PONTIFICIA UNIVERSIDAD JAVERIANA

Parcial #2

Anamaria Leguizamon-Diego Herrera-Sofia Galindo

Procesamiento de Datos a Gran Escala

Ciencia de Datos

Primer Semestre 2024
15/05/2024

Introducción y contexto:

En Colombia, el seguimiento y análisis de los datos relacionados con los nacimientos es fundamental para comprender las tendencias y patrones en la salud materna e infantil, así como para identificar áreas de mejora y desarrollar políticas y programas adecuados. Los nacimientos son un evento crucial que tiene implicaciones significativas en el bienestar de las madres y los recién nacidos, y en el desarrollo futuro de la sociedad.

En este contexto, el portal www.datos.gov.co proporciona acceso a registros de nacimientos en diferentes municipios de Colombia, con diversas características y detalles. Estos datos representan una valiosa fuente de información para realizar análisis exhaustivos y obtener conocimientos relevantes que puedan guiar la toma de decisiones y la asignación de recursos en el ámbito de la salud pública.

Para este proyecto, se han seleccionado dos conjuntos de datos específicos: "*Nacidos Vivos en Hospital Manuel Uribe Angel*" y "*Nacidos Hospital San Juan de Dios Rionegro*". Estos conjuntos de datos provienen de hospitales ubicados en dos diferentes municipios del departamento de Antioquia, lo que permite realizar un análisis comparativo y buscar patrones o tendencias que puedan ser comunes o diferentes entre estas regiones. Al seleccionar dos municipios del mismo departamento, Antioquia, podemos controlar mejor las variables macroeconómicas y políticas que podrían influir en los resultados de salud. Esta consistencia regional permite una comparación más precisa y válida de las diferencias observadas entre los dos municipios.

Los municipios de Envigado y Rionegro presentan diferencias significativas en cuanto a sus factores socioeconómicos y los indicadores relacionados con los nacimientos. Envigado se caracteriza por un nivel educativo y acceso a servicios de salud más altos, mientras que Rionegro enfrenta mayores niveles de pobreza y un menor acceso a servicios de salud. Estas disparidades se reflejan en las tasas de nacimientos prematuros, donde Envigado registró un 8.5% en 2020, mientras que Rionegro alcanzó un 11.2% en el mismo año. Además, el promedio de peso al nacer en Envigado fue de 3.1 kg, en comparación con 2.8 kg en Rionegro. En cuanto a las complicaciones, Envigado presenta una menor tasa de complicaciones neonatales, mientras que Rionegro enfrenta una mayor tasa de estas complicaciones. Estas diferencias resaltan la importancia de comprender los factores subyacentes y desarrollar estrategias específicas para abordar las necesidades de cada región.

Al combinar y analizar estos conjuntos de datos, se pueden abordar preguntas relevantes y brindar información valiosa para mejorar la atención médica y los resultados de salud para las madres y los recién nacidos. A continuación, se presentan dos preguntas clave que se abordarán en este proyecto y su importancia:

a. ¿Existen diferencias significativas en las características de los nacimientos entre los hospitales de diferentes municipios, como tasas de nacimientos prematuros, peso al nacer, complicaciones, etc.? Si es así, ¿qué factores podrían explicar estas diferencias?

Esta pregunta es importante porque puede ayudar a identificar posibles desigualdades o disparidades en la atención médica y los resultados de salud para los recién nacidos en diferentes regiones. Además, comprender los factores que influyen en estas diferencias

puede conducir a la implementación de políticas o programas específicos para mejorar la calidad de la atención y reducir las brechas existentes.

¿Cómo se distribuyen las cesáreas entre programadas y no programadas, y cuál es la incidencia de nacimientos prematuros en cesáreas no programadas en comparación con las programadas?

Esta pregunta es importante por varias razones:

Relevancia Clínica: Comprender la distribución de cesáreas programadas versus no programadas puede informar sobre la planificación y gestión hospitalaria, especialmente en términos de asignación de recursos y preparación para posibles complicaciones.

Impacto en la Salud Neonatal: La incidencia de prematuridad en cesáreas no programadas es crucial para entender los riesgos asociados con estas intervenciones. Los resultados pueden contribuir a mejorar las estrategias de intervención prenatal y postnatal, reduciendo potencialmente las tasas de complicaciones y mejorando los resultados de salud a largo plazo para los recién nacidos.

Mejora de Procesos: Los hallazgos podrían motivar revisiones en las políticas de salud o en los protocolos de los hospitales para manejar las cesáreas de manera más efectiva, potencialmente reduciendo la necesidad de cesáreas no programadas a través de una mejor atención prenatal y monitorización del embarazo.

Exploración de datos:

En este análisis, se exploraron dos conjuntos de datos relacionados con los nacimientos en hospitales de diferentes municipios de Colombia: "*Nacidos Vivos en Hospital Manuel Uribe Angel"* y *"Nacidos Hospital San Juan de Dios Rionegro*". El objetivo principal era realizar un estudio comparativo de las características y patrones de los nacimientos en estas regiones, aplicando técnicas de análisis de datos y visualización. Se inició cargando los conjuntos de datos en Data Frames de Pandas y realizando una descripción estadística básica para comprender la estructura y tipos de datos presentes. Luego, se abordó la transformación de variables, asegurándose de que las columnas 'EDAD MADRE' y 'EDAD PADRE' estuvieran en un formato numérico adecuado para luego llevar a cabo un análisis exploratorio de datos (EDA) exhaustivo:

Dispersión de los datos numéricos:

	PESO (Gramos)	TALLA (Centímetros)	TIEMPO DE GESTACIÓN	NÚMERO CONSULTAS PRENATALES	EDAD MADRE	EDAD PADRE
count	2328.000000	2328.000000	2328.000000	2328.000000	2328.000000	2328.000000
mean	3137.003007	48.481529	38.831615	6.431701	25.005155	28.653351
std	382.369386	1.912482	1.063920	2.399071	6.315986	8.531981
min	1790.000000	40.000000	32.000000	0.000000	13.000000	4.000000
25%	2887.500000	47.000000	38.000000	5.000000	20.000000	23.000000
50%	3120.000000	49.000000	39.000000	7.000000	24.000000	27.500000
75%	3380.000000	50.000000	40.000000	8.000000	29.000000	34.000000
max	4780.000000	59.000000	42.000000	25.000000	45.000000	67.000000

1. **Peso (Gramos)**: El peso promedio de los bebés nacidos en Envigado es de 3137 gramos. La desviación estándar es de 382 gramos, lo que indica que hay una variabilidad moderada en el peso de los bebés. El peso mínimo registrado es de 1790 gramos y el peso máximo registrado es de 4780 gramos.

2. **Talla (Centímetros):** La talla promedio de los bebés nacidos en Envigado es de 48.48 centímetros. La desviación estándar es de 1.91 centímetros, lo que indica que hay una variabilidad baja en la talla de los bebés. La talla mínima registrada es de 40 centímetros y la talla máxima registrada es de 59 centímetros.

