Universidad de Valparaíso
CHILE
Facultad de Ciencias
Instituto de Estadística

Modelos explicables de Aprendizaje Automático para la predicción de complicaciones perinatales en mujeres embarazadas con diabetes

Trabajo de título presentado por:
Macarena Alejandra Giovanetti Riotti

Trabajo de titulación para optar al título de:
Ingeniero Estadístico

Profesor Guía:
Rodrigo Salas, Ph.D.

Profesores Co-Guía:
Fabian Pardo, Ph.D.
Ayleen Bertini

Valparaíso, Chile, 20 de Diciembre de 2021

Resumen

La diabetes es considerada uno de los desafíos sanitarios de mayor crecimiento del siglo XXI, esto debido al aumento exponencial de esta afección en los últimos 20 años. A continuación se realizará un estudio a 11.100 gestantes que padecen diabetes atendidas entre los años 2015 al 2021 en el hospital San Camilo de la región de Valparaíso, en la provincia de San Felipe. Esto con el objetivo de conocer el impacto de variables clínicas y sociodemográficas en la predicción de las complicaciones perinatales a través de modelos explicables de *machine learning*.

Del estudio realizado se obtuvo que las variables como el hipotiroidismo, la edad y la obesidad son los factores de riesgo que presentan una mayor asociación con las pacientes diabéticas. Por otro lado, el mejor clasificador para la predicción en la mayoría de las complicaciones resulto ser el bosque aleatorio. Por último, fue posible conocer mediante el método de SHAP el aumento de la probabilidad de presencia de una complicación a través de la interacción entre las variables regresoras.

Algunas Palabras

Primero que todo quiero dedicarle el presente proyecto a mi hijo Franco y a mi pequeño hermano Máximo, quienes me han sido el pilar fundamental para sacar adelante esta carrera universitaria y quienes me motivan a seguir avanzado.
Por otro lado, quisiera agradecer a mi familia, en primer lugar a mis padres Renzo y Marcela por quienes inicie este periodo universitario y los cuales siempre me han incentivado para llegar a ser profesional. A mis profesores, quienes semana tras semanas me han ayudado con el avance de este proyecto de titulación
Por último, me gustaría agradecer el apoyo que me han brindado en muchos aspectos a lo largo de estos 6 años de carrera, además de la comprensión y el cariño entregado, en primer lugar a mi pareja César, mis hermanas Yasminne, Giuliana, Esmeralda y a mi amiga Tamara.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Actualmente en Chile y según las últimas publicaciones del ministerio de salud, 1 de cada 10 chilenos padecen diabetes. Ésta enfermedad ha sido considerada una epidemia mundial de la cual los diagnosticados se han triplicado en los últimos años (Ministerio de salud [MINSAL], 2017). Por otro lado, existen pocos estudios nacionales asociados a la diabetes mellitus gestacional, uno de los más recientes publicado por Garmendia el cual fue realizado en el hospital Sotero del Río llego a la conclusión que la prevalencia de diabetes gestacional se triplico entre los años 2002 y 2015 en el país (Garmendia, 2019). Las proyecciones realizadas por la federación internacional de diabetes indican que estos números podrían seguir en un aumento considerable y para el año 2030 se espera que 578 millones padezcan esta enfermedad, por lo que estudiar los diversos tipos de diabetes (diabetes tipo I, diabetes tipo II, diabetes gestacional), con la aplicación de técnicas de aprendizaje automático puede tener implicaciones positivas para la reducción de costos, la efectividad de intervenciones como tratamientos y acciones preventivas.

La diabetes mellitus es una afección grave y de largo plazo (o "crónica") que ocurre cuando los niveles de glucosa en la sangre de una persona son altos, porque su cuerpo no puede producir o no utiliza de manera eficaz la insulina. La insulina es una hormona indispensable que se produce en el páncreas la cual permite que la glucosa del torrente circulatorio ingrese en las células del cuerpo, donde se convierte en energía. Además, es fundamental para el metabolismo de las proteínas y las grasas. La falta de insulina o la incapacidad de las células para responder a ella deriva en altos niveles de glucosa en sangre (hiperglucemia), el cual es un indicador clínico de la diabetes. Si no se controla el déficit de insulina a largo plazo, muchos de los órganos del cuerpo pueden resultar dañados, lo que derivaría en complicaciones de la salud incapacitantes y potencialmente mortales (Organización panamerica de salud [OMS] 2006,).

El valle del Aconcagua, perteneciente a la región de Valparaíso y situado entre las provincias de San Felipe y Los Andes, posee una población de 265,320 según el último CENSO del año 2017. El hospital más importante de la red de salud Aconcagua es el hospital San Camilo, el cual solo en el año 2020 presentó un total de 1,784 nacimientos. Este proyecto tiene como finalidad ser una ayuda para las pacientes embarazadas residentes en el valle, mediante el estudio de libro de nacimientos anonimizado del hospital San Camilo. La hipótesis en la presente tesis sugiere que las complicaciones perinatales en mujeres embarazadas con diabetes se pueden predecir mediante la aplicación de técnicas de aprendizaje automático con modelos explicables. Dado lo anterior, se tiene que evaluar los métodos explicables de aprendizaje automático para predecir complicaciones perinatales en gestantes con diabetes, los cuales permiten comprender las decisiones tomadas por

el modelo de aprendizaje automático.

Para llevar a cabo este procedimiento, se desprende: (i) determinar cuáles son las variables que representan mayor asociación con las mujeres embarazadas con diabetes, (ii) aplicar métodos de *machine learning* para la predicción de complicaciones perinatales en gestantes con diabetes, (iii) comprender la interacción de las variables clínicas y sociodemográficas que impactan en las complicaciones perinatales en mujeres embarazadas con diabetes mediante modelos explicativos de *machine learning*. En base a lo anterior, la metodología propuesta comprende aplicar técnicas de selección de atributos junto con la creación de *cluster* para agrupar variables con características similares que permitan reducir la dimensionalidad de la base de datos, y utilizar arboles de decisión y comparación de clasificadores (SVM, análisis de discriminante lineal, bosque aleatorio y MLP) para la predicción de las complicaciones perinatales. Para interpretar los resultados arrojados por el modelo, se aplicará el método de SHAP con el cual se busca conocer las características más importantes, el impacto y la interacción de variables.

Capítulo 2

Estado del arte

Una de las amenazas más serias para la salud mundial es la diabetes, esto debido a que quienes la padecen están en riesgo de desarrollar un conjunto de complicaciones graves y potencialmente mortales, lo cual conlleva una reducida calidad de vida y produce un estrés excesivo para la familia, además de una creciente necesidad de atención médica. Es por ello que si la diabetes y sus complicaciones no se tratan a tiempo y de manera adecuada los ingresos hospitalarios pueden ser frecuentes y la muerte prematura. A nivel mundial, esta enfermedad es una de las diez principales causas de fallecimiento, en la actualidad alrededor de 463 millones de adultos entre 20 y 79 la padecen, esto representa el $9,3\,\%$ de la población mundial en este grupo etario. Además se prevé que la cantidad total aumente a 578 millones ($10,2\,\%$) para el año 2030 y si la tendencia continúa 700 millones de adultos tendrán diabetes para el año 2045 (FID, 2019).

La diabetes es una afección definida principalmente por el nivel de hiperglucemia, es decir, por la cantidad excesiva de glucosa en la sangre lo cual a su vez da lugar a riesgo de daño microvascular (retinopatía, nefropatía y neuropatía). También se asocia con una menor esperanza de vida, morbilidad significativa debido a complicaciones microvasculares específicas relacionadas con la diabetes, aumento del riesgo de complicaciones macrovasculares (cardiopatía isquémica, accidente cerebrovascular y enfermedad vascular periférica), y disminución de la calidad de vida (OMS, 2006). Uno de los tipos de esta enfermedad es la diabetes mellitus gestacional (DMG) la cual es una complicación común durante el embarazo que afecta hasta al $15\,\%$ de las mujeres embarazadas en todo el mundo. La hiperglucemia no es, por sí sola, potencialmente mortal para las mujeres embarazadas, pero puede ser perjudicial para el feto y provocar complicaciones como: muerte fetal, parto prematuro, macrosomía, hiperinsulinemia fetal e hipoglucemia neonatal clínica (Chiefari et al., 2017).

Con el desarrollo continuo de la tecnología informática y la tecnología de redes, se ha mejorado enormemente la capacidad de utilizar diversos datos o conocimientos recopilados en diferentes orígenes y diferentes dispositivos, por lo que también se ha mejorado la capacidad de aplicar estos datos y conocimientos para resolver problemas y tomar decisiones médicas (Zhihan, 2020). El Aprendizaje Automático de Máquinas (*Machine Learning*, ML por sus siglas en inglés) es una técnica de entrenamiento de una máquina para reconocer patrones utilizando datos y un algoritmo. La precisión de la predicción de la máquina aumenta con los datos y la complejidad de las reglas introducidas en la máquina, la salud es una de las aplicaciones líderes en ML debido a que posee datos masivos y enormes. Este avance en la tecnología ayuda a los profesionales sanitarios a analizar los datos y les ayuda en la toma correcta de decisiones (Chowriappa et al., 2014). Los medios tradicionales de

detección tienen sus limitaciones y defectos, la adopción de herramientas de extracción de datos y la adaptación de la inteligencia de máquinas es producir un enfoque de diagnóstico predictivo que ofrece solución a la tarea, que los medios tradicionales no ofrecen además de resultados a bajo costo (Adimabua y Ekurume, 2021).

Se han realizado diversos estudios sobre la DMG con la aplicación de ML y otras técnicas utilizadas para ayudar al control e identificación temprana de esta enfermedad y sus complicaciones en distintos lugares del mundo. Uno de los últimos estudios realiza la comparación de distintos métodos de predicción de la DMG basándose en datos de registros médicos electrónicos de los cuales utiliza la selección de elementos de datos importantes, la reducción de dimensionalidad, la concentración del conjunto de datos y otras técnicas para el pre procesamiento de los datos, para posteriormente predecir la DMG mediante distintos modelos de clasificación y realizando una comparación entre las métricas de desempeño obtenidas de cada uno de los clasificadores (Liu et al., 2021).

Otro trabajo publicado en Marzo del presente año, al igual que el anteriormente mencionado, realiza un estudio con ML para la predicción temprana DMG pero esta vez en la población China, mediante un conjunto de datos de pacientes con registros médicos electrónicos obstétricos en 2017, específicamente del hospital de salud materno infantil Peace de la Universidad Jiao Tong de Shangai. Se utilizaron cuatro métodos de predicción: Regresión logística, k vecinos mas próximos (*K-Nearest Neighbors, kNN*), Máquina de vectores de soporte (*Support vector machine, SVM*), y red neuronal profunda, con la finalidad de comparar los resultados de los clasificador y observar cual tenía mejor comportamiento para predecir la DMG en el primer trimestre del embarazo (Wu, et al., 2021).

Otro de los estudios realizados compara el rendimiento de 8 métodos mas comunes de aprendizaje automático, entre ellos *XGBoost* , Árboles de decisión (*Decision Tree*) y bosque aleatorio *Random Forest* con regresiones logísticas tradicionales, utilizando ambos métodos para la predicción de la DMG mediante datos clínicos de un hospital terciario en China, Los modelos se compararon en función de las métricas de discriminación y calibración (Ye, et al., 2020).

Por otro lado, la creciente evidencia indica que la contaminación del aire es capaz de alterar el sistema inmunológico y por lo tanto, podría estar asociada con la aparición de diabetes tipo 1, es por ello que se realizó un estudio acerca de posibles vínculos de la diabetes tipo 1 con la exposición ambiental en la población del sur de Israel, caracterizada por un clima cálido, seco y frecuentes tormentas de polvo. Este se llevo a cabo mediante un estudio de casos y controles anidado dentro de la población de la cohorte de recién nacidos entregados y tratados en el Centro Médico de la Universidad de Soroka (CMUS) entre los años 2001 y 2018. Para la evaluación de resultados se utilizó un modelo de regresión logística condicional para evaluar la asociación entre la diabetes tipo 1 y la probabilidad de exposición a factores ambientales (Taha, et al., 2021).

La tasa de mortalidad materna es mayor en los partos por cesárea que en los partos naturales, además arrastra varias otras complicaciones debido a la intervención quirúrgica en el parto. Es por ello que científicos de la India realizaron un análisis para determinar si existe asociación entre el parto por cesárea y DMG, utilizando cuatro técnicas de ML para clasificación. Los datos fueron extraídos de la encuesta de evaluación de riesgos del embarazo, la cual fue recopilada de los centros de control y prevención de la India (Siddegowda & Puttabuddi, 2020).

Entre las publicaciones realizadas recientemente en Chile, es posible destacar un estudio de académicos de la Universidad de Concepción los cuales mediante técnicas de inteligencia artificial (IA), descubrieron la relación existente entre la alteración de las hormonas tiroideas y el desarrollo de diabetes gestacional en el primer semestre del embarazo, información que permitiría prevenir de manera efectiva y oportuna el desarrollo de la enfermedad. Se analizaron distintas características de la hormona tiroidea en conjunto con otras variables, medidas en cohortes de poblaciones de embarazada sanas y otras afectadas por diabetes gestacional diagnosticadas con el método tradicional de la PTGO (prueba de tolerancia de la glucosa oral). Se reclutaron 39 mujeres embarazadas con DMG en la ciudad de Concepción (Chile), a las que se les dio seguimiento desde las 12 a las 28 semanas de gestación, posteriormente se analizaron 29 variables mediante ML utilizando Análisis de componentes principales (ACP) para el reconocimiento de patrones (Araya, et al., 2021).

Capítulo 3

Machine Learning

3.1. ¿Qué es el aprendizaje automático?

El aprendizaje automático es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a los ordenadores aprender sin ser programados de manera explícita (Samuel, 1959). Esta disciplina ha mostrado enormes mejoras en los últimos 20 años, esto se ve reflejado en la implementación de métodos de *machine learning* (ML) en diversas áreas como por ejemplo en la medicina, la ecología, las finanzas y muchas otras (Kononenko & Kukar, 2007).

Se puede definir ampliamente como métodos computacionales que utilizan la experiencia para mejorar el rendimiento o hacer predicciones más precisas. En esta situación, la experiencia se refiere a la información pasada disponible para el investigador, que normalmente toma la forma de datos electrónicos recopilados y puestos a disposición para su análisis. Éstos datos pueden contener diversos tipos de información, sin embargo, en todos los casos su calidad y tamaño son cruciales para el éxito de las predicciones obtenidas. Las garantías del aprendizaje teórico de un algoritmo dependerán de la complejidad del concepto, las clases consideradas y el tamaño de la muestra de formación. En otras palabras el éxito de un algoritmo depende de los datos utilizados (Samuel, 1959).

El aprendizaje de las máquinas está intrínsecamente relacionado con el análisis de datos y la estadística. De manera más general, los métodos de ML son basados en datos que combinan conceptos fundamentales en de la ciencia informática con ideas de estadística, probabilidad y optimización (Mohri et al, 2018). Su principio básico es el modelado automático de procesos que han generado los datos recopilados. Aprender de los datos da como resultado reglas, funciones, relaciones, sistemas de ecuaciones, distribuciones de probabilidad, y otros conocimientos tales como reglas de decisión, árboles de decisión, regresiones, redes bayesianas, redes neuronales, etc. Los modelos explican los datos y pueden usarse para respaldar decisiones relacionadas con el mismo proceso subyacente (Kononenko y Kukar, 2007).

3.2. Minería de datos

La disciplina denominada Minería de Datos estudia métodos y algoritmos que permiten la extracción automática de información sintetizada que permite caracterizar las relaciones escondidas en la gran cantidad de datos. También se requiere que la información obtenida posea capacidad predictiva, facilitando así el análisis de los datos de forma eficiente. Bajo la denominación de "minería de datos" se han agrupado recientemente diversas técnicas estadísticas y del aprendizaje automático (Inteligencia Artificial) enfocadas, principalmente, a la visualización, análisis, y modelización de información de bases de datos masivas (Beltrán, s.f).

3.2.1. Taxonomía de los modelos de minería de datos

Es útil distinguir entre dos tipos principales de modelos en la minería de datos, uno orientado a la verificación (el sistema verifica la hipótesis del usuario) y otro orientado al descubrimiento (el sistema encuentra nuevas reglas y patrones de forma autónoma), cabe destacar que cada tipo tiene su propia metodología.

Los métodos de descubrimiento que identifican automáticamente patrones en los datos incluyen tanto métodos de predicción como de descripción. Los métodos de descripción se centran en comprender el funcionamiento de los datos subyacentes, mientras que los métodos orientados a la predicción pretenden construir un modelo de comportamiento para obtener muestras nuevas y para predecir los valores de una o más variables relacionadas con la muestra. Mientras que la mayoría de las técnicas orientadas al descubrimiento utilizan el aprendizaje inductivo como se ha comentado anteriormente, los métodos de verificación evalúan una hipótesis propuesta por una fuente externa, como un experto. Estas técnicas incluyen los métodos más comunes de la estadística tradicional, como la prueba de bondad de ajuste, la prueba t de medias y el análisis de la varianza. La principal diferencia se centra en que la minería de datos tiene como uno de sus objetivos la identificación de modelos mientras que los métodos estadísticos suelen centrarse en la estimación del modelo (Rokach & Maimon, 2014).

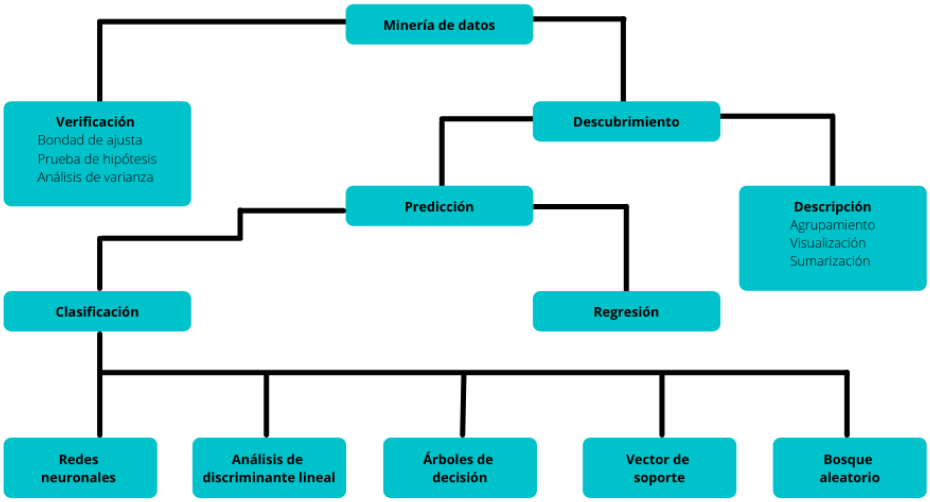

Figura 3.1: Taxonomía Minería de datos (Rokach & Maimon, 2014).

Capítulo 4

Interpretabilidad y Explicabilidad en Machine Learning

Los conceptos de explicabilidad e interpretabilidad aparecen como respuesta a la búsqueda de la comprensión de los modelos de inteligencia artificial, donde ambos términos se encuentran estrechamente relacionados pero que sin embargo, tienen significados diferentes.

Se define la interpretabilidad como la habilidad de explicar en términos comprensibles para un ser humano. En ML se refiere al grado en que se puede observar una causa y su efecto dentro de un sistema, es decir, en qué medida se puede explicar el resultado de un modelo dado un cambio en los datos de entrada. Con eso en mente, decimos que un modelo es interpretable si es capaz de ser entendido por los humanos por sí solo (Doshi-Velez & Kim, 2017). ,

Los modelos de ML pueden ser asombrosamente buenos para realizar predicciones con altos valores de precisión, pero que a menudo no pueden dar explicaciones para sus pronósticos en términos que los humanos puedan entender fácilmente. Las características de las que extraen conclusiones pueden ser tan numerosas y sus cálculos tan complejos, que los investigadores pueden encontrar imposible establecer exactamente por qué un algoritmo está tomando una decisión. En algunos casos, no es de vital importancia conocer porque el algoritmo tomó una decisión sino mas bien, es suficiente saber que el rendimiento predictivo en un conjunto de datos de prueba fue bueno. Sin embargo, en otros casos conocer que es lo que realizó el modelo para llegar a esa conclusión puede ayudar a aprender más sobre el problema, los datos y la razón por la que un modelo puede fallar. Es posible que algunos modelos no requieran explicaciones porque se utilizan en un entorno de bajo riesgo, lo que significa que un error no tendrá consecuencias graves o el método ya se ha estudiado y evaluado exhaustivamente. Sin embargo, en muchas situaciones el modelo también debe explicar cómo llegó a la predicción esto ya que una predicción correcta solo resuelve parcialmente el problema original (Velez et al., 2017).

La explicabilidad se refiere al grado en el que el comportamiento del modelo se puede explicar en términos humanos, considerando tanto el resultado como todo el proceso de la toma de decisión. Esto incluye explicar los datos utilizados y cómo se obtienen los resultados, cómo se producen los *outputs* a través de los *inputs* y la intención con la que el sistema afecta a las partes involucradas (Gandhi, 2019). Por tanto el concepto de explicabilidad podría entenderse como un concepto más amplio, con un objetivo más ambicioso que la interpretabilidad (Vitoriano, 2007).

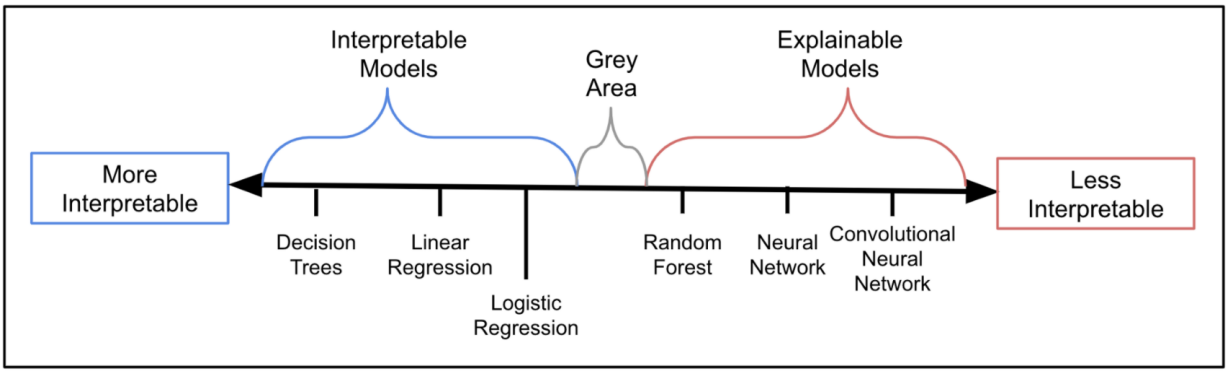

Figura 4.1: Espectro interpretable (Sullivan, 2017).

4.1. Modelos Interpretables

4.1.1. Regresión

Regresión Lineal:

"La técnica de regresión proporciona los medios legítimos a través de los cuales pueden establecerse asociaciones entre las variables de interés en las cuales la relación usual no es causal" (G. Canavos,1988).

Un modelo de regresión lineal predice la variable objetivo como una suma ponderada de las entradas de características, la linealidad de la relación aprendida facilita la interpretación. Estos modelos han sido utilizados durante mucho tiempo por estadísticos, informáticos y otras personas que abordan problemas cuantitativos.

Los modelos lineales se pueden utilizar para modelar la dependencia de una variable objetivo de regresión y de algunas características. Las relaciones son lineales y se pueden escribir para una sola instancia i de la siguiente forma:

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_{pag} X_{pag} + \epsilon$$

El resultado previsto de una instancia es una suma ponderada de sus p características. Los parámetros (β_j) representan las ponderaciones o coeficientes de las características aprendidas. El primer peso de la suma (β_0) se llama intercepto y no se multiplica por una característica. El error aleatorio (ϵ) es la diferencia entre la predicción y el resultado real. Los errores siguen generalmente una distribución gaussiana, lo que significa que se cometen errores tanto con signo negativo como positivo (Molnar, 2018).

Regresión Logística:

Se tienen k observaciones independientes, $y_1, ..., y_k$, la variable aleatoria dependiente denotada Y_i se asume que posee una distribución binomial (Rodríguez, 2007):

$$Y_i \sim B(n_i, \pi_i) \tag{4.1}$$

donde n_i es el denominador binomial y π_i es la probabilidad. Se asume además que el *logit* de la probabilidad subyacente i es una función lineal de los predictores dada por (Rodríguez, 2007):

$$logit(\pi_i) = x_i'\beta \tag{4.2}$$

donde x_i es un vector de covariables y β es un vector de coeficientes de regresión. Esto define la estructura sistemática del modelo. Las ecuaciones definidas anteriormente son una generalización de un modelo lineal con respuesta binomial y una función de enlace *logit* (Rodríguez, 2007).

El coeficiente de regresión β puede interpretarse de forma similar a un modelo lineal, teniendo en cuenta que el lado izquierdo es una función *logit* en lugar de una media. Por lo tanto, β_j representa el cambio en el logit de la probabilidad asociada con un cambio de unidad en el j-ésimo predictor que mantiene constantes todos los demás predictores.

Si bien, interpretar resultados en escala logit es complejo, tiene la ventaja que es posible facilitar estos resultados mediante la interpretación de los odds los cuales se encuentran aplicando exponencial en ambos lados de la ecuación (4.2) se encuentra que el Odds para la unidad i-ésima estpa dado dado por:

$$\frac{\pi_i}{1-\pi_i} = \exp x_i'\beta \tag{4.3}$$

Finalmente, esta expresión (4.3) define el modelo multiplicativo para los odds.

4.1.2. Árboles de decisiones

Los árboles de decisión desarrollados por Breiman, Freidman. Olshen y Stone en 1984, son una técnica de minería de datos, de aprendizaje inductivo supervisado no paramétrico, a partir de observaciones y construcciones lógicas. Durante el aprendizaje inductivo se adquieren conocimientos que representan un árbol de decisión en donde puede abordar soluciones a problemas ya sea de predicción, segmentación o de clasificación (Medina & Ñique , 2017).

Para entender como se forma un árbol de decisión debemos tener en cuenta que está representado por un conjunto de nodos, hojas y ramas que poseen sus propias funciones (las cuales serán vistas más adelante), en donde, el nodo inicial o principal es por el cual se inicia el proceso de clasificación representando toda la población o muestra. Los nodos siguientes que representan los extremos de la cadena se le conocen como nodos hojas. Las ramas nos muestra los distintos caminos que se pueden tomar mediante la decisión o bien un evento aleatorio (Medina & Ñique , 2017).

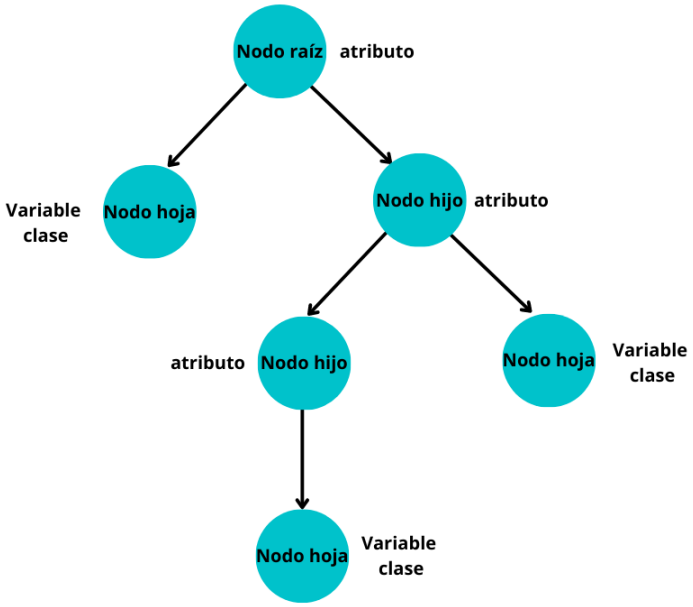

Figura 4.2: Estructura principal árboles de decisión (Medina & Ñique , 2017).

Árbol de decisión para regresión

Los árboles de regresión son el subtipo de árboles de predicción que se aplica cuando la variable respuesta es continua. En términos generales, en el entrenamiento de un árbol de regresión, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. Cuando se quiere predecir una nueva observación, se recorre el árbol acorde al valor de sus predictores hasta alcanzar uno de los nodos terminales. La predicción del árbol es la media de la variable respuesta de las observaciones de entrenamiento que están en ese mismo nodo terminal. La idea básica es combinar árboles de decisión y regresión lineal para pronosticar atributo de destino numérico basado en un conjunto de atributos de entrada. Estos métodos realizan la inducción mediante un eficiente algoritmo de partición recursiva (Amat, 2020). La elección de la mejor división en cada nodo del árbol suele estar guiada por un criterio de error de mínimos cuadrados (Rokach & Maimon, 2014).

El entrenamiento del árbol se divide en dos etapas:

Etapa 1. División sucesiva del espacio de los predictores generando regiones no solapantes (nodos terminales) $R_1, R_2, R_3, ..., R_j$. Aunque, desde el punto de vista teórico las regiones podrían tener cualquier forma, si se limitan a regiones rectangulares (de múltiples dimensiones), se simplifica en gran medida el proceso de construcción y se facilita la interpretación.

Etapa 2. Predicción de la variable respuesta en cada región.

En los árboles de regresión, el criterio empleado con más frecuencia para identificar las divisiones es el *Residual Sum of Squares* (Suma residual de cuadrados, RSS). El objetivo es encontrar las J regiones $(R_1, ..., R_j)$ que minimizan el RSS total:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{4.4}$$

donde $\widehat{y}_{R_j}$ es la media de la variable respuesta en el reglón R_j. En otras palabras, se busca una distribución de regiones tal que, la sumatoria de las desviaciones al cuadrado entre las observaciones y la media de la región a la que pertenecen sea lo menor posible (Amat, 2020).

Árbol de decisión para clasificación.

Los árboles se utilizan para clasificar un objeto o una instancia en un conjunto predefinido de clases en función de su valores de atributos. En otras palabras, se toman casos en grupos o pronostican valores de una variable dependiente (criterio) basada en valores independientes (predictoras). Los árboles de clasificación son frecuentemente utilizados en campos aplicados como finanzas, marketing, ingeniería y medicina (Rokach & Maimon, 2014).

El Algoritmo no es paramétrico y crea árboles binarios a partir de datos descritos por características tanto continuas como discretas. Para características continuas, se consideran todas las posibles separaciones binarias en los intervalos $(-\infty, a]$ y (a, ∞). Por otro lado, para variables discretas el análisis se refiere a todas las posibles divisiones del conjunto de símbolos en dos subconjuntos disjuntos y complementarios (Grabczewski, 2014).

Para construir un árbol de clasificación y dado que la variable respuesta es cualitativa, no se puede emplear la suma residual como criterio de selección de las divisiones óptimas. Sin embargo, existen varias alternativas, todas ellas con el objetivo de encontrar nodos lo más puros/homogéneos posible. El criterio de división aplicado para medir la calidad de las condiciones de prueba es quizás el elemento que tiene el impacto más significativo en la efectividad y expresividad de un clasificador. Estos criterios pueden medir la impureza de la partición, estimar algún otro valor discriminante o evaluar algún valor de costo. Además, considerando la presencia de incertidumbre y ambigüedad en la información, se debe incluir una medida blanda en un criterio de división (River et al., 2021). Entre las funciones de criterio división más empleadas es posible mencionar:

- Índice de Gini:

  Cuantifica la varianza total en el conjunto de las K clases del nodo m, es decir, mide la pureza del nodo a través de la función:

  $$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{4.5}$$

  cuando $\hat{p}_{mk}$ es cercano a 0 o a 1 (el nodo contiene en su mayoría observaciones de una sola clase), el término $\hat{p}_{mk}(1 - \hat{p}_{mk})$ es muy pequeño. Como consecuencia, cuanto mayor sea la pureza del nodo, menor es el valor del índice Gini G (Amat, 2020).

- Ganancia de Información:

  La ganancia de información (*Information Gain* en inglés) o también llamada entropía cruzada es una propiedad matemática que mide la aleatoriedad y la incertidumbre sobre el resultado de una variable aleatoria. Cuanto menor sea la entropía, más predecible será el resultado de la variable aleatoria. Esta entropía usada convencionalmente viene dada por (Gonen et al., 2020):

$$H = -\sum_{i=1}^{n} p_i log(p_i), \tag{4.6}$$

donde p_i es la función de masa de probabilidad del i-ésimo resultado de la variable de clase y K se considera como una constante que normaliza las unidades de información de acuerdo con la base logarítmica utilizada (Gonen et al., 2020).

Los árboles de decisiones son propensos a caer en el sobreajuste (*overffiting*), entre los métodos utilizados para evitar este problema está el de la validación cruzada la cual tiene como finalidad (i) averiguar hasta qué complejidad debe entrenar el modelo o (ii) ajustar los parámetros del modelo. En la validación cruzada, dividimos el conjunto de datos D en dos particiones, es decir, conjunto de entrenamiento denotado por T y conjunto de prueba denotado por R donde la unión de estos dos subconjuntos es el conjunto de datos completo y la intersección de ellos es el conjunto vacío. La T se utiliza para entrenar el modelo. Después de que el modelo es entrenado, la R se utiliza para probar el rendimiento del modelo. Uno de los métodos más conocido de validación cruzada es el de K-Fold (Ghojogh & Crowley, 2019).

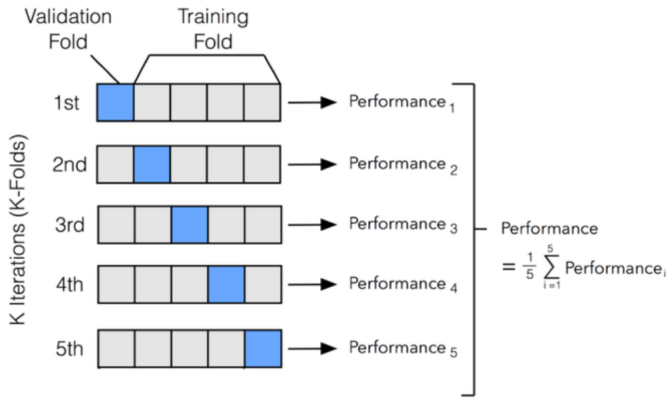

Figura 4.3: Validación cruzada de tipo K-fold (Shen, 2020).

La validación cruzada de K-fold incluye $k-1$ grupos los cuales se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración (como se observa en la figura 4.3). El proceso genera k estimaciones del error cuyo promedio se emplea como estimación final (Ghojogh & Crowley, 2019).

4.2. Modelos Explicables

4.2.1. Método de SHAP

Los valores de Shapley o método de SHAP (Explicaciones aditivas de Shapley) es un método para explicar predicciones individuales propuesto originalmente como un concepto de compensación de la teoría de juegos cooperativos (Shapley, 1959). El uso de la teoría de juegos establece tres formas para realizar el modelamiento de un escenario real: *extensiva*, *estratégica* y de *coalición*. Las dos primeras solo son aplicables a juegos no cooperativos, en donde prima el interés en el beneficio

propio, sin importar el resultado de los demás jugadores. Por otro lado, la tercera forma (*coalición*), es aplicable exclusivamente a juegos de tipo cooperativo, el cual corresponde a un juego en el cual dos o más jugadores no compiten entre sí, sino que por el contrario, trabajan de manera conjunta para conseguir el mismo objetivo y por lo tanto, ganan o pierden como un grupo. Adicionalmente, el hecho de trabajar cooperativamente o de establecer coaliciones entre dos o más jugadores, aumenta la probabilidad de obtener una ganacia superior en comparación a lo que se obtiene de forma individual (Vesga et al. 2015).

El objetivo de SHAP es explicar la predicción de una instancia x calculando la contribución de cada característica a la predicción. Los valores de las características de una instancia de datos actúan como jugadores en una coalición, por lo tanto los valores de Shapley nos dicen cómo distribuir equitativamente la predicción entre las funciones. Para llevar a cabo este procedimiento y ahondando en la matemática del modelo, una innovación que SHAP aporta es que la explicación del valor de Shapley se representa como un método de atribución de características aditivas en un modelo lineal (Molnar, 2018):

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j, \tag{4.7}$$

donde g es la explicación del modelo, $z' \in \{0,1\}^M$ es la coalición del vector, M es el máximo tamaño de la coalición y $\phi_j \in \mathbb{R}$ es la característica atribuida para j (Molnar, 2018).

De manera más general, los valores de Shapley cumplen una serie de propiedades útiles que permiten comprender mejor cómo el modelo usa sus características para brindar una respuesta confiable en un proceso complejo de toma de decisiones (Giucidi & Raffineti, 2021).

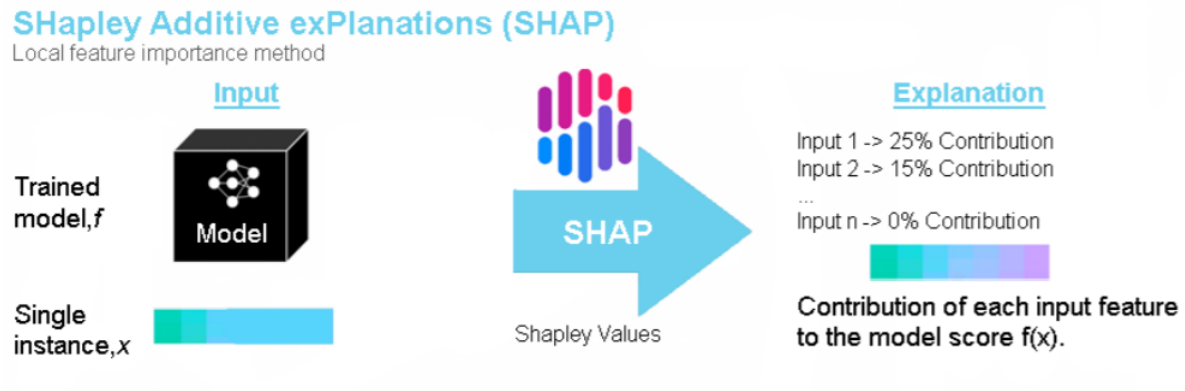

Figura 4.4: Método de SHAP (Knapic et al., 2021).

Capítulo 5

Clasificadores de machine learning

La clasificación es una forma de aprendizaje supervisado, que asigna datos de entrada a datos de salida basándose en muchos pares (entrada-salida) de ejemplos determinados durante una fase de entrenamiento.

Usando la clasificación, las características relacionadas con un conjunto de observaciones de ejemplo se pueden usar para entrenar una función de decisión que genera asignaciones de clase comúnmente llamadas *etiquetas* con una precisión determinada. Estas características pueden ser muy diversas en función de cualquier objeto, desde datos de neuroimágenes funcionales hasta publicaciones de redes sociales. Una vez que este clasificador se ha creado en función de estas características, puede adjuntar automáticamente etiquetas de clase a observaciones nuevas y no vistas utilizando los patrones establecidos anteriormente (Pisner et al., 2020).

5.1. Support Vector Machine

SVM pertenece a técnicas de aprendizaje supervisado ya que uno debe realizar un aprendizaje sin distribución porque no hay información sobre las funciones subyacentes de probabilidad conjunta, la única información disponible es un conjunto de datos de entrenamiento $D = (x_i, y_i) \in X \times Y, i = 1, ..., l$, donde l representa el número de entrenamientos pares de datos y, por lo tanto, es igual al tamaño del conjunto de datos de entrenamiento D. A veces y_i es denotada d_i donde d representa la variable objetivo (Kacprzyk, 2005).

Los SVM son llamados los modelos "no paramétricos", esto no en el sentido que su modelo no contiene parámetros en lo absoluto sino que a diferencia de la estadística clásica, el parámetro no está pre-definido y su número depende de los datos de entrenamiento usados. En otras palabras, los parámetros son impulsados por los datos de tal manera que coincidan la capacidad del modelo con la complejidad de los datos. Este es un paradigma básico para la minimización del riesgo estructural impulsados por Vapnik y Chervonenkis para llevar al aprendizaje de un nuevo algoritmo (Kacprzyk, 2005).

Según Janusz Kacprzyk, hay dos enfoques constructivos principales para la creación de un modelo con buena propiedad de generalización:

1. Eligiendo una estructura apropiada del modelo (orden de polinomios, número de neuronas HL, número de reglas en el modelo de lógica difusa) y, manteniendo el error de estimación (intervalo

de confianza o también conocido como varianza del modelo) de ésta manera, se minimiza el error de entrenamiento (es decir, riesgo empírico).

2. mantener el valor del error de entrenamiento fijo (igual a cero o igual a algún nivel aceptable) y minimizar el intervalo de confianza.

"*Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión mayor (si los puntos de entrada están en R^2 entonces son mapeados por la SVM a R^3) y encuentra un hiperplano que los separe y maximice el margen m entre las clases en este espacio*" (Betancourt, 2005). Esto es posible apreciarlo mas claramente en la siguiente imagen:

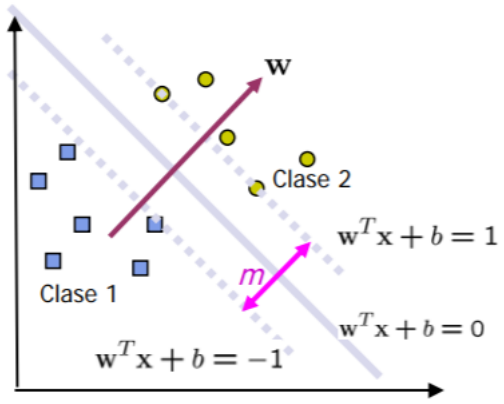

Figura 5.1: Frontera de desición SVM (Betancourt, 2005).

El objetivo de SVM es producir un modelo (basado en los datos de entrenamiento) que prediga los valores objetivos de los datos de prueba dando solo los atributos de datos de prueba. Dado un conjunto de entrenamiento para pares de etiqueta de instancia, las máquinas de vectores de soporte requiere la solución del siguiente problema de optimización (Boser et al., 1995):

$$\min_{(\omega,b,\epsilon)} \frac{1}{2} w^T w + C\Sigma_{i=1}^{1} \epsilon_i^2 \tag{5.1}$$

Sujeto a:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \epsilon_i, \epsilon_i \geq 0. \tag{5.2}$$

SVM encuentra un hiperplano de separación lineal con el margen máximo en este espacio dimensional superior. $C > 0$ es el parámetro de penalización del término de error. Adémas, $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ es llamado el kernel de la función. Se pueden encontrar 4 kernels básicos (Awad & Khanna, 2015).:

- 1. Lineal: $K(x_i, x_j) = x_i^T x_j$.
- 2. Polinomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$.
- 3. Función de base radial (radial basis function(RBF)): $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \gamma > 0$
- 4. Sigmoide: $K(x_i, x_j) = tanh(\gamma x_i^T x_j + r)$,

donde γ, r y d son los parámetros del kernel (Awad & Khanna, 2015).

5.2. Bosque Aleatorio

Breiman (2001) propuso una definición general de bosques de árboles de decisión como una colección de clasificadores de árboles construidos con respecto a vectores aleatorios. Para k árboles, un vector aleatorio Θ_k se genera de forma independiente de los vectores aleatorios pasados $\Theta_1, ..., \Theta_{k-1}$ pero con la misma distribución. Luego se siembra un árbol utilizando el conjunto de entrenamiento y Θ_k, lo que da como resultado un clasificador $h(x, \Theta_k)$, donde x es un vector de entrada (Breiman, 2001).
Después de que se genera una gran cantidad de árboles, se selecciona la clase más popular. Llamamos estos procedimientos bosques aleatorios (Breiman, 2001).

Definición: *un bosque aleatorio es un clasificador que consiste en una colección de árboles estructurados* $\{h(\boldsymbol{x}, \Theta_k), k = 1, ...\}$ *donde* $\{\Theta_k\}$ *son vectores aleatorios independientes e idénticamente distribuidos (i.i.d.) y cada árbol emite un voto unitario para la clase más popular en la entrada* $\boldsymbol{x}$ (Breiman, 2001).

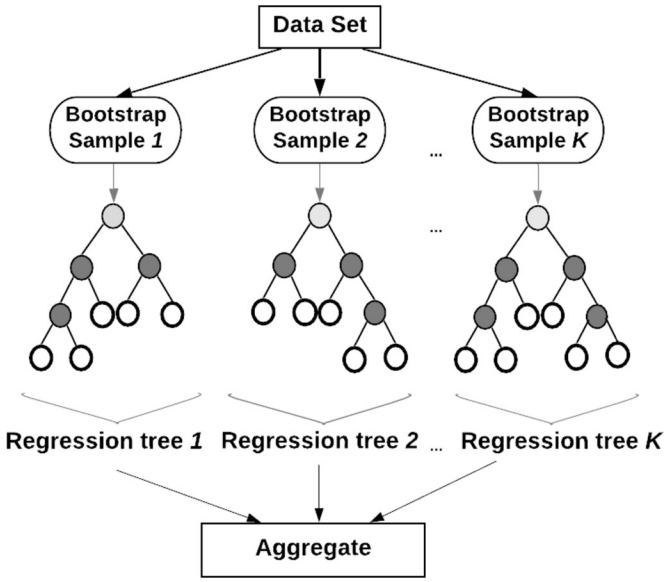

Figura 5.2: Estructura general para bosques aleatorios (Makariou et al. 2021).

En primera instancia, para el proceso de generación de bosques aleatorios se utiliza el muestreo de arranque. En particular, de un conjunto de datos se toman muestras con reemplazo, cada una de ellas del mismo tamaño que el conjunto de datos original. La segunda etapa es el desarrollo de árboles de clasificación, donde de cada muestra de arranque , se crean K árboles de clasificación los cuales se cultivan usando partición recursiva. En cada nivel del proceso de partición recursiva, el mejor predictor para realizar la división se considera en función de una submuestra aleatoria nueva, cada vez, del conjunto completo de predictores denotados como m_{try} (Makariou et al., 2021).

La mejor división se elige examinando todos los predictores posibles en esta submuestra y todos los puntos de corte posibles según su capacidad para minimizar la suma de cuadrados residual para el árbol resultante. Un árbol deja de crecer cuando se alcanza un número mínimo de observaciones en un nodo dado, pero en general los árboles que componen el bosque aleatorio crecen por completo

y no se podan. Al construir estos K árboles, efectivamente obtenemos K estimadores de la función $f(h_1, h_2, ..., h_k)$. El promedio de estos estimadores individuales $h_{en} = \frac{1}{K}\sum_{k=1}^{K} h_k(x_n)$ es el bosque aleatorio (Makariou et al., 2021).

A partir de la descripción del proceso anterior, es evidente que hay tres parámetros cuyos valores deben fijarse antes del desarrollo de poda aleatorio; para ello, debe fijarse el número de árboles cultivados, el tamaño del nodo y el número de variables seleccionadas al azar en cada división. Cada uno de ellos controla respectivamente el tamaño del bosque, el tamaño del árbol individual y un aspecto de la aleatoriedad dentro del árbol. Hay ciertos valores predeterminados que se han sugerido después de experimentos empíricos en varios conjuntos de datos, pero se puede usar una estrategia de optimización de ajuste con respecto al rendimiento de la predicción para seleccionar los valores más adecuados específicamente para el conjunto de datos en estudio (Probst et al., 2018).

5.3. Perceptrón multicapa

El campo de las Redes Neuronales ha surgido de diversas fuentes, que van desde la fascinación de la humanidad con la comprensión y la emulación del cerebro humano, a aspectos más amplios de copia de habilidades humanas como el habla y el uso del lenguaje. También a la práctica comercial, disciplinas científicas y de ingeniería de reconocimiento, modelado y predicción de patrones. Combinando múltiples capas ocultas y funciones de activación no lineales, los modelos de redes pueden aprender prácticamente cualquier patrón (Opela et al., 2021).

Los investigadores descubrieron que al combinar múltiples capas ocultas, la red puede aprender relaciones mucho más complejas entre los predictores y la variable respuesta. A esta estructura se le conoce como perceptrón multicapa o *multilayer perceptron* (MLP). Su estructura consta de varias capas de neuronas ocultas, cada neurona está conectada a todas las neuronas de la capa anterior y a las de la capa posterior como se aprecia en la figura 5.3. Aunque no es estrictamente necesario, todas las neuronas que forman parte de una misma capa suelen emplear la misma función de activación (Amat, 2021) .

La activación de la neurona de McCulloch-Pitts se ha generalizado de la siguiente forma:

$$y_i = f_j(\sum w_{ji} X_i) \tag{5.3}$$

donde la función de activación f_j puede ser cualquier función no lineal. Los nodos han sido divididos en una capa de entrada I y una capa de salida O.

"El nivel de umbral o sesgo de la ecuación (5.1) se ha incluido en la suma, con el supuesto de un componente extra en el vector X cuyo valor se fija en 1. Rosenblatt estudió las capacidades de grupos de neuronas en un capa única y, por tanto, todos actúan sobre los mismos vectores de entrada. Esta estructura se denominó Perceptrón y Rosenblatt propuso la regla de aprendizaje de Perceptron para aprender pesos adecuados para problemas de clasificación. Cuando f es una función de umbral estricto, es decir, salta de forma discontinua de un valor límite inferior a uno superior, la ecuación (6.2) define una función no lineal a través de un hiperplano en el espacio de atributos; con una función de activación de umbral la salida de la neurona es simplemente 1 en un lado del hiperplano y 0 en el otro. Cuando se combinan en una estructura de perceptrón, las neuronas

pueden segmentar el atribuir el espacio en regiones, y esto forma la base de la capacidad de las redes de perceptrones para realizar la clasificación" (Michie et al., 1994).

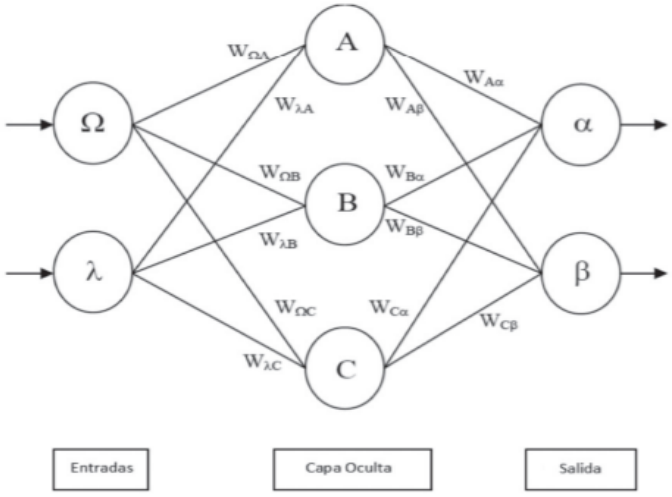

Figura 5.3: Estructura Perceptrón multipaca (Mercado et al., 2015).

5.4. Análisis discriminante lineal

El análisis de discriminante lineal (ADL) es un método empírico para la clasificación basado en vectores de atributos. Un hiperplano (línea en dos dimensiones, plano en tres dimensiones, etc.) en el espacio de atributos de dimensión 0 se elige para separar también las clases conocidas como sea posible. Los puntos se clasifican según el lado del hiperplano sobre el que caen (Alpaydin, 2014).

En el caso de dos clases, asumiendo que $\bar{x}, \bar{x}_1$ y $\bar{x}_2$ son respectivamente las medias de los vectores de atributos en general y para las dos clases. Supongamos que se nos da un conjunto de coeficientes $a_1, ..., a_p$ y llamando a la particular combinación lineal de atributos $g(x) = \sum a_j x_j$ el discriminante entre las clases (Alpaydin, 2014).

Ahora que los discriminantes para las dos clases difieren tanto como sea posible, y una medida para esto es la diferencia $g(\bar{x}_1) - g(\bar{x}_2)$ entre los discriminantes medios para las dos clases divididas por la desviación estándar de los discriminantes S_g, se obtiene la siguiente medida de discriminación (Murphy, 2012):

$$\frac{g(\bar{x}_1) - g(\bar{x}_2)}{S_g} \tag{5.4}$$

Esta medida de discriminación está relacionada con una estimación de error de clasificación basado en la suposición de una distribución normal multivariable para $g(x)$ (note que esta es una suposición más débil a que x tiene una distribución normal). Establecemos la línea divisoria entre las dos clases dado por el punto medio entre los medios de las dos clases. Entonces podemos estimar la probabilidad de error de clasificación para una clase como la probabilidad de que la variable aleatoria normal $g(x)$ para esa clase está en el lado equivocado de la línea divisoria, i.e. el lado equivocado de (Murphy, 2012):

$$\frac{g(\bar{x}_1) + g(\bar{x_2})}{2} \tag{5.5}$$

donde se asume que $g(\bar{x}_1) - g(\bar{x}_2)$ es negativo. Si las clases no son de igual tamaño, o si, como es muy frecuente en el caso, la varianza de $g(x)$ no es la misma para las dos clases, la línea divisoria se dibuja mejor en algún punto distinto del punto medio (Murphy, 2012)..

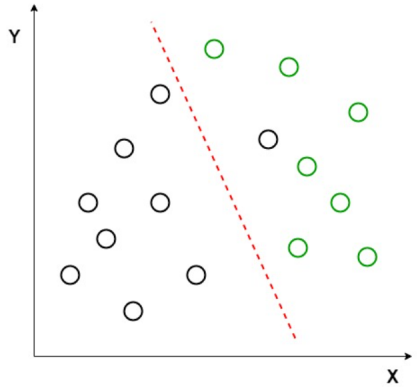

Figura 5.4: Clasificador de Análisis de discriminante lineal.

5.5. Métricas de desempeño

"Las métricas de rendimiento juegan un papel muy importante en la clasificación, donde se busca discriminar diferentes algoritmos Machine y Deep Learning, con la finalidad de facilitar la elección del mejor algoritmo dependiendo del objetivo de investigación" (Danjuma, 2015).

La matriz de confusión es uno de las medidores de rendimiento más utilizados el cual compara las clases predichas y las columnas de clase real, sus resultados se expresan como se presentan la figura 5.5 (Borja et al., 2020).

	Clasificador 1	
	Clase 1: Positivo	**Clase 2: Negativo**
Clase 1: Positivo	f11 = Verdadero positivo	f10 = Falso negativo
Clase 2: Negativo	f01 = Falso positivo	f00 = Verdadero negativo

Figura 5.5: Composición matriz de confusión para variables dicótomas (Borja et al., 2020).

Métrica	Fórmula	Descripción
Exactitud	$\frac{f_{11}+f_{00}}{N}$	Proporción de clasificaciones correctamente realizadas.
Sensibilidad	$\frac{f_{11}}{f_{11}+f_{01}}$	Proporción de Casos positivos correctamente clasificados.
Precisión	$\frac{f_{11}}{f_{11}+f_{10}}$	Proporción de casos mal clasificados.
Medida F	$2 * \frac{precision*sensibilidad}{precision+sensibilidad}$	Medida armónica de las métricas precisión y sensibilidad.
Error	$\frac{f_{10}}{f_{10}+f_{00}}$	Error tipo I

Cuadro 5.1: Fórmulas métricas de desempeño (Borja et al., 2020).

Capítulo 6

Metodología

En el presente capítulo se dará a conocer información de la base de datos y las técnicas implementadas para el desarrollo de los modelos. Además, se analizará los software y librerías utilizadas para el estudio y modelación de los datos.

6.1. Software

Para el estudio y aplicación de modelos en el conjunto se utilizaron 3 software principales los cuales fueron *Python* en su interfaz de *jupyter notebook*, *RStudio* y *STATA16*. Dentro de los múltiples paquetes y algoritmos utilizados se pueden mencionar las siguientes librerías del software principal:

- Math: este modulo proporciona funciones matemáticas definidas por el estándar *C*. Algunas de las funciones que proporciona son: de potencia y logarítmicas, trigonométricas, conversión angular, funciones hiperbólicas, constantes, etc. Es una de las librerías más importantes para el procesamiento de datos a alto nivel de Python. La estructura básica de pandas, series (unidimensional) y DataFrame (bidimensional).
- Numpy: Ofrece soporte para matriz N-dimensional y con funciones matemáticas para operar en las matrices. Se pueden definir tipos de datos arbitrarios, permitiendo que se integre una amplia variedad de bases de datos.
- Matplotlib.pyplot: diseñado para realizar figuras interactivas y gráficos programáticos.
- Sklearn: biblioteca para aprendizaje automático. Cuenta con algoritmos de clasificación, regresión, clustering, etc.
- Train test split: aplicación para dividir matrices en subconjuntos de pruebas y entrenamientos aleatorios.
- StandardScaler: aplicación para estandarizar características eliminando la media y escalando la varianza.
- RandomForestClassifier: clasificador de bosque aleatorio para ajustar varios clasificadores de árboles de decisión en varias submuestras del conjunto de entrenamiento.
- DecisionTreeClassifier: clasificador para árboles de decisión.

- KFold: divide el conjunto de datos en k pliegues consecutivos (sin mezclar de forma predeterminada).
- SVC: librería para la implementación de support vector classifier.
- LinearDiscriminantAnalysis: clasificador con un límite de decisión lineal, generado ajustando densidades condicionales de clase a los datos y utilizando la regla de Bayes.
- MLPClassifier: perceptrón multicapa para clasificación.
- Classification report: crea un informe para las métricas de la clasificación, con *precisión*, *recall* y *f1-score*.
- Confusion matrix: calcula la matriz de confusión para evaluar la precisión de una clasificación.
- Accuracy score: métrica que calcula la precisión del subconjunto con las etiquetas predichas.
- Shap: es una librería para realizar Inteligencia Artificial Explicable. Utiliza cálculos del campo de la teoría de juegos para averiguar qué variables tienen más influencia en las predicciones de las técnicas de machine learning.
- GridSearchCV: permite seleccionar los valores de los hiperparametros para un modelo y conjunto de datos.

Para RStudio se utilizaron las librerías:

- Car: posee herramientas para la aplicación de pruebas de regresión.
- lmtest: una colección de pruebas, conjuntos de datos y ejemplos para la verificación de diagnóstico en modelos de regresión.
- Readxl: permite obtener los datos de Excel y en R.
- Skimr: está diseñado para proporcionar estadísticas resumidas sobre variables en marcos de datos, tibbles, tablas de datos y vectores.
- Rpart.plot: partición recursiva para árboles de clasificación, regresión y supervivencia.
- Texreg: conversión de la salida de regresión R a tablas LaTeX o HTML.

Por último, cabe destacar que para STATA no es necesario ir implementando librerías sino ir cargando los datos y aplicando los comandos que ya vienen previamente cargados en el software.

6.2. Conjunto de datos y agrupación

El conjunto de datos a utilizar cuanta con la aprobación del comité de ética el cual fue extraído y digitalizado del libro de partos del Hospital San Camilo, Provincia de San Felipe, Región de Valparaíso. Contiene la información de 11,198 pacientes anonimizadas atendidas en el centro hospitalario desde Enero del 2015 a Agosto del 2021, excluyendo el año 2020 que no presenta registros en el conjunto de datos. Posee un total de 121 variables entre las cuales se encuentra información de la madre, información del recién nacido y variables clínicas.

Para el análisis del conjunto de datos se buscó en primera instancia disminuir la dimensionalidad mediante técnicas de *cluster* y Análisis de componentes principales. Sin embargo, no es viable aplicarlo debido a la naturaleza de las variables y principalmente por la pérdida de la interpretabilidad debido a la agrupación de las variables. Debido a que el conjunto de datos posee una gran cantidad de variables y a que no es posible agrupar mediante técnicas de *clusters*, se optó por la creación de 2 grandes grupos, separados en variables previas y posteriores al parto (ver las tablas 6.1, 6.2, 6.3, 6.4). Vale decir, que en las variables binarias, el valor 1 indica la presencia de la complicación mientras que el 0 la ausencia.

Variables pre-parto		
Variable	Significado	Valores
Edad	Edad de la madre	valores del 1 al 5.
E. Civil	Estado Civil	valores del 1 al 4.
Escolaridad	Nivel escolar	valores del 1 al 3.
Previsión	Tipo de previsión	valores del 1 al 4.
Tramo	Tramo de la previsión	valores del 1 al 6.
Paridad	Cantidad de hijos previos al parto	valores del 0 al 11.
I	E. Nutricional al inicio del embarazo	valores del 1 al 4
OB	Obesidad	Binaria
SHE	Síndrome Hipertensivo del Embarazo	Binaria
HPT	Hipotiroidismo	Binaria
MCO	Malas Condiciones Obstétricas	Binaria
SEQ	Solicitud de Esterilización materna	Binaria
SIFL	Sífilis	Binaria
RCIU	Restricción de Crecimiento Intrauterino	Binaria
IVC	Inserción Velamentosa de Cordón	Binaria
CIE	Colestasia Intrahepática del Embarazo	Binaria
PES	Preeclampsia Severa	Binaria
INDFAL	Inducción Fallida/Frustrada	Binaria
METR	Metrorragia en el embarazo	Binaria
DPPNI	Desprendimiento Prematuro de Placenta Normoincerta	Binaria
OS	Occipito Sacra	Binaria
NUDV	Nudo Umbilical Verdadero	Binaria
PE	Preeclampsia	Binaria
INFO	Infección Ovular	Binaria
SMC	Solicitud Materna de Cesárea	Binaria
PHA	Polihidramnios	Binaria

Cuadro 6.1: Variables Pre-Parto, Grupo 1.

Variables pre-parto		
Variable	Significado	Valores
OB	Obesidad	Binaria
SHE	Síndrome Hipertensivo del Embarazo	Binaria
HPT	Hipotiroidismo	Binaria
MCO	Malas Condiciones Obstétricas	Binaria
SEQ	Solicitud de Esterilización materna	Binaria
SIFL	Sífilis	Binaria
RCIU	Restricción de Crecimiento Intrauterino	Binaria
IVC	Inserción Velamentosa de Cordón	Binaria
CIE	Colestasia Intrahepática del Embarazo	Binaria
PES	Preeclampsia Severa	Binaria
INDFAL	Inducción Fallida/Frustrada	Binaria
METR	Metrorragia en el embarazo	Binaria
DPPNI	Desprendimiento Prematuro de Placenta Normoincerta	Binaria
OS	Occipito Sacra	Binaria
NUDV	Nudo Umbilical Verdadero	Binaria
PE	Preeclampsia	Binaria
INFO	Infección Ovular	Binaria
SMC	Solicitud Materna de Cesárea	Binaria
PHA	Polihidramnios	Binaria
OHA	Oligohidramnios	Binaria
PRODR	Pródromos	Binaria
OBM	Obesidad Mórbida	Binaria
DUFP	Deterioro Unidad Fetoplacentaria	Binaria
DALT	Doppler Alterado	Binaria
FDP	Falta de Descenso de Presentación	Binaria
INDFAL	Inducción Fallida/Frustrada	Binaria
CHG(+)	Chagas	Binaria
ENC	Embarazo No Controlado	Binaria
INUT	Inercia Uterina	Binaria
DP	Deterioro Progresivo	Binaria
HELLP	Síndrome de hellp	Binaria
HTA	Hipertensión Arterial	Binaria
PP	Placenta Previa	Binaria
Multiparidad	Más de un hijo.	Binaria
PEM	Preeclampsia Moderada	Binaria
COVID	Paciente con COVID	Binaria
SGB	Estreptococo del grupo B	Binaria
ANM	Anemia	Binaria
TTC	Test de Tolerancia a las Contracciones	Binaria
CORIO	Corioamnionitis	Binaria

Cuadro 6.2: Variables Pre-Parto, Grupo 1.

Variables Parto y Post-Parto		
Variable	Significado	Valores
EG(Decimal)	Edad gestacional decimal	valores del 20.0 al 42.0
E	Estado nutricional al final del embarazo	valores del 1 al 4.
Tipo	Tipo de parto	valores del 1 al 2.
Nacido	Feto nacido Vivo/Fallecido	valores del 1 al 2.
Sexo	Sexo del recién nacido(RN)	valores del 1 al 2.
Peso	Peso del RN	valores desde 465gr. a 5670gr.
Talla	Talla del RN	valores desde 25cm. a 59.5cm.
I. Ponderal	Índice ponderal del RN	valores desde 1.18 a 5.80.
CC	Centímetros craneales del RN	valores desde 19cm. a 40cm.
APGAR 1	Primer examen de Apgar al RN	valores de 0-10
APGAR 2	Segundo examen de Apgar al RN	Valores de 0-10
EGP	Edad Gestacional dada por el médico	Valores de 20 a 42
CES	Cesárea	Binaria
CAA	Circular al Abdomen	Binaria
CACI	Circular Al Cuello Irreductible	Binaria
MEC+	Meconio +	Binaria
SFA	Sufrimiento fetal agudo	Binaria
GEG	Grande para la Edad Gestacional	Binaria
RH(-)	Grupo Sanguíneo	Binaria
LATM	Laterocidencia de Mano	Binaria
TPD	Trabajo de Parto Detenido	Binaria
DCP	Desproporción Cefalopelvica	Binaria
SDPP	Síndrome de Parto Prematuro	Binaria
PEXP	Expulsión	Binaria
MEC++	Meconio ++	Binaria
MRTN	Mortinato	Binaria
PMI	Pujo Materno Inefectivo	Binaria

Cuadro 6.3: Variables parto y post-parto, grupo 2.

Variables Parto y Post-Parto		
Variable	Significado	Valores
Peso	Peso del RN	valores desde 465gr. a 5670gr.
Talla	Talla del RN	valores desde 25cm. a 59.5cm.
I. Ponderal	Índice ponderal del RN	valores desde 1.18 a 5.80.
CC	Centímetros craneales del RN	valores desde 19cm. a 40cm.
DIRP	Directo a Parto	Binaria
FCP	Fórceps	Binaria
DILAC	Dilatación Completa	Binaria
PEG	Pequeño para la Edad Gestacional	Binaria
TAQF	Taquicardia Fetal	Binaria
MEC	Meconio Fetal	Binaria
PEIN	Pelvis Infundibuliforme	Binaria
DDC	Displasia de Caderas	Binaria
RPLAC	Retención de Placenta Post Parto	Binaria
SHCG	Shock Cardiogénico	Binaria
MRTM	Muerte Materna	Binaria
MRTNEO	Muerte Neonatal	Binaria
SDPO	Síndrome de Potter	Binaria
MALF	Polimalformado	Binaria
POD	Podálica	Binaria
SDD	Síndrome de Down	Binaria
CRDF	Cardiopatía Fetal	Binaria
HT	Histerectomía	Binaria
Prematuro	RN Prematuro	Binaria
Macrosomía	RN más de 4000gr.	Binaria
BCF	Bradicardia Fetal	Binaria
POD	Podálica	Binaria
GEM	Gemelos	Binaria
EGD	Edad Gestacional Dudodas	Binaria

Cuadro 6.4: Variables parto y post-parto, grupo 2.

En las tablas presentadas anteriormente se aprecian todas las variables contenidas en el conjunto de datos, además del significado de cada abreviatura y el tipo de variable.
Para la aplicación del modelo en algunos software (como RStudio) el manejo de variables categóricas es automático. Sin embargo, tanto en Python como en STATA 16, las variables deben estar cuantificadas para la aplicación de la regresión logística, por lo cual, en las figura 6.5 y 6.6 se entregan los valores correspondientes a cada variable.

Variables cuantificadas				
Variable	Estado Civil	Escolaridad	Previsión	Tramo
Valores finales	Casada = 1 Conviviente = 2 Soltera = 3 Viuda = 4	Básica 1 Media= 2 Superior= 3	Dipreca= 1 Fonasa:2 Isapre: 3 SP= 4	Sin información=1 A= 2 B= 3 C= 4 D= 5 E= 6 PAD= 7 Particular= 8

Cuadro 6.5: Variables cuantificadas

Variables cuantificadas				
Variable	Est. Nutricional	Tipo de parto	Nacido	Sexo
Valores finales	Enflaquecida= 1 Normal= 2 Obesidad= 3 Sobrepeso= 4	Cesárea = 1 P. Vaginal= 2	Fallecido= 1 Vivo = 2	Femenino= 1 Masculino = 2

Cuadro 6.6: Variables cuantificadas

6.3. Regresión Logística

Para la aplicación de la regresión logística se aplicaron múltiples modelos con la misma variable dependiente la cual está señalada en el conjunto de datos como "DGTrelacionada a las pacientes que padecen diabetes tanto gestacional como previo al embarazo.

En primera instancia, se aplicaron 2 regresiones con los grupos conformados de las tablas 6.1 y 6.3 buscando las variables que presentan asociación con las pacientes diabéticas. Como los grupos están compuestos por una gran cantidad de variables, se realizaron múltiples regresiones posteriores que contenían las variables que arrojaron un nivel de signifancia considerable, es decir, un valor p menor a 0.05 que indica una relación la variable dependiente, en este caso con las pacientes diabéticas. Teniendo los resultados de las distintas regresiones se utilizó indicadores de los cuales se puede mencionar el criterio de información Akaike (*Akaike information criterion*, AIC por sus siglas en inglés) y el criterio de información bayesiano (*bayesian information criterion*, BIC por sus siglas en inglés) con los cuales se seleccionó el mejor modelo de regresión logística para el conjunto de datos.

Vale decir que se utilizo más de una regresión para el grupo 2 debido a que existen problemas de multicolinealidad, lo cual fue resuelto aplicando de forma individual de las variables que presentan una alta correlacionación en regresiones distintas. En RStudio es necesario separar las variables con problemas de multicolinealidad que pueden provocar la entrega errónea de resultados en la regresión debido a que se asume que se está entregando 2 veces la misma información con las variables correlacionadas mientra que en STATA se anulan automáticamente las variables con una muestra

no significativa y con problemas de multicolinealidad.

Una vez seleccionados los mejores modelos se realizó la verificación de los supuestos de regresión donde se confirmó la independencia y la homocedasticidad de los residuos mediante los test de Breush-Pagan y Durbin-Watson. Además se buscó el desempeño predictivo del modelo para evaluar el comportamiento de las regresiones.

6.4. Árboles de decisión

Mediante los resultados obtenidos de la regresión logística se extrajo del conjunto de datos las gestantes con diabetes de las cuales se creó un nuevo conjunto de datos con un total de 878 pacientes.

Para la aplicación de los modelos de machine learning, las variables significativas del grupo 1, es decir, los factores de riesgos que presentaron una asociación directa con las pacientes que padecen diabetes fueron seleccionadas como variables regresoras, mientras que los resultados del grupo 2 serán utilizadas como variables dependientes para realizar las predicciones.

Debido a que se redujo el tamaño del conjunto y se extrajo información únicamente de gestantes con diabetes, la distribución de presencia o ausencia de las complicación cambió en relación al conjunto de datos inicial. Esto se ve directamente afectado en cada una de las variables dependientes seleccionadas de la regresión provocando un desequilibrio de la variable respuesta donde predominan la ausencia de la complicación. Para el caso particular de los árboles de decisión, el modelo posee un comando que permite balancear la variable respuesta directamente en la aplicación del clasificador mediante *class_weight*, donde se le entrega la proporción del balance que se quiere utilizar. Esto dependerá de cada variable debido a que todas poseen una cantidad distinta de resultados. A modo de ejemplo se puede mencionar una variable que presenta el doble de 0 en comparación con 1 (ausencia o presencia de la complicación). Por lo tanto, la relación queda compuesta por la cantidad total de ausencia de la complicación dividido en la cantidad total de presencia de la complicación. En el ejemplo mencionado daría una proporción de $1:2$, cuyo resultado es el que se le entrega al comando y automáticamente genera el balance en la variable respuesta el cual se puede observar en el nodo principal. Los árboles de decisiones son propensos a caer en el sobreajuste (comúnmente conocido como *overfitting*). Es por ello la importancia de analizar los resultados del modelo para evitar este problema. El overfitting se analizó mediante la técnica de validación cruzada de forma gráfica y, a través de métricas de desempeño, los resultados de entrenamiento vs los resultados del test.

6.5. Comparación de clasificadores

Se aplicaron 4 clasificadores diferentes los cuales fueron SVM, ADL, bosque aleatorio y perceptrón multicapa para cada da una de las variables dependientes con los cuales se busca conocer el clasificador que presenta los mejores resultados para la predicción mediante el análisis de sus métricas de desempeño compuestas por la exactitud enfocada a los casos correctamente clasificados, la precisión se encarga de evaluar los casos positivos que se evaluaron erróneamente, la sensibilidad del modelo como evaluador de falsos negativos, la medida F la cual combina los resultados de la precisión, la sensibilidad y el error,. Todo lo anterior reforzado gráficamente con una matriz de confusión.

Por otro lado, si bien en los árboles de decisiones existe un comando que balancea automáticamente los datos, no es posible aplicarlo en todos los clasificadores. Es por ello que se utiliza una técnica distinta para la comparación de los clasificadores. Para solucionar esta situación se aplicó un método llamado *Oversampling* (en español, sobremuestreo) el cual implica duplicar aleatoriamente ejemplos de la clase minoritaria y agregarlos al conjunto de datos de entrenamiento. Por lo tanto, se buscaron los mejores hiperparámetros para los clasificadores de forma que los resultados sean basados en el mejor desempeño del clasificador respectivo.

6.6. Método de SHAP

Finalmente, se utilizó el método de SHAP como modelo explicable con el clasificador que presentó los mejores resultados en las métricas de desempeño para la predicción. Para este modelo, los resultados se obtienen de forma visual mediante gráficos en su mayoría interactivos en los cuales se conoce el impacto y la interacción de las variables para un aumento o descenso en la probabilidad de ocurrencia para un evento de interés.

Capítulo 7

Experimentos

7.1. Regresión Logística

Variables Pre-Parto				
Variable	Odds Ratio	P-valor	Error Estándar	[IC 95 %]
Edad	1.389	0.0003	0.006	1.291; 1.492
OB	1.732	0.00024	0.133	1.490; 1.995
SHE	1.87	0.00058	0.247	1.336; 1.922
HPT	6.283	0.00016	1.006	1.552; 1.548
MCO	3.237	0.00011	0.874	1.839; 5.663
SEQ	3.584	0.00249	1.919	1.545; 8.272
Multiparidad	1.479	0.00021	0.128	1.831; 1.316
METR	2.69	0.011	0.953	1.247; 5.766
PHA	2.299	0.031	0.885	1.174; 4.865
OHA	1.75	0.037	0.448	1.231; 3.057
PEM	3.022	0.032	0.448	1.484; 1.784

Cuadro 7.1: Variables significativas obtenidas de la regresión logística con mejor desempeño para el grupo 1.

Variables Parto y Post Parto				
Variable	Odds Ratio	P-valor	Error estándar	[IC 95 %]
Talla	1.16	0.029	0.076	1.017;1.126
CES	1.40	0.000	0.126	1.172;1.560
GEG	1.47	0.001	0.152	1.092;1.655
LATM	2.879	0.000	0.344	1.127;3.031
HT	4.19	0.004	1.010	1.324;2.854
Prematuro	1.34	0.004	0.144	1.135;1.892
Macrosomía	2.67	0.000	0.196	1.064;1.839
MEC	3.27	0.000	0.074	1.104;1.540
RPM	1.47	0.009	0.221	1.100; 1.982

Cuadro 7.2: Variables significativas obtenidas de la regresión logística con mejor desempeño para el grupo 2.

En los resultados de la regresión logística de las tablas 7.1 y 7.2 se puede deducir que algunas variables presentan una mayor asociación con la afección como por ejemplo el hipotiroidismo, ya que es uno de los factores de riesgo que presenta un odd ratio mayor el cual indica que esta característica aumenta 6.28 veces la probabilidad de padecer diabetes. Además, se muestran variables que ya son conocidas como factores de riesgo directo de esta enfermedad como son la edad y la obesidad.

También es posible identificar las variables protectoras las cuales presentan un odd ratio menor a 1. Éstas son el estado civil, el tramo y la previsión las cuales no están presentes en las tablas ya que sólo se muestran los factores de riesgo. Éstas serán utilizadas en la predicción con el mejor clasificador para conocer en detalle su impacto mediante el modelo explicable.

Para comprobar la correcta aplicación y validar los resultados de la regresión, se verificaron los supuestos de la regresión logística. Para la independencia de los residuos, se realizó el test de Breush-Pagan del cual se obtuvo un valor p sobre 0.05 lo cual indica que no existe evidencia muestral suficiente para rechazar la hipótesis nula sobre heterocedasticidad. Por lo tanto, se comprueba que los residuos poseen varianza constante. Además, se aplicó el test de Durwin-Watson para estudiar la independencia de los residuos. Se comprobó con un valor p menor a 0.05 que se cumple la independencia residual, por lo cual los resultados de la regresión logística son considerados positivos y se puede proceder a aplicar los demás modelos.

Es necesario recordar que los resultados obtenidos en el grupo 1 serán utilizadas como variables regresoras para la predicción mientras que los resultados del grupo 2 serán cada una de las variables dependientes, es decir, las complicaciones que se van a predecir.

7.2. Prematuro

7.2.1. Árbol de decisión

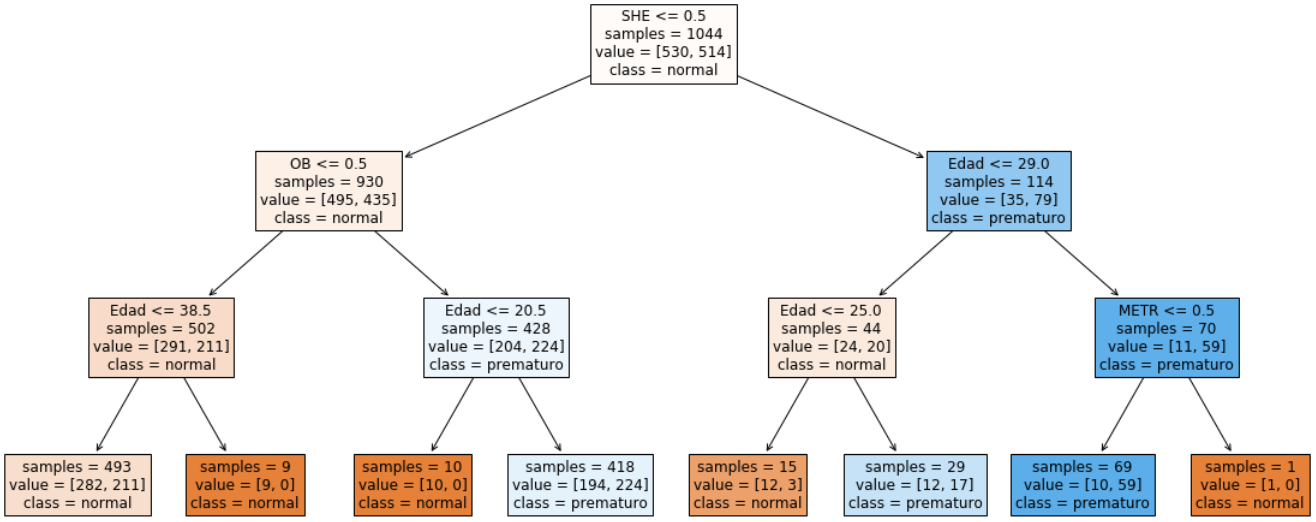

Figura 7.1: Árbol de decisión para clasificación de parto prematuro

En la figura 7.1 se aprecia un árbol de 3 hojas donde en 6 nodos el parto puede ser clasificado como prematuro. Tomando una rama de guía, se puede decir que si la paciente padece de hipertensión, es altamente probable que tenga un parto prematuro a no ser que tenga una edad menor a 25 años.

Para la aplicación de los árboles se verificó que la cantidad de hojas sea la adecuada para lograr el mejor desempeño del árbol evitando el sobreajuste mediante la validación cruzada (esta se puede apreciar gráficamente en la figura 7.24) y la selección de los mejores hiperparamétros.

Exactitud	Precisión	Sensibilidad	Medida F	Error
0.63	0.62	0.68	0.65	0.36

Cuadro 7.3: Medidas de desempeño para el árbol de decisión en parto prematuro.

7.2.2. Mejor Clasificador

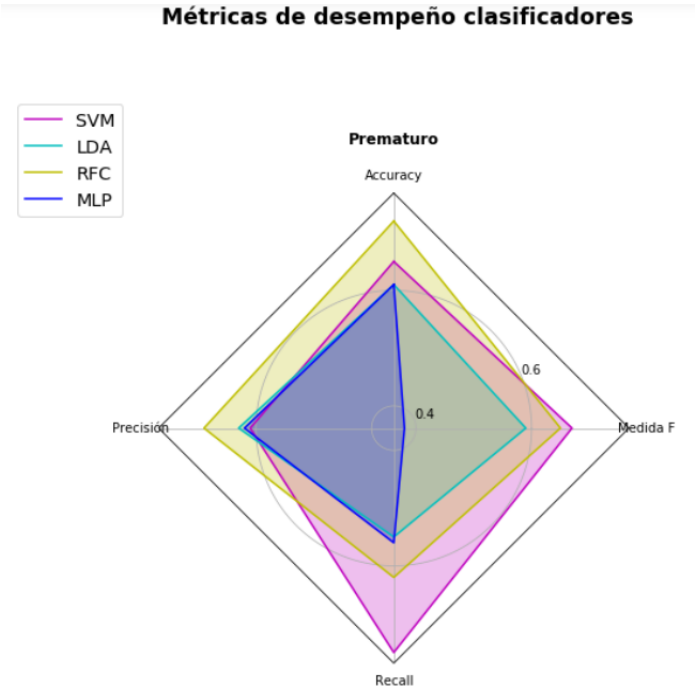

Figura 7.2: Comparación de métricas de desempeño para la predicción de parto prematuro.

Comparación de clasificadores variable parto prematuro				
Métrica	SVM	LDA	RFC	MLP
Exactitud	0.73	0.61	0.72	0.61
Precisión	0.72	0.63	0.69	0.62
Sensibilidad	0.60	0.55	0.62	0.56
Medida F	0.65	0.59	0.65	0.38

Cuadro 7.4: Comparación de métricas de desempeño en clasificadores para la predicción de parto prematuro.

Para la selección del mejor clasificador, tanto SVM como Bosque Aleatorio poseen buenos desempeños. Sin embargo, en general SVM obtiene mejores resultados con una exactitud del 73 %.

	precision	recall	f1-score	support
0	0.70	0.59	0.64	175
1	0.65	0.74	0.69	174

Figura 7.3: Medidas de desempeño bosque aleatorio para parto prematuro.

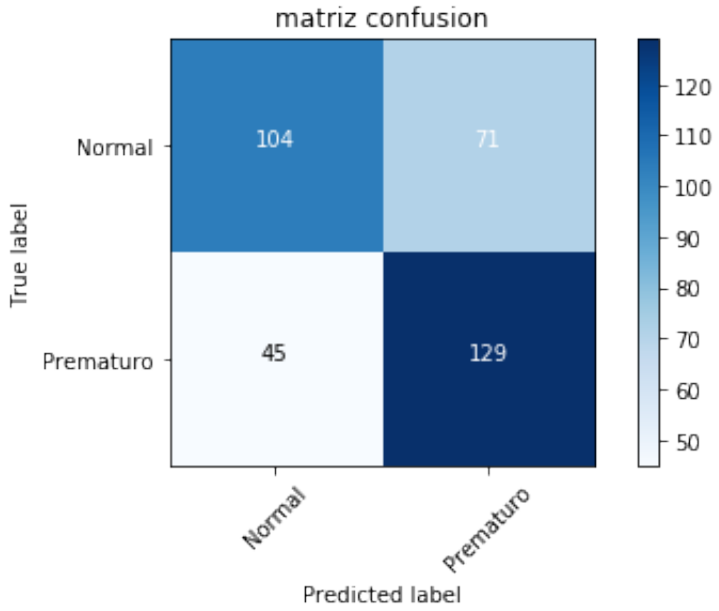

Figura 7.4: Matriz de confusión para predicción de parto prematuro mediante clasificador de máquinas de soporte.

En la figura 7.4 se observa que los falsos positivos son más alto que los falsos negativos con una cantidad de 71 vs 45 respectivamente. Pese a ello se clasifican bien 104 recién nacidos con su etapa gestacional completa y 129 con nacimientos prematuros.

7.2.3. Método de SHAP

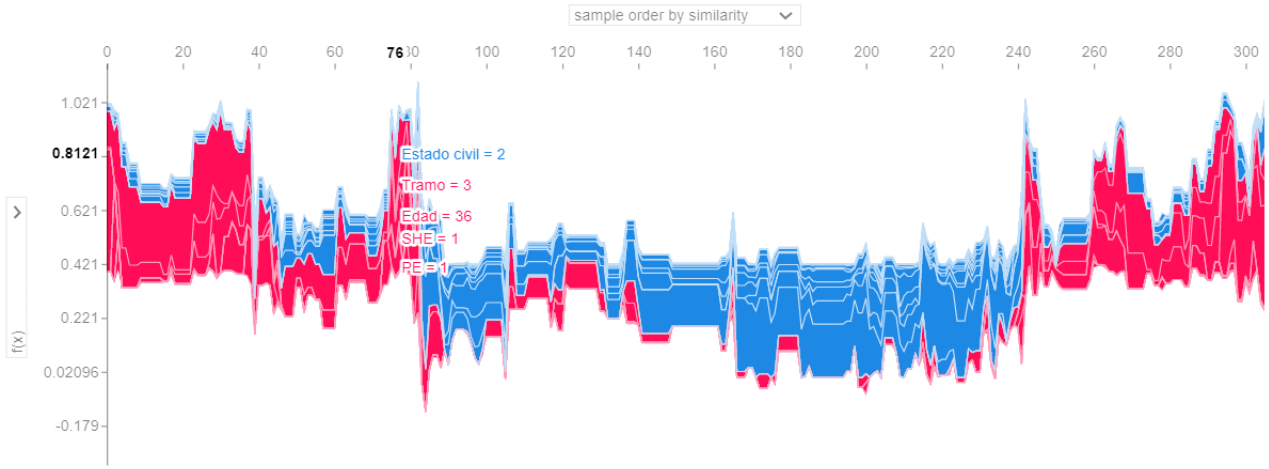

Figura 7.5: Método de SHAP en la predicción de parto prematuro.

La figura 7.5 se observa que el síndrome hipertensivo en el embarazo en conjunto con una edad avanzada y la preeclampsia aumenta el riesgo de tener un parto prematuro.

Figura 7.6: Método de SHAP en la predicción de parto prematuro.

En la figura 7.6,se aprecia que si no presenta hipotiroidismo ni síndrome hipertensivo en el embarazo a una edad de 26 años perteneciente al tramo A de FONASA la probabilidad que el parto sea prematuro son bajas.

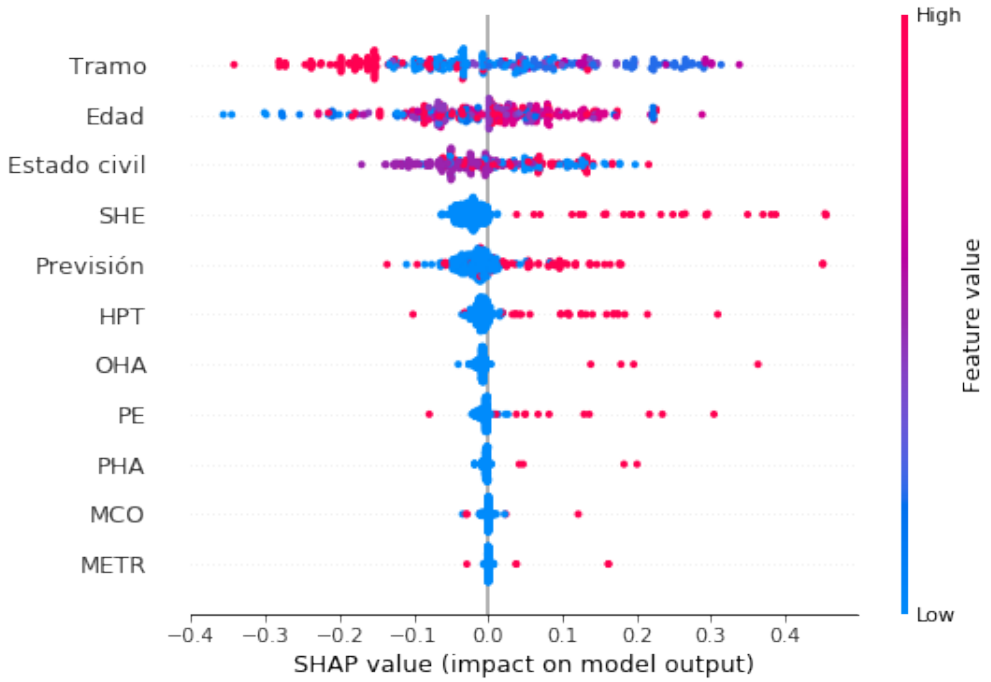

Figura 7.7: Influencia de variables en la predicción de parto prematuro

El gráfico de la figura 7.7 arroja que el tramo de salud puede influir en la disminución del parto prematuro. Por otro lado, la mayoría de las variables se correlacionan positivamente en el aumento de un parto prematuro.

7.3. Meconio

7.3.1. Árbol de decisión

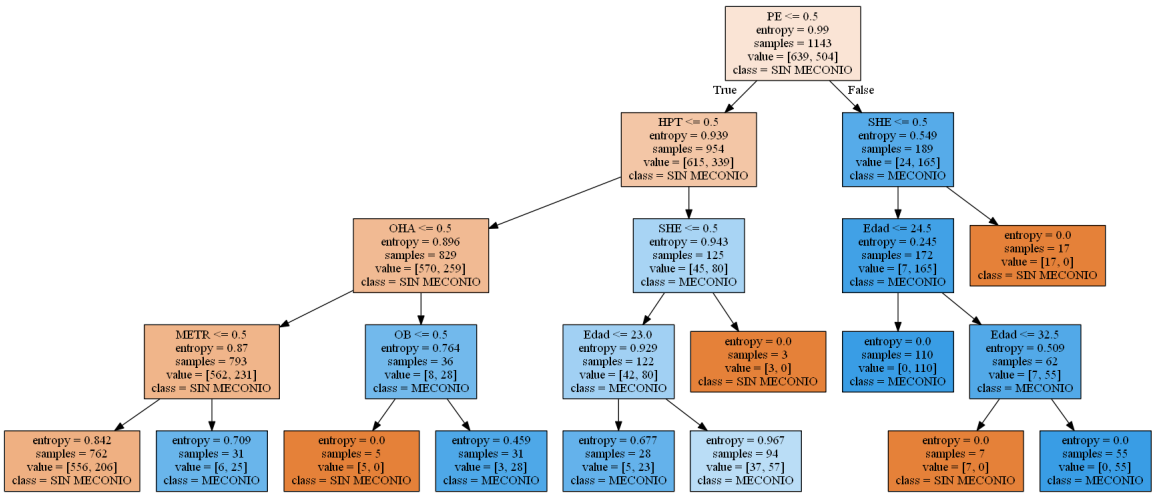

Figura 7.8: Árbol de decisión para Meconio.

En esta nueva variable dependiente vemos que la cantidad de hojas varía, esto ya que cada variable posee diferentes valores y características que entregan mayor información al modelo. El árbol de decisión de la figura 7.8 posee un total de cuatro hojas y para esta variable en específico toma como nodo principal la preeclampsia, lo cual indica que es la variable que está entregando más información al modelo.
Posteriormente, se realiza la división de que si tiene o no preeclampsia y, dependiendo de ello baja al siguiente nivel, el cual se divide en hipotiroidismo y síndrome hipertensivo en el embarazo.
Si se elige una de las ramas, es posible deducir por ejemplo que si la paciente padece de preeclampsia, además tiena una edad mayor a los 24 años y tiene hipertensión las probabilidades de padecer meconio son altas.

Exactitud	Precisión	Sensibilidad	Medida F	Error
0.78	0.80	0.65	0.72	0.21

Cuadro 7.5: Medidas de desempeño Decision Tree Meconio

7.3.2. Mejor Clasificador

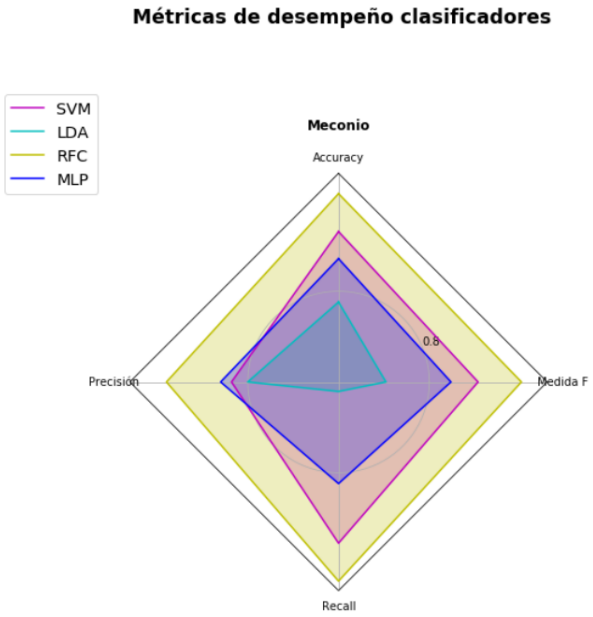

Figura 7.9: Comparación de métricas de desempeño para la clasificación de Meconio

Comparación de clasificadores variable meconio				
Métrica	SVM	LDA	RFC	MLP
Exactitud	0.91	0.78	0.98	0.86
Precisión	0.83	0.80	0.95	0.85
Sensibilidad	0.93	0.65	0.99	0.82
Medida F	0.89	0.72	0.97	0.84

Cuadro 7.6: Comparación de métricas de desempeño en clasificadores para la predicción de meconio.

	precision	recall	f1-score	support
0	1.00	0.97	0.98	212
1	0.96	1.00	0.98	165

Figura 7.10: Medidas de desempeño Bosque aleatorio para clasificación de Meconio

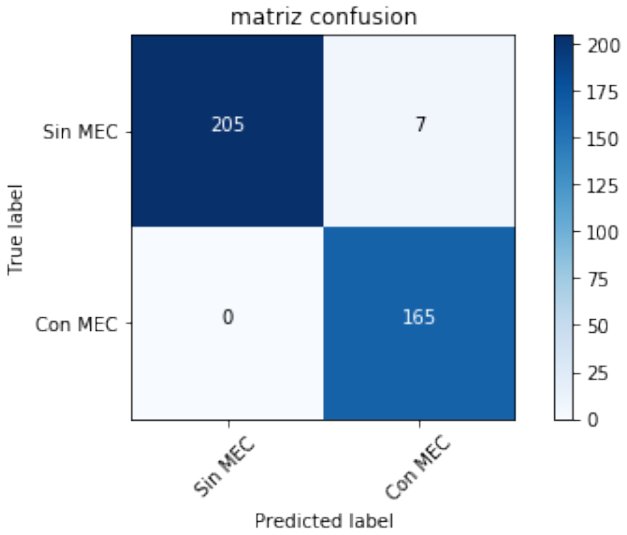

Figura 7.11: Matriz de confusión para clasificación de Meconio

En general los clasificadores utilizados presentaron desempeños aceptables para esta variable dependiente. Sin embargo, según los resultados obtenidos, el clasificador que obtuvo las mejores métricas fue bosque aleatorio, el cual está señalado en la figura 7.9 como RFC (*Random Forest Classifier*). Además, se comprueba la eficacia del modelo con la matriz de confusión la cual sólo clasifica erróneamente 7 situaciones sin meconio, mientras que en pacientes con Meconio la predicción no presenta fallos.

7.3.3. Método de SHAP

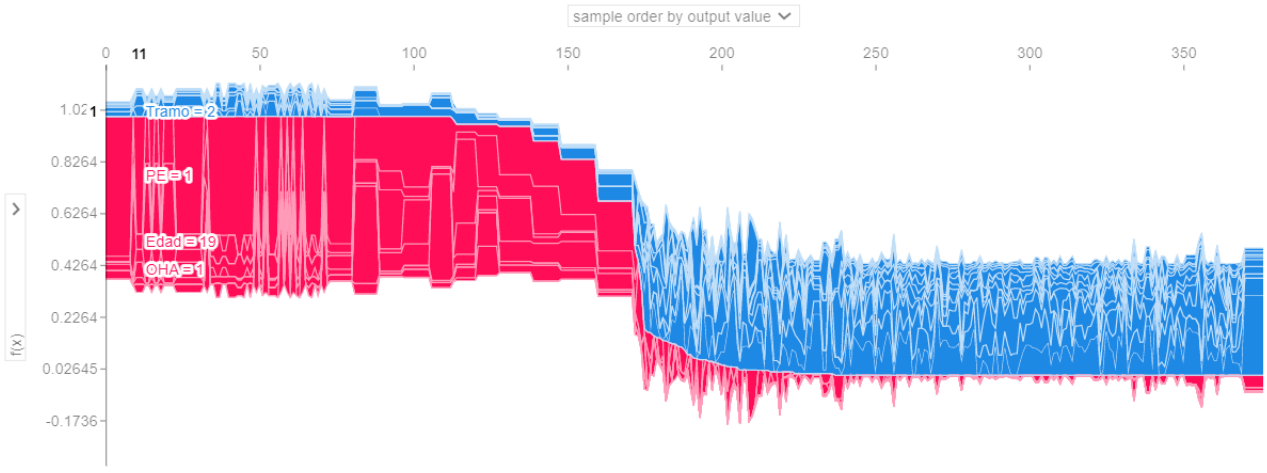

Figura 7.12: Método de SHAP para la predicción de Meconio.

En los gráficos del método de SHAP se van entregando las distintas interacciones de las características, donde las variables que aparecen con rosa ayudan en el incremento de la probabilidad y por ende en la presencia de la complicación, mientras que las variables en azul provocan el efecto

contrario causando la disminución de la probabilidad y disminuyendo la aparición de la complicación.
En la figura 7.12 se puede observar uno de los múltiples resultados para la predicción del Meconio. En el gráfico se aprecia que una paciente pese a ser joven (19 años) si presenta preeclampsia y oligohidramnios es altamente probable que tenga meconio. Además, en esta paciente el tramo actúa como variable protectora. El gráfico presenta 370 pacientes aproximadamente que equivale a la muestra del test del clasificador y obtiene las combinaciones probables de predicción para cada una de ellas.

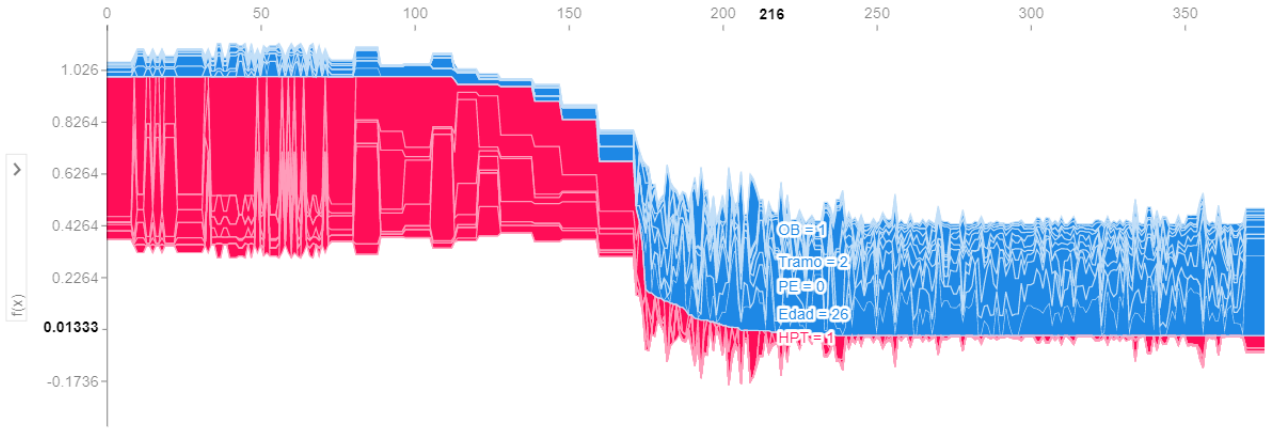

Figura 7.13: Método de SHAP a la predicción de Meconio

Por otro lado, en la figura 7.13 vemos que una paciente de 26 años con obesidad, perteneciente al tramo A de FONASA y sin preeclampsia, tiene baja probabilidad de Meconio. Sin embargo, el hipotiroidismo es una de las variables que aumenta el riesgo.

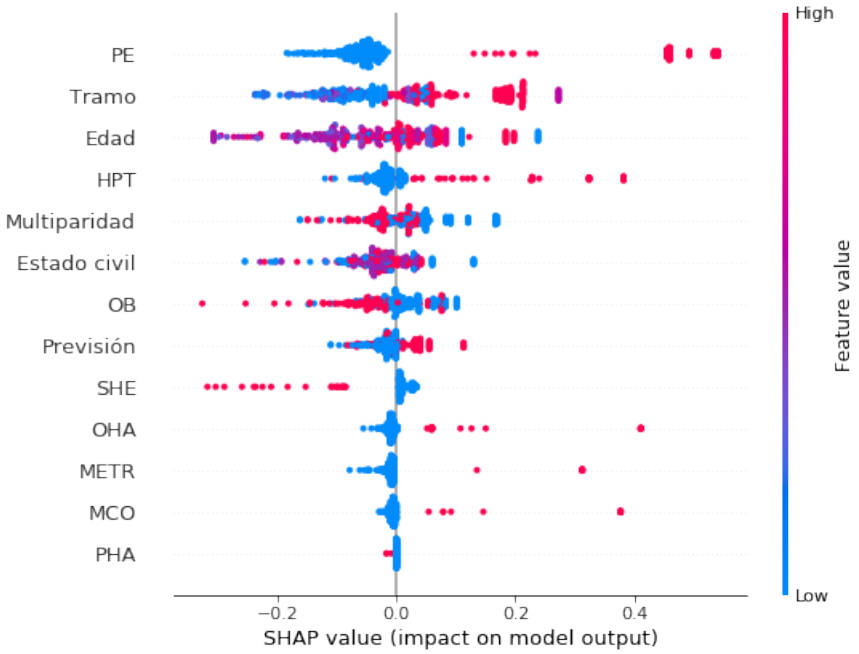

Figura 7.14: Método de SHAP para la predicción de Meconio

Por último, en la figura 7.14 se observa la influencia de cada variable en la predicción. Los puntos rojos para el lado positivo indican que esas variables aumentan la probabilidad de meconio, como por ejemplo la preeclampsia, el hipotiroidismo. También se ven variables sociodemográficas como el tramo y la previsión.

7.4. Cesárea

7.4.1. Árbol de decisión

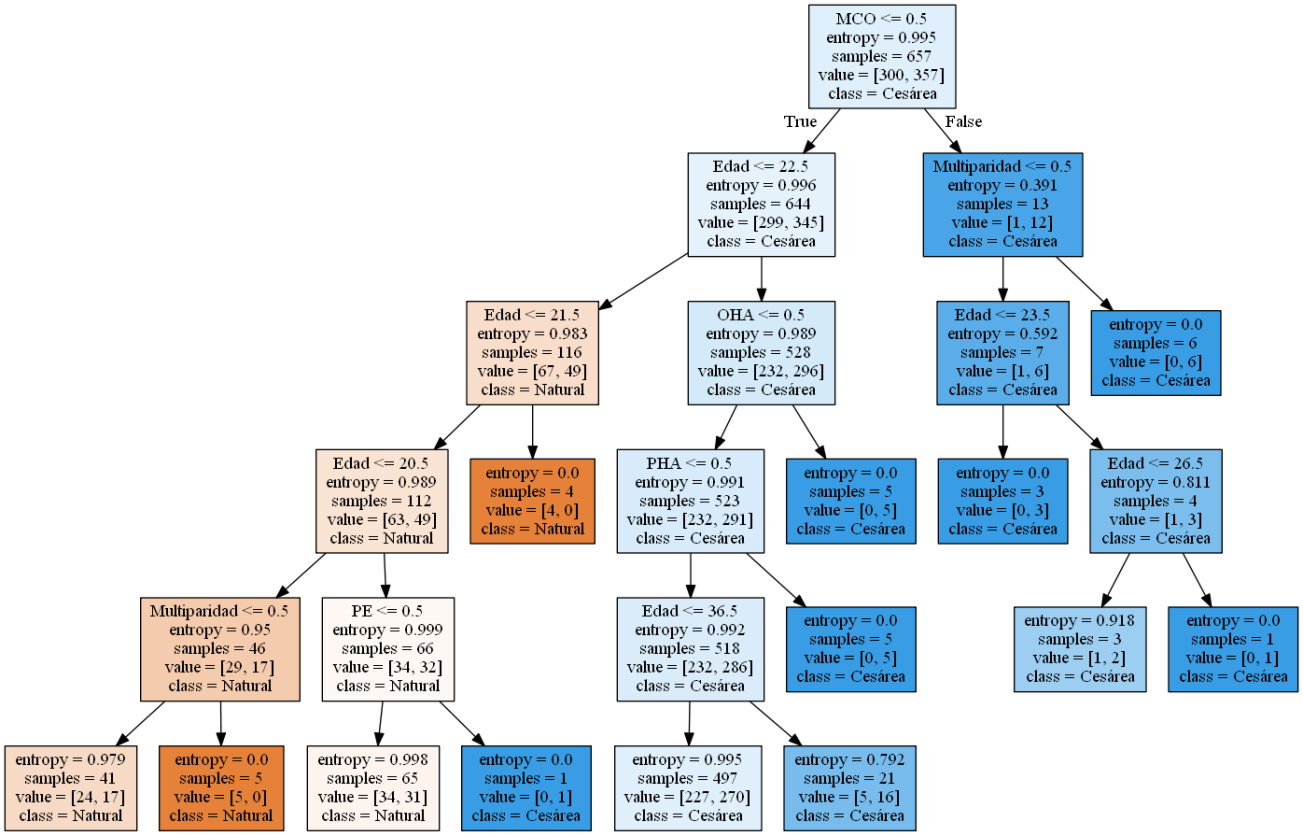

Figura 7.15: Árbol de decisión variable CES

La variable cesárea fue una de las más difíciles de analizar, esto debido a que en esta situación puntual se presentan más casos en partos naturales. En el árbol realizado se observa que las malas condiciones obstétricas es el factor más importante para el modelo y por ende es la característica que entrega más información al árbol. Además, de las 5 hojas sólo en 8 nodos el resultado sería clasificado como parto natural. Por otro lado, se aprecia que los factores como la edad y la multiparidad son las variables que están más relacionadas con el tipo de parto.

Exactitud	Precisión	Sensibilidad	Medida F	Error
0.62	0.60	0.94	0.73	0.33

Cuadro 7.7: Medidas de desempeño del árbol de decisión para parto por cesárea.

7.4.2. Mejor Clasificador

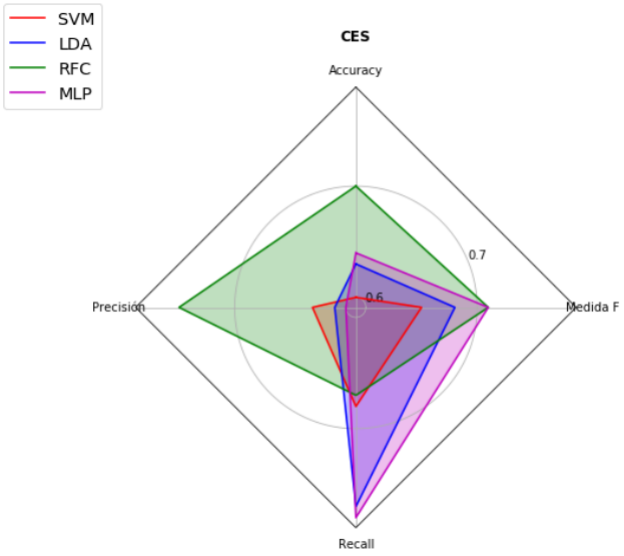

Figura 7.16: Comparación de métricas de desempeño para la clasificación de CES.

Comparación de clasificadores variable CES				
Métrica	SVM	LDA	RFC	MLP
Exactitud	0.60	0.63	0.70	0.64
Precisión	0.63	0.61	0.75	0.60
Sensibilidad	0.68	0.77	0.67	0.78
Medida F	0.65	0.68	0.71	0.71

Cuadro 7.8: Comparación de métricas de desempeño en clasificadores para la predicción de parto por cesárea.

	precision	recall	f1-score	support
0	0.64	0.72	0.68	72
1	0.75	0.68	0.71	90

Figura 7.17: Métricas de desempeño con bosque aleatorio para la predicción de cesárea

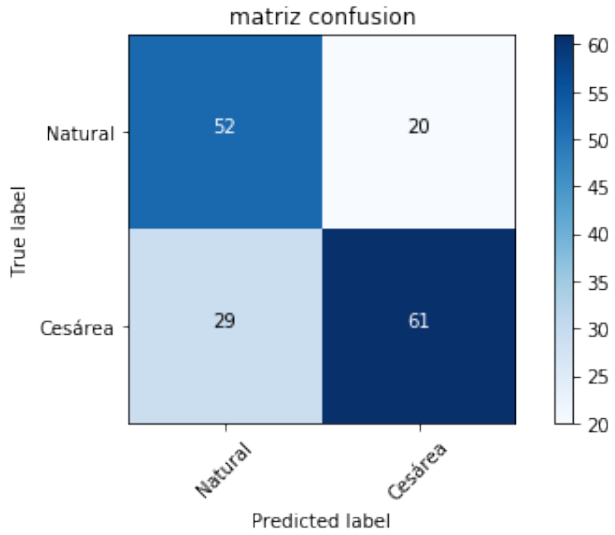

Figura 7.18: Matriz de confusión para predicción de cesárea.

De la matriz de confusión se puede deducir que los casos correctamente clasificados son 61 y 52 en pacientes con parto natural y cesárea, respectivamente. Mientras que los clasificados erróneamente son 20 y 29.

En la figura 7.16 se observa que el comportamiento de las métricas para cada clasificador varia considerablemente. En la comparación existen 2 clasificadores que poseen una alta sensibilidad (LDA, MLP). Sin embargo, se eligió al mejor clasificador a aquel que en todas las métricas presentó resultados significativos. Es por ello que para la variable de parto por cesárea el bosque aleatorio es el mejor clasificador con un 70 % de exactitud.

7.4.3. Método de SHAP

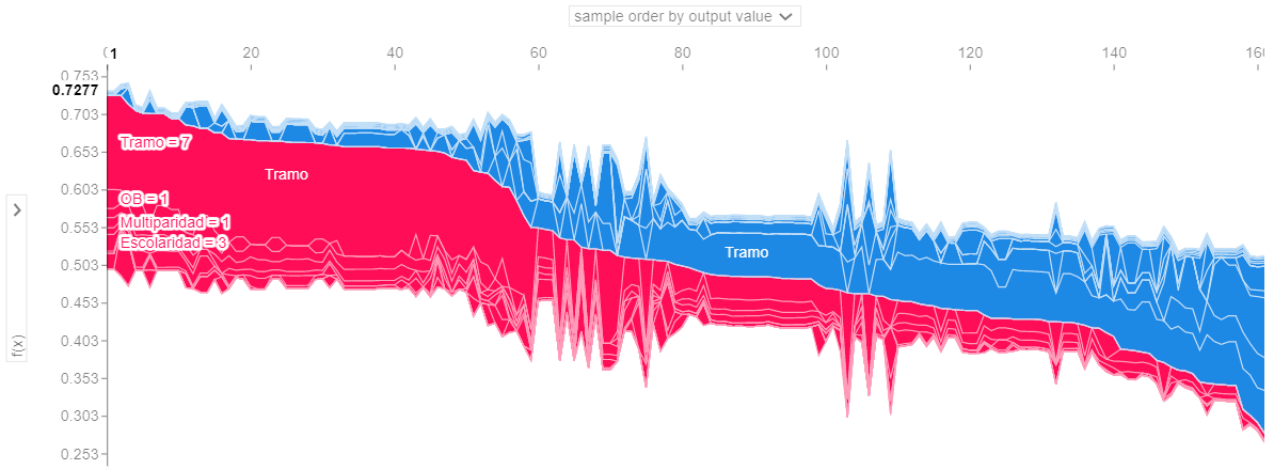

Figura 7.19: Método de SHAP para la predicción de cesárea

En la figura 7.19 se destaca la variable *tramo* tanto como variable de incremento y de disminución de probabilidad. Esto es debido a que considera como factor de incremento de cesárea a las gestantes

que se atendieron de forma particular o mediante el programa *PAD* (Pago Asociado a Diagnóstico). Además, el modelo asocia al parto natural a las mujeres de estratos socioeconómicos más bajos pertenecientes a grupos de FONASA A y B. Según el método de SHAP las variables como la obesidad y la multiparidad en conjunto con el tramo son factores que influyen en un parto por cesárea.

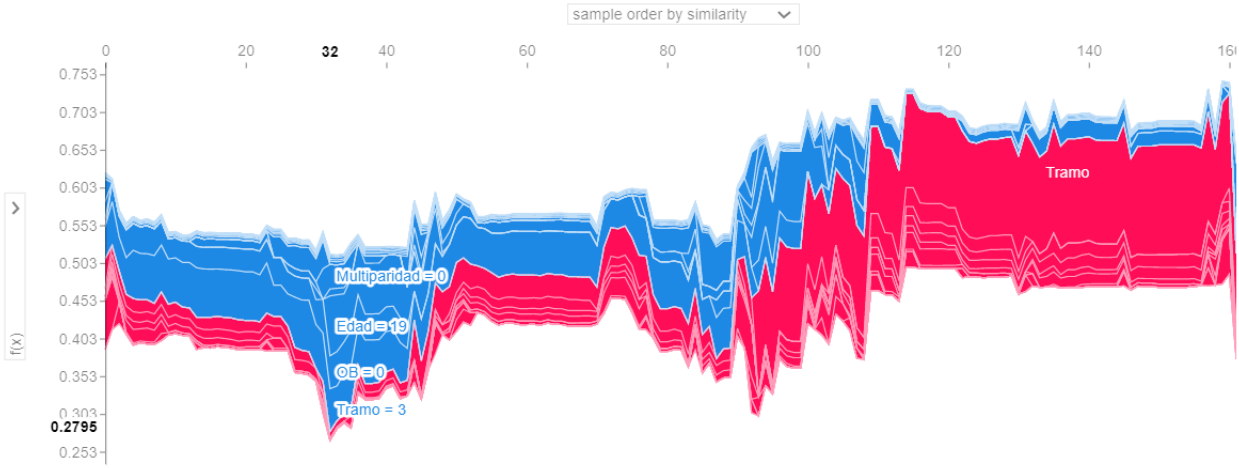

Figura 7.20: Método de SHAP para la predicción de parto por cesárea

Ahora viendo la probabilidad más baja de un parto por cesárea vemos los factores inversos a los mencionados anteriormente. Es decir, en la figura 7.20 podemos ver que una gestante sin multiparidad, sin obesidad, joven y perteneciente a FONASA B, es poco probable que tenga un parto por cesárea.

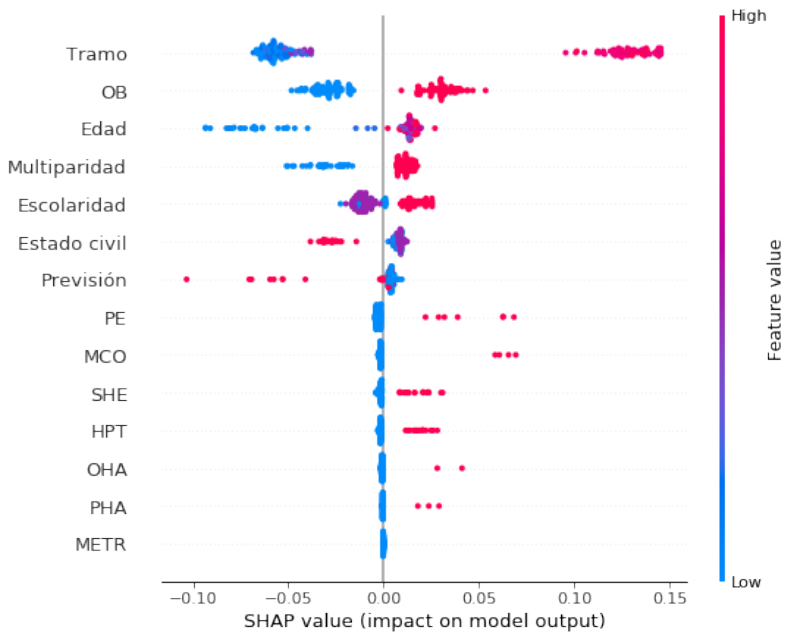

Figura 7.21: Método de SHAP para la predicción de cesárea mediante gráfico de influencia.

Figura 7.22: Método de SHAP para la predicción de cesárea mediante gráfico de impacto de características.

Por último, vemos que las figuras 7.21 y 7.22 confirman lo indicado por el gráfico anterior por lo cual factores como la edad, la multiparidad, el tramo socioeconómico, el nivel de escolaridad y la obesidad. Además, son las variables que presentan una correlación más alta en el aumento de la probabilidad de un parto por cesárea.

7.5. Grande para la edad gestacional

7.5.1. Árbol de decisión

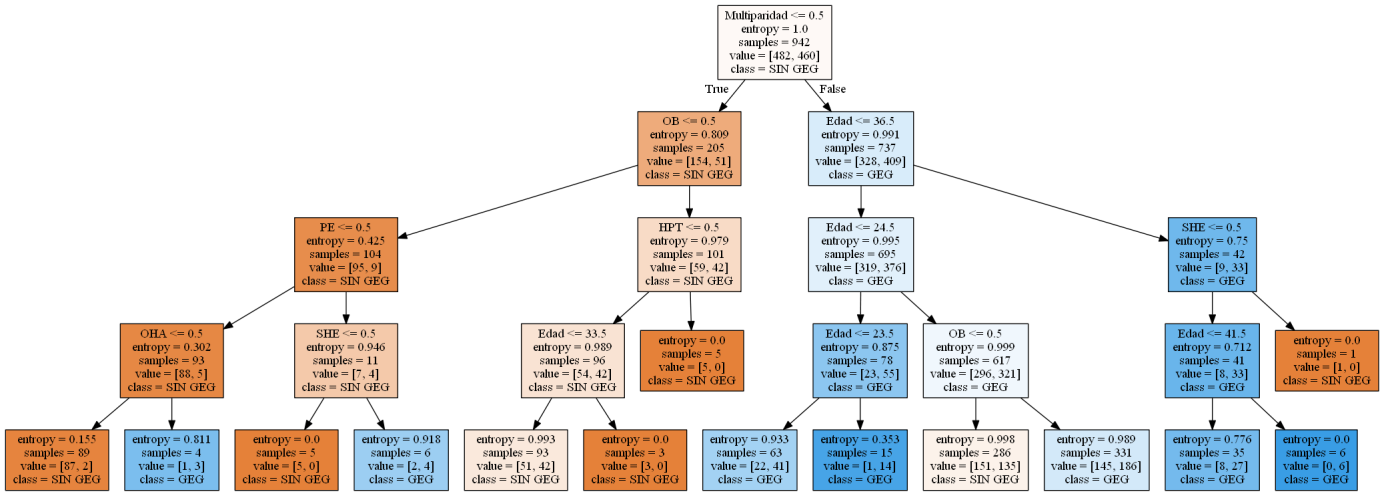

Figura 7.23: Árbol de decisión para la predicción de grande para la edad gestacional.

Para esta variable, la división de las hojas y el desempeño del árbol fue complejo debido a que presenta una baja relación con las variables clínicas, no así con las variables sociodemográficas. Debido a esto, los mejores rendimientos parten de cuatro hojas en adelante. Sin embargo, no se utilizó el árbol con más hojas debido a que dificulta la visualización y el modelo pierde interpretabilidad. El nodo principal está dado por la multiparidad, vemos que el árbol consta de 4 hojas y 27 nodos de los cuales 12 terminan con la complicación "GEG".
Es posible apreciar que si la paciente diabética tiene hijos previos es altamente probable que el bebé nazca grande para la edad gestacional, a eso se le suman otros factores de riesgo como la edad, la obesidad y la hipertensión.

Exactitud	Precisión	Sensibilidad	Medida F	Error
0.62	0.61	0.55	0.57	0.35

Cuadro 7.9: Medidas de desempeño de árbol de decisión para la variable grande para la edad gestacional.

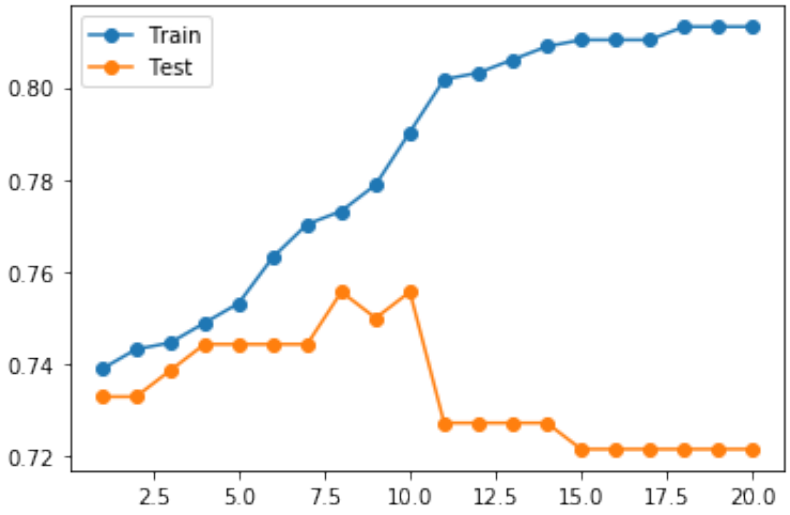

Figura 7.24: Validación cruzada para decision tree de variable GEG.

7.5.2. Mejor Clasificador

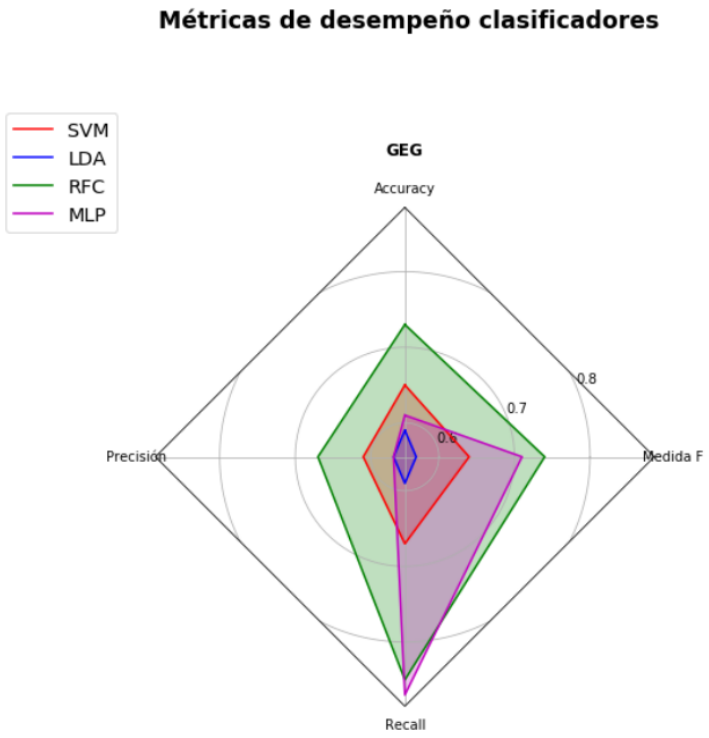

Figura 7.25: Comparación de métricas de desempeño para la clasificación de GEG.

Comparación de clasificadores variable GEG				
Métrica	SVM	LDA	RFC	MLP
Exactitud	0.65	0.59	0.73	0.61
Precisión	0.61	0.57	0.67	0.57
Sensibilidad	0.67	0.59	0.85	0.87
Medida F	0.64	0.57	0.74	0.71

Cuadro 7.10: Comparación de métricas de desempeño en clasificadores para la predicción de GEG.

	precision	recall	f1-score	support
0	0.82	0.62	0.71	165
1	0.67	0.85	0.75	149

Figura 7.26: Métricas de desempeño con bosque aleatorio para la predicción de GEG.

En la comparación de los clasificadores se aprecia una diferencia significativa entre los clasificadores. El desempeño más bajo lo presenta el análisis de discriminante lineal y el mejor clasificador

fue el bosque aleatorio con una exactitud del 72 %. Por otro lado, la sensibilidad alta se aprecia en la matriz de confusión donde clasifica erróneamente sólo 22 casos sin GEG.

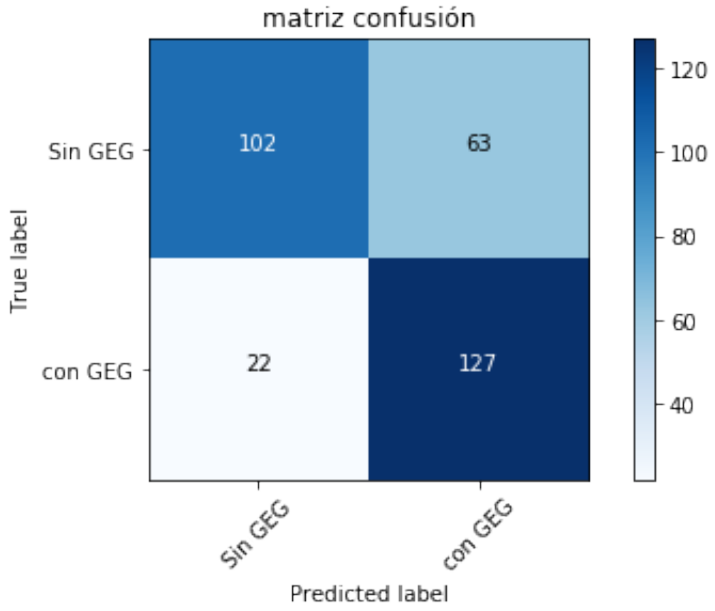

Figura 7.27: Matriz de confusión con bosque aleatorio para la predicción de GEG.

En los recién nacidos grandes para la edad gestacional, como se observa en la figura 7.27, los falsos positivos aumentan un poco más en comparación con otras variables a 63 casos. Sin embargo, la clasificación correcta, tanto en los pacientes sin GEG y con GEG, esto debido a que el tamaño de entrenamiento para está variable es mayor.

7.5.3. Método de SHAP

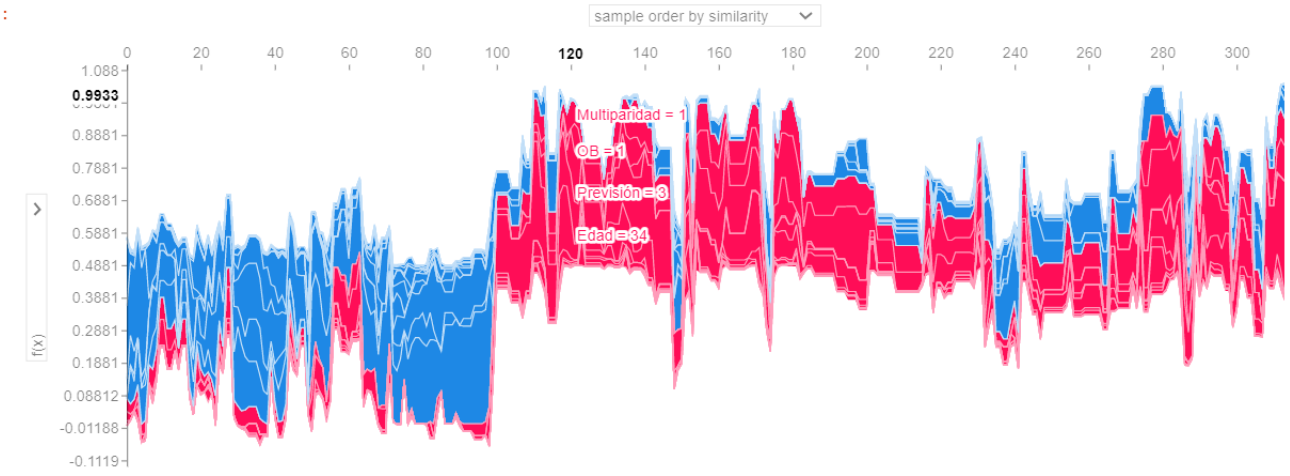

Figura 7.28: Método de SHAP para la predicción de GEG con variables de riesgo.

En la figura 7.28 la combinación de variables como la multiparidad, la obesidad y la edad de la paciente actúa como factores de aumento de la complicación con una probabilidad de 0,727 de que

el feto nazca grande para su edad gestacional.

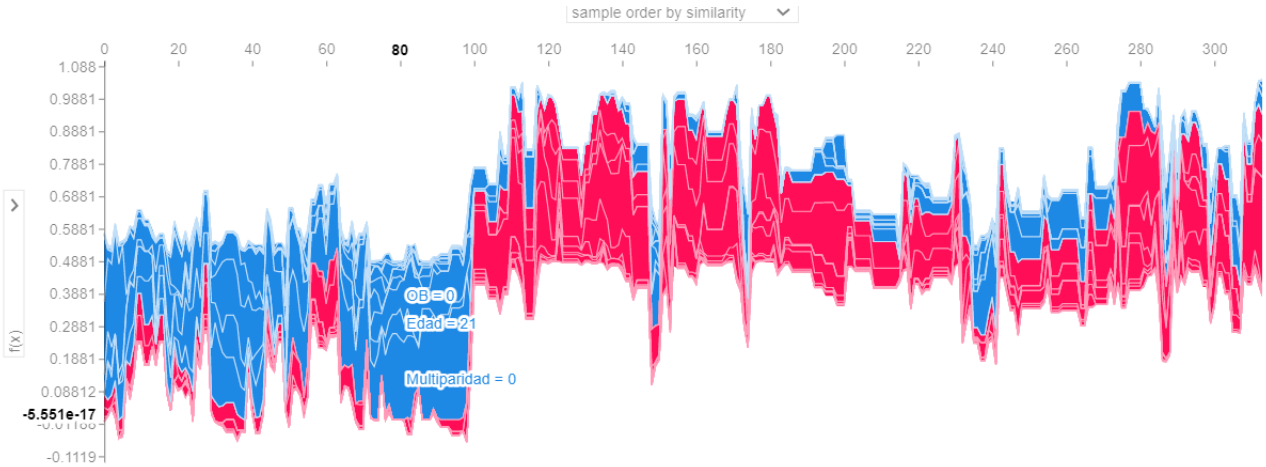

Figura 7.29: Método de SHAP a la predicción de GEG con variables protectoras.

Por el contrario, una paciente que no presenta obesidad ni a sufrido partos previos y además es joven, la probabilidad de que el feto nazca grande para la edad gestacional son considerablemente bajas.

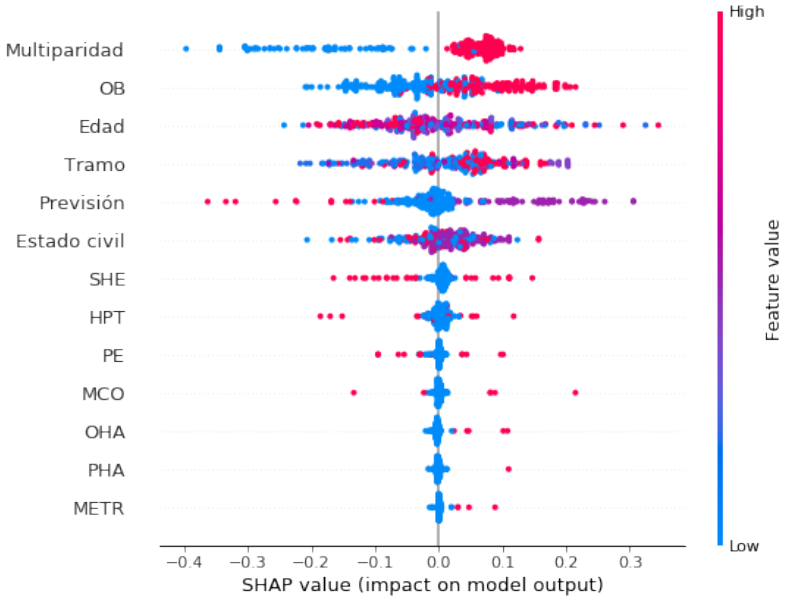

Figura 7.30: Método de SHAP a la predicción de GEG.

Para finalizar con esta variable respuesta, en la figura 7.30 de impacto de SHAP se observa que las variables multiparidad y obesidad son la que generan mayor impacto en la aparición de la complicación.

7.6. Histerectomía

7.6.1. Árbol de decisión

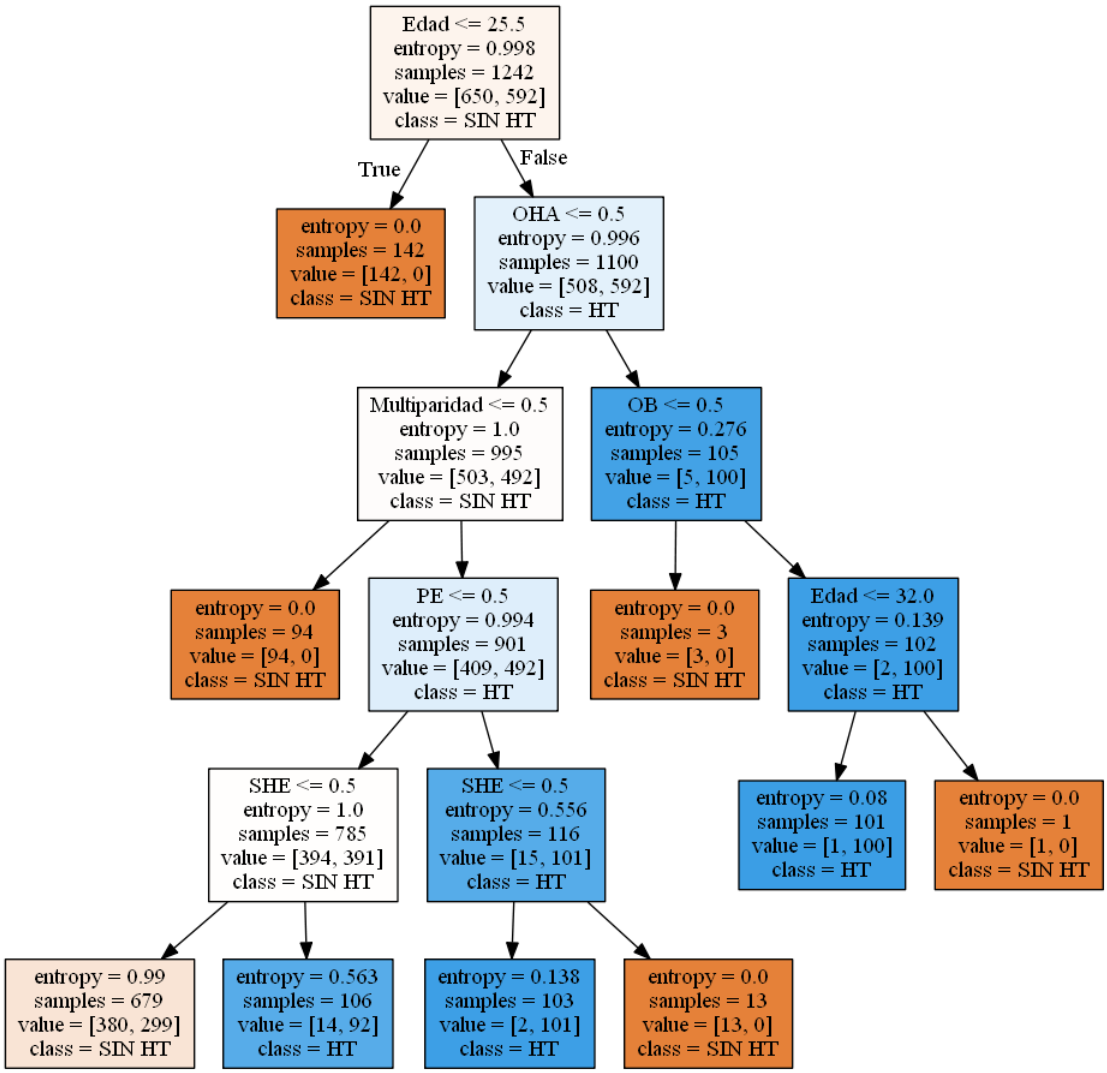

Figura 7.31: Árbol de decisión variable Histerectomía.

En la histerectomía la muestra era más pequeña que el resto de las variables. En el árbol que se muestra en la figura 7.31 vemos que si la gestante tiene menos de 25 años, con la muestra actual del conjunto de datos, no tendría histerectomía. Sin embargo, teniendo más de 25 años y además padecer de oligohidramnios, obesidad, multiparidad y preeclampsia el riego de tener histerectomía aumenta considerablemente.

Exactitud	Precisión	Sensibilidad	Medida F	Error
0.75	0.91	0.51	0.66	0.24

Cuadro 7.11: Medidas de desempeño de árboles de decisión para histerectomía.

7.6.2. Mejor Clasificador

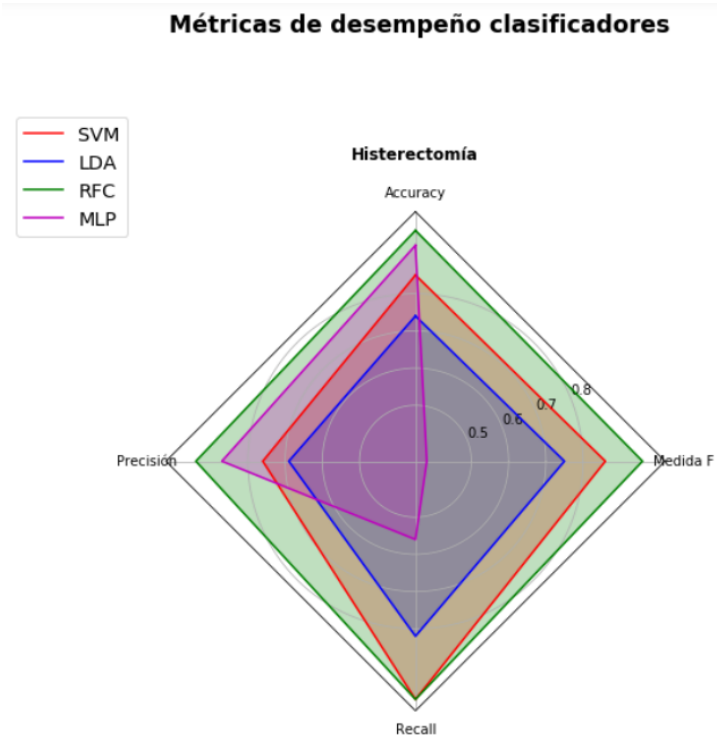

Figura 7.32: Comparación de métricas de desempeño para la clasificación de HT.

Comparación de clasificadores variable HT.				
Métrica	SVM	LDA	RFC	MLP
Exactitud	0.85	0.74	0.97	0.93
Precisión	0.76	0.69	0.94	0.87
Sensibilidad	0.99	0.82	0.99	0.56
Medida F	0.86	0.75	0.96	0.38

Cuadro 7.12: Comparación de clasificadores para la predicción histerectomía.

	precision	recall	f1-score	support
0	1.00	0.98	0.99	210
1	0.98	1.00	0.99	205

Figura 7.33: Métricas de desempeño con bosque aleatorio para la predicción de HT.

El desempeño de la mayoría de los clasificadores fue bueno. Pese a ello el que arrojó peores resultados fue MLP mientras que el clasificador con mejor desempeño fue nuevamente bosque

aleatorio.

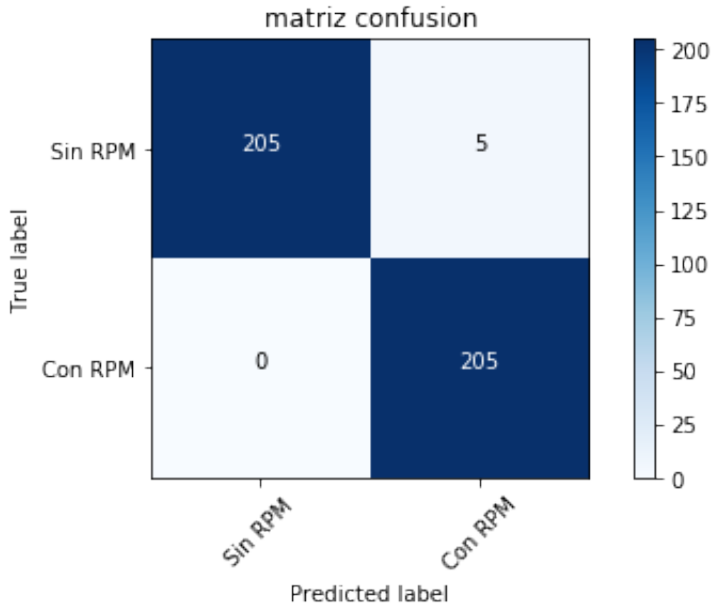

Figura 7.34: Matriz de confusión de bosque aleatorio para predicción de histerectomía.

Se comprueba el buen desempeño del clasificador en la matriz de confusión de la figura 7.34 donde sólo en 5 casos se realiza una clasificación errónea.

7.6.3. Método de SHAP

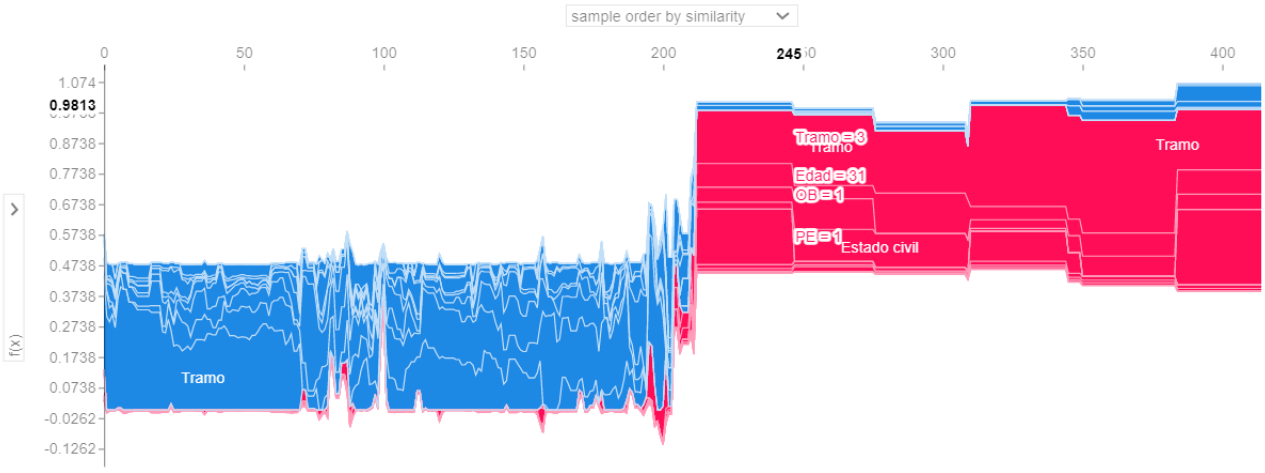

Figura 7.35: Método de SHAP a la predicción de histerectomía.

Dentro de los factores que aumentan más la probabilidad de histerectomía es posible encontrar la obesidad, la edad mayor a 26 años, la preeclampsia, el nivel socioeconómico y el estado civil de la paciente.

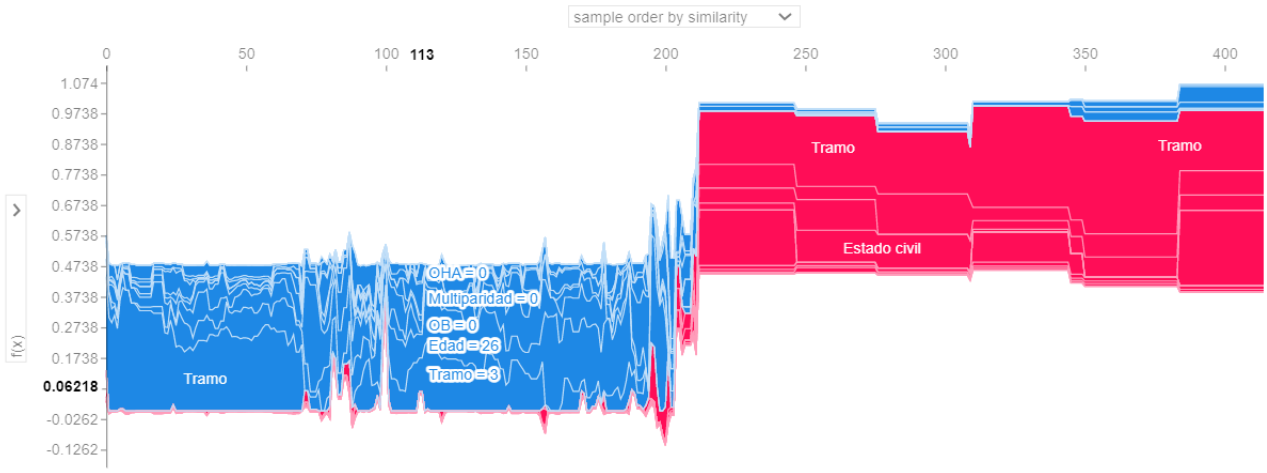

Figura 7.36: Método de SHAP a la predicción de Histerectomía.

En caso contrario se observa que una gestante joven sin oligohidramnio, sin obesidad ni hijos previos al parto, posee una baja probabilidad de histerectomía.

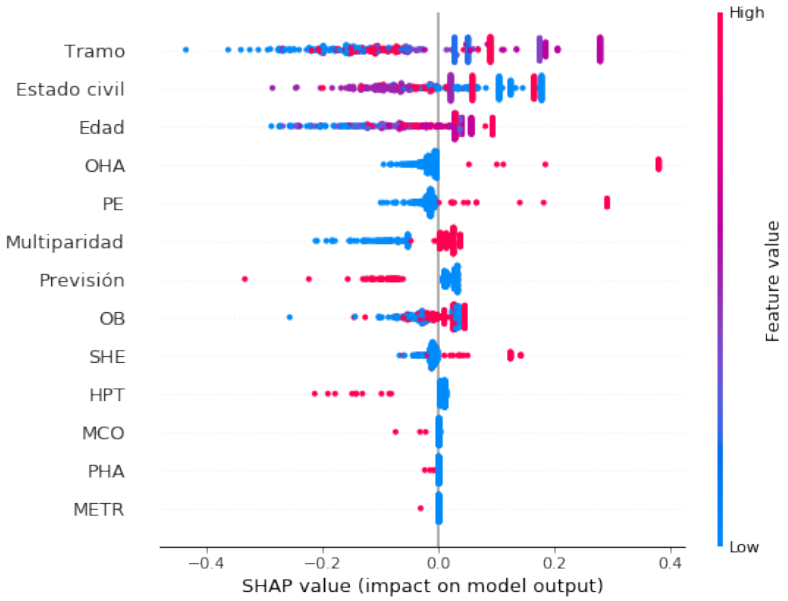

Figura 7.37: Método de SHAP a la predicción de Histerectomía.

Para finalizar con esta variable respuesta, en la figura 7.37 se observa que las variables de mayor impacto en el modelo son la multiparidad, la obesidad, el tramo asociado al nivel socioeconómico, la edad, el síndrome hipertensivo en el embarazo y el estado civil, principalmente en pacientes casadas o convivientes.

7.7. Ruptura Prematura de Membrana

7.7.1. Árbol de decisión

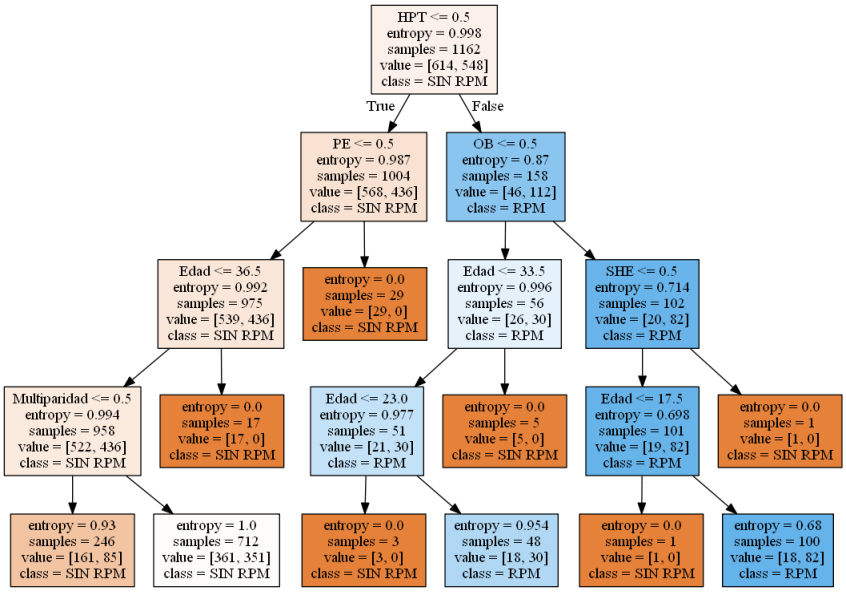

Figura 7.38: Árbol de decisión para la ruptura prematura de membrana.

Para la ruptura prematura de membrana uno de los factores de riesgo más importantes es el hipotiroidismo. En un árbol de 4 hojas en 7 nodos la clasificación termina con ruptura prematura de membrana como por ejemplo en la situación de una paciente con hipotiroidismo en conjunto con síndrome hipertensivo en el embarazo y una edad mayor a 17 años.

Exactitud	Precisión	Sensibilidad	Medida F	Error
0.60	0.56	0.82	0.67	0.29

Cuadro 7.13: Medidas de desempeño de árbol de decisión para la ruptura de membrana.

7.7.2. Mejor Clasificador

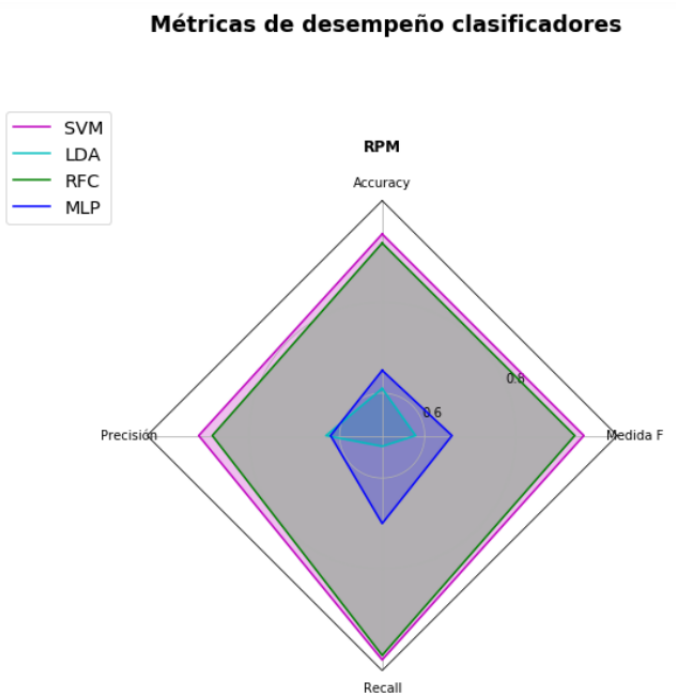

Figura 7.39: Comparación de métricas de desempeño para la clasificación de RPM.

Comparación de clasificadores variable RPM				
Métrica	SVM	LDA	RFC	MLP
Exactitud	0.95	0.61	0.93	0.65
Precisión	0.91	0.63	0.88	0.62
Sensibilidad	0.99	0.53	0.99	0.70
Medida F	0.95	0.62	0.70	0.66

Cuadro 7.14: Comparación de métricas de desempeño en clasificadores para la predicción de ruptura prematura de membrana.

	precision	recall	f1-score	support
0	1.00	0.91	0.95	158
1	0.91	1.00	0.95	152

Figura 7.40: Medidas de desempeño SVM para la predicción de RPM

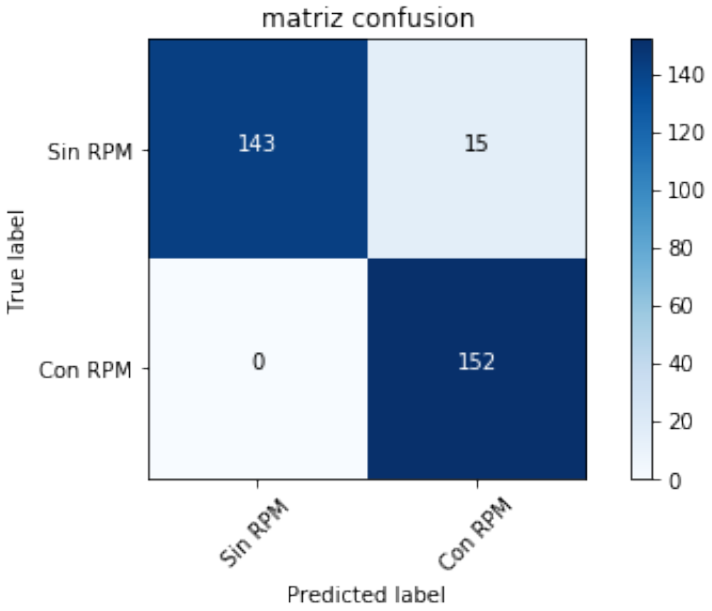

Figura 7.41: Matriz de confusión para predicción de RPM.

Por una leve diferencia en las métricas de desempeño el clasificador con mejores resultados fue SVM pese a que el bosque aleatorio también obtuvo resultados significativos. Además, se observa los bajos rendimientos para LDA y MLP para esta variable dependiente. Por otro lado, la matriz de confusión indica que solo en 15 situaciones se clasificaron erróneamente los pacientes con ruptura prematura de membrana.

7.7.3. Método de SHAP

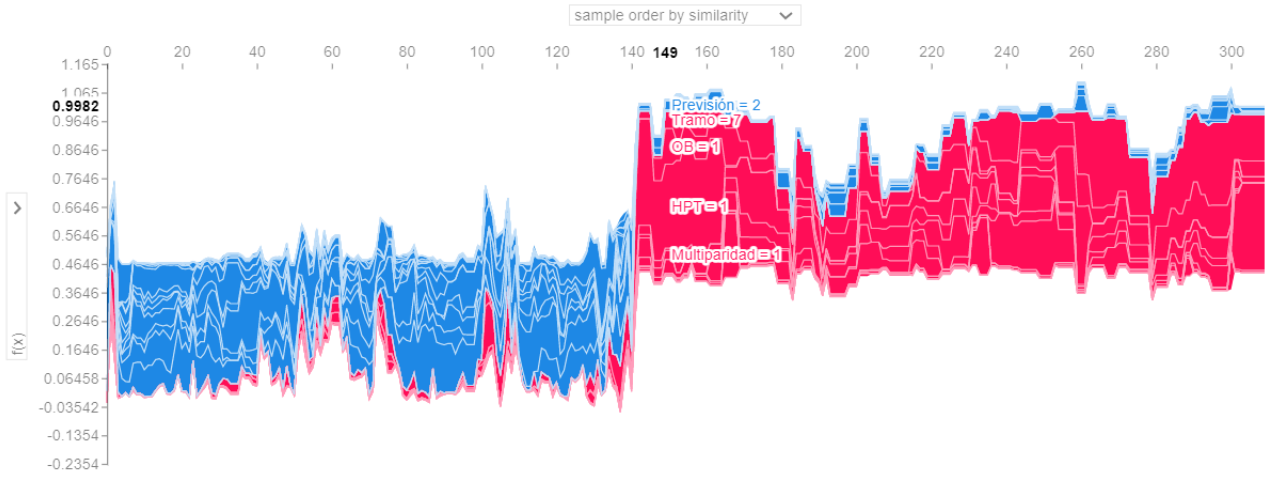

Figura 7.42: Método de SHAP a la predicción de RPM

En la figura 7.42 es posible apreciar que si la paciente sufre de obesidad en conjunto con hipotiroidismo y además tuvo partos previos, lo más probable es que pueda sufrir ruptura prematura de membrana.

Figura 7.43: Método de SHAP a la predicción de RPM

Ahora evaluando desde una perspectiva contraria, la probabilidad de RPM disminuye cuando la paciente es joven, no posee obesidad ni multiparidad.

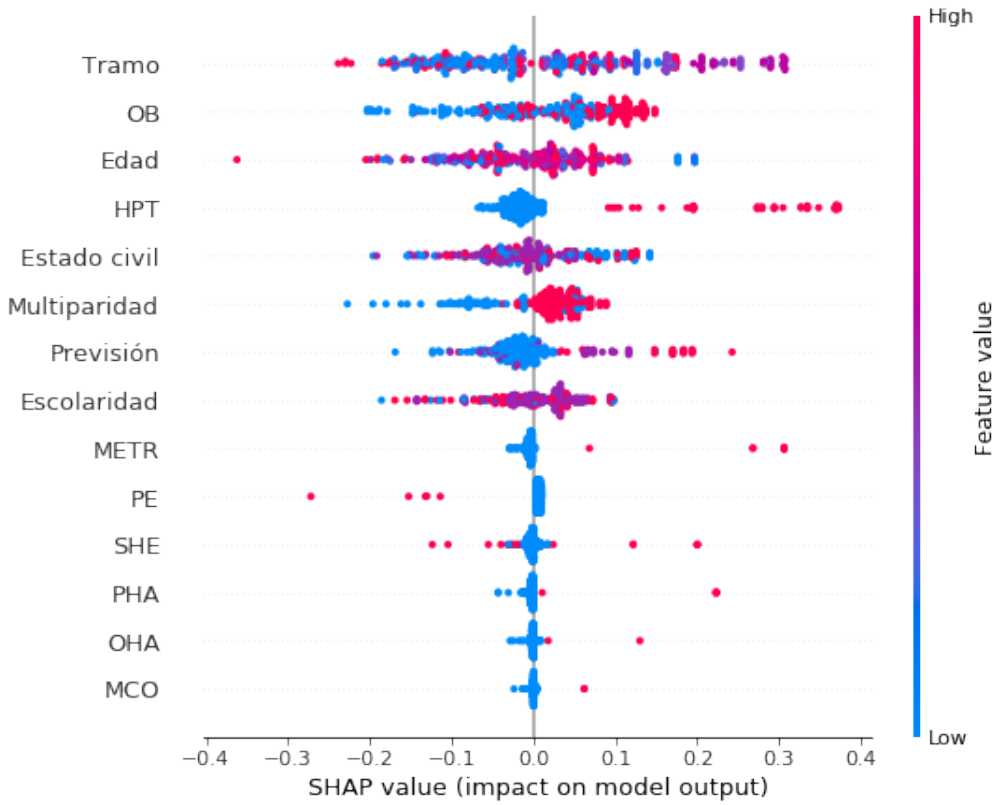

Figura 7.44: Método de SHAP a la predicción de Meconio

En la figura 7.44 es posible observar que la obesidad, el hipotiroidismo, la edad y la multiparidad son los factores que están mayormente correlacionados con el aumento en la probabilidad de ruptura prematura de membrana.

7.8. Macrosomía

7.8.1. Árbol de decisión

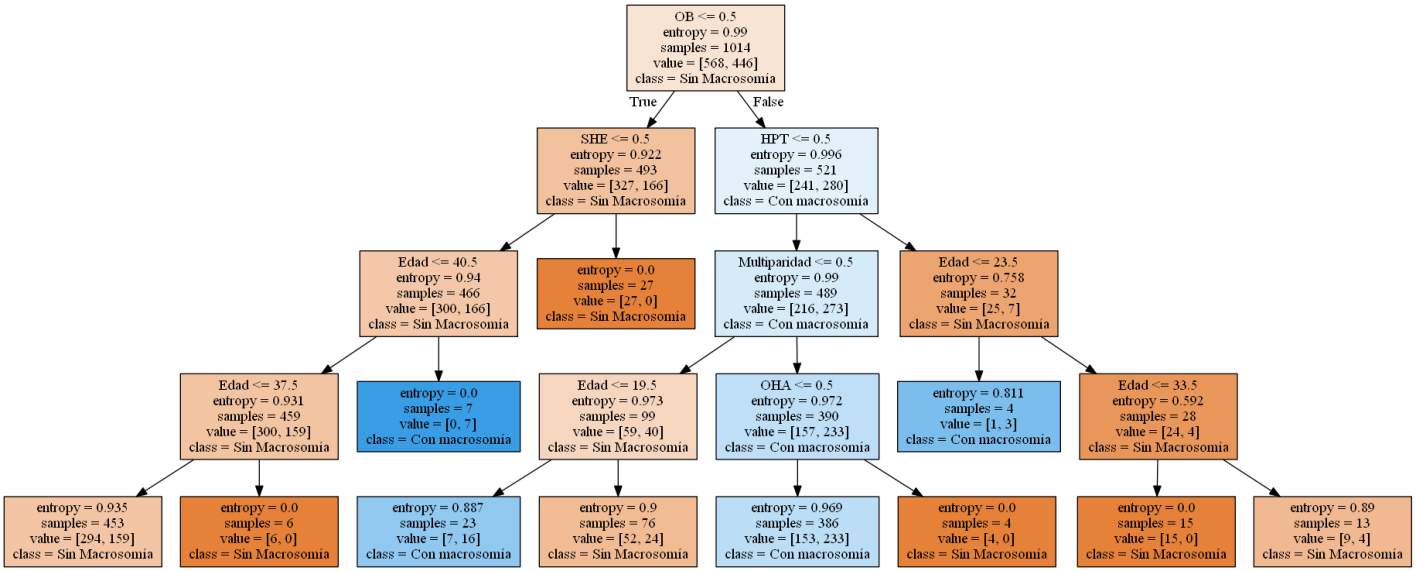

Figura 7.45: Árbol de decisión para Macrosomía.

Para finalizar se encuentra la variable respuesta *Macrosomía* la cual está indicada en recién nacidos de un peso mayor a 4,000 gramos. Para está variable se aprecia que el nodo principal está compuesto por el factor de obesidad.
El árbol cuenta de 4 hojas de las cuales en 7 nodos la clasificación finaliza en macrosomía. Un ejemplo de ello son las pacientes con obesidad, hipotiroidismo, multiparidad y oligohidramnios. Por otro lado, se aprecia que pese a no padecer obesidad ni síndrome hipertensivo en el embarazo si la gestante tiene una edad avanzada mayor a 40 años también es posible que el bebé padezca de macrosomía.

Exactitud	Precisión	Sensibilidad	Medida F	Error
0.66	0.62	0.58	0.60	0.33

Cuadro 7.15: Medidas de desempeño de árbol de decisión para macrosomía.

7.8.2. Mejor Clasificador

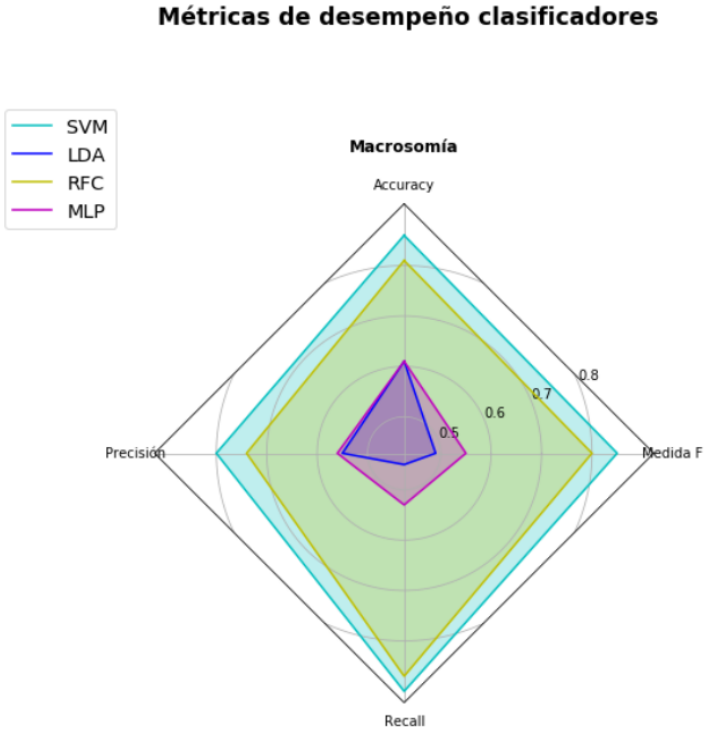

Figura 7.46: Comparación de métricas de desempeño para la predicción de macrosomía.

Comparación de clasificadores variable Macrosomía.				
Métrica	SVM	LDA	RFC	MLP
Exactitud	0.86	0.61	0.81	0.61
Precisión	0.80	0.55	0.74	0.56
Sensibilidad	0.90	0.45	0.87	0.53
Medida F	0.85	0.49	0.80	0.55

Cuadro 7.16: Comparación de métricas de desempeño en clasificadores para la predicción de macrosomía.

	precision	recall	f1-score	support
0	0.92	0.83	0.87	191
1	0.81	0.91	0.85	148

Figura 7.47: Métricas de desempeño con SVM para la predicción de macrosomía .

Al igual que en variables anteriormente mencionadas, los desempeños más bajos se dan con los clasificadores LDA y MLP mientras que los resultados más altos se obtuvieron con SVM y bosque

aleatorio de los cuales el primero presenta las mejores métricas.

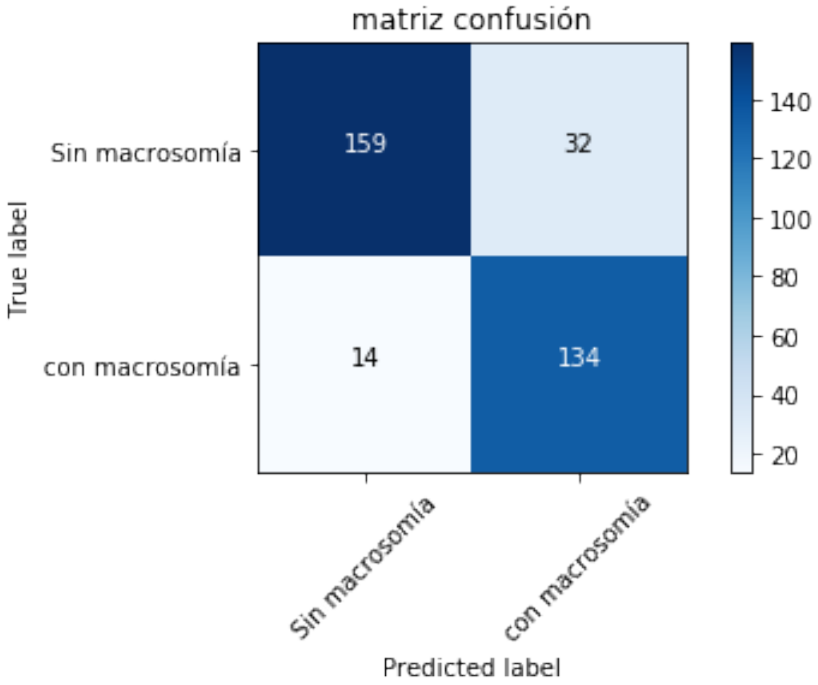

Figura 7.48: Matriz de confusión para predicción de macrosomía.

7.8.3. Método de SHAP

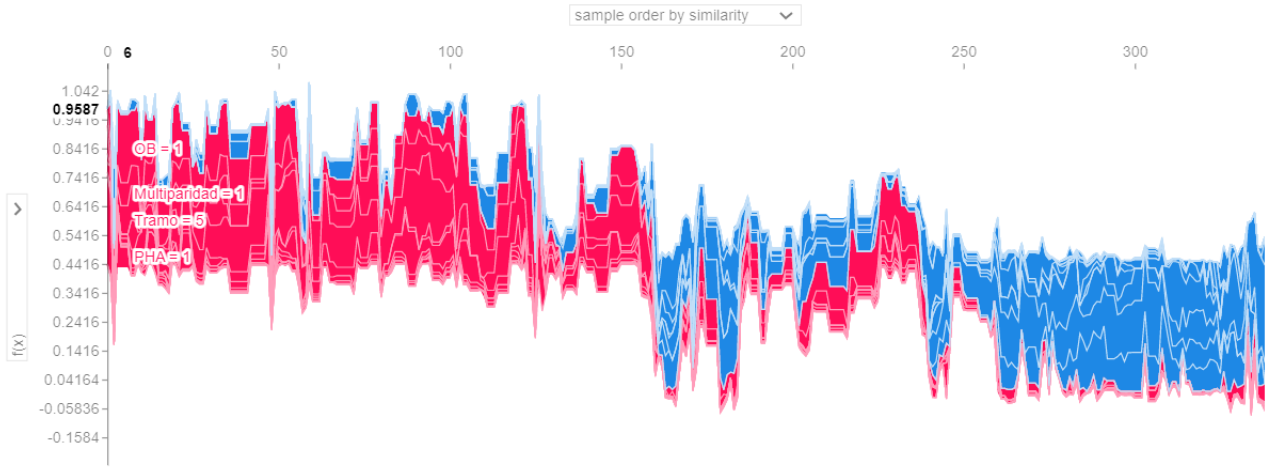

Figura 7.49: Método de SHAP a la predicción de macrosomía.

En el caso de la predicción de macrosomía fetal la interacción de los factores es más variable que en características anteriores. En el caso puntual presente en la figura 7.49 se puede ver como la obesidad sigue siendo un factor de aumento en la predicción junto con la multiparidad y el polihidramnios con una probabilidad del 0,958.

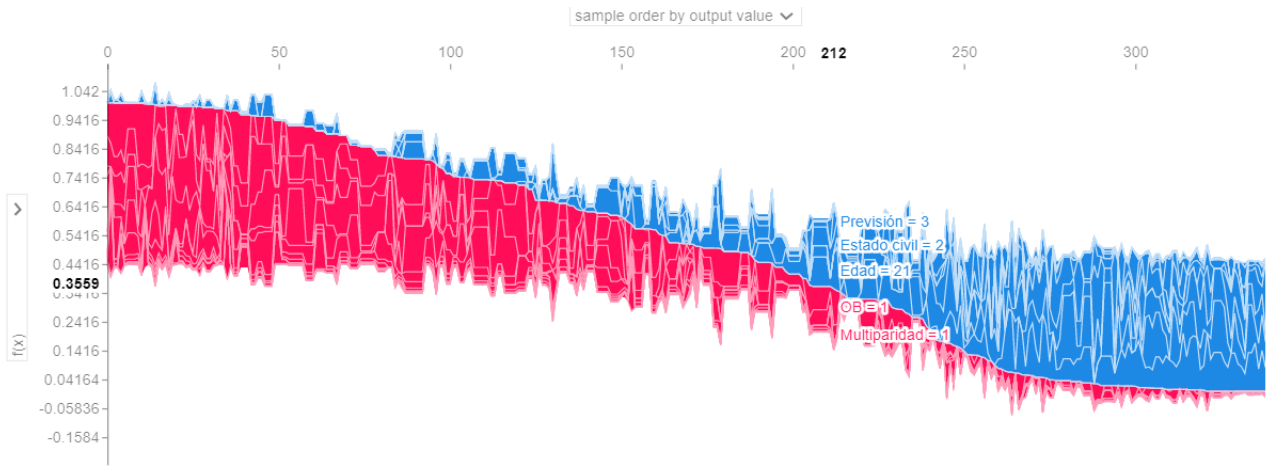

Figura 7.50: Método de SHAP a la predicción de macrosomía.

Por otro lado, las variables sociodemográficas actúan como factores de disminución de probabilidad. En la figura 7.50, como se mencionó anteriormente se repite el factor de obesidad y multiparidad como variables de riesgo y por ende de aumento en el peso del modelo al momento de la predicción aunque que pese a estar presentes ambas características en una gestante joven y con previsión de ISAPRE la probabilidad de macrosomía es de 0,355.

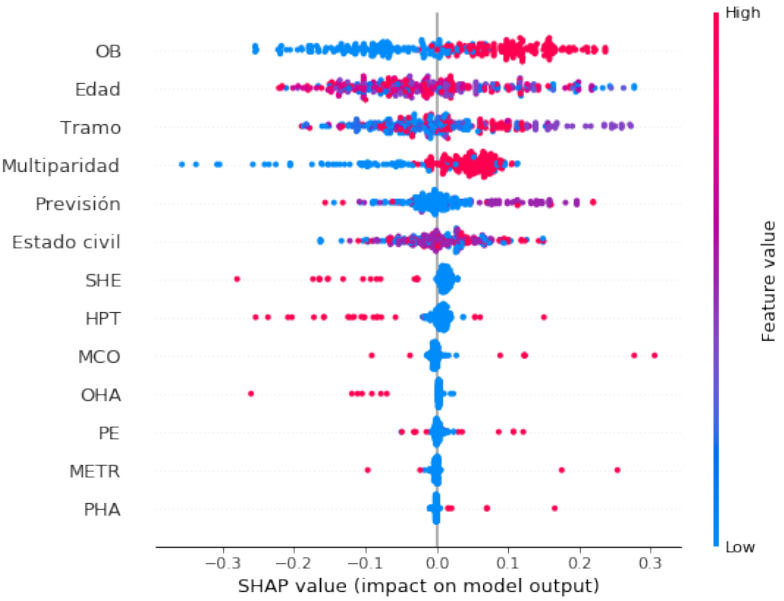

Figura 7.51: Método de SHAP a la predicción de macrosomía.

Por último, se identificaron las variables de mayor impacto asociadas a la predicción de la macrosomía en la figura 7.51 donde se observa nuevamente que la obesidad es uno de los factores influyentes al momento de la predicción de la complicación en conjunto con la edad, la multiparidad y otras complicaciones clínicas como la preeclampsia y las malas condiciones obstétricas.

Capítulo 8

Conclusión

Del trabajo realizado se identificaron, en primera instancia, las variables que se asocian directamente con las pacientes diabéticas. Además, fue posible apreciar que cada una de las variables dependientes necesitaban un tratamiento diferente tanto para el balanceo, como en el tamaño de entrenamiento. Así mismo iban surgiendo distintos tipos de problemas en la aplicación de los modelos, como por ejemplo, el overfitting, más aún cuando la muestra es pequeña y en clasificadores como los árboles de decisión. Pese a ello y mediante la técnica de *pruning* en conjunto con el balaceo de los datos y la aplicación de validación cruzada, fue posible aplicar los árboles de decisiones sin caer en el sobreajuste y con un buen desempeño.

Por otro lado, se dedujo que la predicción con bosque aleatorio presentó los mejores resultados en casi todas las variables esto debido principalmente a la naturaleza de los datos y a que en su mayoría se tenía variables binarias. Cabe recalcar que tanto para SVM como para bosque aleatorio, la mejora en el desempeño del modelo dependía en gran parte de un correcto ajuste de los hiperparamétros.

Uno de los objetivos principales es conocer la interacción de las variables al momento de la predicción. La importancia de ello viene dada por lo difícil que es interpretar los modelos compuestos por "cajas negras"que predominan en machine learning. Sin embargo, mediante la implementación del método de SHAP se pudo conocer el impacto y la interacción de las variables clínicas y sociodemográficas, está información es realmente útil principalmente en el área de la medicina para observar el comportamiento de los factores de riesgo en el embarazo. Se puede deducir que en términos generales, la obesidad y la edad son las variables que más se repetían en la mayoría de las predicciones. No obstante, cada variable tenía factores de alto impacto distintos al momento de la predicción dependiendo de la naturaleza de la variable. Las variables que más se asociaban al momento del parto, como por ejemplo la cesárea y la histerectomía, presentaban un mayor impacto con variables sociodemográficas en comparación con otras complicaciones clínicas.

En términos generales, el trabajo realizado cumplió con todos sus objetivos y se obtuvieron resultados que buscan ser un aporte para el área de la medicina, los cuales tienen la finalidad de reducir las complicaciones sufridas en el embarazo ayudando a prevenir futuros problemas que pueden estar relacionados y de los cuales es posible tener una predicción adecuada que permita conocer la interacción de las variables en las pacientes con diabetes.

Capítulo 9

Referencias

- Adimabua, A., & Ekurume, E. (2021).Predictive Intelligent Decision Support Model in Forecasting of the Diabetes Pandemic Using a Reinforcement Deep Learning Approach,*Education and Management Engineering.* 40-48. doi:$https://10,5815/ijeme$,2021,02,05

- Alpaydin, E.(2014). Introdyction to machine learning, third edition, *Massachusetts Institute of Technology, 3rd edition.*

- Araya, J., Rodríguez, A., Lagos-San Martin, K., Mennickent, D., Guitiérrez-Vega, S., Ortega, B., Valderrama, B., Gonzalez, M., Farías, M., & Guzmán, E. (2021). Maternal thyroid profile in first and second trimester of pregnancy is correlated with gestational diabetes mellitus through machine learning,*Placenta (103)*, 82-85.
doi: https://doi.org/10.1016/j.placenta.2020.10.015

- Bertini, A., Chabert, S., Sobrevia, L., Pardo, F., & Salas, R.(2021). *Using machine learning to predict complications in pregnancy: a systematic review., Chile.* Metabolic Diseased Research Laboratory.

- Borja, R., Monleón, A., Rodellar, J. (2020). Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning. *Revista Ibérica de sistemas y tecnologías de información.* pp: 172-184,.

- Chiefari, M., Arcidiacono, B., Foti, D., & Brunetti, A. (2017). Gestational diabetes mellitus: an updated overview. *J Endocrinol Invest.*899-909. doi: $10,1007/s40618 - 016 - 0607 - 5$

- Chowriappa, P., Dua, S., & Todorov, Y. (2014). Introduction to machine learning in healthcare informatics, *Machine Learning in Healthcare Informatic* . 1-2.
doi: $https://doi.org/10,1007/978 - 3 - 642 - 40017 - 9_1$

- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous Science of Interpratable Machine Learning. *Stat ML, V2. doi: arXiv:1702.08608*

- Gandhi, P. (2019). Explainable Artificial Intelligence. Recuperado de:
$https://www.kdnuggets.com/2019/01/explainable - ai.html$

- Garmendia, M., Mondschein, S., Montiel, B., & Kusanovic, J. (2019). Trends and predictors of gestational diabetes mellitus in Chile.*International journal of gynecology obstetrics.* 210-2018.
doi: 10.1002/ijgo.13023.

- Ghojogh, B. & Crowley, M.(2019). The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and boosting: Tutorial. *Machine Learning* , doi: $arXiv$: 1905,12787v1.
- Giucidi, P. & Raffineti, E.(2021). Shapley-Lorenz eXplainable Artificial Intelligence,*Expert Systems with Applications* , vol. 167, doi: $https$: $//doi.org/$10,1016/$j.eswa$,2020,114104
- Grabczewski, K.(2014). Meta-learning in decision tree induction, Vol 489, pp. 16.
- International Diabetes Federation. (2019). *Atlas de la diabetes de la IDF*. Novena edición 2019. 4-30.
- Instituto nacional de estadísticas. (2018). *Resultados CENSO 2017*. http://resultados.censo2017.cl/
- Joseph, A.(2019). Shapley Regressions: A framework for statistical inference on machine learning models. *Bank of England and King's college London*. No. 2019/7. recuperdo de :$Github.com/Bank-of-England/Shapleyregressions$.
- Knapi, S., Malhi, A., Saluja, R., Framling, K., (2021). Explainable Artificial Intelligence for Human Decision Support System in the Medical,*Machine Learning and Knowledge Extraction* doi: 10,3390/$make$3030037.
- Konenenko, I. & Kukar, M. (2007). Machine learning and data mining: Introduction to principles and algorithms. *Horwood Publishing Chichester*.
- Liu, Y., Yu, Z., & Sun, H. (2021). Prediction Method of Gestational Diabetes Based on Electronic Medical Record Data, *Journal of Healthcare Engineering.* *doi: httpps* : $//doi.org/$10,1155/2021/6672072
- Lopez, R., Reich, J., Mezura, E., Cruz, M. (2021). Inducción de árboles de decisión como modelos de clasificación mediante metaheurísticas, *Swarm and Evolutionary Computation*, doi: $https$: $//doi.org/$10,1016/$j.swevo$,2021,101006.
- Makariou, P., Barrieu, Y. (2021).Un enfoque aleatorio basado en el bosque para predecir los diferenciales en el mercado primario de bonos catastróficos. Matemáticas y Economía Volumen 101, Parte B, pp: 140-162.
- Mercado, D., Pedraza, L., Martínez, E.(2015). perceptron multipaca: Comparison of Neural Network applied to prediction of Time Series.
- Michie, D., Spiegelhalter, D., Taylor, C.(1994). Machine learning, neural and statistical classification, *University of Strathclyde*, doi: $https$: //10,1,1,27,355
- Miller, T.(2017). Hacia una ciencia rigurosa del aprendizaje automático interpretable, *ML: 1-13*. doi: $arXiv$: 1706,07269.
- Ministerio de Salud de Chile. (2017). *Día Mundial de la Diabetes*. Recuperado de https://www.minsal.cl/dia-mundial-de-la-diabetes/.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning. *The MIT press, second edition.*

- Molnar, C. (2018). Interpretable Machine Learning, A guide for making Black Box Models Explainable. *.Lean Publishing*. Recuperado de: http://ganj-ie.iust.ac.ir:8081/images/6/69/Interpretable-machine-learning.pdf
- Murphy, K.(2012). Machine Learning ,A Probabilistic Perspective, 2012 Massachusetts Institute of Technology. *Massachusetts Institute of Technology*.
- Opěl, P., Schindler, I., Kawulok, P., Kawulok, R., Rusz, S., Navrátil, H.(2021). Varias redes neuronales artificiales basadas en funciones de perceptrón multicapa y base radial en el proceso de descripción de una curva de flujo caliente Los enlaces de autor abren el panel de superposición, No. 20, Vol 14, pp:1837-1847 doi: $https://doi.org/10,1016/j.jmrt$,2021,07,100
- Samuel, A. (1959). Eight-move opening utilizing generalization learning. *IMB journal*(pp. 218- 220).
- Shapley, L., & Rigby, F.(1959). Equilibrium points in games with vector payoffs. *doi: https://doi.org/10.1002/nav.3800060107*
- Shen, Z. (2020). 3 min of Machine Learning: Cross Vaildation, url: $https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-vaildation/$
- Siddegowda, N., & Puttabuddi, M. (2020).Analysis of Association between Caesarean Delivery and Gestational Diabetes Mellitus Using Machine Learning , *Proceedings of Engineering and Technology Innovation. (15)*.08-15. doi:$https://doi.org/10,46604/peti$,2020,4740
- Sullivan, C.(2017). Interpretable vs Explainable Machine Learning, *Towards Data Science*, url: $https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48$
- Taha, A., Haim, A., Karakis, I., Shashar, S., Biederko, R., Shtein, A., Hershkowits E.,& Novack, L.(2021). *Air pollution and meteorological conditions during gestation and type 1 diabetes in offspring*. Environment International. doi:$https://doi.org/10,1016/j.envint$,2021,106546
- Rodríguez, G. (2007). Logit Models for Binary Data. *Princeton University, doi:* $10,1007/978-0-387-73186-5-9$
- Probst, P., Bischl, B., Boulesteix, A. (2018). Sintonización: importancia de los hiperparámetros de los algoritmos de aprendizaje automático preimpresión. doi: $arXiv$: 1802,09596.
- Vitoriano, B. (2007). Teoría de la decisión: Decisión con incertidumbre, Decisión Multicriterio y Teoría de juegos. *Complutense Madrir*.
- World Helth Organization. (2006). *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia*.Report of a WHO Consultation.
- Ye, Y., Xiong, Y., Zhou, Q., Wu, J., Li, X. & Xiao, X.(2020).Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning, *The Journal of Clinical Endocrinology & Metabolism (106)*. 1191-1205. descargado de: $https://https://academic.oup.com/jcem/article/106/3/e1191/6031346$

- Ye, Y., Xiong, Y., Zhou, Q., Wu, J., Li, X. & Xiao, X.(2020).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.* doi:$https://doi.org/10,1155/2020/4168340$
- Yu, G., Jin, M., Huang, M., Aimuzi, R., Zheng, T., Nian, M., Tian, Y., Wang, W., Luo, Z., Shen, L., Wang, X., Du, Q., Xu, W., & Zhang, J.(2021).Environmental exposure to perfluoroalkyl substances in early pregnancy, maternal glucose homeostasis and the risk of gestational diabetes: A prospective cohort study, *Environment International. doi: https* : $//doi.org/10,1016/j.envint,2021,106621$
- Zhihan, L.(2020). Security of Internet of Things edge devices,*Software: Practice and Experience / Early View.* 1-11 . doi: $https://doi.org/10,1002/spe,2806$