



Universidad  
del Cauca

**DESARROLLO DE UN MODELO PREDICTIVO A PARTIR DE ESPECTROS DE MASAS MALDI-TOF DE SUEROS SANGUÍNEOS COMO MÉTODO DE DIAGNÓSTICO DE PREECLAMPSIA.**

**LAURA CAMILA MACA CASTRO**  
Grupo de Investigación en Procesos Electroquímicos  
**GIPEL**

Universidad del Cauca  
Facultad de Ciencias Naturales, Exactas y de la Educación  
Programa de Química  
Popayán, Colombia  
2024

**DESARROLLO DE UN MODELO PREDICTIVO A PARTIR DE ESPECTROS DE MASAS MALDI-TOF DE SUEROS SANGUÍNEOS COMO MÉTODO DE DIAGNÓSTICO DE PREECLAMPSIA.**

**Trabajo de grado en la modalidad de  
investigación Presentado como requisito para optar  
al título de Química.**

**Director**

**Dr. German Cuervo Ochoa  
Grupo de investigación en procesos Electroquímicos  
(GIPEL)**

**Línea de investigación**

**Desarrollo de métodos analíticos para la determinación de sustancias**

**Asesor**

**Ph.D Enrique Mejía Ospino  
Laboratorio de Espectroscopia Atómica y Molecular  
(LEAM)**

**Colaboradores**

**Fernanda Amezquita, Qca.  
Norma Serrano, Ms.D.  
Paula Bautista, Ph. D.  
Yuly Andrea Prada, Ph.D Qca.  
Tania Gutiérrez, Ph.D Qca.  
Claudia Colmenares, Ph. D.**

**Universidad del Cauca  
Facultad de Ciencias Naturales, Exactas y de la Educación  
Programa de Química  
Popayán, Colombia  
2024**

## **Dedicatoria**

A mis padres, Ana María Castro y Manuel Bolívar Maca, por su valiosa crianza, apoyo incondicional, amor y paciencia; a mis hermanos Ernesto, Konchita, Andrés, Manuela y nuestro angelito Matías por simplemente estar en mi corazón les dedico este logro como muestra de que pueden lograr lo que se propongan. También a mis abuelos Flor y Antonio por sus cuidados y consejos.

A Fabián Ramírez Chavarro por su compañía, alegría, ánimos, paciencia, enseñanzas, amor y ejemplo de persistencia.

A todos mis profes por formarme en esta maravillosa Ciencia.

A mi persona, por la disciplina, dedicación y responsabilidad

## Agradecimientos

Quiero agradecer profundamente a la vida y mis padres.

A la profe Tania Gutiérrez junto con el profe German Cuervo por la inversión de su tiempo en la dirección de este trabajo y por todo su apoyo en las actividades del Semillero de Investigación en Procesos Electroquímicos y Separativos (SIPRES) y el Grupo de Investigación en Procesos Electroquímicos (GIPEL) de la Universidad del Cauca.

A el profesor Enrique Mejía por brindarme su confianza en el manejo del Espectrómetro de Masas con Ionización-Desorción Láser Asistida por Matriz con Tiempo de Vuelo (MALDI-TOF-MS) del Laboratorio de Espectroscopía Atómica y Molecular (LEAM) del Parque Tecnológico de Guatiguará de la Universidad Industrial de Santander (UIS).

A la Dra. Yuly Andrea Prada por su grato acompañamiento, valiosas enseñanzas en la realización de esta investigación y el compartir de experiencias.

A Fernanda Amézquita, profesional en Química y joven Investigadora, por ser mi compañera de investigación, por su paciencia, amistad y enseñanzas durante la estadía en Bucaramanga.

A la Dra. Paula Katherine Bautista por su acogida, confianza, apoyo y enseñanzas en la Fundación Cardiovascular de Colombia (FCV).

A la Dra. Norma Serrano, investigadora líder del Estudio Genómica y Preeclampsia (Genpe) de la FCV, por su colaboración en la disposición de las muestras de suero sanguíneo a trabajar en esta investigación.

A Ian Soler por sus enseñanzas en el uso adecuado del equipo MALDI-TOF-MS, su paciencia y ayuda cuando se bloqueaba la Target.

A Vanessa y Yenny Velandia que nos introdujeron al trabajo en el LEAM.

A los integrantes del Grupo de Investigación en Bioquímica y Microbiología (GIBIM) de la UIS

A la profesional Amparo Martínez, por su compartir de conocimientos y guía en el manejo del biobanco en el Centro Tecnológico Empresarial (CTE) de la FCV.

A la Universidad del Cauca, por admitirme estudiar Química.

A la Organización de Estados Iberoamericanos (OEI), el Ministerio de Ciencias de Colombia (MINCIENCIAS) y el programa de Mas Mujer, Mas Ciencia y Mas Equidad (+M+C+E) por la oportunidad de realizar la pasantía de investigación en la ciudad de Floridablanca-Santander.

**Nota de aceptación**

**Director**

---

Dr. Germán Cuervo Ochoa

**Jurado**

---

Ph.D Paola Andrea Gómez Buitrago

**Jurado**

---

Ph.D Fernando José Hernández Blanco

**Fecha y hora de sustentación: 17 de mayo del 2024 a las 9:00 am**

Colombia-Cauca-Popayán-Universidad del Cauca.

# Contenido

<b>1.</b>	<b>RESUMEN .....</b>	<b>14</b>
<b>2.</b>	<b>PALABRAS CLAVE .....</b>	<b>15</b>
<b>3.</b>	<b>ABSTRACT .....</b>	<b>16</b>
<b>4.</b>	<b>GRAPHICAL ABSTRACT .....</b>	<b>17</b>
<b>5.</b>	<b>INTRODUCCIÓN .....</b>	<b>18</b>
<b>6.</b>	<b>PLANTEAMIENTO DEL PROBLEMA .....</b>	<b>20</b>
<b>7.</b>	<b>OBJETIVOS.....</b>	<b>21</b>
7.1	OBJETIVO GENERAL. ....	21
7.2	OBJETIVOS ESPECÍFICOS. ....	21
<b>8.</b>	<b>MARCO TEÓRICO Y ANTECEDENTES.....</b>	<b>22</b>
8.1	EL EMBARAZO EN UNA MUJER.....	22
8.2	PREECLAMPSIA .....	23
8.2.1	<i>Teorías más relevantes del origen de la preeclampsia.....</i>	<i>24</i>
8.2.2	<i>Diagnóstico, tratamiento y prevención actual de la preeclampsia. ....</i>	<i>26</i>
8.3	MUESTREO DE MUESTRAS BIOLÓGICAS- SUERO SANGUÍNEO. ....	28
8.4	BIOMARCADORES PROTEÓMICOS.....	28
8.5	ESPECTROMETRÍA DE MASAS MALDI-TOF EN LA PROTEÓMICA PARA EL DIAGNÓSTICO DE ENFERMEDADES. ....	30
8.5.1	<i>Principio de MALDI-TOF-MS. ....</i>	<i>31</i>
8.6	METODOLOGÍA PARA PREPARACIÓN DE MUESTRAS ASISTIDA POR FILTRO (FASP). ....	34
8.7	CIENCIA DE DATOS. ....	35
8.7.1	<i>Machine Learning y manejo de datos MALDI-TOF-MS. ....</i>	<i>36</i>
8.8	MODELOS DE APRENDIZAJE UTILIZADOS.....	38
8.8.1	<i>Técnica de análisis de componentes principales, modelo no supervisado. ....</i>	<i>40</i>
8.8.2	<i>Regresión logística, modelo supervisado. ....</i>	<i>41</i>
8.8.3	<i>Máquina de vectores de soporte, modelo supervisado.....</i>	<i>42</i>
8.8.4	<i>Árboles aleatorios o Random Forest, modelo supervisado. ....</i>	<i>42</i>
8.8.5	<i>Refuerzo extremo de gradiente, modelo supervisado. ....</i>	<i>43</i>
8.9	MATRICES DE CONFUSIÓN. ....	44
<b>9.</b>	<b>METODOLOGÍA EXPERIMENTAL.....</b>	<b>45</b>
9.1	EQUIPOS .....	45
9.2	MATERIALES .....	45
9.3	REACTIVOS.....	45
9.4	SOFTWARES Y PROGRAMAS .....	45
9.5	ETAPAS GENERALES DE LA METODOLOGÍA.....	45
9.5.1	<i>Etapa 1 .....</i>	<i>46</i>
9.5.2	<i>Etapa 2 .....</i>	<i>49</i>
9.5.3	<i>Ajuste de condiciones para MALDI-TOF-MS .....</i>	<i>49</i>
9.5.4	<i>Etapa 3 .....</i>	<i>53</i>
9.5.5	<i>Cuantificación de proteínas .....</i>	<i>55</i>
<b>10.</b>	<b>RESULTADOS Y ANÁLISIS .....</b>	<b>57</b>
10.1	ETAPA 1.....	57
10.2	ETAPA 2 .....	60

10.3	CUANTIFICACIÓN DE PROTEÍNAS .....	68
10.4	ETAPA 3 .....	72
<b>11.</b>	<b>ANEXOS.....</b>	<b>85</b>
<b>12.</b>	<b>CONCLUSIONES.....</b>	<b>86</b>
<b>13.</b>	<b>BIBLIOGRAFIA.....</b>	<b>88</b>

## LISTA DE FIGURAS

<b>Figura 1.</b> Mecanismo de peroxidación lipídica por radicales libres. _____	25
<b>Figura 2.</b> Diagnósticos, tratamientos y prevenciones de la PE actualmente. _____	27
<b>Figura 3.</b> Ciencias ómicas en relación con el dogma central de la biología molecular. _____	29
<b>Figura 4.</b> Partes fundamentales de un espectrómetro de masas MALDI-TOF _____	31
<b>Figura 5.</b> Etapas de trabajo en ciencia de datos. _____	35
<b>Figura 6.</b> Diagrama de Venn de las disciplinas de la ciencia de datos, entre ellas el Machine Learning. _____	36
<b>Figura 7.</b> Técnicas de minado de datos. _____	37
<b>Figura 8.</b> Secuencia de procesamiento de datos de espectros de masas para en modelo predictivo. _____	38
<b>Figura 9.</b> Ejemplificación gráfica de un modelo predictivo. _____	39
<b>Figura 10.</b> Explicación esquemática de método de aprendizaje no supervisado de análisis de componentes principales PCA. _____	41
<b>Figura 11.</b> Comparación explicativa del modelo random forest. _____	43
<b>Figura 12.</b> Representación explicativa del modelo refuerzo extremo de gradiente. _____	43
<b>Figura 13.</b> Representación general de una matriz de confusión para una clasificación binaria. _____	44
<b>Figura 14.</b> Etapas generales de la metodología. _____	46
<b>Figura 15.</b> Flujograma de etapa 1. Digestión enzimática por metodología FASP. _____	49
<b>Figura 16.</b> Metodología limpieza de placa Target MALDI-TOF-MS _____	51
<b>Figura 17.</b> Preparación de muestras de fragmentos peptídicos para el análisis MALDI-TOF-MS. _____	52
<b>Figura 18.</b> Modo de sembrado o deposición de la mezcla muestra-matriz en la placa. _____	52
<b>Figura 19.</b> Resumen del cuadernillo de código que representa la construcción del modelo predictivo de PE. _____	54
<b>Figura 20.</b> Criocajas con 164 muestras de suero sanguíneo con y sin PE y descongelamiento controlado de 24 muestras por día. _____	57
<b>Figura 21.</b> Mecanismo de acción del DTT en la metodología FASP. _____	59
<b>Figura 22.</b> Mecanismo de reacción de IAA en la metodología FASP. _____	59
<b>Figura 23.</b> Fragmentos peptídicos después de SpeedVac. Apariencia cristalina. _____	61
<b>Figura 24.</b> Colección de imágenes de sembrados de matrices en la placa de MALDI-TOF-MS. _____	62
<b>Figura 25.</b> Apariencia de sembrados en la placa MALDI-TOF. _____	63
<b>Figura 26.</b> Espectros de masas MALDI-TOF bajo las condiciones establecidas. a) matriz HCCA, b) patrón Vapreotida. _____	64
<b>Figura 27.</b> Espectro de la muestra # 101 con etiqueta de caso (sí). _____	65
<b>Figura 28.</b> Espectro de la muestra # 11 con etiqueta de caso (sí). _____	66
<b>Figura 29.</b> Espectro de la muestra # 13 con etiqueta de control (no). _____	66
<b>Figura 30.</b> Espectro de la muestra # 99 con etiqueta de control (no). _____	67
<b>Figura 31.</b> Apariencia de las microplacas para la cuantificación de sueros originales. _____	68
<b>Figura 32.</b> Apariencia de microplaca para la cuantificación de fragmentos peptídicos en TA30. _____	68
<b>Figura 33.</b> Reacciones del principio de detección de proteínas por el método de Kit de BCA. _____	69
<b>Figura 34.</b> Curva de cuantificación para sueros sanguíneos originales en agua. _____	70
<b>Figura 35.</b> Curva de cuantificación de fragmentos peptídicos en TA30. _____	70
<b>Figura 36.</b> Visualización de la matriz de datos. 194049 puntos de (intensidad, m/z) por 164 espectros. _____	72
<b>Figura 37.</b> Datos de intensidades normalizadas de espectros de casos y controles de PE. _____	73
<b>Figura 38.</b> Diagrama de Codo para la elección de componentes principales. _____	74
<b>Figura 39.</b> Diagrama de PCA para reducción de dimensionalidad a 2 componentes principales. _____	74
<b>Figura 40.</b> Diagrama de pares para relacionar los 8 componentes principales. _____	76
<b>Figura 41.</b> Ilustración de hiperplanos del modelo SVM en conjunto de entrenamiento. _____	77
<b>Figura 42.</b> Matrices de confusión para modelo supervisado. _____	80
<b>Figura 43.</b> Gráfico de barras de comparación de métricas por clase de cada modelo supervisado. _____	82
<b>Figura 44.</b> Selección de características a partir del modelo RF. _____	83
<b>Figura 45.</b> Selección de características a partir del modelo XGBOOST. _____	83
<b>Figura 46.</b> Huella peptídica de intensidades relativas para los pacientes de casos PE. _____	84



## LISTA DE TABLAS

<b>Tabla 1.</b> Clasificación de los trastornos hipertensivos en el embarazo. _____	23
<b>Tabla 2.</b> Algunos ejemplos de compuestos orgánicos utilizados como matriz en MALDI-TOF-MS. _____	32
<b>Tabla 3.</b> Funciones y Características de la matriz en MALDI-TOF-MS. _____	32
<b>Tabla 4.</b> Comparación de los enfoques del aprendizaje automatizado: aprendizaje supervisado y no supervisado. _____	37
<b>Tabla 5.</b> Cálculos para la preparación se soluciones de digestión – FASP. _____	47
<b>Tabla 6.</b> Factores que afectan el análisis MALDI-TOF-MS _____	50
<b>Tabla 7.</b> Variables de respuesta de los factores que afectan el análisis MS-MALDI-TOF. _____	50
<b>Tabla 8.</b> Librerías utilizadas en el cuadernillo para el modelo predictivo PE. _____	53
<b>Tabla 9.</b> Esquema de dilución para el protocolo estándar en tubo de ensayo y procedimiento de microplaca. _____	56
<b>Tabla 10.</b> Comparación de concentración de proteínas en sueros sanguíneos intactos y fragmentos peptídicos después de digestión enzimática disueltos en TA30. _____	71
<b>Tabla 11.</b> Comparación de exactitud de los modelos predictivos supervisados. _____	79
<b>Tabla 12.</b> Reporte de clasificación por clases de modelos supervisados. _____	79

## LISTA DE ECUACIONES

<b>Ecuación 1.</b> Energía cinética de un ion adquirida por un láser. _____	33
<b>Ecuación 2.</b> Energía cinética de una masa en movimiento. _____	33
<b>Ecuación 3.</b> Transformación de energía cinética, aplicado a la ionización MALDI. _____	33
<b>Ecuación 4.</b> Descripción del movimiento rectilíneo uniforme. _____	33
<b>Ecuación 5.</b> Relación masa/carga. _____	34
<b>Ecuación 6.</b> Ecuación sigmoide o función logística general. _____	41
<b>Ecuación 7.</b> Ecuación de regresión logística. _____	41
<b>Ecuación 8.</b> Función de decisión en general. _____	42
<b>Ecuación 9.</b> Función de decisión para el modelo SVM. _____	42
<b>Ecuación 10.</b> Función de kernel radial para el modelo SVM. _____	42
<b>Ecuación 11.</b> Precisión, métrica de evaluación por clases. _____	44
<b>Ecuación 12.</b> Sensibilidad, métrica de evaluación por clases. _____	44
<b>Ecuación 13.</b> Puntuación F1. métrica de evaluación combinada por clases. _____	44
<b>Ecuación 14.</b> Cálculo de solución TA30 para re-suspensión se los fragmentos peptídicos para MALDI-TOF-MS. _____	51
<b>Ecuación 15.</b> Formula de volumen de solución de trabajo en con el kit de cuantificación de proteínas. _____	55
<b>Ecuación 16.</b> Exactitud. Métrica global de evaluación de los modelos supervisados. _____	78
<b>Ecuación 17.</b> Ejemplo de cálculo de la métrica precisión para casos en el modelo SVM. _____	81
<b>Ecuación 18.</b> Ejemplo de cálculo de métrica precisión para controles en el modelo SVM. _____	81
<b>Ecuación 19.</b> Ejemplo de cálculo de métrica sensibilidad para casos en el modelo RL. _____	81
<b>Ecuación 20.</b> Ejemplo de cálculo de métrica sensibilidad para controles en el modelo RL. _____	81
<b>Ecuación 21.</b> Ejemplo de cálculo de puntuación F1 para casos del modelo RF. _____	82
<b>Ecuación 22.</b> Ejemplo de cálculo de puntuación f1 para controles del modelo RF. _____	82

## LISTA DE ABREVIATURAS

ABREVIATURA	SIGNIFICADO
ACN	Acetonitrilo
AS	Ácido Sinapínico
BD	Base de Datos
CD	Ciencia de Datos
CTE	Centro tecnológico empresarial
D	Dilución
DANE	Departamento Administrativo Nacional de Estadística
DBH	Ácido 2,5- dihidroxibenzoico
DOE	Design Of Experiments / Diseño de Experimentos
MALDI-TOF-MS	Matrix-Assisted Laser Desorption ionization-Time Of Flight- Mass Spectrometry / Espectrometría de Masas- Ionización Desorción por Laser Asistida por Matriz-Analizador de Tiempo de Vuelo
ESI	Electrospray Ionization / Ionización por Electrospray
ET-1	Endotelina 1
FASP	Filter Aided Sample Preparation / Preparación de Muestras Asistida con Filtro
FCV	Fundación Cardiovascular de Colombia
FSH	Hormona Folículo estimulante
GenPE	Genómica y Preeclampsia
GIBIM	Grupo de investigación en bioquímica y Microbiología
GIPEL	Grupo de investigación en procesos electroquímicos
HCCA	Ácido $\alpha$ -ciano-4- hidroxicinámico
HELLP	Hemolysis Elevated Liver enzymes and Low Platelet count / Hemólisis microangiopática, Enzimas hepáticas Elevadas, Bajo recuento de plaquetas, trombocitopenia
IA	Inteligencia Artificial
IAA	Iodo Acetamida
LDL	Low Density Lipoproteins / Lipoproteínas de baja densidad
LEAM	Laboratorio de Espectroscopía Atómica y Molecular
LH	Hormona luteinizante
MBT.par	Micro bio tools para analizar microorganismos
+M+C+E	Mas mujer, más ciencia y más equidad
MED-FASP	Preparación de muestras asistida con filtro -Digestión multi enzimática
MD	Minado de datos
MINCIENCIAS	Ministerio de Ciencias de Colombia
ML	Machine Learning / Aprendizaje automatizado
MWCO	Molecular Weight Cutoff/ Membrana de corte de peso molecular
NIH	National Institutes of Health/ Institutos Nacionales de Salud
OEI	Organización de Estados Iberoamericanos
OMS	Organización Mundial de la Salud
PAD	Presión arterial diastólica
PAS	Presión arterial sistólica

<b>PASD</b>	Presión arterial sistólica y diastólica
<b>PCA</b>	Principal Component Analysis / Análisis de componentes principales
<b>PE</b>	Preeclampsia
<b>RANME</b>	Real Academia Nacional De Medicina De España
<b>RF</b>	Random Forest / Árboles aleatorios
<b>RL</b>	Regresión logística
<b>SIPRES</b>	Semillero de Investigación en Procesos Electroquímicos y Separativos
<b>S.S</b>	Suero sanguíneo
<b>SVM</b>	Support Vector Machine / Máquina vector de soporte
<b>TFA</b>	Trifluoroacetic acid / Ácido trifluoroacético
<b>THAE</b>	Trastornos hipertensivos asociados al embarazo
<b>.txt</b>	Archivo tipo texto
<b>UIS</b>	Universidad Industrial de Santander
<b>XGBOOST</b>	Refuerzo de gradiente extremo

## LISTA DE SIMBOLOS

SÍMBOLO	SIGNIFICADO
<i>Ec</i>	Energía cinética de un ion
<i>Z</i>	Carga del ion desorbido
<i>e</i>	Carga de un electrón
<i>V</i>	Voltaje de la placa del tubo
<i>Ek</i>	Energía cinética de una masa
<i>m</i>	Masa
<i>v</i>	Velocidad
<i>x</i>	Posición o cualquier dato
<i>t</i>	Tiempo
m/z	Relación masa-carga
M	Molar
mmHg	Milímetros de mercurio
mg	Miligramos
g	Gramos
h	Horas
min	Minutos
kDa	Kilo Dalton
°C	Grados centígrados
rpm	Revoluciones por minuto
V	Volumen
Vf	Volumen final
C	Concentración
PM	Peso molecular
d	Densidad
P	Probabilidad
N	Número de datos

## 1. RESUMEN

En este documento se presentan los resultados del trabajo de investigación realizado durante una pasantía de 6 meses, el cual fue encaminada a brindar aportes para el diagnóstico de la preeclampsia usando técnicas de Espectrometría de Masas con Ionización-Desorción por Láser Asistida por Matriz acoplado a analizador de Tiempo de Vuelo y herramientas computacionales de Aprendizaje Automatizado para el análisis de datos.

La preeclampsia, es una enfermedad grave que puede causar una morbilidad y mortalidad significativas tanto para la madre como para el bebé si no se trata o diagnostica a tiempo. Esta enfermedad, también conlleva riesgos a largo plazo para la salud de la mujer como mayor riesgo de enfermedades cardiovasculares, entre otras en etapas tardías de la vida. Dadas las consecuencias potencialmente graves de esta afección, la investigación sobre la preeclampsia desde una perspectiva clínica y química es esencial para comprender mejor sus mecanismos subyacentes e identificar los factores de riesgo. Por lo tanto, en este trabajo se propuso comparar el perfil proteómico de 164 muestras de suero sanguíneo de pacientes sanas y con preeclampsia. Las muestras se digestaron con Tripsina por 18 horas a 37 grados centígrados, mediante un método preparativo conocido como preparación de muestras asistida por filtro de exclusión de tamaño de tres kilodalton (3 kDa), recolectando los fragmentos proteicos del sobrenadante, los cuales, fueron secados al vacío y analizados mediante Espectrometría de Masas con Ionización-Desorción láser asistida por Matriz acoplado a analizador de Tiempo de Vuelo. Las condiciones de trabajo fueron estudiadas estadísticamente por un diseño factorial mixto 3x3x2 empleando una re- suspensión de 50 microlitros (uL), una dilución de 1:10 uL y se usó el ácido  $\alpha$ -Ciano-4-hidroxicinámico como matriz de ionización depositando la mezcla muestra-matriz en doble capa de sembrado en la placa metálica. El modo de trayectoria de los iones fue lineal. El rango de detección fue de 500 a 6000 relación masa/carga ( $m/z$ ). Los patrones para la digestión enzimática y el análisis espectrométrico fueron los estándares de suero de albumina bovino y el péptido Vapreotida, respectivamente. Los espectros obtenidos fueron preprocesados en el software Flexanalysis y analizados mediante aprendizaje automático en el Navegador Anaconda utilizando lenguaje Python trayendo modelos de aprendizaje no supervisado (análisis de componentes principales) y supervisado (maquina vector de soporte, regresión logística, bosques aleatorios y refuerzo de gradiente extremo) obteniendo algoritmos de aprendizaje en un 80 % de muestras de entrenamiento y métricas de evaluación como exactitud, precisión y sensibilidad en un conjunto de 20 % de prueba. También se analizaron matrices de confusión con el fin de categorizar los modelos por clases de casos y controles. Se encontró que el mejor modelo fue maquina vectores de soporte con una exactitud del 88 % de predictibilidad.

## **2. PALABRAS CLAVE**

Preeclampsia; Suero sanguíneo; Fragmentos peptídicos; Huella peptídica; Espectrometría de masas MALDI-TOF; Machine Learning; Modelo predictivo.

