Universidad de Cienfuegos Carlos Rafael Rodríguez

Facultad de Ingeniería

Trabajo de Diploma para optar por el título de Ingeniero Informático

SELECCIÓN DE RASGOS INFLUYENTES EN LA PREMATURIDAD Y DETERMINACIÓN DE UN MODELO PARA PREDECIR EL PARTO PRE-TÉRMINO

Autor:

Lilian María Tay González

Tutores:

Lic. Ciro Rodríguez León

Msc. Yailem Arencibia Rodríguez del Rey

Consultante:

Dra. Aymara Rodríguez Márquez

Cienfuegos, Cuba

Curso 2012-2013

Declaración de autoría

Yo Lilian María Tay González, me declaro como único autor del trabajo de diploma titulado "*Selección de rasgos influyentes en la prematuridad y determinación de un modelo para predecir el parto pre-término*", y autorizo al Departamento de Informática de la Facultad de Ingeniería en la Universidad de Cienfuegos "Carlos Rafael Rodríguez" para que hagan el uso que estimen pertinente con este trabajo.

Para que así conste firmo dicha declaración a los 10 días del mes de Junio del año 2013.

Lilian María Tay González.

Nombre completo del autor

Lic. Ciro Rodríguez León

MSc. Yailem Arencibia Rodríguez del Rey

Nombre completo de los tutores

Los abajo firmantes certificamos que el presente trabajo ha sido revisado según acuerdo de la dirección de nuestro centro y el mismo cumple los requisitos que debe tener un trabajo de esta envergadura referente a la temática señalada.

_____ _____

Firma de la Tutora Firma del Tutor

_____ _____

Firma ICT Firma Vicedecano

Pensamiento

“No a nosotros, oh Jehová, no a nosotros,

Sino a tu nombre da gloria,

Por tu misericordia, por tu verdad.”

La Biblia, Salmos 115: 1

Agradecimientos

- Ante todo dejo constancia de mi gratitud a Dios, y sincero reconocimiento de su providencia necesaria y eficaz durante todo este tiempo de mi carrera estudiantil.
- A mi esposo Dariel también doy gracias, pues me ha sido de gran apoyo y aliento.
- Cuidado y dedicación digna de resaltar han tenido mis amados padres, por tanto empeño le doy gracias.
- A la Familia de mi esposo (Padres y hermano) por su oportuno apoyo y esfuerzo.
- A mis tutores Ciro y Yailem que bastante tiempo pudieron invertir en mí y aportaron sus conocimientos para que esta investigación se realizara.
- A la Dra. Aymara y su familia por su oportuna aparición y por brindarme toda la información necesaria.
- Agradezco a mis hermanos maternos: Jorgito y esposa, Alain y su esposa Thays (que también es mi amiga) que siempre están al tanto de mi vida.
- A mi hermanito paterno: Juan Pablo y su mamá por su disposición.
- A mis tías que se preocupan tanto por mí.
- A mi sobrino Aarón por su llegada a finales del año anterior.
- A mis queridos hermanos y amigos porque ha sido de consuelo el apoyo que me han brindado.
- A mi vecina Vanesa y a su hija Sofi que me brindan su tiempo y se preocupan por mí.
- A todos los profesores que impartieron sus materias los cinco años de la carrera.

Resumen

Esta investigación, titulada: **"*Selección de rasgos influyentes en la prematuridad y determinación de un modelo para predecir el parto pre-término*"**, constituye una alternativa para mejorar el diagnóstico del parto pre-término en el hospital provincial de Cienfuegos. El empleo de técnicas de inteligencia artificial, permite mejorar las estrategias existentes para la prevención y manejo de este problema de salud en los diferentes niveles de atención médica.

Dentro de la IA, las técnicas de selección de atributos (análisis factorial, CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, WrapperSubsetEval) y los métodos de clasificación supervisados (K-vecinos más cercanos, árbol de decisión: J48, perceptrón multicapa y máquina de soporte vectorial) fueron los seleccionados para determinar los factores de riesgo de prematuridad y obtener el mejor modelo para predecir el parto pre-término respectivamente.

La experimentación se desarrolla en *Statistical Product and Serviced Solution* (SPSS) y *Waikato Environment for Knowledge Analysis* (WEKA). Consiste en probar con diferentes variantes para cada método, modificando sus parámetros hasta obtener las mejores.

Los rasgos influyentes en la prematuridad, según la experimentación y el criterio del experto, son (en orden de importancia): parto pre-término anterior, hábito de fumar, ciur, edad materna, diabetes gestacional, embarazo múltiple, gestorragia e infección vaginal.

El modelo que se propone para predecir el parto pre-término en el hospital provincial de Cienfuegos, con 81.0% de clasificaciones correctas y 31.5% de falsos negativos, es el SMO, con los siguientes parámetros: buildLogisticModels: verdadero; debug: verdadero; Epsilon: 1.0E-12; filterType: normalizar; Kernel: NormalizedPolyKernel; numFolds: -1; randomSeed: 1 y toleranceParameter: 0.0010.

Índice de Contenido

Índice de tablas

Índice de figuras

Introducción

El parto es un proceso caracterizado por cambios moleculares en el miometrio[1], cérvix[2] y otros tejidos gestacionales. El cuello del útero que es firme durante todo el estado no gestacional y durante todo el embarazo hasta cerca del momento del parto, se ablanda y se dilata a la vez que el cuerpo uterino se contrae y expulsa al feto.[1]

El número de receptores a la oxitocina[3] en el miometrio y en la decidua (endometrio en el embarazo) aumenta más de 100 veces durante este y alcanza el máximo al principio del trabajo de parto. Se conoce que la oxitocina aumenta las contracciones uterinas de varias formas, no solo haciendo que se contraigan las células musculares lisas sino estimulando la producción de prostaglandinas en la decidua que a su vez aumentan las contracciones inducidas por la oxitocina, comprobándose que ambos procesos son necesarios para que se produzca un trabajo de parto normal. [1]

Durante el período de gestación de la vida humana existen diversos riesgos. Dentro de estos riesgos se encuentra el parto pre-término, que se efectúa antes de las 37 semanas de gestación.

Los nacimientos prematuros pueden dividirse en dos categorías: aquellos que son espontáneos por inicio precoz del parto o ruptura prematura de las membranas y aquellos que son inducidos. Los partos prematuros inducidos pueden ocurrir cuando la salud de la madre o del feto está en peligro, como con la preclampsia (presión arterial peligrosamente alta durante el embarazo); por comodidad del médico, de la partera o de la madre; o por un error en la fecha del parto.[2]

[1] Miometrio: es la capa muscular intermedia, entre la serosa peritoneal y la mucosa glandular (endometrio), que constituye el grueso del espesor de la pared del cuerpo uterino.

[2] Cérvix: es la parte inferior del útero, situada en el fondo de la vagina, flexible, delgada y de unos tres centímetros de longitud.

[3] Oxitocina: es una hormona relacionada con los patrones sexuales y con la conducta maternal y paternal que actúa también como neurotransmisor en el cerebro.

El neonato pre-término representa la primera causa de mortalidad perinatal y la primera causa de ingreso a la terapia intensiva neonatal. Representa, aproximadamente el 11.1% de todos los nacimientos. Existen complicaciones del periodo neonatal inmediato y secuelas a largo plazo con predilección del sistema nervioso central. No se conoce la causa del fenómeno desencadenante del parto pre-término (PP) espontáneo, por consiguiente, las medidas terapéuticas no han dado los resultados esperados, por lo que su frecuencia no ha disminuido. Parece ser que solamente las medidas preventivas, la identificación de algunos factores de riesgo, y algunas mediciones de laboratorio permitirán iniciar tratamientos o tomar medidas en edades tempranas del embarazo y, probablemente, disminuir las complicaciones, las secuelas y la mortalidad.[1][3]

Según la Organización Mundial de la Salud (OMS), factor de riesgo es toda característica o circunstancia determinable de una persona o de un grupo de personas, que según los conocimientos que poseen, está asociada a un riesgo anormal de aparición o evolución de un proceso patológico o de afectación desfavorable de tal proceso. Aclarado esto, es importante resaltar que un factor de riesgo no significa causa, es simplemente una herramienta para calificar la posibilidad de daño y el riesgo es la probabilidad de sufrirlo. [4]

El número de nacimientos prematuros está aumentando internacionalmente. Unos 50 millones de nacimientos en el mundo aún ocurren en los hogares y muchos bebés mueren sin certificados de nacimiento o de muerte. En países de altos ingresos, el aumento en el número de nacimientos prematuros está vinculado con el número de mujeres mayores teniendo bebés y el aumento en el consumo de drogas de fertilidad, resultando en embarazos múltiples. En algunos países desarrollados los partos médicamente inducidos innecesariamente y las cesáreas antes de término también han aumentado los nacimientos prematuros. En muchos países de bajos ingresos las principales causas de los nacimientos prematuros incluyen infecciones, malaria, VIH y altas tasas de embarazo adolescente. En países ricos y pobres, muchos nacimientos prematuros siguen siendo inexplicables. [2]

Parte inicial del manejo del parto pre-término es catalogar estrictamente el caso, estableciendo para ello la edad gestacional, determinada tanto por la clínica como por los estudios ecográficos realizados durante la gestación si los hubiera. El examen clínico abarcará el estado general y clínico de la madre y el bienestar fetal, para tomar las medidas necesarias.[5]

A raíz de esta situación internacional muchos estudios se centran hoy en la búsqueda de factores de riesgo. Entre aquellos factores que han sido identificados se encuentran: bajo peso, obesidad, diabetes, hipertensión, fumar, infecciones, edad materna (menores de 17 o mayores de 40), genética, embarazo múltiple (gemelos, trillizos o mayor), y los embarazos demasiado seguidos. [2]

Investigadores como Iams [6], Meis [7] y Astolfi [8], han encontrado asociaciones entre el riesgo aumentado de parto pre-término en los extremos de la edad materna (menos de 20 y más de 35 años).

Por otro lado Astolfi [8], Parker [9] y Peacock [10] han reportado el nivel socioeconómico y educativo como factores de riesgo.

Según Rosell [11] en muchas mujeres en trabajo de parto pre-término no es posible establecer su etiología de ahí que este autor defienda que lo más importante no es solo conocer los factores de riesgo sino también la detección temprana del trabajo de parto pre-término.

En Cuba se llevan a cabo acciones preventivas encaminadas a disminuir este indicador. A pesar de los índices elevados de prematuridad en todo el mundo, del año 2005 al 2010, el nacimiento pre-término en el territorio se encontraba entre el 4% y el 4,8% del total de nacimientos [12]. A pesar de esto el criterio de los especialistas obstétricos es que este problema va en aumento.

Cienfuegos no queda exento de esta problemática internacional. En las estadísticas hospitalarias no se lleva un control del total de prematuros que nacen en la actualidad y por otro lado, se desconoce cómo repercuten los diferentes factores de riesgo en el desencadenamiento del parto antes de término en la provincia.

Por otro lado el número de investigaciones en el campo de la medicina que emplean técnicas de inteligencia artificial cada día toma mayor auge con resultados satisfactorios. En la amplia gama de aplicaciones de estas técnicas se encuentran: la interpretación de imágenes médicas, sistemas expertos en apoyo al diagnóstico médico, la monitorización y control en las unidades de cuidados intensivos, diseño de prótesis, diseño de fármacos, sistemas tutores inteligentes para diversos aspectos de la medicina.

Muchos investigadores emplean estas técnicas para determinar rasgos relevantes u obtener modelos de pronóstico médico que permita aumentar la efectividad del diagnóstico médico y el tratamiento. Tal es el caso de:

- Selección de atributos en una base de datos que contenía variables involucradas en el estado nutricional de niños de 6 a 11 años por Lic. Roxana Martín Ramos, MSc. Rosa María Ramos Palmero, Dr. Ricardo Grau Ávalos, Dra. María Matilde García Lorenzo[13].

- "Modelo para el diagnóstico, tratamiento y seguimiento de pacientes con hipertensión arterial" por Alfredo Morales, Raykenler Yzquierdo y otros colaboradores.[14]

- "Modelo de redes neuronales artificiales para el diagnóstico de asma" por Young Moon Chae y Seung Hee Ho [15]

Teniendo en cuenta todo lo anterior se identifica como **problema a resolver:** la necesidad de determinar los factores de riesgo que más influyen en la prematuridad y un modelo para predecir el parto pre-término.

Se define como **objeto de estudio:** el proceso de embarazo en pacientes con riesgo de prematuridad, como **campo de acción**: el pronóstico del parto pre-término, aplicando técnicas de inteligencia artificial.

Se plantea como **idea a defender** que: la determinación de los factores de riesgo más influyentes en la prematuridad y la obtención de un modelo para predecir el parto pre-término, utilizando técnicas de inteligencia artificial, permite mejorar las

estrategias existentes para la prevención y manejo de este problema de salud en los diferentes niveles de atención médica.

La labor científico investigativa está encaminada a dar cumplimiento al siguiente **objetivo general:**

- Determinar los factores de riesgo influyentes en la prematuridad y proponer un modelo para el pronóstico del parto pre-término.

Partiendo del objetivo anterior se plantean los siguientes **objetivos específicos:**

- Estudiar el estado del arte sobre los factores de riesgo de prematuridad, el pronóstico del parto pre-término y la aplicación de técnicas de inteligencia artificial frente a problemas de clasificación.
- Aplicar diferentes técnicas para seleccionar los rasgos o factores de riesgo más influyentes en la prematuridad.
- Experimentar con diferentes modelos supervisados para la clasificación de nuevos casos con riesgo de parto pre-término.
- Comparar los resultados obtenidos en la experimentación para concluir los rasgos más influyentes en la prematuridad y proponer el mejor modelo para su predicción.

La justificación de la investigación está dada por la necesidad de encontrar un modelo que permita un pronóstico correcto del parto pre-término en embarazadas y determinar las variables predictivas que ofrezcan una mayor certeza en la toma de decisiones. De esta manera se contribuye a mejorar la calidad del diagnóstico médico y las medidas de prevención.

El trabajo se estructura de la siguiente forma: introducción, tres capítulos, conclusiones y recomendaciones.

Capítulo I: Fundamentación teórica. En este capítulo se exponen los conceptos asociados al dominio del problema y el estado del arte sobre los factores de riesgo que más influyen en la prematuridad y el proceso de pronóstico del parto pre-término.

También se abordan las técnicas de inteligencia artificial y aplicaciones de estas en el campo de la salud asociadas a la predicción de nuevos casos.

Capítulo II: Diseño del experimento. Se presenta la base de casos y la organización de la experimentación en las herramientas seleccionadas, mostrando los resultados alcanzados por los diferentes métodos de selección de rasgos y modelos de clasificación supervisada utilizados.

Capítulo III: Análisis de los resultados. Se comparan los resultados obtenidos en la experimentación con los diferentes métodos para llegar a conclusiones con respecto a los rasgos que más influyen en la prematuridad y el mejor modelo para la predicción del parto pre-término.

1 Capítulo I: Fundamentación teórica.

El presente capítulo expone los principales aspectos asociados al dominio del problema. Se mencionan diversas aplicaciones en la medicina que hacen uso de técnicas de inteligencia artificial y se aborda el tema de la selección de rasgos y clasificación supervisados de forma general. Por último se especifican los métodos que serán usados para la solución del problema planteado y las razones por las cuales serán empleados.

1.1 Definición de prematuridad.

La Organización Mundial de la Salud, considera el nacimiento pre-término como aquel que se efectúa antes de las 37 semanas de gestación.[2]

Mientras más corto sea el período del embarazo mayor es el riesgo de padecer serias complicaciones. Los bebés prematuros tienen un alto riesgo de muerte en sus primeros años de vida o pueden llegar a desarrollar graves problemas de salud.

1.2 Incidencia del parto pre-término e investigaciones relacionadas en el ámbito internacional.

Durante gran parte del siglo XX el nacimiento de niños pre-término, era un hecho impredecible e inevitable. Los esfuerzos estaban centrados en disminuir las consecuencias que ello generaba, en lugar de prevenir su ocurrencia. Esto mejoró los resultados neonatales, pero con un alto costo para los niños y sus familias. Actualmente las investigaciones se enfocan en la prevención del parto pre-término.[16]

Un informe de la fundación “March of Dimes”, la OMS y otras organizaciones reporta que: cada año, unos 15 millones de bebés en el mundo, más de uno en 10 nacimientos, nacen demasiado pronto. Más de un millón de estos bebés mueren poco después del nacimiento; muchos otros sufren algún tipo de discapacidad física, neurológica o educativa.[2]

A pesar de los avances tecnológicos y estudios que se enfocan en la búsqueda de factores de riesgos que pueden aumentar la posibilidad de que ocurra el parto pre-término, no se ha logrado frenar el incremento constante de la prematuridad.

Los nacimientos prematuros representan casi la mitad de todas las muertes de recién nacidos en el mundo según la Dra. Joy Lawn.[2]

El problema de los nacimientos prematuros no se limita a los países de bajos ingresos. Los Estados Unidos y Brasil se ubican entre los 10 países con mayor número de nacimientos prematuros. Por ejemplo, en los Estados Unidos, alrededor del 12 %, o más de uno en nueve de todos los nacimientos, son prematuros.[2]

Los países con el mayor número de nacimientos prematuros son: India –3.519.100; China –1.172.300; Nigeria – 773.600; Pakistán –748.100; Indonesia –675.700; Estados Unidos –517.400; Bangladesh –424.100; Filipinas –348.900; República Democrática del Congo –341.400; y Brasil –279.300.[2]

Los 10 países con las mayores tasas de nacimientos prematuros por cada 100 nacimientos son: Malawi –18.1 por cada 100; Comoras y Congo –16.7; Zimbawue –16.6; Guinea Ecuatorial –16.5; Mozambique –16.4; Gabón –16.3; Pakistán –15.8; Indonesia –15.5; y Mauritania –15.4. Estos países contrastan con los 11 países con las tasas más bajas de nacimientos prematuros: Belarús –4.1; Ecuador –5.1; Letonia –5.3; Finlandia, Croacia y Samoa –5.5; Lituania y Estonia –5.7; Barbados/Antigua –5.8; Japón –5.9. [2]

Según la Dra. Joy Lawn, unos 50 millones de nacimientos en el mundo aún ocurren en los hogares y muchos bebés mueren sin certificados de nacimiento o de muerte.[2]

En países de altos ingresos, el aumento en el número de nacimientos prematuros está vinculado con el número de mujeres mayores teniendo bebés y el aumento en el consumo de drogas de fertilidad, resultando en embarazos múltiples. En algunos países desarrollados, los partos médicamente inducidos innecesariamente y las cesáreas antes de término también han aumentado los nacimientos prematuros. En

muchos países de bajos ingresos, las principales causas de los nacimientos prematuros incluyen infecciones, malaria, VIH y altas tasas de embarazo adolescente. En países ricos y pobres, muchos nacimientos prematuros siguen siendo inexplicables.[2]

Se han identificado varios factores de riesgo del nacimiento prematuro, incluyendo una historia previa de nacimiento prematuro, bajo peso, obesidad, diabetes, hipertensión, fumar, infecciones, edad materna, genética, embarazo múltiple, y los embarazos demasiado seguidos. [2]

En la actualidad las investigaciones en el tema están relacionadas con la búsqueda de aquellos factores de riesgos que aumentan la posibilidad del parto pre-término. En estos estudios se destacan autores como Iams [6], Meis [7] y Astolfi [8], que han encontrado asociaciones entre el riesgo aumentado de parto pre-término en los extremos de la edad materna (menos de 20 y más de 35 años).

