Anexo 1 Análisis exploratorio estadístico

Seminario de investigacion

Estudiantes:

Jeshua David Junca Rojas

Oscar Ivan Echeverria Marrungo

Javier Callejas Cardozo

Técnicas de Análisis de Datos

La identificación de patrones y la selección de variables relevantes para la predicción de MME se llevarán a cabo mediante diversas técnicas estadísticas, abordando el objetivo 2. Las siguientes etapas se proponen para alcanzar estos objetivos

Preparación de los Datos

En esta sección se realiza la carga del dataset, su exploración inicial y el preprocesamiento necesario para adecuarlo al modelo de machine learning.

Por librerias necesarias se trabajara con Python 3.10

Diccionario

Este diccionario de datos describe cada una de las columnas presentes en el archivo CSV utilizado para el análisis exploratorio y el desarrollo del modelo de clasificación de riesgo de morbilidad materna extrema. La base de datos cuenta con 88.992 registros y las siguientes variables:

- **TIPO DE DOC**: Tipo de registro o caso, que puede indicar la categoría o modalidad de atención.
- **DOCUMENTO**: Número de documento de identificación de la paciente.
- **FUM**: Fecha de la Última Menstruación, punto de partida para estimar la edad gestacional.
- **FPP**: Fecha Probable de Parto, calculada a partir de la FUM.
- **SEMANA_GESTACIONAL**: Número de semanas de gestación al momento del registro.
- **EDAD**: Edad de la paciente en años.
- **MAYOR_35**: Indicador (por ejemplo, 1/0 o Sí/No) que señala si la paciente tiene 35 años o más.
- **NUMEROS_PARTOS_CESARIAS**: Total de partos y cesáreas realizados previamente.
- **ABORTO**: Número de abortos previos.
- **VIVOS**: Número de partos que resultaron en nacidos vivos.
- **MUERTOS**: Número de partos que resultaron en muertes (intrauterinas o neonatales).
- **RIESGO_PREECLAMPSIA**: Indicador de riesgo para desarrollar preeclampsia.
- **NUMEROS_CONTROLES_PRENATALES**: Cantidad de controles o visitas prenatales realizados.
- **IMC**: Índice de Masa Corporal de la paciente.
- **RIESGO**: Valor o clasificación de riesgo general derivado de la combinación de indicadores clínicos.
- **CONSULTA_URGENCIA_ULTIMOS_30_DIAS**: Número de consultas de urgencia realizadas en los últimos 30 días.
- **NACIONALIDAD_PROCEDENCIA**: Nacionalidad o procedencia de la paciente.
- **CODIGO_OCUPACION**: Código que clasifica la ocupación de la paciente.

- **NIVEL_EDUCATIVO**: Nivel educativo alcanzado.
- **AFIC_GRUPO_ETNICO**: Afiliación o pertenencia a un grupo étnico.
- **AFIN_NIVEL_SISBEN**: Nivel del SISBEN asignado, referente a la clasificación socioeconómica.
- **AFIN_GRUPO_POBLACIONAL**: Clasificación según el grupo poblacional al que pertenece la paciente.
- **AFIC_ZONA**: Zona de afiliación o residencia (por ejemplo, urbana o rural).
- **TIPO_DE_CASO**: Clasificación del caso según criterios predefinidos.
- **COD_MUNICIPIO**: Código del municipio de residencia o atención.
- **DIFERENCIA_FIP_FUM**: Diferencia en días entre la Fecha de Inicio del Parto (FIP) y la FUM, que puede ayudar a validar la cronología de eventos.
- **HIPERTENSION**: Indicador de presencia de hipertensión en la paciente.
- **VIH_MATERNO_CONFIRMADO**: Indicador que confirma la infección por VIH en la paciente.
- **TAMIZAJE_SIFILIS**: Resultado o indicación de la realización del tamizaje para sífilis.
- **TAMIZAJE_VIH**: Resultado o indicación de la realización del tamizaje para VIH.
- **TAMIZAJE_HEPATITIS**: Resultado o indicación de la realización del tamizaje para hepatitis.
- **DIAGNOSTICOS**: Diagnósticos clínicos asociados al caso, que pueden incluir complicaciones o condiciones preexistentes.
- **HEMOGLOBINA**: Nivel de hemoglobina medido, utilizado como indicador de anemia.
- **FECHA_HB**: Fecha en la que se realizó la medición de hemoglobina.
- **GLUCOSA_PRE**: Nivel de glucosa en ayunas o preprandial.
- **GLUCOSA_1_HORA**: Nivel de glucosa medido a 1 hora tras la administración de glucosa.
- **GLUCOSA_2_HORA**: Nivel de glucosa medido a 2 horas tras la administración de glucosa.
- **FECHA_GLUCOSA**: Fecha en la que se realizó la prueba de glucosa.
- **HEMORRAGIA**: Indicador de la ocurrencia de hemorragia durante el embarazo o el parto.
- **POLIHIDRAMNIOS**: Indicador de la presencia de polihidramnios (exceso de líquido amniótico).
- **MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA**: Registro de antecedentes de mortalidad perinatal o neonatal tardía.
- **TRIMESTRE**: Trimestre del embarazo en el que se encuentra la paciente.
- **ETIQUETA_MORBILIDAD**: Etiqueta objetivo que indica la presencia o riesgo de morbilidad materna extrema, basada en criterios predefinidos.

Definición del Problema y Objetivos

Contextualización

La **morbilidad materna extrema (MME)** se refiere a complicaciones graves durante el embarazo, parto o posparto que **ponen en riesgo la vida de la madre**. Este concepto abarca desde emergencias obstétricas hasta situaciones clínicas severas que requieren atención médica intensiva. La detección temprana es fundamental para intervenir de manera oportuna, reducir las tasas de mortalidad materna y optimizar la asignación de recursos en salud.

Relevancia:
Identificar a las gestantes en riesgo permite mejorar la atención médica, priorizar intervenciones y asignar recursos de forma eficiente, lo cual tiene un impacto significativo en la reducción de la mortalidad y morbilidad materna.

Variables de Interés

Variable Objetivo

- `ETIQUETA_MORBILIDAD`**:**
  Indica la presencia (**1**) o ausencia (**0**) de morbilidad materna extrema.

Variables Predictoras

Las variables predictoras se agrupan en tres categorías principales:

1. Demográficas y Clínicas

- **Demográficas:**
  - `EDAD`, `MAYOR_35`: Información sobre la edad y si la gestante supera los 35 años.
- **Clínicas:**
  - `SEMANA_GESTACIONAL`: Edad gestacional al momento del análisis.
  - `NUMEROS_PARTOS_CESARIAS`, `ABORTO`, `VIVOS`, `MUERTOS`: Historial obstétrico.
  - `IMC`, `HEMOGLOBINA`: Indicadores del estado nutricional y de salud general.

2. Atención y Seguimiento

- `NUMEROS_CONTROLES_PRENATALES`: Número total de controles prenatales realizados.
- `CONSULTA_URGENCIA_ULTIMOS_30_DIAS`: Registro de consultas de urgencia recientes, indicador de complicaciones.

3. Socioeconómicas y de Contexto

- **Socioeconómicas:**
  - `NACIONALIDAD_PROCEDENCIA`, `CODIGO_OCUPACION`, `NIVEL_EDUCATIVO`
- **Contexto y Vulnerabilidad:**
  - `AFIC_GRUPO_ETNICO`, `AFIN_NIVEL_SISBEN`, `AFIN_GRUPO_POBLACIONAL`, `AFIC_ZONA`

Explicación breve de los datos:

La población de estudio comprenderá a mujeres afiliadas a la EPS MUTUAL SER ESS, residentes en el departamento de Bolívar, que hayan estado embarazadas durante el periodo 2022-2024. Se realizará un muestreo censal, incluyendo todos los registros que cumplan con los criterios de inclusión y exclusión definidos (por ejemplo, datos completos, edad fértil, ausencia de inconsistencias críticas, entre otros).

Preparación de los Datos

En esta sección se realiza la carga del dataset, su exploración inicial y el preprocesamiento necesario para adecuarlo al modelo de machine learning.

Para iniciar, se requiere cargar las librarías necesarias, en caso de necesitar otras puede agregarlas a la celda.

In [211…

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from sklearn.model_selection import train_test_split
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.metrics import Precision, Recall
from tensorflow.keras import Input
from sklearn.metrics import precision_recall_curve, auc, average_precision_score
from sklearn.metrics import precision_score, recall_score, f1_score
import numpy as np
from sklearn.metrics import classification_report
# 📈 Estadística
from scipy.stats import chi2_contingency
# Opcional para que las gráficas se vean mejor
sns.set(style="whitegrid")
%matplotlib inline
```

Se importa el archivo y se revisan los datos contenidos en MME.csv

In [151…
```
df=pd.read_csv('../Datos/MME.csv')
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_35500\1177704023.py:1: DtypeWarning: Columns (1,8,9,10,27) have mixed types. Specify dtype option on import or set low_memory=False.
  df=pd.read_csv('../Datos/MME.csv')
```

Se revisan los 3 primeros y 3 últimos registros del dataset, así también como las colunas, índices, tipos de registros y obtenemos una muestra aleatoria para conocer más los datos a analizar y conocer su comportamiento

In [152…
```
display("primeros 3 registros",df.head(3))
display("3 últimos registros",df.tail(3))
display("Columnas:",df.columns)
display("Índices:",df.index)
display("Tipos de registros en el DataFrame",df.dtypes)
display("Muestra aleatoria de 3 registros",df.sample(3))
```

'primeros 3 registros'

	IDENTIFICACION	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB	GLUC
0	DOC-2910047637871544806	1002392059.0	22.0	19/05/2023	23/02/2024	0.0	0	0.0	0	0	...	Z321	12.7	19/05/2023	
1	DOC-8011058928252310222	1041973230.0	21.0	04/07/2023	09/04/2024	0.0	0	0.0	1	0	...	Z321	14.0	04/07/2023	
2	DOC-3730659741295613587	1050973231.0	28.0	19/01/2023	26/10/2023	0.0	0	1.0	0	0	...	Z321	NaN	NaN	

3 rows × 36 columns

'3 últimos registros'

	IDENTIFICACION	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB
88994	DOC-2557039733480117944	1042350131.0	35.0	26/02/2023	03/12/2023	9.9	0	3.0	0	3	...	Z359 Z321 Z358	11.0	NaN
88995	DOC-3200517063775557572	1002303770.0	25.0	02/12/2021	08/09/2022	9.9	0	2.0	0	2	...	Z359 Z321 Z358	12.0	NaN
88996	DOC-9200819619636388640	1048265360.0	30.0	25/02/2023	02/12/2023	9.9	0	2.0	0	2	...	Z359 Z321 Z358	11.0	NaN

3 rows × 36 columns

```
'Columnas:'
Index(['IDENTIFICACION', 'DOCUMENTO', 'EDAD', 'FUM', 'FPP',
       'SEMANA_GESTACIONAL', 'MAYOR_35', 'NUMEROS_PARTOS_CESARIAS', 'ABORTO',
       'VIVOS', 'MUERTOS', 'RIESGO_PREECLAMPSIA',
       'NUMEROS_CONTROLES_PRENATALES', 'IMC', 'RIESGO',
       'CONSULTA_URGENCIA_ULTIMOS_30_DIAS', 'NACIONALIDAD_PROCEDENCIA',
       'TRABAJA_DURANTE_PARTO', 'NIVEL_EDUCATIVO', 'AFIC_GRUPO_ETNICO',
       'AFIN_NIVEL_SISBEN', 'AFIN_GRUPO_POBLACIONAL', 'AFIC_ZONA',
       'COD_MUNICIPIO', 'HIPERTENSION', 'VIH', 'DIAGNOSTICOS', 'HEMOGLOBINA',
       'FECHA_HB', 'GLUCOSA_PRE', 'FECHA_GLUCOSA', 'HEMORRAGIA',
       'MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA', 'TRIMESTRE',
       'TIPO_DE_CASO', 'ETIQUETA_MORBILIDAD'],
      dtype='object')
'Índices:'
```

```
RangeIndex(start=0, stop=88997, step=1)
'Tipos de registros en el DataFrame'
IDENTIFICACION                                object
DOCUMENTO                                     object
EDAD                                         float64
FUM                                           object
FPP                                           object
SEMANA_GESTACIONAL                           float64
MAYOR_35                                       int64
NUMEROS_PARTOS_CESARIAS                      float64
ABORTO                                        object
VIVOS                                         object
MUERTOS                                       object
RIESGO_PREECLAMPSIA                           object
NUMEROS_CONTROLES_PRENATALES                   int64
IMC                                          float64
RIESGO                                        object
CONSULTA_URGENCIA_ULTIMOS_30_DIAS             object
NACIONALIDAD_PROCEDENCIA                     float64
TRABAJA_DURANTE_PARTO                         object
NIVEL_EDUCATIVO                              float64
AFIC_GRUPO_ETNICO                            float64
AFIN_NIVEL_SISBEN                            float64
AFIN_GRUPO_POBLACIONAL                       float64
AFIC_ZONA                                     object
COD_MUNICIPIO                                float64
HIPERTENSION                                  object
VIH                                           object
DIAGNOSTICOS                                  object
HEMOGLOBINA                                   object
FECHA_HB                                      object
GLUCOSA_PRE                                  float64
FECHA_GLUCOSA                                 object
HEMORRAGIA                                    object
MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA     int64
TRIMESTRE                                     object
TIPO_DE_CASO                                 float64
ETIQUETA_MORBILIDAD                            int64
dtype: object
'Muestra aleatoria de 3 registros'
```

	IDENTIFICACION	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB
58771	DOC-4868724176366271411	1068815555	26.0	02/06/2023	09/03/2024	35.0	0	2.0	0	2	...	Z359 Z359 Z321	11.5	07/09/2023
2811	DOC-105436624031431945	55247532.0	43.0	15/12/2024	21/09/2025	10.0	1	5.0	0	5	...	Z359 Z321 Z358	11.0	20/02/2025
38771	DOC-4405163666560249847	1143138683	33.0	10/10/2023	16/07/2024	23.0	0	2.0	0	2	...	Z359 Z321 Z358	12.0	04/03/2024

3 rows × 36 columns

Revisamos la cantidad de columnas y filas del dataset

In [153…

