UNIVERSIDAD AUTÓNOMA DE GUERRERO

UNIDAD ACADÉMICA DE INGENIERÍA

TRABAJO DE INVESTIGACIÓN

USO DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE COMORBILIDAD PERINATAL EN EMBARAZOS DE ALTO RIESGO

QUE PRESENTA

ING. FRANCISCO ABAD NAVA

PARA OBTENER EL TÍTULO DE

MAESTRÍA EN INGENIERÍA PARA LA INNOVACIÓN Y DESARROLLO TECNOLÓGICO, OPCIÓN TERMINAL: TIC´S

DIRECTOR DE TRABAJO DE INVESTIGACIÓN
DR. GUSTAVO ADOLFO ALONSO SILVERIO

CO-ASESOR
DR. IRIS PAOLA GUZMÁN GUZMÁN

CHILPANCINGO, GUERRERO, AGOSTO DE 2019

CONTRAPORTADA
DEDICATORIA

AGRADECIMIENTOS

ÍNDICE

INTRODUCCIÓN

JUSTIFICACIÓN

Durante el periodo de gestación, una mujer puede presentar diversas complicaciones relacionadas a comorbilidades perinatales, por ejemplo, Preeclampsia, Macrosomía, Bajo Peso o Diabetes gestacional, que en el peor de los casos pueden poner en riesgo la vida de la madre o del neonato (Esparza et al., 2018). El monitoreo y diagnóstico oportuno de alteraciones metabólicas o clínicas pueden evitar el desarrollo de escenarios adversos asociados con estas comorbilidades perinatales.

Estas comorbilidades han sido asociadas a factores de riesgo tales como exceso de peso materno, características sociodemográficas, nutricionales, incluso genéticas, sin embargo, la mayoría de los estudios en los cuales se intenta determinar cuáles son los factores determinantes para cada patología, normalmente se abordan la asociación de variables en modelos de regresión logística univariada, pero no desde la perspectiva de interacción multivariable, (Akhtar et al., 2018).

Según Martínez et al.,(2015), un embarazo de alto riesgo es aquel en el que existen probabilidades de muerte por parte de algún integrante del binomio o en su defecto el desarrollar alguna enfermedad. Según la Organización Mundial de la Salud (OMS), una mujer embarazada en un país en vías de desarrollo tiene 100 a 200 veces más riesgo de morir que una mujer en un país desarrollado.

En la actualidad el uso de herramientas tecnológicas, como los algoritmos de aprendizaje automático o redes neuronales, podrían ser implementados en la predicción de comorbilidades, realizando análisis a los principales factores de riesgo usando combinaciones multivariable para desarrollar herramientas tecnológicas de apoyo en la práctica clínica del personal médico, (Kumru et el., 2016).

En este trabajo de investigación se plantea el uso de algoritmos de aprendizaje automático y de plataformas disponibles de software para el proceso de datos, en la construcción de algoritmos predictores de morbimortalidad perinatal, así como el diseño de herramientas tecnológicas de bajo costo y elevada utilidad para profesionales del sector salud.

ALCANCES

Se obtendrá una metodología como herramienta auxiliar en la determinación de los factores de riesgo asociados con la morbilidad perinatal, a través de la implementación de inteligencia artificial.

OBJETIVOS

Objetivo General

Modelar datos de mujeres con embarazos de alto riesgo usando algoritmos de aprendizaje automático, para identificar las combinaciones de factores asociados al riesgo de presentar comorbilidad perinatal.

Objetivos Específicos:

1.- Identificar los principales factores de riesgo reportados en el estado del arte
2.- Comparar el rendimiento de diferentes clasificadores usando combinaciones de factores de riesgo
3.- Implementar el mejor clasificador en una aplicación del tipo prospectiva
4.- Realizar pilotajes con nuevos datos, para de validar la herramienta

CAPÍTULO I. MARCO TEÓRICO E HIPÓTESIS

Los hábitos alimenticios de la población mexicana se caracterizan por el consumo de una dieta rica en grasas y en azucares, que aunado a factores sociodemográficos que colocan a la población en desigualdad de oportunidad, posicionan a México como uno de los países con mayor obesidad en el mundo (Romo et al., 2017).

Se tienen registros que alrededor del 90 % de los casos de diabetes tipo 2 (DT2) se relacionan a la obesidad, siendo la prevalencia y la incidencia de casos mayor en la población femenina en rangos de edad reproductiva, mujeres que en estado de gestación resulta complejo mantener un peso o ganancia de peso adecuado (Hamann, 2017).

1.1 Periodo de gestación y riesgo de diabetes gestacional

El periodo de gestación es una etapa de desafío en la vida de las mujeres, puesto que pone a prueba su cuerpo para adaptarse a uno o varios cambios que se presentan. Estos cambios van desde aumento de la circulación sanguínea, mayor esfuerzo muscular e inclusive un cambio en la estructura ósea. Todos estos cambios representan dificultades si no están preparadas adecuadamente. No sólo el estado físico cambia sino también el estado anímico, un aumento del estrés, la ansiedad, etc., son respuestas que deben atenderse (Nereu et al., 2016). Según Nava et al., (2011), el embarazo es considerado por su naturaleza un estado diabetogénico e iniciarlo con problemas de sobrepeso u obesidad origina un aumento a la resistencia de la insulina, lo que ocasiona agotamiento de la capacidad de las células ß para secretar la insulina requerida por este proceso, y aumenta considerablemente el riesgo de desarrollar otro tipo de comorbilidades.

Una de las principales comorbilidades en el embarazo es la diabetes gestacional (DG); comúnmente, la diabetes es considerada una enfermedad crónica que aparece cuando el páncreas no puede producir suficiente insulina, una hormona clave en el proceso de regular la concentración de glucosa en la sangre, por lo que la baja capacidad para responder a la insulina o su producción conduce a hiperglicemia (OMS, 2016).

Sin embargo, la diabetes gestacional refiere al estado hiperglicémico que se reconoce por primera vez durante el embarazo, y que puede o no revertirse posterior al parto (Carriel, 2017). La desaparición de la DG post parto se logrará mayormente en aquellas mujeres que siguen las recomendaciones de los especialistas. Sin embargo, la presencia de diabetes en el embarazo representa mayores riesgos durante o después del parto para el binomio madre-neonato en comparación con un embarazo normal (OMS, 2016). Las comorbilidades que se pueden presentar son muerte materna, preeclampsia, sangrado mayor a 400mL, EHE, óbito, SDR, macrosomía y bajo peso neonatal (Zaragoza *et al*., 2017, en proceso de publicación). Es por ello la importancia de un diagnóstico y seguimiento oportuno. Otras investigaciones afirman que la diabetes gestacional afecta entre el 3-6 % de las mujeres gestantes (Carriel, 2017). Este mismo autor apoya la hipótesis de que el diagnóstico oportuno es vital para darle un seguimiento y tratar de mantener los niveles de glucosa dentro del rango aceptable.

Existen diferentes factores de riesgo que pueden desencadenar en diabetes gestacional, estos van desde sobre peso, obesidad, antecedentes familiares, hábitos alimenticios, hábitos de higiene, la edad en la que se presenta el periodo de gestación y los embarazos previos, entre otros (Trujillo, 2016). El tratamiento a seguir en las mujeres que han sido diagnosticadas con diabetes gestacional es cuidar su alimentación, así como hacer ejercicio, con el objetivo de regular sus niveles de glucosa en sangre, y de acuerdo a las guías del tratamiento de la diabetes, el tratamiento con metformina o insulina puede ser necesario, aunque es importante considerar los efectos adversos relacionados al tratamiento (Trujillo, 2016).

Diversos autores están de acuerdo en que la diabetes gestacional es un padecimiento para tomar serias consideraciones, que de no hacerlas conlleva a poner en peligro la vida de la madre o del neonato durante el periodo de gestación o durante el parto y existe suficiente información para afirmar que afecta la salud del binomio aún después del parto.

1.2 Estados Hipertensivos del Embarazo Los estados hipertensivos del embarazo (EHE) son una de las complicaciones con mayor repercusión en la salud materna, y continúan siendo una de las cuatro primeras causas de mortalidad en la mujer embarazada, tanto en países desarrollados como en vías de desarrollo. Son la principal causa de morbilidad potencialmente grave, generalmente transitoria, pero con riesgo de secuelas permanentes como las alteraciones neurológicas, hepáticas, hematológicas o renales (González et al., 2013). Un estado hipertensivo, se diagnostica cuando las cifras tensionales están por encima de 140/90mm de Hg, a partir de la semana 20 de gestación, en una paciente previamente normotensa, sin proteinuria (Huarte et al., 2009).

Los trastornos hipertensivos en el embarazo han recibido diferentes denominaciones, como toxemia gravídica, gestosis e hipertensión gestacional (Beltrán et al., 2014). La EHE es una entidad compleja y multisistémica, diversos modelos han intentado explicar su patogénesis. Una de la hipótesis postula que se debe a una respuesta inmune de la madre, ante el estímulo alogénico del feto y la reducción de perfusión de oxígeno placentario por vasoespasmo arterial, provocando una invasión anormal de tejido trofoblástico en la pared uterina, en la semana 12-13 de gestación (Beltrán et al., 2014). Según la guía rápida publicada por el Instituto Mexicano del Seguro Social (2017) la enfermedad hipertensiva del embarazo se clasifica en:

- Hipertensión Preexistente
  - Con condición co-morbida
  - Con evidencia de Preeclampsia
- Hipertensión Gestacional
  - Con condición co-morbida
- Preeclampsia
  - Preeclampsia con datos de severidad
  - Efecto hipertensivo transitorio
  - Efecto hipertensivo de “la bata blanca”
  - Efecto hipertensivo enmascarado

1.3 Preeclampsia

La preeclampsia es uno de los trastornos hipertensivos del embarazo que representa elevados índices de mortalidad materna en países en vías de desarrollo y se caracteriza por presentar una presión arterial igual o superior a 160/110 mm Hg, y síntomas como cefalea, visión borrosa, fosfenos, dolor en el flanco derecho, vómito, proteinuria (Beltrán et al., 2014), incluso edema, que aparece después de la semana 20 de gestación (Barrera et al., 2013).

La preeclampsia es considerada una enfermedad grave; que puede evolucionar a eclampsia, síndrome de HELLP y finalmente la muerte (Camacho et al., 2015). Se estima que la preeclampsia afecta entre el 3-7% de todas las gestaciones, principalmente en los países en vías de desarrollo, en los cuales ocasiona el 15% de las muertes maternas en comparación con los países industrializados (Valdivia 2016). Si bien la muerte neonatal es una de las principales complicaciones asociadas a preeclampsia, también pueden ocurrir nacimientos pretérminos, restricciones en el crecimiento fetal y muerte materna (Rodríguez, 2017), siendo está la de mayor impacto en salud.

La preeclampsia es un problema de salud a nivel mundial, sin embargo, los países en vías de desarrollo son los mayormente afectados. Un estudio realizado en Perú reportó que la preeclampsia es una de las principales causas (40 a 60%) de desprendimientos de la placenta (Nuñez, 2017). En la misma población se encontró que de los hijos nacidos de madres con preeclampsia severa, síndrome de HELLP y eclampsia, 21.4% fueron prematuros, 28% tuvieron bajo peso al nacer y el 31% fue pequeño en función de su edad gestacional. En México, de acuerdo a la OMS, entre el 10 y el 14% de las embarazadas sufre preeclampsia, con una tasa de decesos de 4 mil muertes maternas y 20 mil de neonatos cada año, lo que la convierte en la primera causa de muerte materna, muerte fetal u óbito en el país.

Si bien la preeclampsia es un padecimiento común del embarazo, en presencia de DG puede ser de mayor riesgo. En un estudio realizado en el Hospital de la Madre y el Niño en la ciudad de Chilpancingo Guerrero, México, un grupo de 96 mujeres con DG, atendidas en el año 2015, se encontró una prevalencia de preeclampsia de 11.5% (Zaragoza et al., 2017).

1.4 Óbito

El Óbito ha sido definido por la OMS como la muerte del producto dentro del vientre de la madre, antes de que éste sea expulsado o extraído de su gestante, o al momento de la labor de parto por alguna circunstancia que involucre muchas veces al equipo que realiza la tarea de asistencia obstétrica a la madre (Moreno, 2016). El óbito afecta a 3,6 millones de familias a nivel mundial, las mujeres son las más afectadas cuando sucede óbito en su embarazo (Díaz, 2017). El óbito es una de las más devastadoras circunstancias, tanto para la madre como para los especialistas que atienden el caso (Moreno, 2016). Según Díaz (2017), uno de los factores a los cuales se relaciona el óbito fetal es a la disminución o ausencia del flujo sanguíneo hacia el útero, o se debe a que la placenta se ha desprendido de manera anticipada.

Huerta et al. (2017), en un estudio llevado a cabo entre los años 2014 y 2015 en el Instituto Mexicano del Seguro Social, en el Hospital General Regional No.17 de Quintana, Roo, reportaron que de 7170 partos, 43 casos fueron óbito, la mayoría con edad gestacional pretérmino (33.6±4.7 semanas), esto es que seis de cada 1000 nacidos vivos terminan en muerte fetal y que además la población de mujeres con una edad superior a 35 años es considerada de alto riesgo para muerte fetal.

Si se considera la tasa de muerte fetal a nivel nacional, en México, ésta representa el 18.5 y 20.8 por 1000 nacidos vivos. La muerte fetal conocida como óbito, presenta dificultad de aceptación en el diagnóstico, siendo traumática tanto para la madre, como para la familia y el obstetra, quien debe tomar decisiones inmediatas con adecuada comunicación posterior al diagnóstico. El porcentaje de muertes fetales sin aparente explicación ronda entre el 21 al 50%. En la actualidad ni con la autopsia, o el examen histológico del cordón umbilical, placenta o membranas, logran identificar la causa del deceso (Huerta et al., 2017). El óbito representa un escenario completamente desbastador, y muchas pueden ser prevenibles.

1.5 Síndrome de Deficiencia Respiratoria

El síndrome de deficiencia respiratoria (SDR), también conocido como distrés respiratorio, es causa de más de la mitad de las condiciones patológicas del recién nacido. La dificultad respiratoria se diagnostica clínicamente por la presencia de al menos dos de los siguientes criterios: cuando la frecuencia respiratoria es > 60/min, las retracciones (subcostal, xifoides esternal e intercostal), ensanchamiento de las alas de la nariz, estridor espiratorio y cianosis (Contreras, 2017).

Según Quiroga (2014), el SDR, también conocido como enfermedad de la membrana hialina (EMH), es una de las patologías respiratorias más comunes entre los recién nacidos, más preocupante en aquellos neonatos nacidos a pretérmino. Es causada por la carencia de surfactante, la inmadurez de la estructura anatómica pulmonar y por la incapacidad de mantener una respiración efectiva. Se tienen registros de SDR a comienzos del siglo XX, cuando fue descrito por primera vez por Hochheim para representar el líquido amniótico que era aspirado; los obstetras de la época, así como los pediatras vieron con asombro estás observaciones de los especialistas, pero no fue hasta 1950 cuando se logró distinguir un patrón reticulogranular en la atelectasia neonatal generalizada en los recién nacidos que aspiraban líquido amniótico. Con los trabajos de Clements y Brown(agregar año) se comenzó a demostrar que la baja tensión superficial en los pulmones era indispensable para que éstos desempeñarán adecuadamente sus funciones (Quiroga, 2014).

En 1959, con los trabajos de Avery y Mead, se comprendió completamente la importancia de los hallazgos. En su trabajo "*Surface properties in relation to atelectasis and hyaline membrane disease",* dieron evidencia contundente de que los pulmones de los neonatos con SDR carecían de un material en el alvéolo, el cual es la sustancia (surfactante pulmonar) responsable de la baja tensión superficial (Quiroga, 2014). El SDR es un trastorno de desarrollo que comienza luego del nacimiento en neonatos nacidos a pretérmino, con pulmones inmaduros incapaces de secretar surfactante, se categoriza como una enfermedad compleja y además por atelectasias alveolares difusas en el pulmón, causada principalmente por la carencia de surfactante (Quiroga, 2014).

Las patologías respiratorias son muy frecuentes en los neonatos y representan las principales causas de morbilidad y mortalidad, entre el 2 y 3% de los recién nacidos a

término y en los nacidos en pretérmino representan el 20%. Cuando un feto es pretérmino, si no fallece puede presentar diferentes incapacidades que perjudican la calidad de vida del niño (Poma, 2017).

1.6 Macrosomía

La importancia de la clasificación de los neonatos al nacer radica en determinar el nivel de cuidado según el riesgo de mortalidad y morbilidad neonatal, esto se hace según la relación entre el peso y la edad gestacional (Zavala, 2009). La macrosomía es el término usado para describir a un neonato con peso excesivo (Gonzales, 2012), según Zaragoza et al. (2017) un neonato presenta macrosomía cuando su peso es igual o mayor a 4,000 g, aunque Ramas (2013), expresa que de acuerdo a diversos autores están la definición más acertada es aquella que considera la edad gestacional del feto y el peso neonatal por arriba del percentil 90. La macrosomía neonatal le depara al individuo un futuro complicado a que son candidatos para padecer diversas complicaciones de salud relacionadas al sobrepeso y diabetes en la vida adulta.

En México la macrosomía presenta aproximadamente una prevalencia de 5.4% de los nacimientos, y se asocia factores de riesgo demográficos, fisiológicos, metabólicos, incluso genéticos (García et al., 2016. Entre los factores maternos que influyen para que un neonato presente macrosomía, sobresalen el sobrepeso preconcepcional, la presencia de diabetes gestacional, la estatura materna y edad de la madre, inclusive ser multípara y tener antecedentes de neonatos macrosómicos (Ávila, 2013).

La macrosomía no es fácil de detectar durante el embarazo, a través del ultrasonido se podría detectar un trastorno como éste, sin embargo, para tener buenos resultados en el ultrasonido se requiere personal capacitado para realizarlo. Cuando un feto es diagnosticado con macrosomía, éste puede presentar complicaciones durante el parto, que van desde hemorragias obstétricas en la madre hasta impacto de la cabeza fetal y distocia de hombros entre otros (García et al., 2016). Sin embargo, no sólo el exceso de peso neonatal representa un factor relacionado a la morbilidad perinatal sino también el bajo peso.

1.7 Bajo Peso Neonatal

La Asociación Americana de Pediatría y la OMS establecen que un neonato presenta bajo peso cuando éste es <2,500 g (Zaragoza et al., 2017; Daza et al., 2009). Los neonatos que presentan bajo peso tienen hasta 5 veces más probabilidades de fallecer entre el mes y el primer año de vida, en comparación con los recién nacidos, con peso dentro del rango normal (Flores, 2018), por lo que al nacer el peso constituye uno de los indicadores más útiles para evaluar los resultados de la atención prenatal (Flores, 2018).

Daza et al. (2009), también documentan que los recién nacidos que presentan bajo peso tienen mayores probabilidades de perecer durante los primeros meses de vida; estos investigadores documentaron que las probabilidades son de 5 a 30 veces más que un niño nacido con peso normal. Otros autores refuerzan la hipótesis de que el bajo peso al nacer es un estado delicado. Álvarez (2001), en su trabajo “Repercusión de los Factores de Riesgo en el Bajo Peso al Nacer”, documentó que la mortalidad durante el primer año de vida es 14 veces mayor en niños con antecedentes de bajo peso neonatal con respecto a los que nacen a término y con peso normal. Algunos factores de riesgo asociados al bajo peso del recién nacido son: IMC materno ≥ a 27 kg/m², edad gestacional pretérmino, exposición a humo/polvos, mala higiene, alta concentración de colesterol, edad materna ≥ a 35 años entre otros (Zaragoza et al., (2017). Sin embargo, la predicción de alguno de estos eventos relacionados al embarazo a pesar de conocer múltiples factores de riesgo representa un reto.

1.8 Minería de datos usando métodos de clasificación

La minería de datos es un procedimiento que se aplica a cantidades grandes de datos no triviales con el objetivo de encontrar patrones de comportamiento dentro de ese conjunto de datos. La minería de datos abarca todo un conjunto de técnicas; mediante los modelos extraídos se aborda la solución a problemas de predicción, clasificación y segmentación (Corso, 2009). Puede visualizarse también como un proceso analítico, diseñado para explorar grandes cantidades de datos, con el objetivo de encontrar relaciones entre las diferentes variables, para aplicarlas a nuevos conjuntos de datos (Martínez, 2003).

La minería de datos nació por la idea de aprovechar la gran cantidad de datos que se generan en el comercio, la banca o la sanidad, pero también por la potencia de las

nuevas computadoras para poder realizar un análisis sobre esos datos (Vallejo et al., 2018). Surgió a principios de los 80's en los Estados Unidos de Norte América, como una necesidad de su administración hacendaria para detectar fraudes en la omisión y evasión de pago de impuestos, haciendo uso de lógica difusa, y técnicas de reconocimiento de patrones (Vallejo et al., 2018).

La minería de datos o cómo suele llamarse "*data mining*" es un conjunto de técnicas que hace uso de las tecnologías actuales, con el propósito de encontrar patrones de comportamiento repetitivo, tendencias o reglas en el set de datos no trivial, esto es de manera automática (Hernández, 2017). Según Corso et al., (2014) la minería de datos se define como el proceso de exploración y fundamentalmente el análisis, usando medios automáticos o semiautomáticos para poder manejar grandes volúmenes de información, teniendo como objetivo extraer conocimiento, éste puede darse de manera relacional, en patrones o reglas que son inferidas del conjunto de datos.

Cada conjunto de datos contiene sus propios requisitos, esto es que la técnica para procesar un set de datos A, puede no ser la misma para procesar un conjunto de datos B, por lo tanto, el analista debe conocer bien la naturaleza de los datos y establecer criterios basados en la literatura existente para ese conjunto de datos (Corso et al., 2014). Vallejo et al., (2018), documentan que la minería de datos es un proceso a través del cual se puede identificar información relevante extraída de grandes volúmenes de datos, con el propósito de descubrir patrones y tendencias estructurando la información obtenida. La minería de datos contempla el procesamiento de grandes volúmenes de información, estos datos pueden ser de diversas naturalezas, Los modelos que se pueden generar a través de la minería de datos pueden ser de tipo predictivo o descriptivo (Corso et al., 2014).

Los modelos predictivos tienen como objetivo estimar valores futuros o desconocidos de variables de interés, las cuales se denominan variables objetivo o dependientes, usando otras variables. Los modelos descriptivos tratan de identificar comportamientos que expliquen o resuman los datos, éstos últimos sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos, esta es la diferencia principal (Corso et al., 2014).

Existen otros tipos de técnicas importantes en la minería de datos, éstas son técnicas de agrupamiento y técnicas de asociación. La primera es una técnica de análisis exploratorio de datos que se utiliza principalmente para resolver problemas de clasificación. Consiste en ordenar objetos en grupos, de tal manera que el grado de similitud entre los miembros del mismo grupo sea más fuerte que el de los miembros de grupos diferentes. En las técnicas de asociación, se dice que se utilizan principalmente para buscar, entre el conjunto de datos, reglas que revelan la naturaleza de las relaciones o asociaciones entre datos de las entidades (Corso et al., 2014). En la figura 1.1 se muestran las principales técnicas de minería de datos.

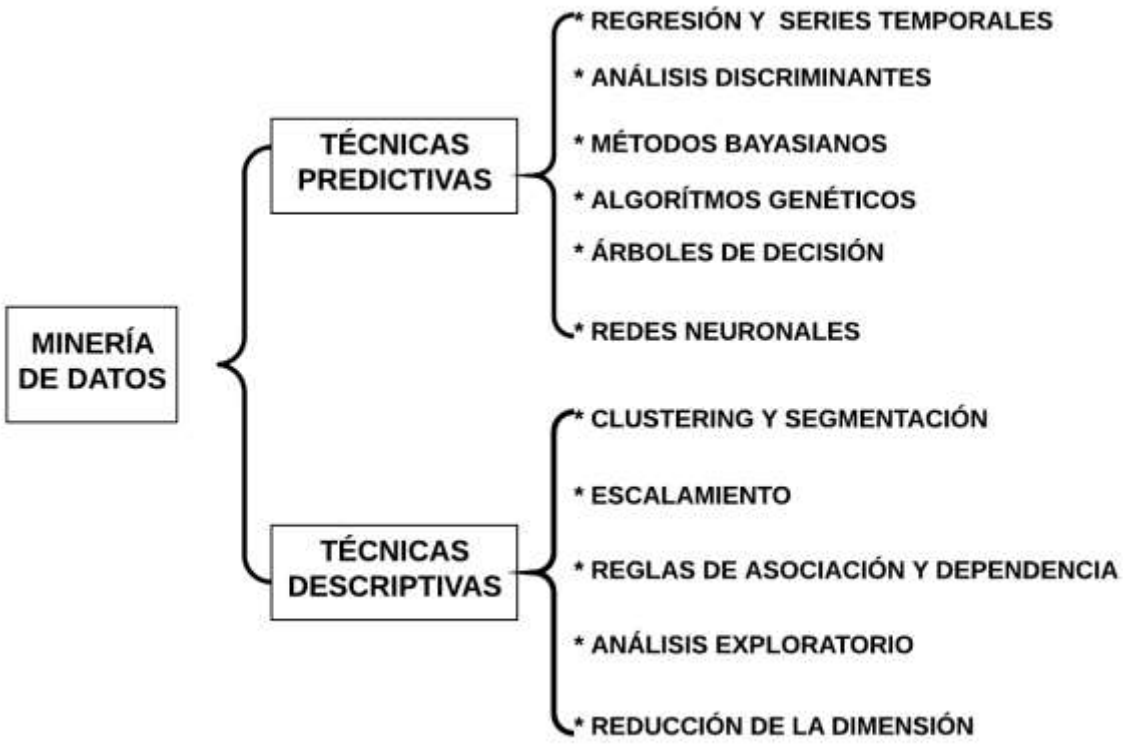


Figura 1.1 Tipos de técnicas de minería de datos

Entre las etapas del proceso de minería de datos resalta la selección del conjunto de datos a procesar, la selección del algoritmo o conjunto de éstos, el entrenamiento con una porción del conjunto de datos, la aplicación, la interpretación a los resultados y al final la aplicación que se le dé a las variables resultantes.

1.9 WEKA

El acrónimo Weka (del inglés *Waikato Environment for Knowledge Analysis*), es un software desarrollado por la Universidad Waikato en Nueva Zelanda en el año de 1993, fue desarrollado en el lenguaje de programación java, con el propósito de ser una herramienta en el proceso de minería de datos.

Weka contiene una amplia colección de algoritmos para realizar análisis de datos y modelado predictivo, y no sólo eso, contiene diversas herramientas que permiten ver la estructura de los datos, mezclar algoritmos y permite encontrar las variables que más peso aportan durante el proceso de clasificación. Esto se logra con la herramienta de selección de atributos. Éstas son sólo algunas de la mucha herramienta que ofrece este software con licencia GPL, que puede ser utilizado sin infringir en alguna falta administrativa (Hernández, 2017).

Weka es una colección de algoritmos de aprendizaje automático y herramientas de minería de datos, fue diseñado para hacer pruebas rápidas y efectivas a grandes volúmenes de datos, ofrece amplio soporte para procesos completos con datos experimentales, eso incluye esquemas de visualización de datos de entrada y salida. Weka contiene una interfaz gráfica amigable e intuitiva que permite ser usado aun cuando no se es experto en el uso de este software. Como ya se mencionó, Weka contiene diversas herramientas que permiten procesar grandes cantidades de datos de diferente naturaleza. También permite hacer uso de herramientas de filtrado de datos, trabajar con clases desbalanceadas y con ausencia de datos en instancias del archivo arff (del inglés, Attribute-Relation File Format), (Frank et al., 2016). En la figura 1.2 se muestra la interfaz principal de Weka al cargar un conjunto de datos.

Figura 1.2 Ventana principal del software Weka

Weka utiliza un tipo de archivo como formato nativo, sin embargo, permite utilizar otro tipo de archivos, como lo son los archivos CSV (del inglés *comma-separated values*). El archivo “arff” contiene una estructura definida y es la que se sigue cuando se estructura un archivo para ser procesado en Weka. En las figuras 1.3 y 1.4 se muestran la apariencia que toma un archivo “arrf”, cuando éste ya está listo para ser procesado.

Cabecera del archivo: La cabecera del archivo incluye la definición del nombre de la relación de datos que contiene el archivo, éste debe ser de tipo cadena, el formato correcto que sigue es el siguiente:

@relation <nombre_de_la_relacion> <tipo_de_dato> Ver figura 1.3

```
@relation 'HIIDTWEKA-weka.filters.unsupervised.instance.ImageFilter.ColorLayoutFilter-DC:\\Program Files\\Weka-3-7-
weka.filters.unsupervised.attribute.Remove-R1'

@attribute edadMat       real
@attribute escolar       real
@attribute ocupacion     real
@attribute antFamDm2     real
@attribute parenDm2      real
@attribute antFamHta     real
@attribute parentHta     real
@attribute higiene       real
@attribute contaminacion real
@attribute pesoMat       real
@attribute tallaMat      real
@attribute IMCMat        real
@attribute IMCat         real
@attribute TAS           real
@attribute TAD           real
@attribute SDG           real
@attribute TxContDG      real
@attribute TotGestas     real
@attribute TotPartos     real
@attribute TotCesareas   real
@attribute TotAbortos    real
@attribute eritrocitos   real
@attribute HBm           real
@attribute plaquetasMat  real
@attribute glucosaMat    real
@attribute colTotMat     real
@attribute PhOrina       real
@attribute gluOrinaMat   real
@attribute edadMatCat2   real
@attribute gestasMatCat  real
@attribute TASAlter      real
@attribute TADAlter      real
@attribute TASAlter2     real
@attribute abortPrevCat2 real
@attribute IMC27         real
@attribute edadCat35     real
@attribute edadCat25     real
```

Figura 1.3. Cabecera de un archivo “arff”

Tipo de datos: Es el apartado en el cual se declaran el tipo de atributos que contendrá el archivo, así como el nombre de la variable que se utilizará durante el proceso, tiene el siguiente formato:

@attribute <nombre-del-atributo> <tipo>

El tipo de dato que se pueden utilizar son, de tipo numérico, entero, tipo fecha, tipo cadena y enumerado, ver figura 1.4 y 1.5.