Otro de los factores de riesgo para la prematuridad que ha sido reportado el bajo nivel socioeconómico y educativo. Esta afirmación ha sido sustentada mediante estudios epidemiológicos de investigadores como Astolfi [8], Parker [9] y Peacock [10] que han buscado las asociaciones entre el nivel económico, tipo de seguridad social y nivel educativo, en poblaciones que acceden a instituciones de carácter público.

Sobre la asociación entre las pacientes con parto pre-término y control prenatal inadecuado, estudios como el de Ballard [17], Mercer y colaboradores [18], Bakketeig [19] y Adams [20], reflejan que el antecedente de parto prematuro ha sido descrito como el principal factor de riesgo en multíparas[4].

1.3 Incidencia del parto pre-término en Cuba e investigaciones nacionales.

La prematuridad debería ser hoy en día una de las prioridades de salud de los gobiernos, ya que como acuñan las estadísticas su prevalencia es elevada, además

[4] Multípara: la mujer que ha tenido varios partos.

de su aumento constante y las consecuencias que produce, afectando la calidad de vida del niño prematuro y de sus familiares.

En Cuba antes del año 2000, el nacimiento pre-término representaba del 8 al 9% de los partos y estaba asociado al 75% de la mortalidad perinatal.[21]

A pesar de un ligero aumento de la incidencia de la prematuridad en nuestro país, hasta el año 2005 esta se mantenía por debajo del 3,0% de los nacidos vivos, como resultado de las acciones preventivas encaminadas a disminuir este indicador.[22]

Del año 2005 al 2010 y comparado con los años anteriores, hubo un incremento en el número de nacimientos de niños pre-término que representaron entre el 4 y el 4,8% del total de nacimientos.[12]

Dentro de las investigaciones realizadas en Cuba se destaca un estudio prospectivo descriptivo en el hospital Ginecobstétrico Provincial Docente de Camagüey “Ana Betancourt de Mora”, desarrollado por los doctores: Dra. Maitté Vera López, Dr. Frank Alberto Castillo Fernández y Dra. Noris Navas Ábalos, para determinar la repercusión del parto pre-término desde enero a diciembre del año 2004. Como resultado se obtuvo que la infección vaginal es la enfermedad asociada más frecuente, además la depresión respiratoria transitoria y neumonía congénita constituyó la principal morbilidad de los neonatos; mientras que la anemia y la dehiscencia de la rafia la principal causa de las maternas.[23]

Las doctoras: MSc. Dra. Lázara Miriam Ortega Figueroa, Dra. Ana Bertha Álvarez Pineda, MSc. Dr. Yahavivi Águila Nogueira y la Dra. Mercedes I. Viera Hernández en el Hospital "Profesor Eusebio Hernández" en La Habana, realizaron un estudio retrospectivo descriptivo de enero a septiembre de 2010 para la detección de infección por Mycoplasma en las gestantes con riesgo de parto pre-término. Los resultados mostraron que el 67% de las pacientes presentaban infección moderada o severa mientras que los exudados vaginales simples fueron negativos en un 71,5%, el antimicrobiano más utilizado fue la eritromicina.[24]

1.4 La prematuridad en la provincia de Cienfuegos.

En la provincia de Cienfuegos los especialistas de las atenciones primaria y secundaria de la salud desempeñan una labor en conjunto, donde la salud de las gestantes y del recién nacido es una prioridad del sector. Una de las tareas que se llevan a cabo es la de dar seguimiento a las embarazadas desde las 18 semanas de gestación, período en el cual se les realiza un ultrasonido transvaginal[5] para prevenir partos precoces. Aquellas gestantes de alto riesgo se mantienen en observación, ya sea en ingreso domiciliario o en hogares maternos.

A pesar de esto Cienfuegos no queda exento de esta problemática internacional. En las estadísticas hospitalarias no se lleva un control del total de prematuros que nacen en la actualidad y por otro lado, se desconoce cómo repercuten los diferentes factores de riesgo en el desencadenamiento del parto antes de término en la provincia.

1.5 Introducción a la inteligencia artificial.

Existen varias definiciones de la inteligencia artificial (IA) hecha por diferentes autores. A continuación se presentan algunas de estas organizadas en cuatro categorías: [25]

- sistemas que piensan como los humanos.
- sistemas que actúan como humanos.
    - Según R. Kurzweil la IA es el arte de crear máquinas con capacidad de realizar funciones que realizadas por personas requieren de inteligencia.[26]
    - Según E. Rich y K. Knight la IA es el estudio de cómo lograr que las computadoras realicen tareas que, por el momento, los humanos hacen mejor.[27]
- sistemas que piensan racionalmente.
- sistemas que actúan de forma racional.

[5] Ultrasonido transvaginal: técnica que permite medir la longitud del cérvix.

- Según Stubblefied la IA es la rama de la ciencia de la computación que se ocupa de la automatización de la conducta inteligente.[28]
- Según R. Schalkoff la IA es el campo de estudio que se enfoca a la explicación y emulación de la conducta inteligente en función de procesos computacionales. [29]

En esta investigación se sigue la definición presentada por Schalkoff, 1990.

1.5.1 Características de la IA.

Alguna de las características que presenta la IA según Pablo David Santiago son: [30]

- uso de símbolos no matemáticos.
- el comportamiento de los programas no es descrito explícitamente por el algoritmo.
- el razonamiento basado en el conocimiento.
- aplicabilidad a datos y problemas mal estructurados.

Las principales corrientes en IA son el cognitivismo y el conexionismo. El enfoque conexionista es muy diferente al enfoque simbólico tradicional pero ambos comparten el mismo problema: intentar estudiar los difíciles aspectos de búsqueda, la representación del conocimiento y el aprendizaje.

Para dar soluciones a los problemas que aborda la IA existen diferentes técnicas. Una técnica de IA es un método que utiliza conocimiento representado de tal forma que: el conocimiento represente las generalizaciones.[31]

Existen dos clases importantes de técnicas de IA: métodos para representación y el uso del conocimiento; métodos para conducir la búsqueda heurística. Estos dos aspectos interactúan fuertemente entre ellos. La elección del marco de representación del conocimiento determina el tipo de método para solucionar problemas que pueden aplicarse.[32]

1.5.2 Aplicaciones de la IA.

Hay un conjunto muy abierto de campos de aplicación de la IA. Los siguientes son sólo ejemplos, y no una lista completa: [33]

- IA en la medicina, que incluye la interpretación de imágenes médicas, sistemas expertos de apoyo al diagnóstico médico, la monitorización y control en las unidades de cuidados intensivos, diseño de prótesis, diseño de fármacos, sistemas tutores inteligentes para diversos aspectos de la medicina.
- IA en la robótica, que incluye la visión, el control de motores, el aprendizaje, la planificación, la comunicación lingüística, el comportamiento cooperativo.
- IA en muchos aspectos de la ingeniería, diagnóstico de fallos, sistemas inteligentes de control, sistemas inteligentes de fabricación, ayuda inteligente al diseño, sistemas integrados de ventas, diseño, producción, mantenimiento, herramientas de configuración expertas (por ejemplo, garantizando que el personal de ventas no vendan un sistema que no funciona). En la ingeniería de software incluye síntesis de programas, verificación, depuración, prueba y monitorización de software.
- IA en interfaces y sistemas de "ayuda", ya que las computadoras se usan para más y más aplicaciones que implican la interacción con los seres humanos, hay cada vez más presiones para construir máquinas más fáciles de utilizar para los no expertos. Un enfoque a este problema consiste en dotar a las máquinas de más inteligencia para que puedan guiar o asesorar a los usuarios. Este tipo de aplicación puede incluir muchos de los sub-campos de la sección previa.
- IA en la gestión de la información: esto incluye el uso de la IA en la minería de datos, el rastreo web, filtrado de correo, etc.
- IA en la biología: hay muchos problemas complicados en biología donde se están desarrollando sistemas informáticos más o menos inteligentes, por ejemplo, análisis de ADN, predicción de la estructura de plegado de moléculas complejas, la predicción, la elaboración de modelos de procesos biológicos, evolución, desarrollo de embriones, comportamientos de los distintos organismos.

1.6 La inteligencia artificial en la medicina.

La inteligencia artificial constituye un tema vigente en el desarrollo de las investigaciones actuales. Además de la existencia de herramientas computacionales que permiten la aplicación de distintas técnicas de IA a diversos problemas. Cada día aumenta el número de investigaciones asociadas a estas técnicas y aplicadas a la medicina con resultados relevantes. Tal es el caso de los ejemplos que presentaremos a continuación.

Los investigadores: Lic. Roxana Martín Ramos, MSc. Rosa María Ramos Palmero, Dr. Ricardo Grau Ávalos, Dra. María Matilde García Lorenzo[13] llevaron a cabo un trabajo donde se aplicaron métodos de selección de atributos a una base de datos que contenía variables involucradas en el estado nutricional de niños de 6 a 11 años, con el propósito de precisar cuál de los métodos aplicados determinaba los factores que más aportaron a la evaluación nutricional. Pudo comprobarse que la selección de atributos brindó los factores más relevantes.

Otra investigación fue realizada por: Ing. Rolando Acosta Sánchez, Dr. Alejandro Rosete Suárez, Lic. Alfredo Rodríguez Díaz [34] donde se aplicaron técnicas de Minería de Datos a los resultados de una encuesta realizada en la localidad de Jaruco, Provincia Habana, Cuba; a pacientes identificados por el personal médico como diabéticos o con riesgo de padecer la enfermedad. El trabajo permitió identificar y describir las características principales de los datos a emplear para clasificar a los pacientes de la localidad de Jaruco. Se logró identificar los atributos relevantes para la investigación.

Por otro lado Young Moon Chae y Seung Hee Ho [15] realizaron un trabajo donde se aplicaron tres modelos al diagnóstico de asma. Se estableció una comparación entre ellos y se notó que las capacidades diagnósticas para los tres modelos de conocimiento difirieron. La red neuronal tuvo la mejor tasa global (92%) de predicción y la mejor tasa de predicción para el asma (96%); el análisis discriminatorio tuvo la mejor tasa de predicción para el poco asma (80%).

También es el caso de la aplicación de un modelo de red neuronal para clasificar los sujetos en las categorías “normal” y “con enfermedad Alzheimer”, en pruebas de generalización la red clasificó correctamente el 92% de los casos y superó a los métodos estadísticos clásicos como el análisis discriminante. [35]

1.7 Aprendizaje.

De acuerdo con la definición dada por Michalski en 1986, aprender es la habilidad de adquirir nuevo conocimiento, desarrollar habilidades para analizar y evaluar problemas a través de métodos y técnicas, así como también por medio de la experiencia propia; siendo un requisito que el resultado del aprendizaje sea entendible para un hombre. [36]

También, el aprendizaje puede definirse como alguno de los siguientes procesos [37]:

- adquisición de nuevo conocimiento.
- desarrollo de habilidades motoras y cognoscitivas a través de la instrucción o la práctica.
- organización de nuevo conocimiento dentro de representaciones generales y efectivas.
- descubrimiento de nuevos factores y teorías a través de la observación y la experimentación.

De esta forma, el conocimiento humano puede ser expresado a la lógica en conjunción con la notación matemática, para crear máquinas que razonen y propongan un modelo de solución ante un problema dado. Desde el comienzo de la historia humana, el aprender ha sido una característica propia de cada individuo que ha contribuido a fundamentar las bases del desarrollo, ya que cuando se aprende se adquiere el conocimiento por medio del estudio, el ejercicio o la experiencia. [37]

1.7.1 Tipo de aprendizaje.

El aprendizaje se lleva a cabo de manera automática ya que es un proceso tan sencillo para el hombre que él mismo no percibe cómo se realiza y todo lo que ello implica. En cualquier proceso de aprendizaje, el aprendiz aplica el conocimiento

poseído a la información que le llega, procedente de un maestro o del entorno, para obtener nuevo conocimiento, que es almacenado para poder ser usado posteriormente. [38]

Dependiendo del esfuerzo requerido por el aprendiz (o número de inferencias que necesita sobre la información que tiene disponible) han sido identificadas varias estrategias. Una clasificación ya "clásica" de los diferentes tipos de aprendizaje más importantes, en orden creciente de esfuerzo inferencial por parte del aprendiz, es [36]:

- **aprendizaje por instrucción:** el sistema de aprendizaje adquiere el nuevo conocimiento a través de la información proporcionada por un maestro, pero no la copia directamente en memoria, sino que selecciona los datos más relevantes y/o los transforma a una forma de representación más apropiada.
- **aprendizaje por deducción:** partiendo del conocimiento suministrado y/o poseído, se deduce el nuevo conocimiento, es decir, se transforma el conocimiento existente mediante una función preservadora de la verdad.
- **aprendizaje por inducción:** el sistema de aprendizaje aplica la inducción a los hechos u observaciones suministrados, para obtener nuevo conocimiento. La inferencia inductiva no preserva la verdad del conocimiento, sólo su falsedad; es decir, si se parte de hechos falsos, el conocimiento adquirido por inducción será falso, pero si los hechos son verdaderos, el conocimiento inducido será válido con cierta probabilidad (y no con certeza absoluta, como ocurre con la deducción). Hay dos tipos de aprendizaje inductivo y de los cuales se hablará más adelante: aprendizaje con ejemplos o Supervisado y Aprendizaje por observación y descubrimiento o no Supervisado.

1.8 Selección de rasgos o atributos.

Existen dos principales enfoques para la reducción dimensional de los datos: extracción de rasgos (Feature Extraction) y selección de rasgos (Feature Selection); mientras que el primero de estos busca pasar de un espacio multidimensional de rasgos a uno menor realizando transformaciones sobre los mismos, el segundo

busca la obtención de un subconjunto de rasgos que no sólo reduce dimensionalmente el problema, sino que además, mejora la precisión en la toma de decisiones; de ahí que esta técnica también sea considerada como una alternativa combinatoria de optimización. [39]

Un proceso de selección de atributos se divide en cuatro etapas [40]: en la primera se determina el posible conjunto de atributos para realizar la representación del problema; en la segunda se evalúa el subconjunto de atributos generado en el paso anterior; el chequeo de si el subconjunto seleccionado satisface el criterio de detección de búsqueda se realiza en la tercera etapa; y por último, en la etapa cuarta se verifica la calidad del subconjunto de atributos que se determinó.

Atendiendo a la función de evaluación usada, los procedimientos de selección de atributos se pueden clasificar en tres categorías [41]:

- Filters: en estos métodos el procedimiento de selección es realizado independientemente de la función de evaluación.
- wrappers: son los que combinan la búsqueda en el espacio de atributos con el algoritmo de aprendizaje, evaluando los conjuntos de atributos y escogiendo el más adecuado. [42]
- híbridos: estos modelos toman las ventajas que aportan los modelos filters y wrappers, aprovechando sus diferentes criterios de evaluación en diferentes etapas de búsqueda.[43]

Independientemente de la función de evaluación elegida, los métodos de selección de atributos deben realizar una búsqueda entre los distintos candidatos de subconjuntos de atributos, pudiendo ser [41]: completa que garantiza encontrar el resultado óptimo [44], secuencial que puede que no encuentre el subconjunto óptimo o aleatoria que tampoco asegura el óptimo [45].

De esta forma se pueden manipular rápidamente grandes volúmenes de datos al extraer de ellos solo la información necesaria que los describa sin perder la calidad del sistema, esta representa el problema de encontrar un subconjunto óptimo de características, es decir rasgos o atributos en una base de datos, tales que se pueda

generar un clasificador con la mayor calidad posible a través de un algoritmo inductivo. [39]

1.9 Clasificación.

La clasificación es una tarea básica en el análisis de datos y en el reconocimiento de patrones[46]. Es el medio por el que se ordena el conocimiento [47].

Según Alberto Téllez, la clasificación puede ser formalizada como la tarea de aproximar una función objetivo desconocida (que describe cómo instancias del problema deben ser clasificadas de acuerdo a un experto en el dominio) por medio de una función llamada el clasificador. Comúnmente cada instancia es representada como una lista de valores característicos, conocidos como atributos. [48]

Según investigadores como: D. Michie, D.J. Spiegelhalter y C.C. Taylor [49] la clasificación tiene dos significados distintos:

- se pueden dar un conjunto de observaciones con el objetivo de establecer la existencia de clases o cluster en la información. Conocido como aprendizaje no supervisado o clustering.
- se puede conocer con certeza la existencia de algunas clases y el objetivo es encontrar una regla (procedimiento) con el cual se pueda clasificar cada nueva observación dentro de alguna de estas clases existentes. Conocido como aprendizaje supervisado.

En esta investigación se asume la definición de clasificación dada por D. Michie, D.J. Spiegelhalter y C.C. Taylor, donde se hace referencia al aprendizaje supervisado, en el cual el procedimiento de clasificación se hace a partir de información dada que está correctamente clasificada.

Los contextos en los que el tema de la clasificación es fundamental incluyen, por ejemplo: procedimientos mecánicos para clasificación de cartas basándose en códigos postales leídos en máquinas y en el diagnóstico preliminar de la enfermedad de un paciente con el fin de seleccionar el tratamiento inmediato mientras se esperan los resultados definitivos de las pruebas. De hecho, algunos de los problemas más urgentes que se plantean en la ciencia, la industria y el comercio pueden

considerarse como problemas de clasificación o problemas de decisión usando complejos y a menudo gran cantidad de datos. [49]

1.9.1 Métodos de clasificación.

En esta investigación se adoptaron las perspectivas en la clasificación dadas por D. Michie, D.J. Spiegelhalter y C.C. Taylor donde una gran variedad de técnicas han sido tomadas en cuenta en este tema. Pero pueden ser identificadas tres principales líneas de investigación: [49]

- técnicas estadísticas
- aprendizaje automatizado
- redes neuronales

Estas han involucrado diferentes grupos de profesionales y académicos, enfatizado en diferentes aspectos. Todos los grupos, sin embargo, tienen algunos objetivos en común y han logrado derivar procedimientos que podrían: [49]

- no solo igualar, sino superar el comportamiento de un decisor humano, pero además tienen la ventaja de su consistencia y, hasta cierto punto, su carácter explícito.
- manejar una amplia variedad de problemas, y dados una cierta cantidad de datos, ser extremadamente general.
- ser empleados en escenarios prácticos con probado éxito.