```
df.shape    # imprime el número de columnas y filas del DataFrame
```

Out[153… (88997, 36)

Vemos información general del dataset

In [154…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 88997 entries, 0 to 88996
Data columns (total 36 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   IDENTIFICACION                            88997 non-null  object
 1   DOCUMENTO                                 88997 non-null  object
 2   EDAD                                      88886 non-null  float64
 3   FUM                                       88997 non-null  object
 4   FPP                                       88997 non-null  object
 5   SEMANA_GESTACIONAL                        88997 non-null  float64
 6   MAYOR_35                                  88997 non-null  int64
 7   NUMEROS_PARTOS_CESARIAS                   88995 non-null  float64
 8   ABORTO                                    88997 non-null  object
 9   VIVOS                                     88997 non-null  object
 10  MUERTOS                                   88997 non-null  object
 11  RIESGO_PREECLAMPSIA                       49873 non-null  object
 12  NUMEROS_CONTROLES_PRENATALES              88997 non-null  int64
 13  IMC                                       88995 non-null  float64
 14  RIESGO                                    88997 non-null  object
 15  CONSULTA_URGENCIA_ULTIMOS_30_DIAS         49873 non-null  object
 16  NACIONALIDAD_PROCEDENCIA                  49873 non-null  float64
 17  TRABAJA_DURANTE_PARTO                     88997 non-null  object
 18  NIVEL_EDUCATIVO                           49873 non-null  float64
 19  AFIC_GRUPO_ETNICO                         88885 non-null  float64
 20  AFIN_NIVEL_SISBEN                         88886 non-null  float64
 21  AFIN_GRUPO_POBLACIONAL                    88886 non-null  float64
 22  AFIC_ZONA                                 88885 non-null  object
 23  COD_MUNICIPIO                             88883 non-null  float64
 24  HIPERTENSION                              88997 non-null  object
 25  VIH                                       88997 non-null  object
 26  DIAGNOSTICOS                              88997 non-null  object
 27  HEMOGLOBINA                               87401 non-null  object
 28  FECHA_HB                                  52861 non-null  object
 29  GLUCOSA_PRE                               50364 non-null  float64
 30  FECHA_GLUCOSA                             12847 non-null  object
 31  HEMORRAGIA                                88997 non-null  object
 32  MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA  88997 non-null  int64
 33  TRIMESTRE                                 88970 non-null  object
 34  TIPO_DE_CASO                              49873 non-null  float64
 35  ETIQUETA_MORBILIDAD                       88997 non-null  int64
dtypes: float64(12), int64(4), object(20)
memory usage: 24.4+ MB
```

Eliminar filas donde la edad es nula, tambien menor a 11 años y se elimina al columna IDENTIFICACION ya que no se usara pra el modelo y es dato sensible

In [155…
```
df = df[df['EDAD'].notna() & (df['EDAD'] >= 11)].drop(columns=['IDENTIFICACION'])
```

In [156…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 88884 entries, 0 to 88996
Data columns (total 35 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   DOCUMENTO                                 88884 non-null  object
 1   EDAD                                      88884 non-null  float64
 2   FUM                                       88884 non-null  object
 3   FPP                                       88884 non-null  object
 4   SEMANA_GESTACIONAL                        88884 non-null  float64
 5   MAYOR_35                                  88884 non-null  int64
 6   NUMEROS_PARTOS_CESARIAS                   88882 non-null  float64
 7   ABORTO                                    88884 non-null  object
 8   VIVOS                                     88884 non-null  object
 9   MUERTOS                                   88884 non-null  object
 10  RIESGO_PREECLAMPSIA                       49819 non-null  object
 11  NUMEROS_CONTROLES_PRENATALES              88884 non-null  int64
 12  IMC                                       88882 non-null  float64
 13  RIESGO                                    88884 non-null  object
 14  CONSULTA_URGENCIA_ULTIMOS_30_DIAS         49819 non-null  object
 15  NACIONALIDAD_PROCEDENCIA                  49819 non-null  float64
 16  TRABAJA_DURANTE_PARTO                     88884 non-null  object
 17  NIVEL_EDUCATIVO                           49819 non-null  float64
 18  AFIC_GRUPO_ETNICO                         88883 non-null  float64
 19  AFIN_NIVEL_SISBEN                         88884 non-null  float64
 20  AFIN_GRUPO_POBLACIONAL                    88884 non-null  float64
 21  AFIC_ZONA                                 88883 non-null  object
 22  COD_MUNICIPIO                             88779 non-null  float64
 23  HIPERTENSION                              88884 non-null  object
 24  VIH                                       88884 non-null  object
 25  DIAGNOSTICOS                              88884 non-null  object
 26  HEMOGLOBINA                               87295 non-null  object
 27  FECHA_HB                                  52799 non-null  object
 28  GLUCOSA_PRE                               50312 non-null  float64
 29  FECHA_GLUCOSA                             12834 non-null  object
 30  HEMORRAGIA                                88884 non-null  object
 31  MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA  88884 non-null  int64
 32  TRIMESTRE                                 88857 non-null  object
 33  TIPO_DE_CASO                              49819 non-null  float64
 34  ETIQUETA_MORBILIDAD                       88884 non-null  int64
dtypes: float64(12), int64(4), object(19)
memory usage: 24.4+ MB
```

Se verifica si existen menores de 12 años

In [158…
```
menores_12 = df[df['EDAD'] < 12]
print(f"Número de registros con edad menor a 12 años: {len(menores_12)}")
```

```
Número de registros con edad menor a 12 años: 0
```

Se consultan top cinco de mayor edad, pra verificar datos anomalos

In [159…
```
df['EDAD'].dropna().sort_values(ascending=False).unique()[:5]
```

Out[159…
```
array([55., 54., 53., 52., 51.])
```

Datos duplicados

se analizó si el dataset contiene valores duplicados

In [160…
```
#Análisis de filas duplicadas en la DB, se calculó las filas antes de eliminar las repetidas
print(f'Tamaño del set antes de eliminar las filas repetidas: {df.shape}')

#Eliminamos filas duplicadas de la DB
df = df.drop_duplicates()
#Impresión del tamaño del set luego de eliminar las duplicadas

print(f'Tamaño del set después de eliminar las filas repetidas: {df.shape}')
#Se evidencia que no se encuentran valores duplicados
df
```

```
Tamaño del set antes de eliminar las filas repetidas: (88884, 35)
Tamaño del set después de eliminar las filas repetidas: (88883, 35)
```

Out[160…

	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	MUERTOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB	GLUCOSA
0	1002392059.0	22.0	19/05/2023	23/02/2024	0.0	0	0.0	0	0	0	...	Z321	12.7	19/05/2023	
1	1041973230.0	21.0	04/07/2023	09/04/2024	0.0	0	0.0	1	0	0	...	Z321	14.0	04/07/2023	
2	1050973231.0	28.0	19/01/2023	26/10/2023	0.0	0	1.0	0	0	0	...	Z321	NaN	NaN	
3	1044919777.0	35.0	04/07/2023	09/04/2024	0.0	0	0.0	0	0	0	...	Z321	13.0	04/07/2023	
4	1050945968.0	21.0	09/05/2023	13/02/2024	0.0	0	0.0	0	0	0	...	Z321	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
88992	1043198728.0	28.0	04/09/2022	11/06/2023	9.9	0	0.0	0	0	0	...	Z359 Z321 Z358	12.0	NaN	
88993	1045170170.0	31.0	13/09/2022	20/06/2023	9.9	0	2.0	0	2	0	...	Z359 Z321 Z358	12.0	NaN	
88994	1042350131.0	35.0	26/02/2023	03/12/2023	9.9	0	3.0	0	3	0	...	Z359 Z321 Z358	11.0	NaN	
88995	1002303770.0	25.0	02/12/2021	08/09/2022	9.9	0	2.0	0	2	0	...	Z359 Z321 Z358	12.0	NaN	
88996	1048265360.0	30.0	25/02/2023	02/12/2023	9.9	0	2.0	0	2	0	...	Z359 Z321 Z358	11.0	NaN	

88883 rows × 35 columns

Análisis de valores nulos

Se calcula la cantidad de valores nulos por columna

In [161…
```
# Identificar valores nulos
display(df.isnull().sum())
```

```
DOCUMENTO                                    0
EDAD                                         0
FUM                                          0
FPP                                          0
SEMANA_GESTACIONAL                           0
MAYOR_35                                     0
NUMEROS_PARTOS_CESARIAS                      2
ABORTO                                       0
VIVOS                                        0
MUERTOS                                      0
RIESGO_PREECLAMPSIA                      39064
NUMEROS_CONTROLES_PRENATALES                 0
IMC                                          2
RIESGO                                       0
CONSULTA_URGENCIA_ULTIMOS_30_DIAS        39064
NACIONALIDAD_PROCEDENCIA                 39064
TRABAJA_DURANTE_PARTO                        0
NIVEL_EDUCATIVO                          39064
AFIC_GRUPO_ETNICO                            1
AFIN_NIVEL_SISBEN                            0
AFIN_GRUPO_POBLACIONAL                       0
AFIC_ZONA                                    1
COD_MUNICIPIO                              105
HIPERTENSION                                 0
VIH                                          0
DIAGNOSTICOS                                 0
HEMOGLOBINA                               1589
FECHA_HB                                 36085
GLUCOSA_PRE                              38572
FECHA_GLUCOSA                            76049
HEMORRAGIA                                   0
MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA  0
TRIMESTRE                                   27
TIPO_DE_CASO                             39064
ETIQUETA_MORBILIDAD                          0
dtype: int64
```

Se calcula la cantidad de valores nulos por columna

In [162…

```python
# cantidad de valores nulos (datos faltantes) y su respectivo porcentaje en la base de datos

# cálculo de nulos por columna
nulos_por_columna = df.isnull().sum()
columnas_con_nulos = nulos_por_columna[nulos_por_columna > 0]

# cálculo de porcentaje de los valores nulos por columna
porcentaje_nulos = (columnas_con_nulos / len(df)) * 100
porcentaje_nulos = porcentaje_nulos.round(2)
columnas_con_nulos = pd.DataFrame({
    'Cantidad de Nulos': columnas_con_nulos,
    'Porcentaje de Nulos (%)': porcentaje_nulos
})
columnas_con_nulos
```

Out[162...

	Cantidad de Nulos	Porcentaje de Nulos (%)
NUMEROS_PARTOS_CESARIAS	2	0.00
RIESGO_PREECLAMPSIA	39064	43.95
IMC	2	0.00
CONSULTA_URGENCIA_ULTIMOS_30_DIAS	39064	43.95
NACIONALIDAD_PROCEDENCIA	39064	43.95
NIVEL_EDUCATIVO	39064	43.95
AFIC_GRUPO_ETNICO	1	0.00
AFIC_ZONA	1	0.00
COD_MUNICIPIO	105	0.12
HEMOGLOBINA	1589	1.79
FECHA_HB	36085	40.60
GLUCOSA_PRE	38572	43.40
FECHA_GLUCOSA	76049	85.56
TRIMESTRE	27	0.03
TIPO_DE_CASO	39064	43.95

✅ Variables sin problemas de nulidad (< 1%) Estas variables tienen un porcentaje despreciable de datos faltantes y pueden ser utilizadas sin imputación o con una imputación directa:

NUMEROS_PARTOS_CESARIAS (0.00%)

IMC (0.00%)

AFIC_GRUPO_ETNICO (0.00%)

AFIC_ZONA (0.00%)

COD_MUNICIPIO (0.12%)

Estas variables están limpias y listas para modelar.

Las variables con porcentaje mayor a 1 son normales, ya que es normal que no se tenga el dato debido a que son exámenes de laboratorio rutinarios.

Cálculo de valores faltantes

In [163...