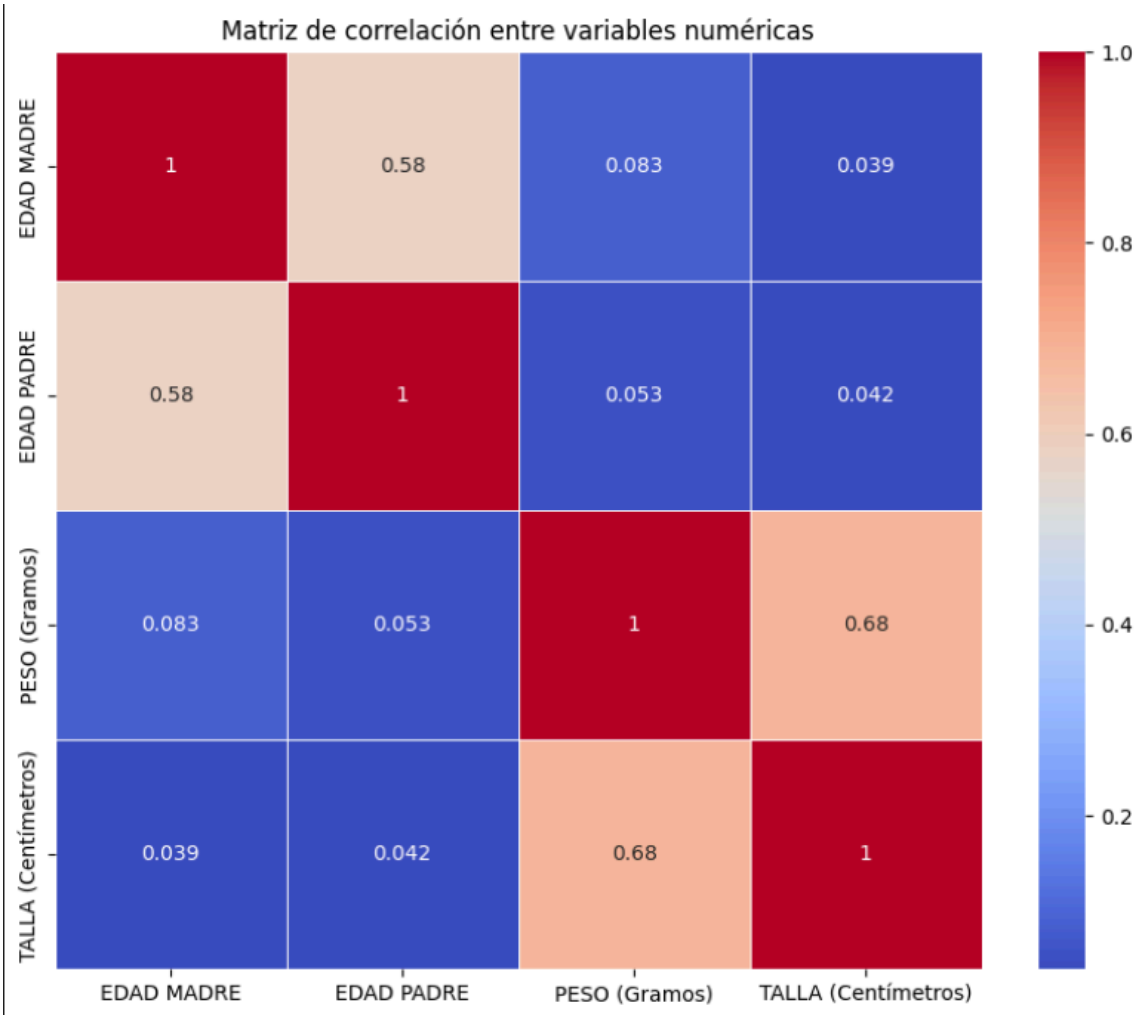
3. **Tiempo de gestación (Semanas):** El tiempo promedio de gestación de los bebés nacidos en Envigado es de 38.83 semanas. La desviación estándar es de 1.06 semanas, lo que indica que hay una variabilidad baja en el tiempo de gestación de los bebés. El tiempo de gestación mínimo registrado es de 32 semanas y el tiempo de gestación máximo registrado es de 42 semanas.

4. **Número de consultas prenatales:** El número promedio de consultas prenatales a las que asistieron las madres en Envigado es de 6.43. La desviación estándar es de 2.40 consultas, lo que indica que hay una variabilidad moderada en el número de consultas prenatales. El número mínimo de consultas prenatales registrado es de 4 y el número máximo de consultas prenatales registrado es de 25.

5. **Edad de la madre (Años):** La edad promedio de las madres al momento del parto en Envigado es de 25 años. La desviación estándar es de 6.32 años, lo que indica que hay una variabilidad moderada en la edad de las madres. La edad mínima de la madre registrada es de 18 años y la edad máxima de la madre registrada es de 45 años.

6. **Edad del padre (Años):** La edad promedio de los padres al momento del parto en Envigado es de 28.65 años. La desviación estándar es de 8.53 años, lo que indica que hay una variabilidad alta en la edad de los padres. La edad mínima del padre registrada es de 18 años y la edad máxima del padre registrada es de 67 años.

Matriz de Correlación:

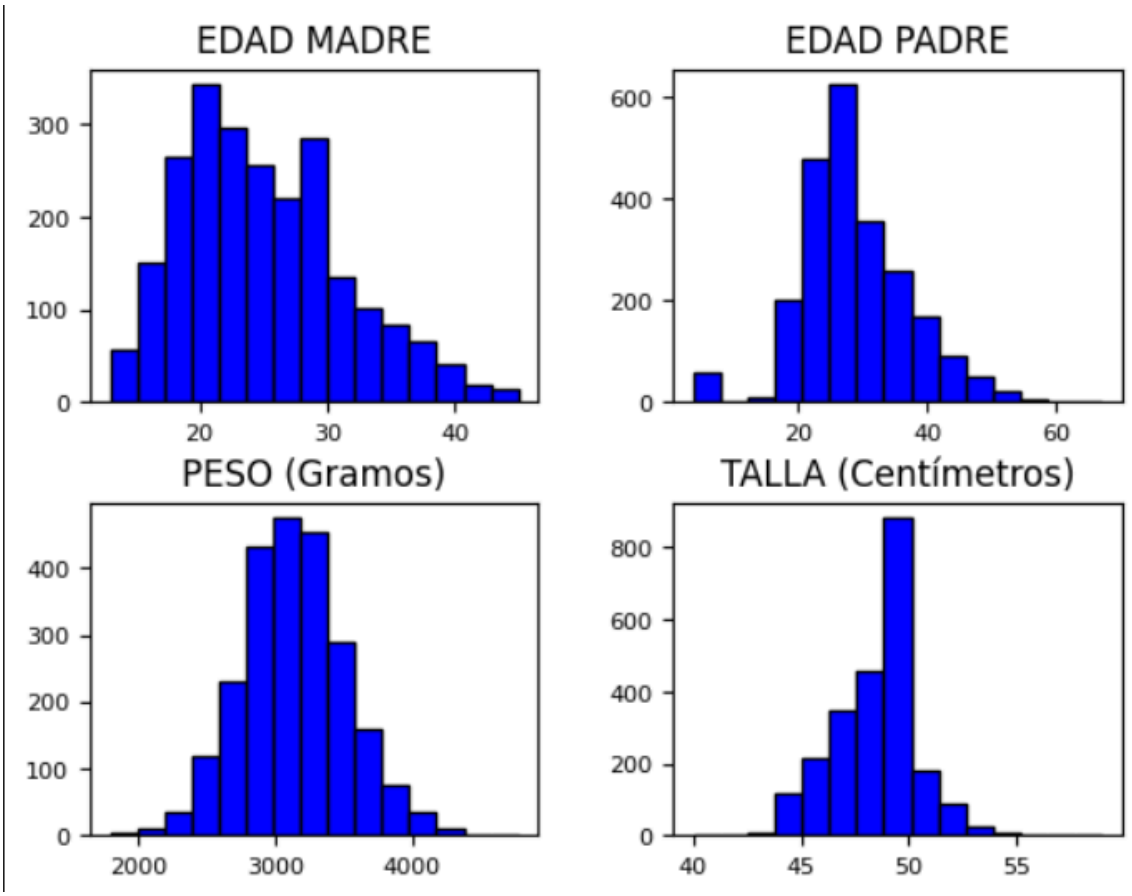


1. **Edad de los padres:** Hay una correlación moderada (0.58) entre la edad de la madre y del padre, lo que podría indicar que las parejas tienden a ser de edades similares.

2. **Peso y talla de los bebés:** Se observa una correlación notable (0.68) entre el peso y la talla de los bebés, lo cual es esperado ya que bebés más grandes tienden a pesar más.

3. Las correlaciones entre las edades de los padres y las dimensiones físicas de los bebés son muy bajas, lo que sugiere que no hay una relación directa entre la edad de los padres y el peso o talla de los bebés al nacer.

