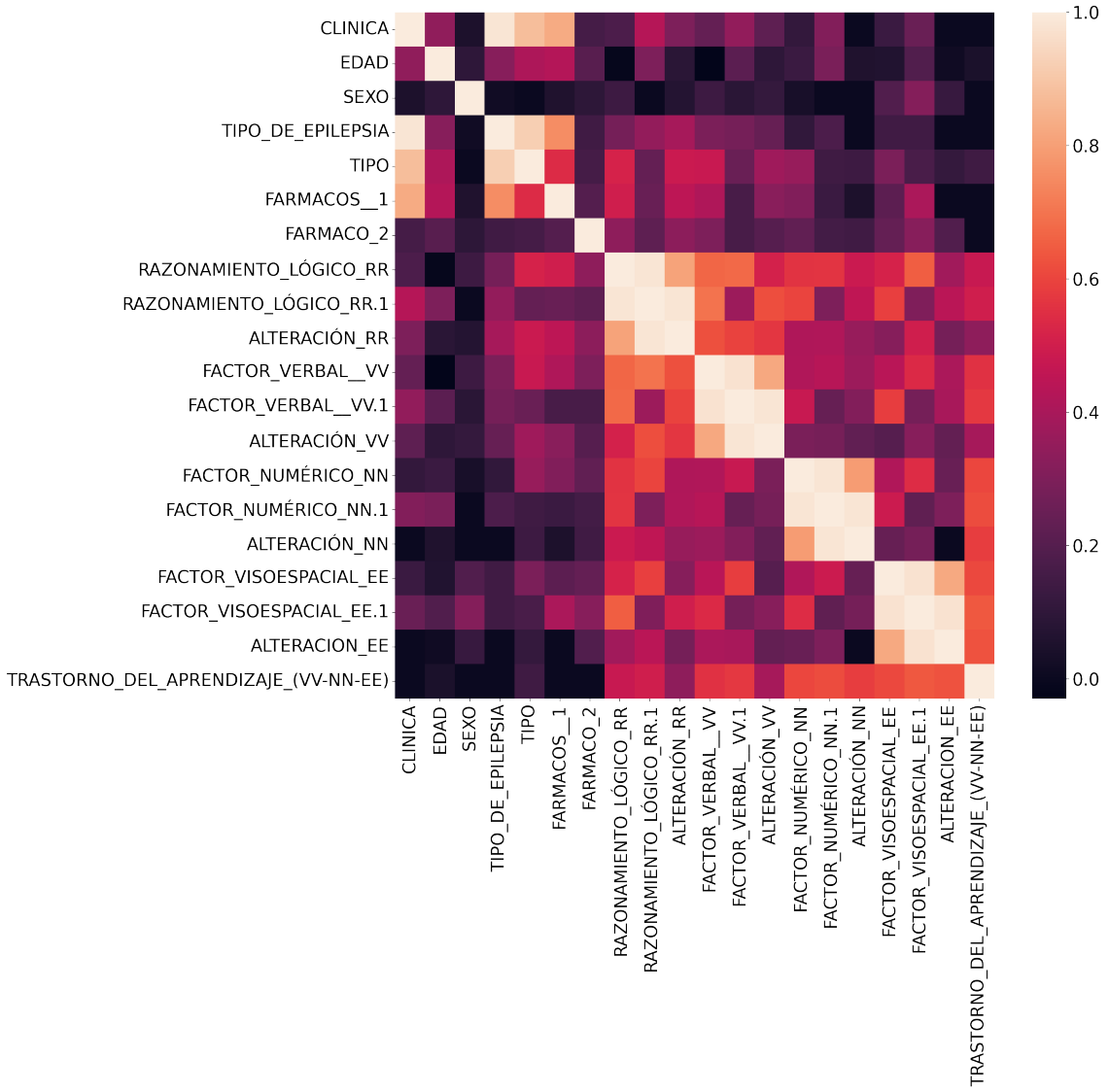

Figura 2.1.1: *Diagrama de correlaciones*

En la figura 2.1.8 se puede observar como no hay relación final entre trastorno de aprendizaje y las variables iniciales, siendo **TIPO** la variable que más influye en las caracterizaciones. Esto se ve acentuado cuando se realiza el análisis de relevancia de variables para *TRASTORNO DEL APRENDIZAJE (VV NN EE)*:

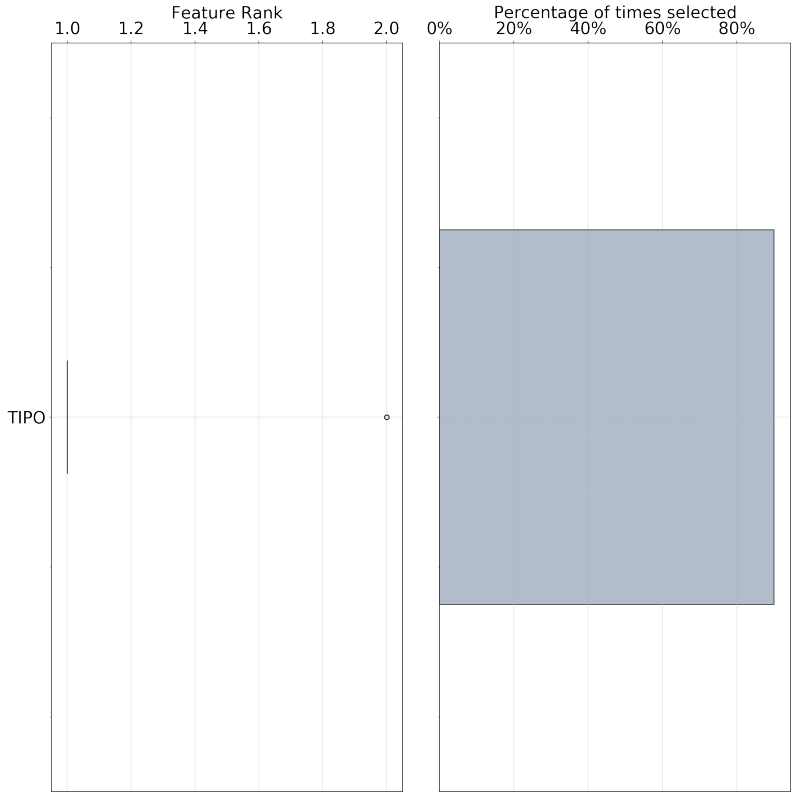

Figura 2.1.2: *Análisis de relevancia de variables TRASTORNO DEL APRENDIZAJE (VV NN EE)*

Podemos ver entonces, como es el comportamiento para la distribución de **TIPO** caracterizado por *TRASTORNO DEL APRENDIZAJE (VV NN EE)*.

