ESTUDIO COMPARATIVO DE TÉCNICAS DE *MACHINE LEARNING* PARA LA DETERMINACIÓN DE EMBARAZOS PRE-TÉRMINO A PARTIR DEL ELECTROHISTEROGRAMA

ANGELA STEPHANYA CAIPE GORDILLO

JORGE ARMANDO MUÑOZ ROSERO

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA
SAN JUAN DE PASTO
2017

ESTUDIO COMPARATIVO DE TÉCNICAS DE *MACHINE LEARNING* PARA LA DETERMINACIÓN DE EMBARAZOS PRE-TÉRMINO A PARTIR DEL ELECTROHISTEROGRAMA

ANGELA STEPHANYA CAIPE GORDILLO

JORGE ARMANDO MUÑOZ ROSERO

TRABAJO DE GRADO PARA OPTAR EL TÍTULO DE INGENIERO ELECTRÓNICO

DIRECTOR

PhD. DIEGO HERNÁN PELUFFO ORDÓÑEZ

INGENIERO ELECTRÓNICO

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA
SAN JUAN PASTO
2017

NOTA DE RESPONSABILIDAD

“La Universidad de Nariño no se hace responsable por las opiniones o resultados obtenidos en el presente trabajo y para su publicación priman las normas sobre el derecho de autor.”

Acuerdo 1. Artículo 324. Octubre 11 de 1966, emanado del Honorable Consejo Directivo de la Universidad de Nariño.

NOTA DE ACEPTACIÓN:

_____
_____
_____
_____
_____
_____
_____
_____

_____
Firma del presidente del jurado

_____
Firma del jurado

_____
Firma del jurado

San Juan de Pasto, 10 de mayo del 2017

DEDICATORIAS

“A mis padres, hermanos y sobrina. A mis padres que siempre me han apoyado a seguir mis sueños quienes con su consejo y confianza me dan la fortaleza para continuar. Mis hermanos por ser mi apoyo incondicional y a mi sobrina por llenar de alegría mis días con sus tiernas palabras. Los amo.”

Angela Stephanya Caipe Gordillo

“A mis padres y hermano, por su amor incondicional, por ser la guía, motor y el pilar de mi vida, que gracias a su motivación y consejo me han permitido cosechar logros.”

Jorge Armando Muñoz Rosero

AGRADECIMIENTOS

"*A Dios, en primer lugar, sin él nada de esto hubiera sido posible. A mis padres Jorge y Lidia, quienes han estado para mí en todo momento y me han impulsado a seguir adelante. A mis hermanos y sobrina por estar pendiente de mí, por sus palabras y consejo. Al profesor Diego Peluffo, por creer en las capacidades de sus estudiantes, quien con su motivación, conocimiento y palabras ha fomentado un espíritu investigativo en cada uno de nosotros. Y finalmente a todos mis profesores, gracias por prepararnos para un futuro competitivo no solo como profesionales sino también como personas."*

Angela Stephanya Caipe Gordillo

"*A Dios, por todas las bendiciones que he recibido. A mis padres Ricardo y Beatriz por todo el apoyo que me han brindado a lo largo de mi vida, por ser la razón de levantarme cada día para seguir adelante y por darme la oportunidad de ser un profesional. A mi maestro Diego Peluffo que gracias a su humildad y sencillez ha sido fuente de inspiración y apoyo para recorrer los caminos del conocimiento.*"

Jorge Armando Muñoz Rosero

RESUMEN

El embarazo pre-término ocurre cuando se presenta labor de parto antes de la semana treinta y siete de gestación, siendo una de las principales causas de mortalidad y morbilidad en la población infantil. A pesar de que existen diversos factores que indiquen riesgo de un trabajo de parto prematuro, éste se puede producir espontáneamente sin la necesidad de ningún síntoma o factor indicativo. En el mundo, se estima que alrededor de quince millones de bebés prematuros nacen al año, una cifra que va en aumento y que además tiene impacto mayor en los países en desarrollo.

El cambio de la actividad eléctrica uterina durante el embarazo, es un hecho; por tanto, realizar un estudio comparativo de registros sobre la actividad eléctrica, mediante electrohisterografía, es una estrategia potencial al momento de clasificar entre el trabajo de parto pre-término falso o verdadero. En la actualidad, se usa sistemas computarizados basados en técnicas de *machine learning* para determinar la ocurrencia de un embarazo pre-término a partir de registros electrohisterograficos. No obstante, aún no existen métodos definitivos para caracterizar y clasificar estos registros.

En este trabajo, se plantea realizar el diseño e implementación de una metodología de comparación para el diagnóstico de un trabajo de parto prematuro. Para ello se desarrolla un estudio comparativo de diferentes métodos de *machine learning* considerando técnicas supervisadas y no supervisadas de reconocimiento de patrones, con el fin de desarrollar un sistema computarizado eficiente de clasificación de registros electrohisterograficos que permita predecir eficientemente esta condición.

ABSTRACT

Pre-term happens when the labor occurs before thirty-seven weeks of gestation, it is one of the leading causes of mortality and morbility in children. Despite there are several factors that indicate its risk, preterm labor can be produced spontaneously without any symptom or indicative factor in the world it is estimated that around fifteen milliones pre-term babies are born each year, and this amount having a greater impact in developing countries.

The change of uterine electrical activity during pregnancy is a real fact, therefore, a study of records of uterine electrical activity -through electrohysterography- is a potential discriminator between the true and false preterm labor. Currently, computerized pre-term systems have been introduced, which take advantage of electrohysterografic records.

To estimate the probability of pregnancy, systems based on machine learning techniques are used, however, still there are not definitive methods to characterize and classify these records, in order to effectively predict pregnancy preterm.

This work is aimed at the design and implementation of a comparison methodology for diagnosing risk of preterm labor. To do so, a comparative study of different methods of machine learning is performed by considering supervised and unsupervised pattern recognition techniques, in order to develop an effecient computarized system of electrohysterografic record classification able to identify the risk of pre-term delivery.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

INTRODUCCIÓN

En el mundo se estima que al año alrededor de quince millones de bebés nacen prematuramente, una cifra que va en aumento y tiene mayor impacto en los países en desarrollo (las consecuencias son más sentidas en las clases socio-económicas desfavorecidas [1]).

El embarazo pre-término ocurre cuando se presenta labor de parto antes de la semana treinta y siete de gestación. Entre las principales consecuencias que conlleva un nacimiento prematuro se encuentra un alto riesgo de mortalidad en los primeros años de vida [1] (alrededor de un 80%, tanto en países desarrollados como en desarrollo). Además, es una de las principales causas de morbilidad perinatal grave, que incluye problemas respiratorios, enterocolitis necrotizante, hemorragia intraventricular, discapacidades a largo plazo como la parálisis cerebral, ceguera y perdida de la audición [1] [2] [3].

En la actualidad existen estrategias que incluyen el análisis de diversos factores indicativos de riesgo de labor de parto prematuro, sin embargo, no son determinantes al momento de predecir un trabajo de parto pre-término, y además, éste se puede producir espontáneamente sin haberse presentado ningún síntoma o factor indicativo.

Diversos estudios han demostrado que existe un cambio de la actividad eléctrica uterina durante el embarazo, asimismo, se evidencia una relación como factor discriminante potencial entre un parto pre-término y uno a término completo [17]. Entonces, el electrohisterograma (EHG) es un procedimiento no invasivo que se lleva a cabo para detectar cambios bioeléctricos y que registra la actividad eléctrica del útero.

Con respecto al párrafo anterior, las señales obtenidas son caracterizadas teniendo en cuenta la magnitud y duración de las contracciones uterinas, de este modo, se permite ver la evolución del comportamiento fisiológico y patofisiológico del trabajo de parto, es así que, el hecho de realizar un estudio comparativo de registros de la actividad eléctrica uterina, mediante electrohisterografía, representa un elemento clave en la detección de un parto prematuro [4] [5] [6].

Con lo anterior, se aprecia que a pesar de ser un campo de estudio relativamente nuevo, se han presentado diversos estudios de señales de electrohisterografía en

los cuales se evalúa las características de dichas señales, y han generado resultados prometedores, aunque no definitivos ni generalizantes. En efecto existen aún diversos problemas.

En este trabajo de grado se analizó características propias para señales EHG, técnicas de selección de características, reducción de dimensión y clasificación con el objetivo de mejorar la clasificación de señales EHG.

En este trabajo se aborda una metodología que incluye: Pre-procesamiento de las señales EHG, técnicas de selección de características (tales como *ranking,* sub-conjuntos y búsqueda exhaustiva) técnicas de reducción de dimensión convexas y no convexas (basadas principalmente en el escalamiento clásico) y, finalmente, técnicas supervisadas y no supervisadas de clasificación.

Acorde a lo anterior, se realiza la comparación de algoritmos de clasificación supervisada mediante familias que incluyen clasificadores lineales y de alto grado polinomial, clasificadores basados en densidad normal y clasificación no lineal. También, se utilizó una técnica tradicional no supervisada conocida como k-medias [7] [8].

En este trabajo se presenta los resultados de un estudio comparativo, consiste consiste en evaluar el desempeño de combinaciones entre técnicas de selección de características, técnicas de reducción de dimensión y de técnicas de clasificación, que presenten un equilibrio entre efectividad, costo computacional y fácil interpretación del concepto fisiológico al momento de realizar la clasificación de registros EHG, en diagnósticos de riesgo de embarazo pre-termino, siendo hasta el momento el mejor resultado de la mediana del error 18.75%.

1. DESCRIPCIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

En la actualidad, para la valoración de patologías y el cuidado prenatal, la medicina requiere de un diagnóstico de alta efectividad para llevar a cabo de forma apropiada los procedimientos de tratamiento por medio de métodos no invasivos. Estos métodos se basan en mediciones superficiales de señales eléctricas uterinas y por tanto no implican realizar mediciones directamente sobre el cuello uterino [4] [6].

Específicamente, en el caso de riesgo de trabajo de parto prematuro, las señales de electromiografía uterinas también conocidos como señales electrohisterograficas o señales EHG [4] [6] permiten, detectar cambios en la excitabilidad celular y acoplamiento necesarios para el trabajo de parto y tiene altos valores predictivos de parto prematuro que otros métodos disponibles en la actualidad [5]. Un modelo ampliamente utilizado en clasificación de señales es el *machine learning* que, por medio de técnicas de clasificación supervisada y no supervisada, busca una comparación entre diferentes modelos matemáticos. Aunque ya se han desarrollado algunos estudios con registros EHG por algunos tipos de modelos matemáticos lineales y no lineales [6] [4], se debe tener presente que mediante las técnicas antes mostradas la garantía de efectividad en el diagnóstico de riesgo de trabajo de parto prematuro sigue siendo un problema abierto.

1.2. JUSTIFICACIÓN

Según la organización mundial de la salud (OMS), una de las principales causas de morbilidad y mortalidad en niños menores de cinco años es debido al parto prematuro, es por esta razón que existe la necesidad de garantizar un diagnóstico de riesgo de parto pre-término, eficiente y oportuno.

En la actualidad, gracias al progreso de la tecnología y la complejidad del diagnóstico de parto pre-término, se utiliza sistemas computacionales. Por tanto, se desarrolla este trabajo que tiene como objetivo principal identificar métodos computacionales que brinden una mayor efectividad en el procesado de las señales y así generar un diagnóstico soporte en esta área. Se debe agregar que se cuenta con el recurso humano y técnico para llevar a cabo esta investigación.

En este orden de ideas, y considerando que el *machine learning* permite clasificar información de forma automatizada, es adecuado realizar un estudio comparativo de técnicas supervisadas y no supervisadas por medio de aplicada a señales EHG para el diagnóstico de riesgo de parto pre-término en términos de efectividad, costo computacional e interpretación del concepto fisiológico

1.3. CONTRIBUCIÓN DE LA INVESTIGACIÓN

Para las mujeres gestantes tanto como para la medicina, en especial las ramas encargadas de un seguimiento peri-natal, como la ginecología, la obstétrica y la medicina neonatal, sería muy valiosa la información que arrojaría el estudio que se pretende realizar en este proyecto, el cual consiste en definir modelos matemáticos y computacionales por medio de técnicas supervisadas y no supervisadas que presenten equilibrio entre efectividad, costo computacional en señales EHG e interpretación del concepto fisiológico, al momento de brindar un diagnóstico con riesgo de un parto a pre-término. Reconociendo la importancia de esta información, y de las consecuencias que implicaría en el crecimiento y desarrollo del neonato, para que de esta manera, pueda ser tratado, según las consideraciones del especialista, mejorando la calidad de vida de las madres gestantes y sus bebés.

Este trabajo podría representar un aporte a la comunidad científica y académica en el área de procesamiento de señales biomédicas, específicamente, en el análisis de señales EHG; así como, a nivel formal y técnico, las formulaciones matemáticas, los algoritmos a explorar y desarrollar, representarían una contribución en la investigación básica de la misma área. Es importante que el desarrollo de esta investigación cuenta con el respaldo del Grupo de Investigación en Ingeniería y Electrónica de la Universidad de Nariño (GIIEE) y los productos académicos desarrollados benefician al mismo grupo.

1.4. ORGANIZACIÓN DE ESTE DOCUMENTO

Este documento está dividido en 7 capítulos principales nombrados de la siguiente manera: Introducción, Descripción del problema, Objetivos, Marco teórico, Metodología, Marco experimental, Resultados, Conclusiones y Trabajo Futuro.

En el capítulo 1, se presenta el planteamiento del problema, la justificación del trabajo y las contribuciones de la investigación

En el capítulo 2, se presenta el objetivo general y los objetivos específicos planteados al comienzo de esta tesis.

En el capítulo 3, se describe el marco teórico en el cual se desarrolla esta investigación.

En el capítulo 4, se describe la metodología propuesta para el desarrollo de esta investigación.

En el capítulo 5, se presenta el marco experimental donde se da a conocer el tipo de pruebas empleadas.

En el capítulo 6, se presenta los resultados obtenidos en la investigación que se realizada.

Por último, en el capítulo 7, se presenta las conclusiones producto de la investigación realizada además del posible trabajo futuro, en busca de mejorar los resultados obtenidos que contribuyan al desarrollo de nuevas investigaciones en el tema.

2. OBJETIVOS

2.1. OBJETIVO GENERAL

Realizar un estudio comparativo de técnicas de *machine learning* para determinar embarazos con riesgo de parto prematuro a partir de registros electrohisterográficos

2.2. OBJETIVOS ESPECÍFICOS

- Establecer un conjunto de características que represente adecuadamente los registros electrohisterográficos con el fin de generar separabilidad entre registros de embarazo a término y pre-término.

- Seleccionar métodos representativos de *machine learning* de clasificación supervisada y no supervisada para ser aplicados en señales EHG.

- Diseñar una estrategia de comparación de los métodos seleccionados con el fin de proponer un sistema de *machine learning* que presente equilibrio entre efectividad, costo computacional y la interpretación del concepto fisiológico para la predicción de embarazos pre-término.

3. MARCO TEÓRICO

El cambio del potencial eléctrico producido en el útero, puede ser registrado en señales EHG (registros electrohisterograficos), y por tanto, el análisis de señales EHG constituye una fuente significativa de información en la predicción del parto pre-término [4], el cual, representa una de las causas principales de muerte en recién nacidos. Aunque, existen diferentes estrategias para la predicción del parto pre-término, EHG ha demostrado ser una alternativa llamativa debido a que es un procedimiento no invasivo.

En este capítulo se menciona algunos conceptos preliminares orientados al diagnóstico de parto pre-término a través de señales EHG. Específicamente la sección 3.1 presenta el contexto fisiológico y en 3.2 sistemas computaciones para la predicción de un parto pre-término.

3.1. CONTEXTO FISIOLÓGICO

3.1.1. Labor Pre-término

Cuando se presenta labor de parto antes de la semana 37 de gestación, se conoce como parto pre-término. Por otra parte, se considera un parto normal cuando se presenta labor en la semana 40 de gestación. Según la Organización Mundial de la Salud (OMS), cada año nacen quince millones de bebés prematuros, una cifra que está creciendo, teniendo en cuenta la problemática social en el mundo.

Cabe mencionar que más del 60% de los nacimientos prematuros se producen en África y Asia meridional, sin embargo, se trata de un verdadero problema mundial. Según la OMS, en alrededor de 184 países estudiados, la tasa de nacimiento prematuro varió ente el 5% y el 18% de los recién nacidos, más de un millón de niños prematuros mueren cada año. En la Tabla 1, se muestra los 10 países con mayor número de nacimientos prematuros para el año 2010.

En la actualidad, se reconoce que el parto pre-término es un síndrome heterogéneo que puede ser tratado entre un 37% y 75 % de los casos [9] aunque se han establecido estrategias basadas en sistemas de puntuación que involucran diversos factores. En la práctica se presentan valores predictivos positivos entre el 17% y 34%.

Tabla 1. Países con mayor número de nacimiento pre-término año 2010. Fuente: Organización mundial de la salud (OMS) [1]

País	Nacidos Pre-término
India	3519100
China	1172300
Nigeria	773600
Pakistán	748100
Indonesia	675700
Estados Unidos de América	517400
Bangladesh	424100
Filipinas	348900
República Dominicana Del Congo	341400
Brasil	279300

En la tabla 1 se presenta la información de los países con mayor tasa de natalidad prematura para el año 2010, como se puede observar Estados Unidos de América, presenta una alta tasa, siendo este dato de interés, puesto que este país es de un alto nivel socio-económico.

3.1.2. Causas y consecuencias de labor pre-término

El parto prematuro se produce por diversos factores. La mayoría de los partos pre-término ocurre espontáneamente, aunque también, en algunos casos se desencadena una inducción precoz de las contracciones uterinas o parto por cesárea, por razones médicas o no [1].

Los factores de riesgo (causas) del parto pre-término incluyen embarazos múltiples, infecciones y enfermedades crónicas, abortos anteriores, nacimientos prematuros anteriores, influencia genética, la edad, abuso doméstico, violencia familiar, estrés, nivel socio-económico, abuso de sustancias tóxicas y peso corporal inadecuado [1] [3].

El nacimiento pre-término es la causa principal de muerte en los primeros 5 años de vida y es una causa del 80% de la morbilidad perinatal. Entre las posibles consecuencias están: El síndrome de dificultad respiratoria, la enterocolitis necrotizante, hemorragia intraventricular y discapacidades a largo plazo, como la

[1] http://www.who.int/mediacentre/factsheets/fs363/es

parálisis cerebral, ceguera, pérdida de la audición y el desarrollo cognitivo del individuo [1].

La tasa de supervivencia presenta una gran brecha en los países del mundo. Por ejemplo, en países de ingresos bajos, la mitad de bebés a las 32 semanas mueren a causa de no recibir cuidados sencillos y eficaces; en contraste, en países de altos ingresos, prácticamente la totalidad de ellos sobrevive.

3.1.3. Índices clínicos del trabajo de parto pre-término

Existen diversas estrategias al momento de valorar un posible parto pre-término, entre los cuales se encuentran: Cambios en el cuello uterino y contracciones en el útero [9].

- **Cambios en el cuello:** Se ha observado que 6 semanas antes del parto, sea pre-término o no, hay cambios cervicales, a partir de los cuales se obtuvo una sensibilidad de 0 a 50% para predecir un parto pre-término desde las 34 semanas gestacionales. La principal desventaja en este tipo de valoración, es la inconstancia relativa al cuello uterino [9].
- **Contracciones uterinas:** Se observa un incremento progresivo de las contracciones 5 semanas antes del parto, con una sensibilidad de 57 a 86% de predicción del parto pre-término.

La electrohisterografía, es una técnica relativamente nueva que se encuentra aún en desarrollo, la cual, explica la existencia de descargas eléctricas espontáneas en el músculo uterino. Este músculo envía disparos intermitentes de potenciales de acción en espiga, de forma que las espigas son múltiples y coordinadas para generar contracciones fuertes y mantenidas. Estas señales son conocidas como señales electromiográficas [9].

3.1.4. Electrohisterografía

Existen dos técnicas usadas en la práctica para el monitoreo de las contracciones uterinas que pueden ser de tipo invasivas y no invasivas. Las técnicas invasivas son procedimientos donde un cuerpo extraño se introduce en el organismo, el catéter de presión intrauterino (IUP) es uno de ellos, por otra parte, las técnicas no invasivas son procedimientos que no involucran instrumentos que rompen la piel o penetran físicamente en el cuerpo, un ejemplo es el tocógrafo fetal o toco instrumento biomédico no invasivo que se encarga de monitorear estado fetal y contracciones

uterinas en trabajo de parto. En general la técnica invasiva presenta una gran desventaja al incrementarse el riesgo de infecciones que puede ocasionar problemas tanto en la madre como en el feto [10] [11], mientras que la segunda técnica al no ser invasiva es segura, pero presenta una baja sensibilidad y exactitud.

Las contracciones uterinas son causadas por la actividad eléctrica en forma de potenciales de acción, de esta forma tanto contracción como relajación del musculo uterino se encuentran asociadas con la actividad eléctrica.

El electrohisterograma (EHG) es el registro de la actividad eléctrica muscular producida en el útero. Este registro se realiza en la superficie abdominal, no es invasivo y ha sido una técnica alternativa para la caracterización en el cambio de actividad eléctrica producida en el útero asociado al progreso del embarazo y a la labor de parto. Es por esto que EHG, puede brindar información completa para un análisis subsecuente [7] [8].

3.2. SISTEMAS COMPUTACIONALES PARA LA PREDICCIÓN DE PARTO PRE-TÉRMINO

En esta sección se describe, los métodos utilizados en la caracterización, la selección y extracción de características de los registros EHG de embarazos a término y pre-término, de manera que, se obtenga un conjunto compacto de características que proporcione toda información necesaria sobre del estado de la paciente. Todas las técnicas consideradas son etapas de un sistema de *machine learning.*

3.2.1. *Machine learning* y reconocimiento de patrones

Machine learning se traduce como aprendizaje automático o aprendizaje de máquina y es una rama del campo de la inteligencia artificial que estudia como las máquinas pueden tomar decisiones a partir de generalizaciones o emulaciones de los seres vivos (bioinspirados).

El reconocimiento de patrones hace parte de *machine learning* y se ocupa de la clasificación (pre-término y término) de objetos o muestras (señales EHG correspondientes a cada paciente) a partir de mediciones o atributos

(características) con enfoques supervisados o no supervisados (con o sin información apriori).