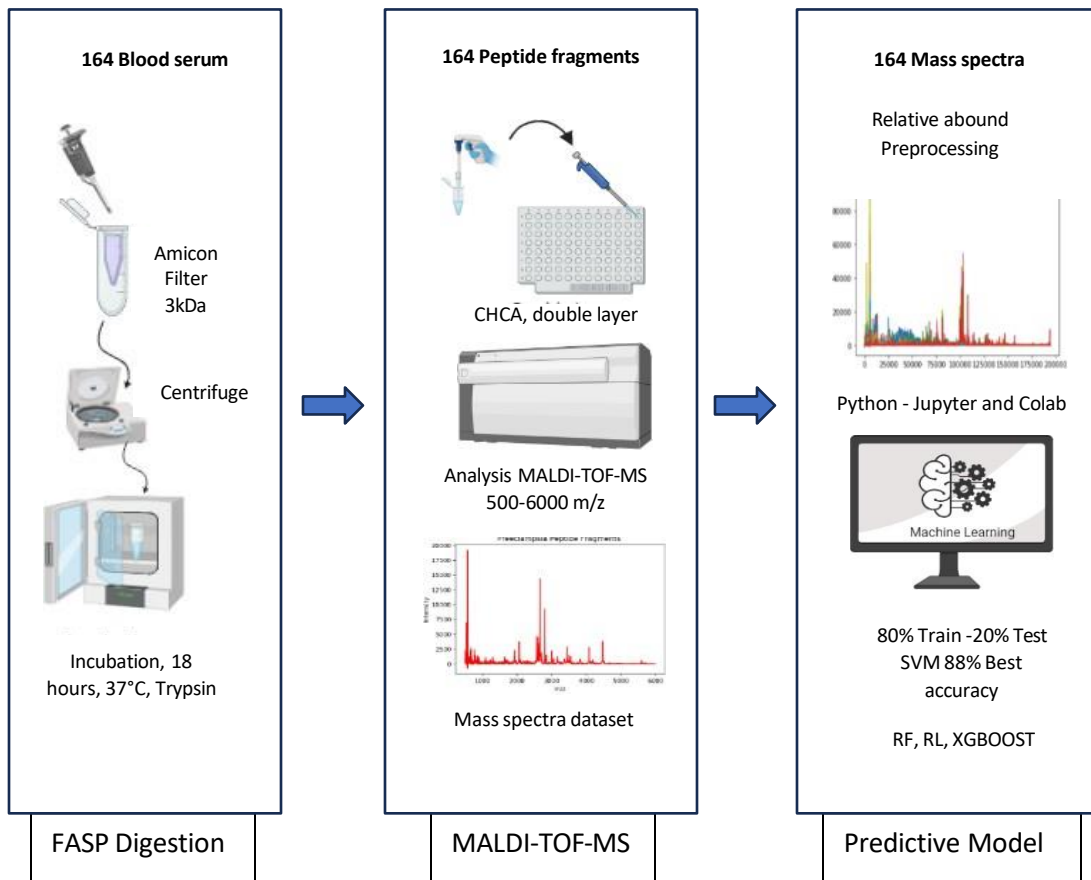
### 3. ABSTRACT

This paper presents the results of the research work performed during a 6-month internship, which was aimed at providing input for the diagnosis of preeclampsia using Mass Spectrometry techniques with Matrix-Assisted Laser Ionization-Desorption Laser Scanning coupled to Time-of-Flight analyzer and Machine Learning computational tools for data analysis.

Preeclampsia is a serious disease that can cause significant morbidity and mortality for both mother and baby if not treated or diagnosed in time. This disease also carries long-term risks to women's health such as increased risk of cardiovascular disease, among others in later stages of life. Given the potentially serious consequences of this condition, research on preeclampsia from a clinical and chemical perspective is essential to better understand its underlying mechanisms and identify risk factors. Therefore, in this work we set out to compare the proteomic profile of 164 blood serum samples from healthy and preeclampsia patients. The samples were digested with Trypsin for 18 hours at 37 degrees Celsius, using a preparative method known as three kilodalton (3 kDa) size exclusion filter-assisted sample preparation, collecting the protein fragments from the supernatant, which were dried under vacuum and analyzed by Matrix Assisted Laser Ionization-Desorption Mass Spectrometry coupled to a Time-of-Flight analyzer. The working conditions were statistically studied by a 3x3x2 mixed factorial design employing a 50 microliter (uL) re-suspension, 1:10 uL dilution and  $\alpha$ -Cyano-4-hydroxycinnamic acid was used as ionization matrix by depositing the sample-matrix mixture in double seeding layer on the metal plate. The ion trajectory mode was linear. The detection range was 500 to 6000 mass-to-charge ratio (m/z). The standards for enzymatic digestion and spectrometric analysis were bovine albumin serum standards and Vapreotide peptide, respectively. The obtained spectra were preprocessed in Flexanalysis software and analyzed by machine learning in Anaconda Browser using Python language bringing unsupervised (principal component analysis) and supervised (support vector machine, logistic regression, random forests, and extreme gradient boosting) learning algorithms in 80 % of training samples and evaluation metrics such as accuracy, precision, and sensitivity in a 20 % test set. Confusion matrices were also analyzed to categorize the models by classes of cases and controls. The best model was found to be support vector machine with an accuracy of 88 % of predictability.



#### 4. GRAPHICAL ABSTRACT



## 5. INTRODUCCIÓN

Según Karrar [1], la preeclampsia (PE) es una enfermedad hipertensiva que se presenta durante el embarazo y abarca del 2 al 8% de todas las complicaciones que puede tener una mujer en este estado. El diagnóstico de la PE hoy en día está dado por la identificación de los síntomas que suelen aparecer después de las 20 semanas de gestación. Estos son presión arterial sistólica (PAS) mayor a 140 mmHg y presión arterial diastólica (PAD) mayor a 90 mmHg. En una mujer con este trastorno estas cifras llegan a 160 y 110 mmHg respectivamente. Otros síntomas incluyen hinchazón en las piernas, dolores fuertes de cabeza, visión borrosa, dificultades para respirar y daños clínicamente estudiados en órganos como disfunción hepática, dolor epigástrico o síntomas de reflujo, náuseas, trombocitopenia manifestada con hematomas en la piel y por no mencionar más, aumento de excreción de proteína en la orina conocida como proteinuria que esté en el límite mayor a 300 mg/24 h. Aunque son varios síntomas que se presentan y pueden ser confundidos con otras enfermedades, la PE es multisistémica y no se tiene conocimiento etiológico esclarecido. Es por eso por lo que la PE es una enfermedad conocida por sus propuestas teóricas [2].

Hasta ahora la explicación más aceptada es la de una implantación inadecuada de la placenta, pero sigue siendo una fuente de investigaciones en medicina y por supuesto en las ciencias básicas como la química y biología. La PE es un trastorno anterior a su complicación, eclampsia, por lo que puede ser identificada tempranamente y tratada para que no haya peores consecuencias. La importancia del estudio de esta enfermedad y el soporte de esta investigación está dada por las cifras actualizadas al año 2023 por el instituto nacional de salud (NIH) reportando que 50.000 mujeres y 500.000 fetos mueren en todo el mundo [1], lo que hace indispensable el hecho de diagnosticar tempranamente la enfermedad.

Actualmente, el interés en el estudio de la PE está ligada a la identificación de biomoléculas que puedan relacionarse en los procesos bioquímicos y que están asociados con diferentes biomarcadores moleculares extraídos de diferentes muestras biológicas. En esta investigación se usan los sueros sanguíneos (S.S) de pacientes casos y controles.

Por otro lado, analíticamente se desarrollan metodologías para la eliminación de estos interferentes. En este caso, los S.S se sometieron a la metodología de preparación de muestras asistida por un filtro (FASP) de tamaño de poro de 3 kDa que funciona como un reactor o recipiente para realizar la digestión enzimática pasando de la estructura compleja de las proteínas a fragmentos peptídicos sin necesidad de realizar partes del procedimiento en diferentes recipientes, preparando la muestra para analizarla por espectrometría de masas [3]. En este ámbito, resalta la espectrometría de masas con ionización- láser asistida por matriz con analizador de tiempo de vuelo (MALDI-TOF-MS) que es una técnica de ionización suave adecuada para la identificación de las proteínas y sus derivados [4].

Así mismo, gracias a la interdisciplinariedad que puede tener esta investigación y teniendo en cuenta que la MALDI-TOF-MS recolecta muchos datos que en ocasiones los seres humanos no tienen la capacidad de procesar al mismo tiempo y correctamente, conlleva a la oportunidad de intervenir en la ciencia de datos (CD) acercándose al Aprendizaje Automático o Machine Learning, (ML) por sus siglas en inglés, que son algoritmos de

enfoque computacional que se especializan en buscar patrones que sean modelados con el fin de distinguir entre grandes cantidades de datos sin necesidad de saber que variables afectan en la diferenciación que pueden aportar a un diagnóstico, como lo es el objetivo de esta investigación. En este sentido, el ML en conjunto con la química pueden trabajar en pro de dar un análisis de datos de mayor precisión y rapidez [5], sin llegar a la discusión típica de la controversia actual de qué profesión es la indispensable y cual no. Por lo cual, más allá de los objetivos de la investigación, mostrar como la inteligencia artificial (IA) puede ayudar a resolver problemas y despertar el interés por saber más de estas técnicas de análisis y dar un buen resultado a tal finalidad. Aclarando que, por supuesto no se pretende encontrar la explicación científica del origen de la enfermedad si no sentar algunas bases de posibles investigaciones químicas aprovechando la versatilidad de la IA y ML. Por esto, el propósito de este trabajo es presentar el desarrollo de un conjunto de algoritmos que represente un modelo de diagnóstico para la PE, desde luego enlazando con la parte química que corresponde a la obtención de los datos que son los espectros de masas MALDI-TOF.

En este contexto, fue de vital importancia que la investigación realizada en instituciones de educación superior en áreas como la bioquímica fuera aplicada en ámbitos clínicos reales a través de redes de investigación traslacional e interdisciplinar con el fin de encontrar alternativas en la solución de problemas de salud pública del país como lo es el diagnóstico oportuno de PE. Por lo tanto, el presente trabajo logró una alianza científica entre la Universidad del Cauca, la Fundación Cardiovascular de Colombia (FCV) y la Universidad Industrial de Santander (UIS) apoyada y financiada por la Organización de Estados iberoamericanos (OEI) y por el Ministerio de Ciencia y Tecnología con el programa de “Mas Mujer, Mas Ciencia y Mas equidad”.

## 6. PLANTEAMIENTO DEL PROBLEMA

Los trastornos hipertensivos asociados al embarazo (THAE) han sido y siguen siendo hoy en día una preocupación en todo el mundo. De acuerdo con la organización mundial de la salud (OMS) es la causa de más del 20 % de las muertes maternas [6], [7]. En Colombia, un estudio realizado por el departamento administrativo nacional de estadística (DANE) [8] mostró que para el año 2021 los casos de morbilidad materna aumentaron a un 39,9% por cada 1000 nacidos vivos, equivalentes a casi 10.000 mujeres embarazadas que enfermaron por alguna razón; dentro de la mitad de los casos que se clasificaron un 12,3 % de los registros eran causados por THAE [9], en el que la PE, es uno de los más estudiados y se desconoce su origen y cómo evoluciona, pero se sabe que sus síntomas aparecen después de las 20 semanas de gestación, y suelen consistir en dolores fuertes de cabeza, hinchazón en las piernas y ritmo del corazón acelerado. Al ser una enfermedad multifactorial, los síntomas pueden ser confundidos con hipertensión y otras cardiopatías por lo cual las pacientes no reciben el cuidado y atención adecuada. En casos aislados algunas mujeres asintomáticas solo manifiestan síntomas de eclampsia al momento del parto, lo cual aumenta significativamente la muerte materno-fetal. Por lo general, clínicamente se tienen exámenes rutinarios de diagnóstico, como lo es la determinación de proteinuria (>300 mg/24 h) y monitoreo de la presión arterial sistólica/diastólica (PASD) (>140 mm Hg y >90 mmHg) respectivamente y en caso de estar fuera de los niveles normales se remite inmediatamente a el control médico; sin embargo, estos síntomas no son específicos para PE [10].

Desde un enfoque científico, las ciencias biológicas y químicas son aplicadas constantemente a mejorar la salud y elevar la calidad de vida de la comunidad, siendo el desarrollo de métodos de diagnóstico de enfermedades una de las líneas de investigación de mayor impacto en el ámbito clínico. La comprensión en los mecanismos bioquímicos involucrados en la aparición de enfermedades, así como el estudio de los factores asociados a dichas enfermedades ha permitido alcanzar una mayor eficiencia en los servicios de salud. La búsqueda de nuevas metodologías, rápidas, simples y confiables para el estudio de patologías complejas como el cáncer, la diabetes [11], desórdenes metabólicos [12], enfermedades cognitivas [13] y la preeclampsia [14], está basada principalmente, en la identificación y detección de biomarcadores presentes en muestras biológicas (sangre, orina y saliva) y así comparar y diferenciar el perfil molecular (metabólico/proteómico) de los pacientes.

De acuerdo con lo anterior, la recolección y tratamiento de datos químicos pueden ser aprovechados con ML que hoy en día sobresale en varios campos mostrando las ventajas que se tienen a la hora de manejar grandes cantidades de datos y aplicando métodos de modelado basado en algoritmos con enfoque computacional y teoría matemática, que se pueden utilizar para extraer un patrón de características específicas comparables en las huellas peptídicas a partir de los datos de espectros de masas.

Es por lo anterior que esta investigación pretende dar un aporte al estudio de una enfermedad de salud pública enlazando técnicas químicas y de la ciencia de datos para que pueda trascender en el ámbito social y científico.

## **7. OBJETIVOS.**

### **7.1 Objetivo general.**

Construir un modelo predictivo para el diagnóstico de preeclampsia basado en espectrometría de masas MALDI-TOF y aprendizaje automatizado Machine Learning.

### **7.2 Objetivos específicos.**

Ajustar las condiciones para la obtención de adecuados espectros de masas por MALDI-TOF.

Adquirir los espectros de masas de 164 muestras de sueros sanguíneos de pacientes con PE y pacientes de control.

Formular un modelo predictivo basado en los espectros de masas obtenidos por MALDI-TOF-MS usando herramientas de machine Learning.

## 8. MARCO TEÓRICO Y ANTECEDENTES

### 8.1 El embarazo en una mujer.

Dentro de las etapas de la vida de algunas mujeres se encuentra el embarazo, así se le llama al periodo en el cual un feto se desarrolla en el útero de la mujer convirtiéndola en madre. Un embarazo normal suele durar aproximadamente 40 semanas o 9 meses, su inicio es desde el último periodo menstrual hasta el momento del parto. El personal médico divide el embarazo en tres etapas, primer, segundo y tercer trimestre; también hacen referencia a estas fases con un rango de semanas, de 1 a 12, de 13 a 28 y de 29 a 40, respectivamente [15]. Además, según la evolución y crecimiento celular antes del nacimiento del bebé se divide en 3 periodos que son germinal o pre- embrionario (1-2 semanas), embrionario (3-8 semanas), y el fetal (9-40 semanas) [16].

En cada trimestre la mujer experimenta cambios físicos y químicos en su cuerpo debido al crecimiento natural del feto, puede sufrir diferentes síntomas cambiantes, que deben ser monitoreados por médicos, La rama de la medicina que se dedica a la atención y el tratamiento de las madres antes, durante y después del nacimiento del feto es la obstetricia y también se complementa con la ginecología. Al ser estas las especialidades médicas en tratar a las mujeres en condición de embarazo son estos los profesionales que pueden diagnosticar por primera vez patologías o complicaciones que se presenten en el cuerpo de la mujer y/o su bebé, como la PE por medio de exámenes clínicos. Pero bioquímicamente es todo un conjunto de secreción de hormonas femeninas como: Hormona liberadora hipotalámica (GnRH): se produce en el hipotálamo y libera gonadotropinas, consta de una secuencia de 10 aminoácidos (Glu-His-Trp-Ser-Tyr-Gly-Leu-Arg-Pro-Gly-NH<sub>2</sub>), viaja pulsativamente hacia la hipófisis. Hormonas adenohipofisarias: estas son dos, la foliculoestimulante (FSH) y la luteinizante (LH). Hormonas ováricas: también son dos, estrógeno y progesterona, sintetizadas en principio a partir del colesterol en una mujer no gestante y se encargan de producirse cíclicamente para la expulsión del ovulo. En una gestante son sintetizadas en la placenta por esteroides endógenos secretados por la glándula suprarrenal, permiten el estiramiento del útero y lactancia. Gonadotropina coriónica humana: secretada por las células embrionarias en un pico máximo hasta las 12 semanas del embarazo, impidiendo la menstruación y favoreciendo el sostenimiento del feto en el útero [17].

No obstante, en el embarazo otros factores hormonales pueden formarse en la placenta y en general todas las glándulas endocrinas no sexuales responden químicamente al embarazo por la carga metabólica. Cuando se presentan anomalías en la función principal de la placenta, que es el transporte de nutrientes al feto, aparecen otros factores como citocinas inflamatorias y péptidos con función antiangiogénica (inhibe formación de vasos sanguíneos) que pueden afectar los tejidos de la mujer. La incógnita de la PE es el desconocimiento de la identificación y función exacta de dichos factores, es por esta razón que se define hasta el momento como una alteración relacionada con la placenta.

Aunque en este trabajo no se estudian los mecanismos metabólicos de las diferentes moléculas en el cuerpo de una mujer es indispensable reconocer que las alteraciones bioquímicas asociadas al embarazo pueden contribuir a la explicación de la PE y que investigaciones posteriores a esta en lo posible puedan relacionarla con tal objetivo.

## 8.2 Preeclampsia

En la antigüedad, las primeras descripciones de complicaciones de mujeres en el embarazo eran confundidas con epilepsia [18] y en Francia surgieron los primeros textos acerca de la eclampsia [19],[20] y fue con el paso del tiempo del estudio y monitoreo de la emergencia en el parto que se escribió sobre la condición pre a la eclampsia.

Actualmente, se sabe que la PE es un síndrome hipertensivo y es este el más importante dentro de las complicaciones obstétricas, la hipertensión durante el embarazo es uno de los problemas más importantes de estudio y no se sabe a ciencia cierta cuál es la razón de que estar en estado de embarazo produzca cambios en la tensión [21]. Es pertinente aclarar algunos términos que por la incierta naturaleza de la PE ha tendido a confusiones en el campo médico y por ende en el ámbito científico y que se establecen en la **Tabla 1**.

**Tabla 1.** Clasificación de los trastornos hipertensivos en el embarazo.

TRASTORNOS HIPERTENSIVOS EN EL EMBARAZO	
TRASTORNO	DESCRIPCION Y CARACTERISTICAS
Hipertensión gestacional	También llamada hipertensión transitoria. No se desarrolla PE. Ausencia de proteinuria. Se resuelve antes de las 12 semanas postparto. PAS $\geq$ 140 mmHg o PAD $\geq$ 90 mmHg. Algunas veces molestia epigástrica. Trombocitopenia.
Síndrome preeclampsia (requerimientos mínimos)	PAS $\geq$ 140 mmHg o PAD $\geq$ 90 mmHg después de 20 semanas de gestación. Proteinuria $\geq$ 300 mg/24 h o $\geq$ 1+ con tira reactiva Creatinina sérica $\geq$ 30 mg/mmol
Síndrome de preeclampsia (severa)	PAS $\geq$ 160 mmHg o PAD $\geq$ 110 mmHg después de 20 semanas de gestación. Proteinuria $\geq$ 2,0 g/24 h o $\geq$ 2+ con tira reactiva. Creatinina sérica $\geq$ 30 mg/mmol. Plaquetas $<$ 150000 mm <sup>3</sup> . Cefalea, visión borrosa, vomito, papiledema, fosfenos, dolor en flanco derecho, hipersensibilidad hepática, elevación de lipoproteínas de baja densidad (LDL), Hemólisis microangiopática, enzimas hepáticas elevadas, bajo recuento de plaquetas, trombocitopenia (síndrome de HELLP).
Eclampsia	Convulsiones que no pueden atribuirse a otras causas en una mujer con PE, es la complicación de la PE severa. Hiperreflexia, convulsiones, cefalea, alteraciones visuales, edema pulmonar, aparece antes del décimo día postparto.

Preeclampsia superpuesta a hipertensión crónica	Proteinuria de inicio reciente $\geq 300$ mg en mujeres hipertensas, pero sin proteinuria antes de las 20 semanas de gestación, aumento súbito de la presión arterial.
Hipertensión crónica	PAS $\geq 140$ mmHg o PAD $\geq 90$ mmHg antes de 20 semanas de gestación, no atribuible a enfermedad trofoblástica gestacional.

Tomado y adaptado de Williams [21], Beltrán [22]

El diccionario de términos médicos de la real academia nacional de medicina de España (RANME) [23] define la PE como:

*“Síndrome clínico y heterogéneo caracterizado por la existencia de un daño endotelial que precede al diagnóstico clínico. Es específico de la gestación humana y está relacionado con una placentación anormal que conduce a una mayor capacidad de la perfusión placentaria. La hipoperfusión relativa produce daño endotelial y aumento de estrés oxidativo y provoca una respuesta inflamatoria sistémica, alteraciones en la angiogénesis y disfunción endotelial generalizada asociada a vasoespasmos.”*

Explicando un poco esta definición, la PE se considera un síndrome clínico heterogéneo debido a la variedad de síntomas que se pueden analizar por expertos médicos, como se muestra en la **Tabla 1** e incluso pueden tener aún más variaciones dependiendo de otros factores como la edad y enfermedades previamente adquiridas de los progenitores entre otras [24]. Por otro lado, el endotelio es la capa que recubre el interior de los vasos sanguíneos y ayuda a la coagulación de la sangre, por ende, el flujo de esta interviene en procesos inflamatorios y reparación de los tejidos [25]. Debido a que el proceso de la formación de la placenta es a partir de las células del embrión y se incrusta en el útero, lo que surge en algunas ocasiones es que se implanta en lugares no favorables para el suministro de nutrientes y oxígeno necesarios para el crecimiento adecuado del feto, lo que puede conducir a una implantación anormal de la placenta, o en la mayoría de los casos, ni siquiera implantarse ya que podría conllevar al aborto atendido del embrión; aunque no es la única razón, aún se estudia por qué se da este proceso defectuosamente lo que conduce a que no exista buen flujo de sangre a través de aquella pared, a este paso de sangre a través de la placenta se le conoce como perfusión placentaria y la disminución del flujo se le llama hipoperfusión relativa, esto provoca daño en el endotelio y desencadenamiento de estrés en este tejido, lo que contrasta a un desequilibrio entre los radicales libres y moléculas con potencial poder antioxidante, al cual el cuerpo responde con inflamación sistémica alterando absolutamente todo el organismo de la mujer y el feto, algunas de las consecuencias ya fueron nombradas en la **Tabla 1**. Del mismo modo, son de ayuda para el diagnóstico no solo de la PE si no de los THAE en general.

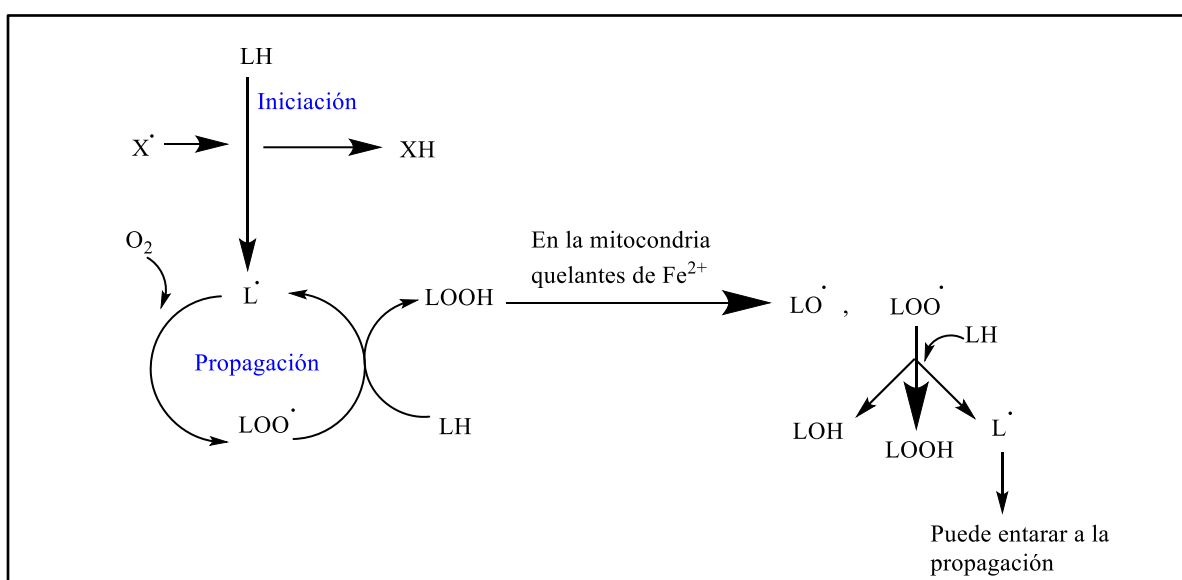
### 8.2.1 Teorías más relevantes del origen de la preeclampsia.

La PE es llamada la enfermedad de las mil teorías. A partir de observaciones se han sacado varias hipótesis, en las cuales algunos casos se ajustan, sin embargo, no se llega a evidencia concisa sobre el origen de la PE. A continuación, se citan varias teorías propuestas por diferentes autores.