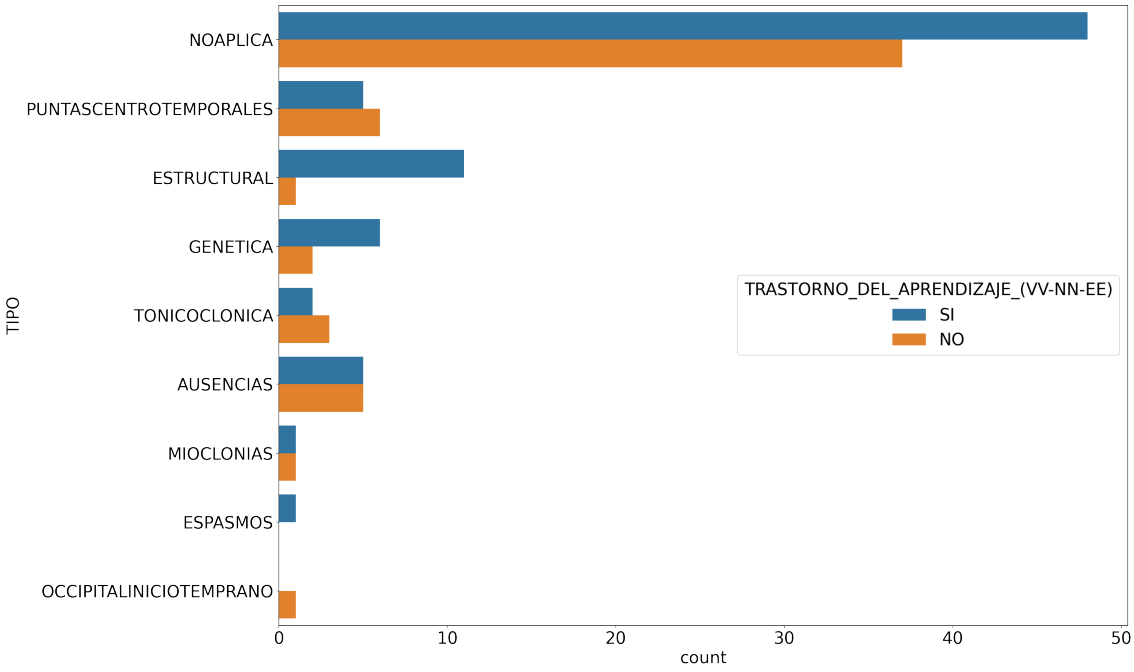

Figura 2.1.3: *comportamiento para la distribución de **TIPO** caracterizado por TRASTORNO DEL APRENDIZAJE (VV NN EE)*

En la figura 2.1.3 se puede observar la distribución y si se encuentra trastorno de aprendizaje según la distribución del tipo de Epilepsia. Claramente, es más probable que el paciente tenga trastorno de aprendizaje si su tipo de epilepsia es **Estructural y Genética**, sin embargo y debido a la poca cantidad de datos no se puede definir este comportamiento como tendencia.