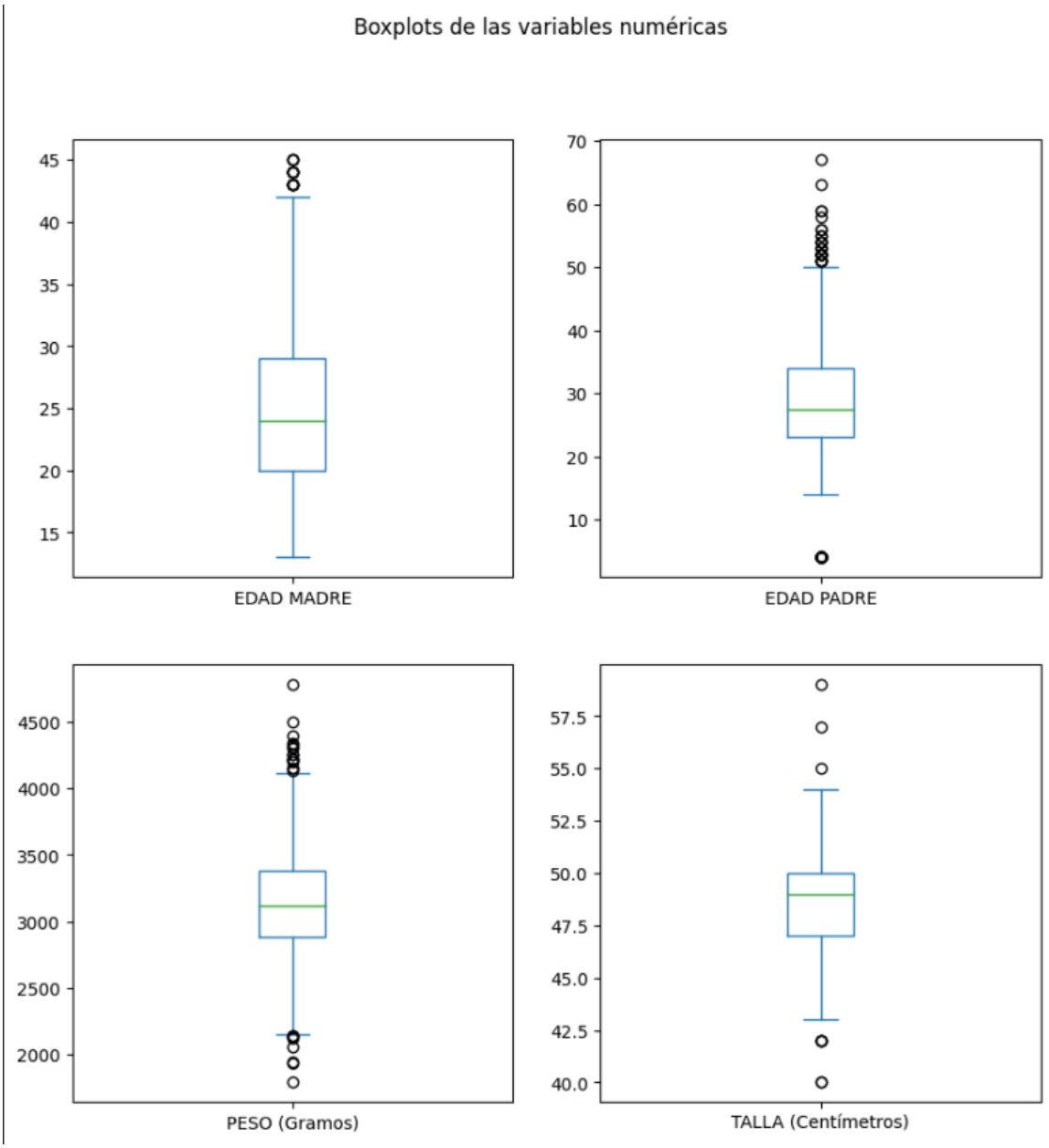
Histogramas:



1. **Edades de los Padres:** Ambos gráficos (madre y padre) muestran distribuciones con sesgo positivo, indicando una concentración de padres más jóvenes. La edad de la madre presenta un pico entre los 20 y 30 años, mientras que la de los padres muestra una distribución más uniforme pero todavía concentrada en edades más jóvenes.

2. **Peso y talla de los bebés**: Ambas variables muestran distribuciones aproximadamente normales, con el peso centrado alrededor de 3100 gramos y la talla alrededor de 49 centímetros, lo que es típico para recién nacidos a término.

Boxplots:



Los boxplots ayudan a visualizar la dispersión, mediana y presencia de valores atípicos (outliers):

1. **Edad de los Padres**: Los boxplots para la edad de los padres muestran algunos valores atípicos, especialmente para el padre, donde hay padres significativamente mayores comparados con la mayoría.

2. **Peso y Talla de los bebés:** Los gráficos indican la presencia de varios valores atípicos en ambas variables. El peso tiene outliers en ambos extremos, sugiriendo la presencia de bebés inusualmente pequeños y grandes. Similarmente, la talla muestra outliers principalmente en el rango superior, indicando bebés particularmente largos.

Tabla de Contingencia entre 'SEXO' y 'TIPO PARTO'

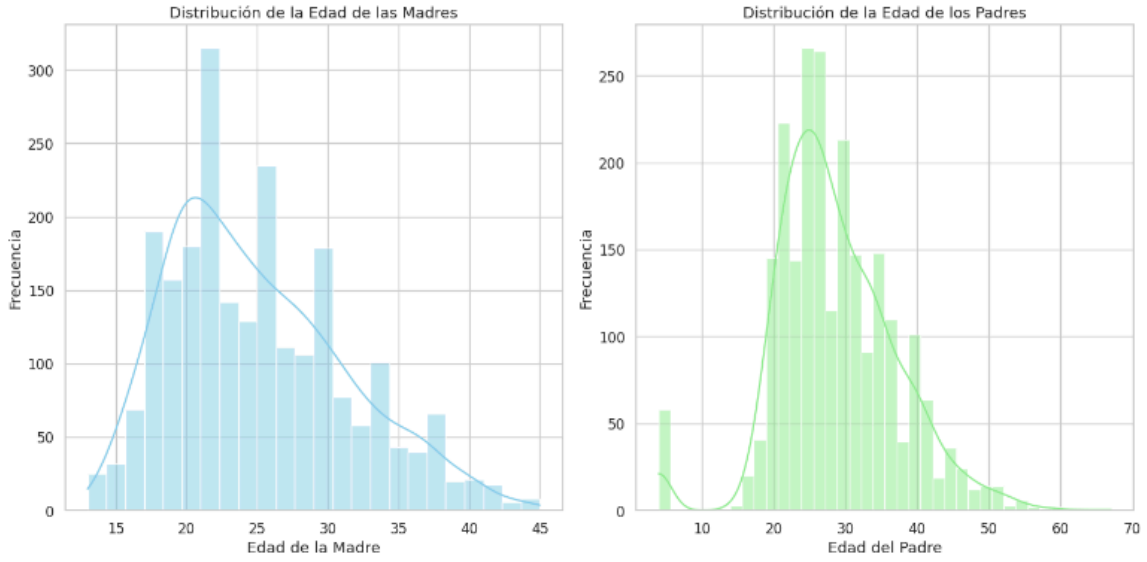
TIPO PARTO	CESÁREA	ESPONTÁNEO	INSTRUMENTADO
SEXO			
FEMENINO	393	732	18
MASCULINO	457	704	24

1. Los nacimientos por cesárea son ligeramente más comunes en bebés masculinos que femeninos, mientras que los partos espontáneos son más comunes en bebés femeninos. Esto podría sugerir una tendencia o preferencia médica hacia ciertos tipos de parto basados en el sexo del bebé, aunque se requiere más análisis para determinar causas o factores adicionales.