3.2.2. Caracterización

Si se entiende el útero como un sistema bioeléctrico, el cual está compuesto por miles de millones de células que generan un potencial eléctrico, y que a su vez es extraído en forma de registros de señales bioeléctricas, se puede tratar como un sistema complejo y dinámico. Por lo tanto el objetivo de esta etapa es caracterizar este sistema dinámico, de forma que las características también llamadas atributos, sean lo más dicientes a la hora de diferenciar entre un parto pre-término y uno a término.

Como las señales de los registros de EHG son de tipo vibratorio y no estacionario, no es recomendable tomar su morfología como un factor discriminante. Así que se realiza una caracterización en el dominio del tiempo, frecuencia y tiempo-frecuencia como se destaca en investigaciones recientes [7] [12].

3.2.3. Selección y extracción de características

Debido al gran volumen de información que se obtiene en la caracterización, es recomendable representar los datos en una dimensión más baja. Se realiza selección y extracción de características. Esto se hace, con el fin de evitar tres problemas:

- La existencia de atributos irrelevantes.
- La existencia de atributos redundantes.
- La denominada “maldición de la dimensionalidad”.

Los cuales se relacionan con problemas de sobreaprendizaje y de generación de modelos confusos. Al final de esta etapa se obtendrá un conjunto compacto de características ricas en información relevante y discriminante.

El hecho de aplicar la selección de características es determinante para un análisis eficiente, un conjunto de datos que contiene más información de la realmente útil, necesita un uso elevado de procesamiento y memoria que se traduce como alto coste computacional durante el proceso de entrenamiento y análisis.

La tarea de escoger un método no es trivial, ya que existen diversos factores a tener en cuenta como lo son: el rendimiento del sistema, la calidad y cantidad de

información del conjunto de datos, al momento de realizar la clasificación. Existen diferentes técnicas de selección, en este trabajo se destacan ranking y subconjuntos [13]:

- ***Ranking:*** Proporciona una lista de características ordenadas según alguna medida de evaluación, de esta forma a cada característica se le asigna un peso y se ordenan partiendo desde la relevancia con respecto a una característica dominante, las primeras características de la lista conforman el subconjunto final. Algunas de las ventajas de este método de selección están orientadas a las selecciones de un número de características, también, establece que características son importantes y cuáles no, teniendo en cuenta el orden de relevancia entre ellas [14].
- **Subconjuntos:** Recorre un espacio de búsqueda de subconjuntos de características, evaluando subconjuntos completos de características, aunque no recorre un espacio entero, que correspondería a otra técnica de selección de características conocida como búsqueda exhaustiva, sino que evalúa aquellos subconjuntos que sean más prometedores, de esta forma se evalúa el subconjunto de manera conjunta [15].

Desde el punto de vista de la minería de datos, realizar selección y extracción de características es importante debido a la presencia de:

- **Información irrelevante:** Características que no aportan información al sistema, generando problemas de sobre-aprendizaje y son causa de modelos resultantes confusos.

- **Información redundante:** Características que se relacionan (linealmente) entre ellas, desempeñando la misma función. Esto genera dificultades en los algoritmos de aprendizaje.

- **“Maldición de la dimensionalidad”:** El número de características es considerablemente mayor al número de datos, siendo que cada característica representa una dimensión, además, se encuentra directamente relacionado con el nivel de complejidad del modelo resultante [13].

3.2.4. Clasificación

La clasificación de los atributos es la parte más determinante en esta investigación, ya que el propósito de esta sección es agrupar la información del problema biclase planteado, y así dar a conocer que tan buenos son los modelos computacionales a la hora de agruparlos, es por esto que se utilizaron varios algoritmos de clasificación como: Los supervisados y no supervisados [16].

Análisis de clasificación supervisado: En la clasificación supervisada, los algoritmos de clasificación tienen ya un conocimiento del conjunto estudiado, por lo que le es necesario utilizar datos previamente etiquetados con el fin de agrupar en el conjunto o clase que corresponda.

Análisis no supervisado: En el análisis no supervisado se cuenta con objetos que tienen un conjunto de atributos, de las que no se sabe a qué clase o categoría pertenecen, es decir, no se dispone de valores que comparen la información resultante del proceso con la conocida, por eso la finalidad de los métodos no supervisados es agrupar entre tipo de objetos, utilizando un área de entrenamiento disponible.

4. METODOLOGÍA

La metodología propuesta en este estudio comparativo incluye etapas de pre-procesamiento, caracterización, selección de características, reducción de dimensión y clasificación de registros. En la Figura 1 se muestra el diagrama que resume la metodología utilizada en esta investigación, de izquierda a derecha, así: La descripción de la base de datos de señales EHG (sección 4.1), el pre-procesamiento (sección 4.2), caracterización (sección 4.3), selección de características (sección 4.4), reducción de dimensión (sección 4.5), y finalmente, clasificación (sección 4.6).

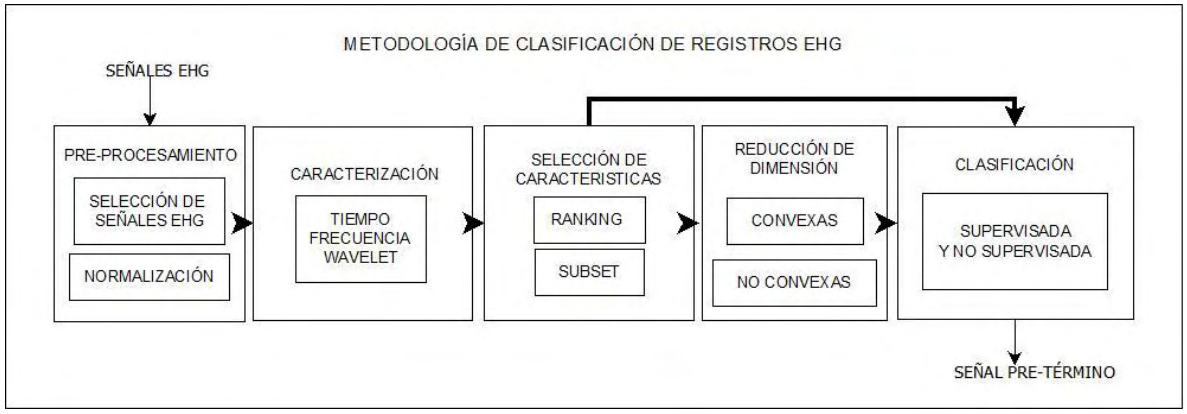


Figura 1. Etapas de la metodología de este estudio. **Fuente:** Esta investigación.

4.1. REGISTROS EHG

En el banco de señales de PHYSIONET se encuentra la base de datos TPDB EHG, recolectada entre 1997 hasta 2006 por el Departamento de Obstetricia y Ginecología, del Centro Médico de Ljubljana, Ljubljana[2]. La base de datos cuenta con registros recolectados de una población general. Asimismo como pacientes admitidos por el hospital con diagnóstico de labor de parto prematuro inminente. La duración de cada registro es aproximadamente 30 minutos, cada señal ha sido digitalizada a 20 muestras por segundo, es decir, con una frecuencia de muestreo de 20 Hz [4].

Cada registro cuenta con la información obtenida por tres canales producidos por una diferencia de potencial en cuatro electrodos superficiales. Los registros fueron recolectados de la superficie abdominal usando cuatro electrodos superficiales de $AgCl_2$, estos electrodos fueron dispuestos en dos filas horizontales, ubicados

[2] https://physionet.org/physiobank/database/tpehgdb/

simétricamente por debajo y encima del ombligo, apartados 7 cm. En la unión de los electrodos fue utilizado un protocolo especial con el fin de mejorar las mediciones [4]. Esto se puede apreciar en la Figura 2.

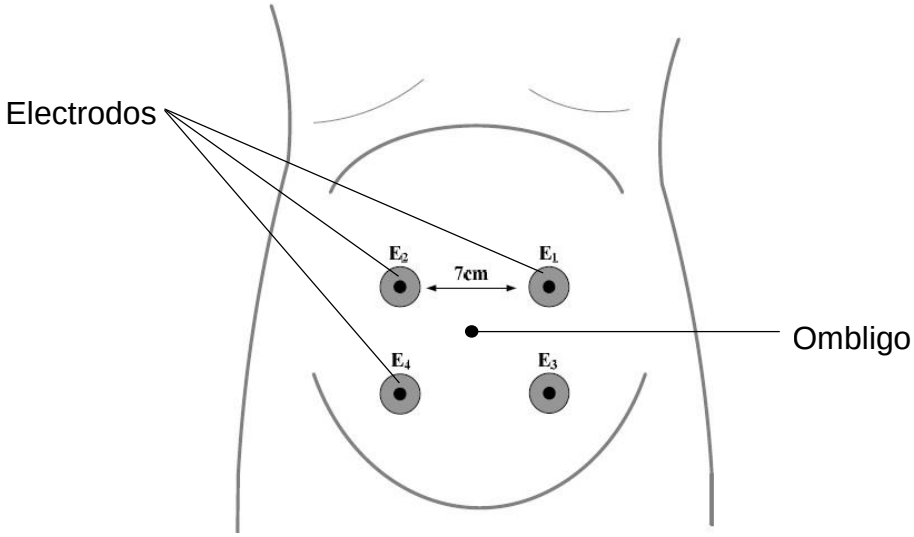


Figura 2. Disposición de electrodos superficiales en el abdomen. Ubicación centrada en el ombligo de cada paciente. **Fuente:** [10]

El primer electrodo (E1) se ubica 3.5 cm por encima del ombligo y 3.5 cm hacia la derecha del ombligo, el segundo electrodo (E2) está ubicado 3.5 cm por encima y 3.5 cm a la izquierda del ombligo, el tercer electrodo se encuentra ubicado a 3.5 cm por debajo del ombligo y 3.5 cm hacia la derecha y finalmente el cuarto electrodo se encuentra ubicado 3.5 cm por debajo del ombligo y 3.5 cm hacia la izquierda [4]. De acuerdo con el protocolo, la resistencia más baja entre los electrodos ha sido menor que 100 Kilo-ohmios (KΩ), las señales adquiridas fueron medidas con la diferencia de potenciales entre estos electrodos, como se observa en las ecuaciones (4.1), (4.2) y (4.3), así:

- Primer canal:

$$S_1(t) = E_2(t) - E_1(t). \quad (4.1)$$

- Segundo canal:

$$S_2(t) = E_2(t) - E_3(t). \quad (4.2)$$

- Tercer canal:

$$S_3(t) = E_4(t) - E_3(t). \quad (4.3)$$

Teniendo en cuenta los efectos producidos por diferentes fenómenos que aportan información innecesaria o ruido que deterioran la calidad de la señal a la hora de realizar el procesamiento de las señales, es conveniente realizar una etapa de pre-procesado de las señales, en la cual se incluye un conjunto de operaciones que toman como entrada las magnitudes medidas por los sensores, para el caso de estudio dichos sensores son electrodos, y se presenta su salida en valores numéricos susceptibles de ser procesados e interpretados, de esta forma, anterior al muestreo se dispuso a filtrar la información contenida en cada canal usando un filtro analógico Butterworth de tripolo con una ancho de banda de 0 hasta 5 Hz [4], con una resolución de 16 bits y con un rango de amplitud de ± 2.5 milivoltios (mv).

Además, considerando el alcance de la investigación y aspectos anteriormente mencionados, algunos errores de grabación fueron inevitables, como, por ejemplo, la falta de datos adjuntos, la pérdida de la señal o conexión interrumpida entre la piel y otros electrodos o falta de actividad eléctrica.

De un total de 1211 registros, fueron escogidos 300 puesto que se realizó una inspección exhaustiva y se rechazaron aquellos archivos de mujeres en embarazo que no contenían actividad eléctrica o contenían ruido excesivo y que terminaron en parto inducido. De los 300 registros seleccionados 262 corresponden a embarazos concluidos en labor de término completo y los 38 restantes con labor de parto pre-término.

4.2. PRE-PROCESAMIENTO

En el caso de esta investigación y para mayor facilidad del modelo matemático. Se decidió comenzar el estudio exploratorio con un número de muestras iguales en cada caso. Entonces, se utilizaron 38 muestras con labor de parto a término completo escogidas aleatoreamente y 38 con labor pre-término.

El resultado del análisis de una señal ha de estar influenciado en gran medida por la selección de los filtros digitales utilizados para eliminar el ruido de las señales anterior a su procesamiento. Se realizó un filtrado de bandas, en este caso se utilizaron tres anchos de banda: 0.08-4Hz, 0.05-4 Hz, 0.2-4 Hz, se reconoció además que las señales uterinas eléctricas contienen rangos de 0 a 5 Hz [4] [17] [18] [19] [20]. Se elige filtros digitales Butterworth en cuatripolo, puesto que tienen una respuesta en frecuencia suave y son computacionalmente no intensivos, sin embrago, presentan un inconveniente debido al cambio de fase. Afortunadamente, el cambio de fase se puede eliminar filtrando dos veces la señal en diferentes direcciones, hacia adelante y luego hacia atrás, de esta forma se obtiene una señal bien filtrada, con fase de desplazamiento en cero.

En la base de datos se ha utilizado los siguientes filtros digitales de cuatripolo aplicados bidireccionalmente a cada señal con los siguientes pasos de banda:

- **Banda 1.** Banda de paso: 0.08-4 Hz, se ha estudiado en investigaciones anteriores [17] [21], en las cuales se ha concluido que, debido al ruido en las frecuencias más bajas debido al estiramiento de la piel y respiración, a menudo frecuente en señales EHG se trabajó con filtros con una mayor frecuencia de corte.

- **Banda 2.** Banda de paso: 0.3-4 Hz, esta banda tanto como la primera y la tercera fueron caso de estudio en [4], aunque mostró buenos resultados, no fueron los mejores con en el caso de la banda 3.

- **Banda 3.** Banda de paso: 0.3-3 Hz, según [4], los mejores resultados se presentaron gracias a este filtro pasa bandas y es por eso que se decidió realizar este estudio comparativo aplicándolo.

De igual manera, se realizó revisión de artículos y se obtuvo que la información del canal 3, es la más favorable para varios casos de estudios similares al de este trabajo de grado [4].

Prosiguiendo con el tema, en la base de datos, cada registro contiene la información de 12 señales, cada señal contiene 30 minutos de grabación aproximadamente, de esta forma. Cada registro se fracciona en 12 partes, cada una corresponde a cada señal en su versión discreta (vector), ordenadas, como se observa en la matriz **R.** En la ecuación (4.4), se muestra la disposición del registro en virtud de las señales que contiene. Como se indicó anteriormente, la señal que se eligió para realizar esta investigación es $\boldsymbol{S}_{3F3}$[22].

$$\boldsymbol{R} = [\boldsymbol{S}_1\ \boldsymbol{S}_{1F1}\ \boldsymbol{S}_{1F2}\ \boldsymbol{S}_{1F3}\ \boldsymbol{S}_2\ \boldsymbol{S}_{2F1}\ \boldsymbol{S}_{2F2}\ \boldsymbol{S}_{2F3}\ \boldsymbol{S}_3\ \boldsymbol{S}_{3F1}\ \boldsymbol{S}_{3F2}\ \boldsymbol{S}_{3F3}\ ], \qquad (4.4)$$

donde, $\boldsymbol{S}_i$ representa una matriz compuestas por las señales correspondientes a cada paciente además:

- Primer canal, sin filtrar ($\boldsymbol{S}_1$).
- Primer canal, filtro pasa bandas Butterworth de 0.08 Hz a 4 HZ ($\boldsymbol{S}_{1F1}$).
- Primer canal, filtro pasa bandas Butterworth de 0.3 Hz a 3 HZ ($\boldsymbol{S}_{1F2}$).
- Primer canal, filtro pasa bandas Butterworth de 0.3 Hz a 4 HZ ($\boldsymbol{S}_{1F3}$).
- Segundo canal, sin filtrar ($\boldsymbol{S}_2$).
- Segundo canal, filtro pasa bandas Butterworth de 0.08 Hz a 4 HZ ($\boldsymbol{S}_{2F1}$).
- Segundo canal, filtro pasa bandas Butterworth de 0.3 Hz a 3 HZ ($\boldsymbol{S}_{2F2}$).
- Segundo canal, filtro pasa bandas Butterworth de 0.3 Hz a 4 HZ ($\boldsymbol{S}_{2F3}$).
- Segundo canal, sin filtrar ($\boldsymbol{S}_3$).
- Segundo canal, filtro pasa bandas Butterworth de 0.08 Hz a 4 HZ ($\boldsymbol{S}_{3F1}$).

- Segundo canal, filtro pasa bandas Butterworth de 0.3 Hz a 3 HZ ($\boldsymbol{S}_{3F2}$).
- Segundo canal, filtro pasa bandas Butterworth de 0.3 Hz a 4 HZ ($\boldsymbol{S}_{3F3}$).

Seguidamente, se han recortado 180 segundos al comienzo y al final de cada señal con el fin de evitar efectos transitorios. Por último, se tomaron 1.394 segundos para cada señal, es decir, cada registro tiene un total de 27.880 muestras, a partir de este punto se inicia la caracterización para cada señal. En la Figura 3, se muestra una señal de labor término en contraste a una señal pre-término después de realizar el pre-procesamiento.

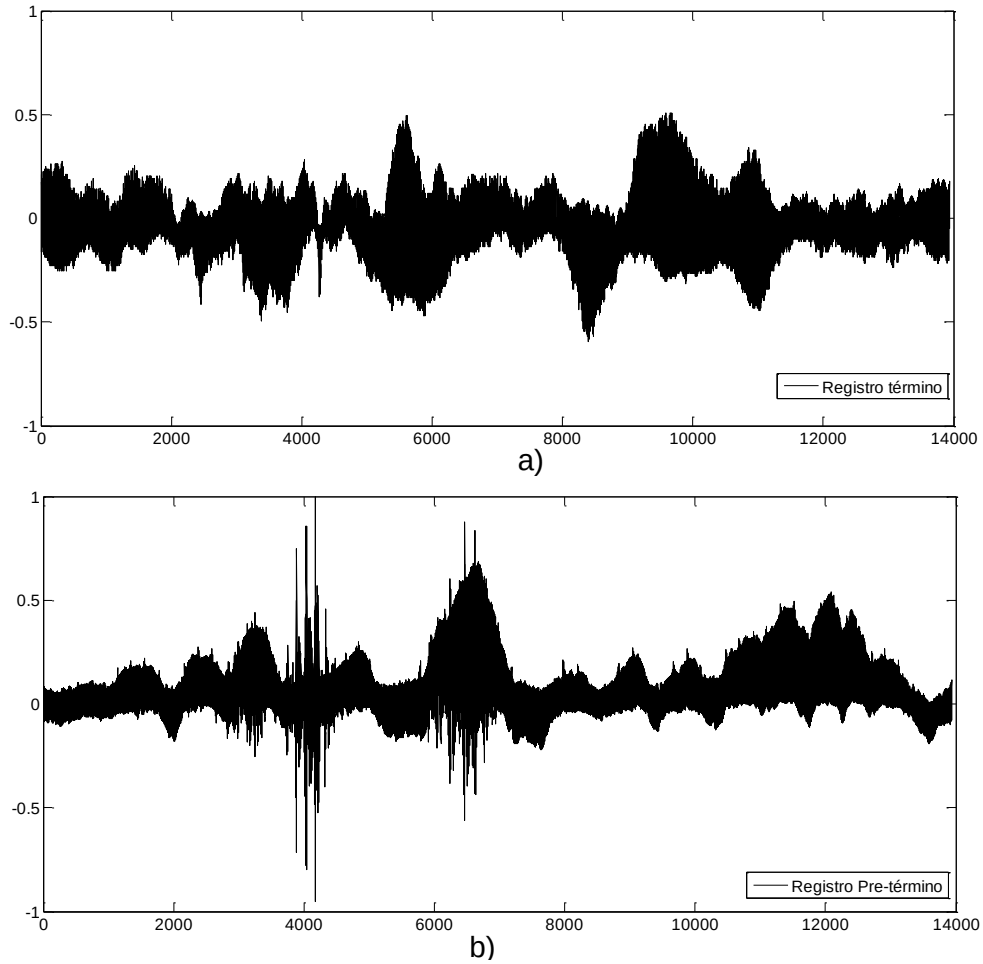


Figura 3. Comparación de registros EHG. a) Señal EHG término, b) señal EHG pre-término del registro $\boldsymbol{S}_{3F3}$. Se grafica los primeros 13.940 puntos correspondientes a 697 segundos de cada registro con una amplitud de voltaje normalizado. **Fuente:** Esta investigación.

4.3. CARACTERIZACIÓN DE SEÑALES

Si se entiende el útero como un sistema bioeléctrico, el cual está compuesto por miles de millones de células, que generan un potencial eléctrico que a su vez es extraído como un registro de señales bioeléctricas, es posible tratarlo con un sistema complejo y dinámico.

Teniendo en cuenta investigaciones anteriores como en [10]. Se realiza una caracterización de las señales, la cual proporciona un conjunto de valores más compacto y representativo, como útil. Se hizo un estudio similar a la hora de caracterizar los registros de esta investigación, el análisis en tiempo, frecuencia y tiempo-frecuencia, como por ejemplo las componentes Wavelet, que fueron muy útiles para el desarrollo de la caracterización de los registros [15].