- **Teoría 1:** según Redman y colaboradores [26] en el 2009, la PE es un trastorno de 2 etapas:



- 1. Placentación deficiente:** es la etapa en la que no se presentan síntomas y se caracteriza porque el trofoblasto se remodela defectuosamente lo que hace que no llegue suficiente oxígeno por las arterias uterinas.
- 2. Estrés oxidativo placentario:** por la mala circulación materna se producen factores que reaccionan con inflamación y activación endotelial, pero esta etapa también puede asociarse a otras enfermedades como enfermedad cardíaca, diabetes y obesidad. La disfunción endotelial se debe al estado activo extremo de leucocitos en la circulación materna, este estado produce interleucinas que contribuyen a la formación de peróxidos lipídicos que propagan la cadena de formación de radicales, lesionando el endotelio; este argumento también lo defiende Negre [27] en el 2022 y Opichka [28] en el 2021.



LH=ácido graso poliinsaturado con un hidrogeno metílico H.  $X^\bullet$  = radical libre de poder oxidativo.  $L^\bullet$  = radical lipídico.  $XH$ = especie potencial radicalario.  $LOO^\bullet$  = radical peroxilipídico.  $LOOH$  =peróxido lipídico.  $LOH$ = ácido graso hidroxilado.  
Tomado y adaptado de [29]

**Figura 1.** Mecanismo de peroxidación lipídica por radicales libres.

El estrés oxidativo es un conjunto de reacciones de oxidación periódica, la

**Figura 1** muestra el mecanismo cuando un ácido graso poliinsaturado reacciona con un radical iniciador para formar un radical lipídico que espontáneamente interactúa con el oxígeno para formar un radical peroxilipídico que consecuentemente reacciona con otra molécula de ácido graso presente en el medio desencadenando la propagación de radicales en las mitocondrias de células como los leucocitos y eritrocitos ricas en especies susceptibles al ataque radicalario como los quelantes de ion férrico como la hemoglobina que estimulan este proceso donde la consecuencia es el consumo de paredes celulares que son ricas en ácidos grasos [29].

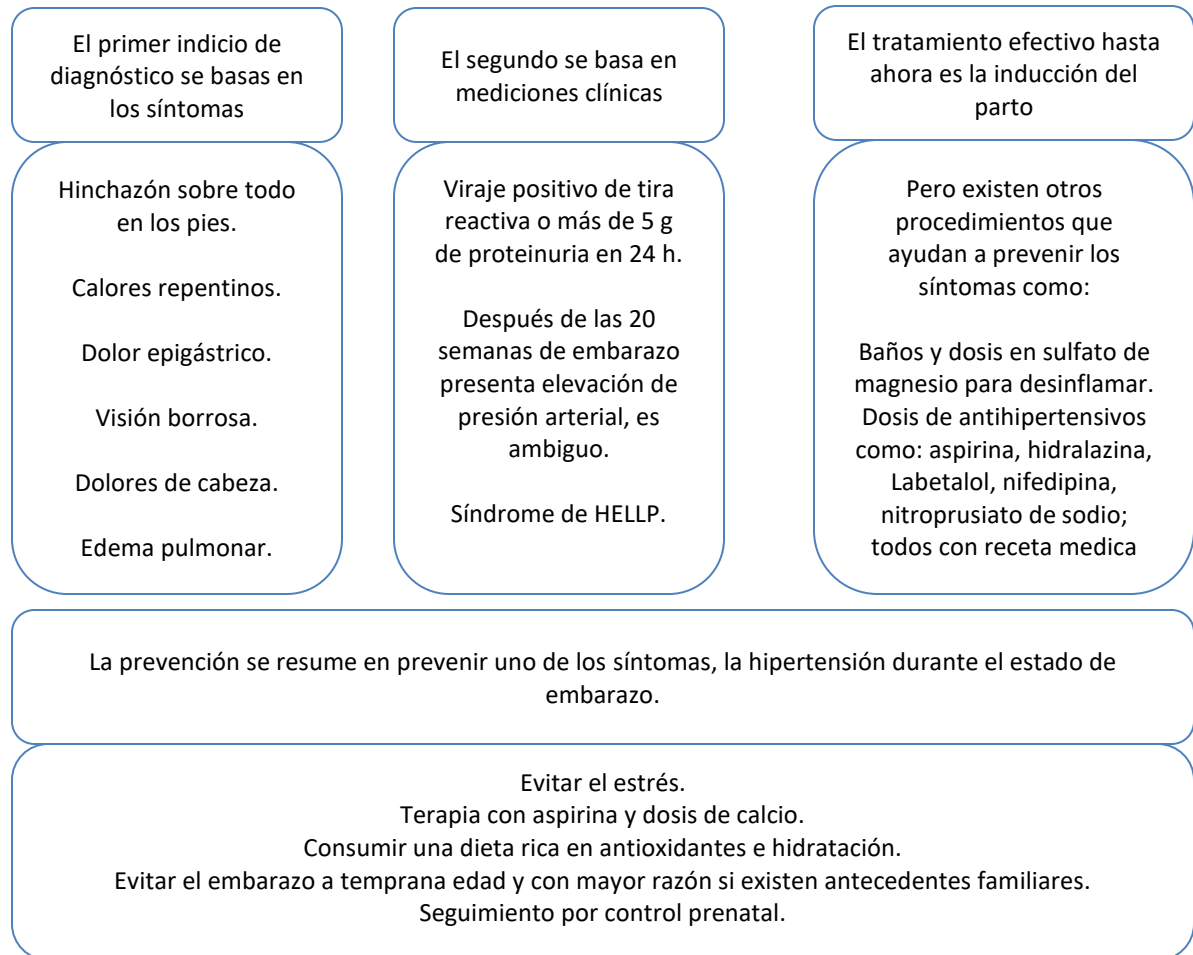
- **Teoría 2:** otra teoría va ligada a la inmunología, Labarrere [30] en 1988 señala que el lecho placentario sufre cambios demostrando que se rechaza al injerto, uno de esos cambios puede ser la aterosclerosis que es el estrechamiento de los vasos sanguíneos por el aumento de las células lipídicas. Lo mismo ocurre cuando se sabe que el embarazo se trata de un síndrome congénito, como la trisomía 13, el cuerpo de la mujer puede expresar antígenos con el fin de interrumpir el proceso donde la placenta se inflama y aumenta el riesgo de sufrir PE.
- **Teoría 3:** en 1990, Newman [31] señaló que la nutrición es un factor importante en el embarazo, por lo que la otra teoría va ligada a que el bajo consumo de alimentos ricos en antioxidantes como la vitamina C, E, Betacarotenos, aceites omegas 3, calcio y exceso de sodio proveniente de la sal de mesa, NaCl podrían potenciar el aumento de una presión sanguínea y por ende la PE.
- **Teoría 4:** Ariza [32] en el 2007 señaló que la insuficiencia de óxido nítrico, vasodilatador poderoso en el cuerpo humano, hace que se presente menor flujo sanguíneo conduciendo a la inflamación. A su vez la disminución de este puede relacionarse con la inhibición de su síntesis a partir del aminoácido L-Arginina, o por un agente vasoconstrictor como las endotelinas que son péptidos de 21 aminoácidos. Esto está relacionado con la endotelina ET-1 que aumenta su concentración en personas normotensas y aún más en mujeres con PE y su tratamiento se realiza con sulfato de magnesio.

### 8.2.2 Diagnóstico, tratamiento y prevención actual de la preeclampsia.

Debido a que la PE es una complicación desde hace mucho tiempo, en la actualidad se ha logrado encontrar un diagnóstico clínico, tratamiento y recomendaciones de prevención aclarando que han sido hallazgos empíricos que han demostrado una efectividad ante este trastorno y evitar llegar a la eclampsia como emergencia.

Para el diagnóstico, algunos métodos son criticados ya que no funcionan en la totalidad de los casos y es común encontrar opiniones contradictorias. Según Chesley [33] pueden ser correctos en aproximadamente la mitad de los casos. La

**Figura 2** muestra un esquema resumen de los diagnósticos, tratamientos y prevenciones ante la PE. Hablar de prevención ante una enfermedad en la que no se sabe con exactitud su origen suele ser subjetivo.



Tomado y adaptado de Sibai [34], Wagner [35].

**Figura 2.** Diagnósticos, tratamientos y prevenciones de la PE actualmente.

La mayoría de las formas de diagnóstico, prevención y tratamiento de la PE conlleva la recolección y manejo de muestras biológicas que pueden contener componentes orgánicos y biológico por lo que todos los estudios en los que se utilicen estas muestras deben contar con métodos de tratamiento y almacenamiento de altísima precaución para garantizar que la muestra se conserve lo más cercana a las condiciones en las que se extrajo del organismo.

### 8.3 Muestreo de muestras biológicas- Suero sanguíneo.

Generalmente en investigaciones con enfoque clínico en el que se estudian enfermedades, se trabaja con pacientes que son seleccionados con fines comparativos. Por lo general, las muestras biológicas a estudiar son fluidos corporales, orina, mucosas, sangre, entre otras. Escoger alguna de ellas para una investigación científica depende de la enfermedad y la facilidad de obtener la muestra.

Sanchez y colaboradores [36] definen los biobancos como “instituciones públicas o privadas sin ánimo de lucro dedicadas a la recolección, el procesamiento, el almacenamiento y la distribución de especímenes biológicos humanos, junto a los datos asociados con esas muestras”. Es de gran importancia conocer este tipo de espacios que ayudan a que surjan investigaciones a lo largo del tiempo y que se pueda dar espera entre los procesos de recolección de muestras y el análisis de estas para planear mucho mejor los proyectos que se tienen en perspectiva.

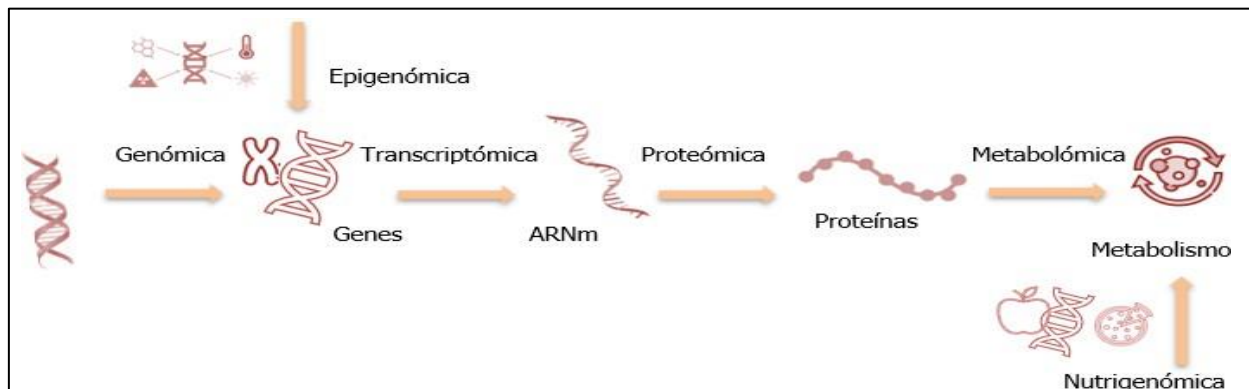
Existen diferentes propósitos por lo que se puede crear un biobanco, pero el principal objetivo a los biobancos que están anclados a hospitales como el del complejo hospitalario de la FCV, pretenden apoyar la investigación tanto en los grupos del propio hospital como de los grupos nacionales e internacionales. La solicitud tanto de muestras como datos de los donantes de las muestras deben ser previamente evaluadas por un comité ético y legal antes de entregarse al investigador principal. En esta investigación se tienen reservados los datos de todas las mujeres con PE y pacientes de control, sin embargo, en el anexo 0 se muestran documentos relacionados con el muestreo de los S.S.

Por otro lado, la sangre tiene varias fracciones que pueden ser separadas con ayuda de reactivos contenidos en los tubos de recolección al vacío, vacutainer, identificados por código de color dependiendo del propósito de la prueba, o bien por centrifugación [37],[38],[39]. Esta es una de las técnicas más utilizadas en la separación de los componentes de la sangre; destacando el suero y plasma.

Cuando se centrifuga sin ningún aditivo se obtiene el S.S, por lo general para estudios de investigación se extrae del tubo de tapa color rojo y el plasma en tubos de tapa color verde que contienen Heparina (sódica, de litio o amonio) [40]; ambos fluidos son similares en apariencia de líquidos amarillentos y diferenciados por la composición como coagulantes naturales, y fibrinógenos presentes en el plasma y no en el suero.

### 8.4 Biomarcadores proteómicos.

Con el crecimiento de los biobancos y el interés por el dogma central de la biología molecular nacieron las ciencias ómicas, **Figura 3** que son de gran ayuda clínica sabiendo que la información genética se relaciona con cambios epigenéticos, la transcripción del ADN a ARNm, transcriptómica, traducción a proteínas, sustratos y metabolitos generando muchos datos que necesitan ser relacionados entre sí. [31].



Tomado y adaptado de Gutiérrez [41].

**Figura 3.** Ciencias ómicas en relación con el dogma central de la biología molecular.

El S.S, aunque es producto de la separación por centrifugación contiene información íntima del ser vivo [42], en este sentido, puede ser comparado con el agua ya que toda sustancia que entre en contacto con ella deja una huella, así mismo sucede con el fluido corporal, la sangre guarda la huella de sustancias que se producen en todo el cuerpo al circular por él; pudiendo determinar su concentración química de biomoléculas, metabolitos, estructura de las mismas y predicción de enfermedades mucho antes de que se presenten síntomas clínicos, y como en su mayoría contiene material proteico, es por eso que la proteómica es de gran ayuda para detectar enfermedades complejas, en este campo es pionera la detección de cáncer que es bien conocido por atacar órganos con un crecimiento acelerado y exagerado de células malignas que forman tumores dirigiendo a un paciente a la muerte, así pues, la tarea de la proteómica en este caso es encontrar biomarcadores, con ayuda de métodos que suelen consistir en realizar primero una electroforesis o cromatografía para la separación de proteínas y luego identificarlas por MS [42], los resultados se reúnen con el fin de analizar los espectros de masas y proponer moléculas, aquí se suelen utilizar bases de datos de identificación de péptidos o posibles biomarcadores. Hoy en día es un reto para la investigación clínica y bioquímica estructurar metodologías para la identificación de biomarcadores en proteómica y demás ciencias ómicas.

Las proteínas, son biomoléculas. biopolímeros, macromoléculas, que se encuentran conformadas por alfa-aminoácidos, estos son compuestos con una estructura caracterizada por 3 partes, un extremo carboxílico, un extremo amínico y una cadena carbonada, esta unidad monomérica permite la interacción con más aminoácidos por medio de enlaces covalentes para la formación de cadenas que establecen péptidos o cadenas de gran longitud de aminoácidos que combinan interacciones electrostáticas para formar estructuras superiores y otros tipos de enlaces químicos otorgando una conformación compleja, confiriéndole propiedades y diferentes funciones que cumplen óptimamente a determinadas condiciones de temperatura, pH y presión dentro de un organismo o sistema bioquímico, además pueden estar relacionadas con otro tipo de biomoléculas generando una clasificación de proteínas conjugadas como núcleo, lipo, fosfo metalo y glucoproteína, por dar los ejemplos más generales.

El tamaño de las proteínas puede ser variable, este depende del número, tipo y organización de residuos de aminoácidos además de los grupos adicionales con los que pueden tener afinidad, pero por lo general se consideran proteínas de pesos moleculares en ordenes de  $10^3$  a  $10^6$  Da [43].

Uno de los retos del análisis proteómico es separar biomarcadores ya que suelen estar en baja concentración y las proteínas en abundancia como la albumina con casi el 50 % de la composición del S.S [42], pueden presentar afinidad, absorbiéndolos a la hora de tratar de eliminarlas, además los fragmentos de interés deben llegar a la identificación lo más íntegros posibles para así poder relacionarlos con su expresión de origen y la enfermedad.

### **8.5 Espectrometría de masas MALDI-TOF en la proteómica para el diagnóstico de enfermedades.**

En la especialidad de la proteómica se emplean dos enfoques principales para el análisis de proteínas: el enfoque ascendente (*Bottom-up*) y el enfoque descendente (*Top-down*). Ambos enfoques culminan en el análisis de proteínas mediante MS. Zhang. [41] diferencia entre estos. En el enfoque *Bottom-up*, las proteínas se descomponen en péptidos más pequeños mediante digestión enzimática utilizando por ejemplo la tripsina. Estos péptidos se separan por cromatografía y se analizan por MS, lo que permite obtener espectros que contienen señales correspondientes a pequeñas partes de un conjunto de proteínas. Por otro lado, en el enfoque *Top-Down*, las proteínas intactas se separan y analizan directamente por MS, sin modificar su estructura terciaria. Esto permite obtener un espectro menos modificado y preserva mejor el conocer como es el comportamiento del analito en la muestra biológica.

A pesar de que ambos enfoques comienzan de manera diferente en el tratamiento de la muestra, ambos convergen en el análisis por MS. Carrera [42] expresa que ambos enfoques son compatibles y pueden integrarse de tal forma que pueda cumplir el objetivo de la investigación en específico. En esta investigación se propone un tratamiento de fragmentos peptídicos, pero sin llegar a su separación e identificación, estaríamos hablando de la integración del enfoque *Bottom up* para la cualificación de características de los espectros de masas por métodos computacionales.

La MS tanto en la química como en la medicina es una técnica rápida, robusta y sensible, estas características son propias de un diagnóstico ante enfermedades que necesitan ser prontamente atendidas. Swiner [43] resalta que el tiempo en el diagnóstico es para otros un tiempo de espera, donde la gravedad de una enfermedad conduce a que la veracidad de los resultados debe ser alta y que el “cuello de botella” en la determinación, como señala el mismo autor, es el pretratamiento y manejo de la muestra biológica, por eso se opta por técnicas automatizadas que realicen el proceso sin errores y lleguen a proporcionar informes. La MS que maneja un equilibrio entre el tiempo, costo y calidad de los resultados es la apropiada en cuanto a la explicación de enfermedades. Zhang [41] también cita a varios autores [44], [45] en los que utilizan la MS acoplada a cromatografía líquida investigando fragmentos peptídicos para asociarlos a enfermedades como el síndrome de Down y autismo por medio de diferentes metodologías para encontrar péptidos o proteínas por medio del hallazgo de posibles biomarcadores [46]. Por lo que la PE no sería una excepción para aplicarla.

Otro ejemplo es la búsqueda de biomarcadores en la enfermedad del cáncer, campo donde se han realizado muchos estudios, pero la mayoría de los biomarcadores encontrados son equívocos para el personal médico, sin embargo, hace una década, Petricoin [47] predijo que la MS conduciría a que las solicitudes medicas serian

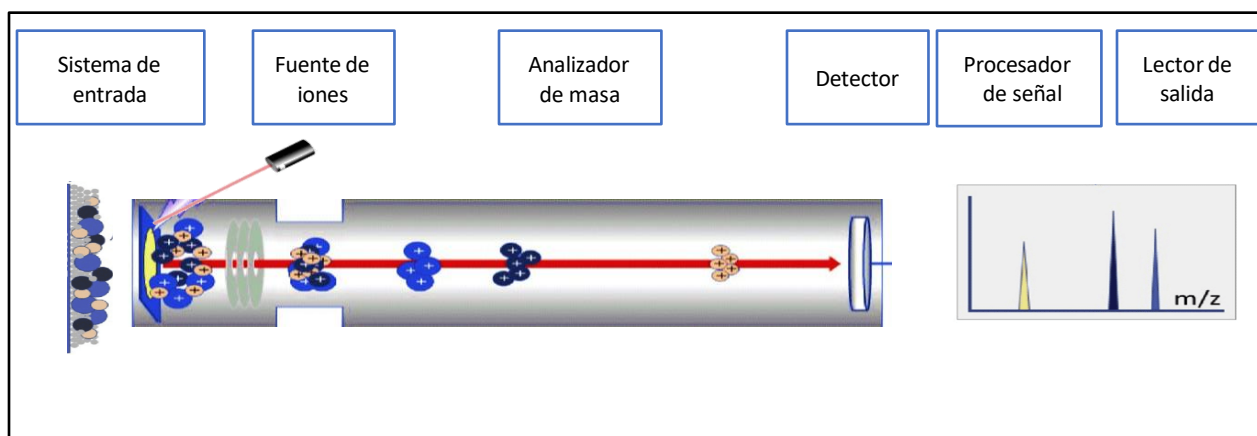
satisfactoriamente resueltas con esta técnica siendo esta, la plataforma futura para el diagnóstico de enfermedades haciendo uso de 3 pasos importantes: proceso para recolectar los datos en MS, selección de algoritmos para el tratamiento de datos e identificar las moléculas correspondientes al patrón de diagnóstico.

De este modo, podría comentar que un sistema como el cuerpo humano que está compuesto para trabajar en conjunto, en la afección de cualquier enfermedad producirá cambios que se podrían detectar por MS.

### 8.5.1. Principio de MALDI-TOF-MS.

La MS ha tenido una larga evolución, pero hoy en día en campo bioquímico se destacan las que proporcionan una ionización suave como la ionización por electrospray (ESI) [48] y MALDI [49] que proporcionan el cuidado de la información de las moléculas termolábiles. MALDI es la forma de ionización y el conjunto MALDI-TOF-MS detecta, en principio, moléculas grandes no volátiles y lábiles que pesen más de 500 Da, valor que antes, era imposible de aislar en medio acuoso y mucho menos cargarlas o ionizarlas para un análisis espectro métrico.

En la **Figura 4** El proceso se basa en la ablación-desorción de la co-cristalización (mezcla con evaporación de solvente) entre una matriz que tenga la capacidad de absorber energía de un láser con el analito en condiciones ácidas sobre una placa metálica (Target) de acuerdo con diferentes estilos de preparación de la muestra. Como se trata de compuestos orgánicos y la mayoría son sólidos cristalinos, la forma de cristalización es distinta y la manera en la que se depositan en el objetivo interactuando con el analito cambia de una matriz a otra. Por lo que existen protocolos de preparación de muestras específicas para MALDI-TOF-MS como los que se citan a continuación [50], [51], [52], [53], [54], [55].



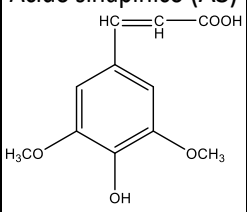
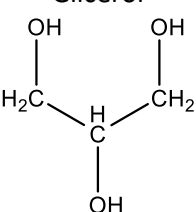
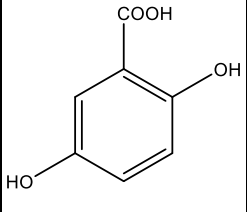
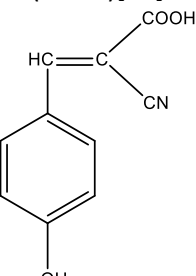
Tomado y adaptado de [56] Skoog [57], Bevnnon [58], Vestal [59], Gross [60] y Gomis [61].

**Figura 4.** Partes fundamentales de un espectrómetro de masas MALDI-TOF

El láser puede ser de longitud de onda en la zona infrarroja o en la zona ultravioleta que apunta hacia el objetivo con un ángulo de  $90^\circ$ . El impacto del láser excita la matriz y desprende partículas para separarse en el analizador TOF. Ya que todo el análisis se realiza en unos cuantos nanosegundos, los mecanismos moleculares son teorías que se acercan a dar la explicación de cómo la matriz interfiere para formar los iones por intercambio de

protones con las proteínas. La matriz no es estándar y se escoge bajo los parámetros de cuánta energía puede absorber y que moléculas se van a analizar, ver **Tabla 2**, pero en general las funciones de la matriz se enlistan en la **Tabla 3** [62].

**Tabla 2.** Algunos ejemplos de compuestos orgánicos utilizados como matriz en MALDI-TOF-MS.

TIPOS DE MATRICES				
Nombre y estructura	Peso molecular (g/mol)	Nombre y estructura	Peso molecular (g/mol)	Aplicación
<p>Ácido sinapínico (AS)</p> 	224,2	<p>Glicerol</p> 	92,1	Proteínas y péptidos
<p>Ácido 2,5-dihidroxibenzoico (DBH)</p> 	154,1	<p>Ácido α-ciano-4-hidroxicinámico (HCCA)[54]</p> 	189,2	<p>Proteínas Péptidos Carbohidratos Polímeros sintéticos (DBH)</p> <p>Proteínas, péptidos (HCCA)</p>

Tomado y adaptado de Duong [62]

**Tabla 3.** Funciones y Características de la matriz en MALDI-TOF-MS.

Funciones y características de la matriz de MALDI-TOF-MS	
FUNCION	CARACTERISTICA
Proporcionar una fuente de absorción de luz láser (energía).	De naturaleza ácida
Transferir energía del láser a la muestra de forma controlada	De bajo peso molecular comparado con polipéptidos y proteínas
Aislar las moléculas individuales de la muestra para evitar agregaciones	

Por otro lado, el principio de detección cuando los iones generados por MALDI son analizados por TOF; se cimienta en el tiempo que se demoran los iones desde que se forman hasta la llegada al detector por medio de un osciloscopio de almacenamiento digital. La calibración del tiempo se realiza internamente con la división de la



luz láser hacia un fotodiodo para que no haya retrasos en las señales eléctricas.