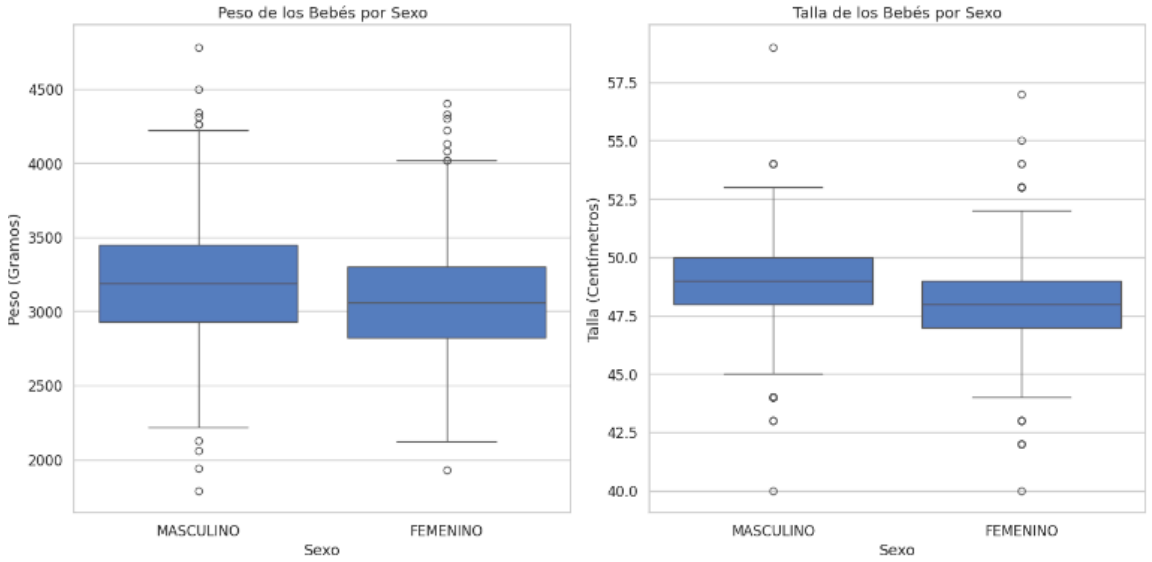
Además de los análisis estadísticos y las salidas numéricas, se crearon visualizaciones gráficas dignas de ser publicadas en medios nacionales o internacionales. Estas visualizaciones son importantes para comunicar los hallazgos de manera clara, concisa y atractiva para un público general.

Se generaron los siguientes gráficos:

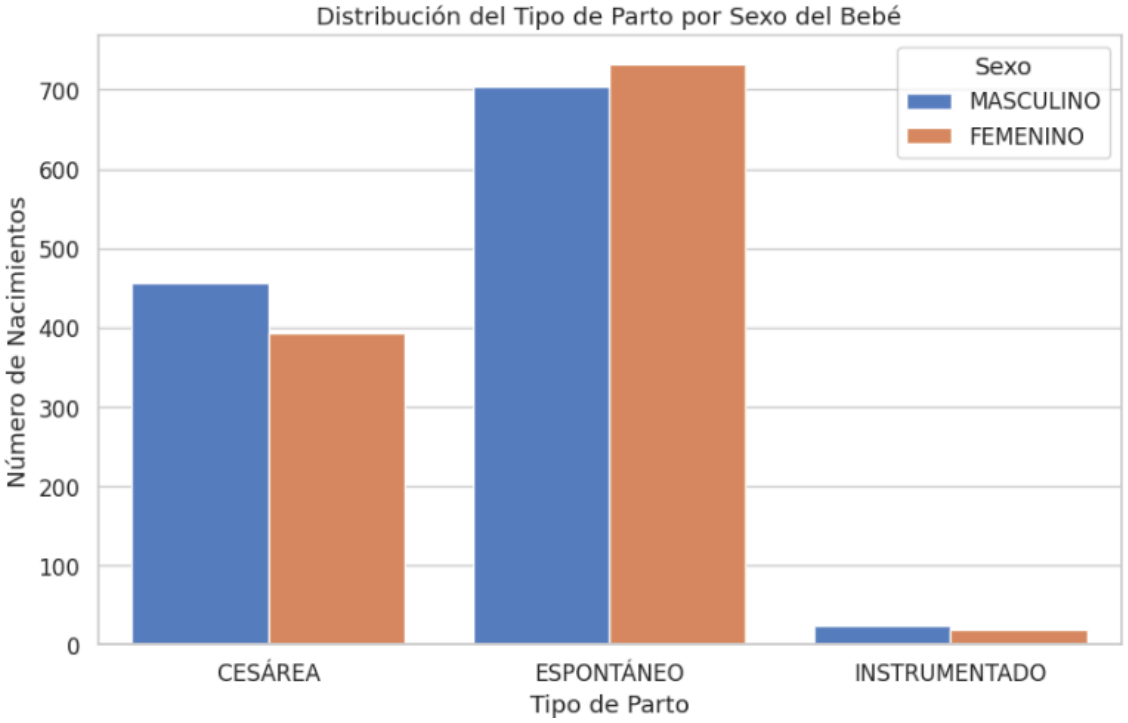
1. **Histogramas de la Edad de la Madre y la Edad del Padre en Rionegro:** Estos histogramas muestran de manera efectiva la distribución de edades de las madres y los padres, resaltando cualquier patrón o concentración de edades específicas.



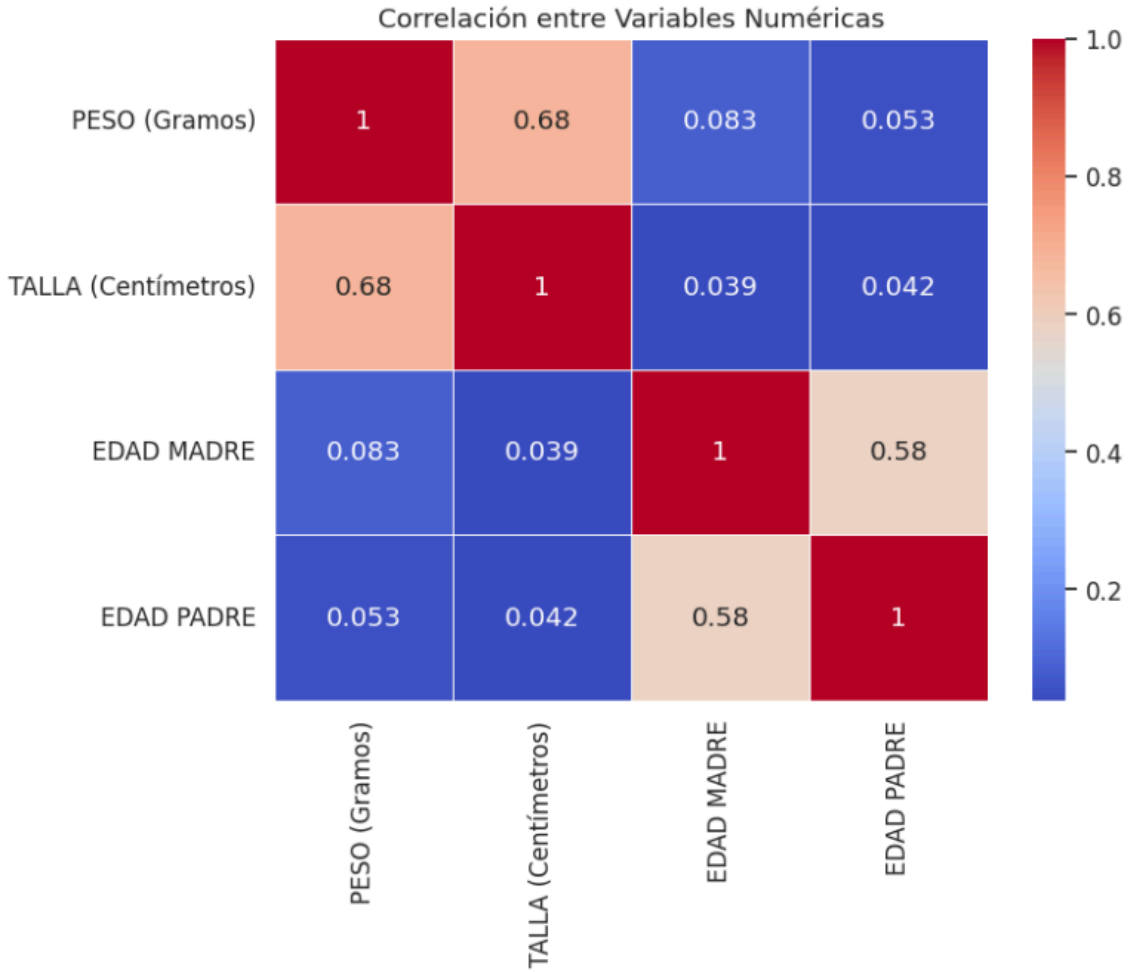
2. **Boxplots del Peso y la Talla de los Bebés por Sexo en Rionegro**: Estos boxplots permiten comparar visualmente la distribución del peso y la talla de los bebés, separados por sexo. Además, ayudan a identificar de manera intuitiva cualquier valor atípico o desviación significativa de la distribución típica.