Figura 1.4 estructura central de un archivo “arff”

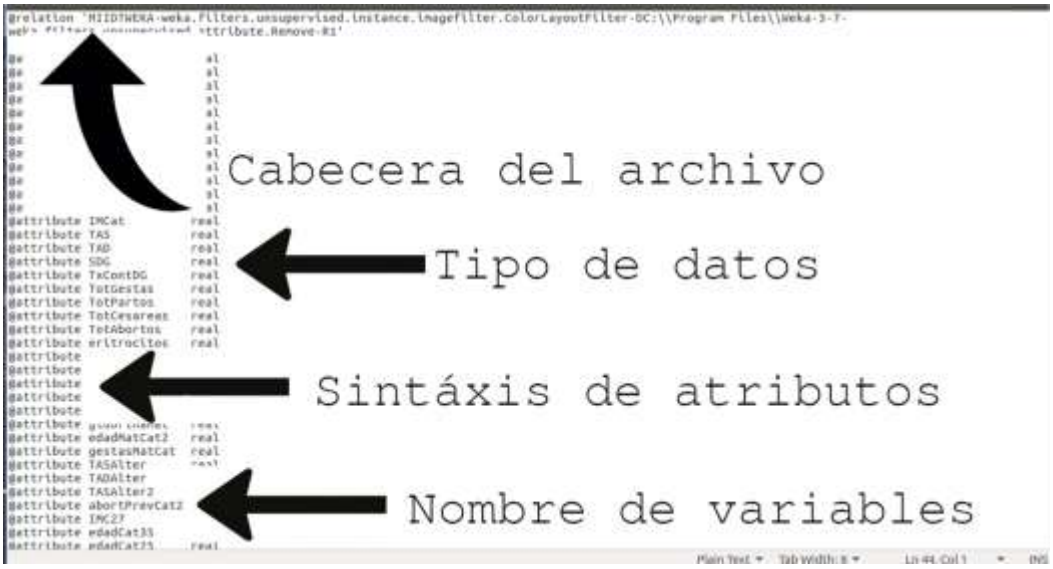


Figura 1.5. Cabecera, los tipos de datos y la sintaxis de un archivo arff

Sección de datos: En esta sección se colocan todos los valores que conforman el universo de datos a procesar, la cantidad de elementos debe coincidir con el total de atributos declarados en el apartado anterior, de lo contrario indicará un error. La clase a la que pertenece el registro siempre se declara al final del mismo registro, ver figura 1.6

Figura 1.6. Entrada de los datos y el espacio que ocupan los datos

1.10 Clasificadores

Existen dos tipos de clasificadores para trabajar con minería de datos Hernández (2017); clasificación supervisada y clasificación no supervisada; en la clasificación supervisada, parte del universo de los datos se conoce la clase a que pertenece y se ocupa un set de entrenamiento o conjunto de aprendizaje, este tipo de clasificación es la que ha sido ampliamente aplicada a diversos ámbitos, entre ellos en el cuidado de la salud para el diagnóstico de enfermedades por citar un ejemplo (Corso, 2009).

Otros autores escriben que en los clasificadores supervisados se tiene un conocimiento a priori y para la tarea de clasificar un objeto dentro de una categoría o clase es necesario contar con modelos ya clasificados, es decir, objetos agrupados que tienen similitudes. En este tipo de clasificadores se pueden diferenciar dos tipos de fases; en la primera se tiene un conjunto de elementos para entrenamiento o aprendizaje y otro para realizar una prueba o test de validación de lo aprendido por el

clasificador, con toda esta información se construye un modelo o regla general para la clasificación (Hernández, 2017).

En clasificación no supervisada, se tiene un conjunto de datos descrito por un conjunto de características propias de los datos, sin conocer a que clase pertenece cada uno de ellos (Corso, 2009). Para los clasificadores no supervisados no se cuenta con conocimiento previo, por lo que se tiene un área de entrenamiento disponible para la clasificación o comúnmente llamada *clustering*, también en la clasificación no supervisada se cuenta con objetos o muestras que presentan un conjunto de características de las cuales se desconoce a qué clase o categoría pertenecen, entonces, en eso radica el propósito de descubrir grupos de objetos que tengan características similares (Hernández, 2017).

En la figura 1.7 se enlistan los clasificadores supervisados más usados

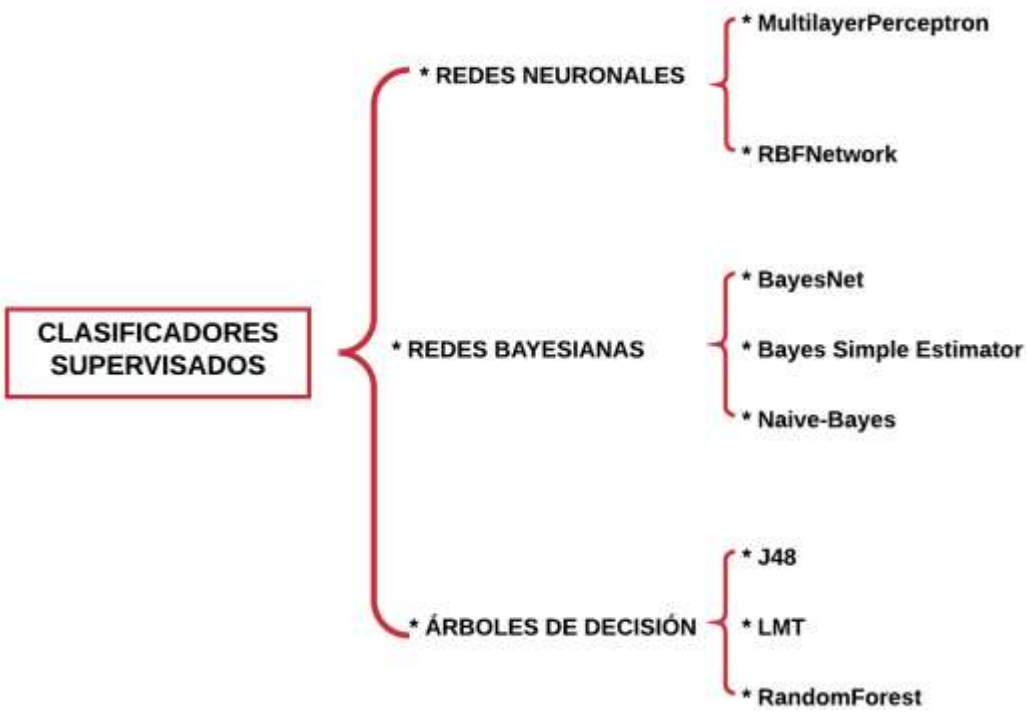


Figura 1.7. Principales clasificadores supervisados

Existe una gran cantidad de clasificadores contenidos en el software Weka. Haciendo uso del software y dependiendo de la naturaleza del conjunto de datos a procesar, se puede elegir el que mejor se desempeñe. A continuación, se describen algunos de los

clasificadores que pueden ser utilizados en Weka (Hernández, 2017). En la tabla 1.1 se describen los principales tipos de clasificadores.

Tabla 1.1 Descripción de los principales tipos de clasificadores supervisados (Hernández, 2017).

Nombre	Descripcíon	Clasificador
Redes neuronales	Inspirado en el comportamiento de las neuronas en el cerebro humano, una red neuronal es utilizada para reconocer patrones de comportamiento, esto incluye imágenes, manuscritos y secuencias de tiempo, es capaz de aprender y mejorar su desempeño.	MultilayerPerceptron RBFNetwork
Redes bayesianas	Este tipo de clasificación son una representación gráfica de las dependencias para el razonamiento pirobalística, en este caso los nodos son variables aleatorias y los arcos representan dependencia directa entre las variables.	BayesNet Bayes Simple Estimador NaiveBayesMultinomialUpdateable Naive-Bayes
Árboles de decisión	Son un conjunto de condiciones o reglas organizadas en estructura jerárquica, esto significa que existe una decisión final que predomina siguiendo las condiciones desde la raíz.	J48 LMT RandomForest DecisionStump HoeffdingTree RandomTree REPTree

MultilayerPerceptron: Es un tipo de red neuronal constituida por nodos, que componen la capa de entrada, así como un conjunto de una o más capas ocultas de neuronas y una capa de salida.

J48: Se trata de un refinamiento del modelo generado a OneR. Existe una mejora moderada y dará la posibilidad de acierto ligeramente superior.

El algoritmo J48 es una implementación del existente C4.5 el cual es un modelo de árbol de decisión, esta implementación produce modelos de árboles para posteriormente tomar una decisión. El J48 genera árboles, cuyos nodos evalúan la existencia o significado de rasgos individuales de la clase (Bhargava et al.,2013).

Los árboles de decisión están construidos de manera descendente eligiendo el atributo más apropiado, se utiliza una medida de la teoría de la información para evaluar las características, lo que proporciona una indicación del “poder de clasificación” de cada una de ellas. Una vez que se elige una característica, los datos de entrenamiento se dividen en subconjuntos correspondientes a diferentes valores de la característica seleccionada, este proceso se repite para cada subconjunto, hasta que una gran proporción de las instancias en cada uno de ellos pertenecen a una sola

clase. El árbol de decisión es un algoritmo que normalmente aprende un alto conjunto de reglas de precisión (Kaur et al.,2014).

LMT: Combina modelos de regresión logística con árboles de inducción. Consiste en una estructura de árbol de decisión estándar con funciones de regresión logística en las hojas.

RandomForest: Este clasificador combina grandes cantidades de árboles de decisión independientes, probados sobre los conjuntos de datos aleatorios con distribución similar.

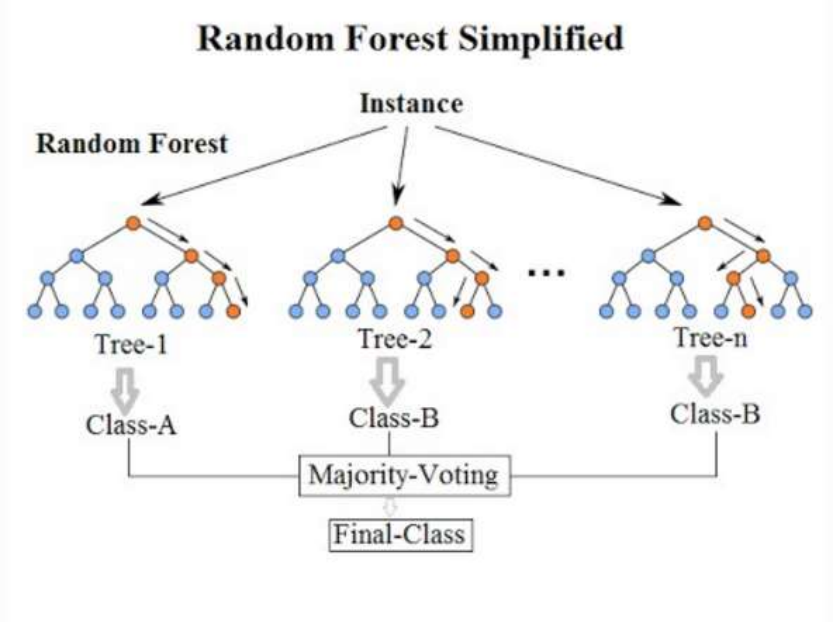


Figura 1.8. Representación gráfica del algoritmo RandomForest

Este clasificador aplica el siguiente proceso:

- Selecciona individuos al azar para crear diferentes conjuntos de datos.
- Crea un árbol de decisión con cada conjunto de datos, con lo cual obtiene diferentes árboles.
- Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad.
- Predice los nuevos datos usando el voto mayoritario, donde clasificará como positivo si la mayoría de los árboles predicen la observación como positiva.

Existen algunos algoritmos que utilizan la experiencia o conocimiento previo como técnica para clasificar una clase, dentro de los algoritmos que hacen uso de esta técnica se encuentran los conocidos como “perezosos”. Una nueva instancia se compara con el resto de la base de casos a través de una medida de similitud o de distancia. Esta nueva clase será categorizada dentro del grupo en donde su cercanía sea menor, a este método se le conoce como “vecino más cercano” (Díaz et al., 2015).

Dentro de este grupo de algoritmos se encuentra el LBR que trabaja para conjuntos de pruebas pequeñas, debido a que cada instancia de prueba selecciona un conjunto de atributos para los cuales la supuesta independencia no debe ser hecha, los demás son tratados como independientes de cada una de las clases dadas y el conjunto de atributos seleccionado.
Existen algoritmos que utilizan funciones de distancia o casos para clasificar una instancia nueva o que no formó parte del aprendizaje. Las instancias son almacenadas en memoria tanto como sea necesario, así cuando una instancia nueva es presentada al modelo se intenta relacionar ésta con las instancias almacenadas, buscando que sean lo más similares posibles. Estos tipos de clasificadores son muy efectivos para trabajar con tipos de datos no estándar (Román,2011).

KStar: K-star o K * es un clasificador basado en instancias. La clase de una instancia de prueba se basa en el entrenamiento de instancias similares a él, según lo determinado por alguna función de similitud. Se diferencia de otros algoritmos basados en instancias en que éste utiliza una función de distancia basada en la entropía. Los algoritmos basados en instancias comparan con una base de datos de ejemplos preclasificados (Cleary et al., 1995).

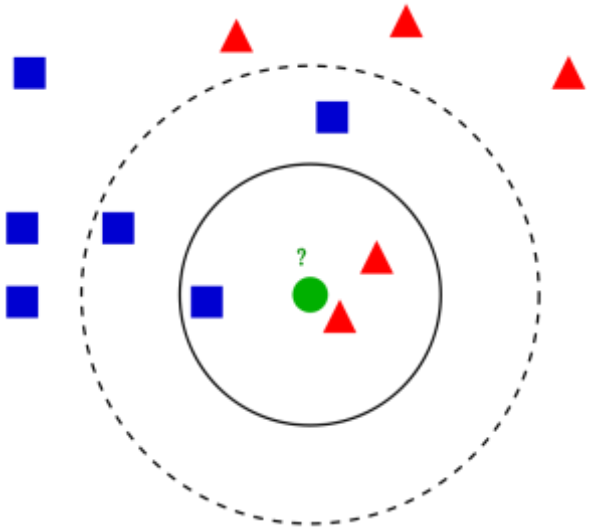


Figura 1.9 Representación gráfica de un algoritmo basado en métricas de distancia

Los componentes correspondientes de un algoritmo basado en instancia son la función de distancia que determina qué tan similares son las dos instancias, y la función de clasificación que especifica cómo las similitudes de la instancia producen una clasificación final para la nueva instancia. Este algoritmo utiliza una medida entrópica, basada en la probabilidad de transformar una instancia en otra al elegir aleatoriamente entre todas las posibles transformaciones. Usar la entropía como un medidor para una distancia de instancia es muy beneficioso y la teoría de la información ayuda a calcular la distancia entre las instancias. La complejidad de una transformación de una instancia en otra es en realidad la distancia entre las instancias (Mahmood et al., 2013).

Algoritmos Bayesianos: Una red bayesiana es un modelo gráfico de probabilidades que representa un conjunto de variables o características y sus dependencias probabilísticas. Puede calcular la distribución de probabilidad para cualquier subconjunto de variables de la red, dado los valores o distribuciones de las variables restantes. Este tipo de clasificador no es muy sensible a los cambios de sus parámetros, ya que se basa en información de toda la base, lo cual hace que pequeños cambios en la base no sean necesariamente significativos. Las Redes Bayesianas representan el conocimiento cualitativo del modelo mediante un grafo dirigido acíclico. Este conocimiento se articula mediante las relaciones de

independencia/dependencia de los atributos o variables que componen el modelo (Roman, 2011).

1.11 Evaluación del desempeño de un clasificador

El software Weka ofrece cuatro modos de pruebas para poder ser aplicados al conjunto de datos (Bouckaert et al., 2010).

Use training set: Con esta opción el clasificador es evaluado en qué tan bueno es su desempeño para predecir las clases con las que entrenó, esto es usando el total de datos disponibles y posteriormente aplicando su aprendizaje sobre el mismo set para prueba.

Supplied test set: El clasificador es evaluado qué tan bueno es su desempeño para predicer la clase, pero esta vez aplicando a un archivo independiente.

Cross-validation: El clasificador es evaluado según su capacidad de clasificación cruzada. Se dividirán las instancias en tantas carpetas como indica el parámetro "Folds", y en cada evaluación se toman las instancias de cada carpeta como datos de test, en las cuales el algoritmo aplicará lo aprendido, y el resto de los datos lo usará como set de entrenamiento para construir el modelo. Se calculará el promedio de los errores de todas las ejecuciones y ese será el resultado del desempeño del clasificador usando esta opción, ver figura 1.8.

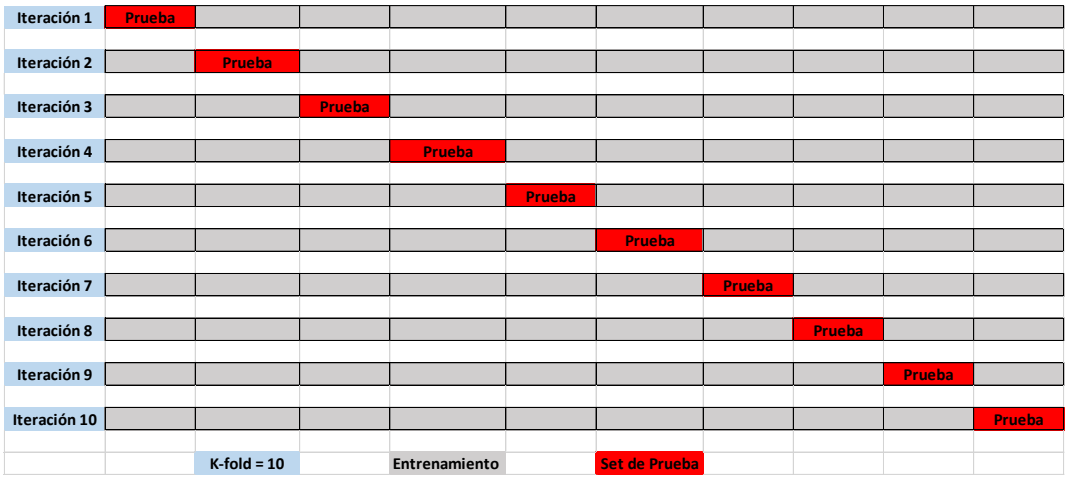


Figura 1.10 Representación gráfica de la técnica de validación cruzada, con K-fold =10.

Percentage split: El clasificador es evaluado según su desempeño para predecir usando un porcentaje de datos para entrenar y el resto lo utiliza para aplicar lo aprendido. El dato se especifica en %, si es 50% eso significa que utilizará la mitad del set de datos para entrenar y el resto de prueba, si es 65% ese porcentaje utilizará para entrenar y el 35% lo usará como set de prueba.

1.12 Matriz de confusión

En el área de la inteligencia artificial, la matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Entre los múltiples beneficios que esta herramienta ofrece es la facilidad para ver si el sistema está confundiendo dos clases.

Si en los datos de entrada el número de muestras de clases diferentes cambia mucho la tasa de error del clasificador no es representativa de lo bien que realiza la tarea el clasificador. La matriz de confusión es particularmente útil cuando en un problema de clasificación se tiene un desbalanceo de clases, una clase mayoritaria de hasta 3 a 1 con respecto a la clase minoritaria, si sólo se utiliza la exactitud o la curva ROC (del inglés, *Receiver Operating Characteristic curve*) para validar el desempeño del sistema, se puede tener una concepción errónea del desempeño, en la exactitud se podrían obtener valores aceptables, lo cual no significa que el algoritmo esté distinguiendo correctamente entre las clases, por lo tanto, es recomendable utilizar la sensibilidad, especificidad para poder tener una idea más real del desempeño de los algoritmos (Patil et al., 2016).

A continuación se describen los 4 posibles resultados que se pueden obtener en un problema de clasificación binaria (Loyola et al., 2014).

Positivos verdaderos (TP): Estos son los valores positivos predichos correctamente, lo que significa que el valor de la clase real es sí y el valor de la clase predicha también es sí. P.ej. si el valor de clase real indica que este pasajero sobrevivió y la clase predicha le dice lo mismo.

Negativos verdaderos (TN): Estos son los valores negativos predichos correctamente, lo que significa que el valor de la clase real es no y el valor de la clase predicha también es no. P.ej. si la clase real dice que este pasajero no sobrevivió y la clase pronosticada le dice lo mismo.
Falsos positivos y falsos negativos, estos valores se producen cuando su clase real se contradice con la clase predicha.

Falsos positivos (FP): Esto sucede cuando la clase real es no y la clase predicha es sí. P.ej. si la clase real dice que este pasajero no sobrevivió, pero la clase pronosticada le dice que este pasajero sobrevivirá.

Falsos negativos (FN): Esto sucede cuando la clase real es sí, pero la clase predicha es no. P.ej. si el valor real de la clase indica que este pasajero sobrevivió y la clase pronosticada le dice que el pasajero morirá.

1.13 Métricas de validación: Sensibilidad, Especificidad, Exactitud, Curva ROC, F-measure y G-mean

Medir el desempeño de un clasificador es de suma importancia cuando se habla de cuestiones médicas se espera que los resultados sean lo más veraces posible, cuando la predicción de un clasificador sea positivo, es decir, que una persona sea diagnosticada como enferma se espera que el margen de error sea el mínimo posible, eso también se espera cuando el resultado sea negativo, pero si una persona es diagnosticada como negativa y en realidad tiene dicha enfermedad, es definitivamente un problema de consideración (Hernández, 2017).

A continuación, se explica la dicotomía de la exactitud de un clasificador.

Cuando se trata de realizar un diagnóstico o aplicar una clasificación sobre el estado de una persona siendo este un estado de salud, se espera una gran precisión para poder tomar decisiones futuras con respecto al estado clasificado, en cuestiones médicas suele utilizarse dos tipos de estados; estar o no estar enfermo, respuesta positiva o negativa, generalmente la exactitud diagnosticada se expresa como sensibilidad y especificidad (López, 1998).

Sensibilidad: Es la probabilidad de obtener un resultado positivo cuando el individuo tiene la enfermedad. Mide su capacidad para detectar la enfermedad cuando está presente, ecuación 1.

$$\textbf{Sensibilidad} = \frac{enfermospositivos}{totalenfermos} = \frac{VP}{VP + FN} \qquad (1)$$

Especificidad: Indica la probabilidad de obtener un resultado negativo cuando el individuo no tiene la enfermedad. Mide su capacidad para descartar la enfermedad cuando ésta no está presente, ecuación 2.

$$\textbf{Especificidad} = \frac{Sanosnegativos}{totalsanos} = \frac{VN}{VN + FP} \qquad (2)$$

El escenario ideal sería tener una sensibilidad y especificidad lo más cercano al 100%. Esto constituye una excepción, pero se debe dudar de todas aquellas pruebas donde los resultados sean menores a 80% (Hernández, 2017).

Precisión: Se refiere a la dispersión del conjunto de valores obtenidos de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Una medida común de la variabilidad es la desviación estándar de las mediciones y la precisión se puede estimar como una función de ella. Es importante resaltar que la automatización de diferentes pruebas o técnicas puede producir un aumento de la precisión. Esto se debe a que, con dicha automatización, lo que logramos es una disminución de los errores manuales o su corrección inmediata. No hay que confundir resolución con precisión, (Mahmood et al., 2013), ecuación 3.

$$\textbf{Precisión} = \frac{VP}{VP + FN} \qquad (3)$$

Exactitud: Se refiere a cuán cerca del valor real se encuentra el valor medido. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. Cuanto menor es el sesgo más exacto es una estimación. Cuando se expresa la exactitud de un resultado, se expresa mediante el error absoluto que es la diferencia entre el valor experimental y el valor verdadero, ecuación 4.

$$\textbf{Exactitud} = \frac{VP + VN}{totalmuestra} \qquad (4)$$

F1-score o F-measure: Media armónica entre *recall* y *precision*.

El puntaje F1 es el promedio ponderado de precisión y recall. Por lo tanto, esta puntuación tiene en cuenta tanto los falsos positivos como los falsos negativos. Intuitivamente, no es tan fácil de entender como la precisión, pero la F1 suele ser más útil que la precisión, especialmente si tiene una distribución de clases desigual. La precisión funciona mejor si los falsos positivos y los falsos negativos tienen un costo similar. Si el costo de los falsos positivos y los falsos negativos es muy diferente, es mejor tener en cuenta tanto la precisión como la retirada, (Mahmood et al., 2013) ecuación 5.

$$\textbf{F1 score} = \frac{2VP}{2VP + FP + FN} \qquad (5)$$

G-mean o G-measure: Media geométrica entre recall y precisión, ecuación 6.

$$\textbf{G mean} = \sqrt{Recall * Precision} \qquad (6)$$

1.14 Curva ROC

La curva ROC (del inglés, *receiver operating characteristic curve*) o (Característica Operativa del Receptor), es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente de (y antes de especificar) el coste de la distribución de las dos clases sobre las que se decide. La curva ROC es también independiente de la distribución de las clases en la población (en diagnóstico, la prevalencia de una enfermedad en la población).

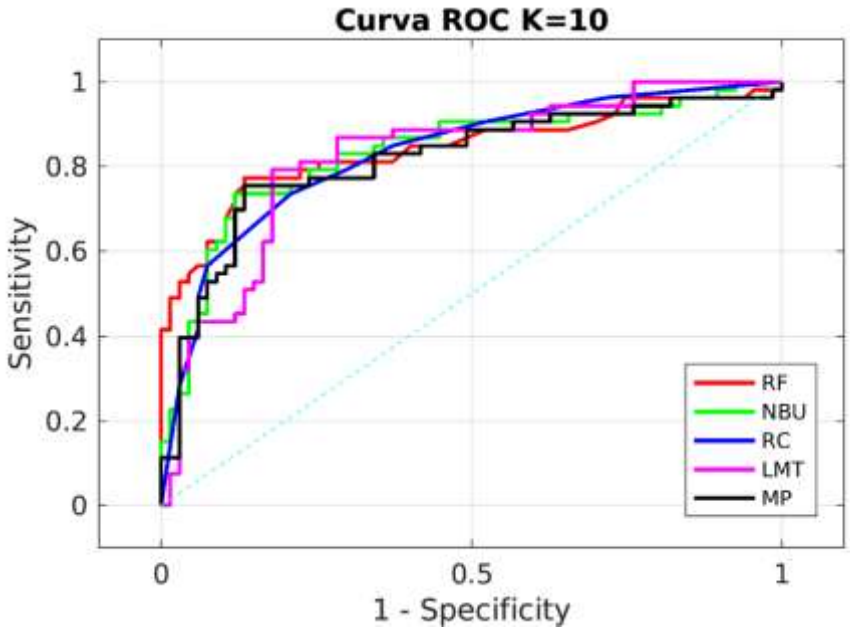


Figura 1.11. Ejemplo de una curva ROC mostrando el rendimiento de 5 algoritmos

Según Valle (2017), se define como una herramienta estadística utilizada para determinar, dentro de un objeto de estudio, la clase o categoría a la que pertenece un elemento, en otras palabras, determinar si un individuo pertenece a una de las dos clases posibles, presencia o ausencia de una enfermedad. Hernández (2017) menciona que la curva ROC puede entenderse como un gráfico en el cual se representan en el eje “y” la fracción de los verdaderos positivos o sensibilidad, y en el eje “x” se sitúan la fracción de los falsos positivos, o lo que es lo mismo 1 - la especificidad.

La curva ROC cuantifica la capacidad de un indicador para discriminar entre enfermos y sanos.

Es una representación gráfica en la cual se representan los valores de especificidad y sensibilidad así como las dos variantes de error que puede tener el sistema clasificador, en la medicina el término de dicotomía es usado para hacer énfasis en el conjunto de población que tiene o no tiene cierta enfermedad, por ejemplo en el grupo de población de estudio habrá personas clasificadas como enfermas y que están enfermas (VP), así como personas clasificadas como enfermas pero que en realidad no lo están (FP), también habrá un conjunto de personas que son determinadas como sanas (VN), y por consiguiente personas que son consideradas como sanas cuando en realidad no lo están (FN), siendo éste el escenario más delicado y peligroso, puesto

que si el sistema lo clasifica como saludable cuando en realidad está enfermo representa un peligro porque genera confianza de falsa salud, esto se agrava si la enfermedad no presenta síntomas (Cerda, 2012).
La curva ROC representa en cada punto de la gráfica un par de S/1-E que corresponde al nivel de decisión, lo ideal es que no exista solapamiento o eso se espera en un sistema con un excelente desempeño, los valores de la curva que más se acerquen a 1 representan poca solapación, es decir el área en el cual convergen los datos de FP y FN son pequeñas (Hernández, 2017).