```
# cálculo del porcentaje de valores faltantes en todo el conjunto de datos

porcentaje_variables_val_faltantes = df.isnull().sum().sum() / df.size * 100
print("\nPorcentaje total de valores faltantes en el dataset: {:.2f}%".format(porcentaje_variables_val_faltantes))
```

```
Porcentaje total de valores faltantes en el dataset: 11.18%
```

Manejo de datos perdidos o esperados

In [164…
```
df.isnull().any(axis=0) # Indicador de valores nulos en una columna
```

Out[164…
```
DOCUMENTO                                  False
EDAD                                       False
FUM                                        False
FPP                                        False
SEMANA_GESTACIONAL                         False
MAYOR_35                                   False
NUMEROS_PARTOS_CESARIAS                     True
ABORTO                                     False
VIVOS                                      False
MUERTOS                                    False
RIESGO_PREECLAMPSIA                         True
NUMEROS_CONTROLES_PRENATALES               False
IMC                                         True
RIESGO                                     False
CONSULTA_URGENCIA_ULTIMOS_30_DIAS           True
NACIONALIDAD_PROCEDENCIA                    True
TRABAJA_DURANTE_PARTO                      False
NIVEL_EDUCATIVO                             True
AFIC_GRUPO_ETNICO                           True
AFIN_NIVEL_SISBEN                          False
AFIN_GRUPO_POBLACIONAL                     False
AFIC_ZONA                                   True
COD_MUNICIPIO                               True
HIPERTENSION                               False
VIH                                        False
DIAGNOSTICOS                               False
HEMOGLOBINA                                 True
FECHA_HB                                    True
GLUCOSA_PRE                                 True
FECHA_GLUCOSA                               True
HEMORRAGIA                                 False
MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA    False
TRIMESTRE                                   True
TIPO_DE_CASO                                True
ETIQUETA_MORBILIDAD                        False
dtype: bool
```

In [165…
```
perd=df.isnull().any(axis=1) # Para saber si hay NaN en los datos de cada variable:
```

In [166…
```
perd
```

Out[166…
```
0        True
1        True
2        True
3        True
4        True
         ...
88992    True
88993    True
88994    True
88995    True
88996    True
Length: 88883, dtype: bool
```

In [167…
```
# Filtrar y mostrar solo las filas donde 'perd' es True
df_con_nulos = df[perd]
print(df_con_nulos)
```

```
          DOCUMENTO  EDAD         FUM         FPP  SEMANA_GESTACIONAL  \
0      1002392059.0  22.0  19/05/2023  23/02/2024                 0.0
1      1041973230.0  21.0  04/07/2023  09/04/2024                 0.0
2      1050973231.0  28.0  19/01/2023  26/10/2023                 0.0
3      1044919777.0  35.0  04/07/2023  09/04/2024                 0.0
4      1050945968.0  21.0  09/05/2023  13/02/2024                 0.0
...             ...   ...         ...         ...                 ...
88992  1043198728.0  28.0  04/09/2022  11/06/2023                 9.9
88993  1045170170.0  31.0  13/09/2022  20/06/2023                 9.9
88994  1042350131.0  35.0  26/02/2023  03/12/2023                 9.9
88995  1002303770.0  25.0  02/12/2021  08/09/2022                 9.9
88996  1048265360.0  30.0  25/02/2023  02/12/2023                 9.9

       MAYOR_35  NUMEROS_PARTOS_CESARIAS ABORTO VIVOS MUERTOS  ...  \
0             0                      0.0      0     0       0  ...
1             0                      0.0      1     0       0  ...
2             0                      1.0      0     0       0  ...
3             0                      0.0      0     0       0  ...
4             0                      0.0      0     0       0  ...
...         ...                      ...    ...   ...     ...  ...
88992         0                      0.0      0     0       0  ...
88993         0                      2.0      0     2       0  ...
88994         0                      3.0      0     3       0  ...
88995         0                      2.0      0     2       0  ...
88996         0                      2.0      0     2       0  ...

             DIAGNOSTICOS  HEMOGLOBINA    FECHA_HB GLUCOSA_PRE FECHA_GLUCOSA  \
0              Z321              12.7  19/05/2023         NaN           NaN
1              Z321              14.0  04/07/2023         NaN           NaN
2              Z321               NaN         NaN         NaN           NaN
3              Z321              13.0  04/07/2023         NaN           NaN
4              Z321               NaN         NaN         NaN           NaN
...                 ...          ...         ...         ...           ...
88992  Z359 Z321 Z358             12.0         NaN        75.0           NaN
88993  Z359 Z321 Z358             12.0         NaN        75.0           NaN
88994  Z359 Z321 Z358             11.0         NaN        55.0           NaN
88995  Z359 Z321 Z358             12.0         NaN        66.0           NaN
88996  Z359 Z321 Z358             11.0         NaN        55.0           NaN

       HEMORRAGIA MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA  \
0              NO                                          0
1              NO                                          0
2              NO                                          0
3              NO                                          0
4              NO                                          0
...           ...                                        ...
88992          NO                                          0
88993          NO                                          0
88994          NO                                          0
88995          NO                                          0
88996          NO                                          0

              TRIMESTRE  TIPO_DE_CASO  ETIQUETA_MORBILIDAD
0                   NaN           NaN                    0
1                   NaN           NaN                    0
2                   NaN           NaN                    0
3                   NaN           NaN                    1
4                   NaN           NaN                    0
...                 ...           ...                  ...
88992  PRIMER_TRIMESTRE          21.0                    0
```

```
88993  PRIMER_TRIMESTRE          21.0                   0
88994  PRIMER_TRIMESTRE          21.0                   0
88995  PRIMER_TRIMESTRE          21.0                   0
88996  PRIMER_TRIMESTRE          21.0                   0

[84262 rows x 35 columns]
```

Devolverá todas las filas que tienen al menos un valor nulo

In [168…
```
df[perd] # devolverá todas las filas que tienen al menos un valor nulo
```

Out[168…

	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	MUERTOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB	GLUCOSA
0	1002392059.0	22.0	19/05/2023	23/02/2024	0.0	0	0.0	0	0	0	...	Z321	12.7	19/05/2023	
1	1041973230.0	21.0	04/07/2023	09/04/2024	0.0	0	0.0	1	0	0	...	Z321	14.0	04/07/2023	
2	1050973231.0	28.0	19/01/2023	26/10/2023	0.0	0	1.0	0	0	0	...	Z321	NaN	NaN	
3	1044919777.0	35.0	04/07/2023	09/04/2024	0.0	0	0.0	0	0	0	...	Z321	13.0	04/07/2023	
4	1050945968.0	21.0	09/05/2023	13/02/2024	0.0	0	0.0	0	0	0	...	Z321	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
88992	1043198728.0	28.0	04/09/2022	11/06/2023	9.9	0	0.0	0	0	0	...	Z359 Z321 Z358	12.0	NaN	
88993	1045170170.0	31.0	13/09/2022	20/06/2023	9.9	0	2.0	0	2	0	...	Z359 Z321 Z358	12.0	NaN	
88994	1042350131.0	35.0	26/02/2023	03/12/2023	9.9	0	3.0	0	3	0	...	Z359 Z321 Z358	11.0	NaN	
88995	1002303770.0	25.0	02/12/2021	08/09/2022	9.9	0	2.0	0	2	0	...	Z359 Z321 Z358	12.0	NaN	
88996	1048265360.0	30.0	25/02/2023	02/12/2023	9.9	0	2.0	0	2	0	...	Z359 Z321 Z358	11.0	NaN	

84262 rows × 35 columns

Data set sin duplicados

In [169…
```
df.head()
```

Out[169…

	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	MUERTOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB	GLUCOSA_PRE
0	1002392059.0	22.0	19/05/2023	23/02/2024	0.0	0	0.0	0	0	0	...	Z321	12.7	19/05/2023	NaN
1	1041973230.0	21.0	04/07/2023	09/04/2024	0.0	0	0.0	1	0	0	...	Z321	14.0	04/07/2023	NaN
2	1050973231.0	28.0	19/01/2023	26/10/2023	0.0	0	1.0	0	0	0	...	Z321	NaN	NaN	NaN
3	1044919777.0	35.0	04/07/2023	09/04/2024	0.0	0	0.0	0	0	0	...	Z321	13.0	04/07/2023	NaN
4	1050945968.0	21.0	09/05/2023	13/02/2024	0.0	0	0.0	0	0	0	...	Z321	NaN	NaN	NaN

5 rows × 35 columns

Frecuencia de cada valor en una columna ETIQUETA_MORBILIDAD

In [170…
```
# frecuencia de cada valor en una columna Class
df['ETIQUETA_MORBILIDAD'].value_counts(dropna=False)
```

Out[170…
```
ETIQUETA_MORBILIDAD
0    84486
1     4397
Name: count, dtype: int64
```

Cálculo de estadísticas

In [171…
```
df.describe(include='all')
```

Out[171…

	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	MUERTOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB
count	88883	88883.000000	88883	88883	88883.000000	88883.000000	88881.000000	88883.0	88883.0	88883.0	...	88883	87294.0	52798
unique	88879	NaN	1492	1506	NaN	NaN	NaN	24.0	30.0	381.0	...	5498	249.0	1364
top	5305637	NaN	28/12/2021	04/10/2022	NaN	NaN	NaN	0.0	0.0	0.0	...	Z359	12.0	11/02/2022
freq	2	NaN	422	226	NaN	NaN	NaN	40355.0	34719.0	45136.0	...	13123	19359.0	141
mean	NaN	27.536706	NaN	NaN	26.732906	0.080510	1.039964	NaN	NaN	NaN	...	NaN	NaN	NaN
std	NaN	6.370880	NaN	NaN	564.968799	0.272083	1.282235	NaN	NaN	NaN	...	NaN	NaN	NaN
min	NaN	12.000000	NaN	NaN	0.000000	0.000000	0.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
25%	NaN	23.000000	NaN	NaN	10.000000	0.000000	0.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
50%	NaN	27.000000	NaN	NaN	17.000000	0.000000	1.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
75%	NaN	32.000000	NaN	NaN	29.000000	0.000000	2.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
max	NaN	55.000000	NaN	NaN	45104.000000	1.000000	22.000000	NaN	NaN	NaN	...	NaN	NaN	NaN

11 rows × 35 columns

Imputar datos en semana gestacional con el promedio

In [172…
```
# Calcular la media solo de las semanas gestacionales válidas (≤ 42)
media_valida = df[df['SEMANA_GESTACIONAL'] <= 42]['SEMANA_GESTACIONAL'].mean()

# Imputar la media donde los valores sean mayores a 42
df.loc[df['SEMANA_GESTACIONAL'] > 42, 'SEMANA_GESTACIONAL'] = media_valida
```

Verificacion de la imputacion

In [173…
```
df.describe(include='all')
```

Out[173...

	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	MUERTOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB
count	88883	88883.000000	88883	88883	88883.000000	88883.000000	88881.000000	88883.0	88883.0	88883.0	...	88883	87294.0	52798
unique	88879	NaN	1492	1506	NaN	NaN	NaN	24.0	30.0	381.0	...	5498	249.0	1364
top	5305637	NaN	28/12/2021	04/10/2022	NaN	NaN	NaN	0.0	0.0	0.0	...	Z359	12.0	11/02/2022
freq	2	NaN	422	226	NaN	NaN	NaN	40355.0	34719.0	45136.0	...	13123	19359.0	141
mean	NaN	27.536706	NaN	NaN	19.620330	0.080510	1.039964	NaN	NaN	NaN	...	NaN	NaN	NaN
std	NaN	6.370880	NaN	NaN	10.436621	0.272083	1.282235	NaN	NaN	NaN	...	NaN	NaN	NaN
min	NaN	12.000000	NaN	NaN	0.000000	0.000000	0.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
25%	NaN	23.000000	NaN	NaN	10.000000	0.000000	0.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
50%	NaN	27.000000	NaN	NaN	17.000000	0.000000	1.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
75%	NaN	32.000000	NaN	NaN	29.000000	0.000000	2.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
max	NaN	55.000000	NaN	NaN	42.000000	1.000000	22.000000	NaN	NaN	NaN	...	NaN	NaN	NaN

11 rows × 35 columns

Eliminar partos / Cesarias mayores a 10

In [174...
```
df = df[df['NUMEROS_PARTOS_CESARIAS'] <= 10]
```

Se procede con la verificacion

In [175...
```
df.describe(include='all')
```

Out[175...

	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	MUERTOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECHA_HB
count	88841	88841.000000	88841	88841	88841.000000	88841.000000	88841.000000	88841.0	88841.0	88841.0	...	88841	87252.0	52771
unique	88837	NaN	1492	1506	NaN	NaN	NaN	23.0	25.0	379.0	...	5495	249.0	1364
top	1052524839	NaN	28/12/2021	04/10/2022	NaN	NaN	NaN	0.0	0.0	0.0	...	Z359	12.0	11/02/2022
freq	2	NaN	422	226	NaN	NaN	NaN	40338.0	34719.0	45118.0	...	13120	19347.0	141
mean	NaN	27.533729	NaN	NaN	19.619583	0.080346	1.033656	NaN	NaN	NaN	...	NaN	NaN	NaN
std	NaN	6.369065	NaN	NaN	10.436656	0.271829	1.245249	NaN	NaN	NaN	...	NaN	NaN	NaN
min	NaN	12.000000	NaN	NaN	0.000000	0.000000	0.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
25%	NaN	23.000000	NaN	NaN	10.000000	0.000000	0.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
50%	NaN	27.000000	NaN	NaN	17.000000	0.000000	1.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
75%	NaN	32.000000	NaN	NaN	29.000000	0.000000	2.000000	NaN	NaN	NaN	...	NaN	NaN	NaN
max	NaN	55.000000	NaN	NaN	42.000000	1.000000	10.000000	NaN	NaN	NaN	...	NaN	NaN	NaN

11 rows × 35 columns

Eliminar filas con datos anomalos en la columna ABORTO

In [176…
```
valores_invalidos = ['o', '2 / 1 MOLAR', '01/01/1900']
df = df[~df['ABORTO'].isin(valores_invalidos)]
```

Eliminar filas cuando vivos sea mayor a 22

In [178…
```
# Asegurar que la columna VIVOS sea numérica
df['VIVOS'] = pd.to_numeric(df['VIVOS'], errors='coerce')

# Eliminar filas donde VIVOS > 21
df = df[df['VIVOS'] <= 21]
```

Calcular la media de IMC considerando solo valores válidos (entre 10 y 70)

Reemplazar valores anómalos con la media

In [179…
```
# Calcular la media de IMC considerando solo valores válidos (entre 10 y 70)
media_imc = df[(df['IMC'] > 12) & (df['IMC'] < 70)]['IMC'].mean()

# Reemplazar valores anómalos con la media
df.loc[(df['IMC'] <= 12) | (df['IMC'] >= 70), 'IMC'] = media_imc
```

In [181…
```
print("Resumen estadístico de la columna IMC para verificar la imputacion:")
print(f"Promedio (mean): {df['IMC'].mean():.2f}")
print(f"Mediana (median): {df['IMC'].median():.2f}")
print(f"Mínimo (min): {df['IMC'].min():.2f}")
print(f"Máximo (max): {df['IMC'].max():.2f}")
```