Los iones ya desorbidos pasan inmediatamente al vacío entre placas a las cuales se les aplica un potencial de 25 kV y una rejilla de 23 kV, esto hace que sientan en total una diferencia de potencial relativamente de 2 kV, lo que permite que todos los iones mantengan una misma energía de partida. El análisis del sistema tiene dos momentos: el primero es la energía cinética que adquieren los iones por los rayos del láser representado en la **Ecuación 1** y el segundo es la energía cinética con la que se mueven a través del tubo descrita por la **Ecuación 2**. Donde:  $E_c$  = energía cinética de un ion,  $Z$  = carga del ion desorbido,  $e$  = carga de un electrón,  $V$  = voltaje de la placa del tubo,  $E_k$  = energía cinética de una masa,  $m$  = masa,  $v$  = velocidad,  $x$  = posición y  $t$  = tiempo.

$$E_c = Z \cdot e \cdot V$$

**Ecuación 1.** Energía cinética de un ion adquirida por un láser.

$$E_k = \frac{1}{2} m v^2$$

**Ecuación 2.** Energía cinética de una masa en movimiento.

La ley de transformación de la energía se cumple en este caso por lo que es pertinente comparar las dos energías.

$$E_c = E_k$$

$$Z \cdot e \cdot V = \frac{1}{2} m v^2$$

**Ecuación 3.** Transformación de energía cinética, aplicado a la ionización MALDI.

Por otro lado, la velocidad de la masa:

$$x = v \cdot t$$
$$v = \frac{x}{t}$$

**Ecuación 4.** Descripción del movimiento rectilíneo uniforme.

Se puede reemplazar en la **Ecuación 3**, la velocidad de la **Ecuación 4**, en la ecuación de igualdad de energías, para obtener la relación masa-carga ( $m/z$ ), como se describe en procedimiento para obtener la **Ecuación 5** y observar que la masa de la molécula univalentemente cargada tiene relación proporcional con el tiempo que se demore al recorrer una determinada distancia [62].

$$Z \cdot e \cdot V = \frac{1}{2} m \left( \frac{x}{t} \right)^2$$

$$\left(\frac{t}{x}\right)^2 \cdot e \cdot V = \frac{1}{2} m/Z$$

$$t^2 \cdot \frac{2 \cdot Z \cdot e \cdot V}{x^2} = m$$

$$t^2 \cdot \frac{2 \cdot e \cdot V}{x^2} = m/Z$$

**Ecuación 5.** Relación masa/carga.

Las unidades de masas molecular son los Dalton, que en MS está relacionado con la relación m/z, ambos suelen utilizarse para referirse a las señales en el eje x del espectro, pero no indican lo mismo. Los Da indican el peso total de la molécula cargada y la m/z señala la masa dividida en la carga eléctrica de la molécula, por lo que no siempre una relación m/z corresponderá en valor numérico a la masa de la molécula. Depende de la información que se quiera destacar.

De la técnica se resalta que pueden llegar al detector moléculas de alto peso molecular con tan solo una carga y sin excesiva fragmentación, además se necesita poca muestra para realizar el análisis. El punto negativo es que la energía cinética del láser puede convertirse en energía vibracional que supere la energía de enlace y formar fragmentos, lo que dificulta la señal de un pico ion molecular cuando se pretende enlazar la información para secuencias peptídicas. Además, la ionización de los distintos péptidos puede variar en varios ordenes de magnitud.

En este contexto, la técnica de ionización MALDI emerge como la preferida para la detección de proteínas debido a sus destacadas sensibilidad, resolución y especificidad, especialmente en la identificación de proteínas hasta 10,000 m/z, se destaca por emplear matrices de ionización de bajo peso molecular compatibles con analitos termolábiles y macromoléculas en general. La preparación de la muestra para su análisis en el espectrómetro es rápida y sencilla, requiriendo de una cantidad mínima de muestra, además es tolerante a ciertas impurezas como sales que pueden estar presentes en la muestra biológica y la calidad de los resultados no se ve comprometida ya que MALDI-TOF-MS puede alcanzar intensidades de señal de hasta 10<sup>9</sup> órdenes de magnitud lo que garantiza una excelente relación señal-ruido.

Además, MALDI-TOF se destaca por la velocidad con la que se revela el espectro, lo que agiliza el proceso de análisis y permite obtener resultados de manera eficiente. Estas características hacen de que sea tenida en cuenta dentro de las primeras opciones como técnica para análisis de muestras biológicas como es el caso de este trabajo.

**8.6 Metodología para preparación de muestras asistida por filtro (FASP).**

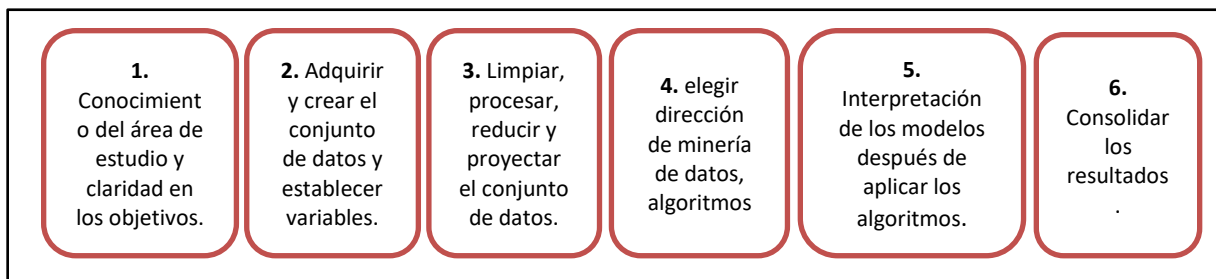
La caracterización de péptidos en MS depende en gran medida del primer paso de cualquier análisis como la

naturaleza, origen y preparación de la muestra [63], que es, por lo general, el que más tiempo consume y el más propenso a errores. Una preparación adecuada de la muestra facilita la fragmentación de las proteínas, condición que suele ser utilizada en MS. Separar las proteínas en varias fracciones puede mejorar la profundidad analítica al reducir la complejidad de la muestra [64].

El desarrollo de FASP [65] combina las ventajas de las estrategias de digestión en gel y en solución [66], [67] que consiste en realizar los pasos regulares de una digestión enzimática (desnaturalización, reducción, alquilación y tratamiento con enzima) de manera seguida sin tener que trasladar fracciones después de los lavados para eliminar impurezas, sales, detergentes, grasas, entre otras. Para este método se utilizan los filtros Amicon que contienen una membrana de corte de peso molecular (MWCO), por sus siglas en inglés, de diferentes tamaños [68], [69], que utilizan la fuerza centrífuga para separar dos fracciones: sobrenadante que contiene la porción importante o de interés y filtrado que contendría la parte a descartar, sin embargo, las membranas de Amicon contienen cierto porcentaje de retención por lo que, en algunos casos, es pertinente estudiar ambas fracciones. Cabe señalar que uno de los factores que pueden influir es la concentración de la muestra, así como la enzima proteolítica. En este caso se empleó la tripsina que es la más utilizada en investigaciones, es un enzima peptidasa que hidroliza el extremo C a los residuos de Lys y Arg [70], esto la hace muy específica en las condiciones óptimas.

## 8.7 Ciencia de datos.

La complejidad de los datos de MS y no solo de esta técnica sino de todas las formas de obtener información biológica hizo que los científicos apelaran a refuerzos informáticos. El término es la bioinformática que según el NIH “es una subdisciplina científica que recopila, almacena, analiza y discierne información biológica” [71]. En el desarrollo de la bioinformática se hace uso de la ciencia de datos que se caracteriza por manipular grandes cantidades de datos que es algo común en la investigación científica [72]. Hoy en día el poder de la CD se basa en aprovechar los datos obtenidos para extraerles la mayor información posible e inferir conclusiones que a simple vista un humano no hubiera podido con la exactitud y precisión en un tiempo razonable. El camino recomendado por García [73] para hacer uso de la ciencia de datos se muestra en la **Figura 5**.

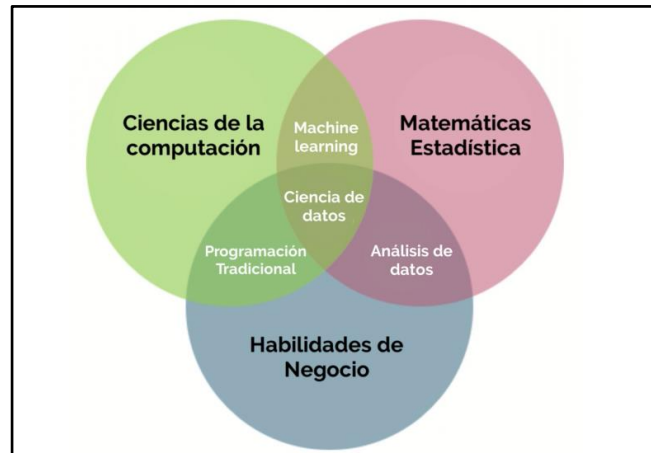


Obtenido y adaptado de García [73].  
**Figura 5.** Etapas de trabajo en ciencia de datos.

### 8.7.1 Machine Learning y manejo de datos MALDI-TOF-MS.

La interdisciplinariedad de la CD hace que otras ciencias como las exactas y biomédicas se empapen de dicha rama. Una subdisciplina de la CD es el ML o aprendizaje automatizado, ver **Figura 6** que es definido, por ejemplo, por Samuel citado por Ramírez en el 2021 [74] como la capacidad de indicarle a un computador cómo aprender sin ser programados explícitamente. El ML brinda técnicas para crear relaciones, reglas, patrones, o resúmenes que sean útiles para el conocimiento. Dicha actividad se conoce como modelado y pueden ser de dos tipos: modelos descriptivos y modelos predictivos [73], los primeros solo pretenden formar un conjunto de datos más conocido como bases de datos (BD) y describirlo como su nombre lo dice; en cambio los predictivos pretenden usar dichos conjuntos de datos para aproximar posibles valores futuros. Para ello existen tareas que luego pueden ser representadas por funciones como las siguientes:

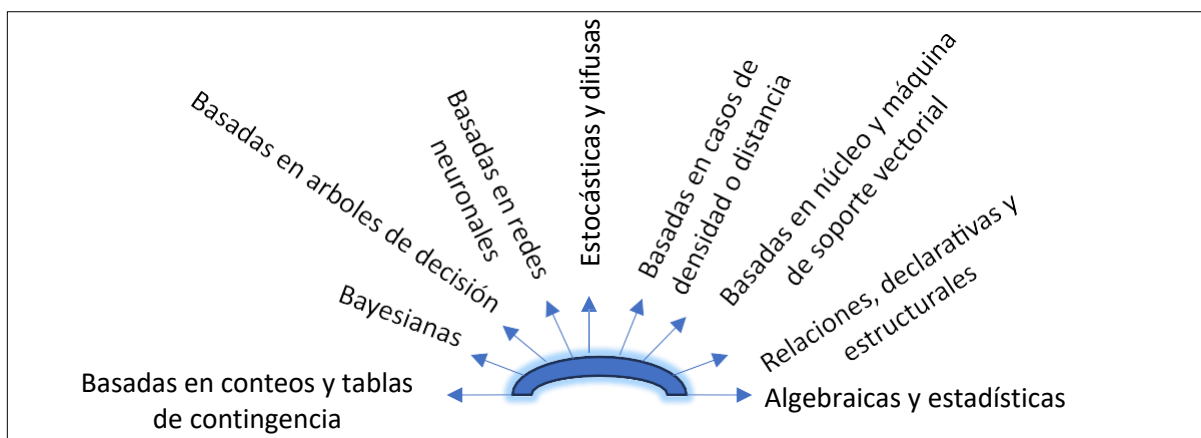
- Descriptivas (Agrupamiento, correlaciones y factorizaciones, asociación y dependencias funcionales)
- Predictivas (Clasificación, categorización, priorización, correlaciones y regresiones).



Tomado y adaptado de Méndez [75].

**Figura 6.** Diagrama de Venn de las disciplinas de la ciencia de datos, entre ellas el Machine Learning.

Existen muchas técnicas y métodos para realizar dichas tareas [76], entre ellas está el análisis estadístico descriptivo típico como mediana, moda, media, percentiles, tipificado o normalizados, valores atípicos, etc, que suele ser el primero que se realiza y luego viene lo que se conoce como técnicas de minado de datos o Data Mining (MD) [77], como se muestra en la **Figura 7** [73]. El MD tiene como objetivo transformar datos en conocimiento útil. En este sentido, Es indispensable conocer se necesita un lenguaje que ayude a manipularlos de manera más eficiente. Existen muchos lenguajes de programación que se usan con tal fin, entre ellos se encuentra Python, R, C++, JAVA entre otros [78],[79]. Hoy en día Python ha tomado ventaja por su facilidad de términos, creación de algoritmos con parámetros brevemente explicados y la interacción con diferentes librerías que proporcionan las herramientas y funciones para el desarrollo de una serie de instrucciones (programa/código) sea más sencillo. Por otro lado, existen técnicas de MD en relación con la IA que se subdividen en dos categorías, supervisados o predictivos y no supervisados o descriptivos donde las diferencias se muestran en la **Tabla 4**.



**Figura 7.** Técnicas de minado de datos.

**Tabla 4.** Comparación de los enfoques del aprendizaje automatizado: aprendizaje supervisado y no supervisado.

Enfoques de la inteligencia artificial—Machine Learning-minado de datos	
Aprendizaje supervisado	Aprendizaje no supervisado
Es un enfoque del ML que usa conjuntos de datos etiquetados	Es un enfoque de ML que usa conjuntos de datos sin etiquetar
El conjunto de datos se usa con el fin de entrenar o supervisar algoritmos para posteriormente predecir	El conjunto de datos se usa con el fin de descubrir patrones ocultos en los datos sin necesidad de intervención humana o sin supervisión
En minería de datos se enfrenta a la tarea de clasificación y regresión o correlación (para cada tarea hay muchos algoritmos que se pueden aplicar)	En minería de datos se enfrenta a las tareas de agrupación, asociación, reducción de dimensionalidad
Los datos de entrada y salida se encuentran etiquetados	Los datos de entrada y salida no están etiquetados
Los modelos con este enfoque tienden a ser más precisos que los no supervisados	Aunque sean no supervisados aun cuentan con cierta intervención humana para validar las variables de salida
Los modelos son más simples computacionalmente.	Los modelos son computacionalmente complejos por que necesitan muchos datos para entrenar
Pueden llevar mucho tiempo y experiencia en las variables de entrada y salida	Pueden tener resultados inexactos a menos de que se evalúe humanamente las variables de salida

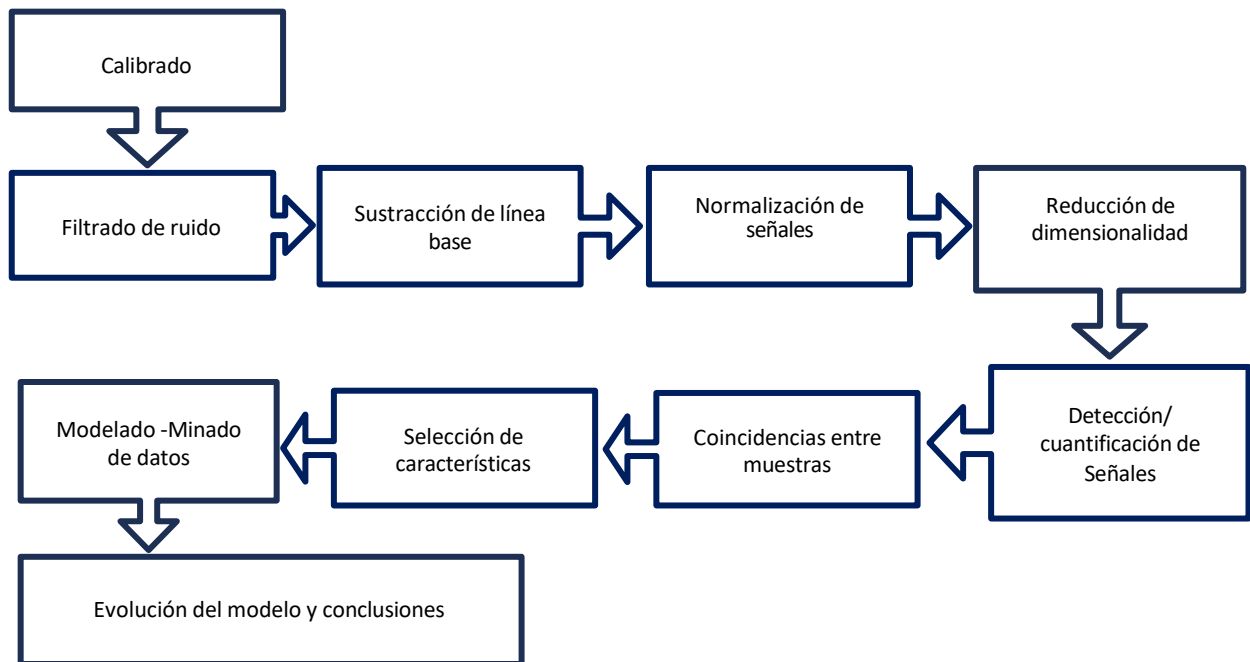
Obtenido y adaptado de Delua [80].

Existen varias formas de evaluar el rendimiento de un modelo predictivo pero lo principal consiste en dividir en primera instancia los datos en dos conjuntos. Entrenamiento y prueba o evaluación. De los cuales se pueden inferir parámetros de evaluación como la exactitud, precisión entre otras.

Por último, específicamente datos como los que se confieren a espectros de masas tienen una serie de limitaciones como lo es la cantidad de señales asociadas a fragmentos peptídicos; el ruido asociado a impurezas e incluso hablando de MALDI-TOF-MS la matriz estaría afectando también. Sin contar las modificaciones post-transduccionales o todas las posibles reacciones que sufren las proteínas antes de ser analizadas. Algunas veces

el número de datos suele ser muy grande para trabajarlos en un computador corriente limitando el almacenamiento del dispositivo por lo que muchas veces se prefiere reducir la cantidad de datos que es algo importante para sacar conclusiones representativas. Sabiendo estas limitaciones es donde puede entrar a trabajar lo anterior mencionado de CD-ML-MD con Python en donde se pueden manejar muchos datos en poco tiempo.

Un espectro de masas está compuesto por señales de  $m/z$  en el eje x y cada punto tiene su imagen en eje y representada por la intensidad que prácticamente se asocia con la abundancia de los iones detectados, la altura de las señales puede dar un indicio de selección de características o patrones en el caso proteómico ayuda para encontrar posibles biomarcadores, entendidos en el ámbito de péptidos relacionados con la enfermedad de la PE [81]. El manejo de datos de un espectro de masas consiste en exportarlos en otro formato por ejemplo en texto (.txt) donde se obtienen los datos en texto plano de cada punto que conforma el espectro, así que para cada valor de  $m/z$  no se obtendrá solo un valor de abundancia relativa de un ion, si no una serie creciente de puntos para formar la señal determinada. Así pueden manipularse cada variable ya sea en relación o por separado. Para realizar un modelo predictivo usando datos de espectros de masas se sigue la siguiente secuencia de trabajo ver **Figura 8**. Después de la evaluación y conclusiones y dependiendo de la trascendencia de la investigación, se debería hacer la respectiva validación médica.



Obtenido y adaptado de Thomas [82] y Coombes [83].

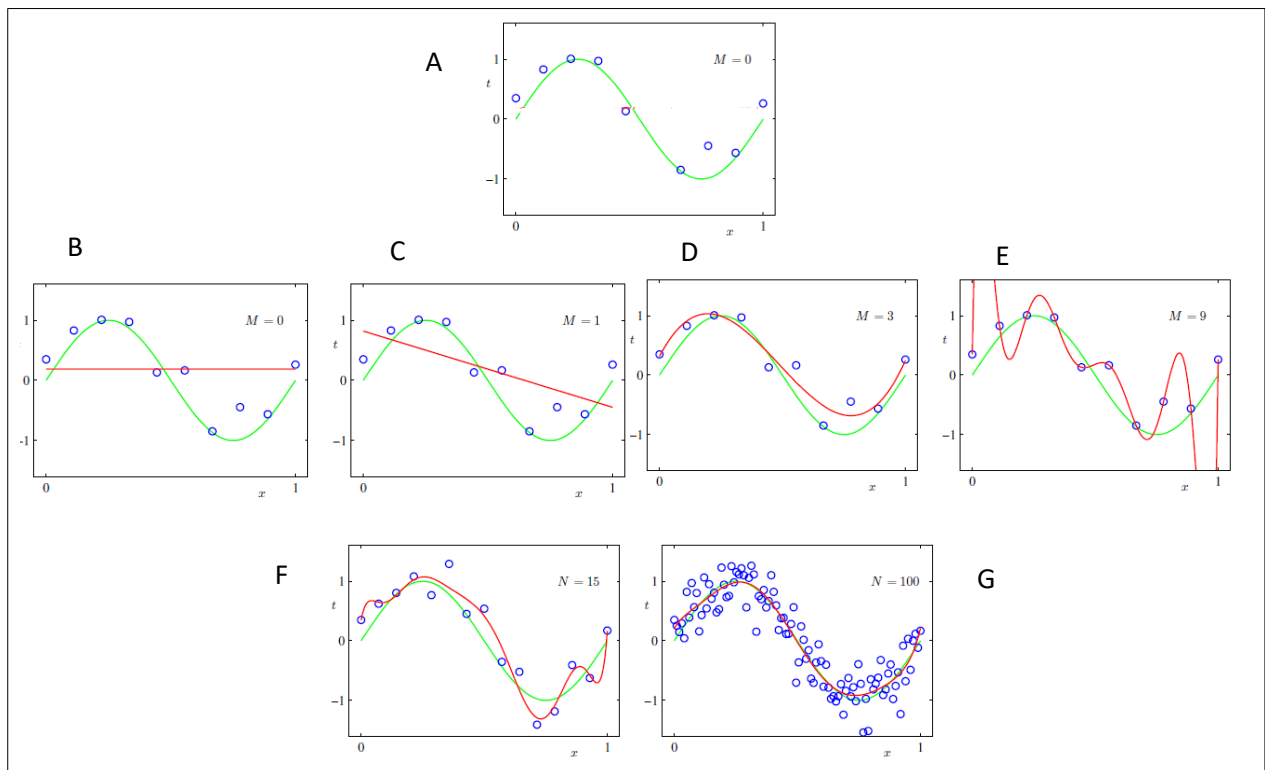
**Figura 8.** Secuencia de procesado de datos de espectros de masas para en modelo predictivo.

### 8.8 Modelos de aprendizaje utilizados.

Cuando se dispone de un conjunto de datos, resulta crucial aprovecharlos para extraer información relevante de las observaciones realizadas en su momento. Es importante reconocer que algunos conjuntos de datos son más

complejos que otros, lo que implica que los métodos y enfoques utilizados para su análisis y tratamiento también deben ser más sofisticados.

Una tarea común en el análisis de datos es modelar el comportamiento inherente a los mismos. En otras palabras, se trata de aplicar y ajustar parámetros de funciones matemáticas de forma que estas se ajusten de manera precisa a las observaciones recopiladas. El objetivo final de este proceso es que, cuando se presente una nueva observación o se agregue un nuevo dato al conjunto el modelo sea capaz de predecir su comportamiento con facilidad y precisión. Si lo vemos gráficamente, con un ejemplo en la **Figura 9** supongamos que se recolectaron datos (representados por círculos azules) que relacionan las variables tiempo ( $t$ ) y distancia ( $x$ ), la curva verde representaría el modelo ideal al que se espera llegar con la aplicación de una función matemática, la cual tiene una forma de seno. Para llegar a este podríamos ensayar diferentes polinomios de diferentes órdenes (curvas de color rojo) y mirar si se ajustan a los datos, ese ajuste de parámetros puede interpretarse como un entrenamiento del modelo, donde se están dando las características de los datos a una función para que esta aprenda. Por su puesto, entre más datos se suministren el entrenamiento será más preciso hasta llegar a un modelo casi exacto, ya que el modelo ideal no existe debido a que por lo general los datos se adquieren experimentalmente casi siempre con intervención humana, el cual cuenta con un error inherente imposible de ignorar.



A: datos, B-E: aplicación de funciones polinomiales de diferentes ordenes o modelos aplicados, el que mejor se ajusta a los datos de ejemplo es D, F-G: comparación de mejor ajuste respecto a la cantidad de datos.

Tomado de Bishop [84].

**Figura 9.** Ejemplificación gráfica de un modelo predictivo.

Ahora bien, con esta breve explicación sobre cómo funcionan los modelos de aprendizaje o entrenamiento, se puede pensar en la relación de etiquetas-datos en la diferencia entre el entrenamiento supervisado y no supervisado. En el primero como su nombre lo indica se usan datos etiquetados donde ya se conoce la salida deseada o clase a la que pertenece, sirve para realizar predicciones; en caso contrario, en los modelos no supervisados no se cuentan con las etiquetas y el objetivo es encontrar patrones o estructuras subyacentes en los datos sin ninguna guía externa. Ambos pueden usarse con fines de comparación.

Los modelos de aprendizaje y entrenamiento que se utilizaron en esta investigación fueron: análisis de componentes principales (PCA) por siglas en inglés, regresión logística (RL), *random forest* (RF) y refuerzo extremo de gradiente (XGBOOST) por sus siglas en inglés.