Las características que representan los registros de EHG son:

Área bajo la curva (A**)**: Corresponde al área de cada registro de EHG representando así el valor total de un grupo de muestras tomadas en cada uno. La expresión matemática se muestra en la ecuación (4.5), donde $\boldsymbol{N}$ es el número de las muestras tomadas y $|\boldsymbol{x_n}|$ los valores de cada muestra [4] [23]*,*

$$\mathrm{A} = \sum_{\mathrm{n}=1}^{\mathrm{N}} |(\mathrm{x_n})|. \tag{4.5}$$

Media ($\bar{X}$): La media o también llamada promedio, es el valor obtenido al sumar todos los datos y dividir el resultado entre el número total de datos, expresada en la ecuación (4.6). $\bar{X}$ es el promedio de la señal, N es el tamaño de cada registro que para EHG es de 27880 muestras, en $|x_n|$ se toma el valor de cada muestra siempre positivos [4] [23].

$$\bar{X} = \frac{1}{N}\sum_{n=0}^{N} |x_n|. \tag{4.6}$$

Raíz Media cuadrática (VRMS): Se refiere al cálculo del comportamiento dinámico de un sistema, es una buena opción a la hora de caracterizar los registros de EHG. En la ecuación (4.7) se expresa matemáticamente VRMS, donde N es el tamaño del registro y x_n el valor para cada muestra [4] [23]:

$$VRMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}. \tag{4.7}$$

Varianza (V^2**)**: Es utilizado para la caracterización de la amplitud de la distribución de los valores instantáneos de señales aleatorias alrededor del valor medio, identificando la media de las desviaciones cuadráticas de una variable de carácter aleatorio. La ecuación (4.8), muestra la definición de la varianza en variable discreta donde N es el tamaño del registro de EHG, $\mathrm{x_i}$ el valor de cada muestra y $\bar{\mathrm{x}}$ la media de cada registro [4] [23]:

$$\mathrm{V^2} = \frac{\sum_{\mathrm{i=1}}^{N}(\mathrm{x_i} - \bar{\mathrm{x}})^2}{N-1}. \tag{4.8}$$

Desviación estándar (σ**)** : Representa una medida de distribución de los valores instantáneos alrededor del valor medio, dicho en otras palabras, nos indica cuánto pueden alejarse los valores respecto a la mediana. Matemáticamente es representada por la raíz cuadra de la varianza como se observa en la ecuación (4.9) [4] [23]:

$$\sigma = \sqrt{\frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i} - \bar{\mathrm{x}})^2}{\mathrm{n}-1}}. \tag{4.9}$$

Entropía ($\mathrm{H(X)}$**):** Cuantifica el grado de irregularidad presente en la señal, de tal forma que si la señal es perfectamente regular y predecible, el valor de la entropía será mínima, y en el caso contrario si la señal es muy compleja, irregular e impredecible su entropía será máxima. Su fórmula matemática se expresa en la ecuación (4.10), donde $\mathrm{H(X)}$ representa la medida de información de la fuente y $\mathrm{p(x)}$ es la probabilidad del conjunto de registros [4] [23]:

$$\mathrm{H(X)} = -\sum_{\mathrm{x\in A}} \mathrm{p(x)log_2 p(x)}. \tag{4.10}$$

Frecuencia pico (FP**):** Rango de frecuencia donde los registros de EHG tienen mayor potencia. La ecuación (10) muestra la expresión matemática de la frecuencia pico [4] [23]:

$$\mathrm{FP} = \arg\left(\frac{f_s}{N} max_{i=0}^{N-1} P(i)\right), \tag{4.11}$$

donde f_s es la frecuencia de muestreo, N es el número de muestras y $P(i)$ es el i-ésimo pico.

Transformada de Wavelet: Existen dos tipos de señales, las estacionarias y las no estacionarias. En el caso del análisis de señales estacionarias, donde la frecuencia no varía, la trasformada de Fourier es una muy buena herramienta para análisis de estas señales, porque la descompone es sus componentes sinusoidales.

Sin embargo, para el análisis de los registros de EHG, los cuales tienen un comportamiento no estacionario, Fourier no es recomendable debido a que se perderá información. Entonces, la trasformada de Wavelet representa una alternativa como herramienta para la descomposición y reconstrucción de señales de este tipo. La trasformada de Wavelet analiza señales que presentan cambios abruptos en sus componentes de tiempo-frecuencia. En la Figura 4, se presenta de forma gráfica el análisis por trasformada de Fourier y transformada de wavelet.

Dentro de los usos de esta poderosa herramienta se incluye, además del análisis local de señales no estacionarias, el análisis de señales electrocardiográficas, sísmicas, de sonido, de radar, así como también, es utilizada para la compresión y procesamiento de imágenes y el reconocimiento de patrones [15] [23].

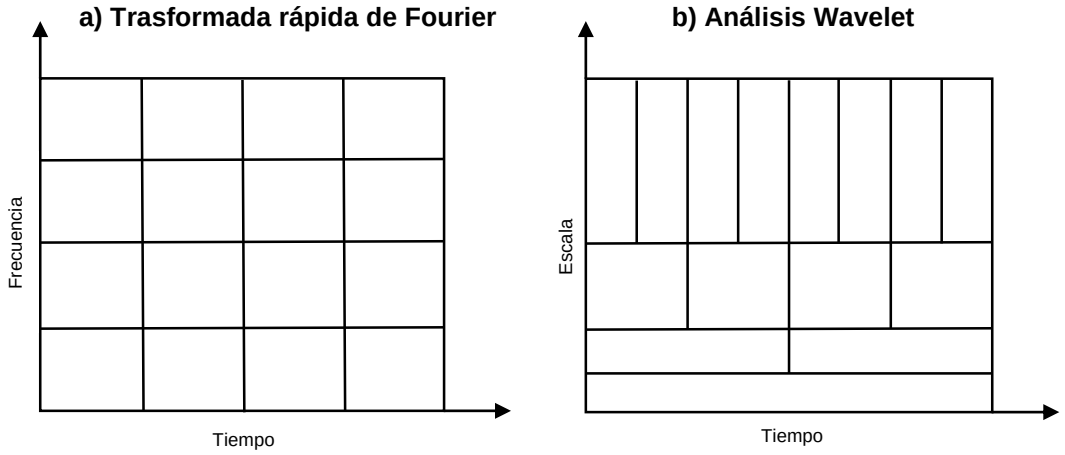


Figura 4. Representación de las transformadas de Fourier y Wavelet. En (a) se muestra la forma gráfica de representar la trasformada rápida de Fourier STFT (tiempo-frecuencia), en (b) se muestra una forma gráfica del análisis en Wavelet (tiempo-escala). **Fuente:** [48].

La Transformada Wavelet no es solamente local en tiempo, sino también en frecuencia. A diferencia de Fourier, en donde las funciones base son senos y cosenos de duración infinita, en el análisis Wavelet la base son funciones localizadas en frecuencia (dilatación) y en tiempo (traslación). Una Wavelet es una "pequeña onda" de duración limitada, es decir, su energía está concentrada en el tiempo alrededor de un punto. Se puede observar en la *Figura 5*, una señal seno normal, en contraste con una Wavelet.

a) Función sinusoidal **b) Función madre Daubechie**

Figura 5. Función sinusoidal y función madre Daubechie. Comparación de a) una señal sinusoidal normal y b) una onda Wavelet. **Fuente:** [48]

Por la complejidad de los registros, debido a la variabilidad en forma continua de los parámetros de escala como de traslación, es indispensable contar con una herramienta que permita la discretización de esta. Por lo anterior se pasará de un mapeo continuo a un espectro o conjunto finito de valores, redefiniendo la integral por una aproximación con una sumatoria, como es expresa en la ecuación (4.12). La discretización permite representar una señal en términos de funciones elementales $c\ y\ \varphi$ acompañadas de coeficientes,

$$f(t) = \sum_{\lambda=0} c_\lambda \varphi_\lambda, \tag{4.12}$$

en los sistemas wavelet, las wavelet madre $\Psi(t)$ traen consigo unas funciones de escala $\Phi(t)$, las primeras son las encargadas de representar los detalles finos de la función, mientras que las funciones de escala realizan una aproximación, entonces es posible hacer una representación de la señal $f(t)$ como una sumatoria de funciones de Wavelet y funciones de escala, expresada en la ecuación (4.13):

$$\boldsymbol{f}(t) = \sum_{k=0} c_{j_0,k} \Phi_{j_0,k}(t) + \sum_{j=j_0} \sum_{k=0} d_{jk} \Psi_{j_0,k}(t) \qquad j,k \in Z^+, \tag{4.13}$$

Donde c_j son los coeficientes de escala, d_j son los coeficientes de wavelet y j_0 nos entrega el espacio inicial que será el espacio de menor resolución, dependiendo de este j_0 es que el resto de los índices seguirán corriendo [55] [].

El objetivo del análisis multi-resolución que genera la transformada de wavelet, consiste en expandir una señal, en una base de funciones cuyas propiedades tiempo-frecuencia se adapten a la estructura local de la señal. Esta trasformada

permite obtener el desarrollo de una señal en una base ortonormal de funciones wavelets [4] [22] [23].

En la caracterización se utiliza la transformada de wavelet correspondiente a la función madre *Symlet 5*, en [24] se observa que es la función más representativa para este tipo de señales.

En la Figura 6, se observa la familia de la función *Symlet* a la cual corresponde la función madre utilizada en la investigación correspondiente. Según [24], el nivel de descomposición apropiados para este tipo de señales no estacionarias es de 5.

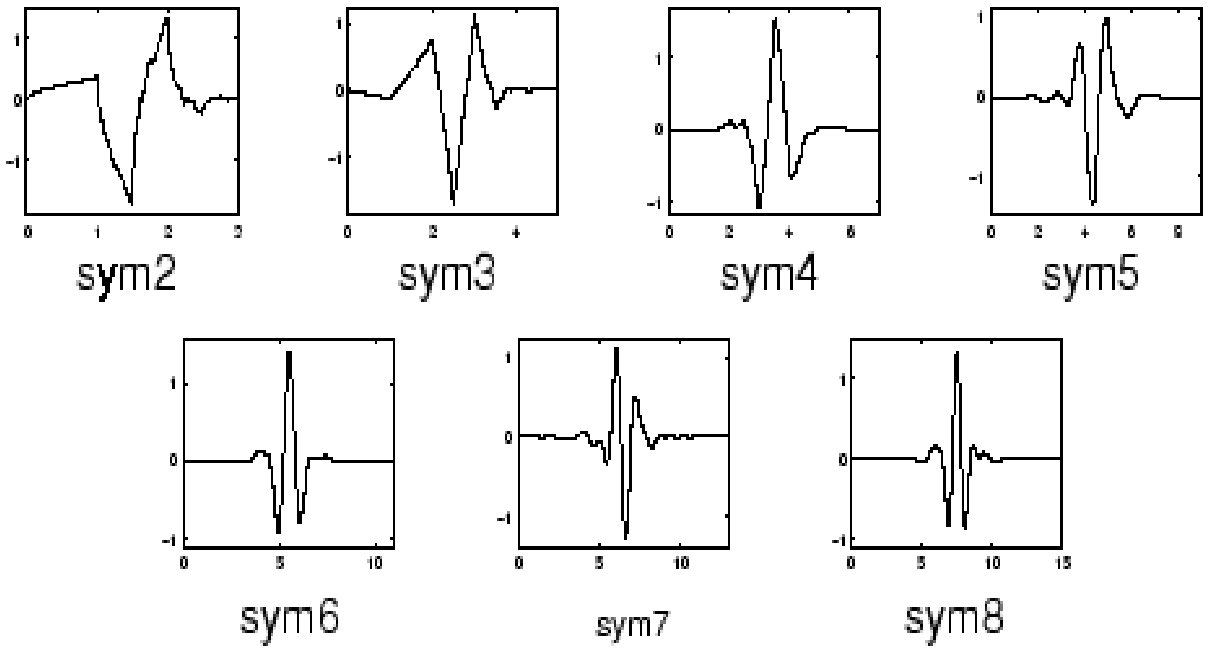


Figura 6. Familia de funciones madre Wavelet *Symlet.* La función madre *Symlet* 5 es apropiada para el análisis de esta investigación. **Fuente:** [24]

En la Figura 7**,** se muestra la descomposición en cinco niveles de una señal pre-término, utilizando la función madre sym 5. Como se puede observar cada nivel descomposición y aproximación se deriva del anterior.

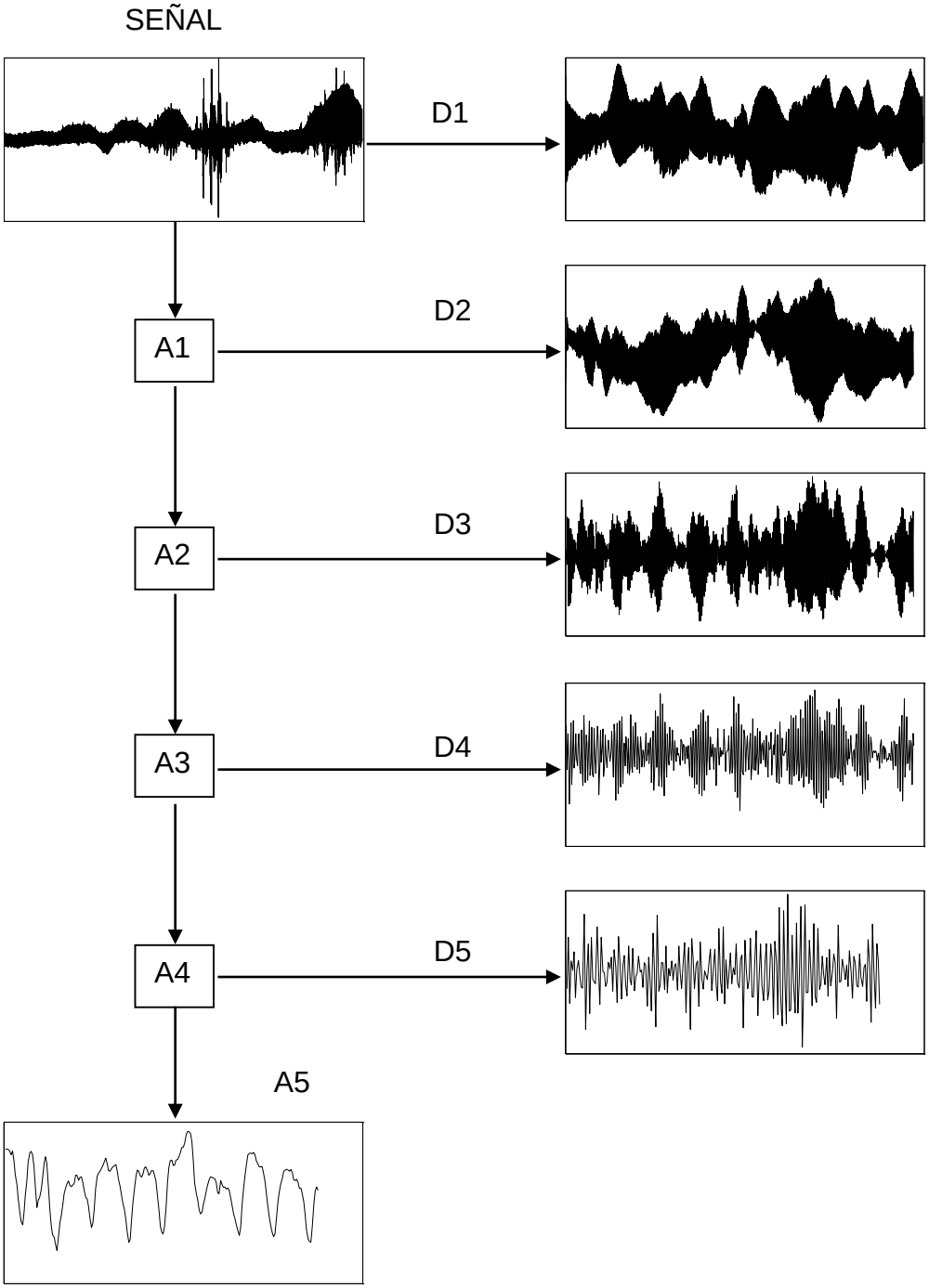
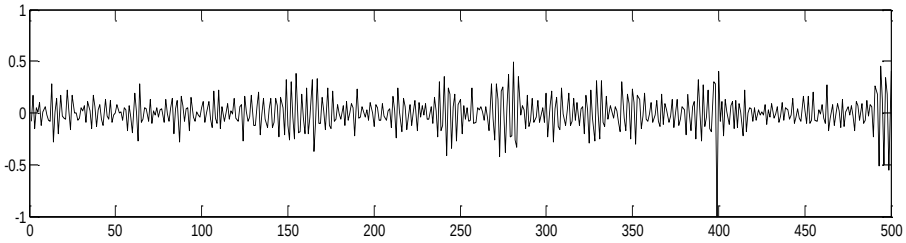
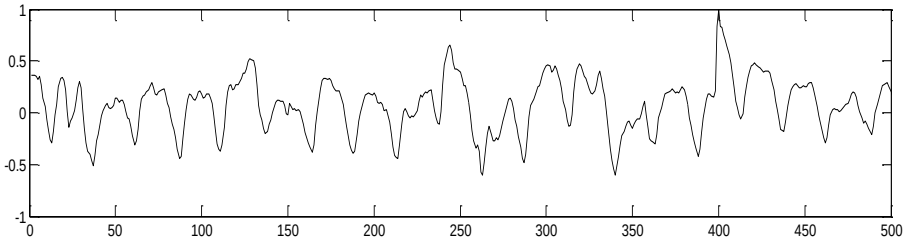


Figura 7. Descomposición Wavelet de una señal EHG Pre-término con una función madre *Symlet* 5. En esta imagen se puede observar que D1, D2, D3, D4 y D5 son los componentes de detalle correspondientes a cada nivel nombrado, de manera similar, A1, A2, A3, A4 y A5, son los componentes de aproximación para cada nivel **Fuente:** Esta investigación

En consecuencia y con el fin se observar una comparación entre las dos clases de señales se presenta en la Figura 8 los coeficientes de detalle y de aproximación correspondientes a un registro EHG pre-término con el quinto nivel de descomposición.

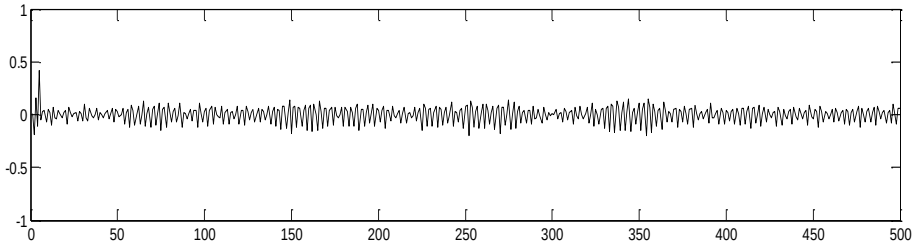


a) Componente de detalle señal pre-término

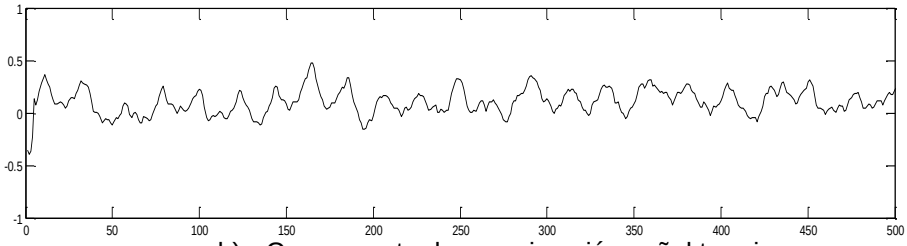


b) Componente de aproximación señal pre-término

Figura 8. Componentes de detalle y aproximación de una señal EHG pre-termino. a) Componente de detalle. b) Componente de aproximación. Correspondientes a 500 muestras y quinto nivel de descomposición de Wavelet con función madre Symlet 5 de una señal EHG pre-término. **Fuente:** Esta investigación.



a) Componente de detalle señal termino



b) Componente de aproximación señal termino

Figura 9. Componentes de detalle y aproximación de una señal EHG a término. a) Componente de detalle. b) Componente de aproximación. 500 muestras

correspondientes al quinto nivel de descomposición de Wavelet con función madre Symlet 5 de una señal EHG término. **Fuente:** Esta investigación.

En la Figura 9 se muestra los coeficientes de detalle y de aproximación correspondientes a un registro EHG término con el quinto nivel de descomposición. Estas señales se obtuvieron a partir de los registros mostrados en la Figura 3. Consecuente a lo descrito anteriormente, se encuentra estimadores para los componentes de descomposición, equivalentes a las características en tiempo-frecuencia mencionadas en esta sección.

Finalmente, en la Tabla 2 se resume las características que son utilizadas en esta investigación, utilizadas tanto en las señales correspondientes a cada registro como a los niveles de descomposición de cada señal.

Tabla 2. Formulación matemática de las características representativas para EHG.

Característica	**Formulación Matemática**
Área bajo la Curva	$I = \sum_{n=1}^{N} \lvert(x_n)\rvert$
Media	$\bar{X} = \frac{1}{N}\sum_{n=0}^{N} \lvert x_n \rvert$
Raíz Media cuadrática (VRMS)	$VRMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$
Varianza	$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$
Desviación estándar	$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
Entropía	$H(X) = -\sum_{x \in A} p(x) log_2 p(x)$
Frecuencia Pico	$\arg\left(\frac{f_s}{N} max_{i=0}^{N-1} P(i)\right)$
Componentes Wavelet	$\boldsymbol{f}(t) = \sum_k \sum_j C_{jk} \Phi(t) + \sum_k \sum_j d_{jk} \Psi(t)$

4.4. SELECCIÓN DE CARACTERÍSTICAS

En el estudio realizado, cada señal cuenta con 27.880 características; la selección de características se lleva a cabo con el método mejor primero correspondiente a ranking y decisión de árboles para subconjuntos. Seguidamente para depurar las características obtenidas se utilizó la técnica de búsqueda exhaustiva.

- **Mejor primero:** Dado las características $A_1, A_2, \dots, A_3$, se evalúa cada A_i de manera independiente, calculando medias de correlación entre cada característica de la matriz $\boldsymbol{x}$ y la componente de la matriz $\boldsymbol{y}$**,** correspondiente a la clase. Se denota en forma matemática en la ecuación (4.14) [25]:

$$I(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{i}}\sum_{\boldsymbol{j}} p(\boldsymbol{x}=i,\boldsymbol{y}=j) log\left[\frac{p(\boldsymbol{x}=i,\boldsymbol{y}=j)}{p(\boldsymbol{x}=i)p(\boldsymbol{y}=j)}\right], \qquad (4.14)$$

  donde i son los valores de las características de la matriz $\boldsymbol{x}, j$ son los valores de la matriz de clase **y**.

- **Decisión de árboles (algoritmo j48):** Este algoritmo construye arboles de decisión a partir de un conjunto de ejemplos, constituidos por un conjunto de atributos y un clasificador; con el fin de elegir que atributos y en qué orden aparecen en el árbol se utiliza como función de evaluación la minimización de la entropía [26].