```
Resumen estadístico de la columna IMC para verificar la imputacion:
Promedio (mean): 25.33
Mediana (median): 24.75
Mínimo (min): 12.02
Máximo (max): 69.85
```

In [184…
```
df.describe(include='all')
```

Out[184...

	DOCUMENTO	EDAD	FUM	FPP	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	ABORTO	VIVOS	MUERTOS	...	DIAGNOSTICOS	HEMOGLOBINA	FECH
count	88836	88836.000000	88836	88836	88836.000000	88836.000000	88836.000000	88836.0	88836.000000	88836.0	...	88836	87247.0	5
unique	88832	NaN	1492	1506	NaN	NaN	NaN	21.0	NaN	379.0	...	5495	249.0	
top	1052524839	NaN	28/12/2021	04/10/2022	NaN	NaN	NaN	0.0	NaN	0.0	...	Z359	12.0	11/02/
freq	2	NaN	422	226	NaN	NaN	NaN	40338.0	NaN	45118.0	...	13119	19346.0	
mean	NaN	27.533669	NaN	NaN	19.619966	0.080350	1.033680	NaN	0.885182	NaN	...	NaN	NaN	
std	NaN	6.369135	NaN	NaN	10.436767	0.271836	1.245275	NaN	1.132157	NaN	...	NaN	NaN	
min	NaN	12.000000	NaN	NaN	0.000000	0.000000	0.000000	NaN	0.000000	NaN	...	NaN	NaN	
25%	NaN	23.000000	NaN	NaN	10.000000	0.000000	0.000000	NaN	0.000000	NaN	...	NaN	NaN	
50%	NaN	27.000000	NaN	NaN	17.000000	0.000000	1.000000	NaN	1.000000	NaN	...	NaN	NaN	
75%	NaN	32.000000	NaN	NaN	29.000000	0.000000	2.000000	NaN	1.000000	NaN	...	NaN	NaN	
max	NaN	55.000000	NaN	NaN	42.000000	1.000000	10.000000	NaN	10.000000	NaN	...	NaN	NaN	

11 rows × 35 columns

Elimina datos anomalos de la columna HEMOGLOBINA

In [191...
```python
#ELIMNA DATOS ANOMALOS DE LA COLOMNA HEMOGLOBINA
valores_invalidos = ['11..1', '8.3/ 9.2', '11.4/10.9', '9.2/10.5/9.9', 'NEG', '29/09/2022']
df = df[~df['HEMOGLOBINA'].isin(valores_invalidos)]
```

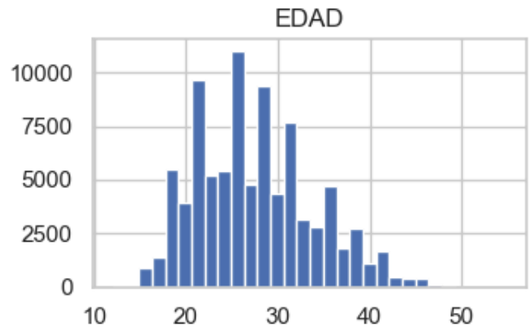
In [ ]:
```python
# Lista de valores inválidos
valores_invalidos = ['11..1', '8.3/ 9.2', '11.4/10.9', '9.2/10.5/9.9', 'NEG', '29/09/2022']

# Verificar si alguno sigue presente
valores_restantes = df[df['HEMOGLOBINA'].isin(valores_invalidos)]

# Mostrar resultados
if valores_restantes.empty:
    print("✅ La columna HEMOGLOBINA ya no contiene ninguno de los valores inválidos.")
else:
    print(" Aún existen valores inválidos en HEMOGLOBINA:")
    display(valores_restantes)
```

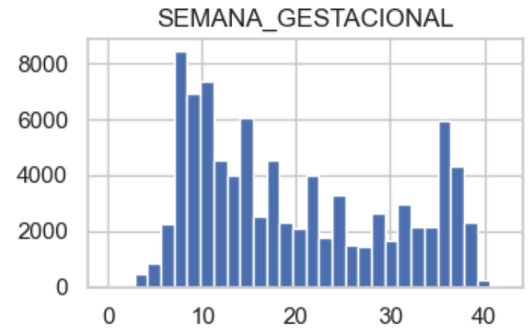
```
✅ La columna HEMOGLOBINA ya no contiene ninguno de los valores inválidos.
```

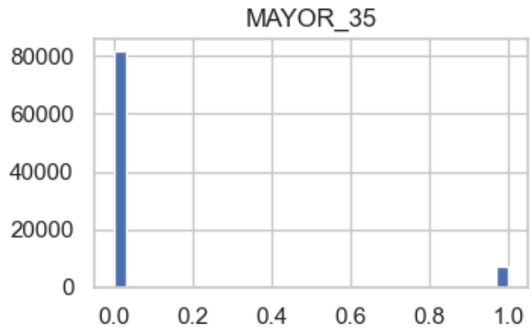
Histogramas de variables numéricas

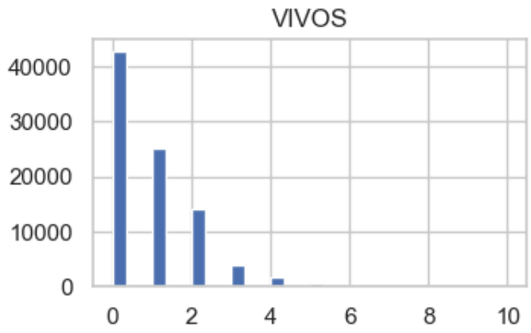
In [185...
```python
df.select_dtypes(include=['float64', 'int64']).hist(figsize=(15, 12), bins=30)
plt.suptitle('Histogramas de variables numéricas', fontsize=16)
plt.tight_layout()
plt.show()
```

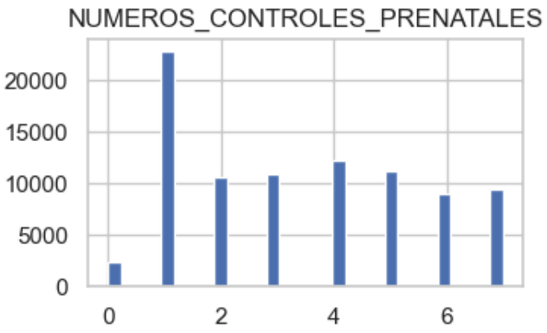
Histogramas de variables numéricas

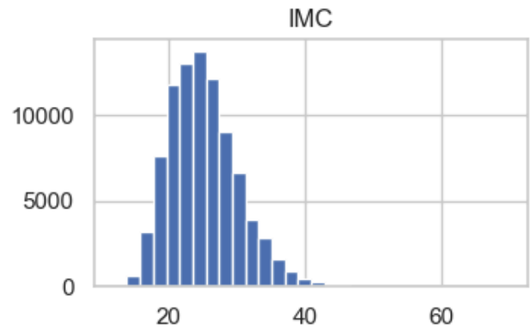


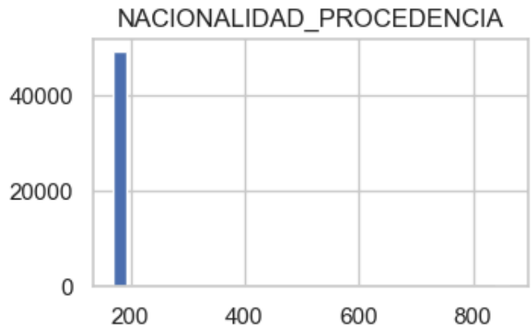


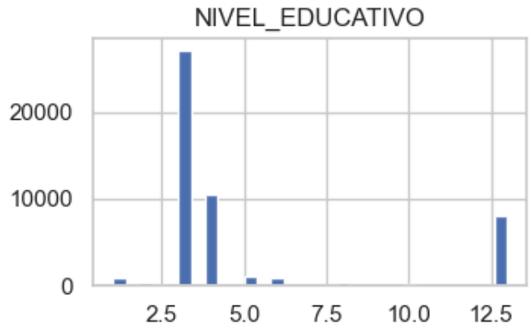


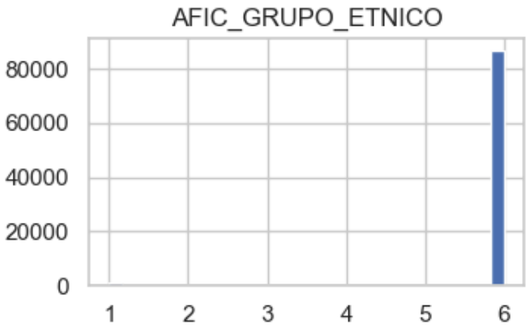


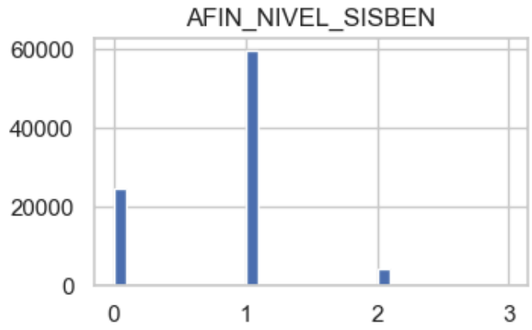


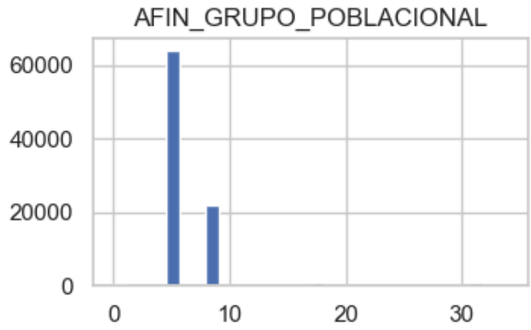


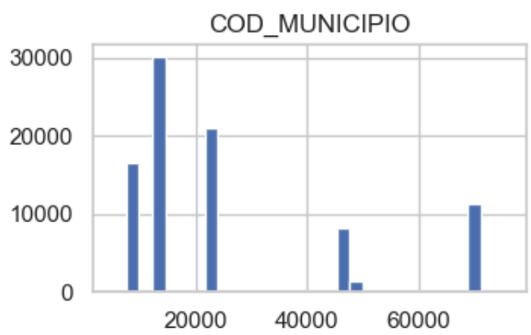


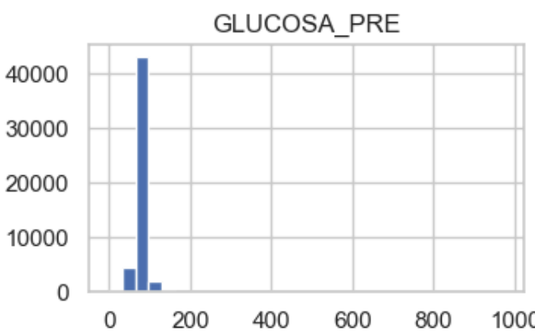


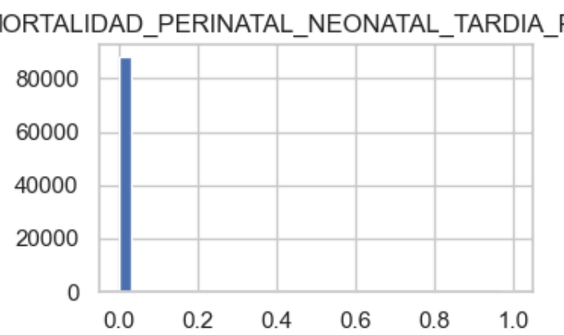


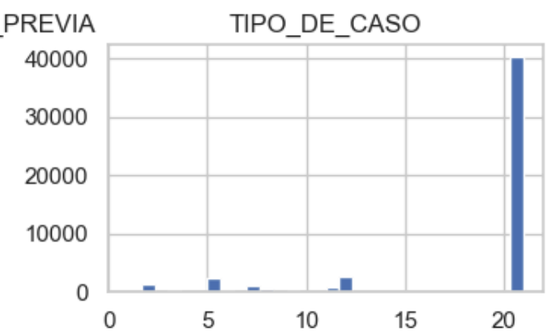


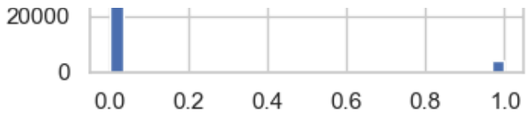












Análisis de los histogramas de variables numéricas.

Con el objetivo de comprender mejor la distribución y comportamiento de las variables numéricas en el dataset, se desarrolló una serie de histogramas que permitieron identificar patrones clave, valores atípicos, y posibles necesidades de transformación para el posterior modelado.

🔷 EDAD La edad de las gestantes se concentra principalmente entre los 20 y 35 años, lo cual es esperable clínicamente. Existe un sesgo leve a la izquierda, y aunque hay algunos casos mayores de 40, son considerablemente menos frecuentes.

🔷 SEMANA_GESTACIONAL Se observa una mayor frecuencia entre las semanas 10 y 30, con una caída gradual hacia las semanas finales. Algunos valores extremos cercanos a 0 o mayores a 42 semanas probablemente deben ser revisados por errores o necesidades de imputación.

🔷 MAYOR_35 Al ser una variable categórica binaria, como se esperaba, la mayoría de las gestantes tienen menos de 35 años. Este desbalance puede afectar el análisis si no se trata adecuadamente.

🔷 NUMEROS_PARTOS_CESARIAS y VIVOS Ambas variables presentan una distribución altamente sesgada a la derecha. La mayoría de las mujeres tienen entre 0 y 2 partos/cesáreas o hijos vivos, lo cual es consistente con la realidad demográfica.

🔷 NUMEROS_CONTROLES_PRENATALES La distribución es casi uniforme, lo que podría indicar una buena cobertura de atención prenatal o simplemente un registro sistemático que agrupa los controles en intervalos iguales.

🔷 IMC El índice de masa corporal tiene una forma aproximadamente normal, centrada en los valores considerados saludables. Esto es positivo porque facilita su uso en modelos predictivos sin necesidad de transformaciones.

🔷 Variables socioeconómicas y de afiliación (NACIONALIDAD_PROCEDENCIA, NIVEL_EDUCATIVO, AFIN_*) Estas variables muestran distribuciones categóricas muy sesgadas o agrupadas en pocos valores dominantes. Por ejemplo, AFIN_NIVEL_SISBEN y AFIC_GRUPO_ETNICO tienen clases claramente predominantes, lo cual puede ser útil pero también podría inducir sesgos si no se maneja bien en modelos supervisados.

🔷 GLUCOSA_PRE Esta variable presenta valores atípicos extremos (incluso cercanos a 1000), lo cual es clínicamente improbable. Definitivamente requiere limpieza o truncamiento para no distorsionar el modelo.

🔷 MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA Casi todos los registros tienen valor 0, con muy pocos casos positivos. Esto indica que es una variable altamente desbalanceada, lo cual puede ser relevante desde lo clínico, pero difícil de modelar si no se aplican técnicas específicas.

🔷 TIPO_DE_CASO Presenta múltiples valores, pero algunos son muy poco frecuentes. Puede ser conveniente agrupar categorías raras para simplificar el análisis.

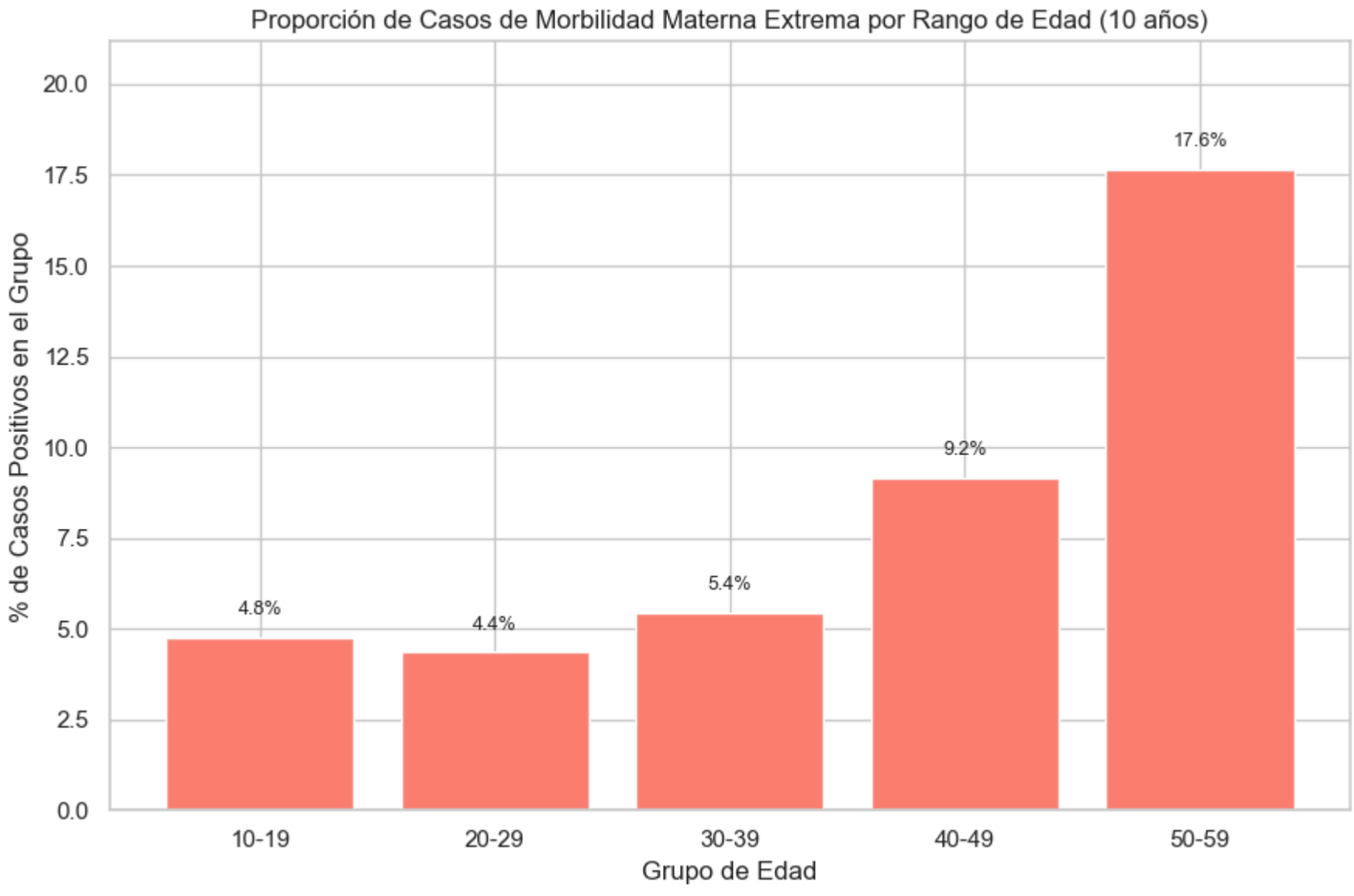
🔷 ETIQUETA_MORBILIDAD Como es la variable objetivo, confirmo que tiene un fuerte desbalance: la gran mayoría de los casos no presenta morbilidad materna extrema. Esto refuerza la necesidad de aplicar técnicas como oversampling, undersampling o el uso de métricas específicas (recall, F1-score, AUC) durante la fase de modelado.

In [207…