1.9.1.1 Métodos de clasificación basados en técnicas estadísticas.

Dos fases principales de trabajo sobre la clasificación pueden ser identificadas dentro de la comunidad estadística. La primera, la fase “clásica”, concentrada en las derivadas de los primeros trabajos de Fisher sobre discriminación lineal. La segunda, la fase “moderna” que explota modelos de clases más flexibles, mucho de los cuales logran proporcionar una estimación de la distribución de probabilidades conjunta de los atributos de cada clase, la cual puede a su vez proveer una regla de clasificación. [49]

Dentro de los métodos de estadística clásica se encuentran: [49]

- discriminante lineal:
  - discriminante lineal por cuadrado menor
  - caso especial de dos clases
  - discriminante lineal por probabilidad máxima
  - más que dos clases
- discriminante cuadrático:
  - discriminante cuadrático-detalles programación
  - normalización y cálcuos suavizados
  - elección de parámetros normalizados
- discriminante lógico:
- discriminante lógico-detalles programación
- reglas de Bayes

Dentro de los métodos de estadística moderna se hallan: [50]

- estimación de densidad
- vecino más cercano
- Naive Bayes
- red causal
- otros acercamientos recientes:
  - Ace
  - Mars

1.9.1.2 Métodos de clasificación basados aprendizaje automatizado de árboles y reglas.

El aprendizaje automatizado (ML por sus siglas en inglés) se utiliza generalmente para abarcar procedimientos computacionales basados en operaciones lógicas o binarias, que aprenden una tarea a partir de una serie de ejemplos. Aquí nos

ocupamos acerca de la clasificación, y es discutible lo que podría caer bajo la categoría que aglutina el concepto de Aprendizaje Automatizado. Se ha enfatizado mucho la atención en los enfoques del árbol de decisión, en los cuales la clasificación resulta de una sucesión de pasos lógicos. Ellos son capaces de representar el problema más complejo dados los suficientes datos (pero esto puede significar una enorme cantidad de datos). Otras técnicas, tales como los algoritmos genéticos y procedimientos de lógica inductiva (ILP por sus siglas en inglés), están actualmente bajo un activo desarrollo y en principio nos permitirían tratar con tipos de datos más generales, incluyendo aquellos casos en los cuales el número y el tipo de atributo puede variar, y donde niveles adicionales de aprendizaje se superponen, con una estructura jerárquica de atributos, clases y otros. [49]

El aprendizaje automatizado persigue como objetivo generar expresiones de clasificación suficientemente simples para ser comprendidas fácilmente por los humanos. Ellos deben imitar el razonamiento humano suficientemente como para proveer una visión de cómo funciona el proceso de decisión. Al igual que los enfoques estadísticos, el conocimiento contextual (precedente) puede ser explotado en su desarrollo, pero la operación se asume sin la intervención del criterio (juicio) humano. [49]

Algoritmos de aprendizaje: supervisado y no supervisado

Sobre la base de que a las máquinas hay que indicarles cómo aprender, ya que si no se logra que una máquina sea capaz de desarrollar sus habilidades entonces el proceso de aprendizaje no se estará llevando a cabo, sino que solo será una secuencia repetitiva, se ofrecerán dos tipos importantes de ML: supervisado y no supervisado. [37]

- **Aprendizaje supervisado:** el algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Constituye un algoritmo de aprendizaje basado en ejemplos donde el nuevo conocimiento es inducido a partir de una serie de ejemplos y contraejemplos. Un ejemplo de este tipo de algoritmo es el problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de vectores

utilizando una entre varias categorías (clases). La base de conocimiento del sistema está formada por ejemplos de etiquetados anteriores. [37]

- **Aprendizaje no supervisado:** todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Constituye un tipo de aprendizaje por observación y descubrimiento, donde el sistema de aprendizaje analiza una serie de entidades y determina que algunas tienen características comunes, por lo que pueden ser agrupadas formando un concepto. [37]

Entre los métodos de Aprendizaje automatizado de reglas y árboles se encuentran los siguientes: [49]

- reglas y árboles para datos
    - de lo particular a lo general: un paradigma de aprendizaje por reglas
    - árboles de decisión
    - de lo general a lo particular: inducción por árboles
    - stopping rules y class probability trees
    - splitting criteria
    - getting a “right-sized tree”
- statlog’s ML algorithms
    - aprendizaje con árbol: further features of C4.5
    - NewID
    - AC2
    - Further features of CART
    - Cal5
    - Bayes tree
    - Rule-learning algorithms: CN2
    - ITrule
- más allá de la barrera de complejidad
    - árboles con reglas
    - construir nuevos atributos

- límites inherentes de la propuesta - aprendizaje de nivel

1.9.1.3 Métodos de clasificación basados en Redes Neuronales Artificiales (RNA).

El campo de las redes neuronales ha surgido de diversas fuentes, que abarcan desde la fascinación del ser humano por entender y simular el funcionamiento del cerebro humano, así como temas más amplios de copiar las capacidades o aptitudes del ser humano tales como el habla y uso del lenguaje, hasta la práctica comercial, científica y la disciplina ingenieril del reconocimiento de patrones, modelación y predicción. La carrera de la tecnología es una potente fuerza conductora para investigadores, tanto de academia como de la industria, en muchos campos de la ciencia y la ingeniería. En las redes neuronales, al igual que en el aprendizaje automatizado, el entusiasmo por el progreso de la tecnología está suplementado por el reto de reproducir inteligencia en sí misma. [50]

Una amplia gama de técnicas puede caer bajo este acápite, pero, las redes neuronales consisten en capas de nodos interconectados, donde cada nodo produce una función no lineal de su entrada. La entrada de datos a un nodo puede provenir de otros nodos o de datos propios del mismo nodo. También, algunos nodos se identifican con la salida de la red. La red completa representa por tanto un conjunto muy complejo de interdependencias las cuales pueden incorporar cualquier grado de no-linealidad, permitiéndose que puedan ser modeladas funciones muy generales. [50]

En las más sencillas de las redes, los datos de salida provenientes de un nodo son introducidos en otro nodo de tal manera que los mensajes se propaguen a través de las capas de nodos interconectados. Comportamientos más complejos pueden ser modelados por redes en las cuales los nodos que producen el resultado final son conectados con nodos previos, y entonces el sistema tiene las características de un sistema altamente no-lineal con retroalimentación. Se ha planteado que las redes neuronales hasta cierto punto reflejan el comportamiento de las redes de neuronas en el cerebro. [50]

Los enfoques de las redes neuronales combinan la complejidad de algunas técnicas estadísticas con el objetivo del aprendizaje automatizado de imitar la inteligencia humana: sin embargo esto se hace a un nivel más inconsciente y de esta manera no hay recursos acompañantes para hacer que los conceptos aprendidos sean transparentes para el usuario. [50]

Entre los métodos de redes neuronales se encuentran:

- redes supervisadas para la clasificación
    - perceptrón y perceptrón multicapa
    - redes de función base radial
    - redes de propagación hacia delante

1.10 ¿Por qué utilizar técnicas de selección de atributos y clasificación supervisada para dar solución al problema planteado?

La utilidad que tienen las técnicas de clasificación es evidente en todo el mundo, ya que cada día aumenta el número de publicaciones sobre aplicaciones de dichas técnicas de IA en distintas áreas, incluyendo la medicina donde se han obtenido buenos resultados.

Serán usadas las técnicas de clasificación supervisada porque se tiene un conjunto de 687 casos (mujeres que parieron) de los que se conocen características como: edad materna, paridad, abortos, parto pre-término anterior, hábito de fumar, ingestión de bebidas alcohólicas, escolaridad, embarazo múltiple, infección vaginal, infección urinaria, hipertensión, diabetes mellitus, asma, enfermedades asociadas, diabetes gestacionaria, preclancia, ciur, gestorragia, enfermedades propias y además se conoce la clase, que representa la categoría pre-término o a-término, a la que corresponde cada caso.

Para la solución del problema planteado se han seleccionado los siguientes métodos de selección de atributos.

Análisis factorial: es una de las técnicas empleadas para la selección de rasgos es el análisis factorial ya que es una técnica de reducción de dimensionalidad de los datos que sirve para encontrar grupos homogéneos de variables a partir de un

conjunto de variables. Los grupos homogéneos lo conforman las variables que correlacionan mucho entre si y procurando, inicialmente, que unos grupos sean independientes de otros. [51]

Métodos de evaluación y métodos de búsqueda: donde el método de evaluación (atribute evaluator) es la función que determina la calidad del conjunto de atributos para discriminar la clase. Podemos distinguir dos tipos de métodos de evaluación: [52]

- Uno de ellos son los que directamente utilizan un clasificador específico para medir la calidad del subconjunto de atributos a través de la tasa de error del clasificador. Se denomina métodos wrapper, ya que envuelven al clasificador para explorar la mejor selección de atributos que optimiza sus prestaciones. Necesitan un proceso completo de entrenamiento y 9 evaluaciones en cada caso de búsqueda, por eso son muy costosos.
- El otro tipo no utiliza este clasificador específico. Dentro de este tipo está el método CfsSubsetEval, que calcula la correlación de la clase con cada atributo, y elimina atributos que tienen una correlación muy alta como atributos redundantes.

En esta investigación se emplearán los siguientes métodos de evaluación: **CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, WrapperSubsetEval.**

El método de búsqueda (Search Method) es la forma de realizar la búsqueda de conjuntos. Si se quiere realizar la evaluación exhaustiva de todos los subconjuntos, aparece un problema combinatorio inabordable en cuanto crece el número de atributos. Para ello aparecen estas estrategias que permiten realizar la búsqueda de una forma más eficiente. [52]

- Uno de estos métodos, que se caracteriza por su rapidez, es el **ForwardSelection**. Se trata de un método de búsqueda sub-óptima en escalada. El procedimiento es el siguiente: se elige primero el mejor atributo, después añade el siguiente atributo que más aporta y continúa así hasta llegar a la situación en la que añadir un nuevo atributo empeora la situación. [52]

- Otro de este tipo de métodos sería el **BestSearch**, que permite buscar interacciones entre atributos más complejas. Su procedimiento es ir analizando lo que mejora y empeora un grupo de atributos al añadir elementos, con la posibilidad de hacer retrocesos para explorar con más detalle.
- Otro de estos métodos que podríamos utilizar es el **ExhaustiveSearch**, que enumera todas las posibilidades y las evalúa para seleccionar la mejor. [52]

También fueron seleccionados los siguientes métodos de clasificación supervisados:

Algoritmo J48 (C4.5): este algoritmo, forma parte también de los algoritmos basados en árboles de decisión, al igual que el algoritmo anterior. La característica fundamental de este algoritmo es que incorpora una poda del árbol de clasificación una vez que éste ha sido inducido, es decir, una vez construido el árbol de decisión, se podan aquellas ramas del árbol con menor capacidad predictiva. Este algoritmo es una mejora de ID3, también basado en árboles, donde el criterio escogido para seleccionar la variable más informativa está basado en el concepto de cantidad de información mutua entre dicha variable y la variable clase. [53]

Para construir el árbol, ID3 usa una aproximación descendente que da preferencia a los árboles pequeños sobre los grandes. El nodo raíz es seleccionado por encontrar el atributo más valioso en el conjunto de entrenamiento, el que mejor clasifica las instancias; la búsqueda se realiza por medio de una prueba estadística que mide qué tan bien un atributo separa el conjunto de entrenamiento de acuerdo a las clases. Una vez que la raíz es seleccionada, se agrega una rama desde la raíz para cada posible valor del atributo correspondiente, y el conjunto de entrenamiento es ordenado en los nodos apropiados, cada nodo contiene los ejemplos que cumplen la restricción de la rama anterior. Para seleccionar el atributo más valioso en cada punto del árbol, se repite el proceso entero usando el conjunto de entrenamiento asociado con el nodo. De manera que cuando una nueva instancia necesita ser clasificada, los atributos especificados por los nodos son evaluados iniciando por el nodo raíz, entonces de manera descendente se recorren las ramas del árbol que corresponden a los valores de los atributos en la instancia dada, el proceso se repite

hasta que una hoja es alcanzada, y es en este punto donde la etiqueta asociada a la hoja es asignada a la nueva instancia como su categoría. [48]

El perceptrón multicapa (MLP): es una red neuronal artificial (RNA) formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón (también llamado perceptrón simple). El perceptrón multicapa puede ser totalmente o localmente conectado. En el primer caso cada salida de una neurona de la capa "i" es entrada de todas las neuronas de la capa "i+1", mientras que en el segundo cada neurona de la capa "i" es entrada de una serie de neuronas (región) de la capa "i+1".

Las capas pueden clasificarse en tres tipos [54]:

- capa de entrada: constituida por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento.
- capas ocultas: formadas por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores.
- capa de salida: neuronas cuyos valores de salida se corresponden con las salidas de toda la red.

Este modelo se entrena mediante el algoritmo de retro-propagación (backpropagation) de errores o variantes, este soluciona el problema de entrenar los nodos de las capas ocultas y en la medida que se entrena la red los nodos de las capas intermedias se organizan de tal modo que los distintos nodos aprenden a reconocer distintas características del espacio total de entradas. [39]

El MLP es reconocido como la mejor red neuronal para solucionar un problema de clasificación supervisada a partir de ejemplos. [55]

k-Vecinos más cercanos (k-NN, por sus siglas en inglés): este algoritmo está basado en instancias, por ello consiste únicamente en almacenar los datos presentados. Cuando una nueva instancia es encontrada, un conjunto de instancias similares relacionadas es devuelto desde la memoria y usado para clasificar la instancia consultada. Se trata, por tanto, de un algoritmo del método *lazy learning*. Este método de aprendizaje se basa en que los módulos de clasificación mantienen en memoria una selección de ejemplos sin crear ningún tipo de abstracción en forma

de reglas o de árboles de decisión (de ahí su nombre, *lazy*, perezosos). Cada vez que una nueva instancia es encontrada, se calcula su relación con los ejemplos previamente guardados con el propósito de asignar un valor de la función objetivo para la nueva instancia. [53]

La idea básica sobre la que se fundamenta este algoritmo es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. [53]

Máquinas de soporte vectorial (*Support Vector Machine,* SVM), la que concretamente, fundamenta las decisiones de clasificación, no basadas en todo el conjunto de datos, sino en un número finito y reducido de casos, que constituyen los "vectores soporte" [56].

Una máquina de soporte vectorial aprende la superficie de decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un *kernel* Gaussiano u otro tipo de *kernel* a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento. [57]

Las SVM son básicamente clasificadores para 2 clases. El entrenamiento es relativamente fácil. No hay óptimo local, como en las redes neuronales. Se escalan relativamente bien para datos en espacios dimensionales altos. El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente. Datos no tradicionales como cadenas de caracteres y árboles pueden ser usados como entrada a la SVM, en vez de vectores de características. [57]

1.11 Conclusiones parciales

- El parto pre-término es un problema internacional que se incrementa y afecta también a Cuba y a la provincia de Cienfuegos. A pesar de investigaciones realizadas en diversos países no se conocen las causas, pero existen factores de riesgo que influyen en la prematuridad.
- En la medicina cada vez es más frecuente las investigaciones que emplean técnicas de IA, obteniéndose resultados favorables.
- Dentro de la IA, las técnicas de selección de atributos (análisis factorial, CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, WrapperSubsetEval) y los métodos de clasificación supervisados (IBK, J48, MLP y SMO) fueron los seleccionados para determinar los factores de riesgo y obtener el mejor modelo de clasificación respectivamente.

2 Capítulo II: Diseño del experimento.

En este capítulo se presenta la base de casos, describiendo las variables y la explicando la preparación de los datos. Se presentan las dos herramientas a emplear en la experimentación: SPSS y Weka, así como la conversión de los datos a los formatos establecidos por estos .sav y .arff respectivamente. Se abordan teóricamente los diferentes métodos de selección de atributos y clasificación supervisada utilizados durante la experimentación en cada herramienta. Por último se presenta la configuración de parámetros de cada una de las técnicas, con todas las variantes probadas, hasta determinar la mejor variante de cada método.

2.1 Base de casos, descripción de las variables y preparación de los datos.

Para la realización de este trabajo se recopila la información que conforma la base de casos, tomada de las historias clínicas de todas las pacientes cuyo parto tuvo lugar en el Hospital General Universitario-Docente “Dr. Gustavo Aldereguía Lima” de Cienfuegos, en el periodo del 1ro de enero de 2012 al 31 de diciembre de 2012.

Se conforman dos grupos:

- grupo estudio: pacientes en las que el parto ocurra antes de las 37 semanas.
- grupo control: pacientes que tengan su parto después de las 37 semanas.

La base de casos tiene un total de 687 pacientes. De cada caso se tienen 19 variables que constituyen todos aquellos factores considerados como riesgosos para el parto pre-término. Estas variables fueron elegidas según el criterio del equipo de especialistas en obstetricia trabajadores del hospital Gustavo Aldereguía Lima de Cienfuegos, quienes apoyados en su experiencia y diferentes estudios como los que ya se han mencionado anteriormente, determinaron un total de factores de riesgo que condicionan el parto pre-término pero necesitan saber cuáles de estos son los que mayor influencia tienen.

Las variables son descritas en la siguiente tabla:

Variable	Tipo	Operacionalización.

		Escala	Descripción
Edad materna	Ordinal	Menos de 19 años Entre 20 y 34 años Mayor de 35 años	Según años de edad.
Paridad	Nominal	Nulípara Multípara	Según el antecedente de partos previos
Abortos	Nominal	Espontáneo Provocado No abortos Espontáneos y provocados	Antecedentes de abortos previos
Parto pre-término anterior	Nominal	Sí No	Antecedentes de partos pre-término anteriores
Hábito de fumar	Nominal	Sí No	Según refiera el paciente
Ingestión de bebidas alcohólicas	Ordinal	Si Bebedor ligero Bebedor moderado Bebedor severo	Según refiera el paciente
Escolaridad	Ordinal	Sexto grado Noveno grado Doce grado Universitario	Según grado de escolaridad alcanzado
Embarazo múltiple	Nominal	Si No	Según la presencia de más de un feto en el embarazo actual
Infección vaginal	Nominal	Si No	Según la presencia de infección vaginal ya sea por leucorrea o exudado vaginal
Infección urinaria	Nominal	Si No	Según la presencia de infección del tracto urinario ya sea por síntomas clínicos o uro-

			cultivo
Hipertensión	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
Diabetes mellitus	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
Asma	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
Enfermedades asociadas	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
Diabetes gestacional	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
Preclancia	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
ciur	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
Gestorragia	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación
Enfermedades propias	Nominal	Sí No	Según el antecedente de enfermedades crónicas asociadas a la gestación

Tabla 1. Descripción de las variables de estudio.

Durante la preparación del conjunto de entrenamiento se realizó una limpieza en los datos modificándose así el número de instancias (mujeres paridas) que se tenía inicialmente por la presencia de valores perdidos. Los casos que tenían más de una variable con valores perdidos no se tomaron en cuenta quedando un total de 585

casos de los 587 casos originales. Del total de instancias resultantes 203 pertenecen a la clase pre-término y 382 a la clase a-término.

Frecuencias de las variables: en las siguientes figuras se muestra la distribución de los casos en cada una de las variables.

