Parcial II Procesamiento de Datos

Pontificia Universidad Javeriana

Procesamiento de Datos a Gran Escala

Juan David López Becerra

Nicolas Samuel Martin Vasquez

Juan Diego Gonzalez

13/05/2024

Descripción

Para realizar un análisis comparativo detallado sobre los patrones de nacimiento entre dos ciudades colombianas con contextos demográficos distintos, se escogieron dos datasets representativos de Bucaramanga y Río Negro. La elección de estos conjuntos de datos responde a la intención de comparar un dataset grande con uno más pequeño, lo que proporciona una oportunidad única para explorar cómo las tendencias y los patrones pueden variar entre muestras de diferentes tamaños y cómo esto puede influir en las conclusiones estadísticas que se pueden derivar.

Bucaramanga, siendo una ciudad más grande y un importante centro urbano, contrasta con Río Negro que, aunque también significativo, es menor en tamaño y población. Analizar estos dos contextos permite entender mejor cómo los factores locales pueden influir en aspectos como la salud materna, prácticas de nacimiento y demografía. Los datasets incluyen variables críticas como el peso al nacer, la talla, el tiempo de gestación y las consultas prenatales, entre otros, permitiendo revelar patrones importantes y posibles áreas de intervención para políticas de salud pública.

La combinación de estos datasets en un solo marco de datos unificado facilita la comparación directa y el análisis conjunto, enriqueciendo la comprensión de las dinámicas específicas de cada área. Este enfoque no solo contribuye al diseño y la implementación de medidas de salud pública más efectivas y personalizadas, sino que también permite aplicar técnicas avanzadas de análisis de datos para descubrir insights que podrían no ser evidentes a simple vista.

Dataframe Rio Negro

Atributos Dataset

Se hace una breve explicación de cada uno:

1. **DEPARTAMENTO:** Nombre del departamento donde se registra el nacimiento.
2. **MUNICIPIO:** Nombre del municipio donde ocurre el nacimiento.
3. **ÁREA NACIMIENTO:** Área donde ocurre el nacimiento (urbano o rural).
4. **SEXO:** Sexo del recién nacido.
5. **PESO (Gramos):** Peso del bebé al nacer, expresado en gramos.
6. **TALLA (Centímetros):** Estatura del bebé al nacer, expresada en centímetros.
7. **FECHA NACIMIENTO:** Fecha exacta del nacimiento.
8. **TIEMPO DE GESTACIÓN**: Duración del embarazo expresada en semanas.

9. **NÚMERO CONSULTAS PRENATALES:** Cantidad de consultas médicas prenatales asistidas por la madre durante el embarazo.
10. **TIPO PARTO:** Método por el cual se realizó el parto (natural, cesárea, etc.).
11. **MULTIPLICIDAD EMBARAZO:** Indica si fue un embarazo simple o múltiple (gemelos, trillizos, etc.).
12. **EDAD MADRE:** Edad de la madre al momento del parto.
13. **EDAD PADRE:** Edad del padre al momento del parto.

Dataframe Bucaramanga

Atributos Dataset

Se hace una breve explicación de cada uno:

- **DEPARTAMENTO:** Departamento que reporta el nacido vivo.
- **MUNICIPIO:** Municipio que reporta el nacido vivo.
- **ÁREA NACIMIENTO:** Área donde nació la persona: Urbana o Rural.
- **SITIO NACIMIENTO: Sitio de nacimiento:** INSTITUCIÓN DE SALUD, OTRO SITIO O EL DOMICILIO.
- **CODIGO INSTITUCION:** Código de la institución IPS.
- **SEXO:** Características fisiológicas del nacido vivo: MASCULINO, FEMENINO, Sin información.
- **PESO (Gramos):** Primera medida de peso del recién nacido, tomada preferiblemente dentro de la primera hora de vida.
- **TALLA (Centímetros):** Estatura del recién nacido, medida desde la planta del pie hasta el vértice de la cabeza.
- **HORA NACIMIENTO:** Hora de nacimiento del nacido vivo.
- **PARTO ATENDIDO POR:** Parto atendido por: MÉDICO, PROMOTOR(A) DE SALUD, ENFERMERO (A), OTRA PERSONA, PARTERA O AUXILIAR DE ENFERMERÍA.
- **TIEMPO DE GESTACIÓN (Semanas):** Duración del embarazo en semanas.
- **NÚMERO CONSULTAS PRENATALES:** Número de consultas prenatales.
- **TIPO PARTO:** Tipo de parto experimentado**.**
- **MULTIPLICIDAD EMBARAZO:** Multiplicidad del embarazo de la madre.
- **APGAR1:** APGAR1 del nacido vivo.
- **APGAR2:** APGAR2 del nacido vivo.
- **GRUPO SANGUÍNEO:** Grupo sanguíneo del feto o recién nacido.
- **FACTOR RH:** Factor RH del feto o recién nacido.
- **PERTENENCIA ÉTNICA:** Pertenencia étnica del feto o recién nacido.
- **GRUPO INDIGENA:** Grupo indígena del nacido vivo.
- **EDAD MADRE:** Edad de la madre.
- **ESTADO CONYUGAL MADRE:** Estado conyugal de la madre.
- **NIVEL EDUCATIVO MADRE:** Último nivel educativo alcanzado por la madre.

- **ÚLTIMO AÑO APROBADO MADRE:** Último año cursado por la madre en su nivel educativo.
- **PAÍS RESIDENCIA:** País de residencia del nacido vivo.
- **departamento_residencia:** Departamento de residencia del nacido vivo.
- **MUNICIPIO RESIDENCIA:** Municipio de residencia del nacido vivo.
- **ÁREA RESIDENCIA:** Área de residencia del nacido vivo.
- **LOCALIDAD:** Localidad del nacido vivo**.**
- **BARRIO:** Barrio de residencia de la madre gestante.
- **CENTRO POBLADO:** Centro poblado del nacido vivo.
- **RURAL DISPERSO:** Rural disperso del nacido vivo.
- **NÚMERO HIJOS NACIDOS VIVOS:** Número total de hijos nacidos vivos durante las edades reproductivas de la madre.
- **FECHA ANTERIOR HIJO NACIDO VIVO:** Fecha del anterior hijo nacido vivo.
- **NÚMERO EMBARAZOS:** Número de embarazos que ha tenido la madre gestante.
- **RÉGIMEN SEGURIDAD:** Tipo de régimen de seguridad: SUBSIDIADO, CONTRIBUTIVO, NO ASEGURADO, EXCEPCIÓN y ESPECIAL.
- **NOMBRE ADMINISTRADORA:** Nombre de la administradora del nacido vivo.
- **EDAD PADRE:** Edad del padre.
- **NIVEL EDUCATIVO PADRE:** Último nivel educativo alcanzado por el padre.
- **ÚLTIMO AÑO APROBADO PADRE:** Último año cursado por el padre en su nivel educativo.
- **COMUNA:** Subdivisión administrativa menor, equivalente al municipio o concejo.
- **NOMCOMUNA:** Nombre de la comuna.
- **NUM NOMBCOMUNA:** Número asociado a la comuna.
- **GRUPO EDAD MADRE:** Clasificación de los grupos etarios según el Dane y la OMS para la madre.
- **CURSO DE VIDA MADRE:** Enfoque que aborda los momentos continuos de la vida, reconociendo la interacción de diferentes factores a lo largo del curso de la vida**.**
- **GRUPO EDAD PADRE:** Clasificación de los grupos etarios según el Dane y la OMS para el padre.
- **CURSO DE VIDA PADRE:** Enfoque que aborda los momentos del continuo de la vida del padre.
- **AÑO:** Año de nacimiento del nacido vivo.
- **BARRIO_VER:** Barrio verificado del nacido vivo.
- **ciudad_geo:** Ciudad georreferenciada del nacido vivo.
- **ORDEN: Consecutivo autonumérico.**
- **DÍA SEMANA: Día de la semana en que nació la persona.**

- **MES: Mes en que nació la persona.**

Unión dataset

```
# Renombrar las columnas en df para que coincidan con los nombres en dff
df_renamed = df.rename(columns={
    'TIEMPO DE GESTACIÓN (Semanas)': 'TIEMPO DE GESTACIÓN',
    'NUMERO CONSULTAS PRENATALES': 'NÚMERO CONSULTAS PRENATALES'
})

# Definir las columnas comunes a ambos DataFrames
columns_common = [
    'DEPARTAMENTO', 'MUNICIPIO', 'AREA NACIMIENTO', 'SEXO', 'PESO (Gramos)',
    'TALLA (Centímetros)', 'TIEMPO DE GESTACIÓN', 'NÚMERO CONSULTAS PRENATALES',
    'TIPO PARTO', 'MULTIPLICIDAD EMBARAZO', 'EDAD MADRE', 'EDAD PADRE'
]

# Seleccionar solo las columnas comunes en ambos DataFrames
df_final1 = df_renamed[columns_common]
dff_final1 = dff[columns_common]

# Concatenar los DataFrames
newdf = pd.concat([df_final1, dff_final1], ignore_index=True)

# Verificar los resultados
print("Dimensiones del DataFrame final:", newdf.shape)
print(newdf.head())
```

Código 1. código para concatenar los atributos similares

Tablas del data frame concatenado

```
Tipos de datos de las columnas después de la corrección:
DEPARTAMENTO                 object
MUNICIPIO                    object
AREA NACIMIENTO              object
SEXO                         object
PESO (Gramos)               float64
TALLA (Centímetros)         float64
TIEMPO DE GESTACIÓN         float64
NÚMERO CONSULTAS PRENATALES float64
TIPO PARTO                   object
MULTIPLICIDAD EMBARAZO       object
EDAD MADRE                  float64
EDAD PADRE                  float64
```

Tabla 1. Info atributos data frame concatenado

```
Valores nulos por columna después de la imputación:
DEPARTAMENTO                  0
MUNICIPIO                     0
AREA NACIMIENTO               0
SEXO                          0
PESO (Gramos)                 0
TALLA (Centímetros)           0
TIEMPO DE GESTACIÓN           0
NÚMERO CONSULTAS PRENATALES   0
TIPO PARTO                    2
MULTIPLICIDAD EMBARAZO        2
EDAD MADRE                    0
EDAD PADRE                    0
dtype: int64
```

Tabla 2. Verificación limpieza data frame concatenado

Preguntas

Pregunta 1: ¿Cómo cambia la salud de los bebés entre Bucaramanga y Rionegro?

Importancia: Comparar los indicadores de salud neonatal entre Bucaramanga y Rio Negro es esencial para identificar diferencias en la atención medica y condiciones socioeconómicas que afectan a los recién nacidos. Esta comparación ayudará a desarrollar estrategias específicas para mejorar la atención prenatal y neonatal en cada región, y a implementar intervenciones localizadas que optimicen los resultados de salud de los recién nacidos y mejoren los procesos de atención materna e infantil

Pregunta 2: ¿Qué factores están asociados con los nacimientos prematuros (período de gestación menor a 37 semanas) en Bucaramanga y Río Negro?