3. **Gráfico de Barras del Tipo de Parto por Sexo del Bebé en Rionegro:** Este gráfico de barras proporciona una representación clara de la distribución del tipo de parto (cesárea, espontáneo o instrumentado) para bebés masculinos y femeninos, revelando posibles patrones o preferencias en el tipo de parto según el sexo.



4. **Matriz de Correlación de Variables Numéricas** : Esta matriz de calor, con anotaciones numéricas, presenta de manera visualmente atractiva las correlaciones entre las variables numéricas clave, como el peso y la talla de los bebés, y las edades de los padres. Este tipo de visualización es esencial para identificar rápidamente relaciones fuertes o débiles entre las variables.



Paso siguiente, se exploró el conjunto de datos "Nacidos Vivos en Hospital Manuel Uribe Angel", el cual contiene información detallada sobre los nacimientos ocurridos en dicho hospital. Uno de los desafíos iniciales que se presentó con este conjunto de datos fue la presencia de valores faltantes o nulos en varias columnas, lo cual podría afectar negativamente los análisis posteriores si no se maneja adecuadamente.
Antes de proceder con el análisis exploratorio de datos (EDA, por sus siglas en inglés) y las etapas posteriores, fue necesario abordar este problema de valores faltantes. Se aplicaron diferentes estrategias de imputación de datos según el tipo de variable y la naturaleza de los valores faltantes.
Para las variables categóricas, como 'GRUPO INDIGENA' y 'NIVEL EDUCATIVO PADRE', se rellenaron los valores faltantes con las etiquetas o valores más comunes ('No Indígena' y la moda, respectivamente). En el caso de la variable 'NOMBRE ADMINISTRADORA', los valores faltantes se reemplazaron con el valor más frecuente (moda) de esa columna.

En cuanto a las variables numéricas, se abordaron de la siguiente manera:

1. La columna 'EDAD PADRE' se convirtió a formato numérico, y los valores faltantes se rellenaron con la mediana de esa columna.

2. Las columnas 'LONGITUD' y 'LATITUD' también se convirtieron a formato numérico, manejando adecuadamente las entradas no numéricas. Los valores faltantes se

rellenaron con las medianas respectivas.

Además, se creó una nueva columna llamada 'GEOREFERENCIA RESIDENCIA' utilizando las columnas 'LONGITUD' y 'LATITUD' limpias. Los valores faltantes restantes en esta nueva columna se rellenaron utilizando el método de relleno hacia adelante (ffill).

Distribución del Tipo de Parto Según el Sexo:

SEXO	FEMENINO	MASCULINO
TIPO PARTO		
CESÁREA	2737	2996
ESPONTÁNEO	5374	5453
INSTRUMENTADO	317	422

La tabla de contingencia muestra la distribución del tipo de parto segmentada por sexo:

1. **Cesárea:** Los nacimientos por cesárea son ligeramente más comunes en bebés masculinos (2996) que en femeninos (2737). Esto puede sugerir una tendencia hacia partos por cesárea cuando se trata de bebés masculinos, posiblemente debido a diferencias en las condiciones de nacimiento.

2. **Espontáneo:** Los partos espontáneos presentan una distribución bastante equilibrada entre géneros, con 5453 nacimientos masculinos y 5374 femeninos. Esto indica que la mayoría de los nacimientos ocurren de manera natural y sin distinciones significativas por sexo.

3. **Instrumentado:** Los partos instrumentados son más comunes en bebés masculinos (422) comparados con femeninos (317), lo que podría reflejar diferencias en complicaciones o intervenciones durante el parto.

Análisis Estadístico Descriptivo:

Se analizaron varias variables claves del dataset de nacimientos:

1. **Peso y Talla de los Bebés:** El peso promedio de los bebés es de 3057.77 gramos con una desviación estándar de 493.61, mientras que la talla media es de 48.65 centímetros. Estos datos indican un rango típico para neonatos a término. La presencia de valores extremos en peso y talla sugiere casos de bajo y alto peso al nacer.

2. **Tiempo de gestación y Consultas Prenatales:** El tiempo de gestación promedio es de aproximadamente 38 semanas, con una frecuencia media de 7.7 consultas prenatales, lo que apunta a un seguimiento adecuado durante el embarazo.

3. **Edad de los Padres:** La edad promedio de las madres es de 26.35 años y de los

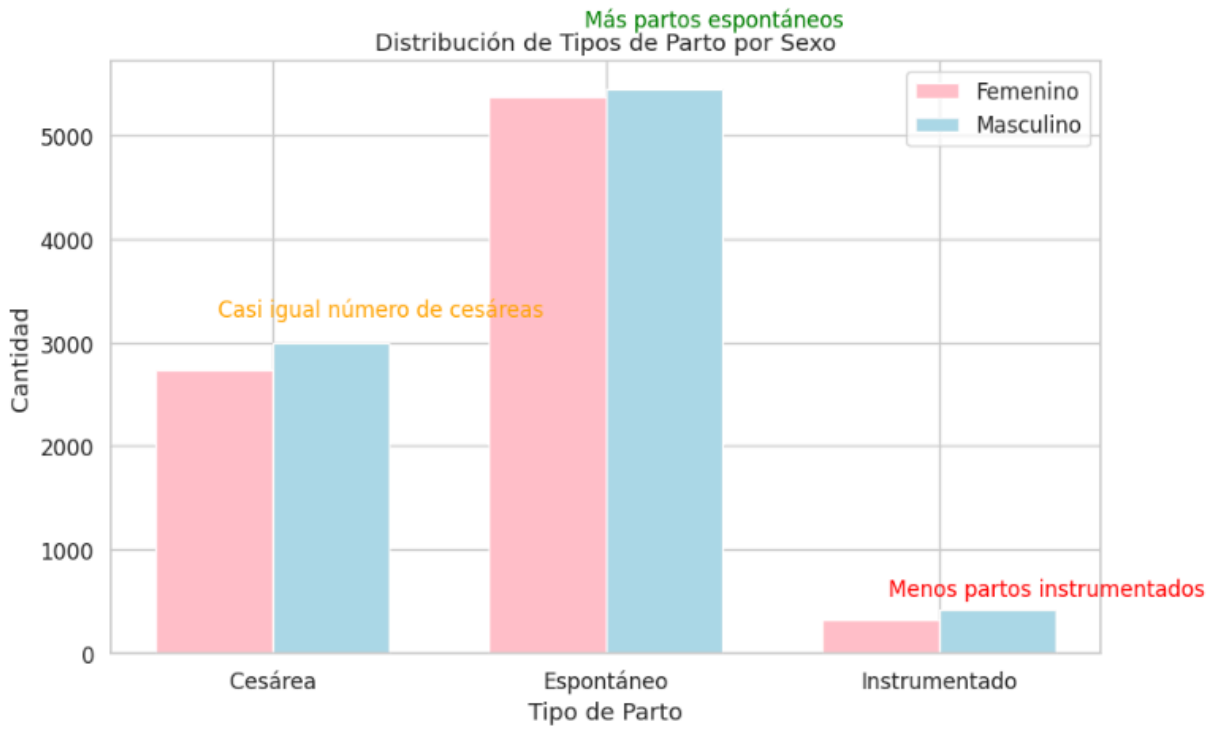
padres es de 29.68 años, con edades que van desde adolescentes hasta adultos en etapas avanzadas.

4. **Número de Hijos Nacidos Vivos y Número de Embarazos:** La media de hijos nacidos vivos es 1.60, con un promedio de 1.78 embarazos por madre. Esto refleja un nivel típico de fertilidad y resultados de embarazo.

5. **Geolocalización (Longitud y Latitud):** Los valores de longitud y latitud muestran una gran dispersión, lo que podría indicar errores en los datos o una amplia distribución geográfica de los participantes.