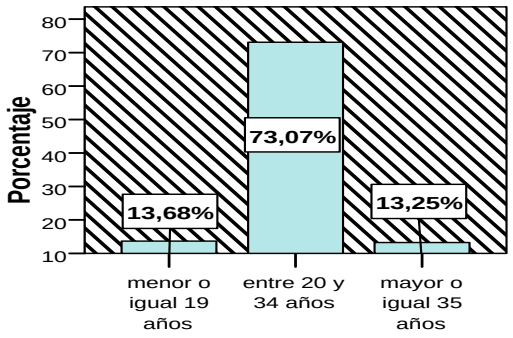


Figura 1. Edad materna

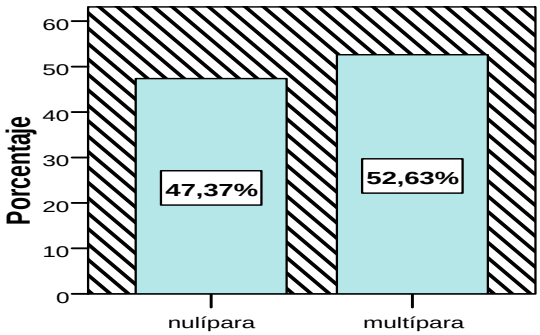


Figura 2. Paridad

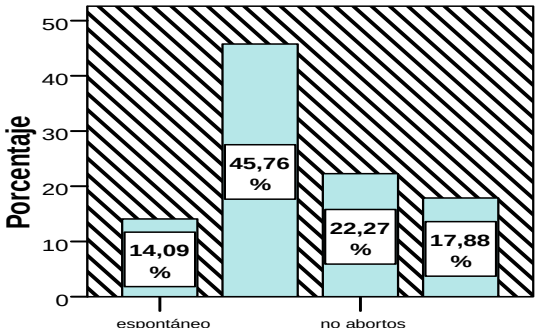


Figura 3. Abortos

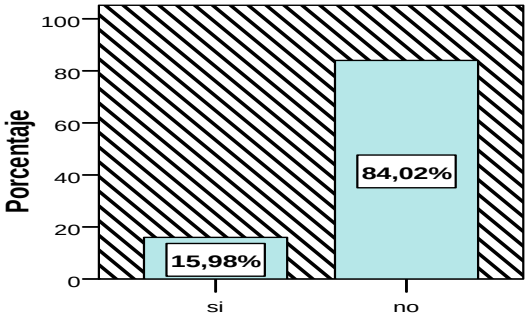


Figura 4. Parto pre-término anterior

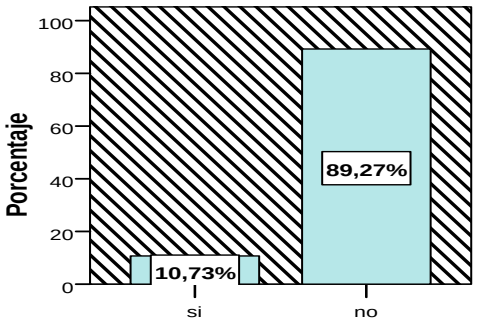


Figura 5. Hábito de fumar

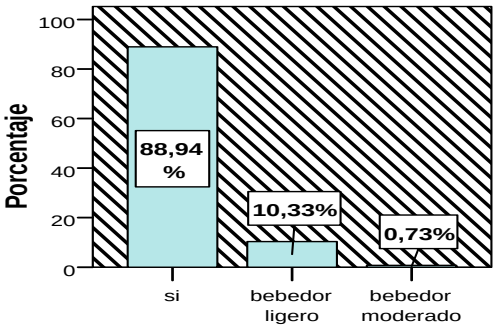


Figura 6. Ingestión de alcohol

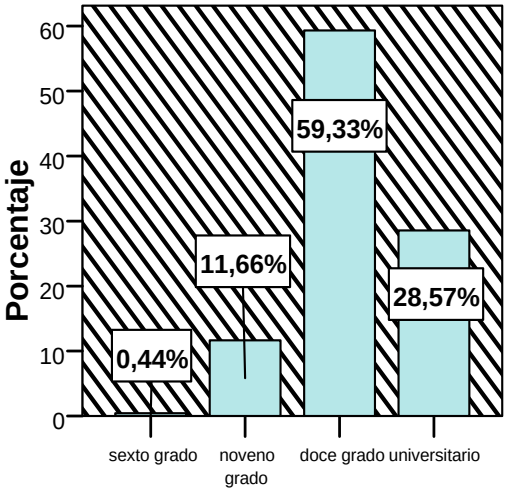


Figura 7. Escolaridad

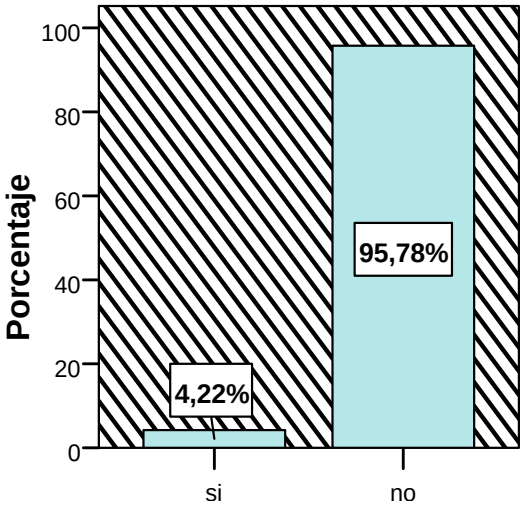


Figura 8. Embarazo múltiple

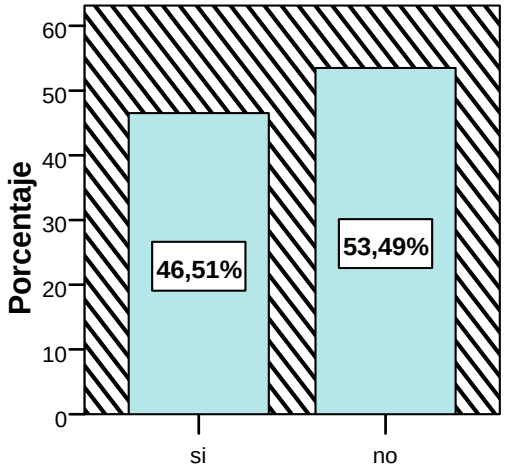


Figura 9. Infección vaginal

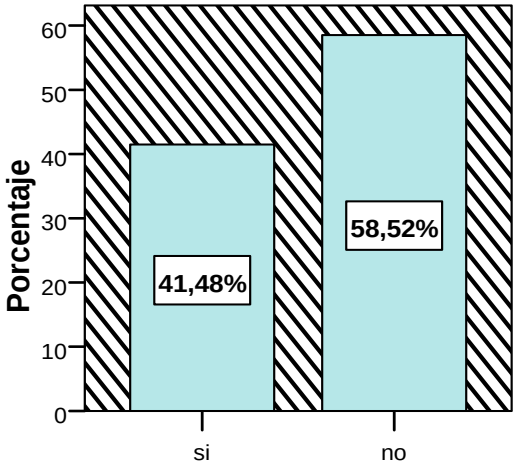


Figura 10. Infección urinaria

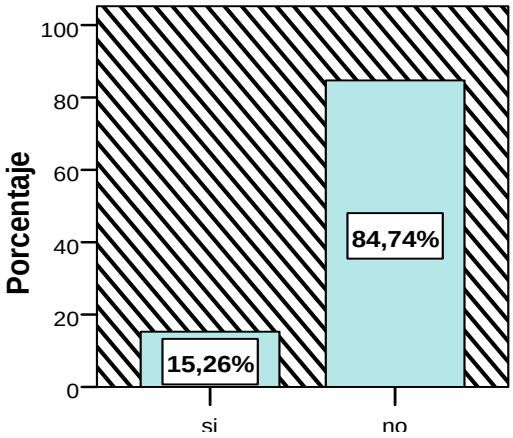


Figura 11. Hipertensión

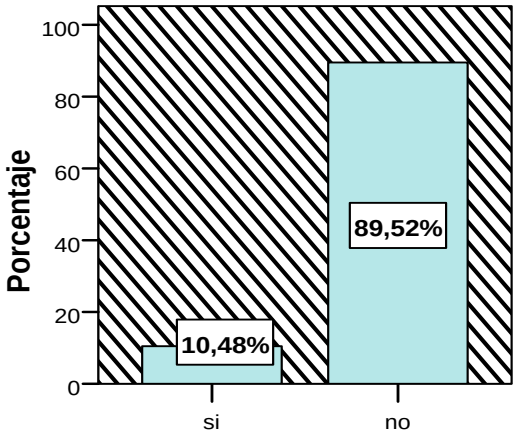


Figura 12. Diabetes mellitus

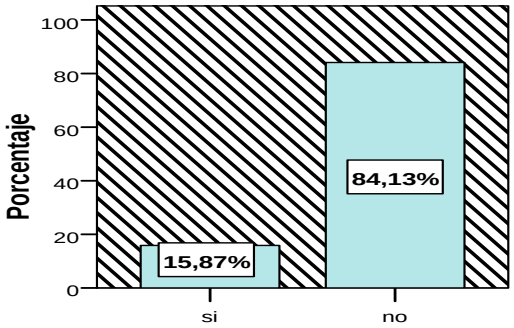


Figura 13. Asma

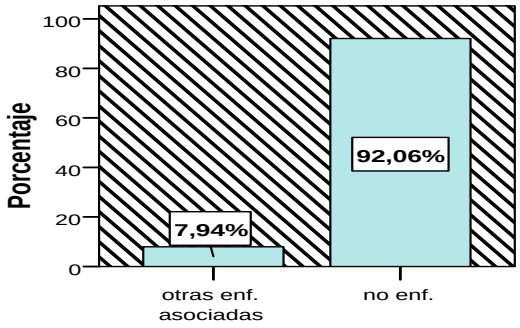


Figura 14. Enfermedades asociadas

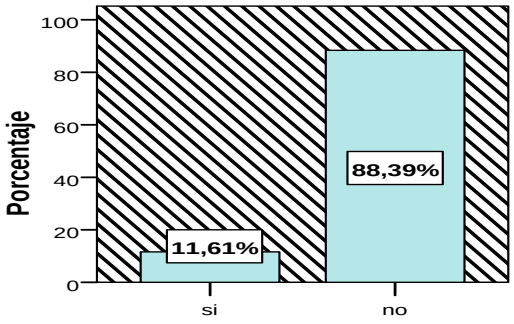


Figura 15. Diabetes gestacional

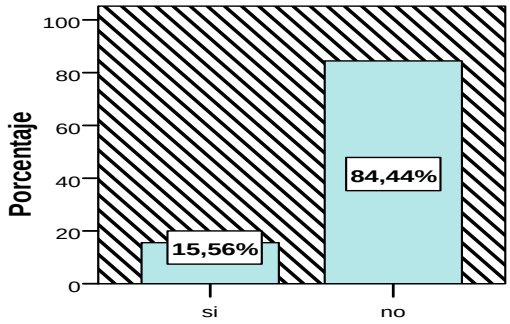


Figura 16. Preclancia

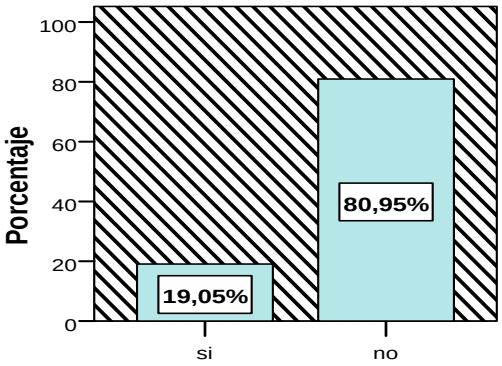


Figura 17. Ciur

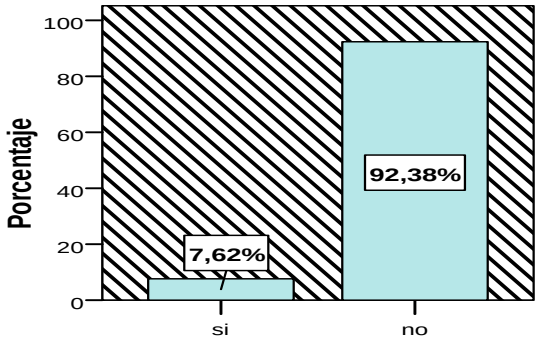


Figura 18. Gestorragia

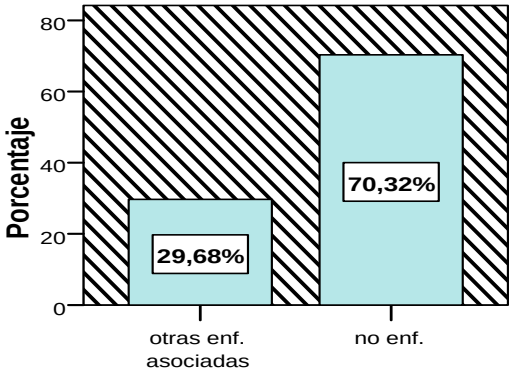


Figura 19. Enfermedades propias

2.2 Herramientas utilizadas en la experimentación

Partiendo de los métodos que fueron propuestos en el capítulo I se seleccionaron las herramientas que posibilitaban el uso de estos métodos para la solución del problema planteado. Estas herramientas son: SPSS versión 15.0 y Weka versión 3.7.5.

2.2.1 SPSS versión 15.0

Es un conjunto de potentes herramientas de tratamiento de datos y análisis estadístico. Este programa cuenta con ocho tipos de ventanas donde las dos principales son el editor de datos y el visor de resultados. Además dispone de un editor de tablas, uno de gráficos y uno de texto. Hace posible la gestión de datos y la obtención de informes. Permite trabajar con archivos de datos de diversos formatos. Las selecciones de menú y cuadros de dialogo permiten realizar análisis complejos sin necesidad de teclear ni una sola línea de sintaxis de comandos. [51]

Entre las muchas funcionalidades que provee la herramienta en esta investigación es usado el módulo de reducción de datos que permite la aplicación del análisis factorial, el cual tiene implementado varios métodos de extracción.

2.2.2 Weka versión 3.7.5

La plataforma Weka, es un paquete de software libre, desarrollado en el lenguaje de programación Java y se distribuye bajo los términos de la licencia GNU. Corre prácticamente sobre cualquier plataforma y ha sido probado en los sistemas operativos Linux, Windows y Macintosh (Witten y Frank, 2005). Weka incluye métodos para diferentes problemas de minería de datos: regresión, clasificación, agrupamiento y los archivos que utiliza como bases de casos tienen extensión “.arff”. [56]

En esta investigación se emplean las funcionalidades que ofrece el entorno explorador del programa con los siguientes paneles de trabajo:

- El panel "Preprocess"
- El panel "Classify"

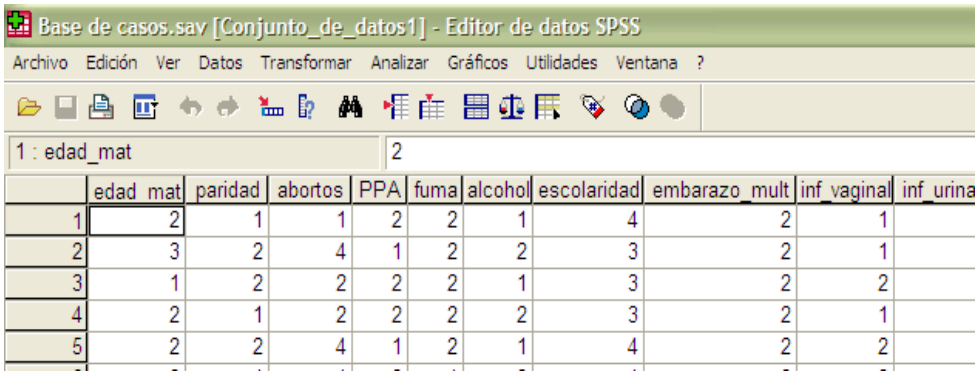
- El panel "Selected attributes"

2.3 Conversión de los datos a los formatos de SPSS (.sav) y Weka (.arff)

Partiendo de la descripción de la base de casos que aparece en el capítulo II epígrafe 2.1, se conforma el fichero .sav para el análisis correspondiente en el SPSS (Figura 2 y 3) y el fichero .arff para usar en la herramientas Weka (Figura 4).

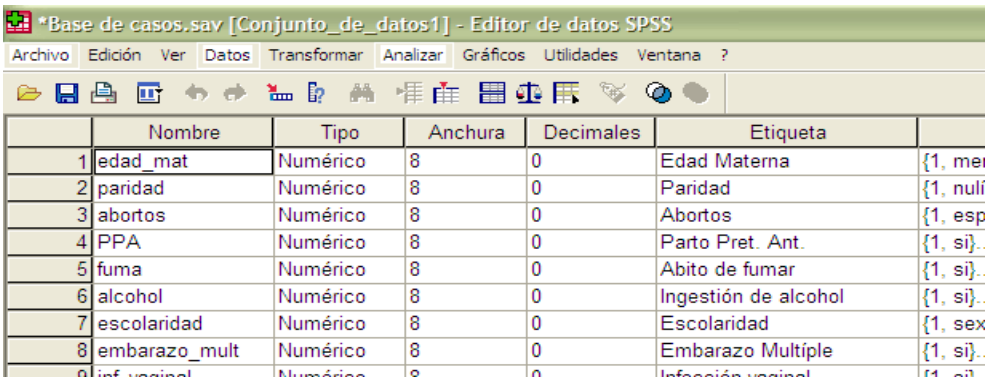
2.3.1 La base de casos en el SPSS.

El tipo de variable en el SPSS se define como: numérico, fecha, cadena y otros. Para definir una variable numérica es necesario asignarle uno de los siguientes niveles de medida: escala, ordinal o nominal.



	edad_mat	paridad	abortos	PPA	fuma	alcohol	escolaridad	embarazo_mult	inf_vaginal	inf_urina
1	2	1	1	2	2	1	4	2	1	
2	3	2	4	1	2	2	3	2	1	
3	1	2	2	2	2	1	3	2	2	
4	2	1	2	2	2	2	3	2	1	
5	2	2	4	1	2	1	4	2	2	

Figura 20. Vista de datos del SPSS



	Nombre	Tipo	Anchura	Decimales	Etiqueta	
1	edad_mat	Numérico	8	0	Edad Materna	{1, mer
2	paridad	Numérico	8	0	Paridad	{1, nuli
3	abortos	Numérico	8	0	Abortos	{1, esp
4	PPA	Numérico	8	0	Parto Pret. Ant.	{1, si}..
5	fuma	Numérico	8	0	Abito de fumar	{1, si}..
6	alcohol	Numérico	8	0	Ingestión de alcohol	{1, si}..
7	escolaridad	Numérico	8	0	Escolaridad	{1, sex
8	embarazo_mult	Numérico	8	0	Embarazo Multíple	{1, si}..

Figura 21. Vista de variables del SPSS

2.3.2 La base de casos en Weka

En la herramienta Weka los atributos pueden ser principalmente de dos tipos: numéricos de tipo real o entero (real o integer), y simbólicos (especificando los valores posibles que pueden tomar entre llaves).

```
@RELATION datos_mujeres_paridas

@ATTRIBUTE edad_mat {1,2,3}
@ATTRIBUTE paridad INTEGER [1,2]
@ATTRIBUTE abortos INTEGER [1,4]
@ATTRIBUTE ppa INTEGER [1,2]
@ATTRIBUTE fuma INTEGER [1,2]
@ATTRIBUTE alcohol {1,2,3,4}
@ATTRIBUTE escolaridad {1,2,3,4}
@ATTRIBUTE embarazo_mult INTEGER [1,2]
@ATTRIBUTE inf_vaginal INTEGER [1,2]
@ATTRIBUTE inf_urinaria INTEGER [1,2]
@ATTRIBUTE hipertension INTEGER [1,2]
@ATTRIBUTE diabetes_mell INTEGER [1,2]
@ATTRIBUTE asma INTEGER [1,2]
@ATTRIBUTE enf_asociadas INTEGER [1,2]
@ATTRIBUTE diabetes_gest INTEGER [1,2]
@ATTRIBUTE preclancia INTEGER [1,2]
@ATTRIBUTE ciur INTEGER [1,2]
@ATTRIBUTE gestorragia INTEGER [1,2]
@ATTRIBUTE enf_propias INTEGER [1,2]
@ATTRIBUTE class {1,2}
%
@DATA
%
% Instances (585):
%
2,1,1,2,2,1,4,2,1,1,2,2,2,2,2,2,2,2,2,1
2,1,2,2,2,2,3,2,1,2,2,2,2,2,2,2,1,2,2,1
1,1,3,2,1,2,2,2,1,1,1,2,2,2,1,1,2,2,2,1
```

Figura 22. Datos convertidos al formato de Weka(arff)

Al abrir la base de casos en la herramienta Weka se puede observar determinada información así como:

- nombre de cada atributo incluyendo la clase.
- total de instancias: 585
- total de atributos: 20 (incluyendo la clase)
- total de instancias por cada Clase: 203 perteneciente a la clase pre-término (color azul) y 382 perteneciente a la clase a-término.

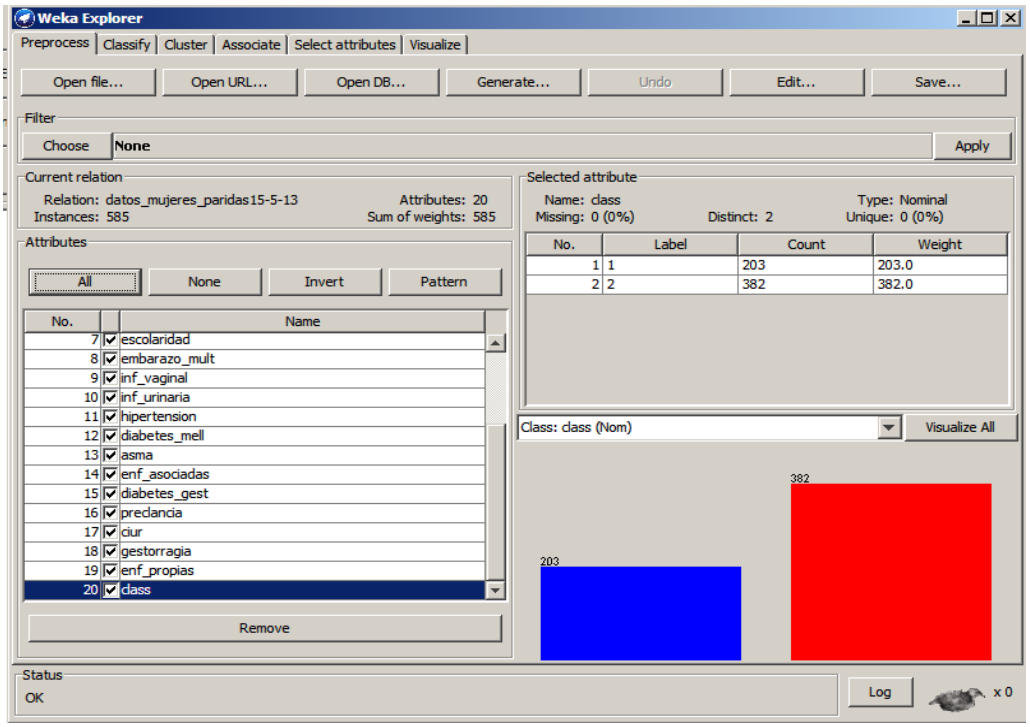


Figura 23. Interface explorador de Weka con la base de casos abierta

2.4 Selección de rasgos:

Con el fin de determinar los rasgos que influyen en el parto pre-término se aplicaron algunas de las técnicas de selección de atributos que brindan las herramientas SPSS y Weka. Esto permite establecer una comparación de los resultados que se obtengan en cada método usado.

2.4.1 Análisis factorial en SPSS

En esta investigación se aplica el análisis factorial comentado en el capítulo I epígrafe 1.10 usando los siguientes dos métodos de extracción:

1. **Componentes principales.** Método de extracción donde los factores obtenidos son los auto-vectores de la matriz de correlaciones re-escalados. [51]
2. **Factorización de ejes principales.** Método de estimación iterativo en el que, como estimación inicial de la comunalidad, la matriz de correlaciones original se reduce sustituyendo los unos de su diagonal por las estimaciones de la correlación múltiple al cuadrado entre cada variable y todas las demás. La matriz

reducida se auto-descompone y se corrigen las estimaciones iniciales de la comunalidad por las nuevas estimaciones resultantes. El proceso continúa hasta que no existe diferencia entre dos pasos sucesivos o se alcanza alguno de los criterios de parada. [51]

Para esto es necesario seleccionar primeramente las variables e introducir las especificaciones relacionadas con descriptivos, extracción, rotación, puntuaciones y opciones.

Descriptivos: permite establecer los estadístico descriptivos más relevantes en relación a las variables seleccionadas.
Extracción: permite especificar el método para la estimación del modelo factorial.
Rotación: permite seleccionar el método de rotación factorial.
Puntuaciones: permite especificar las opciones posibles en el cálculo y almacenamiento de las puntuaciones factoriales.

Para la aplicación del análisis factorial se realizaron las especificaciones que se muestran a continuación en la tabla 2:

Parámetros	Componentes principales	Factorización de ejes principales
Descriptivos	Solución inicial	Solución inicial
Extracción	Matriz de correlaciones Auto-valores mayores que 1 Solución factorial sin rotar Gráfico de sedimentación	Matriz de correlaciones Auto-valores mayores que 1 Solución factorial sin rotar Gráfico de sedimentación
Rotación	Varimax Solución rotada	Varimax Solución rotada
Puntuaciones	Guardar como variable Método: regresión	Guardar como variable Método: regresión

Tabla 2. Especificaciones para realizar el análisis de componentes.

Como resultados del análisis de componentes principales se conformaron los componentes mostrándose 8 por la condición de que solo se toman los auto-valores mayores que 1. Este resultado que coincide con el número de factores mostrados por el análisis usando factorización de ejes principales.

2.4.2 Métodos de evaluación y búsqueda en Weka:

A continuación se comenta brevemente la funcionalidad de los métodos de evaluación implementados en la herramienta.

CfsSubsetEval: evalúa el valor del conjunto de atributos teniendo en cuenta la capacidad de predicción individual de cada atributo con el grado de redundancia entre ellos. Es preferido por el método el subconjunto de atributos que está muy correlacionado con la clase, mientras que la inter-correlación entre ellos sea baja. Entre los parámetros que requiere se encuentran:

- locallyPredictive: que identifica localmente los atributos predictivos. Iterativamente adiciona atributos que tengan la más alta correlación con la clase.
- missingSeparate: trata el valor perdido como un valor separado.

GainRatioAttributeEval: evalúa el valor de un atributo para medir la ganancia (aumento) del radio con respecto a la clase. Requiere el parámetro:

- missingMerge: distribuye cuentas para los valores perdidos. Las cuentas son distribuidas a través de otros valores en proporción con su frecuencia.

InfoGainAttributeEval: evalúa el valor de un atributo midiendo la ganancia (obtención) de la información con relación a la clase. Los parámetros que requiere son:

- missingMerge

OneRAttributeEval: evalúa el valor de un atributo usando el clasificador OneR.

- evalUsingTrainingData: usa los datos de entrenamiento para evaluar los atributos en vez de la validación cruzada.
- folds: introduce el número de iteraciones para la validación cruzada.
- minimumBucketSize: mínimo número de objetos en un cubo.

- seed: introduce la semilla para el uso en la validación cruzada.

WrapperSubsetEval: evalúa sets de atributo usando un esquema de aprendizaje. Utiliza la validación cruzada para estimar la exactitud del esquema de aprendizaje para un conjunto de atributos.

- classifier: clasificador a usar para estimar la exactitud de subconjuntos.
- evaluationMeasure: la medida usada para evaluar las combinaciones de atributos.
- folds: numero de las iteraciones de “xval” a usar cuando sea estimada la exactitud del subconjunto.

Los métodos de búsqueda utilizados por los de evaluación son los siguientes:
BestFirst: Busca el espacio del subconjunto de atributos el “insaciable” hillclimbing incrementado con una facilidad de búsqueda de retroceso. Establece el número de nodos consecutivos no superados permitiendo controlar el nivel de búsqueda de retroceso hecho. Best-first puede comenzar con el conjunto de atributos vacío y buscar hacia delante, o con el conjunto de atributos lleno y buscar hacia detrás, o comenzar en un punto y buscar en ambas direcciones. Contiene los siguientes parámetros:

- direction: introduce la dirección de búsqueda.
- lookupCacheSize: introduce la longitud máxima de la búsqueda caché del subconjunto evaluado.
- searchTermination: fija la cantidad de retrocesos.
- startSet: introduce el punto de comienzo de la búsqueda.

Ranker: Ordena los atributos por sus evaluaciones individuales. Tiene los siguientes parámetros:

- generateRanking: si se selecciona, Ranker es solo capaz de generar atributos ordenados.
- numToSelect: especifica el número de atributos a conservar.
- startSet: introduce el punto de comienzo de la búsqueda.
- threshold: introduce el umbral por el cual los atributos pueden ser descartados.

GreedyStepwise: hace la búsqueda hacia delante o hacia atrás en el espacio del subconjunto de atributos. Puede comenzar con ningún o todos los atributos o por un punto arbitrario en el espacio. Se detiene cuando es saciado por cualquier atributo restante dando como resultado una disminución en la evaluación. Puede producir una lista ordenada por rango de atributos. Contiene los siguientes parámetros:

- generateRanking: seleccionar verdadero si es requerida una lista ordenada por rango.
- numToSelect: especifica el número de atributos a conservar.
- searchBackwards: búsqueda hacia atrás.
- startSet: introduce el punto de comienzo de la búsqueda
- threshold: introduce el umbral por el cual los atributos pueden ser descartados.

A continuación se muestran las distintas variantes aplicadas para la selección de rasgos empleando la herramienta Weka.

Método de evaluación	**Método de búsqueda**	**Parámetros del método de búsqueda**	**Atributos seleccionados**
CfsSubset	BestFirst	direction: hacia adelante lookupCacheSize: 1 searchTermination: 5	6
GainRatio Attribute	Ranker	generateRanking: Verdadero numToSelect: -1 threshold: -1.7976931348623157E308	12
InfoGain Attribute	Ranker	generateRanking: Verdadero numToSelect: -1 threshold: -1.7976931348623157E308	11

OneR Attribute	Ranker	generateRanking: Verdadero numToSelect: -1 threshold: -1.7976931348623157E308	Todos
Wrapper Subset	Greedy Stepwise	generateRanking: Falso numToSelect: -1 searchBackwards: Verdadero threshold: -1.7976931348623157E308	6

Tabla 3. Métodos de selección de atributos con los resultados obtenidos.

En los resultados obtenidos luego de realizar la selección de atributos en Weka se visualiza que los rasgos: parto pre-término anterior, alcohol, infección vaginal, preclancia y ciur fueron seleccionados por todos los métodos aplicados. Otras variables como: embarazo múltiple, infección urinaria fueron seleccionados por cuatro de los métodos de selección.

2.5 Clasificación supervisada.

Con el fin de obtener un modelo para predecir el parto pre-término se aplicaron técnicas de clasificación supervisada con las funcionalidades que provee la herramienta Weka. Los métodos utilizados están comentados en el capítulo I epígrafe 1.10.

IBK: es el clasificador k-vecinos más cercanos. Puede seleccionar el valor apropiado de K basado en la validación cruzada. Puede también hacer el coeficiente de distancia. Tiene además los siguientes parámetros:

- KNN: el número de vecinos más cercanos al caso que será clasificado.
- crossValidate: se usará para seleccionar el mejor valor de k.
- debug: si es verdadero, el clasificador puede mostrar información adicional en la consola.
- distanceWeighting: obtiene la distancia ponderando el método usado.
- meanSquared: promedio cuadrático.
- nearestNeighbourSearchAlgorithm: el vecino más cercano busca el algoritmo a usar.

J48: clasificador para generar un árbol de decisiones C4.5 podado o sin podar. Tiene como parámetros los siguientes:

- binarySplits: en la construcción de los árboles si se utiliza "*splits"* binario se dividen los atributos nominales.
- confidenceFactor: el factor de confianza utilizado para la poda.
- debug: si es verdadero, el clasificador puede emitir información adicional en la consola.
- minNumObj: el número mínimo de instancias por hoja.
- nunFolds: determina la cantidad de datos utilizados para la poda de reducción de error. Unas veces se utiliza para la poda, el resto para el cultivo del árbol.
- reducedErrorPruning: si se reduce el error de poda se utiliza en lugar de la poda C.4.5.
- saveInstanceData: si se guardan los datos del entrenamiento para la visualización
- seed: la semilla usada para aleatorizar los datos cuando se utiliza la reducción del error en la poda.
- subtreeRaising: si se considera el sub-árbol levantando la operación cuando la poda.
- unpruned: si la poda se lleva a cabo.
- useLaplace: si cuenta con menos hojas, se alisan sobre la base de Laplace.

MultilayerPerceptron: este clasificador usa propagación hacia detrás para clasificar instancias. Esta red puede ser construida manualmente, creada por un algoritmo o por ambos. La red puede ser monitoreada o modificada durante el tiempo de entrenamiento. Los parámetros que contiene son:

- GUI: permite que la aplicación dibuje la red neuronal.
- autoBuild: autoconstruye la red neuronal (cantidad de capas ocultas y cantidad de neuronas en cada una de estas).
- debug: muestra información adicional en consola.
- hiddenLayers: cantidad de neuronas por cada capa oculta que se adiciona a la red. Valores enteros separados por coma.
- learningRate: tasa de aprendizaje.

- momentum: impulso aplicado a los pesos durante la actualización.
- nominalToBinaryFilter: procesa las instancias con el filtro. Ayuda a mejorar el rendimiento de la red si hay atributos nominales en los datos.
- normalizeAttributes: normaliza los atributos. Ayuda a mejorar el rendimiento de la red. Para esto la clase no tiene que ser numérica, también normaliza atributos nominales (valores entre -1 y 1).
- normalizeNumericClass: normaliza la clase si esta es numérica. Ayuda a mejorar el rendimiento de la red. Normaliza la clase entre -1 y 1. Este es un proceso interno, los valores mostrados son dados en los rangos originales.
- randomSeed: inicializa un número aleatorio el cual es usado para ubicar los pesos iniciales en las conexiones entre los nodos, así como para barajar los datos de entrenamiento.
- reset: permite a la red neuronal reiniciarse con un bajo ritmo de entrenamiento. Si la red diverge de la respuesta automáticamente se reinicia la red con un bajo ritmo de entrenamiento y comienza el entrenamiento de nuevo. Esta opción es factible sólo si la interface gráfica no está activada.
- trainingTime: ritmo de entrenamiento.
- validationThreshold: umbral

SMO: implementación de John Platt del algoritmo de optimización secuencial mínima para entrenar un clasificador de soporte vectorial.

- buildLogisticModels: si coloca modelos logísticos en las salidas.
- debug: si se selecciona verdadero, el clasificador puede devolver información adicional in la consola.
- epsilon: utilizar el epsilon para redondear el error.
- filterType: determina cómo los datos serán transformados.
- kernel: determina el kernel (el vector de soporte) a usar.
- numFolds: el numero de iteraciones para la validación cruzada
- randomSeed: número aleatorio de la semilla para la validación cruzada.
- toleranceParameter: el parámetro de tolerancia (no debe ser cambiado).

2.5.1 Configuración de los parámetros del IBK.

Se realizaron 4 combinaciones de parámetros con el fin de obtener el mejor modelo de clasificación. Estas combinaciones se muestran a continuación en la tabla 4 y los resultados de cada una de las variantes aparecen en el Anexos a.

Parámetros	Variante 1	Variante 2	Variante 3	Variante 4
KNN	25	1	5	5
crossValidate	Verdadero	Falso	Verdadero	Verdadero
debug	Verdadero	Falso	Verdadero	Verdadero
distanceWeighting	1/distancia	No	1-distancia	1/distancia
meanSquared	Falso	Falso	Verdadero	Verdadero
nearestNeighbourSearchAlgorithm	LinearNNSearch	EuclideanDistance	MinkowskDistance	LinearNNSearch

Tabla 4. Combinaciones de los parámetros de IBK.

Según los resultados obtenidos la mejor variante probada de este modelo fue la variante tres por tener el mejor por ciento de clasificaciones correctas y el menor porciento de falsos a-términos.

2.5.2 Configuración de los parámetros del J48:

Se realizaron 4 combinaciones de parámetros con el fin de obtener el mejor modelo de clasificación. Estas combinaciones se muestran a continuación en la tabla 5 y los resultados de cada una de las variantes aparecen en el Anexos b.

Parámetros	Variante 1	Variante 2	Variante 3	Variante 4
binarySplits	Verdadero	Verdadero	Verdadero	Verdadero
confidenceFactor	0.25	0.25	0.29	0.01
debug	Falso	Verdadero	Verdadero	Verdadero
minNumObj	2	50	3	2
nunFolds	3	10	5	3
reducedErrorPruning	Falso	Verdadero	Verdadero	Verdadero
saveInstanceData	Falso	Falso	Falso	Verdadero
seed	1	0	0	0
subtreeRaising	Verdadero	Verdadero	Verdadero	Falso
unpruned	Falso	Falso	Falso	Falso
useLaplace	Falso	Verdadero	Verdadero	Verdadero

Tabla 5. Combinaciones de los parámetros de J48.

Según los resultados obtenidos la mejor variante probada de este modelo fue la variante cuatro por tener el mejor por ciento de clasificaciones correctas y el menor porciento de falsos a-términos.

2.5.3 Configuración de los parámetros del MLP:

Se realizaron 4 combinaciones de parámetros con el fin de obtener el mejor modelo de clasificación. Estas combinaciones se muestran a continuación en la tabla 6 y los resultados de cada una de las variantes aparecen en el Anexos c.

Parámetros	Variante 1	Variante 2	Variante 3	Variante 4
GUI	Falso	Falso	Falso	Falso
autoBuild	Verdadero	Verdadero	Verdadero	Verdadero
debug	Falso	Verdadero	Falso	Verdadero
hiddenLayers	2	2	1	2
learningRate	0.6	0.6	0.3	0.3
momentum	0.2	0.2	0.2	0.2
nominalToBinaryFilter	Verdadero	Verdadero	Verdadero	Verdadero
normalizeAttributes	Verdadero	Verdadero	Verdadero	Verdadero
normalizeNumericClass	Verdadero	Verdadero	Verdadero	Verdadero
reset	Falso	Falso	Verdadero	Falso
randomSeed	0	1	0	0
trainingTime	203	203	500	203
validationThreshold	200	50	20	200

Tabla 6. Combinaciones de los parámetros del MLP.

Según los resultados obtenidos la mejor variante probada de este modelo fue la variante cuatro por tener el mejor por ciento de clasificaciones correctas y el menor porciento de falsos a-términos.

2.5.4 Configuración de los parámetros del SMO:

Se realizaron 7 combinaciones de parámetros con el fin de obtener el mejor modelo de clasificación. Estas combinaciones se muestran a continuación en la tabla 7 y los resultados de cada una de las variantes aparecen en el Anexos d.

Parámetros	Variante 1	Variante 2	Variante 3	Variante 4
buildLogisticModels	Verdadero	Verdadero	Falso	Falso
debug	Falso	Verdadero	Verdadero	Falso
epsilon	1.01	1.0E-12	1.0E-12	1.0E-5
filterType	Normalizar	Normalizar	Normalizar	Normalizar
kernel	Normalized PolyKernel	Normalized PolyKernel	Normalized PolyKernel	Normalized PolyKernel
numFolds	10	-1	10	20
randomSeed	1	1	1	1
toleranceParameter	0.1	0.0010	0.0010	1.0E-4

Tabla 7. Combinaciones de los parámetros del algoritmo SMO.

Según los resultados obtenidos la mejor variante probada de este modelo fue la variante dos por tener el mejor por ciento de clasificaciones correctas y el menor porciento de falsos a-términos.

2.6 Conclusiones parciales.

- La selección de atributos se realizó en SPSS y WEKA:

- El análisis factorial en SPSS por los métodos: análisis de ejes principales y componentes principales resultaron 8 factores y 8 componentes respectivamente.
- La selección de atributos en Weka por el método CfsSubsetEval obtuvo 6 rasgos, por el GainRatioAttributeEval obtuvo 12, por el InfoGainAttributeEval obtuvo 11 y por el WrapperSubsetEval obtuvo 6.

- La clasificación supervisada realizada en WEKA consistió en experimentar con 4 variantes por cada uno de los métodos: IBK, J48, MLP, y SMO. Obteniéndose como las mejores: la variante 3 del IBK, la variante 4 del J48, la variante 4 del MLP, y la variante 2 del SMO.

3 Capítulo III: Análisis de los resultados

En este capítulo se presenta el análisis de los resultados obtenidos en la selección de rasgos y la clasificación supervisada. Se explican los criterios a seguir en la selección de los rasgos influyentes y el mejor modelo para la predicción del parto pre-término. Por último se comparan las mejores variantes obtenidas de cada modelo de clasificación, proponiéndose el mejor y se concluye cuáles son los rasgos más influyentes en la prematuridad en la provincia de Cienfuegos.

3.1 Métodos de evaluación del clasificador

El resultado de aplicar un algoritmo de clasificación se evalúa comparando la clase predicha con la clase real de las instancias.

Existen diversos modos de realizar la evaluación entre ellos se encuentran:

- **use training set**: evaluación del clasificador sobre el mismo conjunto sobre el que se construye el modelo predictivo para determinar el error, que en este caso se denomina “error de resustitución”.
- **supplied test set**: esta opción evalúa sobre un conjunto independiente. Permite cargar un conjunto nuevo de datos. Sobre cada dato se puede realizar una predicción de clase para contar los errores.
- **cross-Validation**: evaluación con validación 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, y el resto como datos de entrenamiento para construir el modelo. Los errores calculados serán el promedio de todas las ejecuciones.
- **percentage split**: se dividen los datos en dos grupos, de acuerdo con el porcentaje indicado. El valor indicado es el porcentaje de instancias para construir el modelo, que seguidamente es evaluado sobre las que se han dejado aparte.

En esta investigación se emplea como método de evaluación del clasificador a cross validation con el parámetro “Folds” igual a 10

3.2 Criterios de selección

Para la selección de las variables más influyentes en la prematuridad son valorados los siguientes criterios:

- Selección según los métodos: se especifican como rasgos más influyentes en la prematuridad las variables que fueron seleccionadas por más de un método y se establece un orden entre ellas.
- Selección según criterio de experto: se analizan con el especialista en obstetricia los resultados obtenidos con el criterio anterior para concluir cuáles factores influyen en mayor medida en la prematuridad.

Para la selección del mejor modelo de clasificación se tienen en cuenta los siguientes indicadores:

- por ciento de clasificaciones correctas: seleccionar el modelo que contenga el mayor por ciento de instancias bien clasificadas.
- número de falsos negativos: seleccionar el modelo con el menor número de instancias que fueron clasificadas como a-término cuando en realidad pertenecen a la clase pre-termino.

3.3 Resultados obtenidos en el análisis factorial en el SPSS.

Como resultado de la aplicación del análisis factorial se obtuvo la siguiente información:

Comunalidades: contiene las comunalidades asignadas inicialmente a las variables (inicial) y las comunalidades reproducidas por la solución factorial (extracción). La comunalidad de una variable es la proporción de su varianza que puede ser explicada por el modelo factorial obtenido.

Porcentaje de varianza explicada: se ofrece un listado de auto-valores de la matriz varianzas-covarianzas y del porcentaje de varianza que representa cada uno de ellos. Los auto-valores expresan la cantidad de la varianza total que está explicada por cada factor; y los porcentajes de varianza explicada asociados a cada factor se

obtienen dividiendo su correspondiente auto-valor por la suma de los auto-valores (la cual coincide con el número de variables).

Solución factorial: contiene las correlaciones entre las variables originales y cada uno de los factores.

3.3.1 Método componentes principales.

En la tabla se muestran las comunalidades obtenidas usando como método de extracción: componentes principales.

	Inicial	**Extracción**
Edad materna	1,000	,649
Paridad	1,000	,752
Abortos	1,000	,490
Parto pret. ant.	1,000	,766
Hábito de fumar	1,000	,598
Ingestión de alcohol	1,000	,630
Escolaridad	1,000	,704
Embarazo múltiple	1,000	,684
Infección vaginal	1,000	,507
Infección urinaria	1,000	,188
Hipertensión	1,000	,679
Diabetes mellitus	1,000	,631
Asma	1,000	,611
Enfermedades asociadas	1,000	,519
Diabetes gestacional	1,000	,531
Preclancia	1,000	,590
Ciur	1,000	,644
Gestorragia	1,000	,785
Enf. propias	1,000	,733

Tabla 8. Comunalidades iniciales y finales

Se puede observar que las variables abortos, infección urinaria y enfermedades asociadas son las que peor quedan explicadas. Todas las demás quedan explicadas por el modelo en más de un 50%.

	Auto-valores iniciales			**Sumas de las saturaciones al cuadrado de la extracción**		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,232	11,748	11,748	2,232	11,748	11,748
2	1,843	9,698	21,446	1,843	9,698	21,446
3	1,594	8,392	29,837	1,594	8,392	29,837
4	1,437	7,562	37,399	1,437	7,562	37,399
5	1,274	6,706	44,105	1,274	6,706	44,105
6	1,190	6,262	50,367	1,190	6,262	50,367
7	1,104	5,809	56,176	1,104	5,809	56,176
8	1,016	5,348	61,524	1,016	5,348	61,524
9	,982	5,170	66,694			
10	,929	4,891	71,584			
11	,880	4,634	76,219			
12	,801	4,214	80,432			
13	,730	3,840	84,272			
14	,664	3,492	87,764			
15	,582	3,063	90,827			
16	,522	2,750	93,577			
17	,485	2,553	96,130			
18	,375	1,976	98,107			
19	,360	1,893	100,000			

Tabla 9. Varianza total explicada

Se puede observar que las ocho primeras componentes son las que más peso tienen a la hora de explicar los datos. En total explican un 61,52%.

La siguiente tabla muestra la correlación que existe entre las variables por cada componente.

	Componente							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Edad materna	-,463	,209	,219	,533	,176	,165	-,024	,023
Paridad	-,558	-,246	,406	-,144	-,250	,007	-,244	,269
Abortos	-,143	-,524	,217	-,101	,031	,259	,202	-,172
Parto pret. ant.	,613	,406	-,336	,161	,125	-3,10E-005	,048	-,262
Hábito de fumar	-,149	,713	,165	-,044	-,088	-,081	-,074	,135
Ingestión de alcohol	,227	-,664	-,211	,143	,241	,034	-,068	,093
Escolaridad	-,059	,320	,175	,523	-,058	,512	,163	-,041
Embarazo múltiple	-,112	,009	,196	-,215	,573	,412	,257	-,152
Infección vaginal	-,078	,299	,098	-,307	,250	,289	-,185	,356
Infección urinaria	-,004	,198	-,222	-,134	,243	,035	-,106	-,101
Hipertensión	,551	,106	-,120	-,187	-,416	,227	,112	,279
Diabetes mellitus	,381	-,094	,057	-,254	-,073	,375	,387	,337
Asma	-,199	,106	,260	-,109	-,194	-,304	,558	-,198
Enfermedades asociadas	-,078	,115	-,203	-,143	,491	-,212	-,074	,382
Diabetes gestacional	,022	-,172	-,265	,442	-,266	,283	-,252	,147

Preclancia	,465	-,151	,432	,176	,288	-,105	-,181	-,085
Ciur	,466	,121	,600	-,146	-,054	-,053	-,155	-,030
Gestorragia	-,061	-,043	-,085	,408	,154	-,334	,499	,470
Enf. propias	-,531	,077	-,522	-,309	-,087	,192	,059	-,171

Tabla 10. Matriz de componentes

Las variables que describen con mayor por ciento cada componente, se muestran en la siguiente tabla:

Componentes	Variables más relacionadas
1	Parto pre-término anterior
2	Hábito de fumar
3	Ciur
4	Edad materna
5	Embarazo múltiple
6	Abortos
7	Asma
8	Infección vaginal

Tabla 11. Componentes resultantes de análisis factorial con componentes principales

3.3.2 Método factorización del eje principal.

En la tabla se muestran las comunalidades obtenidas usando como método de extracción: factorización del eje principal.

	Inicial	Extracción
Edad materna	,255	,382

Paridad	,373	,495
Abortos	,150	,205
Parto pret. ant.	,418	,776
Hábito de fumar	,280	,458
Ingestión de alcohol	,297	,494
Escolaridad	,191	,566
Embarazo múltiple	,114	,438
Infección vaginal	,079	,129
Infección urinaria	,047	,035
Hipertensión	,243	,666
Diabetes mellitus	,136	,175
Asma	,080	,225
Enfermedades asociadas	,048	,100
Diabetes gestacional	,127	,131
Preclancia	,251	,280
Ciur	,288	,401
Gestorragia	,113	,389
Enf. propias	,356	,767

Tabla 12. Comunalidades iniciales y finales

Se puede observar que las variables parto pre-término anterior, escolaridad, hipertensión y enfermedades propias son las que mejor quedan explicadas por el modelo en más de un 50%.

	Autovalores iniciales			**Sumas de las saturaciones al cuadrado de la extracción**		
Factor	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,232	11,748	11,748	1,789	9,414	9,414
2	1,843	9,698	21,446	1,295	6,816	16,230
3	1,594	8,392	29,837	1,100	5,789	22,019

4	1,437	7,562	37,399	,870	4,581	26,600
5	1,274	6,706	44,105	,682	3,587	30,187
6	1,190	6,262	50,367	,609	3,206	33,393
7	1,104	5,809	56,176	,436	2,297	35,689
8	1,016	5,348	61,524	,331	1,743	37,432
9	,982	5,170	66,694			
10	,929	4,891	71,584			
11	,880	4,634	76,219			
12	,801	4,214	80,432			
13	,730	3,840	84,272			
14	,664	3,492	87,764			
15	,582	3,063	90,827			
16	,522	2,750	93,577			
17	,485	2,553	96,130			
18	,375	1,976	98,107			
19	,360	1,893	100,000			

Tabla 13: Varianza total explicada

Se puede observar que los ocho primeros factores son los que más peso tienen a la hora de explicar los datos. En total explican un 61,52%.

La siguiente tabla muestra la correlación que existe entre las variables por cada factor.

	Factor							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Edad materna	-,348	,215	,253	,364	,070	,080	-,057	,055
Paridad	-,516	-,202	,293	-,218	,179	-,044	-,107	,095
Abortos	-,141	-,344	,066	-,026	-,002	,175	,038	-,172
Parto pret. ant.	,666	,395	-,272	,226	-,149	,001	-,130	-,112

Hábito de fumar	-,075	,583	,212	-,148	-,052	-,149	,034	,142
Ingestión de alcohol	,151	-,579	-,200	,263	-,016	,074	-,042	,139
Escolaridad	-,012	,324	,262	,367	,346	,367	-,022	-,048
Embarazo múltiple	-,100	,006	,098	-,021	-,378	,483	,201	,030
Infección vaginal	-,043	,170	,057	-,146	-,121	,122	,049	,204
Infección urinaria	,009	,120	-,107	-,003	-,074	,017	-,024	,054
Hipertensión	,521	,044	-,133	-,339	,449	,124	,181	,097
Diabetes mellitus	,254	-,094	-,018	-,157	,079	,206	,167	,017
Asma	-,136	,079	,138	-,106	-,004	-,101	,186	-,354
Enfermedades asociadas	-,046	,068	-,097	,011	-,191	-,039	,086	,195
Diabetes gestacional	,020	-,072	-,106	,201	,243	-,015	-,095	,072
Preclancia	,339	-,187	,297	,069	-,167	,011	-,093	-,005
Ciur	,378	-,015	,421	-,243	-,047	,038	-,122	-,053
Gestorragia	-,035	-,028	,021	,332	,042	-,240	,465	,030
Enf. propias	-,530	,217	-,609	-,158	,033	,175	-,069	-,086

Tabla 14. Matriz factorial

Las variables que describen con el por ciento más elevado cada factor, se muestran en la siguiente tabla:

Factores	Variables más relacionadas
1	Parto pre-término anterior
2	Hábito de fumar
3	Ciur

4	Edad materna
5	Diabetes gestacional
6	Embarazo múltiple
7	Gestorragia
8	Infección vaginal

Tabla 15. Factores resultantes de análisis factorial con factorización del eje principal.

3.4 Resultados obtenidos en la selección de atributos en Weka.

A continuación se muestran las variables seleccionadas por los métodos aplicados en Weka:

CfsSubsetEval	**GainRatio AttributeEval**	**InfoGain AttributeEval**	**Wrapper SubsetEval**
1. ppa	1. emb_múltiple	1. preclancia	1. ppa
2. emb_múltiple	2. preclancia	2. ppa	2. alcohol
3. alcohol	3. ppa	3. ciur	3. inf_vaginal
4. inf_vaginal	4. ciur	4. inf_urinaria	4. preclancia
5. inf_urinaria	5. inf_urinaria	5. emb_múltiple	5. ciur
6. preclancia	6. alcohol	6. inf_vaginal	6. enf_propias
7. ciur	7. inf_vaginal	7. alcohol	
	8. fuma	8. edad_mat	
	9. abortos	9. fuma	
	10. edad_mat	10. escolaridad	
	11. escolaridad	11. abortos	
	12. enf_asociadas		

Tabla 16. Rasgos seleccionados por los distintos métodos aplicados en Weka

3.5 Resultados obtenidos en la aplicación de las técnicas de clasificación

Luego de crearse los modelos y validar los mismos el programa devuelve información detallada para cada uno de los modelos construidos. Entre estos aspectos se encuentran los siguientes:

Estadístico kappa: es un índice que compara el nivel de coincidencia entre varios expertos con el nivel de coincidencia que se podría dar por casualidad.

En realidad, puede verse como el nivel de coincidencia entre los expertos, al cual se le resta el que se da por casualidad, para tener una medida totalmente indicativa. Valores de +1 indican total coincidencia (el valor ideal), valores de 0 indican no más coincidencia de la que puede esperarse por casualidad y valores de -1 total desacuerdo.

Medidas TP y FP: la medida TP hace referencia a true positive (observaciones que el modelo clasificó correctamente para cada clase), y FP a false positives (observaciones que el modelo clasificó como de esa clase cuando en realidad no lo eran).

Precisión (Precision): se refere a la fracción de ejemplares que se han clasificado como de la clase correspondiente y que, en realidad, son de esa clase. Por tanto:

Precisión = TP / (TP + FP´)

Sensibilidad (recall): se refiere a la fracción de ejemplos de la clase de todo el conjunto que se clasifican correctamente.

Recall = TP / (TP + FN`)

Combinación (f - measure): combinación de ambas:

f - measure = (2 × Precision × Recall) / (Precision + Recall`)

Matriz de confusión: es la última información que aparece. En esta, se muestra en forma de matriz, en qué han consistido, concretamente, las equivocaciones del clasificador ya que en m [i, j] se tiene el número de ejemplares que son de la clase i y

se han clasificado como de la clase j. Por tanto, es de esperar que la diagonal principal contenga enteros altos y el resto de celdas números próximos a cero o cero.

3.5.1 MLP:

Correctly Classified Instances 467 79.8291 %

Incorrectly Classified Instances 118 20.1709 %

Kappa statistic 0.54

Clase 1:

TP Rate 0.64

FP Rate 0.118

Precision Recall 0.743

F-Measure 0.64

Clase 2:

TP Rate 0.882

FP Rate 0.36

Precision Recall 0.822

F-Measure 0.882

Matriz de confusión:

a b <-- classified as

130 73 | a = 1

45 337 | b = 2

En la matriz de confusión la letra “a” representa la clase pre-término y la letra “b” la clase a-término.

La matriz de confusión muestra dónde se producen los errores. Se visualiza que el clasificador asigna 73 instancias erróneamente a la clase “b” y 45

instancias a la clase “a”. Por tanto la clase pre-término es la que peor clasifica el algoritmo MLP.

3.5.2 IBK:

Correctly Classified Instances 466 79.6581 %

Incorrectly Classified Instances 119 20.3419 %

Kappa statistic 0.5356

Clase 1:

TP Rate 0.635

FP Rate 0.118

Precision Recall 0.741

F-Measure 0.635

Clase 2:

TP Rate 0.882

FP Rate 0.365

Precision Recall 0.82

F-Measure 0.882

Matriz de confusión:

a b <-- classified as

129 74 | a = 1

45 337 | b = 2

Se visualiza que el clasificador asigna 74 instancias erróneamente a la clase “b” y 45 instancias a la clase “a”. Por tanto la clase pre-término es la que peor clasifica el algoritmo IBK

3.5.3 J48:

Correctly Classified Instances 445 76.0684 %

Incorrectly Classified Instances 140 23.9316 %

Kappa statistic 0.4569

Clase 1:

TP Rate 0.596

FP Rate 0.152

Precision Recall 0.676

F-Measure 0.596

Clase 2:

TP Rate 0.848

FP Rate 0.404

Precision Recall 0.798

F-Measure 0.848

Matriz de confusión:

a b <-- classified as

121 82 | a = 1

58 324 | b = 2

Se visualiza que el clasificador asigna 82 instancias erróneamente a la clase “b” y 58 instancias a la clase “a”. Por tanto la clase pre-término es la que peor clasifica el algoritmo J48

3.5.4 SMO:

Correctly Classified Instances 474 81.0256 %

Incorrectly Classified Instances 111 18.9744 %

Kappa statistic 0.5729

Clase 1:

TP Rate 0.685

FP Rate 0.123

Precision Recall 0.747

F-Measure 0.685

Clase 2:

TP Rate 0.877

FP Rate 0.315

Precision Recall 0.84

F-Measure 0.877

Matriz de confusión:

a b <-- classified as

139 64 | a = 1

47 335 | b = 2

Se visualiza que el clasificador asigna 64 instancias erróneamente a la clase “b” y 47 instancias a la clase “a”. Por tanto la clase pre-término es la que peor clasifica el algoritmo SMO

3.6 Rasgos influyentes en la prematuridad.

Es preciso señalar que tres del total de rasgos fueron seleccionados por todos los métodos empleados. Estor son: parto pre-término anterior, ciur e infección vaginal.

Según el primer criterio para la selección de variables presentado en el epígrafe 3.2 de este capítulo se muestran a continuación los rasgos más influyentes (por orden de importancia) resultantes de todos los métodos de selección (análisis factorial, CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, y por el WrapperSubsetEval). No se tiene en cuenta el resultado obtenido por el método GainRatioAttributeEval porque seleccionó todos los rasgos.

Los rasgos que los métodos seleccionaron como más influyentes en la prematuridad son: parto pre-término anterior, ciur, infección vaginal, fuma, edad materna,

embarazo múltiple, escolaridad, preclancia, alcohol, abortos, gestorragia, diabetes gestacional.

El especialista obstétrico concluyó que de los resultados obtenidos con el criterio anterior, la selección que mejor concuerda con la teoría y experiencia médicas es la resultante del análisis factorial por el método factorización de ejes principales. Determinando como rasgos más influyentes en la prematuridad: parto pre-termino anterior, hábito de fumar, ciur, edad materna, diabetes gestacional, embarazo, múltiple, gestorragia e infección vaginal.

3.7 Propuesta del mejor modelo de clasificación supervisado para la predicción de parto pre-término.

La selección del mejor modelo de clasificación supervisada está dada por los criterios para la selección presentados en el epígrafe 3.2 de este capítulo, es decir por el mayor por ciento de clasificaciones correctas y el menor por ciento de falsos negativos (las instancias de la clase pre-término que clasificó erróneamente como a-término).

Modelo	% Clasificaciones correctas	Falsos negativos
MLP	79.8 %	12.5 %
IBK	79.7 %	12.7 %
J48	76.1 %	14.0 %
SMO	**81.0 %**	**10.9 %**

Tabla 17. Resultados de los mejores modelos.

El mayor porciento de clasificaciones correctas y el menor porciento de falsos negativos son criterios fundamentales, ya que el modelo resultante debe tener buena precisión al predecir el parto pre-término en nuevos casos. Como muestra la tabla anterior el modelo SMO es el mejor para la predicción del parto pre-término en nuevos casos. Esto se evidencia por tener el mayor por ciento de clasificaciones correctas y el menor por ciento de falsos negativos.

3.8 Conclusiones parciales

- Los rasgos influyentes en la prematuridad, según la experimentación y el criterio de experto, son (en orden de importancia): ppanterior, fuma, ciur, edad materna, diabetes gestacional, embarazo múltiple, gestorragia, infección vaginal.
- El modelo seleccionado para predecir el parto pre-término en el hospital provincial de Cienfuegos (según el por ciento de clasificaciones correctas y el número de falsos negativos) es el SMO, con los siguientes parámetros:
  - buildLogisticModels: Verdadero
  - debug: Verdadero
  - Epsilon: 1.0E-12
  - filterType: Normalizar
  - Kernel: NormalizedPolyKernel
  - numFolds: -1
  - randomSeed: 1
  - toleranceParameter: 0.0010

Conclusiones

- El parto pre-término es un problema internacional que se incrementa y afecta también a Cuba y a Cienfuegos, resultando apropiado utilizar técnicas de IA como: selección de atributos y clasificación supervisada para determinar los factores de riesgo influyentes en la prematuridad y el mejor modelo para predecir el parto pre-término respectivamente.
- La selección de atributos se realizó con los métodos: análisis de ejes principales y componentes principales (SPSS), resultando de ambos 8 variables. Además se experimentó en Weka, por el método CfsSubsetEval obtuvo 6 rasgos, por el GainRatioAttributeEval obtuvo 12, por el InfoGainAttributeEval obtuvo 11 y por el WrapperSubsetEval obtuvo 6.
- La clasificación supervisada realizada en WEKA consistió en experimentar con 4 variantes por cada uno de los métodos: IBK, J48, MLP, y SMO. Obteniéndose como las mejores: la variante 3 del IBK, la variante 4 del J48, la variante 4 del MLP, y la variante 2 del SMO.
- Los rasgos influyentes en la prematuridad, según la experimentación y el criterio del experto, son (en orden de importancia): ppanterior, fuma, ciur, edad materna, diabetes gestacional, embarazo múltiple, gestorragia, infección vaginal.
- El modelo que se propone para predecir el parto pre-término en el hospital provincial de Cienfuegos, con 81.0% de clasificaciones correctas y 31.5% de falsos negativos es el SMO, con los siguientes parámetros: buildLogisticModels: verdadero; debug: verdadero; Epsilon: 1.0E-12; filterType: normalizar; Kernel: NormalizedPolyKernel; numFolds: -1; randomSeed: 1 y toleranceParameter: 0.0010.

Recomendaciones

1. Ampliar la base de casos, incluir datos de diferentes años con el objetivo de mejorar la clasificación y realizar otros análisis interesantes.

2. Experimentar con métodos multiclasificadores o metaclasificadores para valorar los resultados de la clasificación obtenida por estos en la misma base de casos.

3. Implementar los resultados obtenidos de esta investigación para obtener un sistema informático portable, que pueda ser explotado por los especialistas en obstetricia en el diagnóstico del parto pre-término.

Glosario

Asma: es una enfermedad crónica del sistema respiratorio.

Cérvix: es la parte inferior del útero, situada en el fondo de la vagina, flexible, delgada y de unos tres centímetros de longitud.

Ciur: es un término médico que describe el retraso del crecimiento del feto.

Decidua: es la interface materna del embrión y participa en el intercambio de gases, nutrientes y productos de desecho.

Diabetes mellitus: es un grupo de trastornos metabólicos, que afecta a diferentes órganos y tejidos, dura toda la vida y se caracteriza por un aumento de los niveles de glucosa en la sangre.

Diabetes mellitus gestacional: es una forma de diabetes mellitus inducida por el embarazo.

Estrógenos: son hormonas sexuales esteroideas de tipo femenino principalmente, producidos por los ovarios y, en menores cantidades, por las glándulas adrenales.

Hipertensión arterial: es una condición médica caracterizada por un incremento continuo de las cifras de presión arterial.

Miometrio: es la capa muscular intermedia, entre la serosa peritoneal y la mucosa glandular (endometrio), que constituye el grueso del espesor de la pared del cuerpo uterino.

Neonato: es un bebe que tiene 27 días o menos desde su nacimiento, bien sea por parto o por cesárea.

Oxitocina: es una hormona relacionada con los patrones sexuales y con la conducta maternal y paternal que actúa también como neurotransmisor en el cerebro.

Prostaglandinas: son un conjunto de sustancias derivadas de los ácidos grasos de 20 carbonos.

Referencias bibliográficas

[1] WF Ganong, «Tratado de Fisiología Médica.».
[2] «OMS | Informe de Acción Global sobre Nacimientos Prematuros», 06-May-2013. [Online]. Available: http://www.who.int/pmnch/media/news/2012/preterm_birth_report/es/index3.html. [Accessed: 06-May-2013].
[3] Ángel García Alonso López, Sergio Rosales Ortiz, y Guillermo Jiménez Solís, «DIAGNÓSTICO Y MANEJO DEL PARTO PRETÉRMINO».
[4] Ministerio de Salud Pública, «Norma técnica para la elaboración de las guías de buenas prácticas clínicas». 2005.
[5] José Pacheco, «PARTO PRETÉRMINO: TRATAMIENTO Y LAS EVIDENCIAS», *Rev Per Ginecol Obstet*, n°. 54, págs. 24-32, 2008.
[6] JD Iams, *Preterm birth*, 3o ed. New York: Churchill Livingstone: , 1996.
[7] PJ Meis, R Michielutte, TJ Peters, HB Wells, RE Sands, y EC Coles, *Factors associated with preterm birth in Cardiff, Wales: II*. 1995.
[8] P Astolfi y LA Zonta, *Risks of preterm delivery and association with maternal age, birth order and fetal genderr. Hum Reprod*. 1999.
[9] JD Parker, KC Schoendorf, y JL Kiely, *Associations between measures of socio-economic status and low birthweight, small for gestational age and premature delivery in the United States. Ann Epidemiol*. 1994.
[10] JL Peacock, JM Bland, y HR Anderson, *Preterm delivery: effects of socioeconomic factors, psychological stress, smoking, alcohol and caffeine. BMJ*. 1995.
[11] LE Rosell, A Calzado, y L Torres, «Imnportancia cuantificada de los síntomas sutiles de amenaza de parto pretérmino», *Revista Cubana medicina genera integra*, págs. 275-9, 2000.
[12] Lelyem Marcell Rodríguez y Victoria Esther González Ramírez, «Relación de las citoquinas proinflamatorias con la corioamnionitis subclínica y el parto pretérmino», *Rev Cubana Obstet Ginecol*, vol. 37, n°. 4, Dic-2011.
[13] Roxana Martín Ramos, Rosa María Ramos Palmero, Ricardo Grau Ávalos, y María Matilde García Lorenzo, «Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños». 2007.
[14] Alfredo Morales, Raykenler Yzquierdo, Ernesto Medina, René Lazo, y Pedro Y. Piñedo, «Modelo para el diagnóstico, tratamiento y seguimiento de pacientes con hipertensión arterial», vol. 1, n°. 3, págs. p. 48-57, 2007.
[15] Young Moon Chae y Seung Hee Ho, «ScienceDirect.com - Expert Systems with Applications - Comparison of alternative knowledge model for the diagnosis of asthma», 23-Mar-2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417496000577. [Accessed: 23-Mar-2007].
[16] L Muglia y M Katz, «El enigma del parto pretérmino espontáneo. Existiría un factor genético-ambiental implicado en el parto pretérmino espontáneo.»
[17] JL Ballard, JC Khoury, K Wedig, L Wang, BL Eilers Walsman, y R Lipp, *New Ballard Score, expanded to include extremely premature infants*. 1991.
[18] BM Mercer, RL Goldenberg, A Das, AH Moawad, JD Iams, y PJ Meis, *The preterm prediction study: a clinical risk assessment system. Am J Obstet Gynecol*.

1996.
[19] LS Bakketeig y HJ Hoffman, *Epidemiology of preterm birth: results from a longitudinal study of births in Norway*. 1981.
[20] MM Adams, LD Elam-Evans, HG Wilson, y DA Gilbertz, *Rates and factors associated with recurrence of preterm delivery*. 2000.
[21] Rosell Juarte E (último) y Monzón Torres L, «Importancia cuantificada de los síntomas sutiles de amenaza de parto pretérmino.», *Rev Cubana Med Gen Integr*, 2000.
[22] Morilla Guzmán, Tamayo Pérez, y Fernández Braojos, «Enfermedad de la membrana hialina en Cuba», *Rev Cubana Pediatr*, 2007.
[23] Dra. Maitté Vera López, Dr. Frank Alberto Castillo Fernández (último), y Noris Navas Ábalos, «REPERCUSIÓN DEL PARTO PRETÉRMINO», 2006.
[24] Lázara Miriam Ortega Figueroa, Ana Bertha Álvarez Pineda, Yahavivi Águila Nogueira, y Mercedes I. Viera Hernández, «Detección de infección por Mycoplasma en las gestantes con riesgo de parto pretérmino», vol. 38, Jun-2012.
[25] Stuart J. Russell y Peter Norvig, *Artificial Intelligence A Modern Approach*. Shirley McGuire.
[26] R. Kurzweil, *The age of intelligent machines*. 1990.
[27] E. Rich y K. Knight, *Artificial Intelligence*, 2o ed. 1994.
[28] G. Lugar y W. Stubblefied, *Artificial intelligence: Structures and strategies for complex problem solving*. 1993.
[29] R. Schalkoff, *Artificial intelligence: An engineering approach*. 1990.
[30] «Inteligencia artificial - Monografias.com», 12-Nov-2008. [Online]. Available: http://www.monografias.com/trabajos16/inteligencia-artificial-historia/inteligencia-artificial-historia.shtml. [Accessed: 12-Nov-2008].
[31] «¿Qué es la inteligencia artificial - Monografias.com», 12-Nov-2008. [Online]. Available: http://www.monografias.com/trabajos7/inar/inar.shtml. [Accessed: 12-Nov-2008].
[32] E. R. Kevin Knight, *Inteligencia Artificial*, Segunda.
[33] Aaron Sloman, «ARTIFICIAL INTELLIGENCE. AN ILLUSTRATIVE OVERVIEW».
[34] Rolando Acosta Sánchez., Alejandro Rosete Suárez., y Alfredo Rodríguez Díaz., «Predicción de pacientes diabéticos. Preprocesado para Minería de Datos», *Revista cubana de informática médica*.
[35] Alberto Delgado, «Aplicación de las Redes Neuronales en Medicina», 1999.
[36] R. S. MICHALSKI, *Understanding the Nature of Learning: Issues and Research Directions*. California, 1986.
[37] Humberto Chaviano Arteaga, «La Programación Lógica Inductiva para el Aprendizaje Automatizado de conceptos.», Universidad Central «Marta Abreu» de Las Villas, Santa Clara, 2011.
[38] GÓMEZ A, *Inducción del conocimiento con incertidumbre en Bases de Datos. Relaciones Borrosas*. Madrid, España, 1998.
[39] Zaily Alfonso Cuencio, «Aplicación de Técnicas de Clasificación de Inteligencia Artificial en el Pronóstico de Temperaturas en el Centro Meteorológico de Cienfuegos», Trabajo Diploma, Carlos Rafael Rodríguez, Cienfuegos, 2010.
[40] M. D. H. Liu, *Feature Selection for Classification. Intelligent Data Analysis, Elsevier*, vol. 1. 1997.

[41] H. L. Lei Yu Toward, *Integrating Feature Selection Algorithms for Classification and Clustering*, vol. 17. 2005.
[42] R. Kohavi y G.H. John, *Wrappers for Feature Subset Selection. Artificial Intelligence*, vol. 97. 1997.
[43] S. Das Filters, *Wrappers and a Boosting Based Hybrid for Feature Selection*. 2001.
[44] J. Doak, «An Evaluation of Feature Selection Methods and Their Apllication to Computer Security», Univ. of California at Davis, Technical report, 1992.
[45] H. L. H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston: , 1998.
[46] «COMPARACIÓN DE MODELOS DE CLASIFICACIÓN AUTOMÁTICA DE PATRONES TEXTURALES DE MINERALES PRESENTES EN LOS CARBONES COLOMBIANOS», vol. 72, n°. 146, Ago. 2005.
[47] G Booch, *Análisis y Diseño Orientado a Objetos con Aplicaciones*, Segunda Edición. Addison Wesley Iberoamericana, 1996.
[48] ALBERTO TÉLLEZ VALERO, «Extracción de Información con Algoritmos de Clasificación», 2005.
[49] D. M. D.J. Spiegelhalter y C.C. Taylor, Eds., *Machine Learning, Neural and Statistical Classification*. 1994.
[50] D. M. D.J. Spiegelhalter y C.C. Taylor, Eds., *Machine Learning, Neural and StatisticalClassification*. 1994.
[51] Dra. Josefa Marín Fernández, *Prácticas de ordenador con SPSS para Windows*. .
[52] María García Jiménez y Aránzazu Álvarez Sierra, «Análisis de Datos en WEKA – Pruebas de Selectividad».
[53] A. B. V. Zaida Cebrián Jiménez, *Inteligencia en redes de comunicaciones. Practica WEKA. Diagnóstico cardiología*.
[54] Nimali Shanika Liyanage, «Modelo para el pronóstico del consumo eléctrico en el Hotel Jagua», Trabajo Diploma, Universidad de Cienfuegos “Carlos Rafael Rodríguez, 2010.
[55] C. Borgelt y R. Kruse, *Neural Networks: Working Group Neural Networks and Fuzzy Systems*.
[56] I. BONET, «Modelo para la clasificación de secuencias, en problemas de la bioinformática, usando técnicas de inteligencia artificial», Tesis de doctorado, Universidad Central «Marta Abreu» de Las Villas, 2008.
[57] GUSTAVO A. BETANCOURT, «LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs)», n°. 27, Abr. 2005.

Bibliografía

[1] R. Kohavi y G.H. John, *Wrappers for Feature Subset Selection. Artificial Intelligence*, vol. 97. 1997.
[2] S. Das Filters, *Wrappers and a Boosting Based Hybrid for Feature Selection*. 2001.
[3] S. MANSOURI ALGHALANDIS (último) y GH. NOZAD ALAMDARI, «WELDING DEFECT PATTERN RECOGNITION IN RADIOGRAPHIC IMAGES OF GAS PIPELINES USING ADAPTIVE FEATURE EXTRACTION METHOD AND NEURAL NETWORK CLASSIFIER», Amsterdam, 2006.
[4] «Use of Neural Networks to Model Complex Immunogenetic Associations of Disease: Human Leukocyte Antigen Impact on the Progression of Human Immunodeficiency Virus Infection», 23-Mar-2007. [Online]. Available: http://aje.oxfordjournals.org/content/147/5/464.short. [Accessed: 23-Mar-2007].
[5] R. S. MICHALSKI, *Understanding the Nature of Learning: Issues and Research Directions*. California: , 1986.
[6] MICHALSKI, *Understanding the Nature of Learning: Issues and Research Directions*. 1986.
[7] WF Ganong, «Tratado de Fisiología Médica.» .
[8] BM Mercer, RL Goldenberg, A Das, AH Moawad, JD Iams, y PJ Meis, *The preterm prediction study: a clinical risk assessment system. Am J Obstet Gynecol*. 1996.
[9] RL Goldenberg, *The management of preterm labor*. 2002.
[10] R. Kurzweil, *The age of intelligent machines*. 1990.
[11] A Williams, «Sección VII Complicaciones del embarazo», in *Obstetric*, 21o ed., Mc Graw- Hill, 2001, págs. 592-623.
[12] Young Moon Chae y Seung Hee Ho, «ScienceDirect.com - Expert Systems with Applications - Comparison of alternative knowledge model for the diagnosis of asthma», 23-Mar-2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417496000577. [Accessed: 23-Mar-2007].
[13] P Astolfi y LA Zonta, *Risks of preterm delivery and association with maternal age, birth order and fetal genderr. Hum Reprod*. 1999.
[14] ER Pschirrer y M Monga, *Risk factors for preterm labor*. 2000.
[15] Dra. Maitté Vera López, Dr. Frank Alberto Castillo Fernández (último), y Noris Navas Ábalos, «REPERCUSIÓN DEL PARTO PRETÉRMINO», 2006.
[16] Lelyem Marcell Rodríguez y Victoria Esther González Ramírez, «Relación de las citoquinas proinflamatorias con la corioamnionitis subclínica y el parto pretérmino», *Rev Cubana Obstet Ginecol*, vol. 37, n^{o}. 4, Dic-2011.
[17] Damián Jorge Matich, *Redes Neuronales: Conceptos Básicos y Aplicaciones.* 2001.
[18] Bonifacio Martín del Brío y Alfredo Sanz Molina, *Redes Neuronales y Sistemas Difusos*, Segunda. .
[19] Luis Alonso Romero y Teodoro Calonge Cano, «Redes Neuronales y Reconocimiento de Patrones.», .
[20] «Redes neuronales artificiales Fundamentos, modelos y aplicaciones - Monografias.com», 12-Nov-2008. [Online]. Available:

http://www.monografias.com/trabajos12/redneur/redneur.shtml. [Accessed: 12-Nov-2008].
[21] Alfonso Pitarque, Juan Francisco Roy, y Juan Carlos Ruiz, *Redes neurales vs modelos estadísticos: Simulaciones sobre tareas de predicción y clasificación*. .
[22] MM Adams, LD Elam-Evans, HG Wilson, y DA Gilbertz, *Rates and factors associated with recurrence of preterm delivery*. 2000.
[23] Janny Hermosa Morell Díaz, «Propuesta de técnica de Inteligencia Artificial para la detección de anomalías en la red de datos de la Universidad de Cienfuegos.», Carlos Rafael Rodríguez, 2009.
[24] RK Creasy y JD Iams, *Preterm labor and delivery. En: Creasy RK, Resnik R (eds). Maternal-Fetal Medicine.*, 4o ed. Philadelphia: WB Saunders Co: , 1999.
[25] JL Peacock, JM Bland, y HR Anderson, *Preterm delivery: effects of socioeconomic factors, psychological stress, smoking, alcohol and caffeine. BMJ*. 1995.
[26] K Haram y J Helge, «Preterm delivery: an overview». 2003.
[27] JD Iams, *Preterm birth*, 3o ed. New York: Churchill Livingstone: , 1996.
[28] Rolando Acosta Sánchez., Alejandro Rosete Suárez., y Alfredo Rodríguez Díaz, «Predicción de pacientes diabéticos. Preprocesado para Minería de Datos», *Revista cubana de informática médica*.
[29] Dra. Josefa Marín Fernández, *Prácticas de ordenador con SPSS para Windows*.
[30] José Pacheco, «PARTO PRETÉRMINO: TRATAMIENTO Y LAS EVIDENCIAS», *Rev Per Ginecol Obstet*, n^{o}. 54, págs. 24-32, 2008.
[31] «OMS | Informe de Acción Global sobre Nacimientos Prematuros», 06-May-2013. [Online]. Available: http://www.who.int/pmnch/media/news/2012/preterm_birth_report/es/index3.html. [Accessed: 06-May-2013].
[32] Ministerio de Salud Pública, «Norma técnica para la elaboración de las guías de buenas prácticas clínicas». 2005.
[33] JL Ballard, JC Khoury, K Wedig, L Wang, BL Eilers Walsman, y R Lipp, *New Ballard Score, expanded to include extremely premature infants*. 1991.
[34] C. Borgelt y R. Kruse, *Neural Networks: Working Group Neural Networks and Fuzzy Systems*. .
[35] I. BONET, «Modelo para la clasificación de secuencias, en problemas de la bioinformática, usando técnicas de inteligencia artificial», Tesis de doctorado, Universidad Central «Marta Abreu» de Las Villas, 2008.
[36] Nimali Shanika Liyanage, «Modelo para el pronóstico del consumo eléctrico en el Hotel Jagua», Trabajo Diploma, Universidad de Cienfuegos “Carlos Rafael Rodríguez, 2010.
[37] Alfredo Morales, Raykenler Yzquierdo, Ernesto Medina, René Lazo, y Pedro Y. Piñedo, «Modelo para el diagnóstico, tratamiento y seguimiento de pacientes con hipertensión arterial», vol. 1, n^{o}. 3, págs. p. 48-57, 2007.
[38] Alejandro González Delgado, «Modelo de pronóstico alternativo para el estado al egreso de pacientes con enfermedades cerebro-vasculares en el hospital provincial de Cienfuegos», Grado, Carlos Rafael Rodríguez, Cienfuegos, 2008.
[39] María Castellano Méndez, «Modelación estadística con Redes Neuronales. Aplicaciones a la Hidrología, Aerobiología y Modelización de Procesos», Doctoral, De Coruña, 2009.