Importancia: Identificar los factores asociados con los nacimientos prematuros en Bucaramanga y Río Negro es crucial para reducir la morbilidad y mortalidad neonatal. Conocer estos factores permitirá diseñar intervenciones específicas y asignar recursos de manera más efectiva para apoyar a las mujeres embarazadas de alto riesgo. Además, los hallazgos pueden utilizarse para educar a las futuras madres sobre los riesgos y las medidas preventivas asociadas con los nacimientos prematuros, mejorando la salud neonatal y los resultados de los embarazos en ambas regiones

Exploración de los datos y visualizaciones

Df De Bucaramanga

Tabla de información de atributos

#	Column	Non-Null Count	Dtype
0	DEPARTAMENTO	65577 non-null	object
1	MUNICIPIO	65577 non-null	object
2	AREA NACIMIENTO	0 non-null	float64
3	SITIO NACIMIENTO	65575 non-null	object
4	CODIGO INSTITUCION	65549 non-null	float64
5	SEXO	0 non-null	float64
6	PESO (Gramos)	65570 non-null	float64
7	TALLA (Centímetros)	65570 non-null	float64
8	HORA NACIMIENTO	65574 non-null	object
9	PARTO ATENDIDO POR	65575 non-null	object
10	TIEMPO DE GESTACIÓN (Semanas)	65567 non-null	float64
11	NUMERO CONSULTAS PRENATALES	65567 non-null	float64
12	TIPO PARTO	65575 non-null	object
13	MULTIPLICIDAD EMBARAZO	65575 non-null	object
14	APGAR1	65504 non-null	float64
15	APGAR2	65504 non-null	float64
16	GRUPO SANGUINEO	65430 non-null	object
17	FACTOR RH	0 non-null	float64
18	PERTENENCIA ÉTNICA	65575 non-null	object
19	GRUPO INDIGENA	20 non-null	object
20	EDAD MADRE	65349 non-null	float64
21	ESTADO CONYUGAL MADRE	65349 non-null	object
22	NIVEL EDUCATIVO MADRE	65191 non-null	object
23	ULTIMO AÑO APROBADO MADRE	63970 non-null	float64
24	PAIS RESIDENCIA	65191 non-null	object
25	departamento_residencia	65191 non-null	object
26	MUNICIPIO RESIDENCIA	65191 non-null	object
27	AREA RESIDENCIA	0 non-null	float64
28	LOCALIDAD	22271 non-null	object
29	BARRIO	63183 non-null	object
30	CENTRO POBLADO	0 non-null	float64
31	RURAL DISPERSO	1736 non-null	object
32	NÚMERO HIJOS NACIDOS VIVOS	64964 non-null	float64
33	FECHA ANTERIOR HIJO NACIDO VIVO	34296 non-null	object
34	NÚMERO EMBARAZOS	64964 non-null	float64
35	RÉGIMEN SEGURIDAD	64964 non-null	object
36	NOMBRE ADMINISTRADORA	60651 non-null	object
37	EDAD PADRE	64780 non-null	float64
38	NIVEL EDUCATIVO PADRE	64955 non-null	object
39	ULTIMO AÑO APROBADO PADRE	60729 non-null	float64
40	COMUNA	64964 non-null	object
41	NOMCOMUNA	47908 non-null	object
42	NUM NOMBCOMUNA	47951 non-null	object
43	GRUPO EDAD MADRE	64964 non-null	object
44	CURSO DE VIDA MADRE	64964 non-null	object
45	GRUPO EDAD PADRE	64780 non-null	object
46	CURSO DE VIDA PADRE	64780 non-null	object
47	AÑO	64964 non-null	float64
48	BARRIO_VER	64964 non-null	object
49	ciudad_geo	46594 non-null	object
50	ORDEN	57904 non-null	float64
51	DIA SEMANA	64964 non-null	object
52	MES	64964 non-null	object

dtypes: float64(20), object(33)