Según Loyola et al., (2014), es una medida que puede ser utilizada para evaluar el desempeño de los clasificadores supervisados, y es especialmente confiable en problemas donde existe desbalanceo de clases, debido a que en estas gráficas se puede visualizar el equilibrio costo-beneficio; cualquier clasificador no puede aumentar el número de VP sin aumentar los FP, ecuación 7.

$$\textbf{AUC} = \frac{1+tVP-tFP}{2} \qquad (7)$$

Mean absolute error: Es la diferencia entre el valor de la medida y el valor tomado como exacto, este resultado puede ser negativo o positivo. El promedio del error absoluto es la suma de todos los errores de clasificación en todas las muestras divide entre el total de muestras. Por tanto, el clasificador que dé como resultado una cifra mayor a 0.1 define como error de clasificación alto, por lo tanto, no debe considerarse frente a los que arrojen una cifra menor (Hernández, 2017).

1.15 Sistema operativo Android y el lenguaje de programación Java

Android es un sistema operativo basado en el núcleo Linux. Fue diseñado principalmente para dispositivos móviles con pantalla táctil como: teléfonos inteligentes, tabletas y relojes inteligentes, televisores y hasta para algunos tipos de automóviles. Inicialmente fue desarrollado por Android Inc., empresa que Google respaldó económicamente y más tarde compro en 2005. Este sistema operativo fue presentado en 2007 junto la fundación del Open Handset Alliance (un consorcio de compañías de *hardware*, *software* y telecomunicaciones).

El primer móvil con el sistema operativo Android fue el HTC Dream y se vendió en octubre de 2008. Android es el sistema operativo móvil más utilizado del mundo, con una cuota de mercado superior al 80 % al año 2017, muy por encima de IOS.

Android es un conjunto de herramientas y aplicaciones vinculadas a una distribución Linux para dispositivos móviles. Por sí solo no es un Sistema Operativo Android es de código abierto, gratuito y no requiere pago de licencias.

Pero Android es más que un sistema operativo, su éxito se debe principalmente a que es Open Source, lo que significa que el código del sistema operativo está disponible en la red, y cualquier usuario con conocimiento de programación puede obtenerlo y hacer cualquier tipo de modificaciones. Debido a esto millones de internautas pueden aportas mejoras y posteriormente lanzar una versión mejorada. Por otra parte, el que sea Open Source lo convierte también en una de sus debilidades, pues al igual que se pueden hacer mejoras, un usuario podría encontrar puertas falsas a la seguridad y utilizarlas para fines no tan adecuados.

La versión básica de Android es conocida como Android Open Source Project (AOSP). El 25 de junio de 2014 en la Conferencia de Desarrolladores Google I/O, Google mostró una evolución de la marca Android, con el fin de unificar tanto el *hardware* como el *software* y ampliar mercados.

El 17 de mayo de 2017, se presentó Android Go. Una versión más ligera del sistema operativo para ayudar a que la mitad del mundo sin smartphone consiga uno en menos de cinco años. Incluye versiones especiales de sus aplicaciones donde el consumo de datos se reduce al máximo.

El lenguaje de programación java es el lenguaje oficial para desarrollar aplicaciones para el sistema operativo Android, fue elegido por ser el más potente y versátil de todos los lenguajes en el mercado. Java es un lenguaje de programación de propósito general, concurrente, orientado a objetos. Fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible. El objetivo es permitir que los desarrolladores de aplicaciones escriban el programa una vez y lo ejecuten en cualquier dispositivo (conocido en inglés como WORA, o "write once, run anywhere"), lo que permite programarlo en una plataforma y no tiene que ser recompilado para correr en otra.

Java es, a partir de 2012, uno de los lenguajes de programación más populares en uso, particularmente para aplicaciones de cliente-servidor de web, con unos diez millones de usuarios reportados

El lenguaje de programación Java fue originalmente desarrollado por James Gosling, de Sun Microsystems (constituida en 1982 y posteriormente adquirida el 27 de enero de 2010 por la compañía Oracle. Las aplicaciones de Java son compiladas a bytecode, y puede ejecutarse en cualquier máquina virtual Java (JVM) sin importar la arquitectura de la computadora subyacente (Tomás, 2012).

1.16 Lenguaje de programación Python

Python, pertenece al grupo de los lenguajes de programación y puede ser clasificado como un lenguaje interpretado, de alto nivel, multiplataforma, de tipado dinámico y multiparadigma. Python provee reglas de estilos, a fin de poder escribir código fuente más legible y de manera estandarizada. Estas reglas de estilo, son definidas a través de la Python Enhancement Proposal Nº 8 (PEP 8) (Duque, 2017), figura 1.9 logotipo del lenguaje de programación Python.

Figura 1.12 Logo de la plataforma del lenguaje de programación Python

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, usa tipado dinámico y es multiplataforma.

El lenguaje de programación Python permite hacer uso de una amplia variedad de librerías con las cuales es posible realizar diversos procedimientos. Es posible utilizar librerías tales como OpenCV, con la que se logra un poderoso procesamiento a imágenes digitales. Python es un lenguaje de programación de uso general, versátil y popular. Puede usarse para todo, desde desarrollo web hasta desarrollo de software y aplicaciones científicas.

El intérprete de Python es Open Source y esa es otra de sus grandes ventajas, cualquier usuario puede descargar el editor y rápidamente empezar a realizar procesos complejos. Otra de las grandes ventajas es que tiene soporte para los tres sistemas operativos más utilizados; Windows, Mac y Linux, por lo cual su practicidad aumenta. Al igual que en los lenguajes de programación como Java, existen ID´S de desarrollo que hacen aún más fácil el trabajo de realizar scripts en el lenguaje Python.

Python hace uso de librerías especiales para poder mostrar gráficas y entender mejor los resultados de un trabajo de investigación. A continuación, se describen brevemente algunas de las librerías más populares, utilizadas en este trabajo de investigación como parte de diversas pruebas que se llevaron a cabo.

Numpy: Es la librería por excelencia, y la más elemental en Python, aquella que no debe falta y de las primeras en aprender a utilizar los métodos de los que dispone. Proporciona una gran cantidad de métodos para trabajar con arreglos y matrices.

Pandas: Pandas nos permite crear Series (marcos unidimensionales) y los DataFrames (marcos bidimensionales). Con todos los métodos que ofrece Pandas, es posible visualizar y trabajar con datos de manera muy fácil.

Matplotlib: Es una librería que permite implementar gráficas para los datos resultantes. Es posible realizarlo en segunda y tercera dimensión, lo cual resulta muy práctico.

Scikit-Learn: Scikit-Learn es una librería que permite entrenar modelos de Aprendizaje Automático, tales como: Random Forests, SVM, Regresión Lineal y Logística, k-Nearest y mucho más.

TensorFlow: La librería estelar en este trabajo de investigación. Es una biblioteca de software libre que se utiliza para realizar cálculos numéricos mediante diagramas de flujo de datos. Los nodos de los diagramas representan operaciones matemáticas y las aristas reflejan las matrices de datos multidimensionales (tensores) comunicadas entre ellas. Gracias a la flexibilidad de la arquitectura, solo se necesita una API para desplegar el sistema informático de una o varias CPU o GPU en un escritorio, servidor o dispositivo móvil. Su objetivo era realizar investigaciones en el campo del aprendizaje automático y las redes neuronales profundas. A pesar de que este era su propósito inicial, se trata de un sistema lo bastante general como para poder aplicarse en muchos otros campos.

Keras: Es una librería escrita en Python, diseñada específicamente para hacer experimentos con redes neuronales. Permite crear prototipos rápidamente y de manera fácil. Es una librería que proporciona de manera limpia y sencilla la creación de una gama de modelos de aprendizaje profundo, encima de otras librerías: TensorFlow, Theano o CNTK. Esta librería fue desarrollada por François Chollet, un ingeniero de Google, él es el responsable también del mantenimiento de la misma. El código ha sido liberado bajo la licencia permisiva del MIT.

1.17 Redes neuronales

Las redes neuronales desde una perspectiva computacional son otra forma de emular ciertas características propias de los humanos, como la capacidad de memorizar y de asociar hechos. Si se examinan con atención aquellos problemas que no pueden expresarse a través de un algoritmo, se observará que todos ellos tienen una característica en común: la experiencia.

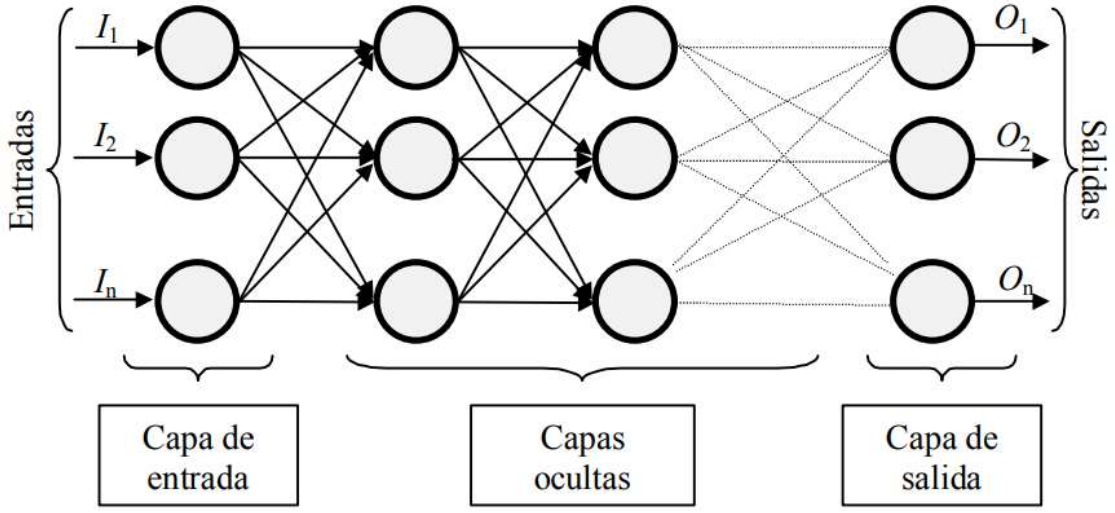


Figura 1.13 Elementos básicos de una red neuronal artificial

Los seres humanos son capaces de resolver situaciones haciendo uso de la experiencia acumulada. Así, parece claro que una forma de aproximarse al problema consista en la construcción de sistemas que sean capaces de reproducir esta característica humana. Las redes neuronales no son más que un modelo artificial y simplificado del cerebro humano, que es el ejemplo más perfecto del que disponemos para un sistema que es capaz de adquirir conocimiento a través de la experiencia (Antona, 2017). Una red neuronal es un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona.

Tratando de imitar el proceso que sucede en la neurona en donde, esta es estimulada o excitada a través de sus entradas (inputs) y cuando se alcanza un cierto umbral, la neurona se dispara o activa, pasando una señal hacia el axón. Actualmente se conoce que estos procesos son el resultado de eventos electroquímicos (Saji et al., 2015). Alan Turing. Fue el primero en estudiar el cerebro como una forma de ver el mundo de la computación. Sin embargo, los primeros teóricos que concibieron los fundamentos de la computación neuronal fueron Warren McCulloch, un neurofisiólogo, y Walter Pitts, un matemático, quienes, en 1943, lanzaron una teoría acerca de la forma de trabajar de las neuronas (Un Cálculo Lógico de la Inminente Idea de la Actividad Nerviosa - Boletín de Matemática Biofísica 5: 115-133). En este trabajo modelaron una red neuronal simple mediante circuitos eléctricos.

Debido a su constitución y a sus fundamentos, las redes neuronales artificiales presentan un gran número de características semejantes a las del cerebro. Por ejemplo, una de ellas es la capacidad de aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante. Las redes neuronales ofrecen numerosas ventajas y por esta y muchas otras razones las redes neuronales están siendo aplicadas en múltiples áreas (Ruiz et al., 2001).

A continuación, se describen algunas de las características de una red neuronal (Román, 2011):

Aprendizaje Adaptativo: Tienen la capacidad de aprender para realizar tareas basadas en un entrenamiento o en una experiencia inicial.

Auto-organización: Tiene la capacidad para crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.

Tolerancia a fallos: La destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.

Operación en tiempo real: Los cómputos neuronales pueden ser realizados en paralelo; para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.

1.18 Planteamiento de las hipótesis

El uso de algoritmos de aprendizaje automático ayudará a predecir comorbilidad perinatal y reducir riesgos en embarazos de alto riesgo.

CAPÍTULO II. ESTADO DEL ARTE

La diabetes gestacional es un trastorno metabólico al cual se de prestar especial atención, el diagnóstico oportuno puede hacer la diferencia para poder lograr un embarazo normal y sobre todo para que la madre y en neonato tengan una excelente salud (Ceñedo et al., 2017). Según Ceñedo et al., (2017), el diagnóstico de la diabetes gestacional fue descrito originalmente por O'Sullivan y Mahan en el año de 1979, esto se hizo en base a un estudio estadístico que incluía la presencia de dos o más tomas de glucemia considerando un estándar de la media.

Diversos autores documentan que las comorbilidades que más afectan a los pacientes con diabetes gestacional son de factible tratamiento, de tener un diagnóstico oportuno, y a partir de ello dar un seguimiento de cuidados personalizados a cada paciente con el objetivo de brindar un buen servicio en el cuidado de la salud y evitar escenarios adversos durante el parto y posterior al mismo.

A continuación, se presentan algunos de los trabajos documentados que se han realizado con respecto a la diabetes gestacional y las comorbilidades, así como los principales factores de riesgo asociados a cada comorbilidad, también trabajos en los cuales se hace uso de algoritmos de aprendizaje automático en la predicción de tipos de diabetes.

2.1 Frecuencia de obesidad y su relación con algunas complicaciones maternas y perinatales en una comunidad indígena.

Los autores Valdés et al., (2015), mencionan que es preocupante la manera en como se ha incrementado el número de personas que presentan sobrepeso y obesidad, en el año 2008 la OMS reportaba que en el mundo la cifra rondaba los 500 millones de adultos obesos, y se estimaba que en el 2015 la cifra alcanzaría los 700 millones.

En este trabajo descriptivo transversal acerca de las complicaciones que ocasionan el sobrepeso y la obesidad en mujeres en periodo de gestación, realizado en el año 2013, teniendo como muestra de estudio gestantes de la comunidad mistika de "El muelle", municipio de Puerto Cabezas en Nicaragua, usando como fuente de

información primaria las historias clínicas obstétricas que se le elaboraron para cada paciente.

El objetivo de este trabajo fue determinar la frecuencia de obesidad pregestacional, y su relación con algunas complicaciones maternas y perinatales en la comunidad mistika de “El muelle”.

Participaron 166 gestantes pertenecientes a la comunidad antes mencionada se analizaron las variables siguientes: edad, peso y talla en el momento de la captación, presión arterial, niveles de glucemia y de hemoglobina, resultados de los exámenes de orina, edad gestacional (EG) al parto, vía y tipo de parto, así como el peso al nacer. Para evaluar el estado nutricional se utilizó el IMC, que se calculó mediante la fórmula siguiente: peso = (kg)/talla (m^2). Se consideró la presencia de obesidad cuando el IMC fue ≥ 30 $Kg/m^{2;}$ sobrepeso, entre 25 y 29,9 Kg/m^2; normopeso, entre 18,5 y 24,9 Kg/m^2; y bajo peso, < 18,5 Kg/m^2.

Se tomaron en cuenta las normas y protocolos para la atención y cuidado vigentes en ese país, para determinar las complicaciones maternas y perinatales, las cuales fueron las siguientes: Anemia, Infección urinaria, toxemia gravídica, Diabetes Mellitus Gestacional (DMG), Parto pretérmino, Macrosomía neonatal y Bajo peso.

Se obtuvieron distribuciones de frecuencia como son números absolutos y porcentajes de las variables cualitativas, y para las variables cuantitativas medidas de tendencia central y de dispersión tales como: la media y la desviación estándar. Se empleó la prueba de chi cuadrado para comprobar la hipótesis sobre la relación que pudiera existir entre las variables cualitativas, y se asumió un valor de $p< 0,05$ para la significación estadística. Se determinó la fuerza de asociación para cada complicación estimando la razón de productos cruzados odds-ratio (OR), con un intervalo de confianza (IC) del 95 %. El resultado del OR se valoró de acuerdo con los siguientes criterios:

Si > 1, variable SI constituye un factor de riesgo.
Si = 1, variable NO constituye un factor de riesgo.
Si < 1, variable constituye un factor de protección.

En este trabajo de investigación hicieron uso del software Epidat 3.1. para la tarea de procesamiento estadístico de cada historial clínico.

Los resultados que obtuvieron 40 de las gestantes (24,1 %), iniciaron el embarazo con obesidad. De ellas, 26 (15,6 %), con obesidad grado I; 10 (6,1 %) con grado II, y 4 (2,4 %) con grado III. En general, el 87,5 % de las embarazadas con obesidad pregestacional presentaron alguna complicación materna o perinatal; mientras en las no obesas se observaron en el 59,5 % (OR: 4,76, IC: 1,74-12,96, p= 0,0011). La obesidad elevó significativamente el riesgo de presentar diabetes mellitus gestacional (OR: 5,03, IC: 2,03-12,4, p= 0,0002), macrosomía (OR: 8,06, IC: 2,56-25,36, p= 0,0001) y cesárea (OR: 5,13, IC: 1,53-17,22, p= 0,0040).

La frecuencia de obesidad en la población obstétrica de la comunidad indígena de “El Muelle” es elevada e incrementa el riesgo de complicaciones maternas y perinatales como diabetes mellitus gestacional, macrosomía y cesárea.

2.2 Factores asociados a mortalidad en recién nacidos prematuros con enfermedad de membrana hialina en el Hospital Nacional Sergio E. Bernales, mayo 2015 – mayo 2017

En este trabajo de investigación llevado a cabo en Perú, se realizó un estudio observacional, retrospectivo, analítico de tipo casos controles, que comprendió el periodo de mayo del 2015 a mayo del 2017. Se revisó el historial clínico correspondientes al servicio de Neonatología en el archivo del Hospital Nacional Sergio E. Bernales, en el cual fueron registrados los datos en una ficha de recolección de datos. Se utilizó la fórmula de cálculo de la muestra por número de casos y controles diferentes, utilizando un *odds ratio* de 3 y una frecuencia de 0.4, con una relación de controles y casos de 2 a 1. Para realizar el procesamiento de los datos obtenidos se utilizó el software IBM SPSS V24. Este trabajo tuvo como objetivo principal determinar los factores asociados a mortalidad en recién nacidos prematuros con enfermedad de membrana hialina.

La población de estudio fueron 123 historias clínicas se encontró un mayor grado de mortalidad en pacientes prematuros con edad gestacional menor a 34 semanas,

también existió una mayor mortalidad en pacientes con un peso <1500 gr (80,5 %), encontrándose una asociación estadísticamente significativa (p: 0,000), entre un peso menor a 1500 gr y una mortalidad por enfermedad de membrana hialina. La mortalidad por enfermedad de membrana hialina fue mayor en asociación a APGAR al minuto menor o igual a 6 (85,4%), siendo la asociación estadísticamente significativa entre casos y controles y la variable APGAR al minuto (p=0,000; OR= 5,556; IC95%=2,110 – 14,630). Asimismo, la mortalidad fue similar, presencia de ruptura prematura de membranas (48,8%) o en ausencia de ella (51,2%). Se evaluó la mortalidad por enfermedad de membrana hialina en relación a la presencia de infección de tercer trimestre del embarazo, encontrándose que la relación no es estadísticamente significativa (p=0,168; IC95%=0,259 – 1,269), al igual que en el caso de comorbilidades maternas (p=0,751; IC95%=0,463 – 2,909).

En conclusión, en este trabajo de investigación se documentó que los principales factores asociados a mortalidad por enfermedad de membrana hialina en pacientes prematuros fueron; sexo, grado de prematuridad, bajo peso al nacer, APGAR menor o igual a 6 y ruptura prematura de membranas. Este equipo también concluye que se deben aplicar medidas de control con el propósito de disminuir la presencia de los factores antes descritos (Risco, 2018).

2.3 Morbilidad del hijo de madre con diabetes gestacional, en el Instituto Nacional de Perinatología Isidro Espinosa de los Reyes

En este trabajo de investigación se realizó un trabajo de observación comparativo y analítico, el objetivo de esta investigación fue conocer la morbilidad general desarrollada por los hijos de madres con diabetes gestacional, en comparación con hijos de madres sin esta patología en la población atendida en el Instituto Nacional de Perinatología Isidro Espinosa de los Reyes ubicado en la ciudad de México. En este trabajo participaron un total de 288 pacientes, de las cuales 144 presentaban diabetes gestacional y el otro 50% no la presentaba. También se incluyeron en el estudio los hijos de las madres con y sin diabetes puesto que era necesario documentar las comorbilidades desarrolladas en cada conjunto, en el caso de que las hubiera.

En este estudio se incluyeron variables sociodemográficas, así como antecedentes familiares con alguna complicación de salud, se investigaron antecedentes maternos y morbilidad desarrollada por el neonato. El análisis estadístico incluyó: medidas de tendencia central (media, desviación estándar e intervalo de confianza) para las variables demográficas, chi cuadrada y riesgo relativo, con intervalo de confianza del 95% para la comparación de ambas poblaciones.

La conclusión a la que se llegó con esta investigación fue que, en la población estudiada, los antecedentes de familiares con diabetes mellitus, obesidad y muerte fetal en la madre gestante son factores de riesgo para el desarrollo de DG. La morbilidad del hijo de madre con DG está por encima de lo referido en la literatura internacional esto debido probablemente por ser un centro de concentración de tercer nivel de atención. Las alteraciones respiratorias predominaron, específicamente el síndrome de adaptación pulmonar (Delgado, 2011).

2.4 Multilayer Perceptron Application for Diabetes Mellitus Prediction in Pregnancy Care

En este trabajo de investigación se utiliza una red neuronal conocida como Multilayer perceptron o Perceptron multicapa, la cual es un tipo de red neuronal artificial ANN (del inglés Artificial Neural Networks), y ha demostrado la capacidad de resolver problemas que son linealmente separables, la utilizan en este trabajo con el objetivo de predecir la diabetes mellitus gestacional. Los autores dan un panorama global acerca de la prevalencia de esta enfermedad y los sectores de la población, así como el momento en el que aparece o es diagnosticada.

Según los autores de este trabajo, cualquier mujer puede desarrollar diabetes mellitus en el embarazo. Algunos grupos de mujeres presentan mayor riesgo, como grupo racial de ascendencia africana, hispana, india o asiática. También documentaron que los factores de riesgo para esta enfermedad son los siguientes: edad mayor de 25 años, antecedentes familiares de diabetes, diabetes gestacional previa, embarazos previos entre otros.

Esta investigación se utiliza la técnica de evaluación, llamada k-fold cross-validación, para la evaluación del desempeño de la propuesta basada en ANN. En este método, k subconjuntos dividen la base de datos de diabetes. Se conserva un subconjunto para entrenamiento del algoritmo y el porcentaje restante como conjunto de prueba.

Los resultados muestran que este enfoque alcanzó una precisión de 0.74, Recall 0.741, F-measure de 0.741, y ROC área de 0.779. Según los autores los resultados que obtuvieron muestran que este método es un excelente predictor de esta enfermedad, y que esta contribución ofrece una herramienta de inteligencia computacional capaz de identificar los casos de riesgo durante el embarazo y, por lo tanto, reducir las posibles secuelas tanto para la mujer embarazada como para el feto (Moreira et al., 2018).

2.5 Factores de riesgo asociados a la morbi-mortalidad perinatal en mujeres con diabetes gestacional del Sur de México

Este trabajo de investigación realizado por un equipo de académicos especialistas de la facultad de Ciencias Químicas de la AUGro., en el cuál se realizó un estudio transversal en el hospital de la madre y en niño de la ciudad de Chilpancingo Guerrero, México, con el objetivo de evaluar los factores de riesgo asociados a la morbi-mortalidad perinatal en mujeres con diabetes gestacional, con este trabajo se pudo conocer la prevalencia de la DG durante el año 2015 en esta zona del territorio Mexicano, así como las principales variables de comorbilidad perinatal, se documentaron 7 principales comorbilidades, para el caso de la madre: Enfermedad hipertensiva del Embarazo(EHE), Preeclampsia, sangrado >400 mL, y para los neonatos: Óbito, Macrosomía, Síndrome de Deficiencia Respiratoria (SDR) y bajo peso entre otros más, sin embargo éstos son los más recurrentes para el binomio madre-neonato. Durante este año en ese hospital se atendieron 4190 nacimientos de los cuales 96 pacientes presentaron diabetes gestacional por lo tanto la prevalencia de la DG es alrededor de 2.3%, los datos que se obtuvieron fueron procesados usando el software STATA v.12.0, y se analizó el valor de aporte por cada factor de riesgo, es decir se determinó qué tanto afecta el que una madre tenga un antecedente de diabetes o de hipertensión en su familia por ejemplo, y el impacto que esto tiene al momento de presentar diabetes gestacional.

Este trabajo de investigación concluye con los siguiente. La elevada tasa de morbi-mortalidad en mujeres con diabetes gestacional es asociada a factores de riesgo tradicionales, clínicos y metabólicos que son de factible prevención, y que pueden ser tratados y monitoreados por personal de salud que participan en el cuidado de la salud en el embarazo (Zaragoza et al., 2017).

2.6 Maternal risk factors for hypertensive disorders in pregnancy: a multivariate approach

En este trabajo de investigación los autores buscan desarrollar algoritmos que puedan ser capaces de predecir la aparición de un desorden hipertensivo conocido como Preeclampsia, dividen en tres momentos a este padecimiento; preeclampsia temprana, preeclampsia tardía e hipertensión gestacional. Es un estudio prospectivo que cuenta con una muestra amplia de casos. Según los autores su muestra fue de 8,366 casos de los cuales 37 casos fueron con preeclampsia temprana, 128 con preeclampsia tardía, 140 con hipertensión gestacional y 8,061 casos que no presentaron síntomas por preeclampsia temprana o por hipertensión gestacional.

Realizaron un análisis multivariado de factores de la historia materna y compararon el rendimiento estimado de algoritmos en la predicción temprana de preeclampsia, preeclampsia tardía e hipertensión gestacional. Se realizó un estudio prospectivo que se llevó a cabo entre marzo de 2006 y noviembre de 2007.

Se utilizaron características como: edad de la madre, ascendencia racial, hábito de fumar, el tipo de método utilizado para concebir, el historial médico para conocer si padece algún padecimiento como hipertensión crónica, diabetes mellitus, entre algunas otras características, se le preguntó a la madre si consume algún tipo de medicamento como; antidepresivos, medicamentos para la hipertensión anti-epilépticos, aspirinas, esteroides entre otras más. También se incluyen características obstétricas para conocer la paridad de la gestante.
Utilizaron regresión logística para determinar ¿Cuál de los factores entre las características maternas, historia médica y obstétrica tuvo una contribución significativa en la predicción de la Preeclampsia temprana, Preeclampsia tardía e Hipertensión gestacional? La medición del rendimiento fue determinada por la curva

ROC. Utilizaron los paquetes de software estadístico SPSS 15.0 (SPSS Inc., Chicago, IL, EE. UU.).

Según los autores, en el grupo de preeclampsia temprana, hubo una mayor prevalencia de mujeres de ascendencia de raza negra, hipertensos crónicos, mujeres con Preeclampsia temprana en sus embarazos previos, también aquellas con asistencia para concebir y en donde su IMC fue alto. En el grupo de Preeclampsia tardía, hubo una mayor prevalencia de mujeres negras, mujeres con una maternidad y la historia de la Preeclampsia temprana, y aquellas con IMC alto. En el grupo de Hipertensión gestacional, hubo una mayor prevalencia de mujeres con una maternidad e historia de la Preeclampsia temprana, y con IMC alto.
Los resultados con respecto al área bajo la curva para la detección de preeclampsia temprana fueron de (0.794), para Preeclampsia tardía y para la Hipertensión gestacional fue de 0.796 respectivamente.

CAPÍTULO III. DESARROLLO DE LA METODOLOGÍA PROPUESTA

3.1 Metodología para la obtención de datos

Se realizo un estudio transversal retrospectivo a un total de 4000 expedientes clínicos revisados en dos hospitales de segundo nivel de la ciudad de Chilpancingo Guerrero México. Éstos son el Hospital General Raymundo Abarca Alarcón y en el Hospital de la Madre y el niño Guerrerense, encontrando 318 casos de estudio, 53 casos que fueron etiquetados con preeclampsia, 57 fueron productos con bajo peso, 27 neonatos clasificados con peso alto para la edad gestacional, 96 casos de diabetes gestacional y 86 casos de embarazos de control, la fecha de ocurrencia de las comorbilidades fue del año 2013 al 2019. La información se obtuvo principalmente del expediente clínico de la paciente, así como la hoja de codificación para el recién nacido, se recabaron datos previos a la labor de parto y post resolución. Entre las variables evaluadas se consideró el antecedente familiar de enfermedades crónico-degenerativas, estrato socioeconómico, tipo de parto (características cualitativas), edad, peso, tensión arterial, talla, Índice de Masa Corporal (IMC: peso/talla^2) y condición bioquímica de la gestante (características cuantitativas).

Se evaluó al nacimiento las mediciones antropométricas del neonato, a partir del cual y de acuerdo a los criterios de la Asociación Americana de Pediatría se definió bajo peso (<2500 g) y macrosomía (≥4000 g). Además, se evaluó el APGAR, considerando la escala Silverman/ Andersen, la presencia de malformaciones.

A continuación, se describen los principales apartados del expediente clínico de donde se extrajeron las variables que participaron en el estudio.

En el apartado “Hoja de hospitalización” se obtuvieron datos como el peso y la talla materna.