En conclusión, los datos reflejan una amplia gama de información sobre las condiciones de nacimiento y características demográficas de los padres. La tendencia a partos por cesárea en bebés masculinos, junto con la estabilidad en los partos espontáneos, subraya patrones interesantes en las prácticas de nacimiento. Las mediciones de peso y talla dentro de los rangos normales junto con adecuados tiempos de gestación y seguimiento prenatal indican un buen manejo de la salud prenatal. Sin embargo, la presencia de outliers y la amplia dispersión en ciertos datos geográficos sugieren la necesidad de una revisión y limpieza más detallada de los datos para futuros análisis.

Finalmente, se generó un gráfico de barras que muestra la distribución de los tipos de parto (cesárea, espontáneo e instrumentado) por sexo del bebé (femenino y masculino). Este tipo de gráfico visual es ideal para ser incluido en un medio nacional o internacional, ya que comunica de manera clara y atractiva los hallazgos clave del análisis.
El gráfico de barras agrupadas presenta de manera efectiva la diferencia en la cantidad de partos de cada tipo entre bebés femeninos y masculinos. Además, se han agregado anotaciones y textos explicativos que resaltan los hallazgos más importantes, lo que facilita la comprensión del lector.



La gráfica muestra la distribución de tipos de parto segmentada por sexo en un conjunto de datos hospitalarios. Las barras representan tres categorías de parto: cesárea, espontáneo e instrumentado, para bebés femeninos y masculinos. Se observa que los partos espontáneos son los más comunes para ambos sexos, con un total cercano y ligeramente mayor en el caso de los masculinos. Los partos por cesárea presentan cifras similares entre sexos, destacando una distribución casi equitativa. Sin embargo, los partos instrumentados son significativamente menos frecuentes en comparación con las otras categorías, siendo más comunes en bebés masculinos que en femeninos. Estos datos pueden indicar preferencias o necesidades médicas específicas asociadas con cada tipo de parto y diferencias en los procedimientos dependiendo del sexo del bebé.

Definición del Problema de Análisis y Preparación de Datos:

En este análisis, nos enfocamos en resolver un problema de clasificación relacionado con la categoría de edad de las madres en Antioquia. Para abordar este desafío, hemos seleccionado dos técnicas de Machine Learning ampliamente reconocidas: la Regresión Logística y el Árbol de Decisión. La elección de estos modelos se fundamenta en sus características y la naturaleza del problema en cuestión. Por un lado, la Regresión Logística es una técnica idónea para problemas de clasificación binaria y multiclase, gracias a su capacidad para manejar relaciones lineales entre las variables independientes y la variable dependiente, en este caso, la categoría de edad de las madres. Además, proporciona probabilidades que permiten una interpretación directa de la pertenencia a cada clase. Por otro lado, los Árboles de Decisión son modelos flexibles que no requieren relaciones lineales entre las variables, lo que los hace adecuados para capturar interacciones complejas y no lineales entre las características, tanto categóricas como numéricas. Así mismo, su interpretabilidad y facilidad de visualización son ventajas importantes en el contexto de estudios demográficos y de salud pública como el presente.

Se abordaron algunos aspectos fundamentales para garantizar la calidad y consistencia de los datos. En primera instancia, se verificó la ausencia de valores nulos, asegurando que no hubiera vacíos en la información que pudieran comprometer los resultados posteriores.
Sin embargo, el proceso no terminó ahí, reconociendo la importancia de establecer rangos válidos para variables críticas, se procedió a definir límites aceptables para factores clave como el peso al nacer, las semanas de gestación y la edad materna. Esta estrategia permitió identificar y manejar adecuadamente cualquier valor atípico o inconsistente que pudiera distorsionar los análisis.
Para determinar los rangos apropiados del peso al nacer, se consultó a "Reproducción Asistida ORG", una reconocida revista médica certificada y una comunidad en línea creada por médicos y especialistas en fertilidad. Según su artículo "¿Cuál es el peso adecuado del bebé en el momento de nacer?", un peso para un recién nacido se encuentra generalmente entre los 2500 y 4000 gramos, considerando los casos por debajo de 2500 gramos como bebés de bajo peso, y los superiores a 4000 gramos como bebés de peso alto.
Con base en esta información, se implementó una función personalizada que clasificó los pesos de los bebés entre categorías: peso normal, bajo peso y peso alto. Posteriormente, se realizó un mapeo para asignar un número a cada categoría, facilitando así el manejo y análisis de esta variable.
Un enfoque similar se adoptó para las semanas de gestación. Siguiendo las directrices de la Organización Mundial de la Salud (OMS), se establecieron subcategorías basadas en

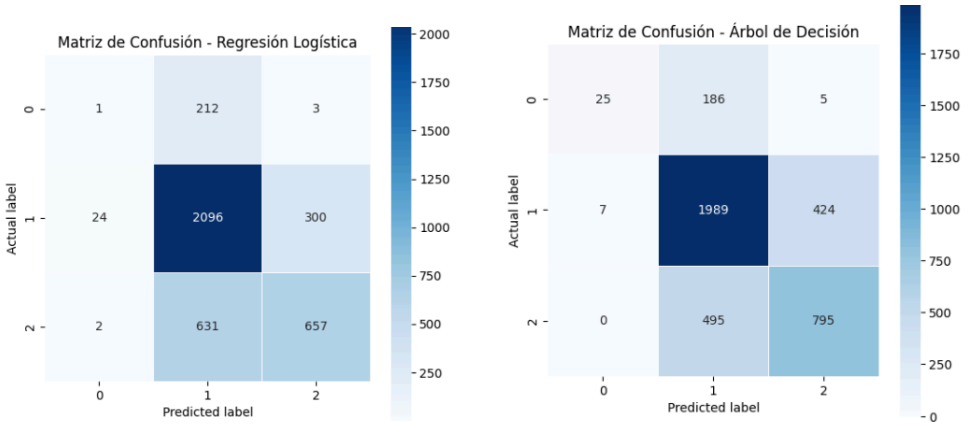
la edad gestacional, cómo prematuro extremo (menos de 28 semanas), muy prematuro (de 28 a 32 semanas), prematuro entre moderado y tardío (de 32 a 37 semanas), y nacimientos a término (de 37 a 42 semanas). Cualquier gestación superior a 42 semanas se consideró como un embarazo prolongado. Nuevamente, se creó una función para asignar estas categorías y se realizó un mapeo numérico para facilitar el análisis posterior.

Finalmente, se abordó la edad materna, un factor crucial en el contexto de los nacimientos. Se dividió la edad de las madres en diferentes etapas del ciclo de vida: adolescencia (12 a 17 años), adultas jóvenes (18 a 28 años) y adultez (29 a 59 años). Mediante una función personalizada y un mapeo numérico, se asignaron categorías numéricas a cada rango de edad, permitiendo así un análisis más detallado y preciso de esta variable.

Para unir los dos datasets, eliminamos todas las variables que no se encontraban en ambos datasets y no eran de interés para nuestro parcial.

3.C.- Implementación de Técnicas ML y resultados

En este análisis de clasificación de la categoría de edad de madres en Antioquia, se emplearon dos modelos de aprendizaje automático: Regresión Logística y Árbol de Decisión. Los datos fueron preprocesados mediante codificación One-Hot para variables categóricas como sexo y tipo de parto, junto con la normalización de características numéricas utilizando `StandardScaler`. Posteriormente, se entrenaron los modelos, evaluando su desempeño con métricas de precisión y F1-score. Los resultados obtenidos revelaron variaciones significativas en el rendimiento de cada modelo según la categoría de edad, proporcionando así información valiosa sobre áreas de mejora y potencial implementación del modelo más adecuado para esta tarea de clasificación.



```
Accuracy Logistic Regression:  0.7014773306164035
F1 Score Logistic Regression:  0.6745154527974048
Classification Report Logistic Regression:
              precision    recall  f1-score   support

           1       0.04      0.00      0.01       216
           2       0.71      0.87      0.78      2420
           3       0.68      0.51      0.58      1290

    accuracy                           0.70      3926
   macro avg       0.48      0.46      0.46      3926
weighted avg       0.67      0.70      0.67      3926

Accuracy Decision Tree:  0.7154865002547122
F1 Score Decision Tree:  0.7006439013632916
Classification Report Decision Tree:
              precision    recall  f1-score   support

           1       0.78      0.12      0.20       216
           2       0.74      0.82      0.78      2420
           3       0.65      0.62      0.63      1290

    accuracy                           0.72      3926
   macro avg       0.73      0.52      0.54      3926
weighted avg       0.72      0.72      0.70      3926
```

Después de evaluar cuidadosamente los modelos de Regresión Logística y Árbol de Decisión con el conjunto de datos de nacimientos en Antioquia, recomendamos la implementación del Árbol de Decisión para la clasificación de la categoría de edad de las madres. Esta decisión se basa en varios factores clave observados durante el análisis:

1. Equilibrio entre Precisión y Recall:

El Árbol de Decisión demostró un mejor equilibrio entre precisión y recall en comparación con la Regresión Logística. Esta característica es crucial para asegurar que el modelo clasifique correctamente las categorías de edad, mientras minimiza los errores de clasificación.