Podemos extrapolar este ejercicio para cada una de las alteraciones

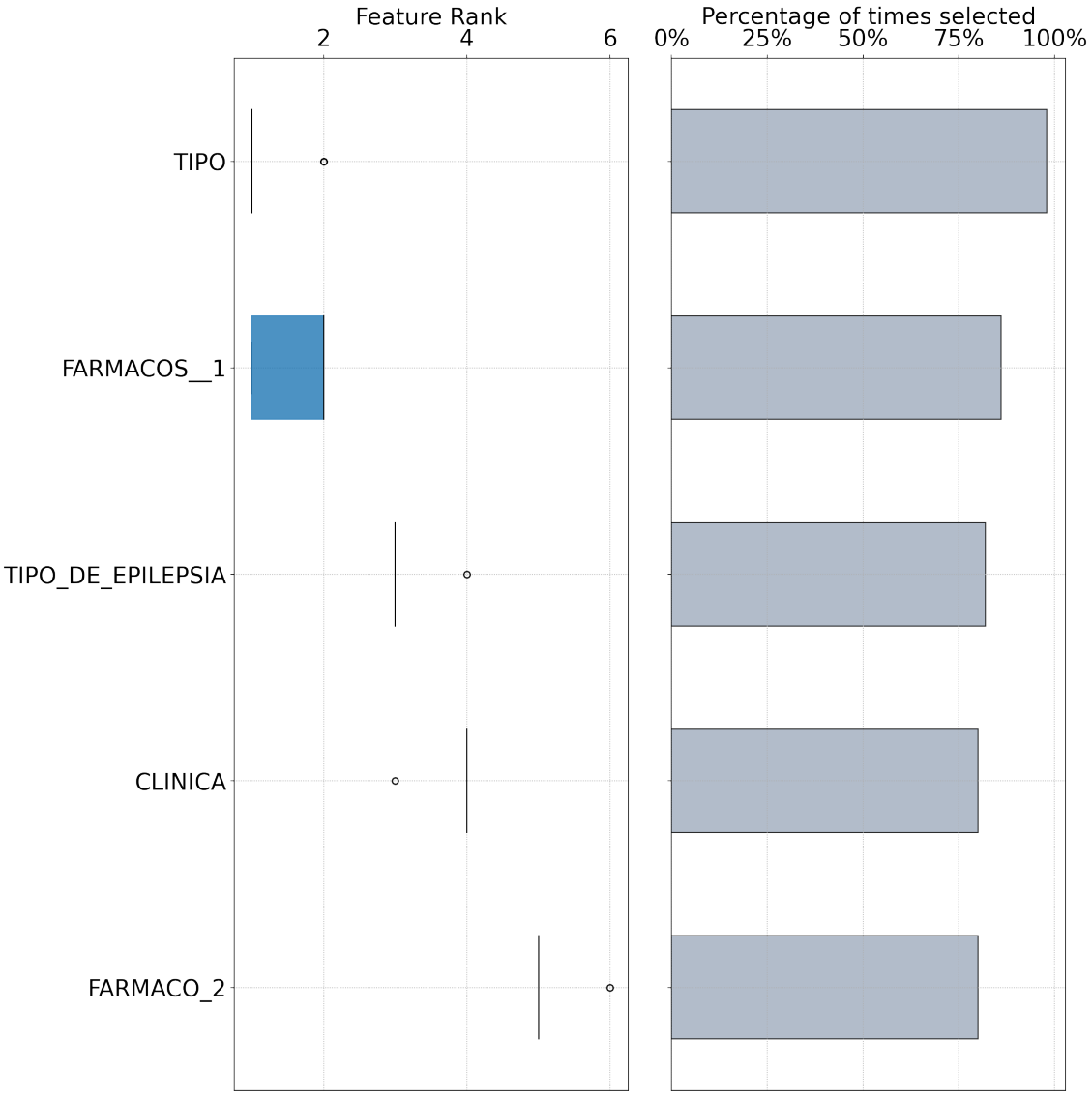

Figura 2.1.4: *Análisis de relevancia de variables ALTERACIÓN RR*

Las principal característica es tipo de epilepsia, por lo que podemos observar este comportamiento:

Alteración RR

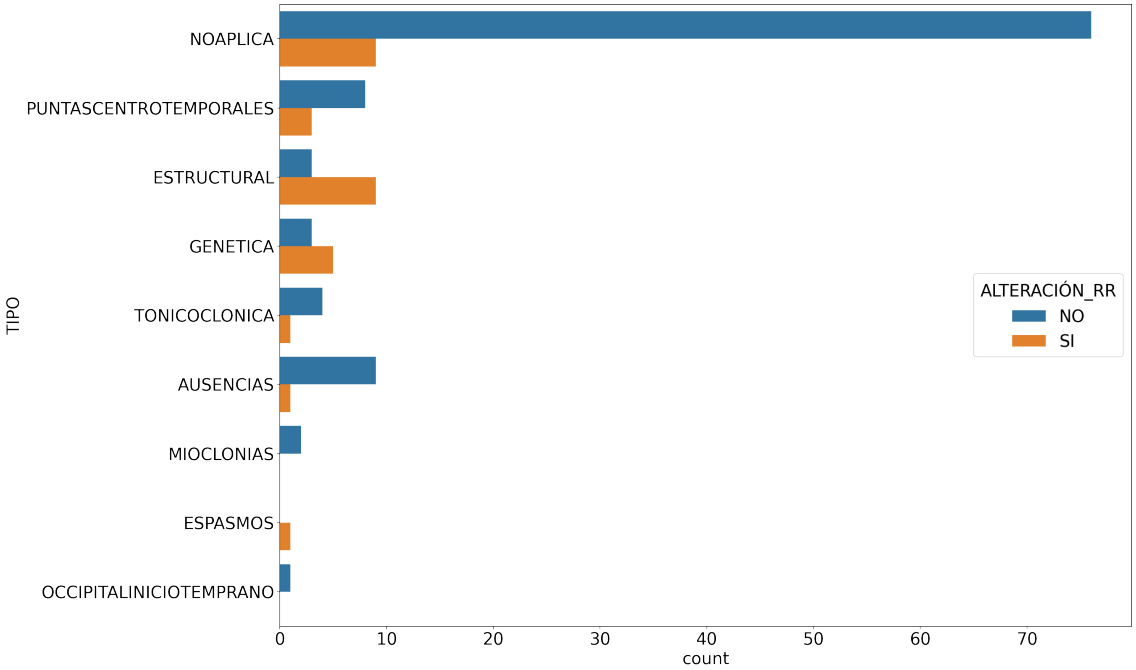

Figura 2.1.5: *comportamiento para la distribución de **TIPO** caracterizado por ALTERACIÓN RR*

En el tipo de epilepsia Genética y estructural es más probable que el paciente tenga alteración de tipo RR.
Para los demás tipos de alteraciones, se encontró que la única variable relevante es **TIPO**

