Procesos Gaussianos Heterogéneos de Múltiple Salida para la Predicción del Bajo Peso al Nacer en Medellín.

Diego Alejandro Salazar Blandón

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Hernán Felipe García Arias, Doctor (PhD)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	(Salazar Blandon, 2023)
Referencia **Estilo APA 7 (2020)**	Salazar Blandón, D. A. (2023). *Procesos Gaussianos Heterogéneos de Múltiple Salida para la Predicción del Bajo Peso al Nacer en Medellín.* Trabajo de grado especialización. Universidad de Antioquia, Medellín, Colombia.

creative commons CC BY NC SA

Especialización en Analítica y Ciencia de Datos, Cohorte V.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

A mi amada esposa Ana María y a mi querida hija Azucena. Mi mayor fuente de motivación y alegría.

Agradecimientos

A mi asesor Hernán Felipe García Arias por su paciencia y disposición para compartir sus conocimientos. Su mentoría ha sido una fuente constante de inspiración y motivación.

Al Doctor Juan José Giraldo por su orientación y valiosos aportes en la implementación técnica de los modelos HETMOGP.

Tabla de contenido

Resumen 1

Abstract 2

1. Introducción 3

1.1. Planteamiento del problema 4

1.2. Aproximación desde la analítica de datos 5

1.3. Origen de los datos 7

2. Objetivos 8

2.1. Objetivo general 8

2.2. Objetivos específicos 8

3. Datos 9

3.1. Datos originales 9

3.2. Datsets 12

4. Proceso de analítica 17

4.1. Pipeline principal 17

4.2. Preprocesamiento 18

4.3. Modelos 18

4.4. Métricas 24

5. Metodología 26

5.1. Enfoques generales para modelos de única salida 26

5.1.1 Baseline 26

5.1.2 Validación 27

5.1.3 Iteraciones y Evolución 27

5.1.4 Herramientas 28

5.2. Enfoques generales basados en proceso Gaussianos de múltiple salida 28
5.2.1 Baseline 28
5.2.2 Iteraciones y Evolución 29
5.2.3 Validación 29
5.2.4 Herramientas 29
6. Resultados y discusión 31
6.1. Modelos de ML para la predicción del peso recién nacido. 31
6.1.1. Ajuste de los parámetros y tratamiento de datos SVM y RL 32
6.1.2 Validación Cruzada SVM y RL 33
6.2. Modelos de ML para la clasificación del bajo peso al nacer. 35
6.2.1. Ajuste de parámetros y tratamiento de datos Random Forest y LR 36
6.2.2. Validación Cruzada Random Forest y LR 37
6.3. Modelo de Múltiple Salida HETMOGP 38
7. Conclusiones 43
8. Trabajos futuros 44
Referencias 45

Lista de figuras

Figura 1: Tasa de nacimientos según bajo peso x cada 1.000 nacidos vivos. Total nacional – Años completos 2015-2020 y año corrido 2021pr. 5

Figura 2: Ficha de nacido vivo.....11

Figura 3: Gráficos de barras para grupo 1 de variables cualitativas vs bajo peso al nacer.....13

Figura 4: Gráficos de barras para grupo 2 de variables cualitativas vs bajo peso al nacer.....13

Figura 5: Diagramas de dispersión y distribución para el grupo 1 variables cuantitativas.....14

Figura 6: Diagramas de dispersión y distribución para el grupo 2 variables cuantitativas.....15

Figura 7: Diagramas de cajas y bigotes para variables cuantitativas (estandarizadas) según clasificación de bajo peso al nacer. 16

Figura 8: Flujo de trabajo.....17

Figura 9: En (a) se presentan tres funciones extraídas al azar de un GP a priori y en (b) se presenta las funciones condicionadas a cinco puntos observados.....19

Figura 10: Esquema gráfico de un GP de regresión.....21

Figura 11: Representación de una sola neurona y una aplicación de una red neuronal artificial para predecir los niveles de azúcar en la sangre en ayunas en función de variables clínicas y demográficas. 23

Figura 12: Ejemplo hipotético de como un árbol de decisión predice la interacción entre proteínas. 23

Figura 13: Representación de una clasificación de datos en dos dimensiones que son discriminados luego de una transformación de SVM en un espacio de tres dimensiones. 24

Figura 14: Comparación de pesos reales versus predichos por los modelos de ML 32

Figura 15: Dispersión de las métricas en validación cruzada para predicción del peso del recién nacido 34

Figura 16: Curvas ROC de los desempeños de diferentes modelos de ML para la clasificación del bajo peso al nacer 36

Figura 17: Dispersión de las métricas en validación cruzada en la clasificación del bajo peso al nacer. 38

Figura 18:Prediccion del peso del recién nacido y bajo peso al nacer según modelo HETMOGP 41

Figura 19: Optimización configuración del experimento 4 en un HETMOGP 42

Lista de tablas

Tabla 1: Información del archivo que contiene conjunto de datos 9

Tabla 2: Descripción y naturaleza de las variables en el conjunto 10

Tabla 3: Variables incluidas en el análisis 12

Tabla 4: Parámetros por defecto usados en las modelos de única salida para la tarea de regresión y de clasificación. 27

Tabla 5: Desempeño de diferentes técnicas de ML en la predicción del peso del recién nacido .31

Tabla 6: Métricas luego de ajuste y transformación de los datos en la predicción del peso del recién nacido. 33

Tabla 7: Métricas por pliegue en validación cruzada en predicción del peso del recién nacido...33

Tabla 8: Desempeño de diferentes técnicas de ML en la clasificación del bajo peso al nacer 35

Tabla 9: Métricas luego de ajuste y transformación de los datos en la clasificación del bajo peso al nacer. 36

Tabla 10: Métricas por pliegue en validación cruzada en la clasificación del bajo peso al nacer.37

Tabla 11: Análisis de relevancia de características mediante un LGBM.....39

Tabla 12: Variación de parámetros y desempeño de diferentes modelos HETMOGP para predicción del peso del recién nacido y la clasificación del bajo peso al nacer. 40

Siglas, acrónimos y abreviaturas

BPN	Bajo Peso al Nacer
OMS	Organización Mundial de la Salud
ML	Machine Learning o Aprendizaje de Máquina
VA	Variable Aleatoria
GP	Gaussian Process o Proceso Gaussiano
MOGP	Multi-Output Gaussian Process o Proceso Gaussiano de Múltiple Salida
HETMOGP	Heterogeneous Multi-Output Gaussian Process o Proceso Gaussiano Heterogéneos de Múltiple Salida
g	Gramos
SVM	Support Vector Machine o Máquinas de Soporte Vectorial
KNN	K-Nearest Neighbors o K-Vecinos más Cercanos
ComplementNB	Complement Naive Bayes Classifier o Complemento para Clasificador de Bayes Ingenuo.
RanForest	Random Forest o Bosques Aleatorios
MPL	Multi-layer Perceptron Classifier o Clasificador de Perceptrones Multicapa.
RL	Regresión Lineal
LG	Regresión Logística
PCA	Análisis de Componentes Principales
ROC	Receiver Operating Characteristic o Característica Operativa del Receptor
LGBM	Light Gradient Boost Machines o Máquinas de Refuerzo de Gradiente de Luz
MSE	Mean Squared Error o Error Cuadrático Medio

Resumen

El bajo peso al nacer (BPN) es un importante indicador de salud pública en la atención de la madre y el recién nacido, debido a la relación que tiene con las condiciones de vida, la salud de la población y de los cuidados médicos a la mujer gestante. En esta investigación se implementan diferentes modelos Machine Learning para la predicción independiente y conjunta del peso en gramos y clasificación de bajo peso (Si/No) al momento de nacer, usando información que se registra al nacimiento relacionada con variables propias del recién nacido y de la madre. Con los modelos de única salida (tarea de regresión) se logra un desempeño promedio de R^2 y MAE de 0.6748 ±0.0076 y 299.7097 ±3.8662 respectivamente usando un modelo de SVM con kernel polinomial de grado 2, en tanto en la clasificación por medio de Random Forest alcanza un promedio de accuracy y AUC de 0.8349±0.0105 y 0.8933±0.0088. El modelo HETMOGP logra desempeños similares en R^2 y accuracy, al abordar las dos tareas al tiempo lo que lo posiciona como una alternativa más eficiente en la predicción. Se recomienda la implementación del modelo HETMOGP para anticiparse al BPN y diseñar políticas de cuidado y atención prenatal que permitan reducir la incidencia de este evento en la ciudad.

Palabras clave: Recién Nacido de Bajo Peso, Atención prenatal, Aprendizaje de Múltiple Tarea, Aprendizaje Automático.

Repositorio GitHub

https://github.com/dasb03/MONOGRAFIA/tree/main

Abstract

Low birth weight (BPN) is an important public health indicator in maternal and newborn care, due to the relationship it has with living conditions, the health of the population and the medical care to the pregnant woman. In this research, different Machine Learning models are implemented for the independent and joint prediction of weight in grams and low weight classification (Yes/No) at birth, using information recorded at birth related to the newborn and the mother's own variables. With the single output models (regression task) an average performance of R^2 and MAE of 0.6748 ±0.0076 and 299.7097 ±3.8662 respectively is achieved using a SVM model with polynomial kernel of degree 2, while in the classification by Random Forest it reaches an average accuracy and AUC of 0.8349±0.0105 and 0.89333±0.0088. The HETMOGP model achieves similar performances in R^2 and accuracy, addressing both tasks at the same time, which positions it as a more efficient alternative in prediction. The implementation of the HETMOGP model is recommended to anticipate BPN and design prenatal care and attention policies to reduce the incidence of this event in the city.

Keywords: Low Birth Weight Newborn, Prenatal Care, Multi-task Learning, Machine Learning.

1. Introducción

Los avances tecnológicos de las últimas décadas han enfocado grandes esfuerzos en lograr simular la inteligencia humana por medio de máquinas (Inteligencia Artificial); el Machine Learning o Aprendizaje de Máquina es una rama especifica que busca emular por medio de algoritmos computacionales la capacidad que tienen los humanos de aprender con la experiencia que se puede materializar en este contexto en grandes volúmenes de datos (Bi et al., 2019).

Esta perspectiva de aprendizaje trajo consigo enormes retos para el análisis estadístico tradicional basado en supuestos teóricos derivados de los fundamentos matemáticos, que subyacen en la mayoría de sus técnicas; lo cual ha implicado un cambio de paradigma al privilegiar el poder predictivo de un modelo de análisis sobre el cumplimiento de supuestos teóricos (Wiemken & Kelley, 2020). Esta considerable libertad permitió importantes avances en el Machine Learning, el cual se caracteriza por agrupar un conjunto de técnicas estadísticas tradicionales y algoritmos computacionales en busca de optimizar el análisis de datos para la toma de decisiones y cuyo alcance desbordó su intencionalidad inicial y abrió la puerta para que sea aplicado en diferentes y diversas áreas del conocimiento como la Medicina (Núñez Reiz et al., 2019) o el Derecho (Salinas de León, 2020) por mencionar algunas. La implementación del Machine Learning en cualquier disciplina particular implica cuestionamientos teóricos sobre dichas técnicas que avalen su uso, donde el poder predictivo por sí solo ya no es suficiente, con el agravante de que muchas de estas técnicas computacionales han evolucionado a tal punto que se convierten en una caja negra que siempre ofrece resultados, pero donde no se entiende muy bien lo que pasa al interior como sí ocurre con la mayoría de las técnicas estadísticas convencionales.

A pesar de lo anterior en el área de la salud en general, el uso de técnicas Machine Learning se abre camino poco a poco para análisis de datos, con un alto potencial para transformar ciencias epidemiológicas (Wiemken & Kelley, 2020). No obstante, y paralelo a esto, es relevante avanzar en el estudio del alcance y las limitaciones de dichas técnicas que brinden mayor seguridad para su uso en este contexto particular y permitan por ejemplo distinguir cual algoritmo es más efectivo en la predicción de eventos en salud o que características en los datos mejoran el desempeño de uno u otro modelo y así poder aportar a la formalización y estandarización de su uso en el área de la salud. El presente estudio busca aportar en esta dirección tomando un evento en salud importante

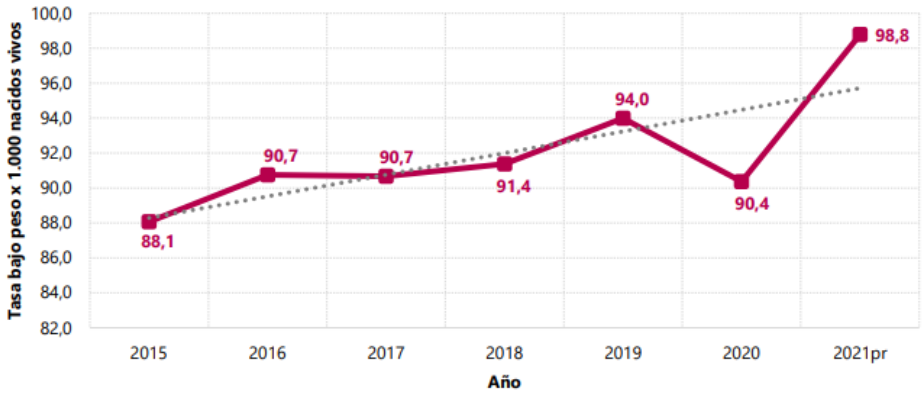
como es el bajo peso al nacer en un contexto local y usar una técnica novedosa de ML como lo son los Procesos Gaussiano Heterogéneos de Múltiples Salidas o HETMOGP (*Heterogeneous Multi-Output Gaussian Process*) para entrenar un modelo que permita predecir al tiempo el peso del recién nacido y el bajo peso al nacer y comparar los resultados con técnicas estadísticas tradicionales de única salida y más usadas en salud como la Regresión Lineal o Logística y otros técnicas el ML como las Máquinas de Soporte Vectorial (SVM) y Arboles de Decisión.

1.1. Planteamiento del problema

La Organización Mundial de la Salud (OMS) define el Bajo Peso al Nacer (BPN) como aquel recién nacido que tiene un peso inferior o igual a 2.500g sin importar su edad gestacional o motivo del bajo peso. Esta definición se basa en observaciones epidemiológicas según las cuales los niños que pesan menos de 2.500g tienen aproximadamente 20 veces más probabilidades de morir que los bebés con un peso mayor (World Health Organization & United Nations Children's Fund, 2004) .

El BPN es una de las principales causas de mortalidad neonatal y también se asocia con problemas posteriores a esta etapa a mediano y largo plazo. Los niños con bajo peso presentan mayor incidencia de déficit neurológico, alteraciones del crecimiento, deficiencias cognitivas y enfermedades crónicas no transmisibles. Entre las causas del BPN se ha encontrado evidencia de relación con factores sociodemográficos asociados a los padres (Victoria et al., 2009), la cantidad de consultas prenatales entre otros. Es por esto, que se considera un importante indicador de salud pública en la atención a la madre y el recién nacido, debido a la relación que tiene con las condiciones de vida, la salud de la población y de los cuidados médicos a la mujer gestante. Según cifras oficiales en Colombia este indicador de salud ha mostrado una tendencia al alza en los últimos años (ver figura 1). En este sentido contar un modelo que permita anticiparse al BPN identificando los factores riesgo que se pueden intervenir permitiría reducir las tasas de mortalidad infantil y mejorar los indicadores de salud.

***Figura 1**: Tasa de nacimientos según bajo peso x cada 1.000 nacidos vivos. Total nacional – Años completos 2015-2020 y año corrido 2021pr.*



Fuente: DANE – Estadísticas Vitales.

Diferentes estudios han abordado la predicción del BPN, desde diferentes perspectivas. En este sentido Victoria et al., 2009 usaron variables morfológicas y fisiológicas de madre, encontrando que los predictores más relevantes fueron la circunferencia del brazo, la pantorrilla entre el segundo y el tercer trimestre, el peso de la madre en el primero y segundo trimestre usando regresión lineal y arboles de decisión, se reporta un R^2 de 0.245. Por su parte Agudelo Pérez et al., (2017) se basaron en variables sociodemográficas y factores de la madre y del recién nacido; encontraron como características relevantes el nivel educativo de la madre y asistir a más de 4 controles prenatales. Emplearon un modelo de regresión logística y no reportaron indicadores de bondad y ajuste del modelo. En otro estudio se empleó la información contenida en los certificados de nacido vivo y se aplicó un modelo de regresión logística para determinar los predictores asociados. Entre sus hallazgos se destacan como variables predictoras el sexo del bebé, nacer por cesárea, el nivel educativo de la madre y la cantidad de controles prenatales; nuevamente no se reportan métricas de desempeño del modelo (Márquez-Beltrán et al., 2013).

1.2. Aproximación desde la analítica de datos

La investigación en las áreas que componen las ciencias de salud a menudo está orientada a encontrar la causalidad de un fenómeno de interés, para luego generar intervenciones que permitan mejorar sus condiciones. Uno de los caminos para lograr este fin; se encuentra en el procesamiento de los datos que permitan un modelamiento matemático, con la intención de

predecir una variable de interés en función de otras características con sospecha de estar relacionadas con la variable a predecir (Mora Pineda, 2022) y así poder encontrar evidencia que permita orientar la causa y el efecto en las variables predictoras, al tiempo que se usa el modelo para prevenir futuros casos y así poder diseñar estrategias de cuidado y prevención.

El modelamiento estadístico en salud clásico tiene dos grandes representantes en función de la naturaleza de la variable respuesta, la Regresión Lineal (RL) para el caso de variables cuantitativas y la Regresión Logística (LR) para cualitativas. En las últimas décadas, el avance de las ciencias de la computación ha logrado entre otras cosas tener una mayor capacidad de almacenamiento y análisis, lo que ha permitido diseñar e implementar nuevas técnicas de modelamiento basadas en procesos algorítmicos conocidos como Aprendizaje de Máquinas o *Machine Learning* (ML), las cuales generan modelos predictivos donde en algunos de los casos el efecto o evidencia de la causalidad se pierde, por lo complejo del procesamiento. Sin embargo, todas las técnicas que conforman el ML permiten crear un modelo para predecir eventos con el fin de poder intervenir en función del pronóstico.

Entre el abanico de técnicas de ML se distinguen enfoques que buscan la predicción de una única salida cuantitativa como es el caso de RL o cualitativa como en la LR, por lo que se clasifican como de única salida o de una sola tarea y un enfoque *Multi-Task* o de múltiple tarea que busca aprovechar la información contenida en los datos para predecir de forma conjunta múltiples salidas relacionadas.

Esta investigación centra su atención en un evento en salud en particular, el bajo peso al nacer y pretende comparar el desempeño de técnicas de ML de múltiple salida y única salida en cuanto a la predicción de este evento, dejando de lado la interpretabilidad de las variables que pueden ofrecer los modelos de regresión clásico. Es decir, el foco de la comparabilidad está en el poder predictivo del modelo más que en su interpretación.

Según Xu et al., (2020) el aprendizaje supervisado tradicional de única salida ha demostrado a lo largo de su historia buenos resultados en la predicción de eventos, es uno de los más usados para realizar predicciones rápidas y precisas. Sin embargo, las demandas y las

complejidades de las aplicaciones de la actualidad requieren un cambio de paradigma, ya que la toma de decisiones a menudo no se basa en una única salida y se requiere usar al tiempo un conjunto de datos para predecir de manera simultánea diferentes variables. En este contexto surge lo que se conoce como los modelos de múltiple salida. Los procesos Gaussiano de múltiple salida o MOGP, son caso particular que apunta en esta dirección; parten del supuesto básico de correlación entre las variables a predecir (inicialmente se consideran cuantitativas) y aprovechan la información de las variables de entrada para realizar predicciones más precisas que si se modelaran de forma independiente (Liu et al., 2018).

Los HETMOGP propuestos por Moreno Muñoz et al., (2018) son una extensión de los MOGP donde las variables son de diversa naturaleza, es decir, se pueden combinar al tiempo salidas reales, binarias o categóricas. En ellos se supone que cada salida tiene su propia función de verosimilitud y se usa un proceso Gaussiano vectorial a priori para modelar conjuntamente los parámetros de todas las verosimilitudes como funciones latentes mediante una función de covarianza. Reciente J. J. Giraldo et al., 2022 proponen usar en la implementación de los HETMOGP procesos Gaussianos encadenados (Saul et al., 2016) que introducen variables latentes mediante el uso de un proceso de convolución (Scharf et al., 2018).

En conclusión, se pretende con esta investigación entrenar y evaluar un modelo **HETMOGP** usando la implementación realizada por J. J. Giraldo et al., 2022 para predecir al tiempo el bajo peso al nacer (en gramos) y bajo peso al nacer (Si/No) como variable dicotómica, en función de variables relacionadas como: características sociodemográficas de los padres, semanas de gestación y numero de controles prenatales y comparar su desempeño con otras técnicas de ML de única salida.

1.3. Origen de los datos

Los datos están publicados en MeData (repositorio de datos oficiales de la Alcaldía de Medellín) tienen licencia *Creative Commons (*CC), está compuesta por más de 250.000 registros y 40 variables de los nacimientos en el municipio de Medellín, comprendidos en los años 2012 al 2021 (http://medata.gov.co/dataset/nacimientos). El evento de interés de esta base de datos es el peso del recién nacido.

2. Objetivos

2.1.Objetivo general

Desarrollar un modelo de múltiples salidas heterogéneo para predecir de manera simultánea el peso del recién nacido y la etiqueta bajo peso al nacer (Si/No) en una muestra de recién nacidos en Medellín entre 2012 y 2021.

2.2.Objetivos específicos

- Diseñar una estrategia de aprendizaje probabilístico basada en procesos Gaussianos heterogéneos de múltiples salidas para la predicción del bajo peso.
- Entrenar un modelo de procesos Gaussianos heterogéneo de múltiples salidas utilizando enfoques de inferencia aproximada.
- Evaluar la validez predictiva del modelo de múltiples salidas frente a técnicas comúnmente usadas en el estado del arte.

3. Datos

3.1. Datos originales

Los datos originales están disponibles en MeData (Alcaldía de Medellín, 2019) repositorio institucional de la ciudad de Medellín que busca la apropiación y apertura y uso de datos como herramienta de gobierno, acción ciudadana y toma de decisiones. La metainformación de los datos se describe en la tabla 1.

Tabla 1: *Información del archivo que contiene conjunto de datos*

Campo	Valor
Dependencias	Salud
Fecha de modificación	10/04/2023
Fecha de publicación	24/11/2020
Frecuencia	Anualmente
Identificador	Nacimientos
Estándar de datos	http://www.mintic.gov.co
Cobertura temporal	De domingo, enero 1, 2012 - 00:00 hasta lunes, diciembre 31, 2021- 00:00
Idioma	Español (Colombia)
Licencia	https://creativecommons.org/licenses/by-sa/4.0/legalcode
Granularidad	Municipal
Autor	Secretaría de Salud Unidad de Gestión de la Información y el conocimiento - Salud Pública
contacto	medata@medellin.gov.co
Nivel de Acceso	Público
Tema POD	Salud
tipo MIME	text/csv
tamaño de archivo	58.21 MB
tipo de recurso	file upload
marca de tiempo	15/09/2023

Fuente: MeData

El fichero original contine información relacionada con los nacimientos en el Municipio de Medellín entre los años 2012 a 2021, cuenta con 272526 registros y 40 variables que se muestran en la tabla 2.

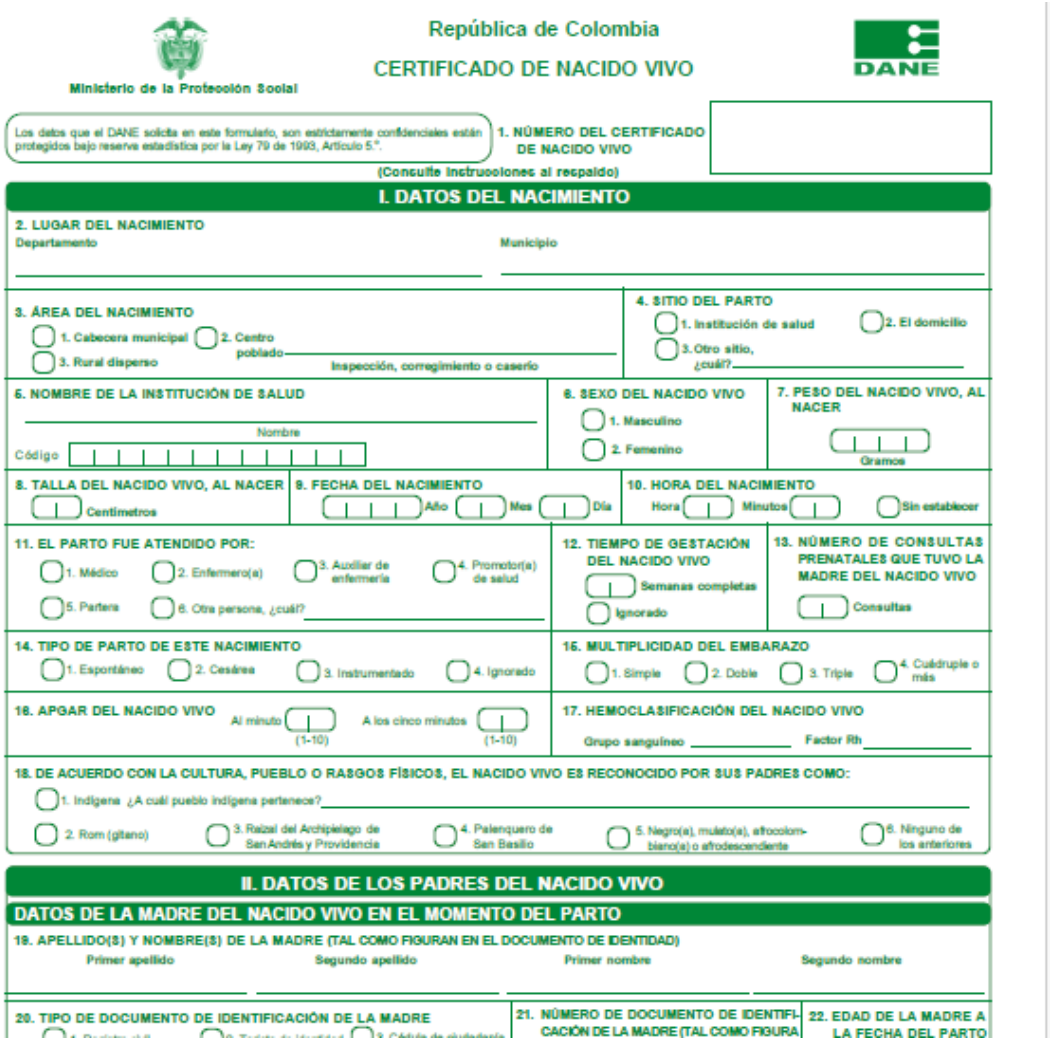
Tabla 2: *Descripción y naturaleza de las variables en el conjunto*

Nombre	Tipo	Descripción
ID	number	Número identificador
AREANAC	string	Área del Nacimiento
COD_INSP	string	Centro poblado del nacimiento (inspección, corregimiento o caserío)
SIT_PARTO	string	Sitio de parto
OTRO_SIT	string	Otro sitio, ¿cuál?
NOM_INST	string	Nombre de la institución de salud
COD_INST	string	Código de la institución de salud
SEXO	string	Sexo del nacido vivo
PESO_NAC	number	Peso del nacido vivo, al nacer
TALLA_NAC	number	Talla del nacido vivo, al nacer
FECHA_NAC	date	Fecha del nacimiento
ANO	string	Año de la ocurrencia
MES	string	Mes de la ocurrencia
ATEN_PAR	string	El parto fue atendido por
OTRPARATX	string	El parto fue atendido por otra persona, ¿Cuál?
T_GES	string	Tiempo de gestación del nacido vivo
NUMCONSUL	string	Número de consultas prenatales que tuvo la madre del nacido vivo
TIPO_PARTO	string	Tipo de parto de este nacimiento
MUL_PARTO	string	Multiplicidad del embarazo
IDHEMOCLAS	string	Hemoclasificación del nacido vivo: Grupo Sanguíneo
IDFACTORRH	string	Hemoclasificación del nacido vivo: Factor RH
IDPERTET	string	De acuerdo con la cultura, pueblo o rasgos físicos, se reconocía como
NOM_PUEB	string	¿A cuál pueblo indígena pertenece?
EDAD_MADRE	number	Edad de la madre a la fecha del parto
EST_CIVM	string	Estado conyugal de la madre
NIV_EDUM	string	Último año de estudio que aprobó la madre
CODPTORE	string	Departamento de residencia habitual de la madre
CODMUNRE	string	Municipio de residencia habitual de la madre
COD_BARRIRES	string	Barrio de residencia del fallecido
N_HIJOSV	number	Número de hijos nacidos vivos que ha tenido la madre
FECHA_NACM	date	Fecha de nacimiento del anterior hijo nacido vivo
N_EMB	number	Número de embarazos, incluido el presente
SEG_SOCIAL	string	Régimen de seguridad social en salud de la madre
IDCLASADMI	string	Entidad Administradora en Salud a la que pertenece la madre
NOMCLASAD	string	Nombre de la Administradora en Salud, a la que pertenece la madre
CODCLASAD	string	Código de la Administradora en Salud, a la que pertenece la madre
EDAD_PADRE	number	Edad del padre a la fecha del nacimiento de este hijo
NIV_EDUP	string	Último año de estudio que aprobó el padre
BARRIO_RES	string	Barrio de residencia de la madre del nacido vivo
COMUNA_RES	string	Comuna de residencia de la madre del nacido vivo

Fuente: MEdata

Las etiquetas de las variables originales publicadas fueron complementadas con la información de la ficha original (ver figura 2) donde según el Ministerio de la Protección Social, 2010 es un documento que registra la información básica del nacimiento del menor. Debe ser diligenciado por el médico o la enfermera que asistió el parto; en las áreas del país en donde no exista este recurso humano, las auxiliares de enfermería y las promotoras de salud están autorizadas para diligenciarlo. Es un requisito básico para luego tramitar el Registro Civil de Nacimiento.

Figura 2: *Ficha de nacido vivo*



Ministerio de la Protección Social

República de Colombia

CERTIFICADO DE NACIDO VIVO

DANE

Los datos que el DANE solicita en este formulario, son estrictamente confidenciales están protegidos bajo reserva estadística por la Ley 79 de 1993, Artículo 5.°.

1. NÚMERO DEL CERTIFICADO DE NACIDO VIVO

(Consulte instrucciones al respaldo)

I. DATOS DEL NACIMIENTO

2. LUGAR DEL NACIMIENTO
Departamento
Municipio

3. ÁREA DEL NACIMIENTO
1. Cabecera municipal
2. Centro poblado
3. Rural disperso
Inspección, corregimiento o caserío

4. SITIO DEL PARTO
1. Institución de salud
2. El domicilio
3. Otro sitio, ¿cuál?

5. NOMBRE DE LA INSTITUCIÓN DE SALUD
Nombre
Código

6. SEXO DEL NACIDO VIVO
1. Masculino
2. Femenino

7. PESO DEL NACIDO VIVO, AL NACER
Gramos

8. TALLA DEL NACIDO VIVO, AL NACER
Centímetros

9. FECHA DEL NACIMIENTO
Año Mes Día

10. HORA DEL NACIMIENTO
Hora Minutos
Sin establecer

11. EL PARTO FUE ATENDIDO POR:
1. Médico
2. Enfermero(a)
3. Auxiliar de enfermería
4. Promotor(a) de salud
5. Partera
6. Otra persona, ¿cuál?

12. TIEMPO DE GESTACIÓN DEL NACIDO VIVO
Semanas completas
Ignorado

13. NÚMERO DE CONSULTAS PRENATALES QUE TUVO LA MADRE DEL NACIDO VIVO
Consultas

14. TIPO DE PARTO DE ESTE NACIMIENTO
1. Espontáneo
2. Cesárea
3. Instrumentado
4. Ignorado

15. MULTIPLICIDAD DEL EMBARAZO
1. Simple
2. Doble
3. Triple
4. Cuádruple o más

16. APGAR DEL NACIDO VIVO
Al minuto (1-10)
A los cinco minutos (1-10)

17. HEMOCLASIFICACIÓN DEL NACIDO VIVO
Grupo sanguíneo
Factor Rh

18. DE ACUERDO CON LA CULTURA, PUEBLO O RASGOS FÍSICOS, EL NACIDO VIVO ES RECONOCIDO POR SUS PADRES COMO:
1. Indígena ¿A cuál pueblo indígena pertenece?
2. Rom (gitano)
3. Raizal del Archipiélago de San Andrés y Providencia
4. Palenquero de San Basilio
5. Negro(a), mulato(a), afrocolombiano(a) o afrodescendiente
6. Ninguno de los anteriores

II. DATOS DE LOS PADRES DEL NACIDO VIVO

DATOS DE LA MADRE DEL NACIDO VIVO EN EL MOMENTO DEL PARTO

19. APELLIDO(S) Y NOMBRE(S) DE LA MADRE (TAL COMO FIGURAN EN EL DOCUMENTO DE IDENTIDAD)
Primer apellido
Segundo apellido
Primer nombre
Segundo nombre

20. TIPO DE DOCUMENTO DE IDENTIFICACIÓN DE LA MADRE

21. NÚMERO DE DOCUMENTO DE IDENTIFICACIÓN DE LA MADRE (TAL COMO FIGURA

22. EDAD DE LA MADRE A LA FECHA DEL PARTO

Fuente: formato_nacido_vivo.pdf (sanandres.gov.co)

3.2. Datsets

Los datos originales están conformados por 40 variables y 272.526 registros. Luego de una selección inicial de variables según relevancia clínica y utilidad en el modelo, una limpieza inicial del conjunto de datos, eliminando registros incompletos, duplicados y atípicos se tiene una base con 20 características y 146.801 de los cuales tan solo 10.9 % (16.055 nacimientos) tienen un bajo peso, por lo que se tiene un desbalanceo importante en la variable de interés por lo que se realizó un muestreo aleatorio en la categoría sin bajo peso al nacer para tener así una base datos balanceada de 32.110 registro y 22 variables que se describen en la tabla 3.

Tabla 3: *Variables incluidas en el análisis*

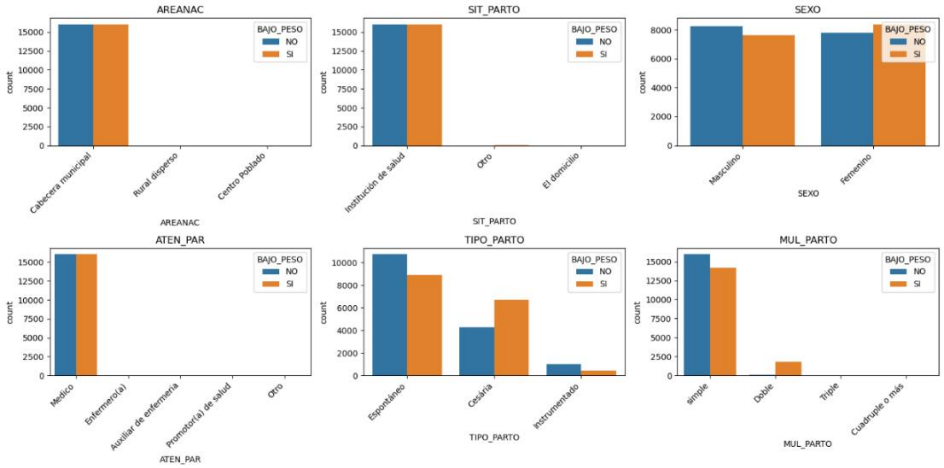
Nombre	**Descripción**
AREANAC	Area del Nacimiento
SIT_PARTO	Sitio de parto
SEXO	Sexo del nacido vivo
MES	Mes de la ocurrencia
ATEN_PAR	El parto fue atendido por
T_GES	Tiempo de gestación del nacido vivo
NUMCONSUL	Número de consultas prenatales que tuvo la madre del nacido vivo
TIPO_PARTO	Tipo de parto de este nacimiento
MUL_PARTO	Multiplicidad del embarazo
IDHEMOCLAS	Hemoclasificación del nacido vivo: Grupo Sanguíneo
IDFACTORRH	Hemoclasificación del nacido vivo: Factor RH
EDAD_MADRE	Edad de la madre a la fecha del parto
EST_CIVM	Estado conyugal de la madre
NIV_EDUM	Último año de estudio que aprobó la madre
N_HIJOSV	Número de hijos nacidos vivos que ha tenido la madre
N_EMB	Número de embarazos, incluido el presente
SEG_SOCIAL	Régimen de seguridad social en salud de la madre
EDAD_PADRE	Edad del padre a la fecha del nacimiento de este hijo
NIV_EDUP	Último año de estudio que aprobó el padre
COMUNA_RES	Comuna de residencia de la madre del nacido vivo
PESO_NAC	**Peso del nacido vivo, al nacer**
BAJO_PESO	**Si el peso del bebé es inferior a 2500g se clasifica como bajo peso**

Donde las variables a predecir son el peso del recién nacido en gramos **PESO_NAC** y la clasificación del bajo peso al nacer Si/No, **BAJO_PESO**. Las demás variables son consideradas como variables predictoras para los modelos.

3.3.Analítica descriptiva

Para el análisis descriptivo inicial de los datos las variables predictoras fueron agrupadas según su naturaleza cuantitativa o descriptiva. En cada caso se exploraron distribuciones según las variables de interés en busca de indicios de relaciones o posibles correlaciones entre las variables.

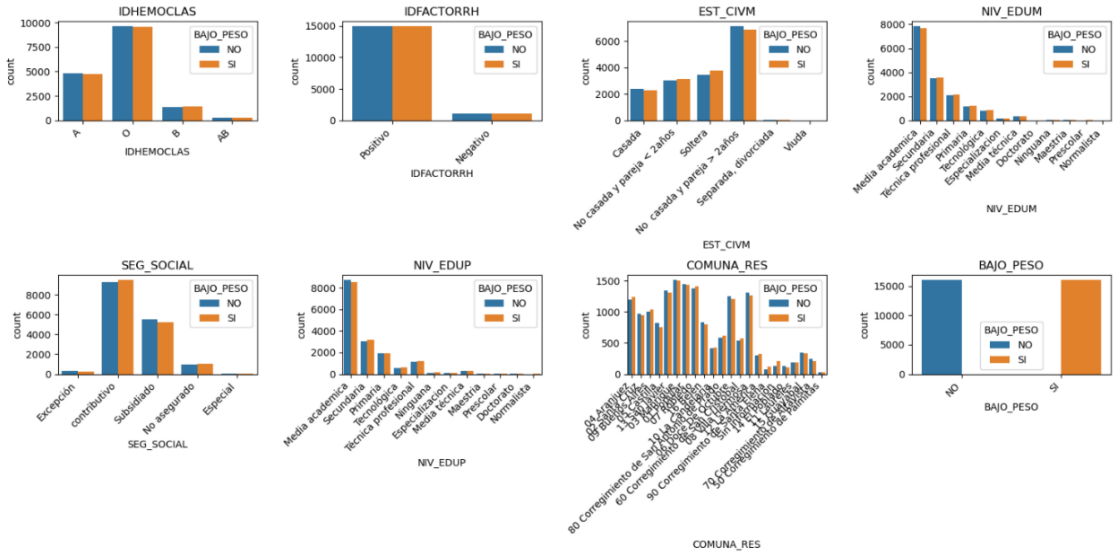
Figura 3: *Gráficos de barras para grupo 1 de variables cualitativas vs bajo peso al nacer*



Fuente: Elaboración propia

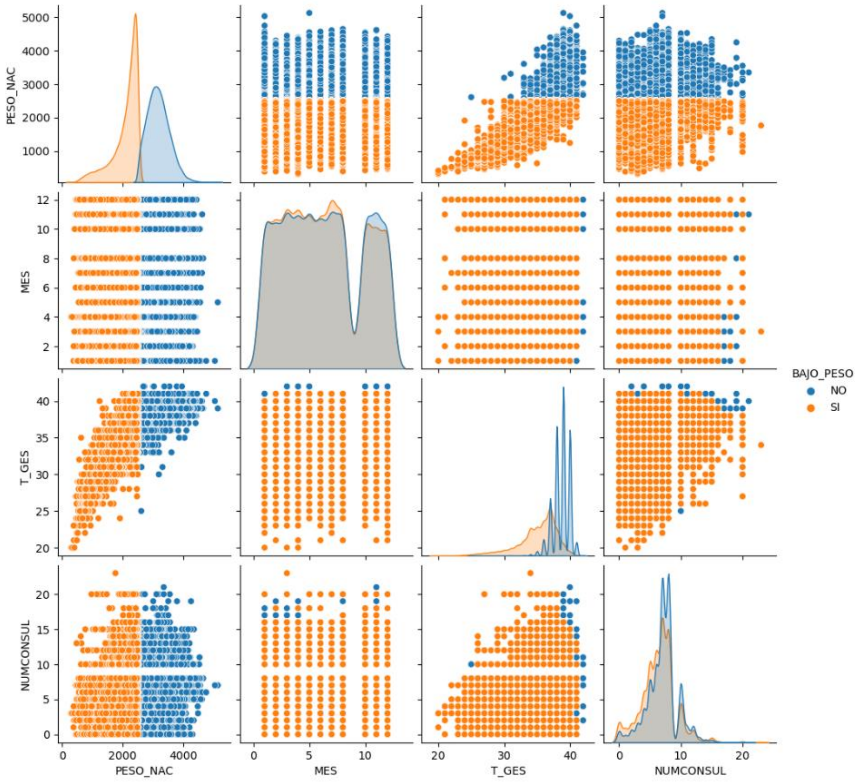
De los gráficos de las presentados en las figuras 4 y 5 resalta como el bajo peso al nacer tiene una mayor incidencia en el sexo femenino, en tipos de partos espontáneos y en tipo de sangre O+. Así mismo se reportan mayor número de casos cuando la madre reside en comunas como Manrique, Popular o Robledo.

Figura 4*: Gráficos de barras para grupo 2 de variables cualitativas vs bajo peso al nacer*



Fuente: Elaboración propia

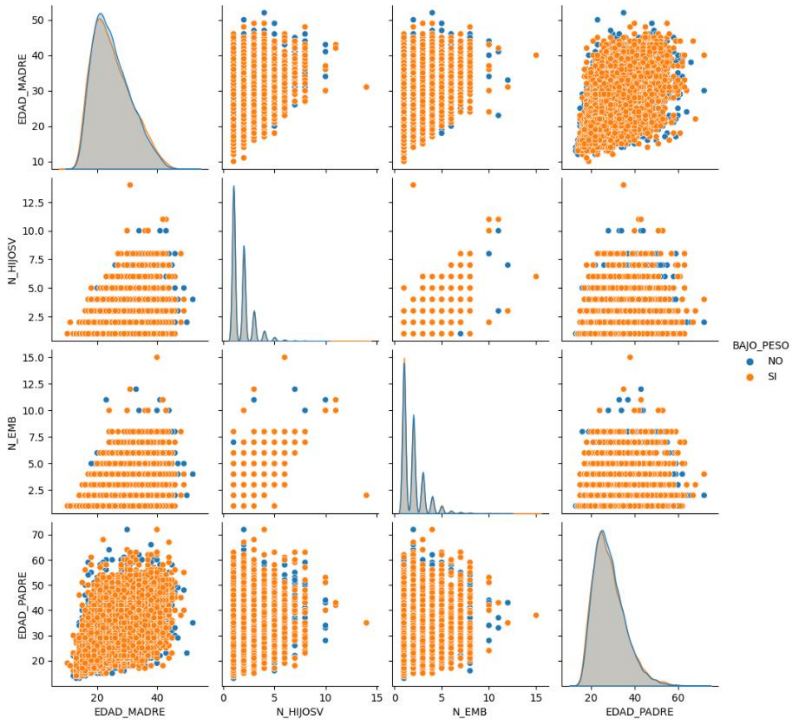
Figura 5: *Diagramas de dispersión y distribución para el grupo 1 variables cuantitativas.*



Fuente: Elaboración propia

En las figuras 5 y 6 se presentan los gráficos de dispersión y distribución del grupo de variables cuantitativas, en ellos se puede apreciar como las semanas de gestación tienen una aparente correlación positiva con el peso del bebé. La distribución de los nacimientos con bajo peso es mayor entre los meses de marzo y agosto y se no observan indicios de correlación entre las variables predictoras, lo cual es deseable para el modelamiento en la mayoría de las técnicas de ML, para evitar problemas de multicolinealidad o redundancia que afecten la veracidad de las métricas de desempeño.

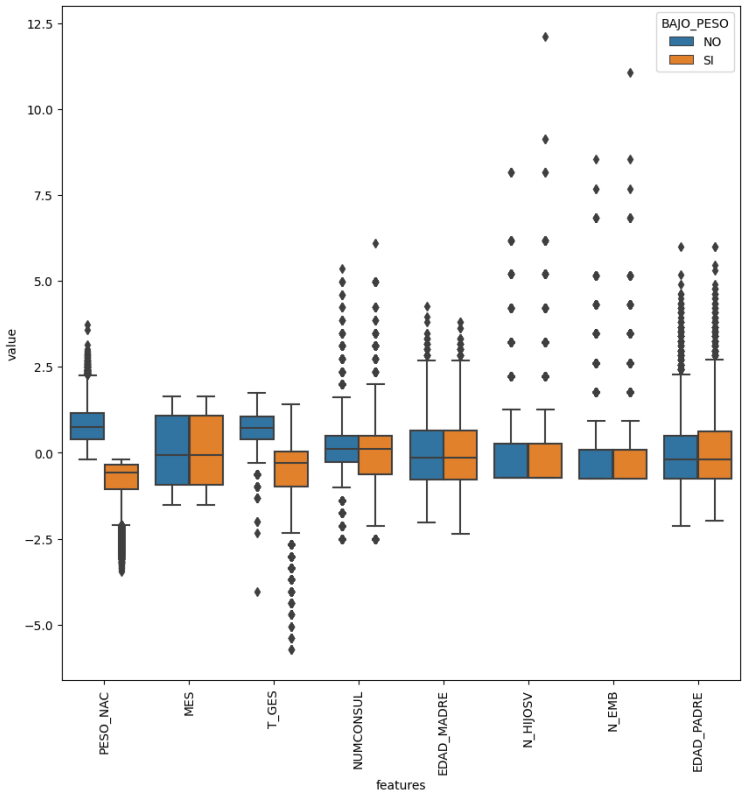
Figura 6: *Diagramas de dispersión y distribución para el grupo 2 variables cuantitativas.*



Fuente: Elaboración propia

En la figura 7 se observa nuevamente como las semanas de gestación tienen una incidencia en la clasificación del bajo peso al nacer, en tanto las cajas de los diagramas es superior para nacimientos con más semanas de gestación. Es decir, los nacimientos que no fueron clasificados como bajo peso tienen más semanas de gestación en comparación con los que nacieron con bajo peso.

Figura 7: *Diagramas de cajas y bigotes para variables cuantitativas (estandarizadas) según clasificación de bajo peso al nacer.*



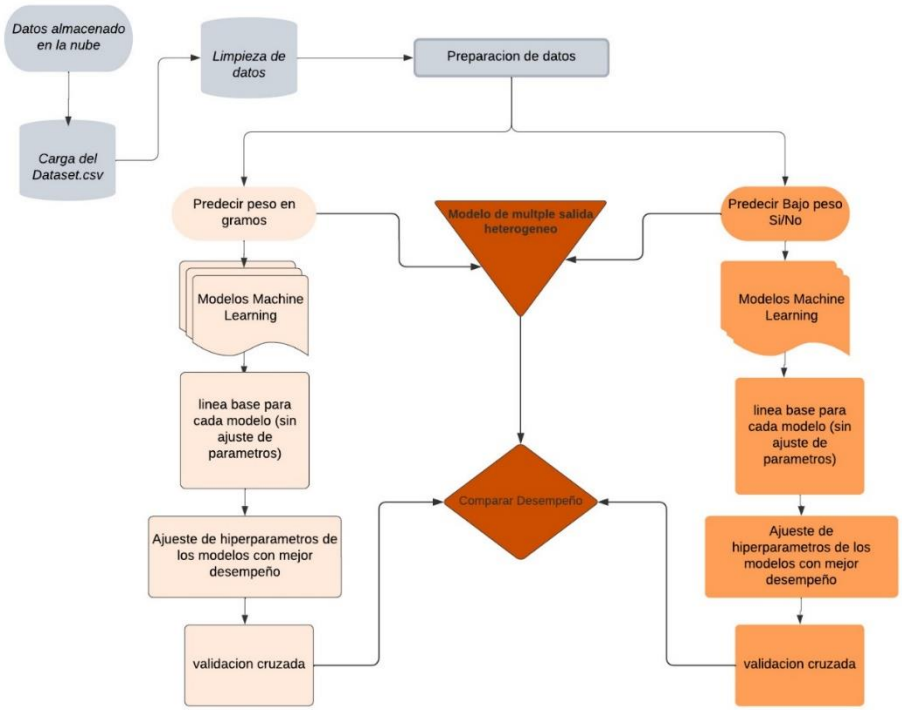
Fuente: Elaboración propia

4. Proceso de analítica

4.1. Pipeline principal

El flujo de trabajo parte de los datos almacenados en la nube (repositorio institucional MeData) luego de descargar los datos se procede a limpieza de estos eliminando duplicados, datos incompletos e inconsistencias. Se preparan los datos identificando la naturaleza de cada variable junto con sus etiquetas según la ficha del nacido vivo. En este punto el trabajo se divide en dos frentes: la predicción del bajo peso al nacer en gramos y la clasificación según bajo peso (Si/No), en cada caso se prueban modelos tradicionales de única salida usando la mayoría de los parámetros por defectos y se evalúa el desempeño del modelo empleando una división aleatoria de los datos 70/30 para entrenamiento y validación dando como resultado la línea base de desempeño.

Figura 8: *Flujo de trabajo*



Fuente: Elaboración propia

Los dos métodos con mejor desempeño en la predicción del peso del bebé y en la clasificación del bajo peso según la línea base, son seleccionados para realizar un ajuste de parámetros y validación cruzada. Paralelo a este proceso se entrena y valida un modelo ML de múltiple salida

para predecir el peso y clasificación del bajo peso al tiempo y se comparan los resultados como se plantea en la figura 8.

4.2. Preprocesamiento

Los datos fueron filtrados eliminando duplicados, incompletos y atípicos. Luego se de este paso se realizó una selección de características según relevancia clínica y posibilidad de ser medida antes del del nacimiento del bebé, para darle sentido a los modelos predictivos. Posterior a este paso se realizó un balance de variables según característica minoritaria de bajo peso, se seleccionó de manera aleatoria una muestra de igual tamaño en la característica sin bajo peso al nacer.

Antes de entrenar los modelos se le aplico una transformación *one-hot encoding*, para crear variables *dummies* o ficticias de ceros y unos que reflejen de forma numérica la variabilidad de las categorías, en el caso particular del modelo HETMOGP, dada la complejidad en la optimización de la función de pérdida se usó *Label Encoding* (asigna un valor numérico a cada categoría de la variable) para no incrementar la dimensionalidad del problema. Las características numéricas fueron transformadas usando método MinMaxScaler(), para lograr equilibrio en las escalas numéricas de las variables. Es decir, se espera que variables como edad y número de consultas tengan una escala similar luego de la transformación.

4.3.Modelos

Se presenta una breve descripción sobre los conceptos teóricos relacionados con algunos de los modelos implementados.

4.3.1 HETMOGP

Para tener una idea de las nociones básicas detrás de los Procesos Gaussiano Heterogéneos de Múltiple Salida, es necesario definir algunos conceptos previos tomando como referencia el texto *Gaussian Processes for Machine Learning* (Rasmussen & Williams, 2006) y las notas clase del curso *Machine Learning for Intellingnet Systems* de la Universidad de Cornell (Weinberger, 2018).

Definición 1: Un Proceso Gaussiano (GP) es una colección infinita de variables aleatorias (VA), donde cualquier subconjunto finito de VA tiene una distribución Gaussiana multivariada.

De la definición se infiere que todo proceso Gaussiano está completamente especificado por su media y su covarianza. De donde, si $f(\boldsymbol{x})$ es un GP con media $m(\boldsymbol{x})$ y función de covarianza $k(\boldsymbol{x}, \boldsymbol{x}')$ se tiene por definición que:

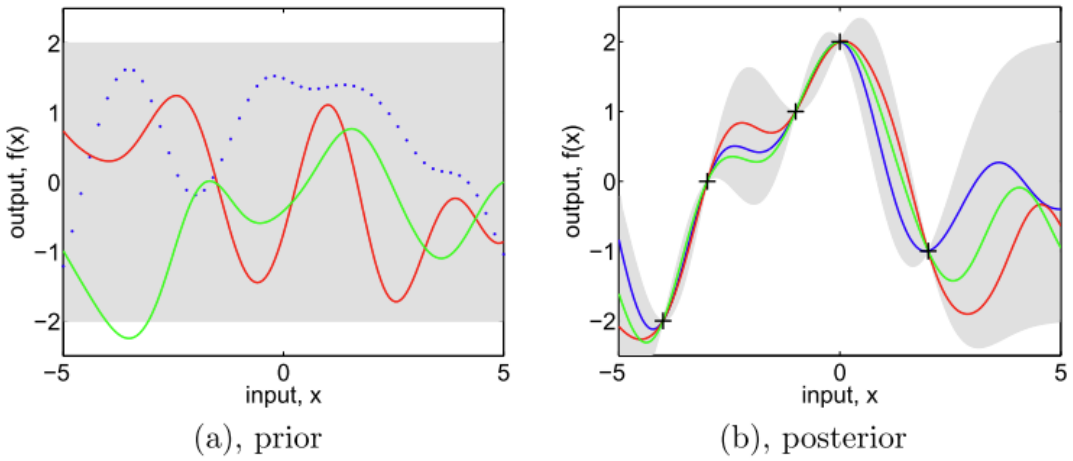
$$m(x) = \mathbb{E}[f(x)],$$
$$k(x, x') = \mathbb{E}\big[\big(f(x) - m(x)\big)\big(f(x') - m(x')\big)\big].$$

En adelante se usa la siguiente notación para referirse a $f(\boldsymbol{x})$ como GP.

$$f(x) \sim \mathcal{GP}\big(m(x), k(x, x')\big).$$

Ahora, con el fin de modelar la distribución predictiva de un nuevo elemento dado el conjunto de entrenamiento y una nueva observación $P(f_* \mid x_*, D)$ podemos utilizar un enfoque bayesiano mediante el uso de un GP a priori: $P(f \mid x) \sim \mathcal{N}(\mu, \Sigma)$ y condicionar a los datos de entrenamiento D para modelar la distribución conjunta de $f = f(X)$ (vector de observaciones de entrenamiento) y $f_* = f(x_*)$ (predicción en la entrada de prueba) ver figura 9.

Figura 9: *En (a) se presentan tres funciones extraídas al azar de un GP a priori y en (b) se presenta las funciones condicionadas a cinco puntos observados.*



Fuente: Gaussian processes for machine learning (Rasmussen & Williams, 2006)

Definición 2: **Regresión de Proceso Gaussiano**. Suponga que antes de observar las etiquetas de entrenamiento, estas se extraen de la distribución Gaussiana a priori de media cero, es decir

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}(0, \Sigma).$$

(La media cero es siempre posible restando la media de la muestra. Todas las etiquetas de entrenamiento y prueba se extraen de una distribución Gaussiana de dimensión $(n + m)$, donde n es el número de puntos de entrenamiento y m es el número de puntos de prueba). De donde se puede inferir que las etiquetas de entrenamiento, $y_1, \dots, y_n$, son muestras de $Y_1, \dots, Y_n$

Ahora es necesario considerar las siguientes propiedades para Σ :

1. $\Sigma_{ij} = E\left((Y_i - \mu_i)(Y_j - \mu_j)\right)$.

2. Σ es siempre semidefinido positivo.

3. $\Sigma_{ii} = Varianza(Y_i)$ siempre que $\Sigma_{ii} > 0$.

4 Si Y_i y Y_j son muy independientes, es decir, x_i es muy diferente de x_j, entonces

$$\Sigma_{ij} = \Sigma_{ji} = 0.$$

5. Si x_i y x_j son similares entonces

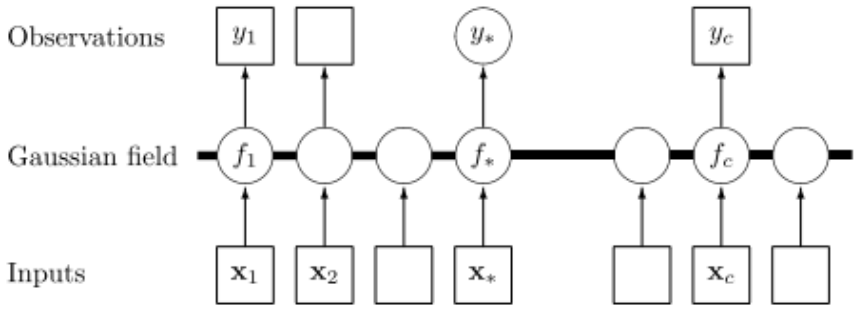
$$\Sigma_{ij} = \Sigma_{ji} > 0.$$

Note que lo anterior es similar a la matriz del kernel en las SVMs. De donde si se supone que efectivamente $\Sigma_{ij} = K(x_i, x_j)$ y se usa por ejemplo un kernel RBF (también conocido como "kernel exponencial cuadrado"), se tiene que $\Sigma_{ij} = \tau e^{\frac{-|x_i - x_j|^2}{\sigma^2}}$ y si fuese un kernel polinómico, $\Sigma_{ij} = \tau(1 + x_i^\top x_j)^d$. Luego, se evidencia que Σ se puede descomponer como $\begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}$, siendo K es la matriz del kernel de entrenamiento, K_* es la matriz del kernel de entrenamiento-prueba, $K_*^\top$ es la matriz del kernel de prueba-entrenamiento y K_{**} es la matriz del kernel de prueba. La distribución condicional de los valores (sin ruido) de la función latente f se puede escribir como:

$$f_* \mid (Y_1 = y_1, \dots, Y_n = y_n, x_1, \dots, x_n, x_t) \sim \mathcal{N}(K_*^\top K^{-1} y, K_{**} - K_*^\top K^{-1} K_*),$$

donde las matrices del kernel K_*, K_{**}, K son funciones de $x_1, \dots, x_n, x_*$ (ver figura 10)

***Figura 10**: Esquema gráfico de un GP de regresión.*



Fuente: Gaussian processes for machine learning (Rasmussen & Williams, 2006)

Definición 3: **Modelo de corregionalización**

Dado un problema de regresión para funciones $f: R^D \rightarrow R^P$ y un conjunto de datos de la forma $(X_1, f_1), \ldots, (X_P, f_P)$ (donde, no necesariamente se observan todas las salidas para una ubicación de entrada en particular). Suponga que cada f_i tiene una distribución de ruido diferente asignando una probabilidad diferente a cada una. Si f se modela como un GP, con un kernel de la forma:

$$cov\left(f_i(X), f_j(X')\right) = k(X, X') \cdot B[i, j].$$

Donde la covarianza de la i-esima función de X y de la j-esima function de X' es un kernel aplicado a X y X', para la (i, j) -esima entrada de la matriz B de dimensión $P \times P$ y definida positiva. Se tiene entonces lo que se conoce como Modelo Intrínseco de Corregionalizacion (ICM) (Bonilla et al., 2008)

Para asegurarse de que B es definida positiva se define:

$$B = WW^{\top} + diag(\kappa).$$

Definición 4: **Proceso de convolución para generar los parámetros de las funciones latentes**

Suponga que en modelo de corregionalización una función de salida particular $f_j(X)$ depende de todos los valores continuos de una función latente derivada de un GP $\{u_q(\cdot)\}_{q=1}^{Q}$ de la siguiente manera

$$f_j(X) = \sum_{q=1}^{Q} \sum_{i=1}^{R_q} \int_{\chi} G_{i,q}^{i}(x - r') u_q^i(\mathrm{r}')\mathrm{dr}' .$$

Donde $u_q^i(x)$ son muestras independiente idénticamente distribuidoras de GP $u_q(\cdot) \sim \mathcal{GP}\left(0, k_q(\cdot,\cdot)\right)$ y cada $G_{i,q}^i(\cdot)$ representa un kernel suavizado.

Definición 5: **Proceso Gaussiano encadenado**

Considere la distribución de $p(y_n|f(x_n))$ consiste en encadenar un GP previo a cada parámetro de la verosimilitud para formular un modelo conjunto como sigue

$$p(y,f) = p(y|f)p(f) = \prod_{n=1}^{N} p(y_n|\psi_1(x_n),\cdots,\psi_J(x_n))p(f).$$

Donde cada $\psi_j(x_n) = \alpha(f_j(x_n))$ representa el j-esimo parámetro de la verosimilitud encadenado al priori de un GP latente $f_j(x) \sim \mathcal{GP}\big(m(x), k(x,x')\big)$, a través de una función de enlace $\alpha(\cdot)$ y J representa el número de parámetros de la verosimilitud.

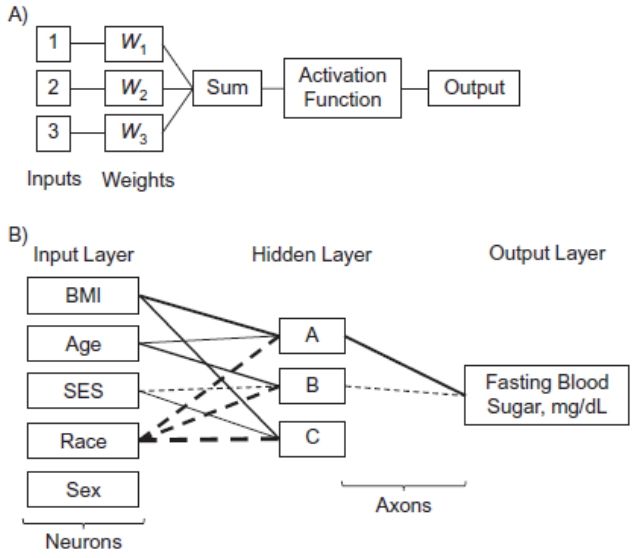
En tanto en esta investigación se usa la implementación de HETMOGP propuesta por (J. J. Giraldo et al., 2022) basados en un GP encadenado usando un modelo lineal de corregionalizacion basado en un proceso de convolución.

4.3.2 Otros modelos de ML

Redes Neuronales Artificiales o Artificial Neural Networks

Las redes neuronales artificiales se inspiran en el comportamiento de las redes neuronales biológicas. La analogía que subyace en esta comparación es que un grupo de neuronas interconectadas a través de vías de señalización complejas pueden ser utilizadas para analizar interacciones entre un grupo de covariables medibles para predecir un resultado. Las neuronas en capas adyacentes se comunican con otras a través de funciones de activación, que convierten la ponderación de las entradas de una neurona en una salida (figura 11). La suma ponderada de las entradas de una neurona es algo análogo a los coeficientes en las regresiones lineales o logísticas (Bi et al., 2019).

Figura 11: *Representación de una sola neurona y una aplicación de una red neuronal artificial para predecir los niveles de azúcar en la sangre en ayunas en función de variables clínicas y demográficas.*

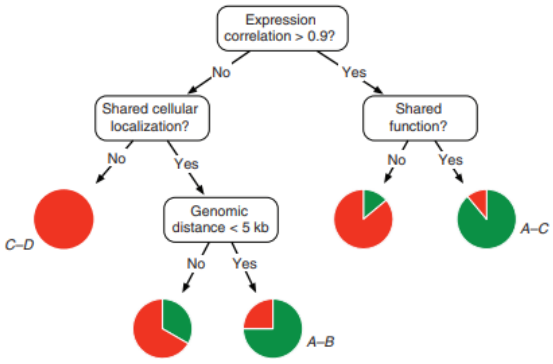


Fuente: What is machine learning? A primer for the epidemiologist (Bi et al., 2019)

Arboles de Decisión o *Decision Trees*

Los Árboles de Decisión fueron uno de los primeros modelos de *Machine Learning* en implementarse. La idea central de esta técnica es analizar que covariable separa mejor la variable a predecir (rama) y luego replica la idea de manera sucesiva sobre las demás covariables y así predecir reglas de decisión en función de un orden jerárquico las covariables o ramas (Kingsford & Salzberg, 2008). Una estructura sencilla se ilustra en la figura 12.

Figura 12: *Ejemplo hipotético de como un árbol de decisión predice la interacción entre proteínas.*

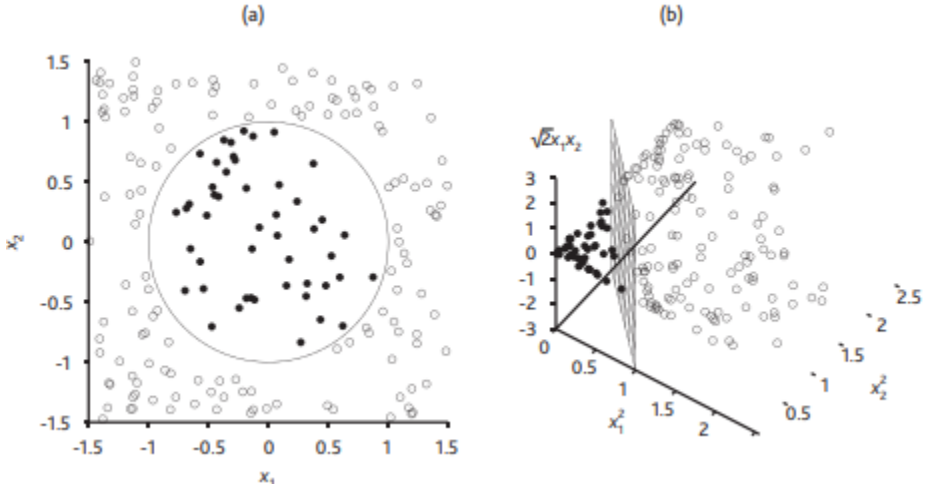


Fuente: What are decision trees? (Kingsford & Salzberg, 2008)

Máquinas de Soporte Vectorial o Support Vector Machines

Las Máquinas de soporte vectorial basan su idea de clasificación en el principio de que todo conjunto de datos es linealmente separable en un espacio de dimensión mayor y deriva su regla de clasificación en la proyección del hiperplano lineal de separación en el espacio original de los datos (Salazar et al., 2012) (ver figura 13)

***Figura 13**: Representación de una clasificación de datos en dos dimensiones que son discriminados luego de una transformación de SVM en un espacio de tres dimensiones.*



Fuente: Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?(Salazar et al., 2012)

4.4.Métricas

Para el caso de la predicción del peso del bebé (modelos predictivos de regresión) se contempla usar el MAE (Error Medio Absoluto) definido por la formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n-1}|y_i - \hat{y}_i|$$

donde n es el número de muestras de los datos de prueba y i es la i-ésima muestra de los datos de prueba. Y el coeficiente de determinación R^2 definido por

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})}$$

Para evaluar el desempeño de los modelos de clasificación se usarán el *accuracy* o exactitud definida como:

$$Accuracy = \frac{NPC}{NPT}$$

Donde NPC representa el número total de predicciones correctas (verdaderos positivos + verdaderos negativos) y NPT es la cantidad total de predicciones (verdaderos positivos + verdaderos negativos + falsos positivos + falsos negativos). Y el AUC que se define el área bajo la curva ROC, la cual es una representación gráfica de la sensibilidad (verdaderos positivos/total positivos) frente a la especificidad (verdaderos negativos / total de negativos) del sistema de clasificación binario.

5. Metodología

El enfoque metodológico, parte de considerar el desempeño de los modelos tradicionales de ML en la predicción de cada salida (peso en gramos y clasificación de bajo peso) de forma independiente, como línea base o punto de partida para la evaluación del modelo HETMOGP. Por consiguiente, esta sección se divide en dos apartados donde se describen los procesos metodológicos para los modelos de ML de única salida y de múltiple salida.

Una vez definida una línea base del desempeño con los modelos de ML convencionales, se procede a entrenar y ajustar un Proceso Gaussiano Heterogéneos de Múltiple Salida. Finalmente, se comparan resultados y se analiza si el modelo de múltiple salida ofrece mejores o iguales resultados, frente a técnicas de la línea base.

5.1. Enfoques generales para modelos de única salida

Se usan técnicas convencionales de ML de única salida real o binaria como la Regresión Lineal y Logística, Arboles de Decisión y SVM donde se incluyen todas las variables en los modelos de regresión y clasificación luego del preprocesamiento descrito en la sección 4.2.

5.1.1 Baseline

Los experimentos iniciales parten de una evaluación de las métricas luego de un entrenamiento y validación convencional 70/30, usando la mayoría de los parámetros por defecto (ver tabla 4) de cada método y todas las variables predictoras. Luego los dos modelos con mejor desempeño según las métricas se les aplica una búsqueda por cuadricula o *grid search* para ajustar de parámetros y evaluar si se logra una mejoría en las métricas. En caso de que los métodos seleccionados no dependan de parámetros como en la Regresión Lineal o Logística se realiza tratamiento de datos, transformaciones en la variable de salida o reducción de dimensionalidad por PCA.

***Tabla 4**: Parámetros por defecto usados en las modelos de única salida para la tarea de regresión y de clasificación.*

Regresión	
Método	**Parámetros**
Regresión Lineal	`fit_intercept=True, normalize=False, copy_X=True, n_jobs=None`
Árboles de Decisión	`criterion='mse', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1,min_weight_fraction_leaf=0.0`
SVM	`C=1000, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=1e-3, random_state=None`
Clasificación	
Regresión Logística	`penalty='l2', dual=False, tol=1e-4, C=1.0, fit_intercept=True, intercept_scaling=1, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto'`
SVM	`C=1.0, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=1e-3, random_state=None`
KNN	`n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None`
ComplNB	`alpha=1.0, fit_prior=True, class_prior=None, norm=False`
Arboles de Decisión	`criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0`
RandForest	`n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', random_state=None`
MLPClassifier	`hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, max_iter=200, shuffle=True, random_state=None, tol=1e-4`

Fuente: Elaboración propia

El desempeño de los cuatro modelos (dos para regresión y dos para clasificación), conforman la línea base del enfoque única salida, para cada tarea.

5.1.2 Validación

Una vez seleccionados y ajustados los parámetros de los métodos con mejor desempeño se procede a validarlo, mediante la evaluación de la incertidumbre los resultados realizando validación cruzada.

5.1.3 Iteraciones y Evolución

En las iteraciones de la validación cruzada se consideran todos los datos para forman de manera aleatoria 10 pliegues donde cada pliegue forma una división aleatoria de datos de entrenamiento y validación. En cada iteración o pliegue se reportan las métricas y al final el promedio y desviación estándar de los 10 pliegues constituyen la línea de comparación para los modelos de múltiple salida.

5.1.4 Herramientas

En esta fase se usa el entorno de Anaconda y Jupyter. Anaconda, es una distribución de código abierto, para la gestión de paquetes y la creación de ambientes virtuales, lo cual facilita la configuración y reproducibilidad de los experimentos realizados. Por su parte, Jupyter, es un entorno interactivo basado en *web*, que permitirte la creación de cuadernos (*notebooks*) interactivos que integraron código, visualizaciones y texto explicativo. Esta combinación de Anaconda y Jupyter proporciona una plataforma robusta y eficiente para llevar a cabo los procesos de exploración de datos, implementación de modelos y presentación de resultados de manera clara y ordenada.

Para la implementación de los modelo se usan: Librerías, *Pandas, Numpy* y *Scikit-learn*, y para la visualización *Matplotlib* y *Seaborn.*

5.2.Enfoques generales basados en proceso Gaussianos de múltiple salida

Si bien para generar la línea base, en los escenarios de única salida se usan todas las variables, para el enfoque basado en múltiple salida dada la complejidad computacional en la optimización en la función de pérdida; se realiza un análisis de importancia de características previo al entrenamiento por medio de un *Light Gradient Boost Machines*, o LGBM, ideal cuando se tiene bases de datos de gran tamaño con mezcla de variables cuantitativas y cultivas (Ke et al., 2017) .

5.2.1 Baseline

En primer lugar, se realiza una selección de característica utilizando un enfoque de predicción binaria del cual se seleccionan las características con mayor relevancia y que juntas logran una importancia acumulada de al menos el 80%. Luego, se procede a entrenar y validar un primer experimento con una configuración inicial de parámetros donde se selecciona un modelo con dos salidas: una de tipo Gaussiana heterogénea que según J. J. Giraldo et al., 2022 permite modelar de manera flexible salidas de tipo continuo (como es el caso de la regresión). Para la segunda salida ya se tiene, una variable binaria para el bajo peso se elige una salida de tipo Bernoulli. Adicionalmente, según pruebas iniciales de ajuste parámetros y tiempos de ejecución,

se optó por el optimizador Adam (Estimación Adaptativa de Momentos) por sus beneficios de rápida convergencia. Finalmente, algunos parámetros de configuración son:

- 250 instancias para cada lote (*Minibatch*)
- Se seleccionan 50 puntos inducidos los cuales permiten estimar el posterior a partir de variables latentes informativas (Puntos inducidos)
- 3000 iteraciones
- Dimensionalidad de un espacio latente que captura la estructura subyacente en los datos (Q=3)

5.2.2 Iteraciones y Evolución

Luego de generar la línea base se realizan experimentos donde se varían los parámetros relacionados con el número de iteraciones, puntos inducidos, tamaño de los lotes, dimensionalidad del espacio latente. En cada caso se reporta el error cuadrático medio normalizado como métrica combinada del desempeño de predicción del peso del recién nacido (regresión) y predicción del bajo peso al nacer (clasificación). Así mismo, se presenta para cada escenario de salida el R^2 y la exactitud como métricas independientes para poder comparar con los resultados de única salida.

5.2.3 Validación

Como validación complementaria de los resultados obtenidos se propone variar el optimizador junto con los demás parámetros, con el fin de valorar si el desempeño mejora o es estable a este tipo de cambios.

5.2.4 Herramientas

Para este caso todas las pruebas se corrieron en *Google Collaboratory*, el cual es un proyecto de código abierto basado en Jupyter que proporciona una máquina virtual para la ejecución de código Python en la versión 3.7.13. Se hace uso de las siguientes librerías para la implementación de los HETMOG: *Gpy, toolbox* de los procesos Gaussianos, *Climin* el cual es un *toolbox* de optimización, las implementaciones de Moreno Muñoz et al., 2018 *hetmogp* y *likelihoods* que permiten la construcción del modelo de múltiple salida y de J. J. Giraldo et al., 2022 se usa

convhetmogp_vik_fullcov para la modificación del proceso Gaussiano encadenado y para el tratamiento de datos y visualización: Pandas, *Numpy* , *Matplotlib* y *Seaborn*.

6. Resultados y discusión

En este capítulo se inicia presentando la línea base del desempeño de diferentes modelos de ML de única salida real (peso del recién nacido en gramos) o binaria (clasificación del bajo peso al nacer Si/No) usando la mayoría de los parámetros por defecto de cada técnica y división 70/30 para el entrenamiento y validación. Posterior a esto se eligen los dos métodos con mejores métricas para ser entrenados ajustando los parámetros o realizando tratamiento a los datos según corresponda.

Una vez se presenta la línea base de las métricas de desempeño de las técnicas de única salida se procede a entrenar y validar un modelo de múltiple salida para predecir de manera conjunta el peso del recién nacido y la clasificación del bajo peso al nacer. Finalmente se presentan una discusión y comparación de los resultados obtenidos.

6.1.Modelos de ML para la predicción del peso recién nacido.

En tabla 5 se evidencia como los métodos como mayor R^2 y menor MAE son la Regresión Lineal y las Máquinas de Soporte Vectorial con un kernel polinomial de segundo grado. Sin embargo, los Árboles de Decisión y los demás kernel del SVM ofrecen resultados muy cercanos a los de mejor desempeño, excepto por las SMV con un kernel sigmoidal con el cual se tiene el peor desempeño según las métricas seleccionadas.

Tabla 5: *Desempeño de diferentes técnicas de ML en la predicción del peso del recién nacido*

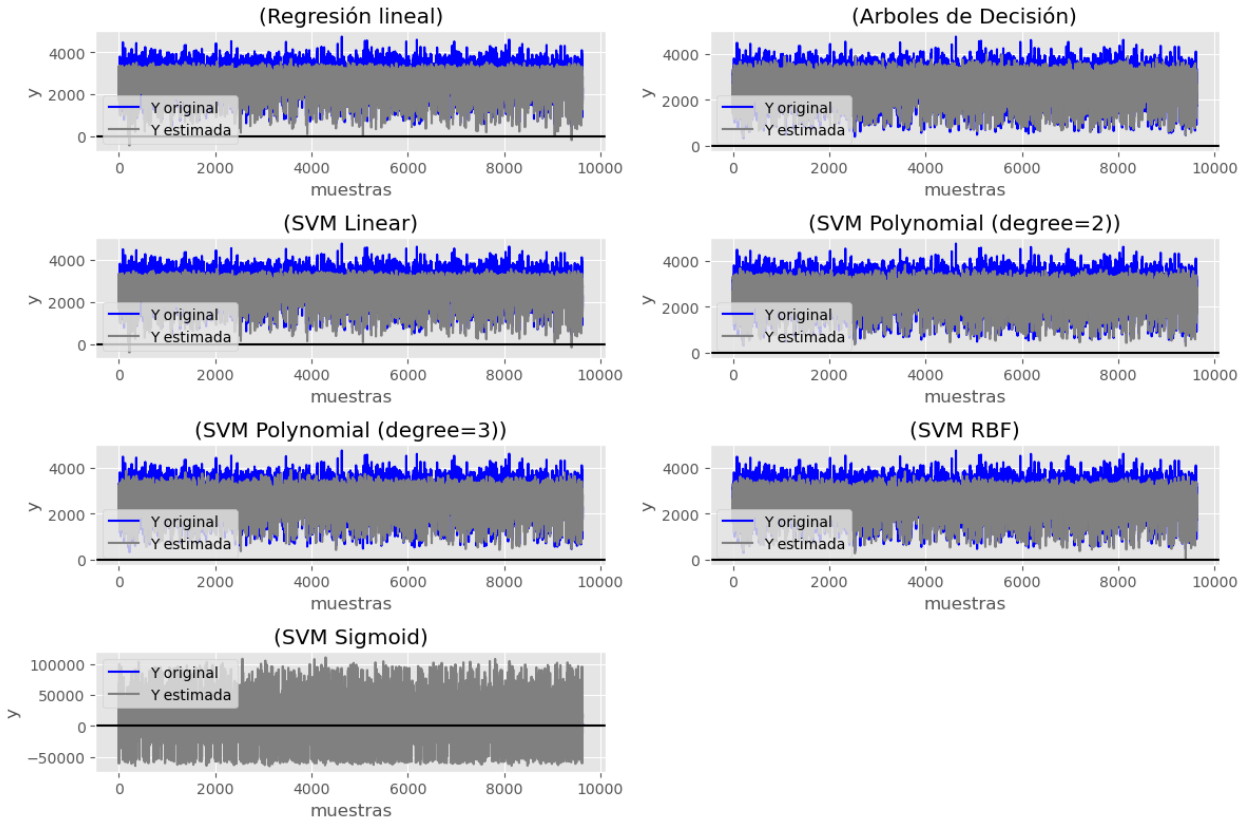
Método	R^2	MAE
Regresión Lineal	0.6720	300.89
Árboles de Decisión	0.6423	307.56
SVMLineal	0.6703	301.01
SVMPoly 2	0.6718	299.13
SVMPoly 3	0.6604	304.41
SVMRBF	0.6659	302.57
SVMSigmoidal	-20.940.388	22896.10

Fuente: Elaboración propia

Este comportamiento se confirma al observan en la Figura 14 como las predicciones (franja gris) de la Regresión Lineal, los Árboles de decisión y las SVM tienden a adaptarse de manera aceptable a los valores reales (franja azul). Llama la atención de este gráfico que los modelos les

cuestan más predecir de forma correcta los pesos de bebés cercanos a los 4000 gramos. Nuevamente se confirma en el gráfico los malos resultados del kernel sigmoidal.

Figura 14: *Comparación de pesos reales versus predichos por los modelos de ML*



Fuente: Elaboración propia

6.1.1. Ajuste de los parámetros y tratamiento de datos SVM y RL

En este apartado se eligen los dos modelos que en la sección anterior arrojaron las mejores métricas y se procede en el caso de las SVM con kernel polinomial se segundo grado a realizar exploración variando parámetro de regularización C en conjunto de valores $[0.001, 0.01, 0.1, 0.3, 0.5, 1, 2, 3, 4, 5, 7, 10, 100, 1000]$ realizando una validación cruzada en cada caso con 10 capas. Para el caso de la Regresión Lineal (RL) se consideró realizar una reducción de dimensionalidad mediante un Análisis de Componentes Principales y una transformación raíz cubica en la variable de respuesta (peso del recién nacido). Los resultados se presentan en la tabla 6.

Tabla 6: *Métricas luego de ajuste y transformación de los datos en la predicción del peso del recién nacido.*

Método	R^2	MAE
SVMPoly 2 C=1000	0.6718	299.13
RL (PCA)	0.5497	356.768
RL (raíz cúbica)	0.7309	0.5205*

Fuente: Elaboración propia *Las unidades de esta métrica no corresponden a la de la variable original

Luego del ajuste se observa que se tienen resultados similares para las SVM y un menor valor para RL (PCA), por lo que se mantienen la configuración inicial para las técnicas seleccionadas con mejor desempeño. Es importante mencionar que, aunque al transformar el peso del recién nacido aplicando una raíz cúbica se obtiene una aparente mejoría en el MAE, esta mejoría no es real ya que no es comparable con resultados de los demás métodos ya que las unidades son diferentes a las de la variable original (g vs $\sqrt[3]{g}$).

6.1.2 Validación Cruzada SVM y RL

Se realiza una validación de los métodos seleccionados mediante validación cruzada con enfoque de *k -fold*, donde el conjunto de entrenamiento se divide en k conjuntos más pequeños para lograr evaluación más confiable del desempeño del modelo. En la tabla 7 se presentan los resultados usando 10 pliegues; se reporta las métricas en cada paso.

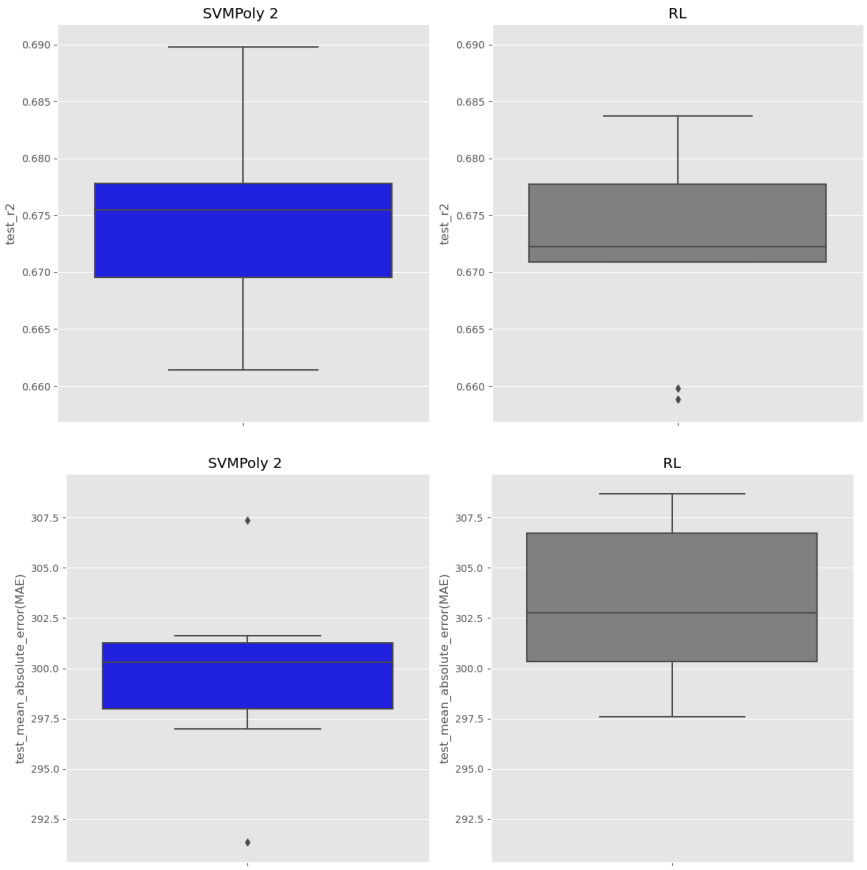
Tabla 7*: Métricas por pliegue en validación cruzada en predicción del peso del recién nacido.*

Pliegue	SVMPoly 2		RL	
	R^2	MAE	R^2	MAE
0	0.6659	300.6994	0.6773	308.6692
1	0.6772	300.7613	0.6588	300.6684
2	0.6747	301.4534	0.6710	300.2499
3	0.6839	296.9724	0.6598	302.9539
4	0.6763	307.3474	0.6728	304.6405
5	0.6614	301.6329	0.6779	302.6015
6	0.6898	291.3544	0.6837	298.8328
7	0.6683	299.9261	0.6717	307.8840
8	0.6733	297.4958	0.6709	307.3931
9	0.6780	299.4542	0.6800	297.5858

Fuente: Elaboración propia

Se observa de la tabla 6 y figura 15 poca variabilidad de un pliegue a otro, lo que valida el resultado inicial de los métodos. De donde se puede afirmar que el desempeño promedio de R^2 y MAE es de 0.6748 ±0.0076 y 299.7097 ±3.8662 para las SVM con kernel polinomial de grado 2 y de 0.6723 ±0.007 y 303.14791 ±3.7095 para la Regresión Lineal. Es decir, los métodos seleccionados son capaces de explicar el 67% de la variabilidad del peso del recién nacido con un error en la estimación cercana a los 300g.

***Figura 15**: Dispersión de las métricas en validación cruzada para predicción del peso del recién nacido*



Fuente: Elaboración propia

Resultados superiores a los reportados por un estudio realizado en Indonesia donde usaron variables sociodemográficas de la madre para predecir el peso del bebé usando otros modelos que lograron explicar menos del 1% de la variabilidad del peso del recién nacido (Aini & Kurniawan, 2023), y al estudio realizado por Tela et al., 2019 quienes reportan una variabilidad explicada del 42.2% en un modelo de Regresión Lineal que incluye información de la madre relacionada con su

peso, edad y dieta. En tanto Giraldo et al., 2023 reportan resultados similares a los obtenidos en este estudio ($R^2 = 0.65$) para un modelo de Regresión Lineal a partir de variables sociodemográficas y relacionadas con la atención a la madre; en un estudio de una cohorte de gestantes del Valle de Cauca en Colombia.

6.2. Modelos de ML para la clasificación del bajo peso al nacer.

Frente a los modelos de clasificación los Random Forest ofrecen la mayor exactitud y área bajo la curva, seguida de cerca por las SVM lineal y regresión logística. Los peores desempeños se observan en K-vecinos más cercanos y el complemento para clasificador de Bayes ingenuo (ver tabla 8)

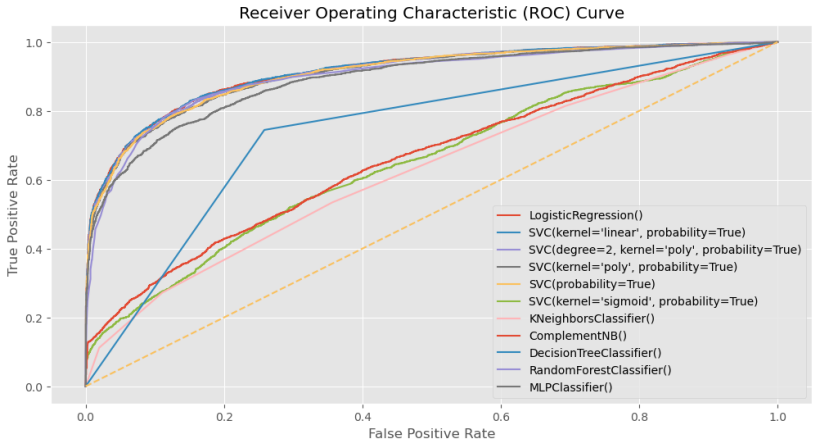
Tabla 8: *Desempeño de diferentes técnicas de ML en la clasificación del bajo peso al nacer*

Método	Accuracy	AUC
Regresión Logística	0.8362	0.8367
SVMlin	0.8339	0.8344
SVMpoly2	0.8310	0.8317
SVMpoly3	0.8267	0.8273
SVMrbf	0.8278	0.8287
SVMsig	0.6012	0.6010
KNN	0.5880	0.5888
ComplNB	0.6106	0.6138
Arboles de Decisión	0.7432	0.7432
RandForest	0.8371	0.8373
MLPClassifier	0.8050	0.8051

Fuente: Elaboración propia

En la figura 16 se puede apreciar como tan solo los modelos KNN, SVMsig, Arboles de decisión y ComplNB ofrecen pobre desempeño, para las demás técnicas las curvas ROC están muy cercanas con áreas bajo la curva superiores en todos los casos a 0.80. Lo es un buen indicador de la capacidad que tiene los modelos distinguir correctamente entre los niños con clasificación de bajo peso y los que no, por lo que se espera que la tasa de falsos positivos y falsos negativos en la clasificación sea baja.

Figura 16: *Curvas ROC de los desempeños de diferentes modelos de ML para la clasificación del bajo peso al nacer*



Fuente: Elaboración propia

6.2.1. Ajuste de parámetros y tratamiento de datos Random Forest y LR

En el caso de los Random Forest se consideraron los siguientes parámetros para la búsqueda de los mejores hiperpamentros con validación cruzada:

```
param_grid = {'n_estimators' : [100,150],
              'max_features' : [5, 7, 9],
              'max_depth'    : [None, 3, 10, 20],
              'criterion'    : ['gini', 'entropy']
                }
```

Teniendo como mejores resultados 'criterion': 'gini', 'max_depth': 10, 'max_features': 7, 'n_estimators': 100. En el caso de la Regresión Logística se consideró nuevamente usar una reducción de dimensionalidad mediante un PCA. Los resultados de las métricas con estas modificaciones se reportan en la tabla 9.

Tabla 9*: Métricas luego de ajuste y transformación de los datos en la clasificación del bajo peso al nacer.*

Método	Accuracy	AUC
RandForest	0.8403	0.8404
LR (PCA)	0.7763	0.7764

Fuente: Elaboración propia

Se observa una leve mejoría en las métricas con la combinación de parámetros descritos anteriormente. Sin embargo, la reducción de dimensionalidad por PCA hace que se tenga una reducción en las métricas comparado con el resultado inicial.

6.2.2. Validación Cruzada Random Forest y LR

Se realiza validación cruzada para combinación de parámetros obtenidos en la sección anterior y se mantiene la dimensión original de los datos en la LR. Los resultados de cada pliegue se muestran en la tabla 10

***Tabla 10**: Métricas por pliegue en validación cruzada en la clasificación del bajo peso al nacer.*

Pliegue	RandForest		LR	
	Accuracy	AUC	Accuracy	AUC
0	0.8150	0.8829	0.8153	0.8936
1	0.8159	0.8833	0.8097	0.8957
2	0.8337	0.8908	0.8275	0.9100
3	0.8415	0.9042	0.8396	0.9135
4	0.8427	0.8979	0.8412	0.9117
5	0.8384	0.8931	0.8371	0.9117
6	0.8359	0.8892	0.8312	0.9061
7	0.8321	0.8831	0.8244	0.8995
8	0.8486	0.9093	0.8465	0.9169
9	0.8452	0.8999	0.8381	0.9105

Fuente: Elaboración propia

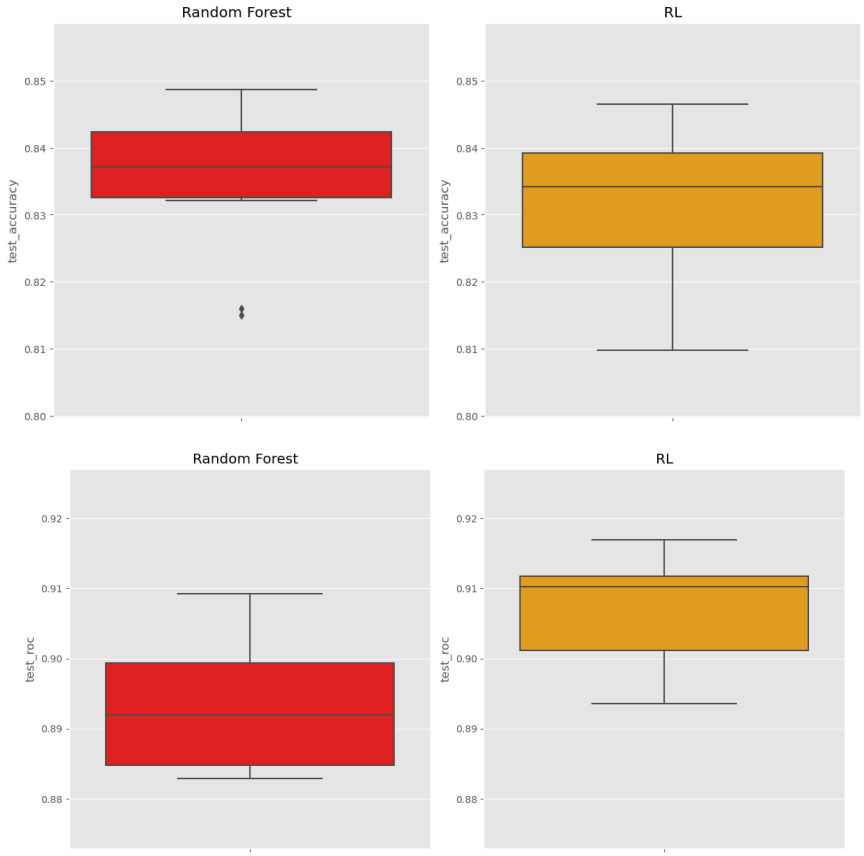
En la predicción del bajo peso al nacer, se observa que los métodos seleccionados ofrecen métricas más estables en cada pliegue (ver figura 17) con un promedio de *accuracy* y AUC de 0.8349±0.0105 y 0.8933±0.0088 para el caso del RandForest y 0.8310±0.0112 y 0.8933±0.0075 para LR. La poca variabilidad en la validación cruzada permite afirmar que los métodos seleccionados tienen una exactitud superior al 80% en la predicción de la clasificación del peso al nacer.

Los resultados con un Random Forest son superiores, la proporción de predicciones correctas realizadas por el modelo con respecto al total de casos evaluados es de 83%. Además, el área bajo la curva ROC es de 0.89, lo que sugiere un buen rendimiento en la capacidad del modelo para discriminar entre las clases recién nacidos con y sin bajo peso.

Los resultados encontrados en este estudio logran métricas superiores a los que reporta Cobas Planchez et al., 2020 quienes usan: un modelo de Regresión Logística, variables clínicas y demográficas de la gestante para predecir el bajo peso al nacer con un $ROC = 0.759$. Y similares

a los obtenidos por Rigo et al., 2021 quienes por medio de un Regresión Logística logran encontrar asociación entre variables relacionadoras con la madre y su atención prenatal y reportan un $ROC = 0.81$.

Figura 17: *Dispersión de las métricas en validación cruzada en la clasificación del bajo peso al nacer.*



Fuente: Elaboración propia

6.3. Modelo de Múltiple Salida HETMOGP

Luego de tener un referente sobre el comportamiento de las métricas, al abordar la tarea de regresión y clasificación en la predicción del peso y el bajo peso al nacer de forma independiente, se procede a evaluar un modelo de múltiple salida HETMOGP que permita modelar de forma conjunta ambas salidas. Un enfoque de aprendizaje múltiple tarea usa la información contenida en los datos de manera inductiva, suponiendo que lo que se aprende para cada tarea puede ayudar a mejorar el aprendizaje en las otras (Caruana et al., 1997). De esta manera se optimiza el proceso de aprendizaje y se puede dar respuesta a múltiples problemas con la misma información.

Como se mencionó en la sección 5, antes de entrenar el modelo de múltiple salida se realiza un análisis de relevancia de características con un enfoque LGBM. En la tabla 11 se presenta los resultados de dicho análisis, en ella se puede observar que las variables que más aportan información son las semanas de gestación, la edad de la madre, el sexo del recién nacido, el tener un embarazo múltiple y el número de consultas prenatales.

***Tabla 11**: Análisis de relevancia de características mediante un LGBM*

id	Característica	Relevancia	Relevancia normalizada	Relevancia acumulada
1	T_GES	39.0	0.200000	0.200000
2	EDAD_MADRE	27.0	0.138462	0.338462
3	SEXO	25.0	0.128205	0.466667
4	MUL_PARTO	19.0	0.097436	0.564103
5	NUMCONSUL	18.5	0.094872	0.658974
6	TIPO_PARTO	13.5	0.069231	0.728205
7	MES	11.5	0.058974	0.787179
8	SEG_SOCIAL	10.0	0.051282	0.838462
9	COMUNA_RES	8.5	0.043590	0.882051
10	NIV_EDUP	7.5	0.038462	0.920513
11	EST_CIVM	6.5	0.033333	0.953846
12	IDHEMOCLAS	4.0	0.020513	0.974359
13	NIV_EDUM	4.0	0.020513	0.994872
14	IDFACTORRH	1.0	0.005128	1.000000
15	SIT_PARTO	0.0	0.000000	1.000000
16	ATEN_PAR	0.0	0.000000	1.000000
17	AREANAC	0.0	0.000000	1.000000

Fuente: Elaboración propia

Según el análisis de relevancia se seleccionan para entrenar y evaluar un modelo HETMOGP las primeras 8 variables; las cuales logran una relevancia acumulada superior al 80% (ver tabla 10).

Ahora como ya se mencionó en secciones anteriores, un modelo HETMOGP es un caso particular de modelado multitarea o de múltiple salida propuesto por Moreno Muñoz et al., 2018 que se basa en procesos Gaussianos y permite modelar de manera conjunta varias salidas de diferente naturaleza, como es el caso de la predicción del peso del bebé y del bajo peso nacer. En la tabla 12 se presentan los resultados de diferentes experimentos donde se varían diferentes parámetros propios de la implementación del modelo y se reporta el desempeño combinado de la tarea regresión y clasificación mediante el Error Cuadrático Medio (MSE) normalizado y las métricas independientes por tarea R^2 para regresión y $accuracy$ para clasificación.

***Tabla 12**: Variación de parámetros y desempeño de diferentes modelos HETMOGP para predicción del peso del recién nacido y la clasificación del bajo peso al nacer.*

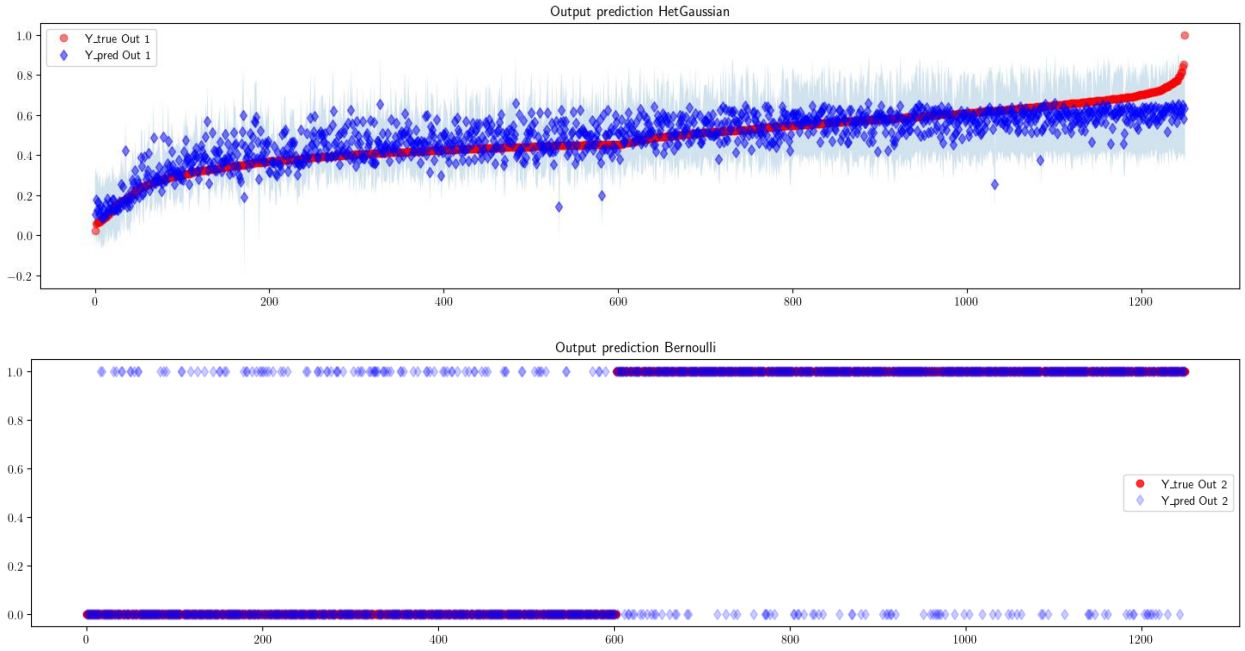
	Variables	**Experimentos**							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Parámetros modelo	Q	2	2	2	3	5	5	5	5
	Optimizador	Adam	Adam	Adam	Adam	Adam	vo	vo	vo
	Minibatch	250	250	250	250	250	250	250	250
	Puntos inducidos	50	150	25	25	25	25	25	25
	Iteraciones	3000	3000	3000	3000	3000	2000	3000	5000
Desempeño	MSE normalizado	0.834	0.834	0.838	0.815	0.811	0.823	0.823	0.821
	R^2	0.663	0.652	0.657	0.671	0.677	0.667	0.666	0.661
	Accuracy	0.843	0.842	0.836	0.845	0.842	0.843	0.844	0.844

Fuente: Elaboración propia. **Adam:** Estimación Adaptativa de Momentos, **vo:** optimización variacional

Se observa en los experimentos realizados al entrenar un validar un modelo HETMOGP, estabilidad en el desempeño a medida que se varían los parámetros del modelo. En todos los escenarios se obtienen niveles MSE y Accuracy superiores al 80% y una varianza explicada del peso del bebé superior al 66%. El mejor desempeño se logra con los parámetros seleccionados en el experimento 4; en la figura 18 se presenta la predicción del modelo para las dos variables de salida: peso del recién nacido y clasificación del bajo peso al nacer, ella se observa nuevamente como los pesos predichos por el modelo (rombos azules en la figura) se ajustan o están cercanos a los valores reales del peso del recién nacido (puntos rojos en la figura). Llama la atención al igual que en los modelos de única salida como el ajuste de la predicción es menor cuando el peso del

recién nacido es alto. Así mismo, se puede evidenciar en el gráfico como cuando se quiere contrastar la predicción de la clasificación del bajo peso al nacer son relativamente pocos los rombos azules o predichos que no coinciden con la clasificación real o puntos rojos en la figura.

***Figura 18**:Prediccion del peso del recién nacido y bajo peso al nacer según modelo HETMOGP*

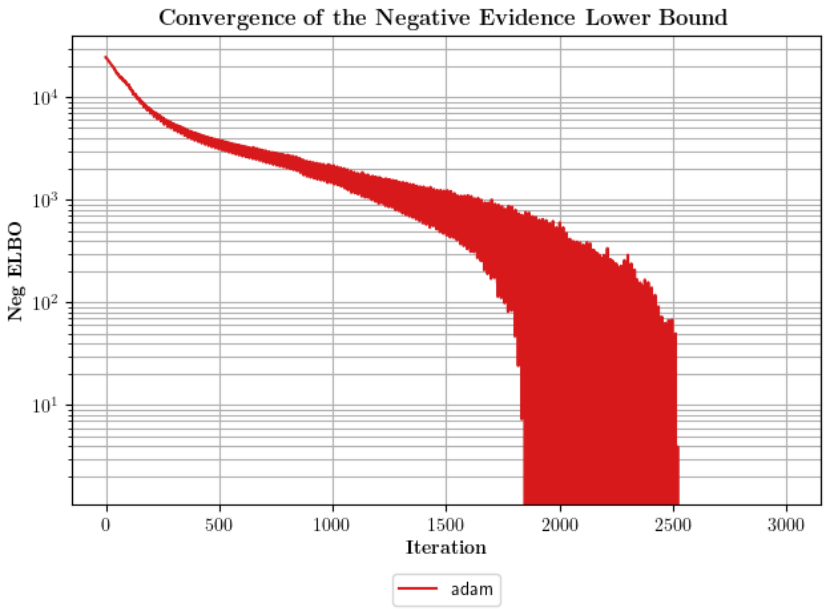


Fuente: Elaboración propia.

Los resultados son cercanos con los valores hallados al enfrentar cada tarea de forma independiente con métodos de única salida donde R^2=0.6748 y error en la estimación de 299.7g para el caso de las SVM y la exactitud de la clasificación fue de 83% con un Random Forest lo cual a su vez es consisten hallazgos de Martínez et al., 2021 quienes usando variables fetales mediante ecografía lograron entrenar un GP de única salida para predecir el peso al nacer con margen de error cercano a los 300 gramos.

En la implementación de estos modelos resalta la manera eficiente en la que la función de pérdida converge a través de las iteraciones. En la figura 19 se presenta el caso particular de la configuración del experimento 4, de ella se puede inferir que el modelo aprende de manera efectiva los patrones en los datos sin un consumo exagerado de tiempo ni recursos computacionales.

***Figura 19**: Optimización configuración del experimento 4 en un HETMOGP*



Fuente: Elaboración propia.

Es importante resaltar como con único modelo de múltiple salida se logra equiparar el desempeño de dos modelos entrenados y evaluados de forma independiente para realizar dos tareas relacionadas, más aún si se tiene en cuenta que en esta investigación se usó un número inferior información a la incluida para realizar cada tarea por separado.

7. Conclusiones

En esta monografía se exploran modelos de ML convencionales de uso corriente en la literatura de única salida real o binaria, adicionalmente se propone un modelo de múltiples tareas en un contexto probabilístico para la predicción conjunta de una variable binaria (clasificación de bajo peso) y una continua (peso del recién nacido) de donde se puede concluir:

- Los modelos tradicionales de única salida ofrecen buenos resultados tanto en la predicción del peso del recién nacido, como en la clasificación del bajo peso al nacer. En regresión logran capturar un porcentaje razonable (superior al 65%) de variabilidad del peso con un error cercano a los 300g, y en la clasificación se obtiene una exactitud y AUC en la mayoría de los casos por encima del 80%. Lo anterior deja en evidencia no solo el buen desempeño de los métodos sino la utilidad práctica de la información recolectada en la ficha del nacido vivo para la predicción del peso y del bajo peso al nacer.
- En términos prácticos contar modelos que permitan predecir el bajo peso en función de variables relacionas con aspectos sociodemográficos de los padres y propias de la madre y su embarazo, permite anticiparse al evento del bajo peso al nacer para diseñar estrategias diferenciadas de atención y cuidado en el proceso de gestación y posterior nacimiento.
- Anticiparse al bajo peso al nacer permite preparar de mejor manera a los padres sobre los cuidados y seguimientos especiales que requiere un recién nacido con esta condición. Contar con tiempo adicional para educar a los padres, los hace cuidadores más competentes para impactar la expectativa de vida, la salud y el bienestar de los niños y niñas.
- Los resultados logran posicionar al modelo HETMOGP como alternativa viable, consistente e innovadora para predicción del peso y la clasificación de bajo peso del recién nacido en Medellín. Estos logran equiparar de una manera más eficiente el desempeño de dos modelos de ML de uso corriente en la literatura, en tanto usan menor información (puntos inducidos y espacio latente) para reproducir al tiempo los resultados de dos modelos por separado, lo cual reduce la posibilidad de error en las estimaciones, ya que se usa la misma covarianza para producir dos salidas, mientras que cuando se hace uso de modelos basados cada uno en covarianzas diferentes podrían producir de forma independiente resultados contradictorios, como que un recién nacido tenga una clasificación de bajo peso y la estimación de su peso sea un normal en gramos.

8. Trabajos futuros

- Es importante en futuras investigaciones lograr incluir en la fase de entrenamiento y evaluación las demás variables consignadas en la ficha del nacido vivo, ya que información publicada en MedData es parcial. Esto podría mejorar el desempeño de los modelos y exhibir otras variables complementarias que podrían ser relevantes en la predicción del peso del recién nacido. Así mismo, seria relevante combinar dicha información con otras variables ecográficas fetales y clínicas de la madre en futuras investigaciones.
- Es relevante continuar evaluando el desempeño de los modelos en una fase de implementación, donde nuevos casos permitan poner a prueba el desempeño real de los modelos.
- Incluir en los modelos datos del recién nacido a nivel nacional con el fin de extrapolar los resultados obtenidos a nivel local.
- Explorar la utilidad del modelo HETMOGP para predecir otros eventos en salud, de tal manera que se pueda extenderse su uso.

Referencias

Agudelo Pérez, S., Maldonado Calderón, M., Plazas Vargas, M., Gutiérrez Soto, I., Gómez, Á., & Díaz Quijano. Diana. (2017). Relación entre factores sociodemográficos y el bajo peso al nacer en una clínica universitaria en Cundinamarca (Colombia). *Salud Uninorte*, *33*(2), 86–97. http://www.scielo.org.co/pdf/sun/v33n2/2011-7531-sun-33-02-00086.pdf

Aini, Y. N., & Kurniawan, F. E. (2023). The Maternal Sociodemographic Determinants of Low Birth Weight in Indonesia. *Kemas*, *18*(4), 536–545. https://doi.org/10.15294/kemas.v18i4.42006

Alcaldía de Medellín. (2019, septiembre 24). *MEData*. Datasets. http://medata.gov.co/dataset/bajo-peso-al-nacer

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology*, *188*(12), 2222–2239. https://doi.org/10.1093/aje/kwz189

Bonilla, E. V, Chai, K. M. A., & Williams, C. K. I. (2008). Multi-task Gaussian Process Prediction. *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. https://papers.nips.cc/paper_files/paper/2007/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf

Caruana, R., Pratt, L., & Thrun, S. (1997). Multitask Learning. *Machine Learning*, *28*, 41–75. https://link.springer.com/article/10.1023/a:1007379606734

Cobas Planchez, L., Emelda Navarro Garcia, Y., & Mezquia de Pedro, N. (2020). Escala pronóstica de bajo peso al nacer en gestantes del municipio Guanabacoa. *Rev haban cienc méd*, *19*(6). http://www.revhabanera.sld.cu/index.php/rhab/article/view/3130

Giraldo, J. J., Zhang, J., & Alvarez, M. A. (2022). Correlated Chained Gaussian Processes for Modelling Citizens Mobility Using a Zero-Inflated Poisson Likelihood. *IEEE Transactions on Intelligent Transportation Systems*, *23*(11), 20337–20351. https://doi.org/10.1109/TITS.2022.3171730

Giraldo, S., Chattes, N., Estrada, E., Ramirez, K., & Quintero, J. (2023). *Factores asociados al bajo peso al nacer a término cohorte de gestantes de una aseguradora de salud del valle del cauca 2020-2021*. https://doi.org/10.1590/SciELOPreprints.6695

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. https://github.com/Microsoft/LightGBM.

Kingsford, C., & Salzberg, S. L. (2008). *What are decision trees?* http://www.nature.com/naturebiotechnology

Liu, H., Cai, J., & Ong, Y. S. (2018). Remarks on multi-output Gaussian process regression. *Knowledge-Based Systems*, *144*, 102–121. https://doi.org/10.1016/j.knosys.2017.12.034

Márquez-Beltrán, M. F. R., Vargas-Hernández, J. E., Quiroga-Villalobos, E. F., & Pinzón-Villate, G. Y. (2013). Análisis del bajo peso al nacer en Colombia. En *Rev. salud pública* (Vol. 15, Número 4).

Martínez, M., Sánchez, J., & Pérez, J. (2021). *Aprendizaje Computacional Para La Estimación Automática Del Peso Al Nacimiento Usando Variables Multimodales Materno-Fetales*. https://doi.org/10.24254/CNIB.21.3

Ministerio de la Protección Social. (2010). *Manual de Usuario de los Módulos de Nacimientos y Defunciones, aplicación NDE del Proyecto Registro Unico de Afiliados al Sistema de la Protección Social-RUAF-Ministerio de la Protección Social República de Colombia*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/OT/Manual-usuarios-RUAF-ND.pdf

Mora Pineda, J. (2022). Predictive models in health based on machine learning. *Revista Médica Clinica Las Condes*, *33*(6), 583–590. https://doi.org/10.1016/j.rmclc.2022.11.002

Moreno Muñoz, P., Artés Rodríguez, A., & Álvarez, M. A. (2018). *Heterogeneous Multi-output Gaussian Process Prediction*. https://papers.nips.cc/paper_files/paper/2018/file/165a59f7cf3b5c4396ba65953d679f17-Paper.pdf

Núñez Reiz, A., Armengol de la Hoz, M. A., & Sánchez García, M. (2019). Big Data Analysis and Machine Learning in Intensive Care Units. En *Medicina Intensiva* (Vol. 43, Número 7, pp. 416–426). Ediciones Doyma, S.L. https://doi.org/10.1016/j.medin.2018.10.007

Rasmussen, C. Edward., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press. https://pubmed.ncbi.nlm.nih.gov/30856197/

Rigo, C. N. K., De Oliveira Almeida, N., Almeida, R. M., & Varnier, R. (2021). Association between Prenatal Care Adequacy Indexes and Low Birth Weight Outcome. *Revista Brasileira de Ginecologia e Obstetricia*, *43*(4), 256–263. https://doi.org/10.1055/s-0041-1728779

Salazar, D. A., Iván Vélez, J., & Salazar, J. C. (2012). Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? *Revista Colombiana de Estadística*, *35*(2), 223–237.

Salinas de León, A. (2020). Sobre la implementación de algoritmos de machine learning en las ciencias penales y sus implicaciones jurídica. *Revista Mexicana de Ciencias Penales*, *12*, 192–204. https://revistaciencias.inacipe.gob.mx/index.php/02/article/view/373

Saul, A. D., Hensman, J., Vehtari, A., & Lawrence, N. D. (2016). *Chained Gaussian Processes*. https://proceedings.mlr.press/v51/saul16.pdf

Scharf, H. R., Hooten, M. B., Johnson, D. S., & Durban, J. W. (2018). Process convolution approaches for modeling interacting trajectories. *Environmetrics*, *29*(3). https://doi.org/10.1002/env.2487

Tela, F. G., Bezabih, A. M., & Adhanu, A. K. (2019). Effect of pregnancy weight gain on infant birth weight among mothers attending antenatal care from private clinics in Mekelle City, Northern Ethiopia: A facility based follow-up study. *PLoS ONE*, *14*(3). https://doi.org/10.1371/journal.pone.0212424

Victoria, M., Rincón, B., Gallestey, J. B., & Jiménez, R. (2009). La predicción del bajo peso y del peso insuficiente al nacer mediante la antropometría materna. *Hacia la Promoción de la Salud*, *14*(1), 35–53. http://www.scielo.org.co/pdf/hpsal/v14n1/v14n1a03.pdf

Weinberger, K. (2018). *Machine Learning for Intelligent Systems*. Universidad de Cornell. https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote15.html

Wiemken, T. L., & Kelley, R. R. (2020). Machine Learning in Epidemiology and Health Outcomes Research. *Annu. Rev. Public Health*, *41*, 21–36. https://doi.org/10.1146/annurev-publhealth

World Health Organization, & United Nations Children's Fund. (2004). *Low birthweight country, regional and global estimates. World Health Organization*. https://apps.who.int/iris/handle/10665/43184.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y. S., Gong, C., & Shen, X. (2020). Survey on Multi-Output Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(7), 2409–2429. https://doi.org/10.1109/TNNLS.2019.2945133