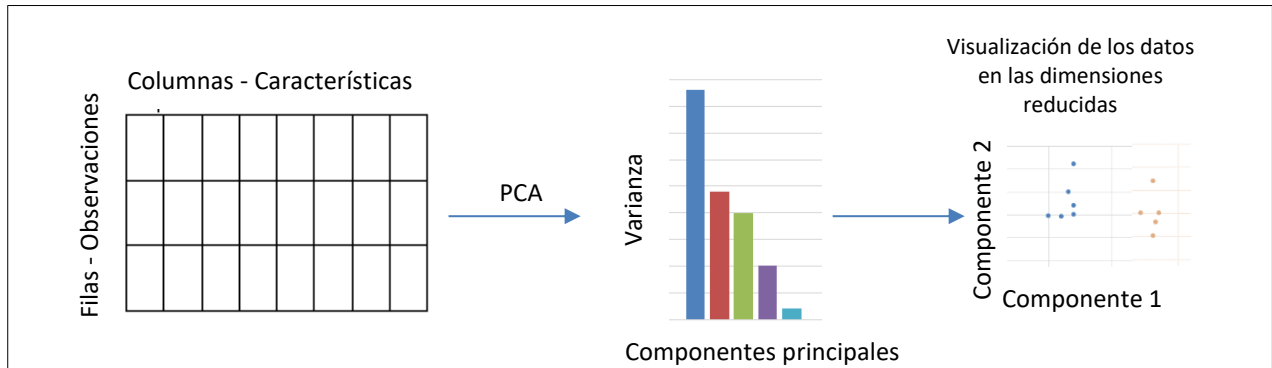
### 8.8.1 Técnica de análisis de componentes principales, modelo no supervisado.

El análisis de componentes principales o más conocido como PCA es una técnica de reducción de dimensionalidad y permite representar el conjunto de datos en un espacio de menor dimensión, manteniendo la mayor parte de la variabilidad presente en los datos originales, es útil para aplicar a datos que tengan muchas características o variables y hagan difícil la visualización de una separación en los datos originales, es ahí donde la función de PCA proyecta los datos en un nuevo conjunto de dimensiones capturando la mayor parte de variabilidad, a los cuales se les llama componentes principales y están ordenados por la cantidad de variabilidad que explican lo que permite seleccionar las dimensiones más importantes para el análisis de los datos. [84]

En la **Figura 10** por ejemplo, si se tiene una tabla que agrupa los datos (*Data set*) que tiene las columnas y filas, estas representan las características y observaciones respectivamente, al aplicar el PCA se pueden visualizar los datos en otra dimensión y con suerte ver una separación.

El PCA se basa en la varianza explicada, que indica cuánta información se puede atribuir a cada uno de los componentes principales, si bien la varianza total o de un 100% sería ideal, sin embargo, esta se explica dejando el mayor número de componentes principales y se logra teniendo en cuenta cada una de las características del conjunto de datos. Por defecto no es objetivo maximizar el número de componentes hasta el número de características, al contrario, se pretende disminuir las dimensionalidades y que la varianza sea la más alta posible. Esto implica entonces que a medida que se reducen los componentes se pierde información o en otras palabras varianza [85].





**Figura 10.** Explicación esquemática de método de aprendizaje no supervisado de análisis de componentes principales PCA.

Por otro lado, los modelos de aprendizaje supervisado pueden tener la función de clasificación o regresiones, por supuesto teniendo en cuenta las etiquetas, que pueden comprender 2 (binaria) o más (multiclase) tipos de etiquetas o clases.

### 8.8.2 Regresión logística, modelo supervisado.

La RL que es un modelo lineal de clasificación, que se usa para predecir la probabilidad de pertenencia a una clase binaria; en este caso consiste en aplicar la función logística quien modela la relación entre variables. La función logística también es conocida como función sigmoide que tiene la forma de la **Ecuación 6** [86].

$$f(x) = \frac{1}{1 + e^{-x}}$$

**Ecuación 6.** Ecuación sigmoide o función logística general.

Lo particular de esta ecuación es que recibe cualquier valor  $x$  y lo transforma a un valor entre 0 y 1; por eso es útil para aplicarla en contexto donde se tengan clasificaciones binarias o de dos clases. En este trabajo se aplica ya que se tienen 2 tipos de datos, casos y controles respecto a la PE. La predicción se basa en que se pueda interpretar el resultado como la probabilidad de que un punto o un dato se encuentre en alguna de las clases. En términos de probabilidad se ve en la **Ecuación 7** donde  $p$  representa la probabilidad de que la variable dependiente y sea máxima dado un valor para la variable independiente  $x$ , los exponentes  $bnxn$  son coeficientes que estiman el modelo y dependen del número de datos que se tengan.

$$p(y = 1|x) = \frac{1}{1 + e^{-(b_0 + b_1x_1 + \dots + b_nx_n)}}$$

**Ecuación 7.** Ecuación de regresión logística.

### 8.8.3 Máquina de vectores de soporte, modelo supervisado.

El modelo de SVM usa tanto la clasificación como la regresión para buscar y encontrar el hiperplano óptimo que maximice la distancia entre puntos más cercanos de las diferentes clases que mejor las separe en un espacio multidimensional. Es de gran utilidad cuando los datos no son lineales, además puede usar diferentes *kernel* que son funciones matemáticas que se aplican para mapear los datos en un espacio de características de mayor dimensión y llevarlos a otra dimensión transformándolos por ejemplo a planos (polinomios) o superficies (*kernel* radial); buscando la distancia más grande entre el hiperplano y los puntos de datos más cercanos a cada dato; estos son conocidos como vectores de soporte y son importante para definir el hiperplano. este último suele mostrar imágenes de superficies extrañas cuando los parámetros no son bien ajustados o simplemente porque a partir de la tercera dimensión la percepción grafica se pierde.

SVM se basa en una función de decisión que en términos generales tiene la forma de la **Ecuación 8** donde  $W$  es el vector de pesos normal al hiperplano de separación,  $X$  es el vector de características del punto de datos y  $b$  es el termino de sesgo que desplaza el hiper plano separación desde el origen hacia el lado del hiperplano que caiga el punto  $x$  esto en caso de que sea un mapeo lineal. Cuando se decide cambiar de mapeo, por ejemplo, a radial o gaussiano la fórmula de decisión cambia a la **Ecuación 9** donde:  $\alpha_i$  son los multiplicadores de Lagrange obtenidos durante el proceso de entrenamiento,  $y_i$  son las etiquetas de clase correspondientes a los vectores de soporte  $x_i$  y  $K(x_i, x)$  es la función de *kernel*, ver **Ecuación 10** radial donde  $r$  es un parámetro ajustable [87].

$$f(x) = \pm(W \cdot X + b)$$

**Ecuación 8.** Función de decisión en general.

$$f(x) = \pm \left( \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right)$$

**Ecuación 9.** Función de decisión para el modelo SVM.

$$K(x_i, x) = \exp(-r \|x_i, x\|^2)$$

**Ecuación 10.** Función de *kernel* radial para el modelo SVM.

### 8.8.4 Árboles aleatorios o Random Forest, modelo supervisado.

Ahora, RF es un modelo que es basado en árboles de decisión para la clasificación que son algoritmos mucho más complejos en comparación con RL o SVM. Los árboles de decisión son estructuras de tipo árbol, ver

**Figura 11** donde en cada nodo interno se representa una característica y las ramas representan la decisión basada en esas características y la hoja representaría el resultado. Lo relevante es que las características y datos

los escoge aleatoriamente para garantizar que se puedan crear diferentes arboles individuales que estén menos correlacionados favoreciendo el hallazgo de diferencias o anomalías en el conjunto de datos sobre todo cuando tienen muchas características y una alta dimensionalidad para el resultado de la clasificación.

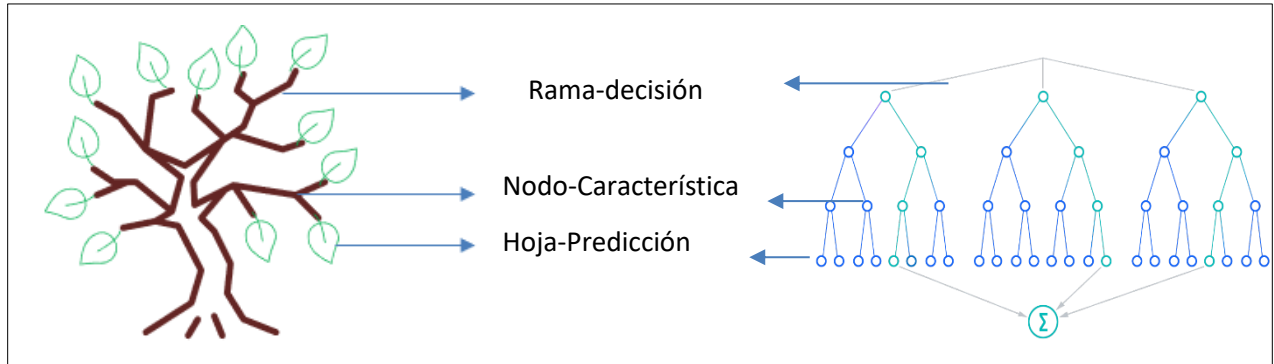
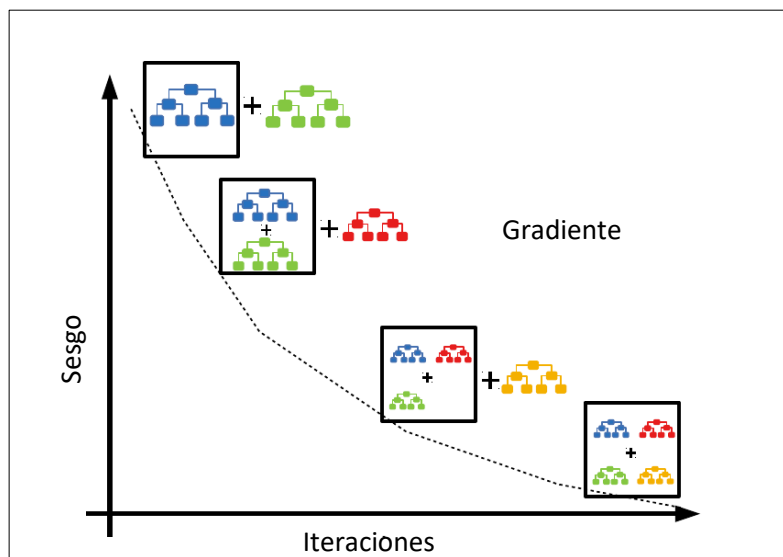


Figura 11. Comparación explicativa del modelo random forest.

### 8.8.5 Refuerzo extremo de gradiente, modelo supervisado.

Otro modelo basado en Árboles de decisión es el XGBOOST, ver **Figura 12** que entrena varios árboles en secuencia tratando de corregir errores del modelo anterior, de esta forma cada árbol resultante se especializa en predecir que los modelos anteriores no pudieron resolver aplicando una función de pérdida haciéndola cada vez más pequeña representado por un gradiente. Este es útil para manejar conjuntos de datos desbalanceados y de gran cantidad [88].



Tomado y adaptado de Pal [88].

Figura 12. Representación explicativa del modelo refuerzo extremo de gradiente.

La diferencia entre RF y XGBOOST radica en la manera como ensamblan los árboles de decisión, RF construye varios árboles independientes y combina la predicción de cada uno de ellos promediando el resultado, a este tipo de ensamblaje se le conoce como “*bagging*” en cambio XGBOOST utiliza el ensamblaje “*boosting*” que es construir en secuencia los árboles de decisión tratando de corregir los errores siempre del árbol anterior [88].

### 8.9 Matrices de confusión.

Luego están las matrices de confusión que son modelos, pero son herramientas que ayudan a evaluar el rendimiento de los modelos supervisados o de clasificación. Prácticamente es una tabla o matriz que muestra la frecuencia con la que un modelo hace predicciones correctas e incorrectas respecto a cada clase que exista en el conjunto de datos. El tamaño de la matriz depende de las clases si es binaria tendrá un tamaño de 2 filas por 2 columnas y así sucesivamente. La siguiente **Figura 13** muestra la forma general para entender la matriz de confusión ante un problema de clasificación de 2 clases.

Verdaderos Controles Verdaderos Negativos Verdaderos 0	Falsos Casos Falsos positivos Falsos 1
Falsos Controles Falsos Negativos Falsos 0	Verdaderos Casos Verdaderos positivos Verdaderos 1

**Figura 13.** Representación general de una matriz de confusión para una clasificación binaria.

La matriz de confusión arroja métricas evaluativas como precisión en la **Ecuación 11**, sensibilidad en la **Ecuación 12** y puntuación F1 en la **Ecuación 13**.

$$\text{Precisión} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos Positivos}}$$

**Ecuación 11.** Precisión, métrica de evaluación por clases.

$$\text{Sensibilidad} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos Negativos}}$$

**Ecuación 12.** Sensibilidad, métrica de evaluación por clases.

$$\text{Puntuación F1} = \frac{2 \cdot \text{Precisión} \cdot \text{Sensibilidad}}{\text{precisión} + \text{Sensibilidad}}$$

**Ecuación 13.** Puntuación F1. métrica de evaluación combinada por clases.

En consecuencia, en este trabajo haremos uso de las técnicas de ML y MD con el fin de procesar espectros de masas de MALDI-TOF de S.S de pacientes casos y control de PE. El reto será entonces poder manejarlos bajo plataformas que operen el lenguaje de programación Python e interpretar los resultados del modelado para indicar características importantes en los espectros de masas obtenidos a partir de S.S de casos y controles.

## 9. METODOLOGÍA EXPERIMENTAL

### 9.1 Equipos

pH metro HANNA/Hi 2221 Calibration Check pH/ORP Meter; Incubadora/BINDER; Balanza analítica OHAUS/Pioneer; Microcentrífuga Thermo scientific IEC CL31R Multispeed; Ultrasonido Elma/ E 30H Elmasonic; Secado al vacío / Speed Vac Thermo scientific/Savant SpeedVac SPD120 Vacuum Concentrator; Mono/multi – Micropipetas 10-200-1000 uL Brand y Eppendorf; Lector de microplacas Thermo scientific/Varioskan Flash; Vortex BOECO Germany/ Vortex V1 plus; Espectrómetro de masas MALDI-TOF Bruker Daltonics/ultraflexxtreme; Cabina de extracción Esco/Frontier iso cide; Ultracongelador; Computador de mesa HP/Compaq /A2205wg; Computador portátil ASUS/ X407UA-BUV385.

### 9.2 Materiales

164 filtros Amicon de corte de peso molecular de 3kDa; 400 Tubos de microcentrífuga Eppendorf de 0,5 mL; 330 Tubos de reacción Eppendorf de 2 mL; 3700 puntas de micropipetas/Corning 10-200-1000 uL; 1 Vidrio reloj; 4 Tubos falcon de plástico graduado, 1 espátula; 2 Beacker; 1 recipiente de icopor; 4 gradillas para microtubos; 4 Microplacas de 96 pocillos transparentes a UV; 1 Frasco lavador; Papel indicador; 3 criocajasde 100 puestos.

### 9.3 Reactivos

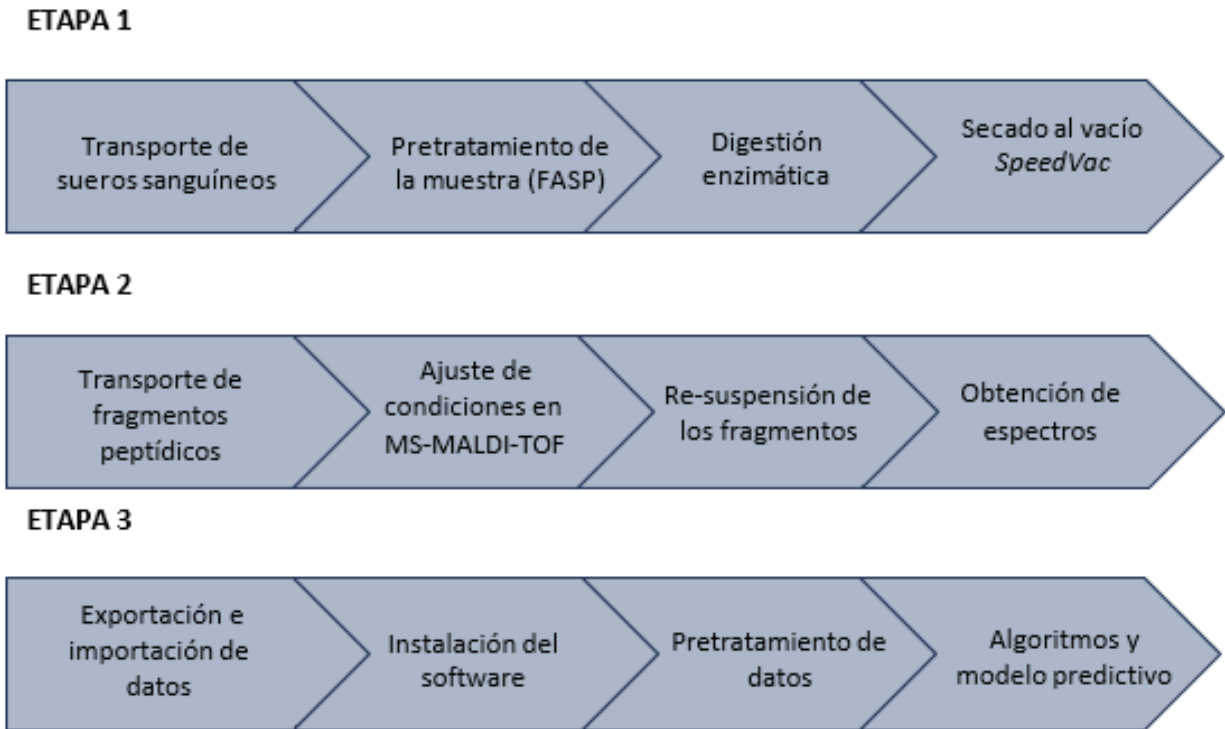
Sueros sanguíneos; Tripsina grado masas [89]; Ditiotreitól (DTT); Iodoacetamida (IAA); Ácido 2,5-dihidroxibenzoico (DBH); Trisaminometano (TRIS); Kit de BCA Reactivo A: ácido bicincónico; Reactivo B: sulfato de cobre; Patrón: suero Albumina bovino (BSA) [90]; Vapreotida grado masas; Ácido trifluoro acético (TFA); acetronitrilo (ACN), Ácido acético; Isopropanol; Ácido  $\alpha$ -ciano-4- hidroxicinámico (HCCA); agua tipo 1; Ácido sinapínico (AS).

### 9.4 Softwares y programas

Flex análisis-3.3; Minitab- 21,1; Anaconda -3.0; Google colab; Biorender.

### 9.5 Etapas generales de la metodología

La metodología se puede entender en tres etapas generales, como se muestra en la **Figura 14** la primera consistió en la adquisición de muestras y la realización de la digestión enzimática; la segunda consistió en la obtención de los espectros de masas MALDI-TOF y la última en la realización del modelo predictivo y aplicación de algoritmos.



**Figura 14.** Etapas generales de la metodología.

### 9.5.1 Etapa 1

Las 164 muestras de S.S fueron entregadas en 2 criocajas para ser transportadas desde el Biobanco de la FCV en la ciudad de Floridablanca hasta el laboratorio GIBIM en las instalaciones principales de la UIS. Debido al cuidado que deben tener las muestras biológicas, estas se rodearon de geles refrigerantes dentro de un termostato por aproximadamente 30 minutos, cabe resaltar que este fue el primer ciclo de descongelación de las muestras, las cuales fueron refrigeradas por una semana en una nevera especial para muestras biológicas a una temperatura de -50 °C.

Las muestras fueron parte del estudio colombiano de Genética y preeclampsia (GenPE) [9] que se realizó en colaboración de diferentes hospitales de Colombia durante los años 2000 al 2012 los criterios generales fueron los siguientes y el cuestionario dado por la FCV y acta de consentimiento ético se encuentran en el **anexo 0**:

1. Primigestantes.
2. Sin historia de enfermedades crónicas.
3. Menores a 26 años.
4. Casos: mujeres diagnosticadas con PE después de las 20 semanas de gestación.
5. Controles: mujeres que al término del embarazo (más de 37 semanas) no presentaran síntomas de PE.

En principio, las muestras fueron tratadas bajo la metodología FASP [65] para su pretratamiento que se esquematiza en la **Figura 15**. Se trataron 24 muestras de S.S al día ya que estas eran las posiciones totales de la ultracentrífuga. Para ello se descongelaron controladamente bajo cama de hielo, este proceso demoró 20 min.

Posteriormente, se procedió a preparar las siguientes soluciones, su respectivos cálculos y cantidades se encuentran en la **Tabla 5**.

- Solución Buffer: Tris-HCl 0,1 M a pH 8,5.
- Solución DTT: Ditiotreitól 0,05 M en UA.
- Solución UB: Buffer Tris-HCl 0,05 M a pH 8,5.
- Solución de tripsina: 20 ug/200 uL en ácido acético 0,1 M.
- Solución UA: 8 M urea en Buffer Tris-HCl 0,1 M. pH 8,5.
- Solución IAA: Iodoacetamida 0,05 M en UA.
- Solución HA: ácido acético 0,1 M en agua.

**Tabla 5.** Cálculos para la preparación se soluciones de digestión – FASP.

Cantidades utilizadas para la preparación de soluciones en la digestión FASP					
<b>Solución Buffer</b>		<b>Solución UA</b>		<b>Solución DTT</b>	
A partir de solución madre 1,5 M		C	8 M	C	0,05 M
C1	1,5 M	V	50 mL	V	1 mL
V2	14,0 mL	PM	60,06 g/mol	PM	154,25 g/mol
C2	0,1 M	g urea = C.V.PM	24,024 g	g DTT= C.V.PM	0,00771 g
V1= C2. V2/ C1	0,933 mL			<b>Uso inmediato post-preparado</b>	
<b>Solución IAA</b>		<b>Solución UB</b>		<b>Solución HA</b>	
C	0,05	A partir de solución madre 1,5 M		C	0,1 M
V	1	C1	1,5 M	Vf	30 mL
PM	184,96	V2	14,0 mL	PM	60,05 g/mol
g IAA= C.V.PM	0,009248 g	C2	0,05 M	d	1,05 g/mL
<b>Uso inmediato post-preparado</b>		V1= C2. V2/ C1	0,466 mL	V= C.Vf.PM/d	0,172 mL

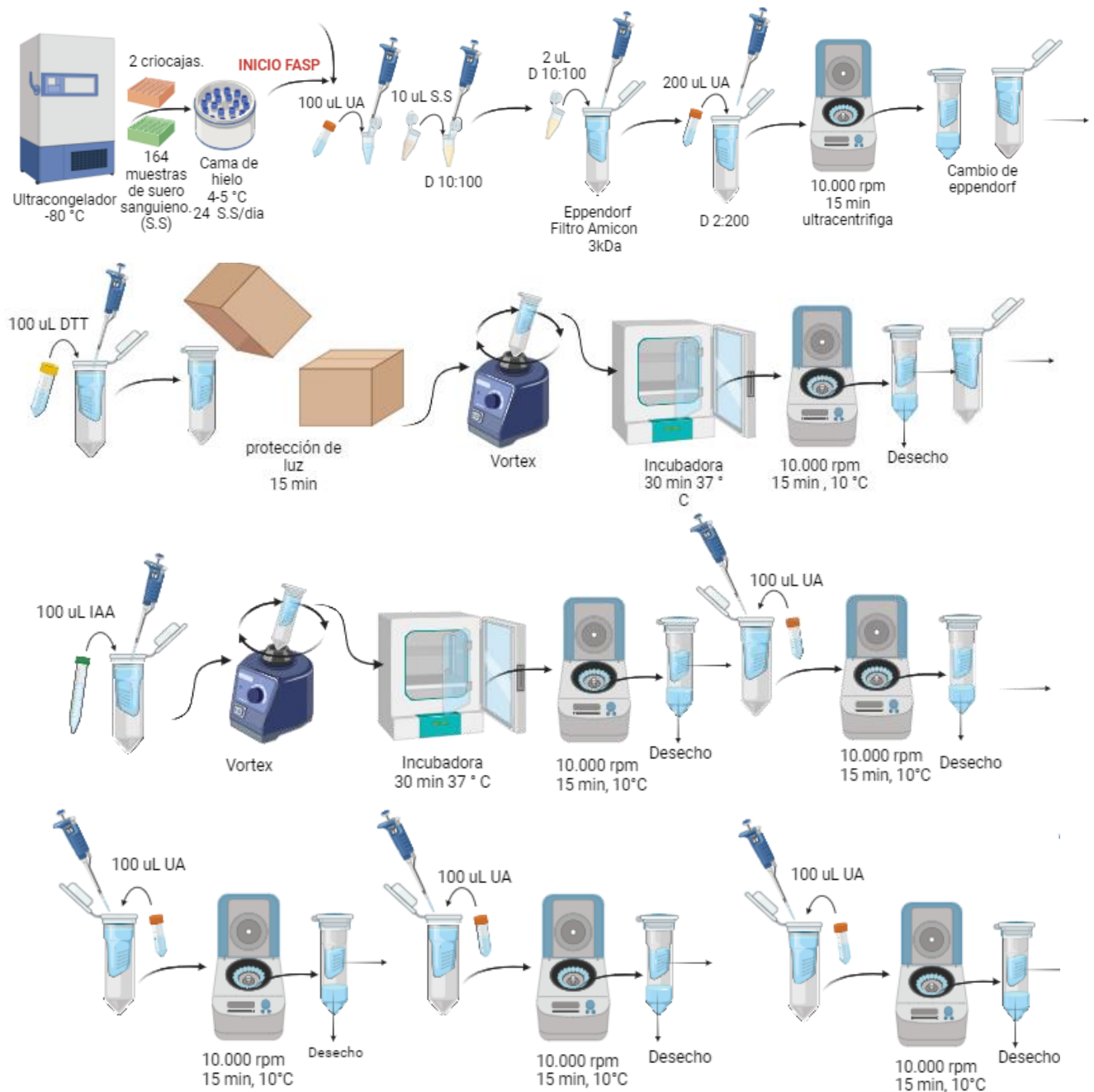
Dichas soluciones fueron utilizadas con el fin de generar fragmentos peptídicos de la siguiente manera de acuerdo con la metodología [65]:

Se alistaron los tubos Eppendorf con el filtro Amicon con tamaño de poro 3 kDa MWCO debidamente rotulados consecutivamente del 1 al 164. Primero se realizó una dilución de 10 µL de suero sanguíneo en 100 µL de buffer UA, de los cuales, solo se extrajeron 2 uL para disolverlos en 200 uL del mismo buffer, estos fueron centrifugados a 10000 rpm por 15 min y 10°C. Se obtenían filtrados al fondo del tubo los cuales eran desechados y se continuaba trabajando con el sobrenadante. El resto de la D 10:100, fue almacenada en la nevera a -50°C para realizar la cuantificación de proteínas con el Kit de BCA y las muestras originales fueron devueltas al biobanco.