[40] CÉSAR HERVÁS MARTINEZ, «MÉTODOS DE CLASIFICACIÓN NO LINEAL UTILIZANDO REDES NEURONALES». .
[41] V. E. Gómez C, «Medición ecográfica transvaginal del cuello uterino en la predicción del parto pretérmino espontáneo en el Instituto Materno Perinatal durante el año 2002», Universidad Nacional Mayor de San Marcos, 2004.
[42] D. M. D.J. Spiegelhalter y C.C. Taylor, Eds., *Machine Learning, Neural and Statistical Classification*. 1994.
[43] GUSTAVO A. BETANCOURT, «LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs)», n°. 27, Abr. 2005.
[44] Humberto Chaviano Arteaga, «La Programación Lógica Inductiva para el Aprendizaje Automatizado de conceptos.», Universidad Central «Marta Abreu» de Las Villas, Santa Clara, 2011.
[45] A. B. V. Zaida Cebrián Jiménez, *Inteligencia en redes de comunicaciones. Practica WEKA. Diagnóstico cardiología*. .
[46] «Inteligencia artificial - Monografias.com», 12-Nov-2008. [Online]. Available: http://www.monografias.com/trabajos16/inteligencia-artificial-historia/inteligencia-artificial-historia.shtml. [Accessed: 12-Nov-2008].
[47] E. R. Kevin Knight, *Inteligencia Artificial*, Segunda. .
[48] H. L. Lei Yu Toward, *Integrating Feature Selection Algorithms for Classification and Clustering*, vol. 17. 2005.
[49] S. K. G.Syrkos, C. B. A.Dounis, y T. C. D. Papachristos, «INTEGRATED AUTOMATED SYSTEM FOR THE IMPLEMENTATION OF SKIN TESTS AND ALLERGY DIAGNOSIS». .
[50] GÓMEZ A, *Inducción del conocimiento con incertidumbre en Bases de Datos. Relaciones Borrosas*. Madrid, España: , 1998.
[51] Rosell Juarte E y Monzón Torres L, «Importancia cuantificada de los síntomas sutiles de amenaza de parto pretérmino.», *Rev Cubana Med Gen Integr*, 2000.
[52] LE Rosell, A Calzado, y L Torres, «Imnportancia cuantificada de los síntomas sutiles de amenaza de parto pretérmino», *Revista Cubana medicina genera integra*, págs. 275-9, 2000.
[53] «Fundamentos de Data Mining y sus Aplicaciones en la Gerencia Integrada de Yacimientos». .
[54] H. L. H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston: , 1998.
[55] M. D. H. Liu, *Feature Selection for Classification. Intelligent Data Analysis, Elsevier*, vol. 1. 1997.
[56] PJ Meis, R Michielutte, TJ Peters, HB Wells, RE Sands, y EC Coles, *Factors associated with preterm birth in Cardiff, Wales: II*. 1995.
[57] Norma Soledad Riva Reategui, «Factores de riesgo para parto pretérmino espontáneo en gestantes adolescentes del Hospital de apoyo N° 2 Yarinacocha-Pucallpa», Universidad Mayor de san Marcos, Lima-Perú, 2004.
[58] Juvenal Calderón Guillén, Genaro Vega Malagón, Jorge Velásquez Tlapanco, Régulo Morales Carrera, y Alfredo Jesús Vega Malagón, «Factores de riesgo materno asociados al parto pretérmino», vol. 43, n°. 4, 2005.
[59] W Villamonte, N Lam, y E Ojeda, «Factores de riesgo del parto pretérmino. Instituto Materno Perinatal.», 2001.

[60] ALBERTO TÉLLEZ VALERO, «Extracción de Información con Algoritmos de Clasificación», 2005.
[61] Vincenzo Berghella, Jason K Baxter, Nancy W, y Hendrix, «Evaluación ecográfica del cuello del útero para la prevención del parto prematuro (Revision Cochrane traducida)», 2009.
[62] LS Bakketeig y HJ Hoffman, *Epidemiology of preterm birth: results from a longitudinal study of births in Norway*. 1981.
[63] CV Ananth y AM Vintzileos, *Epidemiology of preterm birth and its clinical subtypes*. 2006.
[64] RL Goldenberg, Iams JD, y Romero R, *Epidemiology and causes of preterm birth.* 2008.
[65] Vásquez G, Ramírez J, y Villar A, *Epidemiología del parto pretérmino en el Hospital San Bartolomé de Lima*. Lima – Perú: , Enero 94 – Diciembre 98.
[66] Morilla Guzmán, Tamayo Pérez, y Fernández Braojos, «Enfermedad de la membrana hialina en Cuba», *Rev Cubana Pediatr*, 2007.
[67] F Althabe, G Carroli, R Lede, JM Belizán, y OH Althabe, «El parto pretérmino: detección de riesgos y tratamientos preventivos.», 2003.
[68] L Muglia y M Katz, «El enigma del parto pretérmino espontáneo. Existiría un factor genético-ambiental implicado en el parto pretérmino espontáneo.»
[69] Ángel García Alonso López, Sergio Rosales Ortiz, y Guillermo Jiménez Solís, «DIAGNÓSTICO Y MANEJO DEL PARTO PRETÉRMINO». .
[70] Lázara Miriam Ortega Figueroa, Ana Bertha Álvarez Pineda, Yahavivi Águila Nogueira, y Mercedes I. Viera Hernández, «Detección de infección por Mycoplasma en las gestantes con riesgo de parto pretérmino», vol. 38, Jun-2012.
[71] Diana Lourdes Pernús Alonso, «Desarrollo de un nuevo método de Aprendizaje de la Ordenación, basado en el algoritmo de optimización global Supernova», Grado, Carlos Rafael Rodríguez, Cienfuegos, 2012.
[72] Dr. Rafael Bello Pérez, *CURSO INTRODUCTORIO A LAS REDES NEURONALES ARTIFICIALES*. 1993.
[73] A. A. R. Manuel Cué Brugueras, «Comportamiento del asma bronquial en Cuba e importancia de la prevención de las enfermedades alérgicas en infantes», 2006.
[74] «COMPARACIÓN DE MODELOS DE CLASIFICACIÓN AUTOMÁTICA DE PATRONES TEXTURALES DE MINERALES PRESENTES EN LOS CARBONES COLOMBIANOS», vol. 72, n^{o}. 146, Ago. 2005.
[75] J. A. C. Laura B. Martinez y Yuliana V. Puerta, «Clasificador Difuso para Diagnóstico de Enfermedades», Dic-2010.
[76] «Características de las redes neuronales», 22-Mar-2007. [Online]. Available: http://thales.cica.es/rd/Recursos/rd98/TecInfo/07/capitulo3.html. [Accessed: 22-Mar-2007].
[77] Guillermina Salazar de Dugarte, Pedro Faneite Antique, y Francis Pineda de Molina, «Capacidad de la alfa-fetoproteína en suero materno como marcador predictivo de parto pretérmino», *Revista de Obstetricia y Ginecología de Venezuela*, vol. 69, n^{o}. 4, Dic-2009.
[78] María L Sanz, «Avances en el Diagnóstico in Vitro de las Enfermedades Alérgicas.», 2009.

[79] JD Parker, KC Schoendorf, y JL Kiely, *Associations between measures of socio-economic status and low birthweight, small for gestational age and premature delivery in the United States. Ann Epidemiol*. 1994.
[80] G. Lugar y W. Stubblefied, *Artificial intelligence: Structures and strategies for complex problem solving*. 1993.
[81] R. Schalkoff, *Artificial intelligence: An engineering approach*. 1990.
[82] Aaron Sloman, «ARTIFICIAL INTELLIGENCE. AN ILLUSTRATIVE OVERVIEW». .
[83] Stuart J. Russell y Peter Norvig, *Artificial Intelligence A Modern Approach*. Shirley McGuire.
[84] E. Rich y K. Knight, *Artificial Intelligence*, 2o ed. 1994.
[85] Alejandro Rodríguez González, «Aplicación de tecnologías semánticas para la creación de sistemas inteligentes de diagnóstico diferencial de alta sensibilidad en medicina», Doctoral, Universidad de Murcia, 2012.
[86] Zaily Alfonso Cuencio, «Aplicación de Técnicas de Clasificación de Inteligencia Artificial en el Pronóstico de Temperaturas en el Centro Meteorológico de Cienfuegos», Trabajo Diploma, Carlos Rafael Rodríguez, Cienfuegos, 2010.
[87] Roxana Martín Ramos, Rosa María Ramos Palmero, Ricardo Grau Ávalos, y María Matilde García Lorenzo, «Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños». 2007.
[88] Alberto Delgado, «Aplicación de las Redes Neuronales en Medicina», 1999.
[89] G Booch, *Análisis y Diseño Orientado a Objetos con Aplicaciones*, Segunda Edición. Addison Wesley Iberoamericana, 1996.
[90] María García Jiménez y Aránzazu Álvarez Sierra, «Análisis de Datos en WEKA – Pruebas de Selectividad».
[91] J. Doak, «An Evaluation of Feature Selection Methods and Their Apllication to Computer Security», Univ. of California at Davis, Technical report, 1992.
[92] «¿Qué es la inteligencia artificial - Monografias.com», 12-Nov-2008. [Online]. Available: http://www.monografias.com/trabajos7/inar/inar.shtml. [Accessed: 12-Nov-2008].

Anexos

Anexo a: “Variantes del método de clasificación IBK”.

Variante 1:

=== Summary ===

Correctly Classified Instances	389	66.4957 %
Incorrectly Classified Instances	196	33.5043 %
Kappa statistic	0.0907	
Mean absolute error	0.3844	
Root mean squared error	0.485	
Relative absolute error	84.7964 %	
Root relative squared error	101.8758 %	
Coverage of cases (0.95 level)	95.2137 %	
Mean rel. region size (0.95 level)	91.7094 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
24	179	a = 1
17	365	b = 2

Variante 2:

=== Summary ===

Correctly Classified Instances	434	74.188 %
Incorrectly Classified Instances	151	25.812 %
Kappa statistic	0.4065	
Mean absolute error	0.263	
Root mean squared error	0.4879	

Relative absolute error 58.0076 %

Root relative squared error 102.4959 %

Coverage of cases (0.95 level) 78.6325 %

Mean rel. region size (0.95 level) 55.4701 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

110 93 | a = 1

58 324 | b = 2

Variante 3:

=== Summary ===

Correctly Classified Instances 449 76.7521 %

Incorrectly Classified Instances 136 23.2479 %

Kappa statistic 0.4377

Mean absolute error 0.284

Root mean squared error 0.4191

Relative absolute error 62.6487 %

Root relative squared error 88.0482 %

Coverage of cases (0.95 level) 94.0171 %

Mean rel. region size (0.95 level) 78.6325 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

97 106 | a = 1

30 352 | b = 2

Variante 4:

=== Summary ===

Correctly Classified Instances	443	75.7265 %
Incorrectly Classified Instances	142	24.2735 %
Kappa statistic	0.4023	
Mean absolute error	0.2939	
Root mean squared error	0.4156	
Relative absolute error	64.8197 %	
Root relative squared error	87.302 %	
Coverage of cases (0.95 level)	96.0684 %	
Mean rel. region size (0.95 level)	84.1026 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
87	116	a = 1
26	356	b = 2

Anexo b: "Variantes del método de clasificación J48".

Variante 1:

=== Summary ===

Correctly Classified Instances	440	75.2137 %
Incorrectly Classified Instances	145	24.7863 %
Kappa statistic	0.4341	
Mean absolute error	0.3173	
Root mean squared error	0.4432	
Relative absolute error	69.9973 %	
Root relative squared error	93.0957 %	
Coverage of cases (0.95 level)	95.8974 %	
Mean rel. region size (0.95 level)	93.5897 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
116	87	a = 1
58	324	b = 2

Variante 2:

=== Summary ===

Correctly Classified Instances	432	73.8462 %
Incorrectly Classified Instances	153	26.1538 %
Kappa statistic	0.373	
Mean absolute error	0.3849	
Root mean squared error	0.4384	
Relative absolute error	84.9072 %	

Root relative squared error 92.0861 %

Coverage of cases (0.95 level) 100 %

Mean rel. region size (0.95 level) 100 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

92 111 | a = 1

42 340 | b = 2

Variante 3:

=== Summary ===

Correctly Classified Instances 440 75.2137 %

Incorrectly Classified Instances 145 24.7863 %

Kappa statistic 0.4273

Mean absolute error 0.3445

Root mean squared error 0.4305

Relative absolute error 75.9756 %

Root relative squared error 90.4353 %

Coverage of cases (0.95 level) 99.4872 %

Mean rel. region size (0.95 level) 97.7778 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

111 92 | a = 1

53 329 | b = 2

Variante 4:

=== Summary ===

Correctly Classified Instances	445	76.0684 %
Incorrectly Classified Instances	140	23.9316 %
Kappa statistic	0.4569	
Mean absolute error	0.3256	
Root mean squared error	0.4257	
Relative absolute error	71.8089 %	
Root relative squared error	89.4378 %	
Coverage of cases (0.95 level)	99.3162 %	
Mean rel. region size (0.95 level)	96.2393 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
121	82	a = 1
58	324	b = 2

Anexo c: “Variantes del método de clasificación MLP”.

Variante 1:

=== Summary ===

Correctly Classified Instances	459	78.4615 %
Incorrectly Classified Instances	126	21.5385 %
Kappa statistic	0.5169	
Mean absolute error	0.2302	
Root mean squared error	0.4342	
Relative absolute error	50.7844 %	
Root relative squared error	91.2128 %	
Coverage of cases (0.95 level)	86.1538 %	
Mean rel. region size (0.95 level)	59.9145 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
133	70	a = 1
56	326	b = 2

Variante 2:

=== Summary ===

Correctly Classified Instances	459	78.4615 %
Incorrectly Classified Instances	126	21.5385 %
Kappa statistic	0.5181	
Mean absolute error	0.2782	
Root mean squared error	0.412	
Relative absolute error	61.3514 %	

Root relative squared error 86.5462 %

Coverage of cases (0.95 level) 95.7265 %

Mean rel. region size (0.95 level) 86.2393 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

134 69 | a = 1

57 325 | b = 2

Variante 3:

=== Summary ===

Correctly Classified Instances 447 76.4103 %

Incorrectly Classified Instances 138 23.5897 %

Kappa statistic 0.4734

Mean absolute error 0.2401

Root mean squared error 0.4534

Relative absolute error 52.9653 %

Root relative squared error 95.2391 %

Coverage of cases (0.95 level) 85.4701 %

Mean rel. region size (0.95 level) 60.7692 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

129 74 | a = 1

64 318 | b = 2

Variante 4:

=== Summary ===

Correctly Classified Instances	467	79.8291 %
Incorrectly Classified Instances	118	20.1709 %
Kappa statistic	0.54	
Mean absolute error	0.2704	
Root mean squared error	0.376	
Relative absolute error	59.6469 %	
Root relative squared error	78.9879 %	
Coverage of cases (0.95 level)	99.4872 %	
Mean rel. region size (0.95 level)	91.8803 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
130	73	a = 1
45	337	b = 2

Anexo d: "Variantes del método de clasificación SMO".

Variante 1:

=== Summary ===

Correctly Classified Instances	468	80 %
Incorrectly Classified Instances	117	20 %
Kappa statistic	0.5334	
Mean absolute error	0.2861	
Root mean squared error	0.3815	
Relative absolute error	63.1062 %	
Root relative squared error	80.1504 %	
Coverage of cases (0.95 level)	99.8291 %	
Mean rel. region size (0.95 level)	95.2137 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
121	82	a = 1
35	347	b = 2

Variante 2:

=== Summary ===

Correctly Classified Instances	474	81.0256 %
Incorrectly Classified Instances	111	18.9744 %
Kappa statistic	0.5729	
Mean absolute error	0.248	
Root mean squared error	0.3853	
Relative absolute error	54.6979 %	

Root relative squared error 80.9453 %

Coverage of cases (0.95 level) 97.094 %

Mean rel. region size (0.95 level) 80.4274 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

139 64 | a = 1

47 335 | b = 2

Variante 3:

=== Summary ===

Correctly Classified Instances 469 80.1709 %

Incorrectly Classified Instances 116 19.8291 %

Kappa statistic 0.5287

Mean absolute error 0.1983

Root mean squared error 0.4453

Relative absolute error 43.7368 %

Root relative squared error 93.5444 %

Coverage of cases (0.95 level) 80.1709 %

Mean rel. region size (0.95 level) 50 %

Total Number of Instances 585

=== Confusion Matrix ===

a b <-- classified as

114 89 | a = 1

27 355 | b = 2

Variante 4:

=== Summary ===

Correctly Classified Instances	463	79.1453 %
Incorrectly Classified Instances	122	20.8547 %
Kappa statistic	0.5116	
Mean absolute error	0.2085	
Root mean squared error	0.4567	
Relative absolute error	45.999 %	
Root relative squared error	95.9332 %	
Coverage of cases (0.95 level)	79.1453 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	585	

=== Confusion Matrix ===

a	b	<-- classified as
117	86	a = 1
36	346	b = 2