UNIVERSIDAD TECNOLÓGICA DE BOLÍVAR

FACULTAD DE INGENIERÍAS

Título: Predicción Temprana de Morbilidad Materna Extrema Usando Aprendizaje Automático

Autor: Eugenia Luz Arrieta Rodríguez

Jurado

Jurado

Director: Juan Carlos Martínez Santos

Cartagena, Julio de 2017

Predicción Temprana de Morbilidad Materna Extrema Usando Aprendizaje Automático

Eugenia Luz Arrieta Rodriguez
Director: Juan Carlos Martínez Santos

Universidad Tecnológica de Bolívar
Facultad de Ingenierías
Programa de Ingeniería de Sistemas
Cartagena

Julio de 2017

Predicción Temprana de Morbilidad Materna Extrema Usando Aprendizaje Automático

Eugenia Luz Arrieta Rodriguez

Trabajo de grado para optar al título de

Magister en Ingeniería

Director: Juan Carlos Martínez Santos

Universidad Tecnológica de Bolívar
Facultad de Ingenierías
Cartagena

Julio de 2017

Resumen

La Morbilidad Materna Extrema (MME) es un problema de salud pública en el mundo. Ocurre durante el embarazo, parto, o puerperio, esta condición pone en riesgo la vida de la mujer y del bebé. Esta condición es muy difícil de detectar en una etapa temprana. En respuesta a lo anterior, este trabajo propone la utilización de teécnicas de Aprendizaje Automático, que se consideran más relevantes en un entorno biomédico, usando el aprendizaje para predecir el nivel de riesgo de morbilidad materna extrema en pacientes durante el embarazo. La población estudiada corresponde a las mujeres que embarazadas reciben atención prenatal y conclusión de su embarazo en la E.S.E Clínica de Maternidad Rafael Calvo (CMRC) en Cartagena, Colombia.

Para la construcción de esta herramienta inicialmente se recolectó la información más importante de las gestantes de control prenatal a través de formularios. Con esta información se creó una base de datos para el proyecto que incluyó datos de 2015 y 2016. Debido a la cantidad de variables del estudio se decide hacer un proceso de filtrado de variables al conjunto de entrenamiento que corresponde al 80 % de la muestra, quedando solo con las variables que contribuyan sensiblemente al buen desempeño del clasificador. La técnica de filtrado utilizada es selección de características que determinan la correlación entre atributos y la variable que se desea predecir. La herramienta fue construida usando varias técnicas de aprendizaje automático como son la regresión logística, regresión logística polinomial y máquinas de soporte vectorial.

Se desarrollaron algoritmos procedimentales en lenguaje Python haciendo uso de las librerias *sklearn*, para clasificación entre dos clases. Para el proceso de entrenamiento se utilizó el laboratorio de computación de alto rendimiento (HPCLab) de la Universidad Tecnológica de Bolívar, debido a la demanda de recursos del entrenamiento. Se aplicó la validación cruzada y el error de pruebas para determinar la técnica de aprendizaje supervisado que mejor se adaptara al problema de predicción temprana de Morbilidad Materna Extrema. Se seleccionó la técnica de Regresión Logística para el set de datos de primer y segundo trimestre obteniendo como resultado un clasificador con 97 % de sensibilidad, 51.8 % de precisión y 67.7 % medida F. Para el set de datos del tercer trimestre se seleccionó Máquinas de Soporte Vectorial con resultados de 100 % de sensibilidad y 27 % de precisión.

Durante la realización de este proyecto de investigación se realizó la publicación del artículo llamado "Early Prediction of Severe Maternal Morbidity Using Machine Learning Techniques"[21], que fue presentado en el congreso iberoamericano de inteligencia artificial (IBERAMIA) en noviembre 2016. En este se presentó el análisis estadístico de los datos del año 2015 y la implementación la técnica de Regresión Logística para la predicción de la Morbilidad Materna Extrema.

También fue presentado el trabajo titulado "Predicción temprana de Morbilidad Materna Extrema Usando Machine Learning", en el VII Festival internacional de la ciencia y la cultura XII Encuentro de investigadores de RIESCAR (Red de Instituciones de Educación Superior del Caribe) en octubre 2016.

Agradecimientos

Agradezco a Dios por brindarme esta oportunidad y la fortaleza para lograr esta meta. A mis padres y a mis hijos Tomás y Santiago por su apoyo incondicional.

A la E.S.E Clínica de Maternidad Rafael Calvo y al equipo de personas que apoyaron y permitieron el desarrollo del proyecto en esta institución. Al Laboratorio de Computación de alto desempeño (HPCLab) de la Universidad Tecnológica de Bolívar. Al Centro de Investigación para la Salud Materna, Perinatal y de la Mujer de la E.S.E Clínica de Maternidad Rafael Calvo.

A mi director de tesis Juan Carlos Martínez Santos por su constancia y aporte de conocimientos de acuerdo a su amplia experiencia investigativa. A William Caicedo por su disposición y colaboración como experto en aprendizaje automático en el campo de la medicina. Agradecimientos especiales a Iván Baños Delgado y Luz Estella Robles por su apoyo incondiconal y valiosos aportes para mi crecimiento académico.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1
Introducción

1.1. Planteamiento del Problema

A pesar de lograr avances en la salud materna, las complicaciones relacionadas con la gestación siguen siendo un importante problema de salud pública en el mundo. Cada año mueren aproximadamente 500.000 mujeres durante la gestación, el parto o el puerperio [12]. Se presentan cerca de 50 millones de problemas en salud materna anualmente y aproximadamente 300 millones de mujeres sufren a corto y largo plazo, de enfermedades y lesiones relacionadas con el embarazo, el parto y el puerperio [17][24].

Para el caso particular de la E.S.E Clínica de maternidad Rafael Calvo, institución donde se realiza el estudio, las cifras de 2014 y 2015 muestran que el promedio de atenciones entre partos, cesáreas y legrados son de 10 mil casos al año. Esta institución realiza controles del embarazo a aproximadamente 3.000 pacientes, mediante consultas especializada de Ginecología y Obstetricia, de las cuales cerca del 34 % (alrededor de 1.000) concluyen su embarazo en la E.S.E clínica de Maternidad Rafael Calvo.

Entre 2015 y mediados de 2016 se presentaron alrededor de 1.500 casos de MME debido a trastornos hipertensivos, preeclampsia y eclampsia. De estos casos 191 corresponden a pacientes con embarazos controlados en la CMRC. Es decir, que cerca del 20 % de las pacientes controladas, con atención final del embarazo en CMRC llegaron a clasificarse con MME entre 2015 y 2016.

El problema por el cual no se ha logrado la reducción significativa de la Morbilidad Materna Extrema en la CMRC se divide en dos partes. Primero, factor humano, durante el proceso de atención por valoración Gineco-Obstétrica en los controles prenatales resulta difícil para el personal médico la detección oportuna de los factores de riesgo para casos potenciales de morbilidad materna extrema (MME), aún cuando estos cuentan con los conocimientos y se siguen los protocolos indicados. Las causas van desde la duración de la atención, la cantidad de pacientes citados, hasta los compromisos adicionales del personal médico en otras entidades. Esto lleva a complicaciones posteriores, poniendo en riesgo la vida de la madre y el bebe. Segundo, Historia Clínica, actualmente la E.S.E Clínica de Maternidad Rafael Calvo (CMRC) no cuenta con una guía o historia clínica completa que contenga las variables que evalúan factores de riesgo para morbilidad propuestas por la organización mundial de la salud (OMS) [17], la Organización Panamericana de la salud (OPS) [5] y el Centro Latino Americano de Perinatología (CLAP) [22].

1.2. Pregunta de investigación

Teniendo en cuenta lo anterior es necesario preguntar ¿ En que medida la implementación de las técnicas de aprendizaje automático supervisado facilitan el proceso de predicción de riesgo de MME, así como la identificación de variables que se asocian a esta condición, a través del diseño y ejecución de una herramienta tecnológica?

1.3. Objetivos

Construir una herramienta para predicción de riesgos de morbilidad materna extrema basada en Técnicas de aprendizaje automático.

Para lograr el anterior objetivo se proponen los siguientes objetivos específicos:

- Teniendo en cuenta los estudios anteriores, identificar el grupo de variables que se asocian a la ocurrencia de la morbilidad materna extrema, con el fin de seleccionar el subconjunto de variables que históricamente han tenido mayor impacto en la ocurrencia de MME en la población objeto de estudio.

- Construir un aplicativo que facilite el proceso de recolección de la información de los controles prenatales, para contar solo con los datos que se relacionan directamente con las variables en estudio.

- Diseñar el conjunto de pruebas, a través de las técnicas estadísticas que permitirán re definir el grupo de variables que tienen mayor incidencia en la ocurrencia de MME.

- Diseñar e implementar algoritmos de aprendizaje supervisado automático que permitan obtener el nivel deseado de sensibilidad y precisión.

- Establecer un comparativo entre las técnicas (de aprendizaje supervisadas automática) implementadas, que midan la precisión de cada algoritmo a nivel de sensibilidad y precisión.

- Diseñar el conjunto de pruebas que permitan validarla herramienta diseñada.

1.4. Justificación

Dentro del Plan de Acción 2012- 2017 para acelerar la reducción de la mortalidad y morbilidad materna extrema (OPS/OMS) [5][17], se plantea la necesidad del fortalecimiento de los sistemas de información y vigilancia de la salud materna en los países de la región (Latinoamérica). Se establecen, dentro de los indicadores de monitoreo y evaluación, el registro sistemático de la morbilidad materna extrema y la medición de los indicadores del evento [4]. Siendo la reducción de la MME una de las metas del milenio y un propósito nacional, se han tomado medidas como la vigilancia epidemiológica que contribuyen a la disminución, al aporte de nuevos conocimientos sobre la base científica del problema, y a la identificación de factores que contribuyen a estos eventos [4].

A pesar de los esfuerzos, no se logra cumplir aún con la meta planteada, haciéndose necesario contar con una herramienta que apoye con la identificación o clasificación de posibles riesgos para MME en forma oportuna, que a su vez permita realizar seguimiento y análisis retrospectivo de las pacientes mórbidas.

Si no se toman estas medidas, se seguirá presentando un alto nivel de MME a nivel local, en nuestra región caribe, a nivel nacional y latinoamericano. En este proyecto se escogerá un modelo estadístico que se ajuste más al tipo de información. Con la integración de este modelo se espera la detección temprana de los casos de

morbilidad y de esta forma ayudar a la intervención oportuna de las pacientes por parte del personal médico. Esto reduciría el riesgo que pueda tener la gestante y el bebe durante esta etapa, y a su vez disminuyen las repercusiones sociales (a nivel de indicadores internacionales) y económicas en el país.

Se debe entender que no se pueden escatimar recursos para la salvar la vida de un ser humano, pero lo mejor es evitar que se den los casos. Algo importante a resaltar es que la experiencia a ganar nos será, a futuro, una base de información para la realización de informes e indicadores como insumo fundamental para la toma oportuna de correctivos y decisiones frente a la atención materna a nivel regional escalable a nivel nacional e internacional.

Con la realización de este proyecto se espera que a futuro la E.S.E CMRC se convierta un referente regional en atención prenatal y prevención de MME. Logrando con esto que cada vez más entidades prestadoras en salud envíen sus pacientes a CMRC, para hacerles el control prenatal y atención final del parto.

Capítulo 2
Marco Teórico

A continuación, se presentan las definiciones de los diferentes términos que ayudarán a enmarcar el proyecto investigativo, esto ayudará a entender mucho más los temas que se tratan en él.

2.1. *Aprendizaje Automático*

El aprendizaje automático o aprendizaje de máquinas (del inglés, "Machine Learning") es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Es decir, trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. En muchas ocasiones el campo de actuación del aprendizaje automático se solapa con el de la estadística computacional, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el aprendizaje automático también se centra en el estudio de la complejidad computacional de los problemas. Este puede ser visto como un intento de automatizar algunas partes del método científico mediante métodos matemáticos [26].

El aprendizaje automático tiene una amplia gama de aplicaciones dentro de las cuales se puede mencionar los motores de búsqueda, diagnósticos médicos, detección de fraude, análisis del mercado, reconocimiento del habla y del lenguaje escrito, juegos y robótica. Esta forma de aprendizaje se relacionada con estadística computacional, que también se centra en la predicción de decisiones mediante el uso de computadoras. Tiene fuertes vínculos con la optimización matemática. Existen dos tipos de aprendizaje para aprendizaje automático: el supervisado y no supervisado.

2.1.1. Aprendizaje No Supervisado

Es un método de aprendizaje automático donde un modelo es ajustado a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que en el aprendizaje supervisado se tiene una clase de respuesta, tiene establecidas las categorías, mientras que el no supervisado no cuenta con categorías de respuesta, este utiliza técnicas de agrupamiento para encontrar la cantidad de clases. El aprendizaje no supervisado puede ser usado junto con la inferencia bayesiana para conseguir probabilidades condicionales, es decir, aprendizaje supervisado. Las técnicas de aprendizaje no supervisado mas comunes son *clustering*, visualización, reducción de dimensionalidad y extracción de características. Este tipo de aprendizaje es frecuentemente usado para la lograr comprender el comportamiento de los datos, es decir, encontrar alguna estructura o forma de organizarlos.

2.1.2. Aprendizaje Supervisado

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de dos vectores: uno que representa los datos de entrada y el otro los resultados deseados.

La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). En la Figura 2.1 se muestra el objetivo del aprendizaje supervisado el cual es crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente.

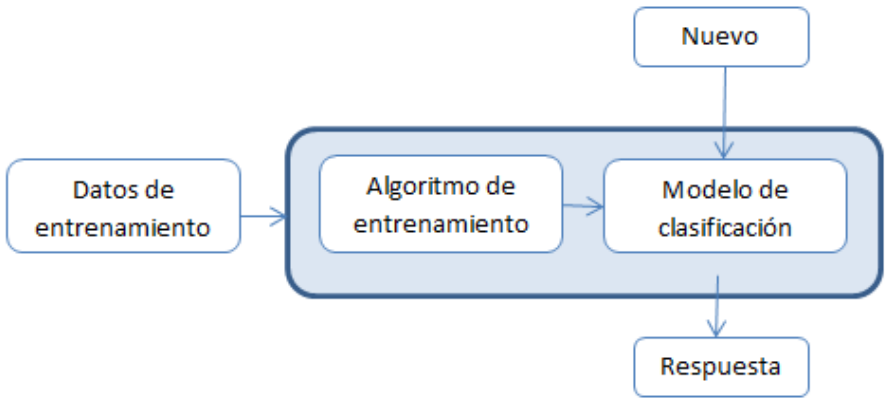

Figura 2.1: Aprendizaje automático

En la Figura 2.2 se presenta una descripción formal del problema de aprendizaje automático, indicando que dado un conjunto de datos de entrenamiento que son los vectores X (variables de entrada) y Y (respuesta), se realiza un proceso de entrenamiento usando un algoritmo de aprendizaje, con esto se logra encontrar la función

$h(x)$ haga una predicción y. Donde la función h se llama hipótesis o modelo.

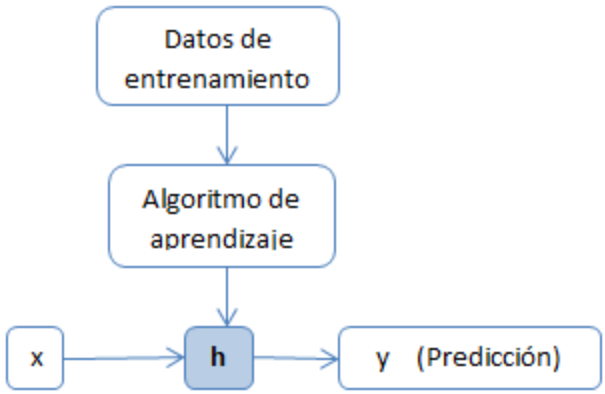

Figura 2.2: Definición del problema de aprendizaje automático

El algoritmo de aprendizaje debe encontrar los parámetros correctos de la función se realiza un proceso llamado entrenamiento o proceso de aprendizaje. Por ejemplo, si se desea enseñar matemáticas a un grupo de estudiantes, se les entrena en el tema mediante ejercicios, de esta misma forma se hace con los clasificadores, primero se les enseña mediante casos de ejemplo.