Alteración VV

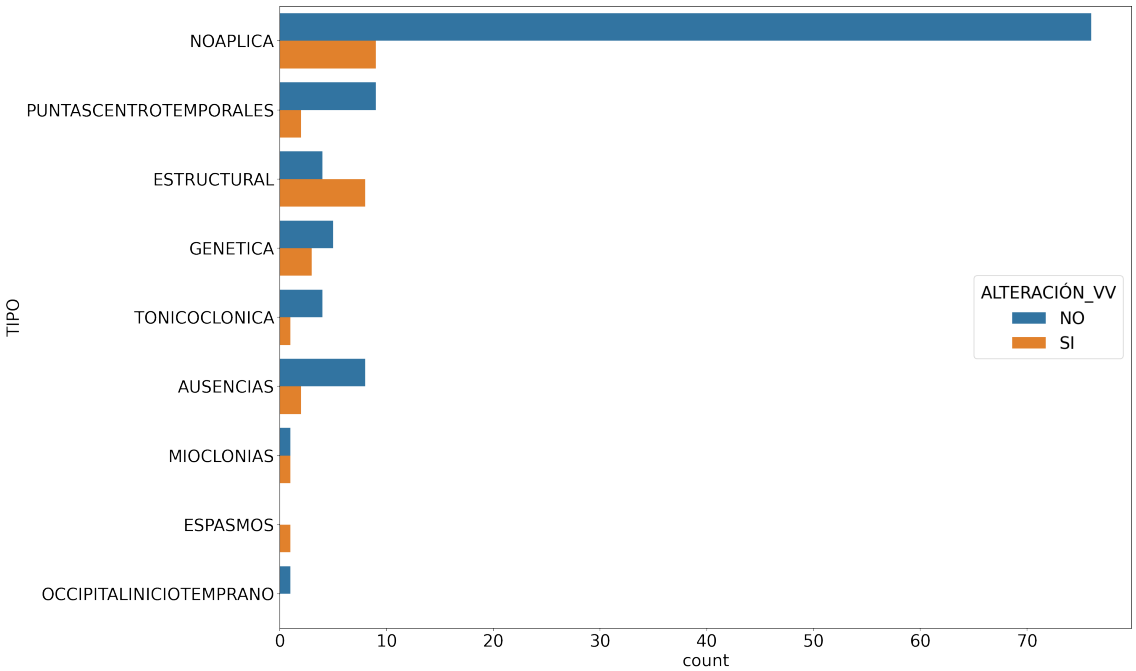

Figura 2.1.6: *comportamiento para la distribución de **TIPO** caracterizado por ALTERACIÓN VV*