- **Búsqueda exhaustiva:** Consiste en realizar las combinaciones que aporten mayor información a la clasificación, evaluándose 2^n conjuntos, donde n, representa el número de atributos [25].

Particularmente, en esta investigación se procedió a realizar la selección inicialmente usando los algoritmos de árbol de decisión y mejor primero, los cuales fueron potencializados usando otro método de selección en cascada, en este caso, búsqueda exhaustiva. De esta forma se obtuvo una matriz de atributos más compacta y apropiada a la hora de representar y clasificar este fenómeno. En la Tabla 3 se registra el número de características resultantes por cada técnica de selección.

Tabla 3. Número de características resultantes para cada técnica de selección utilizada en esta investigación.

Técnica de selección inicial	Características seleccionadas	Técnica de selección en cascada	Características seleccionadas
Mejor primero	33	Búsqueda exhaustiva	6
Árbol de decisión	45	Búsqueda exhaustiva	7

4.5. REDUCCIÓN DE DIMENSIÓN

Con el fin de generar una representación aún más compacta y generar los resultados de clasificación, se emplea además técnicas de reducción de dimensión sobre las características seleccionadas.

En este estudio comparativo se considera técnicas convencionales y no convencionales, algunas de las técnicas que se incluyen en el estudio comparativo son: Análisis de componentes principales (PCA), ISOMAP, empotramiento localmente lineal (LLE), análisis del espacio tangencial local (LTSA), análisis del espacio tangencial local lineal (LLTSA) empotramiento preservado de vecinos (NPE) y mapeo de Sammon (MDS) [27].

4.5.1. Técnicas convexas

Esta categoría de técnicas de reducción de dimensión busca optimizar una función objetivo, que no contenga en sí misma una función óptima local, por lo tanto, el espacio de solución es convexo. Por lo general la mayoría, de las técnicas de reducción de dimensión caen en esta categoría, y se puede apreciar en la ecuación (4.15):

$$\varphi(Y) = \frac{\boldsymbol{Y^T A Y}}{\boldsymbol{Y^T B Y}}, \tag{4.15}$$

donde Y es la matriz reducida de características y los términos A y B son matrices a definir de acuerdo con la naturaleza del método.

La forma de esta función puede ser optimizada para ser solución generalizada de los problemas propios. En las siguientes subsecciones se explicará cómo esta técnica de reducción de dimensión convexa puede dividirse en dos subcategorías, que forman parte de una descomposición propia de una matriz completa (espectro completo) o de una matriz dispersa (espectro disperso).

Espectro Completo: Este tipo de técnicas de reducción de dimensión se realizan mediante la descomposición propia de una matriz completa, la cual captura las covarianzas entre las dimensiones o similitudes pares entre los puntos de los datos. En este estudio comparativo se utilizaron dos técnicas, las cuales corresponden a PCA e ISOMAP.

- **Análisis de Componentes Principales (PCA,** de sus siglas en inglés**):** El PCA es una técnica de reducción de dimensión convexa en espectro completo lineal. Esto quiere decir que se realiza la reducción de dimensión mediante la incorporación de datos en un sub-espacio lineal de dimensionalidad menor. Mediante PCA se construye una representación de baja dimensión de los datos que describe la varianza de estos tanto como sea posible. Esto se logra gracias a que se encuentra una base lineal de dimensionalidad reducida para

los datos, en la cual la varianza en los datos es máxima, de esta forma, se encuentre una correlación lineal M, la cual maximice el rastreo de la función de coste. Dicho de otra forma, esta técnica identifica un primer componente que presente la mayor cantidad de varianza, un segundo componente que tenga la siguiente mayor cantidad y así sucesivamente, en términos matemáticos se expresa en la ecuación (4.16) [27] [28] [29]:

$$\sum \boldsymbol{M}\,(\boldsymbol{X}) = \boldsymbol{\lambda M}, \tag{4.16}$$

donde $\sum \boldsymbol{M}\,(\boldsymbol{X})$ es la matriz de la covarianza de las muestras, $\boldsymbol{X}$ es la matriz de características y $\boldsymbol{\lambda}$ es la matriz de valores propios.

Ahora bien, el problema propio se soluciona para la diagonal principal $\boldsymbol{d}$ de valores propios $\boldsymbol{\lambda}$. La baja dimensionalidad de los datos representados en Y_i de los puntos X_i, se calculan asignándolas en la base lineal M, de esta forma se tiene la ecuación (4.17)

$$\boldsymbol{Y} = \boldsymbol{XM}, \tag{4.17}$$

es válido afirmar que PCA es idéntica a una técnica multidimensional de escalado llamada escalado clásico [30], en el cual, se tiene una matriz de distancia Euclidiana **D**, cuyas entradas d_{ij} representan la distancia euclidiana entre los puntos de alta-dimensinalidad x_i y x_j. Esta técnica encuentra un mapeo lineal M que minimiza la función de costo expresada matemáticamente en la ecuación (4.18):

$$\varphi(Y) = \sum_{ij} \left(d_{ij}^2 - \left\| y_i - y_j \right\|^2 \right), \tag{4.18}$$

donde $\left\| y_i - y_j \right\|^2$ es el cuadrado de la distancia euclidiana entre los puntos de baja dimensionalidad y_i y y_j, y_i esta restringido para que $x_i \boldsymbol{M}$ y $\left\| m_j \right\|^2 = 1$ para $\forall j$. Demostrado en [30] [31], la minimización de la función de costo está dada por la descomposición propia de la matriz $\boldsymbol{K} = \boldsymbol{XX^T}$ de alta dimensionalidad, la entrada de esta matriz pueden ser obtenida en la ecuación (4.19):

$$k_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_{l} d_{il}^2 - \frac{1}{n}\sum_{l} d_{jl}^2 + \frac{1}{n^2}\sum_{lm} d_{lm}^2 \right). \tag{4.19}$$

- **ISOMAP:** Es una técnica que resuelve el problema que presenta PCA, que consiste en retener distancias euclidianas pares y no tiene en cuenta la

distribución de los datos vecinos. A pesar de ser muy parecido al escalamiento clásico, es una técnica de éxito en la reducción de dimensión. La propuesta de ISOMAP es preservar las distancias geodésicas o curvilíneas por pares entre los puntos de datos, es decir la distancia entre dos puntos medidos sobre el colector [27] [34].

Las distancias geodésicas entre los puntos de datos $\boldsymbol{x}_i$ $(i = 1,2, \dots, n)$, se calculan construyendo una gráfica de vecindad G, en la cual cada punto de datos x_i está conectado con sus K vecinos más cercanos $\boldsymbol{x}_{ij}(j = 1,2,..,k)$ en el conjunto de datos $\boldsymbol{X}$, de esta forma el camino más corto entre dos puntos es la estimación de la distancia geodésica que puede ser fácilmente calculada utilizando los algoritmos de trayectoria más corta de Dijkstra o de Floyd [35] [36]. Las distancias geodésicas entre todos los puntos de datos en $\boldsymbol{X}$ se calculan, formando así una matriz por parejas. Las representaciones de dimensiones bajas y_i de los puntos de datos x_i en el espacio de baja dimensión $\boldsymbol{Y}$ se calculan aplicando la escala clásica, explicado en la anterior subsección para la matriz de distancia geodésica resultante. Esta técnica presenta una dificultad, debido a su inestabilidad topológica [35], la cual construye conexiones erróneas en el gráfico de la vecindad G que afecta terriblemente los resultados.

Espectro Disperso: En la anterior subsección se revisaron técnicas de reducción de dimensión convexa que corresponden a un espectro completo, ahora se analizan técnicas de espectro disperso las cuales están enfocadas a la retención de una estructura local de datos, las cuales están enfocadas a la retención de una estructura local de datos, las técnicas que fueron de interés en el estudio comparativo son:

- ***Método de Empotramiento Localmente Lineal (LLE):***
  Empotramiento Local Lineal (LLE) [27] [34], siendo una técnica de reducción de dimensión convexa en espectro disperso, es similar a la técnica de ISOMAP anteriormente descrita, en la cual se construye una gráfica con los puntos de los datos, sin embargo, difiere en que intenta preservar únicamente las propiedades de los datos, además, es menos sensible a los cortocircuitos, puesto que de presentarse solo se verán afectadas un pequeño número de propiedades locales, de esta forma, la preservación de las propiedades locales permite la incorporación exitosa de los colectores no convexos.

  Estas propiedades locales del colector de datos se construyen escribiendo los puntos de datos de alta dimensionalidad como una combinación lineal de sus vecinos más cercanos, ahora, en la representación de datos de baja dimensionalidad, esta técnica intenta retener los pesos de reconstrucción en las combinaciones lineales tan buena como sea posible.

Esta técnica describe los datos x_i, como una combinación lineal $w_{i,}$conocida como pesos de reconstrucción de sus *K* vecinos más cercanos x_{ij}. Es decir, LLE se ajusta a través de un hiperplano mediante los puntos x_i y sus vecinos más cercanos, asumiendo que el colector es localmente lineal, de esta forma, los pesos w_i, son invariantes a la traducción, rotación y reescalado. Esto implica que cualquier mapeo lineal del hiperplano a un espacio menor de la dimensionalidad conserva los pesos de reconstrucción del espacio. En consecuencia, la representación de datos de dimensiones equivale a minimizar la función de coste, tal como se muestra en la ecuación (4.20)

$$\varphi\,(Y) = \sum_i \left\| y_i - \sum_{j=1}^{k} w_{ij} y_{ij} \right\|^2, \qquad (4.20)$$

cuando $\left\|y^{(k)}\right\|^2 = 1\ para\ \forall\ k,$ donde $y^{(k)}$es la k-esima columna de la solución de la matriz $\boldsymbol{Y}$. Para excluir la solución trivial $Y = 0$, es necesaria la restricción sobre las varianzas de las columnas de $\boldsymbol{Y}$.

- **Empotramiento preservado de vecinos (NPE):** El empotramiento preservado de vecinos es una técnica de reducción de dimensión lineal que resulta de una aproximación del algoritmo LLE [27], descrito anteriormente, en el cual la estructura local puede ser representada como una combinación lineal de sus vecinos para cada punto. Ahora bien, si se supone que cada punto tiene K vecinos cercanos, entonces, se puede caracterizar la geometría local dada por los coeficientes lineales que reconstruyen cada punto de datos desde sus vecinos. Los errores de reconstrucción se miden por la función de costo, que se muestra en la ecuación (4.21):

$$\varphi\,(\boldsymbol{W}) = \sum_i \left\| \boldsymbol{x_i} - \sum_j w_{ij} \boldsymbol{x_j} \right\|^2, \qquad (4.21)$$

Considerando el problema de mapeo original de los puntos a la lineal de cada punto sobre la línea que puede ser representada como una combinación lineal de sus vecinos con los coeficientes w_{ij}, donde **y** = $(y_1,\ y_2, \ldots, y_3)^T$como un mapa. De esta forma, se escoge un criterio justo que minimice la siguiente función de costo, expresada en la ecuación (4.22):

$$\varphi\,(\boldsymbol{y}) = \sum_i \left(\boldsymbol{y_i} - \sum_j w_{ij} \boldsymbol{y_j} \right)^2. \qquad (4.22)$$

- ***Análisis del espacio Local Tangencial (LTSA) y Análisis del espacio local Tangencial Lineal (LLTSA):*** Son técnicas que describen las propiedades locales de datos de alta dimensión usando el espacio local tangencial en cada punto de los datos [37], se basan en asumir la linealidad local del colector, de esta forma existe un mapeo lineal desde un punto de los datos de alta dimensionalidad a su espacio local tangente, y demás, que existe una correlación lineal de datos de baja dimensionalidad con el mismo espacio tangente local.

  Estas técnicas buscan simultáneamente las coordenadas de las representaciones de datos de baja dimensión, y para las asignaciones lineales de los puntos de datos de baja dimensión para el espacio local tangencial de los datos de alta dimensión. De acuerdo con lo anterior, se comienza con el análisis para los puntos $\boldsymbol{x}_i$, del espacio local tangencial, hecho esto aplica PCA sobre los K puntos de los datos $\boldsymbol{x}_{i_j}$, que son vecinos de $\boldsymbol{x}_i$, resultando en un mapeo $\boldsymbol{M}_i$ de vecinos $\boldsymbol{x}_i$ para un espacio local tangencial $\boldsymbol{\theta}_i$, además, existe un mapeo lineal $\boldsymbol{L}_i$ de coordenadas $\boldsymbol{\theta}_{i_j}$ de baja dimensión representadas en $\boldsymbol{y}_{i_j}$. De esta forma una representación matemática de ambas técnicas se define en la ecuación (4.23):

$$\min_{\boldsymbol{Y}_i, \boldsymbol{L}_i} \sum_i \|\boldsymbol{Y}_i \boldsymbol{J}_k - \boldsymbol{L}_i \boldsymbol{\theta}_i\|^2, \tag{4.23}$$

  donde $\boldsymbol{J}_k$ es la matriz de centrado, de tamaño K. La solución de $\boldsymbol{Y}$ es mostrada en [37], la cual está formada por vectores propios de una matriz de alineación $\boldsymbol{B}$, correspondiendo a los d valores más pequeños de los vectores propios diferentes de cero. Las entradas de la matriz $\boldsymbol{B}$ se obtuvieron debido a una sumatoria iterativa de todas las matrices $\mathbf{V}_i$, expresada matemáticamente en la ecuación (4.24):

$$\boldsymbol{B}_{N_i N_i} = \boldsymbol{B}_{N_{i-1} N_{i-1}} + \boldsymbol{J}_k \left(\boldsymbol{I} - \mathbf{V}_i \mathbf{V}_i^T\right) \boldsymbol{J}_k, \tag{4.24}$$

  donde $\boldsymbol{N}_i$ es el conjunto de índices de vecinos cercanos del punto $\boldsymbol{x}_i$.

4.5.2. Técnicas no convexas

En esta sección se trabajará una técnica de reducción de dimensión que optimiza una función objetivo que no es convexa, siendo una técnica de escalamiento multidimensional, que forma parte de una alternativa para el escalamiento clásico llamada mapeo de Sammom.

- **Mapeo de Sammon (MDS)**: Adapta una función de coste a la escala clásica, ponderando la contribución de cada par de puntos (i, j) a la función de coste por la inversa de su distancia entre la pareja de puntos en el espacio $\boldsymbol{d}_{ij}$ de alta dimensionalidad, de esta forma, la función de coste asigna aproximadamente el mismo peso a la retención de cada una de las distancias entre las parejas y de esta forma retiene la estructura local de los datos mejor que la escala clásica, matemáticamente la función de coste de Sammon se expresa en la ecuación (4.25):

$$\varphi(Y) = \frac{1}{\sum_{ij} d_{ij}} \sum_{i \neq j} \frac{(d_{ij} - \|y_i - y_j\|)^2}{d_{ij}}, \tag{4.25}$$

donde d_{ij} representa la distancia euclidiana entre los puntos de alta dimensionalidad $\boldsymbol{x}_i$ y $\boldsymbol{x}_j$ [27] [38] [39]. La minimización de la función de costo de Sammon se realiza generalmente utilizando un método Seudo-Newton [40].

4.6. CLASIFICACIÓN

Los métodos de clasificación que se utilizaron fueron los supervisados y no supervisados, en los métodos supervisados es necesario recurrir a expertos, para tener una información a priori para etiquetar el conjunto de entrenamiento (termino y pre-término), mientras que los no supervisados son más flexibles a la hora de etiquetar los atributos, ya que, al contario de los supervisados, no es necesario una información a priori.

4.6.1. Clasificación supervisada

Para la clasificación supervisada, el clasificador tiene ya un conocimiento de cada clase de patrón, por lo que le es necesario utilizar datos previamente etiquetados para poderlos agrupar en el conjunto patrones en la clase que corresponda. Estos clasificadores se dividen en varios grupos dependiendo de su algoritmo de trabajo, aquí se nombran tres subgrupos:

Clasificadores polinomiales de grado lineal y superior:

- **Clasificador lineal por expansión de PCA en los datos conjuntos (PCLDC** por sus siglas en inglés**)**: Su funcionamiento hace uso de análisis de componentes principales (PCA por sus siglas en ingles), técnica que es utilizada para la reducción de dimensión, donde los conjuntos de atributos son analizados de tal forma que se busca una nueva proyección, formando datos que pueden representar mejor dichos conjuntos y así ser clasificados [41].

- **Clasificador Linear logístico (LOGLC** por sus siglas en inglés): este modelo se obtiene de la discriminante de Bayes, siendo un método parcialmente paramétrico para la clasificación de observaciones multivariables $\mathbf{x} = [\,\mathrm{x}_{1,\ldots,}\,\mathrm{x}_{\mathrm{p}}]$, vector de interés de una de varias poblaciones $P_1, \ldots, P_g$. [42] [43]. Matemáticamente se expresa en las ecuaciones (4.26) y (4.27):

$$p(A|\mathrm{x}) = \frac{p(A)p(\mathrm{x}|A)}{p(A)p(\mathrm{x}|A) + p(B)p(\mathrm{x}|B)} = \frac{e^{\mathrm{w}^{\mathrm{T}}\mathrm{x}+w_0}}{1 + e^{\mathrm{w}^{\mathrm{T}}\mathrm{x}+w_0}} = \frac{1}{1 + e^{-\mathrm{w}^{\mathrm{T}}\mathrm{x}-w_0}} \qquad (4.26)$$

$$p(B|\mathrm{x}) = \frac{p(B)p(\mathrm{x}|B)}{p(A)p(\mathrm{x}|A) + p(B)p(\mathrm{x}|B)} = \frac{1}{1 + e^{\mathrm{w}^{\mathrm{T}}\mathrm{x}+w_0}} \qquad (4.27)$$

- **Clasificador lineal KL por expansión de matriz de covarianza (KLLDC** por sus siglas en inglés**):** Calcula la función discriminante lineal para el conjunto de datos. Esto se hace computando el LDC en los datos proyectados sobre los primeros auto-vectores de la matriz de covarianza de las clases. Se usa (Expansión de Karhunen Loeve) [44].

- **Clasificador discrimínate de Fisher:** Este clasificador está desarrollado con el fin de agrupar problemas biclase. Su función hace uso de planos perpendiculares a la dirección en el plano de características, en alguno caso se comporta como un clasificador de Bayes asumiendo covarianzas iguales. Se expresa matemáticamente en la ecuación (4.28):

$$pR(\mathrm{x}) = (\hat{\boldsymbol{\mu}}_{\mathrm{A}} - \hat{\boldsymbol{\mu}}_{\mathrm{B}})^{\mathrm{T}}\hat{\Sigma}^{-1}\mathbf{x} + const, \qquad (4.28)$$

donde $\mathbf{x} = [\,\mathrm{x}_{1,\ldots,}\,\mathrm{x}_{\mathrm{p}}]$, es el vector de interés, $\hat{\boldsymbol{\mu}}_{\mathrm{A}}, \hat{\boldsymbol{\mu}}_{\mathrm{B}}$ son los vectores de medias de la poblaciones $\mathrm{P}_1, \mathrm{P}_2$, y $\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}$ la matriz de covarianza común para ambas poblaciones [44] [45] [46].

Clasificación basada en la densidad normal:

- **Clasificador normal lineal (LDC** de sus siglas en inglés**):** Este clasificador asume que todas las clases se caracterizan por múltiples distribuciones normales con igual matriz de covarianza $\boldsymbol{S}$. Teniendo en cuenta que el caso de estudio es un problema de dos clases, el clasificador LDC [16], es matemáticamente expresado en las ecuaciones (4.31) y (4.32), así:

$$\mathrm{R}(\mathbf{x}) = [\hat{\boldsymbol{\mu}}_{\mathrm{A}} - \hat{\boldsymbol{\mu}}_{\mathrm{B}}]^{\mathrm{T}}\hat{\Sigma}^{-1}\mathbf{x} + const, \tag{4.31}$$

$$const = -\frac{1}{2}\hat{\boldsymbol{\mu}}_{\mathrm{A}}{}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_{\mathrm{A}} + \frac{1}{2}\hat{\boldsymbol{\mu}}_{\mathrm{B}}{}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_{\mathrm{B}} + \log\left[\frac{p(A)}{p(B)}\right], \tag{4.32}$$

donde $\mathbf{x} = [\,\mathrm{x}_{1,\ldots,}\,\mathrm{x}_{\mathrm{p}}]$, es el vector de interés, $\hat{\boldsymbol{\mu}}_{\mathrm{A}}, \hat{\boldsymbol{\mu}}_{\mathrm{B}}$ son los vectores de medias de la poblaciones $\mathrm{P}_1, \mathrm{P}_2$, y $\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}$ la matriz de covarianza común para ambas $\mathrm{P(A)}, \mathrm{P(B)}$ son las probabilidades

- **Clasificador normal cuadrático (QDC** de sus siglas en inglés**)**: Este clasificador asume que las clases tienen múltiples distribuciones normales, pero cada una es caracterizada por una matriz de covarianza diferente. Para el caso de estudio con un problema biclase, el clasificador QDC, se expresa matemáticamente en la ecuación (4.33) y su contante en la ecuación (4.34).

$$R(\mathrm{x}) = -\frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathrm{A}})^{\mathrm{T}}\hat{\Sigma}_{A}^{-1}(\mathrm{x} - \hat{\boldsymbol{\mu}}_{\mathrm{A}}) + \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathrm{B}})^{\mathrm{T}}\hat{\Sigma}_{\mathrm{B}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathrm{B}}) + c, \tag{4.33}$$

$$c = \log\left[\frac{p(A)}{p(B)}\right] + \frac{1}{2}\log\left[\frac{|\hat{\Sigma}_{\mathrm{B}}|}{|\hat{\Sigma}_{\mathrm{A}}|}\right], \tag{4.34}$$

donde $\mathbf{x} = [\,\mathrm{x}_{1,\ldots,}\,\mathrm{x}_{\mathrm{p}}]$, es el vector de interés, $\hat{\boldsymbol{\mu}}_{\mathrm{A}}, \hat{\boldsymbol{\mu}}_{\mathrm{B}}$ son los vectores de medias de la poblaciones $\mathrm{P}_1, \mathrm{P}_2$, y $\hat{\Sigma}_{A}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathrm{A}}), \hat{\Sigma}_{\mathrm{B}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathrm{B}})$ son las matrices de covarianza muéstrales

Clasificadores no lineales:

- **Clasificador de Parzen:** Este clasificador nace del concepto básico de aprendizaje supervisado no paramétrico basado en la estimación de

densidades de probabilidad de Parzen. Dependiendo de la implementación se utilizan los mismos o diferentes Kernels para las clases [46]. El objetivo de este clasificador es obtener estimaciones de densidades de probabilidad condicional $\mathrm{p}(\mathrm{z}|\mathrm{w_k})$, el espacio de medida es patrocinado en un número finito de regiones disyuntas $\mathrm{R_i}$ llamadas cajas y se cuentan las muestras que caen en ellas, siendo la estimación de la densidad de probabilidad dentro de la caja proporcional a tal número, Además $\mathrm{N_{k,i}}$ denota el número de muestras con clase $\mathrm{W_K}$, en [16] matemáticamente se expresa en la ecuación (4.35):

$$\hat{\mathrm{p}}(\mathrm{z}|\mathrm{w_k}) = \frac{\mathrm{N_{k,j}}}{\text{volumen } (\mathrm{R_i}).\,\mathrm{N_k}}. \qquad (4.35)$$

- **Máquina de soporte vectorial (SVC** de sus siglas en inglés**)**: Este se basa en una sólida base teórica llamada minimización del riesgo empírico (ERM), cuyo objetivo es maximizar las distancias de los objetos de entrenamiento al clasificador [46]. Gracias a la función Kernel que implementa, busca un hiperplano que separe los puntos de una clase con otra de la forma más óptima, teniendo en cuenta la característica fundamental, es decir este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de el mismo, siendo previamente proyectado a un espacio de dimensión superior [16] [47]. Dado por su naturaleza, el clasificador SVC es muy funcional en clasificaciones biclase.

4.6.2. Clasificación no supervisada

En el análisis no supervisado o clustering, se cuenta con objetos que tienen un conjunto de atributos de las que no se sabe a qué clase o categoría pertenecen, es decir, que no tenemos una información a priori que relaciona la información que se encuentra presente en ellos, por eso la finalidad de los métodos no supervisados es clasificar o agrupar este tipo de objetos, para ello es necesario un área de entrenamiento disponible. Un método representativo es el método de K-medias.

Clasificador K-medias *(k-means):* Es un método de agrupamiento, que tiene como objeto la separación de N objetos en k grupos, y al final se cumple que los elementos N de cada conjunto k sean similares entre ellos, y en el caso contrario de no ser similares, los elementos k pertenecerán a otro conjunto.

Para hacer el desarrollo de este método de clasificación se debe tener encuenta unos aspectos como lo son:

- Cada grupo debe tener al menos un objeto.
- Cada objeto solo debe pertenecer a un solo grupo.
- La iniciación de centroides que permiten el agrupamiento de los N objetos deben ser puntos aleatorios que no coinciden con una secuencia.

La formulación matemática que expresa el método de K-medias se expresa en la ecuación (4.36).

$$\boldsymbol{E} = \sum_{centros} \sum_{i=1}^{k} |\boldsymbol{p} - \boldsymbol{m_i}|^2, \tag{4.36}$$

donde $\boldsymbol{m_i}$ es la media de los puntos, y $\boldsymbol{p}$ los conjuntos representados por los centroides.

5. MARCO EXPERIMENTAL

En esta se sección de realiza la descripción de las medidas de desempeño y de las pruebas aplicadas a la base de datos (sección 4.1).

5.1. MEDIDAS DE CLASIFICACIÓN

5.1.1. Medidas para clasificación supervisada

Para evaluar la metodología propuesta en esta investigación se tiene en cuenta además de las medianas del error, medidas supervisadas, tales como: sensibilidad (Se), especificidad (Es), porcentaje de clasificación (PC), expresadas en las ecuaciones (5.1), (5.2) y (5.3) respectivamente, índice de Rand (RI), índice ajustado de Rand (ARI), y finalmente, información mutua normalizada (NMI) [51],

$$S_e = \frac{VP}{VP + FN} \tag{5.1}$$

$$E_S = \frac{VN}{VN + FP} \tag{5.2}$$

$$PC = \frac{VN + VP}{VN + VP + FN + FP} \tag{5.3}$$

dado que:

VP: verdaderos positivos o clase pre-término clasificada correctamente.

VN: verdaderos negativos o clase término clasificada correctamente.

FP: falsos positivos o clase pre-término clasificada como término.

FN: falsos negativos o clase término clasificada como pre-término.

INDICE DE RAND (RI): Sea la matriz de características $\boldsymbol{X} = \{x_1, x_2, \dots x_N\}$ de N elementos, además presenta dos particiones $\boldsymbol{V} = \{v_1, \dots v_R\}$ y $\boldsymbol{U} = \{u_1, \dots u_R\}$, tal que, $\cup_{i=1}^{R}\ v_i = X = \cup_{j=1}^{C}\ u_j$ y $v_i \cap v_i' = \emptyset = u_j \cap u_j'$para $1 \leq i \neq i' \leq R$ y $1 \leq j \neq j' \leq C$. Suponiendo que $\boldsymbol{V}$ sea el conjunto de etiquetas reales y $\boldsymbol{U}$ las etiquetas que arroja la clasificación, entonces a es el número de pares de objetos que se ubican en la misma clase de $\boldsymbol{V}$ en el mismo grupo $\boldsymbol{U}$, b el número de pares de objetos ubicados en la misma clase $\boldsymbol{V}$, pero no diferente grupo $\boldsymbol{U}$, c es el número de pares de objetos diferentes a la clase $\boldsymbol{V}$ pero en el mismo grupo $\boldsymbol{U}$, y d es el número de pares de objetos en diferentes clases y diferentes grupos para ambas particiones [51]. RI se define en la ecuación (5.4):

$$RI = \left(\frac{a+d}{a+b+c+d}\right) = \frac{a+d}{\left(\frac{N}{2}\right)}, \tag{5.4}$$

Una dificultad que presenta esta medida radica en que no es garantía de un etiquetado aleatorio, para contrarrestar este problema se define el índice de Rand ajustado (ARI de sus siglas en inglés) [51]. Por esta razón, como medida supervisada para esta investigación se tiene en cuenta el índice ajustado de Rand.

INDICE AJUSTADO DE RAND: El índice ajustado propuesto en [52], asume una distribución hipergeométrica generalizada como un modelo aleatorio. En la tabla 3, se presenta una comparación de la partición de las clases, teniendo los grupos correspondientes a la clase real (v_1, v_2) y los grupos obtenidos tras la metodología que arroja cada clasificador (u_1, u_2),considerando el modelo de esta investigación [51].

Tabla 3. Matriz de confusión para los dos grupos teniendo en cuenta las etiquetas reales en contraste con las resultantes.

REAL-RESULTANTE	$\boldsymbol{u_1}$	$\boldsymbol{u_2}$	**Suma**
$\boldsymbol{v_1}$ $\boldsymbol{v_2}$	n_{11} n_{21}	n_{12} n_{22}	$n_{1.}$ $n_{2.}$
Suma	$n_{.1}$	$n_{.2}$	$n_{..} = N$

Nota: **la información presentada corresponde a una matriz de confusión, donde los valores atribuidos a v_1 y v_2, corresponden a las clases reales (término y pre-término respectivamente), de igual manera, los valores atribuidos a u_1 y u_2, correspondientes a las clases resultantes (teniendo en cuenta la metodología aplicada).**

Teniendo en cuenta la definición de índice ajustado [51] se obtiene las expresiones matemáticas dadas por las ecuaciones (5.5) y (5.6), asi:

$$INDICE\ AJUSTADO = \frac{INDICE - INDICE\ ESPERADO}{MÁX\ INDICE - INDICE\ ESPERADO}. \tag{5.5}$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right]}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \frac{\left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right]}{\binom{N}{2}}}, \tag{5.6}$$

INFORMACIÓN MUTUA NORMALIZADA (NMI): según [51] [53], NMI está construida a partir de conceptos de teoría de la información, donde $I(U,V)$ denota la información mutua entre U y V, y $H(U)$ la entropía de U dadas por las ecuaciones (5.7), (5.8) y (5.9) respectivamente.

$$H(U) = -\sum_{i=1}^{C} \frac{n_{.i}}{N}\, log\, \frac{n_{.i}}{N}, \tag{5.7}$$

$$I(U,V) = -\sum_{i=1}^{C}\sum_{j=1}^{R} \frac{nij}{N} \log\left(\frac{\frac{n_{ij}}{N}}{\frac{n_{.i}\, n_{.j}}{N^2}}\right), \tag{5.8}$$

$$NMI = \frac{I(U,V)}{\sqrt{H(U)H(V)}}. \tag{5.9}$$

5.1.2. Medidas para clasificación no supervisada

CRITERIO DE FISHER, J: Se expresa matemáticamente en la ecuación (5.10)

$$J = \frac{tr\,(\sum_{k=1}^{K}(\overline{\boldsymbol{x}_k} - \overline{\boldsymbol{X}})^{T}(\overline{\boldsymbol{x}_k} - \overline{\boldsymbol{X}}))}{tr\,(\sum_{k=1}^{K}\boldsymbol{S}_k)}, J \in \boldsymbol{R}^{+}, \tag{5.10}$$

donde K es el total de número de grupos, $\overline{\boldsymbol{x}_k}$ es la mediana de k-ésima clase, $\overline{\boldsymbol{X}}$ es la media de $\boldsymbol{X}$ y $\boldsymbol{S}_k$ es la matriz de covarianza asociada a k [51].

SILHOUETTE, S: Se expresa matemáticamente en la ecuación (5.11).

$$S = \frac{1}{N}\sum_{i=1}^{N} \frac{\min_{v_i}\{\boldsymbol{b}_i - a_i\}}{\sup\{a_i, \min\{\boldsymbol{b_i}\}\}}, \varepsilon_s \epsilon [-1,1], \tag{5.11}$$

donde a_i es el promedio de la distancia del i-ésimo punto para otros puntos sin k-ésima clase y el vector $\boldsymbol{b}_i = [b_1^i, \dots, b_K^i]$ [51].

PUREZA (Pu): Aunque es una medida externa (supervisada), se puede utilizar para clasificación no supervisada, con la ayuda de un vector de pertenencia (etiquetas) [54], se define:

$$\text{Pu } (\boldsymbol{\omega}, \mathbf{c}) = \frac{1}{N}\sum_k \max_j \left|w_K \cap c_j\right|, \tag{5.12}$$

donde $\boldsymbol{\omega} = \{w_1, w_2, \dots w_K\}$ son los conjuntos de las agrupaciones y $\mathbf{c} = \{c_1, c_2, \dots c_K\}$ son los conjuntos de las clases.

5.2. PRUEBAS APLICADAS A LA BASE DE DATOS

Las pruebas aplicadas a la base de datos se expresan a continuación. La selección de características por medio de búsqueda exhaustiva se aplicó a las matrices resultantes de mejor primero y de decisión de árboles, luego se aplicó a estas la técnica de selección de reducción de dimensión para, finalmente, realizar la clasificación supervisada y no supervisada.

- **Prueba 1** (P1)**:** Selección de características mejor primero, método de reducción de dimensión PCA, aplicada a la clasificación.
- **Prueba 2** (P2)**:** Selección de características mejor primero, método de reducción de dimensión ISOMAP, aplicada a la clasificación.
- **Prueba 3** (P3)**:** Selección de características mejor primero, método de reducción de dimensión LLE, aplicada a la clasificación.

- **Prueba 4** (P4)**:** Selección de características mejor primero, método de reducción de dimensión NPE, aplicada a la clasificación
- **Prueba 5** (P5)**:** Selección de características mejor primero, método de reducción de dimensión LTSA, aplicada a la clasificación
- **Prueba 6** (P6)**:** Selección de características mejor primero, método de reducción de dimensión LLTSA, aplicada a la clasificación
- **Prueba 7** (P7)**:** Selección de características mejor primero, método de reducción de dimensión MDS, aplicada a la clasificación
- **Prueba 8** (P8)**:** Selección de características mejor primero, sin método de reducción de dimensión, aplicada a la clasificación.
- **Prueba 9** (P9)**:** Selección de características basada en árboles de decisiones, método de reducción de dimensión PCA, aplicada a la clasificación.
- **Prueba 10** (P10)**:** Selección de características basada en árboles de decisiones, método de reducción de dimensión ISOMAP, aplicada a la clasificación
- **Prueba 11** (P11)**:** Selección de características basada en árboles de decisiones, método de reducción de dimensión LLE, aplicada a la clasificación
- **Prueba 12** (P12)**:** Selección de características basada en árboles de decisiones, método de reducción de dimensión NPE, aplicada a la clasificación
- **Prueba 13** (P13)**:** Selección de características basada en árboles de decisiones, método de reducción de dimensión LTSA, aplicada a la clasificación.
- **Prueba 14** (P14)**:** Selección de características basada en árboles de decisiones, método de reducción de dimensión LLTSA, aplicada a la clasificación.
- **Prueba 15** (P15)**:** Selección de características basada en árboles de decisiones, método de reducción de dimensión SDM, aplicada a la clasificación.
- **Prueba 16** (P16)**:** Selección de características basada en árboles de decisiones, sin método de reducción de dimensión, aplicada a la clasificación.

6. RESULTADOS

En este capítulo se describen los resultados obtenidos con la metodología y el marco experimental aplicado. Con el fin presentar una comparación que involucre las técnicas de selección de características, reducción de dimensión y clasificación, anteriormente descritas.

6.1. RESULTADOS DE LAS PRUEBAS APLICADAS

En esta sección se presentan los resultados obtenidos al aplicar a cada prueba un clasificador, con el fin de obtener medidas de comparación, tanto en las pruebas como en el desempeño de las técnicas de clasificación.

Clasificadores evaluados según las pruebas:

- C1: Clasificador lineal por expansión de PCA en los datos conjuntos (PCLDC)
- C2: Clasificador lineal logísticos (LOGLC)
- C3: Clasificador lineal KL por expansión de matriz de covarianza (KLLDC)
- C4: Clasificador de Fisher (FISHERC)
- C5: Clasificador lineal (LDC)
- C6: Clasificador cuadrático (QDC)
- C7: Clasificador de Parzen (PARZENC)
- C8: Máquina de soporte vectorial (SVC)
- C9: Clasificador K-MEDIAS

6.1.1. Resultados a partir de la clasificación supervisada

En esta subsección se presenta los resultados a partir de la clasificación supervisada, teniendo como referente la mediana del error, junto con su desviación estándar, evaluando el clasificador con su correspondiente prueba, además se presenta, medidas de desempeño.

Tabla 4. Porcentaje de las medianas del error de las pruebas aplicadas a los clasificadores.
(Parte I)

	C1	C2	C3	C4
P1	31.25±9.58	31.25±9.86	31.25±9.58	31.25±9.58
P2	31.25±11.43	31.25±11.63	31.25±11.43	31.25±11.43
P3	37.5±10.29	37.5±9.79	37.5±10.29	37.5±10.29
P4	31.25±11.13	31.25±10.82	31.25±11.13	31.25±11.13

P5	43.75±11.46	43.75±10.46	43.75±11.46	43.75±11.28
P6	31.25±11.32	31.25±11.44	31.25±11.32	31.25±11.32
P7	34.38±9.53	37.5±9.5	34.38±9.53	34.38±9.53
P8	37.5±11.44	37.5±11.71	37.5±11.44	37.5±11.44
P9	**18.75±10.77**	**18.75±10.87**	**18.75±10.77**	**18.75±10.77**
P10	37.5±9.98	37.5±10.02	37.5±9.98	37.5±9.98
P11	**18.75±8.91**	**18.75±9.13**	**18.75±8.91**	**18.75±8.91**
P12	**18.75±8.87**	**18.75±8.84**	**18.75±8.87**	**18.75±8.87**
P13	31.25±9.97	31.25±13.11	31.25±9.97	25±9.7
P14	**18.75±9.13**	21.88±9.75	**18.75±9.13**	**18.75±9.13**
P15	25±10.48	31.25±10.48	25±9.87	25±10.48
P16	25±10.1	25±10.58	25±10.1	25±10.1

(Tabla 5, parte II)

	C5	**C6**	**C7**	**C8**
P1	31.25±9.58	37.5±10.98	31.25±9.09	31.25±8.87
P2	31.25±11.43	31.25±10.97	31.25±11.7	31.25±10.89
P3	37.5±10.29	37.5±10.77	31.25±9.85	37.5±10.36
P4	31.25±11.13	31.25±11.43	31.25±10.84	31.25±10.44
P5	43.75±11.46	37.5±12.18	43.75±10.73	46.88±11.32
P6	31.25±11.32	37.5±11.08	31.25±11.32	31.25±10.88
P7	34.38±9.53	37.5±10.68	31.25±10.14	31.25±10.23
P8	37.5±11.44	43.75±11.94	31.25±9.84	31.25±10.9
P9	**18.75±10.**77	25±9.64	25±10.26	25±9.39
P10	37.5±9.98	37.5±9.76	43.75±9.89	43.75±10.52
P11	**18.75±8.91**	25±10.24	**18.75±8.37**	**18.75±9.51**
P12	**18.75±8.87**	25±6.89	25±8.05	21.88±8.5
P13	31.25±9.97	37.5±9.96	31.25±10.23	31.25±9.3
P14	**18.75±9.1**	25±9.8	25±10.46	**18.75±9.5**
P15	25±10.48	25±10.48	31.25±10.13	25±10.51
P16	25±10.1	25±10.42	31.25±10.47	25±9.1

Nota: En la Tabla 5, se registra los resultados de un promedio de las medianas del error obtenidos al aplicar las 16 pruebas a los diferentes clasificadores. (Para una mejor visualización de los datos, la tabla se ha divido en dos partes) **Fuente:** esta investigación.

La información que se obtiene en la tabla 5, muestra que los mejores resultados para la clasificación involucran las pruebas que dependen de técnicas de selección de características de subconjuntos, con respecto a la técnica de selección de características Ranking.

Siguiendo este orden de ideas, se observa que la prueba 11 (P11), puede ser trabajada con casi todos los clasificadores, salvo por el clasificador C6, correspondiente a QDC.

Tabla 5. Porcentaje de la sensibilidad de los clasificadores para cada prueba realizada.

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**
P1	50	50	50	50	50	44.23	50	**50.94**
P2	45.28	45.28	45.28	45.28	45.28	47.37	**50**	44.44
P3	47.27	47.27	47.27	47.27	47.27	44.64	47.27	47.27
P4	49.09	49.09	49.09	49.09	49.09	44.44	42.86	48.15
P5	44.44	48.15	44.44	43.4	44.44	43.4	47.69	52
P6	45.45	45.45	45.45	45.45	45.45	**49.06**	47.62	45.28
P7	46.3	46.3	46.3	46.3	46.3	43.64	47.62	44.23
P8	49.06	49.06	49.06	49.06	49.06	43.86	48.39	**50**
P9	50.79	50.79	50.79	50.79	50.79	47.54	**52.17**	49.18
P10	54.35	54.35	54.35	54.35	54.35	50	**64.71**	60.42
P11	51.61	50.82	51.61	51.61	51.61	50	53.73	**58.62**
P12	50	50	50	50	50	46.67	**51.47**	47.37
P13	51.61	50	51.61	50.79	51.61	56.67	54.69	**64.15**
P14	49.18	50	49.18	49.18	49.18	47.46	**51.43**	47.46
P15	50	49.12	50	50	50	51.61	47.83	49.15
P16	51.72	51.67	51.72	51.72	51.72	54.55	**55.07**	52.54

Nota: En la Tabla 5, se registra el porcentaje de la sensibilidad correspondiente a las pruebas aplicadas a los diferentes clasificadores. **Fuente:** esta investigación.

Los valores de sensibilidad presentes en la Tabla 5, indican una sensibilidad baja para la información resultante del proceso de clasificación. Esto se presenta debido a la naturaleza de las señales tratadas en esta investigación.

Sin embargo, se resalta valores correspondientes a la técnica de selección de características por medio de subconjuntos, particularmente en la prueba 10, en la cual se utilizó como método de reducción de dimensión ISOMAP y como clasificador el de Parzen, con sensibilidad de 64.71%, siendo el valor más alto en comparación con las pruebas realizadas.

Tabla 6. Porcentaje de la especificidad de las pruebas aplicadas a los clasificadores.

	C1	C2	C3	C4	C5	C6	C7	C8
P1	70.27	70.27	70.27	70.27	70.27	67.44	**91.89**	72.22
P2	69.05	69.05	69.05	69.05	69.05	75.00	**78.38**	69.77
P3	72.50	72.50	72.50	72.50	72.50	72.09	**72.50**	72.50
P4	73.68	73.68	73.68	73.68	73.68	69.77	71.11	71.79
P5	69.77	71.79	69.77	68.18	69.77	68.18	**85.00**	68.57
P6	71.43	71.43	71.43	71.43	71.43	71.05	**82.50**	69.05
P7	70.73	70.73	70.73	70.73	70.73	70.45	**82.50**	67.44
P8	71.05	71.05	71.05	71.05	71.05	72.73	**82.05**	70.27
P9	83.78	83.78	83.78	83.78	83.78	78.05	**94.29**	79.49
P10	61.76	61.76	61.76	61.76	61.76	65.79	**78.26**	67.86
P11	83.33	81.08	83.33	83.33	83.33	84.21	**93.94**	85.71
P12	76.32	76.32	76.32	76.32	76.32	76.19	**91.67**	73.17
P13	83.33	81.58	83.33	83.78	83.33	86.67	**90.63**	82.61
P14	79.49	81.58	79.49	79.49	79.49	75.61	**94.44**	75.61
P15	75.68	76.32	75.68	75.68	75.68	85.71	**90.00**	78.95
P16	77.78	80.56	77.78	77.78	77.78	93.75	**96.96**	80.00

Nota: En la Tabla 6, se registra el porcentaje de la especificidad correspondiente a las pruebas aplicadas a los diferentes clasificadores. **Fuente:** esta investigación.

La información obtenida en la Tabla 6, con respecto a la especificidad, indica valores altos en las pruebas realizadas para el clasificador de Parzen al momento de tratar señales de electrohisterografía. Como es evidente en la Tabla 6, para la prueba 16 (P16) aplicada al clasificador se obtiene una especificidad del 96.96%, siendo la mayor en el estudio comparativo.

Tabla 7. Porcentaje de clasificación de las pruebas aplicadas a los clasificadores

	C1	C2	C3	C4	C5	C6	C7	C8
P1	68.42	68.42	68.42	68.42	68.42	68.42	**89.47**	69.74
P2	69.74	69.74	69.74	69.74	69.74	75.00	**76.32**	71.05
P3	72.37	72.37	72.37	72.37	72.37	73.68	72.37	72.37
P4	72.37	72.37	72.37	72.37	72.37	71.05	**73.68**	71.05
P5	71.05	71.05	71.05	69.74	71.05	69.74	**85.53**	65.79
P6	72.37	72.37	72.37	72.37	72.37	69.74	**82.89**	69.74
P7	71.05	71.05	71.05	71.05	71.05	72.37	**82.89**	68.42
P8	69.74	69.74	69.74	69.74	69.74	75.00	**81.58**	68.42
P9	82.89	82.89	82.89	82.89	82.89	80.26	**90.79**	80.26
P10	60.53	60.53	60.53	60.53	60.53	65.79	**67.11**	63.16
P11	81.58	80.26	81.58	81.58	81.58	84.21	**88.16**	76.32

P12	76.32	76.32	76.32	76.32	76.32	78.95	**89.47**	75.00
P13	81.58	81.58	81.58	82.89	81.58	78.95	**84.21**	69.74
P14	80.26	81.58	80.26	80.26	80.26	77.63	**92.11**	77.63
P15	73.68	75.00	73.68	73.68	73.68	81.58	**90.79**	77.63
P16	76.32	78.95	76.32	76.32	76.32	86.84	**90.79**	77.63

Nota: En la Tabla 7, se registra el porcentaje de clasificación correspondiente a las pruebas aplicadas a los diferentes clasificadores. **Fuente:** Esta investigación.

En la Tabla 7, se obtiene altos valores en porcentaje de clasificación. De la información obtenida, es evidente que la clasificación realizada con la técnica de Parzen es de gran interés. Además, se tiene presente que esta clasificación es la mejor para cada prueba realizada.

Asimismo, se puede observar un porcentaje de clasificación que corresponde al 92.11 %, en P14, siendo este valor el más alto en el conjunto de datos presentados.

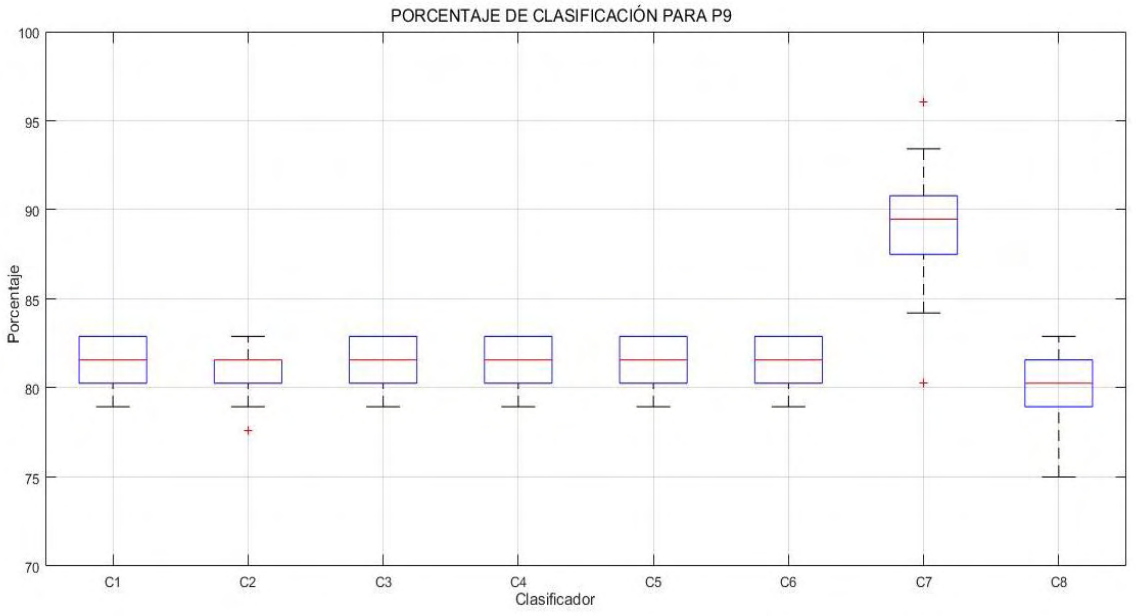


Figura 10. Diagrama de cajas y bigotes del porcentaje de clasificación. **Fuente:** Esta investigación.

En la Figura 10, se observa un diagrama de cajas y bigotes del porcentaje de clasificación obtenido de los 8 clasificadores supervisados aplicados a P9, en esta figura es evidente el desempeño del C7, clasificador de Parzen, obtuvo el mayor valor.

Tabla 8. Índice Ajustado de Rand (ARI) de las pruebas aplicadas a los clasificadores.

	C1	C2	C3	C4	C5	C6	C7	C8
P1	0.14442	0.14442	0.14442	0.14442	0.14442	**0.39078**	**0.39078**	0.16624
P2	0.14445	0.14445	0.14445	0.14445	0.14445	**0.21391**	0.18945	0.10524
P3	0.04348	0.04348	0.04348	0.04348	0.04348	0.04348	0.03142	0.04420
P4	0.14442	0.14442	0.14442	0.14442	0.14442	0.21397	**0.29665**	0.07172
P5	0.07146	0.07146	0.07146	0.07146	0.07146	**0.42535**	0.39104	0.07195
P6	0.16618	0.16618	0.16618	0.16618	0.16618	0.16618	**0.18937**	0.12407
P7	0.12407	0.16618	0.12407	0.12407	0.12407	**0.26730**	0.21397	0.14442
P8	0.10514	0.10511	0.10514	0.10514	0.10514	0.23998	**0.70522**	0.16618
P9	0.32626	0.35781	0.32626	0.32626	0.32626	0.39078	**0.53680**	0.26730
P10	0.04365	0.04365	0.04365	0.04365	0.04365	0.16624	0.12581	0.05809
P11	-0.01062	-0.01062	-0.01062	-0.01062	-0.01062	-0.00491	-0.00119	**0.00461**
P12	0.18937	0.18934	0.18937	0.18937	0.18937	0.39078	**0.75084**	0.26730
P13	0.23998	0.18937	0.23998	0.29605	0.23998	0.21508	**0.75087**	0.10827
P14	-0.00227	-0.00227	-0.00227	-0.00227	-0.00227	0.12413	**0.21408**	0.01177
P15	0.26725	0.21397	0.26725	0.26725	0.26725	0.26725	**0.61818**	0.23998
P16	0.26730	0.23987	0.26730	0.26730	0.26730	0.46108	**0.75091**	0.23998

En la Tabla 8, se observa como medida de valoración el Índice Ajustado de Rand (ARI), con valores de interés asociados a la selección de características por medio de subconjuntos. Además, la clasificación realizada por C7, correspondiente a Parzen y particularmente en P12, P13 Y P16 tiene un valor de 0.75, en escala de 0 a 1 en términos de ARI.

Tabla 9. Información mutua normalizada (MNI) de las pruebas aplicadas a los clasificadores.

	C1	C2	C3	C4	C5	C6	C7	C8
P1	0.11564	0.1156	0.11564	0.11564	0.11564	0.31268	0.31268	0.1318
P2	0.11519	0.1151	0.11519	0.11519	0.11519	0.16951	0.15015	0.0861
P3	0.04257	0.0425	0.04257	0.04257	0.04257	0.04257	0.03250	0.0481
P4	0.11564	0.1156	0.11564	0.11564	0.11564	0.16856	0.25593	0.0656
P5	0.06099	0.0609	0.06099	0.06099	0.06099	0.35073	0.32482	0.0628
P6	0.13173	0.1317	0.13173	0.13173	0.13173	0.13173	0.14931	0.1000
P7	0.10004	0.1327	0.10004	0.10004	0.10004	0.21051	0.16856	0.1156
P8	0.08582	0.0862	0.08582	0.08582	0.08582	0.18971	0.60138	0.1327
P9	0.25807	0.2832	0.25807	0.25807	0.25807	0.31057	0.44047	0.2105
P10	0.04083	0.0408	0.04083	0.04083	0.04083	0.13183	0.12165	0.0561
P11	0.00218	0.0021	0.00218	0.00218	0.00218	0.00434	0.00637	0.0122

P12	0.14931	0.1497	0.14931	0.14931	0.14931	0.31057	0.65170	0.2105
P13	0.18971	0.1493	0.18971	0.23296	0.18971	0.19737	0.66596	0.1291
P14	0.00792	0.0079	0.00792	0.00792	0.00792	0.10001	0.17037	0.0179
P15	0.21147	0.1722	0.21147	0.21147	0.21147	0.21001	0.51860	0.1951
P16	0.21051	0.1889	0.21051	0.21051	0.21051	0.37845	**0.71040**	0.1897

En la Tabla 9, se presenta la información correspondiente a la valoración de la información normalizada mutua (NMI), en este caso, se obtiene como valor de interés 0.71, en una escala de 0 a 1, correspondiente al clasificador de Parzen aplicado a P16, con la técnica de selección de características de subconjuntos y método de reducción de dimensión LLTSA.

6.1.2. Resultados a partir de la clasificación no supervisada

Tabla 10. Medidas de comparación no supervisada aplicada a K-medias.

	FISHER	SILHOUETTE	PUREZA
P1	**0.25378823**	0.45±0.20	0.65789474
P2	0.23474363	0.43±0.22	0.69736842
P3	0.04498641	0.14±0.16	0.60526316
P4	0.0604631	0.16±0.18	0.52631579
P5	0.03819412	0.16±0.18	0.69736842
P6	0.24025032	0.44±0.21	0.67105263
P7	**0.25378823**	**0.45±0.20**	0.65789474
P8	0.24009834	**0.45±0.20**	0.65789474
P9	0.14055664	0.32±0.17	0.57894737
P10	0.13361307	0.29±0.21	0.61842105
P11	0.05178112	0.26±0.19	0.81578947
P12	0.09802456	0.18±0.17	0.78947368
P13	0.06419549	0.41±0.22	**0.96052632**
P14	0.14055664	0.32±0.17	0.57894737
P15	0.14055664	0.32±0.17	0.57894737
P16	0.14055664	0.32±0.17	0.57894737

La información que aporta la Tabla 10, evidencia que no es pertinente realizar una clasificación de tipo no supervisada para señales de tipo EHG, la información presentada contiene los valores de medidas no supervisadas tales como el Índice de Fisher, con máximos en las pruebas P1 y P7, con 25.4%, 45% para SILHOUETTE en las pruebas 7 y 8 y finalmente Pureza de 96% en la prueba P13. Dado la variedad de los datos, este tipo de clasificación no es concluyente.

En conclusión, a partir de estudio, se comprueba que un sistema provisional (no definitivo) de *machine learning* adecuado para clasificar señales de EHG debe constar de:

- Pre-proceso a partir de la limitación de las señales a un lapso establecido, con una normalización de los valores de voltaje.
- Selección de características en cascada a partir de la técnica de selección por subconjuntos basadas en el algoritmo ID3, junto con la técnica de búsqueda exhaustiva con el objetivo de volver aún más compacta la matriz de características resultantes
- Reducción por medio de técnicas representativas convexas y no convexas, en particular, análisis del espacio local lineal tangencial (LLTSA).
- Clasificación usando clasificador de Parzen.

6.2. INTERFAZ DE VISUALIZACIÓN DE RESULTADOS

Los resultados obtenidos de este estudio comparativo son visibles a través de una interfaz desarrollada, donde se encuentran las distintas combinaciones de los resultados: Medianas del error, sensibilidad especificidad entre otros.

La Figura 11 indica la simulación del procedimiento llevado a cabo en esta investigación por medio de selección, aplicado a las técnicas de selección de características, métodos de reducción de dimensión y finalmente la clasificación. La interfaz ofrece la posibilidad de obtener resultados mediante medios gráficos y además el registro de los datos cuantitativos en las medidas de valor.

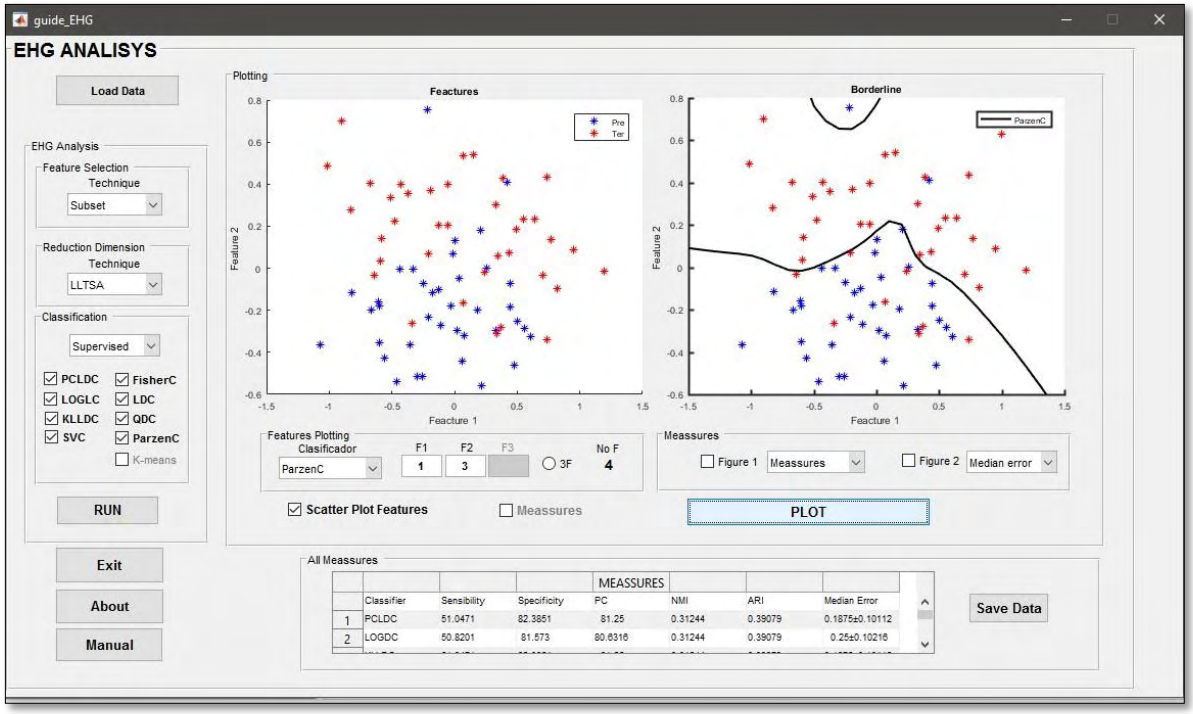


Figura 11. Interfaz de visualización de resultados. En la interfaz se puede visualizar la clasificación de las señales de EHG, en pre-término y término. **Fuente:** esta investigación.

- **Gráficas de las características**

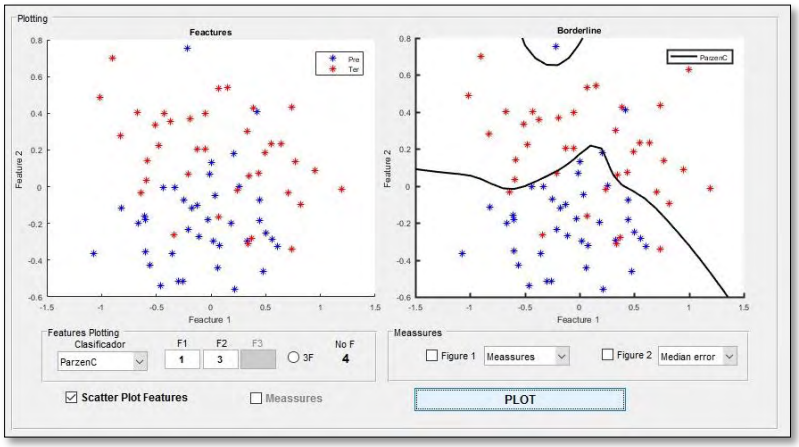


Figura 12. Gráfica de características y trazado de la frontera de decisión. En la figura se visualiza la posición de las clases de estudio respecto a dos características seleccionadas. Además del trazado de la frontera de decisión del clasificador utilizado. **Fuente:** Esta investigación.

En la Figura 12. Se observa la gráfica de las características teniendo en cuenta cada clase, es posible graficar hasta tres características, y en el recuadro subsecuente se observa el trazado de la frontera de decisión del clasificador implementado.

Además, en la interfaz ayuda a visualizar las diferentes medidas de valoración teniendo en cuenta los clasificadores elegidos. En la Figura 13, se observa el porcentaje de clasificación en cada uno de los clasificadores utilizados, siendo C7, el clasificador de Parzen.

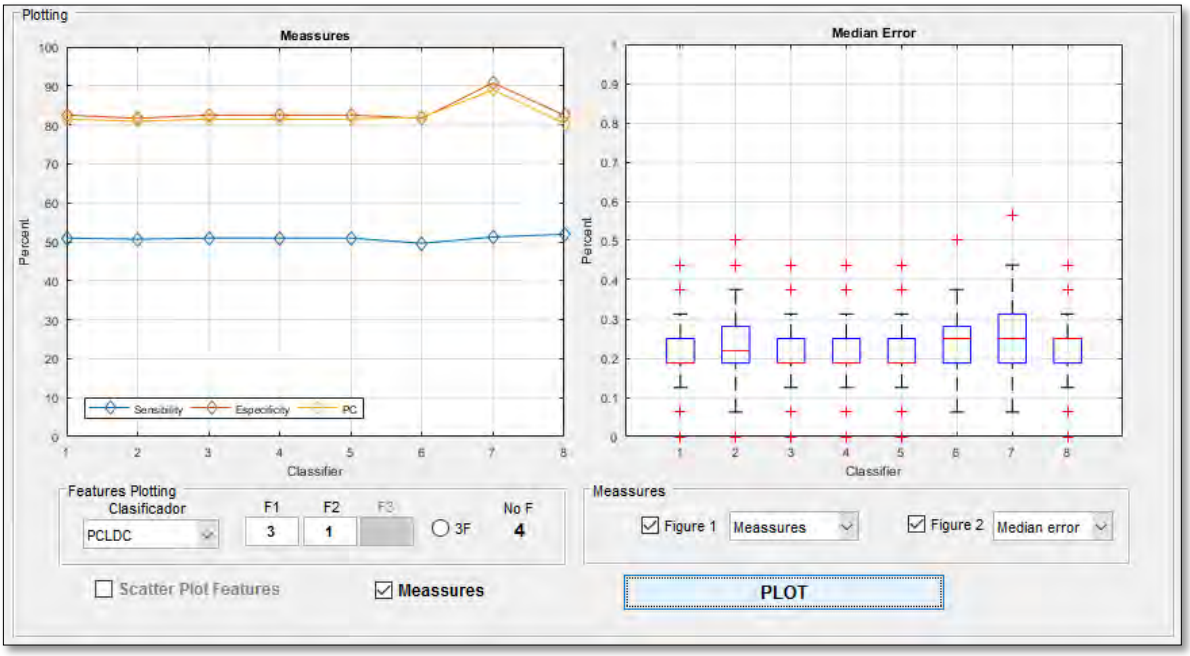


Figura 13. Gráfica de medidas. Especificidad, sensibilidad y porcentaje de clasificación, junto a diagrama de cajas y bigotes de la mediana del error de los clasificadores. **Fuente:** Esta investigación.

En la interfaz se puede realizar el procedimiento del estudio comparativo, llevado a cabo en esta investigación.

Se puede realizar una comparación desde la selección de características, pasando por métodos de reducción de dimensión y finalmente técnicas de clasificación.

Asimismo, según la clasificación que se elija (supervisada o no supervisada), se presenta una tabla con medidas de valoración del proceso de clasificación. Esta tabla puede ser guardada en un registro. En la Figura 14 se puede observar la tabla que presenta las medias de valoración para cada clasificador al momento de ser registradas.

- **Información registrada**

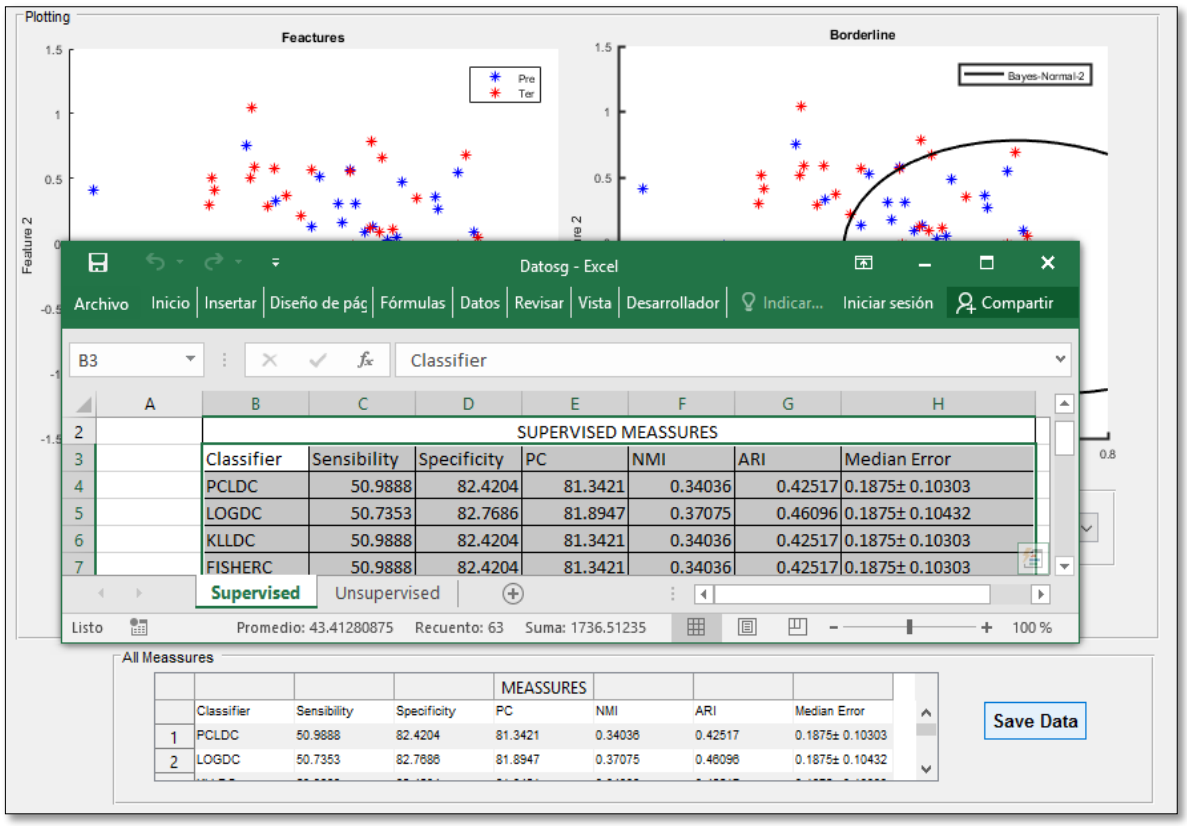


Figura 14. Gráfica de medidas exportadas a Excel. Las medidas de valoración calculadas exportadas a un archivo para futuras aplicaciones. **Fuente:** Esta investigación.

7. CONCLUSIONES Y TRABAJO FUTURO

Los sistemas computarizados juegan un papel relevante en el análisis de un fenómeno y, en el caso particular de datos a nivel biomédico, los cuales permiten automáticamente clasificar y predecir para contribuir a la toma inteligente de decisiones ante situaciones particulares. Asimismo, es motivante y desafiante para jóvenes investigadores, desarrollar soluciones a problemas reales por medio de herramientas computacionales y conocimientos adquiridos en la academia.

La naturaleza de las señales EHG ha representado un reto al momento de realizar el análisis de la información contenida en estos registros. Consecuentemente, la metodología utilizada en este estudio comparativo ha contribuido en la búsqueda de un procedimiento propio para este tipo de señales.

La etapa de selección de características es crucial en la metodología implementada para la clasificación de los datos, asimismo es relevante la técnica de selección que se utilice, es decir, la selección de características basada en subconjuntos favorece a los resultados de la subsecuente tarea de clasificación. Adicionalmente, la reducción de dimensión de los datos favorece al proceso de clasificación, específicamente, contribuye a mejorar hasta un 6% el valor de la mediana del error al momento de ser implementado.

Para el estudio comparativo desarrollado se evalúa técnicas de clasificación supervisada y no supervisada, particularmente, se aprecia que se presenta resultados más favorables para la clasificación supervisada, mientras que en el caso de la clasificación no supervisada los resultados no fueron concluyentes.

En la clasificación supervisada, la técnica del clasificador de Parzen es útil al momento de ser valorado por las medidas implementadas, resaltándose valores promedio de especifidad de 96.96% y porcentaje de clasificación de 92.11% en experimentos de 100 iteraciones.

Como resultado general de esta investigación, se obtiene un sistema de *machine learning* inicial para clasificar señales de EHG que consta de: Pre-proceso a partir de la limitación de las señales a un lapso de tiempo establecido, selección de características en cascada utilizando selección por subconjuntos (*subset*) y búsqueda exhaustiva, reducción de dimensión utilizando el análisis del espacio local tangencial lineal, y clasificación usando clasificador de Parzen.

Este trabajo exploratorio representa una primera aproximación para el diseño de sistemas automáticos de predicción de embarazo pre-término y devela los indicios

y los aspectos clave acerca de los caminos que deberían tomarse para pre-procesar, caracterizar y clasificar señales EHG.

TRABAJO FUTURO

Como trabajo futuro, en vista de que el clasificador de Parzen, que es de tipo probabilístico, fue el que mejor desempeño presentó, se propone estudiar y/o diseñar más clasificadores basados en densidades de probabilidad, con el fin de mejorar el desempeño alcanzado hasta el momento. También, se propone realizar mezcla de clasificadores para aprovechar las características y potencialidades individuales de cada clasificador simultáneamente y alcanzar mejores rendimientos.

Adicionalmente, se espera construirá una base de datos propia realizada con pacientes oriundos de la región de Nariño, como primera fase para realizar un diagnóstico sistematizado de embarazos pre-término. Además de presentar también una interfaz que facilite el uso de la información recolectada, siendo todos estos estudios un aporte de carácter científico y significativo para la región.

REFERENCIAS

[1] Althabe, F., Carroli, G., Lede, R., Belizán, J. M., & Althabe, O. H. (1999). El parto pretérmino: detección de riesgos y tratamientos preventivos.

[2] Roura, L. C. (2006). *Parto prematuro*. Ed. Médica Panamericana.

[3] Ortega Maroto, G. N., & Hinojosa León, Y. A. (2010). *Causas del embarazo en las adolescentes y riesgos de complicaciones en el recién nacido en el área de ginecología y obstetricia del Hospital Provincial General Docente Riobamba en el período enero a julio del 2010* (Bachelor's thesis, Riobamba: Universidad Nacional de Chimborazo).

[4] Fele-Žorž, G., Kavšek, G., Novak-Antolič, Ž., & Jager, F. (2008). A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups. *Medical & biological engineering & computing*, *46*(9), 911-922.

[5] Monteiro, A. V. (2010). *Electrohisterografia dinâmica intra-parto: contribuição para o desenvolvimento de um protótipo* (Doctoral dissertation, Universidade da Beira Interior).

[6] RUBIO, J. A. (2011). Diseño y desarrollo de un sistema para el registro y monitorización de la actividad mioeléctrica uterina.

[7] Trujillo Pulgarín, C. A. *Clasificación basada en la estimación de Parzen en espacios generalizados de disimilitudes= Classification based on the Parzen estimation in generalized dissimilarity spaces* (Doctoral dissertation, Universidad Nacional de Colombia-Sede Manizales).

[8] Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*, *14*, 85-86.

[9] Cristina, L., & Elsa, A. Guía de prácticas clínicas. Amenaza de parto prematuro.

[10] Baghamoradi, S. M. S., Naji, M., & Aryadoost, H. (2011, December). Evaluation of cepstral analysis of EHG signals to prediction of preterm labor. In *Biomedical Engineering (ICBME), 2011 18th Iranian Conference of* (pp. 81-83). IEEE.

[11] Fergus, P., Idowu, I., Hussain, A., & Dobbins, C. (2016). Advanced artificial neural network classification for detecting preterm births using EHG records. *Neurocomputing*, *188*, 42-49.

[12] Paredes, J., Luzardo, E., & Briceño, H. (2005). A Wavelet based method to characterize electrical insulators under partial discharges. *Revista Técnica de la Facultad de Ingeniería. Universidad del Zulia*, *28*(2).

[13] Herrera, F., & Cano, J. R. (2006). Técnicas de reducción de datos en KDD. El uso de Algoritmos Evolutivos para la Selección de Instancias. *Actas del I Seminario Sobre Sistemas Inteligentes (SSI'06), Universidad Rey Juan Carlos, Madrid*, 165-181.

[14] Ruiz, R., Aguilar, J., & Riquelme, J. (2005). Evaluación de rankings de atributos para clasificación. *Universidad de Sevilla, España, Tesis*.

[15] Martínez, G. R. S., & Mejía, J. A. S. (2011). Arboles de decisiones en el diagnóstico de enfermedades cardiovasculares. *Scientia et technica*, *3*(49), 104-109.

[16] Subba Rao, T. (2011). Classification, parameter estimation and state estimation-an engineering approach using MATLAB. *Journal of Time Series Analysis*, *32*(2), 194-194.

[17] Verdenik, I., Pajntar, M., & Leskošek, B. (2001). Uterine electrical activity as predictor of preterm birth in women with preterm contractions. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, *95*(2), 149-153.

[18] Maner, W. L., & Garfield, R. E. (2007). Identification of human term and preterm labor using artificial neural networks on uterine electromyography data. *Annals of biomedical engineering*, *35*(3), 465-473.

[19] Maul, H., Maner, W. L., Olson, G., Saade, G. R., & Garfield, R. E. (2004). Non-invasive transabdominal uterine electromyography correlates with the strength of intrauterine pressure and is predictive of labor and delivery. *The Journal of Maternal-Fetal & Neonatal Medicine*, *15*(5), 297-301.

[20] Devedeux, D., Marque, C., Mansour, S., Germain, G., & Duchêne, J. (1993). Uterine electromyography: a critical review. *American journal of obstetrics and gynecology*, *169*(6), 1636-1653.

[21] Kavšek, G. (2001). *Electromiographic activity of the uterus in threatened preterm delivery* (Doctoral dissertation, Master's Thesis, University of Ljubljana, Medical faculty, Ljubljana).

[22] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet. *Circulation*, *101*(23), e215-e220.

[23] Garcia-Gonzalez, M. T., Charleston-Villalobos, S., Vargas-Garcia, C., Gonzalez-Camarena, R., & Aljama-Corrales, T. (2013, July). Characterization of EHG contractions at term labor by nonlinear analysis. In *Engineering in Medicine*

and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE (pp. 7432-7435). IEEE.

[24] Carre, P., Leman, H., Fernandez, C., & Marque, C. (1998). Denoising of the uterine EHG by an undecimated wavelet transform. *IEEE transactions on biomedical engineering*, *45*(9), 1104-1113.

[25] Ruiz, R., Riquelme, J., & Aguilar-Ruiz, J. (2005). Búsqueda secuencial de subconjuntos de atributos sobre un ranking. *Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA*, 251-260.

[26] Sanchez, R. R. (2006). Heurísticas de selección de atributos para datos de gran dimensionalidad.

[27] Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, *10*, 66-71.

[28] Bengio, Y. (2007). Learning deep architectures for AI (Technical Report 1312). *Université de Montréal, dept. IRO*.

[29] Bengio, Y., & Monperrus, M. (2004, December). Non-Local Manifold Tangent Learning. In *NIPS* (pp. 129-136).

[30] Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323-2326.

[31] Hoi, S. C., Liu, W., Lyu, M. R., & Ma, W. Y. (2006). Learning distance metrics with contextual constraints for image retrieval. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on* (Vol. 2, pp. 2072-2078). IEEE.

[32] Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, *290*(5500), 2319-2323.

[33] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269-271.

[34] Floyd, R. W. (1962). Algorithm 97: shortest path. *Communications of the ACM*, *5*(6), 345.

[35] Balasubramanian, M., & Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, *295*(5552), 7-7.

[36] Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323-2326.

[37] Zhang, Z. Y., & Zha, H. Y. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University (English Edition)*, *8*(4), 406-424.

[38] Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, *100*(5), 401-409.

[39] Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4), 401-419.

[40] Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. CRC press.

[41] Holland, S. M. (2008). Principal components analysis (PCA). *University of Georgia*.

[42] Hardgrave, B. C., Wilson, R. L., & Walstrom, K. A. (1994). Predicting graduate student success: A comparison of neural networks and traditional techniques. *Computers & Operations Research*, *21*(3), 249-263.

[43] Albus, J. E., Anderson, R. H., Brayer, J. M., DeMori, R., Feng, H. Y., Horowitz, S. L., ... & Vamos, T. (2012). *Syntactic pattern recognition, applications* (Vol. 14). Springer Science & Business Media.

[44] Pękalska, E., & Duin, R. P. (2005). *The dissimilarity representation for pattern recognition: foundations and applications* (Vol. 64). World scientific.

[45] Duin, R. P. W. C., Juszczak, P., Paclik, P., Pekalska, E., De Ridder, D., Tax, D., & Verzakov, S. (2010). PRTools 4.1. *A Matlab Toolbox for Pattern Recognition, Software and Documentation downloaded May*.

[46] Estévez Núñez de Prado, I. (2013). Análisis Discriminante: Un estudio de simulación.

[47] Bishop, C. M. (2006). Pattern recognition. *Machine Learning*, *128*, 1-58

[48] Nieto, N., & Rojas, D. M. O. (2008). El uso de la transformada wavelet discreta en la reconstrucción de señales senosoidales. *Scientia et technica*, *1*(38), 381-386.

[49] McDaid, A. F., Greene, D., & Hurley, N. (2011). Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.

[50] Tsuda, K., Kawanabe, M., & Müller, K. R. (2003). Clustering with the Fisher score. *A@ A*, *5*, 6.

[51] Peluffo Ordoñez, D. H. *Agrupamiento espectral de datos dinamicos* (Doctoral dissertation, Universidad Nacional de Colombia-Sede Manizales).

[52] Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193-218.

[53] Romano, S., Vinh, N. X., Bailey, J., & Verspoor, K. (2016). Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research*, *17*(134), 1-32.

[54] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.

[56] Rupérez Cerezo, E. (2012). Wavelets, Universidad Complutense de Madrid.

ANEXOS

Esta sección ha sido destinada a los resultados tangibles logrados con el trabajo realizado en esta investigación. Estos anexos contienen una descripción ampliada de los resultados mencionados en el Capítulo 6.

ANEXO 1. Artículo: International Conference on Information Systems and Computer Science (INSISCOS).

Este anexo contiene el artículo que seleccionado para sustentación en modalidad ponente en el evento INSISCOS desarrollado en Quito del 24 al 26 de noviembre de 2016. El libro de actas se puede encontrar en el siguiente link: http://ingenieria.ute.edu.ec/inciscos/assets/s6/INCISCOS_2016_paper_71.pdf ISBN 978-9978-389-32-4.

Machine learning for the prediction of preterm pregnancy using EHG signals

Angela Stephanya Caipe Gordillo
Departamento de ingeniería electrónica
Universidad de Nariño
Pasto, Colombia
ascaipe@udenar.edu.co

Jorge Armando Muñoz Rosero
Departamento de ingeniería electrónica
Universidad de Nariño
Pasto, Colombia
jorgem@udenar.edu.co

Diego Hernan Peluffo Ordoñez
Departamento de ingeniería electrónica
Universidad Técnica del Norte
Ibarra, Ecuador
dhpeluffo@utn.edu.ec

Abstract- The pre-term pregnancy occurs when labor occurs before 37 weeks of gestation, this fact is a major cause of mortality and morbidity in children, at present. Despite that there are several factors that indicate risk a pre-term delivery, it can be produced without the need of a symptom or indication factor. In the world it is estimated that around 15 million premature babies was born each year, this quantity is growing and also has a greater impact in developing countries. That is why several investigations aimed at solving this problem through a study of records of uterine electrical activity, known as electrohisterography, which represents a great hope when detecting a pre-term pregnancy. By using computerized systems based on techniques of machine learning it is possible to determine the probability of pregnancy preterm as from electrohisterography records, however, there is still no definitive methods to characterize and classify these records This article presents a comparison methodology for the diagnosis of pre-term pregnancy occurs using different supervised Pattern recognition techniques such as feature selection, dimensionality reduction and classification. Considered techniques and reach an average error of 18.75%.

Keywords – Electrohisterography, feature selection, Preterm pregnancy, Supervised classificationIntroduction

En el mundo se estima que alrededor de 15000000 de bebes prematuros nacen al año, una cifra que va en aumento, además, tiene mayor impacto en los países en desarrollo y las consecuencias son más sentidas en las clases socio-económicas desfavorecidas[3][1].

El embarazo pre-término ocurre cuando se presenta labor de parto antes de la semana 37 de gestación, entre las principales consecuencias que conlleva un nacimiento prematuro se encuentra un alto riesgo de mortalidad en los primeros años de vida, además de un mayor grado de morbilidad que incluye problemas respiratorios, dificultades alimenticias, ser más propensos a infecciones graves, problemas cerebrales, dificultades visuales y auditivas. A pesar de que existen diversos factores que indiquen riesgo de un trabajo de parto prematuro, este se puede producir espontáneamente sin haberse presentado ningún síntoma o factor indicativo [1] [2] [3].

Diversos estudios han demostrado el cambio de la actividad eléctrica uterina durante el embarazo, así mismo, existe una relación como factor discriminante potencial entre un parto pre-término y uno a término completo. Siendo el electrohisterograma (EHG) un procedimiento no invasivo que se lleva a cabo para detectar cambios bioeléctricos que se presentan en la actividad muscular uterina en el embrazo, a manera de registros, los cuales son recolectados para la caracterización de la magnitud y la duración de las

[3] http://www.who.int/mediacentre/factsheets/fs363/es/

contracciones uterinas, de este modo permite ver la progresión fisiológica y patofisiológica del trabajo de parto, por lo tanto, realizar un estudio comparativo de registros de la actividad eléctrica uterina, mediante electrohisterografia, representa un elemento clave en la detección de un parto prematuro [4][5][6].

Teniendo en cuenta lo anterior y a pesar de ser un campo de estudio relativamente nuevo, se han presentado diversos estudios de señales de electrohisterografía en los cuales se evalúa las características de estas señales como tal, obteniéndose resultados prometedores, pero no concluyentes hasta el momento.

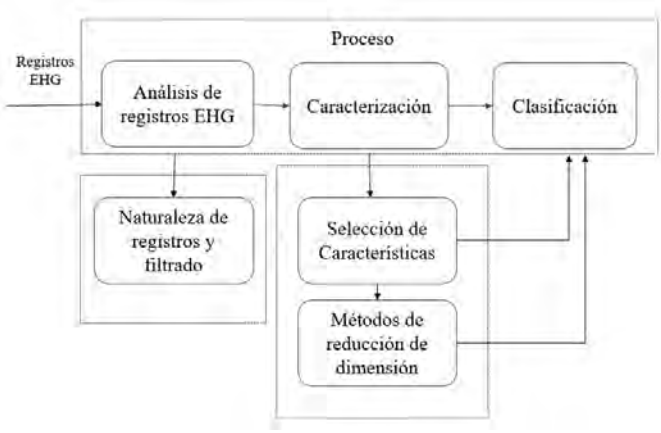
En consecuencia, se analizaron características propias y dicientes para este tipo de señales, así mismo se estudiaron técnicas de selección de características, métodos de reducción de dimensión con el objetivo de optimizar la clasificación y obtener resultados satisfactorios, en el proceso de clasificación sobre espacios de características, fue necesario utilizar técnicas simples de reconocimiento de patrones que los generalicen mejor, al ser menor el número de parámetros que deben ser determinados sobre la base de datos de muestras obtenidas, de esta forma, fueron de especial interés los clasificadores lineales o cuadráticos, fue así que se estudiaron los siguientes clasificadores dentro de este grupo en el espacio de características o basados en reglas de decisión: el clasificador Normal Lineal (LDC)[4] este asume que todas las clases se caracterizan por múltiples distribuciones normales con igual matriz de covarianza, Clasificador Normal Cuadrático (QDC)[2] con múltiples distribuciones normales, pero cada una caracterizada con una matriz de covarianza diferente, clasificadores no lineales como *support vector classifier*, (SVC) que transforma el espacio de características en un espacio con más dimensiones, de manera que las características sean separables, y por último el clasificador de Parzen que obtiene estimaciones de densidades de probabilidad condicional para hacer la clasificación [11][12].

En este trabajo se presentan los primeros resultados de un estudio comparativo el cual consiste en evaluar el desempeño de técnicas de reconocimiento de patrones que presenten un equilibrio entre efectividad, costo computacional y fácil interpretación del concepto fisiológico al momento de realizar la clasificación de registros EHG, en diagnósticos de riesgo de embarazo pre-termino, siendo hasta el momento el mejor resultado de la mediana del error 18.75%.

Este articulo está organizado de la siguiente forma: presentación de registros EHG, así como su caracterización en la Sección II, fase experimental en la Sección III, resultados obtenidos en la Sección IV, conclusiones y trabajo futuro en la Sección V.

En Fig.1. Se presenta un diagrama de bloques que representa el orden del proceso llevado a cabo.

Fig. 1 Diagrama de bloques General



I. Registros EHG

A. Base de datos

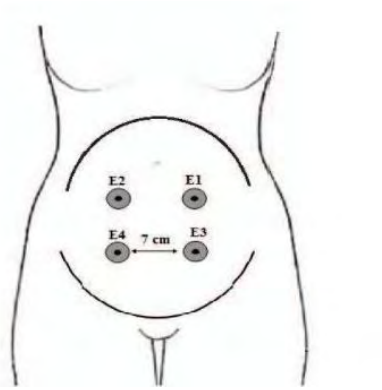
En el banco de señales de PHYSIONET se encuentra la base de datos TPDB EHG, recolectada entre 1997 hasta 2006 por el Departamento de Obstetricia y Ginecología, del Centro Médico de Ljubljana, Ljubljana[5]. Para este estudio se utilizaron 38 muestras de cada clase (termino y pre-termino), la duración de cada registro es aproximadamente 30 minutos, cada señal ha sido digitalizada a 20 muestras por segundo, para cada canal con 16 bits de resolución a un rango de ± 2.5 milivoltios [4]. Cada registro cuenta con la información obtenida por tres canales producidos por una diferencia de potencial en cuatro electrodos

[4] http://bibing.us.es/proyectos/abreproy/70448/fichero/05_Capitulo4.pdf

[5] https://physionet.org/physiobank/database/tpehgdb/

superficiales, la disposición de los electrodos se puede observar en la Fig. 2.

Fig. 2 Posición de los electrodos



Teniendo en cuenta la Fig.2. El primer electrodo (E1) se ubica 3.5 cm debajo y hacia la izquierda del ombligo, el segundo electrodo (E2) está ubicado 3.5 cm por encima y a la izquierda del ombligo, el tercer electrodo se encuentra ubicado a 3.5 cm por encima del ombligo hacia la derecha y finalmente el cuarto electrodo se encuentra ubicado 3.5 cm por debajo del ombligo en la parte derecha [4]. La diferencia de potenciales entre estos electrodos fue registrada en tres canales de la siguiente forma:

- PRIMER CANAL: S1 = E2-E1 (1)
- SEGUNDO CANAL: S2 = E2-E3 (2)
- TERCER CANAL: S3 = E4-E3 (3)

Teniendo en cuenta los efectos producidos por diferentes fenómenos que aportan información innecesaria o ruido que deterioran la calidad de la señal a la hora de realizar el procesamiento de las señales se dispuso a filtrar la información contenida en cada canal teniendo en cuenta que las señales EHG operan mejor en un rango de banda comprendido entre 0 hasta 5 Hz [4], de esta forma, fue necesario realizar un filtrado, las señales EHG han sido filtradas entre 0.3 Hz a 4 Hz utilizando un filtro digital Butterworth, también se han recortado 180 segundos al comienzo y al final de cada señal con el fin de evitar efectos transitorios, por último, se tomaron 1394 segundos para cada señal, es decir, cada registro tiene un total de 27880 muestras.

B. Caracterización de Señales

Si se entiende el útero como un sistema bioeléctrico, el cual está compuesto por miles de millones de células, se puede tratar con un sistema complejo y dinámico, teniendo en cuenta investigaciones anteriores [10] se realiza una caracterización de las señales, teniendo en cuenta un análisis en tiempo, frecuencia y tiempo-frecuencia como lo son las componentes Wavelet [15].

Las características utilizadas en este estudio comparativo se especifican en la Tabla 1.

Tabla1. Características representativas para EHG

Característica	Formulación Matemática
Integral bajo la Curva	$\sum_{n=1}^{N}\lvert(x_n)\rvert$
Media	$\frac{1}{N}\sum_{n=0}^{N}\lvert x_n\rvert$
Raíz Media cuadrática (VRMS)	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}x_n^2}$
Integral Simple Cuadrada	$\sum_{n=0}^{N}\lvert x_n\rvert^2$
Varianza	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
Desviación estándar	$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
Diferencia absoluta de la desviación estándar	$\frac{1}{N-1}\sum_{n=1}^{N-1}(x_{n+1}-x_n)^2$
Longitud Fractal Máxima	$\log 10\left(\sqrt{\sum_{n=1}^{N-1}(x_n-x_{n+1})^2}\right)$
Cambio del promedio de Amplitud	$\frac{1}{N}\sum_{n=1}^{N-1}\lvert x_{n+1}-x_n\rvert$
Entropía	$H=-\sum_{I}pi(log_2 pi)$
Frecuencia Pico	$\arg\left(\frac{f_s}{N}max_{i=0}^{N-1}P(i)\right)$
Frecuencia Media	$i_m\frac{f_s}{N},\sum_{i=0}^{i=i_m}P(i)=\sum_{i=i_m}^{i=N-1}P(i)$
Componentes Wavelet	$\sum_{k}h(k-2n)X(k)$

II. Fase experimental

Una vez identificado los registros a trabajar, se procedió a realizar la extracción de características, el análisis de las mismas, y uso de métodos de selección de características y reducción de dimensión para obtener una mejor representación de los registros de EHG a la hora de clasificarlos.

A. Selección de Características

Los registros de EHG fueron representados por medio de características calculadas en el dominio del tiempo y frecuencia, debido a que la cantidad de características era realmente extensa,

se presentó la necesidad de implementar técnicas de selección de características, de esta forma se obtuvo una matriz de atributos más compacta y apropiada a la hora de representar y clasificar este fenómeno, para ello se eliminaron atributos redundantes e irrelevantes, con el objeto de minimizar problemas relacionados con la maldición de la dimensionalidad, sobre aprendizaje[6] que conlleven a modelos resultantes confusos además de ser nocivos en los algoritmos de aprendizaje. En [10], las características más relevantes fueron, la raíz media cuadrada (VRMS), frecuencia media y no lineales como la muestra de entropía, sin embargo, en este estudio las características que mejor representaron los registros de EHG fueron las componentes de Wavelet, resultados obtenidos mediante el uso de dos métodos de selección de características ranking y sub set. El método *bestfirst* de ranking analiza el desempeño de las características a una a una respecto a una característica principal, este fue implementado y descartado porque los resultados obtenidos no representaron eficientemente los registros EHG, como, el método de *subset* que selecciona características mediante su desempeño en subconjuntos, de esta forma fue necesario recurrir al método de selección por decisión de árboles J48 basados en el algoritmo ID3 [7][8][9], esta afirmación se evidencia en la sección IV.

B. Reducción de dimensión

La reducción de dimensión fue utilizada en la extracción de atributos, de esta forma se buscó obtener características más separables[4]. Para ello se consideraron métodos convencionales y no convencionales que incluyen análisis en el espectro completo y disperso, algunos de ellos son: Análisis de Componentes Principales (PCA), Ajuste Localmente Lineal (LLE) y Análisis del Espacio Tangencial Local Lineal (LLTSA) [16].

- *Análisis de Componentes Principales (PCA)*

Este método identifica un primer componente que presente la mayor cantidad de varianza, un segundo componente que tenga la siguiente mayor cantidad y así sucesivamente, en términos matemáticos [13], [16]:

$$cov(X)M = \lambda M \qquad (4)$$

Donde $cov(X)$ es la matriz de la covarianza de las muestras, X es la matriz de características, además se tiene la matriz principal de vectores propios λ

- *Método de Ajuste Localmente Lineal (LLE)*

LLE asume una estructura localmente lineal en los datos, de forma que cada muestra es susceptible de regresión lineal a partir de sus vecinos más próximos, matemáticamente se expresa:

$$\varepsilon_{II}(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} y_i^T y_{j=tr(YMY^T)} \qquad (5)$$

Donde M es una matriz n x n encontrada como $M = (I - W)^T(I - W)$, y Y contiene el $y_j\,´s$ como sus columnas

- *Análisis del espacio Tangencial Local Lineal (LLTSA)*

Utiliza una técnica lineal para minimizar la función de coste en del Análisis de espacio Tangencial Local (LTSA), partiendo del algoritmo matemático tenemos:

$$\arg min_{x,\theta,Q} \sum_{j=1}^{K} \left\| x_{i_j} - (x + Q\theta_j) \right\|_2^2$$
$$= \arg min_{\theta,Q} \|x_i H_k - (x + Q\theta)\|_2^2 \qquad (6)$$

Donde $H_k = I - \frac{ee^T}{k}, Q$ es una matriz ortonormal del espacio tangencial con d columnas $\theta = [\theta_1, \dots, \theta_k]$, donde θ_j es la coordenada local correspondiente para la base Q.

C. Clasificación

Una vez extraídos los atributos finales se procede a realizar la clasificación a partir métodos supervisados, es decir, se tiene en cuenta un conocimiento a priori para determinar si un objeto está dentro de una u otra categoría. En esta investigación se presentan dos posibles condiciones, por tanto, la clasificación es biclase, dentro del algoritmo de clasificación se propuso realizar el entrenamiento con 30 muestras correspondientes a cada clase y 8 muestras de cada tipo para el test de validación.

Los clasificadores definidos en el espacio de características, que mejor generalizaron las técnicas simples de reconocimiento de patrones utilizados en este estudio fueron:

- *Clasificador normal lineal (LDC)*

Este clasificador asume que todas las clases se caracterizan por múltiples distribuciones normales con igual matriz de covarianza *S.* Teniendo en cuenta que el caso de estudio es un problema de dos clases, el

[6] http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/SELECCION%20DE%20ATRIBUTOS.pdf

clasificador LDC [16], es matemáticamente expresado así:

$$f(x) = \left[x - \frac{\bar{x}_{(1)}+\bar{x}_{(2)}}{2}\right]^t S^{-1}\left(\bar{x}_{(1)} + \bar{x}_{(2)}\right) + 2log\frac{P_{(1)}}{P_{(2)}} \quad (7)$$

Donde:

- $P_{(i)}$, i = 1,2 probabilidades a priori

- *Clasificador normal cuadrático (QCD)*

Este clasificador asume que las clases tienen múltiples distribuciones normales, pero cada una es caracterizada por una matriz de covarianza diferente, para el caso de estudio con un problema biclase, el clasificador QDC, se expresa matemáticamente:

$$f(x) = \sum_{i_1}^{2}(-1)^i \left(x - \bar{x}_{(i)}\right)^t S_{(i)}^{-1}\left(x - \bar{x}_{(i)}\right) + 2\log\frac{|P_{(1)}|}{|P_{(2)}|} + \log\frac{|S_{(1)}|}{|S_{(2)}|} \quad (8)$$

Donde:

- $P_{(i)}$, i = 1,2 probabilidades a priori

- *Sopport vector clasifier (SVC)*

Este clasificador basado en máquinas de vectores de soporte (SVC), busca un hiperplano que separe los puntos de una clase con otra de la forma más óptima, teniendo en cuenta la característica fundamental, es decir este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de el mismo, siendo previamente proyectado a un espacio de dimensión superior, además, están estrechamente relacionados con las redes neuronales [16] [17][18] dado que por naturaleza el clasificador SVC es muy funcional en clasificaciones biclase, ha sido implemento puesto que represento una buena opción para el estudio de los registros de EHG.

- *Clasificador de PARZEN*

El clasificador Parzen nace del concepto básico de aprendizaje supervisado no paramétrico basado en la estimación de densidades de probabilidad de Parzen

El objetivo de este clasificador es obtener estimaciones de densidades de probabilidad condicional $p(z|w_k)$, el espacio de medida es patrocinado en un número finito de regiones disyuntas R_i llamadas cajas y se cuentan las muestras que caen en ellas, siendo la estimación de la densidad de probabilidad dentro de la caja proporcional a tal número, Además $N_{k,i}$ denota el número de muestras con clase W_K, en [16] matemáticamente se expresa:

$$\hat{p}(z|w_k) = \frac{N_{k,j}}{volumen\ (R_i)*N_k} \quad (9)$$

III. Resultados

Con el fin presentar una comparación que involucre las técnicas anteriormente descritas, en esta sección se presentan los resultados obtenidos de los cuatro clasificadores a la hora de determinar un embarazo con labor a término y pretermito.

Los resultados presentados en las tablas 2 y 3 indican las medianas del error y desviación estándar respectivamente utilizando técnicas de selección de características *subset*, comparando el método de reducción de dimensión con el tipo de clasificador, Adicionalmente, se realiza la misma comparación en las tablas 4 y 5 con selección de características en *ranking*

De la Tabla 2 se puede decir que la mejor combinación entre método de reducción y tipo de clasificador es PCA con LDC, presentándose el mejor resultado para la mediana del error en 18.75 %, del mismo modo, se puede observar en la tabla 3, que esta combinación arroja el mejor resultado en la desviación estándar para este valor de mediana de error.

Así mismo es importante resaltar que los métodos de reducción de dimensión, contribuyen en medida a reducir el rango del error presentado.

Tabla 2. Medianas del error Método de reducción vs clasificador con *Subset*

MEDIANAS DEL ERROR				
	LDC	QDC	SVC	PARZENC
PCA	**18.75**	25	25	25
LLE	25	25	25	25
LLTSA	18.75	18.75	18.75	25
SIN METODO	25	25	25	31.25

Tabla 3. Desviación Estándar de Medianas del error Método de reducción vs clasificador con *Subset*

DESVIACION ESTANDAR DE MEDIANAS DEL ERROR				
	LDC	QDC	SVC	PARZENC
PCA	**8.12**	8.27	8.85	8.50
LLE	9.06	7.83	10.12	8.94
LLTSA	9.00	8.31	9.20	9.60
SIN METODO	9.50	11.89	9.50	10.76

Es importante reconocer que el tipo de selección de características influyo dramáticamente en el resultado final de la clasificación como se puede observar en las tablas 4 y 5 y contrastar con las tablas 2 y 3 respectivamente

Tabla 4. Medianas del error Método de reducción vs clasificador con *Ranking*

MEDIANAS DEL ERROR				
	LDC	QDC	SVC	PARZENC
PCA	31.25	31.25	31.25	25
LLE	31.25	31.25	37.5	31.25
LLTSA	37.5	31.25	37.5	31.25
SIN METODO	34.37	37.5	31.25	31.25

Tabla 5. Desviación estándar de Medianas del error Método de reducción vs clasificador con *Ranking*

DESVIACION ESTANDAR DE MEDIANAS DEL ERROR				
	LDC	QDC	SVC	PARZENC
PCA	10.67	10.65	10.79	11.18
LLE	10.04	10.54	10.03	10.48
LLTSA	10.34	9.35	9.91	9.75
SIN METODO	10.49	10.52	10.57	9.69

IV. Conclusiones y trabajo futuro

El papel que juegan los sistemas de información médica en el campo bioeléctrico son relevantes al momento de analizar un fenómeno médico, que contribuya a la solución de una condición crítica presentada en una población y más aún como jóvenes investigadores, es importante para nosotros ser parte de dicha solución.

Teniendo en cuenta que el tema de análisis de señales EHG es relativamente nuevo, es esperanzador y a la vez un reto a mejorar, el obtener un resultado de clasificación con respecto a la mediana del error en 18.75 con una desviación estándar de 8.12 % de una investigación en curso, como se pudo observar y algo que surgió de la investigación es la influencia de las técnicas de selección de características en el resultado final, también cabe anotar el desempeño de los métodos de reducción de dimensión y finalmente el tipo de clasificador que en su conjunto operan para dar un mejor resultado, se espera sin embargo reducir el error, valorando técnicas de diezmado en la señal y evaluar también métodos de clasificación no supervisados.

De esta forma se ve la gran necesidad de avanzar en el desarrollo de esta investigación con el fin de contribuir a menguar los estragos de un embarazo pre-término tanto como a quienes padecen directas

consecuencias, así como también en la salud emocional de las madres gestantes.

Referencias

[1] Althabe, F., Carroli, G., Lede, R., Belizán, J. M., & Althabe, O. H. (1999). El parto pretérmino: detección de riesgos y tratamientos preventivos.

[2] Roura, L. C. (2006). *Parto prematuro*. Ed. Médica Panamericana.

[3] Ortega Maroto, G. N., & Hinojosa León, Y. A. (2010). Causas del embarazo en las adolescentes y riesgos de complicaciones en el recién nacido en el área de ginecología y obstetricia del Hospital Provincial General Docente Riobamba en el período enero a julio del 2010.

[4] Fele-Žorž, G., Kavšek, G., Novak-Antolič, Ž., & Jager, F. (2008). A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups. *Medical & biological engineering & computing*, *46*(9), 911-922.

[5] Monteiro, A. V. (2010). *Electrohisterografia dinâmica intra-parto: contribuição para o desenvolvimento de um protótipo* (Doctoral dissertation, Universidade da Beira Interior).

[6] ALBEROLA RUBIO, J. O. S. É. (2011). Diseño y desarrollo de un sistema para el registro y monitorización de la actividad mioeléctrica uterina.

[7] Alfonso, J. D. L. M. C. DIAGNÓSTICO DE DIABETES UTILIZANDO LOS ALGORITMOS APRIORI Y J48.

[8] Martínez, G. R. S., & Mejía, J. A. S. (2011). Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares. *Scientia et Technica*, *3*(49),

[9] Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M. D. C. G., León, P. P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista médica de la Universidad Veracruzana*, *9*(2), 19-24. 104-109.

[10] Yousefi, J., & Hamilton-Wright, A. (2014). Characterizing EMG data using machine-learning tools. *Computers in biology and medicine*, *51*, 1-13.

[11] Trujillo Pulgarín, C. A. *Clasificación basada en la estimación de Parzen en espacios generalizados de disimilitudes= Classification based on the Parzen estimation in generalized dissimilarity spaces* (Doctoral dissertation, Universidad Nacional de Colombia-Sede Manizales).

[12] Gunn, S. R. (1998). Support vector machines for classification and regression.*ISIS technical report*, *14*.

[13] Fernández, F., & María, A. (2013). Análisis de componentes principales.

[14] Isasi Viñuela, P., & Galván León, I. M. (2004). Redes de neuronas artificiales. *Un Enfoque Práctico, Editorial Pearson Educación SA Madrid España*.

[15] PAREDES, José; LUZARDO, Ender; BRICEÑO, Hildemaro. A Wavelet based method to characterize electrical insulators under partial discharges. *Revista Técnica de la Facultad de Ingeniería. Universidad del Zulia*, 2005, vol. 28, no 2.

[16] Van Der Heijden, F., Duin, R., De Ridder, D., & Tax, D. M. (2005). *Classification, parameter estimation and state estimation: an engineering approach using* MATLAB. John Wiley & Sons.

[17] DUDA, Richard O.; HART, Peter E.; STORK, David G. *Pattern classification*. John Wiley & Sons, 2012.

[18] Pekalska, Elzbieta, and Robert PW Duin. "Foundations and Applications." (2005).

ANEXO 2. Ponencia: INCISCOS

Se realizó una ponencia en la Universidad Tecnológica Equinoccial, sobre la investigación realizada en el presente trabajo de grado.

Official Certificate

Caipe Gordillo Ángela Stephanya

attended the
International Conference on Information Systems and Computer Science **(INCISCOS 2016)**,
held in Quito-Ecuador, November 24-26, 2016
and presented a paper entitled:

Técnicas de Machine Learning para la determinación de embarazos pretermino a partir del electrohisterograma

as a conference speaker

On behalf of the Organizing Committee,

Ing. Juan Bravo, Ph.D.
DEAN FCII-UTE

Ph.D. Sergio Luján Mora
INCISCOS General Chair

www.UTE.edu.ec

Figura 15. Certificados de la ponencia presentada en el evento internacional INCISCOS desarrollado en la Universidad Tecnológica Equinoccial entre 24 y 26 de noviembre del 2016.

ANEXO 3. Ponencia: AUNAR DatavisDay

Se realizó en la universidad autónoma de Nariño, una ponencia sobre la investigación realizada en el presente trabajo de grado.

CORPORACIÓN UNIVERSITARIA AUTÓNOMA DE NARIÑO

OTORGA EL PRESENTE

CERTIFICADO A:

Angela Stephanya Caipe Gordillo

Con documento de identidad No.

PARTICIPO COMO PONENTE CON:

Técnicas de machine learning en predicción de parto prematuro

EN AUNAR DATAVIS DAY

Realizado en las instalaciones de AUNAR, del 25 al 27 de Agosto

Se firma en San Juan de Pasto a los 27 días del mes de Agosto de 2016

Dra. Ingreed Colunge
Rectora

Mg. Constanza Colunge
Vicerrectora Académica

Ing. Milton Cabrera Alvarez
Coordinador Programa
Ingeniería Informática

CORPORACIÓN UNIVERSITARIA AUTÓNOMA DE NARIÑO

OTORGA EL PRESENTE

CERTIFICADO A:

Jorge Armando Muñoz Rosero

Con documento de identidad No.

PARTICIPO COMO PONENTE CON:

Técnicas de machine learning en predicción de parto prematuro

EN AUNAR DATAVIS DAY

Realizado en las instalaciones de AUNAR, del 25 al 27 de Agosto

Se firma en San Juan de Pasto a los 27 días del mes de Agosto de 2016

Dra. Ingreed Colunge
Rectora

Mg. Constanza Colunge
Vicerrectora Académica

Ing. Milton Cabrera Alvarez
Coordinador Programa
Ingeniería Informática

Figura 16. Certificados de la ponencia presentada en el evento binacional Aunar DataVis day desarrollado en la Universidad Autónoma de Nariño entre 25 y 27 de agosto del 2016.

ANEXO 4. PAGINA WEB.

En el desarrollo de este proyecto, se contempla la creación de una página web en google sites, donde se puede encontrar información general de la interfaz desarrollada, así como algunos scripts. Un manual de usuario y un video tutorial que explica el funcionamiento de la interfaz gráfica, con el fin de dar visibilidad al proyecto y fomentar la divulgación de los resultados obtenidos.

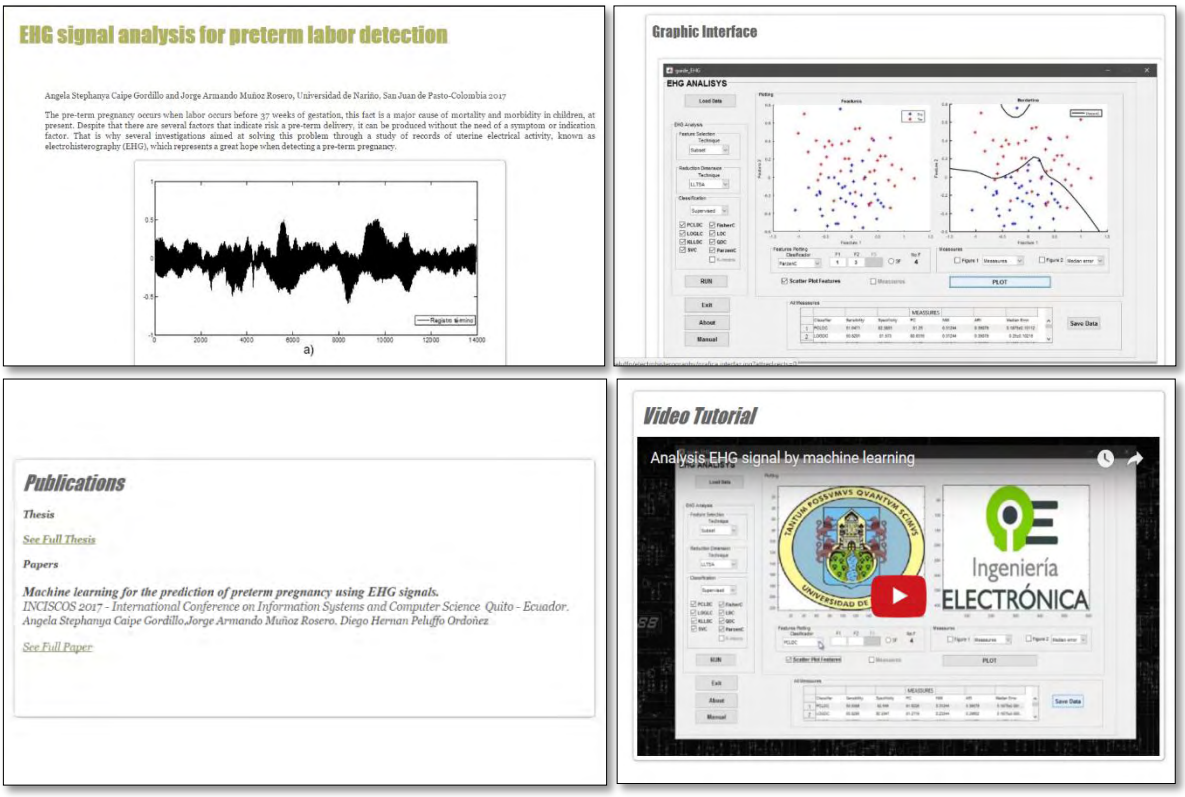


Figura 17. Diseño de la página web donde se encuentra información acerca de la interfaz desarrollada, así como scripts, tutoriales, manuales y publicaciones realizadas.

[3] https://sites.google.com/site/degreethesisdiegopeluffo/electrohisterography