Al igual que en la vida real, a los estudiantes no se les evalúa usando el mismo ejercicio del entrenamiento, si no ejercicios que nunca han visto. De igual forma a los clasificadores tampoco se les evalúa con los mismos casos del proceso de aprendizaje, porque la idea es que pueda predecir correctamente con casos reales. Por ende, el objetivo de un clasificador es que logre predecir casos reales tan bien como lo hizo con los casos de entrenamiento. A esta capacidad de extrapolar correctamente se le llama generalización. Para llegar a lograr la generalización de un clasificador se requiere que en las pruebas el error de predicción sea bajo, a este error se le llama

error de pruebas.

Si el error entrenamiento es pequeño, es muy probable que el de pruebas también lo sea, pero esto no es una garantía para el aprendizaje. Para lograr una buena generalización del modelo se recomienda hacer un buen muestreo, tomando una muestra lo suficientemente grande y representativa de la población. Con esto se logra una garantía probabilística de que el aprendizaje es factible. Es decir, que el error de pruebas está cercano al error de entrenamiento, expresado matemáticamente como se muestra en la Ecuación 2.1.

$$P\left[|v-\mu|>\epsilon\right] \leq 2e^{-2\epsilon^2 N} \tag{2.1}$$

Esta fórmula también llamada desigualdad de Hoeffding, donde v corresponde a la fracción de casos mal clasificados en la muestra (en adelante lo llamaremos error de entrenamiento E_{in}), μ es la fraccion de de casos que el algoritmo clasifica con error (en adelante lo llamaremos error de pruebas E_{out}), N es el tamaño de la muestra, ϵ es un valor que se desea que sea pequeño. Como el objetivo es que la diferencia entre el E_{in} y el $Eout$ sea pequeño, se recomienda aumentar el tamaño de la muestra N. Convirtiendo la desigualdad de Hoeffding en la Ecuación 2.2.

$$P\left[|E_{in}(h)-Eout(h)|>\epsilon\right] \leq 2e^{-2\epsilon^2 N} \tag{2.2}$$

En el aprendizaje automático, la función de predicción la llamaremos f, y a la hipótesis h. Cuando el algoritmo predice correctamente en un punto o para un caso se dice que $f(x) = h(x)$, cuando la predicción es incorrecta $f(x) \neq h(x)$.

Cuando la variable objetivo, la que se está tratando de predecir, es continua se le llama al problema de aprendizaje un problema de regresión. Cuando se predicción esperada está definida por algunos valores discretos (tales como Si o No), lo llamamos un problema de clasificación. Los clasificadores más utilizados son las redes neuronales, las máquinas de soporte vectorial, regresión logística, el algoritmo de los K-vecinos más cercanos, el clasificador bayesiano ingenuo, y los árboles de decisión.

2.2. *Regresión Logística*

La regresión logística ha sido históricamente una herramienta bastante importante y útil para el análisis de datos en investigación clínica y epidemiología. La regresión logística permite discriminar entre dos clases, en términos de un conjunto de variables numéricas, en el papel de predictores [16]. Los objetivos principales de un modelo de regresión logística son:

- Obtener una estimación no sesgada o ajustada de la relación entre la variable dependiente (o resultado) y una variable independiente.
- Evaluar varios factores simultáneamente que estén presumiblemente relacionados de alguna manera (o no) con la variable dependiente.
- Construir un modelo y obtener una ecuación con fines de predicción o cálculo del riesgo.

La regresión logística, a pesar de su nombre, es un modelo lineal para la clasificación en lugar de regresión [16]. La regresión logística es un algoritmo de clasificación lineal ampliamente utilizado en medicina donde una función logística sigmoidea está acoplada a un modelo de regresión lineal. La función utilizada para representar la hipótesis del clasificador está dada por la función sigmoide, cuyos valores de salida son 0 y 1, con una frontera de decisión de 0.5. Los resultados pueden interpretarse como la probabilidad de que la entrada pertenezca a la clase positiva, $p(Y = 1|x; \theta)$, o la negativa $P(y = 0; |x; \theta)$. Donde los resultados se encuentran en el intervalo (0,1), como se muestra en Ecuación 2.3.

$$P(y = 1|x; \theta) + P(y = 0; |x; \theta) = 1 \tag{2.3}$$

Despejando la Ecuación 2.3 se obtiene la Ecuación 2.4 que es la probabilidad de que la entrada pertenezca a la clase negativa.

$$P(y = 0; |x; \theta) = 1 - P(y = 1|x; \theta) \tag{2.4}$$

Cuando la probabilidad de que y sea 1 es mayor de 0.5 entonces el clasificador predice 1. Cuando esta probabilidad es inferior a 0.5 predice 0.

Regresión logística es un algoritmo de clasificación lineal ampliamente utilizado en medicina donde una función logística sigmoidea está acoplada a un modelo de regresión lineal.Para realizar las predicciones el algoritmo utiliza la función sigmoide que se puede ver en la Figura 2.3.

La función sigmoide, $h_\theta(x)$, se basa en unos parámetros θ que son desconocidos y

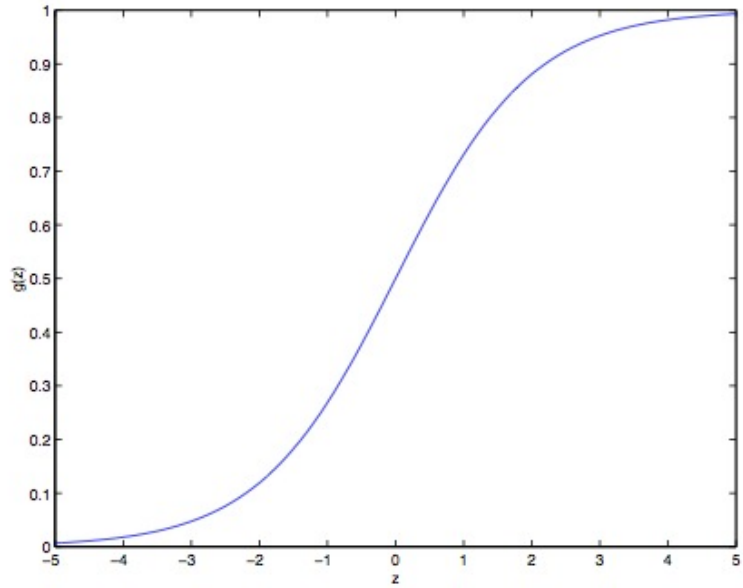

Figura 2.3: Función sigmoide [16]

los valores de cada set de entrenamiento $x_0, x_1, x_2...x_n$. En la Ecuación 2.5 se muestra la fórmula función sigmoide, donde $-\theta^t x$ es el vector de parámetros o pesos de cada una de las variables predictoras.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^t x}} \tag{2.5}$$

Para encontrar la frontera de decisión en problemas que no son linealmente separables se utiliza la función de costos de mínimos cuadrados, esta permite ajustar los parámetros o los θ y lograr encontrar una función que clasifique entre dos clases, en la Ecuación 2.6 se muestra la función de costos.

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{2.6}$$

Donde $h_\theta(x^{(i)})$ corresponde a la predicción del clasificador y $y^{(i)}$ es la respuesta esperada, a la mitad del cuadrado de la diferencia entre ellas se le llama función de

error cuadrático medio, que en adelante se escribirá como *cost*() como se muestra en la Ecuación 2.7.

$$Cost(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{2.7}$$

Para lograr la predicción correcta mediante el uso de la función de *cost*() se hace necesario aplicar el Gradiente Descendente como se muestra en la Ecuación 2.8, este permite encontrar en una función convexa un punto mínimo local óptimo en el plano.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x) & \text{si} \quad y = 1 \\ -\log(1 - h_\theta(x)) & \text{si} \quad y = 0 \end{cases} \tag{2.8}$$

Lo anterior se puede expresar en una linea como se indica en la Ecuación 2.9 :

$$Cost(h_\theta(x), y) = -y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x)) \tag{2.9}$$

Reemplazando la Ecuación 2.7 en la Ecuación 2.6 se consigue la Ecuación 2.10

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} Cost(h_\theta(x^{(i)}), y^{(i)}) \tag{2.10}$$

Finalmente compactando las Ecuaciones 2.10 y 2.9 se consigue la función de costos con gradiente descendente expresada en la Ecuación 2.11.

$$J(\theta) = -\frac{1}{N}[\sum_{i=1}^{N} -y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))] \tag{2.11}$$

Hasta ahora se tiene una función de costos que permite encontrar los valores de los θ para la predicción correcta en un conjunto de datos determinado. Esta función presenta un problema de optimización llamado sobreentrenamiento (*overfitting*), que es el aprendizaje excesivo sobre el set de entrenamiento (memorización) este evita que

se consiga un modelo generalizado que pueda funcionar para cualquier set de datos y no solo en el set de datos que se entrenó. Para resolver este problema de optimización se minimiza la función de costos y se le agrega un factor de regularización que permite controlar la complejidad del modelo. La regularización consiste en reducir la importancia de los parámetros θ viendose modificada la función de costos por la adición de la sumatoria de todos los parámetros θ con un factor llamado parámetro de regularización, λ. Obteniendo como resultado la Ecuación 2.12 [1].

$$min\ -\frac{1}{N}\sum_{i=1}^{N}[y_n \log h_\theta(x) + (1-y_n)\log(1-h_\theta(x))] + \lambda||\theta||^2 \qquad (2.12)$$

Sabiendo que N es el numero de variables, θ son los parámetros de cada variable, y es el vector de respuesta (solo maneja valores binarios (0,1)), y λ es el parámetro de la regularización.

Es posible aumentar las capacidades de clasificador mediante la aplicación de transformaciones polinomiales a las entradas. En cuyo caso, el límite de decisión puede ser no lineal y problemas más difíciles de manejar. En este modelo, las probabilidades que describen los posibles resultados de un único ensayo se modelan mediante una función logística.

2.3. *Máquinas de Soporte Vectorial*

Las máquinas de soporte vectorial en inglés *Support Vector Machine* (SVM) son uno de los métodos de aprendizaje supervisado para problemas de clasificación de dos clases [2]. SVMs tratan problemas linealmente no separables, y buscan separar

los datos con una gran brecha o hiperplano. En la Figura 2.4 se muestra la línea punteada que indica la frontera de decisión, la distancia que existe entre las dos lineas punteadas se llama *margen*, los puntos que caen sobre la frontera de decisión de llaman vectores de soporte. En el ejemplo se puede ver tres puntos (dos ejemplos positivos y uno negativos) que se encuentran en la frontera de decisión.

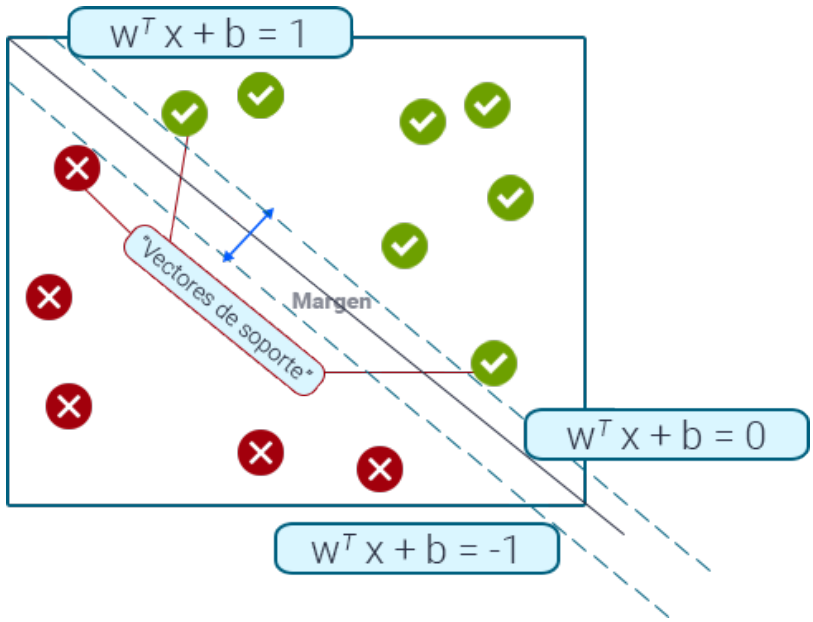

Figura 2.4: Hiperplano de separación óptimo [16].

La función de costo de un SVM busca maximizar el margen o distancia entre la frontera de decisión y los puntos que se desea separar, de esta forma se puede obtener el hiperplano de máximo margen. El hiperplano (frontera) de decisión se define mediante la Ecuación 2.13.

$$w^T x + b = 0 \tag{2.13}$$

Los hiperplanos que caen sobre el margen se muestran en las Ecuaciones 2.14 y 2.15.

$$w^T x + b = 1 \tag{2.14}$$

$$w^T x + b = -1 \tag{2.15}$$

Teniendo en cuenta que el vector de pesos W es perpendicular al plano correspondiente, se define $x^+ = x^- + \lambda w$ como la relación que existe entre del margen M con el vector w para encontrar el hiperplano que maximiza el margen como se muestra en la Figura 2.5.

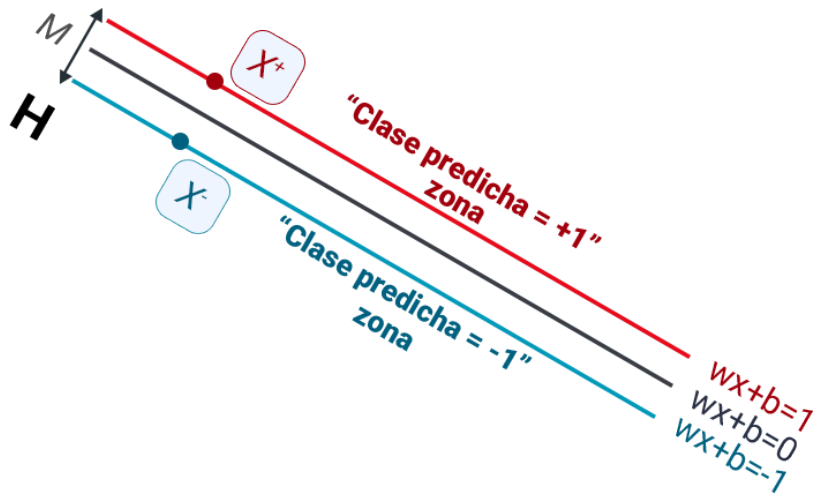

Figura 2.5: Márgenes del hiperplano [16].

El objetivo de un SVM es encontrar el valor de w que que maximice el margen M, conociendo que el margen se define mediante la Ecuación 2.16.

$$M = \frac{2}{||w||} \tag{2.16}$$

El valor del vector w que maximiza el margen M está dado por las Ecuaciones 2.17, 2.18 y 2.19.

$$min \frac{1}{2}||w||^2 \tag{2.17}$$

sujeto a

$$y_i(\vec{x_i} \cdot w + b) \geq 1 \tag{2.18}$$

La anterior formulación asume que los datos son linealmente separables. Cuando no lo son, no es suficiente con maximizar el margen, también se minimizar los errores de clasificación. Lo anterior se convierte en un problema de optimización que se soluciona mediante las Ecuaciones 2.19 y 2.20, donde Los ξ representan variables de holgura, que relajan las restricciones y aportan al objetivo y donde C es una constante lo suficientemente grande, elegida por el usuario, que permite controlar en qué grado influye el término del coste de ejemplos no-separables en la minimización de la norma, es decir, permitirá regular el compromiso entre el grado de sobreajuste del clasificador final y la proporción del número de ejemplos no separables, así, un valor de C muy grande permitiría valores de ξ_i muy pequeños.

$$min w, \xi \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \tag{2.19}$$

sujeto a,

$$y(\vec{x_i} \cdot w + b) \geq 1 - \xi \tag{2.20}$$

Algunas de ventajas de las máquinas de vectores de soporte son:

- Efectiva en espacios de alta dimensión.
- Utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados vectores de soporte), utilizando eficientemente la memoria.