Alteración EE

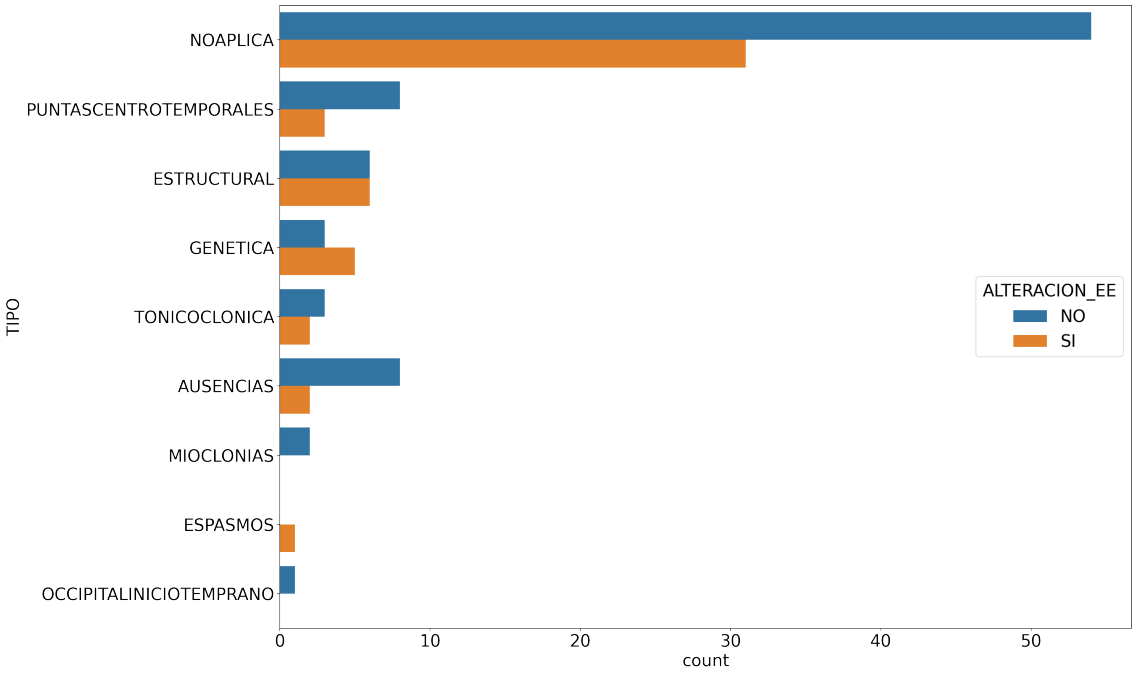

Figura 2.1.7: *comportamiento para la distribución de **TIPO** caracterizado por ALTERACIÓN VV*

Alteración NN

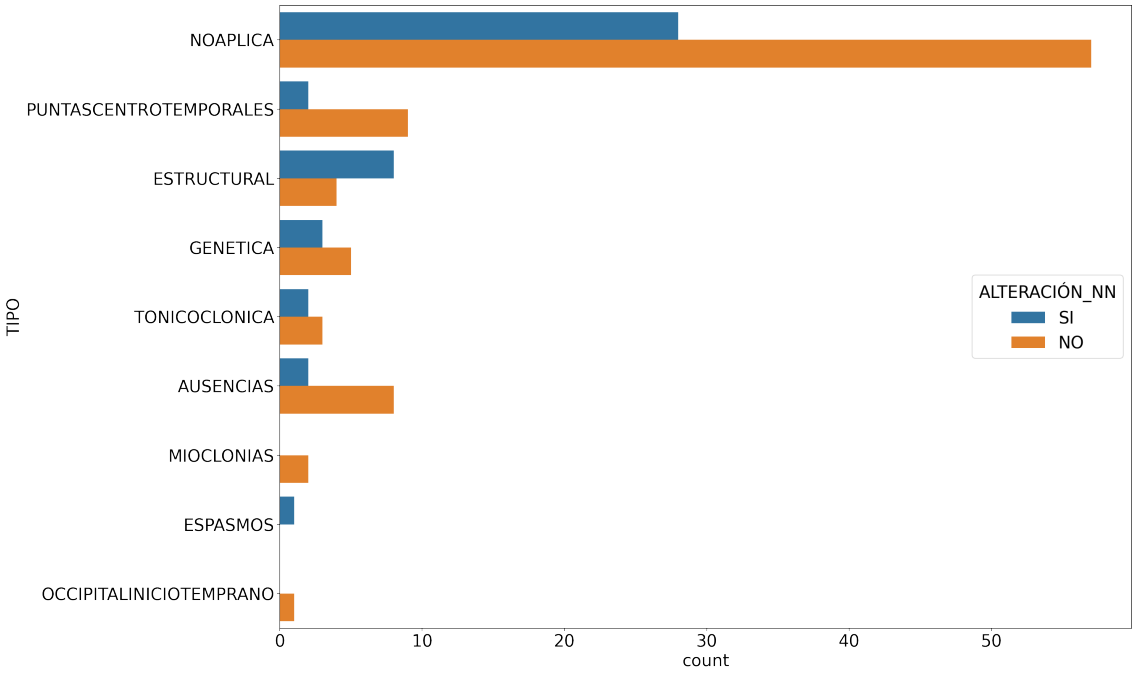

Figura 2.1.8: *comportamiento para la distribución de **TIPO** caracterizado por ALTERACIÓN NN*

2.2 Conclusiones

La cantidad de datos y distribución de las clases de estudio, no es suficiente para construir argumentos sólidos alrededor de tendencias, sin embargo, a continuación se hace lista de los hechos

más relevantes del conjunto, con base en el perfilamiento de datos y el análisis de relevancia de variables se concluye lo siguiente:

- Los valores numéricos de las pruebas de trastorno de aprendizaje están altamente correlacionadas, es decir, si un paciente presenta un valor alto en una de las pruebas, es más probable que en las demás también tenga un puntaje alto.
- No es posible definir si un paciente puede llegar a tener **algún** trastorno VV, NN o EE con base es su caracterización de entrada.
- Individualmente, las alteraciones VV, NN y EE muestran tendencia con el tipo de Epilepsia que el paciente posea.