Imagen 1. información atributos

Gráfica nulos y no nulos en los atributos

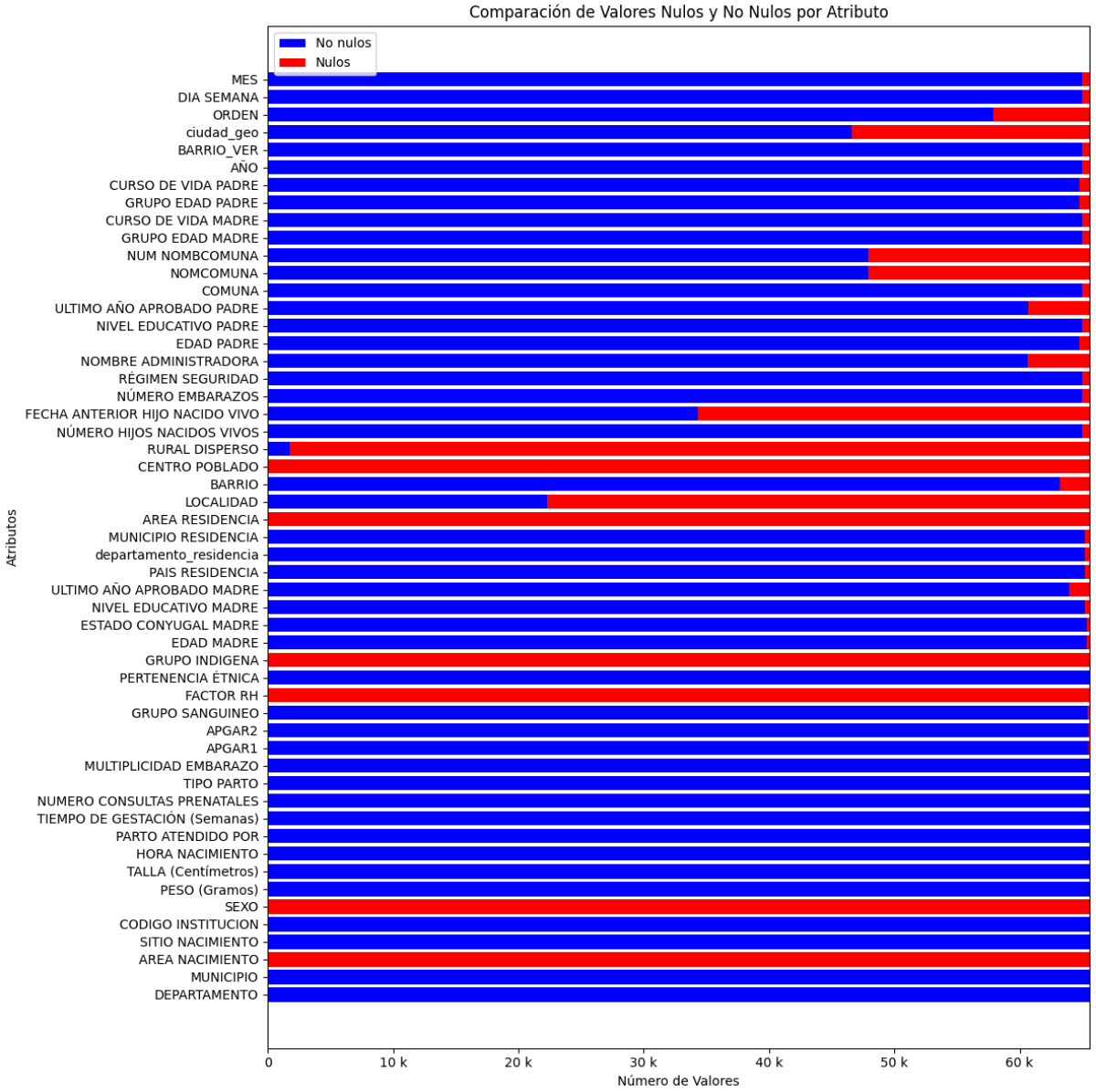

Imagen 2. Valores Nulos vs no Nulos

Mapa de correlación entre todas las variables numéricas

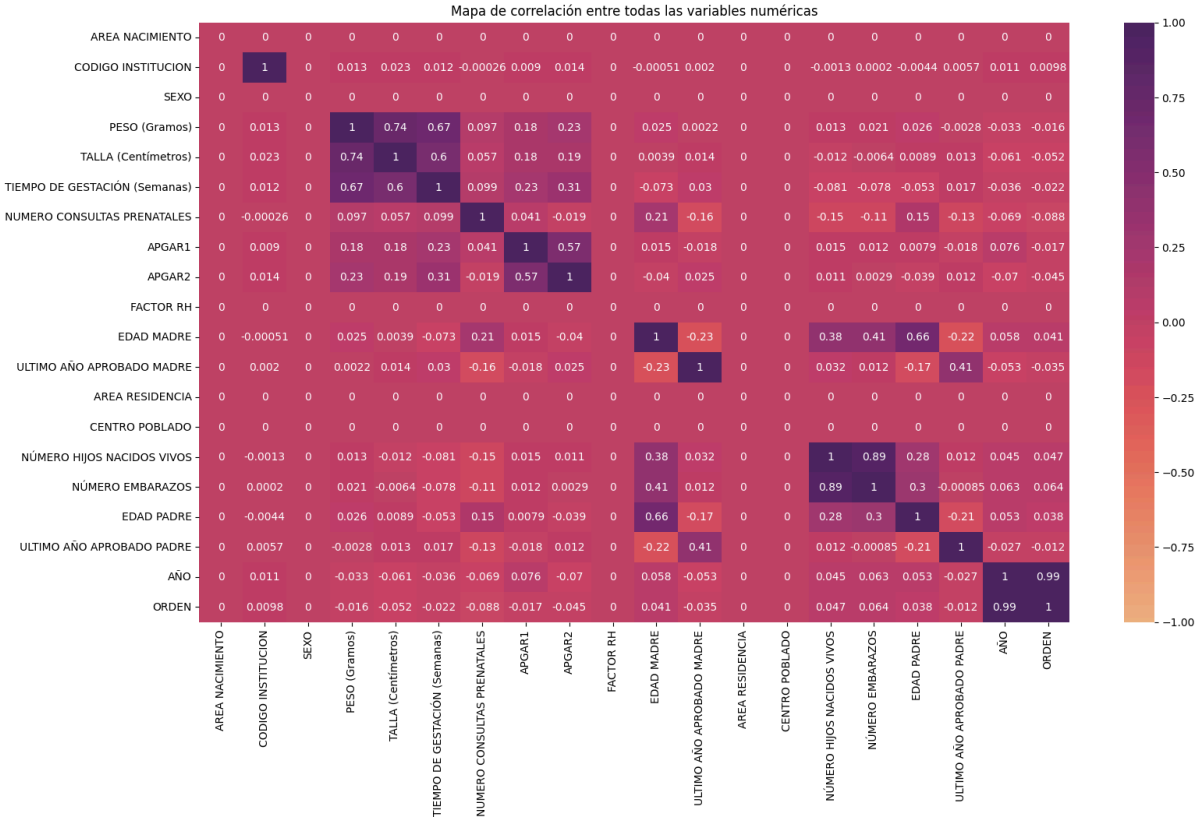

Imagen 3. Mapa de correlaciones

Gráfica del Peso en Gramos

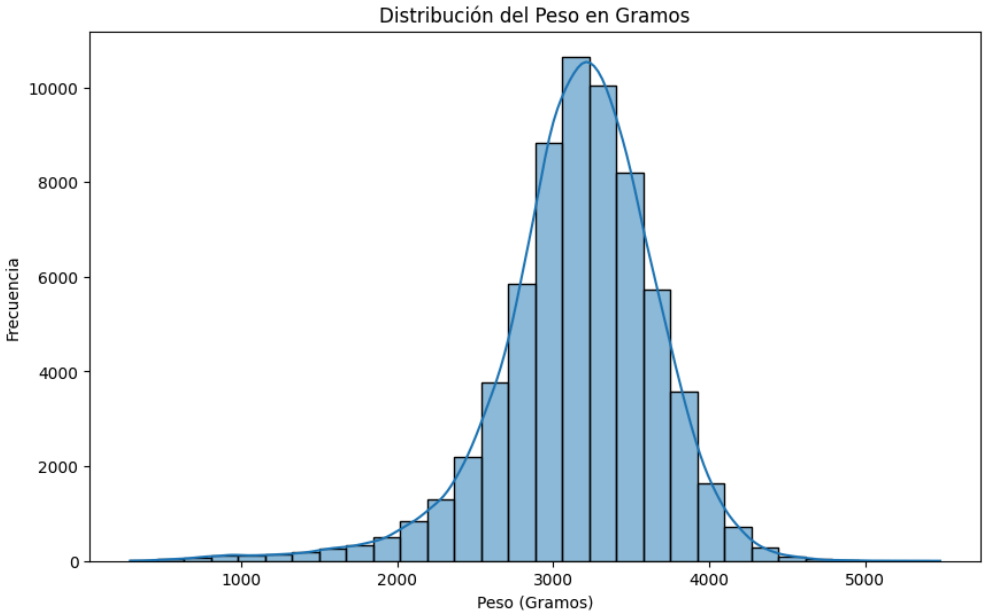

Imagen 4. Gráfico de barras

Gráfica de Sexo por Grupo Sanguíneo

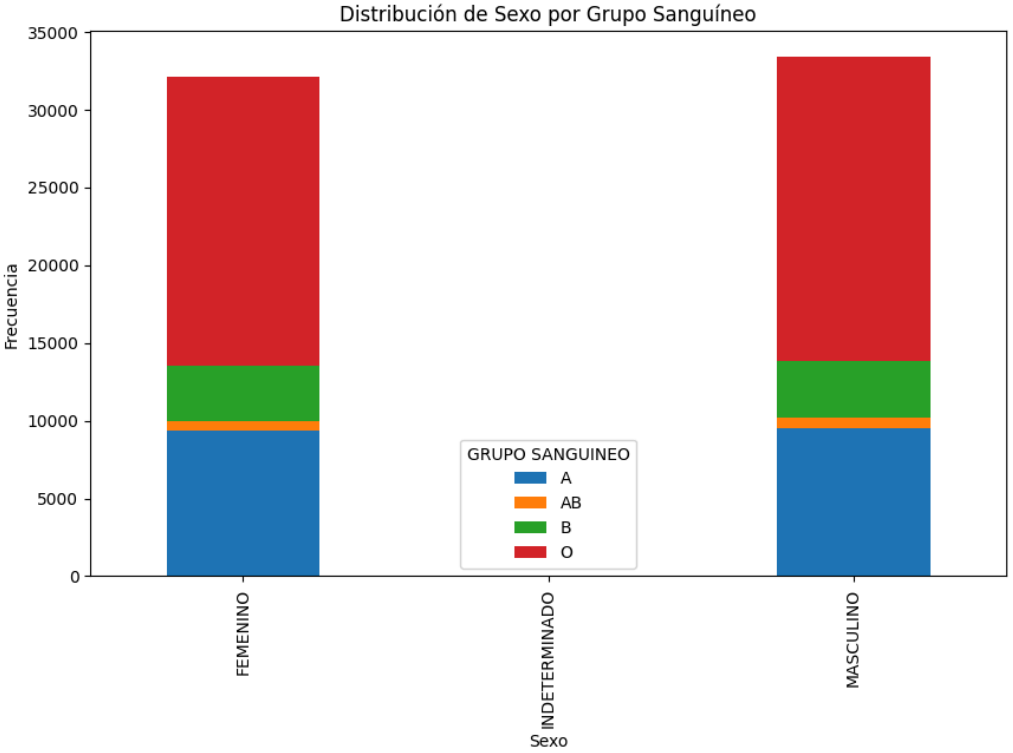

Imagen 5. Gráfico de barras correlacionado

Gráfico Pair plot de algunos atributos relacionados

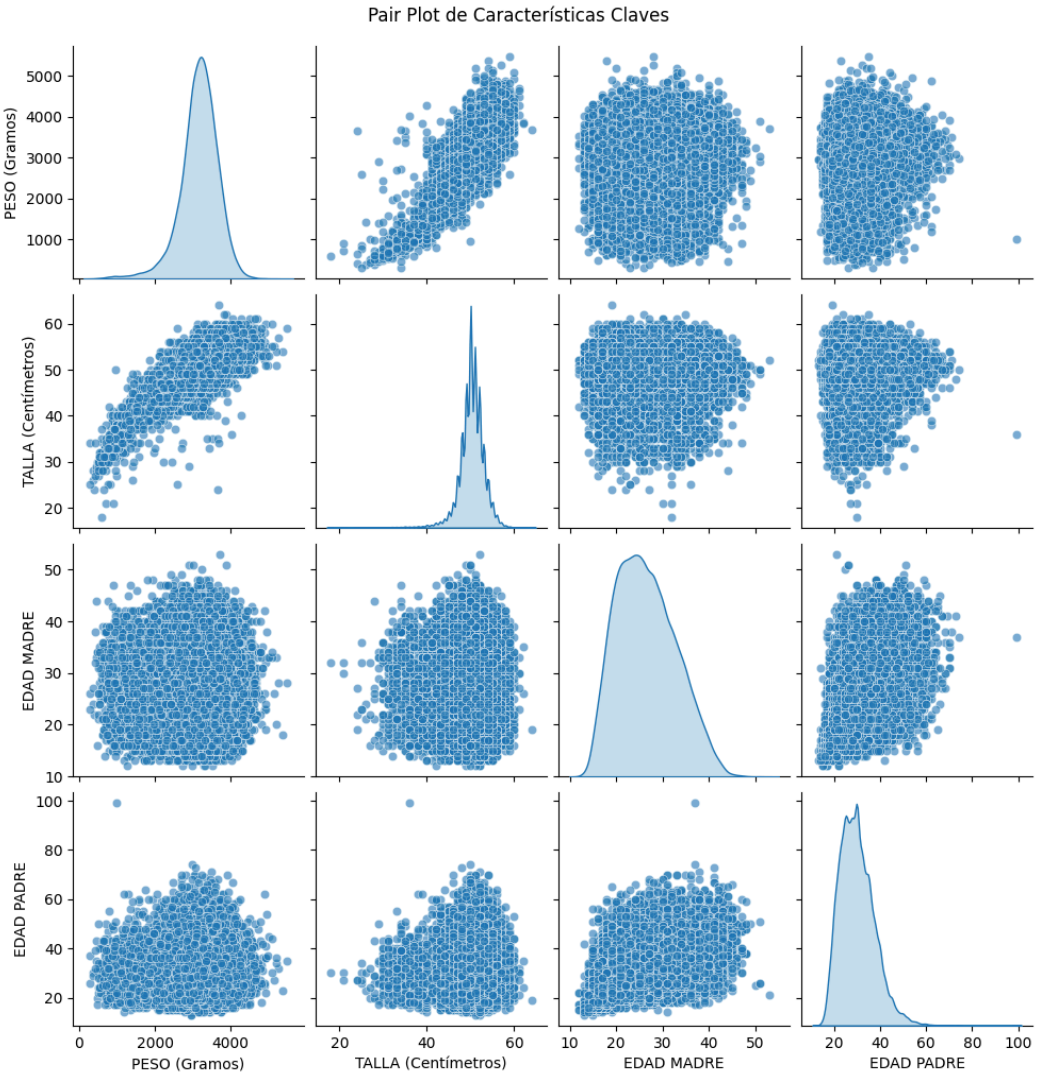

Imagen 6. Gráfico pair plot

Gráfica de Edades de Padres y Madres

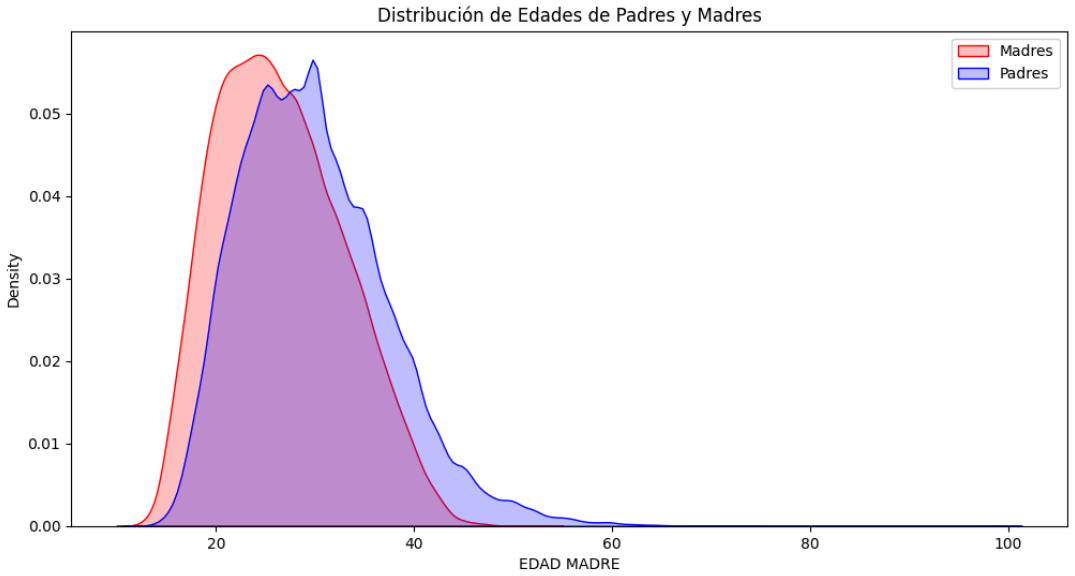

Imagen 7. Gráfico de comparación madre y padres

Gráfica de Nacimientos por Sexo y Mes

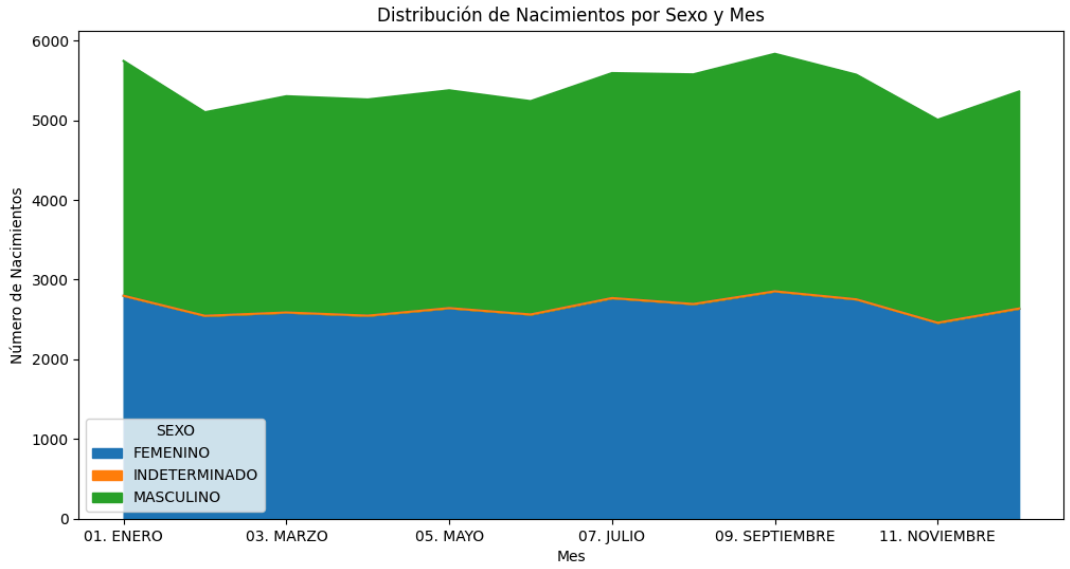

Imagen 8. Gráfico de Distribución de Nacimientos por Sexo y Mes

Df De Rio Negro

Información atributos dataset

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   DEPARTAMENTO                  2074 non-null   object
 1   MUNICIPIO                     2074 non-null   object
 2   AREA NACIMIENTO               2074 non-null   object
 3   SEXO                          2074 non-null   object
 4   PESO (Gramos)                 2074 non-null   int64
 5   TALLA (Centímetros)           2074 non-null   int64
 6   FECHA NACIMIENTO              2074 non-null   object
 7   TIEMPO DE GESTACIÓN           2074 non-null   int64
 8   NÚMERO CONSULTAS PRENATALES   2074 non-null   int64
 9   TIPO PARTO                    2074 non-null   object
 10  MULTIPLICIDAD EMBARAZO        2074 non-null   object
 11  EDAD MADRE                    2074 non-null   object
 12  EDAD PADRE                    2074 non-null   object
dtypes: int64(4), object(9)
```

Imagen 9. Tabla de info

Tabla de contingencia para Tipo de Parto vs Área de Nacimiento

AREA NACIMIENTO	CABECERA MUNICIPAL	RURAL DISPERSO
TIPO PARTO		
CESÁREA	760	3
ESPONTÁNEO	1283	5
INSTRUMENTADO	23	0

Imagen 10. Tabla de contingencia

Gráfico para visualizar datos nulos y no nulos en los atributos

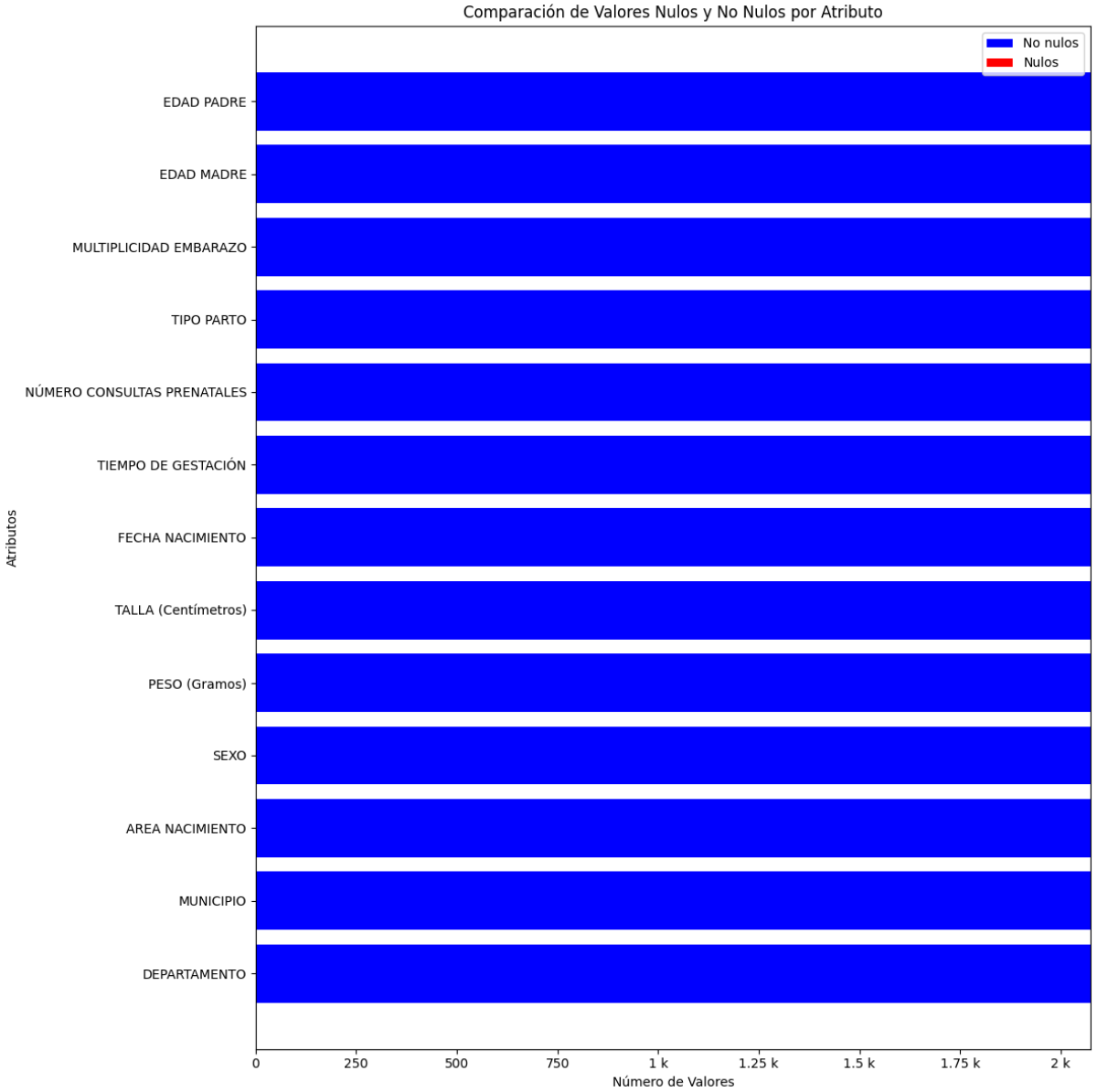

Imagen 11. Gráfico nulos vs no nulos

Mapa de correlación entre todas las variables numéricas

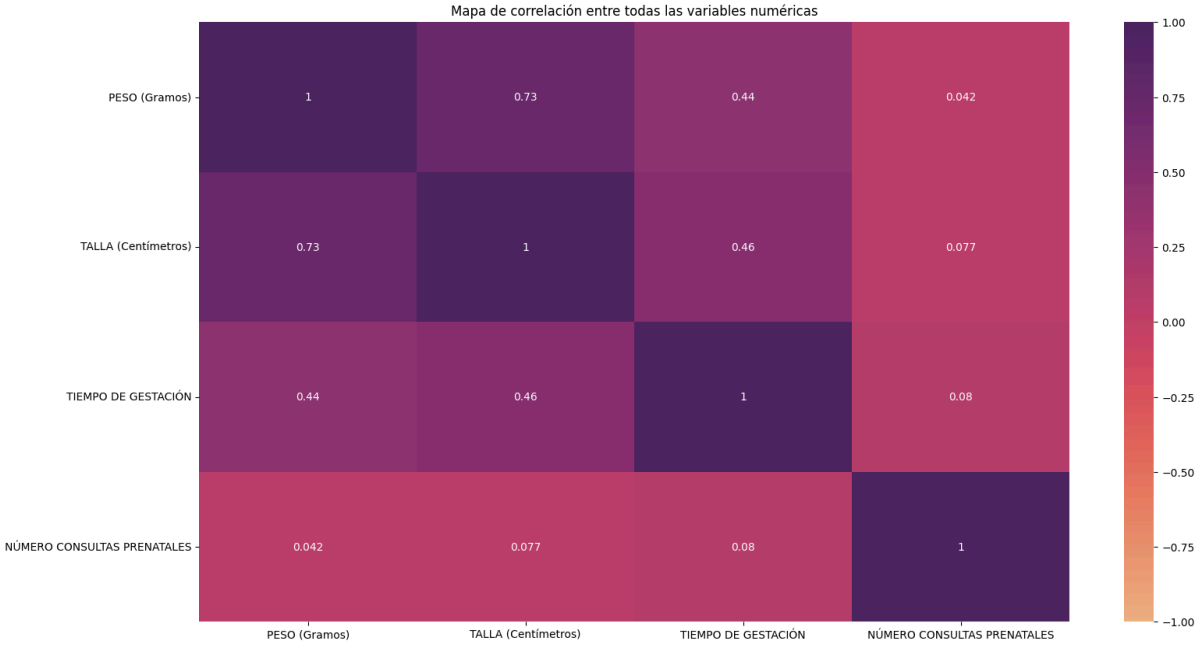

Imagen 12. Matriz de correlación

Grafico Distribución del Peso al Nacer en Río Negro

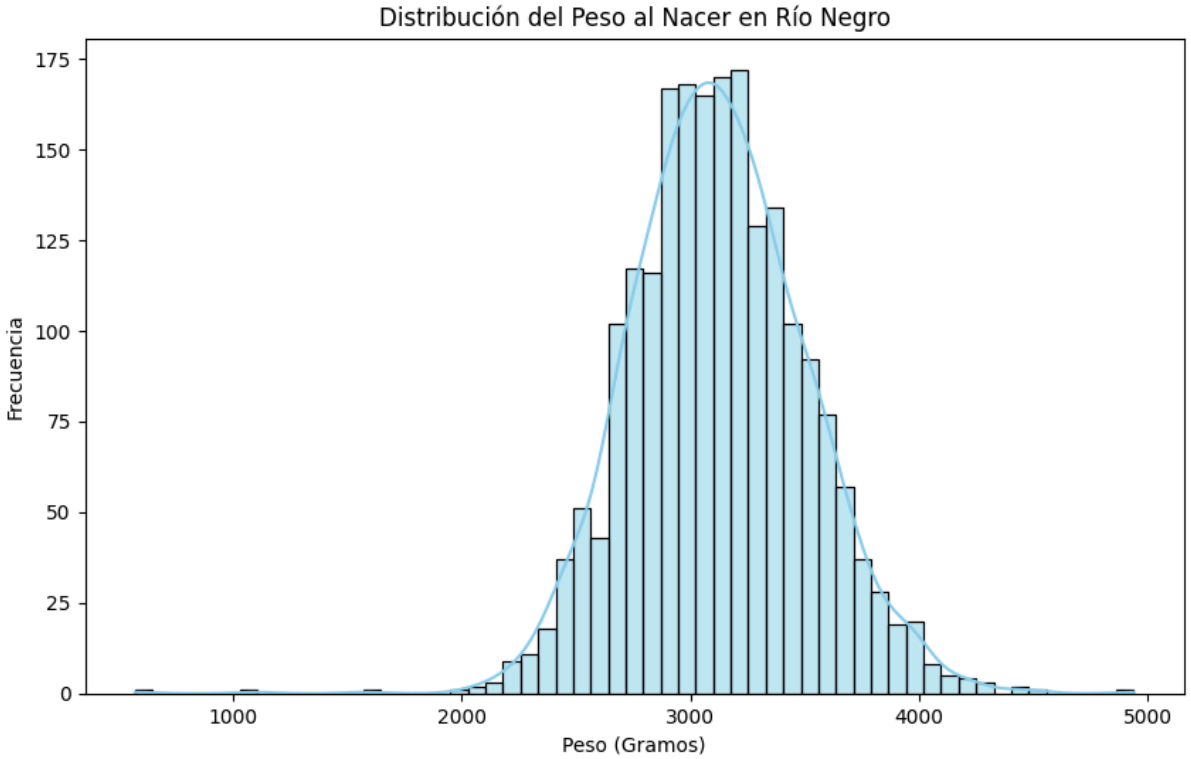

Imagen 13. Gráfico de distribución

Gráfico de dispersión entre Peso al Nacer y Tiempo de Gestación

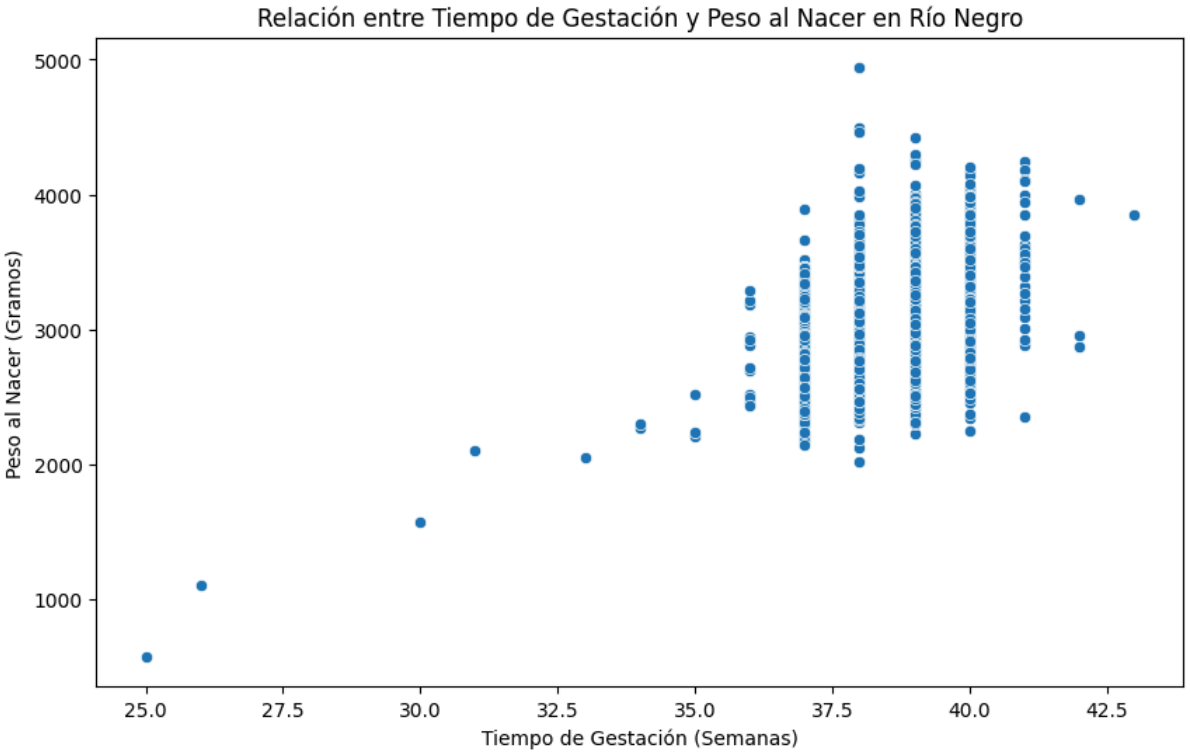

Imagen 14. Gráfico de correlación

Pairplot con Diferenciación por Sexo

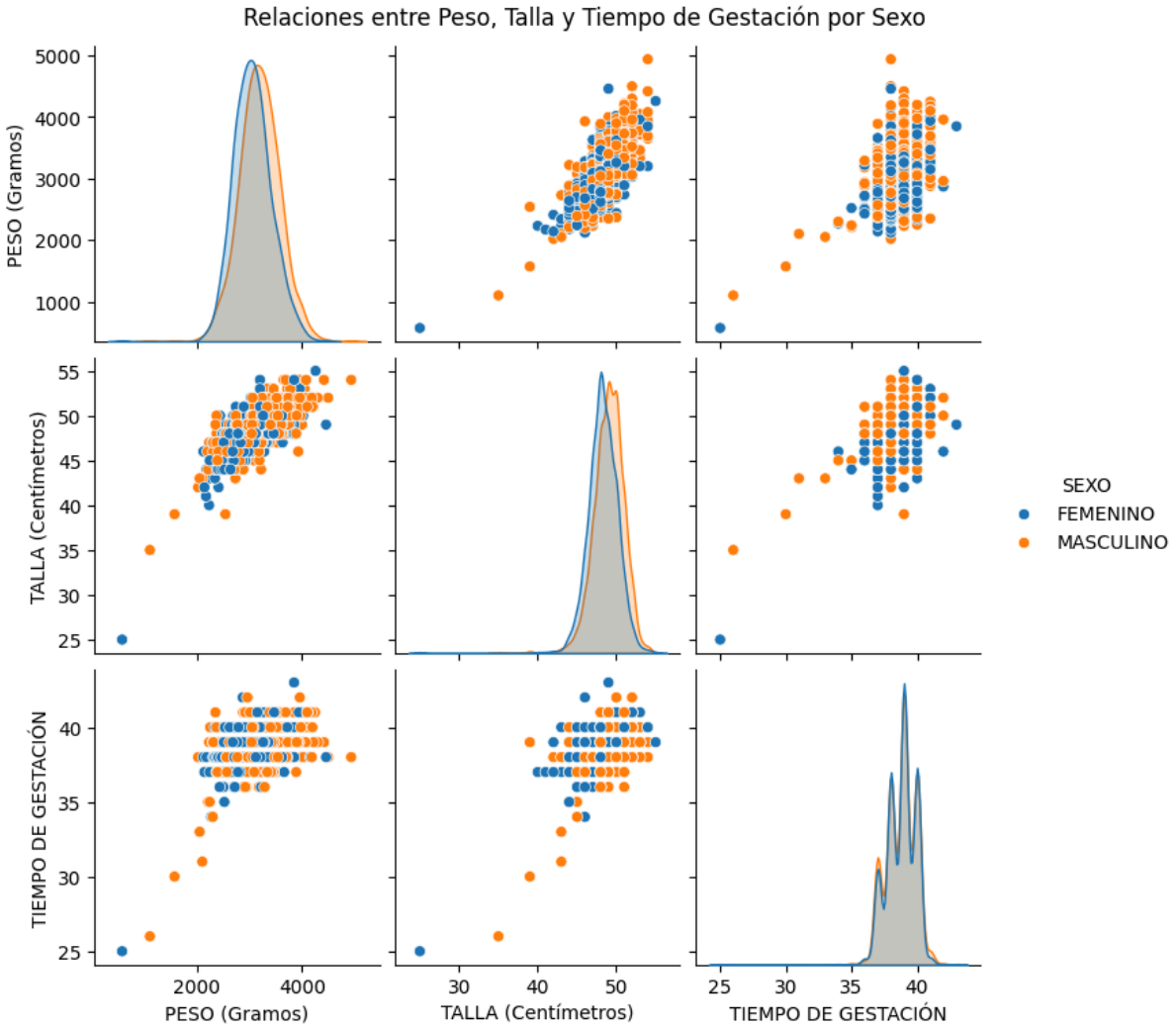

Imagen 15. Gráfico pair plot

Gráfico de Violín para Comparar Distribuciones

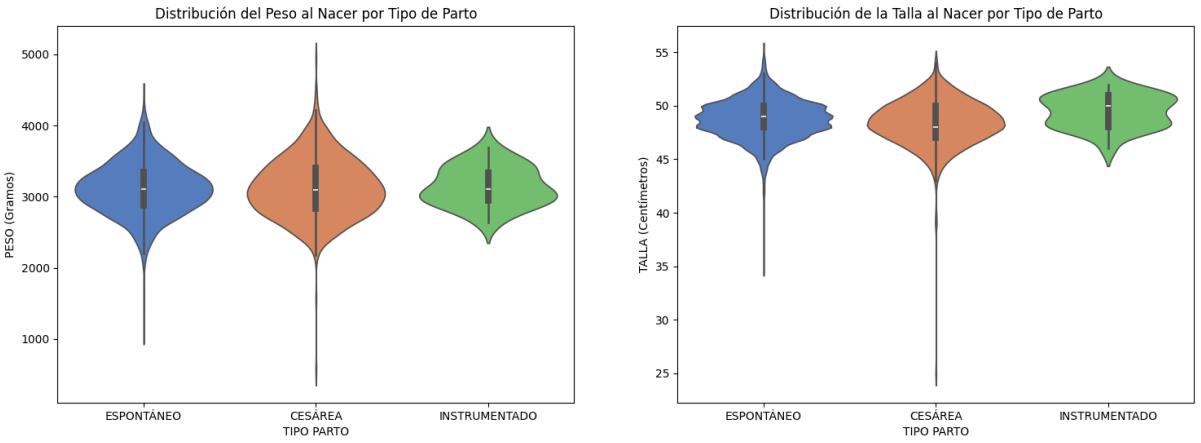

Imagen 16. Gráfico de distribución

Descripción del problema de analítica

El análisis comparativo de los patrones de nacimiento entre las ciudades de Bucaramanga y Río Negro presenta un desafío fundamental para comprender las disparidades en la salud neonatal y las prácticas de atención prenatal. Esto implica investigar cómo diversos factores, como el peso al nacer, la talla, el tiempo de gestación y las consultas prenatales, difieren entre las dos poblaciones, considerando las características demográficas y socioeconómicas únicas de cada región.

El objetivo principal es identificar diferencias significativas en la salud de los recién nacidos entre Bucaramanga y Río Negro, así como los factores asociados con los nacimientos prematuros en ambas ciudades. Esto permitirá desarrollar estrategias específicas para mejorar la atención prenatal y neonatal, así como asignar recursos de manera efectiva para apoyar a las mujeres embarazadas de alto riesgo. Para ello se promueven ciertas tecnicas de analisis para asi conseguir resultados mas asertivos y eficientes:

Técnicas de Análisis Propuestas:

1. Análisis de Clasificación y Regresión:
    - Se pueden emplear técnicas de clasificación para identificar patrones en la salud neonatal y determinar si un bebé nace prematuramente o a término.