```
#Crear rangos de edad por décadas
df['grupo_edad_10'] = pd.cut(df['EDAD'],bins=range(10, 70, 10),right=False,labels=['10-19', '20-29', '30-39', '40-49', '50-59'])

display("Índices:",df['grupo_edad_10'])
# Calcular proporción de casos positivos por grupo de edad
proporcion_positivos = df[df['ETIQUETA_MORBILIDAD'] == 1]['grupo_edad_10'].value_counts(normalize=False).sort_index()
total_por_grupo = df['grupo_edad_10'].value_counts().sort_index()
porcentaje_positivos = (proporcion_positivos / total_por_grupo) * 100
# Plot
fig, ax = plt.subplots(figsize=(9, 6))
bars = ax.bar(porcentaje_positivos.index, porcentaje_positivos.values, color='salmon')
# Título y etiquetas
plt.title('Proporción de Casos de Morbilidad Materna Extrema por Rango de Edad (10 años)')
plt.xlabel('Grupo de Edad')
```

```
plt.ylabel('% de Casos Positivos en el Grupo')
plt.ylim(0, porcentaje_positivos.max() * 1.2)
# Agregar % sobre las barras
for bar in bars:
    yval = bar.get_height()
    ax.text(bar.get_x() + bar.get_width()/2, yval + 0.5, f'{yval:.1f}%', ha='center', va='bottom', fontsize=9)
plt.tight_layout()
plt.show()
```

```
'Índices:'
0        20-29
1        20-29
2        20-29
3        30-39
4        20-29
         ...
88992    20-29
88993    30-39
88994    30-39
88995    20-29
88996    30-39
Name: grupo_edad_10, Length: 88830, dtype: category
Categories (5, object): ['10-19' < '20-29' < '30-39' < '40-49' < '50-59']
```



Análisis de la gráfica: "Proporción de Casos de Morbilidad Materna Extrema por Rango de Edad (10 años)"

La gráfica presenta una visión clara de cómo varía la proporción de casos de morbilidad materna extrema según los grupos de edad en intervalos de 10 años. A partir del análisis, se evidencian patrones importantes que tienen implicaciones directas en la gestión del riesgo obstétrico y en la formulación de políticas públicas en salud materna.

Uno de los aspectos más relevantes es el marcado incremento en la proporción de casos a medida que avanza la edad. Mientras que en los grupos de edad más jóvenes (10-19, 20-29 y 30-39 años) las tasas se mantienen por debajo del 6%, a partir del grupo de 40-49 años la morbilidad se duplica (9.2%) y se dispara en el grupo de 50-59 años, alcanzando un preocupante 17.6%.

Este comportamiento puede estar asociado a diversos factores: en primer lugar, las complicaciones propias de embarazos en edades avanzadas, que incluyen mayor riesgo de hipertensión gestacional, diabetes, preeclampsia y cesáreas de emergencia. Además, las condiciones preexistentes en mujeres mayores pueden agravar el curso del embarazo, generando una mayor vulnerabilidad ante eventos de morbilidad materna extrema.

A pesar de que el grupo de 50-59 años puede tener una baja frecuencia absoluta de embarazos, el hecho de que una proporción tan alta de esos casos presenten complicaciones graves refuerza la necesidad de atención diferenciada y seguimiento especializado para embarazos en mujeres mayores. No se trata simplemente de un fenómeno estadístico, sino de un llamado a priorizar este perfil de riesgo en estrategias de intervención.

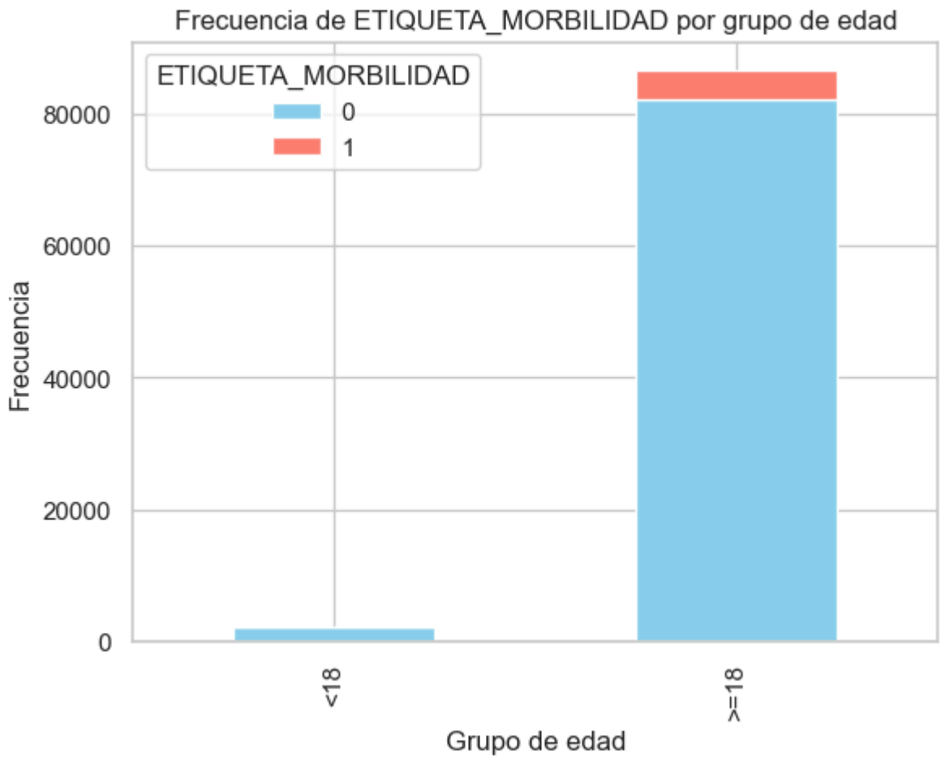
En contraste, los grupos de edad entre 20 y 39 años —considerados tradicionalmente como el rango de mayor fertilidad presentan una menor proporción de complicaciones, lo que respalda la evidencia clínica sobre una menor carga de riesgo en edades reproductivas óptimas.

In [212…

```python
#Clasificar a las gestantes en dos grupos de edad
df['grupo_edad'] = df['EDAD'].apply(lambda x: '<18' if x < 18 else '>=18')

# Calcular la tabla de contingencia entre el grupo de edad y ETIQUETA_MORBILIDAD
contingency_table = pd.crosstab(df['grupo_edad'], df['ETIQUETA_MORBILIDAD'])
print("Tabla de contingencia:")
print(contingency_table)

# Realizar la prueba de chi-cuadrado para evaluar la asociación
chi2, p_value, dof, expected = chi2_contingency(contingency_table)
print("\nPrueba Chi-cuadrado:")
print(f"Chi2: {chi2:.2f}, p-valor: {p_value:.4f}, grados de libertad: {dof}")
print("Frecuencias esperadas:")
print(expected)

# Interpretación de la prueba Chi-cuadrado
alpha = 0.05  # Nivel de significancia
if p_value < alpha:
    interpretation = (
        "La prueba de chi-cuadrado indica que existe una asociación estadísticamente significativa "
        "entre el grupo de edad y la morbilidad materna extrema (p < 0.05). Esto sugiere que la frecuencia "
        "de morbilidad materna extrema difiere significativamente entre las gestantes menores de 18 años y "
        "las de 18 años o más."
    )
else:
    interpretation = (
        "La prueba de chi-cuadrado no indica una asociación estadísticamente significativa entre el grupo de edad "
        "y la morbilidad materna extrema (p >= 0.05). Esto sugiere que la frecuencia de morbilidad materna extrema "
        "no difiere significativamente entre las gestantes menores de 18 años y las de 18 años o más."
    )
print("\nInterpretación de la prueba Chi-cuadrado:")
print(interpretation)
# 4. Gráfico: Diagrama de barras apilado para visualizar la distribución de ETIQUETA_MORBILIDAD por grupo de edad
plt.figure(figsize=(10, 6))
contingency_table.plot(kind='bar', stacked=True, color=['skyblue', 'salmon'])
plt.title("Frecuencia de ETIQUETA_MORBILIDAD por grupo de edad")
plt.xlabel("Grupo de edad")
plt.ylabel("Frecuencia")
```