Del apartado “Nota de ingreso” se obtuvieron las variables de predisposición genética, como antecedentes de hipertensión, antecedente de diabetes mellitus tipo 2 así como el parentesco de los familiares. De este apartado también se obtuvieron variables como el total de gestas, partos, abortos y la cantidad de cesáreas, variables

sociodemográficas como; edad, escolaridad, ocupación, tensión arterial sistólica y diastólica.

Del apartado “Nota inicial del recién nacido” en el cual se puede obtener gran parte de las variables utilizadas que tienen que ver con el estado de salud del recién nacido

- Peso
- Talla
- Perímetro craneal
- Perímetro de tórax
- perímetro abdominal
- Tamaño de pie
- Frecuencia cardiaca
- Frecuencia respiratoria
- APGAR 1’ 5’
- Escala Silverman

De la hoja “valoración inicial gineco-obstétrica (Triage)” se obtuvieron las variables de tensión arterial sistólica (TAS) y la tensión arterial diastólica (TAD).

De la hoja “Estudio sociodemográfico” se obtuvieron las variables; Exposición a contaminantes como el polvo y humo, escolaridad, ocupación

Del apartado “historia clínica” se obtuvieron las variables ocupación, antecedentes personales herederos familiares y las variables obstétricas como el total de partos, cesáreas, abortos, gestas y óbitos, este es el segundo apartado del que es posible extraer estas variables

Del apartado “hoja frontal para diagnóstico y operaciones quirúrgicas” se obtuvieron el tipo de diagnóstico inicial de la paciente, así como el tipo de parto, si fue cesárea o si fue eutócico.

Del apartado “hoja quirúrgica” se puede obtener información como; el sangrado total durante la intervención quirúrgica.

En los casos de preeclampsia se lleva a cabo un estudio de laboratorio que incluye biometría hemática, y estudio general de orina, de los cuales se obtienen las siguientes variables;

Tabla 3.1 Características obtenidas mediante un estudio de laboratorio

No.	Características maternas
1	Eritrocitos
2	Hemoglobina
3	Hematocrito
4	Volumen Globular Medio (vgm)
5	Concentración de Hemoglobina Corpuscular Media (cmhb)
6	cmhbc
7	Distribución de glóbulos rojos (rdw)
8	plaquetas
9	Volumen Plaquetario Medio (vpm)
10	Leucocitos
11	Neutrofilo
12	Linfocito
13	Monocito
14	Eosinofilo
15	Basofilo
16	Glucosa
17	Urea
18	Creatinina
19	Ácido úrico
20	Nitrógeno Ureico en la Sangre(bun)
21	Colesterol
22	Trigliceridos

Para poder tener acceso a los expedientes clínicos fue necesario realizar las solicitudes correspondientes al jefe de enseñanza del Hospital, el Dr. Emmanuel González Abonce, en el anexo 1 se encuentra el documento que autoriza la entrada a archivo clínico.

3.2 Descripción de la metodología propuesta

La parte fundamental de esta investigación son los datos, puesto que en ellos recae la posibilidad de obtener buenos o malos resultados. El set de datos que se utilizó tiene la ventaja de ser situaciones reales y, por lo tanto, su validez es aceptable. Es preciso enfatizar que al tratarse de datos personales es sumamente complicado obtenerlos, y más aun siendo datos médicos.

La metodología general propuesta que se utilizó para llevar a cabo las pruebas necesarias y posteriormente documentar los resultados se muestra en la figura 3.1.

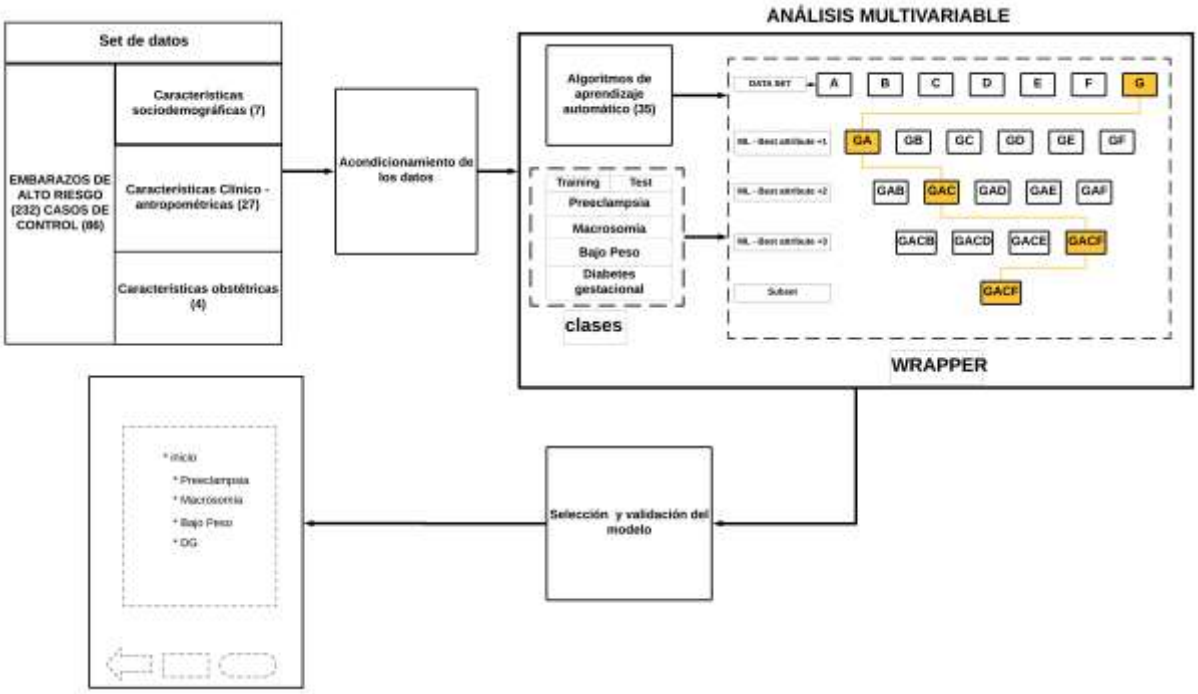


Figura 3.1. Metodología propuesta

La metodología se divide en 2 etapas que a continuación se describen:

Análisis multivariable

Se tiene una base de datos con información de gestantes que presentaron embarazos de alto riesgo, la cual contiene características sociodemográficas, clínico-antropométricas, así como obstétricas, tal como fueron recolectados de los expedientes clínicos. Por lo tanto, para que estos valores puedan ser utilizados en programación es necesario realizar una adecuación, prestando atención a mantener

la naturaleza de estos. La base de datos está contenida en una hoja de cálculo con formato xlsx, los casos de preeclampsia, macrosomía, bajo peso y los embarazos sin complicaciones fueron recolectados del Hospital general Raymundo Abarca Alarcón, la base de datos con embarazos que presentaron DG fue proporcionada por el laboratorio de diabetes y obesidad de la facultada de Ciencias Químico-Biológicas de la Universidad Autónoma de Guerrero. Una vez realizado el preprocesamiento se exporta a un formato “csv”, lo que agrega facilidad para realizar las pruebas necesarias. Se realiza un proceso de adecuación para obtener un archivo numérico.

El método de selección de característica que se utilizó fue el *“wrappers”*, éste, utiliza el algoritmo de aprendizaje automático de interés como una caja negra para calificar subconjuntos de variables de acuerdo con su poder predictivo. Se prueban todas las combinaciones de variables posibles y se lleva un registro del puntaje de predicción, en seguida se muestra aquella combinación que dio el puntaje más alto (Guyon, 2003).

De esta parte de la metodología se obtuvo como resultado aquellos algoritmos que según la naturaleza de los datos se adecúan mejor. Pero también se obtuvieron los conjuntos de características o factores de riesgo predominantes para cada comorbilidad, esto en función de los mejores resultados para obtener qué factores de riesgo sobresalen para la comorbilidad que se esté estudiando.

Se desarrolló un modelo de red neuronal que contenía las características del universo de los datos con los que inicialmente se inició la investigación. Este modelo fue desarrollado, entrenado y exportado del lenguaje de programación Python para posteriormente ser utilizado en una aplicación Android. Esta aplicación contiene este modelo y una interfaz que solicita al usuario el vector de características o valores de factores de riesgo de un nuevo registro, la cantidad de factores es igual al conjunto que resultó del análisis multivariable o incluso puede ser utilizadas las variables iniciales.

En función del aprendizaje, el modelo hace una inferencia para conocer si la clase es positiva o negativa, el porcentaje de exactitud de la inferencia dependerá directamente del tamaño de la muestra con la que el modelo entrenó.

3.3 Adecuación de los datos

La base de datos consiste en un grupo selecto de características de cada paciente, con información de antecedentes familiares de DM2, así como de hipertensión; también hábitos de alimentación e higiene, características clínico-metabólicas, antropométricas y sociodemográficas.

Tabla 3.2 Descripción de las variables utilizadas en el análisis

#	Características		Promedio (±)	Min	Max	Std (±)
1	SOCIODEMOGRÁFICAS	edad	25	13	44	6.50
2		escolaridad	Secundaria	Sin escolar	Lic.	1.12
3		ant_fam_dm2	Sin antece	No	Si	0.47
4		parentesco_dm2	Sin parentezco	No	Madre	1.24
5		ant_fam_hta	Sin antece	No	Si	0.42
6		parentesco_hta	Sin parentezco	No	Padre	1.04
7		Contaminación	Humo	Ninguno	Ambos	0.00
8	ANTROPOMETRICAS	Peso	67	40	105	12.15
9		Talla	2	1	2	0.07
10		IMC	29	18	46	4.95
11		TAS	117	80	220	23.44
12		TAD	74	50		13.41
13	OBSTÉTRICAS	Total Gestas	2	0	11	1.66
14		Total de partos	1	0	11	1.59
15		Total cesareas	0	0	2	0.56
16		Total de Abortos	0	0	2	0.39
17	CLÍNICAS	Eritrocitos	4	2	12	0.65
18		Hemoglobina	0.15	0	1	0.36
19		Hematocrito	37	13	65	4.80
20		vgm	88	62	113	7.03
21		cmhbc	30	19	93	5.89
22		cmhbc	34	30	53	1.66
23		rdw	13	9	23	2.14
24		Plaquetas	227	42	526	69.43
25		vpm	9	7	16	1.05
26		Leucocitos	11	4	33	3.57
27		Neutrofilo	70	6	92	12.29
28		Linfocito	23	6	80	9.82
29		Monocito	5	1	14	2.43
30		Eosinofilo	1	0	10	0.95
31		Basofilo	1	0	7	0.71
32		Glucosa	91	53	326	32.78
33		Urea	17	9	43	5.41
34		Creatinina	1	0	1	0.14
35		Ácido Úrico	5	2	10	1.32
36		bun	8	4	21	2.61
37		Colesterol	235	97	388	52.34
38		Triglicéridos	309	118	1341	125.98

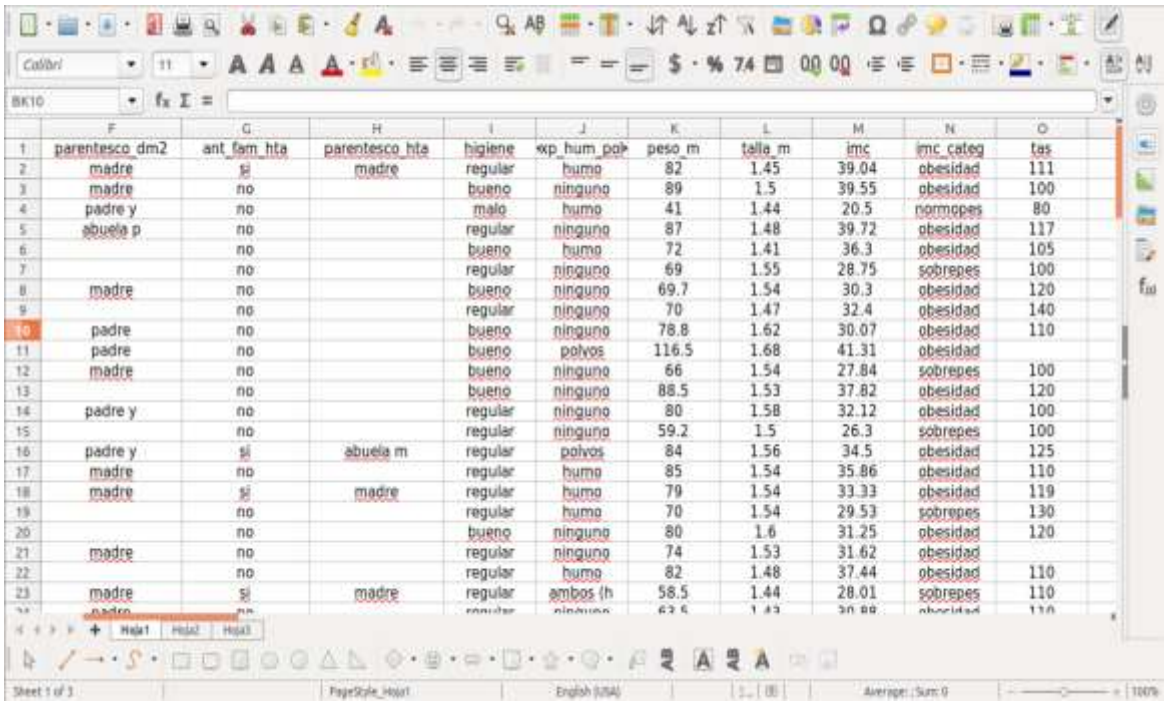
Se llevó a cabo una discriminación de variables con el objetivo de realizar nuevamente las mismas pruebas y observar el comportamiento de los algoritmos cuando el universo de datos disminuye, dicha selección de características estuvo a cargo de la Dra. Irisi Paola Guzmán Guzmán. En la tabla siguiente se muestran las variables que resultaron después de dicha selección, ver tabla 3.3.

Tabla 3.3 Variables seleccionadas para medir el rendimiento de los Algoritmos sin TAS y TAD

	Características maternas	Nombre de la Variable
1	Edad	edad
2	Escolaridad	escolaridad
3	Antecedente Familiar DM2	antFamDM2
4	Parentesco DM2	parFamDM2
5	Antecedente Familiar Hta	antFamHt
6	Parentesco Hta	parFamHt
7	Contaminación	contaminacion
8	Peso	peso
9	Talla	talla
10	IMC	IMCMat
11	TAS	TAS
12	TAD	TAD
13	Total Gestas	totGestas
14	Total Partos	totPartos
15	Total de cesareas	totCesareas
16	Total de abortos	totAbortos
17	Anemia	anemia
18	Plaquetas	plaquetas
19	Glucosa	glucosa
20	Colesterol	colesterol

Los resultados con esta cantidad de variables se encuentran en el anexo 2 de este trabajo.

Para realizar las pruebas necesarias y obtener el rendimiento de los algoritmos de aprendizaje automático es de suma importancia contar con una base de datos adecuada, lo ideal sería tener suficientes registros positivos y negativos con los cuales realizar múltiples pruebas, modificar la cantidad de variables y obtener cuál es el desempeño de cada algoritmo disponible. La Dra. Iris Paola Guzmán Guzmán, adscrita al laboratorio de diabetes y obesidad, fue la encargada de proporcionar los datos de los que disponía. Posteriormente al realizar una estancia en el Hospital General Dr. Raymundo Abarca Alarcón, se logró conseguir datos de embarazos de alto riesgo, así como información de la resolución de los mismos. En los datos iniciales algunos tenían un formato de respuesta tipo cadena de texto, no numérico, esto genera un inconveniente para ser utilizados en algún ambiente de programación, por lo tanto, para poder realizar pruebas con ellos, se adecuaron, lo anterior, sin alterar su significado o su valor numérico, en la figura 3.2 se observa una muestra de los datos crudos.



	F	G	H	I	J	K	L	M	N	O
1	parentesco_dm2	ant_fam_hta	parentesco_hta	higiene	exp_hum_pol	peso_m	talla_m	imc	imc_categ	tas
2	madre	si	madre	regular	humo	82	1.45	39.04	obesidad	111
3	madre	no		bueno	ninguno	89	1.5	39.55	obesidad	100
4	padre y	no		malo	humo	41	1.44	20.5	normopes	80
5	abuela p	no		regular	ninguno	87	1.48	39.72	obesidad	117
6		no		bueno	humo	72	1.41	36.3	obesidad	105
7		no		regular	ninguno	69	1.55	28.75	sobrepes	100
8	madre	no		bueno	ninguno	69.7	1.54	30.3	obesidad	120
9		no		regular	ninguno	70	1.47	32.4	obesidad	140
10	padre	no		bueno	ninguno	78.8	1.62	30.07	obesidad	110
11	padre	no		bueno	polvos	116.5	1.68	41.31	obesidad	
12	madre	no		bueno	ninguno	66	1.54	27.84	sobrepes	100
13		no		bueno	ninguno	88.5	1.53	37.82	obesidad	120
14	padre y	no		regular	ninguno	80	1.58	32.12	obesidad	100
15		no		regular	ninguno	59.2	1.5	26.3	sobrepes	100
16	padre y	si	abuela m	regular	polvos	84	1.56	34.5	obesidad	125
17	madre	no		regular	humo	85	1.54	35.86	obesidad	110
18	madre	si	madre	regular	humo	79	1.54	33.33	obesidad	119
19		no		regular	humo	70	1.54	29.53	sobrepes	130
20		no		bueno	ninguno	80	1.6	31.25	obesidad	120
21	madre	no		regular	ninguno	74	1.53	31.62	obesidad	
22		no		regular	humo	82	1.48	37.44	obesidad	110
23	madre	si	madre	regular	ambos (h	58.5	1.44	28.01	sobrepes	110

Figura 3.2. Datos sin adecuación

La adecuación consistió en establecer un archivo numérico para todos ellos, en los casos en los que éstos fueran tipo cadena de texto se estableció un criterio para colocar un valor numérico representativo; por ejemplo, para la variable “contaminación”, para obtener esta variable se realiza un breve cuestionario directo

en el cual se le pregunta a la gestante o en su defecto al familiar responsable al momento si existe una exposición a contaminantes como humo de leña, polvo u otro tipo. La variable de antecedente familiar de DM2 o Hipertensión fue adecuada de la siguiente manera, primeramente, se documenta si existe algún antecedente de cada una de ellas, y posteriormente se pregunta el parentesco de origen del antecedente.
A continuación, se describen las adecuaciones necesarias para cada una de las variables; el orden que se tomó para describirlas obedece al orden original de la base de datos proporcionada, no por orden de importancia:

Edad materna: La edad es una variable sociodemográfica, por la facilidad para documentarla no requiere profundizar en ella, solo se requiere el valor numérico en años.

Escolaridad: Establece el grado que cursó la gestante; para esta variable se utilizaron las categorías que se muestran en la tabla 3.

Tabla 3.4 Asignación de valores numéricos al grado escolar

No.	Grado escolar	Valor numérico asignado
1	No hay dato	-1
2	Sin estudios	0
3	Primaria	1
4	Secundaria	2
5	Preparatoria	3
6	Licenciatura	4
7	Posgrado	5

Antecedente familiar con diabetes tipo 2: Esta variable se utiliza para conocer si existe algún familiar en el árbol genealógico que haya padecido esta enfermedad; La asignación de valores numéricos es la siguiente: no hay dato = -1, NO = 0 y SI = 1.

Tabla 3.5 Asignación de valores numéricos para antecedentes familiares con DM tipo 2

No.	Antecedente familiar con diabetes mellitus tipo 2	Valor numérico asignado
1	No hay dato	-1
2	NO	0
3	SI	1

Parentesco familiar con diabetes tipo 2: Podría darse el caso que en la variable de antecedente familiar no se cuente con registro, eso es distinto a que no existe un antecedente familiar, por lo tanto, deben ser dos valores numéricos diferentes. Si el valor numérico en la variable antecedente familiar con diabetes mellitus 2 es NO, la respuesta en el parentesco es negativa y se le asigna un valor numérico a esta variable de 0. Esta variable es relevante cuando la respuesta al antecedente es afirmativa, se pregunta cuál es el parentesco que tiene la gestante con el familiar que tuvo diabetes mellitus tipo 2.

Tabla 3.6 Asignación de valores numéricos a parentesco familiar con DM tipo 2

No.	Parentesco familiar con diabetes mellitus 2	Valor numérico asignado
1	No hay dato	-1
2	Negativo	0
3	Abuelos	1
4	Padre	2
5	Madre	4

Antecedente familiar con hipertensión: Con esta variable se busca conocer si existe una predisposición a padecer hipertensión, al igual que la variable de antecedente para diabetes, las posibles respuestas son 3.

Tabla 3.7 Asignación de valores numéricos para antecedentes familiares con Hipertensión

No.	Antecedente familiar con Hipertensión	Valor numérico asignado
1	No hay dato	-1
2	NO	0
3	SI	1

Parentesco familiar con hipertensión: La asignación de valor numérico para las opciones de esta variable son similares al antecedente de diabetes mellitus tipo 2. El valor numérico más alto es asignado a la respuesta afirmativa para el padre.

Tabla 3.8 Asignación de valores numéricos a parentesco familiar con Hipertensión

No.	Parentesco familiar con Hipertensión	Valor numérico asignado
1	No hay dato	-1
2	Negativo	0
3	Abuelos	1
4	Madre	2
5	Padre	4

Exposición a contaminación de humo y polvo: Con esta variable se conoce si la madre en su vida cotidiana está expuesta a factores contaminantes tales como; humo y polvo, si vive en alguna comunidad, probablemente la exposición a humo de leña sea muy frecuente. Pero también puede darse el caso de que este expuesta a contaminación de polvo por la cercanía a alguna trituradora o mina. Para asignar valores numéricos a esta variable se tomó el siguiente criterio:

Tabla 3.9. Asignación de valores numéricos para exposición a contaminación

No.	Exposición a humo y polvo	Valor numérico asignado
1	No hay dato	-1
2	ninguno	0
3	humo	1
4	polvo	2
5	ambos	3

Se tienen registros del **peso** y la **talla** de la gestante, con estas dos variables se calcula el **IMC** usando la fórmula ($imc=peso/talla^2$).

El set de datos contiene información de características de mediciones de presión arterial, **Tensión Arterial Sistólica (TAS)** y **Tensión Arterial Diastólica (TAD)**. Para el caso de estas variables no fue necesario recategorizar, se utilizó su valor numérico bruto que se almacenó en la base de datos.

El set de datos también contiene información obstétrica de las gestantes; se documentan variables tales como el total de gestas que ha tenido (**totGestas**), así como si la resolución de alguna de ellas ha sido a través de cesárea (**totCesareas**) y también información si ha tenido algún aborto (**totAbortos**), también se incluye la variable que describe las ocasiones en que un embarazo ha terminado por parto natural **(totPartos)**.

Con el objetivo de incrementar el universo de estudio, y tomando en cuenta que se tenía disponibles las siguientes variables, su utilizaron como parte del muestreo, ver tabla 3.10.

Tabla 3.10 Variables que se incluyen en el universo de datos

eritrocitos	linfocito
hemoglobina	monocito
hematocrito	eosinofilo
vgm	basofilo
cmhb	glucosa
cmhbc	urea
rdw	creatinina
plaquetas	acidoUri
vpm	bun
leucocitos	colesterol
neutrofilo	trigliceridos

De las anteriores variables no fue necesario una adecuación, debido a que los valores de éstas, es numérico, por lo tanto, se utiliza ese valor.

Para realizar las pruebas necesarias en la comorbilidad de diabetes gestacional fue necesario utilizar un set de variables que estuvieran disponibles en el set datos de embarazos de sin complicaciones, así como en el set de datos de diabetes gestacional, esto debido a que estos dos conjuntos de datos fueron recolectados de dos hospitales distintos, en la tabla 3.11 se muestra un resumen de las variables utilizadas.

Tabla 3.11 Variables utilizadas para la comorbilidad de DG

#	Nombre de variable
1	Edad
2	Escolaridad
3	Antecedente Fam. Con DM2
4	Parentesco Fam. Con DM2
5	Antecedente Fam. Hta.
6	Parentesco Fam. Con Hta.
7	Contaminación
8	Peso
9	Talla
10	Índice de masa corporal (imc)
11	Tensión Arterial Sistólica (TAS)
12	Tensión Arterial Diastólica (TAD)
13	Total de Gestas
14	Total de Partos
15	Total Cesáreas
16	Total de Abortos
17	Eritrocitos
18	Hemoglobina
19	Plaquetas
20	Glucosa
21	Colesterol

Se realizaron pruebas con la siguiente configuración utilizaron las variables antes enlistadas: Con un total de 96 instancias positivas con diabetes gestacional y 86 instancias negativas, es decir embarazos que no presentaron complicaciones durante o en la resolución de este. Se obtuvieron los siguientes resultados que se muestran en la tabla 4.9

3.4 Usando Weka para la selección de atributos

Cuando se tiene los datos listos para ser utilizados, se inicia con una parte importante del trabajo de investigación; modelar los datos para poder obtener los subconjuntos de los factores de riesgo en cada comorbilidad, prescindiendo de aquellos factores menos sobresalientes o redundantes, y utilizar aquellos que según la etapa de selección de características son determinantes. A continuación, se detalla la metodología específica que se utilizó para determinar cuáles son los algoritmos que mejor se desempeñan, y cuáles son los factores de riesgo con los que mejor lo hacen.

Etapa 1: En esta etapa se establecen los nombres de todas las variables que van a ser incluidas en el estudio, para un mejor manejo de la base de datos, aquellas que sean tipo cadena son convertidas a valores numéricos.

Etapa 2: Prueba de algoritmos de aprendizaje automático, utilizando las 23 variables para medir el rendimiento, utilizando métricas de validación; sensibilidad, especificidad, exactitud y curva ROC.

Etapa 3: Las técnicas de selección de características generan subconjuntos de atributos que tienen buena capacidad predictiva. Se basan en el supuesto de que un atributo dado puede tener un mejor poder predictivo cuando se combina con otros, en comparación de cuando se usa solo (Rathore et al., 2014).

En Weka, la pestaña *"Select Atributes"* permite acceder al área de selección atributos. Esta opción permite la búsqueda a través de todas las posibles combinaciones de atributos en los datos para encontrar qué subconjunto funcionan mejor para la predicción.

Existen dos opciones (***Attribute Evaluator***, ***Search Method),*** que se deben elegir para poder realizar la selección de atributos en Weka:

***Attribute Evaluator*:** Este método es el encargado de evaluar cada uno de los casos a los que se le de acceso y le dará a cada atributo un peso específico.

Esta técnica evalúa cada atributo con el conjunto de datos; se evalúa en el contexto de la variable de salida (la clase). Esta técnica permite probar o navegar entre diferentes combinaciones de atributos del conjunto de datos para generar una lista de características seleccionadas, con estas se pretende obtener el mismo o mejor porcentaje de clasificación correcta.

El evaluador de atributos ***WrapperSubsetEval*****,** evalúa mediante el uso de algoritmos de aprendizaje. Este evaluador utiliza la validación cruzada para estimar la precisión del esquema de aprendizaje de un conjunto de atributos.

Search Method**:** El siguiente paso será elegir el método de búsqueda que será el encargado de generar el espacio de pruebas. La opción de "*Search Method*" determina el mérito de atributos individuales o subconjuntos de atributos.

El método de búsqueda ***GreedyStepwise*****,** realiza una búsqueda minuciosa hacia adelante o hacia atrás a través del espacio de los subconjuntos de atributos. Puede comenzar sin o con todos los atributos, o desde un punto arbitrario. Se detiene cuando la adición o eliminación de cualquier atributo restante produce una disminución en la evaluación.

Se genera una tabla de clasificación con todas las variables y una con selección de características para conocer aquellos atributos sobresalientes, así como el porcentaje que se obtiene en las métricas de validación, y además se obtienen las gráficas ROC de las tablas.

3.5 Implementación de un modelo de red neuronal

Dentro de la metodología propuesta se estableció la posibilidad de experimentar con redes neuronales, se utilizó el mismo set de datos y se obtuvieron los valores de las métricas de validación. Esto se llevó a cabo siguiendo la metodología antes descrita. Es importante mencionar que el modelo de red neuronal no se encuentra dentro del set de algoritmos del software Weka. Por lo tanto, se desarrolló y configuró un modelo de red neuronal en el lenguaje de programación Python. El procedimiento de

configuración para generar este modelo de red neuronal, entrenarlo, así como exportarlo se explica en el anexo 2.

3.6 Desarrollo de una aplicación Android

Se desarrolló una aplicación intuitiva para poder validar el modelo, a continuación, se muestra el diagrama de flujo.