    - Para el análisis de regresión, es posible predecir el peso al nacer utilizando variables como tiempo de gestación, número de consultas prenatales, edad de los padres, entre otros.

Comparación de Técnicas:

1. Análisis de Clasificación:
    - Para la clasificación, se pueden comparar algoritmos como Support Vector Machines (SVM) y Random Forest.
    - SVM es una técnica poderosa para separar clases en conjuntos de datos complejos, mientras que Random Forest es robusto y puede manejar características irrelevantes o redundantes.
2. Análisis de Regresión:
    - En el análisis de regresión, se pueden comparar modelos como Random Forest, Gradient Boosting, Linear Regression y Neural Networks.
    - Gradient Boosting y Neural Networks han mostrado buen desempeño en términos de precisión y exactitud, mientras que Linear Regression ofrece tiempos de entrenamiento más rápidos.

Al comparar estas técnicas, buscamos determinar cuál ofrece el mejor equilibrio entre precisión, capacidad explicativa y eficiencia en el entrenamiento. Esto permitirá seleccionar el modelo más adecuado para predecir la salud neonatal y los factores asociados con los nacimientos prematuros en Bucaramanga y Río Negro, facilitando así la implementación de intervenciones efectivas para mejorar la salud materna e infantil en ambas regiones.

Limpieza de los datos

En esta sección, abordaremos la crucial tarea de asegurar la integridad y precisión de los datos recopilados, a través de las siguientes etapas:

1. **Manejo de valores faltantes:** Empezaremos por examinar exhaustivamente todas las columnas en busca de posibles valores faltantes. Una vez identificados, evaluaremos las opciones disponibles para abordar esta situación. Podemos optar por eliminar las filas afectadas, imputar los valores faltantes utilizando técnicas estadísticas como la media o la mediana, o recurrir a métodos más avanzados si la complejidad del problema lo requiere.
2. **Detección de valores atípicos:** Prestaremos especial atención a ciertas variables críticas como peso, talla, tiempo de gestación, número de consultas prenatales, así como las edades de la madre y el padre. Buscaremos identificar cualquier valor que parezca estar significativamente fuera de lo común, lo cual podría indicar errores en la recolección de datos o incluso casos excepcionales. Una vez detectados, evaluaremos si es necesario eliminarlos o imputarlos para mantener la coherencia y fiabilidad de nuestros análisis.
3. **Formateo de fecha:** Es fundamental garantizar que la información temporal, en particular la fecha de nacimiento, esté presentada de manera coherente y adecuada para su posterior análisis. Verificaremos la consistencia del formato de fecha en la columna correspondiente y, de ser necesario, realizaremos las conversiones pertinentes para asegurar su compatibilidad con cualquier análisis de series temporales que se realice posteriormente.
4. **Corrección de errores de entrada:** Finalmente, nos aseguraremos de abordar posibles errores en la entrada de datos, especialmente en variables críticas como sexo, área de nacimiento, tipo de parto, multiplicidad de embarazo entre otros atributos. Mediante una revisión exhaustiva, identificaremos y corregiremos cualquier valor incorrecto o inconsistente, garantizando así la fiabilidad y precisión de nuestros datos para futuros análisis y conclusiones.

Implementación de Técnicas ML y resultados

Entrenamiento de los Modelos de Regresión

En este estudio, se emplearon varias técnicas de Machine Learning para predecir el peso de los bebés al nacer utilizando variables como tiempo de gestación, número de consultas prenatales, edad de los padres, y características del parto. Los modelos evaluados incluyen Random Forest, Gradient Boosting, Linear Regression, y Neural Networks.

Split Train, test: Se determinó un split de 20% para el dataset de entrenamiento.

Bibliotecas y Hiperparámetros Usados

Random Forest: Implementado usando RandomForestRegressor de sklearn.ensemble.

- **Hiperparámetros:** n_estimators=100, random_state=42.

Gradient Boosting: Utilizado a través de GradientBoostingRegressor de sklearn.ensemble.

- **Hiperparámetros**: n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42.

Linear Regression: Aplicado mediante LinearRegression de sklearn.linear_model.

Neural Network: Configurado usando MLPRegressor de sklearn.neural_network.

- **Hiperparámetros:** hidden_layer_sizes=(100,), max_iter=1000, random_state=42.

Resultados de Rendimiento:

Modelo	MAE	MSE	RMSE	R^2	MAPE	Explained Variance	Training Time (s)
Random Forest	322.65	166800.88	408.41	0.346	10.63%	0.346	14.44
Gradient Boosting	287.6	132768.21	364.37	0.479	9.50%	0.479	3
Linear Regression	291.03	136248.58	369.12	0.466	9.65%	0.466	0.07
Neural Network	288.84	135066.33	367.51	0.47	9.48%	0.472	206.61

Interpretación:

Análisis de Rendimiento

1. **Precisión y Exactitud:**
   - Gradient Boosting y Neural Network tienen los mejores valores de MAE, MSE, RMSE y MAPE. Esto indica que son los modelos más precisos y exactos en términos de predicción del peso del bebé.
   - Gradient Boosting tiene el mejor R^2 (0.479), lo que sugiere que este modelo explica mejor la variabilidad en el peso del bebé en comparación con los otros modelos.
2. **Variabilidad Explicada:**
   - Gradient Boosting también tiene el mejor valor de Explained Variance (0.479), que mide la proporción de variabilidad en los datos que es capturada por el modelo. Esto refuerza la conclusión de que Gradient Boosting es el mejor modelo en términos de capacidad explicativa.
3. **Tiempo de Entrenamiento:**

- Linear Regression tiene el tiempo de entrenamiento más rápido (0.071 segundos), seguido de Gradient Boosting (2.999 segundos).
- Neural Network tiene el tiempo de entrenamiento más largo (206.61 segundos), lo que podría ser una desventaja si se requiere entrenamiento rápido o si se trabaja con grandes volúmenes de datos.

Consideraciones Adicionales

1. **Trade-off entre Precisión y Tiempo de Entrenamiento:**
   - Aunque Gradient Boosting y Neural Network ofrecen las mejores métricas de rendimiento, Gradient Boosting es significativamente más rápido de entrenar. Esto lo hace una opción más práctica cuando se necesita un equilibrio entre precisión y eficiencia.
   - Linear Regression, a pesar de ser menos preciso, tiene un tiempo de entrenamiento extremadamente bajo, lo que puede ser útil en aplicaciones donde la rapidez es crítica y se puede tolerar una menor precisión.
2. **Importancia de Características:**
   - Es esencial revisar la importancia de las características para entender mejor qué factores están contribuyendo más a las predicciones del modelo. Esto también puede proporcionar información valiosa para la toma de decisiones en el contexto clínico o de investigación.

Recomendaciones

1. **Modelo Sugerido:**
   - Gradient Boosting es el modelo recomendado, ya que ofrece el mejor equilibrio entre precisión, capacidad explicativa y tiempo de entrenamiento razonable.
2. **Mejoras Futuras:**
   - Optimización de Hiperparámetros: Se pueden realizar búsquedas de hiperparámetros (Grid Search o Random Search) para afinar aún más el rendimiento de los modelos.
   - Ampliación de Características: Considerar la inclusión de más características o la creación de nuevas características derivadas puede mejorar aún más la precisión de las predicciones.
   - Ensemble Methods: Combinar múltiples modelos (por ejemplo, stacking) podría mejorar el rendimiento general.
3. **Evaluación Continua:**
   - Implementar validación cruzada para garantizar la generalización del modelo y evitar sobreajuste.

Conclusión

Los resultados indican que **Gradient Boosting** es el mejor modelo para predecir el peso del bebé, considerando tanto la precisión de las predicciones como la eficiencia en el entrenamiento. Este modelo debería ser el enfoque principal, con potenciales optimizaciones y expansiones en el futuro para mejorar aún más su rendimiento.

Presentación de las Respuestas a las Preguntas Propuestas en la Fase Inicial

Importancia de las Preguntas

Las preguntas planteadas en la fase inicial fueron diseñadas para profundizar en la comprensión de cómo diversos factores prenatales y ambientales pueden afectar la salud de los bebés al nacer en dos municipios específicos: Bucaramanga y Rionegro. Estas preguntas son fundamentales ya que permiten:

- Identificar diferencias regionales en la salud neonatal que podrían ser influenciadas por factores socioeconómicos, de atención médica o ambientales.
- Orientar políticas públicas y prácticas médicas para mejorar los cuidados y resultados durante el embarazo y el parto.
- Evaluar la efectividad de los servicios de atención prenatal, proporcionando una base para mejorar la calidad y accesibilidad de estos servicios.

Las preguntas fueron lo suficientemente amplias para abarcar un tema significativo de salud pública, pero también lo suficientemente concretas para permitir un análisis detallado y resultados palpables, fundamentales para implementar cambios reales y medibles.

Respuestas a las Preguntas Propuestas

¿Cómo cambia la salud de los bebés entre Bucaramanga y Rionegro?

El análisis de los datos recolectados de Bucaramanga y Rionegro mostró diferencias marginales en el peso y la talla al nacer entre ambos municipios, indicando variaciones en el desarrollo fetal que pueden estar influenciadas por diferencias en la calidad de la atención prenatal, nutrición y factores socioeconómicos. El acceso uniforme a la atención prenatal, reflejado en el número similar de consultas prenatales en ambos municipios, sugiere que las políticas actuales son efectivas en proporcionar un nivel básico de cuidado, pero podrían necesitar ajustes para abordar las diferencias específicas encontradas.

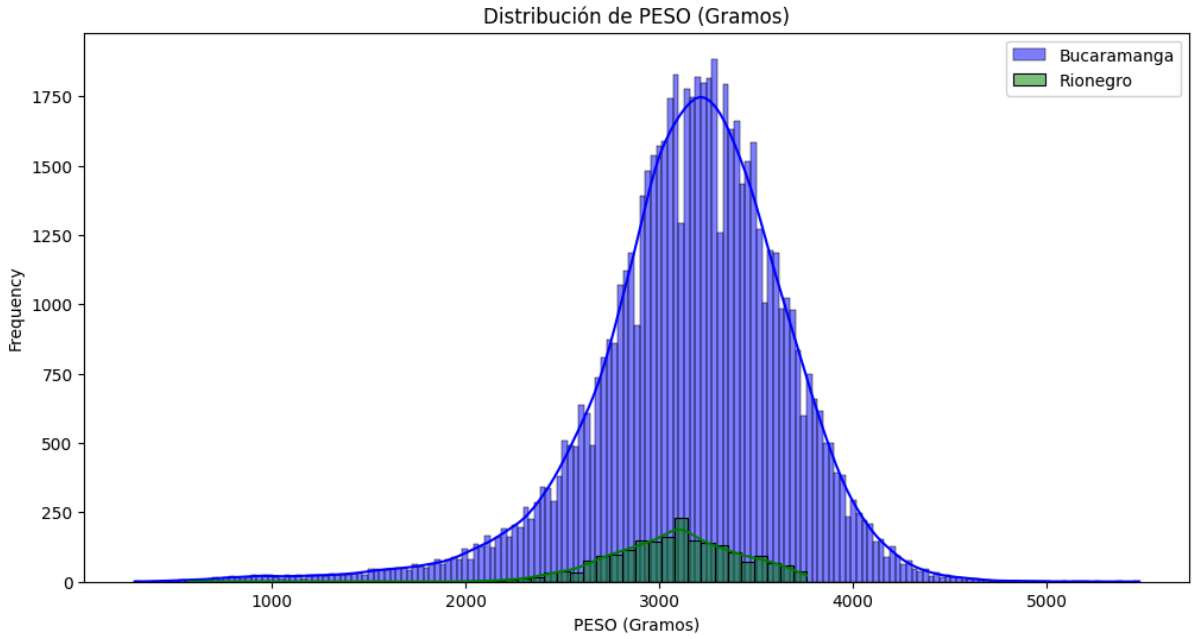

Imagen 17. Comparación de Distribución de peso

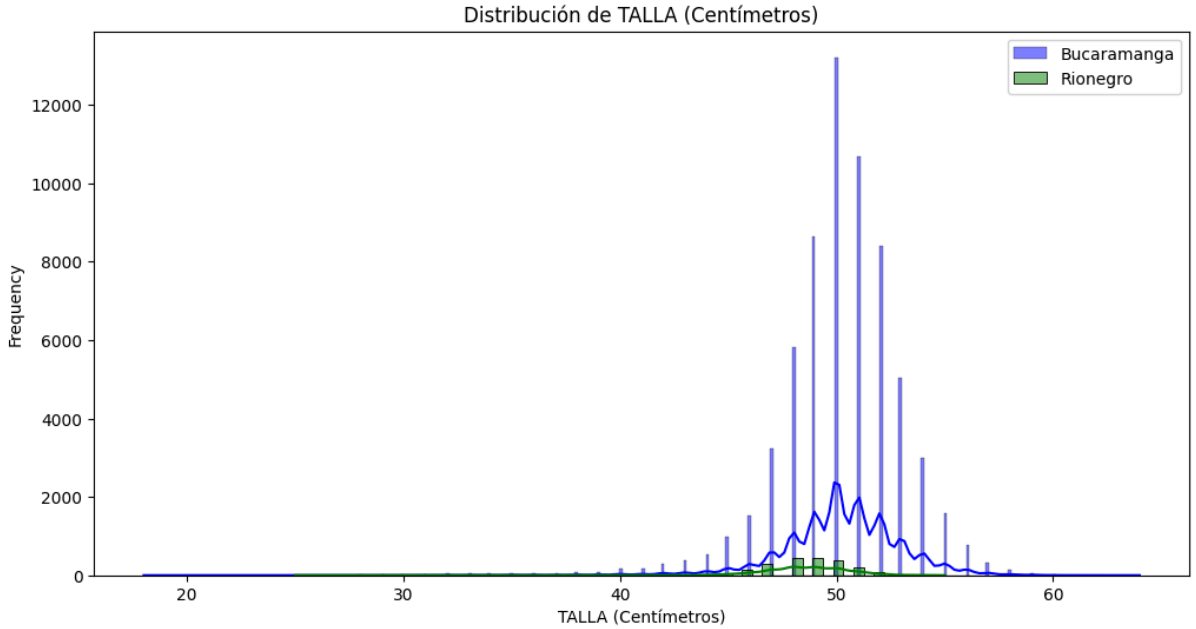

Imagen 18. Comparación de Distribución de talla

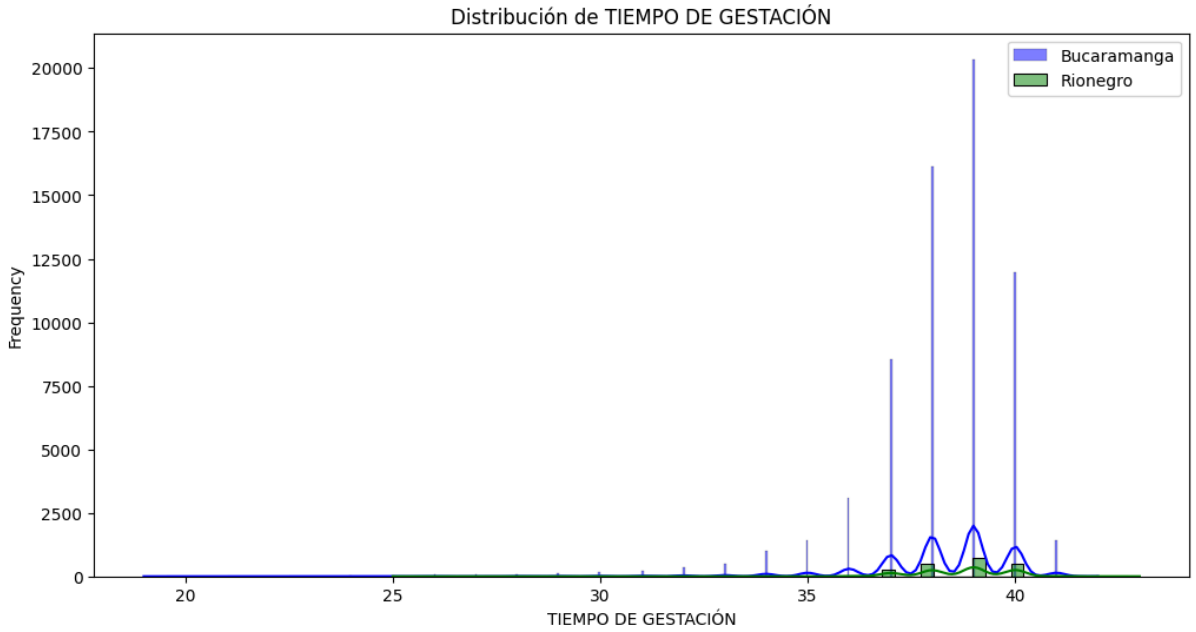

Imagen 18. Comparación de Distribución de tiempo de gestacion

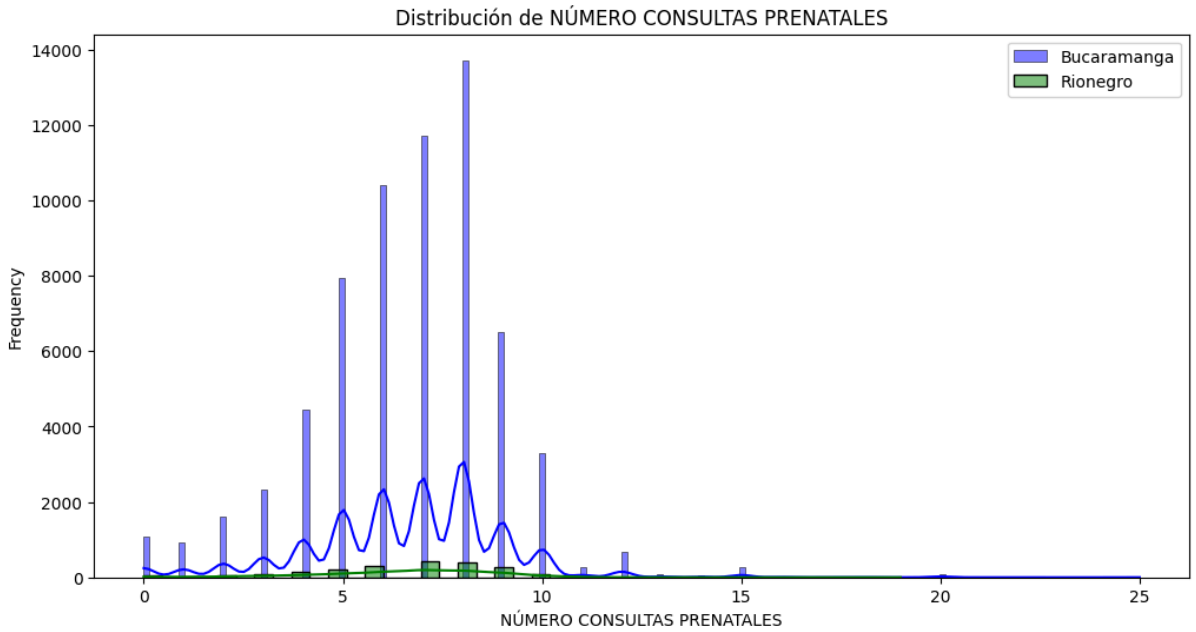

Imagen 19. Comparación de Distribución de número de consultas prenatalesv

Insights Encontrados

Diferencias Marginales en el Peso al Nacer

- Promedio en Bucaramanga: 3157.54 gramos.
- Promedio en Rionegro: 3080.40 gramos.
- La mayor desviación estándar en Bucaramanga indica una mayor variabilidad en el peso al nacer, posiblemente reflejando diferencias en condiciones nutricionales o socioeconómicas.

Superioridad en Talla al Nacer en Bucaramanga

- Promedio en Bucaramanga: 50.16 centímetros.
- Promedio en Rionegro: 48.62 centímetros.
- Un rango más amplio en Bucaramanga sugiere un desarrollo fetal más completo en términos de crecimiento físico en este municipio.

Consistencia en el Tiempo de Gestación en Rionegro

- La duración media del embarazo es similar, con medianas cercanas a 39 semanas.
- Rionegro muestra una menor variabilidad (desviación estándar de 1.17 semanas) comparado con Bucaramanga (1.86 semanas), indicando una mayor uniformidad en la duración de los embarazos.

Acceso Uniforme a la Atención Prenatal

- Número medio de consultas prenatales: Alrededor de 6.6 en ambos municipios.

- La similitud en el número de consultas prenatales entre Bucaramanga y Rionegro muestra que el acceso a la atención prenatal es uniforme, destacando una cobertura equitativa de servicios de salud prenatal.

En general, aunque existen algunas variaciones entre Bucaramanga y Rionegro en términos de peso y talla al nacer, el cuidado prenatal medido por las consultas parece ser uniformemente distribuido, sugiriendo que las diferencias observadas podrían ser influenciadas por factores socioeconómicos o ambientales más que por la calidad de la atención médica prenatal.

¿Qué factores están asociados con los nacimientos prematuros en Bucaramanga y Rionegro?

El estudio reveló que factores como el número de consultas prenatales y las medidas físicas al nacer (peso y talla) están asociados con el tiempo de gestación. En particular, la investigación mostró una correlación positiva débil entre el número de consultas prenatales y un mayor tiempo de gestación, sugiriendo que aumentar el número de estas consultas podría contribuir a reducir la tasa de nacimientos prematuros. Este hallazgo es crucial para mejorar las estrategias de atención prenatal y puede ser un enfoque clave para las intervenciones de salud pública.

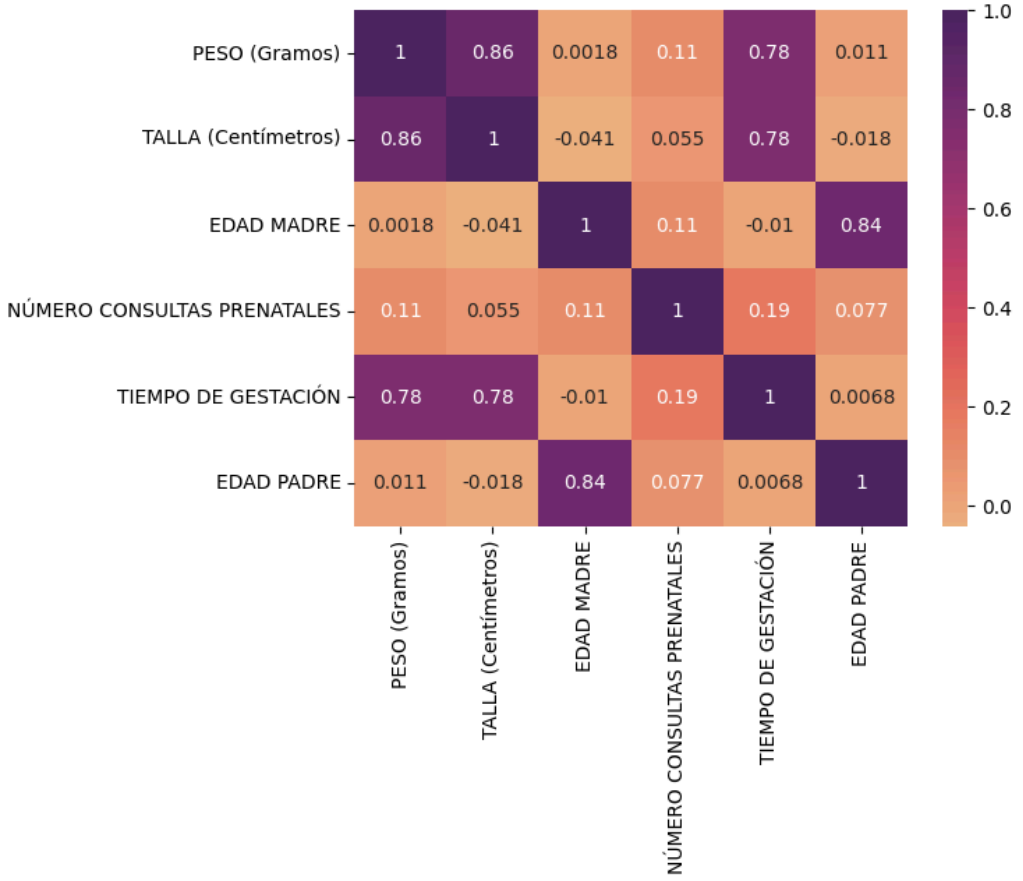

Imagen 20. Matriz de correlación

Factores Asociados con los Nacimientos Prematuros en Bucaramanga y Rionegro

Basado en el análisis de la matriz de correlación, que incluye variables como el peso al nacer, la talla al nacer, el número de consultas prenatales, y el tiempo de gestación, podemos identificar algunos factores asociados con los nacimientos prematuros:

1. Peso y Talla al Nacer

- **Correlación Fuerte:** Tanto el peso como la talla al nacer muestran fuertes correlaciones positivas (0.78) con el tiempo de gestación. Esto indica que menores pesos y tallas al nacer están comúnmente asociados con nacimientos prematuros, reflejando una menor duración del embarazo y menos tiempo para el desarrollo fetal.

2. Número de Consultas Prenatales

- **Correlación Positiva Débil:** Existe una correlación positiva débil (0.19) entre el número de consultas prenatales y el tiempo de gestación. Aunque la correlación no es fuerte, sugiere que un mayor número de consultas podría estar asociado con un aumento en la duración del embarazo, posiblemente debido a una mejor vigilancia y manejo del embarazo.

Recomendaciones Basadas en el Análisis

- Promover el Aumento de las consultas prenatales: Fomentar la asistencia regular a las consultas prenatales puede ayudar a gestionar mejor los riesgos durante el embarazo y prevenir complicaciones que podrían llevar a un parto prematuro.
- Investigar Factores Contribuyentes: Es importante realizar más estudios para entender qué aspectos de las consultas prenatales son más efectivos en extender la duración del embarazo y cómo estos factores podrían variar entre diferentes grupos de población.

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

Trabajo en Equipo y Contribuciones de los Integrantes

El proyecto de análisis comparativo de los patrones de nacimiento entre Bucaramanga y Río Negro fue un esfuerzo colaborativo que requirió la especialización y dedicación de cada uno de los integrantes del equipo. Cada miembro asumió responsabilidades clave que contribuyeron al éxito del proyecto.

Contribuciones de los Integrantes

Juan Diego González - Fase Inicial (3.A)

- Selección de los Conjuntos de Datos: Juan Diego fue responsable de seleccionar los conjuntos de datos adecuados para el análisis. Esto incluyó la identificación de las fuentes de datos relevantes y la evaluación de su calidad y pertinencia para el proyecto.
- Formulación de Preguntas de Investigación: Desarrolló y formuló preguntas de investigación que fueron fundamentales para guiar el análisis. Aseguró que las preguntas fueran lo suficientemente amplias para permitir un análisis significativo y suficientemente concretas para obtener resultados tangibles.
- Exploración Inicial de los Datos: Realizó un análisis exploratorio detallado, revisando la dispersión de los datos, tendencias centrales y sesgos. Este trabajo fue esencial para establecer la base del análisis posterior y para identificar áreas que requerían limpieza y normalización.

Nicolás Samuel Martín - Problema (3.B)

- Definición del Problema de Analítica: Nicolás definió claramente el problema de analítica a resolver, estableciendo si sería de clasificación o regresión y seleccionando las técnicas analíticas apropiadas.
- Solución de Problemas de Calidad de Datos: Abordó eficazmente los problemas de calidad de los datos, como valores faltantes y datos atípicos, aplicando técnicas de limpieza y normalización para asegurar que los datos fueran de alta calidad.
- Creación de Variables Derivadas y Fusión de Datos: Fue responsable de la creación de nuevas variables que enriquecieron el análisis y de la fusión efectiva de los conjuntos de datos, garantizando que la unión fuera coherente y completa.

Juan David López Becerra - Implementación de Técnicas ML y Resultados (3.C)

- Entrenamiento de Modelos de Machine Learning: Juan David se encargó del entrenamiento de los modelos, utilizando diferentes técnicas de machine learning para comparar su desempeño. También fue responsable de ajustar los hiperparámetros y utilizar las bibliotecas adecuadas.
- Evaluación de los Modelos: Evaluó los modelos en la etapa de pruebas utilizando varias métricas de rendimiento, lo que permitió identificar el modelo más efectivo basado en los datos analizados.
- Presentación de Resultados y Conclusiones: Compiló y presentó los resultados finales, asegurando que las respuestas a las preguntas de investigación fueran claras y bien fundamentadas. Además, redactó las conclusiones finales del proyecto, destacando las observaciones clave y recomendaciones para futuras investigaciones.

Conclusión del Trabajo en Equipo

La colaboración entre Juan Diego, Nicolás y Juan David fue fundamental para el éxito del proyecto. Cada miembro no solo cumplió con sus responsabilidades asignadas sino que

también aportó al trabajo colectivo, asegurando que el proyecto no solo cumpliera con los objetivos académicos sino que también ofreciera insights aplicables para mejorar la atención prenatal y neonatal en las regiones estudiadas.

Conclusiones, Observaciones y Recomendaciones sobre el Proyecto

Conclusiones Generales

El proyecto de análisis comparativo de los patrones de nacimiento entre Bucaramanga y Río Negro ha revelado diferencias significativas y patrones interesantes que reflejan las variabilidades sociodemográficas y de atención médica entre ambas ciudades. El uso de técnicas avanzadas de procesamiento de datos y machine learning ha permitido no solo identificar estas diferencias sino también ofrecer perspectivas sobre posibles intervenciones.

Observaciones Clave

1. **Variabilidad en los Datos**: Los conjuntos de datos de Bucaramanga y Río Negro, aunque similares en estructura, mostraron variaciones notables en aspectos como el peso al nacer y el número de consultas prenatales, lo que subraya la importancia de considerar factores locales en estudios de salud pública.
2. **Calidad de Datos**: Durante el proyecto, se observó la necesidad de una limpieza y preparación rigurosa de datos para asegurar la precisión de los análisis. La presencia de valores atípicos y datos faltantes fue un reto que se abordó eficazmente para mejorar la fiabilidad de los resultados.
3. **Uso de Tecnologías de Machine Learning**: La aplicación de varios modelos de regresión y técnicas de análisis estadístico proporcionó una comprensión más profunda de los factores que influyen en los patrones de nacimiento y destacó la utilidad de estas herramientas en el análisis de datos complejos.

Recomendaciones

1. **Fortalecimiento de la Atención Prenatal**: Basado en los hallazgos del proyecto, es crucial mejorar la calidad y frecuencia de la atención prenatal en ambas ciudades. Las políticas deben enfocarse en garantizar que todas las mujeres embarazadas tengan acceso a consultas prenatales regulares y de alta calidad.
2. **Educación y Concienciación**: Desarrollar programas que aumenten la conciencia sobre la importancia de la atención prenatal y educar a las comunidades sobre prácticas de salud materna e infantil óptimas.
3. **Investigación Continua**: Se recomienda continuar con la investigación en esta área para explorar más a fondo las causas específicas de las variaciones observadas y para evaluar la efectividad de las intervenciones implementadas.

4. **Colaboración Intersectorial**: Fomentar una colaboración más estrecha entre los gobiernos locales, instituciones de salud, y organizaciones no gubernamentales para abordar las necesidades identificadas en el estudio de manera coordinada y efectiva.
5. **Uso de Datos en Tiempo Real**: Integrar sistemas de monitoreo que utilicen datos en tiempo real para proporcionar actualizaciones constantes sobre las condiciones de salud materna e infantil, lo que puede facilitar respuestas más rápidas y dirigidas.

Conclusión Final

Este proyecto ha demostrado la capacidad de los análisis de datos a gran escala para proporcionar insights valiosos que pueden informar y mejorar las políticas de salud pública. Las conclusiones obtenidas destacan la necesidad de enfoques personalizados y basados en datos para la atención médica prenatal, lo que eventualmente puede llevar a mejoras significativas en la salud de madres y bebés en Bucaramanga, Río Negro, y potencialmente en otras regiones similares.