```
plt.legend(title="ETIQUETA_MORBILIDAD")
plt.show()
```

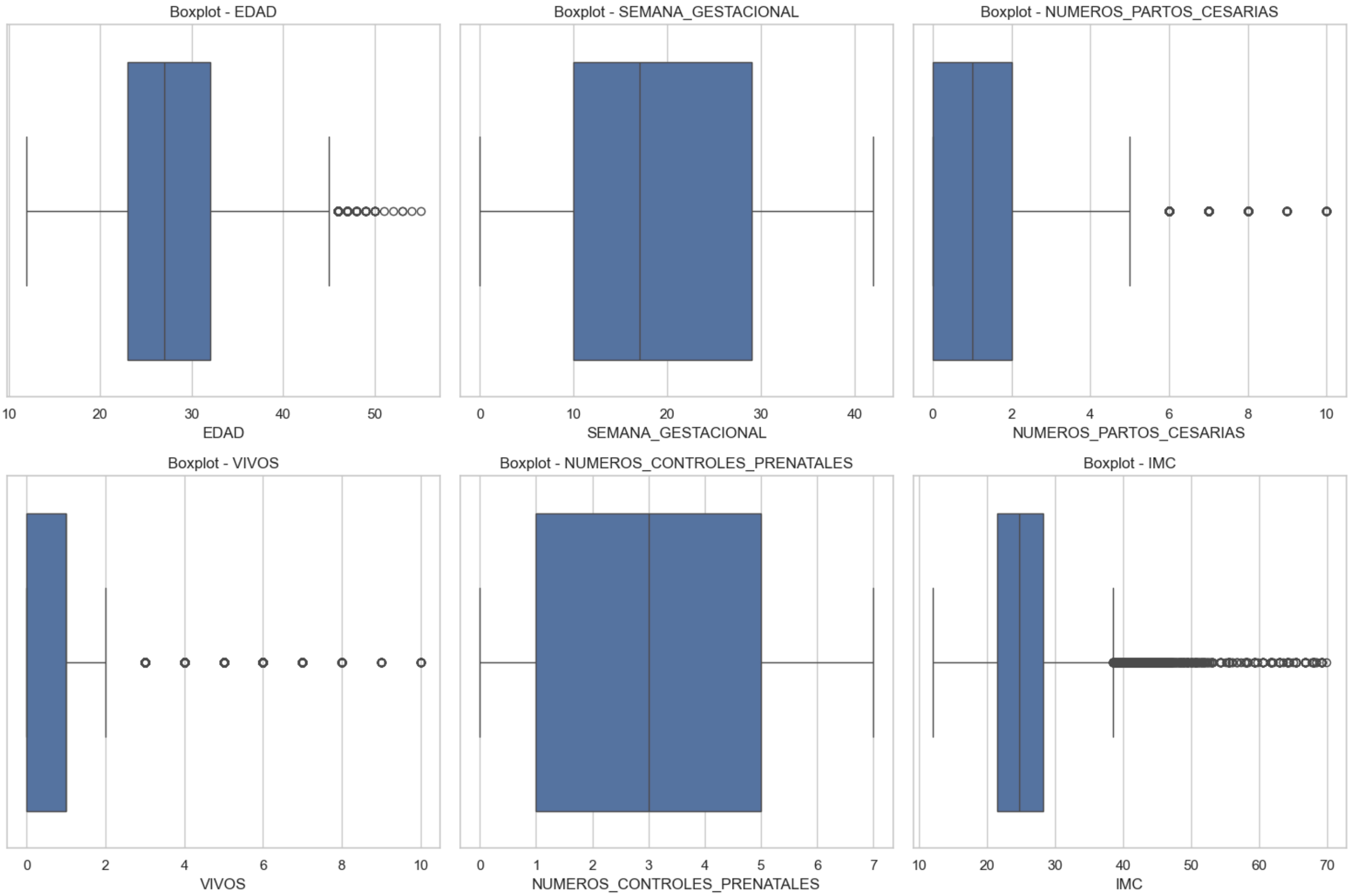
```
Tabla de contingencia:
ETIQUETA_MORBILIDAD      0     1
grupo_edad
<18                   2245   126
>=18                 82188  4271

Prueba Chi-cuadrado:
Chi2: 0.61, p-valor: 0.4348, grados de libertad: 1
Frecuencias esperadas:
[[ 2253.63776877   117.36223123]
 [82179.36223123  4279.63776877]]

Interpretación de la prueba Chi-cuadrado:
La prueba de chi-cuadrado no indica una asociación estadísticamente significativa entre el grupo de edad y la morbilidad materna extrema (p >= 0.05). Esto sugiere que la frecuencia de morbilid
ad materna extrema no difiere significativamente entre las gestantes menores de 18 años y las de 18 años o más.
<Figure size 1000x600 with 0 Axes>
```



Diagramas de caja para detectar outliers

In [186…

```
numeric_cols = df.select_dtypes(include=['float64', 'int64']).columns.drop('MAYOR_35', errors='ignore')

plt.figure(figsize=(15, 10))
for i, col in enumerate(numeric_cols[:6]):  # Muestra solo los primeros 6 por espacio
    plt.subplot(2, 3, i+1)
    sns.boxplot(x=df[col])
    plt.title(f'Boxplot - {col}')
plt.tight_layout()
plt.show()
```



Análisis de Boxplots de Variables Numéricas

Como parte del análisis exploratorio de datos, se generaron boxplots para identificar posibles valores atípicos, rangos intercuartílicos y distribución general de algunas variables numéricas clave en el contexto del embarazo y la morbilidad materna extrema. A continuación presentamos las observaciones más relevantes:

🔷 EDAD El boxplot muestra una distribución centrada entre los 20 y 35 años, con una ligera presencia de outliers hacia la derecha (edades superiores a los 45 años). Esto es esperable, ya que los embarazos en edades avanzadas son menos frecuentes pero clínicamente importantes por su asociación a mayores riesgos.

Aunque hay outliers, No fueron eliminados, ya que representan casos clínicamente relevantes (embarazos en edades extremas).

🔷 SEMANA_GESTACIONAL La variable muestra una distribución simétrica, sin outliers aparentes, lo que se le indica que los registros son razonablemente confiables en términos de semanas de embarazo. Los valores se concentran entre las semanas 10 y 40, como se esperaría.

Esto refuerza que la variable está bien capturada y podría utilizarse directamente sin transformaciones.

🔷 NUMEROS_PARTOS_CESARIAS Hay una clara concentración entre 0 y 3 partos/cesáreas, pero aparecen valores atípicos a partir de 6, llegando hasta 10. Aunque pueden ser reales, también podrían representar casos excepcionales o errores de digitación.

Se puede considerar recortar o imputar valores extremos mayores a 7 para evitar distorsiones en los modelos.

🔷 VIVOS La gran mayoría de los casos está entre 0 y 2 hijos vivos, pero existen varios outliers que superan los 6, llegando hasta 10. Similar al caso anterior, no se han descartado que algunos sean reales, pero podrían requerir validación clínica o truncamiento.

Por ahora, se marcó esta variable como requiere revisión, especialmente si no aporta valor predictivo significativo.

🔷 NUMEROS_CONTROLES_PRENATALES Este indicador muestra una distribución bastante equilibrada entre 0 y 7 controles, sin valores extremos visibles. Esto es positivo, ya que representa consistencia en los registros.

se considera que esta variable está lista para ser usada sin necesidad de transformaciones adicionales.

🔷 IMC (Índice de Masa Corporal) El boxplot del IMC revela una gran cantidad de outliers por encima de 40, alcanzando incluso valores cercanos a 70. Esto podría deberse a errores de digitación o a casos clínicos graves de obesidad mórbida.

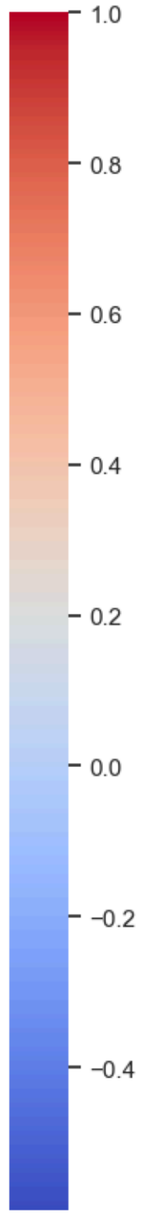
se está considerando imputar o recodificar valores mayores a 60 como "obesidad extrema" o eliminarlos si se confirma que son errores.

In [188…