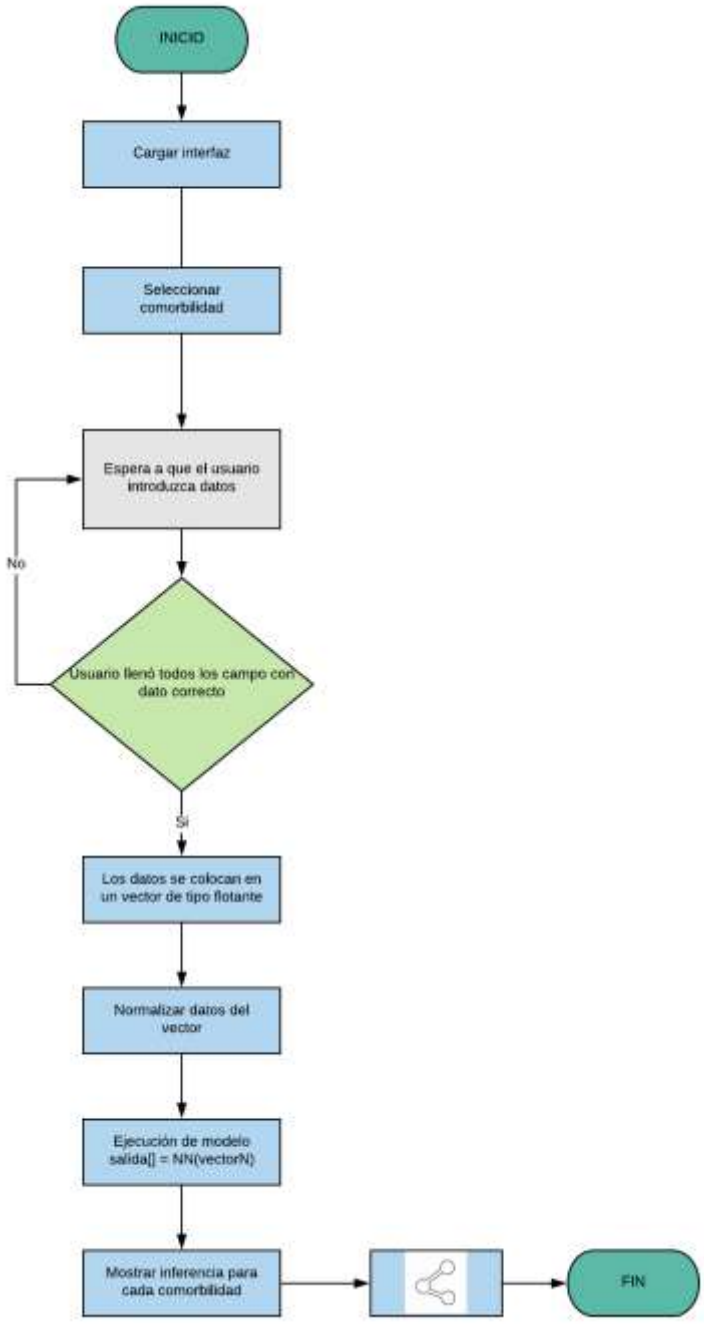


Figura 3.3 Diagrama de flujo de la aplicación.

Inicio: Se ejecuta la aplicación

Cargar interfaz: Se invocan el método onCreate() nativo de Android, en el cual se cargan , las cajas de texto, los botones, las librerías entre otros elementos necesarios para poder visualizar la interfaz previamente diseñada.
Seleccionar comorbilidad:

Espera a que el usuario introduzca datos: Una vez cargada la interfaz, la aplicación espera a que el usuario introduzca todos los datos que le solicita, y valida que cada caja de texto contenga valores de tipo numérico. Además, la aplicación espera a que el usuario decida si utiliza el botón “procesar” o utiliza el botón “limpiar”.

Validar datos: Si el usuario presiona el botón “procesar”, la aplicación manda a ejecutar los siguientes métodos: El método validar, verifica que ninguna caja de texto este vacía, si el usuario no rellena todos los campos y ha presionado, se envía un mensaje de alerta que le indica cuál caja de texto omitió rellenar y lo regresa a la caja en cuestión, esto es repetitivo si el usuario omite la misma caja u otra.

Usuario llenó todos los campos con dato correcto: Este se encarga de tomar todos los valores en las cajas de texto, esto lo logra a través de la invocación de transferencia de información entre la interfaz y el lenguaje Java.

Los datos se colocan en un vector de tipo flotante: Toda la información que se obtiene es ordenada dentro de un vector, el cual tendrá una longitud equivalente al total de los valores solicitados en la interfaz. A continuación, este método invoca al siguiente. Normalizar datos.

Normalizar datos del vector: El método de normalizar contiene dos vectores de características que fueron exportados del entrenamiento en el lenguaje de programación Python. Éstos corresponden a los máximos y mínimos de cada factor de riesgo o característica de cada registro incluidos en el set de datos. Esto se utiliza para normalizar el nuevo vector que el usuario está introduciendo, a través de la fórmula de normalización.

$$X' = a + \frac{(X - Xmin)(b - a)}{Xmax - Xmin} \qquad \textbf{(8)}$$

Una vez normalizados los datos se regresa un vector de datos normalizados.

Ejecución del modelo, salida[ ] = NN(vectorN): Éste importa el modelo NN que se encuentra en una dirección especial del proyecto Android, la carpeta que se crea para almacenarlo tiene por nombre ***"essets",*** se construye un objeto de la clase de "***TensorFlowInferenceInterface***" para poder utilizar los métodos de tensorflow en la aplicación Android y de esta manera poder hacer una clasificación. Se especifica la longitud del vector, en el que están almacenados los valores normalizados, y también se especifica qué tipo de clasificación es, si es binaria o de otra naturaleza, la anterior configuración se establece cuando se crea el objeto, como se muestra a continuación.

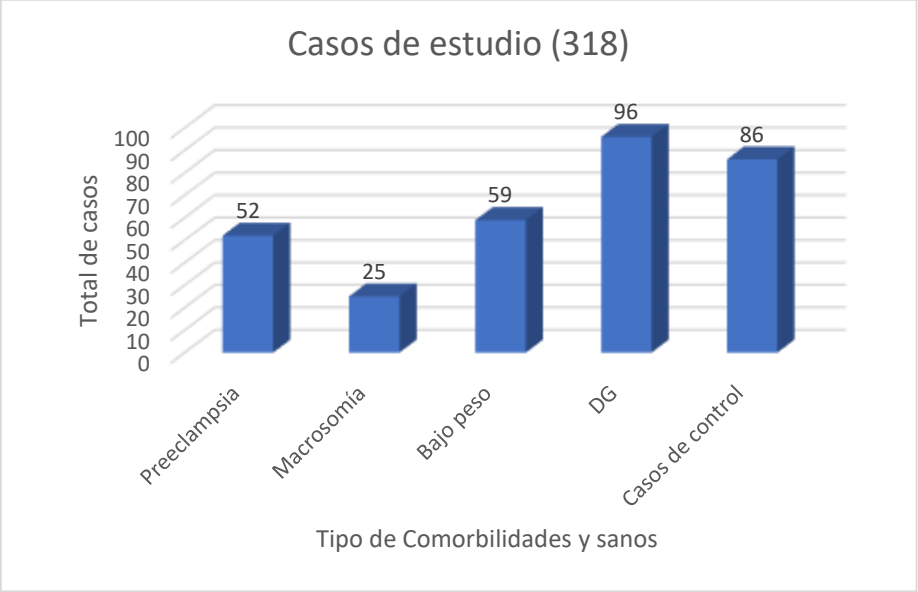
Objeto.feed(input_nodo, vector[ ], 1,longitud);

Mostrar inferencia para cada comorbilidad: Para hacer la inferencia se establece un umbral para decidir si la salida del modelo pertenece a la clase 0 (ausencia de comorbilidad) o a la clase 1 (si existe comorbilidad). La inferencia es mostrada en un mensaje de la aplicación.

Compartir: Se genera un archivo txt con todos los datos de entrada y con el resultado de la inferencia, este puede ser compartido a través de correo o en alguna red social, como what´s App.

CAPÍTULO IV. RESULTADOS

La base de datos que se utilizó para poder llevar a cabo el estudio que se describe en este trabajo de investigación está conformado de la siguiente manera como se muestra en la gráfica 3.1.



Gráfica 3.1 total de casos de estudio

Se puede observar que se utilizan datos de 4 comorbilidades, Preeclampsia, Macrosomía, Bajo peso, Diabetes Gestacional y también se utilizan una muestra de embarazos en los cuales no existieron complicaciones durante o al momento de la resolución.

En la tabla 4.1 se muestra el ejemplo de un registro con datos crudos, tal y como es recolectado de los expedientes clínicos, y en la tabla 4.2 se muestra el mismo registro pero adecuado, utilizando los criterios de la metodología propuesta y descrita en el capítulo III.

Tabla 4.1 valores crudos

#	Característica	Valor
1	edad_m	27
2	escolaridad	Secundaria
3	ant_fam_dm2	No
4	parentesco_dm2	No
5	ant_fam_hta	Si

continuación tabla 4.1

Tabla 4.2 valores adecuados

#	Variable	Valor
1	edad	27
2	escolaridad	2
3	antFamDm2	0
4	parenFamDm2	0
5	antFamHta	1

#	Característica	Valor	#	Variable	Valor
6	parentesco_hta	No especifica	6	parenFamHta	-1
7	Contamiación	No	7	contaminacion	0
8	Peso	60	8	peso	60
9	Talla	1.46	9	talla	1.46
10	IMC	28.1479	10	IMCMat	28.1479
11	TAS	110	11	TAS	110
12	TAD	70	12	TAD	70
13	Total Gestas	3	13	totGestas	3
14	Total de partos	3	14	totPartos	3
15	Total cesareas	0	15	totAbortos	0
16	Total de Abortos	0	16	Total de Partos	0
17	Eritrocitos	3.71	17	eritrocitos	3.71
18	Hemoglobina	0	18	hemoglobina	0
19	Hematocrito	37.87	19	hematocrito	37.87
20	vgm	102	20	vgm	102
21	cmhb	34.16	21	cmhb	34.16
22	cmhbc	33.48	22	cmhbc	33.48
23	rdw	11.5	23	rdw	11.5
24	Plaquetas	264	24	plaquetas	264
25	vpm	8.58	25	vpm	8.58
26	Leucocitos	8.01	26	leucocitos	8.01
27	Neutrofilo	64	27	neutrofilo	64
28	Linfocito	26.5	28	linfocito	26.5
29	Monocito	8.7	29	monocito	8.7
30	Eosinofilo	0.8	30	eosinofilo	0.8
31	Basofilo	0.5	31	basofilo	0.5
32	Glucosa	83	32	glucosa	83
33	Urea	12.26	33	urea	12.26
34	Creatinina	0.64	34	creatinina	0.64
35	Ácido Úrico	5.2	35	acidoUri	5.2
36	bun	9	36	bun	9
37	Colesterol	2.64	37	colesterol	2.64
38	Triglicéridos	597	38	triglicéridos	597

En la siguiente tabla 4.3, se muestran el rendimiento de los 35 algoritmos utilizando el set completo de variables con un total de 53 instancias positivas y 69 instancias negativas, sin aplicar selección de características, esto para la comorbilidad de Preeclampsia.

Tabla 4.3. Rendimiento de los algoritmos sin utilizar selección de características

#	Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
1	RandomCommitte	0.88	0.899	0.932	0.89
2	RandomForest	0.84	0.957	0.942	0.91
3	J Rip	0.82	0.942	0.881	0.89
4	OneR	0.82	0.957	0.888	0.90
5	Bagging	0.8	0.957	0.921	0.89
6	LogitBoost	0.8	0.957	0.948	0.89
7	PART	0.8	0.913	0.865	0.87
8	J48	0.8	0.855	0.837	0.83
9	NaiveBayes	0.78	0.928	0.893	0.87
10	NaiveBayesUpdateable	0.78	0.928	0.893	0.87
11	AttributSelectedClassifier	0.78	0.812	0.799	0.80
12	IterativeClassifierOptimizer	0.78	0.942	0.94	0.87
13	RandomSubSpace	0.78	0.971	0.917	0.89
14	LMT	0.78	0.87	0.895	0.83
15	REPTree	0.78	0.942	0.829	0.87
16	BayesNet	0.76	0.899	0.889	0.84
17	SimpleLogistic	0.76	0.899	0.901	0.84
18	HoeffdingTree	0.76	0.884	0.88	0.83
19	MultilayerPerceptron	0.74	0.942	0.887	0.86
20	SDG	0.74	0.928	0.834	0.85
21	SMO	0.74	0.957	0.848	0.87
22	AdaBoostM1	0.74	0.957	0.937	0.87
23	ClassificationViaRegression	0.74	0.957	0.898	0.87
24	MultiClassClassifierUpdateable	0.74	0.928	0.834	0.85
25	FilteredClassifier	0.72	0.87	0.79	0.81
26	DecisionTable	0.72	0.899	0.81	0.82
27	DecisionStump	0.72	0.957	0.833	0.86
28	RandomTree	0.72	0.797	0.759	0.76
29	Logistic	0.7	0.739	0.744	0.72
30	LWL	0.7	0.957	0.817	0.85
31	MultiClassClassifier	0.7	0.739	0.744	0.72
32	RandomizableFilteredClassifier	0.66	0.826	0.74	0.76
33	IBK	0.54	0.942	0.726	0.77
34	Kstar	0.4	0.957	0.77	0.72
35	VotedPerceptron	0.34	0.913	0.718	0.67

En la tabla 4.4, se encuentran el rendimiento obtenido con las siguientes características. Se utiliza el set completo de variables con 53 instancias positivas y 69 instancias negativas, para este caso se utilizó técnicas de selección de características y 35 algoritmos.

Tabla 4.4 Rendimiento de los algoritmos aplicando selección de características para Preeclampsia

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
1	RandomForest (7)	edad, IMCMat, TAS, totGestas, vpm, neutrofilo, eosinofilo	0.933	0.92	0.9	0.942
2	NaiveBayes (3)	peso, TAS, colesterol	0.929	0.92	0.88	0.957
3	NaiveBayesUpdateable (3)	peso, TAS, colesterol	0.929	0.92	0.88	0.957
4	IBK (4)	parenFamHta, TAS, TAD, cmhb, linfocito	0.897	0.91	0.88	0.928
5	HoeffdingTree (2)	TAS, colesterol	0.94	0.92	0.88	0.957
6	SDG (5)	escolaridad, TAS, cmhb, monocito, acidoUri	0.908	0.92	0.86	0.957
7	MultiClassClassifierUpdateable (5)	escolaridad, TAS, cmhb, monocito, acidoUri	0.908	0.92	0.86	0.957
8	Logistic (4)	TAS, eosinofilo, urea, acidoUri	0.922	0.92	0.84	0.971
9	MultilayerPerceptron (5)	parenFamDm2, TAS, TAD, totAbortos, bun	0.903	0.93	0.84	1
10	Kstar (2)	TAS, totGestas	0.916	0.91	0.84	0.957
11	MultiClassClassifier (4)	TAS, eosinofilo, urea, acidoUri	0.922	0.92	0.84	0.971
12	RandomizableFilteredClassifier (2)	TAS, totAbortos	0.885	0.90	0.84	0.942
13	J48 (4)	TAS, TAD, hemoglobina, neutrofilo	0.886	0.88	0.84	0.913
14	RandomTree (3)	antFamHta, TAS, totGestas	0.883	0.90	0.84	0.942

Continuación tabla 4.4

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
15	SimpleLogistic (1)	TAS	0.922	0.90	0.82	0.957
16	AdaBoostM1 (5)	peso, TAS, TAD, totAbortos, neutrofilo	0.948	0.90	0.82	0.957
17	ClassificationViaRegression (4)	parenFamHta, talla, TAS, acidoUri	0.91	0.90	0.82	0.957
18	IterativeClassifierOptimizer (2)	antFamDm2, TAS	0.851	0.90	0.82	0.957
29	LogitBoost (2)	parenFamDm2, TAS	0.884	0.90	0.82	0.957
20	RandomCommitte (3)	parenFamDm2, TAS, totGestas	0.909	0.89	0.82	0.942
21	OneR (1)	TAS	0.88	0.90	0.82	0.957
22	LMT (1)	TAS	0.922	0.90	0.82	0.957
23	SMO (1)	TAS	0.878	0.89	0.8	0.957
24	AttributSelectedClassifier (1)	TAS	0.842	0.88	0.8	0.942
25	Bagging (1)	TAS	0.899	0.89	0.8	0.957
26	RandomSubSpace (1)	TAS	0.866	0.89	0.8	0.957
27	Jrip (3)	TAS, vgm, neutrofilo	0.883	0.89	0.8	0.957
28	PART (5)	escolaridad, peso, TAS, totAbortos, leucocitos	0.835	0.89	0.8	0.957
29	REPTree (2)	TAS, urea	0.852	0.88	0.78	0.957
30	BayesNet (1)	TAS	0.833	0.86	0.72	0.957
31	LWL (3)	antFamHta, TAS, TAD	0.889	0.88	0.72	1
32	FilteredClassifier (1)	TAS	0.833	0.86	0.72	0.957
33	DecisionTable (1)	TAS	0.834	0.86	0.72	0.957
34	DecisionStump (1)	TAS	0.833	0.86	0.72	0.957
35	VotedPerceptron (4)	edad, TAS, neutrofilo, colesterol	0.86	0.76	0.5	0.942

La tabla 4.3 muestra el rendimiento de los 35 algoritmos para la comorbilidad de Preeclampsia, el algoritmo RandomComitte obtiene el mejor, la sensibilidad es de 0.88, la exactitud de 0.89 y un rendimiento de curva ROC de 0.932, lo sigue de cerca el algoritmo RandomForest con una sensibilidad de 0.84, la exactitud de 0.91 y un rendimiento en la curva ROC de 0.942, todo lo anterior es sin aplicar técnicas de selección de características.

En la tabla 4.4 se muestran los resultados obtenidos siguiendo la metodología propuesta y aplicando técnicas de selección de características descritas en el capítulo III. El algoritmo RandomForest relaciona 7 factores de riesgo como los más sobresalientes durante el proceso de clasificación binaria para la comorbilidad de preeclampsia, con los cuales obtiene un rendimiento de sensibilidad, exactitud y curva ROC de 0.9, 0.92 y 0.933 respectivamente. El rendimiento mejora cuando se aplica selección de características.

En la figura N, se muestra la curva ROC de los 5 algoritmos que obtuvieron el rendimiento más alto en el proceso dicotómico de clasificación, estos resultados son aplicando selección de características.

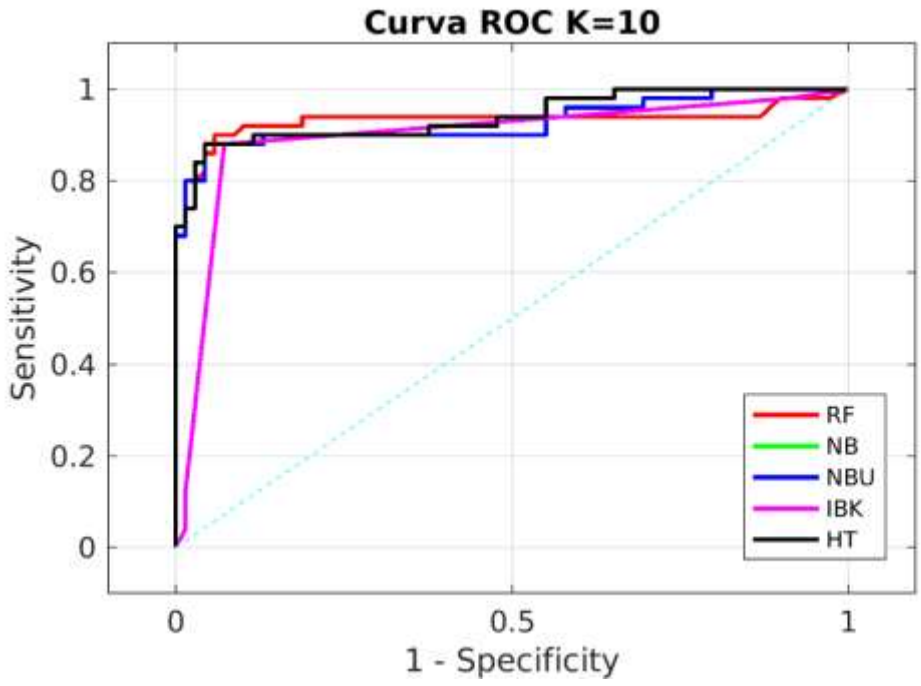


Figura 4.1. Curva ROC de los 5 mejores algoritmos para la comorbilidad de preeclampsia

En la tabla 4.5, se muestra el rendimiento de los 35 algoritmos para la comorbilidad de macrosomía, pero sin aplicar selección de atributos, se encuentran ordenados de forma descendente en función del máximo rendimiento obtenido en la variable se sensibilidad con las siguientes características; se utiliza el set completo de variables y 25 instancias positivas, así como 43 instancias negativas sin selección de atributos.

Tabla 4.5. Rendimiento de los algoritmos para la comorbilidad de macrosomía sin selección de características

#	Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
1	RandomTree	0.72	0.767	0.744	0.75
2	J48	0.64	0.791	0.695	0.74
3	LogitBoost	0.6	0.86	0.782	0.76
4	RandomCommitte	0.6	0.767	0.764	0.71
5	SimpleLogistic	0.52	0.767	0.628	0.68
6	RandomizableFilteredClassifier	0.52	0.721	0.624	0.65
7	PART	0.52	0.814	0.619	0.71
8	AdaBoostM1	0.48	0.791	0.676	0.68
9	ClassificationViaRegression	0.48	0.767	0.646	0.66
10	IterativeClassifierOptimizer	0.48	0.744	0.646	0.65
11	Logistic	0.44	0.674	0.527	0.59
12	MultilayerPerceptron	0.44	0.674	0.521	0.59
13	MultiClassClassifier	0.44	0.674	0.527	0.59
14	NaiveBayes	0.4	0.884	0.657	0.71
15	NaiveBayesUpdateable	0.4	0.884	0.657	0.71
16	LWL	0.4	0.93	0.634	0.74
17	Bagging	0.4	0.791	0.652	0.65
18	RandomForest	0.4	0.814	0.708	0.66
19	SMO	0.36	0.837	0.599	0.66
20	Kstar	0.36	0.837	0.695	0.66
21	OneR	0.36	0.86	0.61	0.68
22	LMT	0.36	0.767	0.614	0.62
23	SDG	0.32	0.721	0.52	0.57
24	IBK	0.32	0.698	0.501	0.56
25	MultiClassClassifierUpdateable	0.32	0.721	0.52	0.57
26	J Rip	0.28	0.767	0.541	0.59
27	AttributSelectedClassifier	0.24	0.791	0.559	0.59
28	DecisionTable	0.24	0.86	0.561	0.63
29	HoeffdingTree	0.24	0.93	0.593	0.68
30	REPTree	0.24	0.884	0.601	0.65
31	RandomSubSpace	0.16	0.93	0.572	0.65
32	VotedPerceptron	0.12	1	0.586	0.68
33	DecisionStump	0.12	0.767	0.401	0.53
34	BayesNet	0.04	0.814	0.363	0.53
35	FilteredClassifier	0.04	0.86	0.402	0.56

En la tabla 4.6 se muestra el rendimiento de los 35 algoritmos para la comorbilidad de macrosomía utilizando el set completo de variables y 25 instancias positivas, así como 43 instancias negativas con selección de características.

Tabla 4.6 Rendimiento de los algoritmos para la comorbilidad de Macrosomía con selección de características

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
1	IBK (3)	neutrofilo, glucosa, colesterol	0.861	89.71	0.84	0.93
2	MultilayerPerceptron (6)	antFamDm2, vgm, cmhb, neutrofilo, bun, colesterol	0.852	89.71	0.76	0.977
3	RandomTree (2)	creatinina, colesterol	0.833	85.29	0.76	0.907
4	LogitBoost (2)	neutrofilo, colesterol	0.783	83.82	0.72	0.907
5	J48 (3)	hemoglobina, urea, acidoUri	0.737	80.88	0.72	0.86
6	IterativeClassifierOptimizer (3)	parenFamDm2, neutrofilo, colesterol	0.767	82.35	0.68	0.907
7	RandomCommitte (4)	antFamHta, basofilo, creatinina, colesterol	0.838	80.88	0.68	0.884
8	SDG (7)	parenFamHta, peso, totCesareas, eritrocitos, neutrofilo, linfocito, colesterol	0.785	82.35	0.64	0.93
9	MultiClassClassifierUpdateable (7)	parenFamHta, peso, totCesareas, eritrocitos, neutrofilo, linfocito, colesterol	0.785	82.35	0.64	0.93
10	RandomizableFilteredClassifier (5)	edad, linfocito, basofilo, creatinina, colesterol	0.741	79.41	0.64	0.884

Continuación tabla 4.6

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
11	RandomForest (2)	creatinina, colesterol	0.814	80.88	0.64	0.907
12	Logistic (3)	linfocito, eosinofilo, colesterol	0.72	82.35	0.6	0.953
13	Kstar (4)	antFamHta, IMCMat, neutrofilo, colesterol	0.789	82.35	0.6	0.953
14	MultiClassClassifier (3)	linfocito, eosinofilo, colesterol	0.72	82.35	0.6	0.953
15	LMT (3)	eritrocitos, neutrofilo, colesterol	0.795	80.88	0.6	0.93
16	NaiveBayes (7)	antFamDm2, parenFamDm2, TAD, neutrofilo, linfocito, glucosa, acidoUri	0.717	82.35	0.56	0.977
17	NaiveBayesUpdateable (7)	antFamDm2, parenFamDm2, TAD, neutrofilo, linfocito, glucosa, acidoUri	0.717	82.35	0.56	0.977
18	LWL (1)	colesterol	0.703	77.94	0.56	0.907
19	Jrip (3)	cmhb, neutrofilo, acidoUri	0.687	77.94	0.56	0.907
20	SMO (8)	antFamDm2, peso, IMCMat, totCesareas, plaquetas, neutrofilo, linfocitos, trigliceridos	0.737	79.41	0.52	0.953
21	OneR (2)	eritrocitos, colesterol	0.69	73.53	0.52	0.86

Continuación tabla 4.6

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
22	PART (3)	linfocito, urea, acidoUri	0.724	73.53	0.52	0.86
23	SimpleLogistic (4)	antFamDm2, eritrocitos, neutrofilo, colesterol	0.757	75.00	0.48	0.907
24						
25	AdaBoostM1 (4)	TAD, totPartos, neutrofilo, acidoUri	0.78	77.94	0.48	0.953
26	Bagging (4)	antFamDm2, hemoglobina, neutrofilo, colesterol	0.768	72.06	0.44	0.884
27	HoeffdingTree (7)	escolaridad, antFamDm2, peso, plaquetas, neutrofilo, linfocito, colesterol	0.767	75.00	0.44	0.93
28	ClassificationViaRegression (3)	antFamDm2, eritrocitos, neutrofilo	0.741	70.59	0.4	0.884
29	REPTree (4)	contaminacion, TAD, creatinina, colesterol	0.653	69.12	0.32	0.907
30	DecisionTable (2)	neutrofilo, acidoUri	0.65	70.59	0.28	0.953
31	VotedPerceptron (3)	escolaridad, linfocito, colesterol	0.59	72.06	0.24	1
32	AttributSelectedClassifier (1)	acidoUri	0.529	72.06	0.24	1
33	RandomSubSpace (5)	edad, linfocito, basofilo, creatinina, colesterol	0.747	69.12	0.24	0.953
34	DecisionStump (1)	colesterol	0.53	72.06	0.24	1

La tabla 4.5 contiene el rendimiento de los algoritmos obtenido para la comorbilidad de macrosomía utilizando la configuración antes descrita. El algoritmo RandomTree obtiene un rendimiento de sensibilidad, exactitud y curva ROC de 0.72, 0.75 y 0.744 respectivamente, esto es sin aplicar técnicas de selección de características.

La tabla 4.6 muestra el rendimiento para la misma comorbilidad, pero aplicando selección de características y se obtiene lo siguiente; el algoritmo IBK relaciona 3 factores de riesgo como los más determinantes para clasificar correctamente entre las dos clases (presencia o ausencia de la comorbilidad).

Los factores asociados son neutrofilo, glucosa y el colesterol, con éstos se obtiene un rendimiento de sensibilidad, exactitud y curva ROC de 0.84, 0.90 y 0.861.

En la figura 4.2, se muestras la curva ROC de los 5 algoritmos que obtuvieron el mejor rendimiento de clasificación, aplicando técnicas de selección de características.

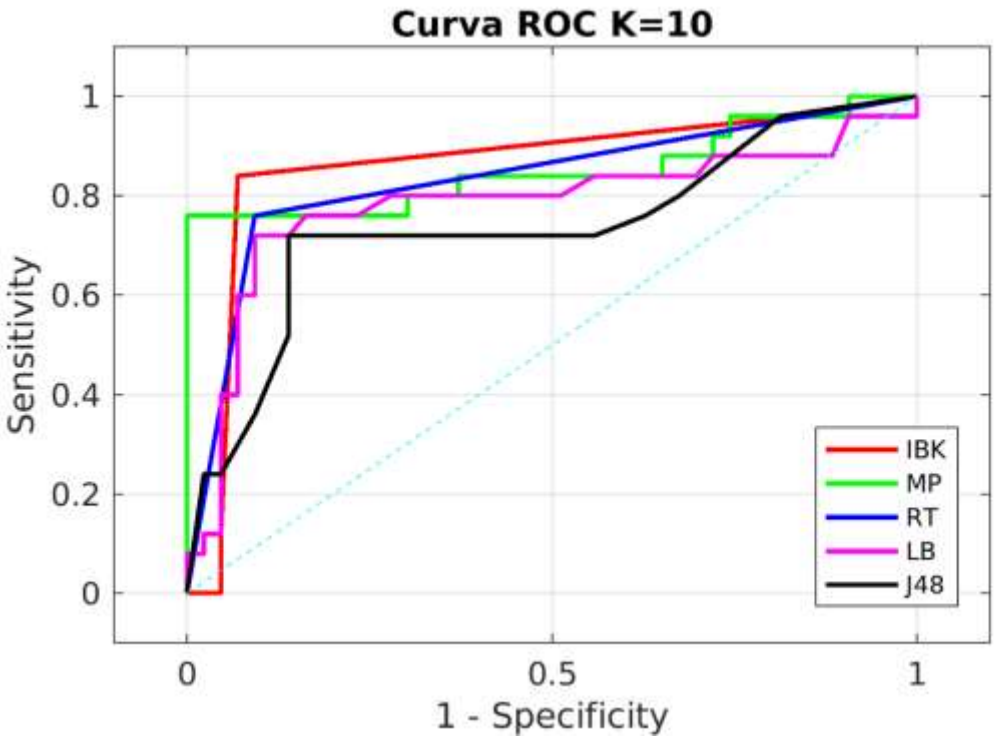


Figura 4.2. Curva ROC de los 5 algoritmos con más alto rendimiento de clasificación para macrosomía.