- Versátil: diferentes funciones del núcleo pueden ser especificados para la función de decisión. Núcleos comunes se proporcionan, pero también es posible especificar núcleos personalizados.

Para el caso de los problemas que no son linealmente separables se recomienda la inclusión de las funciones núcleo o *kernel* tiene como efecto un mapeo de las entradas a un espacio de alta dimensionalidad, donde los datos si serán linealmente separables. La función del núcleo tiene como objetivo separar los vectores de soporte del resto de los datos de entrenamiento, este es un problema de programación cuadrática (QP). En este proyecto se usa la función núcleo RBF (*Radial Base Function*) También es conocido como el núcleo "exponencial". La funciones de *kernel* RBF toman la forma $\exp(-\gamma|x - x'|^2) \cdot \gamma$. Donde γ es una constante de proporcionalidad cuyo rango de valores útiles debe ser estimado para cada aplicación en particular. Otras funciones *kernel* comunmente usadas son: *linear*, *polinomial*, *sigmoide*.

Este núcleo es infinitamente diferenciable, lo que implica que GPs con este kernel como función de covarianza tienen las derivadas cuadradas medias de todos los órdenes, y por lo tanto son muy suaves.

2.4. Técnicas de Selección de Variables

La selección de características es un proceso que consiste en seleccionar un subconjunto de variables relevantes para la construcción de un clasificador o predictor. Es frecuente que en un conjunto de datos existan muchos parámetros que tengan

relación con la variable de respuesta, haciendo difícil el análisis y la predicción. Para ello se plantea la reducción del número de parámetros y obtener solo aquellos que representan mayor variabilidad.

El objetivo principal de la selección de variable es mejorar el rendimiento de la predicción del clasificador y esto a su vez proporciona mayor rapidez del clasificador. La seleccción de características puede ser utilizada para la reducción de la dimensionalidad de la función, ya sea para mejorar los resultados de la precisión de los estimadores o para aumentar su rendimiento en conjuntos de datos de muy alta dimensión.

Teniendo en cuenta un estimador que asigna ponderaciones a las características (por ejemplo, los coeficientes de un modelo lineal), la función recursiva de eliminación (RFE) es para seleccionar funciones, haciendo más y más pequeños el set de datos recursivamente hasta conseguir el número correcto de variables que mejoran los resultados del clasificador.

En la Figura 2.6 se muestra el resultado de aplicar el algoritmo de selección de variables, donde se encuentra el mayor pico de la gráfica es donde se encuentra el número óptimo de variables que se deben seleccionar, para este caso son tres.

El algoritmo también retorna una matriz con verdaderos y falsos que indica si la variables es seleccionada o no, como se muestra en la Figura 2.7.

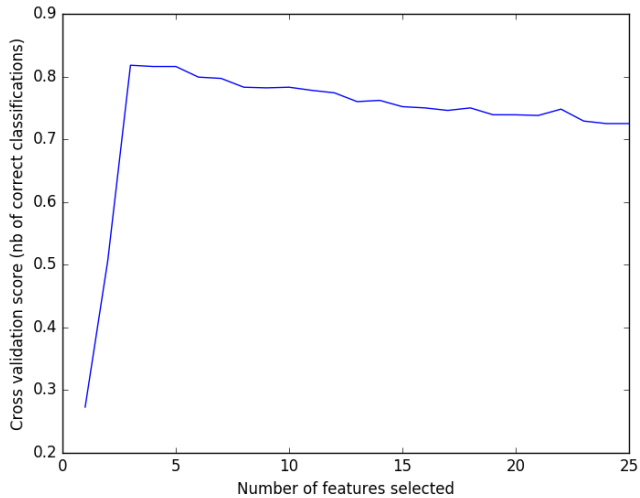

Figura 2.6: Gráfica de selección de variables [19]

```
Número de variables Óptimas : 9
selección de variables:
[False False False  True False False False False False False False False
  True False False False False False False False False False False False
 False False  True False  True  True False  True False False False False
 False False  True False False False False False False False False False
 False  True False False False False False False False  True False False]
Total variables : 60
```

Figura 2.7: Matriz de selección de variables

2.5. *Curva ROC*

En la teoría de detección de señales, una curva ROC (acrónimo de *Receiver Operating Characteristic*, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este gráfico es la representación de la razón o ratio de verdaderos positivos frente a la razón o ratio de falsos positivos [7].

Para comprender la curva ROC es necesario definir los siguientes conceptos:

- **Atributo (campo, variable, característica):** es una cantidad que describe un ejemplo. Un atributo tiene un dominio definido por el tipo de atributo, lo que denota los valores que se pueden tomar por un atributo. Los dominios más comunes son el categórico, nominal, ordinal y continuo. El primero es un número finito de valores discretos; el segundo denota que no hay orden entre los valores; tales como apellidos y colores; el ordinal denota que existe un ordenamiento, tal como en un atributo asumir los valores bajo, medio o alto; el Continuo o cuantitativo es un subconjunto de los números reales, donde hay una diferencia medible entre los valores posibles.

- **Matriz de confusión:** es una matriz que muestra las clasificaciones prevista y las reales. Una matriz de confusión es de tamaño LxL, donde L es el número de diferentes valores de la etiqueta. La matriz de confusión que sigue es para $L = 2$.

Predicho / Real	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

Cuadro 2.1: Matriz de confusión

Donde,

VP: Verdaderos Positivos

VN: Verdaderos Negativos

FP: Falsos Positivos

FN: Falsos Negativos

Los resultados del clasificador se puede interpretar mediante el uso de la gráfica ROC como se muestra en la Figura 2.8.

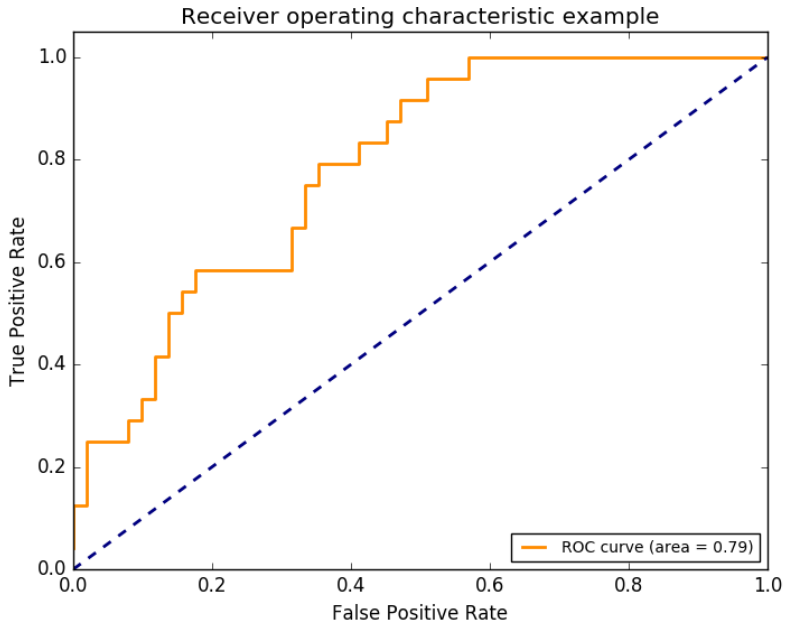

Figura 2.8: Gráfica ROC [7]

- **Sensibilidad (*Recall*)**: también llamada Razón de Verdaderos Positivos (VPR). La sensibilidad indica la capacidad del clasificador para dar como casos positivos los casos realmente positivos (enfermos), útil para detectar pacientes enfermos. La sensibilidad se define mediante la ecuación 2.21.

$$VPR = \frac{VP}{(VP + FN)} \tag{2.21}$$

- ***Precision* (tasa de error):** también llamada *accuracy* en ingles, corresponde a la tasa de predicciones correctas hechas por el modelo sobre un conjunto de datos. La precisión puede verse como una medida de la exactitud o la calidad,

mientras que la sensibilidad es una medida de la integridad. La precisión se suele calcular mediante el uso de un conjunto de pruebas independientes que no fueron utilizada durante el proceso de aprendizaje. Sin embargo para casos en los que el conjunto de datos es pequeño, porque el número de casos es reducido, otras técnicas de estimación de precisión más complejas son utilizadas tales como validación cruzada y rutina de cargas. La formulación matemática para esta medida esta dada porla Ecuación 2.22.

$$precision = \frac{(VP)}{(VP + FP)} \tag{2.22}$$

- **Medida F (F1):** esta medida se considera una medida armónica entre la sensibilidad y la precisión. Es un promedio ponderado entre estas dos medidas como se indica en la Ecuación 2.23 y alcanza su mejor valor cuando tiende a 1 el.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{2.23}$$

- **Especificidad:** también llamada razón de verdaderos negativos. Indica la capacidad del clasificador para dar como casos negativos los casos realmente negativos (sanos). Útil para detectar pacientes sanos. La especificidad se define mediante la Ecuación 2.24.

$$FPR = \frac{VN}{(FP + VN)} \tag{2.24}$$

Generalmente se usan estas medidas para determinar que tan bueno es o no un predictor. En los casos clínicos, por ejemplo, se elige una prueba muy específica

cuando es deseable obtener falsos negativos en lugar de falsos positivos. Se elige una prueba muy sensible cuando se prefiere obtener falsos positivos en lugar de falsos negativos. Es decir, si se desea que el número de enfermos sin detectar sea mínimo. Se elige una prueba muy sensible cuando se prefiere obtener falsos positivos en lugar de falsos negativos. Es decir, quieres que el número de enfermos sin detectar sea mínimo.

2.6. *Scikit-learn*: Aprendizaje Automático en Ptyhon

El lenguaje de programación Python se está estableciendo como uno de los idiomas más populares para computación científica. Gracias a su naturaleza interactiva de alto nivel y su conjunto de librerías, es una opción atractiva para el desarrollo algorítmico y el análisis de datos. *Scikit-learn* es una de las librerías mas populares y usadas que proporciona implementaciones de vanguardia de conocidos algoritmos de aprendizaje, manteniendo una interfaz fácil de usar[19]. Esto responde a la creciente necesidad de análisis de datos de las diferentes disciplinas e industrias, tales como la biología o la física. A continuación se presentan algunas de las ventajas de usar *Scikit-learn*:

- Se distribuye bajo la licencia BSD (que viene del su nombre en inglés Berkeley Software Distribution) que es una licencia de software libre permisiva.
- Incorpora código compilado para eficiencia.
- Depende sólo de la librerías *Numpy* y *scipy* para facilitar la distribución fácil.

- Se centra en la programación imperativa. En esta se describe la programación en términos del estado del programa y sentencias que cambian dicho estado.

2.7. HTCondor

Es un sistema de gestión para la ejecución de tareas por lotes. Al igual que la mayoría de sistemas de proceso por lotes que proporciona un mecanismo de colas, una política de planificación, esquema de prioridades y las clasificaciones de los recursos.

El funcionamiento de este sistema consiste en que los usuarios envíen trabajos ya sea en serie o paralelos a HTCondor. HTCondor los coloca en una cola, decide cuándo y dónde ejecutar los trabajos en base a una política. Monitorea cuidadosamente su progreso, y en última instancia, informa al usuario sobre la terminación[14]. En la Figura 2.9 se muestra la interacción entre los elementos principales de HTCondor que son: nodo maestro, nodo trabajador y universo de ejecución.

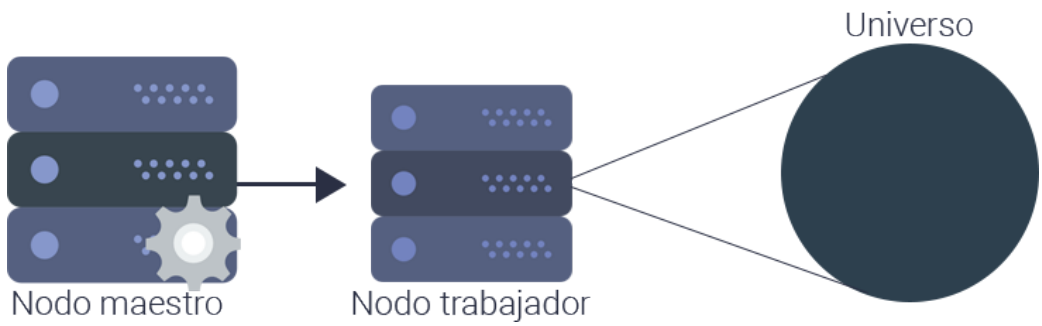

Figura 2.9: Elementos de HTCondor.

El **nodo maestro** es el encargado de administrar y coordinar las tareas que van a ser ejecutadas en el cluster. Esto quiere decir, que es capaz de determinar cuáles nodos esclavos se encuentran disponibles o cuales son ideales para ejecutar una tarea.

El **nodo trabajador** este tipo de nodo es el que está en capacidad de recibir y ejecutar las tareas enviadas por el nodo *submit*.

Para Condor el **universo de ejecución** se define como un conjunto de características que marcarán el entorno de ejecución del trabajo enviado y este depende del tipo de aplicación que se desea ejecutar. Los universos son: *standard*, *vanilla*, *grid*, *java*, *parallel*.

El universo ***standard*** provee confiabilidad y migración de las tareas por medio del mecanismo de *checkpoint*, para que un programa pueda ser enviado por medio de este universo, ha debido ser compilado con las librerías de HTCondor. El universo ***vanilla*** es el universo por defecto si no se especifica, aunque el administrador puede configurar HTCondor para usar otro universo, es el menos restrictivo con los programas que se puedan enviar, ya que acepta cualquier programa escrito en cualquier lenguaje de programación. El universo ***grid*** permite a los usuarios enviar trabajos usando la interfaz de HTCondor, estos trabajos son enviados para ser ejecutados sobre recursos que se encuentren en una grid. El universo ***java*** permite enviar trabajos escritos en este lenguaje. Finalmente el universo ***parallel*** es para enviar programas que requieran múltiples máquinas para ejecutar un solo trabajo o aplicación paralela, para utilizarlo se debe hacer uso del estándar MPI (Message Passing Interface).

En este proyecto se usó el universo *vanilla* de HTCondor, para ejecutar el algoritmo que realiza el proceso de entrenamiento de los clasificadores.

2.8. SQL Server

Microsoft SQL Server es un sistema de manejo de bases de datos del modelo relacional, desarrollado por la empresa Microsoft [23]. El lenguaje de desarrollo utilizado es Transact-SQL (TSQL), una implementación del estándar ANSI del lenguaje SQL, utilizado para manipular y recuperar datos (DML: Lenguaje de Manipulación de Datos), crear tablas y definir relaciones entre ellas (DDL: lenguaje de definición de datos). Algunas de las características de SQL con:

- Soporte de transacciones.
- Soporta procedimientos almacenados.
- Incluye también un entorno gráfico de administración, que permite el uso de comandos DDL y DML gráficamente.
- Permite trabajar en modo cliente-servidor, donde la información y datos se alojan en el servidor y los terminales o clientes de la red sólo acceden a la información.
- Además permite administrar información de otros servidores de datos.

Capítulo 3
Estado del Arte

Durante la última década se han realizado estudios para determinar los factores de riesgo relacionados con las morbilidades maternas y fetales, y la forma de cómo reducir los casos de MME. Estos estudios arrojan resultados similares con respecto a las principales enfermedades que originan la morbilidad, como son la preeclampsia, eclampsia y hemorragia obstétrica.