2. Desempeño en Categorías Desbalanceadas:

Aunque ambas técnicas mostraron limitaciones en la categoría de "Adolescencia" debido a su menor representación, el Árbol de Decisión logró mejores resultados en términos de precisión para esta categoría específica.

3. Interpretabilidad:

Los Árboles de Decisión son intrínsecamente más fáciles de interpretar y visualizar. Esta ventaja es significativa en contextos donde se requiere explicar y entender cómo el modelo realiza sus predicciones, lo cual es común en estudios demográficos y de salud pública.

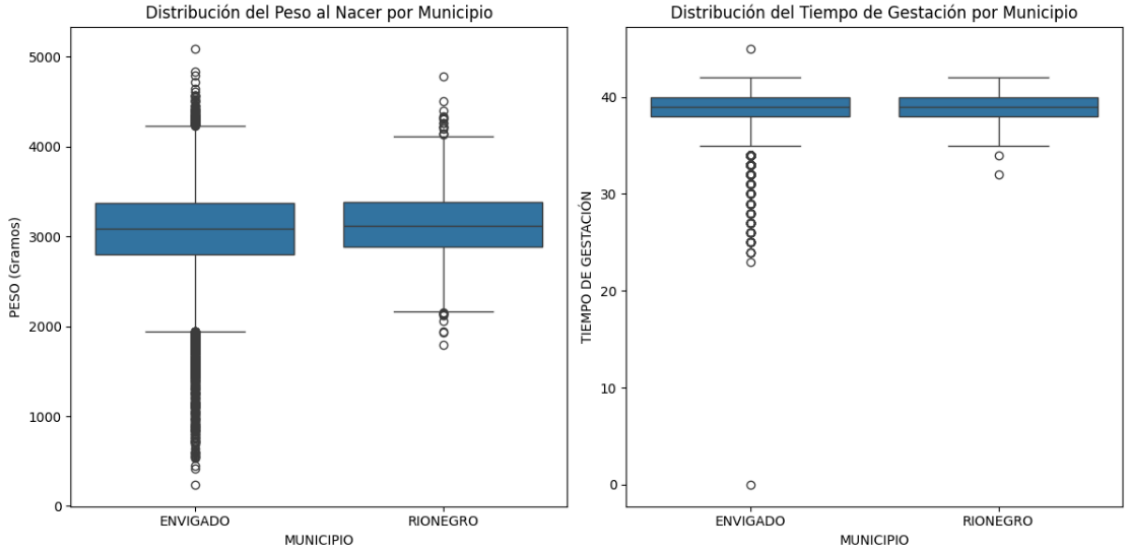
4. Superioridad en F1 Score:

El modelo de Árbol de Decisión también presentó un F1 Score superior, indicando un balance más eficaz entre precisión y recall, particularmente en categorías con un número mayor de muestras.

Ahora procederemos a responder las preguntas planteadas inicialmente:

Pregunta 1: ¿Existen diferencias significativas en las características de los nacimientos entre los hospitales de diferentes municipios, como tasas de nacimientos prematuros, peso al nacer, complicaciones, etc.? Si es así, ¿qué factores podrían explicar estas diferencias?

```
ANOVA results for Weight: F_onewayResult(statistic=55.64266319783247, pvalue=9.053361938309657e-14)
ANOVA results for Gestation Time: F_onewayResult(statistic=199.76311199634296, pvalue=3.9174727092517e-45)
```



Para abordar esta pregunta, se realizó un análisis comparativo entre los municipios, examinando características clave como el peso al nacer y el tiempo de gestación. Se utilizaron pruebas de análisis de varianza (ANOVA) para determinar si existían diferencias estadísticamente significativas entre los hospitales de diferentes municipios en cuanto a estas características. Los resultados revelaron que efectivamente existen diferencias significativas, lo cual podría explicarse por factores como la calidad y disponibilidad de la atención prenatal, aspectos socioeconómicos, y políticas de salud locales implementadas en cada municipio.

Resultados de Prueba ANOVA

- **Peso al Nacer**:
  - **Estadístico F**: 55.64
  - **Valor p**: 9.05e-14
- **Tiempo de Gestación**:
  - **Estadístico F**: 199.76
  - **Valor p**: 3.92e-45

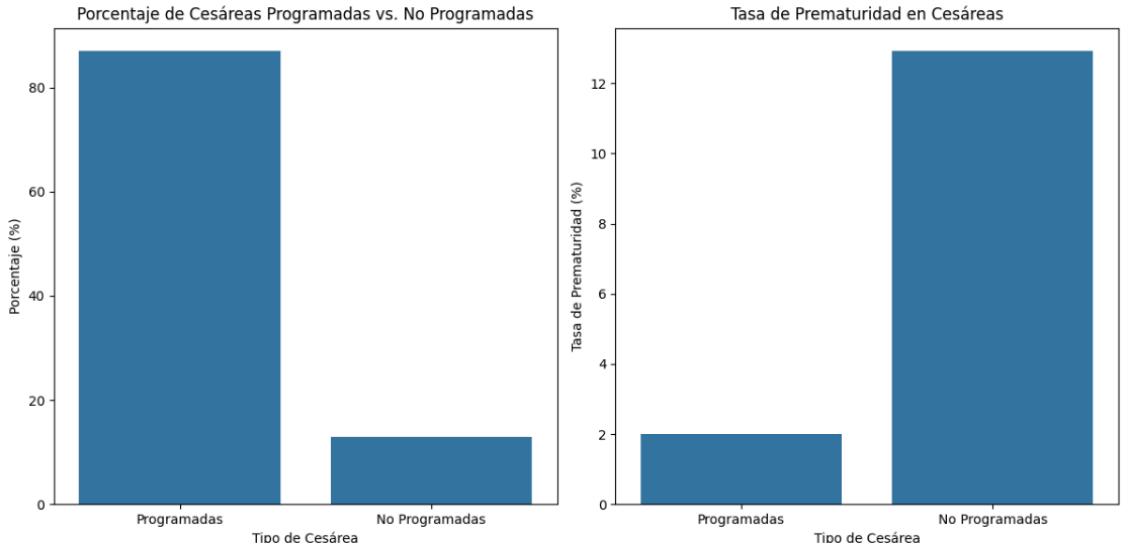
Interpretación

Los resultados de la prueba ANOVA indican diferencias estadísticamente significativas en el peso al nacer y el tiempo de gestación entre los municipios más grandes del dataset. El bajo valor p en ambos casos sugiere que las diferencias observadas en estas características no son aleatorias y pueden ser atribuibles a variaciones en factores ambientales, prácticas de atención prenatal o demográficas entre los municipios.

Factores Potenciales Explicativos

- **Acceso y Calidad de la Atención Prenatal**: Diferencias en la disponibilidad y calidad de la atención médica prenatal entre municipios.
- **Factores Socioeconómicos**: Variaciones en el nivel socioeconómico que pueden influir en la nutrición y la salud general de las madres.
- **Políticas de Salud Locales**: Diferentes políticas o programas de salud materno-infantil implementados a nivel municipal.