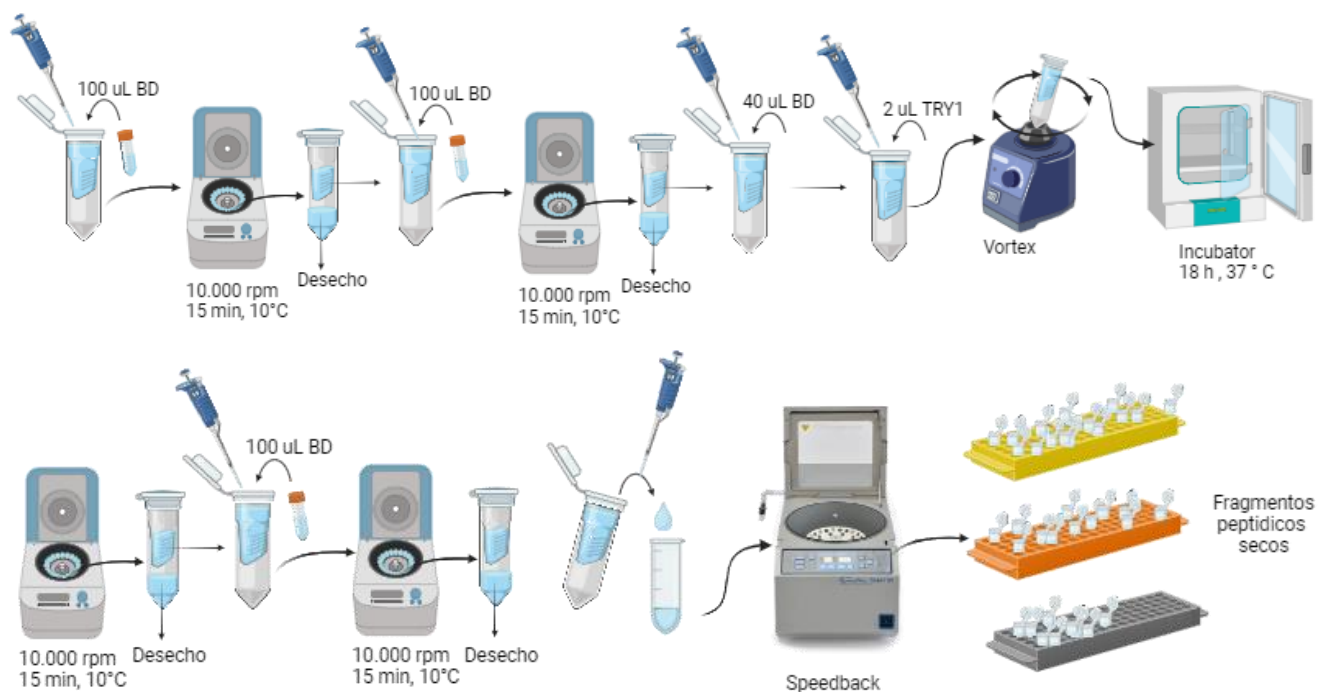
A continuación, se añadieron 100 uL de DTT sobre el filtro y las muestras se incubaron durante 30 min a 37 °C, evitando la interacción con la luz, luego fueron centrifugados a 10000 rpm por 15 min y 10°C. Posteriormente se añadieron 100 uL de IAA, repitiendo el mismo procedimiento que para el DTT. El exceso de reactivos se eliminó por centrifugación y se realizaron tres lavados con 100 uL del buffer UA y posteriormente con el buffer UB.

Enseguida, se añadieron 40  $\mu\text{L}$  de UB al filtro y 2  $\mu\text{L}$  de solución de tripsina a cada muestra, la digestión de las proteínas se llevó a cabo a  $37^\circ\text{C}$  durante 18 horas. Una vez completada la tripsinización, se hicieron lavados sucesivos con 100  $\mu\text{L}$  de la solución buffer UB usando centrifugación en las mismas condiciones mencionadas anteriormente. Vale la pena resaltar que no se registró porcentaje de rendimiento, sin embargo, valores de recuperación son dados por los fabricantes de los tubos Amicon [91]

Por último, los fragmentos de las proteínas digeridas en los sobrenadantes se recuperaron con ayuda de micropipeta y se reconcentraron utilizando el equipo *SpeedVac* a  $38^\circ\text{C}$  y se almacenaron a  $-40^\circ\text{C}$  para su posterior análisis mediante MALDI-TOF-MS. El patrón para la digestión fue el estándar de albumina del kit de BCA al cual se le realizó el mismo procedimiento de las muestras de S.S.







**Figura 15.** Flujograma de etapa 1. Digestión enzimática por metodología FASP.

## 9.5.2 Etapa 2

El transporte de los fragmentos peptídicos hacia el laboratorio LEAM se realizó de la misma manera que los S.S, demoró aproximadamente 1 h y se almacenaron en una nevera a  $-40^{\circ}\text{C}$ .

El análisis por MALDI-TOF-MS depende de muchas variables, entre ellas la naturaleza de la muestra, el tipo de matriz, la relación de concentración entre la matriz y la muestra, la cantidad de muestra y matriz que se coloca en el objetivo, la potencia y longitud de onda del láser, el método de trayectoria de los iones, el rango en el que se hace la detección, la forma y posición como co-cristaliza la muestra en la matriz y como si fuera poco siempre se debe tener en cuenta el error que puede tener el operario. Muchos trabajos de referencia pueden servir de base para formar la metodología, sin embargo, es importante ajustar las condiciones para que, por ejemplo, en un conjunto considerable de muestras como se realizó en este trabajo, la metodología sea lo más concreta posible.

## 9.5.3 Ajuste de condiciones para MALDI-TOF-MS

En cuanto al pretratamiento de los fragmentos peptídicos para MALDI, se tomó como base la experiencia en estudios llevados a cabo anteriormente en el grupo de investigación LEAM [92] y el procedimiento de general del equipo. Sin embargo, fue pertinente realizar ensayos para ajustar las condiciones y adquirir buenos espectros de

masas, teniendo en cuenta que los factores que más afectan son la concentración de material proteico, tipo de matriz a utilizar y la manera de sembrado en la placa. También se escogieron 5 muestras al azar para definir las condiciones que fueron aplicadas como se muestra en la **Tabla 6** subrayadas de color verde.

La **Tabla 7** muestra los criterios para evaluar los factores; el factor de re- suspensión se evaluó directamente agregando consecutivamente los volúmenes indicados en los niveles hasta percibir total dilución de la muestra; en cuanto a los demás factores la variable dependiente se midió en el software flexanalysis y se realizó un Diseño de experimentos (DOE) mixto 3x3x2 en Minitab con el fin de respaldar el experimento. Todo al respecto se muestra con más detalle en el **anexo 1**.

**Tabla 6.** Factores que afectan el análisis MALDI-TOF-MS

Factores que más afectan el análisis en MALDI-TOF-MS			
FACTOR	NIVELES		
Re-suspensión (uL)	10	25	50
Dilución (uL)	1:10	1:50	1:100
Matriz	HCCA	DBH	AS
Tipo de sembrado (uL)	Capa sencilla 0,5 matriz/Muestra	Doble capa 0,5 matriz/0,5 Muestra/0,5 Matriz	

**Tabla 7.** Variables de respuesta de los factores que afectan el análisis MS-MALDI-TOF.

Variables por evaluar asociadas los factores que afectan análisis en MALDI-TOF-MS	
FACTOR	VARIABLES PARA EVALUAR
Re-suspensión (uL)	Suficiencia de volumen
Dilución (uL)	Adquirir un espectro de calidad que tenga el mayor número de señales de masas a una intensidad mínima de 500 unidades dentro del rango de relación m/z (500-6000)
Matriz	
Tipo de sembrado (uL)	

Bajo las condiciones ajustadas, se procedió a re-suspender el total de las muestras y un patrón peptídico denominado Vapreotida para el análisis MALDI-TOF-MS. Adquiriendo espectros en el rango de 500-6000 m/z, con un láser de potencia de 100% y frecuencia 500 Hz. Establecidas en un método por defecto en el equipo denominado “*Micro Bio Tools*” (MBT.par) para detectar microorganismos. Hasta este momento se trabajaron las muestras sin conocimiento de sus etiquetas. En esta etapa las soluciones en las que se trabajaron los péptidos fueron las siguientes:

-Solución HCCA: 20 mg /mL en TA30.

-Solución DBH: 20 mg /mL en TA30.

-Solución AS-ETOH: 20 mg /mL en Etanol.

-Solución AS-TA30: 20 mg /mL en TA30.

Todas las soluciones de las Matrices eran ultra sonicadas antes de su uso por 15 min.

-Solución TA30: (30:70) %v/v de acetonitrilo: ácido trifluoro acético 0,1% en agua. El cálculo se muestra en la **Ecuación 14**.

Para preparar 50 mL de TA30

ACN= 30 %x50 mL = 15 mL de ACN

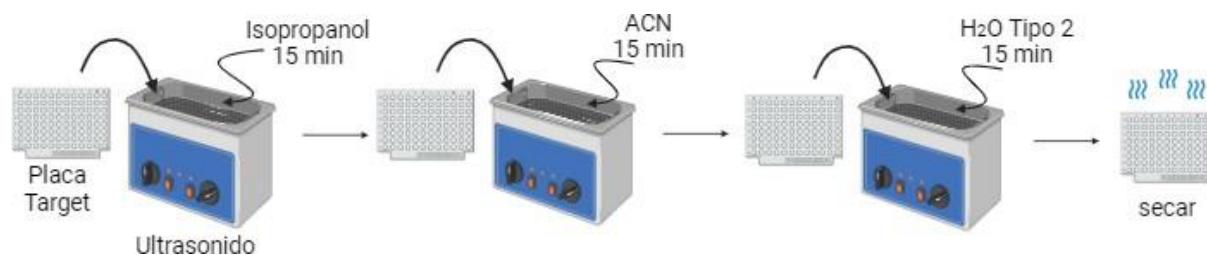
H<sub>2</sub>O (ácida)= 70 %x50 mL = 35 mL de H<sub>2</sub>O (ácida)

se partió de TFA fumante 99%

$V_1 = C_2.V_2/C_1 = 0,1\% \cdot 35 \text{ mL}/99\% = 35,35 \text{ uL}$

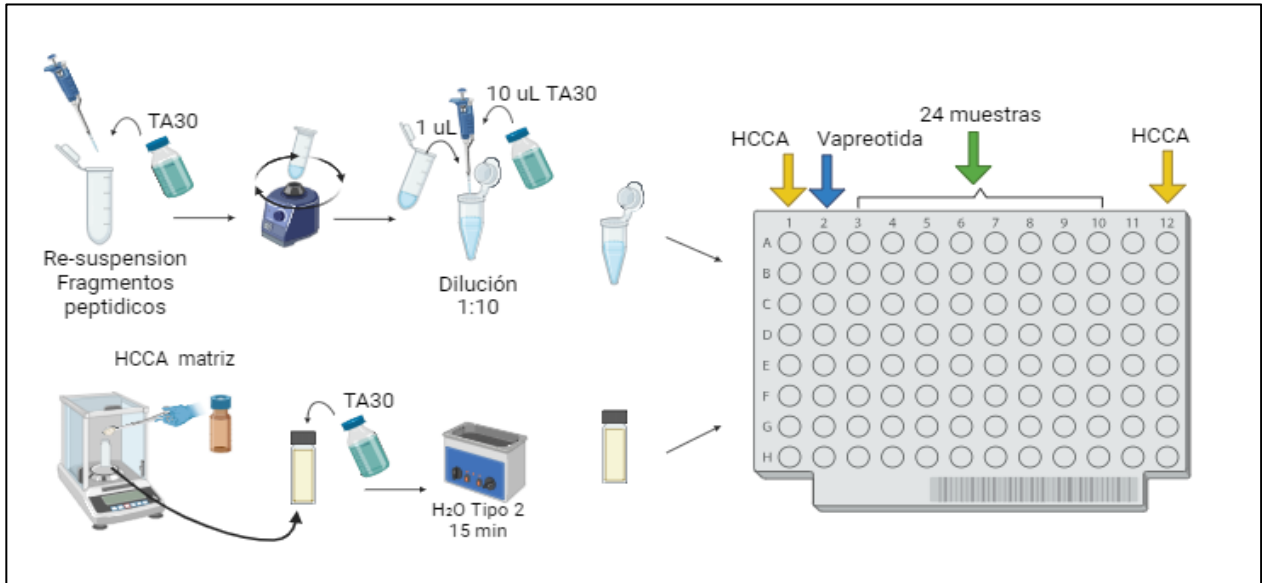
**Ecuación 14.** Cálculo de solución TA30 para re-suspensión de los fragmentos peptídicos para MALDI-TOF-MS.

Además, se siguió el siguiente procedimiento para asegurar la limpieza de la placa Target como se esquematiza en la **Figura 16**.

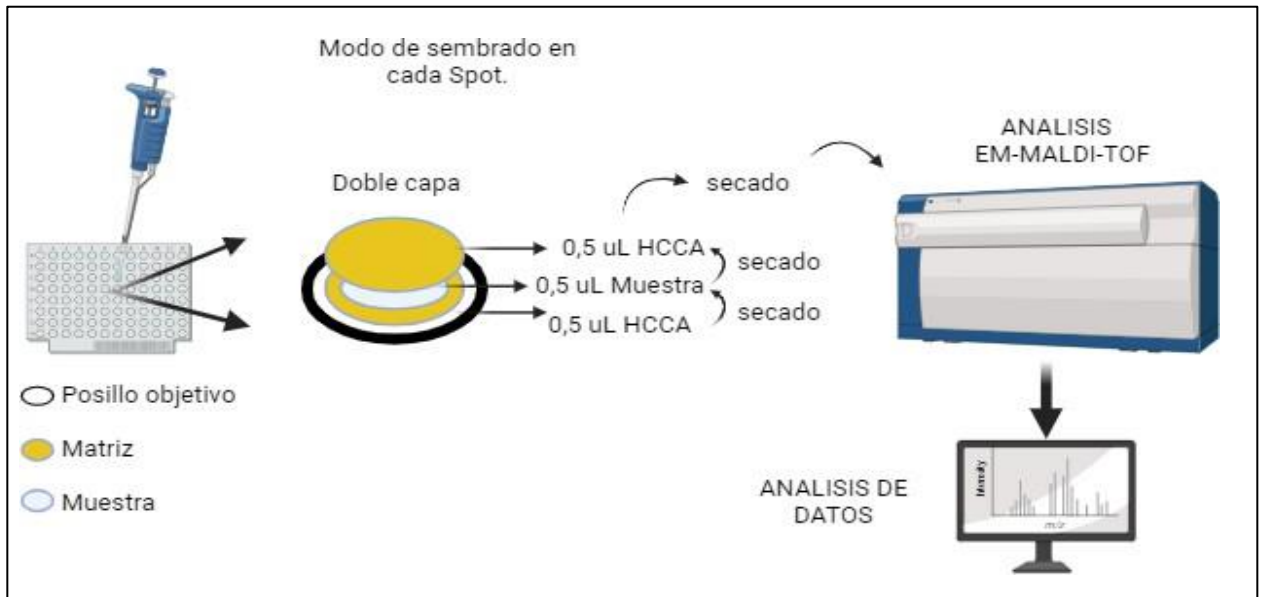


**Figura 16.** Metodología limpieza de placa Target MALDI-TOF-MS

Mientras tanto, se procedió a trabajar las muestras de fragmentos Peptídicos, a estas se le agregaron 50 mL de TA30 y agitadas en un Vortex para garantizar su re-suspensión, luego fueron rotulados respectivamente tubos Eppendorf para una dilución 1 uL:10 uL en TA30 el cual fue utilizado para colocarla en un punto o spot de la placa en forma de doble capa con un volumen 0,5 uL cada capa de matriz y muestra, dejando una pausa de secado entre capa y capa. Esta metodología fue adaptada del protocolo general de muestras dada por Bruker casa comercial del equipo MALDI-TOF-MS. Cada día se analizaron 24 muestras, matriz al inicio y final de la plantación para monitorear la ausencia de impurezas, y patrón Vapreotida. Lo anterior descrito se esquematiza en la **Figura 17** y **Figura 18**



**Figura 17.** Preparación de muestras de fragmentos peptídicos para el análisis MALDI-TOF-MS.



**Figura 18.** Modo de sembrado o deposición de la mezcla muestra-matriz en la placa.

### 9.5.4 Etapa 3

Se estableció el conjunto de los 164 espectros de calidad, *Data set*, los cuales fueron pretratados en el *software Flexanalysis* que consistió en aplicar la línea base, suavizado y exportación en texto plano en formato .txt; también se creó un documento Excel asignando el orden de los espectros y empalme con las etiquetas como SI y NO para casos y controles respectivamente según los registros del Biobanco y así poder leerlos con el *software Anaconda* en un cuadernillo de *Jupyter notebook* en algunas ocasiones en *Google Colab* con lenguaje Python y continuar con su pretratamiento, para ello se utilizaron las librerías que se muestran en **Tabla 8**, y consistió en relacionar cada punto de abundancias relativas de cada espectro en una lista o también llamada *DataFrame* procediendo con la aplicación de métodos de dichas librerías para escalar, normalizar y realizar comparaciones de acuerdo a la variable de abundancias relativas de los casos y controles. Por último, se aplicó PCA, como modelo no supervisado, y como supervisados RL, SVM, RF y XGBOOST con un porcentaje de 80% de entrenamiento y 20 % de prueba para la evaluación de los mismo, además aprovechando la versatilidad de RF y XGBOOST se aplicaron algoritmos de selección de características.

Lo anterior se resume en **Figura 19** y todo el cuadernillo que contiene el código del modelo se puede visualizar en el **anexo 3**.

**Tabla 8.** Librerías utilizadas en el cuadernillo para el modelo predictivo PE.

Biblioteca	Función
Pandas	Manipulación y análisis de datos
Numpy	Arreglos de los datos como matrices
Matplotlib.pyplot	Visualización de datos en gráficas
Scikit learn	Minería de datos y aprendizaje automatizado

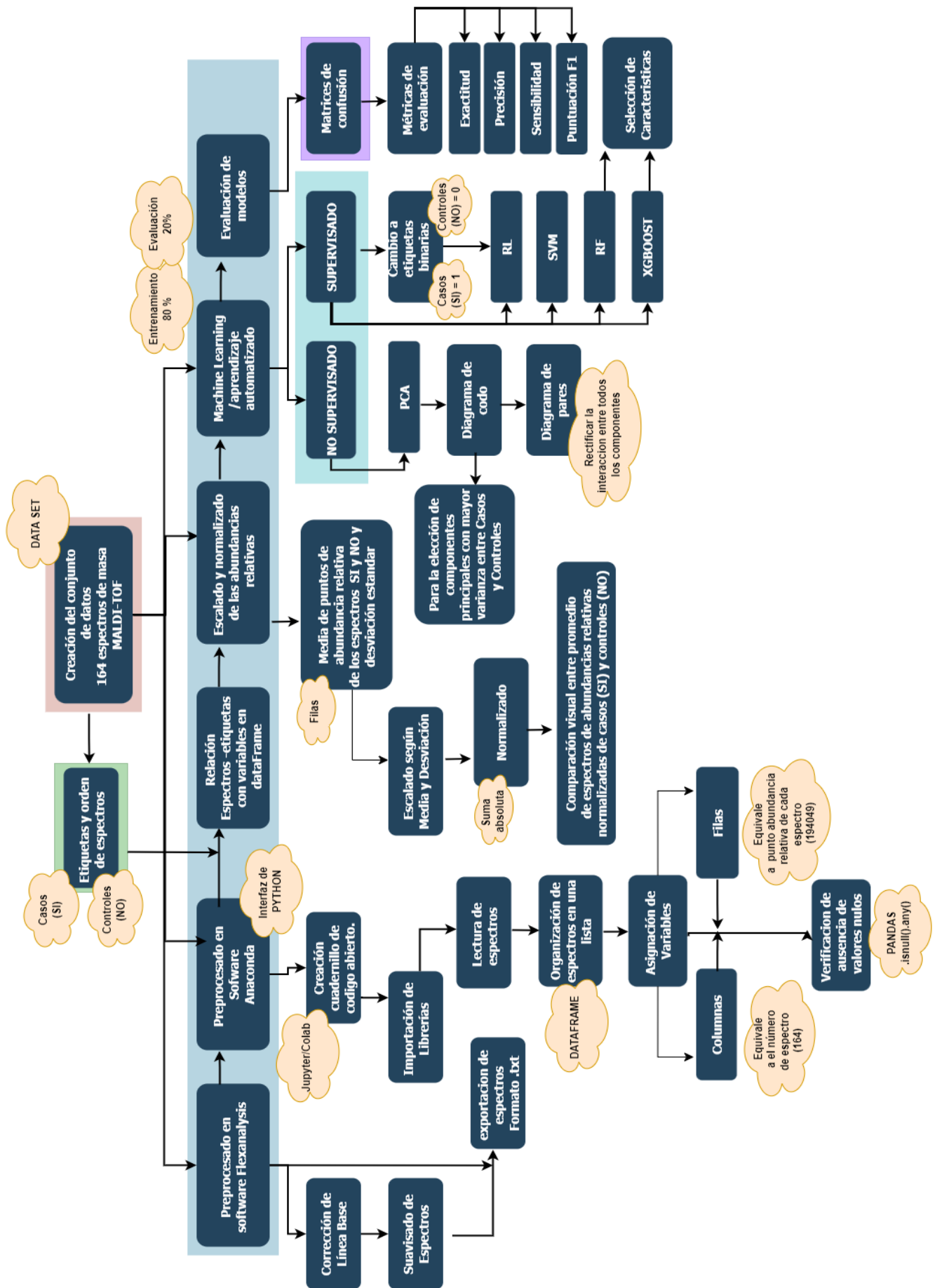


Figura 19. Resumen del cuadernillo de código que representa la construcción del modelo predictivo de PE.

### 9.5.5 Cuantificación de proteínas

En cuanto a la cuantificación proteica se siguió el procedimiento que sugiere el kit de BCA, [93] se escogieron al azar a 78 sueros sanguíneos de la dilución inicial 10:100 y con el fin de comparación se realizó la misma curva, pero cambiando el solvente a TA30 a 26 muestras, de las cuales, 25 se encuentran en intersección con el primer grupo de S.S de las mismas muestras, pero después del procedimiento de digestión y *SpeedVac* que fueron re-suspendidas en 50 uL de TA30. En un rango dinámico de 25-2000 ug/mL, Leídos por el lector de microplacas bajo una longitud de onda de 562 nm estos fueron sometidos a cuantificación.

Se necesitaron las siguientes soluciones para realizar la cuantificación:

-Solución suero original 1:10: 1 uL de S.S en 10 uL de agua tipo 1.

-solución péptidos en TA30: sólido secado por *SpeedVac* después de digestión enzimática disueltos en 50 uL de TA30.

-Soluciones del Kit:

Reactivo A: ácido bicinónico

Reactivo B: sulfato de cobre

Patrón: suero Albumina bovino (BSA)

-Solución de trabajo WR: se preparó a partir de los reactivos que contiene el Kit dependiendo método de cuantificación, en este caso fue por microplaca de 96 pocillos y del número de muestras que se vayan a cuantificar. Se mezclan 50 partes del reactivo A con 1 parte del reactivo B. El cálculo se realizó según la **Ecuación 15**

$$V_{totalWR} = (standar + muestras) \times replicas \times VWR_{pormuestra}$$

**Ecuación 15.** Formula de volumen de solución de trabajo en con el kit de cuantificación de proteínas.

A continuación, se muestran los cálculos:

$$V_{totalWR} = (18 + 78) \times 2 \times 200uL$$

$$V_{totalWR} = 38400uL \cong 40000uL$$

$$proporci3nreactivoA = \frac{50partesA}{51partestotales} = 0,9803$$

$$proporci3nreactivoB = \frac{1parteB}{51partestotales} = 0,0196$$

$$VA = 0,9803 \times 40000uL = 39216uLA$$

$$VB = 0,0196 \times 40000uL = 784uLB$$

Se inició con la preparación de las muestras descongeladas. Con la elaboración de los puntos estándar a partir de una ampolleta de estándar de BSA de acuerdo con el rango de cuantificación de 25 a 2000 ug /mL como se muestra en la **Tabla 9** Las curvas se hicieron por duplicado, el solvente para la curva de cuantificación de S.S fue agua y el solvente para la curva de cuantificación de fragmentos peptídicos fue TA30.