La tabla 4.7 muestra el rendimiento de los algoritmos obtenido para la comorbilidad de bajo peso utilizando la siguiente configuración; se utiliza el set completo de variables, así como 58 instancias positivas y 67 negativas sin selección de características para bajo peso. El algoritmo ClasificationViaRegression obtiene un rendimiento de sensibilidad, exactitud y curva ROC de 0.655, 0.72 y 0.734 respectivamente, todo lo anterior sin aplicar técnicas de selección de características.

Tabla 4.7 Rendimiento de los algoritmos para la comorbilidad de bajo peso

#	Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
1	ClassificationViaRegression	0.655	0.779	0.734	0.72
2	RandomCommitte	0.655	0.676	0.705	0.67
3	RandomForest	0.621	0.824	0.748	0.73
4	J48	0.621	0.765	0.683	0.70
5	RandomizableFilteredClassifier	0.569	0.779	0.696	0.68
6	LogitBoost	0.569	0.706	0.679	0.64
7	LMT	0.552	0.779	0.664	0.67
8	J Rip	0.552	0.838	0.657	0.71
9	RandomTree	0.552	0.75	0.651	0.66
10	RandomSubSpace	0.534	0.882	0.721	0.72
11	Bagging	0.534	0.882	0.698	0.72
12	SDG	0.534	0.838	0.686	0.70
13	MultiClassClassifierUpdateable	0.534	0.838	0.686	0.70
14	SMO	0.517	0.882	0.7	0.71
15	MultilayerPerceptron	0.517	0.794	0.682	0.67
16	PART	0.517	0.618	0.587	0.57
17	SimpleLogistic	0.5	0.838	0.651	0.68
18	REPTree	0.483	0.912	0.711	0.71
19	OneR	0.466	0.721	0.593	0.60
20	NaiveBayes	0.448	0.912	0.79	0.70
21	NaiveBayesUpdateable	0.448	0.912	0.79	0.70
22	IBK	0.448	0.838	0.663	0.66
23	HoeffdingTree	0.431	0.897	0.771	0.68
24	AdaBoostM1	0.431	0.853	0.679	0.66
25	AttributSelectedClassifier	0.431	0.838	0.635	0.65
26	DecisionTable	0.431	0.779	0.623	0.62
27	Logistic	0.431	0.706	0.589	0.58
28	MultiClassClassifier	0.431	0.706	0.589	0.58
29	VotedPerceptron	0.414	0.941	0.689	0.70
30	IterativeClassifierOptimizer	0.379	0.882	0.631	0.65
31	BayesNet	0.379	0.838	0.624	0.63
32	FilteredClassifier	0.362	0.853	0.61	0.63
33	LWL	0.328	0.985	0.669	0.68
34	DecisionStump	0.328	0.985	0.62	0.68
35	Kstar	0.241	0.882	0.59	0.59

En la siguiente tabla 4.8 se muestra el rendimiento de los algoritmos con el set completo de variables, así como 58 instancias positivas y 67 negativas para la comorbilidad de Bajo peso. El algoritmo RandomComitte relaciona dos factores de riesgo como los más determinantes en la clasificación, y obtiene un rendimiento de sensibilidad, exactitud y curva ROC de 0.741, 0.72 y 0.73, aplicando técnicas de selección de características.

Tabla 4.8 Rendimiento de los algoritmos aplicando técnicas de selección de características para bajo peso

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
1	RandomCommitte (2)	antFamHta, colesterol	0.73	0.72	0.741	0.706
2	RandomizableFilteredClassifier (2)	totAbortos, colesterol	0.77	0.74	0.724	0.75
3	MultilayerPerceptron (6)	TAS, totAbortos, neutrofilo, eosinofilo, creatinina, trigliceridos	0.81	0.82	0.707	0.912
4	RandomTree (4)	antFamHta, contaminacion, totAbortos, colesterol	0.73	0.72	0.707	0.735
5	IBK (3)	totAbortos, creatinina, colesterol	0.74	0.73	0.69	0.765
6	PART (5)	contaminacion, TAS, cmhb, rdw, trigliceridos	0.8	0.78	0.672	0.868
7	J48 (5)	contaminacion, TAS, cmhb, neutrofilo, trigliceridos	0.77	0.81	0.672	0.926
8	Jrip (4)	contaminacion, TAS, totCesareas, trigliceridos	0.76	0.76	0.672	0.838
9	SimpleLogistic (6)	antFamDm2, contaminacion, TAS, cmhbc, neutrofilo, colesterol	0.75	0.76	0.672	0.838

Continuación tabla 4.8

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
10	RandomForest (2)	antFamHta, colesterol	0.713	0.71	0.672	0.735
11	LMT (5)	contaminacion, TAS, rdw, neutrofilo, trigliceridos	0.8	0.79	0.655	0.897
12	NaiveBayes (9)	antFamDm2, parenFamDm2, TAS, totAbortos, plaquetas, neutrofilo, acidoUri, colesterol, trigliceridos	0.81	0.81	0.638	0.956
13	NaiveBayesUpdateable (9)	antFamDm2, parenFamDm2, TAS, totAbortos, plaquetas, neutrofilo, acidoUri, colesterol, trigliceridos	0.81	0.81	0.638	0.956
14	Bagging (5)	contaminacion, TAS, hemoglobina, neutrofilo, trigliceridos	0.81	0.77	0.621	0.897
15	ClassificationViaRegression (4)	TAS, neutrofilo, linfocito, trigliceridos	0.76	0.75	0.621	0.868
16	REPTree (8)	antFamDm2, contaminacion, TAS, totPartos, totAbortos, monocito, bun, trigliceridos	0.73	0.71	0.621	0.779
17	RandomSubSpace (4)	TAS, linfocito, glucosa, trigliceridos	0.81	0.75	0.586	0.882
18	SDG (4)	escolaridad, totAbortos, neutrofilo, trigliceridos	0.73	0.75	0.586	0.882
19	MultiClassClassifierUpdateable (4)	escolaridad, totAbortos, neutrofilo, trigliceridos	0.73	0.75	0.586	0.882

Continuación tabla 4.8

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
20	HoeffdingTree (4)	TAS, neutrofilo, glucosa, trigliceridos	0.8	0.78	0.569	0.956
21	SMO (3)	hematocrito, neutrofilo, trigliceridos	0.73	0.74	0.569	0.882
22	Logistic (3)	TAD, neutrofilo, colesterol	0.72	0.73	0.569	0.868
23	MultiClassClassifier (3)	TAD, neutrofilo, colesterol	0.72	0.73	0.569	0.868
24	Kstar (2)	contaminacion, trigliceridos	0.75	0.75	0.534	0.941
25	IterativeClassifierOptimizer (3)	TAS, totAbortos, trigliceridos	0.72	0.72	0.534	0.882
26	LogitBoost (4)	TAS, totAbortos, eosinofilo, trigliceridos	0.7	0.66	0.517	0.779
27	OneR (1)	trigliceridos	0.64	0.65	0.517	0.765
28	AdaBoostM1 (2)	TAS, trigliceridos	0.72	0.71	0.466	0.926
29	LWL (6)	edad, TAS, neutrofilo, acidoUri, colesterol, trigliceridos	0.62	0.71	0.431	0.941
30	AttributSelectedClassifier (3)	TAS, linfocito, trigliceridos	0.68	0.69	0.414	0.926
31	BayesNet (2)	colesterol, trigliceridos	0.62	0.69	0.345	0.985
32	DecisionTable (1)	trigliceridos	0.64	0.68	0.328	0.985
33	FilteredClassifier (1)	trigliceridos	0.62	0.68	0.328	0.985
34	DecisionStump (1)	trigliceridos	0.62	0.68	0.328	0.985
35	VotedPerceptron (4)	TAD, linfocito, bun, trigliceridos	0.69	0.67	0.276	1

En la figura 4.3, se muestran las curvas ROC de los 5 algoritmos que obtuvieron el rendimiento más alto en sensibilidad, para la comorbilidad de bajo peso.

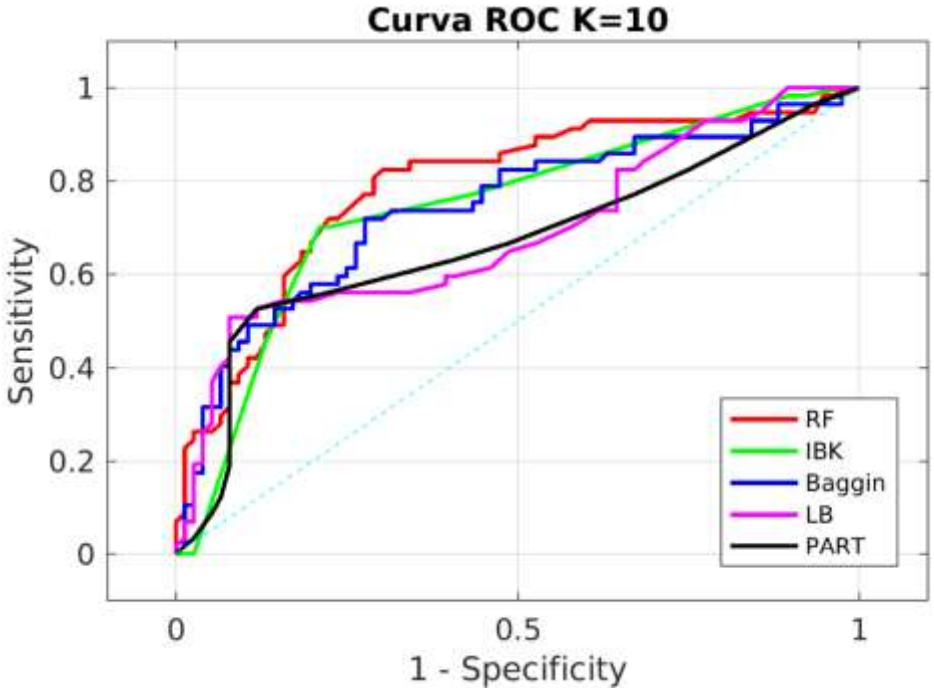


Figura 4.3. Curva ROC de los 5 algoritmos con el mejor rendimiento para la comorbilidad de bajo peso

La tabla 4.9 muestra el rendimiento de los algoritmos con la configuración antes descrita en la tabla N para la comorbilidad de diabetes gestacional. El algoritmo RandomSubSpace obtiene un rendimiento de sensibilidad, exactitud y curva ROC de 0.99, 0.95 y 0.978 respectivamente, lo anterior sin utilizar técnicas de selección de características

Tabla 4.9 Rendimiento de los algoritmos sin selección de atributos para la comorbilidad de DG

#	Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
1	RandomSubSpace	0.99	0.907	0.978	95.05
2	BayesNet	0.979	0.919	0.981	95.05
3	RandomCommitte	0.979	0.919	0.989	95.05
4	PART	0.979	0.953	0.963	96.70
5	AdaBoostM1	0.969	0.919	0.973	94.51
6	J48	0.969	0.942	0.96	95.60
7	OneR	0.958	0.837	0.898	90.11
8	RandomForest	0.958	0.942	0.987	95.05
9	AttributSelectedClassifier	0.948	0.919	0.958	93.41
10	FilteredClassifier	0.948	0.942	0.948	94.51
11	MultilayerPerceptron	0.938	0.965	0.962	95.05
12	ClassificationViaRegression	0.938	0.942	0.971	93.96
13	LogitBoost	0.938	0.907	0.972	92.31
14	IterativeClassifierOptimizer	0.927	0.93	0.968	92.86
15	REPTree	0.927	0.907	0.957	91.76
16	SDG	0.917	0.849	0.883	88.46
17	MultiClassClassifierUpdateable	0.917	0.849	0.883	88.46
18	DecisionTable	0.917	0.872	0.971	89.56
19	J Rip	0.917	0.884	0.945	90.11
20	LMT	0.917	0.93	0.979	92.31
21	Logistic	0.885	0.86	0.896	87.36
22	Bagging	0.885	0.953	0.963	91.76
23	MultiClassClassifier	0.885	0.96	0.896	87.36
24	SMO	0.865	0.826	0.845	84.62
25	RandomTree	0.865	0.86	0.863	86.26
26	SimpleLogistic	0.854	0.826	0.909	84.07
27	LWL	0.854	0.965	0.969	90.66
28	DecisionStump	0.854	0.965	0.872	90.66
29	IBK	0.813	0.884	0.855	84.62
30	HoeffdingTree	0.667	0.837	0.841	74.73
31	NaiveBayes	0.656	0.884	0.861	76.37
32	NaiveBayesUpdateable	0.656	0.884	0.861	76.37
33	VotedPerceptron	0.635	0.721	0.715	67.58
34	RandomizableFilteredClassifier	0.635	0.756	0.709	69.23
35	Kstar	0.583	0.907	0.881	73.63

La tabla 4.10 contiene el rendimiento de los algoritmos para la comorbilidad de DG aplicando selección de características. El algoritmo ClasificationViaRegression relaciona 5 factores de riesgo como los más importantes para clasificar correctamente entre el set de datos. El rendimiento que obtiene este algoritmo con estos 5 factores es el siguiente; 0.99, 0.98 y 0.99 en la sensibilidad, exactitud y curva ROC, respectivamente.

Tabla 4.10 Rendimiento de los algoritmos para DG, aplicando selección de características

	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
1	ClassificationViaRegression (5)	edad, antFamHta, parenFamHta, contaminacion, plaquetas	0.99	0.98	0.99	0.965
2	PART (3)	edad, parenFamHta, contaminacion	0.97	0.98	0.99	0.965
3	J48 (3)	edad, parenFamHta, contaminacion	0.97	0.98	0.99	0.965
4	MultilayerPerceptron (6)	edad, antFamHta, parenFamHta, contaminacion, talla, totAbortos	0.98	0.98	0.979	0.988
5	RandomForest (6)	edad, antFamDm2, antFamHta, parenFamHta, contaminacion, totGestas	0.991	0.98	0.979	0.977
6	RandomizableFilteredClassifier (4)	edad, antFamDm2, antFamHta, parenFamHta	0.98	0.97	0.979	0.953
7	RandomCommitte (5)	edad, antFamDm2, parenFamHta, contaminacion, totPartos	0.99	0.96	0.979	0.93
8	Kstar (4)	edad, parenFamDm2, parenFamHta, contaminacion	0.99	0.97	0.969	0.965
9	IBK (4)	edad, antFamDm2, antFamHta, parenFamHta	0.99	0.96	0.969	0.953
10	LMT (5)	edad, parenFamHta, contaminacion, totCesareas	0.97	0.96	0.969	0.953
11	LogitBoost (5)	edad, antFamDm2, parenFamHta, contaminacion, totGestas	0.99	0.96	0.969	0.942
12	AttributSelectedClassifier (3)	edad, parenFmHta, contaminacion	0.97	0.96	0.969	0.942
13	RandomTree (2)	edad, parenFamHta	0.95	0.96	0.969	0.942

Continuación tabla 4.10

#	Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
14	Jrip (3)	parenFamHta, contamiacion, TAD	0.95	0.96	0.969	0.942
15	BayesNet (3)	edad, parenFamHta, glucosa	0.98	0.96	0.958	0.953
16	REPTree (3)	edad, parenFamHta, contaminacion	0.97	0.96	0.958	0.953
17	RandomSubSpace (3)	edad, parenFamDm2, parenFamHta	0.98	0.95	0.958	0.93
18	MultiClassClassifier (4)	antFamHta, parenFamHta, totCesareas, plaquetas	0.91	0.94	0.958	0.919
19	Logistic (4)	antFamHta, parenFamHta, totCesareas, plaquetas	0.91	0.93	0.958	0.907
20	OneR (1)	parenFamHta	0.9	0.90	0.958	0.837
21	FilteredClassifier (5)	edad, parenFamDm2, parenFamHta, peso, TAS	0.96	0.95	0.948	0.953
22	DecisionTable (4)	edad, parenFamHta, contaminacion, TAS	0.98	0.94	0.948	0.93
23	SimpleLogistic (4)	antFamDm2, antFamHta, parenFamHta, plaquetas	0.91	0.93	0.948	0.919
24	IterativeClassifierOptimizer (4)	edad, antFamHta, parenFamHta, TAD	0.98	0.95	0.938	0.965
25	Bagging (4)	edad, parenFamHta, contaminacion, peso	0.97	0.93	0.927	0.942
26	SDG (3)	antFamHta, parenFamHta	0.92	0.92	0.917	0.93
27	MultiClassClassifierUpdateable (2)	antFamHta, parenFamHta	0.92	0.92	0.917	0.93
28	HoeffdingTree (3)	parenFamDm2, parenFamHta	0.84	0.88	0.917	0.837
29	AdaBoostM1 (3)	edad, parenFamHta, contaminacion	0.94	0.87	0.885	0.849
30	VotedPerceptron (3)	antFamHta, parenFamHta, talla	0.93	0.92	0.875	0.965
31	LWL (3)	antFamDm2, parenFamDm2, parenFamHta	0.97	0.91	0.865	0.965
32	NaiveBayes (1)	parenFamHta	0.87	0.89	0.865	0.919
33	DecisionStump (1)	parenFamHta	0.87	0.91	0.854	0.965
34	SMO (6)	edad, antFamDm2, parentFamDm2, antFamHta, talla, IMCMat	0.78	0.77	0.771	0.779
35	NaiveBayesUpdateable (1)	parenFamHta	0.86	0.76	0.656	0.884

En la figura 4.4, se muestran la Curva ROC de los 5 mejores algoritmos que obtuvieron el rendimiento más alto en sensibilidad, para comorbilidad de diabetes gestacional.

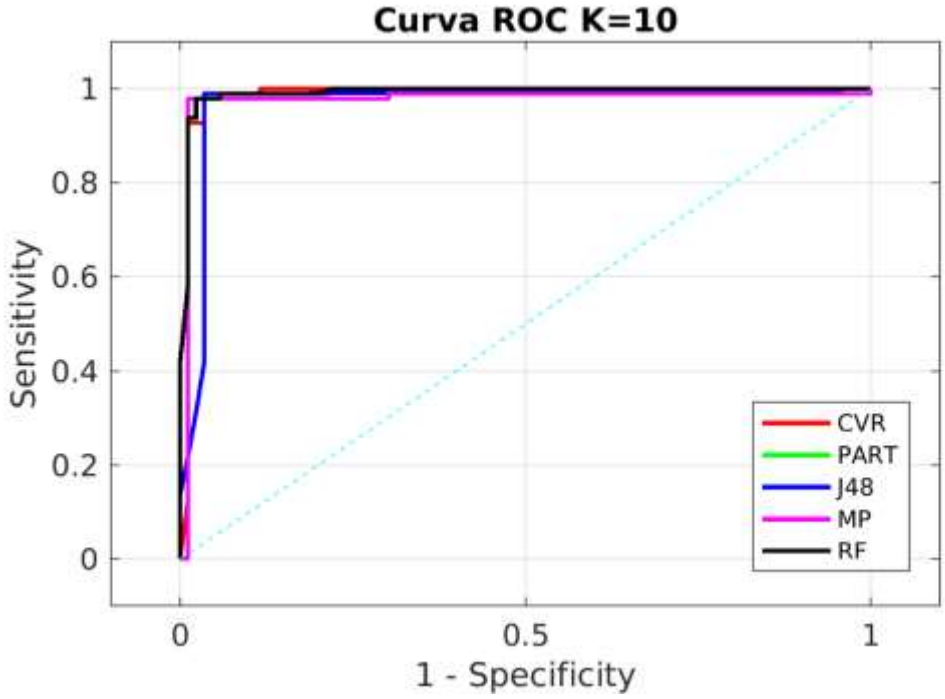


Figura 4.4. Curva ROC de los 5 mejores algoritmos para la comorbilidad de DG.

En la Tabla 4.11 y 4.12, se muestra el rendimiento de los mejores algoritmos para cada comorbilidad, así como los factores de riesgo que mejor se relacionan para obtener dicho rendimiento, haciendo uso de la metodología propuesta con los algoritmos de aprendizaje automático, estos factores resultan ser determinantes para clasificar correctamente entre cada comorbilidad.

Para la comorbilidad de preeclampsia se obtiene que la variable TAS es un factor determinante, esto es bien conocido en el área del cuidado de la salud, al modelar los datos se puede observar que es relacionada por cada algoritmo utilizado. Por ejemplo el mejor rendimiento en exactitud y curva ROC se obtiene con 7 variables; la edad, el IMC, TAS, el total de gestas, vpm,, Neutrofilo, Eosinofilo, con un rendimiento de 0.92 y 0.933, esto con el algoritmo RandomForest respectivamente.

Sin embargo, al no agregar la TAS al modelado se obtienen otros factores sobresalientes. Por ejemplo el grado escolar, específicas como; la concentración de hemoglobina corpuscular media (cmhbc), Amplitud de la Distribución Eritrocitaria (rdw, *Red blood cell Distribution Width*) por sus siglas en inglés, el conteo de plaquetas, el volumen plaquetario medio (vpm), ácido úrico así como el nivel de triglicéridos, usando todos los anteriores como factores representativos se obtiene un porcentaje de 0.841 en la curva ROC y 0.82 en exactitud, con el algoritmo RandomForest, ver tabla 4.11.

En la tabla 4.11 se muestra un resumen de los factores de riesgo importantes para la comorbilidad de macrosomía. Los factores de riesgo son; neutrofilo, glucosa, colesterol, con los cuales se obtiene un rendimiento de exactitud de 0.9 y 0.861 en la curva ROC con el algoritmo el IBK, ver tabla 4.11.

Para la comorbilidad de bajo peso las variables asociadas como factores representativos fueron; el antecedente familiar con DM2, el antecedente familiar con Hta., el parentesco que se tiene con el familiar que padeció Hta., la exposición a contaminantes como el humo de leña o el polvo, el peso materno durante la gestación y la TAS, con los anteriores factores se obtiene un rendimiento en la curva ROC de 0.789, 0.75 en exactitud, esto con el algoritmo RandomForest.

Para la comorbilidad de diabetes gestacional los factores de riesgo que muestran un buen rendimiento son la edad de la gestante, tener antecedentes familiares con Hta., el parentesco con el familiar que padeció HTa., la continua exposición a contaminantes como el humo de leña y el total de plaquetas en la sangre, se obtiene un rendimiento de exactitud de 0.98, y 0.99 en la curva ROC, ver tabla 4.12.

Resumen de factores de riesgo asociados con comorbilidades perinatales en embarazos de alto riesgo, usando algoritmos de aprendizaje automático, en dos Hospitales en la ciudad de Chilpancingo, Guerrero, México, en el 2019, ver tabla 4.11 y 4.12.

Tabla 4.11 Factores de riesgo asociados a las comorbilidades de Preeclampsia y Macrosomía

PREECLAMPSIA Factores de riesgo		MACROSOMÍA Factores de riesgo
Edad	Escolaridad	Neutrofilo
IMCMat	cmhbc	Glucosa
TAS	rdw	Colesterol
Total de gestas	Plaquetas	
Volumen Plaquetar Medio (vpm)	Volumen Plaquetar Medio (vpm)	
Neutrofilo	Ácido Úrico	
Eosinofilo	Triglyceridos	

Rendimiento

Algoritmo:	**RandomForest**				**IBK**			
Configuración:	53 positivas 67 negativas con 36 variables				25 Positivas 43 Negativas con 38 variables			
ROC	**Exactitud**	**Sensibilidad**	**Especificidad**		**ROC**	**Exactitud**	**Sensibilidad**	**Especificidad**
0.933	0.92	0.9	0.942		0.861	0.9	0.84	0.93
0.841	0.82	0.736	0.881					

Tabla 4.12 Factores asociados a las comorbilidades de Bajo peso y DG.

BAJO PESO Factores de riesgo	DIABETES GESTACIONAL Factores de riesgo
Antecedente Familiar Dm2	Edad
antecedente Familiar Hta.	Antecedente Familiar Hta.
Parentesco Familiar Hta.	Parentesco Familiar Hta.
Contaminación	Contaminación
Peso	plaquetas
TAS	

Rendimiento

Algoritmo:	**RandomForest**			**ClassificationViaRegression**			
Configuración:	58 positivas 86 negativas con 20 var			96 Positivas 86 Negativas con 21 variables			
ROC	**Exactitud**	**Sensibilidad**	**Especificidad**	**ROC**	**Exactitud**	**Sensibilidad**	**Especificidad**
0.789	0.75	0.719	0.776	0.99	0.98	0.99	0.965

CAPÍTULO V. DISCUSIONES

En este capítulo de discuten los resultados obtenidos contrastándolos con resultados publicados en la literatura sobre comorbilidades perinatales. La discusión es sobre las 4 comorbilidades que se han abordado en este trabajo, Preeclampsia, Macrosomía, Bajo Peso y Diabetes Gestacional.

Los trabajos revisados en la literatura sobre comorbilidades perinatales la mayoría suelen realizar análisis estadísticos univariable en los cuales se busca relacionar a cada factor por separado con la comorbilidad, sin embargo en el presente trabajo de investigación se propone un análisis multivariable en el cual se modela el set completo de variables e instancias positivas y negativas, utilizando técnicas de selección de características lo que permite omitir y al mismo tiempo resaltar aquellas que son importantes en la clasificación binaria, todo lo anterior haciendo uso de algoritmo de aprendizaje automático.

Preeclampsia

Diversos estudios médicos describen a la preeclampsia como una comorbilidad específica del embarazo, y una de las principales causantes de complicaciones durante la gestación no sólo en México si no a nivel mundial, además estos mismos estudios afirman que no se conoce a ciencia cierta las causas de esta comorbilidad, (Rodríguez, 2017).

Por ejemplo Zaragoza-García et al. (2017), encontró que el antecedente materno con Hipertensión arterial, la cantidad de abortos previos, la exposición a contaminantes como el humo y polvo, el índice de masa corporal, la edad gestacional son factores importantes para desarrollar preeclampsia, esto lo determinaron en un estudio estadístico transversal univariable a datos extraídos de expedientes clínicos donde existió la presencia de diabetes gestacional y preeclampsia.

Según Sanchez-Carrillo et al. (2011), en su trabajo “Factores de riesgo de la preeclampsia severa en gestantes del Hospital Nacional Almanzor Aguinaga Asenjo, de 2006 a abril de 2010”, reportaron factores de riesgo como; en edades maternas menores a 18 y mayores a 35 años durante la gestación, el sobre peso, la obesidad y la nuliparidad.

En el presente trabajo de investigación, los factores o características que más sobresalen para la comorbilidad de preeclampsia; son la tensión arterial sistólica (TAS), algo que es ampliamente conocido en el área del cuidado de la salud, era de esperarse que en esté análisis estadístico resultara también serlo, si no, sería una contradicción, de los 35 algoritmos que se utilizan se puede observar que en todos aparece como factor relevante.

Existen otros factores importantes que junto a la TAS obtienen un alto rendimiento en el presente trabajo de investigación, por ejemplo; la edad materna al momento de la gestación, el índice de masa corporal, la cantidad de gestas previas, y tres factores poco relacionados, esto debido probablemente a que no se han utilizado como objeto de estudio, en el presente estudio se utilizaron características clínicas disponibles al momento de la recolección de los casos con esta comorbilidad, estos son; el volumen plaquetario medio (vpm), y la medición de neutrofilo y eosinófilo con los cuales el algoritmo *RandomForest* muestra un rendimiento de sensibilidad, especificidad, curva ROC y exactitud de 0.9, 0.942, 0.933 y 0.92 respectivamente.

Debido a que la tensión arterial sistólica resulta determinante surgió una pregunta, ¿Qué sucedería si se omitiera esta variable en el conjunto de datos de entrenamiento?, por lo que se realizó nuevamente el modelado prescindiendo de esta variable, se esperaba que ambos resultados coincidieran y que los factores de riesgo continuaran siendo los mismo en ambos entrenamientos.

En el segundo experimento sobresalen la escolaridad como característica sociodemográfica y 6 más que normalmente son variables específicas y que se obtienen con un estudio de laboratorio, éstas son; La concentración de hemoglobina corpuscular media(cmhbc), RDW, por su nombre en inglés (*Red blood cell Distribution Width*), el conteo de plaquetas, el volumen plaquetario medio (vpm), la concentración del ácido úrico y el nivel de triglicéridos. Como se puede observar la única variable que aparece en ambos experimentos es el volumen plaquetario medio (vpm), con los factores de riesgo descritos anteriormente se obtiene un rendimiento de sensibilidad, especificidad, curva ROC y exactitud de 0.736, 0.881, 0.841 y 0.82.

Los resultados del primer experimento muestran similitud en los factores de riesgo obtenidos si se comparan a otros trabajos de investigación, por ejemplo, la edad materna y el IMC, el total de gestas como factores relevantes.

Macrosomía

Los factores de riesgo con los que ha sido relacionada la macrosomía son diversos, por ejemplo, González (2012), menciona que el antecedente de peso materno antes de la gestación mayor a 90 kg, la multiparidad, embarazos prolongados, macrosomía en embarazo previo, diabetes materna entre otros más, son factores de riesgo para desarrollar macrosomía.