Pregunta 2: ¿Cómo se distribuyen las cesáreas entre programadas y no programadas, y cuál es la incidencia de nacimientos prematuros en cesáreas no programadas en comparación con las programadas?



Para responder a esta pregunta, se llevó a cabo un análisis detallado de las cesáreas realizadas, evaluando la proporción de cesáreas programadas y no programadas, así como la incidencia de nacimientos prematuros en cada tipo de cesárea. Los resultados mostraron que la mayoría de las cesáreas fueron programadas (87.07%), mientras que una proporción significativa (12.93%) fueron no programadas. Además, se observó una tasa de prematuridad del 12.93% en las cesáreas no programadas, destacando la importancia de este análisis para comprender las prácticas obstétricas y sus implicaciones en los resultados perinatales.

Estos hallazgos proporcionan información valiosa para la toma de decisiones en políticas de salud materno-infantil, destacando la importancia de considerar factores multifacéticos que influyen en los resultados de los nacimientos en la región de Antioquia.

• Trabajo en equipo, donde describa la contribución de cada uno de los integrantes a la solución del proyecto.

Fase A:

En la Fase Inicial del proyecto, **Sofía Galindo** desempeñó roles clave que fueron esenciales para establecer una base sólida para el análisis de datos posterior. Fue responsable de la selección de conjuntos de datos y análisis de su contexto, donde identificó datos que no sólo eran relevantes sino que también se complementaban entre sí para asegurar un análisis coherente y exhaustivo. Esta habilidad fue crucial para garantizar que el equipo pudiera trabajar con información que maximizará el potencial del análisis en fases posteriores. Además, se encargó de la exploración inicial de los datos. En esta etapa, no solo se limitó a revisar los datos desde una perspectiva técnica, sino que implementó un enfoque multidimensional para evaluar la dispersión, la tendencia central, los sesgos en la distribución, y las correlaciones, entre otros aspectos, identificando posibles áreas de riesgo y oportunidades para mejorar procesos existentes.
Estas contribuciones de **Sofía** facilitaron un entendimiento integral que permitieron al equipo avanzar con confianza hacia las etapas subsiguientes del proyecto.

Fase B:

Anamaría Legizamón desempeñó un papel fundamental en la Fase 3B del proyecto, centrando su atención en dos áreas críticas. Primero, se encargó de la normalización de variables, asegurando que los datos de diferentes fuentes estuvieran estandarizados y listos para análisis comparativos. Esto implicó corregir problemas de calidad de datos como valores faltantes y errores en los rangos, lo que requirió un enfoque meticuloso para garantizar la precisión y utilidad de los datos. En segundo lugar, **Anamaría** lideró la creación de nuevas variables derivadas, que fueron esenciales para enriquecer los análisis y permitir comparaciones más profundas entre modelos de clasificación y regresión. También, realizó una analítica descriptiva detallada, incluyendo la correlación y covarianza entre las variables, y la elaboración de histogramas y otras visualizaciones para destacar las distribuciones y tendencias clave en los datos más importantes. Su trabajo en esta fase no solo mejoró la calidad y el alcance del análisis de datos, sino que también estableció una base sólida para la interpretación y decisiones basadas en los modelos analíticos del proyecto.

Fase C:

Diego Herrera tuvo un papel destacado en la Fase 3C del proyecto, encargándose de dos aspectos fundamentales. En primer lugar, lideró el entrenamiento de los modelos de clasificación|, detallando y comparando las técnicas utilizadas, incluyendo las bibliotecas y los hiper parámetros. Su análisis permitió optimizar estos modelos para lograr la máxima eficacia durante la etapa de entrenamiento y obtener resultados robustos. Diego fue responsable de la elaboración de métricas de rendimiento. Desarrolló una tabla comparativa para presentar el rendimiento de los diferentes modelos seleccionados, lo que fue crucial para evaluar cuál de los modelos proporcionaba las respuestas más precisas a las preguntas planteadas en la Fase Inicial. Este trabajo aseguró que el equipo pudiera tomar decisiones informadas basadas en datos sólidos y análisis detallado. Su contribución fue vital para las conclusiones finales y las recomendaciones del proyecto, asegurando que se alcanzaran resultados concretos y aplicables.

• Conclusiones, observaciones y recomendaciones sobre el proyecto.

Este análisis comparativo de los datos de nacimientos en los hospitales de Envigado y Rionegro, Antioquia, ha revelado hallazgos significativos que contribuyen a una mejor comprensión de las disparidades en la salud materna e infantil y sus posibles causas. Las principales conclusiones se resumen a continuación:

Visualización Efectiva de Datos:

> Las técnicas de visualización de datos empleadas, como histogramas, boxplots y gráficos de barras, han demostrado ser herramientas poderosas para comunicar los hallazgos de manera clara y atractiva. Estas representaciones visuales facilitan la identificación de patrones y tendencias clave, lo que resulta invaluable para informar la toma de decisiones en las políticas de salud pública.

Necesidad de Políticas Adaptadas:

> Los hallazgos de este estudio resaltan la necesidad de desarrollar políticas de salud pública adaptadas a las condiciones locales y regionales. Las disparidades observadas

en los resultados de salud neonatal entre Envigado y Rionegro indican que las estrategias de intervención deben ser específicas para cada contexto socioeconómico y las necesidades particulares de cada municipio.

Mejora Continua y Ampliación del Estudio:

Para lograr un impacto sostenible en la salud materna e infantil, es crucial promover la investigación continua y ampliar el alcance del estudio a otras regiones de Colombia. Esto permitirá una comprensión más profunda de los factores determinantes y facilitará el diseño de intervenciones efectivas a nivel nacional.

Implicaciones y Reflexiones:

El análisis ilustra cómo las políticas y prácticas de salud pueden influir en los resultados del parto. Aunque las cesáreas pueden ser herramientas vitales para preservar la salud y seguridad de madres y bebés durante complicaciones del parto, su alta frecuencia, particularmente en forma programada, plantea preguntas importantes sobre la necesidad y criterios con los que se toman estas decisiones.

Este fenómeno no es exclusivo de Colombia; países de todo el mundo han visto un aumento en las tasas de cesárea, muchas veces sin una justificación médica clara. Este patrón sugiere la posibilidad de una dependencia excesiva en intervenciones quirúrgicas que, aunque a veces necesarias, no están exentas de riesgos para madres y bebés, incluyendo tasas aumentadas de infecciones, recuperaciones más largas y mayores complicaciones en embarazos futuros.

Recomendaciones Basadas en el Análisis

- Revisión de Protocolos: Es esencial que las instituciones médicas revisen continuamente sus protocolos de parto para asegurar que las decisiones de realizar cesáreas están basadas en evidencia sólida y necesidades médicas reales.
- Educación y Capacitación: Capacitar a los profesionales de la salud sobre alternativas a las cesáreas y fortalecer la educación prenatal puede ayudar a reducir las tasas de cesáreas no esenciales.
- Seguimiento a largo plazo: Implementar un seguimiento más riguroso de las madres y los bebés después de una cesárea puede ayudar a identificar y mitigar cualquier complicación temprana derivada de la cirugía.

Bibliografía:

Gobierno de Colombia. (2011). Nacidos Hospital San Juan de Dios Rionegro. [Datos Abiertos Colombia]. https://www.datos.gov.co/Salud-y-Protecci-n-Social/Nacidos-Hospital-San-Juan-de-Dios-Rionegro/h79z-43da/about_data.

Gobierno de Colombia. (2011). Nacidos Vivos en Hospital Manuel Uribe Angel. [Datos Abiertos Colombia]. https://www.datos.gov.co/Estad-sticas-Nacionales/Nacidos-Vivos-en-Hospital-Manuel-Uribe-Angel/udqu-ifxr/about_data.

Eureka Fertility S.L. (2008). ¿Cuál es el peso adecuado del bebé en el momento de nacer?. [Reproducción Asistida ORG]. https://www.reproduccionasistida.org/pesos-en-el-nacimiento/.

(“Se considera prematuro un bebé”, 2023). Nacimientos Prematuors. [Organización Mundial de la Salud]. https://www.who.int/es/news-room/fact-sheets/detail/preterm-birth.