Por ejemplo en estados Unidos, debido al aumento progresivo de los casos de morbilidad materna extrema y mortalidad materna [11], el Departamento de Obstetricia y Ginecología de la Universidad de Columbia Colegio de Médicos y Cirujanos se reunió en 2012 para determinar métodos para ayudar a mejorar los resultados maternos. Como consecuencia de esto se realiza un trabajo cuyo objetivo es estimar la frecuencia de morbilidad materna severa y evaluar sus causas. En este estudio las mujeres fueron clasificados con la morbilidad materna severa de acuerdo con un sistema de puntuación [9] que tiene en cuenta la ocurrencia de transfusión de glóbulos rojos (más de tres unidades), la intubación, la intervención quirúrgica no anticipado, insuficiencia orgánica, y la admisión unidad de cuidados intensivos. Se calculó la frecuencia de morbilidad materna severa y las causas fundamentales. El

análisis multivariable identificó los factores del paciente presentes en la admisión que se asociaron de forma independiente con la morbilidad materna severa. Estos fueron utilizados para desarrollar un modelo de predicción de la morbilidad materna severa [10].

Otro estudio importante es "Hypertensive Disorders and Severe Obstetric Morbidity in the United States" cuya finalidad era examinar las tendencias en las tasas de trastornos hipertensivos en el embarazo [13]. Dentro de estos estudios más destacados en Latinoamérica se menciona el plan de acción para acelerar la reducción de la mortalidad materna y la morbilidad materna grave propuesto por la Red Centro Latino Americano de Perinatología (CLAP) [14]. Finalmente, "Morbilidad materna extrema en el Hospital General Dr. Aurelio Valdivieso" en México. cuyo objetivo fue la identificación de las principales determinantes de la morbilidad obstétrica extrema [8].

Por otro lado, desde un enfoque de ingeniería nos encontranos con frecuencia proyectos que usan aprendizaje automático. Especialmente para resolver problemas de clasificación en la medicina. A continuación se mencionan algunos estudios de predicción de al menos una de las principales enfermedades asociadas a la MME.

En Poon et al [20], los autores presentan un trabajo de predicción temprana de los trastornos hipertensivos durante el primer trimestre de embarazo, mediante la combinación de variables maternas, incluyendo la presión arterial media, índice de pulsatilidad de la arteria uterina, proteína plasmática asociada al embarazo y el factor

de crecimiento placentario en el embarazo temprano.

El estudio cohorte fue realizado en la ciudad de Londres, Reino Unido. La población estuvo constituida por 7.797 embarazos únicos, incluyendo 34 pacientes estudiados que desarrollaron preeclampsia (PE) que requerían el parto antes de las 34 semanas (PE temprana) y 123 con PE tardía, 136 con hipertensión gestacional y 7.504 casos de pacientes (96,3 %) que no se vieron afectadas por PE o hipertensión gestacional.

La PE temprana y tardía PE se asociaron con un aumento de la presión arterial media (1,15 y 1,08) y el índice de pulsatilidad de la arteria uterina (1,53 y 1,23) y la disminución asociada al embarazo proteína plasmática A (0,53 y 0,93) y el crecimiento de la placenta los factores (0,61 y 0,83). Se utilizó un análisis de regresión logística para derivar algoritmos para la predicción de los trastornos hipertensivos. Se estimó, con el algoritmo para la EP temprana, que el 93.1 % corresponde a PE, 35.7 % PE tardia y 18.3 % correspondiente a la hipertensión gestacional. Se pudo detectar con una tasa de falsos positivos del 5 % y que 1 de cada 5 embarazos clasificadas como posibles casos positivas que desarrollaría la hipertensión del embarazo. Este método de detección es muy superior al enfoque tradicional, que se basa enteramente en la historia materna. La regresión logística se utilizó para obtener una tasa de detección alrededor 90 % de preeclampsia temprana, a la tasa de falsos positivos de 5 %.

En Park et al [18], los autores presentan un algoritmo de regresiones logísticas múltiples para predecir el riesgo de preeclampsia en una población australiana. El

modelo predice el riesgo utilizando el tipo de población de la preeclampsia, una variedad de factores demográficos, la presión arterial media materna (MAP), la arteria uterina PI (UTA PI) y la proteína plasmática a asociada al embarazo (PAPP-A), se logra predecir la aparición temprana de preeclampsia en el 95 % de las mujeres a una tasa de falsos positivos del 10 %. Para este estudio Un total de 3.099 mujeres fueron examinadas hasta el parto. 3.066 (98,9 %) mujeres tenían todos los datos para llevar a cabo la investigación del pre-eclampsia disponible. Esto incluyó 3.014 (98,3 %) mujeres con un nacido vivo, donde se calcularon riesgos respecto a la pre-eclampsia. 12 mujeres tuvieron el parto antes de las 34 semanas a causa de la preeclampsia temprana, con una prevalencia de preeclampsia temprana de 1 de cada 256 embarazos.

En Nanda et al [15], los autores presentan un modelo para la predicción de diabetes mellitus gestacional (DMG) en el embarazo durante el primer trimestre , basado en los bio marcadores y algunas características maternas. En el estudio la edad materna, índice de masa corporal, origen racial, antecedentes de DMG y el recién nacido con macrosomía fueron predictores independientes significativos de futuro DMG. En comparación con los controles, el múltiple promedio de la adiponectina mediana normal (0,66; IQR: 0,5-0,9 vs 1,02; IQR: 0,7 a 1,29) y fijadora de hormonas sexuales (SHBG) (0,81; IQR: 0,6 a 1,04 vs 1,02; IQR: 0,8 -1.2) fue menor (P ¡0,05), pero follistatina-como-3 (FSTL3) no fue significativamente diferente. En la detección de DMG por las características maternas, la tasa de detección fue de 61,6 % a una tasa de falsos positivos del 20 %, y la detección aumentó a 74,1 % mediante la adición de

la adiponectina y la SHBG.

En Farran et al. [6], se construyeron varios modelos de clasificación como herramienta para evaluar el riesgo de diabetes, hipertensión y comorbilidad mediante algoritmos de aprendizaje automático en los datos de los centros de atención primaria y hospitales en Kuwait. Se modeló el aumento del riego de que los pacientes diabéticos desarrollaran hipertensión y viceversa. Se comprobó la importancia de la etnia (y nativos vs migrantes extranjeros) y de la utilización de los datos regionales en la evaluación de riesgos. Para este estudio de cohorte retrospectivo, los autores utilizaron cuatro técnicas de aprendizaje automático: regresión logística, k-vecinos más cercanos (k-NN), la reducción de dimensionalidad multifactorial y máquinas de soporte vectorial. El estudio utiliza validación cruzada de cinco veces para obtener una precisión de generalización y errores. Como resultado de este estudio se obtuvo que una precisión de clasificación de $> 85\,\%$ (para la diabetes) y $> 90\,\%$ (para la hipertensión) utilizando parámetros simples no basados en laboratorio. Además, se identifica que las herramientas de evaluación de riesgos sobre la base de modelos de clasificación k-NN son capaces de asignar "alto" riesgo de $75\,\%$ de los pacientes diabéticos y el $94\,\%$ de los pacientes hipertensos. Sólo el $5\,\%$ de los pacientes diabéticos se les asigna riesgo "bajo".

De acuerdo a la literatura revisada, se evidencia que los resultados obtenidos de la aplicación de aprendizaje automático para problemas de clasificación en medicina son bastante satisfactorios. Se encuentran trabajos a nivel internacional relacionados con

la detección temprana de patologías como la diabetes, hipertensión y preeclampsia durante el embarazo. Estas predicciones están basadas en resultados de laboratorios clínicos y patológicos. Algunos de los estudios realizados a nivel internacional se destacan algunos que se mencionaron previamente, estos propone el uso de técnicas como la regresión logística y/o máquinas de soporte vectorial para la predicción de una patología obstétrica en específica, basada los valores de los resultados de la exámenes de laboratorio.

A nivel regional no se evidencian trabajos de predicción temprana de MME, razón por la cual se decide desarrollar un modelo de predicción para morbilidad materna extrema usando técnicas de aprendizaje supervisado para la población de Bolívar que se atiende en la E.S.E Clínica de Maternidad Rafael calvo, teniendo como entradas los factores socio demográficos, antecedentes personales obstétricos y patológicos, antecedentes quirúrgicos, y enfermedades desarrolladas durante el embarazo.

Capítulo 4
Propuesta

4.1. Metodología

Para la formulación y desarrollo de esta investigación se decidió trabajar en la metodología de investigación presentada por Vaishnavi y Kuechler [25] para proyectos de Tecnologías de la Información y Comunicación - TIC. La metodología consiste en una serie de pasos que permitirá llevar a cabo el proyecto de investigación que se propone. Se inicia con la identificación del problema, la propuesta de investigación, el desarrollo y la implementación del diseño propuesto, la evaluación de los artefactos (medidas de desempeño), la presentación de resultados y las conclusiones.

4.2. Descripción de la Metodología

A continuación se detallan los aspectos metodológicos más relevantes de cada paso del proceso utilizado para el cumplimiento de los objetivos planteados.

Para cumplir con el primer objetivo que es la identificación de las variables que se asocian a la ocurrencia de la morbilidad materna extrema, se realizaron las siguientes actividades:

- Entrevistas con algunos ginecólogos y con la epidemióloga de la ESE clínica para determinar cuáles son el grupo de factores de riesgo que deben tomarse como variables de interés para el proyecto.

- Comparación de las variables propuestas por los médicos, las variables relacionadas en protocolo de morbilidad materna extrema emitido por el ministerio de salud en Colombia en el año 2016, y la definición del grupo de factores de riesgo de la Red Latino Americana de Perinatología.

- Documentación con la definición operativa de cada una de las variables seleccionadas

- Se clasifican las variables en dos grupos los antecedentes prenatales y las patologías que ocurren durante la gestación (controles prenatales). Como se aprecia en la Figura 4.1.

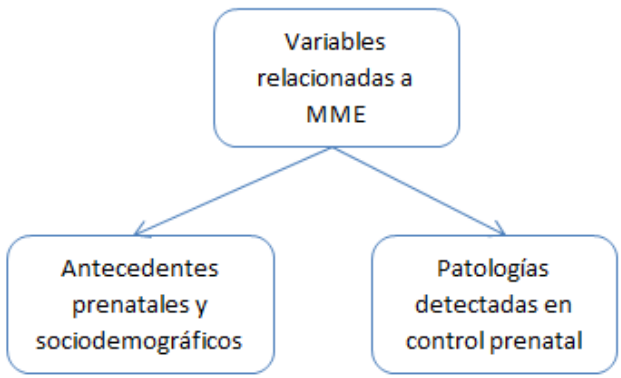

Figura 4.1: Grupos de variables

Para lograr un mecanismo de captura de los datos requeridos para el proyecto, que

corresponde al segundo objetivo se siguen los pasos que se mencionan a continuación.

- Construcción de dos formularios de Google, uno para diligenciamiento de los antecedentes y otro para ingreso de los diagnósticos obtenidos en cada control prenatal.

- Selección de la población que corresponde al grupo de pacientes que se realizaron al menos un tres controles prenatales con la Clínica y que además concluyeron la gestación en esta misma institución.

- Definición del tipo de muestro y determinación del tamaño de la muestra. Esta muestra debe incluir las pacientes que presentaron morbilidad materna extrema y pacientes que no presentaron morbilidad.

- Revisión de las historias clínicas las historias clínicas de control prenatal y diligenciamiento de los datos requeridos para el estudio, en los formularios de google diseñados.

- Depuración de la información consignada a través de los formularios, se construyó la base de datos en la que se subió toda la información capturada en los formularios, esto facilitó la extracción de los datos.

Para lograr el tercer objetivo, que consiste en determinar el subconjunto de variables que tienen incidencia en la ocurrencia de MME de acuerdo a la datos estudiados, se siguieron las siguientes pautas.

- Obtener los datos de la base de datos con el formato correcto para el análisis estadístico, dividiendo los datos por trimeste, es decir, segmentando las consultas del primer trimestre gestacional del segundo y el tercero.

- Realizar un análisis haciendo uso de la estadística descriptiva como son el análisis de frecuencia frecuencia y Anova.

- Filtrado de variables mediante la técnica de selección de variables que permite determinan la correlación entre atributos y la variable que se desea predecir. Como resultado de esto se tendrá el subconjunto de variables que están estrechamente relacionadas con MME y que permitirán maximizar el desempeño del clasificador. Esto contribuye sensiblemente al buen desempeño del clasificador y eliminación del ruido. El proceso de filtrado se debe usar el 80 % de la muestra.

En cumplimiento del objetivo de Implentación de algoritmos de aprendizaje supervisado automático para obtener el nivel de precisión y sensibilidad deseados se realizaron las siguientes actividades:

- Filtrado de variables con las técnicas selección de variables usando regresión logística y máquinas de soporte vectorial, obteniendo como resultado el listado de las variables predictoras.

- Implementación de regresión logística y máquinas de soporte vectorial realizando el proceso de entrenamiento con el 80 % del total de la muestra, evaluando

sólo las variables que tienen relación con la MME. Se validaron los resultados de ambos clasificadores en términos de validación cruzada y error de pruebas.

- Comparación y validación de los modelos resultantes para cada set de datos con cada una de las técnicas seleccionadas, con el 20 % de la muestra restante, estos resultados se evaluaron en términos de precisión y sensibilidad para determinar con cuál de los modelos realiza una clasificación o predicción más acertada.

Para cumplir con el último objetivo que es la construcción de un conjunto de pruebas para validar el modelo de predicción, se ejecutaron las siguientes actividades:

- Se obtuvieron datos de pacientes nuevos, aproximadamente 20 pacientes, tomando sólo los parámetros requeridos por el modelo para la predicción.

- Validación de los resultados con el ginecólogo de la E.S.E Clínica, quienes definen si la herramienta es útil como ayuda en la detección temprana del riesgo de MME de las pacientes de la clínica.

Capítulo 5
Implementación

5.1. Elementos de la Muestra

Este trabajo fue un estudio de cohorte retrospectivo, realizado en la E.S.E Clínica de Maternidad Rafael calvo Castaño, en la ciudad de Cartagena, donde se atienen alrededor de 6.000 y 8.000 cesáreas anualmente.

A continuación se indican los criterios de inclusión de la población seleccionada para este estudio en el Cuadro 5.1.

Cuadro 5.1: Criterios de inclusión

Criterio	Descripción
Género	Femenino
Edad	13 Hasta 45 años
Embarazo Controlado	Si
Número Controles	2 en adelante
Terminación del embarazo	En CMRC

Para un total de 1.338 pacientes entre 2015 y 2016 que cumplían con los criterios de inclusión del proyecto. Para la selección de la muestra se clasificaron los pacientes en dos grupos: pacientes que presentaron MME y pacientes que no tuvieron MME. El primer grupo corresponde a las pacientes que presentaron MME se les confirmó la

Morbilidad teniendo en cuenta los lineamientos del protocolo de vigilancia en salud pública de MME en Colombia. Muchos de los casos de MME fueron reportados al SIVIGILA [4], pero otros se identificaron posteriormente. El segundo grupo son las pacientes que concluyeron su embarazo sin complicaciones obstétricas, es decir, que no tuvieron Morbilidad Materna extrema.

Como el grupo de pacientes con MME representaba un porcentaje bajo de la población, no se tomó aleatorio simple porque con estas condiciones lo mas posible era que la muestra no fuera significativa para el trabajo de investigación. Por tal motivo se decide aplicar un muestreo mixto. Para el primer grupo, se utilizó muestreo aleatorio, y para el segundo grupo, se utilizó el muestreo de conveniencia.

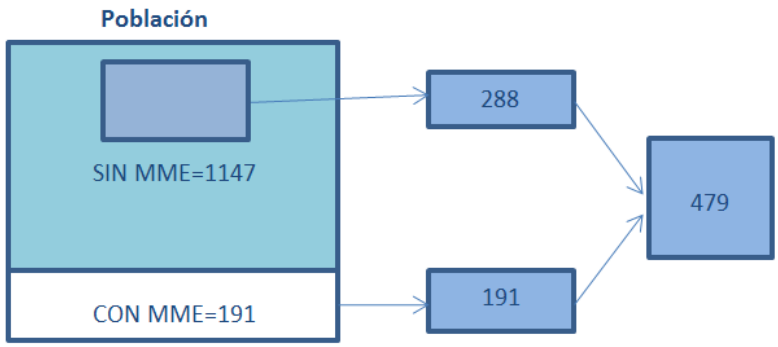

Figura 5.1: Muestreo mixto

En la Figura 5.1 se muestra que de los 1.147 Paciente Sin MME entre 2015 y 2016 se hace el muestreo aleatorio simple con un nivel de confianza del 95 %, un margen de error del 5 % y se obtiene una muestra de 288. La muestra de los pacientes con MME se toma por conveniencia el total de la población, es decir 191 pacientes. Quedando finalmente una muestra de 479 pacientes.

5.2. Construcción de la Base de Datos

Se realizó un análisis retrospectivo de los datos. Se obtuvieron los datos de las pacientes de la muestra, mediante la revisión de las historias clínicas de los controles prenatales. En algunos caso no fue posible obtener toda la información del software porque el formato de este registro clínico no es lo suficientemente específico en lo que refiere a datos importantes para MME.

Se contó con un grupo de estudiantes de medicina que hicieron la lectura de cada historia clínica para organizar los datos que se requerían para el estudio. La construcción del modelo de aprendizaje automático se basó en características o factores de riesgo. Estos factores fueron seleccionados de acuerdo con la caracterización del factor de riesgo descrita por el Centro Latinoamericano de Perinatología (CLAP)[22], en comparación con el protocolo MME de 2015 del Ministerio de Salud y Protección Social de Colombia) [4].

Este listado de variables fueron suministradas por el área de vigilancia en salud pública y el Ginecólogo- Obstetra del centro de investigación de la E.S.E clínica de Maternidad Rafael Calvo. En el Cuadro 5.2 se detallan las variables para la caracterización de la población objeto de estudio, denominadas variables socio demográficas, En el Cuadro 5.3 y 5.4 se indican los datos ginecológicos y antecedentes clínicos que son indicados como variables de riesgo.

Para la recolección de los datos se usó la herramienta de Google Forms para la creación de dos formularios en línea. El primer formulario correspondiente a la

Cuadro 5.2: Variables socio demográficas

Variable	Opciones
Edad	13-45 años
Etnia	Indígena, Palequero, Raizal, Negro, Mulato,Afrodescendiente, Otro
Escolaridad	Primaria, Secundaria,Técnica,
Estrato Socio Económico	Estrato 1, Estrato 2, Estrato 3, Estrato 4, Estrato 5,Estrato 6, Desconocido
Régimen	Contributivo, Subsidiado
Procedencia	Cabecera Municipal, Centro o poblado, Zona rural
Estado Civil	Soltera, Casada, Union libre, Separada, Viuda

Cuadro 5.3: Datos ginecológicos

Variable	Opciones
Paridad	Nulípara, Multípara
Multiplicidad	Simple, Doble, Triple o mas
Edad Gestacional	Trimestre 1,Trimestre 2, Trimestre 3
Periodo intergenésico $<$ 2 Años	Si, No
Micronutrientes	Si,No

información socio demográfica y antecedentes clínicos de las pacientes, el segundo formulario se usó para almacenar los diagnósticos y la edad gestacional descritos por el médico en cada control prenatal. Los valores aceptados para los diagnósticos son los descritos por la codificación de la Clasificación Internacional de Enfermedades (CIE-10)[3].

La información consignada en estos formularios se almacenó en hojas de cálculo Online que fueron consolidados en un solo archivo, seguido de esto se hizo un filtrado

Cuadro 5.4: Antecedentes ginecológicos

Variable	Opciones
Preeclampsia	Si,No
Eclampsia	Si,No
Ruptura uterina	Si,No
Aborto séptico	Si,No
Sepsis o infección sistémica severa	Si,No
Hemorragia obstétrica severa	Si,No
Embarazo ectópico	Si,No
Enfemedad hematológica	Si,No
Enfermedad Oncológica	Si,No
Enfermedad renal	Si,No
Enfermedad gastrointestinal	Si,No
Enfermedad eventos tromboembólicos	Si,No
Enfermedad cardiocerebrovascular	Si,No
Diabetes	Si,No
TORCHS	Si,No
Alcoholismo, tabaquismo o PSA	Si,No
Infección de vías urinarias	Si,No
Enfermedad auto inmune	Si,No

de la información para evitar datos inconsistentes. Una vez validada la información se ingresó en una base de datos SQL SERVER R2 Express con una estructura de cuatro tablas, como se indica en la Figura 5.2.

Para extraer los datos en el formato requerido por los algoritmos se utilizó el lenguaje de búsqueda estructurado (Transact SQL). La consulta se puede ver en el Anexo 1.

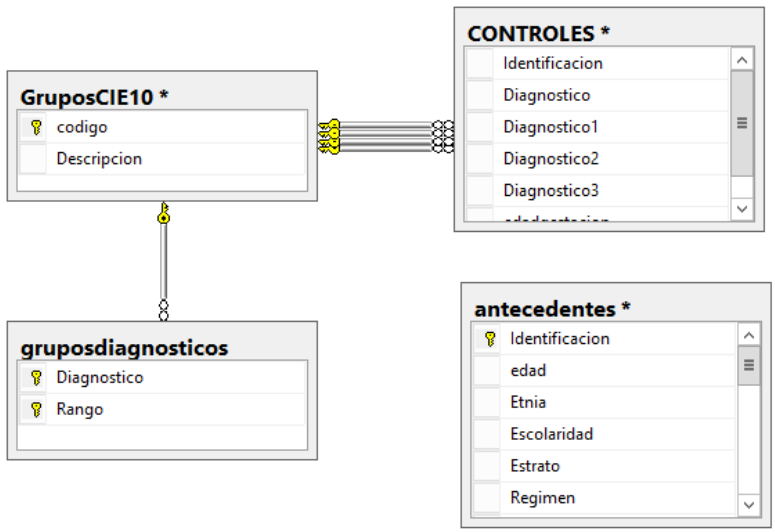

Figura 5.2: Diagrama relacional de base de datos de MME

5.3. Selección de Características

la selección de características es un método que se utiliza para la reducción de las variables en el conjunto de la muestra, para mejorar los resultados del clasificador y/o aumentar el rendimiento del mismo.

Se utilizó la función RFECV de la librería *Sklearn* de Python para la eliminación de las variables. Teniendo en cuenta el objetivo del trabajo que es la detección temprana de la MME se dividen los datos de los controles prenatales en dos grupos, el primero que corresponde a los controles prenatales del primero segundo trimestre del embarazo y el segundo correspondiente al tercer trimestre.

5.3.1. Selección de características con Regresión Logística

El algoritmo usado carga el archivo con los datos y utiliza el 80 % del total de la muestra de forma proporcional. Como el conjunto de diagnósticos con códigos CIE-10 [5] recolectados en las historias de control prenatal era tan grande se deciden agrupar de acuerdo a la categoría superior más cercana. De esta forma se reduce la cantidad de variables a evaluar.

La función RFECV recibe como parámetros:

1. El estimador: regresión logística regularizada con penalización L2, balanceo de clases que equilibra la clase negativa con la positiva.

2. Pasos: eliminación de una variable por iteración.

3. Validación cruzada: estratificada con 5 pliegues, es decir, se divide de forma homogénea el conjunto de datos en 5 subconjuntos.

4. Puntuación: *roc_auc* métrica de calidad para clasificación con validación Cruzada. Calcula el área bajo la curva.

Para este proceso se implementó el algoritmo para la selección de las variables del Anexo 2.

En el conjunto de datos del primer y segundo trimestre se contaba con una muestra de 113 controles prenatales de pacientes que terminaron en MME también llamada clase positiva y 110 controles de pacientes sin MME denominada clase negativa.

Para la selección de variables se utilizaron los datos del primer y segundo trimestre. El resultado obtenido con este método es que 9 de 60 variables se recomiendan como predictores o parámetros para el clasificador que se desea construir, como se puede ver la Figura 5.3, el pico mas alto de la gráfica indica que con esa cantidad de variables se consigue el mejor resultado. El algoritmo entrega como resultado una matriz de falsos y verdaderos que permite saber cuales son las once variables seleccionadas. En el Cuadro 5.5 puede se muestra el nombre de las variables seleccionadas.

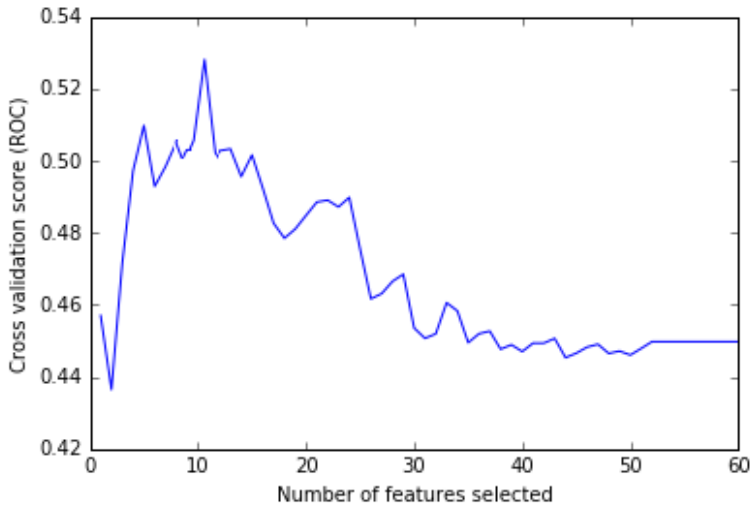

Figura 5.3: Gráfica de RFECV para 1^{er} y 2^{do} trimestre

El conjunto de datos del tercer trimestre contaba con una muestra de 434 controles prenatales de pacientes, de las cuales 165 terminaron en MME (clase positiva) y 269 controles de pacientes sin MME (clase negativa). El mismo algoritmo de selección de variables del Anexo 2, se aplica para los datos de los controles prenatales del tercer trimestre. En la Figura 5.4 se puede observar el resultado de *features selection*.

Cuadro 5.5: Variables predictoras 1^{er} y 2^{do} trimestre RL

Clasificación	Variable
Dato Personal	Edad
Étnia	Palenquero
Régimen	Contributivo
Antecedente	Preeclampsia
Antecedente	Eclampsia
Antecedente	Diabetes
Antecedente	Infección de vías urinarias
Diagnóstico	E20-E35: Trastornos de otras glándulas endocrinas
Diagnóstico	O20-O29: Enfermedades infecciosas del tracto respiratorio
Diagnóstico	N30-N39: Enfermedades que pueden afectar al feto
Diagnóstico	Z30-Z39: Atención médica para la reproducción

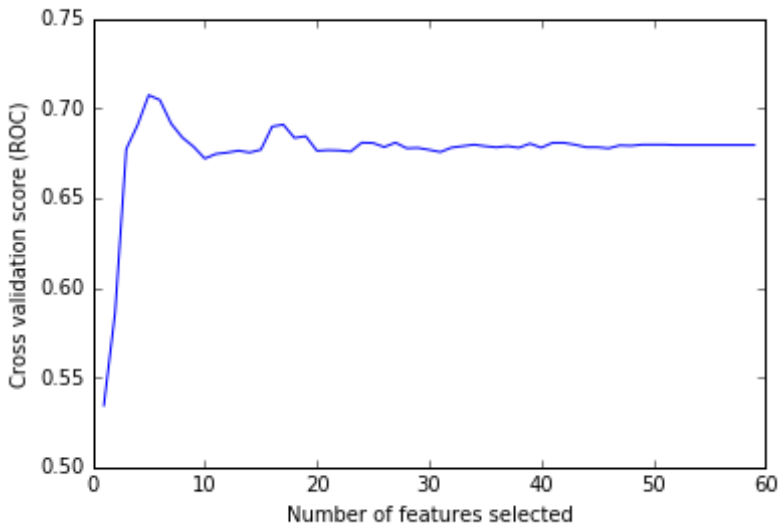

Figura 5.4: Gráfica de RFECV para 3^{er} trimestre

Para este caso el algoritmo seleccionó cinco variables como predictores de 59 variables como se detalla en el Cuadro 5.6

Cuadro 5.6: Variables predictoras 3^{er} trimestre RL

Clasificación	Variable
Dato Personal	Multiplicidad en embarazo
Antecedente	Infección de vías urinarias
Diagnóstico	O10-O16: Edema proteinuria e hipertensión
Diagnóstico	O30-O48: Complicaciones del embarazo que requieren una atención a la madre
Diagnóstico	O60-O75: Complicaciones del embarazo y del parto

5.3.2. Selección de Características con Máquinas de Soporte Vectorial

De la misma forma que el algoritmo de selección de variables, este recibe un archivo con la estructura requerida por la librería de *sklearn*. El Estimador utilizado para la selección de variables es SVM. El algoritmo puede verse en el Anexo 3. Se utiliza el mismo algoritmo con una variación en el parámetro del clasificador, para este caso fue seleccionado como estimador las máquinas de soporte vectorial lineal generalizado. Se aplicó el algoritmo tanto para el primer y segundo trimestre como para el tercer trimestre. Como se resultado de los primeros se obtuvo el grupo de características o variables predictoras como se indica en Figura 5.5.

Para un total de 60 variables el algoritmo hace la eliminación quedando tan solo 15 variables como se indica en el Cuadro 5.8.

Se aplica el nuevamente el algoritmo de selección de características, pero ahora tomando el set de datos del tercer trimestre con 434 casos, de este se toma el 80 %. En la Figura 5.6 se observa el mayor pico cuando el número de variables está entre

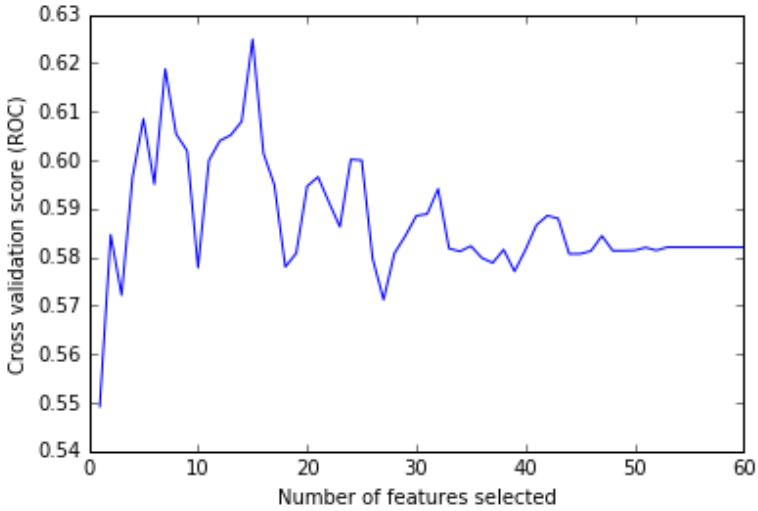

Figura 5.5: Gráfica de RFECV para 1^{er} y 2^{do} trimestre SVM

Cuadro 5.7: Variables predictoras 1^{er} y 2^{do}trimestre SVM

Clasificación	Variable
Dato Personal	Edad
Étnia	Raizal
Estrato socioeconómico	Estrato
Estado marital	Unión libre
Estado marital	Viuda
Dato ginecológico	Paridad
Embarazo multiple	Multiplicidad
Antecedente	Preeclampsia
Antecedente	Eclampsia
Antecedente	Infección de vías urinarias
Diagnóstico	I11-I15: Enfermedad hipertensiva
Diagnóstico	J00-J06: Enfermedades infecciosas del tracto respiratorio
Diagnóstico	N30-N39: Otras enfermedades del sistema urinario
Diagnóstico	O10-O16: Edema proteinuria e hipertensión
Diagnóstico	O20-O29: Enfermedades que pueden afectar al feto

10 y 20.

Para un total de 59 variables el algoritmo hace la eliminación de 45 variables y

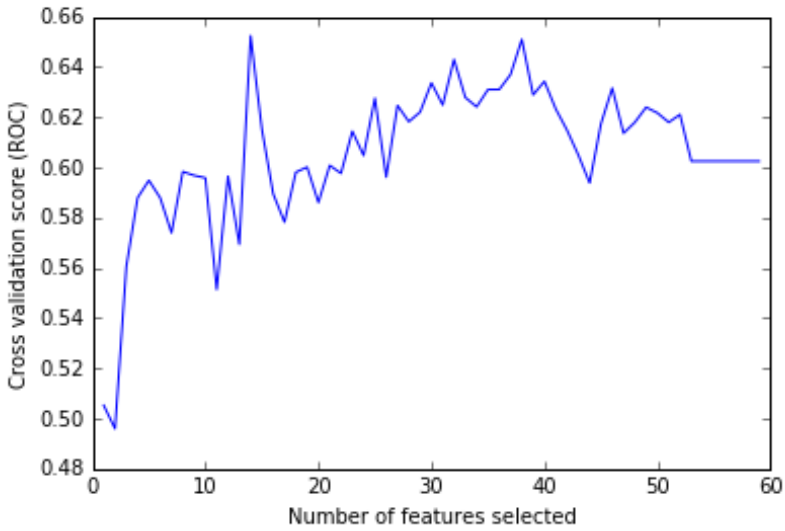

Figura 5.6: Gráfica de RFECV para 3^{er} trimestre SVM

solo se recomiendan 14 como se muestra en el Cuadro 5.8.

Cuadro 5.8: Variables predictoras 3^{er} trimestre SVM

Clasificación	Variable
Étnia	Negro
Régimen	Contributivo
Régimen	No afiliado
Régimen	Subsidiado
Procedencia	Cabecera municipal
Procedencia	Centro o poblado
Estado marital	Separada
Antecedente	Eclampsia
Antecedente	Infección de vías urinarias
Diagnóstico	E00-E07: Trastornos de la glándula tiroides
Diagnóstico	O10-O16: Edema proteinuria e hipertensión
Diagnóstico	O60-O75: Complicaciones del embarazo y del parto
Diagnóstico	Z30-Z39: Atención médica para la reproducción

5.4. Selección de la Técnica

El objetivo de esta fase es identificar cuál es la técnica que mejor se adapta al conjunto de datos, tomando como base para la selección el mejor resultado de la validación cruzada y el menor porcentaje de error en el proceso de pruebas.

5.4.1. Regresión logística

Con el listado de las características seleccionadas que representan variabilidad en la ocurrencia de MME, se genera el archivo final para el entrenamiento del modelo utilizando como estimador la técnica de regresión logística, implementado mediante el uso del paquete *LogisticRegression* de la librería *Sklearn* en Python.

El análisis de los datos se realizó usando el 80 % de los datos, distribuidos de forma proporcional la clase positiva y la negativa. El primer set de datos analizado fue el de primer y segundo trimestre, con nueve parámetros como variables predictoras para el clasificador. Los resultados obtenido se pueden ver en el Cuadro 5.9.

Cuadro 5.9: Métricas para regresión logística 1^{er} y 2^{do} trimestre

Métrica	Resultado
Validación cruzada	63 %
Error de Pruebas	48 %

En la Figura 5.7 se muestra la curva de aprendizaje del clasificador con regresión logística para el conjunto de datos del primer trimestre. Este tipo de gráficas muestran que la puntuación de entrenamiento y la puntuación de validación cruzada son muy

buenas al final. Sin embargo, el puntaje de entrenamiento es muy alto al principio pero disminuye, mientras que la puntuación de validación cruzada es muy baja al principio y luego aumenta.

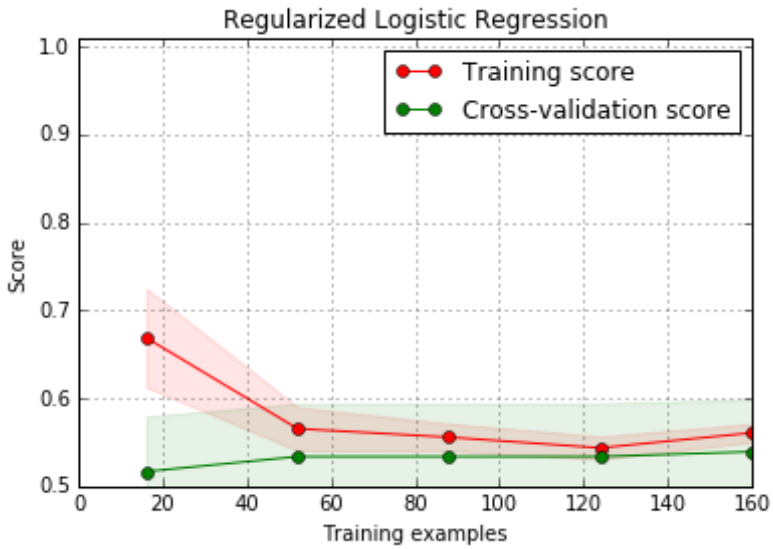

Figura 5.7: Curva de aprendizaje del clasificador regresión logística 1^{er} y 2^{do} trimestre

De la misma forma se aplica el algortimo de regresión logística al set de datos de tercer trimestre con cinco parámetros, el resultado se muestra en el Cuadro 5.10.

Cuadro 5.10: Métricas para regresión logística 3^{er} trimestre

Métrica	Resultado
Validación cruzada	72 %
Error de Pruebas	27 %

En la Figura 5.8 se muestra la curva de aprendizaje del clasificador para el set de datos del tercer trimestre, muestra como notoriamente el puntaje del entrenamiento mejora en la medida que hay mas datos, siendo muy parecido el comportamiento de

la validación cruzada que también aumenta con un mayor número de casos.

Figura 5.8: Curva de aprendizaje del clasificador regresión logística 3^{er} trimestre

5.4.2. Máquinas de Soporte Vectorial

Se probó maquinas de soporte vectorial como clasificador, haciendo uso del conjunto de librerías de *scikit-learn* [19].

Se realizó el análisis aplicando SVM al set de datos del primer y segundo trimestre para verificar el valor de la validación cruzada y el error de pruebas. Los resultados obtenidos para el primer y segundo trimestre se presentan en el Cuadro 5.11.

Cuadro 5.11: Métricas para SVM en 1^{er} y 2^{do} trimestre

Métrica	Resultado
Validación cruzada	61 %
Error de Pruebas	51 %

En la Figura 5.9 se muestra el resultado mediante la curva de aprendizaje del

clasificador SVM para el set de datos del primer y segundo trimestre. En este el comportamiento es casi que constante para ambas medidas, es decir que tanto el puntaje de la validación cruzada como el del entrenamiento no aumenta significativamente en la medida que hay mas datos.

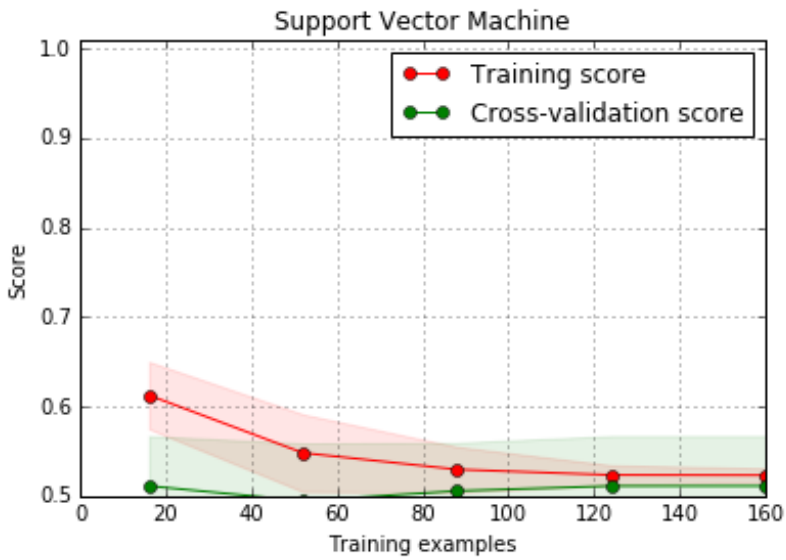

Figura 5.9: Curva de aprendizaje del clasificador SVM en 1^{er} y 2^{do} trimestre

Resultados de aplicar la técnica de SVM al set de datos del tercer trimestre se observa en el Cuadro 5.12.

Cuadro 5.12: Métricas para SVM en 3^{er} trimestre

Métrica	Resultado
Validación cruzada	73 %
Error de Pruebas	31 %

Finalmente, en la Figura 5.10 se muestra la curva de aprendizaje del clasificador SVM para el set de datos del tercer trimestre. El comportamiento de la curva para

el entrenamiento es que mejora los resultados cuando hay mas datos, al igual que la validación cruzada.

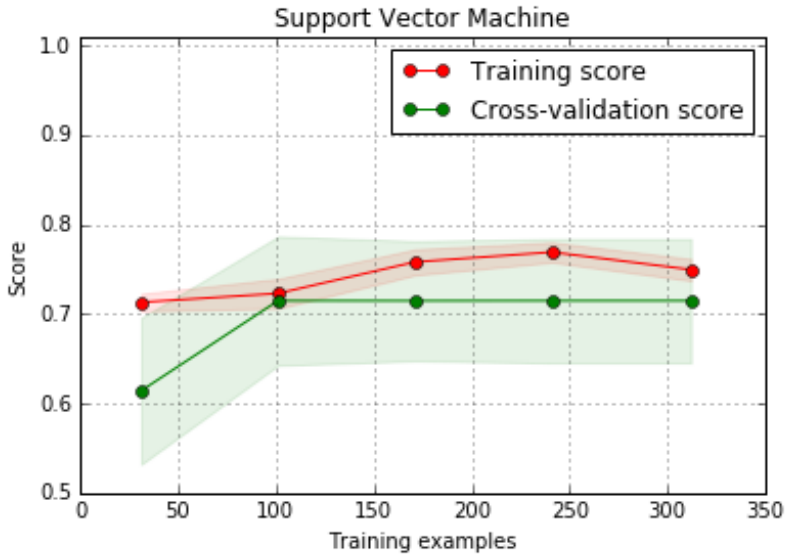

Figura 5.10: Curva de Aprendizaje del clasificador SVM en 3^{er} trimestre

5.4.3. Comparación de técnicas

El resultado del análisis de las dos técnicas se representará en el Cuadro 5.13. El modelo que se selecciona para el primer y segundo trimestre es Regresión Logística tomando como base para decidir, el mejor valor de validación cruzada.

Cuadro 5.13: Comparación de técnicas en 1^{er} y 2^{do} trimestre

Técnica	Parámetros	Validación Cruzada	Eout
Regresión Logística	11	63 %	48 %
Maquinas de Soporte Vectorial	15	61 %	51 %

Se realiza el mismo análisis comparativo entre las técnicas Regresión Logística y SVM aplicadas al set de datos del tercer trimestre como se muestra en el Cuadro 5.14.

El modelo que se selecciona para el tercer trimestre es maquinas de soporte vectorial, basado en el valor de validación cruzada y la capacidad de generalizar (*Eout*).

Cuadro 5.14: Comparación de técnicas para SVM en 3^{er} trimestre

Técnica	Parámetros	Validación Cruzada	Eout
Regresión Logística	5	63 %	48 %
Maquinas de Soporte Vectorial	13	73 %	31 %

5.5. Aplicación de Técnicas de Aprendizaje

Las técnicas de aprendizaje automático aplicadas para este trabajo son la regresión logística y las máquinas de soporte vectorial. Conociendo previamente el conjunto de variables y el resultado de la comparación entre los modelo se procede con la implementación de los clasificadores.

5.5.1. Regresión Logística en primer y segundo trimestre

Finalmente se aplica la técnica de Regresión Logística con el set de datos de Primer y segundo trimestre. Se realiza el proceso de entrenamiento y de pruebas, el proceso de entrenamiento para obtener los parámetros del clasificador que permitan mejorar la predicción, se hace un proceso de búsqueda de los mejores parámetros del clasificador obteniendo el valor del parámetro C que mejor predicción dió fué 0.51051 donde C es el factor de la regularización ($\lambda = 1/C$) y se encontró que las transformaciones polinomiales debían ser de primer grado. El proceso de pruebas para validar la generalización del clasificador en un nuevo set de datos.

El resultado del clasificador puede observarse en la matriz de confusión que se muestra en la Figura 5.11. En esta se puede ver un total de $VP = 85$ (verdaderos positivos), esto significa que se clasificaron como MME a 85 pacientes que realmente tuvieron MME. $VN = 12$ (verdaderos negativos), lo cual indica que a 12 pacientes se les clasificó como sanos o sin riesgo para MME y que en la realidad no tuvieron MME. Por otro lado, $FN = 2$ (falsos negativos) que se traduce a las clasificaciones incorrectas de los pacientes que llegaron a tener MME. Finalmente los falsos positivos $FP = 79$ que son los pacientes a los que se les clasificó con riesgo para MME pero que en realidad no lo tenían.

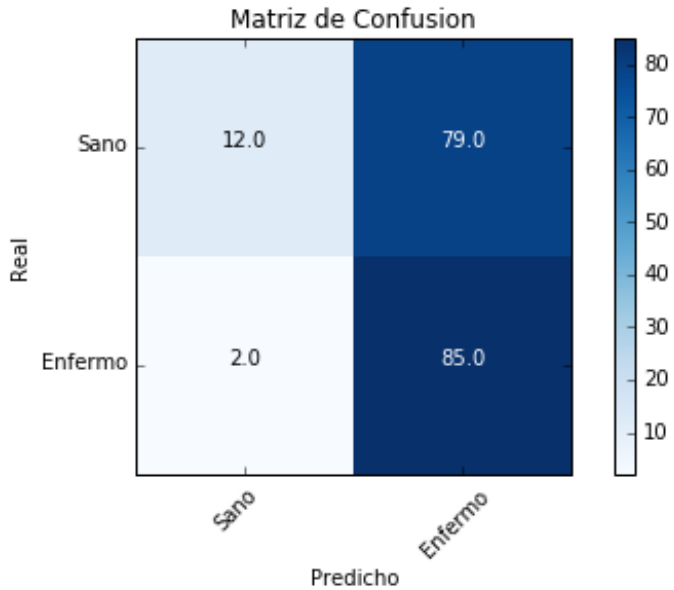

Figura 5.11: Matriz de confusión de 1^{er} y 2^{do} trimestre aplicando regresión logística

Los resultado de la matriz de confusión permiten calcular la precisión, sensibilidad y F1, que son las medidas que permiten evaluar el desempeño del predictor de Morbilidad Materna Extrema de las pacientes del control prenatal de primer y

segundo trimestre, en la E.S.E Clínica de Maternidad Rafael Calvo.Como se puede ver en el Cuadro 5.15.

Cuadro 5.15: Métricas de evaluación para 1^{er} y 2^{do} trimestre

Medida	Resultado
Precisión	51.8 %
Sensibilidad	97.7 %
F1	67.7 %

Otra forma de evaluar un clasificador es mediante la Curva ROC como se muestra en la Figura 5.12, que es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario, en otras palabras es una representación de los la razón de verdaderos positivos (VPR) frente a la razón de falsos positivos (FPR) que es igual a $1 - especificidad$.

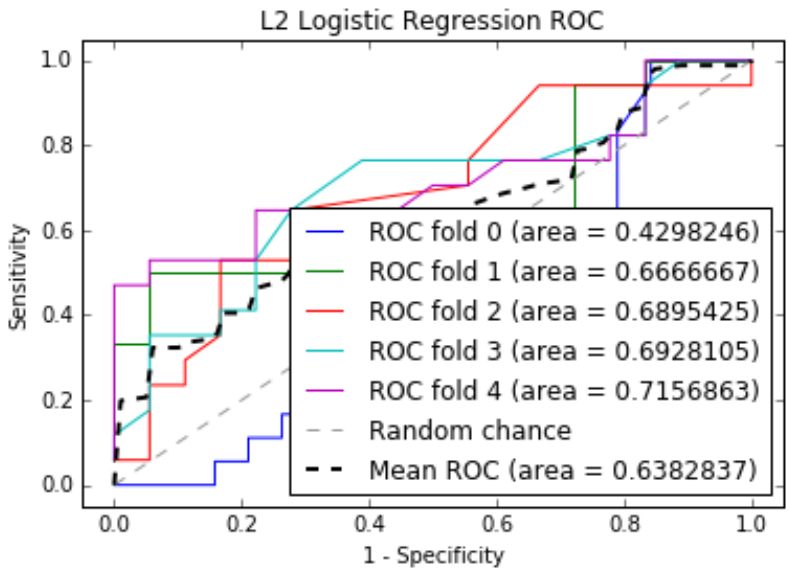

Figura 5.12: Gráfica ROC de 1^{er} y 2^{do} trimestre aplicando regresión logística

En la curva ROC linea punteada es la diagonal que divide el espacio ROC, en este espacio se considera como mejor resultado de la línea punteada hacia arriba, siendo el punto 1.0 el resultado perfecto. En este trabajo se implementó la validación cruzada de cinco pliegues o subconjunto de datos, en la gráfica se muestra el resultado del área bajo la curva para cada pliegue, siendo la línea punteada (más oscura) la media del área bajo la curva de todos los pliegues.

El resultado obtenido de este clasificador se analiza teniendo en cuenta las métricas establecidas como referente teórico por los expertos (Ginecólogos del centro de investigación de la E.S.E Clínica de Maternidad Rafael Calvo).De acuerdo a los expertos el nivel de **sensibilidad** (identificación correcta de las pacientes con riesgo de MME) del predictor debe ser superior al 60 %.

También se establece que la **precisión** (medida de exactitud en términos de verdaderos positivos y todos los predichos como positivos) debe ser mayor al 50 % . De tal forma que la hipótesis de este trabajo se puede expresar de la siguiente forma:

$$H_0 : \begin{cases} sensibilidad \leq 60\,\% \\ precision \leq 50\,\% \end{cases}$$

$$H_i : \begin{cases} sensibilidad > 60\,\% \\ precision > 50\,\% \end{cases}$$

El componente de predicción debía apuntar a detectar el mayor número de pacientes con riesgo de MME, por tal motivo las métricas que se consideraron de mayor

importancia fueron la sensibilidad, seguido de la precisión. Comparando el resultado obtenido del clasificador de regresión logística con respecto a las hipótesis planteadas, se evidencia que se cumple el criterio establecido en la hipótesis alternativa, es decir que el porcentaje de sensibilidad es mayor al 60 % y la precisión es mayor al 50 %.

5.5.2. Máquinas de Soporte Vectorial para tercer trimestre

De acuerdo a los resultados obtenidos en la comparación de las técnicas de aprendizaje supervisado probadas en este trabajo, se decide aplicar un algoritmo de SVM implementado en Python para realizar el entrenamiento y pruebas sobre un set de datos del tercer trimestre del embarazo. Se realizó un proceso de búsqueda de los valores de los parámetros que aportaran mejor desempeño al clasificador, encontrando para C un valor de 0.51051051060540542, donde C controla la compensación entre errores de entrenamiento y los márgenes rígidos, creando así un margen blando o suave que permita algunos errores en la clasificación a la vez que los penaliza. También se encontró para γ un valor de 0.50050050059549545 donde γ es un parámetro que define cuanta influencia tiene un solo ejemplo de entrenamiento, cuanto más grande γ, más cerca deben ser otros ejemplos para ser afectados. Se aplicó Validación cruzada de 5 pliegues, una penalización de diez por cada paciente que fuera realmente de la clase positiva y fuera mal diagnosticado.

El resultado del clasificador SVM se muestra en la Figura 5.13 que es la matriz de confusión. En esta se puede ver un total de $VP = 95$ (verdaderos positivos), esto significa que se clasificaron con riesgo para MME a 95 pacientes que verdaderamente

lo tenían. $VN = 8$ (verdaderos negativos), lo cual indica que a 8 pacientes se les clasificó como sanos o sin riesgo para MME y que en la realidad no tuvieron MME. Por otro lado, $FN = 0$ (falsos negativos) lo que significa que ningún paciente que terminó en MME fue clasificado sin riesgo para MME. Finalmente los falsos positivos $FP = 254$, esto indica que 254 pacientes sin riesgo fueron clasificados con riesgo para MME.

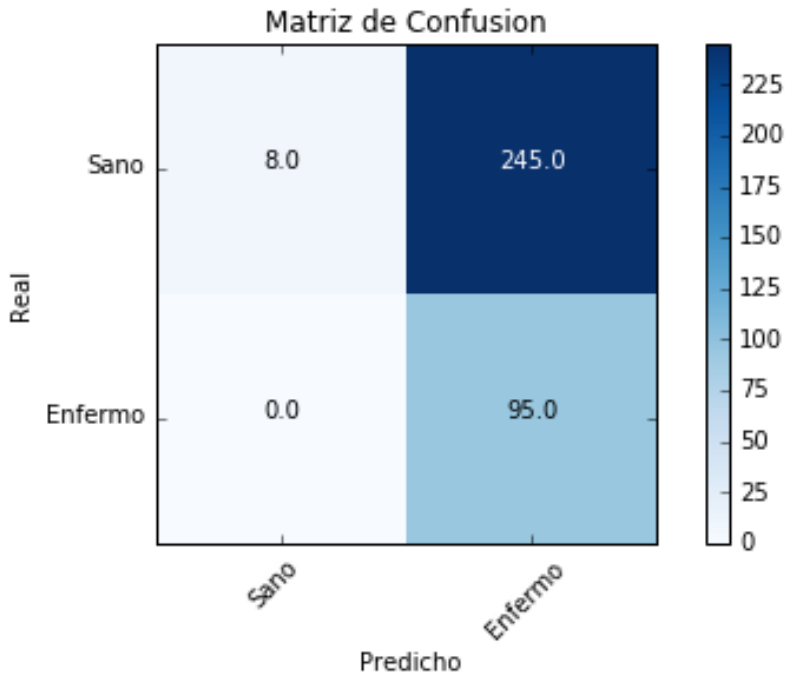

Figura 5.13: Matriz de confusión SVM en 3^{er} trimestre

Con base en la matriz de confusión se calculan las medidas que son de interés para evaluar el desempeño del clasificador de MME, que son la precisión, sensibilidad y medida F1. En el Cuadro 5.15 se muestran los resultados de cada una.

Este resultado indica que el nivel de sensibilidad es el ideal, pero cuando se verifica el nivel de precisión se descarta el clasificador porque no cumple con los dos criterios

Cuadro 5.16: Métricas de evaluación para 3^{er} trimestre

Medida	Resultado
Precisión	27.9 %
Sensibilidad	100 %
F1	43.6 %

establecidos por los expertos que sugieren una sensibilidad superior al 60 % con una precisión por encima del 50 %, este predictor cumple con los criterios establecidos en la hipótesis nula. El resultado de sensibilidad al 100 % indica que todos los pacientes con riesgo potencial de MME serán detectados por el clasificador. Una precisión del 27 % lo cual indica que habrán mucho pacientes que posiblemente no desarrollen MME pero el predictor va a tender a clasificarlos como enfermos.

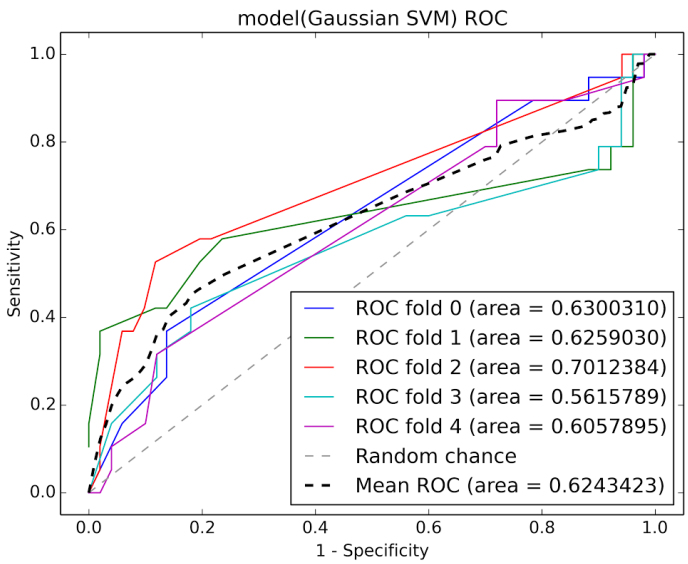

Figura 5.14: Gráfica ROC de 3^{er} trimestre aplicando SVM

De la misma forma como se usó la curva ROC en la regresión logística para evaluar

el desempeño, también se usa para SVM. En anterior Figura 5.14 se representa de forma gráfica la sensibilidad frente a la especificidad.

Este clasificador tiende a detectar casi todas las pacientes en riesgo de MME, esto como resultado del número alto de pacientes que la clínica atiende en el trimestre del embarazo con muchas complicaciones obstétricas. Es importante identificar este tipo de tendencia en los datos porque se logra evidenciar algunos aspectos que podrían aportar en la toma de decisiones que finalmente ayuden a reducir el alto indice de MME que se presenta en la región.

De la misma forma como se analiza el resultado obtenido con el clasificador de regresión es preciso hacer el análisis del predictor de MME con máquinas de soporte vectorial. Para este caso el resultado conseguido en términos de sensibilidad y precisión comparados con la hipótesis de la investigación, se debe rechazar la hipótesis alternativa y se acepta la hipótesis nula.

Capítulo 6
Recomendaciones y Trabajo Futuro

6.1. Recomendaciones

Teniendo en cuenta algunos inconvenientes encontrados durante a realización de este trabajo y con base en los resultados obtenidos se identificaron algunas debilidades que deben fortalecerse para el beneficio de la institución en lo que refiere a disminución de MME, a continuación se presentan los hallazgos y las recomendaciones de los mismos.

- Actualmente la clínica no ha logrado implementar a cabalidad un formato de historias clínicas que vaya enfocado a la atención perinatal, por tal motivo la construcción de la base de datos para el estudio demandó mucho tiempo y esfuerzo. Para poder adaptar un componente de predicción temprana de aprendizaje supervisado se hace necesario lograr el uso de una historia clínica enmarcada en los factores de riesgo para MME en el control prenatal.

- Se identificó que el número de pacientes que asisten a controles de primer y segundo trimestre es menor que las de tercer trimestre, y que las pacientes que presentan mayor número de complicaciones son las que no tienen controles en

los primeros trimestres, si no que asisten al final del embarazo. Por tanto se recomienda sensibilizar y fomentar en la población la realización de controles prenatales desde el principio del embarazo. De esta forma es posible implementar herramientas de detección o identificación temprana de los posibles riesgos durante y después del embarazo.

- Con la implementación de herramientas tecnológicas como estas, junto con el mejoramiento de la atención de las pacientes gestantes, aportariamos como región en el cumplimiento de este propósito nacional -y meta del milenio- como es la disminución de la MME. Estas herramientas de la estadística y el avance tecnológico de las máquinas de computo hace posible cada vez más poder incursionar en el apoyo diagnóstico y la detección de enfermedades.

6.2. Trabajo Futuro

Este trabajo de investigación hace parte de un macro proyecto de la E.S.E Clínica de Maternidad Rafael Calvo que tiene por objetivo mejorar la calidad de la atención prenatal y a su vez disminuir el número de casos de MME mediante el uso de herramientas tecnológicas. A continuación se mencionan algunos de los trabajos futuros que se esperan a partir de este proyecto de investigación:

- Construir un set de datos mas amplio con el cual implementar otras técnicas de aprendizaje automatizado como por ejemplo las redes neuronales, para mejorar cada vez más los resultados de la predicción en el primer y segundo trimestre

de gestación.

- Se espera realizar un proyecto para el desarrollo de un componente independiente de predicción que permita ser acoplado al software de historias clínica, usando servicios web como método de comunicación entre las diferentes plataformas. Esto permitirá ver al médico en tiempo real el porcentaje de riesgo que presente una paciente luego de diligenciar la historia clínica.

Capítulo 7
Conclusiones

La Morbilidad Materna Extrema es una condición que afecta a las mujeres en etapa de gestación, durante el parto o en el puerperio, las pacientes que llegan a tener esta condición pueden tener secuelas incluso a largo plazo. Aún con los avances en salud materna que se han implementado en la región y en el país, estas complicaciones maternas siguen siendo un gran problema en salud pública. Existe un protocolo en salud pública en Colombia que permite clasificar a una paciente con MME. Cada institución está obligada a reportar al sistema de vigilancia en salud pública a las pacientes que han sido catalogadas con al menos un criterio de inclusión para MME. Con esto se logra identificar la tendencia de la población gestante, pero no se logra aminorar el número de pacientes afectadas con tal condición. Por lo anterior las ayudas tecnológicas destinadas a apoyar en el diagnóstico temprano de MME pueden aportar a un mejor cuidado de las maternas. En este trabajo se presentó la validación de dos técnicas de aprendizaje supervisado: Regresión Logística y Máquinas de Soporte Vectorial, aplicadas a los datos de obtenidos en los controles prenatales de la CMRC. Los resultados obtenidos muestran un avance positivo y prometedor en cuanto a usar predictores como apoyo diagnóstico para la detección de riesgo para

MME en el primer y segundo trimestre del embarazo.

Durante la realización del trabajo se obtuvieron resultados que pueden aportar en el mejoramiento de la atención prenatal de la pacientes. A continuación se detallan estas conclusiones:

- En cuanto al análisis de las variables se pude concluir que durante el primer y segundo trimestre del embarazo los principales factores que se encontraron como relacionados a la ocurrencia de morbilidad materna extrema se encuentran en el Cuadro 7.1.

Cuadro 7.1: Factores para MME

Clasificación	Factor
Dato Personal	Edad
Dato Personal	Estrato
Etnia	Palenquero
Dato ginecológico	Multiplicidad
Dato ginecológico	Paridad
Antecedente	Infección de vías urinarias
Antecedente	Preeclampsia
Antecedente	Eclampsia
Antecedente	Diabetes
Diagnóstico	O10-O16: Edema proteinuria e hipertensión
Diagnóstico	E20-E35: Trastornos de otras glándulas endocrinas
Diagnóstico	I11-I15: Enfermedad hipertensiva
Diagnóstico	N30-N39: Otras enfermedades del sistema urinario
Diagnóstico	N30-N39: Otras enfermedades del sistema urinario

- Este proyecto contiene un componente social muy importante como es la detección temprana de MME y ayudar a mejorar los cuidados maternos. Por lo

anterior se hizo mucho esfuerzo en lograr un buen predictor en las fases tempranas del embarazo como son el primer y segundo trimestre. El componente de Predicción debía apuntar a detectar el mayor número de pacientes con riesgo de MME, por tal motivo las métricas que se consideró de mayor importancia fueron la sensibilidad, seguido de la precisión.

- El alto porcentaje de pacientes con complicaciones en el tercer trimestre del embarazo hace que el comportamiento del predictor marque una tendencia a clasificar con riesgo para MME a todas las pacientes que realmente puedan llegar a MME, pero también clasifica con riesgo a muchas pacientes que al final pueden no desarrollan MME. También es importante mencionar que la detección de riesgo para MME en el tercer trimestre brinda menor garantía de evitarlo que si se detecta en los primeros trimestres. Por los cual nuestros esfuerzos deben concentrarse en las fases tempranas del embarazo.

Bibliografía

[1] W. Caicedo-Torres, Á. Paternina, and H. Pinzón. Machine learning models for early dengue severity prediction. *Ibero-American Conference on Artificial Intelligence*, pages 247–258, 2016.

[2] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[3] O. P. de la Salud. *Clasificación estadística internacional de enfermedades y problemas relacionados con la salud: décima revisión: CIE-10*. Pan American Health Org, 1995.

[4] M. Duran. Protocolo de vigilancia en salud publica morbilidad materna extrema. Technical report, Equipo Maternidad Segura,Subdirección de Prevención,Vigilancia y Control en Salud Pública,Instituto Nacional de Salud INS, 2016.

[5] P. Faneite. Taller latinoamericano: una alianza para enfrentar los desafíos, reducir la morbilidad y mortalidad materna y perinatal. flasog/ops. *Rev Obstet Ginecol Venez*, 64(4):223–225, 2004.

[6] B. Farran, A. M. Channanath, K. Behbehani, and T. A. Thanaraj. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from kuwait—a cohort study. *BMJ open*, 3(5):e002457, 2013.

[7] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38, 2004.

[8] R. Fescina. Centro latinoamericano de perinatología, salud de la mujer y reproductiva, clap/smr: programa de trabajo. In *La red perinatal*. Centro Latinoamericano de Perinatología, Salud de la Mujer y Reproductiva, 2006.

[9] S. E. Geller, D. Rosenberg, S. Cox, M. Brown, L. Simonson, and S. Kilpatrick. A scoring system identified near-miss maternal morbidity during pregnancy. *Journal of clinical epidemiology*, 57(7):716–720, 2004.

[10] W. A. Grobman, J. L. Bailit, M. M. Rice, R. J. Wapner, U. M. Reddy, M. W. Varner, J. M. Thorp Jr, K. J. Leveno, S. N. Caritis, J. D. Iams, et al. Frequency of and factors associated with severe maternal morbidity. *Obstetrics and gynecology*, 123(4):804, 2014.

[11] M. C. Hogan, K. J. Foreman, M. Naghavi, S. Y. Ahn, M. Wang, S. M. Makela, A. D. Lopez, R. Lozano, and C. J. Murray. Maternal mortality for 181 countries, 1980–2008: a systematic analysis of progress towards millennium development goal 5. *The lancet*, 375(9726):1609–1623, 2010.

[12] S. Jahan, K. Begum, N. Shaheen, and M. Khandokar. Near-miss/severe acute maternal morbidity (samm): A new concept in maternal care. *Journal of Bangladesh College of Physicians and Surgeons*, 24(1):29–33, 2006.

[13] E. V. Kuklina, C. Ayala, and W. M. Callaghan. Hypertensive disorders and severe obstetric morbidity in the united states. *Obstetrics & Gynecology*, 113(6):1299–1306, 2009.

[14] C. B. Montes. *Evaluación de uso de un cluster oportunista basados en sistemas operativos heterogéneos y HTCondor*. Universidad Tecnológica de Bolívar, 2016.

[15] S. Nanda, M. Savvidou, A. Syngelaki, R. Akolekar, and K. H. Nicolaides. Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. *Prenatal diagnosis*, 31(2):135–141, 2011.

[16] A. Ng. Machine learning: Stanford machine learning course materials.

[17] W. H. Organization and UNICEF. *Revised 1990 estimates of maternal mortality: a new approach*. World Health Organization, 1996.

[18] F. J. Park, C. H. Leung, L. C. Poon, P. F. Williams, S. J. Rothwell, and J. A. Hyett. Clinical evaluation of a first trimester algorithm predicting the risk of hypertensive disease of pregnancy. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 53(6):532–539, 2013.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] L. C. Poon, N. A. Kametas, N. Maiz, R. Akolekar, and K. H. Nicolaides. First-trimester prediction of hypertensive disorders in pregnancy. *Hypertension*, 53(5):812–818, 2009.

[21] E. A. Rodríguez, F. E. Estrada, W. C. Torres, and J. C. M. Santos. Early prediction of severe maternal morbidity using machine learning techniques. In *Ibero-American Conference on Artificial Intelligence*, pages 259–270. Springer, 2016.

[22] R. Schwarcz. La red perinatal: una estrategia de clap para expandir las actividades perinatales en la región. In *Tecnologías perinatales*, pages 53–7. Centro Latinoamericano de Perinatología y Desarrollo Humano, 1990.

[23] S. Server. Documentation. *OpenLink Software Documentation Team*, 2009.

[24] A. O. Tsui, J. N. Wasserheit, J. G. Haaga, et al. Healthy pregnancy and childbearing. 1997.

[25] V. K. Vaishnavi and W. Kuechler. *Design science research methods and patterns: innovating information and communication technology*. Crc Press, 2015.

[26] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

Anexo 1: Consulta SQL

```sql
DECLARE @columnas VARCHAR(MAX)
SELECT @columnas = ''
--primer trimestre >=27
--segundo trimestre 14-27
--primer 1-13

SELECT @columnas =  @columnas + '[' + cast(diagnostico AS
↪  VARCHAR(12)) + '],'
FROM (SELECT RANGO diagnostico
FROM (SELECT A.IDENTIFICACION,diagnostico
          FROM [dbo].[antecedentes] A
           INNER JOIN
           (SELECT identificacion,edadgestacion,
                       rtrim(ltrim(diagnostico)) diagnostico
                FROM [dbo].[Controles]
                WHERE CONVERT(DECIMAL,edadgestacion) <=27
                UNION ALL
                SELECT identificacion,edadgestacion,
                       rtrim(ltrim(diagnostico1)) diagnostico
                FROM [dbo].[Controles]
                WHERE CONVERT(DECIMAL,edadgestacion) <=27
                UNION ALL
                SELECT identificacion,edadgestacion,
                       rtrim(ltrim(diagnostico2)) diagnostico
                FROM [dbo].[Controles]
                WHERE CONVERT(DECIMAL,edadgestacion) <=27
                UNION ALL
                SELECT identificacion,edadgestacion,
                       rtrim(ltrim(diagnostico3)) diagnostico
                FROM [dbo].[Controles]
                WHERE CONVERT(DECIMAL,edadgestacion) <=27
          ) C
                ON C.identificacion=A.identificacion
```

```
                WHERE  diagnostico<>''
                GROUP BY A.identificacion,diagnostico)R
INNER JOIN  dbo.gruposdiagnosticos C ON C.Diagnostico=R.diagnostico
GROUP BY RANGO) as DTM

SET @columnas = LEFT(@columnas,LEN(@columnas)-1)

DECLARE @SQLString NVARCHAR(MAX);

SET @SQLString = N'

SELECT *
FROM
(SELECT DISTINCT A.IDENTIFICACION,Rango
,[edad]
,CASE WHEN [Etnia]=''2.Romgitano'' THEN 1
        ELSE 0        END Romgitano
,CASE WHEN [Etnia]=''3.Raizal'' THEN 1
        ELSE 0  END Raizal
,CASE WHEN [Etnia]=''4.Palenquero'' THEN 1
        ELSE 0 END Palenquero
,CASE WHEN [Etnia]=''5. Negro-mulato-afrodescendiente'' THEN 1
        ELSE 0        END Negro
, CASE  WHEN [Etnia]=''6.Otro'' THEN 1
        ELSE 0        END Otro
,CASE  WHEN [Escolaridad]=''Ninguna'' THEN 1
        ELSE 0        END EscolarNinguna
,CASE  WHEN [Escolaridad]=''Primaria'' THEN 1
        ELSE 0        END Primaria
,CASE  WHEN [Escolaridad]=''Secundaria'' THEN 1
        ELSE 0        END Secundaria
,CASE  WHEN [Escolaridad]=''Tecnica'' THEN 1
        ELSE 0        END Tecnica
,CASE  WHEN [Escolaridad]=''Universitaria'' THEN 1
        ELSE 0        END Universitaria
,[Estrato]
,CASE  WHEN [Regimen]=''Contributivo'' THEN 1
ELSE 0        END Contributivo
,CASE  WHEN [Regimen]=''No afiliado'' THEN 1
ELSE 0        END Noafiliado
,CASE  WHEN [Regimen]=''Subsidiado'' THEN 1
ELSE 0        END Subsidiado
```

```
,CASE  WHEN [Procedencia]=''Cabecera municipal'' THEN 1
ELSE 0        END Cabeceramunicipal
,CASE  WHEN [Procedencia]=''Centro o poblado'' THEN 1
ELSE 0        END Centropoblado
,CASE  WHEN [EstadoCivil]=''Casada'' THEN 1
ELSE 0        END Casada
,CASE  WHEN [EstadoCivil]=''Separada'' THEN 1
ELSE 0        END Separada
,CASE  WHEN [EstadoCivil]=''Soltera'' THEN 1
ELSE 0        END Soltera
,CASE  WHEN [EstadoCivil]=''Union libre'' THEN 1
ELSE 0        END Unionlibre
,CASE  WHEN [EstadoCivil]=''Viuda'' THEN 1
ELSE 0        END Viuda
,[Paridad]
,CASE WHEN [intergenesico]=''Si'' THEN 1
        WHEN [intergenesico]=''No'' THEN 0
        END intergenesico
,CASE WHEN [Multiplicidad]=''No aplica'' THEN 0
        WHEN [Multiplicidad]=''Simple'' THEN 1
        END [Multiplicidad]
,CASE WHEN [micronutrientes]=''Si'' THEN 1
        WHEN [micronutrientes]=''No'' THEN 0
        END [micronutrientes]
,CASE WHEN [Preeclampsia]=''Si'' THEN 1
        WHEN [Preeclampsia]=''No'' THEN 0
        END [Preeclampsia]
,CASE WHEN [Sepsis]=''Si'' THEN 1
        WHEN [Sepsis]=''No'' THEN 0
        END [Sepsis]
,CASE WHEN [Eclampsia]=''Si'' THEN 1
        WHEN [Eclampsia]=''No'' THEN 0
        END [Eclampsia]
,CASE WHEN [Diabetes]=''Si'' THEN 1
        WHEN [Diabetes]=''No'' THEN 0
        END [Diabetes]
,CASE WHEN [TORCHS]=''Si'' THEN 1
        WHEN [TORCHS]=''No'' THEN 0
        END [TORCHS]
,CASE WHEN [vias-Urinarias]=''Si'' THEN 1
        WHEN [vias-Urinarias]=''No'' THEN 0
        END [vias-Urinarias]
```

```
,CASE WHEN [autoinmune]=''Si'' THEN 1
        WHEN [autoinmune]=''No'' THEN 0
        END [autoinmune]
,CASE WHEN [MME]=''Si'' THEN 1
        WHEN [MME]=''No'' THEN 0
        END [MME]
FROM [dbo].antecedentes A
INNER JOIN
(SELECT r.IDENTIFICACION,RANGO
 FROM
 (SELECT A.IDENTIFICACION,diagnostico
   FROM [dbo].[antecedentes] A
   INNER JOIN
   (SELECT identificacion,edadgestacion,
               rtrim(ltrim(diagnostico)) diagnostico
    FROM [dbo].[Controles]
    WHERE CONVERT(DECIMAL,edadgestacion) <=27
        UNION ALL
        SELECT identificacion,edadgestacion,
               rtrim(ltrim(diagnostico1)) diagnostico
        FROM [dbo].[Controles]
        WHERE CONVERT(DECIMAL,edadgestacion) <=27
        UNION ALL
        SELECT identificacion,edadgestacion,
               rtrim(ltrim(diagnostico2)) diagnostico
        FROM [dbo].[Controles]
        WHERE CONVERT(DECIMAL,edadgestacion) <=27
        UNION ALL
        SELECT identificacion,edadgestacion,
               rtrim(ltrim(diagnostico3)) diagnostico
        FROM [dbo].[Controles]
        WHERE CONVERT(DECIMAL,edadgestacion) <=27
        ) C
        ON C.identificacion=A.identificacion
        WHERE  diagnostico<>''''
        GROUP BY A.identificacion,diagnostico

)R
INNER JOIN  dbo.gruposdiagnosticos C ON C.Diagnostico=R.diagnostico
GROUP BY r.IDENTIFICACION,RANGO)C ON
 ↪  C.identificacion=A.identificacion)
AS SourceTable
```

```
158 PIVOT
159 (
160 COUNT(RANGO)
161 FOR RANGO IN (' + @columnas + ')
162 ) AS PivotTable
163 --WHERE MME=''1'''
164 EXECUTE sp_executesql @SQLString
```

Consulta SQL para obtener la información de las pacientes en cada control prenatal y sus antecedentes, con el formato requerido para ejecutar los algoritmos en Python.

Anexo 2: Algoritmo de selección de características LR

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cross_validation import StratifiedKFold
from sklearn import preprocessing
from sklearn.metrics import roc_curve, auc, confusion_matrix
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFECV

f = open("I-II-Trimestre-2015-2016.csv")
f.readline()
data = np.loadtxt(f, delimiter=',')
rows_train=int(len(data)*0.2)
X = data[:-rows_train,range(1,len(data[1,:]))]
y = data[:-rows_train,0]

X_scaled = preprocessing.scale(X)

logReg = LogisticRegression(C=0.0024212178217821782, penalty='l2',
 ↪  class_weight={1: 10})

rfecv = RFECV(estimator=logReg, step=1, cv=StratifiedKFold(y, 5),
              scoring='roc_auc')
rfecv.fit(X_scaled, y)

print("Optimal number of features : %d" % rfecv.n_features_)
print("Selected features mask: ")
print(rfecv.support_)
print len(rfecv.support_)

# Plot number of features VS. cross-validation scores
fig = plt.figure()
```

```
31 plt.xlabel("Number of features selected")
32 plt.ylabel("Cross validation score (ROC)")
33 plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
34 plt.show()
```

Anexo 3: Algoritmo de selección de características SVM

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cross_validation import StratifiedKFold
from sklearn import preprocessing
from sklearn.metrics import roc_curve, auc, confusion_matrix
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFECV



f = open("III-Trimestre-2015-2016.csv")
f.readline()
data = np.loadtxt(f, delimiter=',')
rows_train=int(len(data)*0.2)
X = data[:-rows_train,range(1,len(data[1,:]))]
y = data[:-rows_train,0]


#X = preprocessing.PolynomialFeatures(degree=3,
 ↪  interaction_only=True).fit_transform(X)

#zero mean, unit variance
X_scaled = preprocessing.scale(X)#normalizacion de los datos

logReg = LogisticRegression(C=0.0013212178217821782, penalty='l2',
 ↪  class_weight={1: 10})

estimator = SVR(kernel="linear")
rfecv = RFECV(estimator, step=1, cv=StratifiedKFold(y, 5),
              scoring='roc_auc')
```

```
30 rfecv.fit(X_scaled, y)
31
32 print("Número de variables Óptimas : %d" % rfecv.n_features_)
33 print("selección de variables: ")
34 print(rfecv.support_)
35 print("Total variables : %d" % len(rfecv.support_))
36 print
37
38 # Plot number of features VS. cross-validation scores
39 fig = plt.figure()
40 plt.xlabel("Number of features selected")
41 plt.ylabel("Cross validation score (ROC)")
42 plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
43 plt.show()
44 #fig.savefig('logReg_RFE_result.png', format='png', dpi=500)
```