```
plt.figure(figsize=(14, 10))
correlation = df.corr(numeric_only=True)
sns.heatmap(correlation, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Mapa de calor de correlaciones')
plt.show()
```

Mapa de calor de correlaciones

	EDAD	SEMANA_GESTACIONAL	MAYOR_35	NUMEROS_PARTOS_CESARIAS	VIVOS	NUMEROS_CONTROLES_PRENATALES	IMC	NACIONALIDAD_PROCEDENCIA	NIVEL_EDUCATIVO	AFIC_GRUPO_ETNICO	AFIN_NIVEL_SISBEN	AFIN_GRUPO_POBLACIONAL	COD_MUNICIPIO	GLUCOSA_PRE	MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA	TIPO_DE_CASO	ETIQUETA_MORBILIDAD
EDAD	1.00	-0.07	0.60	0.47	0.43	0.07	0.27	0.00	-0.00	0.01	0.04	-0.01	0.03	0.06	0.01	-0.01	0.04
SEMANA_GESTACIONAL	-0.07	1.00	-0.03	0.05	0.05	-0.10	-0.01	0.03	-0.17	-0.01	-0.00	-0.00	-0.09	-0.03	-0.01	0.08	-0.01
MAYOR_35	0.60	-0.03	1.00	0.27	0.27	0.02	0.11	0.00	-0.00	0.01	0.01	0.00	0.02	0.04	-0.00	-0.02	0.04
NUMEROS_PARTOS_CESARIAS	0.47	0.05	0.27	1.00	0.88	0.05	0.19	0.05	0.09	0.01	-0.01	-0.01	-0.01	0.05	0.04	-0.02	0.01
VIVOS	0.43	0.05	0.27	0.88	1.00	0.08	0.17	0.04	0.04	0.01	-0.03	0.00	0.03	0.06	0.02	-0.05	0.00
NUMEROS_CONTROLES_PRENATALES	0.07	-0.10	0.02	0.05	0.08	1.00	-0.01	0.01	0.12	0.03	-0.03	-0.00	0.06	0.02	0.02	-0.02	0.04
IMC	0.27	-0.01	0.11	0.19	0.17	-0.01	1.00	0.02	0.06	0.02	0.04	-0.01	-0.02	0.10	0.00	-0.02	0.05
NACIONALIDAD_PROCEDENCIA	0.00	0.03	0.00	0.05	0.04	0.01	0.02	1.00	0.01	0.01	0.03	0.04	-0.03	-0.00	-0.01	-0.02	-0.00
NIVEL_EDUCATIVO	-0.00	-0.17	-0.00	0.09	0.04	0.12	0.06	0.01	1.00	0.07	0.15	-0.00	-0.41	-0.05	0.01	0.10	0.06
AFIC_GRUPO_ETNICO	0.01	-0.01	0.01	0.01	0.01	0.03	0.02	0.01	0.07	1.00	0.06	-0.15	-0.03	0.01	0.00	-0.05	0.00
AFIN_NIVEL_SISBEN	0.04	-0.00	0.01	-0.01	-0.03	-0.03	0.04	0.03	0.15	0.06	1.00	-0.59	-0.14	-0.01	0.00	0.03	0.01
AFIN_GRUPO_POBLACIONAL	-0.01	-0.00	0.00	-0.01	0.00	-0.00	-0.01	0.04	-0.00	-0.15	-0.59	1.00	0.05	0.01	-0.00	0.00	-0.01
COD_MUNICIPIO	0.03	-0.09	0.02	-0.01	0.03	0.06	-0.02	-0.03	-0.41	-0.03	-0.14	0.05	1.00	-0.01	-0.00	-0.05	-0.04
GLUCOSA_PRE	0.06	-0.03	0.04	0.05	0.06	0.02	0.10	-0.00	-0.05	0.01	-0.01	0.01	-0.01	1.00	0.00	-0.19	0.02
MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA	0.01	-0.01	-0.00	0.04	0.02	0.02	0.00	-0.01	0.01	0.00	0.00	-0.00	-0.00	0.00	1.00	0.00	0.01
TIPO_DE_CASO	-0.01	0.08	-0.02	-0.02	-0.05	-0.02	-0.02	-0.02	0.10	-0.05	0.03	0.00	-0.05	-0.19	0.00	1.00	-0.01
ETIQUETA_MORBILIDAD	0.04	-0.01	0.04	0.01	0.00	0.04	0.05	-0.00	0.06	0.00	0.01	-0.01	-0.04	0.02	0.01	-0.01	1.00



NUM

MORTALIDAD_PERIN

Análisis del Mapa de Calor de Correlaciones

Como parte del análisis exploratorio de datos y antes de la etapa de selección de características, se elaboró un mapa de calor que muestra las correlaciones de Pearson entre las variables numéricas del dataset, incluyendo la variable objetivo: ETIQUETA_MORBILIDAD.

Este ejercicio fue fundamental para detectar:

Relaciones fuertes entre variables (colinealidad),

Posibles redundancias que podrían distorsionar el modelo,

Y variables con mayor potencial predictivo sobre la condición de morbilidad materna extrema.

Correlaciones más destacadas

NUMEROS_PARTOS_CESARIAS ↔ VIVOS = 0.88 Existe una correlación extremadamente alta entre estas dos variables. Esto es esperable: a mayor número de partos o cesáreas, más probable es que haya más hijos vivos. Sin embargo, esta alta colinealidad indica que probablemente solo una de las dos debería ser incluida en el modelo final para evitar redundancia.

EDAD ↔ MAYOR_35 = 0.60 Esta relación también es lógica, ya que MAYOR_35 es una versión binaria derivada de la edad. Por ende, incluir ambas variables es innecesario. Optaré por conservar EDAD, ya que ofrece mayor granularidad.

NUMEROS_PARTOS_CESARIAS ↔ EDAD = 0.47, y VIVOS ↔ EDAD = 0.43 Estas correlaciones indican una relación moderada: mujeres con mayor edad tienden a haber tenido más partos o hijos. No sorprende, pero es útil para el análisis clínico.

Correlación con la variable objetivo (ETIQUETA_MORBILIDAD)

Lo más relevante para el modelo predictivo es cómo se relacionan las variables explicativas con la morbilidad materna extrema:

Las correlaciones son en su mayoría bajas (≤ 0.05), lo que indica que la predicción de MME no depende fuertemente de una sola variable, sino que requerirá la combinación de múltiples factores (interacciones no lineales, lo cual es común en contextos médicos).

Ejemplos:

EDAD ↔ ETIQUETA_MORBILIDAD: 0.04

IMC ↔ ETIQUETA_MORBILIDAD: 0.05

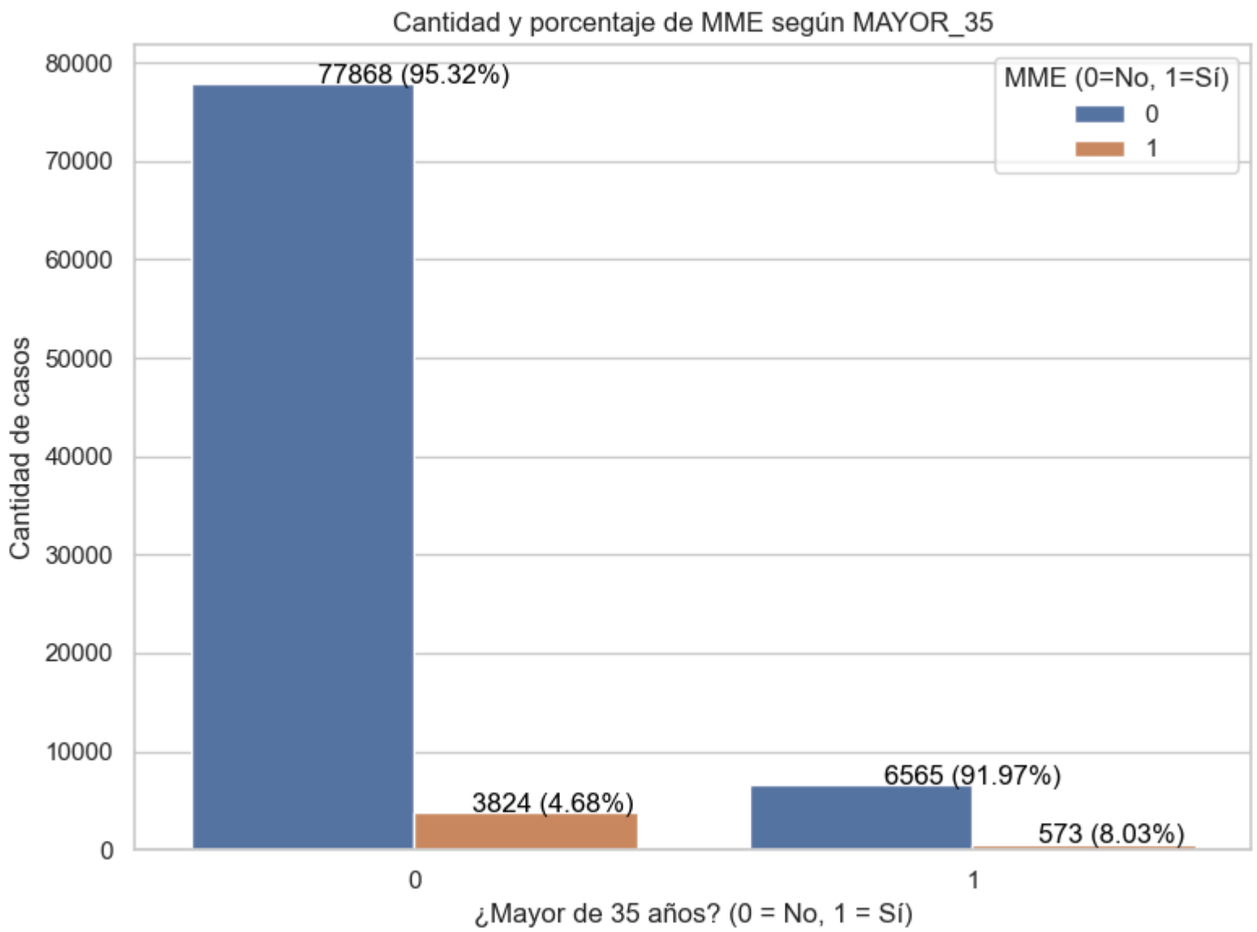
NUMEROS_CONTROLES_PRENATALES: 0.00

Tabla de frecuencias cruzadas

In [193…

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```python
# Crear tabla de frecuencias cruzadas
conteo = df.groupby(['MAYOR_35', 'ETIQUETA_MORBILIDAD']).size().reset_index(name='Cantidad')

# Calcular porcentajes
total_por_grupo = conteo.groupby('MAYOR_35')['Cantidad'].transform('sum')
conteo['Porcentaje'] = (conteo['Cantidad'] / total_por_grupo * 100).round(2)

# Crear gráfico
plt.figure(figsize=(8, 6))
barplot = sns.barplot(data=conteo, x='MAYOR_35', y='Cantidad', hue='ETIQUETA_MORBILIDAD')

# Agregar anotaciones de cantidad y porcentaje
for index, row in conteo.iterrows():
    barplot.text(
        x=index // 2 + (index % 2) * 0.25,
        y=row['Cantidad'] + 50,
        s=f"{int(row['Cantidad'])} ({row['Porcentaje']}%)",
        color='black',
        ha='center'
    )

plt.title('Cantidad y porcentaje de MME según MAYOR_35')
plt.ylabel('Cantidad de casos')
plt.xlabel('¿Mayor de 35 años? (0 = No, 1 = Sí)')
plt.legend(title='MME (0=No, 1=Sí)')
plt.tight_layout()
plt.show()
```



Análisis de MME según la edad materna (>35 años)

Como parte del análisis bivariado entre variables categóricas y la condición de morbilidad materna extrema (ETIQUETA_MORBILIDAD), Se construyó un gráfico de barras apiladas que compara la cantidad y porcentaje de casos de MME según si la gestante era mayor de 35 años (MAYOR_35).

Hallazgos principales

Mujeres menores o iguales a 35 años (MAYOR_35 = 0): Casos sin MME: 77.868 (95.32%)

Casos con MME: 3.824 (4.68%)

Mujeres mayores de 35 años (MAYOR_35 = 1): Casos sin MME: 6.565 (91.97%)

Casos con MME: 573 (8.03%)

Este gráfico evidencia un hallazgo clínicamente significativo: aunque la mayoría de las gestantes están por debajo de los 35 años, el riesgo relativo de MME aumenta notablemente en mujeres mayores de 35 años.

El porcentaje de MME se duplica en mujeres mayores de 35 años (de 4.68% a 8.03%).

Esto sugiere que la edad avanzada es un factor de riesgo relevante y debe ser tenida en cuenta en cualquier modelo predictivo.

Correlación de Pearson (variables numéricas)

In [194…

```python
# Seleccionar variables numéricas
numeric_cols = df.select_dtypes(include=['int64', 'float64']).columns.drop('ETIQUETA_MORBILIDAD', errors='ignore')

# Agregar la variable objetivo al subconjunto
df_corr = df[numeric_cols.tolist() + ['ETIQUETA_MORBILIDAD']]

# Calcular matriz de correlación de Pearson
correlacion_pearson = df_corr.corr(method='pearson')

# Mostrar correlación con la variable objetivo
print("Correlación de Pearson con ETIQUETA_MORBILIDAD:")
print(correlacion_pearson['ETIQUETA_MORBILIDAD'].sort_values(ascending=False))
```

```
Correlación de Pearson con ETIQUETA_MORBILIDAD:
ETIQUETA_MORBILIDAD                           1.000000
NIVEL_EDUCATIVO                               0.056296
IMC                                           0.051311
MAYOR_35                                      0.041940
EDAD                                          0.041932
NUMEROS_CONTROLES_PRENATALES                  0.036776
GLUCOSA_PRE                                   0.017667
AFIN_NIVEL_SISBEN                             0.014186
MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA    0.008082
NUMEROS_PARTOS_CESARIAS                       0.007960
AFIC_GRUPO_ETNICO                             0.002509
VIVOS                                         0.001418
NACIONALIDAD_PROCEDENCIA                     -0.002739
SEMANA_GESTACIONAL                           -0.007952
TIPO_DE_CASO                                 -0.012015
AFIN_GRUPO_POBLACIONAL                       -0.013690
COD_MUNICIPIO                                -0.040140
Name: ETIQUETA_MORBILIDAD, dtype: float64
```

Para identificar variables numéricas con posible poder explicativo sobre la ocurrencia de morbilidad materna extrema (MME), se calculó la correlación de Pearson entre cada variable independiente y la variable objetivo (ETIQUETA_MORBILIDAD). Aunque esta medida refleja únicamente relaciones lineales, fue útil para obtener una visión preliminar.

Variables con mayor correlación positiva NIVEL_EDUCATIVO → 0.056

IMC → 0.051

MAYOR_35 → 0.041

EDAD → 0.041

Estas cuatro variables muestran la correlación más alta con la ocurrencia de MME, aunque los valores siguen siendo bajos (< 0.06). Esto confirma lo que se observa en el mapa de calor: no existe una variable individual fuertemente correlacionada con la morbilidad, lo que sugiere un fenómeno complejo y multifactorial.

Variables con correlación casi nula GLUCOSA_PRE, AFIN_NIVEL_SISBEN, NUMEROS_PARTOS_CESARIAS, VIVOS, etc. tienen correlaciones por debajo de 0.02, lo que indica que no guardan relación lineal directa con la variable objetivo.

Sin embargo, Aún no se han descartado, ya que podrían aportar valor en modelos no lineales o a través de interacciones con otras variables.

Variables con correlación negativa (aunque débil) COD_MUNICIPIO → -0.040

AFIN_GRUPO_POBLACIONAL → -0.013

TIPO_DE_CASO → -0.012

Estas variables presentan una débil asociación inversa con la morbilidad materna. Aunque parecen tener poca influencia individual, las serán almacenadas como candidatas a ser descartadas tras la fase de selección automática de características (RFE, Random Forest).

Este análisis refuerza la hipótesis de que la morbilidad materna extrema no depende de una sola variable fuerte, sino de la combinación de múltiples factores sutiles. Por eso, es fundamental usar modelos que puedan capturar patrones complejos como árboles de decisión, Random Forest o redes neuronales.

Correlación de Spearman (para ordinales o no lineales)

In [195…

```
# Calcular matriz de correlación de Spearman
correlacion_spearman = df_corr.corr(method='spearman')

# Mostrar correlación con la variable objetivo
print("Correlación de Spearman con ETIQUETA_MORBILIDAD:")
print(correlacion_spearman['ETIQUETA_MORBILIDAD'].sort_values(ascending=False))
```

```
Correlación de Spearman con ETIQUETA_MORBILIDAD:
ETIQUETA_MORBILIDAD                          1.000000
NIVEL_EDUCATIVO                              0.052240
IMC                                          0.046408
MAYOR_35                                     0.041940
NUMEROS_CONTROLES_PRENATALES                 0.036819
EDAD                                         0.035451
GLUCOSA_PRE                                  0.015118
AFIN_NIVEL_SISBEN                            0.014223
MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA   0.008082
AFIC_GRUPO_ETNICO                            0.003159
NACIONALIDAD_PROCEDENCIA                    -0.002739
NUMEROS_PARTOS_CESARIAS                     -0.003578
VIVOS                                       -0.006510
SEMANA_GESTACIONAL                          -0.007332
TIPO_DE_CASO                                -0.013141
AFIN_GRUPO_POBLACIONAL                      -0.014618
COD_MUNICIPIO                               -0.045547
Name: ETIQUETA_MORBILIDAD, dtype: float64
```

Análisis de Correlación de Spearman con la Morbilidad Materna Extrema

Como complemento al análisis de Pearson, se decidió calcular también la correlación de Spearman entre las variables predictoras y la variable objetivo (ETIQUETA_MORBILIDAD). Esta métrica es más robusta frente a relaciones no lineales y ordenadas, por lo que se le permite evaluar asociaciones que podrían no haber sido capturadas en el análisis lineal.

Variables con correlación positiva más notable NIVEL_EDUCATIVO → 0.052

IMC → 0.046

MAYOR_35 → 0.041

NUMEROS_CONTROLES_PRENATALES → 0.036

EDAD → 0.035

Estas variables mantienen una tendencia similar a la observada con Pearson, lo cual al pemite dar mayor confianza en su comportamiento. Aunque los coeficientes siguen siendo bajos, hay una señal débil pero persistente de que una mayor edad, índice de masa corporal y nivel educativo podrían estar asociados con mayor riesgo de MME.

Variables con correlación casi nula

GLUCOSA_PRE, AFIN_NIVEL_SISBEN, NUMEROS_PARTOS_CESARIAS, etc., nuevamente muestran muy poca relación individual con la etiqueta.

Sin embargo, algunas de estas variables tienen una importancia clínica potencial y podrían aportar más valor si interactúan con otras variables (por ejemplo, GLUCOSA_PRE combinada con IMC o edad).