El procedimiento para medir las muestras en la microplaca fue el siguiente:

Se agregaron 25 uL de cada punto de la curva o muestra en el pocillo correspondiente, luego, con ayuda de la micropipeta multicanal se agregaron 200 uL de la solución WR, se tapó la microplaca para agitarla cuidadosamente y, por último, se llevó a incubación a 37°C por 30 min para poder introducirlo al lector de microplacas a la longitud de onda recomendada por el kit. Se llenaron 2 microplacas para la cuantificación de S.S originales y una microplaca para los fragmentos peptídicos.

**Tabla 9.** Esquema de dilución para el protocolo estándar en tubo de ensayo y procedimiento de microplaca.

Esquema de dilución para el protocolo estándar en tubo de ensayo y procedimiento de microplaca (intervalo de trabajo 20-2000 ug /mL)			
Vial	Volumen de diluyente (uL)	Volumen y fuente de BSA (uL)	Concentración final de BSA (ug/mL)
A	0	300 del estándar	2000
B	125	375 del estándar	1500
C	325	325 del estándar	1000
D	175	175 de la dilución del vial B	750
E	325	325 de la dilución del vial C	500
F	325	325 de la dilución del vial E	250
G	325	325 de la dilución del vial F	125
H	400	100 de la dilución del vial G	25
I	400	0	0 = BLANCO



## 10. RESULTADOS Y ANÁLISIS

### 10.1 Etapa 1

Esta parte se realizó de acuerdo con la metodología de la etapa 1, que consistió en ejecutar una digestión enzimática a escala ultra. Lo novedoso de realizarlo de esta manera fue que todo el procedimiento referente a cada muestra se realizó en un mismo recipiente, el uso del tubo Amicon de 3 kDa, y la versatilidad con la que se pudieron trabajar las muestras, que después de su descongelamiento controlado como se observa en la **Figura 20** demostró que se puede trabajar con varias muestras a la vez.

La etapa 1 de la metodología va de acuerdo con el primer objetivo de la investigación que pretendía ajustar las condiciones para la obtención de espectros, esto se logró transformando la estructura compleja de las proteínas del S.S hacia fragmentos peptídicos por medio de una digestión enzimática.



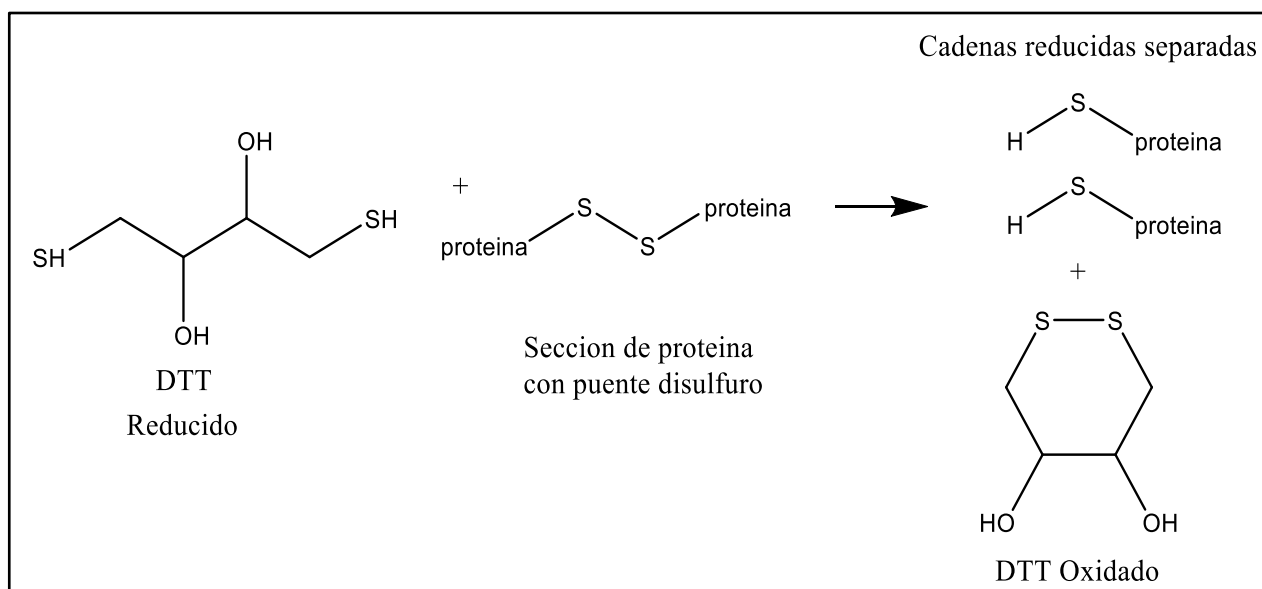
**Figura 20.** Criocajas con 164 muestras de suero sanguíneo con y sin PE y descongelamiento controlado de 24 muestras por día.

Se traen a consideración los cambios químicos que sufren las proteínas del S.S concerniente a la metodología FASP. Aunque se sabe que estuvieron almacenadas a  $-80\text{ }^{\circ}\text{C}$  y hubo precaución en el transporte de estas, no se descarta que durante el ciclo de descongelación hayan sufrido cambios como la desnaturalización o actividad en la composición del suero. La literatura afirma que hay efectos significativos en la integridad de muestras en la evolución del efecto de almacenamiento a largo plazo y ciclos de descongelación [94], [95], [96]. En este caso las muestras estudiadas de S.S fueron recolectadas entre los años 2000 a 2008 en diferentes ciudades de Colombia [9] y analizadas en el 2022 por lo que es pertinente tener este factor en cuenta.

Aparte de lo anterior, el usar un filtro de exclusión de tamaño, pretendía en principio descartar impurezas y fragmentos peptídicos que pesaran menos de 3 kDa, para ello se comenzó con la eliminación de sales, así como también, de la estructura terciaria y cuaternaria de las proteínas. Ya que pueden existir fragmentos que cumplan con la exclusión de tamaño pero que contengan una estructura globular, esto hará que sean impermeables a la membrana del filtro y queden retenidas en la fracción sobrenadante que es la que se quiso estudiar en esta investigación.

En principio el uso de un buffer de Tris HCl saturado de urea a pH 8,5 cumplió la función de solubilizar la muestra de S.S y mantener el equilibrio ácido-base de los aminoácidos ya que se procura no destruir completamente la estructura de cada uno de los residuos ni degradarlos, sino la desnaturalización y limpieza de las proteínas. El uso de una elevada concentración de urea fue con el fin de reducir las proteínas a su estructura secundaria, expandirlas o desplegarlas para que los siguientes reactivos pudieran interactuar de mejor manera con las proteínas. Como el suero en su gran mayoría es acuoso, se puede inferir que la proporción de lípidos en las muestras es mínima, al rededor del 3% [97], lo suficientemente baja como para no tener interferencia en el análisis de MALDI después de una digestión ya que posiblemente pueden ser arrastrados por las cadenas R de algunos residuos aminoácidos. Se pueden estar eliminando metabolitos hidrosolubles y sales en cada centrifugación que se realizó como se muestra en el procedimiento de la **Figura 15**. Aunque varios hallazgos exponen que la urea como desnaturalizante empeora el rendimiento de la digestión de proteínas [98], y si bien, se utiliza muy seguido aún se encuentra en debate el uso de este desnaturalizante [99], [100] por lo que proponen variaciones de esta metodología como la digestión multi enzimática (MED-FASP) [101], con ventajas en la recuperación de fragmentos peptídicos sin desnaturalizante [102].

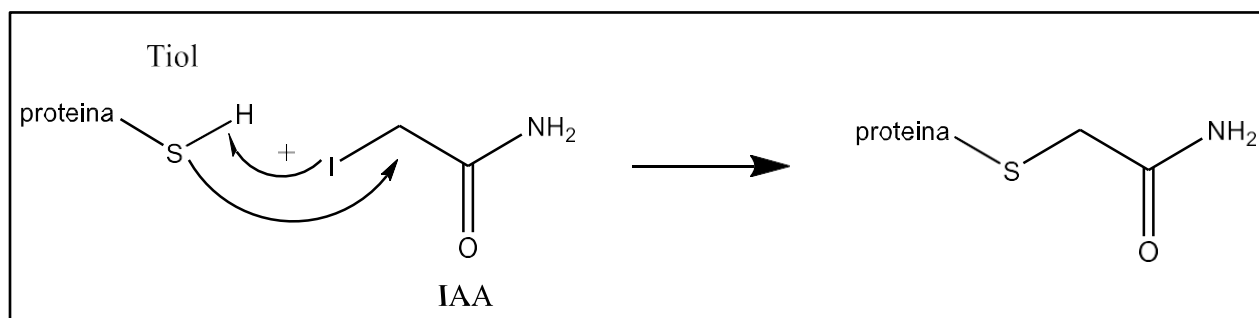
Enseguida, se usó el DTT, el cual es un agente reductor. Muchas de las proteínas contienen el aminoácido cisteína que presenta azufre en su estructura, este elemento al tener varios pares de electrones libres tiende a formar enlaces por hidrógenos, oxígenos, carbonos, incluso con otro átomo de azufre provenientes de otro residuo de cisteína formando puentes disulfuro y provocando bucles y favoreciendo la estructura tridimensional de las cadenas peptídicas. En este paso no se trata de una desnaturalización como en el caso del uso de la urea donde se suprimen los enlaces no covalentes de las conformaciones en hélices alfa o laminas beta; ya que el DTT promueve una reacción de reducción del enlace disulfuro [103] como se muestra en la **Figura 21** el mecanismo de acción se da mediante un proceso de óxido-reducción lo que implica movimiento electrónico para poder separar cadenas reducidas de las porciones que conformaban los puentes disulfuros, y el producto oxidado es un ciclo de 6 miembros, con un enlace disulfuro muy estable inmune a la reducción del mismo.



Tomada y adaptada de García [104] y Santos [105].

**Figura 21.** Mecanismo de acción del DTT en la metodología FASP.

Sin embargo, el grupo tiol o sulfhidrilo de los aminoácidos persiste en la estructura de las proteínas y puede seguir siendo propenso a la formación nuevamente de puentes disulfuro, así que continuando con el análisis del procedimiento de la etapa 1, se continuó con la adición de IAA, la cual ayudó a formar enlaces covalentes irreversibles por medio de una alquilación del tiol formando grupos carboxamida, en este paso la IAA funciona como un grupo protector [106] y en la **Figura 22** se puede ver el mecanismo de como el átomo de azufre actúa como nucleófilo el cual ataca el carbono alfa al carbonilo de la IAA ya que es un carbono que se encuentra altamente desprotegido electrónicamente por acción de la electronegatividad del halógeno y el grupo carbonilo. El yodo al mismo tiempo para recompensar su estabilidad desprende el hidrogeno del grupo sulfhidrilo favoreciendo la formación del enlace carbono-azufre, y debido a que la sustitución se da en un solo paso la reacción biomolecular  $S_N2$ .



Tomado y adaptado de Wísniowski [106] y Hustoft [107].

**Figura 22.** Mecanismo de reacción de IAA en la metodología FASP.

En la digestión por el método FASP, el tubo Amicon, actuó como concentrador y como configuración del reactor químico. Durante cada centrifugación se observó que efectivamente no pasaba la totalidad de la fase superior. Además, el volumen de solución del filtrado desechado era diferente en cada muestra después de cada

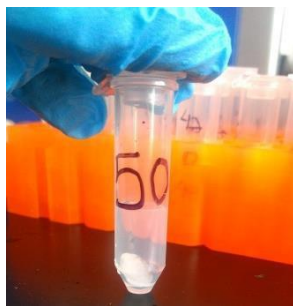
centrifugación. Esto pudo ser debido a que el contenido de proteínas entre distintos sujetos puede variar, dependiendo de diferentes factores, pero precisamente el objetivo del estudio fue encontrar diferencias entre un paciente control y casos que contengan PE. Varios autores resaltan las ventajas de realizar la digestión enzimática en este tipo de filtros porque saben que la extracción, purificación y fraccionamiento de proteínas es un paso crítico en el análisis proteómico [101], [108], [109].

Después del tratamiento con IAA se hicieron lavados repetitivos con buffer sin urea (UB). En ese momento el pretratamiento de la muestra debió completarse para la posterior adición de la tripsina y haber quedado limpia de este compuesto ya que, así como es capaz de desnaturar las proteínas de la muestra, también es capaz de desnaturar una proteína con actividad catalítica, que no era el objetivo; además que el tiempo que se estableció para la digestión fue de 18 horas, tiempo en el que se esperaba que la tripsina actuara completamente sobre las proteínas tratadas en los extremos carbonilo de la Lys y Arg y garantizar su integridad durante ese tiempo fue primordial.

Por último, la eliminación de la tripsina también fue importante realizarlo ya que no se quería encontrar fragmentos peptídicos provenientes de fuentes diferentes a los de la muestra de suero. En este paso, se reconoce que la tripsina cuanta con un peso mayor a 23 kDa, y teniendo en cuenta que se recolectó siempre la fracción sobrenadante del tubo Amicon pudieron haber quedado porciones de la enzima. La recolección del sobrenadante fue un procedimiento relativamente fácil, pero cuidadoso ya que la punta de la micropipeta podría rasgar las membranas del filtro Amicon y arrastrar fibras que pueden quedar concentradas luego del procedimiento con el *SpeedVac*.

## 10.2 Etapa 2

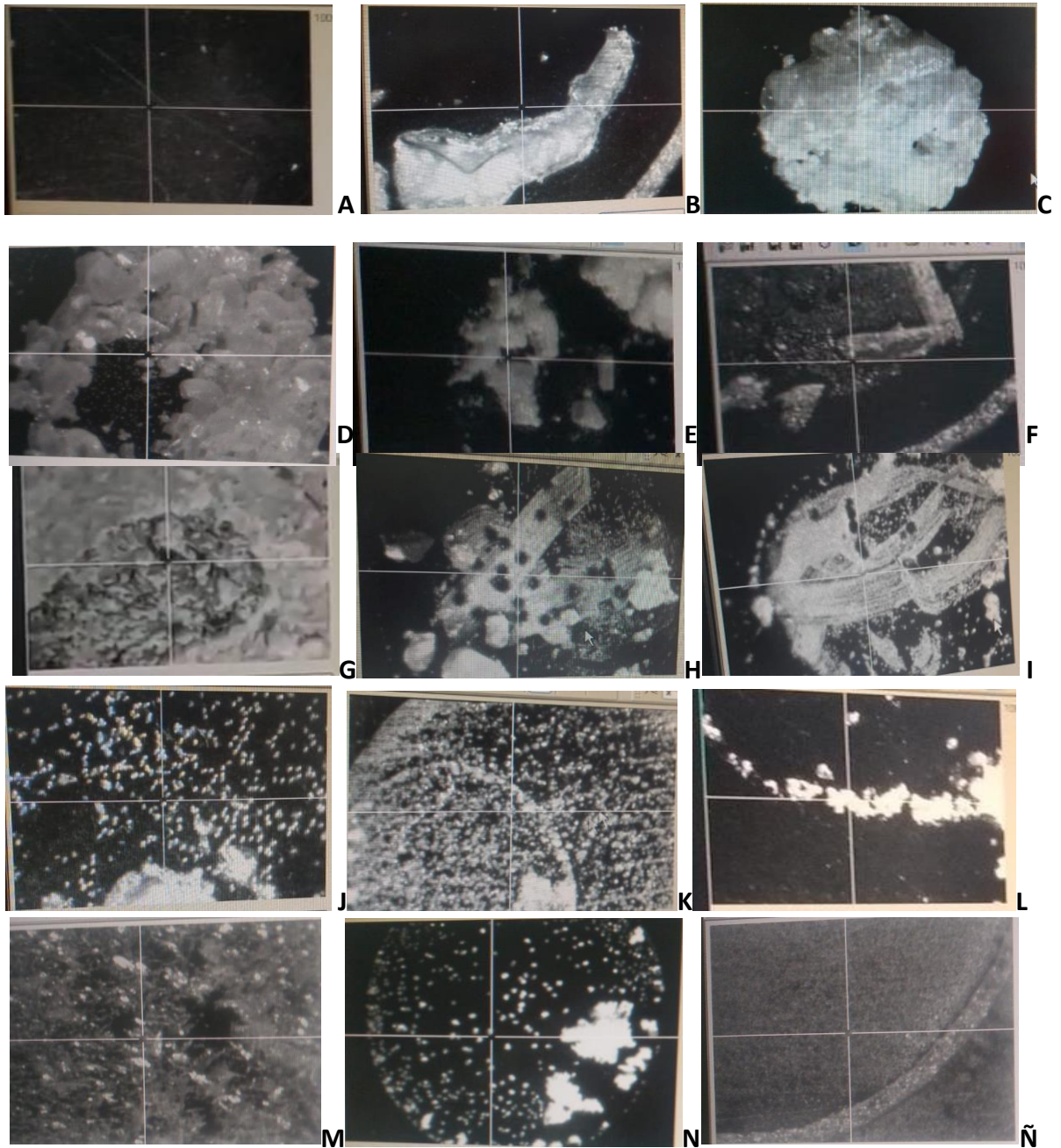
Los fragmentos peptídicos recolectados después del tratamiento al vacío mostraron un aspecto cristalino, se fotografió este para la muestra número 50 en la **Figura 23** pero el resto de las 163 muestras mostraban la misma apariencia a excepción del volumen, que, a través de la observación, sí variaba, pero no fue cuantificado.



**Figura 23.** Fragmentos peptídicos después de SpeedVac. Apariencia cristalina.

En la sección de metodología de la etapa 2, se explicaron los factores que se escogieron con el fin de ajustar dichas condiciones, el análisis del ajuste se encuentra en el **anexo 1** y se resalta que el factor determinante a evaluar fue el mayor número de señales por experimento reflejadas en el espectro; los estadísticos del análisis de datos que se escogieron fueron los que mayor diferencia presentaron respecto al conjunto de experimentos, concluyendo así las mejores condiciones para el pretratamiento de los fragmentos peptídicos. El acto de la medición de la muestra en el Spot es susceptible de sufrir irregularidades, algunas de las cuales fueron experimentadas y se muestran en la **Figura 24**. Los métodos de deposición de gotas sobre las superficies de la placa MALDI-TOF [110] son un fenómeno físico complejo e interesante que combina la nucleación y crecimiento de cristales en una solución saturada y la evaporación de una gota que contiene solutos en contacto con dicha superficie. El anillo que ve en algunos de los recuadros y otros autores [111], [112], [113] se forma porque la línea de contacto de la gota no puede moverse cuando la evaporación elimina el solvente alrededor de la línea, y el soluto es arrastrado hasta la línea de contacto generando patrones [114] y la geometría de cristalización de las matrices [115].

Se quiere resaltar que este proceso de ajustar las condiciones para el análisis espectrométrico demandó tiempo para ensayar diferentes matrices para poder adquirir los espectros de buena calidad que son los datos primordiales de la investigación para que puedan ser modelados posteriormente. Se observó por ejemplo en **Figura 24** que la DBH generaba placas de cristal homogéneas como se muestra en el recuadro M, pero fue difícil de generar fragmentos que pasaran al tubo de tiempo de vuelo con el método escogido, en otras palabras, no se obtuvo espectros de calidad y con señales lo suficientemente predominantes para categorizarlos como aptos para el modelo predictivo. Por otro lado, el AS como en el recuadro Ñ, generó buena dispersión de los cristales y no mostraban gran tamaño, pero al estar preparada en etanol, la primera capa se secaba rápido y brusco, saliendo del límite del spot por muestra, En general la más sencilla de aprender a manipular fue la HCCA, recuadro N.

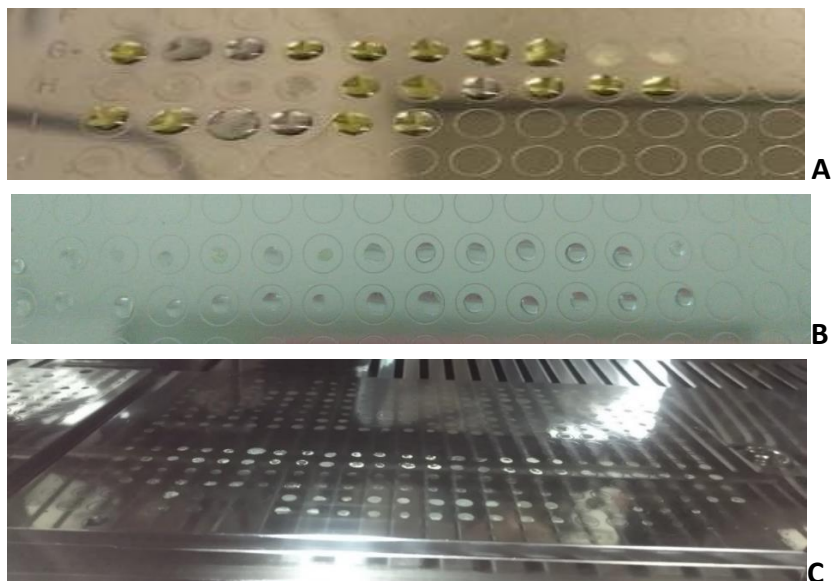


A) Spot vacío. B), C), D), E) y F) representan distintas formas de sobre cristalización tanto en tamaño como en cantidad de matriz. G) exceso de muestra. H) y I) buena dispersión cocrystalización, pero invadida por toque de punta plástica y huella de disparos con láser. J) y K) buena dispersión de matriz y en K se ve la doble capa interceptada. L) movimiento de cristales hacia orilla de la gota. M) DBH cristalizada. N) prototipo de buen sembrado en la placa con matriz HCCA Ñ) prototipo de mal sembrado con Matriz AS.

**Figura 24.** Colección de imágenes de sembrados de matrices en la placa de MALDI-TOF-MS.

En cuanto a la re-suspensión que fue el único factor que no se estudió estadísticamente en el DOE, se observó que al mezclar directamente una alícuota de los fragmentos peptídicos en 50 uL de TA30 con HCCA, estos nunca se secaron completamente, incluso se dejaron 24 horas a lo que resultó una especie de capa viscosa de color amarillo, como se observa en la **Figura 25 A**, esta apariencia es inaceptable para la introducción de la placa en el

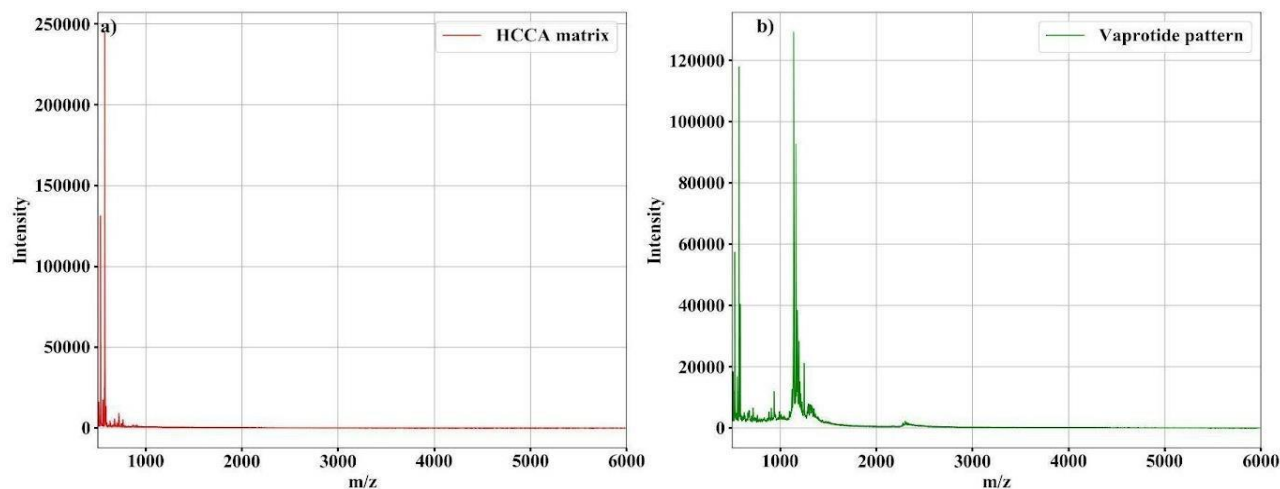
espectrómetro de masas. Esto pudo deberse a que aun esa concentración de proteínas era muy alta, esta prueba impulsó a realizar diluciones y por eso se escogió como un factor en el ajuste de condiciones. En cambio, la **Figura 25 B** muestra la apariencia de una buena relación entre muestra y matriz, no se observa color amarillo y el secado fue bastante rápido no superaban los 20 min a temperatura ambiente En la **Figura 25 C** se demuestra el secado de varios spots listos para introducir al espectrómetro de masas.



A) Mala relación muestra-matriz con apariencia gelatinosa. B) Buena relación muestra-matriz antes de secado. C) Apariencia de buen secado.