Según García et al. (2016), en su trabajo “Factores de riesgo de macrosomía en pacientes sin diabetes mellitus gestacional”, concluyen que los factores de riesgo como, paridad, antecedentes de macrosomía fetal, edad materna y talla materna mayor a 1.70 m, no mostraron diferencias significativas, por el contrario, hubo mayor incidencia de macrosomía en pacientes con factores metabólicos como el sobre peso y el tamiz de diabetes mellitus gestacional alterado.

Zaragoza-García et al. (2017), en un estudio realizado para determinar factores de riesgo que predisponen a desarrollar macrosomía en embarazos con diabetes gestacional encontraron que factores de riesgo tales como un sangrado > 500 mL durante el alumbramiento, primigestación, anemia materna, glucosuria, AF-DT2, AF-HTA, así como el IMC > 30 kg/m^2 no presentan un valor significativo.

En otro trabajo de investigación llevado a cabo por Ávila et al. (2013), en el hospital Civil “Dr. José Macías Hernández” en Tamaulipas, asociaron a factores de riesgo como; edad materna avanzada, ganancia ponderal gestacional > 11 kg, > 2 gestaciones, diabetes gestacional, hijos macrosómicos previos, semanas de gestación > 40 y nivel educativo medio-profesional con el desarrollo de un producto macrosómico.

En el presente trabajo también se utilizan variables similares a las descritas y utilizadas en los trabajos antes mencionados, pero no sobresalen como factores importantes desde el punto de vista de la metodología propuesta. Los factores que

obtienen un buen rendimiento de clasificación son; la concentración de glucosa, el colesterol, así como la concentración de neutrofilo, el algoritmos *IBK* obtiene un rendimiento de sensibilidad, especificidad, curva ROC y exactitud de 0.84, 0.93, 0.861 y 0.9.

Bajo peso

Para la comorbilidad de bajo peso neonatal, existen factores de riesgo que se mencionan en la literatura médica, por ejemplo Zaragoza et al. (2017) en su trabajo "Factores de riesgo asociados a la morbi-mortalidad perinatal en mujeres con diabetes gestacional del Sur de México", concluyeron que los factores de riesgo para que un neonato presente bajo peso tienen que ver principalmente con Estado Hipertensivos en el embarazo (EHE) o Preeclampsia.

En el presente trabajo de investigación multivariable los factores asociados a bajo peso neonatal son; AF-DM2, AF-HTA, parentesco familiar con Hta, la exposición a contaminantes como el humo o polvo por parte de la gestante, el peso de la madre, así como la Tensión arterial sistólica (TAS), se observa que existe una coincidencia entre los factores de riesgo documentados.

El algoritmo *RandomForest* obtiene utilizando los factores antes descritos en este trabajo el siguiente rendimiento en sensibilidad, especificidad, curva ROC y exactitud; 0.719, 0.776, 0.789 y 0.75 respectivamente.

Diabetes gestacional

La diabetes gestacional es un tipo especial de diabetes, se presenta o se diagnostica por primera vez durante el periodo de gestación, de ahí su nombre, en la literatura ha sido asociada con factores de riesgo como el sobre peso y la obesidad materna como factores detonantes.

Medina-Pérez et al., (2017), llevaron a cabo un trabajo de investigación en el cual concluyen que los factores de riesgo que predisponen a desarrollar DG son; sobrepeso, multiparidad, antecedente de óbito, producto con malformaciones genéticas, antecedentes de intolerancia a la glucosa, ganancia de peso materno > 20

kg durante el embarazo, antecedentes de afectaciones obstétricas graves, obesidad, grupo étnicos o raza, edad materna, antecedentes de diabetes mellitus (línea directa), glucosuria, macrosomía previa, diabetes gestacional y abortos previos.

Zaragoza et al., (2017), en su trabajo de investigación menciona que el sobre peso y la obesidad son factores importantes para desarrollar diabetes gestacional.

En el presente trabajo de investigación se encontró que la edad materna, los antecedentes de Hta, el parentesco para la Hta., la exposición a contaminantes como el humo/polvo y el conteo de plaquetas, son factores que obtienen un alto rendimiento de sensibilidad, especificidad curva ROC y exactitud cuando se usan para clasificar entre la presencia o ausencia de la esta condición utilizando algoritmos de aprendizaje automático.

El algoritmo *ClassificationViaRegression* es el algoritmo que muestra el mejor rendimiento aplicando técnicas de selección de características, 0.99 en sensibilidad, lo cual sugiere que con el entrenamiento y la configuración antes mencionada es capaz de distinguir de manera correcta entre las dos clases, presencia o ausencia de la comorbilidad. La especificidad que obtienen es de 0.965, la curva ROC es de 0.99 y obtiene una exactitud de 98 % respectivamente.

Existen dos algoritmos mas que obtiene resultados muy similares, inclusive con solo tres factores de riesgo, lo anterior sugiere que éstos son importantes cuando se trata esta comorbilidad. El algoritmo *PART* relaciona a la edad materna, el parentesco familiar con Hta., y la constante exposición a contaminantes como el humo/polvo con los cuales obtiene un rendimiento de sensibilidad, especificidad, curva ROC y exactitud de 0.99, 0.965, 0.97 y 0.98.

El algoritmo *J48* relaciona los mismos tres factores de riesgo que los dos algoritmos antes mencionados, obteniendo un rendimiento de sensibilidad de 0.99, especificidad de 0.965, curva ROC de 0.97 y una exactitud de 0.98 respectivamente.

CAPÍTULO V. CONCLUSIONES

Los algoritmos de aprendizaje automático podrían ser aplicados en la predicción de comorbilidades perinatales en embarazos de alto riesgo, utilizando características sociodemográficas, clínico-antropométricas, así como obstétricas. En este trabajo de investigación se abordan 4 comorbilidades que podrían presentarse durante un periodo de gestación, si esto sucede complica el mismo periodo o la resolución de dicho periodo.

El modelado de datos que se aplicó en este trabajo de investigación utilizando la metodología propuesta y desarrollada sugiere que puede generarse un modelo predictivo a través del previo entrenamiento o aprendizaje de algoritmos de aprendizaje automático.

Es indispensable realizar un trabajo de investigación longitudinal con el propósito de darle seguimiento y observar el comportamiento y evolución de embarazo, tomar registros de datos y mapear las variables que en este trabajo se utilizan.

Posteriormente con la información obtenida generar un modelo predictivo mucho más exacto y confiable que pueda ser implementado en la predicción temprana de las comorbilidades abordadas en este trabajo de investigación.

REFERENCIAS BIBLIOGRÁFICAS

1.- Esparza-Valencia, D. M., Toro-Ortiz, J. C., Herrera-Ortega, O., & Fernández-Lara, J. A. (2018). Prevalencia de morbilidad materna extrema en un hospital de segundo nivel de San Luis Potosí, México. Ginecología y obstetricia de México, 86(05), 304-312.

2.- Royert, J. M., & Peñate, M. P. (2016). Caracterización de las gestantes de alto riesgo obstétrico (ARO) en el departamento de Sucre (Colombia), 2015. Salud Uninorte, 32(3), 452-460.

3.- Akhtar, F., Li, J., Guan, Y., Imran, A., & Azeem, M. (2018, July). Monitoring Bio-Chemical Indicators Using Machine Learning Techniques for an Effective Large for Gestational Age Prediction Model with Reduced Computational Overhead. In International Conference on Frontier Computing (pp. 130-137). Springer, Singapore.

4.- Kumru, P., Arisoy, R., Erdogdu, E., Demirci, O., Kavrut, M., Ardıc, C., ... & Ertekin, A. (2016). Prediction of gestational diabetes mellitus at first trimester in low-risk pregnancies. Taiwanese Journal of Obstetrics and Gynecology, 55(6), 815-820.

5.- Romo-Romo, A., Almeda-Valdés, P., Brito-Córdova, G. X., & Gómez-Pérez, F. J. (2017). Prevalencia del consumo de edulcorantes no nutritivos (ENN) en una población de pacientes con diabetes en México. Gac Med Mex, 153(2), 61-74.

6.- Hamann (2017). Obesity Update 2017. Diabetologe, 13(5), 331-341. https://doi.org/10.1007/s11428-017-0241-71. Use training set. The classifier is evaluated on how well it predicts the class of the instances it was trained on.

7.- Bjorn, N., Benvinda, M., Neves de Jesus, S., & Casado Morales, M. (2013). Estrategias de relajación durante el período de gestación: beneficios para la salud. Clínica y Salud, 24(2), 77-83.

8.- Nava, P., Garduño, A., Pestaña, S., Santamaría, M., Vázquez, D. A., Camacho, R., & Herrera, J. (2011). Obesidad pregestacional y riesgo de intolerancia a la glucosa en el embarazo y diabetes gestacional. Revista chilena de obstetricia y ginecología, 76(1), 10-14.

9.- OMS (2016). Informe Mundial de la diabetes. Resumen de Orientación, 4 https://doi.org/10.18004/rvspmi/2312-3893/2016.03(02)71-076

10.- Andrade, W. F. C., & Carriel, S. H. (2018). Test de Sullivan. Eficacia en la detección de diabetes gestacional, en mujeres de 18 a 35 años Cantón Buena Fe, Los Ríos. Ecuadorian Science Journal, 2(1), 8-11.

11.- Zaragoza García O., Polanco García J. C., Gutiérrez Pérez I. A., Ramírez M., Parra Rojas , Guzmán Guzmán Iris Paola (2017). Factores de riesgo asociados a la morbi-mortalidad perinatal en mujeres con diabetes gestacional del Sur de México. Chilpancingo de los Bravo.

12.- Trujillo, J. (2016). Critérios diagnósticos y efectividad de intervenciones para el manejo de diabetes gestacional. Revista Cuidarte, 7(2), 1251-1254.

13.- Marrero, J. M. G., Roca, T. Z. O., Almira, R. C., Bauzá, E. B., & Rodríguez, T. G. (2013). Caracterización de la enfermedad hipertensiva gestacional en pacientes de la Policlínica Máximo Gómez Báez. Correo Científico Médico, 17(2).

14.- Huarte, M., Modroño, A., & Larrañaga, C. (2009). Conducta ante los estados hipertensivos del embarazo. In Anales del Sistema Sanitario de Navarra (Vol. 32, pp. 91-103). Gobierno de Navarra. Departamento de Salud.

15.- Chaparro, L. V. B., Benavides, P., Rios, J. A. L., & Herrera, W. O. (2014). Estados hipertensivos en el embarazo: revisión. Revista UDCA Actualidad & Divulgación Científica, 17(2).

16.- Detección, Diagnóstico y Tratamiento de Enfermedades Hipertensivas del Embarazo Guia de Evidencias y Recomendaciones: Guia de Practica Clinica. Mexico, IMSS; 2017.

17.- Barrera-Cruz, A., Mancilla- García, M. ., Román- Maeda, S. ., Rodríguez- Loreto, E., & Villaláz- Ureña, A. (2013). Guía de práctica clínica: Intervenciones de Enfermería en la paciente con preeclampsia. Rev. Enferm. Inst Mex Seguro Soc, 21(2), 91–104. Retrieved from http://www.medigraphic.com/pdfs/enfermeriaimss/eim-2013/eim132f.pdf

18.- Camacho Terceros, L. A., Rodríguez, B., & Carmen, M. (2015). Una mirada clínica al diagnóstico de preeclampsia. Revista Científica Ciencia Médica, 18(1), 50-55.

19- Valdivia Briceño, C. A. (2018). Factores de riesgo perinatales asociados a morbimortalidad perinatal en hijo nacido de madre con preeclampsia severa, síndrome de hellp y eclampsia en el Hospital Santa Rosa durante el año 2016.

20.- Rodríguez-Valenzuela, C. (2017). Actualidades en el manejo de la preeclampsia. Revista Mexicana de Anestesiología, 40(S1), 14-15.

21.- Nuñez Paredes, E. J. (2018). Desprendimiento prematuro de placenta, por pre eclampsia severa caso clínico-Hospital Aplao agosto 2016.

22.- Moreno Perez, J. A. (2016). El indice de masa corporal pregestacional incrementado en nuliparas como factor de riesgo para obito fetal en el Hospital Belen de Trujillo.

23.- Allpas, D., & de María, F. (2018). Desprendimiento prematuro de placenta-obito fetal revisión de caso clínico en un Hospital del Callao Abril 2016.

24.- Huerta Jiménez, O., Pérez Silva, S., De Jesús García, A., Jiménez Báez, M. V., & Sandoval Jurado, L. (2018). Factors related to fetal death in a hospital of second level of care in Cancún, Quintana Roo. Revista CONAMED, 22(1), 5-10.

25.- Contreras (2017) Factores Asociados al Síndrome de Dificultad Respiratoria Neonatal en el Hospital Regional de Ayacucho, periodo enero a Diciembre 2016, UNIVERSIDAD NACIONAL DEL ALTIPLANO, Puno – Perú.

26.- Quiroga, A. (2014). Cuidados al recién nacido con síndrome de dificultad respiratoria. Plan de cuidados de enfermería. Enfermeria Neonatal, 4.

27.- Poma Chuquija, Y. (2018). Comparación entre el uso de CPAP y surfactante pulmonar en el manejo de la dificultad respiratoria en recién nacidos pre termino en el Hospital Regional Miguel Ángel Mariscal Llerena de Ayacucho julio 2016 a junio 2017.

28.- Zavala-González, M. A., Reyes-Díaz, G. K., Posada-Arévalo, S. E., & Jiménez-Balderas, E. A. (2009). Índice de masa corporal en la definición de macrosomía fetal en Cárdenas, Tabasco, México. Salud en tabasco, 15(1), 828-838.

29.- Gonzáles-Tipiana, I. R. (2017). LA MACROSOMIA FETAL: PREVALENCIA, FACTORES DE RIESGO ASOCIADOS Y COMPLICACIONES EN EL HOSPITAL REGIONAL DE ICA, PERU. Revista Médica Panacea, 2(2).

30.- Ramas (2013). UNIVERSIDAD PRIVADA ANTENOR ORREGO INTERCON 2013.

31.- García-de la Torre, J. I., Rodríguez-Valdéz, A., & Delgado-Rosas, A. (2017). Risk factors for fetal macrosomia in patients without gestational diabetes mellitus. Ginecologia y obstetricia de Mexico, 84(03), 164-171.

32.- Reyes, R. Á., Pen, M. H., Cerda, C. I. S., & Ramírez, R. I. C. (2013). Factores de riesgo del recién nacido macrosómico. Pediatría de México, 15(1), 6-11.

33.- Daza, V., Jurado, W., Duarte, D., Gich, I., Sierra-Torres, C. H., & Delgado-Noguera, M. (2009). Bajo peso al nacer: exploración de algunos factores de riesgo en el Hospital Universitario San José en Popayán (Colombia). Revista Colombiana de Obstetricia y Ginecología, 60(2), 124-134.

34.- Román, F., & Kelly, J. (2018). Embarazo adolescente y controles prenatales insuficientes como factores de riesgo para bajo peso al nacer en el Hospital San José de enero a diciembre del 2016.

35.- Fumero, R. A., Cobas, L. R., & Santiago, M. A. (2001). Repercusión de los factores de riesgo en el bajo peso al nacer. Resumed, 14(3), 115-21.

36.- Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. Córdoba: Universidad Tecnológica Nacional, Facultad Regional Córdoba.

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6).

Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22).

37.- Martínez, A. V. (2003). Minería de Datos. Una Introducción. Conciencia Tecnológica, (23).

38.- Vallejo Ballesteros, H. F., Guevara Iñiguez, E., & Medina Velasco, S. R. (2018). Minería de Datos. RECIMUNDO, 2(1 (Esp), 339-349. Recuperado a partir de http://www.recimundo.com/index.php/es/article/view/182

39.- Hernández (2017). Modelado de la Capacidad Funcional Articular de la Mano Usando Algoritmos de Inteligencia Artificial en Pacientes con Artritis Reumatoide, UNIVERSIDAD AUTÓNOMA DE GUERRERO.

40.- Corso, C. L., García, A., Ciceri, L., & Romero, F. (2014). Minería de Datos aplicada a la Detección de factores para la prevención de incidentes informáticos. In XVI Workshop de Investigadores en Ciencias de la Computación.

41.- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

42.- Díaz-Barrios, H., Alemán-Rivas, Y., Cabrera-Hernández, L., Morales-Hernández, A., Chávez-Cárdenas, M. D. C., & Casas-Cardoso, G. M. (2015). Algoritmos de aprendizaje automático para clasificación de Splice Sites en secuencias genómicas. Revista Cubana de Ciencias Informáticas, 9(4), 155-170.

43.- Roman (2011). Minería de Datos en Encuestas de Profesores al fin de Semestre de la Facultad de Ingeniería, UNAM.

44.- Cleary, J. G., & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. In Machine Learning Proceedings 1995 (pp. 108-114). Morgan Kaufmann.

45.- Mahmood, D. Y., & Hussein, M. A. (2013). Intrusion detection system based on K-star classifier and feature set reduction. Int. Organ. Sci. Res. J. Comput. Eng. Vol, 15, 107-112.

46.- Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2010). WEKA Manual for Version 3-7-2. Interface.

47.- Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), 256-261.

48.- González, O. L., Trinidad, J. F. M., Borroto, M. G., Erro, L. E., & Tonantzintla, S. M. (2014). Clasificadores Supervisados basados en Patrones Emer-gentes para Bases de Datos con Clases Desbalanceadas. Reporte Técnico No. CCC-14–004, Coordinación de Ciencias Computacionales, INAOE.

49.- López et al., (1998). Curvas ROC. In Cad Aten Primaria (Vol. 5, pp. 229-235).

Rocío, A. (2017). Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones, 12-16. Retrieved from https://idus.us.es/xmlui/bitstream/handle/11441/63201/Valle Benavides Ana Rocío del TFG.pdf?sequence=1&isAllowed=y

50.- Valle (2017). Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones, UNIVERSIDAD DE SEVILLA, España.

51.- Cerda, J., & Cifuentes, L. (2012). Using ROC curves in clinical investigation: theoretical and practical issues. Revista chilena de infectologia: organo oficial de la Sociedad Chilena de Infectologia, 29(2), 138-141.

52.- Tomás (2012). El gran libro de android. Barcelona España: Alfaomega.

53.- Antona Cortés, C. (2017). Herramientas modernas en redes neuronales: la librería Keras (Bachelor's thesis).

54.- Saji, S. A., & Balachandran, K. (2015, March). Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction. In 2015 International Conference on Advances in Computer Engineering and Applications (pp. 201-206). IEEE.

55.- Ruiz, C. A., & Matich, D. J. (2001). Matich-Redesneuronales.

57.- Andrade, W. F. C., & Carriel, S. H. (2018). Test de Sullivan. Eficacia en la detección de diabetes gestacional, en mujeres de 18 a 35 años Cantón Buena Fe, Los Ríos. Ecuadorian Science Journal, 2(1), 8-11.

58.- Valdés Ramos, E., & Bencosme Rodríguez, N. (2015). Frecuencia de obesidad y su relación con algunas complicaciones maternas y perinatales en una comunidad indígena. Revista Cubana de Endocrinología, 26(3), 0-0.

59.- García, R., & Arnold, J. (2018). Factores asociados a mortalidad en recién nacidos prematuros con enfermedad de membrana hialina en el Hospital Nacional Sergio E. Bernales, mayo 2015–mayo 2017.

60.- Delgado-Becerra, A., Casillas-García, D. M., & Fernández-Carrocera, L. A. (2011). Morbilidad del hijo de madre con diabetes gestacional, en el Instituto Nacional de Perinatología Isidro Espinosa de los Reyes. Perinatología y Reproducción Humana, 25(3), 139-145.

61.- Moreira, M. W., Rodrigues, J. J., Kumar, N., Niu, J., & Sangaiah, A. K. (2017, July). Multilayer Perceptron Application for Diabetes Mellitus Prediction in Pregnancy Care. In International Conference on Frontier Computing (pp. 200-209). Springer, Singapore.

62.- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

63.- Flores-Padilla, L., Solorio-Páez, I. C., Melo-Rey, M. L., & Trejo-Franco, J. (2014). Embarazo y obesidad: riesgo para desarrollo de diabetes gestacional en la frontera norte de México. Gaceta Médica de México, 150(s1), 73-78.

64.- Detección, Diagnóstico y Tratamiento de Enfermedades Hipertensivas del Embarazo. Guía de Rerencia Rápida: Guía de Práctica Clínica. México, CENETEC; Disponible en: http://www.cenetec.salud.gob.mx/contenidos/gpc/catalogoMaestroGPC.htm

65.- López B., I., Álvarez Vega, A. R., Alonso Uría, R. M., Campo González, A., Díaz Aguilar, R., & Amador Morán, R. (2012). Factores de riesgo para complicaciones del recién nacido grande para su edad gestacional. Investigación y Educación En Enfermería, 30(1), 95-100. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=3928357&info=resumen&idioma=SPA

66.- Delgado-Becerra, A., Casillas-García, D. M., & Fernández-Carrocera, L. A. (2011). Morbilidad del hijo de madre con diabetes gestacional, en el Instituto Nacional de Perinatología Isidro Espinosa de los Reyes. Perinatología y Reproducción Humana, 25(3), 139-145.

67.- Rathore, S. S., & Gupta, A. (2014, February). A comparative study of feature-ranking and feature-subset selection techniques for improved fault prediction. In Proceedings of the 7th India Software Engineering Conference (p. 7). ACM.

- 68.- Duque, R. G. (2017). Curso: Python paraTodos. Retrieved from http://curso-python.eugeniabahit.com/sources.tar.gz%0ACurso%0Ahttp://www.iaa.es/python/curso-python-para-principiantes.pdf

- 69.- Guan, C., Qin, S., Ling, W., & Ding, G. (2016). Apparel recommendation system evolution: an empirical review. International Journal of Clothing Science and Technology, 28(6), 854–879. https://doi.org/10.1108/IJCST-09-2015-0100

- 70.- Ingenier, D. E. (2011). Universidad nacional autónoma de méxico.

- 71.- Rathore, S. S., & Gupta, A. (2014). A comparative study of feature-ranking and feature-subset selection techniques for improved fault prediction. Proceedings of the 7th India Software Engineering Conference on - ISEC ’14, 1–10. https://doi.org/10.1145/2590748.2590755

Lista de figuras

Lista de tablas

ANEXOS

Rendimiento de los 35 algoritmos utilizando 20 variables, así como 58 instancias positivas y 70 negativas sin selección de características para la comorbilidad de bajo peso.

Rendimiento de los algoritmos utilizando 20 variables para la comorbilidad de bajo peso sin selección de características

Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
RandomCommitte	0.655	0.6	0.647	0.63
RandomForest	0.603	0.757	0.707	0.69
RandomizableFilteredClassifier	0.586	0.786	0.69	0.70
RandomTree	0.586	0.643	0.615	0.62

LogitBoost	0.569	0.686	0.666	0.63
Bagging	0.534	0.871	0.76	0.72
NaiveBayes	0.517	0.929	0.735	0.74
NaiveBayesUpdateable	0.517	0.929	0.735	0.74
MultilayerPerceptron	0.517	0.543	0.599	0.53
HoeffdingTree	0.517	0.886	0.714	0.72
Logistic	0.483	0.729	0.609	0.62
ClassificationViaRegression	0.483	0.8	0.654	0.66
MultiClassClassifier	0.483	0.729	0.609	0.62
J Rip	0.483	0.743	0.613	0.63
IBK	0.466	0.714	0.619	0.60
PART	0.466	0.557	0.484	0.52
SDG	0.448	0.814	0.631	0.65
SimpleLogistic	0.448	0.743	0.607	0.61
MultiClassClassifierUpdateable	0.448	0.814	0.631	0.65
LMT	0.448	0.743	0.607	0.61
SMO	0.431	0.829	0.63	0.65
RandomSubSpace	0.431	0.871	0.693	0.67
REPTree	0.431	0.757	0.591	0.61
AdaBoostM1	0.414	0.814	0.614	0.63
IterativeClassifierOptimizer	0.414	0.786	0.595	0.62
OneR	0.397	0.614	0.505	0.52
VotedPerceptron	0.379	0.929	0.715	0.68
DecisionTable	0.379	0.886	0.644	0.66
J48	0.379	0.714	0.595	0.56
Kstar	0.345	0.686	0.543	0.53
AttributSelectedClassifier	0.345	0.886	0.642	0.64
BayesNet	0.328	0.929	0.619	0.66
FilteredClassifier	0.328	0.929	0.618	0.66
LWL	0.31	0.971	0.656	0.67
DecisionStump	0.293	0.986	0.6	0.67

Rendimiento de los algoritmos utilizando 20 variables, así como 58 instancias positivas y 70 negativas con selección de características, para bajo peso

Rendimiento de los algoritmos utilizando 20 variables para la comorbilidad de bajo peso

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
IBK (4)	antFamHta, parenFamHta, cotaminacion, colesterol	0.749	0.72	0.776	0.671
RandomCommitte (1)	colesterol	0.686	0.70	0.776	0.643
RandomTree (1)	colesterol	0.686	0.70	0.776	0.643
RandomizableFilteredClassifier (1)	colesterol	0.668	0.71	0.759	0.671
RandomForest (4)	antFamHta, parenFamHta, contaminacion, colesterol	0.719	0.71	0.759	0.671
IterativeClassifierOptimizer (5)	contaminacion, TAS, totPartos, glucosa, colesterol	0.689	0.73	0.621	0.814
LogitBoost (3)	TAS, glucosa, colesterol	0.711	0.72	0.621	0.8
ClassificationViaRegression (4)	contaminacion, TAS, anemia, colesterol	0.726	0.76	0.603	0.886
MultilayerPerceptron (5)	contaminacion, TAS, plaquetas, colesterol	0.759	0.68	0.586	0.757
SimpleLogistic (4)	contaminacion, TAS, glucosa, colesterol	0.719	0.73	0.586	0.843
LMT (3)	contaminacion, TAS, colesterol	0.72	0.73	0.586	0.857

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
NaiveBayes (4)	contaminacion, TAS, totAbortos, colesterol	0.761	0.77	0.569	0.929
NaiveBayesUpdateable (4)	contaminacion, TAS, totAbortos, colesterol	0.761	0.77	0.569	0.929
Bagging (4)	peso, TAS, glucosa, colesterol	0.747	0.70	0.569	0.814
OneR (1)	totGestas	0.612	0.62	0.552	0.671
HoeffdingTree (5)	antFamHta, TAS, anemia, plaquetas, colesterol	0.69	0.75	0.552	0.914
Kstar (5)	contaminacion, TAS, totCesareas, anemia, colesterol	0.703	0.71	0.534	0.857
J48 (3)	TAS, totAbortos, colesterol	0.711	0.73	0.517	0.9
Logistic (3)	antFamDm2, TAD, colesterol	0.685	0.70	0.5	0.857
SDG (4)	antFamDm2, contaminacion, TAS, colesterol	0.714	0.73	0.5	0.929
MultiClassClassifier (3)	antFamDm2, TAD, colesterol	0.685	0.70	0.5	0.857
MultiClassClassifierUpdateable (4)	antFamDm2, contaminacion, TAS, colesterol	0.714	0.73	0.5	0.929
SMO (6)	antFamDm2, contaminacion, TAS, totCesareas, totAbortos, colesterol	0.668	0.69	0.466	0.871

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
RandomSubSpace (4)	antFamDm2, TAS, glucosa, colesterol	0.737	0.70	0.466	0.886
LWL (4)	edad, TAS, glucosa, colesterol	0.672	0.70	0.448	0.914
AdaBoostM1 (5)	parenFamDm2, TAS, plaquetas, glucosa, colesterol	0.645	0.66	0.448	0.829
Jrip (4)	escolaridad, parenFamHta, TAS, colesterol	0.641	0.64	0.448	0.8
PART (3)	talla, TAS, colesterol	0.671	0.70	0.448	0.914
REPTree (3)	contaminacion, TAS, colesterol	0.609	0.64	0.431	0.814
AttributSelectedClassifier (3)	TAS, glucosa, colesterol	0.66	0.66	0.379	0.9
DecisionTable (2)	TAS, colesterol	0.665	0.67	0.379	0.914
BayesNet (2)	TAS, colesterol	0.633	0.66	0.328	0.943
VotedPerceptron (3)	TAS, totGestas, colesterol	0.686	0.66	0.328	0.929
FilteredClassifier (2)	TAS, colesterol	0.633	0.66	0.328	0.943
DecisionStump (1)	colesterol	0.6	0.67	0.293	0.986

Rendimiento de los 35 algoritmos con 20 variables, 58 instancias positivas y 86 negativas sin selección de características, para la comorbilidad de bajo peso.

Rendimiento de los 35 algoritmos para la comorbilidad de bajo peso sin selección de características

Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
RandomCommitte	0.614	0.737	0.73	0.68
Bagging	0.596	0.776	0.753	0.70
RandomForest	0.596	0.803	0.737	0.71
LogitBoost	0.561	0.789	0.627	0.69
J48	0.561	0.763	0.621	0.68
RandomTree	0.561	0.671	0.616	0.62
PART	0.544	0.763	0.643	0.67
J Rip	0.526	0.737	0.624	0.65
SMO	0.474	0.776	0.625	0.65
ClassificationViaRegression	0.456	0.803	0.65	0.65
RandomizableFilteredClassifier	0.456	0.658	0.568	0.57
IBK	0.439	0.697	0.559	0.59
Logistic	0.421	0.75	0.609	0.61
MultilayerPerceptron	0.421	0.658	0.546	0.56
SDG	0.421	0.711	0.566	0.59
IterativeClassifierOptimizer	0.421	0.895	0.673	0.69
MultiClassClassifier	0.421	0.75	0.609	0.61
MultiClassClassifierUpdateable	0.421	0.711	0.566	0.59
LMT	0.404	0.789	0.605	0.62
SimpleLogistic	0.386	0.789	0.581	0.62
Kstar	0.386	0.697	0.552	0.56
REPTree	0.386	0.842	0.611	0.65
OneR	0.368	0.658	0.513	0.53
NaiveBayes	0.351	0.895	0.682	0.66
NaiveBayesUpdateable	0.351	0.895	0.682	0.66
AdaBoostM1	0.351	0.882	0.655	0.65
AttributSelectedClassifier	0.351	0.908	0.628	0.67
RandomSubSpace	0.351	0.882	0.68	0.65
HoeffdingTree	0.351	0.855	0.657	0.64
DecisionTable	0.316	0.882	0.597	0.64
LWL	0.263	0.987	0.7	0.68
BayesNet	0.246	0.987	0.595	0.67
FilteredClassifier	0.246	0.987	0.595	0.67
DecisionStump	0.246	0.987	0.595	0.67
VotedPerceptron	0.193	0.816	0.514	0.55

Rendimiento de los 35 algoritmos utilizando 20 variables, así como 58 instancias positivas y 86 negativas con selección de características.

Rendimiento de los 35 algoritmos para bajo eso con selección de características

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
RandomForest (6)	antFamDm2, antFamHta, parenFamHta, contaminacion, peso, TAS	0.789	75.188	0.719	0.776
IBK (6)	antFamDm2, antFamHta, parenFamHta, contaminacion, peso, TAS	0.742	75.19	0.702	0.789
Bagging (5)	escolaridad, parenFamHta, peso, TAS, totCesareas	0.742	70.68	0.561	0.816
LogitBoost (2)	peso,TAS	0.689	71.43	0.544	0.842
PART (2)	peso, TAS	0.668	72.93	0.526	0.882
AdaBoostM1 (3)	peso, TAS, glucosa	0.664	70.68	0.474	0.882
IterativeClassifierOptimizer (2)	peso, TAS	0.66	72.93	0.474	0.921
Jrip (3)	antFamDm2, peso, TAS	0.659	69.17	0.474	0.855
J48 (3)	peso, TAS, colesterol	0.672	73.68	0.474	0.934
LMT (2)	peso, TAS	0.715	72.18	0.474	0.908
MultilayerPerceptron (4)	antFamDm2,peso, IMCMat, TAS	0.681	71.43	0.456	0.908
REPTree (2)	peso, TAS	0.657	69.92	0.456	0.882
SMO (1)	contaminacion	0.594	61.65	0.439	0.75
NaiveBayes (4)	parenFamDm2, peso, TAS, colesterol	0.714	72.18	0.404	0.961
NaiveBayesUpdateable (4)	parenFamDm2, peso, TAS, colesterol	0.714	72.18	0.404	0.961

RandomSubSpace (2)	peso, TAS	0.624	66.92	0.386	0.882
Clasificador	**Atributos (Factores de riesgo)**	**ROC**	**Exactitud**	**Sensibilidad**	**Especificidad**
SDG (3)	antFamDm2, peso, TAS	0.636	67.67	0.351	0.921
MultiClassClassifierUpdateable (3)	antFamDm2, peso, TAS	0.643	68.42	0.351	0.934
VotedPerceptron (2)	parenFamFm2, contamiacion	0.592	60.15	0.333	0.803
OneR (1)	TAS	0.64	68.42	0.333	0.947
HoeffdingTree (3)	peso, TAS, colesterol	0.718	69.17	0.333	0.961
Logistic (1)	TAS	0.599	66.92	0.316	0.934
MultiClassClassifier (1)	TAS	0.599	66.92	0.316	0.934
RandomCommitte (1)	TAS	0.601	66.17	0.316	0.921
RandomizableFilteredClassifier (1)	TAS	0.607	66.17	0.316	0.921
RandomTree (1)	TAS	0.601	66.17	0.316	0.921
SimpleLogistic (1)	parenFamHta, TAS	0.605	66.92	0.281	0.961
Kstar (3)	antFamDm2, parenFamDm2, TAS	0.601	67.67	0.281	0.974
LWL (2)	TAS, totAbortos	0.602	68.42	0.281	0.987
AttributSelectedClassifier (1)	TAS	0.597	68.42	0.281	0.987
ClassificationViaRegression (1)	TAS	0.605	66.92	0.281	0.961
BayesNet (1)	TAS	0.595	66.92	0.246	0.987
FilteredClassifier (1)	TAS	0.595	66.92	0.246	0.987
DecisionTable (1)	TAS	0.596	66.92	0.246	0.987
DecisionStump (1)	TAS	0.595	66.92	0.246	0.987

Rendimiento de los 35 algoritmos utilizaron 20 variables y 25 instancias positivas, así como 50 instancias negativas sin selección de características.

Rendimiento utilizando 20 variables para la comorbilidad de macrosomía

Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
RandomCommitte	0.68	0.8	0.764	0.76
RandomizableFilteredClassifier	0.68	0.74	0.722	0.72
J48	0.6	0.7	0.673	0.67
LogitBoost	0.52	0.755	0.667	0.67
OneR	0.52	0.8	0.66	0.71
PART	0.52	0.8	0.654	0.71
NaiveBayes	0.48	0.8	0.593	0.69
NaiveBayesUpdateable	0.48	0.8	0.593	0.69
RandomForest	0.48	0.9	0.73	0.76
RandomTree	0.48	0.8	0.64	0.69
MultilayerPerceptron	0.44	0.7	0.507	0.61
Logistic	0.4	0.7	0.485	0.60
IBK	0.4	0.68	0.568	0.59
Bagging	0.4	0.9	0.644	0.73
IterativeClassifierOptimizer	0.4	0.8	0.663	0.67
MultiClassClassifier	0.4	0.7	0.485	0.60
AdaBoostM1	0.36	0.84	0.686	0.68
ClassificationViaRegression	0.36	0.88	0.69	0.71
SDG	0.32	0.88	0.6	0.69
Kstar	0.32	0.76	0.601	0.61
MultiClassClassifierUpdateable	0.32	0.88	0.6	0.69
J Rip	0.32	0.84	0.598	0.67
SMO	0.24	0.92	0.58	0.69
LWL	0.24	0.8	0.42	0.61
RandomSubSpace	0.24	0.96	0.677	0.72
DecisionStump	0.24	1	0.522	0.75
LMT	0.24	0.84	0.534	0.64
SimpleLogistic	0.2	0.84	0.509	0.63
DecisionTable	0.16	0.92	0.464	0.67
AttributSelectedClassifier	0.12	0.92	0.474	0.65
REPTree	0.12	0.84	0.5	0.60
VotedPerceptron	0.08	1	0.496	0.69
HoeffdingTree	0.08	0.96	0.508	0.67
BayesNet	0	1	454	3.12
FilteredClassifier	0	1	0.454	0.67

Rendimiento de los 35 algoritmos utilizando 20 variables y 25 instancias positivas, así como 50 instancias negativas con selección de características para la comorbilidad de macrosomía

Rendimiento de los algoritmos con 20 variables para macrosomía con selección de características para macrosomía

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
IBK (5)	antFamDm2, antFamHta, IMCMat, totCesareas, colesterol	0.857	86.67	0.76	0.92
RandomForest (2)	contaminacion, colesterol	0.817	84.00	0.72	0.9
RandomCommitte (2)	contamiacion, colesterol	0.812	82.67	0.72	0.88
RandomTree (2)	contaminacion, colesterol	0.812	82.67	0.72	0.88
OneR (2)	plaquetas, colesterol	0.77	78.67	0.72	0.82
RandomizableFilteredClassifier (5)	antFamHta, parenFamHta, IMCMat, totAbortos, colesterol	0.796	78.67	0.68	0.84
LWL (3)	totGestas, anemia, colesterol	0.662	78.67	0.56	0.9
Kstar (3)	edad, parenFamDm2, colesterol	0.712	80.00	0.52	0.94
J48 (5)	talla, IMCMat, anemia, plaquetas, colesterol	0.741	80.00	0.48	0.96
LogitBoost (2)	plaquetas, colesterol	0.754	78.67	0.48	0.94
IterativeClassifierOptimizer (2)	plaquetas, colesterol	0.713	77.33	0.44	0.94
AdaBoostM1 (2)	TAD, colesterol	0.689	76.00	0.36	0.96
Bagging (3)	antFamDm2, totAbortos, colesterol	0.744	70.67	0.36	0.88
RandomSubSpace (2)	talla, colesterol	0.75	76.00	0.32	0.98
NaiveBayes (3)	escolaridad, contaminacion, colesterol	0.648	76.00	0.32	0.98
NaiveBayesUpdateable (3)	escolaridad, contaminacion, colesterol	0.648	76.00	0.32	0.98
SDG (6)	contaminacion, talla, IMCMat, totCesareas, plaquetas, colesterol	0.63	73.33	0.32	0.94

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
MultiClassClassifierUpdateable (6)	contaminacion, talla, IMCMat, totCesareas, plaquetas, colesterol	0.63	73.33	0.32	0.94
Jrip (2)	talla, anemia	0.606	73.33	0.32	0.94
HoeffdingTree (1)	colesterol	0.63	72.00	0.28	0.94
MultilayerPerceptron (1)	colesterol	0.552	74.67	0.24	1
SimpleLogistic (1)	colesterol	0.535	74.67	0.24	1
Logistic (1)	colesterol	0.532	74.67	0.24	1
MultiClassClassifier (1)	colesterol	0.532	74.67	0.24	1
LMT (3)	contaminacion, TAS, totCesareas, colesterol	0.524	74.67	0.24	1
DecisionStump (1)	colesterol	0.522	74.67	0.24	1
PART (3)	peso, IMCMat, colesterol	0.614	73.33	0.24	0.98
AttributSelectedClassifier (1)	colesterol	0.53	70.67	0.2	0.96
REPTree (2)	antFamDm2, colesterol	0.564	69.33	0.2	0.94
DecisionTable (1)	colesterol	0.526	72.00	0.16	1
ClassificationViaRegression (3)	edad, anemia, colesterol	0.464	72.00	0.16	1
SMO (1)	colesterol	0.5	66.67	0	1
FilteredClassifier (1)	colesterol	0.454	66.67	0	1

Rendimiento de los 35 algoritmos con 20 variables con un total de 52 instancias positivas y 86 negativas con selección de atributos, para la comorbilidad de preeclampsia, ver tabla 4.6.

Rendimiento utilizando 20 variables para la comorbilidad de Preeclampsia

Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
RandomForest	0.865	0.965	0.917	0.93
Bagging	0.846	0.965	0.907	0.92
IterativeClassifierOptimizer	0.846	0.953	0.902	0.91
RandomCommitte	0.846	0.895	0.908	0.88
OneR	0.846	0.965	0.906	0.92
AttributSelectedClassifier	0.827	0.965	0.865	0.91
FilteredClassifier	0.827	0.965	0.845	0.91
LogitBoost	0.827	0.965	0.925	0.91
RandomTree	0.827	0.779	0.803	0.80
REPTree	0.827	0.965	0.894	0.91
BayesNet	0.808	0.907	0.862	0.87
SimpleLogistic	0.808	0.942	0.922	0.89
AdaBoostM1	0.808	0.953	0.931	0.90
DecisionTable	0.808	0.942	0.875	0.89
J Rip	0.808	0.953	0.855	0.90
DecisionStump	0.808	0.965	0.852	0.91
LMT	0.808	0.942	0.922	0.89
NaiveBayes	0.788	0.919	0.911	0.87
NaiveBayesUpdateable	0.788	0.919	0.911	0.87
Logistic	0.788	0.907	0.885	0.86
SDG	0.788	0.907	0.848	0.86
LWL	0.788	0.965	0.872	0.90
MultiClassClassifier	0.788	0.907	0.885	0.86
MultiClassClassifierUpdateable	0.788	0.907	0.848	0.86
J48	0.788	0.919	0.858	0.87
SMO	0.769	0.942	0.856	0.88
PART	0.769	0.895	0.841	0.85
ClassificationViaRegression	0.75	0.988	0.933	0.90
RandomSubSpace	0.731	0.965	0.933	0.88
MultilayerPerceptron	0.712	0.907	0.891	0.83
HoeffdingTree	0.712	0.919	0.895	0.84
Kstar	0.637	0.907	0.845	0.82
RandomizableFilteredClassifier	0.635	0.895	0.773	0.80
IBK	0.442	0.837	0.646	0.69
VotedPerceptron	0.385	0.884	0.733	0.70

Rendimiento de los 35 algoritmos con las siguientes características; se utilizaron 20 variables, 52 instancias positivas y 86 negativas con selección de atributos. Para esta prueba se utilizaron las variables TAS y TAD, anteriores pruebas se notó que son variables determinantes en la clasificación. Como se puede ver, utilizando tan solo 20 variables y aplicando selección de características se obtienen excelentes resultados, el algoritmo IBK utiliza la variable TAS y el total de cesáreas para clasificar correctamente, obtiene un valor de sensibilidad de 0.885, un resultado muy aceptable, ver tabla 4.7.

Rendimiento utilizando 20 variables para la comorbilidad de Preeclampsia.

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
IBK (2)	TAS, totCesareas	0.91	0.92	0.885	0.942
RandomCommitte (2)	TAS, totCesareas	0.92	0.92	0.885	0.942
RandomizableFilteredClassifier (2)	TAS, totCesareas	0.93	0.92	0.885	0.942
RandomTree (2)	TAS, totCesareas	0.92	0.92	0.885	0.942
NaiveBayes (2)	TAS, antFamHta	0.93	0.91	0.846	0.953
NaiveBayesUpdateable (2)	antFamHta, TAS	0.93	0.91	0.846	0.953
Logistic (2)	TAS, glucosa	0.94	0.92	0.846	0.965
MultilayerPerceptron (1)	TAS	0.92	0.92	0.846	0.965
SDG (1)	TAS	0.91	0.92	0.846	0.965
SimpleLogistic (1)	TAS	0.93	0.92	0.846	0.965
Kstar (3)	contaminacion, TAS, totGestas	0.94	0.93	0.846	0.977
AdaBoostM1 (3)	edad, TAS, TAD	0.92	0.93	0.846	0.977
Bagging (2)	TAS, TAD	0.92	0.92	0.846	0.965
IterativeClassifierOptimizer (1)	TAS	0.92	0.92	0.846	0.965
LogitBoost (1)	TAS	0.93	0.92	0.846	0.965
MultiClassClassifier (2)	TAS, glucosa	0.94	0.92	0.846	0.965

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
MultiClassClassifierUpdateable (1)	TAS	0.91	0.92	0.846	0.965
OneR (1)	TAS	0.91	0.92	0.846	0.965
HoeffdingTree (1)	TAS	0.92	0.91	0.846	0.953
LMT (1)	TAS	0.93	0.92	0.846	0.965
RandomForest (1)	TAS	0.91	0.91	0.846	0.953
J48 (2)	TAS, TAD	0.86	0.91	0.827	0.965
REPTree (3)	antFamHta, peso, TAS	0.91	0.91	0.827	0.953
BayesNet (1)	TAS	0.85	0.91	0.808	0.965
LWL (4)	escolaridad, antFamHta, contaminacion, TAS	0.88	0.91	0.808	0.977
FilteredClassifier (1)	TAS	0.85	0.91	0.808	0.965
RandomSubSpace (1)	TAS	0.87	0.91	0.808	0.965
DecisionTable (1)	TAS	0.86	0.91	0.808	0.965
Jrip (4)	parenFamDm2, parenFamHta, IMCMat, TAS	0.86	0.91	0.808	0.965
PART (1)	TAS	0.85	0.91	0.808	0.965
DecisionStump (1)	TAS	0.85	0.91	0.808	0.965
AttributSelectedClassifier (1)	TAS	0.85	0.91	0.8	0.965
SMO (3)	TAS, TAD, glucosa	0.89	0.92	0.788	1
ClassificationViaRegression (3)	peso, TAS, TAD	0.95	0.89	0.731	0.988
VotedPerceptron (3)	escolaridad, talla, totGestas	0.64	0.67	0.346	0.872

Rendimiento de los 35 algoritmos utilizando 37 variables, 53 instancias positivas y 67 negativas sin selección de atributos. Las variables TAS y TAD son determinantes, por lo tanto, con esta prueba se busca observar cuáles son los resultados que se obtienen si se omiten dichas variables.

Tabla 4.8 Resultados con 37 variables para la comorbilidad de Preeclampisa

Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
DecisionTable	0.774	0.672	0.701	0.72
FilteredClassifier	0.755	0.657	0.681	0.70
OneR	0.717	0.746	0.732	0.73
LWL	0.698	0.742	0.765	0.72
IterativeClassifierOptimizer	0.698	0.716	0.679	0.71
RandomCommitte	0.679	0.761	0.787	0.73
J Rip	0.679	0.612	0.666	0.64
RandomizableFilteredClassifier	0.66	0.866	0.763	0.78
ClassificationViaRegression	0.66	0.821	0.818	0.75
PART	0.66	0.761	0.707	0.72
MultilayerPerceptron	0.66	0.746	0.705	0.71
J48	0.642	0.701	0.722	0.68
RandomForest	0.623	0.776	0.79	0.71
Logistic	0.623	0.776	0.696	0.71
MultiClassClassifier	0.623	0.776	0.696	0.71
RandomTree	0.604	0.716	0.66	0.67
LMT	0.585	0.746	0.715	0.68
SimpleLogistic	0.566	0.701	0.713	0.64
RandomSubSpace	0.547	0.776	0.747	0.68
AdaBoostM1	0.547	0.776	0.721	0.68
BayesNet	0.547	0.746	0.723	0.66
REPTree	0.547	0.627	0.589	0.59
NaiveBayes	0.528	0.851	0.809	0.71
NaiveBayesUpdateable	0.528	0.851	0.809	0.71
LogitBoost	0.528	0.731	0.718	0.64
AttributSelectedClassifier	0.528	0.657	0.615	0.60
HoeffdingTree	0.509	0.866	0.804	0.71
Bagging	0.509	0.821	0.703	0.68
IBK	0.472	0.866	0.663	0.69
DecisionStump	0.472	0.776	0.629	0.64
SDG	0.472	0.701	0.587	0.60
MultiClassClassifierUpdateable	0.472	0.701	0.587	0.60
SMO	0.453	0.806	0.629	0.65
VotedPerceptron	0.321	0.97	0.664	0.68
Kstar	0.321	0.896	0.628	0.64

Rendimiento de 37 variables y 35 algoritmos con un total de 53 instancias positivas y 67 negativas con selección de atributos.

Rendimiento utilizando 37 variables para la comorbilidad de Preeclampsia

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
RandomForest (7)	escolaridad, cmhbc, rdw, plaquetas, vpm, acidoUri, trigliceridos	0.84	0.82	0.736	0.881
NaiveBayesUpdateable (10)	escolaridad, contaminacion, IMCMat, totGestas, totCesareas, neutrofilo, linfocito, glucosa, urea, acidoUri	0.83	0.81	0.717	0.881
RandomCommitte (9)	antFamHta, contaminacion, totGestas, totPartos, rdw, plaquetas, neutrofilo, acidoUri, colesterol	0.83	0.77	0.736	0.791
LMT (4)	antFamHta, totPartos, glucosa, acidoUri	0.82	0.78	0.792	0.776
MultilayerPerceptron (6)	totGestas, hematocrito, vgm, cmhbc, glucosa, acidoUri	0.82	0.81	0.755	0.851
HoeffdingTree (10)	escolaridad, contaminacion, IMCMat, totGestas, totPartos, plaquetas, neutrofilo, glucosa, urea, acidoUri	0.81	0.82	0.698	0.91
LogitBoost (3)	totPartos, urea, acidoUri	0.8	0.78	0.755	0.791
Kstar (2)	totPartos, acidoUri	0.77	0.77	0.755	0.776
AdaBoostM1 (2)	acidoUri, colesterol	0.77	0.75	0.755	0.746
ClassificationViaRegression (3)	glucosa, creatinina, acidoUri	0.76	0.76	0.792	0.731
Bagging (3)	parenFamDm2, IMCMat, acidoUri	0.75	0.72	0.698	0.731

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
J48 (6)	antFamDm2, contaminacion, totPartos, plaquetas, acidoUri, trigliceridos	0.75	0.78	0.736	0.821
BayesNet (2)	acidoUri, trigliceridos	0.74	0.74	0.811	0.687
IterativeClassifierOptimizer (2)	antFamHta, acidoUri	0.74	0.74	0.811	0.687
FilteredClassifier (2)	acidoUri, trigliceridos	0.74	0.74	0.811	0.687
RandomSubSpace (1)	acidoUri	0.74	0.73	0.736	0.716
OneR (1)	acidoUri	0.73	0.73	0.717	0.746
SMO (6)	totGestas, totCesareas, hematocrito, neutrofilo, linfocito, colesterol	0.73	0.76	0.509	0.955
LWL (3)	totGestas, acidoUri, colesterol	0.73	0.77	0.811	0.731
DecisionTable (2)	acidoUri, trigliceridos	0.73	0.74	0.811	0.687
REPTree (3)	parenFamDm2, rdw, acidoUri	0.72	0.73	0.736	0.731
PART (3)	totPartos, acidoUri, colesterol	0.72	0.78	0.774	0.791
Jrip (2)	antFamHta, acidoUri	0.72	0.74	0.792	0.701
NaiveBayes (2)	acidoUri, trigliceridos	0.72	0.64	0.358	0.866
AttributSelectedClassifier (1)	acidoUri	0.71	0.71	0.736	0.687
SDG (4)	totGestas, vpm, neutrofilo, colesterol	0.71	0.73	0.509	0.91
MultiClassClassifierUpdateable (4)	totGestas, vpm, neutrofilo, colesterol	0.71	0.73	0.509	0.91
RandomTree (5)	antFamHta, totGestas, totPartos, cmhbc, acidoUri, bun	0.7	0.71	0.66	0.746
IBK (2)	antFamHta, totGestas	0.67	0.73	0.528	0.881

RandomizableFilteredClassifier (2)	antFamHta, totGestas	0.67	0.73	0.528	0.881

Clasificador	Atributos (Factores de riesgo)	ROC	Exactitud	Sensibilidad	Especificidad
VotedPerceptron (7)	edad, totGestas, vpm, linfocito, glucosa, bun, colesterol	0.66	0.65	0.226	0.985
DecisionStump (1)	acidoUri	0.66	0.72	0.679	0.746
SimpleLogistic (1)	totGestas	0.57	0.68	0.377	0.925
Logistic (2)	totGestas, basofilo	0.55	0.67	0.377	0.896
MultiClassClassifier (2)	totGestas, basofilo	0.55	0.67	0.377	0.896

Se realizaron diversas pruebas buscando los mejores resultados y tratando de evitar en lo posible aquellas variables que de alguna manera representen una inducción a la clase. La siguiente prueba consta de la siguiente configuración: se utilizaron 37 variables, dos variables menos (TAS y TAD), 35 algoritmos con un total de 53 instancias positivas y 86 negativas sin selección de atributos para la comorbilidad de Preeclampsia.

Rendimiento de los algoritmos sin las variables TAS y TAD y con desbalanceo de clases sin selección de características para Preeclampsia

Clasificador	Sensibilidad	Especificidad	ROC	Exactitud
DecisionTable	0.736	0.709	0.707	0.72
FilteredClassifier	0.679	0.767	0.67	0.73
DecisionStump	0.679	0.779	0.656	0.74
LWL	0.642	0.756	0.715	0.71
AttributSelectedClassifier	0.642	0.756	0.683	0.71
J Rip	0.623	0.802	0.676	0.73
BayesNet	0.604	0.767	0.675	0.71
Logistic	0.604	0.721	0.663	0.68
MultiClassClassifier	0.604	0.721	0.663	0.68
J48	0.566	0.814	0.696	0.72
MultilayerPerceptron	0.566	0.767	0.692	0.69
RandomCommitte	0.547	0.779	0.691	0.69
RandomTree	0.547	0.791	0.669	0.70
RandomSubSpace	0.528	0.86	0.743	0.73
ClassificationViaRegression	0.528	0.884	0.741	0.75
LogitBoost	0.528	0.802	0.712	0.70
LMT	0.528	0.837	0.701	0.72
AdaBoostM1	0.528	0.756	0.698	0.67
IterativeClassifierOptimizer	0.528	0.802	0.653	0.70
PART	0.528	0.733	0.61	0.65
Bagging	0.509	0.919	0.755	0.76
RandomForest	0.509	0.919	0.752	0.76
SimpleLogistic	0.509	0.814	0.682	0.70
SDG	0.491	0.814	0.652	0.69
MultiClassClassifierUpdateable	0.491	0.814	0.652	0.69
SMO	0.415	0.919	0.667	0.73
IBK	0.415	0.907	0.667	0.72
RandomizableFilteredClassifier	0.415	0.779	0.593	0.64
OneR	0.396	0.802	0.599	0.65
REPTree	0.358	0.93	0.673	0.71
Kstar	0.358	0.953	0.648	0.73
NaiveBayes	0.321	0.872	0.716	0.66
NaiveBayesUpdateable	0.321	0.872	0.716	0.66
HoeffdingTree	0.189	0.942	0.636	0.65

VotedPerceptron	0.094	0.872	0.604	0.58

Rendimiento de los 35 algoritmos, se utilizaron 37 variables con un total de 53 instancias positivas y 86 negativas aplicando técnicas de selección de características para la comorbilidad de Preeclampsia

Rendimiento de los algoritmos con 37 variables y selección de características para la comorbilidad de Preeclampsia

Clasificador	Atributos (Factores de riesgo)	ROC	Sensibilidad	Especificidad	Exactitud
BayesNet (1)	acidoUri	0.656	0.679	0.779	0.74
LWL (3)	totCesareas, leucocitos, acidoUri	0.719	0.679	0.767	0.73
ClassificationViaRegression (4)	antFamDm2, hematocrito, acidoUri, colesterol	0.724	0.679	0.791	0.75
FilteredClassifier (1)	acidoUri	0.656	0.679	0.779	0.74
IterativeClassifierOptimizer (3)	eritrocitos, eosinofilo, acidoUri	0.715	0.679	0.779	0.74
DecisionTable (1)	acidoUri	0.679	0.679	0.779	0.74
Jrip (3)	acidoUri	0.656	0.679	0.779	0.74
DecisionStump (1)	acidoUri	0.656	0.679	0.779	0.74

Clasificador	Atributos (Factores de riesgo)	ROC	Sensibilidad	Especificidad	Exactitud
AdaBoostM1 (3)	antFamDm2, IMCMat, acidoUri	0.734	0.66	0.779	0.73
Kstar (3)	antFamDm2, totPartos, acidoUri	0.762	0.642	0.849	0.77
LogitBoost (8)	antFamDm2, parenFamDm2, antFamHta, totCesareas, totPartos, cmhb, cmhbc, acidoUri	0.795	0.642	0.884	0.79
REPTree (3)	vpm, acidoUri, colesterol	0.665	0.623	0.802	0.73
AttributSelectedClassifier (1)	acidoUri	0.634	0.604	0.767	0.71
OneR (1)	acidoUri	0.715	0.604	0.826	0.74
LMT (7)	antFamHta, totCesareas, totAbortos, vgm, cmhb, cmhbc, acidoUri	0.813	0.604	0.884	0.78
MultilayerPerceptron (3)	vgm, neutrofilo, acidoUri	0.712	0.566	0.791	0.71
Bagging (5)	edad, antFamHta, contaminacion, plaquetas, acidoUri	0.764	0.566	0.849	0.74

Clasificador	Atributos (Factores de riesgo)	ROC	Sensibilidad	Especificidad	Exactitud
Logistic (8)	antFamDm2, contaminacion, totGestas, totPartos, neutrofilo, monocito, gSangre, glucosa	0.738	0.528	0.872	0.74
MultiClassClassifier (8)	antFamDm2, contaminacion, totGestas, totPartos, neutrofilo, monocito, gSangre, glucosa	0.738	0.528	0.872	0.74
RandomSubSpace (3)	escolaridad, rdw, acidoUri	0.74	0.528	0.872	0.74
J48 (5)	rdw, gSangre, urea, acidoUri, triglicerido	0.619	0.509	0.895	0.75
PART (5)	vgm, rdw, vpm, neutrofilo, acidoUri	0.63	0.491	0.919	0.76
HoeffdingTree (6)	antFamDm2, totCesareas, totPartos, rdw, neutrofilo, urea	0.756	0.472	0.942	0.76
RandomCommitte (2)	totGestas, totPartos	0.633	0.396	0.93	0.73
RandomizableFilteredClassifier (2)	totGestas, totPartos	0.635	0.396	0.93	0.73
RandomTree (2)	totGestas, totPartos	0.633	0.396	0.93	0.73

Clasificador	Atributos (Factores de riesgo)	ROC	Sensibilidad	Especificidad	Exactitud
NaiveBayes (5)	parenFamDm2, talla, totGestas, totCesareas, totPartos	0.654	0.377	0.953	0.73
NaiveBayesUpdateable (5)	parenFamDm2, talla, totGestas, totCesareas, totPartos	0.654	0.377	0.953	0.73
IBK (2)	totGestas, totPartos	0.62	0.377	0.93	0.72
RandomForest (2)	totGestas, totPartos	0.63	0.358	0.953	0.73
SDG (4)	IMCMat, totCesareas, cmhb, acidoUri	0.643	0.321	0.965	0.72
SimpleLogistic (4)	parenFamDm2, totGestas, totPartos, cmhb	0.613	0.321	0.942	0.71
MultiClassClassifierUpdateable (4)	IMCMat, totCesareas, cmhb, acidoUri	0.643	0.321	0.965	0.72
VotedPerceptron (3)	antFamHta, totGestas, rdw	0.591	0.075	0.988	0.64
SMO (1)	0	0	0	0	0.00