Variables con correlación negativa

COD_MUNICIPIO → -0.045

AFIN_GRUPO_POBLACIONAL → -0.014

TIPO_DE_CASO → -0.013

SEMANA_GESTACIONAL → -0.007

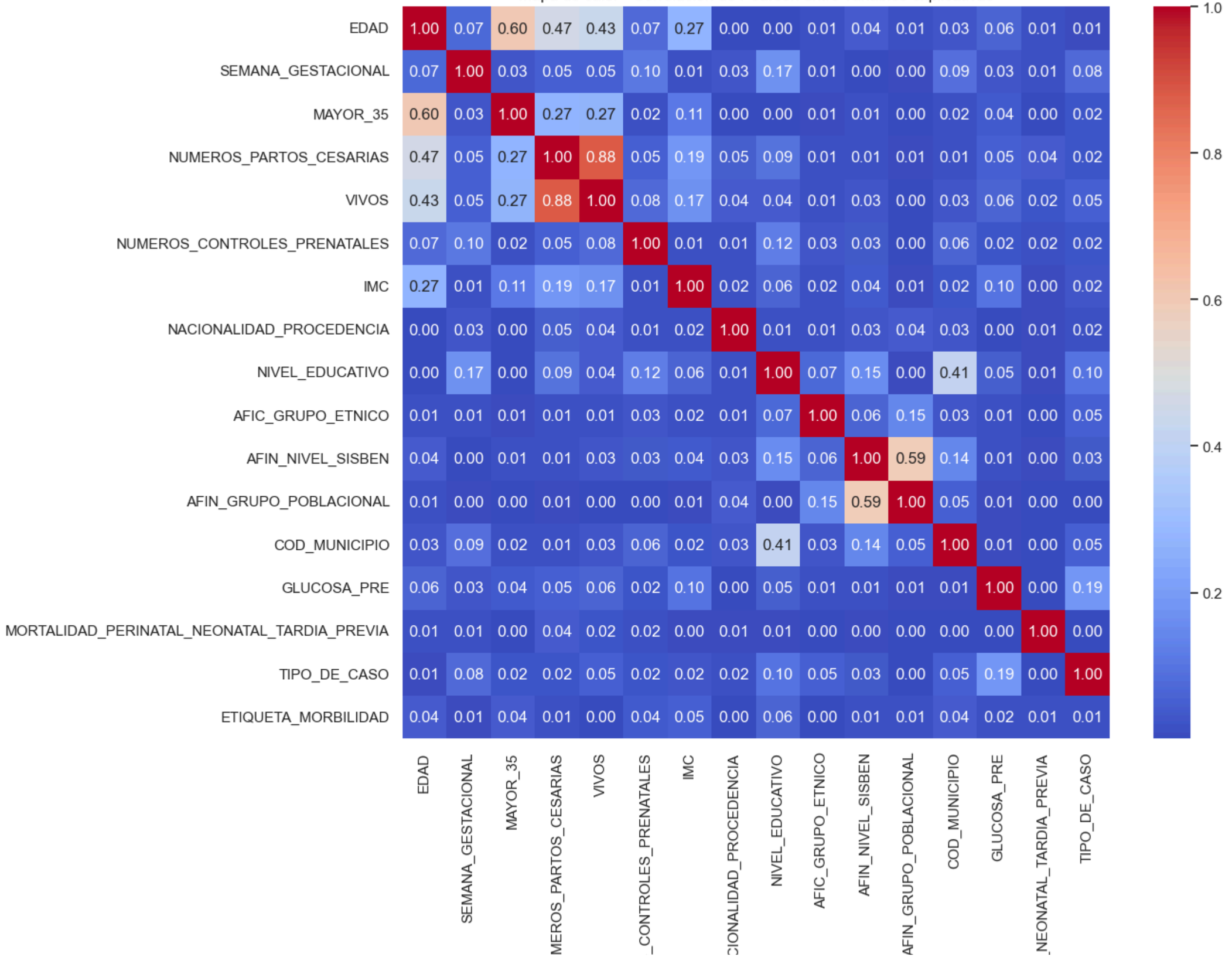
Estas variables presentan una correlación negativa muy débil con la morbilidad materna extrema. No las descarto aún, pero sí las se marcó como candidatas a eliminación en un filtrado automático posterior, si no aportan al desempeño del modelo.

Visualización de colinealidad entre variables (heatmap Pearson)

In [130…

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(12, 10))
sns.heatmap(correlacion_pearson[numeric_cols].abs(), annot=True, fmt=".2f", cmap='coolwarm')
plt.title('Mapa de calor - Correlación de Pearson entre variables explicativas')
plt.show()
```

Mapa de calor - Correlación de Pearson entre variables explicativas



NU

NUMEROS

NA

MORTALIDAD_PERINATAL_

Análisis del Mapa de Calor de Correlación entre Variables Explicativas

Como parte del análisis exploratorio y antes de realizar la selección definitiva de características, se construyó un mapa de calor de correlación de Pearson entre las variables independientes (explicativas) del dataset. Esta herramienta se le permitió detectar colinealidades, relaciones internas entre atributos y posibles redundancias que podrían afectar negativamente el rendimiento del modelo predictivo.

Correlaciones muy altas (colinealidad fuerte)

NUMEROS_PARTOS_CESARIAS ↔ VIVOS = 0.88

NUMEROS_PARTOS_CESARIAS ↔ EDAD = 0.47

VIVOS ↔ EDAD = 0.43

EDAD ↔ MAYOR_35 = 0.60

Estas correlaciones altas se le advierten que incluir ambos campos en el modelo puede introducir redundancia y sobreajuste, especialmente en modelos sensibles a la multicolinealidad (como regresión logística).

Correlaciones moderadas

NUMEROS_CONTROLES_PRENATALES ↔ NUMEROS_PARTOS_CESARIAS = 0.27

IMC ↔ EDAD = 0.27

Estas relaciones no son problemáticas, pero indican que existen asociaciones lógicas entre variables clínicas y demográficas que pueden reforzar patrones en el modelo sin ser redundantes.

Correlaciones bajas o nulas

La mayoría de las variables (como AFIC_GRUPO_ETNICO, TIPO_DE_CASO, MORTALIDAD_PERINATAL_NEONATAL_TARDIA_PREVIA, etc.) presentan correlaciones bajas (< 0.10) con otras variables.

Esto indica que estas variables pueden aportar información complementaria, aunque su valor predictivo individual sea bajo.

En general, la baja colinealidad entre la mayoría de las variables permite dar espacio para trabajar con modelos que se beneficien de diversidad estructural.

Variables con colinealidad interna no útil

AFIN_NIVEL_SISBEN y AFIN_GRUPO_POBLACIONAL → 0.59

AFIN_NIVEL_SISBEN ↔ NIVEL_EDUCATIVO → 0.41

Estas variables parecen capturar información socioeconómica similar. En la selección final se considera conservar solo una de ellas, basándome en su importancia según Random Forest y RFE.

Identificar posibles variables colineales (redundantes)

In [196…

```python
# Umbral para considerar colinealidad (ej. 0.8)
umbral = 0.8
colineales = []

# Recorrer matriz de correlación
for i in range(len(numeric_cols)):
    for j in range(i+1, len(numeric_cols)):
        corr = abs(correlacion_pearson.iloc[i, j])
        if corr > umbral:
            colineales.append((numeric_cols[i], numeric_cols[j], corr))

# Mostrar pares colineales
print("Pares de variables con colinealidad alta (>|0.8|):")
for var1, var2, corr in colineales:
    print(f"{var1} ↔ {var2} = {corr:.2f}")
```

```
Pares de variables con colinealidad alta (>|0.8|):
NUMEROS_PARTOS_CESARIAS ↔ VIVOS = 0.88
```

Análisis Chi-cuadrado

Preprocesamiento: codificar y limpiar

In [197…

```python
from sklearn.preprocessing import LabelEncoder
import pandas as pd

# Copia de trabajo del DataFrame
df_encoded = df.copy()

# Codificar variables categóricas tipo 'object'
for col in df_encoded.select_dtypes(include='object').columns:
    df_encoded[col] = LabelEncoder().fit_transform(df_encoded[col].astype(str))

# Eliminar filas con datos faltantes
df_encoded = df_encoded.dropna()

# Separar variables predictoras y variable objetivo
X = df_encoded.drop(columns=['ETIQUETA_MORBILIDAD'])
y = df_encoded['ETIQUETA_MORBILIDAD'].astype(int)
```

In [198…

```python
from sklearn.feature_selection import SelectKBest, chi2

# Asegurarse de que los valores sean no negativos
X_chi = X.copy()
X_chi[X_chi < 0] = 0

# Selección con Chi-cuadrado
chi_selector = SelectKBest(score_func=chi2, k=10)
chi_selector.fit(X_chi, y)

# Mostrar resultados
chi_scores = pd.DataFrame({
    'Variable': X.columns,
    'Chi2_Score': chi_selector.scores_
}).sort_values(by='Chi2_Score', ascending=False)
```

```
print("Top 10 variables según Chi-cuadrado:")
display(chi_scores.head(10))
```

```
Top 10 variables según Chi-cuadrado:
```

	Variable	Chi2_Score
22	COD_MUNICIPIO	2.453807e+06
25	DIAGNOSTICOS	1.595540e+04
17	NIVEL_EDUCATIVO	5.042337e+02
29	FECHA_GLUCOSA	1.558160e+02
12	IMC	1.056242e+02
0	DOCUMENTO	9.732404e+01
23	HIPERTENSION	7.904757e+01
26	HEMOGLOBINA	6.443793e+01
5	MAYOR_35	6.064847e+01
1	EDAD	5.262430e+01

Análisis del Test Chi-cuadrado para Selección de Características

Como parte del proceso de selección de variables con enfoque estadístico, se aplicó la prueba de Chi-cuadrado para evaluar la asociación entre variables categóricas y la variable objetivo ETIQUETA_MORBILIDAD.

COD_MUNICIPIO aparece como la variable más asociada a la morbilidad, lo cual podría estar reflejando condiciones territoriales, infraestructura médica o acceso a servicios de salud diferenciado por región. Esto se le parece interesante para futuras segmentaciones por área geográfica.

DIAGNOSTICOS y HIPERTENSION son clínicamente muy relevantes. Se evidencia satisfacción verlos con puntajes altos, ya que confirma su utilidad como predictores.

NIVEL_EDUCATIVO también se posiciona alto, reafirmando la dimensión socioeconómica como un factor importante en la salud materna.

IMC y HEMOGLOBINA aparecen bien posicionadas, lo cual valida su inclusión en los modelos desde el punto de vista de la salud física y nutricional de la gestante.

DOCUMENTO aparece en el ranking, pero no debería influir. Esto es probablemente un error o una variable codificada incorrectamente como categórica. Será eliminada en la depuración final.

Este análisis Chi^2 permitió confirmar varias hipótesis clínicas y sociales, y se le dio una guía objetiva para priorizar variables en el proceso de modelado. A pesar de que algunas variables no tienen correlación lineal fuerte, esta técnica ayudó a detectar asociaciones significativas que no se capturan con Pearson o Spearman.

Importancia de variables en Random Forest

In [199…

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X, y)

rf_scores = pd.DataFrame({
    'Variable': X.columns,
    'Importancia_RF': rf.feature_importances_
}).sort_values(by='Importancia_RF', ascending=False)

print("Top 10 variables según Random Forest:")
display(rf_scores.head(10))
```

Top 10 variables según Random Forest:

	Variable	Importancia_RF
12	IMC	0.092115
0	DOCUMENTO	0.088294
3	FPP	0.084889
2	FUM	0.083295
28	GLUCOSA_PRE	0.072913
4	SEMANA_GESTACIONAL	0.065304
1	EDAD	0.062352
22	COD_MUNICIPIO	0.053025
26	HEMOGLOBINA	0.051793
27	FECHA_HB	0.051498

Eliminación Recursiva de Características (RFE)

In [137…

```
from sklearn.feature_selection import RFE

# RFE con Random Forest como estimador
rfe_selector = RFE(estimator=RandomForestClassifier(n_estimators=100, random_state=42), n_features_to_select=10)
rfe_selector.fit(X, y)

rfe_result = pd.DataFrame({
    'Variable': X.columns,
    'Seleccionada_RFE': rfe_selector.support_
})

print("Variables seleccionadas por RFE:")
display(rfe_result[rfe_result['Seleccionada_RFE'] == True])
```

Variables seleccionadas por RFE:

	Variable	Seleccionada_RFE
0	DOCUMENTO	True
1	EDAD	True
2	FUM	True
3	FPP	True
4	SEMANA_GESTACIONAL	True
12	IMC	True
22	COD_MUNICIPIO	True
26	HEMOGLOBINA	True
27	FECHA_HB	True
28	GLUCOSA_PRE	True

Análisis de Importancia de Variables según Random Forest

Con el objetivo de identificar las variables más relevantes para la predicción de morbilidad materna extrema (MME), Se entrenó un modelo de Random Forest y se analizó la importancia relativa de cada atributo. Esta técnica es ideal porque permite detectar patrones no lineales y combinaciones de variables que influyen en la predicción, sin asumir ninguna forma funcional específica.

IMC, GLUCOSA_PRE, HEMOGLOBINA, y EDAD aparecen con alta importancia, lo cual valida la hipótesis clínica de que los factores físicos y metabólicos influyen en la probabilidad de MME.

La presencia de variables temporales como FUM, FPP y FECHA_HB se le parece lógica, ya que reflejan momentos del embarazo en los que pueden surgir complicaciones. Sin embargo, su uso en producción puede requerir estandarización o derivación de nuevas variables como “semanas transcurridas”.

La aparición de DOCUMENTO como una de las variables más importantes es una bandera roja. Es probable que esté actuando como un identificador mal codificado o indirectamente vinculado con otras variables. Se plantea eliminarla del conjunto de datos final o revisar su origen.

La inclusión de COD_MUNICIPIO reafirma lo que Se encontró en Chi^2: la dimensión geográfica tiene influencia, ya sea por diferencias en acceso a salud o condiciones poblacionales.

El análisis de importancia usando Random Forest fue clave para validar hallazgos previos y descubrir nuevas variables relevantes. Este enfoque evidenció que el modelo no se basa únicamente en una sola dimensión (edad o antecedentes clínicos), sino que combina información demográfica, clínica y temporal para construir su predicción.

In [ ]: