UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

DESARROLLO DE MODELOS DE APRENDIZAJE DE MÁQUINAS PARA PREDECIR ENFERMEDADES/CONDICIONES DURANTE EL EMBARAZO

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO

GABRIEL TOMÁS CUBILLOS FUENTES

PROFESOR GUÍA:
CLAUDIO PÉREZ FLORES

PROFESORES CO-GUÍAS:
PABLO ESTÉVEZ VALENCIA
SEBASTIÁN ILLANES LÓPEZ

MIEMBROS DE LA COMISIÓN:
DORIS SÁEZ HUEICHAPAN
MAURICIO CERDA VILLABLANCA

Este trabajo ha sido parcialmente financiado por el proyecto FONDECYT 1231675 de ANID, además del financiamiento basal ANID, AMTC AFB220002 e IMPACT FB210024, y por el Departamento de Ingeniería Eléctrica, Universidad de Chile

SANTIAGO DE CHILE
2024

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
Y MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO
POR: GABRIEL TOMÁS CUBILLOS FUENTES
FECHA: 2024
PROF. GUÍA: CLAUDIO PÉREZ FLORES

DESARROLLO DE MODELOS DE APRENDIZAJE DE MÁQUINAS PARA PREDECIR ENFERMEDADES/CONDICIONES DURANTE EL EMBARAZO

El Aprendizaje de Máquinas (ML) y la Inteligencia Artificial (AI) son temas en auge hoy en día, utilizadas en campos como la medicina. En este contexto, las condiciones del embarazo, como la Diabetes Mellitus Gestacional (GDM) y Grande para la Edad Gestacional (LGA), pueden tener consecuencias graves para la salud materna y fetal.

En esta tesis se desarrollan modelos de ML para detectar GDM y LGA, considerando las limitaciones económicas en sistemas de salud. Se busca que los modelos sean sencillos, utilizando variables fácilmente obtenibles. Se implementaron 12 modelos de ML que fueron entrenados y probados con una base de datos de pacientes de un hospital chileno, se consideraron más de 3000 variaciones de hiperparámetros. Los mejores modelos obtuvieron una sensibilidad y especificidad de 82 % y 75 % respectivamente para GDM y 87 % y 80 % para LGA.

Además, se introduce un método de Aumentación de Datos personalizado e innovador, diseñado para variables específicas con límites y restricciones basada en la experiencia médica. Se realiza una comparación con modelos del estado del arte para evaluar la eficacia de las soluciones propuestas en relación con la investigación existente, lo cual muestra una mejora en el desempeño respecto a modelos propuestos en condiciones similares.

And on the pedestal these words appear:
"My name is Ozymandias, king of kings:
Look on my works, ye Mighty, and despair!"

"Ozymandias Percy Shelley

Agradecimientos

A mis padres, ya que es gracias al esfuerzo de ellos que he logrado todo lo que conseguido. No creo que una simple frase alcance a reflejar todo lo que siento.

A mis familiares, por siempre creer en mi de forma incondicional.

A mi tío Juan, que me hubiese gustado que pudiese haber leído esta tesis.

A mis amigos de la universidad, por el apoyo brindado todos estos años y los buenos momentos que hemos pasado.

A mis amigos del colegio, por años de amistad y apoyo, que espero, sigan así.

Al profesor Claudio Pérez, por la incontable ayuda que me ha entregado y esa "sed"de conocimiento que nos ha impulsado a entregar lo mejor de nosotros.

Al profesor Pablo Estévez, por sus consejos y apoyo en diversos procesos durante mi carrera.

Este trabajo fue parcialmente financiado por el proyecto FONDECYT 1231675 de ANID, además del financiamiento basal ANID, AMTC AFB220002 e IMPACT FB210024, y por el Departamento de Ingeniería Eléctrica, Universidad de Chile

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

1. Introducción

1.1. Motivación

El Aprendizaje de Máquinas (ML) y la Inteligencia Artificial (AI) son temas populares hoy día, debido a la aparición de diversas herramientas como ChatGPT y DALL-E y usados en diversos campos [1, 2]. En medicina, el uso de métodos de ML ha ido en incremento, principalmente en la predicción de enfermedades o patologías usando imágenes médicas, tales como Rayos-X, Resonancia Magnética y Ultrasonido [3, 4]. Sin embargo, su uso con imágenes no siempre es efectivo, debido a ciertos factores, como el equipamiento disponible, el entrenamiento del personal o tomas erróneas [5]. A pesar de esto, existen otras fuentes de información que pueden servir para realizar estas predicciones, las cuales son los registros obtenidos en consultas médicas, como edad, peso, altura, enfermedades, síntomas, y exámenes médicos sanguíneos, los cuales son englobados en el término Registro Médico Electrónico (EHR) [6]. En los últimos años se han desarrollado modelos de ML para predecir o detectar diferentes enfermedades, tales como enfermedades a la piel, hígado, corazón, entre otras [7].

En este marco, existen diversas condiciones/enfermedades que ocurren durante el embarazo, tales como la Diabetes Mellitus Gestacional (GDM) y Grande para la Edad Gestacional (LGA), el primero es definido como cualquier grado de intolerancia a la glucosa durante el embarazo [8, 9]. En el 2017 se estimaba que alrededor del 14 % de los embarazos a nivel mundial fueron afectados por GDM [10]. Según la International Diabetes Federation (IDF), la prevalencia de GDM en la población chilena alcanzó el 9.8 % en el 2021 [11]. GDM está asociado al incremento al riesgo de enfermedades, tanto para la madre, como para el feto en desarrollo [8, 12–14]. Los efectos en la madre asociados a GDM incluyen depresión, desarrollo de Diabetes Mellitus Tipo 2, riesgos de enfermedades hepáticas, renales y cardiacas [15–17]. Mientras que el feto tiene riesgo de desarrollar resistencia a la insulina, macrosomía, nacimiento prematuro, distrés respiratorio, requerir de cuidados intensivos o nacer muerto [15–17].

Por otro lado, LGA es definido como un recién nacido que pesa más que el percentil 90 según su edad gestacional. La prevalencia en general de LGA es cercana a un 10 % en países europeos utilizando tablas de crecimiento nacionales [18], tasas similares son obtenidas en Asia [19, 20]. Brasil tiene la prevalencia más alta con un 17.1 % [21]. En los Estados Unidos, la prevalencia estimada es de un 11.2 %, pero el mayor riesgo de macrosomía (recién nacido con un peso mayor a los 4000 gramos) fue de un 26.0 % para infantes nacidos de madres con pre diabetes mellitus gestacional [22]. La diabetes en el embarazo no es el único factor de riesgo para LGA y macrosomía, pero sí es uno de sus principales factores de estas condiciones [23]. Desafortunadamente, debido a que la prevalencia de la diabetes en el embarazo (GDM) ha ido en aumento mundialmente [24], es esperable un aumento en las tasas de nacimientos LGA.

Infantes con LGA tienen un mayor riesgo de complicaciones perinatales, tales como distocia de hombros, hipoglicemia y una estancia prolongada en la unidad de cuidados intensivos neonatal. LGA también contribuye a un aumento en riesgos para la madre, tales como parto vaginal operativo y hemorragia postparto [25]. Adicionalmente a estas consecuencias de corto plazo, bebés con LGA también son propensos a desarrollar obesidad en la niñez y la infancia y enfermedades metabólicas [26], en el caso de un recién nacido femenino, aumenta el riesgo que a futuro desarrollen GDM, y a su vez, incremente el número de nacimientos LGA.

Estas enfermedades son detectadas por diversos métodos. GDM es diagnosticado mediante un examen de tolerancia a la glucosa (con una carga de 75g de glucosa) entre las 24 y 28 semanas de gestación, con mediciones en ayuna y 2 horas después de consumida la carga. Mientras que LGA, el diagnóstico se realiza al nacer, pero se realiza una predicción anticipada mediante un ultrasonido rutinario en el tercer trimestre de embarazo. Normalmente entre las 32 y 36 semanas de gestación, con la cual, mediante el peso estimado del feto, se predice si tiene LGA. Lo cual tiene como beneficio, evitar complicaciones al momento del parto. Sin embargo, estos métodos tienen desventajas. Por un lado, GDM, su diagnóstico es realizado en una etapa tardía del embarazo, donde el feto ya pudo haber sufrido las consecuencias relacionadas con la enfermedad. Mientras que la predicción por ultrasonido de LGA tiene una alta tasa de detección de negativos (No LGA), cercana al 90 %, pero una baja detección de positivos (LGA), aproximadamente de un 60 %, que también puede acarrear consecuencias negativas, debido a una mala toma de decisiones de cómo proceder en el parto. Estas consecuencias, mencionadas previamente para LGA, pueden ocurrir en caso de ser el feto LGA y diagnosticado equívocamente como negativo. En caso contrario, de ser clasificado como positivo de LGA cuando en realidad es negativo, se puede realizar cesárea, la cual es común en fetos LGA, lo cual es un evento no recomendado por las secuelas que conlleva tanto para el feto como para la madre.

El objetivo de este Trabajo de Tesis es desarrollar modelos de ML que puedan realizar una predicción anticipada de GDM, idealmente en la primera mitad del embarazo y se espera mejorar el desempeño de la detección de LGA utilizando de base los hallazgos obtenidos mediante el ultrasonido. En ambos casos, se considerará la simpleza de obtención de los datos, para que el método desarrollado sea global, evitando así, que sean necesario exámenes adicionales a los solicitados por rutina, es decir, utilizando datos de EHR ya disponibles.

1.2. Estado del Arte

Existen diversos modelos desarrollados para la detección de GDM durante las etapas tempranas del embarazo [27–44], que, de forma similar a lo planteado en la Motivación, pretenden realizar un diagnóstico lo antes posible para prevenir los efectos en la madre y el feto. Los enfoques son diversos, existen modelos que usan variables simples, tales como la edad, historial de GDM previo, familiares de primer grado con diabetes, embarazos múltiples, glucosa en plasma en ayuna (FPG), hemoglobina glicosilada (HBA_{1c}) y triglicéridos [27]. En [28] son utilizadas como variables de entrada, la edad, el índice de masa corporal (BMI) preembarazo, FPG y triglicéridos. Edad, BMI, FPG y el historial de diabetes en la familia son necesarios en [29]. FPG y BMI son las únicas dos variables solicitadas por [30], estas dos

variables sumadas al historial de familiar de diabetes son utilizadas en [31].

Un cuestionario de 9 preguntas es una alternativa planteada como método de recopilación de datos para un método de ML propuesto en [32], este cuestionario, que a juicios de los autores, debería ser fácilmente respondido por la embarazada, incluye preguntas tales como, la edad, peso, altura, número de familiares/relativos con diabetes, enfermedades preexistentes, mayor valor registrado en un test de HBA_{1c}, embarazos previos, y de ser afirmativa esta última, los valores de OGTT en esos embarazos.

También existen aproximaciones sin la necesidad de exámenes sanguíneos, por ejemplo, en [33], son solicitados el BMI preembarazo, la circunferencia del abdomen medido en el primer trimestre de embarazo, edad, el diagnóstico del síndrome de ovario poliquístico (PCOS), gravidez, menstruaciones irregulares y el historial de diabetes en la familia. Otro estudio reporta el uso de un solo examen de sangre [34], FPG, como método de detección, variando solo el umbral de corte, con el cual obtienen diverso niveles de detección.

Otras propuestas requieren de datos más específicos, por ejemplo, [35], además de requerir variables como FPG, BMI preembarazo, edad, y medidas de cintura y cadera, solicitados por otros modelos, requiere adicionalmente de Alanina Aminotransferasa (ALT), presión sanguínea sistólica y diastólica, ingreso promedio y nivel educacional. En [27] se propone un modelo adicional que utiliza 17 variables, las cuales además de las mencionadas previamente, se le añade el valor de las hormonas tiroides, triyodotironina (T3) y tiroxina total (T4), lipoproteínas, apoliproteínas A y B. El estudio de [36] requiere de la edad materna, etnia, historial previo de GDM y de la presión sanguínea promedio, obtenido este último dato en una visita adicional.

En [37] se propone que las variables de edad, peso, BMI, conteo de eritrocitos, plaquetas, leucocitos y neutrófilos, test de prealbúmina, gamma-glutamil transferasa (γ-GT), ácido úrico, bilirrubina directa, hematocritos (HCT), ALT, ferritina y hemoglobina. En [32] es propuesto otra alternativa, la cual utiliza 2355 variables, provenientes de una masiva EHR de Israel, los cuales incluyen datos de hasta dos embarazados pasado para ciertas pacientes y exámenes de laboratorios con una temporalidad similar, las variables más importantes son los OGTT de embarazos previos, la edad, la glucosa en el primer trimestre, parientes de primer grado con Diabetes Mellitus y las pruebas de conteo de células blancas, neutrófilos absolutos y ácido úrico. La edad, altura, BMI, historial obstétrico, el peso al nacer del último embarazo, GDM previo, presión arterial promedio y proteína plasmática A asociada al embarazo (PAPP-A) son las variables requeridas por [38].

El modelo de [39] requiere de 5 variables, edad, nitrógeno ureico en sangre (BUN), la razón entre fibrinógeno y albúmina, la razón entre BUN y creatinina, y la razón entre BUN y albúmina. En [40] se utilizan 73 variables de 489 pacientes, las cuales incluyen varias de las variables frecuentemente utilizadas por otros trabajos, tales como edad, peso, BMI, historial de diabetes en la familia, consumo de tabaco, exámenes de sangre, entre otros, pero adquiridas varias veces en el primer y segundo trimestre, lo cual les ofrece una evolución del paciente en el tiempo. [41] utiliza la edad, BMI preembarazo, historial de GDM, HCT, el volumen de plaquetas promedio, ALT, creatinina, FPG, fibrinógeno, PAPP-A y el tiempo de tromboplastina parcial activada.

El trabajo de [42], propone la realización de un OGTT, con un criterio alternativo a los utilizados internacionalmente, además de un modelo, que añade las variables de edad, BMI e insulina en sangre de 2 horas posprandial.

El modelo con mejor desempeño de [43] utiliza variables recopiladas en múltiples visitas, incluyendo datos básicos de la paciente, como edad, peso antes del embarazo, peso actual, altura, información de preexistencias y embarazos previos, hasta exámenes de sangre, realizados en distintas ocasiones en el primer y segundo trimestre de embarazo. En [44] llegan a la conclusión que la glucosa en sangre en ayuna, HBA_{1c}, triglicéridos y el BMI contribuyen fuertemente a la predicción de GDM.

Para LGA, la forma más común de predecir LGA previo al nacimiento del feto es utilizando el peso fetal estimado (EFW), calculado en el ultrasonido de tercer trimestre, el cual es rutinario. Con este peso fetal y la edad gestacional del feto, se comparan los valores con curvas de crecimiento, las más populares son las de Hadlock 1991 [45], Fetal Medicine Foundation (FMF) [46], INTERGROWTH-21st[47], National Institutes of Child Health and Human Development (NICHD) [48] y la curva de la Organización Mundial de la Salud (WHO) [49]. Estas tablas indican que un feto es sospechoso de ser positivo de LGA si el EFW es mayor al percentil 90 de la curva, según su edad gestacional.

En esta misma línea, existen trabajos, [50, 51], que comparan el rendimiento de predicción de algunas de las curvas. De forma similar, en [52] es utilizado el *Z-Score* de la medida de circunferencia abdominal (AC) del feto basado en un estándar de INTERGROWTH-21st. En [19] son utilizados como entradas, la circunferencia del feto (HC) medido durante el diagnostico de GDM, entre semanas 24 y 30 de gestación, el AC medido a más tardar dos semanas antes del parto y después de las 34 semanas de gestación y el BMI preembarazo, aunque este estudio está restringido a solo pacientes positivos de GDM. En [53] se utiliza una combinación de la curva de crecimiento de la FMF y el valor de (HBA_{1c}) para realizar una predicción de LGA, pero este método está limitado a pacientes con Diabetes Pregestacional. Mientras que lo planteado en [54] usa características maternas, como edad, BMI, etnia, paridad y el consumo de tabaco, el sexo del feto, el cambio de EFW, índice pulsatilidad arteria uterina (UtA-RI) e índice pulsatilidad arteria umbilical (UA-PI), medidos en ultrasonidos ejecutados entre el segundo y tercer trimestre.

Otros modelos de LGA utilizan enfoques sin la necesidad de ultrasonidos, por ejemplo, [55] usa 4 metabolitos medidos en la semana 36 de gestación, utilizando una cromatografía líquida con espectrometría de masas en tándem de alta resolución. Mientras que otros modelos realizan una predicción usando datos recopilados durante las etapas tempranas del embarazo, en [23] utilizan datos obtenidos antes de la semana 28 de gestación, los cuales son: edad, historial familiar de Diabetes Mellitus tipo 2 (T2DM), GDM, historial de macrosomía, colesterol total, triglicéridos, colesterol lipoproteína de baja densidad (LBL-C) y FPG. En [56] se utiliza la altura, edad, BMI, la ganancia de peso durante el embarazo, paridad, etnia, el consumo de tabaco, el historial de diabetes en la familia, la presión arterial sanguínea promedio y el OGTT con datos recopilados en la semana 24 y 32 de gestación. La edad, etnia, paridad, BMI, consumo de tabaco, nivel educacional, el ingreso del hogar, consumo de vegetales, utilización de suplementos vitamínicos, la presencia de enfermedades cardiacas, presión sanguínea sistólica y diastólica, hemoglobina, glucosa, triglicéridos y concentraciones ferritina son las variables requeridas en [57], todas estas recopiladas antes de la semana 21

de embarazo.

El trabajo de [58] presenta dos aproximaciones distintas para realizar la predicción de LGA, por un lado, solo con datos preembarazo y por otro, añadiéndole a estos datos, información recopilada a las 26 semanas de gestación, además separan la detección de mujeres con su primer embarazo, *primiparae*, de mujeres con partos previos, *multiparae*. El mejor rendimiento en las métricas utilizadas en [58] es alcanzado utilizando los datos preembarazo juntos con los obtenidos a las 26 semanas y con mujeres *multiparae*, utilizando como entrada, la edad de la madre, el estatus marital, el quintil de ingreso a nivel de área, el área de residencia (urbano/rural), el consumo de tabaco antes del embarazo, BMI, hipertensión y diabetes pregestacional, gravidez, paridad, GDM en embarazos previos, cesáreas, embarazos prematuros, muertes neonatales, macrosomía previa, nacimientos de fetos con bajo peso, sexo del feto, ganancia de peso a las 26 semanas de gestación, consumo de tabaco y sustancias durante el embarazo, GDM, hipertensión inducida durante el embarazo y desordenes psiquiátricos.

Los modelos propuestos en [59] y en [60] utilizan pocas variables, el primero solo utiliza el BMI, PAPP-A e información sobre multiparidad para realizar un diagnóstico el primer trimestre, mientras que el segundo, utiliza, la edad, el BMI pregestacional, la ganancia de peso durante el embarazo y las Apoliproteínas B y A1, recopilados en exámenes durante el primer trimestre, entre semanas 6 y 14 de gestación.

Los algoritmos/modelos utilizados por los modelos mencionados son diversos, Regresión Logística (LR) es utilizada por [19, 28, 29, 38, 39, 41, 42, 53, 54, 56–60], una red neuronal profunda por [27, 36], un *Random Forest* (RF) es utilizado por [23, 33, 37], una *Gradient Boosting Machine*, es utilizado por [32, 35, 43, 44], [32] utiliza la implementación de *LightGBM*, [35, 43] la implementación de la librería *XGBoost*, una red neuronal recurrente (RNN) del tipo *Long Short-Term Memory* es usada por [40], [31] utiliza una *Recursive Partitioning and Amalgamation* (RECPAM) y [30, 34, 52] utilizan un umbral de decisión según los valores de sus variables (fórmula).

Regresión Logística (LR) es un modelo estadístico que puede ser usado para clasificación y análisis predictivo. LR estima la probabilidad de un ocurrir un evento basado en las variables independientes de la base de datos, esta probabilidad, entre 0 y 1, es obtenida debido al uso de la función logística/sigmoide [61–63]. Los perceptrones multicapas (MLPs), *Deep feedforward networks* o Redes Neuronales, tienen como objetivo el aproximar una función que mapee una entrada y la aproxime a una salida [64]. *Random Forest* (RF) es una combinación de predictores, en que cada uno se entrena a partir del conjunto de entrenamiento, cada uno realiza una clasificación/voto. La clasificación más popular, la elegida por la mayor cantidad de árboles, es la seleccionada por la RF [65]. *Gradient Boosting Machine* (GBM) es una técnica de Aprendizaje de Máquinas que utiliza múltiples algoritmos simples, denominados *weak learners* o *base learner*. La técnica de *Boosting* consiste en que cada vez que un *weak learner* es añadido, este se centra más en los errores que cometieron sus predecesores [66]. Una red neuronal recurrente (RNN) es un red neuronal en que la salida retroalimenta a la entrada, influyendo en el siguiente valor ingresado. RECPAM es un método para construir árboles de regresión a partir de datos que permite el tratamiento explícito de una gran variedad de variables de respuesta. La metodología RECPAM permite encontrar soluciones a dos problemas básicos del análisis de datos: descubrir e identificar clases en los datos (clasificación) y encontrar reglas simples y económicas para asignar individuos a clases (discriminación) [67].

Los modelos del estado del arte son resumidos en la siguiente tabla:

Tabla 1.1: Modelos del estado del arte.

Modelos del estado del arte
Regresión Logística, Red Neuronal Profunda/Perceptrón Multicapas
Random Forest, *Gradient Boosting Machine*
Red Neuronal Recurrente, *Recursive Partitioning and Amalgamation*

1.3. Hipótesis

La hipótesis planteada en esta tesis es si es posible desarrollar modelos de Aprendizaje de Máquinas que puedan realizar un diagnóstico útil de las enfermedades/condiciones de Diabetes Mellitus Gestacional y Grande para la Edad Gestacional, utilizando datos obtenidos en visitas rutinarias de embarazadas, complementando con el uso de técnicas de aprendizaje de máquinas para aumentar el rendimiento de los mismos. Se desprende de esta misma hipótesis los siguientes puntos:

- Es posible desarrollar y entrenar modelos de ML para predecir la diabetes gestacional en la primera mitad del embarazo y de LGA con alta sensibilidad y especificidad, que supere el rendimiento de los modelos ya publicados.
- Es posible identificar las variables más relevantes para la predicción de GDM y LGA. El conjunto de variables resultante debe ser reducido para facilitar la aplicación del modelo en un ambiente clínico.
- Es posible determinar las mejores combinaciones de modelos e hiperparámetros que logren tener el mejor desempeño posible para la predicción de ambas condiciones.
- Mediante el uso de modelos de ML y técnicas de aumentación de datos, es posible alcanzar resultados comparables a aquellos encontrados en otras bases de datos del estado del arte para GDM y LGA.

1.4. Objetivos Generales

Desarrollar modelos de Aprendizaje de Máquinas para la predicción de Diabetes Mellitus Gestacional y Grande para la Edad Gestacional, utilizando como entrada Registros Médicos Electrónicos. Además de identificar las variables más relevantes para la predicción y evaluar el uso de técnicas de aumentación de datos con el fin de aumentar el desempeño de los modelos.

1.5. Objetivos Específicos

Los objetivos específicos del Trabajo de Tesis son los siguientes:

- Desarrollar un modelo de ML para la predicción temprana de Diabetes Gestacional y la predicción de Grande para la Edad Gestacional, utilizando variables comúnmente utilizadas en la consulta médica obstétrica.
- Identificar las variables más relevantes para la predicción temprana de Diabetes Gestacional y para la predicción de Grande para la Edad Gestacional. El conjunto de variables debe ser reducido para que se facilite la aplicación del modelo en la práctica clínica.
- Seleccionar modelo e hiperparámetros mediante una búsqueda que evalúe múltiples combinaciones de estos, teniendo en consideración un buen equilibrio entre una alta sensibilidad y buena especificidad para la predicción de GDM y LGA.
- Evaluar el uso de técnicas de aumentación de datos para mejorar los resultados de predicción de los modelos de ML en la predicción temprana de GDM y la predicción de LGA.
- Aplicar los modelos desarrollados a una base de datos real de pacientes embarazadas y comparar los resultados respecto de lo publicado en el estado del arte, tanto para GDM como para LGA.

1.6. Contribuciones de la Tesis

Las principales contribuciones de este trabajo de tesis es la utilización de modelos de ML en la predicción de dos condiciones/enfermedades en el embarazo (GDM y LGA). En el caso de GDM, la predicción se realiza de forma temprana para poder facilitar el tratamiento preventivo y reducir en gran medida posibles consecuencias negativas para el feto y la embarazada. Para LGA, la principal contribución es obtener resultados mejores a los actuales basados solo en hallazgos en ultrasonidos.

Se utilizaron 12 algoritmos distintos de ML, superior a lo utilizado por otros trabajos del estado del arte. Adicionalmente a su utilización, también se realizan optimizaciones de hiperparámetros estos modelos. Se aplicó un proceso de selección de variables para reducir el uso de datos redundante/innecesarios y mejorar el desempeño de los modelos.

También se propone un método de Aumentación de Datos original basado en la experiencia de especialistas del área de ginecología/obstetricia, el cual considera que los pacientes "creados" por este método tengan sentido en la realidad.

1.7. Estructura de la Tesis

El presente trabajo de tesis está dispuesto en 6 capítulos, incluyendo el actual capítulo de Introducción. En el capítulo 2, Marco Teórico, se exploran las diversas definiciones de las herramientas teóricas utilizadas en el trabajo de tesis, tales como modelos, selección de variables, hiperparámetros, aumentación de datos, entre otras. Además, se incluyen las

definiciones y fórmulas de las métricas utilizadas para evaluar y comparar el desempeño de los modelos con el estado del arte. En el tercer capítulo, Metodología, se presentan los detalles de las bases de datos utilizadas, así como el preprocesamiento requerido, luego se presentan los detalles respecto a los modelos implementados, los hiperparámetros utilizados y se especifican sus procesos de entrenamiento y evaluación. Finalmente, son entregados los detalles de la aumentación de datos implementada, de carácter innovador y dedicado para este problema. El capítulo cuatro, Resultados, como su nombre lo indica, presenta los resultados de la selección de variables para ambas condiciones y los resultados de los modelos, además de una comparación respecto a utilizar o no utilizar la aumentación de datos. Adicionalmente, se presenta una comparación con los modelos del estado del arte de las dos enfermedades. En el capítulo 5, Discusión, se analizan los resultados obtenidos por los modelos y se contrastan con los obtenidos por los modelos del estado del arte. De igual forma, son analizados el método de aumentación de datos propuesto, los modelos utilizados y la selección de variables. En el último capítulo, Conclusiones, se resume lo realizado en este trabajo de tesis. Además, se proponen varias alternativas de trabajo futuro para implementar el trabajo realizado y que impacte en el área de la medicina, posibles mejoras a los modelos desarrollados y opciones diferentes de abordar este trabajo.

2. Marco Teórico

2.1. Aprendizaje Supervisado

El aprendizaje supervisado consiste en una base de datos que contiene características/-variables, x para cada muestra y asociada a una etiqueta, objetivo o clase, y, con esto, un algoritmo aprende a predecir y, a partir de x. El término Aprendizaje Supervisado se original de que la etiqueta y es provista por un "instructor" o "profesor" que le muestra a la máquina que debe hacer. En aprendizaje no supervisado no existe este instructor y por ende no existe y, por lo que la máquina debe aprender por su cuenta a obtener un resultada que "haga sentido" [64].

Dentro de esta categoría de Aprendizaje existen dos tipos de tareas, Clasificación y Regresión. La primera solicita a un programa computacional que indique a cuál de las k categorías pertenece la entrada, por lo tanto y indica una categoría identificado por un código numérico. Sin embargo, también es posible que la salida no sea la categoría como tal, sino, la distribución de probabilidad de que la entrada pertenezca a cada una de las clases. Un ejemplo de este tipo de tareas es el reconocimiento de imágenes. En donde la entrada es una imagen, interpretado por la máquina como un arreglo de valores, que dependiendo de la codificación, pueden ser la intensidad de brillo de cada píxel para cada color, y la salida es un código que identifica a un objeto presente en la imagen, tal como se ilustra en la figura 2.1. La segunda consiste en que la máquina predice un valor numérico, por ejemplo, con datos de casas, predecir su valor de mercado. Este tipo de predicciones son utilizadas ampliamente en el mercado financiero. La diferencia entre ambos es solamente el formato de la salida [64].

Figura 2.1: Ejemplo de una clasificación, en este caso es entregada la clase seleccionada en la parte superior de la imagen, Cerdo, y la probabilidad/confianza de esta predicción, 57.75 %. Fuente: [68].

2.2. Modelos

2.2.1. Algoritmos Naïve Bayes

Los algoritmos de Naïve Bayes son modelos de aprendizaje supervisado basados en la aplicación del teorema de Bayes, con la suposición "Naïve" de independencia condicional entre cada par de características dado el valor de las variables de clase [61]. El teorema de Bayes indica la siguiente relación [61], dado la variable de clase y y el vector de características dependientes x_1 hasta x_n:

$$P(y|x_1, ..., x_n) = \frac{P(y)P(x_1, ..., x_n|y)}{P(x_1, ..., x_n)}. \tag{2.1}$$

Usando la suposición de independencia condicional:

$$P(X_i|y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i|y). \tag{2.2}$$

Para cada i, la relación se simplifica a

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)}. \tag{2.3}$$

Dado que $P(x_1, ..., x_n)$ es constante dado la entrada, es posible simplificarlo de la forma:

$$P(y|x_1, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y). \tag{2.4}$$

Lo cual se simplifica a:

$$\hat{y} = \underset{y}{\text{argmax}}\ P(y)\prod_{i=1}^{n} P(x_i|y). \tag{2.5}$$

Se puede utilizar una estimación a Máxima a Posteriori (MAP) para estimar $P(y)$ y $P(x_i|y)$, el primero es entonces la frecuencia relativa de la clase y en el conjunto de entrenamiento [61].

La diferencia entre los algoritmos Naïve Bayes es principalmente la suposición realizada respecto a la distribución de $P(x_i|y)$ [61], entre las alternativas existentes está, la distribución Multinomial (MNB), la distribución Bernoulli (BNB), la distribución Categórica (CNB), la Gaussiana (GNB), entre otras.

2.2.2. Árboles de Decisión

Un árbol de decisión (DT) es un clasificador expresado como una partición recursiva del espacio de instancia. Los árboles de decisión consisten en nodos que forman un árbol enraizado, que significa que es un árbol dirigido con un nodo llamado raíz que no posee vértices entrantes. El resto de los nodos posee exactamente un vértice entrante. Un nodo con

vértices de salida, es llamado interno o nodo de prueba. El resto de los nodos son llamados hojas, también conocidos como nodos terminales o de decisión. En un árbol de decisión, cada nodo interno divide el espacio de instancia en 2 o más subespacios de acuerdo con cierta función discreta de los valores de los atributos de entrada. En los casos más simples y frecuentes, cada nodo de prueba considera solo un atributo, donde el espacio de instancia es particionado de acuerdo el valor que toma este atributo. En caso de atributo numéricos, sería una condición referente a un rango de valores [69].

Cada hoja es asignada a una clase que representa el valor objetivo más adecuado en base a la decisión tomada, sin embargo, también es posible que el árbol de decisión entregue un vector de probabilidades que indique la probabilidad de que el valor objetivo tome cierto valor. Las instancias son clasificadas "navegando" desde la raíz (arriba) hacia las hojas (abajo), según los valores resultantes en los nodos de prueba [69]. La figura 2.2 describe un árbol de decisión que indica si un alumno aprueba o no un curso, los nodos internos están representados por círculos, la raíz está representada por un cuadrado y las hojas por rombos. Cada nodo tiene etiquetado el atributo que prueba, que en este caso son las notas del alumno, las ramas o vértices están etiquetados con los valores correspondientes.

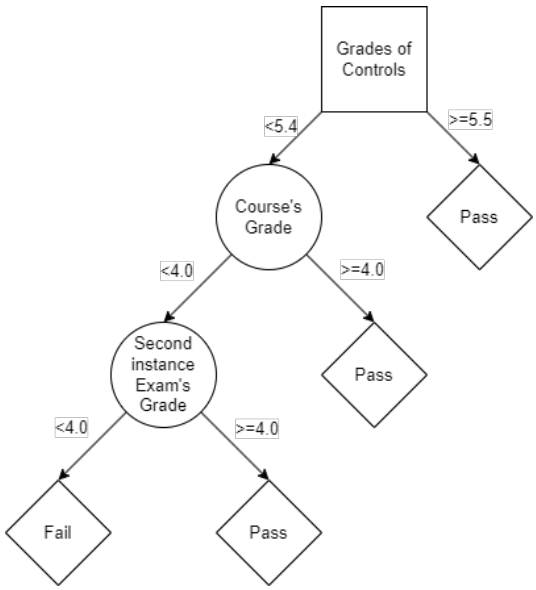

Figura 2.2: Árbol de decisión representando la aprobación de un alumno a un curso.

2.2.3. *Support Vector Machine*

Support Vector Machine (SVM) es un algoritmo de aprendizaje supervisado en el cual el vector de entrada es mapeado a un espacio de características de muy alta dimensionalidad, mediante una función Kernel [70]. Esta transformación permite a las SVMs encontrar una función de decisión, lineal o no lineal, llamado hiperplano, que puede separar de la forma más efectiva los datos. Un hiperplano óptimo es definido si se consigue el máximo margen entre el vector de dos clases (considerando un problema de solo dos clases), el margen es la distancia de separación entre ambas clases [61, 70]. En la figura 2.3 se muestra un ejemplo

de función de decisión en un problema de separación lineal. En el ejemplo se muestran tres bordes de margenes para maximizar la separación de dos clases.

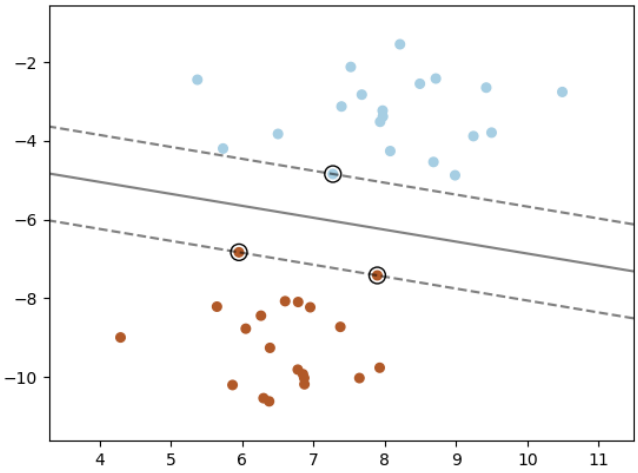

Figura 2.3: Ejemplo de función de decisión en un problema de separación lineal, con tres ejemplos de bordes de márgenes, llamados "*Support Vectors*". Fuente: [61].

Las funciones de Kernel que transforman el espacio de características son diversas, las más comunes son:

- Kernel Lineal: $\langle x, x' \rangle$.
- Kernel Polinomial: $(\gamma \langle x, x' \rangle + r)^d$.
- Kernel *Radial Basis Function* (RBF): $\exp(-\gamma \|x - x'\|^2)$.
- Kernel Sigmoide: $\tanh(\gamma \langle x, x' \rangle + r)$.

Donde γ, r y d son parámetros ajustables, $\langle \cdot, \cdot \rangle$ es la operación producto interno/punto y $\|x - x'\|^2$ es la distancia euclidiana al cuadrado entre dos vectores de características [61].

2.2.4. Perceptrón MultiCapas

Los perceptrones multicapas (MLPs), también conocidos como *Deep feedforward networks* o Redes Neuronales, tienen como objetivo el aproximar una función f^*. Por ejemplo, para un problema de clasificación, $y = f^*(x)$ mapea una entrada x a la clase y. Las MLPs definen un mapeamiento $y = f(x; \theta)$, y aprende los valores de los parámetros θ que resulta en la mejor función de aproximación [64].

Son llamados *feedforward* debido a que la información fluye a través de la función siendo evaluada en x, a través de las computaciones intermedias usadas para definir f, y finalmente

la salida y. No existe una retroalimentación, *feedback*, que conecte la salida consigo mismo [64].

Las MLPs son redes debido a que se puede representar como una composición de diferentes funciones, de la forma $f(x) = f^{(n)}(f^{(n-1)}(...(f^{(2)}(f^{(1)}(x)))))$, donde $f^{(1)}$ es llamada la primera capa de la red, $f^{(2)}$, la segunda capa, y así sucesivamente. El largo de esta cadena es lo que le da profundidad al modelo, y el motivo de porque se llama Aprendizaje Profundo. La capa final es conocida como capa de salida y esta debe dar como resultado un valor cercano a y. El comportamiento de las otras redes no es especificado directamente por los datos de entrenamiento, el algoritmo de aprendizaje solo debe decidir cómo usar estas capas para producir la salida deseada. Debido a que los datos de entrenamiento no muestran como la salida deseada a cada una de estas capas, es que son llamadas capas ocultas. Son llamadas redes neuronales, debido a que están inspiradas en la neurociencia, donde las capas son vectores, donde los elementos de los vectores juegan un rol similar a las neuronas. Estos elementos son conocidos como unidades [64].

Las capas multiplican la entrada x o la salida de la capa anterior, según corresponda, por un vector de peso y le añade un valor conocido como *bias* o sesgo. A la salida de capa oculta se le aplica una función, generalmente no lineal, llamada función de activación, las cuales introducen una relación no lineal entre las entradas, son llamadas funciones de activación ya que deciden si una neurona es activada o no [64, 71]. Las funciones de activación más comunes son ReLU (*Rectified Linear Unit*), Sigmoide y Tangente Hiperbólica [72]. Lo anterior se puede resumir en la ecuación 2.6, donde act es la función de activación, x la entrada que recibe la capa, x' la salida de la capa, W el vector de peso que posee la red y b el *bias* añadido.

$$x' = act(x * W + b). \tag{2.6}$$

Para ajustar los vectores de peso y el *bias* se aplica una optimización basada en gradientes, la cual tiene como objetivo disminuir una función objetivo, conocida como función de pérdida o costo. Este método se aplica utilizando el algoritmo de *back-propagation*, la cual permite que la información de la pérdida fluya hacia atrás a través de la red con el fin de computar estos gradientes [64]. Con los gradientes computados, se puede utilizar un optimizador quien realizara el aprendizaje de la red utilizando los gradiente, los optimizadores más conocidos son *Adam*, *Stochastic Gradient Descent* (SGD) y *RMSprop* [73, 74].

En la figura 2.4 aparece un ejemplo de una MLP con una sola capa oculta, incluyendo *Bias*.

2.2.5. K vecinos más cercanos

El algoritmo K vecinos más cercanos (KNN) es un método de aprendizaje supervisado no paramétrico en que un elemento/objeto es clasificado por un voto mayoritario de sus vecinos [61, 75]. Un elemento es asignado a la clase que tenga la mayor cantidad de representaciones con los vecinos más cercanos de ese punto. La cercanía de un punto a los vecinos es calculada mediante el uso de una métrica de distancia. El número de vecinos cercanos k es un valor entero elegido arbitrariamente por el usuario. Si el número de muestras es alto, se espera

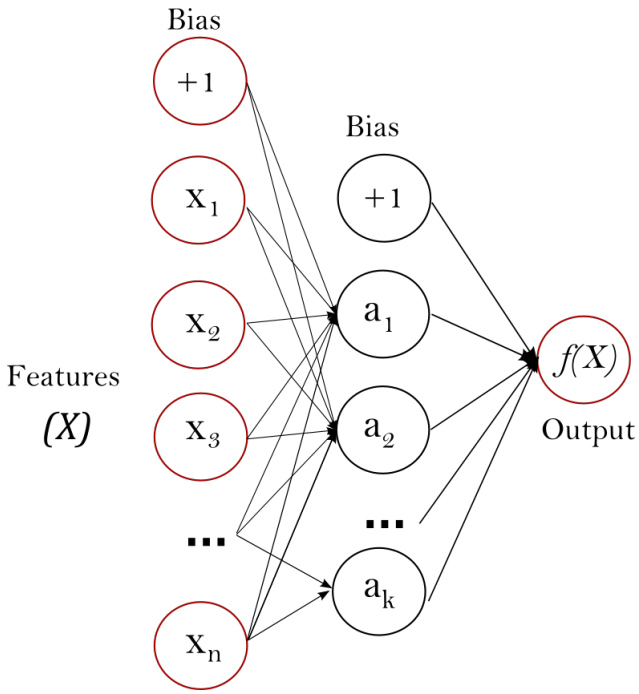

Figura 2.4: Ejemplo de MLP con una solo capa oculta. Fuente: [61].

utilizar un valor elevado para k con el fin de tener un voto mayoritario más fidedigno, pero a su vez se espera que el valor de k sea bajo, en proporción al número de ejemplos, para que los puntos cercanos al elemento a clasificar, x, den un valor estimado acertado de las probabilidades posteriores de la verdadera clase de x [61, 75].

2.2.6. Regresión Logística

Regresión Logística (LR) es un modelo estadístico que puede ser usado para clasificación y análisis predictivo. LR estima la probabilidad de un ocurrir un evento basado en las variables independientes de la base de datos, esta probabilidad, entre 0 y 1, es obtenida debido al uso de la función logística/sigmoide, presente en la ecuación 2.7[61–63].

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}. \tag{2.7}$$

Para LR, x es remplazado por $x_i A + B$, donde x_i son los datos del punto i, A es la pendiente de la curva y B es el intercepto, son A y B los valores que son optimizados en el entrenamiento, el optimizador a utilizar es una variable ajustable [61, 62, 70].

En la figura 2.5 se muestra un ejemplo de una LR mostrando la probabilidad de pasar un examen versus horas de estudio.

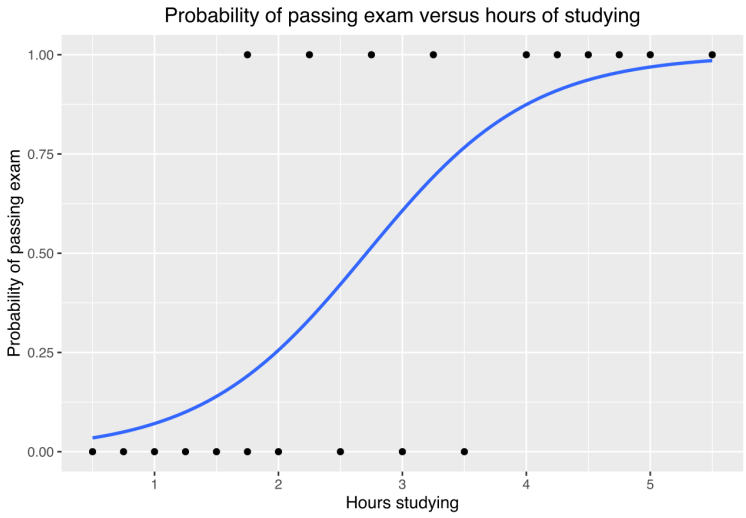

Figura 2.5: Ejemplo de LR mostrando la probabilidad de pasar un examen según horas de estudio. Fuente: [76].

2.2.7. *Random Forest*

Random Forest (RF) es una combinación de predictores de tipo árbol, en que cada árbol "crece" a partir del conjunto de entrenamiento y de un vector aleatorio generado independientemente, y con la misma distribución para todos los árboles del "bosque", luego la clasificación más popular, la elegida por la mayor cantidad de árboles, es la seleccionada por la RF [65]. Este sistema de votación lo convierte en un método de tipo *Ensemble*. Los árboles son comúnmente Árboles de Decisión. Adicionalmente, es posible que cada árbol "crezca" a partir de un muestreo del conjunto de entrenamiento, en vez del conjunto completo, lo cual se denomina *Bootstrap Sample* [61, 65]. El aumentar el número de árboles no produce *Overfitting*, pero produce un valor límite del error de generalización [65].

En la figura 2.6 se ilustra el funcionamiento de una RF, incluido el sistema de votación.

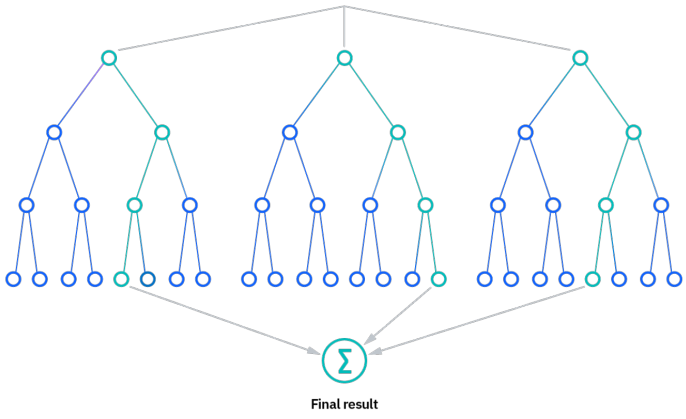

Figura 2.6: Ejemplo de funcionamiento de Random Forest. Fuente: [77].

2.2.7.1. *Balanced Random Forest*

Balanced Random Forest (BRF) es un algoritmo basado en *Random Forest*, el cual toma en consideración el desbalance de las clases. En este algoritmo el *Bootstrap Sample* realizado aplica *under-sample* aleatorio con el fin de balancear las clases en el entrenamiento, en particular, el proceso realiza el *Bootstrap Sample* en la clase minoritaria, posteriormente, elige aleatoriamente un número de casos aleatorios del mismo tamaño que los elegidos para la clase minoritaria, con reemplazo de la clase mayoritaria [78].

2.2.8. *Extra-Trees*

Extremely Randomized Trees también conocido como *Extra-Trees* (ET) es un método de aprendizaje supervisado de tipo *Ensemble*, que, similar a *Random Forest*, utiliza clasificadores de tipo árbol. La principal diferencia entre ambos es la utilización de la aleatoriedad, de ahí su nombre, ya que ET crea nodos de prueba/división de los Árboles de Decisión de forma aleatoria, mientras RF los crea en base a la mejor división posible. Esto puede provocar que ET sea más rápido y obtenga una menor varianza y sesgo respecto a RF. Lo primero debido a que los tiempos de búsqueda de la mejor división posible deberían ser más elevados que uno aleatorio y lo segundo debido a que el modelo no se ve influenciado a ciertos patrones o variables presentes en el conjunto de datos que pueden tender a crear resultados indeseados [79].

2.2.9. *Gradient Boosting Machine*

Gradient Boosting Machine (GBM) es una técnica de Aprendizaje de Máquinas de tipo *Ensemble* que utiliza múltiples algoritmos simples, denominados *weak learners* o *base learner* que usualmente son Árboles de clasificación, i.e., Árboles de Decisión. La técnica de *Boosting* consiste en que cada vez que un *weak learner* es añadido, este se centra más en los errores que cometieron sus predecesores, añadiéndoles más peso a estos errores, y disminuyendo los pesos de los valores correctamente predichos. El nombre *Gradient* proviene de que el optimizador utilizado es Gradiente Descendente, el cual minimiza el error en cada iteración a la mejor dirección (gradiente) que permita reducir el error de predicción [66].

También se utilizan técnicas de regularización, que previenen el sobreajuste, por ejemplo se añade el la técnica de *Shrinkage* que reduce el impacto/contribución de cada nuevo *weak learner* mediante un parámetro llamado tasa de aprendizaje, lo cual se ve expresado en la ecuación 2.8 [66].

$$F_m(x) = F_{m-1}(x) + v \cdot \rho_m h(x; a_m). \tag{2.8}$$

Donde v es la tasa de aprendizaje, x la entrada del modelo, $h(x; a_m)$ es el *weak learner* m creado en base a la entrada y los parámetros a_m, ρ_m es el valor óptimo que se debe ajustar el

modelo, calculado mediante el proceso de gradiente descendente, F_{m-1} es el modelo formado por $m-1$ *weak learners*.

2.2.9.1. *XGBoost* y *LightGBM*

Extreme Gradient Boosting, *XGBoost* (XGB) [80] y *Light Gradient Boosting Machine*, *LightGBM* (LGBM) [81] son dos de las implementaciones más populares de *Gradient Boosting Machine* que añaden la opción de selección de variables, manejo de datos faltantes, *early stopping* y varias opciones de configuración de hiperparámetros.

2.3. Selección de Variables

La selección de variables/características es un proceso importante donde se remueven información irrelevante o redundante entre las entradas, lo cual evita problemas de sobreajuste y permite obtener un mejor rendimiento que utilizar todas las variables, más aún cuando el algoritmo utilizado no tiene forma de lidiar con este tipo de variables que ocasionan estos problemas [82]. Existen diversos métodos o métricas estadísticas para calcular la relevancia de cada variable:

2.3.1. F-Test ANOVA

En Análisis de Varianza (ANOVA), el F-Test o valor F es una medida que calcula la razón entre la variabilidad entre grupos/poblaciones (inter) y dentro del grupo (intra), con el fin de aceptar o rechazar la hipótesis nula, que plantea que el promedio de todos los grupos es igual [83], la ecuación 2.9 presenta la fórmula utilizada para calcular el valor F.

$$F = \frac{\text{Varianza Intergrupos}}{\text{Varianza Intragrupos}}. \tag{2.9}$$

El valor obtenido es comparado con un valor critico de la distribución F para determinar si la diferencia entre los promedios de los grupos es significativa o no. Si el valor F es elevado, entonces, la hipótesis nula es rechazada, existe una diferencia significativa entre los promedios de los grupos, lo contrario sucede si el valor es pequeño [83]. En selección de variables, F-test es utilizado para estimar el grado de dependencia lineal entre dos variables aleatorias [61].

2.3.2. Test Chi cuadrado

Chi cuadrado (χ^2) es un estadístico que examina si dos variables categóricas son independientes, planteando una hipótesis similar a la planteada por F-Test ANOVA, si el estadístico de la prueba sigue una distribución χ^2, entonces existe una independencia entre las variables.

En el caso de selección de variables, una independencia significa que la variable es independiente del objetivo y por ende irrelevante para la clasificación [61].

2.3.3. Información Mutua

Información mutua, también conocida como ganancia de información, es la medida de dependencia mutua entre dos variables, si el valor es cero, ambas variables son independientes, mientras que un mayor valor indica una mayor dependencia, ya que cuantifica cuanta información es obtenida sobre una variable a partir de otra [61].

2.3.4. Métodos basados en Árboles

Los modelos de basados en Árboles, Árboles de Decisión, *Random Forest* y *Balanced Random Forest*, al crear los nodos de los Árboles, utilizan las variables que mejor se adecuen a la labor de clasificación, de acuerdo con fórmulas de pureza, por lo que es posible obtener el nivel de importancia de las variables en la creación de estos clasificadores y utilizarlo para una selección de variables.

2.4. Generalización

Una de las principales ideas de los algoritmos de Aprendizaje de Máquinas es que puedan ofrecer un buen desempeño sobre datos nuevos, "no vistos" con anterioridad, es decir, ejemplos con los que el modelo no se haya entrenado, esta habilidad se le llamada Generalización. Un modelo que tenga mejor rendimiento en métricas, i.e., error, que otro, significa que tiene mejor capacidad de generalización [64].

Para calcular la capacidad de generalización se utiliza un conjunto de prueba/testeo que contiene muestras no contenidas en el conjunto de entrenamiento. Este conjunto de prueba puede ser creado a partir del set de datos completo, realizando particiones de este, o adquiriendo datos externos, esta última opción puede ser un reflejo más fidedigno del mundo real [64, 84].

Con el uso de las particiones aparecen dos temas adicionales, sobreajuste e infrajuste, sobreajuste sucede cuando el error del conjunto de entrenamiento y prueba son muy diferentes, esto debido a que el modelo tiene una capacidad y memoriza las características y patrones del conjunto de entrenamiento, que no pueden no ser útiles para realizar una predicción sobre los datos del conjunto de prueba. Infrajuste ocurre cuando el modelo no logra obtener un error bajo en el conjunto de entrenamiento, lo cual puede estar asociado a una baja capacidad del modelo para crear una función que se ajuste al comportamiento adecuado para realizar la predicción en el conjunto de entrenamiento [64].

Sobreajuste e Infrajuste también pueden deberse a los datos, sobreajuste puede ocurrir si los datos de entrenamiento y prueba no son representativos uno del otro, es decir, que

los datos de cada conjunto no compartan muchas similitudes entre ambos [84]. Por ejemplo, en un clasificador binario de imágenes de perros y gatos, que el conjunto de entrenamiento contenga muchas imágenes de perros y pocas de gatos y el conjunto de prueba sea compuesto mayoritariamente de imágenes de gatos y minoritariamente de gatos, lo cual puede ser corregido realizando una nueva partición. Por otro lado, infrajuste puede suceder si los datos utilizados no son lo suficientemente apropiados para la labor de predicción [84]. Por ejemplo, predecir el ganador de un partido de fútbol utilizando datos climatológicos, en este caso puede existir una relación entre el desempeño de los equipos en base al clima, pero, probablemente, no sea tan estrecha como los datos del rendimiento de ambos equipos en los últimos n partidos. La utilidad de los datos es un riesgo al iniciar una investigación.

2.5. Hiperparámetros

La mayoría de los modelos de Aprendizaje de Máquinas, incluyendo los mencionados en la sección Modelos, poseen hiperparámetros, que son valores ajustables que modifican la capacidad de ajustarse a un problema o a otro. En la figura 2.7 se puede ver el comportamiento de un modelo al variar el hiperparámetro λ, *weight decay*, en un problema de regresión polinomial.

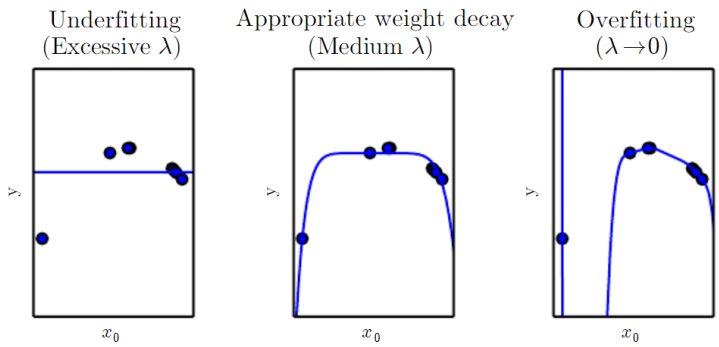

Figura 2.7: Ejemplo de diferentes comportamientos de un modelo según el valor de un hiperparámetro. Fuente: [64].

Debido a esto es necesario optimizar la selección de hiperparámetros, lo cual, por un lado, no puede ser en base al conjunto de prueba, ya que ninguna optimización debe ser en base a estos datos, ya que se crearía un sesgo y no sería un fiel reflejo del mundo real. Tampoco puede ser en base al conjunto de entrenamiento, ya que puede generar problemas de sobreajuste. La solución a esto, crear una nueva partición, llamada conjunto de validación, la cual, similar al conjunto de prueba, deben ser datos que el modelo no debe utilizar para su entrenamiento a la hora de optimizar los hiperparámetros [64].

2.5.1. Búsqueda de Grilla y Validación Cruzada

El número de hiperparámetros puede variar entre modelos. Sin embargo, los modelos más complejos normalmente poseen una alta cantidad, produciéndose un número elevado de

combinaciones, la búsqueda de la combinación de óptima de estas se denomina búsqueda de grilla. Es necesario un conjunto de validación para realizar esta optimización. No obstante, existe la posibilidad de que el conjunto de validación no sea necesario, utilizando la técnica de Validación Cruzada (CV). Esta consiste en realizar diversos entrenamientos y pruebas subdividiendo el conjunto de entrenamiento, en un conjunto de entrenamiento (nuevo) y un conjunto de validación, pero en cada entrenamiento, estas particiones son distintas. La más común la Validación Cruzada de k-hojas, la cual el conjunto se divide en k particiones, sin repetición, siendo, en cada iteración, un conjunto de entrenamiento compuesto por $k-1$ particiones distintas, la partición restante es usada como conjunto de validación, testeo [61, 64]. En la figura 2.8 se ejemplifica esto con k igual a 5.

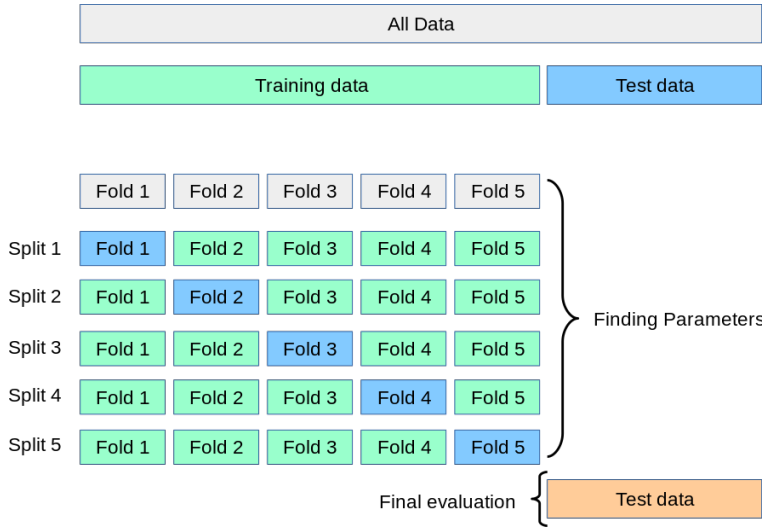

Figura 2.8: Ejemplo de Validación Cruzada de 5-hojas. Fuente: [61].

A pesar de los beneficios, uno de los importantes puntos en contra de CV es su elevado costo computacional, ya que se debe entrenar varias veces un modelo, sumado a la búsqueda de grilla, esto aumenta aún más [61].

Todo el proceso de entrenamiento y optimización de hiperparámetros puede ser resumido en el diagrama de la figura 2.9.

2.6. Desbalance de Clases y Datos faltantes

El desbalance de clases es cuando la cantidad de muestras/ejemplos respecto a una o varias clases es muy inferior a otra clase/s [85]. Por ejemplo, esto ocurre en medicina, donde la prevalencia de ciertas enfermedades en general afecta menos del 20 % de la población. Por lo que al recopilar información al respecto, es esperable que aproximadamente ese porcentaje de la muestra seleccionada tenga la condición (clase positiva) y el porcentaje restante, y mayoritario, no lo tenga (clase negativa), esto puede acarrear en consecuencias negativas en el entrenamiento, ya que el desbalance puede provocar sobreajuste. Para lidiar con este problema existen diversas soluciones, las más comunes son el submuestreo y sobremuestreo. Submuestreo consiste en tomar menos muestras de la clase mayoritaria, para emparejar la

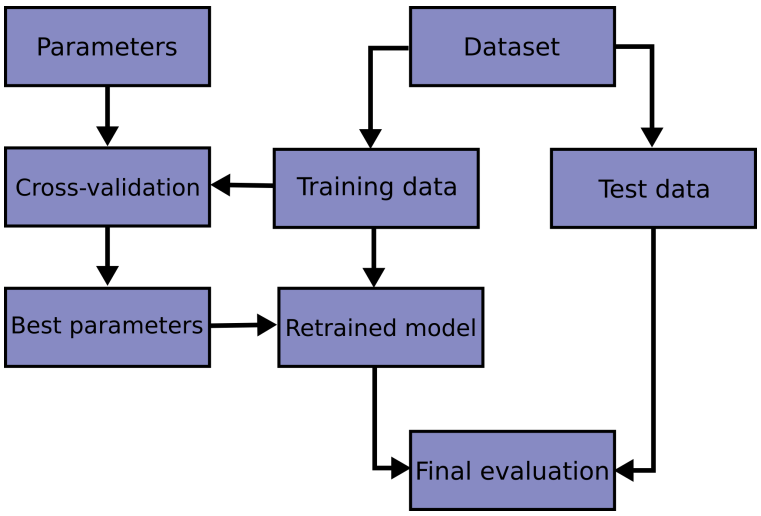

Figura 2.9: Diagrama de flujo del proceso de entrenamiento y optimización de hiperparámetros. Fuente: [61].

cantidad de ejemplos de las clases, mientras que sobremuestreo es generar nuevas muestras de la clase minoritaria para igualar las muestras de la clase mayoritaria, esta creación de nuevas muestras puede ser muestreando con repetición y de forma aleatoria en la clase minoritaria, o de forma sintética, creando ejemplos similares a los originales [85].

Los datos faltantes, como su nombre lo indica, es cuando faltan datos de una o más variable, en una o más muestras, provocado por la falta de disponibilidad de los datos, mal ingreso al registrar los datos, corrupción en los datos, entre otros. A parte de ser un problema por la pérdida de información, ciertos modelos no toleran la ausencia de datos, por lo que deben ser preprocesados. Entre las alternativas existentes están, eliminar las muestras con datos faltantes, eliminar las variables con datos faltantes, común si la cantidad pérdida es elevada, reemplazar los valores faltantes, puede ser por el promedio o mediana de la variable u otra opción [61]. Sin embargo, esta última opción no siempre puede ser aplicada, ya que se asume que es correcto que el valor con el que se reemplaza es representativo y tiene lógica, lo cual puede no ocurrir, si esta opción se utiliza y es incorrecto reemplazar por un valor, el modelo entrenado puede funcionar de forma incorrecta debido a posibles contradicciones en los datos [86].

2.7. Aumentación de Datos

Aumentación de Datos (DA) es un método comúnmente utilizado en aprendizaje de máquinas que consiste en la creación de nuevas muestras a partir de datos ya existentes, con el fin de mejorar el desempeño de los modelos en el entrenamiento, este método es popular en tareas de imágenes, ya que se pueden crear nuevas imágenes alterando su brillo, contraste, saturación, colores o incluso rotándolas respecto a su posición original. En datos tabulares, este proceso modifica los valores de las variables para producir nuevas muestras [87–89]. En la figura 2.10 se muestra un ejemplo de DA donde se aplican diversas técnicas con el fin de crear nuevas muestras.

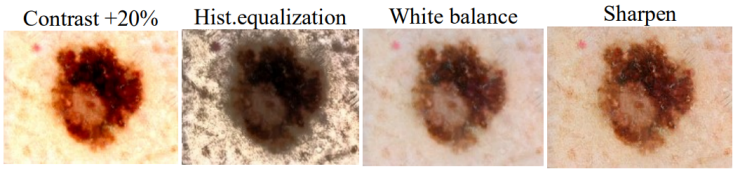

Figura 2.10: Ejemplo de Aumentación de Datos. Fuente: [89].

2.8. Transformación de datos

Las variables de un conjunto de datos pueden venir en distintas escalas. Por ejemplo, en el caso ficticio de un data set de viajes de una aerolínea, la edad varía entre 0 y 100 años, mientras que los kilómetros recorridos puede alcanzar valores desde 0 hasta una cantidad alta, mientras que si el cliente pertenece a club de viajero de frecuente es una variable categórica/binaria, 0 si no pertenece, 1 si pertenece. No obstante, todas estas variables deberían tener, a priori, un impacto similar en la aplicación que se quiera desarrollar, independiente de los valores que pueden alcanzar, para esto es necesario transformar los datos (preprocesamiento) a un rango de valores similar entre ellos, existen diversas alternativas, dos de las más frecuentes son estandarización y escalamiento de mínimo-máximo [61].

Estandarización, también conocido normalización o *Z-score*, consiste en restarle al valor el promedio de la variable y dividirlo por la desviación estándar [61]. En la ecuación 2.10 se ve la fórmula de estandarización, donde z es valor transformado, x es el valor a transformar, μ es el promedio de la población o variable, σ su desviación estándar.

$$z = \frac{x - \mu}{\sigma}. \tag{2.10}$$

Escalamiento de mínimo-máximo transforma los valores entre dos rangos, típicamente entre 0 y 1, esto se logra restando el valor, x, con el mínimo de la variable, x_{min}, y dividiéndole por la diferencia entre el máximo, x_{max}, y el mínimo, obteniéndose el valor escalado, x_{trans} [61]. Lo anterior está representado en la ecuación 2.11.

$$x_{trans} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{2.11}$$

Algunos algoritmos de Aprendizaje de Máquinas requieren de este tipo de preprocesamiento para funcionar [61].

2.9. Métricas

Diversas métricas existen para evaluar el desempeño de los modelos según la tarea que deben abordar, al tratarse de un problema de clasificación y binario (clase negativa, no posee la condición, positiva, posee la condición/enfermedad) se utilizan diversas métricas entre las cuales están:

2.9.1. *Accuracy*

Accuracy, exactitud, mide la proporción de casos clasificados correctamente respecto al total. La fórmula para tareas binarias es mostrada en la ecuación 2.12, donde TP, *True Positives*, Verdaderos Positivos, es la cantidad de casos positivos etiquetados correctamente. TN, *True Negatives*, Verdaderos Negativos, es la cantidad de casos negativos etiquetados correctamente. FN, *False Negatives*, Falsos Negativos, es la cantidad de casos positivos etiquetados incorrectamente, es decir, como negativos. FP, *False Positives*, Falsos Positivos, es la cantidad de casos negativos etiquetados incorrectamente. El valor de *Accuracy* puede variar entre 0 y 1, donde 0 significa que toda la clasificación realizada es incorrecta, mientras que 1 significa que todos los casos fueron etiquetados correctamente, por ende, un valor más elevado es mejor. A pesar de esto, uno de los problemas de esta medida, es que no toma en cuenta el desbalance de clase, por lo que, si una clase es muy mayoritaria respecto a otra, se podría tener una alto *Accuracy*, a pesar de fallar en la clasificación de la clase minoritaria [90].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}. \tag{2.12}$$

2.9.2. *Sensitivity*

Sensitivity, sensibilidad, también conocido como tasa de verdaderos positivo (TPR) o *Recall*, en clasificación binaria, es la tasa de positivos detectados correctamente, que se puede interpretar como la probabilidad de que, dado que se clasificó como positivo, es realmente positivo. Los valores posibles van entre 0 y 1, con 0 igual a que ningún caso positivo fue detectado correctamente y 1 equivalente a que todos los casos positivos fueron detectados correctamente. Esta métrica es muy útil en medicina, ya que si se posee un valor elevado esto significa que, de ser etiquetado como negativo, difícilmente será un caso positivo mal clasificado, por lo que es útil para medir el desempeño de pruebas médicas. Sin embargo, una *Sensitivity* igual a 1, por sí solo no es totalmente útil, ya que si todos los casos, positivos o negativos, son clasificados como positivos, teniéndose esta métrica igual a 1, también se tendría una tasa de falsos positivos igual a 1, con lo que no se podría distinguir entre positivos y negativos, por lo que es necesario complementar esta métrica [91]. La fórmula utilizada para calcular esta métrica aparece en la ecuación 2.13.

$$Sensitivity = \frac{TP}{TP + FN}. \tag{2.13}$$

2.9.3. *Specificity*

Specificity, especificidad, tasa de verdaderos negativos (TPR). En clasificación binaria, es la tasa de negativos detectados correctamente, por lo que podría denominarse como la "Sensibilidad" de la clase negativa, de hecho en problemas multiclase, es calculado el *Recall* de cada clase con la fórmula de sensibilidad, solo que reemplazando los valores según la

clase a medir. Por lo tanto la ecuación de esta métrica es idéntica a la de sensibilidad, solo que tomando en consideración los negativos (ver ecuación 2.14). La interpretación que se le puede atribuir a esta métrica es, dado que se detectó como negativo, que probabilidad es que realmente sea negativo. Similar a sensibilidad, esta métrica también varía entre 0 y 1 y posee el mismo problema que esa métrica, valor igual a 1 no significa necesariamente una buena predicción, ya que puede tener una alta tasa de falsos negativos, por lo que también es buena idea complementar esta métrica con otras [91].

$$Specificity = \frac{TN}{TN + FP}. \tag{2.14}$$

2.9.4. *Recall Macro*

Recall Macro es el promedio del *Recall* de cada clase, en problemas binarios, sensibilidad y especificidad son el *recall* de la clase positiva y negativa respectivamente, por lo tanto, esta métrica es el promedio de ambas métricas. Esta métrica tiene la utilidad de que brinda una idea de que tan bien clasifica las diferentes clases. Debido a que es compuesta por otras métricas que poseen un rango entre 0 y 1, esta métrica también posee ese rango, donde un valor más elevado es mejor [92].

$$Recall\ Macro = \frac{Sensitivity + Specificity}{2}. \tag{2.15}$$

2.9.5. AUCROC

La curva ROC, *Receiver Operating Characteristics*, es una curva usada comúnmente en toma de decisiones médicas, aprendizaje de máquinas y minería de datos [93]. Esta curva bidimensional representa en el eje Y la tasa de verdaderos positivo, Sensibilidad y en el eje X, la tasa de falsos positivos, lo que sería equivalente a 1-Especificidad. Esta curva se crea con los distintos umbrales de decisión que varían las métricas de sensibilidad y especificidad, ya que un umbral de decisión, que varía entre 0 y 1, indica con que valor de salida de un modelo, en general una probabilidad o puntuación, también entre 0 y 1, una entrada es clasificada como positiva o negativa. Por ejemplo, si la salida de un modelo para un caso entrega el valor de 0.6, con un umbral de 0.4, el caso es clasificado como positivo, pero, para un umbral de 0.8, este caso es considerado negativo [93].

Para comparar múltiples clasificadores, es difícil comparar un elemento visual, las curvas, lo idóneo es reducir estas curvas a un valor escalar. Un método común es calcular el área bajo la curva (AUC) de la curva ROC, lo que se denomina, AUCROC o AUC [93, 94]. Al ser un área de una porción de una unidad cuadrada, este valor varía entre 0 y 1, sin embargo, debido a que una predicción aleatoria debería producir una línea diagonal entre (0,0) y (1,1), 1 de especificidad, 0 de sensibilidad y 0 de especificidad y 1 de sensibilidad respectivamente, tiene un área de 0.5, ningún clasificador realista debería tener un valor menor a 0.5, de lo contrario, se podría tener un mejor desempeño "haciendo caso" contrario a lo que indica el

clasificador [93, 94].

Esta métrica es ampliamente utilizada por gran parte de los trabajos del estado del arte.

La elección de modelos utilizados en la tesis se debe a que en el análisis del estado del arte se determina que han sido los más exitosos en problemas similares. Los modelos basados en árboles, como *Random Forest*, *Extra-Trees* y *Gradient Boosting Machine*, Regresión Logística y MLP son particularmente escogidos por ser utilizados por varios modelos del estado del arte, obteniéndose buenos resultados. Modelos como Naïve Bayes, árboles de decisión, K vecinos más cercanos y SVM son escogidos por ser modelos no tan complejos que tienen un buen rendimiento. Específicamente, los diversos modelos mencionados permiten analizar el desempeño de múltiples algoritmos que pueden presentar comportarse de forma distinta en los problemas abordados. A priori, no existen un método superior al resto. Estos modelos pueden ser optimizados para maximizar su rendimiento.

La selección de variables y la aumentación de datos ha tenido un buen desempeño en tareas de ML [82, 89]. Los métodos de selección de variables presentados permiten elegir distintas combinaciones de variables que permiten por un lado mejorar el desempeño de los modelos y por otro reducir el número de variables para aumentar las posibilidades de que los modelos resultantes sean utilizados en ambientes clínicos. La aumentación de datos es un método popular en ML que permite mejorar el desempeño de los modelos creados, aumentando el número de datos con los que se entrena el modelo.

3. Metodología

3.1. Bases de Datos

Las bases de datos utilizadas fueron obtenidas de la unidad de obstetricia y medicina fetal del Hospital Parroquial de San Bernardo, Santiago, Chile, proveniente de pacientes embarazadas atendidas en ese centro asistencial. La base de datos para predecir GDM y la de LGA son distintas, a pesar de compartir algunos registros en común. Los datos utilizados para GDM provienen de registros recopilados entre el 2019 y 2022, mientras que los registros para LGA abarcan embarazados ocurridos entre el 2016 y 2022.

El diagnóstico de GDM, en ambas bases de datos, fue realizado utilizando el criterio de propuesto por la *International Association of Diabetes and Pregnancy Study Group* (IADPSG) [95], basado en los resultados reportados por en el estudio de *Hyperglycaemia and Adverse Pregnancy Outcomes* (HAPO) [96], que se convirtió en el estándar recomendado por la Organización Mundial de la Salud (WHO) en 2013 [97]. Este criterio IADPSG/HAPO/WHO 2013 para la diabetes gestacional contempla el diagnóstico la realización de una prueba de tolerancia a la glucosa (OGTT) con una carga de 75g de glucosa, este examen es efectuado en el segundo trimestre de embarazo, normalmente entre las semanas 24 y 28 de gestación, para ser considerado positivo, se debe tener un registro de glicemia en ayuna mayor o igual a 92 mg/dL o mayor o igual a 153 mg/dL en la medición realizada 2 horas después de consumida la carga (postprandial) [95–97].

Para el diagnóstico de LGA, fue utilizado el peso al nacer del feto, utilizando las curvas recomendadas por la Sociedad Chilena de Pediatría (SOCHIPE), Curvas de Alarcón-Pittaluga [98]. Con el ultrasonido rutinario efectuado en el tercer trimestre de gestación se efectuó la predicción rutinaria de LGA, utilizando la curva de crecimiento propuesta por Hadlock y compañía en 1991 [45], utilizada a nivel internacional.

Las bases de datos están compuestas de información recopilada en diversas instancias del embarazo. La primera visita materna, la última visita materna, el ultrasonido rutinario en el tercer trimestre de embarazo, con el que se efectúan las predicciones de LGA utilizando el criterio de Hadlock 1991 [45]. Los exámenes de glicemia rutinarios, efectuados en el primer trimestre y en el segundo trimestre, con este último se realiza el diagnóstico de GDM. Dentro de la información contenida en estas bases de datos se destaca, la edad y altura de la embarazada, peso y BMI en ambas visitas, datos de embarazos previos, consumo de drogas licitas e ilícitas, condiciones médicas preexistentes e información sobre el ultrasonido de tercer trimestre y el parto.

Por razones de confidencialidad, estos datos no están disponibles públicamente.

3.1.1. Preprocesamiento

Ambas bases de datos contienen información adicional irrelevantes para la realización de la predicción, pero algunas con posibles utilidades en tareas estadísticas, tales como información postparto, identificación, dentro de la base de datos, fecha del parto, entre otros. Por ende, se eliminan como entradas para los modelos predictivos. Las bases de datos contienen datos incompletos, por lo que se remueven a los pacientes con registros incompletos, ya que reemplazar los valores faltantes por otros valores pueden provocar sesgo, además en aplicaciones similares también se efectúa este proceso [23, 27, 29].

Ya que para GDM se busca una predicción temprana, se limita la fecha de la primera visita médica hasta 20 semanas de gestación, como se muestra en la figura 3.1, el primer control de las pacientes luego de este filtro varía entre las 4 y 20 semanas de gestación, concentrándose principalmente entre las semanas 7 y 14.

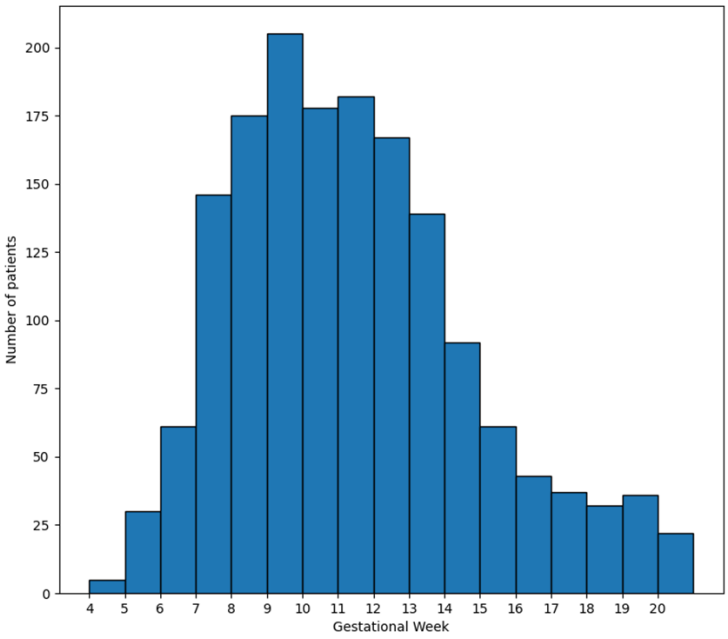

Figura 3.1: Histograma que muestra el número de pacientes según la fecha de la primera visita gestacional para GDM.

Para GDM se eliminan también a los pacientes diagnosticados con Diabetes Mellitus antes del embarazo, Diabetes Mellitus Pregestacional (PGDM), ya que por definición son mutuamente excluyentes, así como toda variable recopilada posterior a las 20 semanas de gestación.

Para LGA los criterios a aplicar son, primera visita maternal hasta las 15 semanas de gestación y última visita posterior a las 24 semanas de embarazo, se eliminan pacientes con fetos con edad gestacional dudosa y solo son incluidos embarazos *singleton*, con un solo feto.

Con los criterios previamente mencionados, los registros disponibles para GDM y LGA son respectivamente 1611 y 1802, de los 1611, 1382 son negativos para GDM y 229 positivos, mientras que estos valores para LGA son 1514 y 288 respectivamente. Teniéndose 30 variables

disponibles para GDM y 59 para LGA.

Las variables no binarias para las entradas de los modelos, por ejemplo, edad, peso, BMI, son estandarizadas/normalizadas, es decir, el valor fue restado por el promedio y dividido por la desviación de la variable, ambas medidas calculadas en base al conjunto de entrenamiento. La base de datos es dividida en tres particiones, conjunto de entrenamiento, correspondiente al 70 % del total de muestras, conjunto de validación, equivalente al 10 % de la base de datos y conjunto de prueba, correspondiente al 20 % restante.

En el anexo 6.1, tablas 6.1 y 6.3, se encuentran las variables clínicas de GDM y LGA, junto con su valor promedio, rango intercuartil, mínimo y máximo para variables no binarias y en porcentaje para variables binarias, además se adiciona la fecha de adquisición, expresada en semanas de gestación (GW). No todas las variables son utilizadas para la predicción, varias de ellas aparecen con fines estadísticos y referenciales. En la figura 6.1 aparecen las distribuciones de las principales variables continuas de ambas condiciones.

3.2. Modelos predictivos e hiperparámetros

Doce modelos son implementados y sus hiperparámetros optimizados para alcanzar el mayor rendimiento de predicción, estos modelos son los de Gaussian Naïve Bayes (GNB), Bernoulli Naïve Bayes (BNB), Árbol de Decisión (DT), *Support Vector Machine* (SVM), Perceptrón Multi-Capas (MLP), K Vecinos más Cercanos (KNN), Regresión Logística (LR), *Random Forest* (RF), *Extra Trees* (ET), implementados usando la librería *Scikit-Learn* [61], *Balanced Random Forest* (BRF), utilizando la librería *Imbalanced-Learn* [99] y *Gradient Boosting Machine*, en sus implementaciones de *Extreme Gradient Boosting* (XGB) [80] y *Light Gradient Boosting Machine* (LGBM) [81], utilizando el lenguaje de programación *Python*.

Estos modelos son seleccionados debido a que son utilizados por varios modelos del estado del arte, obteniéndose los mejores resultados en el estado del arte en tareas de clasificación. Principalmente Random Forest, Balanced Random Forest, Extra-Trees, Gradient Boosting Machine (XGB y LGBM), Regresión Logística y MLP. Los cinco modelos restantes, que pueden ser considerados modelos más simples o populares en su tiempo, pero efectivos en ciertas tareas, son utilizados de igual forma ya que pueden tener mejor desempeño que los otros modelos, ya que modelos más complejos no aseguran un mejor rendimiento. A priori, no existe un modelo superior al resto. Cabe destacar que otros trabajos del estado del arte no analizan muchos modelos diferentes, presentando en general solo un modelo, solo algunos presentan a lo más de 5 modelos, por lo que esto también es un elemento diferenciador de nuestro trabajo.

Se analizan sobre 3000 combinaciones de hiperparámetros, como por ejemplo, el tipo de kernel de SVM, el número de capas y neuronas de MLP o la cantidad de estimadores de los modelos tipo *Ensemble*. El listado de variables utilizado, con los nombres dados por las librerías, los valores y que modelos usan cada variable están en el anexo 6.1, tabla 6.6.

3.3. Aumentación de Datos

Con el fin de aumentar el desempeño de los modelos, se genera un método de Aumentación de Datos (DA) innovador dedicado para estas condiciones/enfermedades y específico para ciertas variables. A diferencia de otros métodos, el propuesto es limitado solo a variables que su alteración posea sentido y que los valores continuos sean realistas, como, por ejemplo, edad o peso, y no variables como número de partos, ya que no es factible tener 1.4 partos, pero si tener 24.2 años.

La creación de nuevos registros, o pacientes, es a partir de los datos de un paciente original. Los nuevos pacientes deben tener un valor en las variables alteradas dentro de un cierto rango de distancia respecto al paciente original para ser creados. Estos rangos son provistos por un especialista en Obstetricia/Ginecología. Las variables a modificar son, edad materna, altura, peso y BMI, para GDM, el peso y BMI de la primera visita, para LGA, de ambas visitas, los test de glicemia, para GDM solo el test de glicemia en ayuna (1TFG), mientras que para LGA, también fueron incluidos los OGTT, ayuna y 2 horas postprandial, y la diferencia de peso entre ambas visitas, aunque solo para LGA.

Los rangos propuestos por el especialista son:

- Edad: Pacientes creados deben estar dentro del rango de ± 2 años respecto al original.
- Tests de glicemia: ± 5 mg/dL respecto a lo obtenido por el paciente original.
- Altura: ± 3 cm de diferencia máxima entre los nuevos pacientes y el original.
- Peso: El peso del nuevo paciente debe estar dentro de un rango de diferencia de ± 5 kg en comparación a las mediciones del paciente original.

Al ser BMI y la diferencia de pesos, variables dependientes de otras mediante fórmulas, estas no son modificadas directamente, sino que se adaptan a la variación de las otras, es decir, son recalculadas a los cambios de las otras variables.

También se implementan restricciones adicionales para añadirle más "realismo" a los datos creados. Al BMI del paciente creado debe estar en la misma categoría que el paciente original, la categoría asignada a cada paciente es calculada en base a las recomendaciones propuestas por la WHO [100], si el nuevo paciente pertenece a una categoría distinta, este no es creado. Para 1TFG, solo son modificados sus valores si el valor original del paciente se encuentra entre 66 y 94 mg/dL o sobre 105 mg/dL, esto por recomendaciones del especialista, mientras que los OGTT solo poseen la restricción de que los nuevos valores no deben cambiar por si solos el diagnóstico de GDM.

Adicional a estos rangos, se realizan experimentos con rangos más limitados que los propuestos por el especialista, los cuales se denominan rangos limitados, mientras que los rangos propuestos por el especialista serán designados con el nombre de rangos originales. El objetivo de los rangos limitados es ofrecer una versión con un menor grado de alteración que los rangos originales. Ambos rangos son mostrados en la tabla 3.1.

Tabla 3.1: Rangos de Aumentación de Datos propuestos.

Rango DA	Edad (Años)	Altura (cm)	Variables de Peso (kg)	BMI	Diferencia de Peso	Exámenes de Glucosa
Rango Original	±2	±3	±5	*	**	±5
Rango Limitado	±1	±1	±2	*	**	±1

*: Recalculado en base a la fórmula de BMI según la modificación de Peso y Altura.
**: Recalculado en base a la diferencia entre los pesos alterados.

Los valores añadidos a cada paciente a los nuevos pacientes son valores aleatorios dentro del rango especificado, e.g., de ±5 % del valor original, y limitados en base a los rangos propuestos por el especialista.

3.4. Entrenamiento y Evaluación de modelos

Con el fin de obtener el mejor rendimiento, se realiza una búsqueda de grilla con las múltiples variaciones de hiperparámetros, es decir, se busca las combinaciones de hiperparámetros que permitan maximizar el desempeño de los modelos. Para esta labor se ejecuta una Validación Cruzada de 5 hojas sobre el conjunto de entrenamiento para cada combinación de hiperparámetros de los diferentes modelos, lo cual consiste en fraccionar el conjunto de entrenamiento en 5 partes, donde 4 partes actúan como conjunto de entrenamiento y la parte restante como conjunto de prueba, este procedimiento se efectúa 5 veces, variando en cada oportunidad el conjunto de prueba seleccionado.

Otro componente que juega un rol clave es la entrada (variables) utilizada por los modelos. Para esto se realiza una selección de variables, usando cuatro métodos, F-test de ANOVA, Test Chi cuadrado, Información Mutua y las variables con mayor relevancia según el modelo de Balanced Random Forest, especificados en la sección de Selección de Variables. Esta selección de variables permite reconocer cuales son las variables que poseen un papel más determinante en la predicción de las condiciones a predecir. Por lo que es posible disminuir el número de variables requeridas por los modelos sin sacrificar rendimiento, inclusive aumentándolo si es que existen variables redundantes o no útiles.

Los métodos de F-test de ANOVA, Test Chi cuadrado e información mutua, fueron utilizadas sus implementaciones de Scikit-Learn [61], versión 1.2.2. En el caso del método de información mutua, este utiliza una función basada en los métodos no paramétricos de estimación de entropía a partir de las distancias calculadas desde los K vecinos más cercanos, planteados en [101, 102].

Las variables elegidas por estos métodos son evaluadas en el mismo proceso de búsqueda de grilla, como si fueran un hiperparámetro más. En particular para GDM, la selección de variables es limitada a solo 12 variables para obtener modelos que requiera de un número bajo de entradas, tampoco se analiza el desempeño de los modelos con las variables seleccionadas por BRF, debido a su alta similitud con las variables seleccionadas por los otros métodos. Para LGA, el número de variables seleccionadas para su evaluación son múltiples de 5 (ej.:

5, 10, 15, etc.) de forma preliminar.

Son escogidos los mejores 15 % modelos evaluados en base a la métrica AUCROC, esta métrica es escogida debido a que permite efectuar una comparación entre modelos y no es dependiente de umbral de decisión como otras métricas. Los mejores modelos son evaluados en el conjunto de validación, no utilizado previamente, con el fin de obtener los umbrales de decisión que posteriormente se utilizarán en el conjunto de prueba para calcular las métricas de *Accuracy*, *Sensitivity*, *Specificity* y *Recall Macro*. Los umbrales de decisión son valores que un modelo utiliza para asignar un valor como positivo o negativo, en base a la probabilidad que esta entrega, por ejemplo, si la probabilidad que entrega un modelo en base a una es 0.6, con un umbral de decisión de 0.5 ese valor es positivo, sin embargo, con un umbral de 0.8, esa entrada se clasifica como negativa. Por lo tanto, influye en las métricas mencionadas.

Los modelos son evaluados en el conjunto de prueba para medir su nivel de generalización. Cabe destacar que este conjunto, ni sus datos que le componen, son utilizados para algunos de procesos de selección previos ni en el entrenamiento de los modelos. Los modelos evaluados en el conjunto de prueba son entrenados utilizando una combinación del conjunto de entrenamiento y validación, a este "nuevo" conjunto de entrenamiento se le aplica ambas opciones de DA mencionado en la sección anterior, aunque también existe la opción sin DA.

Los mejores modelos son escogidos en base a las diferentes métricas ya mencionadas, no obstante, se le entrega mayor relevancia a la métrica de Sensibilidad, debido a que se prioriza el evitar los errores de Falsos Negativos.

Todo este proceso esta representado en el diagrama de la figura 3.2. En la parte superior se observa que a la Validación Cruzada se ingresan los hiperpárametros y los resultados de la selección de variables. Esta Validación Cruzada se realiza utilizando el conjunto de entrenamiento. Los mejores modelos (ubicados en el centro del diagrama) son evaluados con el conjunto de validación para obtener los umbrales de decisión. Finalmente, los modelos son reentrenados utilizando el conjunto de entrenamiento y validación, evaluándose en el conjunto de prueba.

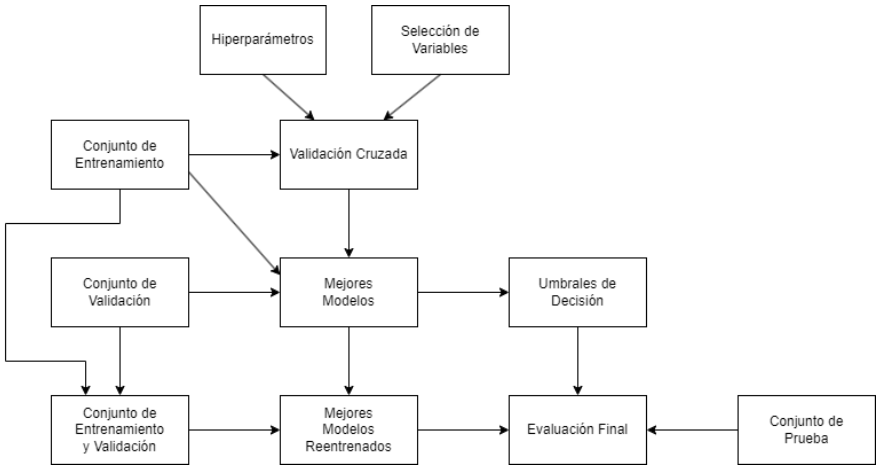

Figura 3.2: Diagrama de flujo del proceso de entrenamiento y evaluación de modelos.

4. Resultados

4.1. Selección de variables

4.1.1. GDM

Las 12 variables más relevantes seleccionadas por los 4 métodos de selección para la clasificación de GDM son mostradas en la tabla 4.1. Entre paréntesis aparecen los valores de importancia determinado por cada métrica, cada una posee una escala diferente. Solo los valores de BRF suman 1.

4.1.2. LGA

La selección de variables para LGA es realizada, inicialmente, utilizando un número múltiplo de 5, por ejemplo 5, 10 o 15 variables. Los mejores modelos, con mejor desempeño en balance entre Sensibilidad y Especificidad, (*Recall Macro*), requieren de entre 10 y 15 variables, por lo que se realiza una búsqueda adicional, entre 6 y 14 variables. Con la cual, se encuentra un nuevo mejor modelo, el cual necesita de 11 variables. Las 11 variables más relevantes para LGA son presentadas en la tabla 4.2. Entre paréntesis aparecen los valores de importancia determinado por cada métrica, cada una posee una escala diferente. Solo los valores de BRF suman 1.

4.2. Rendimiento Modelos GDM

En la tabla 4.3 se muestran los mejores 4 modelos por nivel de Sensibilidad para GDM, con un mínimo de 0.7949 en esta métrica. La tabla incluye las columnas de número de modelo, tipo de algoritmo/modelo, número de variables de entrada, el uso o no de aumentación de datos, con "No" igual a que DA no fue utilizado para este modelo, "RO" si es que se utiliza el Rango Original propuesto por el Experto y "RL" si es utilizado el Rango Limitado. También las métricas de *Accuracy*, *Sensitivity*, *Specificity*, *Recall Macro* y AUCROC, todas calculadas en el conjunto de prueba. En negrita están los modelos con mayor nivel de Especificidad para cada nivel de Sensibilidad. El conjunto de prueba consiste en 284 pacientes negativos de GDM y 39 positivos.

Tabla 4.1: Doce variables más relevantes para la predicción de GDM.

Ranking	F-Test ANOVA	Chi Cuadrado	Información Mutua	BRF
1	1TFG (176.68)	1TFG (226.28)	1TFG (0.060)	1TFG (0.204)
2	BMI (65.60)	Peso Materno (179.76)	BMI (0.048)	BMI (0.154)
3	Peso Materno (59.68)	BMI (84.12)	Edad (0.36)	Peso Materno (0.136)
4	Edad (57.80)	Edad (70.26)	Uso de Drogas Antihipertensivas (0.019)	Edad (0.127)
5	Hipertensión Crónica (24.52)	Gravidez (30.21)	Peso Materno (0.019)	Altura (0.103)
6	Gravidez (22.56)	Hipertensión Crónica (22.59)	Enfermedad Inflamatoria Intestinal (0.014)	Gravidez (0.041)
7	Uso de Drogas Antihipertensivas (19.02)	Paridad (18.90)	Uso de Drogas Ilícitas (0.013)	Paridad (0.037)
8	Paridad (16.83)	Uso de Drogas Antipertensivas (17.77)	Enfermedades Renales Crónicas (0.012)	Partos Vaginales (0.032)
9	Resistencia a la Insulina (13.56)	Abortos (14.44)	Enfermedad al Tracto Urinario (0.011)	Abortos (0.025)
10	Hipotiroidismo (13.50)	Partos Vaginales (14.01)	Resistencia a la Insulina (0.010)	Partos por Cesárea (0.020)
11	Partos Vaginales (10.13)	Resistencia a la Insulina (13.04)	Desordenes Psiquiátricos (0.010)	Hipotiroidismo (0.016)
12	Abortos (8.09)	Hipotiroidismo (12.69)	Enfermedades Cardiacas (0.010)	Hipertensión Crónica (0.011)

1TFG: Test de Glicemia Primer Trimestre. Peso Materno: Peso Primer Control.

Tabla 4.2: Once variables más relevantes para la predicción de LGA.

Ranking	F-Test ANOVA	Chi Cuadrado	Información Mutua	BRF
1	Hadlock >p95 % (339.24)	EFW (19121.88)	Hadlock >p95 % (0.079)	Hadlock >p75 % (0.079)
2	Hadlock >p90 % (271.54)	OGTT 2 Horas (488.25)	Hadlock >p90 % (0.079)	EFW (0.075)
3	Hadlock >p75 % (240.89)	Peso Materno Último Control (279.12)	Hadlock >p75 % (0.076)	Peso Materno Último control (0.073)
4	Peso Materno Último Control (99.15)	Hadlock >p95 % (238.15)	BMI Último Control (0.047)	BMI Último Control (0.068)
5	EFW (92.40)	Peso Materno Primer Control (237.04)	Peso Materno Último Control (0.041)	Diferencia de Peso (0.059)
6	BMI Último Control (86.16)	Hadlock >p90 % (185.40)	BMI Primer Control (0.036)	BMI Primer Control (0.057)
7	Peso Materno Primer Control (68.85)	Hadlock >p75 % (132.50)	EFW (0.033)	Peso Materno Primer Control (0.057)
8	BMI Primer Control (58.04)	BMI Último Control (81.71)	Hadlock <p25 % (0.031)	OGTT 2 Horas (0.057)
9	OGTT 2 Horas (54.56)	OGTT Ayuna (74.43)	Edad (0.022)	OGTT Ayuna (0.051)
10	OGTT Ayuna (40.13)	BMI Primer Control (69.50)	Desordenes Ginecológicos (0.021)	Altura (0.050)
11	Hadlock <p25 % (35.63)	Tratamiento Diabetes Mellitus (53.73)	Polihidramnios (0.021)	Hadlock >p95 % (0.049)

EFW: Peso Fetal Estimado, Ultrasonido Tercer Trimestre.

Tabla 4.3: Tabla de Resultados de GDM.

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
1	**MLP**	**12**	**No**	**0.3994**	**1**	**0.3169**	**0.6585**	**0.8189**
2	MLP	10	RO	0.3715	1	0.2852	0.6426	0.7741
3	MLP	11	RL	0.3715	1	0.2852	0.6426	0.7890
4	MLP	11	RL	0.3653	1	0.2782	0.6391	0.7874
5	**MLP**	**8**	**RL**	**0.5511**	**0.9744**	**0.4930**	**0.7337**	**0.8002**
6	SVM	5	RL	0.5480	0.9744	0.4894	0.7319	0.8161
7	SVM	5	RL	0.5480	0.9744	0.4894	0.7319	0.8161
8	MLP	4	RO	0.5387	0.9744	0.4789	0.7266	0.8052
9	**SVM**	**5**	**RO**	**0.6068**	**0.9487**	**0.5599**	**0.7543**	**0.8234**
10	MLP	4	RO	0.5759	0.9487	0.5246	0.7367	0.8159
11	MLP	3	No	0.5728	0.9487	0.5211	0.7349	0.8165
12	MLP	4	RL	0.5728	0.9487	0.5211	0.7349	0.8082
13	**SVM**	**5**	**RO**	**0.6130**	**0.9231**	**0.5704**	**0.7468**	**0.8234**
14	MLP	6	No	0.6006	0.9231	0.5563	0.7397	0.8221
15	MLP	8	RO	0.6006	0.9231	0.5563	0.7397	0.8183
16	LR	3	RO	0.6006	0.9231	0.5563	0.7397	0.8159
17	**MLP**	**5**	**RL**	**0.6594**	**0.8974**	**0.6268**	**0.7621**	**0.8199**
18	MLP	5	No	0.6594	0.8974	0.6268	0.7621	0.8146
19	MLP	5	RL	0.6563	0.8974	0.6232	0.7603	0.8178
20	MLP	7	RL	0.6563	0.8974	0.6232	0.7603	0.8118
21	**MLP**	**7**	**RL**	**0.6873**	**0.8718**	**0.6620**	**0.7669**	**0.8160**
22	MLP	10	RL	0.6811	0.8718	0.6549	0.7634	0.8078
23	MLP	9	RL	0.6780	0.8718	0.6514	0.7616	0.8137
24	MLP	9	RO	0.6749	0.8718	0.6479	0.7598	0.8137
25	**MLP**	**6**	**RL**	**0.7090**	**0.8462**	**0.6901**	**0.7681**	**0.8142**
26	MLP	9	RO	0.7090	0.8462	0.6901	0.7681	0.8022
27	MLP	10	No	0.7028	0.8462	0.6831	0.7646	0.8063
28	MLP	9	RO	0.7028	0.8462	0.6831	0.7646	0.8022
29	**SVM**	**12**	**No**	**0.7554**	**0.8205**	**0.7465**	**0.7835**	**0.8135**
30	SVM	12	No	0.7461	0.8205	0.7359	0.7782	0.8135
31	SVM	7	RL	0.7399	0.8205	0.7289	0.7747	0.8143
32	SVM	7	RL	0.7368	0.8205	0.7254	0.7729	0.8143
33	**SVM**	**7**	**RL**	**0.7399**	**0.7949**	**0.7324**	**0.7636**	**0.8143**
34	SVM	10	RL	0.7337	0.7949	0.7254	0.7601	0.8173
35	MLP	5	RO	0.7276	0.7949	0.7183	0.7566	0.8120
36	MLP	9	RO	0.7245	0.7949	0.7148	0.7548	0.8068

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

En la tabla 4.3 son mostrados los modelos número 1 al 16 poseen una sensibilidad superior a 0.9231, mientras que los modelos número 17 al 36, sobre 0.7949, pero inferiores a 0.9231

de sensibilidad. Como se mencionó previamente, se le da un alto énfasis en modelos con alta sensibilidad para poder minimizar los Falsos Negativos, ya que la detección de GDM, y su eventual respectivo tratamiento, puede prevenir las serias consecuencias de GDM en la madre y el bebé que pueden ocurrir incluso varios años posteriores al embarazo.

La tabla 4.3 ofrece distintos modelos con distintos valores de sensibilidad y especificidad, esto con el fin de que se pueda escoger un rango de detección deseado. Por ejemplo, se podría escoger el modelo número 17, con una sensibilidad 0.8974 que requiere de solo 5 variables (1TFG, edad, BMI, Peso Materno y Gravidez), con un *Accuracy* de 0.6594, especificidad de 0.6268, *Recall Macro* de 0.7621 y AUCROC de 0.8234. También se podría escoger el modelo número 29, el cual es el modelo con mejor *Recall Macro* de los presentados, por lo que posee el mejor balance entre las métricas de *Sensitivity* y *Specificity*, que puede ser útil si es que un especialista médico prefiere tener un buen nivel de clasificación tanto de pacientes sin GDM como con GDM.

La figura 4.1 muestra dos ángulos distintos de la misma superficie, que corresponde a los diferentes modelos disponibles, para distintos niveles de errores, la suma de Falsos Positivos y Negativos, Verdaderos positivos y número de variables. Esta figura muestra la posibilidad de escoger diversos modelos, con posibilidades como alta sensibilidad (bajo FN) o alta especificidad (bajo FP) e ir variando la complejidad del modelo en base al número de variables. En esta superficie, los puntos rojos representan a los mejores modelos destacados en negrita de la tabla 4.3 con sensibilidad superior a 0.92 (modelos número 1, 5, 9 y 13), y los puntos amarillos representan los mejores modelos destacados en negrita en la misma tabla, pero con una sensibilidad inferior a 0.92, pero mayor a 0.79 (modelos número 17, 21, 25, 29 y 33).

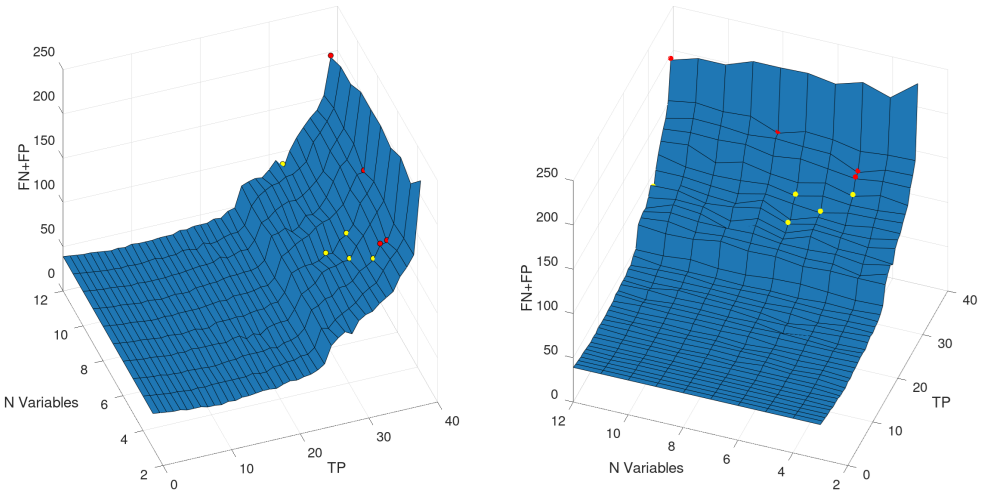

Figura 4.1: Superficie con todos los modelos disponibles de GDM, incluyendo varios valores de hiperparámetros, para varios niveles de error (FP+FN), Verdaderos Positivos (TP) y número de variables. Los puntos rojos representan los mejores modelos con sensibilidad sobre 0.9231 y los amarillos los mejores modelos con sensibilidad menor a 0.9231, pero superior a 0.7949, en negrita en la tabla 4.3.

La figura 4.2 muestra las curvas ROC para cada uno de los 9 mejores modelos para cada nivel de sensibilidad, iniciando con sensibilidad 1 (a), hasta sensibilidad de 0.79 (d). Los mejores modelos son los marcados con negrita en la tabla 4.3. (a) muestra las curvas ROC

de los mejores modelos con sensibilidad 1 (MLP 12 Variables), 0.9744 (MLP 8 Variables) y 0.9487 (SVM 5 Variables). (b) las curvas ROC de los mejores modelos con sensibilidad 0.9231 (SVM 5 Variables), 0.8974 (MLP 5 Variables) y 0.8718 (MLP 7 Variables) y (c) las curvas ROC de los mejores modelos con sensibilidad 0.8462 (MLP 6 Variables), 0.8205 (SVM 12 Variables), 0.7949 (SVM 7 Variables). La figura 4.2 (d) muestra el modelo número 29 de la tabla 4.3, que posee el mejor *Recall Macro* (gris, SVM 12 Variables), en comparación con el mismo modelo aplicando DA (cían) y el modelo número 33 (rosa, SVM 7 Variables), que posee una sensibilidad similar, pero requiriendo de un número inferior de variables.

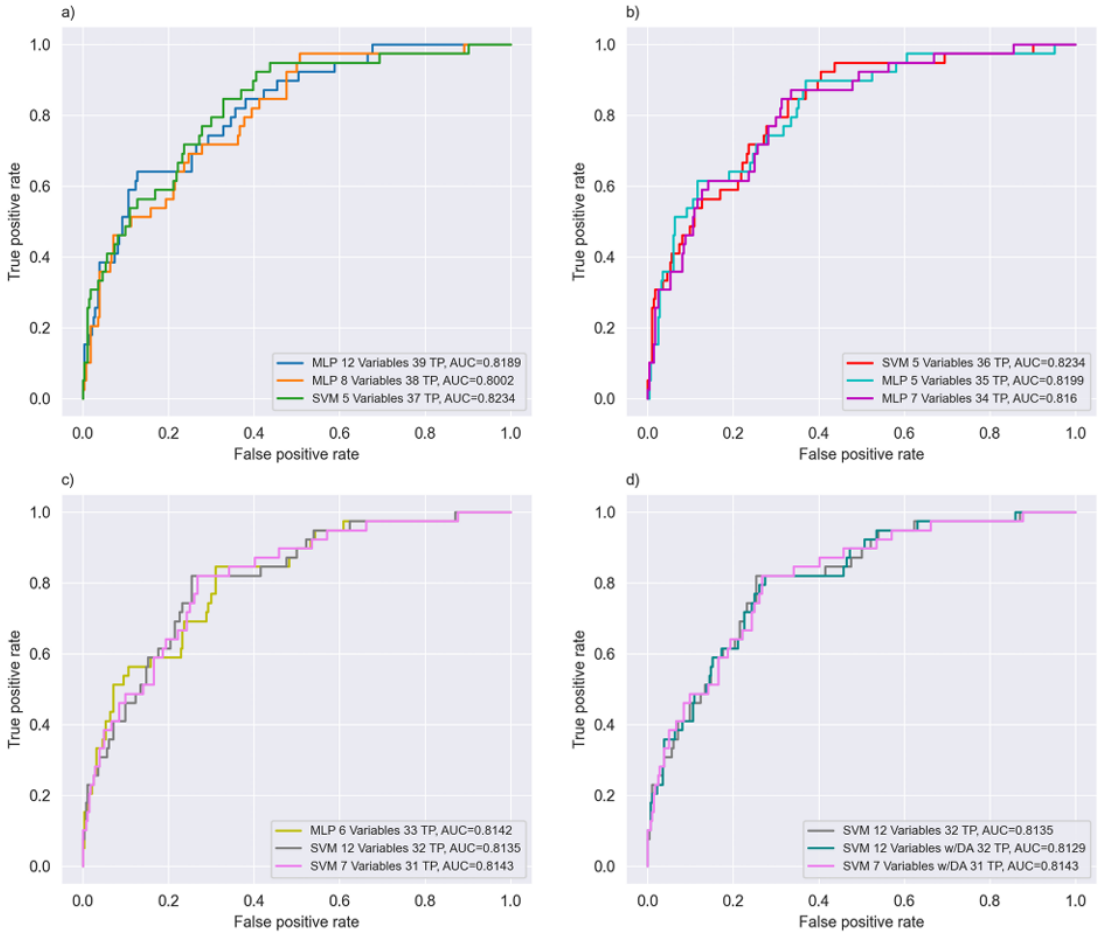

Figura 4.2: Curvas ROC de los mejores modelos de GDM. a) Mejores modelos con sensibilidad 1 , 0.9744 y 0.9487. b) Mejores modelos con sensibilidad 0.9231, 0.8974 y 0.8718. c) Mejores modelos con sensibilidad 0.8462, 0.8205, 0.7949. d) Curva ROC modelo número 29 (gris), tabla 4.3, en comparación con el mismo modelo con DA (cían) y modelo número 33 (rosa).

En la tabla 4.4 muestra los mejores modelos para los diferentes niveles de sensibilidad analizados, pero con más de 12 variables de entradas. Es posible observar que los modelos 38, 42, 43 y 45 alcanzan un mejor desempeño en comparación con los modelos con misma sensibilidad presentados en la tabla 4.3. No obstante, el número de variables requerida es más del doble, por ejemplo, el modelo 25 requiere de 6 variables de entradas, mientras que el modelo 43 requiere de 15, ambas con el mismo nivel de sensibilidad, aumentando la especificidad en solo 2.01 %.

Tabla 4.4: Tabla de Resultados de los mejores modelos de GDM, con un número de variables mayor a 12.

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
37	MLP	15	RL	0.3003	1	0.2042	0.6021	0.8210
38	SVM	15	No	0.5697	0.9744	0.5141	0.7442	0.7872
39	MLP	13	RL	0.5820	0.9487	0.5317	0.7402	0.8093
40	SVM	15	No	0.6099	0.9231	0.5669	0.7450	0.7872
41	MLP	13	No	0.6409	0.8974	0.6056	0.7515	0.8152
42	MLP	14	RL	0.7059	0.8718	0.6831	0.7774	0.7968
43	MLP	15	RL	0.7212	0.8462	0.7042	0.7752	0.7988
44	SVM	15	RO	0.7337	0.8205	0.7218	0.7712	0.8125
45	SVM	15	RO	0.7461	0.7949	0.7394	0.7672	0.8125

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

4.2.1. DA versus no DA

En la tabla 4.5 se muestra un comparación de rendimiento entre los mejores modelos de la tabla 4.3 con Aumentación de Datos y sin esta técnica, para los modelos que no usan DA, se aplicó de igual forma esta técnica a modo de comparación. La comparación se realiza con el mismo nivel de sensibilidad.

4.2.2. Modelos optimizados versus modelos sin optimizar

En la tabla 4.6 aparece una comparación entre el desempeño de los mejores modelos de la tabla 4.3 versus los mejores modelos sin optimizar. Para estos últimos se consideraron el caso de usar las 5 variables utilizadas por los modelos 9, 13 y 17, las 7 variables utilizadas por el modelo 33, las 12 variables utilizadas por los modelos 1 y 29 y todas las variables disponibles (30). Este último caso no tendría optimización alguna.

En la figura 4.3 aparece un gráfico que muestra los errores (FP+FN) en función de los verdaderos positivos para los mejores modelos de la tabla 4.3 (azul), los mejores modelos sin optimizar utilizando todas las variables (rojo), los mejores modelos utilizando las 5 variables utilizadas por los modelos 9, 13 y 17 (verde), los mejores modelos utilizando las 7 variables del modelo 33 (naranja) y las 12 variables de los modelo 1 y 29 (morado). En celeste aparece los mejores modelos sin optimizar de cualquiera de las 4 opciones previas.

Tabla 4.5: Tabla de comparación de rendimiento entre modelos GDM con Aumentación de Datos versus sin DA.

NO	Tipo de Modelo	Número de Variables	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
1 DA RO	MLP	12	0.3313	1	0.2394	0.6197	0.7505
1 No DA	MLP	12	0.3994	1	0.3169	0.6585	0.8189
5 DA RL	MLP	8	0.5511	0.9744	0.4930	0.7337	0.8002
5 No DA	MLP	8	0.4303	0.9744	0.3556	0.6650	0.8172
9 DA RO	SVM	5	0.6068	0.9487	0.5599	0.7543	0.8234
9 No DA	SVM	5	0.4396	0.9487	0.3697	0.6592	0.8221
13 DA RO	SVM	5	0.6130	0.9231	0.5704	0.7468	0.8234
13 No DA	SVM	5	0.5913	0.9231	0.5458	0.7344	0.8221
17 DA RL	MLP	5	0.6594	0.8974	0.6268	0.7621	0.8199
17 No DA	MLP	5	0.5944	0.8974	0.5528	0.7251	0.8202
21 DA RL	MLP	7	0.6873	0.8718	0.6620	0.7669	0.8160
21 No DA*	MLP	7	0.4985	0.9487	0.4367	0.6927	0.7905
21 No DA*	MLP	7	0.5635	0.8462	0.5246	0.6854	0.7905
25 DA RL	MLP	6	0.7090	0.8462	0.6901	0.7681	0.8142
25 No DA	MLP	6	0.6099	0.8462	0.5775	0.7118	0.8156
29 DA RL	SVM	12	0.7368	0.8205	0.7254	0.7729	0.8129
29 No DA	SVM	12	0.7554	0.8205	0.7465	0.7835	0.8135
33 DA RL	SVM	7	0.7399	0.7949	0.7324	0.7636	0.8143
33 No DA*	SVM	7	0.5635	0.8205	0.5282	0.6743	0.7852
33 No DA*	SVM	7	0.6161	0.7692	0.5954	0.6822	0.7852

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

*: Obtenido con el valor de sensibilidad más cercano, en base a los umbrales de decisión obtenidos con el conjunto de validación.

Tabla 4.6: Tabla de mejores resultados GDM versus modelos sin optimizar.

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
1	**MLP**	**12**	**No**	**0.3994**	**1**	**0.3169**	**0.6585**	**0.8189**
-	LGBM	5	-	0.2539	1	0.1514	0.5757	0.7742
-	MLP	7	-	0.2632	1	0.1620	0.5810	0.7820
-	BRF	12	-	0.2941	1	0.1972	0.5986	0.7587
-	LGBM	32	-	0.3065	1	0.2113	0.6056	0.7701
5	**MLP**	**8**	**RL**	**0.5511**	**0.9744**	**0.4930**	**0.7337**	**0.8002**
-	LR	5	-	0.4675	0.9744	0.3979	0.6861	0.8157
-	LR	7	-	0.4675	0.9744	0.4155	0.6949	0.8126
-	LR	12	-	0.4861	0.9744	0.4190	0.6967	0.8046
-	LR	32	-	0.3313	0.9744	0.2430	0.6087	0.7962
9	**SVM**	**5**	**RO**	**0.6068**	**0.9487**	**0.5599**	**0.7543**	**0.8234**
-	LR	5	-	0.5015	0.9487	0.4401	0.6944	0.8157
-	LR	7	-	0.5015	0.9487	0.4401	0.6944	0.8126
-	MLP	12	-	0.3746	0.9487	0.2958	0.6222	0.7657
-	LR	32	-	0.4923	0.9487	0.4296	0.6891	0.7962
13	**SVM**	**5**	**RO**	**0.6130**	**0.9231**	**0.5704**	**0.7468**	**0.8234**
-	MLP	5	-	0.4520	0.9231	0.3873	0.6552	0.7963
-	MLP	7	-	0.4551	0.9231	0.3908	0.6570	0.7820
-	LR	12	-	0.5325	0.9231	0.4789	0.7010	0.8046
-	LR	32	-	0.5139	0.9231	0.4577	0.6904	0.7962
17	**MLP**	**5**	**RL**	**0.6594**	**0.8974**	**0.6268**	**0.7621**	**0.8199**
-	LR	5	-	0.5789	0.8974	0.5352	0.7163	0.8157
-	LR	7	-	0.5604	0.8974	0.5141	0.7058	0.8126
-	GNB	12	-	0.4799	0.8974	0.4225	0.6600	0.7282
-	LGBM	32	-	0.5263	0.8974	0.4754	0.6864	0.7701
21	**MLP**	**7**	**RL**	**0.6873**	**0.8718**	**0.6620**	**0.7669**	**0.8160**
-	LR	5	-	0.5913	0.8718	0.5528	0.7123	0.8157
-	GNB	7	-	0.5418	0.8718	0.4965	0.6841	0.7583
-	BRF	12	-	0.4675	0.8718	0.4120	0.6419	0.7587
-	LR	32	-	0.6006	0.8718	0.5634	0.7176	0.7962
25	**MLP**	**6**	**RL**	**0.7090**	**0.8462**	**0.6901**	**0.7681**	**0.8142**
-	LR	5	-	0.6161	0.8462	0.5845	0.7153	0.8157
-	LR	7	-	0.6100	0.8462	0.5775	0.7118	0.8126
-	LGBM	12	-	0.5356	0.8462	0.4930	0.6696	0.7795
-	XGB	32	-	0.4582	0.8462	0.4049	0.6255	0.7226

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

Tabla 4.7: Continuación Tabla 4.6

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
29	**SVM**	**12**	**No**	**0.7554**	**0.8205**	**0.7465**	**0.7835**	**0.8135**
-	GNB	5	-	0.5604	0.8205	0.5246	0.6726	0.7820
-	XGB	7	-	0.5789	0.8205	0.5458	0.6831	0.7642
-	LGBM	12	-	0.6037	0.8205	0.5739	0.6972	0.7795
-	LGBM	32	-	0.5542	0.8205	0.5176	0.6691	0.7491
33	**SVM**	**7**	**RL**	**0.7399**	**0.7949**	**0.7324**	**0.7636**	**0.8143**
-	LGBM	5	-	0.6502	0.7949	0.6303	0.7126	0.7742
-	XGB	7	-	0.6068	0.7949	0.5810	0.6879	0.7642
-	LR	12	-	0.6409	0.7949	0.6197	0.7073	0.8046
-	GNB	32	-	0.5232	0.7949	0.4859	0.6404	0.6622

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

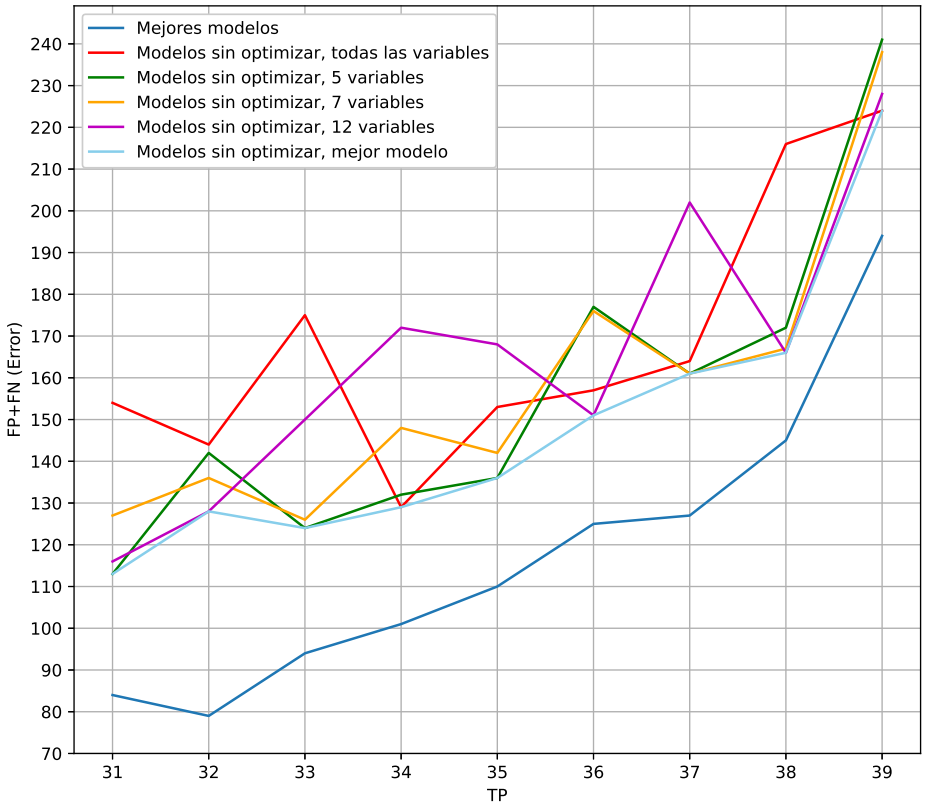

Figura 4.3: Errores (FP+FN) en función de los verdaderos positivos, en azul los mejores modelos de la tabla 4.3, los mejores modelos sin optimizar utilizando todas las variables (rojo), los mejores modelos utilizando las 5 variables utilizadas por los modelos 9, 13 y 17 (verde), los mejores modelos utilizando las 7 variables del modelo 33 (naranja) y las 12 variables de los modelo 1 y 29 (morado). En celeste aparece los mejores modelos sin optimizar de cualquiera de las 4 opciones previas de la tabla 4.6.

4.3. Rendimiento Modelos LGA

En la tabla 4.8 aparece el rendimiento de utilizar solo la variable de Hadlock >p90 %, la cual se utiliza actualmente como predictora de LGA, en el conjunto de prueba. Esta información se usa como referencia para posteriormente comparar el desempeño de los modelos propuesto con la alternativa que existe actualmente. El conjunto de prueba de LGA posee 306 pacientes con fetos sin LGA y 55 pacientes con fetos con LGA.

Tabla 4.8: Rendimiento de Hadlock >p90 %.

Variable	Accuracy	Sensitivity	Specificity	Recall Macro	AUCROC
Hadlock >p90 %	0.8421	0.5818	0.8889	0.7354	0.7354

En la tabla 4.9 se muestran los mejores modelos por nivel de Sensibilidad para LGA. La tabla incluye las columnas de número de modelo, tipo de algoritmo/modelo, número de variables de entrada, el uso o no de aumentación de datos, con "No" igual a que DA no fue utilizado para este modelo, "RO" si es que se utiliza el Rango Original propuesto por el Experto y "RL" si es utilizado el Rango Limitado, así como las métricas de *Accuracy*, *Sensitivity*, *Specificity*, *Recall Macro* y AUCROC, todas calculadas en el conjunto de prueba.

Tabla 4.9: Tabla de Resultados de LGA.

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
1	MLP	25	RL	0.5152	1	0.4281	0.7141	0.8652
2	MLP	30	RL	0.5263	0.9818	0.4444	0.7131	0.8305
3	MLP	15	RL	0.5900	0.9636	0.5229	0.7433	0.8125
4	MLP	15	RO	0.6482	0.9455	0.5948	0.7701	0.8482
5	MLP	15	RL	0.6787	0.9273	0.6340	0.7806	0.8224
6	MLP	15	RL	0.6981	0.9091	0.6601	0.7846	0.8404
7	MLP	10	RL	0.7396	0.8909	0.7124	0.8017	0.8397
8	**BRF**	**10**	**RO**	**0.8033**	**0.8727**	**0.7909**	**0.8318**	**0.8404**
9	BRF	10	RO	0.8061	0.8545	0.7974	0.8260	0.8404
10	**BRF**	**10**	**RO**	**0.8255**	**0.8364**	**0.8235**	**0.8299**	**0.8404**
11	**BRF**	**15**	**RO**	**0.8338**	**0.8182**	**0.8366**	**0.8274**	**0.8412**
12	**BRF**	**15**	**RO**	**0.8393**	**0.8000**	**0.8464**	**0.8232**	**0.8409**
13	**BRF**	**15**	**RO**	**0.8504**	**0.7455**	**0.8693**	**0.8074**	**0.8351**
14	MLP	15	No	0.8587	0.7091	0.8856	0.7974	0.8461
15	MLP	15	No	0.8864	0.6545	0.9281	0.7913	0.8419
16	MLP	15	RO	0.8920	0.5818	0.9477	0.7648	0.8389

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

Los mejores modelos, basados en la métrica de *Recall Macro*, son los modelos número 8, 10, 11, 12 y 13, en negrita en la tabla 4.9. Los modelos 8 y 10 son iguales, ambos son

BRF con los mismo hiperparámetros y variables de entrada, solo difieren en el umbral de decisión seleccionado. Debido a que estos mejores modelos utilizan entre 10 y 15 variables, pero el primer análisis de variables solo fue para un número de variables múltiplo de 5, se realiza un segundo análisis, entre 6 y 14 variables con el fin de encontrar un modelo con mejor desempeño. Los resultados de esta búsqueda aparecen en la tabla 4.10 y de forma análoga a la tabla 4.9, los mejores modelos están destacados con negrita.

Tabla 4.10: Tabla de Resultados de LGA de modelos con 6 a 14 variables.

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
17	MLP	8	RL	0.5014	1	0.4118	0.7059	0.8419
18	BRF	8	RO	0.5319	0.9818	0.4510	0.7164	0.8325
19	MLP	9	RO	0.5817	0.9636	0.5131	0.7384	0.8172
20	MLP	14	RO	0.6565	0.9455	0.6046	0.7750	0.8277
21	MLP	14	RO	0.6787	0.9273	0.6340	0.7806	0.8277
22	BRF	11	RO	0.6898	0.9091	0.6503	0.7797	0.8308
23	BRF	11	RO	0.7729	0.8909	0.7516	0.8213	0.8373
24	**BRF**	**11**	**RO**	**0.8144**	**0.8727**	**0.8039**	**0.8383**	**0.8423**
25	**BRF**	**11**	**RO**	**0.8199**	**0.8545**	**0.8137**	**0.8341**	**0.8423**
26	**BRF**	**14**	**RO**	**0.8366**	**0.8364**	**0.8366**	**0.8365**	**0.8321**
27	**BRF**	**11**	**RO**	**0.8449**	**0.8182**	**0.8497**	**0.8339**	**0.8450**
28	BRF	7	RO	0.8393	0.8000	0.8464	0.8232	0.8297
29	BRF	8	RL	0.8504	0.7455	0.8693	0.8074	0.8403
30	LR	14	RO	0.8670	0.7091	0.8954	0.8023	0.8294
31	MLP	12	RO	0.8837	0.6545	0.9248	0.7897	0.8361
32	MLP	11	RO	0.8892	0.5818	0.9444	0.7631	0.8346

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

De igual forma que para GDM, aunque las tablas 4.9 y 4.10 muestran modelos con una alta sensibilidad, pero con varios modelos con alta especificidad añadidos, existen múltiples alternativas de modelos y es posible escoger diferentes niveles de predicción en base a las métricas de sensibilidad y especificidad, según los requerimientos que posea un especialista de la salud. En la figura 4.4 es presentada la superficie, con dos puntos de vista distintos, correspondiente a los diferentes modelos disponibles para LGA, para distintos niveles de error (FP+FN), Verdaderos Positivos y número de variables, el número de variables es resultado del primer análisis de variable, es decir, solo considera múltiplos de 5. Además el gráfico muestra hasta solo 30 variables, esto debido a que una mayor cantidad no presenta un mejor desempeño y solo complejiza la adquisición de datos. Los puntos rojos representan los mejores modelos destacados en la tabla 4.9, modelos número 8, 10, 11, 12 y 13.

La figura 4.5, de igual manera que la figura 4.4, muestra la superficie de modelos generados para LGA, pero para el segundo análisis de variables, entre 6 y 14 variables de entrada. Los puntos amarillos en la figura simbolizan los mejores modelos en negrita de la tabla 4.10, de los modelos número 24, 25, 26 y 27.

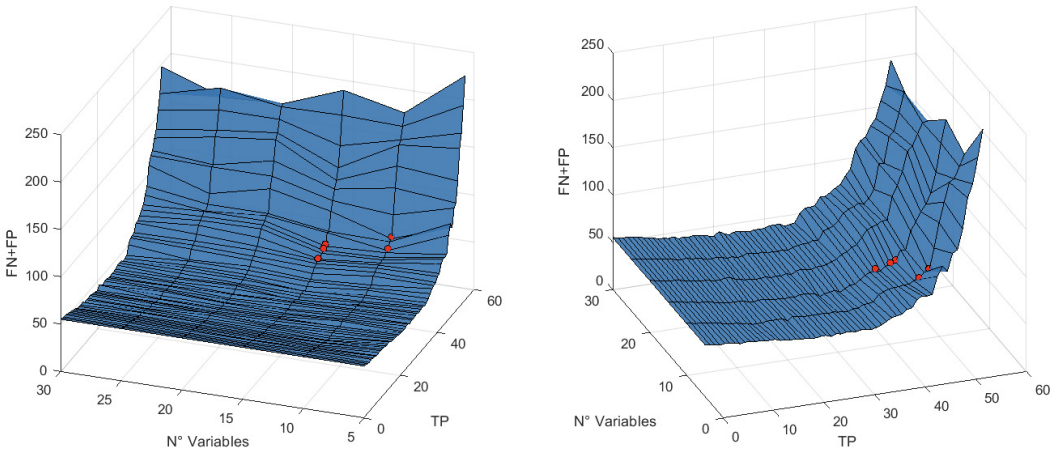

Figura 4.4: Superficie de todos los modelos de LGA, considerando un número de variables múltiple de 5. Los puntos rojos representan los mejores modelos destacados en la tabla 4.9 (modelos número 8, 10, 11, 12 y 13).

Adicionalmente, se desarrollaron modelos en los cuales las variables de percentil de Hadlock fueron reemplazados por la fecha del ultrasonido, en semanas gestacionales, la cual, junto con EFW, deberian poder formar una curva personalizada para la población chilena y ver, por un lado, la utilidad de las variables Hadlock. Por otro, si es posible reemplazar esta curva por una versión "a la medida" de la población utilizada. De esto, el modelo más destacable es un modelo con 10 variables que usa aumentación de datos del tipo RO, el cual alcanza un *Accuracy* del 0.9086 (el mayor de todos los modelos), con una sensibilidad de 0.5818 y una especificidad de 0.9673, con un *Recall Macro* y AUCROC, de 0.7746 y 0.8649, respectivamente. Este modelo es catalogado como el modelo 33.

La figura 4.6 muestra las curvas ROC de todos los modelos presentados en la tabla 4.9, subfiguras a), b) y c), y los que aparecen en la tabla 4.10, subfiguras d), e) y f). La subfigura a) muestra las curvas ROC con sensibilidad 1, 0.9818, 0.9636, 0.9455, 0.9273 de la tabla 4.9. b) las curvas ROC de los modelos con sensibilidad 0.9091, 0.8909, 0.8727, 0.8545, 0.8364, 0.8182 y 0.8000 de la tabla 4.9. c) presenta los modelos con sensibilidad 0.7455, 0.7091, 0.6545 y 0.5818 de la tabla tabla 4.9. La subfigura d) muestra las curvas ROC de los modelos con sensibilidad 1, 0.9818, 0.9636, 0.9455, 0.9273 y 0.9091 de la tabla 4.10. e) los modelos con sensibilidad 0.8909, 0.8727, 0.8545, 0.8364, 0.8182 y 0.8000 de la tabla 4.10. f) presenta las curvas ROC de los modelos con sensibilidad 0.7455, 0.7091, 0.6545 y 0.5818 presentes en la tabla tabla 4.10 y el modelo 33 (MLP 10 variables).

Cabe destacar que los modelos 8, 9 y 10 son idénticos, BRF, mismos hiperparámetros, variables y tipo de DA, difiriendo solamente en los umbrales de decisión, motivo por el que poseen misma AUCROC, pero distintos valores en el resto de las métricas. Lo mismo sucede con los modelos 20 y 21, MLPs con 14 variables, y los modelos 24 y 25, BRF 11 variables.

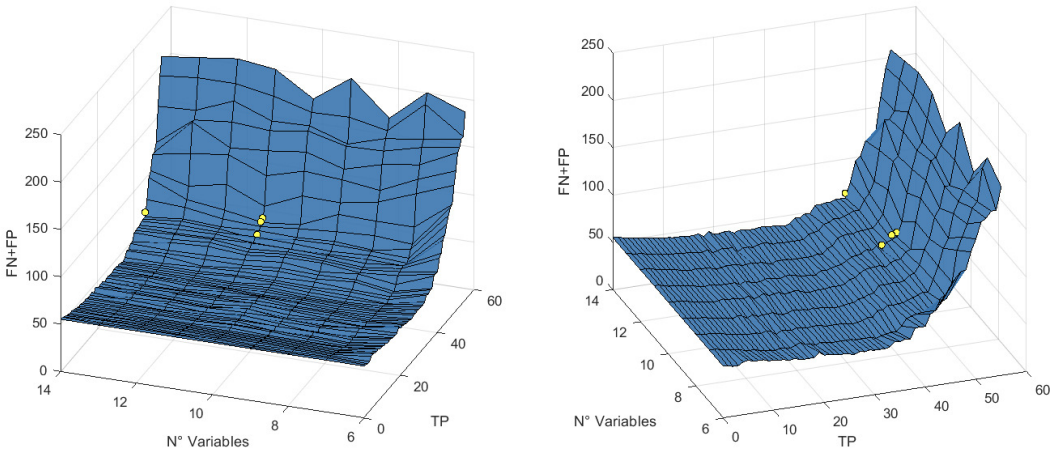

Figura 4.5: Superficie de todos los modelos de LGA, con número de variables entre 6 y 14. Los puntos amarillos representan los mejores modelos destacados en la tabla 4.10 (modelos número 24, 25, 26 y 27).

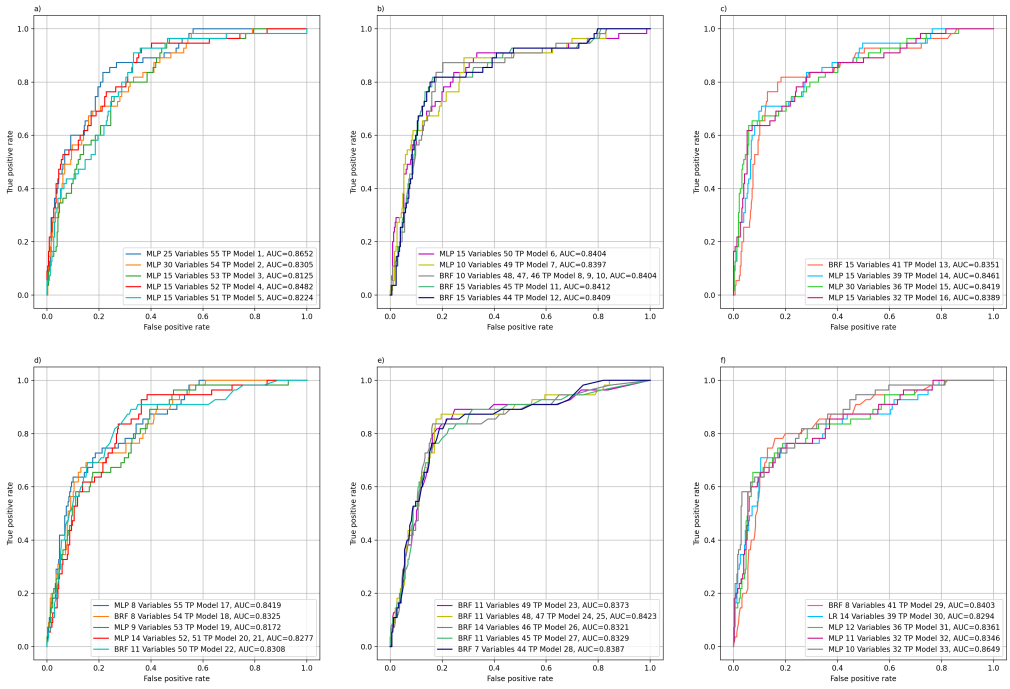

Figura 4.6: Curvas ROC de los mejores modelos de LGA. a) Curvas ROC de los modelos presentados en tabla 4.9 con sensibilidades de 1 (MLP 25 variables), 0.9818 (MLP 30 Variables), 0.9636 (MLP 15 Variables), 0.9455 (MLP 15 Variables) y 0.9273 (MLP 15 Variables). b) Curvas ROC de los modelos presentados en tabla 4.9 con sensibilidades de 0.9091 (MLP 15 Variables), 0.8909 (MLP 10 Variables) y 0.8727, 0.8545 y 0.8364 (BRF 10 Variables), 0.8182 (BRF 15 Variables) y 0.8000 (BRF 15 Variables). c) Curvas ROC de los modelos presentados en tabla 4.9 con sensibilidades de 0.7455 (BRF 15 Variables), 0.7091 (MLP 15 Variables), 0.6545 (MLP 30 Variables) y 0.5818 (MLP 15 Variables). d) Curvas ROC de los modelos presentados en la tabla 4.10 con sensibilidades de 1 (MLP 8 variables), 0.9818 (BRF 8 Variables), 0.9636 (MLP 9 Variables), 0.9455 y 0.9273 (MLP 14 Variables) y 0.9091 (BRF 11 Variables).(e) Curvas ROC de los modelos presentados en la tabla 4.10 con sensibilidades de 0.8909 (BRF 11 Variables), 0.8727 y 0.8545 (BRF 11 Variables), 0.8364 (BRF 14 Variables), 0.8182 (BRF 11 Variables), 0.8000 (BRF 7 Variables). (f) Curvas ROC de los modelos presentados en la tabla 4.10 con sensibilidades de 0.7455 (BRF 8 Variables), 0.7091 (LR 14 Variables), 0.6545 (MLP 12 Variables) y 0.5818 (MLP 11 Variables) y el modelo 33 (MLP 10 variables).

4.3.1. DA versus no DA

En la tabla 4.11 se muestra un comparación de rendimiento entre los modelos LGA de las tablas 4.9 y 4.10 con aumentación de datos y sin esta técnica, para los modelos que no usan DA. La comparación se realiza con el mismo nivel de sensibilidad.

Tabla 4.11: Tabla de comparación de rendimiento entre modelos LGA con Aumentación de Datos versus sin DA.

NO	Tipo de Modelo	Número de Variables	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
1 DA RL	MLP	25	0.5152	1	0.4281	0.7141	0.8652
1 No DA	MLP	25	0.3573	1	0.2418	0.6209	0.8372
2 DA RL	MLP	30	0.5263	0.9818	0.4444	0.7131	0.8305
2 No DA	MLP	30	0.3684	0.9818	0.2582	0.6200	0.8419
3 DA RL	MLP	15	0.5900	0.9636	0.5229	0.7433	0.8125
3 No DA	MLP	15	0.2936	0.9636	0.1732	0.5684	0.8242
4 DA RO	MLP	15	0.6482	0.9455	0.5948	0.7701	0.8482
4 No DA*	MLP	15	0.4072	0.9273	0.2778	0.6207	0.8317
5 DA RL	MLP	15	0.6787	0.9273	0.6340	0.7806	0.8224
5 No DA*	MLP	15	0.4515	0.9091	0.3693	0.6392	0.8277
6 DA RL	MLP	15	0.6981	0.9091	0.6601	0.7846	0.8404
6 No DA	MLP	15	0.6343	0.9091	0.5850	0.7470	0.8466
7 DA RL	MLP	10	0.7396	0.8909	0.7124	0.8017	0.8397
7 No DA	MLP	10	0.6898	0.8909	0.6536	0.7723	0.8458
8 DA RO	BRF	10	0.8033	0.8727	0.7909	0.8318	0.8404
8 No DA	BRF	10	0.7368	0.8727	0.7124	0.7926	0.8395
9 DA RO	BRF	10	0.8061	0.8545	0.7974	0.8260	0.8404
9 No DA*	BRF	10	0.7368	0.8727	0.7124	0.7926	0.8395
10 DA RO	BRF	10	0.8255	0.8364	0.8235	0.8299	0.8404
10 No DA*	BRF	10	0.7452	0.8000	0.7353	0.7676	0.8395
11 DA RO	BRF	15	0.8338	0.8182	0.8366	0.8274	0.8412
11 No DA	BRF	15	0.7396	0.8182	0.7255	0.7718	0.8442
12 DA RO	BRF	15	0.8393	0.8000	0.8464	0.8232	0.8409
12 No DA*	BRF	15	0.7812	0.7818	0.7810	0.7814	0.8418
13 DA RO	BRF	15	0.8504	0.7455	0.8693	0.8074	0.8351
13 No DA	BRF	15	0.8089	0.7455	0.8203	0.7829	0.8452
14 DA RO	MLP	15	0.8393	0.7091	0.8627	0.7859	0.8397
14 No DA	MLP	15	0.8587	0.7091	0.8856	0.7974	0.8461
15 DA RO	MLP	30	0.8504	0.6545	0.8856	0.7701	0.8461
15 No DA	MLP	30	0.8864	0.6545	0.9281	0.7913	0.8419

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

*: Obtenido con el valor de sensibilidad más cercano, en base a los umbrales de decisión obtenidos con el conjunto de validación.

Tabla 4.12: Continuación Tabla 4.11.

NO	Tipo de Modelo	Número de Variables	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
16 DA RO	MLP	15	0.8920	0.5818	0.9477	0.7648	0.8389
16 No DA*	MLP	15	0.8837	0.6000	0.9346	0.7673	0.8363
17 DA RL	MLP	8	0.5014	1	0.4118	0.7059	0.8419
17 No DA	MLP	8	0.4294	1	0.3268	0.6634	0.8513
18 DA RO	BRF	8	0.5319	0.9818	0.4510	0.7164	0.8325
18 No DA	BRF	8	0.3518	0.9818	0.2386	0.6101	0.8406
19 DA RO	MLP	9	0.5817	0.9636	0.5131	0.7384	0.8172
19 No DA	MLP	9	0.4321	0.9636	0.3366	0.6501	0.8452
20 DA RO	MLP	14	0.6565	0.9455	0.6046	0.7750	0.8277
20 No DA	MLP	14	0.4626	0.9455	0.3758	0.6606	0.8385
21 DA RO	MLP	14	0.6787	0.9273	0.6340	0.7806	0.8277
21 No DA	MLP	14	0.4931	0.9273	0.4150	0.6712	0.8385
22 DA RO	BRF	11	0.6898	0.9091	0.6503	0.7797	0.8308
22 No DA	BRF	11	0.6011	0.9091	0.5458	0.7274	0.8345
23 DA RO	BRF	11	0.7729	0.8909	0.7516	0.8213	0.8373
23 No DA	BRF	11	0.6288	0.8909	0.5817	0.7363	0.8279
24 DA RO	BRF	11	0.8144	0.8727	0.8039	0.8383	0.8423
24 No DA	BRF	11	0.7313	0.8727	0.7059	0.7893	0.8394
25 DA RO	BRF	11	0.8199	0.8545	0.8137	0.8341	0.8423
25 No DA*	BRF	11	0.7424	0.8364	0.7255	0.7809	0.8394
26 DA RO	BRF	14	0.8366	0.8364	0.8366	0.8365	0.8321
26 No DA	BRF	14	0.7507	0.8364	0.7353	0.7858	0.8321
27 DA RO	BRF	11	0.8449	0.8182	0.8497	0.8339	0.8450
27 No DA	BRF	11	0.7258	0.8182	0.7092	0.7637	0.8338
28 DA RO	BRF	7	0.8393	0.8000	0.8464	0.8232	0.8297
28 No DA	BRF	7	0.7562	0.8000	0.7484	0.7742	0.8346
29 DA RL	BRF	8	0.8504	0.7455	0.8693	0.8074	0.8403
29 No DA*	BRF	8	0.8144	0.7636	0.8235	0.7936	0.8351
30 DA RO	LR	14	0.8670	0.7091	0.8954	0.8023	0.8294
30 No DA	LR	14	0.8089	0.7091	0.8268	0.7679	0.8381
31 DA RO	MLP	12	0.8837	0.6545	0.9248	0.7897	0.8361
31 No DA	MLP	12	0.8587	0.6545	0.8954	0.7750	0.8343
32 DA RO	MLP	11	0.8892	0.5818	0.9444	0.7631	0.8346
32 No DA*	MLP	11	0.8781	0.6000	0.9281	0.7641	0.8345
33 DA RO	MLP	10	0.9086	0.5818	0.9673	0.7746	0.8649
33 No DA*	MLP	10	0.8670	0.6000	0.9150	0.7575	0.8618

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

*: Obtenido con el valor de sensibilidad más cercano, en base a los umbrales de decisión obtenidos con el conjunto de validación.

4.3.2. Modelos optimizados versus modelos sin optimizar

En la tabla 4.13 aparece una comparación entre el desempeño de los modelos de la tabla 4.10 versus los mejores modelos sin optimizar, con una sensibilidad superior a 0.8. Para estos últimos se consideraron el caso de usar las 11 variables utilizadas por los modelos 24, 25 y 27, las 14 variables utilizadas por el modelo 26, y todas las variables disponibles (59). De igual forma que para GDM, este último caso no tendría optimización alguna.

En la figura 4.7 aparece un gráfico que muestra los errores (FP+FN) en función de los verdaderos positivos, en azul los modelos de la tabla 4.10, los mejores modelos sin optimizar utilizando todas las variables (rojo), los mejores modelos utilizando las 5 variables utilizadas por los modelos 9, 13 y 17 (verde), los mejores modelos utilizando las 7 variables del modelo 33 (naranja) y las 12 variables del modelo 29 (morado). En celeste aparece los mejores modelos sin optimizar de cualquiera de las 3 opciones previas.

Tabla 4.13: Tabla de resultados LGA (tabla 4.10) versus modelos sin optimizar, sobre 0.8 de sensibilidad.

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
17	MLP	8	RL	0.5014	1	0.4118	0.7059	0.8419
-	MLP	11	-	0.3490	1	0.2320	0.6160	0.8352
-	RF	14	-	0.4294	1	0.3268	0.6634	0.8501
-	RF	59	-	0.2881	1	0.1601	0.5801	0.8368
18	BRF	8	RO	0.5319	0.9818	0.4510	0.7164	0.8325
-	LR	11	-	0.4238	0.9818	0.3235	0.6527	0.8400
-	LR	14	-	0.4238	0.9818	0.3235	0.6527	0.8400
-	LGBM	59	-	0.3352	0.9818	0.2190	0.6004	0.8175
19	MLP	9	RO	0.5817	0.9636	0.5131	0.7384	0.8172
-	ET	11	-	0.4848	0.9636	0.3987	0.6812	0.8452
-	LR	14	-	0.4349	0.9636	0.3399	0.6518	0.8389
-	RF	59	-	0.4377	0.9636	0.3431	0.6534	0.8368
20	MLP	14	RO	0.6565	0.9455	0.6046	0.7750	0.8277
-	ET	11	-	0.5374	0.9455	0.4641	0.7048	0.8452
-	RF	14	-	0.4931	0.9455	0.4118	0.6786	0.8501
-	BRF	59	-	0.4820	0.9455	0.3987	0.6721	0.8215
21	MLP	14	RO	0.6787	0.9273	0.6340	0.7806	0.8277
-	BRF	11	-	0.5540	0.9273	0.4869	0.7071	0.8379
-	RF	14	-	0.5679	0.9273	0.5033	0.7153	0.8501
-	SVM	59	-	0.4709	0.9273	0.3889	0.6581	0.8196

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

Tabla 4.14: Continuación Tabla 4.13

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
22	BRF	11	RO	0.6898	0.9091	0.6503	0.7797	0.8308
-	BRF	11	-	0.6233	0.9091	0.5719	0.7405	0.8379
-	MLP	14	-	0.6565	0.9091	0.6111	0.7601	0.8406
-	RF	59	-	0.5623	0.9091	0.5000	0.7045	0.8368
23	BRF	11	RO	0.7729	0.8909	0.7516	0.8213	0.8373
-	ET	11	-	0.6759	0.8909	0.6373	0.7641	0.8452
-	MLP	14	-	0.6925	0.8909	0.6569	0.7739	0.8406
-	ET	59	-	0.5928	0.8909	0.5392	0.7151	0.8099
24	**BRF**	**11**	**RO**	**0.8144**	**0.8727**	**0.8039**	**0.8383**	**0.8423**
-	ET	11	-	0.7313	0.8727	0.7059	0.7893	0.8452
-	RF	14	-	0.7285	0.8727	0.7026	0.7877	0.8501
-	LR	59	-	0.6704	0.8727	0.6340	0.7534	0.8242
25	**BRF**	**11**	**RO**	**0.8199**	**0.8545**	**0.8137**	**0.8341**	**0.8423**
-	ET	11	-	0.7645	0.8545	0.7484	0.8015	0.8452
-	RF	14	-	0.7368	0.8545	0.7157	0.7851	0.8501
-	LR	59	-	0.7147	0.8545	0.6895	0.7720	0.8242
26	**BRF**	**14**	**RO**	**0.8366**	**0.8364**	**0.8366**	**0.8365**	**0.8321**
-	RF	11	-	0.7313	0.8364	0.7124	0.7744	0.8283
-	ET	14	-	0.7812	0.8364	0.7712	0.8038	0.8277
-	LR	59	-	0.7175	0.8364	0.6961	0.7662	0.8242
27	**BRF**	**11**	**RO**	**0.8449**	**0.8182**	**0.8497**	**0.8339**	**0.8450**
-	ET	11	-	0.7784	0.8182	0.7712	0.7947	0.8452
-	ET	14	-	0.7922	0.8182	0.7876	0.8029	0.8277
-	LR	59	-	0.7396	0.8182	0.7255	0.7718	0.8242
28	BRF	7	RO	0.8393	0.8000	0.8464	0.8232	0.8297
-	ET	11	-	0.7950	0.8000	0.7941	0.7971	0.8452
-	RF	14	-	0.7922	0.8000	0.7909	0.7955	0.8452
-	LR	59	-	0.7645	0.8000	0.7582	0.7791	0.8242

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto.

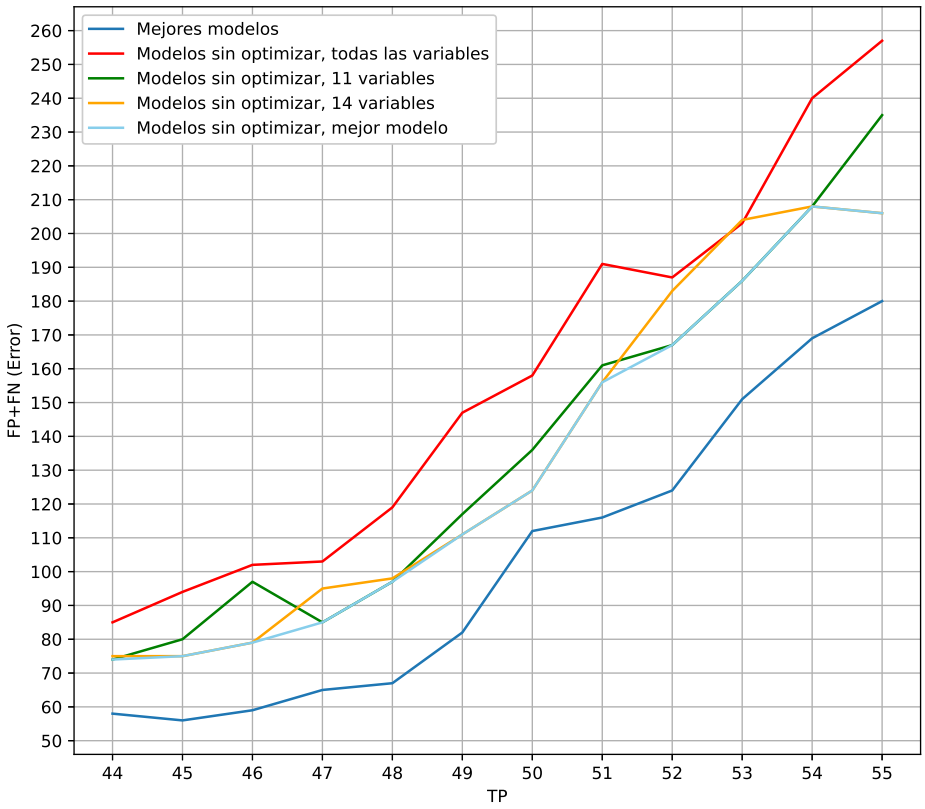

Figura 4.7: Errores (FP+FN) en función de los verdaderos positivos, en azul los modelos de la tabla 4.10, los mejores modelos sin optimizar utilizando todas las variables (rojo), los mejores modelos utilizando las 5 variables utilizadas por los modelos 9, 13 y 17 (verde), los mejores modelos utilizando las 7 variables del modelo 33 (naranja) y las 12 variables del modelo 29 (morado). En celeste aparece los mejores modelos sin optimizar de cualquiera de las 3 opciones previas de la tabla 4.13.

4.4. Comparación con Modelos del estado del arte

4.4.1. GDM

La limitada disponibilidad de los data set de los trabajos previos dificultan una comparación directa de rendimiento entre los modelos y los del estado del arte [27–44]. A pesar de esto, es posible realizar una comparación general comparando los resultados de las diferentes métricas obtenidos en otros trabajos. No obstante, a la hora de realizar la comparación, es necesario tener en cuenta otros factores importantes, tales como las características de la población analizada y el criterio de diagnóstico de GDM, los cuales varían entre países y regiones de los cuales provienen los datos utilizados en los estudios. La tabla 4.15 muestra una comparación de los resultados de los mejores modelos presentados en este trabajo contra los modelos del estado del arte con variables de entrada con una complejidad de adquisición similar a las utilizadas por los modelos presentados y con un criterio de diagnóstico de GDM igual o similar al utilizado en este trabajo (IADPSG/HAPO/WHO 2013) [27–34, 36].

Otros modelos [27, 32, 35, 37–44] que presentan variables más complejas de adquirir no son añadidas a la tabla 4.15.

Tabla 4.15: Resultados de los mejores modelos con distintos rangos de sensibilidad en comparación con los modelos de la literatura con similares variables de entrada y criterio de diagnóstico de GDM.

Modelos	Accuracy	Sensitivity	Specificity	Recall Macro	AUCROC
DNN, 7 Variables [27]	-	0.7	0.69	0.695*	0.77
LR, 5 Variables Continuas [28]	-	0.61	0.80	0.705*	0.766
LGBM. 9 preguntas (Variables) [32]	-	-	-	-	0.799
RF, 6 Variables [33]	0.789	0.651	0.813	0.732*	0.777
LR, 4 Variables [29]	-	-	-	-	0.70
1 Variable ** [34]	-	0.490	0.676	0.583*	0.608
RECPAM, 3 Variables [31]	-	0.89	0.40	0.645*	-
2 Variables ** [30]	-	0.51	0.81	0.660*	0.71
NN, 4 Variables, IADPSG Criteria [36]	-	-	-	-	0.73
Modelo 1 GDM	0.3994	1	0.3169	0.6585	0.8189
Modelo 5 GDM	0.5511	0.9744	0.4930	0.7337	0.8002
Modelo 9 GDM	0.6068	0.9487	0.5599	0.7543	0.8234
Modelo 13 GDM	0.6130	0.9231	0.5704	0.7468	0.8234
Modelo 17 GDM	0.6594	0.8974	0.6268	0.7621	0.8199
Modelo 21 GDM	0.6873	0.8718	0.6620	0.7669	0.8160
Modelo 25 GDM	0.7090	0.8462	0.6901	0.7681	0.8142
Modelo 29 GDM	0.7554	0.8205	0.7465	0.7835	0.8135
Modelo 33 GDM	0.7399	0.7949	0.7324	0.7636	0.8143

DNN: *Deep Neural Network*. NN: *Neural Network*.
*: Valores calculados en base a la fórmula de Recall Macro. **: Modelo determinístico.

La tabla 4.16 ofrece una lista de variables de entradas usado por cada uno de los mejores modelos de GDM, incluyendo los modelos usados en la comparación.

Tabla 4.16: Variables de entrada de cada modelo incluido en la comparación y los mejores modelos de GDM.

Modelos	Variables
DNN, 7 Variables [27]	Edad, GDM previo, Historial de diabetes en relativos de primer grado, Multiples embarazos, FPG, HBA_{1c}, Triglicéridos.
LR, 5 Variables Continuas [28]	Edad, BMI pre-embarazo, FPG y Triglicéridos.
LGBM. 9 preguntas (Variables) [32]	Edad, Peso y Altura, Historial de diabetes en relativos de primer grado, Colesterol alto, Abortos espontáneos, PCOS, Pre-diabetes, Enfermedades Coronarias, GDM o alta presión arterial antes del embarazo actual, HBA_{1c}, Nacimientos previos (Sí o No), si es Sí, número de veces, y FPG o OGTT en esos embarazos si es que están disponibles.
RF, 6 Variables [33]	Edad, BMI pre-embarazo, Circunferencia abdominal primer trimestre, Gravidez, PCOS, menstruación irregular, Historial familiar de diabetes.
LR, 4 Variables [29]	Edad, BMI, FPG, Historial de diabetes en familiares de primer grado.
1 Variable * [34]	FPG.
RECPAM, 3 Variables [31]	BMI, FPG, Historial de diabetes en familiares de primer grado.
2 Variables * [30]	BMI, FPG.
NN, 4 Variables, IADPSG Criteria [36]	Presión arterial promedio, Edad, Historial de GDM previo, Etnia.
Modelo 1, 29 GDM	Edad, Peso, BMI, Uso de drogas ilícitas, Enfermedades cardiacas, Enfermedades recurrentes del tracto urinario, Desordenes psiquiátricos, Enfermedades crónicas al riñón, Enfermedad inflamatoria intestinal, Resistencia a la Insulina, Uso de drogas antihipertensivas, FPG.
Modelo 5 GDM	Edad, Peso, BMI, Uso de drogas ilícitas, Enfermedades crónicas al riñón, Enfermedad inflamatoria intestinal, Uso de drogas antihipertensivas, FPG.
Modelo 9, 13, 17 GDM	Edad, Peso, BMI, Gravidez, FPG.
Modelo 21 GDM	Edad, Peso, BMI, Gravidez, Paridad, Hipertensión Crónica, FPG.
Modelo 25 GDM	Edad, Peso, BMI, Enfermedad inflamatoria intestinal, Uso de drogas antihipertensivas, FPG.
Modelo 33 GDM	Edad, Peso, BMI, Gravidez, Hipertensión Crónica, Uso de drogas antihipertensivas, FPG.

DNN: *Deep Neural Network*. NN: *Neural Network*.
*: Modelo determinístico.

4.4.2. LGA

Similar a lo ocurrido con GDM, los conjuntos de datos utilizados en los trabajos publicados de LGA también son limitados a su utilización pública debido al carácter sensible que poseen este tipo de datos [19, 23, 50–60]. Más aún, cada estudio realizado utiliza diferentes curvas para realizar el diagnóstico de LGA al nacer, sumado a las diferencias entre las características de la población, hace que la comparación entre trabajos sea aún más difícil. No obstante, en la tabla 4.17 es presentada una comparación de rendimiento entre trabajos previos con set de entradas similares, es decir, con datos recopilados en el tercer trimestre de embarazo, principalmente de ultrasonidos [19, 50–54].

Es importante recalcar que esta comparación es solo de carácter referencial, ya que los criterios de LGA son variados entre trabajos y ciertos modelos son limitados a cierto tipo de población, como [19] y [53], que sus trabajos apuntan solo a población con GDM y PGDM respectivamente. Los modelos [23, 55–60] no fueron incluidos en la comparación, ya sea por la diferencia en la temporalidad de las entradas, solo primer y/o segundo trimestre, o porque los datos utilizados no provienen de exámenes rutinarios.

Tabla 4.17: Resultados de los mejores modelos junto con los modelos de la literatura con similares variables de entrada para LGA.

Modelos	Accuracy	Sensitivity	Specificity	Recall Macro	AUCROC
1 Variable **, NC MSUH [50]	-	0.738	0.80	0.769*	0.83
1 Variable **, FMF [51]	0.8444+	0.772	0.854	0.813*	0.81
3 Variables, LR, GDM [19]	0.867	0.811	0.874	0.8425*	0.916
2 Variables, LR, PGDM [53]	-	0.786	0.800	0.793*	0.85
7 Variables, LR [54]	-	0.43	0.90	0.665*	0.76
Modelo 33 LGA	0.9086	0.5818	0.9673	0.7746	0.8649
Modelo 24 LGA	0.8144	0.8727	0.8039	0.8383	0.8423
Modelo 25 LGA	0.8199	0.8545	0.8137	0.8341	0.8423
Modelo 26 LGA	0.8366	0.8364	0.8366	0.8365	0.8321
Modelo 27 LGA	0.8449	0.8182	0.8497	0.8339	0.8450

*: Valores calculados en base a la fórmula de Recall Macro. **: Modelo determinístico.
+: Inferido.

En la tabla 4.18 es mostrado la lista de variables de entrada de los modelos destacados de la tabla 4.10 y el modelo 13 de la tabla 4.9, usados en la comparación.

Tabla 4.18: Variables de entrada de cada modelo incluido en la comparación y los mejores modelos de LGA.

Modelos	Variables
1 Variable *, NC MSUH [50]	Percentil de peso estimado.
1 Variable *, FMF [51]	EFW.
3 Variables, LR, GDM [19]	Z-score de Circunferencia de la cabeza en las semanas 24 a 30 de gestación, Z-score de Circunferencia abdominal dentro de 2 semanas antes del parto y al menos después de las 34 semanas de gestación, BMI pre-embarazo. Solo pacientes con GDM.
2 Variables, LR, PGDM [53]	EFW, HBA_{1c}. Solo Pacientes con PGDM.
7 Variables, LR [54]	EFW, Edad, BMI, Etnia, Paridad, Consumo de tabaco, Sexo fetal.
Modelo 33 LGA	Peso y BMI primer control, Peso y BMI último control, Ganancia de peso, Altura, EFW, OGTT ayuna y 2 horas postprandial, US Date.
Modelo 24 LGA	Edad, BMI primer control, Peso y BMI último control, Desordenes ginecológicos, Polihidramnios, EFW, Hadlock. <p25 %, Hadlock >p75 %, Hadlock >p90 %, Hadlock >p95 %.
Modelo 25 LGA	Edad, BMI primer control, Peso y BMI último control, Desordenes ginecológicos, Polihidramnios, EFW, Hadlock <p25 %, Hadlock >p75 %, Hadlock >p90 %, Hadlock >p95 %.
Modelo 26 LGA	Edad, BMI primer control, Peso y BMI último control, Uso de Alcohol, Desordenes ginecológicos, Resistencia a la Insulina, Eclampsia, Polihidramnios, EFW, Hadlock <p25 %, Hadlock >p75 %, Hadlock >p90 %, Hadlock >p95 %.
Modelo 27 LGA	Edad, BMI primer control, Peso y BMI último control, Desordenes ginecológicos, Polihidramnios, EFW, Hadlock. <p25 %, Hadlock >p75 %, Hadlock >p90 %, Hadlock >p95 %.

*: Modelo determinístico.

4.4.2.1. Comparación con Hadlock

La tabla 4.19 muestra una comparación entre la variable Hadlock >p90 % y los modelos con el mismo nivel de sensibilidad y especificidad, con el mismo nivel de sensibilidad son los modelos 13 y 33 y con el mismo nivel de especificidad, los modelos 34 y 35, además de modelos con un menor número de errores, modelo 29 y 30. Cabe destacar que los modelos 34 y 35 no aparecen en tablas previas.

Tabla 4.19: Tabla de comparación de rendimiento entre Hadlock >p90 % y los modelos LGA.

NO	Tipo de Modelo	Número de Variables	DA	Accuracy	Sensitivity	Specificity	Recall Macro	AUC ROC
-	HLK >p90 %	1	-	0.8421	0.5818	0.8889	0.7354	0.7354
16	MLP	6	RO	0.8920	0.5818	0.9477	0.7648	0.8389
29	BRF	8	RL	0.8504	0.7455	0.8693	0.8074	0.8403
30	LR	14	RO	0.8670	0.7091	0.8954	0.8023	0.8294
33	MLP	10	RO	0.9086	0.5818	0.9673	0.7746	0.8649
34	XGB	14	RO	0.8615	0.7091	0.8889	0.7990	0.8264
35	LR	10	RL	0.8615	0.7091	0.8889	0.7990	0.8552

NO: Número de Modelo. RL: Aumentación de Datos con rango Limitado. RO: Aumentación de Datos con rango Experto. HLK: Hadlock.

5. Discusión

5.1. GDM

En el presente trabajo, uno de los mejores modelos con capacidad de detectar con buenos niveles los pacientes con y sin GDM es el modelo 29 de la tabla 4.3. Este modelo alcanza una sensibilidad de 82 % y una especificidad de 74 %. Sin embargo, la metodología propuesta permite encontrar diversos modelos para un nivel de predicción deseado. De esta forma, le deja disponible al profesional de la salud la posibilidad de elegir el punto de operación deseado. Las variables utilizadas por todos los modelos presentados son provenientes de datos adquiridos comúnmente en las etapas tempranas del embarazo, durante las visitas de cuidados prenatales con ginecólogos/obstetras, siendo solicitados de forma rutinaria y que requieren de cierta noción de enfermedades/condiciones preexistentes por parte de la madre, lo cual facilitaría el uso de modelos en la práctica clínica. Otro importante factor a destacar es la relevancia de la métrica sensibilidad, a la cual se le atribuyó un valor elevado al realizar el análisis, ya que lo ideal es siempre evitar cualquier consecuencia severa que pueda provocar GDM tanto a la madre como al feto, que pueden impactar incluso años posteriores al embarazo. El poseer una elevada sensibilidad, puede producir un *trade-off* con especificidad, teniéndose una mayor cantidad de Falsos Positivos, lo cual podría reducirse con exámenes adicionales, sin embargo, esto viene asociado a un costo, que puede ser prohibitivo para ciertas poblaciones. A pesar de poder tener un elevado número de Falsos Positivos, en muchos casos, el principal tratamiento para GDM involucra la aplicación combinada de una dieta y ejercicio, que no debería tener impacto negativo en la salud de la madre y el feto. Las variables más importantes seleccionadas por los métodos utilizados son los relacionados con el metabolismo de la glucosa, glicemia en ayuna en el primer trimestre (1TFG/FPG), estado físico, peso y BMI, edad, hipertensión e información sobre embarazos previos, los cuales son factores de riesgo asociados generalmente a GDM.

La falta de disponibilidad de bases de datos de trabajos publicados respecto a GDM provoca que sea difícil realizar una comparación directa entre los modelos creados y los resultados obtenidos por los modelos del estado del arte. No obstante, en la tabla 4.15, se muestra una comparación, entre los modelos propuestos en esta tesis y otros modelos de la literatura, para lo cual es importante recalcar que las características de las poblaciones, criterios de diagnósticos y tipos de datos utilizados no son iguales entre países/regiones de los cuales provienen los distintos estudios revisados, lo cual hace que solo pueda realizarse una comparación referencial con las métricas mostrada. La tabla 4.15 presenta solo modelos con entradas de complejidad similar a los de este trabajo, y con criterios de GDM similares a los utilizados, los modelos que no cumplan con este requisito no fueron incluidos. En particular, es interesante comparar el modelo 33, SVM de 7 variables con DA RL (tabla 4.15) y el modelo propuesto por Wu y colegas [27] (tabla 4.15), donde el modelo 33 posee una sensibilidad un

13.55 % mayor, con un incremento de un 6.14 % en especificidad. El modelo 17, MLP de 5 Variables con DA RO (tabla 4.15) versus lo planteado por Pintaudi et al. [31], donde ambos modelos obtienen una sensibilidad similar, pero el modelo 17 posee un significativo incremento en la especificidad de un 56.70 %. Un criterio diferente para el diagnóstico de GDM es utilizado por Kumar y asociados [36], utilizando el antiguo criterio recomendado por la WHO, WHO 1999, en el cual, los valores de OGTT de ayuna y 2 horas postprandial deben ser igual o mayores a 126 mg/dL y/o 140 mg/dL respectivamente para que un paciente sea considerado como positivo de GDM. No obstante, ellos plantean otro modelo usando el mismo criterio de diagnóstico de GDM utilizado en este trabajo, IADPSG/HAPO/WHO 2013, obteniendo un AUCROC promedio de 0.73, en una validación cruzada estratificada de 5 hojas.

En la tabla 4.16 se presentan las variables de entrada de los mejores modelos de GDM y los utilizados para realizar la comparación. Como se mencionó previamente, algunas de las mejores soluciones propuestas requieren de variables de fácil adquisición y en baja cantidad, por ejemplo, los modelos 9, 13 y 17, solo requieren de edad, peso, BMI, gravidez y FPG. Las variables más repetidas entre los modelos propuestos son edad, peso, BMI y FPG, lo cual también es consistente con publicaciones previas [27–34, 36]. Los modelos del estado del arte incluidos en la comparación, a pesar de requerir variables menos complejas que otros modelos no incluidos, muchos de ellos requieren de, exámenes extras, triglicéridos, HBA_{1c}, colesterol, visitas adicionales, medición de presión arterial, ultrasonidos, los cuales pueden poseer costos extras asociados o sean difícil su adquisición por temas de agenda/tiempo, o información de embarazos previos, los cuales no siempre están disponibles.

5.2. LGA

En el primer análisis de variables realizado para LGA, realizado con un muestreo de variables de 5 en 5, entregaba que los mejores modelos requerían en general entre 10 y 15 variables (tabla 4.9. Con los mejores resultados, en base al balance de las métricas de sensibilidad y especificidad, *Recall Macro*, eran obtenidos por modelos *Balanced Random Forest* de 10 variables, por lo que fue realizado una exploración de variables entre 6 y 14 variables, en búsqueda de modelos no analizados que pudieran superar el desempeño de los modelos. Los resultados de estos nuevos modelos aparecen en la tabla 4.10. Es posible notar que los modelos número 18, 20, 21, 23-30 son mejores que sus contrapartes con misma sensibilidad presentes en la tabla 4.9. En estos modelos se utilizó como criterio de mejor, un rendimiento superior en métricas o alcanzar el mismo rendimiento, pero, con una cantidad inferior de variables de entrada requeridas. En particular, el nuevo mejor modelo es el modelo 24, que igual al previo mejor modelo, modelo 8, también es una BRF y con la misma técnica de DA, la única diferencia es la necesidad de una variable extra, polihidramnios, la cual implica una mejora en el nivel de especificidad. No obstante, también vale la pena recalcar el desempeño de los modelos 25, 26 y 27, que poseen todos un *Recall Macro* superior a 0.83 por lo que los hace modelos muy balanceados, es especial el modelo 26, que obtiene un valor de métrica virtualmente idéntico en todas ellas, menos AUCROC. De forma similar a lo planteado para GDM, la sensibilidad juega un rol clave, al ser la métrica que indica la cantidad de pacientes positivos de LGA detectados correctamente, se persigue maximizar lo más posible esta

métrica para evitar los Falsos Negativos. Las variables con más relevancia para la predicción de LGA, según los métodos utilizados, son las relacionadas a los ultrasonidos de tercer trimestre, EFW y Hadlocks y condición física de la madre, peso y BMI, que son factores conocidos asociados a LGA. Las variables requeridas, adicionales a las obtenidas con el ultrasonido rutinario de tercer trimestre, no deberían suponer un desafío en su obtención, ya que son en su mayoría características de la madre, extraíbles en controles médicos, y condiciones preexistentes, lo que juega como un punto a favor para su posible aplicación en instituciones médicas, ya que no sería necesario intervenciones adicionales.

El análisis de un alto nivel de especificidad también es interesante, ya que, si se mantiene la sensibilidad, un aumento en esta métrica reduce posibles cesáreas innecesarias que pueden causar complicaciones adicionales a la madre y al recién nacido, tales como infecciones, incremento del riesgo de distrés respiratorio, dificultades en futuros embarazos, etc. Ya que ambos tipos de errores FP y FN, conllevan posibles complicaciones, es tarea de un médico el determinar un equilibro.

Al comparar el rendimiento de solo usar el Hadlock >p90 % como método predictor ("Caso base", tabla 4.8), con el mejor modelo, Modelo 21, BRF 11 variables (tabla 4.10), es posible apreciar que el uso de las variables adicionales que requiere este modelo produce una mejora sustancial, incrementando el valor de la sensibilidad un 50 %, pero reduciendo la especificidad solo un 9.56 %. Una comparación directa también se muestra en la tabla 4.19, donde modelos con el mismo nivel de sensibilidad y especificidad son comparados con el "Caso base". Por ejemplo, el modelo 33, tiene un incremento de un 8.82 % de especificidad con el mismo nivel de sensibilidad de Hadlock. Los modelos 16, 33, 34 y 35 fueron incluidos en este análisis exclusivamente para ofrecer una comparación más directa con el "Caso base", los primeros dos con igual sensibilidad que el "Caso base", ambos modelos ofrecen un mayor nivel de especificidad, sobre un 94 %, en comparación, al aproximado 89 % de especificidad alcanzado por el "Caso base", lo que nos muestra que la utilización de más variables tiene un efecto positivo en el desempeño de los modelos, en comparación a los métodos actuales de predicción de LGA. Mientras que los modelos 34 y 35, ambos alcanzan una mejora de sensibilidad de 21.88 %, mientras se mantiene el mismo nivel de especificidad. Los modelos 29 y 30 aparecen como alternativas con mayor sensibilidad y con un mejor *Accuracy* y con un incremento en el desempeño general del modelo, principalmente el modelo 30, el cual presenta un rendimiento superior en todas las métricas que Hadlock >p90 %.

Aunque no es posible realizar una comparación con los modelos de la literatura, principalmente, por la diferencia de criterios con los que son clasificados los recién nacidos con LGA. Criterios que varían en cada país, como si se puede realizar con GDM, donde a pesar de las diferencias poblacionales, muchos modelos presentados si comparten un mismo tipo de criterio de diagnóstico, de igual forma en la tabla 4.17 se muestran los modelos de la literatura con entradas similares a los desarrollados, tanto en temporalidad, como en complejidad de adquisición, junto con los modelos destacados de LGA, con el fin de obtener una referencia con otros trabajos similares. La tabla 4.18 muestra las variables de todos los modelos incluidos en la tabla 4.17. Al juntar la información de ambas tablas, es posible ver que los modelos con mejor desempeño de la literatura son los modelos [19] y [53], con buen desempeño en general, principalmente el primero, el cual de hecho supera los resultados obtenidos por los modelos de este trabajo, no obstante, estos modelos poseen las limitaciones de estar enfocados solo en ciertos nichos poblacionales, el modelo desarrollado por Kim y colegas [19] es limitado solo a

pacientes positivas de GDM, mientras que el modelo de Kiefer y compañía [53], está limitado solo a pacientes con PGDM, por lo que no pueden ser aplicados a la población general en caso de no poseer esos requerimientos, mientras que los modelos desarrollados en este trabajo están pensados para toda la población.

5.3. Aumentación de Datos

El novedoso método de Aumentación de Datos, "a medida", desarrollado en colaboración con expertos médicos del área ginecológica/obstétrica, tuvo un efecto muy positivo en el desempeño de los modelos. En GDM, como se puede apreciar en las tablas 4.3 y 4.5, el uso de la técnica de DA propuesta ayudó a aumentar la especificidad hasta un 51.43 % y el AUCROC hasta en un 3.70 % para los casos con mismo nivel de sensibilidad. Además, los mejores modelos para cada nivel de sensibilidad fueron obtenidos con el uso de uno de los rangos de DA en 7 de 9 casos, mientras que en caso opuesto fueron en 2 de 9 casos. En LGA, el efecto fue más notorio, en la tabla 4.11 se puede ver que el uso de DA aumentó la especificidad hasta un 201.91 % y el AUCROC hasta un 3.34 % para los mismos niveles de sensibilidad y es especialmente interesar notar que la mejora promedio de especificidad con DA, con el mismo nivel de sensibilidad, es de un 33.05 %. Solo dos modelos de LGA 14 y 15 poseen un rendimiento inferior con DA. Respecto a los rangos de DA utilizados para las variables, se puede notar que en LGA es predominante el uso del rango propuesto por el experto, RO, mientras que en GDM, es levemente más común el uso del rango limitado, RL.

5.4. Modelos

El uso de 12 diferentes modelos tuvo como objetivo el poder utilizar modelos con distintos niveles de capacidad, bajo la incertidumbre *a priori* de que tan complejo es la tarea de clasificación, por lo que se procedió a incluir algoritmos más "sencillos" como los de Naïve Bayes y Árboles de Decisión, hasta clasificadores más sofisticados como los del tipo *Ensemble*, *Random Forest*, *Gradient Boosting*, y Perceptrón Multi-Capas. En particular para GDM, los modelos dominantes fueron los MLP y *Support Vector Machine*, por otro lado, en LGA, *Balanced Random Forest* y MLP dominaron el panorama de clasificación.BRF y sus modificaciones para afrontar el desbalance de clases, tuvieron una importante relevancia al generar los mejores modelos de LGA, principalmente, los modelos más balanceados en cuanto a detección de ambas clases. No obstante, este éxito no fue obtenido en GDM, lo cual se puede deber a las variables de entradas. Los modelos de *Gradient Boosting*, XGB y LGBM, a pesar de no aparecer en ninguna de las tablas, tuvieron un buen desempeño, pero no el suficiente para superar los modelos seleccionados. Los modelos de *Extra-Trees* y K vecinos más cercanos, no tuvieron un buen desempeño. El buen rendimiento en general de MLP en ambos problemas se condice con la popularidad que tiene este algoritmo en los problemas de clasificación.

5.5. Selección de Variables

Los principales motivos del uso de métodos de selección de variables son el poder crear modelos simples para poder tener una mayor posibilidad de ser aplicado en ambientes clínicos y tener un mejor desempeño de los modelos en comparación al utilizar todas las variables. Como se puede ver en las tablas de resultados de ambas condiciones, la mayoría de los modelos propuestos requieren una cantidad significativamente menor de variables en comparación con el total disponible. En particular para GDM, el aumento de variables (tabla 4.4), incluso con un aumento de más del doble de variables, el modelo 43, 15 variables, en comparación al modelo 25, solo obtiene una mejora del 2.01 % de especificidad, lo cual probablemente no tenga un impacto tan significativo frente al aumento de la complejidad requerida para alcanzar ese valor. En LGA, los mejores resultados dispuestos en la tabla 4.17 incluyeron todos los modelos con un número de variables múltiplo de 5, incluyendo el total, 59, por lo que, si en esa tabla no apareció un modelo de mayor cantidad de variables, es porque métricamente no era superior o, de ser igual, necesitaba un mayor número de entradas. Por lo tanto, es probable que ciertas entradas no aporten información para la predicción de ninguna de las dos condiciones analizadas.

Las principales variables seleccionadas para GDM, tales como 1TFG, BMI, Peso Materno y Edad, utilizadas por todos los modelos del estado del arte analizados, a excepción de [27], el cual no utiliza el BMI y Peso Materno. En [39] se utiliza la edad acompañada de variables sanguíneas. Por lo que la selección se condice con lo reportado por otros estudios. El resto de las variables como preexistencias, información de embarazos previos e información cardíaca, no son analizadas por todos los estudios debido a que no son recopiladas en los estudios. Sin embargo, en los estudios que si utilizan este tipo de información generalmente si es utilizada por sus modelos [27, 32, 35, 36, 38, 40].

Con LGA sucede algo similar que con GDM, las variables más predominantes son las relacionadas con el ultrasonido (EFW y variables obtenidas a partir de la curva Hadlock), lo que sucede de igual forma con todos los estudios analizados con una fecha de predicción similar, a excepción de [55], que presentan una alternativa al ultrasonido. Por otro lado, las variables de peso y de diabetes gestacional (OGTT) son variables utilizadas por modelos que buscan realizar una predicción temprana [23, 56–58].

5.6. Optimización de modelos

Al comparar el desempeño de los modelos sin optimizar con los modelos optimizados, presentes en las tablas 4.6 y 4.13 y las figuras 4.3 y 4.7. Es posible ver una mejora promedio de 30.65 % en la disminución de errores (FP+FN) por parte de los modelos optimizados en contraste a los modelos sin optimización y con todas las variables, en el mismo nivel de sensibilidad. Esto en el nivel de sensibilidad analizado. Se alcanza el *peak* en el nivel de sensibilidad igual a 0.8462, con una mejoría de 46.29 %. Con selección de variables este valor se disminuye, aunque no deja de ser importante. Para 5 variables el error promedio es de 24.72 %, para 7 variables 26.36 % y para 12 variables igual a 28.99 %.

Con LGA, se analizó el caso con sensibilidad igual o superior a 0.8. Los resultados son similares a los de LGA, en comparación con modelos sin ninguna optimización, los modelos presentados disminuyen los errores (FP+FN) en un 35.53 % promedio. Al seleccionar variables, la disminución de error es de 25.62 y 23.96 % respectivamente respecto a modelos con 11 y 14 variables, sin optimización de hiperparámetros.

6. Conclusiones

En este trabajo de tesis se desarrollaron un amplio abanico de modelos de aprendizaje de máquinas con el fin de predecir las enfermedades/condiciones de Diabetes Mellitus Gestacional (GDM) y Grande para la Edad Gestacional (LGA). Estos modelos fueron desarrollados utilizando doce algoritmos diferentes de aprendizaje de máquinas, a los cuales se les optimizaron sus hiperparámetros con el fin de alcanzar el mejor desempeño predictivo. La creación de una amplia gama de modelos fue con la idea en mente de ofrecer a especialistas de la salud diversos niveles de predicción de cada clase, negativa, no tiene la condición, y positiva, tiene la condición. Para que luego sea este profesional el que tome la decisión de que nivel de predicción de cada clase es el más adecuado a su juicio. A su vez, tiene la opción de variar la complejidad del modelo, al modificar el número de entradas requeridas para la utilización de estos, teniendo en consideración que un mayor número puede resultar en variables más complejas de obtener.

Se aplicaron diversas herramientas y técnicas comunes en aprendizaje de máquinas con el fin de aumentar el poder predictivo de los modelos, tales como Normalización de las entradas, Búsqueda de Grilla, Validación Cruzada, Transformación de Datos, Selección de Variables y Aumentación de Datos. Cada una de estas técnicas cumple un papel importante en el proceso de obtención de resultados. No obstante, es destacable el rol de Selección de Variables y Aumentación de Datos. El primero, cumple un doble papel, ya que, por un lado, permite una mejora en el rendimiento de los modelos al descartar variables innecesarias o que generen ruido en la predicción, y por otro, permite la concepción de modelos más simples, lo que favorece la posibilidad de implementar los modelos como aplicación clínica. Mientras que Aumentación de Datos posibilita una mejora en el desempeño del entrenamiento de los modelos, lo que influye, a su vez, en un aumento en el rendimiento en la evaluación del modelo. En particular para este trabajo, se desarrolló un método de Aumentación de Datos personalizado e innovador para estas condiciones y enfermedades, el cual está diseñado específicamente para ciertas variables, con valores límite y restricciones, provistos por un especialista en Obstetricia/Ginecología, para que los valores de los pacientes creados con este método tengan sentido, sean realistas y no contradigan los datos originales. Cabe destacar que este método de Aumentación es original y no hay publicaciones previas con un método similar.

Al comparar los mejores modelos obtenidos para GDM, es posible apreciar, en general, una mejora de rendimiento en comparación con modelos provenientes de la literatura, cuando las variables de entrada y los criterios de diagnóstico de GDM son similares. La comparación de modelos para LGA es más compleja, debido a los diversos criterios para LGA que, en general, son personalizados para cada país. Sin embargo, al comparar los modelos desarrollados con el usar solo peso fetal estimado, mediante la curva de Hadlock, es posible observar una mejora significativa al aplicar más variables.

Se puede concluir de este trabajo, que fue posible crear, desarrollar y evaluar modelos de aprendizaje de máquinas para la predicción de Diabetes Mellitus Gestacional y Grande para la Edad Gestacional, utilizando datos provenientes de Registros Médicos Electrónicos, que fuesen superior a los ya descritos en el estado del arte. Los modelos desarrollados pudieron aprender de los datos y permiten realizar diagnósticos de las condiciones, para GDM, los modelos permiten realizar una predicción anticipada y los datos utilizados abren la posibilidad de implementar estos modelos en un ámbito médico como métodos de soporte de decisión, para LGA, los modelos permitieron una mejora significativa en comparación a solo usar una variable del ultrasonido rutinario de tercer trimestre, las variables utilizadas también permite que sea posible su utilización médica. Como fue mencionado previamente, el uso de diversas herramientas y técnicas utilizadas normalmente para estas labores permitieron, algunas, alcanzar modelos de buen desempeño y otras, aumentar el rendimiento de los modelos ya desarrollados.

Los aportes y resultados más importantes relacionados con la predicción temprana de GDM fueron publicados en un artículo de una revista WoS [103]. Las contribuciones y resultados de LGA fueron enviadas a una revista, actualmente se encuentra en revisión.

6.1. Trabajo Futuro

Como trabajo futuro queda la implementación de los modelos en un ambiente clínico que es un objetivo a largo plazo, ya que los modelos deben tener una fase de pruebas en un centro asistencial, para verificar el correcto funcionamiento de ellos. Primero habría que seleccionar uno o dos modelos de interés por parte de los profesionales médicos, en base a los requerimientos de sensibilidad y especificidad que posean. El trabajo realizado ofrece múltiples posibilidades a elegir por parte de la contraparte médica.

Una deseable segunda fase de pruebas, en diversos centros médicos del país, para garantizar que no exista sesgo alguno por la población utilizada, y luego una implementación y análisis respecto a las contribuciones que tiene este modelo en la salud pública. Estos procesos pueden tardar años.

Una posible mejora para los modelos desarrollados para GDM es la utilización de datos adicionales. Aunque suene contradictorio con lo propuesto en este trabajo, la utilización de otros exámenes, por ejemplo, sanguíneos, pueden contribuir a un mejor desempeño de los modelos. En caso de ser demostrada la relevancia de estos nuevos datos, es posible cambiar las políticas públicas y sugerir estos exámenes como rutinarios. Sin embargo, este proceso puede verse truncado por temas económicos o de burocracia.

Una alternativa para LGA es realizar un método de predicción con datos tempranos, similar a GDM, pero sería interesante plantear el uso de ultrasonido en primer y segundo trimestre, como otros modelos de la literatura proponen. No obstante, esta solución también posee el problema de que son requeridos intervenciones adicionales que tienen asociado un costo adicional, lo que complica su implementación, lo cual es una limitación en gran parte de los centros asistenciales.

Otra posible opción de trabajo futuro es el análisis de otras enfermedades, no necesariamente del embarazo, que se puedan abordar con los algoritmos utilizados en este trabajo o similares. También es posible utilizar otros modelos para la predicción de GDM y LGA, como los *transformers* que poseen variantes para datos tabulares.

Bibliografía

[1] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, pp. 487–492, 2018. [Online]. Available: https://doi.org/10.1038/nature25988

[2] J. E. Zambrano, D. P. Benalcazar, C. A. Perez, and K. W. Bowyer, "Iris Recognition Using Low-Level CNN Layers Without Training and Single Matching," *IEEE Access*, vol. 10, pp. 41 276–41 286, 2022.

[3] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, pp. 1559–1567, 2018. [Online]. Available: https://doi.org/10.1038/s41591-018-0177-5

[4] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Scientific Reports*, vol. 10, p. 19549, 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76550-z

[5] H. Sun, A. Wu, M. Lu, and S. Cao, "Liability, risks, and recommendations for ultrasound use in the diagnosis of obstetrics diseases," *Heliyon*, vol. 9, p. e21829, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023090370

[6] F. Mohsen, H. R. H. Al-Absi, N. A. Yousri, N. E. Hajj, and Z. Shah, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," *npj Digital Medicine*, vol. 6, p. 197, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00933-5

[7] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 8459–8486, 2023. [Online]. Available: https://doi.org/10.1007/s12652-021-03612-z

[8] American Diabetes Association Professional Practice Committee, "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022," *Diabetes Care*, vol. 45, pp. S17–S38, 7 2021. [Online]. Available: https://doi.org/10.2337/dc22-S002

[9] E. M. Wendland, M. R. Torloni, M. Falavigna, J. Trujillo, M. A. Dode, M. A. Campos, B. B. Duncan, and M. I. Schmidt, "Gestational diabetes and pregnancy

outcomes - a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria," *BMC Pregnancy and Childbirth*, vol. 12, p. 23, 2012. [Online]. Available: https://doi.org/10.1186/1471-2393-12-23

[10] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 138, pp. 271–281, 4 2018, doi: 10.1016/j.diabres.2018.02.023. [Online]. Available: https://doi.org/10.1016/j.diabres.2018.02.023

[11] International Diabetes Federation, "Chile Diabetes report 2000 - 2045," accessed: 2023-07-18. [Online]. Available: https://diabetesatlas.org/data/en/country/41/cl.html

[12] S. S. Casagrande, B. Linder, and C. C. Cowie, "Prevalence of gestational diabetes and subsequent Type 2 diabetes among U.S. women," *Diabetes Research and Clinical Practice*, vol. 141, pp. 200–208, 7 2018, doi: 10.1016/j.diabres.2018.05.010. [Online]. Available: https://doi.org/10.1016/j.diabres.2018.05.010

[13] L. P. Lowe, B. E. Metzger, A. R. Dyer, J. Lowe, D. R. McCance, T. R. J. Lappin, E. R. Trimble, D. R. Coustan, D. R. Hadden, M. Hod, J. J. N. Oats, B. Persson, and for the HAPO Study Cooperative Research Group, "Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study: Associations of maternal A1C and glucose with pregnancy outcomes," *Diabetes Care*, vol. 35, pp. 574–580, 7 2012. [Online]. Available: https://doi.org/10.2337/dc11-1687

[14] J. P. Vandorsten, W. C. Dodson, W. A. E. M. A, Grobman, J. M. Guise, B. M. Mercer, H. L. Minkoff, B. Poindexter, L. A. Prosser, G. F. Sawaya, J. R. Scott, L. S. R. M, Smith, A. Thomas, and A. T. N. Tita, "NIH consensus development conference: diagnosing gestational diabetes mellitus," *NIH Consens. State Sci. Statements*, vol. 29, pp. 1–31, 3 2013.

[15] J. Fu and R. Retnakaran, "The life course perspective of gestational diabetes: An opportunity for the prevention of diabetes and heart disease in women," *eClinicalMedicine*, vol. 45, 3 2022, doi: 10.1016/j.eclinm.2022.101294. [Online]. Available: https://doi.org/10.1016/j.eclinm.2022.101294

[16] J. F. Plows, J. L. Stanley, P. N. Baker, C. M. Reynolds, and M. H. Vickers, "The Pathophysiology of Gestational Diabetes Mellitus," *International Journal of Molecular Sciences*, vol. 19, 2018. [Online]. Available: https://www.mdpi.com/1422-0067/19/11/3342

[17] A. Sweeting, J. Wong, H. R. Murphy, and G. P. Ross, "A Clinical Update on Gestational Diabetes Mellitus," *Endocrine Reviews*, vol. 43, pp. 763–793, 7 2022. [Online]. Available: https://doi.org/10.1210/endrev/bnac003

[18] A. Hocquette, M. Durox, R. Wood, K. Klungsøyr, K. Szamotulska, S. Berrut, T. Rihs, T. Kyprianou, L. Sakkeus, A. Lecomte, I. Zile, S. Alexander, J. Klimont, H. Barros, M. Gatt, J. Isakova, B. Blondel, M. Gissler, and J. Zeitlin,

"International versus national growth charts for identifying small and large-for-gestational age newborns: A population-based study in 15 european countries," *The Lancet Regional Health - Europe*, vol. 8, p. 100167, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666776221001447

[19] H.-S. Kim, S.-Y. Oh, G. J. Cho, S.-J. Choi, S. C. Hong, J.-Y. Kwon, and H. S. Kwon, "A Predictive Model for Large-for-Gestational-Age Infants among Korean Women with Gestational Diabetes Mellitus Using Maternal Characteristics and Fetal Biometric Parameters," *Journal of Clinical Medicine*, vol. 11, 2022. [Online]. Available: https://www.mdpi.com/2077-0383/11/17/4951

[20] L. Harvey, R. van Elburg, and E. M. van der Beek, "Macrosomia and large for gestational age in asia: One size does not fit all," *Journal of Obstetrics and Gynaecology Research*, vol. 47, pp. 1929–1945, 2021. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/jog.14787

[21] I. R. Falcão, R. de Cássia Ribeiro-Silva, M. F. de Almeida, R. L. Fiaccone, N. J. Silva, E. S. Paixao, M. Y. Ichihara, L. C. Rodrigues, and M. L. Barreto, "Factors associated with small- and large-for-gestational-age in socioeconomically vulnerable individuals in the 100 million brazilian cohort," *The American Journal of Clinical Nutrition*, vol. 114, pp. 109–116, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002916522003100

[22] G.-R. Yang, T. D. Dye, and D. Li, "Effects of pre-gestational diabetes mellitus and gestational diabetes mellitus on macrosomia and birth defects in upstate new york," *Diabetes Research and Clinical Practice*, vol. 155, 9 2019, doi: 10.1016/j.diabres.2019.107811. [Online]. Available: https://doi.org/10.1016/j.diabres.2019.107811

[23] N. Wang, H. Guo, Y. Jing, Y. Zhang, B. Sun, X. Pan, H. Chen, J. Xu, M. Wang, X. Chen, L. Song, and W. Cui, "Development and validation of risk prediction models for large for gestational age infants using logistic regression and two machine learning algorithms," *Journal of Diabetes*, vol. 15, pp. 338–348, 4 2023, https://doi.org/10.1111/1753-0407.13375. [Online]. Available: https://doi.org/10.1111/1753-0407.13375

[24] M. Saeedi, Y. Cao, H. Fadl, H. Gustafson, and D. Simmons, "Increasing prevalence of gestational diabetes mellitus when implementing the iadpsg criteria: A systematic review and meta-analysis," *Diabetes Research and Clinical Practice*, vol. 172, p. 108642, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168822720308998

[25] S. L. Boulet, G. R. Alexander, H. M. Salihu, and M. Pass, "Macrosomic births in the united states: Determinants, outcomes, and proposed grades of risk," *American Journal of Obstetrics & Gynecology*, vol. 188, pp. 1372–1378, 5 2003, doi: 10.1067/mob.2003.302. [Online]. Available: https://doi.org/10.1067/mob.2003.302

[26] C. M. Boney, A. Verma, R. Tucker, and B. R. Vohr, "Metabolic Syndrome in Childhood: Association With Birth Weight, Maternal Obesity, and Gestational

Diabetes Mellitus," *Pediatrics*, vol. 115, pp. e290–e296, 10 2005. [Online]. Available: https://doi.org/10.1542/peds.2004-1808

[27] Y.-T. Wu, C.-J. Zhang, B. W. Mol, A. Kawai, C. Li, L. Chen, Y. Wang, J.-Z. Sheng, J.-X. Fan, Y. Shi, and H.-F. Huang, "Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning," *The Journal of Clinical Endocrinology & Metabolism*, vol. 106, pp. e1191–e1205, 3 2021. [Online]. Available: https://doi.org/10.1210/clinem/dgaa899

[28] T. Zheng, W. Ye, X. Wang, X. Li, J. Zhang, J. Little, L. Zhou, and L. Zhang, "A simple model to predict risk of gestational diabetes mellitus from 8 to 20 weeks of gestation in Chinese women," *BMC Pregnancy and Childbirth*, vol. 19, p. 252, 2019. [Online]. Available: https://doi.org/10.1186/s12884-019-2374-8

[29] F. Guo, S. Yang, Y. Zhang, X. Yang, C. Zhang, and J. Fan, "Nomogram for prediction of gestational diabetes mellitus in urban, Chinese, pregnant women," *BMC Pregnancy and Childbirth*, vol. 20, p. 43, 2020. [Online]. Available: https://doi.org/10.1186/s12884-019-2703-y

[30] Y. Pan, J. Hu, and S. Zhong, "The joint prediction model of pBMI and eFBG in predicting gestational diabetes mellitus," *Journal of International Medical Research*, vol. 48, p. 0300060519889199, 12 2019, doi: 10.1177/0300060519889199. [Online]. Available: https://doi.org/10.1177/0300060519889199

[31] B. Pintaudi, G. D. Vieste, F. Corrado, G. Lucisano, F. Pellegrini, L. Giunta, A. Nicolucci, R. D'Anna, and A. D. Benedetto, "Improvement of selective screening strategy for gestational diabetes through a more accurate definition of high-risk groups," *European Journal of Endocrinology*, vol. 170, pp. 87–93, 1 2014. [Online]. Available: https://doi.org/10.1530/EJE-13-0759

[32] N. S. Artzi, S. Shilo, E. Hadar, H. Rossman, S. Barbash-Hazan, A. Ben-Haroush, R. D. Balicer, B. Feldman, A. Wiznitzer, and E. Segal, "Prediction of gestational diabetes based on nationwide electronic health records," *Nature Medicine*, vol. 26, pp. 71–76, 2020. [Online]. Available: https://doi.org/10.1038/s41591-019-0724-8

[33] J. Wang, B. Lv, X. Chen, Y. Pan, K. Chen, Y. Zhang, Q. Li, L. Wei, and Y. Liu, "An early model to predict the risk of gestational diabetes mellitus in the absence of blood examination indexes: application in primary health care centres," *BMC Pregnancy and Childbirth*, vol. 21, p. 814, 2021. [Online]. Available: https://doi.org/10.1186/s12884-021-04295-2

[34] J.-N. Tong, Y.-X. Chen, X.-N. Guan, K. Liu, A.-Q. Yin, H.-F. Zhang, L.-L. Wu, and J.-M. Niu, "Association between the cut-off value of the first trimester fasting plasma glucose level and gestational diabetes mellitus: a retrospective study from southern China," *BMC Pregnancy and Childbirth*, vol. 22, p. 540, 2022. [Online]. Available: https://doi.org/10.1186/s12884-022-04874-x

[35] H. Liu, J. Li, J. Leng, H. Wang, J. Liu, W. Li, H. Liu, S. Wang, J. Ma, J. C. N. Chan, Z. Yu, G. Hu, C. Li, and X. Yang, "Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China," *Diabetes/Metabolism*

Research and Reviews, vol. 37, p. e3397, 7 2021, doi: 10.1002/dmrr.3397. [Online]. Available: https://doi.org/10.1002/dmrr.3397

[36] M. Kumar, L. Chen, K. Tan, L. T. Ang, C. Ho, G. Wong, S. E. Soh, K. H. Tan, J. K. Y. Chan, K. M. Godfrey, S. yng Chan, M. F. F. Chong, J. E. Connolly, Y. S. Chong, J. G. Eriksson, M. Feng, and N. Karnani, "Population-centric risk prediction modeling for gestational diabetes mellitus: A machine learning approach," *Diabetes Research and Clinical Practice*, vol. 185, 3 2022, doi: 10.1016/j.diabres.2022.109237. [Online]. Available: https://doi.org/10.1016/j.diabres.2022.109237

[37] Y. Wu, S. Ma, Y. Wang, F. Chen, F. Zhu, W. Sun, W. Shen, J. Zhang, and H. Chen, "A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women," *Diabetes Research and Clinical Practice*, vol. 179, 9 2021, doi: 10.1016/j.diabres.2021.109001. [Online]. Available: https://doi.org/10.1016/j.diabres.2021.109001

[38] L. Shen, D. S. Sahota, P. Chaemsaithong, W. T. Tse, M. Y. Chung, J. K. H. Ip, T. Y. Leung, and L. C. Y. Poon, "First Trimester Screening for Gestational Diabetes Mellitus with Maternal Factors and Biomarkers," *Fetal Diagnosis and Therapy*, vol. 49, pp. 256–264, 6 2022. [Online]. Available: https://doi.org/10.1159/000525384

[39] L. Li, Q. Zhu, Z. Wang, Y. Tao, H. Liu, F. Tang, S.-M. Liu, and Y. Zhang, "Establishment and validation of a predictive nomogram for gestational diabetes mellitus during early pregnancy term: A retrospective study," *Frontiers in Endocrinology*, vol. 14, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fendo.2023.1087994

[40] B. Kurt, B. Gürlek, S. Keskin, S. Özdemir, Özlem Karadeniz, İlknur Buçan Kırkbir, T. Kurt, S. Ünsal, C. Kart, N. Baki, and K. Turhan, "Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques," *Medical & Biological Engineering & Computing*, vol. 61, pp. 1649–1660, 2023. [Online]. Available: https://doi.org/10.1007/s11517-023-02800-7

[41] Y. Wei, A. He, C. Tang, H. Liu, L. Li, X. Yang, X. Wang, F. Shen, J. Liu, J. Li, and R. Li, "Risk prediction models of gestational diabetes mellitus before 16 gestational weeks," *BMC Pregnancy and Childbirth*, vol. 22, p. 889, 2022. [Online]. Available: https://doi.org/10.1186/s12884-022-05219-4

[42] S. Wu, L. Li, K.-L. Hu, S. Wang, R. Zhang, R. Chen, L. Liu, D. Wang, M. Pan, B. Zhu, Y. Wang, C. Yuan, and D. Zhang, "A Prediction Model of Gestational Diabetes Mellitus Based on OGTT in Early Pregnancy: A Prospective Cohort Study," *The Journal of Clinical Endocrinology & Metabolism*, vol. 108, pp. 1998–2006, 8 2023. [Online]. Available: https://doi.org/10.1210/clinem/dgad052

[43] B. S. Kang, S. U. Lee, S. Hong, S. K. Choi, J. E. Shin, J. H. Wie, Y. S. Jo, Y. H. Kim, K. Kil, Y. H. Chung, K. Jung, H. Hong, I. Y. Park, and H. S. Ko, "Prediction of gestational diabetes mellitus in asian women using machine learning algorithms," *Scientific Reports*, vol. 13, p. 13356, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-39680-8

[44] Y. Ye, Y. Xiong, Q. Zhou, J. Wu, X. Li, and X. Xiao, "Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: A retrospective cohort study," *Journal of Diabetes Research*, vol. 2020, p. 4168340, 2020. [Online]. Available: https://doi.org/10.1155/2020/4168340

[45] F. P. Hadlock, R. B. Harrist, and J. Martinez-Poyer, "In utero analysis of fetal growth: a sonographic weight standard." *Radiology*, vol. 181, pp. 129–133, 10 1991, doi: 10.1148/radiology.181.1.1887021. [Online]. Available: https://doi.org/10.1148/radiology.181.1.1887021

[46] K. H. Nicolaides, D. Wright, A. Syngelaki, A. Wright, and R. Akolekar, "Fetal Medicine Foundation fetal and neonatal population weight charts," *Ultrasound in Obstetrics & Gynecology*, vol. 52, pp. 44–51, 7 2018, https://doi.org/10.1002/uog.19073. [Online]. Available: https://doi.org/10.1002/uog.19073

[47] J. Stirnemann, J. Villar, L. J. Salomon, E. Ohuma, P. Ruyan, D. G. Altman, F. Nosten, R. Craik, S. Munim, L. C. Ismail, F. C. Barros, A. Lambert, S. Norris, M. Carvalho, Y. A. Jaffer, J. A. Noble, E. Bertino, M. G. Gravett, M. Purwar, C. G. Victora, R. Uauy, Z. Bhutta, S. Kennedy, and A. T. Papageorghiou, "International estimated fetal weight standards of the INTERGROWTH-21st Project," *Ultrasound in Obstetrics & Gynecology*, vol. 49, pp. 478–486, 4 2017, https://doi.org/10.1002/uog.17347. [Online]. Available: https://doi.org/10.1002/uog.17347

[48] G. M. B. Louis, J. Grewal, P. S. Albert, A. Sciscione, D. A. Wing, W. A. Grobman, R. B. Newman, R. Wapner, M. E. D'Alton, D. Skupski, M. P. Nageotte, A. C. Ranzini, J. Owen, E. K. Chien, S. Craigo, M. L. Hediger, S. Kim, C. Zhang, and K. L. Grantz, "Racial/ethnic standards for fetal growth: the NICHD Fetal Growth Studies," *American Journal of Obstetrics & Gynecology*, vol. 213, pp. 449.e1–449.e41, 10 2015, doi: 10.1016/j.ajog.2015.08.032. [Online]. Available: https://doi.org/10.1016/j.ajog.2015.08.032

[49] T. Kiserud, G. Piaggio, G. Carroli, M. Widmer, J. Carvalho, L. N. Jensen, D. Giordano, J. G. Cecatti, H. A. Aleem, S. A. Talegawkar, A. Benachi, A. Diemert, A. T. Kitoto, J. Thinkhamrop, P. Lumbiganon, A. Tabor, A. Kriplani, R. G. Perez, K. Hecher, M. A. Hanson, A. M. Gülmezoglu, and L. D. Platt, "The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight," *PLOS Medicine*, vol. 14, pp. e1 002 220–, 1 2017. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002220

[50] R. Savirón-Cornudella, L. M. Esteban, R. Aznar-Gimeno, P. D. Pérez, F. R. Pérez-López, B. Castán-Larraz, G. Sanz, and M. Tajada-Duaso, "Prediction of Large for Gestational Age by Ultrasound at 35 Weeks and Impact of Ultrasound-Delivery Interval: Comparison of 6 Standards," *Fetal Diagnosis and Therapy*, vol. 48, pp. 15–23, 9 2020. [Online]. Available: https://doi.org/10.1159/000510020

[51] J. R. Duncan, L. Odibo, E. A. Hoover, and A. O. Odibo, "Prediction of Large-for-Gestational-Age Neonates by Different Growth Standards," *Journal of Ultrasound in Medicine*, vol. 40, pp. 963–970, 5 2021, https://doi.org/10.1002/jum.15470. [Online]. Available: https://doi.org/10.1002/jum.15470

[52] J. Caradeux, E. Eixarch, E. Mazarico, T. R. Basuki, E. Gratacós, and F. Figueras, "Second- to Third-Trimester Longitudinal Growth Assessment for the Prediction of Largeness for Gestational Age and Macrosomia in an Unselected Population," *Fetal Diagnosis and Therapy*, vol. 43, pp. 284–290, 7 2017. [Online]. Available: https://doi.org/10.1159/000477460

[53] M. K. Kiefer, M. M. Finneran, C. A. Ware, P. Foy, S. F. Thung, S. G. Gabbe, M. B. Landon, W. A. Grobman, and K. K. Venkatesh, "Prediction of large-for-gestational-age infant by fetal growth charts and hemoglobin A1c level in pregnancy complicated by pregestational diabetes," *Ultrasound in Obstetrics & Gynecology*, vol. 60, pp. 751–758, 12 2022, https://doi.org/10.1002/uog.26071. [Online]. Available: https://doi.org/10.1002/uog.26071

[54] J. S. Erkamp, E. Voerman, E. A. P. Steegers, A. G. M. G. J. Mulders, I. K. M. Reiss, L. Duijts, V. W. V. Jaddoe, and R. Gaillard, "Second and third trimester fetal ultrasound population screening for risks of preterm birth and small-size and large-size for gestational age at birth: a population-based prospective cohort study," *BMC Medicine*, vol. 18, p. 63, 2020. [Online]. Available: https://doi.org/10.1186/s12916-020-01540-x

[55] U. Sovio, N. Goulding, N. McBride, E. Cook, F. Gaccioli, D. S. Charnock-Jones, D. A. Lawlor, and G. C. S. Smith, "A Maternal Serum Metabolite Ratio Predicts Large for Gestational Age Infants at Term: A Prospective Cohort Study," *The Journal of Clinical Endocrinology & Metabolism*, vol. 107, pp. e1588–e1597, 4 2022. [Online]. Available: https://doi.org/10.1210/clinem/dgab842

[56] K. S. Gibbons, A. M. Z. Chang, R. C. W. Ma, W. H. Tam, P. M. Catalano, D. A. Sacks, J. Lowe, and H. D. McIntyre, "Prediction of large-for-gestational age infants in relation to hyperglycemia in pregnancy - A comparison of statistical models," *Diabetes Research and Clinical Practice*, vol. 178, 8 2021, doi: 10.1016/j.diabres.2021.108975. [Online]. Available: https://doi.org/10.1016/j.diabres.2021.108975

[57] R. J. Wahab, V. W. V. Jaddoe, D. van Klaveren, M. J. Vermeulen, I. K. M. Reiss, E. A. P. Steegers, and R. Gaillard, "Preconception and early-pregnancy risk prediction for birth complications: development of prediction models within a population-based prospective cohort," *BMC Pregnancy and Childbirth*, vol. 22, p. 165, 2022. [Online]. Available: https://doi.org/10.1186/s12884-022-04497-2

[58] S. Kuhle, B. Maguire, H. Zhang, D. Hamilton, A. C. Allen, K. S. Joseph, and V. M. Allen, "Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study," *BMC Pregnancy and Childbirth*, vol. 18, p. 333, 2018. [Online]. Available: https://doi.org/10.1186/s12884-018-1971-2

[59] F. Monari, D. Menichini, L. S. Bascio, G. Grandi, F. Banchelli, I. Neri, R. D'Amico, and F. Facchinetti, "A first trimester prediction model for large for gestational age infants: a preliminary study," *BMC Pregnancy and Childbirth*, vol. 21, p. 654, 2021. [Online]. Available: https://doi.org/10.1186/s12884-021-04127-3

[60] Z. Wang, Y. Peng, S. Mao, L. Zhang, and Y. Guo, "The correlation between blood-lipid ratio in the first trimester and large-for-gestational-age infants," *Lipids in Health and Disease*, vol. 22, p. 18, 2023. [Online]. Available: https://doi.org/10.1186/s12944-023-01781-8

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[62] IBM, "What is logistic regression?" accessed: 2023-07-18. [Online]. Available: https://www.ibm.com/topics/logistic-regression

[63] J. Cramer, "The origins of logistic regression," *SSRN Electronic Journal*, 2003.

[64] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[65] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[66] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, pp. 1189–1232, 10 2001. [Online]. Available: https://doi.org/10.1214/aos/1013203451

[67] A. Ciampi, "Classification and discrimination: the recpam approach," in *Compstat*, R. Dutter and W. Grossmann, Eds. Heidelberg: Physica-Verlag HD, 1994, pp. 129–147.

[68] Anima Naturalis, "Lo que no conoces sobre los cerdos," accessed: 2023-07-18. [Online]. Available: https://www.animanaturalis.org/p/1127/lo-que-no-conoces-sobre-los-cerdos

[69] L. Rokach and O. Maimon, *Decision Trees*. Boston, MA: Springer US, 2005, pp. 165–192. [Online]. Available: https://doi.org/10.1007/0-387-25465-X_9

[70] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[71] C. Perez, G. Gonzalez, L. Medina, and F. Galdames, "Linear Versus Nonlinear Neural Modeling for 2-D Pattern Recognition," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, pp. 955–964, 11 2005.

[72] T. Szandala, "Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks," *CoRR*, vol. abs/2010.09458, 2020. [Online]. Available: https://arxiv.org/abs/2010.09458

[73] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl, "On Empirical Comparisons of Optimizers for Deep Learning," *CoRR*, vol. abs/1910.05446, 2019. [Online]. Available: http://arxiv.org/abs/1910.05446

[74] R.-Y. Sun, "Optimization for Deep Learning: An Overview," *Journal of the Operations Research Society of China*, vol. 8, pp. 249–294, 2020. [Online]. Available: https://doi.org/10.1007/s40305-020-00309-6

[75] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 1967.

[76] Canley, "Exam pass logistic regression," accessed: 2023-07-18. [Online]. Available: https://commons.wikimedia.org/wiki/File:Exam_pass_logistic_curve.svg

[77] IBM, "What is random forest?" accessed: 2023-07-18. [Online]. Available: https://www.ibm.com/topics/random-forest

[78] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," *University of California, Berkeley*, 01 2004.

[79] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3–42, 2006. [Online]. Available: https://doi.org/10.1007/s10994-006-6226-1

[80] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[81] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[82] J. E. Tapia, C. A. Perez, and K. W. Bowyer, "Gender Classification From the Same Iris Code Used for Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 1760–1770, 8 2016.

[83] D. S. Moore, G. P. McCabe, and B. A. Craig, *Introduction to the practice of statistics*, 6th ed. W.H. Freeman, 2009.

[84] IBM, "What is overfitting?" accessed: 2023-07-18. [Online]. Available: https://www.ibm.com/topics/overfitting

[85] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," *CoRR*, vol. abs/2108.00071, 2021. [Online]. Available: https://arxiv.org/abs/2108.00071

[86] F. Cismondi, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein, "Missing data in medical databases: Impute, delete or classify?" *Artificial Intelligence in Medicine*, vol. 58, pp. 63–72, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365713000055

[87] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, p. 60, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[88] D. A. Montecino, C. A. Perez, and K. W. Bowyer, "Two-Level Genetic Algorithm for Evolving Convolutional Neural Networks for Pattern Recognition," *IEEE Access*, vol. 9, pp. 126 856–126 872, 2021.

[89] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 2018, pp. 117–122.

[90] Google, "Classification: Accuracy," accessed: 2023-07-18. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy

[91] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice," *Frontiers in Public Health*, vol. 5, 2017. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2017.00307

[92] J. Opitz, "From Bias and Prevalence to Macro F1, Kappa, and MCC: A structured overview of metrics for multi-class evaluation," in *Heidelberg University*, 2022. [Online]. Available: https://www.cl.uni-heidelberg.de/~opitz/pdf/metric_overview.pdf

[93] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, rOC Analysis in Pattern Recognition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X

[94] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320396001422

[95] International Association of Diabetes and Pregnancy Study Groups Consensus Panel, "International Association of Diabetes and Pregnancy Study Groups Recommendations on the Diagnosis and Classification of Hyperglycemia in Pregnancy," *Diabetes Care*, vol. 33, pp. 676–682, 3 2010. [Online]. Available: https://doi.org/10.2337/dc09-1848

[96] D. R. Coustan, L. P. Lowe, B. E. Metzger, and A. R. Dyer, "The Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study: paving the way for new diagnostic criteria for gestational diabetes mellitus," *American Journal of Obstetrics & Gynecology*, vol. 202, pp. 654.e1–654.e6, 6 2010, doi: 10.1016/j.ajog.2010.04.006. [Online]. Available: https://doi.org/10.1016/j.ajog.2010.04.006

[97] World Health Organization, "Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy: A World Health Organization Guideline," *Diabetes Research and Clinical Practice*, vol. 103, pp. 341–363, 3 2014, doi: 10.1016/j.diabres.2013.10.012. [Online]. Available: https://doi.org/10.1016/j.diabres.2013.10.012

[98] M. Milad A, J. M. Novoa P, J. Fabres B, M. M. Samamé M, and C. Aspillaga M, "Recomendación sobre Curvas de Crecimiento Intrauterino," *Revista chilena de pediatría*, vol. 81, pp. 264–274, 2010.

[99] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[100] World Health Organization, "A healthy lifestyle - WHO recommendations," 5 2010, accessed: 2023-07-18. [Online]. Available: https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations

[101] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, Jun. 2004. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.69.066138

[102] B. C. Ross, "Mutual information between discrete and continuous data sets," *PLOS ONE*, vol. 9, no. 2, pp. 1–5, 02 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0087357

[103] G. Cubillos, M. Monckeberg, A. Plaza, M. Morgan, P. A. Estevez, M. Choolani, M. W. Kemp, S. E. Illanes, and C. A. Perez, "Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy," *BMC Pregnancy and Childbirth*, vol. 23, p. 469, 2023. [Online]. Available: https://doi.org/10.1186/s12884-023-05766-4

Anexos

A. Base de Datos

Tabla 6.1: Variables clínicas GDM.

Variable	Mujeres sin GDM (n=1382) Pro/IQR/Min/Max	Mujeres con GDM (n=229) Pro/IQR/Min/Max
Edad	27.64/(23-32)/14/48	31.11/(27-36)/17/46
Número de Fetos	1.01/(1-1)/1/2	1.02 (1-1)/1/2
Peso Materno (primer control) [kg]	71.62/(60-81)/39/160	81.77/(69-92)/46/137
Altura [m]	1.59/(1.55-1.63)/1.37/1.86	1.59/(1.55-1.63)/1.45/1.76
BMI (primer control) [kg/m^2]	28.18/(24.03-31.64) 15.43/59.49	32.17/(28.16-35.83) 18.20/57.02
Gravidez	1.24/(0-2)/0/10	1.69/(0-2)/0/7
Paridad	1.02/(0-2)/0/7	1.38/(0-2)/0/6
Abortos	0.22/(0-0)/0/9	0.32/(0-0)/0/6
Partos vaginales	0.79/(0-1)/0/7	1.03/(0-2)/0/6
Partos por cesárea	0.22/(0-0)/0/4	0.34/(0-1)/0/3
Mortinatas	0.01(0-0)/0/1	0.03/(0-0)/0/3
Glicemia ayuna primer trimestre (1TFG/FPG) [mg/dL]	77.22/(72-83)/50/116	87.12/(80-93)/62/124
OGTT Ayuna [mg/dL]*	74.28/(69-81)/45/91	95.48/(86-101)/61/220
OGTT 2 horas [mg/dL]*	99.39/(84-114)/43/152	142.87/(120-171)/66/252
	(%)	(%)
Consumo de Tabaco Pre-embarazo	7.74	11.79
Consumo de Alcohol Pre-embarazo	3.62	4.80
Consumo de Drogas Ilícitas Pre-embarazo	2.89	0.87
Enfermedades Cardíacas	0.65	0.44

Pro: Promedio. IQR: Rango intercuartil. Min: Mínimo. Max: Máximo.

*: Variables no utilizadas en la predicción.
Adquisición de los datos entre las semanas 4 y 20, excepto 1TFG (4-12), OGTTs (24-28).

Tabla 6.2: Continuación Tabla 6.1.

Variable	Mujeres sin GDM (n=1382) (%)	Mujeres con GDM (n=229) (%)
Enfermedades Biliares	1.01	2.18
Enfermedades Recurrentes del Tracto Urinario	2.32	4.80
Lupus Eritematoso Sistémico / Síndrome de Anticuerpos Antifosfolípidos	0.14	0.44
Desordenes Psiquiátricos	1.88	3.49
Desordenes Endocrinos	0.36	0.87
Enfermedades Renales Crónicas	0.36	0.00
Epilepsia	1.09	0.44
Enfermedad Inflamatoria Intestinal	0.07	0.44
Enfermedad Crónica Pulmonar	2.31	3.05
Desordenes Ginecológicos	3.40	7.42
Resistencia a la Insulina	2.46	6.99
Hipotiroidismo	4.05	9.17
Hipertensión crónica	4.70	12.66
Uso de drogas antihipertensivas	3.55	10.04

Adquisición de los datos entre las semanas 4 y 20.

Tabla 6.3: Variables clínicas LGA.

Variable	Fetos sin LGA (n=1514) Pro/IQR/Min/Max	Fetos con LGA (n=288) Pro/IQR/Min/Max
Edad	28.01/(23-32)/14/48	29.53/(25-34)/17/44
Peso Materno (primer control) [kg]	71.50/(60-81)/31/138	81.63/(69.38-93)/44/147
Peso Materno (último control) [kg]	82.18/(71.5-91)/45/150	93.85/(81.30-103)/59/165
Altura [m]	1.59/(1.55-1.63)/1.36/1.85	1.61/(1.56-1.65)/1.45/1.77
BMI (primer control) [kg/m^2]	28.29/(24.12-31.90) 14.35/53.13	31.65/(26.99-35.20) 19.04/57.42
BMI (último control) [kg/m^2]	32.52/(28.62-35.81) 18.97/56.46	36.39/(32.42-39.79) 23.18/64.45
Variación de peso	10.68/(7-14)/0.3/38	12.22/(7-16)/0.4/48
Gravidez	1.25/(0-2)/0/10	1.63/(1-2)/0/8
Paridad	1.00/(0-2)/0/7	1.29/(0-2)/0/6
Abortos	0.25/(0-0)/0/6	0.31/(0-0)/0/6
Partos vaginales	0.78/(0-1)/0/7	0.91/(0-2)/0/5
Partos por cesárea	0.22/(0-0)/0/3	0.39/(0-1)/0/4
Mortinatas	0.02/(0-0)/0/1	0.03/(0-0)/0/3
Glicemia ayuna primer trimestre (1TFG/FPG) [mg/dL]	78.75/(72-84)/51/201	81.57/(74-87)/53/140
OGTT Ayuna [mg/dL]	77.02/(70-83)/46/201	82.56/(73-89)/54/131
OGTT 2 horas [mg/dL]	105.70/(86-119.75)/43/343	122.31/(95.75-141)/45/249
	(%)	(%)
GDM	11.69	25.35
Consumo de Tabaco Pre-embarazo	7.46	11.46
Consumo de Alcohol Pre-embarazo	3.30	4.51
Consumo de Drogas Ilícitas Pre-embarazo	2.56	1.74
Enfermedades Cardíacas	0.79	0.35
Enfermedades Hepáticas	0.20	0.00
Enfermedades Biliares	1.19	1.74
Enfermedades Recurrentes del Tracto Urinario	2.97	2.08

Pro: Promedio. IQR: Rango intercuartil. Min: Mínimo. Max: Máximo.
Adquisición de los datos de preexistencias y primera visita entre las semanas 4 y 15, datos de la última visita, entre semana 24 y 41. 1TFG y GDM (4-12), OGTTs (24-28).

Tabla 6.4: Continuación Tabla 6.3.

Variable	Fetos sin LGA (n=1514) (%)	Fetos con LGA (n=288) (%)
Enfermedades Tromboembólicas	0.13	0.00
Lupus Eritematoso Sistémico/ Síndrome de Anticuerpos Antifosfolípidos	0.20	0.00
Desordenes Psiquiátricos	1.98	4.51
Desordenes Endocrinos	0.92	0.00
Enfermedades Renales Crónicas	0.33	0.00
Epilepsia	0.79	0.69
Enfermedad Inflamatoria Intestinal	0.07	0.00
Enfermedad Crónica Pulmonar	2.44	4.17
Desordenes Ginecológicos	5.28	4.51
Diabetes Mellitus Pre-Gestacional	0.59	1.39
Tratamiento de Diabetes Mellitus	*	*
Resistencia a la Insulina	3.57	5.90
Hipotiroidismo	4.82	5.55
Desorden hipertensivo en el embarazo	12.35	16.32
Hipertensión crónica	5.55	7.99
Uso de drogas antihipertensivas	4.16	4.51
Uso de aspirina	5.55	4.51
Preeclampsia	4.43	5.56
Preeclampsia Moderada	2.64	4.86
Preeclampsia Severa	1.85	0.35
Síndrome HELLP	0.33	0.00
Eclampsia	0.13	0.00
Hipertensión Gestacional	3.30	4.17
Polihidramnios	0.73	4.86
Oligohidramnios	3.83	1.04
Malformaciones	1.06	0.35
Presentación Fetal	*	*

IQR: Rango intercuartil.

*: Variable Categórica.

Adquisición de los datos de preexistencias entre las semanas 4 y 15, datos sobre tratamiento de la diabetes, entre semana 4 y 28. Desorden hipertensivo y uso de drogas antihipertensivas entre semana 4 y 41. Preeclampsias, Eclampsia, Síndrome HELLP, Hipertensión Gestacional y Malformaciones (22-41). Poli y Oligohidramnios entre semana 22 y 41. Presentación Fetal obtenido entre semanas 28 y 41.

Tabla 6.5: Continuación Tabla 6.3.

Variable	Fetos sin LGA (n=1514) Pro/IQR/Min/Max	Fetos con LGA (n=288) Pro/IQR/Min/Max
Peso Fetal Estimado (EFW)	2458.13/(2053-2945.75) 500/4748	3003.42/(2442.25-3633) 1040/5189
Hadlock <p3 %	1.39	0.35
Hadlock <p5 %	2.71	0.35
Hadlock <p10 %	4.36	0.35
Hadlock <p25 %	14.40	0.69
Hadlock >p75 %	26.29	75.00
Hadlock >p90 %	10.24	53.47
Hadlock >p95 %	5.09	43.05
SGA*	11.82	0.00
SGA Severo*	3.50	0.00
AGA*	88.18	0.00
Macrosomía*	1.12	60.76
Primera Visita (Semanas) *	9.99/(8-12)/4/15	9.88/(8-12)/4/15
Última Visita (Semanas) *	36.70/(36-38)/25/41	36.68/(36-38)/26/40

IQR: Rango intercuartil

*: Variables Estadísticas, no usadas para la predicción.

Percentiles Hadlock calculados usando la curva de Hadlock 1991 [45] con el EFW y la edad gestacional al momento del Ultrasonido.

SGA: Pequeño para la edad gestacional. AGA: Adecuado para la edad gestacional. Peso fetal estimado, Hadlocks obtenidos entre semanas 28 y 41.

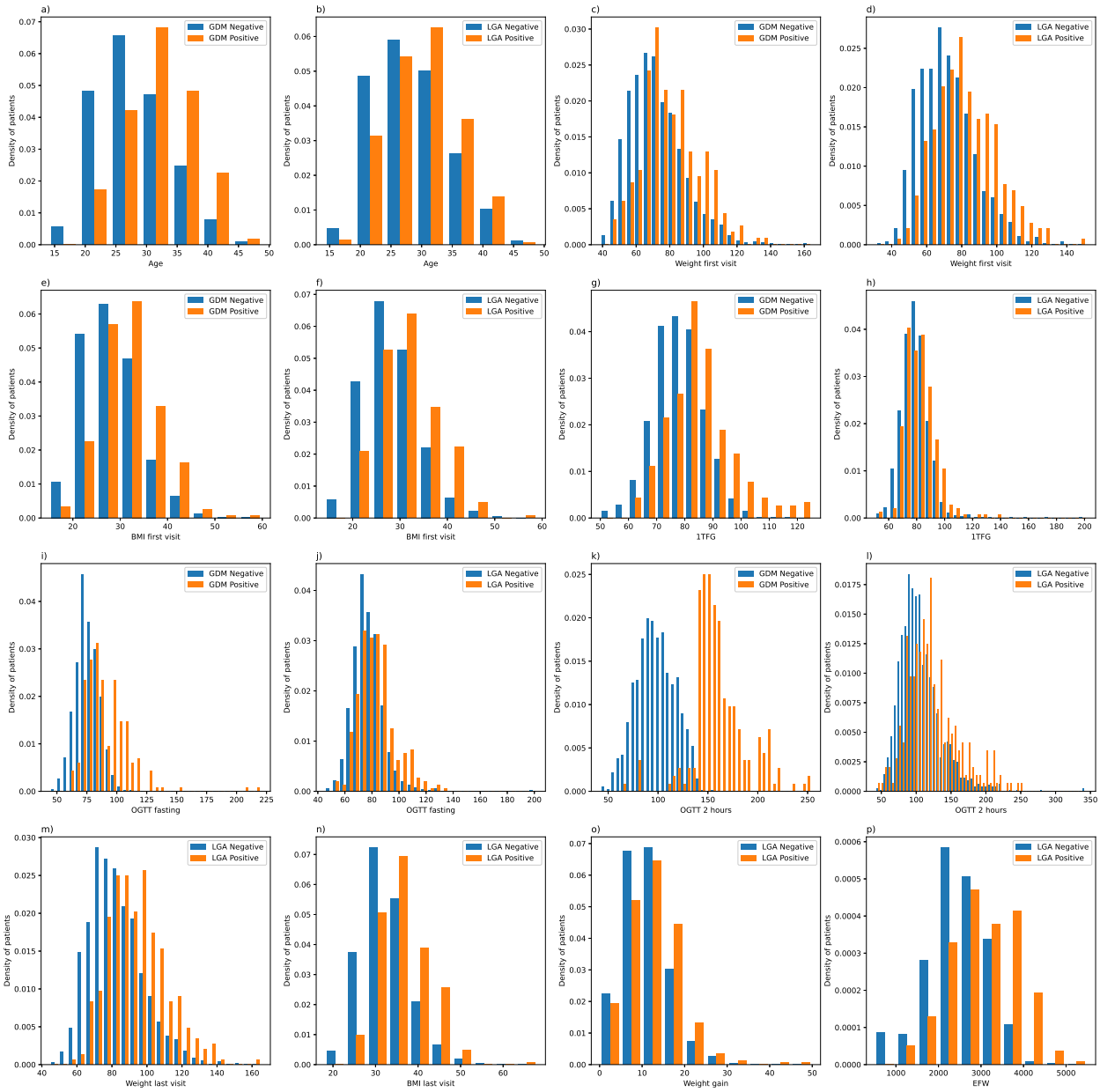

Figura 6.1: Distribuciones de diversas variables continuas para las condiciones de GDM y LGA.

B. Hiperparámetros

Tabla 6.6: Hiperparámetros usados para cada modelo

Hiperparámetro	Usado por	Rango/Valor [límite inferior, límite superior]
"var_smoothing"	Gaussian Naïve Bayes	[1e-10, 1e-7]
"alpha"	Bernoulli Naïve Bayes	[1e-10, 1]
"criterion"	DT, RF, ET, BRF	"gini", "entropy"
"max_depth"	DT, XGB	[1, 20]
"max_leaf_nodes"	Árbol de Decisión	[6, 384]
"splitter"	Árbol de Decisión	"best", "random"
"kernel"	SVM	"linear", "poly", "rbf", "sigmoid"
"degree"	SVM	[1, 3]
"C"	SVM, LR	[0.0001, 10]
"solver"	Perceptrón Multi-Capas	"sgd" , "adam"
"hidden_layer_sizes"	Perceptrón Multi-Capas	[8, 256]*
"activation"	Perceptrón Multi-Capas	"logistic", "tanh", "relu"
"learning_rate_init"	Perceptrón Multi-Capas	[0.001, 0.1]
"max_iter"	Perceptrón Multi-Capas	20000
"early_stopping	Perceptrón Multi-Capas	True, False
"learning_rate	Perceptrón Multi-Capas	"constant", "invscaling", "adaptative"
"algorithm"	KNN	"auto", "ball_tree", "kd_tree", "brute"
"leaf_size"	KNN	[1, 30]
"p"	KNN	[1, 4]
"n_neighbors"	KNN	[1, 25]
"solver"	Regresión Logística	"newton-cg", "lbfgs", "liblinear", "sag", "saga"
"n_estimators"	RF, ET, BRF, XGB, LGBM	[10, 2000]
"eta"	XGB	[0.001, 0.3]
"booster"	XGB	"gbtree", "gblinear", "dart"
"gamma"	XGB	[0, 1]
"boosting"	LGBM	"gbdt", "rf", "dart", "goss"
"learning_rate"	LGBM	[0.001, 0.1]

*: Número de capas, entre 0 y 10