**Figura 25.** Apariencia de sembrados en la placa MALDI-TOF.

Por lo mencionado anteriormente se estableció el conjunto de espectros de masas incluyendo los de la matriz y el patrón para verificar que el método estuviera funcionando correctamente. Pero para el modelo predictivo de PE solo se utilizaron los 164 espectros correspondientes al número de sueros o pacientes, sin involucrar patrón ni matriz. a continuación, se muestran los espectros obtenidos de la matriz HCCA en la **Figura 26 a)** y el patrón de verificación de masas en la **Figura 26 b).**



**Figura 26.** Espectros de masas MALDI-TOF bajo las condiciones establecidas. a) matriz HCCA, b) patrón Vaprotida.

El espectro de la matriz permitió monitorear su probable contaminación por uso en el momento de hacer contacto con los spots y extraer la alícuotas de matriz del frasco mientras se realizaba el sembrado de doble capa y se percataba por la apariencia de la señal característica cada vez que se analizaba un conjunto de muestras. Teniendo en cuenta que el peso molecular de la matriz HCCA es 224,21 g/mol; se esperaría que no se vieran señales por debajo del límite inferior que se eligió de 500 m/z, también se debe considerar la formación del dímero de la matriz, el cual representaría el doble del valor de 448,62 g/mol que al ser ionizado por un H<sup>+</sup> se detectaría en un valor de 449,62 m/z, podríamos aproximarlos a 450 m/z, sin embargo, alrededor de los 500 a 600 m/z se ven señales arrinconadas. También la matriz puede contener impurezas de fabrica que aparecerían como un error sistemático en cada espectro que se tomó, otra opción es la formación de aductos de la matriz con sales como la unión con iones salinos e incluso con los solventes que se usaron para analizarla en el equipo que también son de naturaleza orgánica.

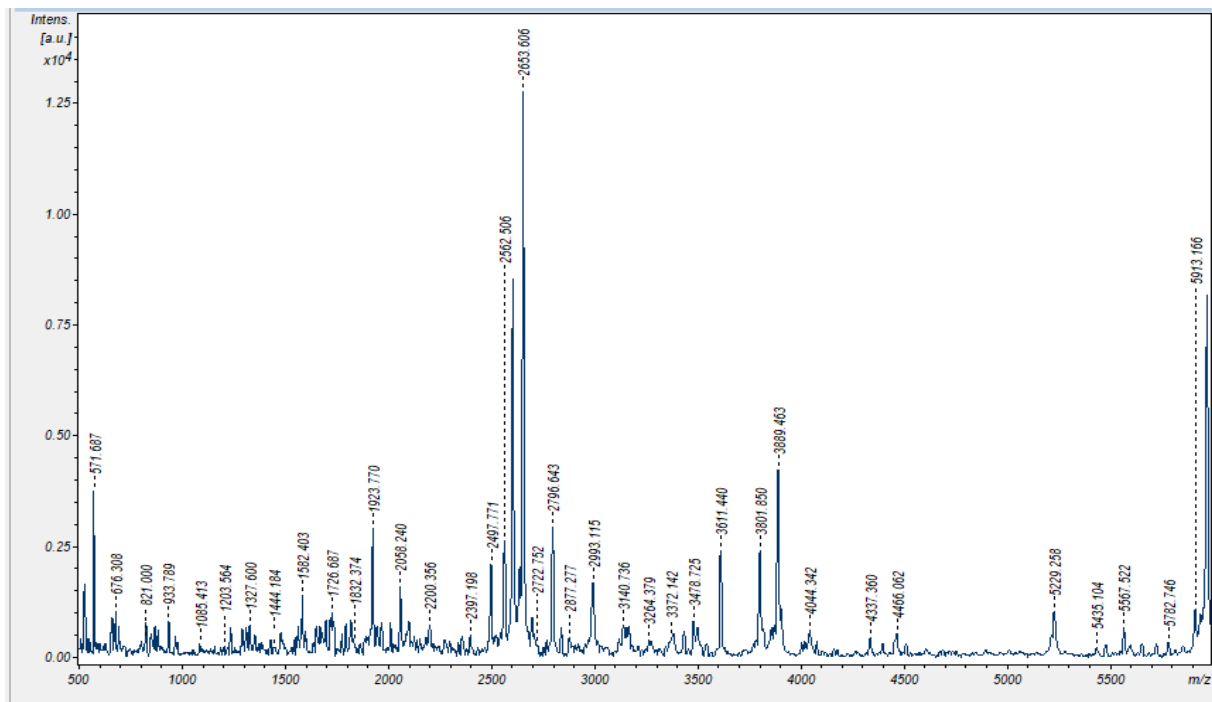
Lo mismo puede comprobarse con el espectro del patrón de verificación, el péptido Vaprotida tiene un peso molecular de 1131,4 g/mol, el espectro de este compuesto presenta más señales que el de la matriz por lo que puede ser posible pensar que se fragmente, sin embargo, señal de base para dicho espectro presento un valor de 1139,64 m/z, que, aunque no es un valor muy alejado del teórico, también podría justificar analizando la estructura del péptido ya que contiene grupos imidazol, amino terminal, puente disulfuros y fenoles que pueden ser susceptibles a protonarse en medios ácidos como lo es el solvente de TA30, el cual contiene un porcentaje de ácido acético para favorecer dicho mecanismo.

Por otro lado, se aclara que no necesariamente dos espectros de una misma muestra, tomados en las mismas condiciones deban ser exactamente iguales, ya que puede haber corrimientos significativos como no significativos de las señales; esto depende del spot donde se tomó la muestra, exactamente de la topografía del spot cuando quedan aglomeradas o muy delgadas.



Aunque se conoce la estructura de la matriz y se tiene una idea de la composición de las muestras, de lo que se tiene real certeza es de que la matriz y energía del láser favorecen la ionización de las moléculas, sin embargo, según Karas M. (2003) nunca se ha investigado la existencia de moléculas peptídicas cargadas en los cristales de la matriz ni el estado de carga de los iones en la cocrystalización [116]. Lo anterior hace que haya una incertidumbre en el identificar las masas exactas de iones de interés sin una separación previa del analito. También puede provocar los corrimientos o explicar por qué espectros de una misma muestra pueden variar levemente en los valores de  $m/z$ , Dichos corrimientos en el eje  $x$  o de  $m/z$  hace que sea una variable fluctuante para realizar un modelo predictivo, por eso los datos que se usan en el modelo son las intensidades, que representan la abundancia relativa de la proporción de péptidos ionizados en una muestra en específico.

En cuanto a la apariencia de los espectros de las muestras se observa que la mayoría llegó a la intensidad deseada que era 4 órdenes de magnitud, sin embargo, algunos que quedaron en el límite de este rango bajaron su intensidad al aplicarles la sustracción de línea base y la función de suavizado; este proceso mejora la relación S/N por lo cual, la tendencia del conteo de señales se vio disminuida y algunos espectros bajaron a una intensidad 3 órdenes de magnitud. El soporte de los espectros de toda la investigación se muestra en el **anexo 2**. Y a colisión se exponen unos ejemplos en la **Figura 27** y **Figura 28** para muestras positivas o casos ante la enfermedad de PE y para muestras negativas o controles en la **Figura 29** y **Figura 30**.



**Figura 27.** Espectro de la muestra # 101 con etiqueta de caso (sí).

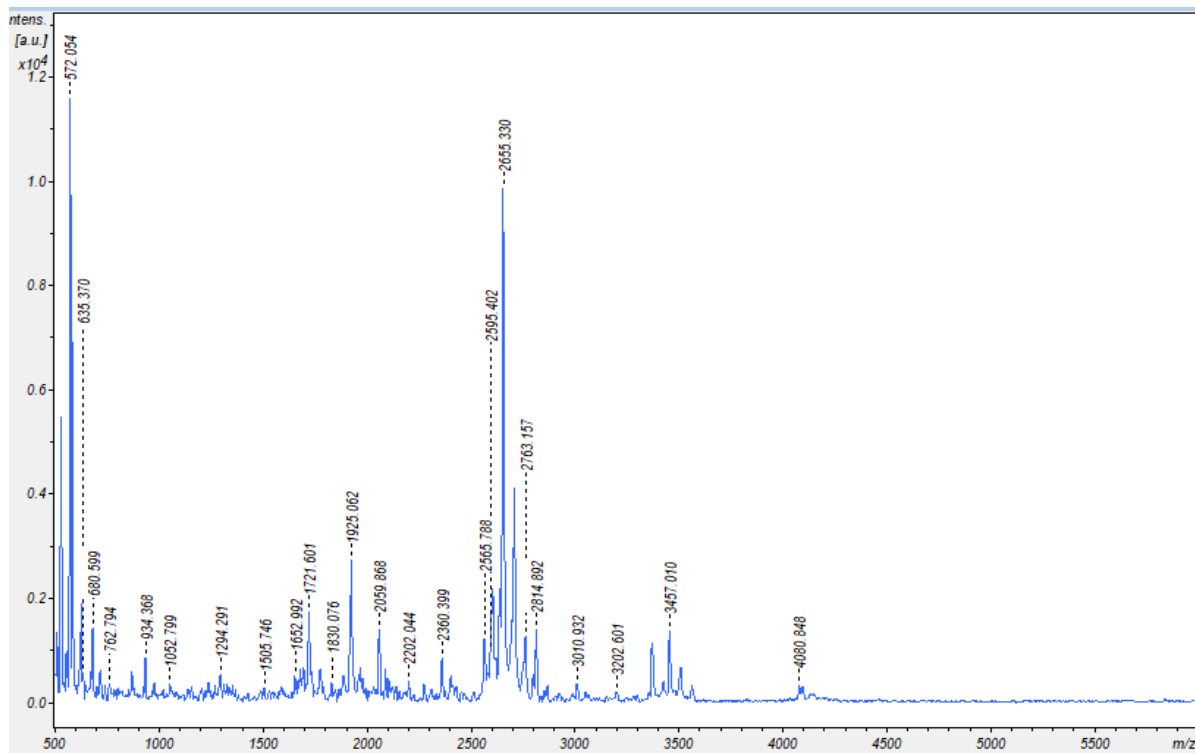


Figura 28. Espectro de la muestra # 11 con etiqueta de caso (sí).

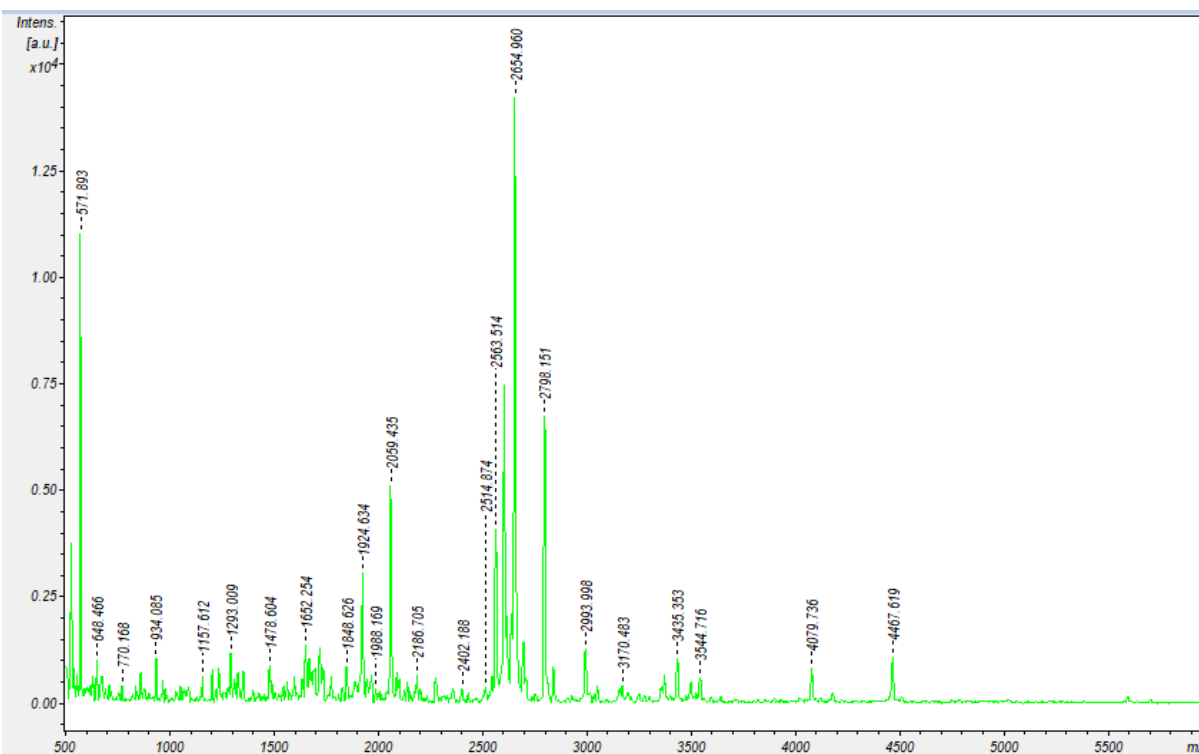
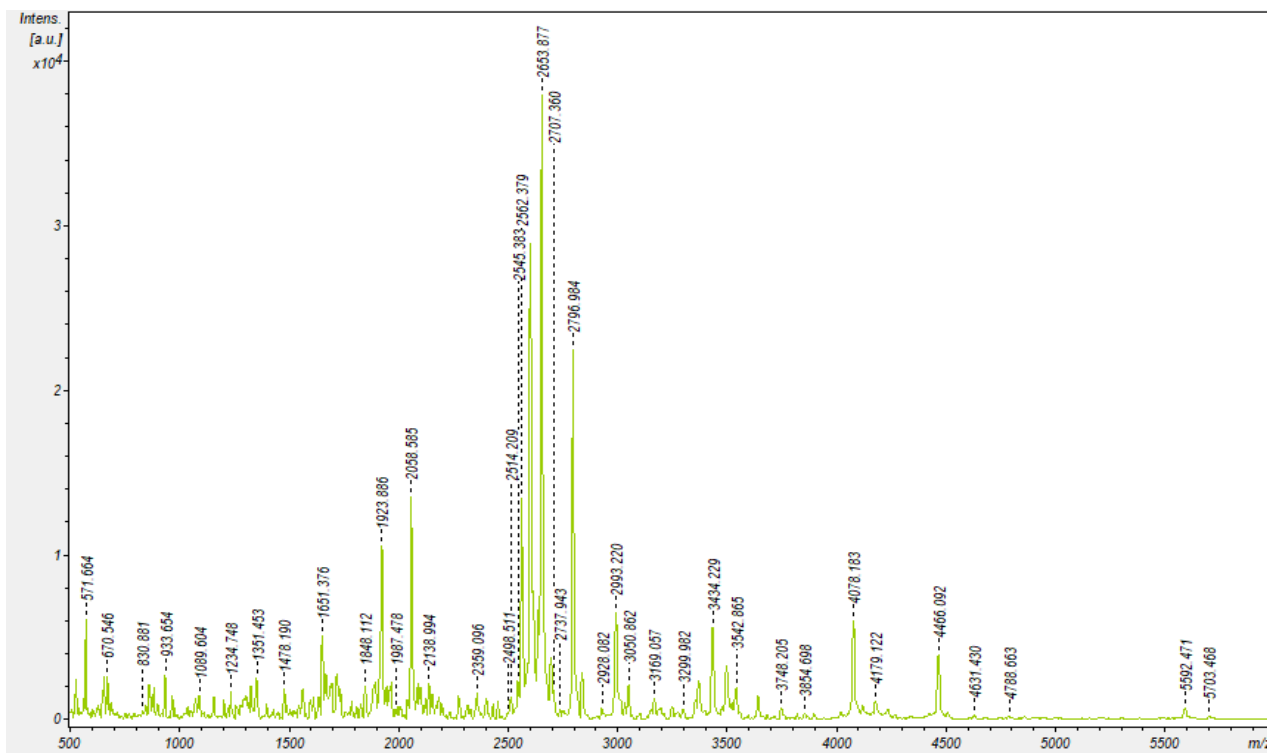


Figura 29. Espectro de la muestra # 13 con etiqueta de control (no).



**Figura 30.** Espectro de la muestra # 99 con etiqueta de control (no).

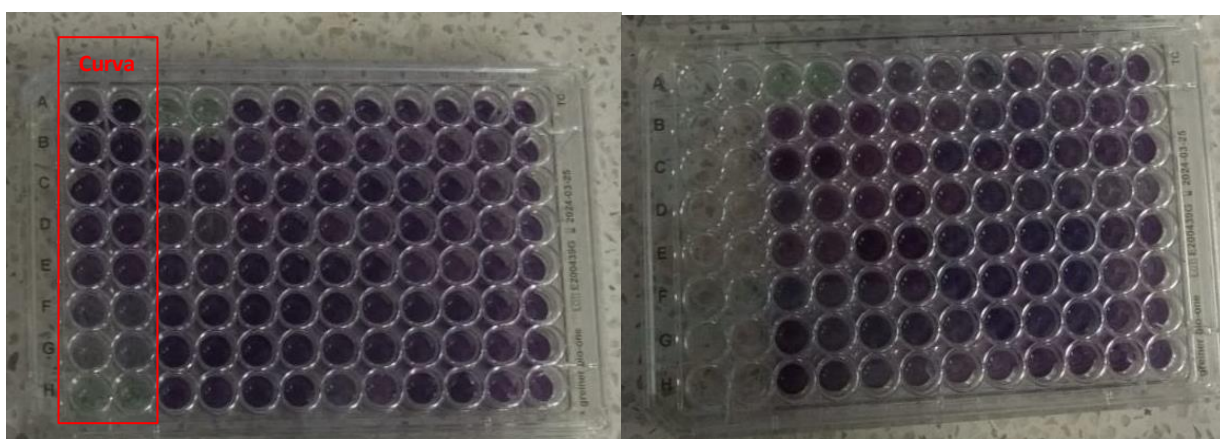
Los espectros entre casos y controles, en realidad no mostraron diferencias que se puedan detectar a simple vista, ni en el conteo general de señales, al contrario, en general el total de los espectros de masas presentan similitudes como: las señales de la matriz alrededor de 500 m/z, señales de baja intensidad en el rango de 600 a 2300 m/z, señales solapadas y de gran intensidad al redor de 2400 a 3000 m/z, donde en la mayoría de las muestras esta es la sección donde se encuentra la señal más alta. También en el rango de 3100 a 5000 m/z señales de baja intensidad que en muchos espectros son confundidos con ruido y en algunas excepciones de 5000 a 6000 m/z se encontraron señales que con poca frecuencia aparecían.

De acuerdo con el uso del filtro MWCO de 3 kDa en el análisis de MALDI-TOF-MS de los sobrenadantes de la centrifugación se esperaban mayor abundancia de picos por encima de los 3000 m/z suponiendo que los polipéptidos se ionizaran con una carga de 1<sup>+</sup>. Al encontrar picos por debajo de este valor de m/z es preciso pensar que posiblemente la ionización de algunos fragmentos fue al menos de 2<sup>+</sup> o en su defecto, algunos péptidos que cumplían con el corte de peso molecular quedaron retenidos en el sobrenadante por presentar alguna disposición tridimensional. Tampoco se descarta la posibilidad de que algunas señales sean pertenecientes a fragmentos de tripsina. O en su irregularidad, presencia de metabolitos ya que la tripsinización es un proceso específico que hace incisión en extremos carbonilos de los aminoácidos Lys y Arg.

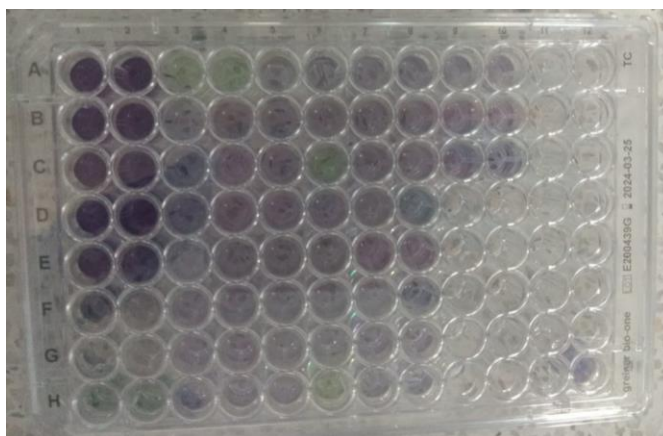
Debido a las diferencias indetectables visualmente, se procedió a realizar la búsqueda de posibles intensidades relativas de las señales que marquen diferencia general y con base a ellos poder proponer un modelo predictivo para la PE, haciendo uso de técnicas de análisis de datos y ML con lenguaje Python como se explica en la etapa 3 que antes de pasar a ello se procederá a mostrar la etapa de cuantificación de los S.S.

### 10.3 Cuantificación de proteínas

Como se mencionó en la metodología, con el fin de comparación de concentración proteica entre S.S al inicio (D 10:100) y al final de su tratamiento (solución fragmentos peptídicos en TA30). Se empleó el kit de cuantificación BCA. En el **anexo 4** se muestra más en detalle el cálculo en una hoja de Excel automatizada para realizar la solución WR y esquematizar la posición de las muestras según las coordenadas de la microplaca y los datos que respaldan la cuantificación de cada una de las muestras que se utilizaron con tal fin. Para la cuantificación de los S.S originales se ocuparon 2 microplacas como se ve en la **Figura 31** y para los fragmentos en TA30 en la **Figura 32**.



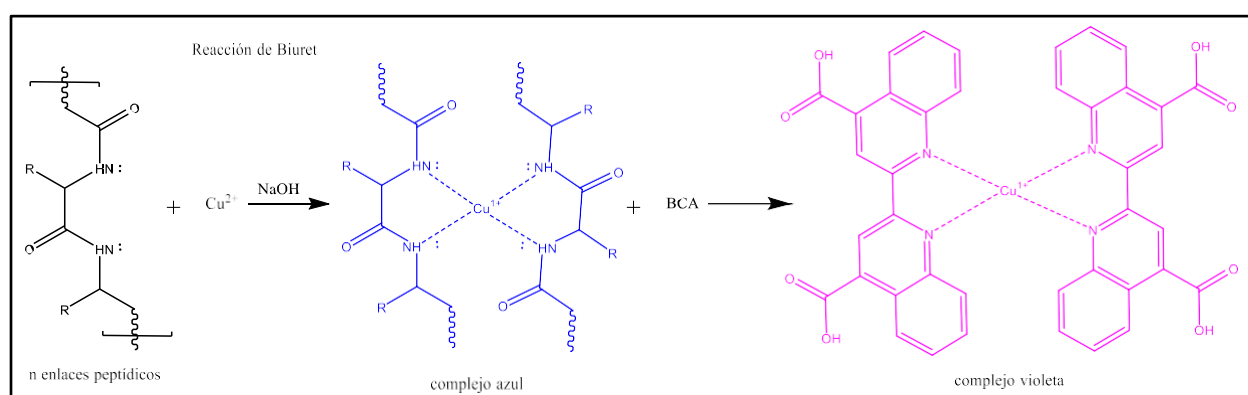
**Figura 31.** Apariencia de las microplacas para la cuantificación de sueros originales.



**Figura 32.** Apariencia de microplaca para la cuantificación de fragmentos peptídicos en TA30.

Con el color purpura se puede comprobar que el kit está funcionando correctamente. Además, da un indicio de qué tan concentradas están ciertas muestras respecto de otras, por ejemplo, en la **Figura 31** en la microplaca de la izquierda, las columnas 1 y 2 contienen los puntos de la curva de mayor a menor concentración del estándar (BSA) de la fila A hasta la H donde se observa la degradación del color purpura.

En este ensayo se llevan a cabo dos reacciones, la primera es la reacción de Biuret, donde las proteínas o al menos dos enlaces peptídicos pueden reducir el cobre del estado cúprico ( $\text{Cu}^{2+}$ ) a cuproso ( $\text{Cu}^{1+}$ ) en medio alcalino con tartrato sódico potásico. La coloración de este sistema es azul. Posteriormente la interacción del ion cuproso con dos moléculas de BCA forma otro complejo que al ser irradiado por luz visible de longitud de onda de 562 nm provoca que la intensidad de absorbancia sea proporcional a la cantidad de proteína en la muestra. Lo que pretende esta reacción es garantizar un límite inferior de detección. En la **Figura 33** se esquematiza el proceso descrito anteriormente y se observa cómo el sistema conjugado del complejo con BCA aumenta considerablemente en comparación al obtenido por la reacción de Biuret, esto hace que los fotones absorbidos exciten a los electrones y estos últimos se deslocalicen alrededor de los orbitales  $\pi$  y proporcione más estabilidad al complejo en el tiempo de detección. Así que, puede decirse que entre más intenso el color, mayor es la concentración del complejo, que fue formado gracias a la acción de reducción de las proteínas hacia el cobre que por cierto es favorecido por el tiempo de espera de incubación para que se transfiera el electrón al ion cúprico. Lo que se puede verificar entre la Figura 28 y Figura 29 donde se comparan las cuervas de cuantificación para determinar las concentraciones los S.S o sin tratamientos y las muestras de fragmentos peptídicos después de realizar la digestión donde se lavaron porciones proteicas varias veces de las muestras originales, respectivamente.



Obtenida y adaptada de Otieno [117] y Fisher Scientific [118].

**Figura 33.** Reacciones del principio de detección de proteínas por el método de Kit de BCA.

Como se mencionó en la metodología, fueron 25 muestras en total que se pudieron comparar ante los dos tratamientos y el resto se encuentra en el **anexo 4**. En la **Figura 34 y Figura 35** se muestran las curvas con mejor  $R^2$  que se utilizaron para la cuantificación de cada grupo de muestras y en la **Tabla 10** el cambio en la concentración de las muestras antes y después de la digestión enzimática.

Es posible que dentro del contenido de muestras de fragmentos peptídicos haya, además del material proteico, compuestos como restos de tripsina, IAA, DTT y urea que se considerarían impurezas. Aunque pueden estar afectando la cuantificación, el efecto será el mismo en todo el conjunto de muestras. En cuanto al uso de TA30 como solvente para realizar la curva, el kit BCA no limita el método para realización de la curva en el mismo solvente en que se encuentran las muestras, es por eso por lo que se procedió a realizarlo de tal forma, y así poder comparar.

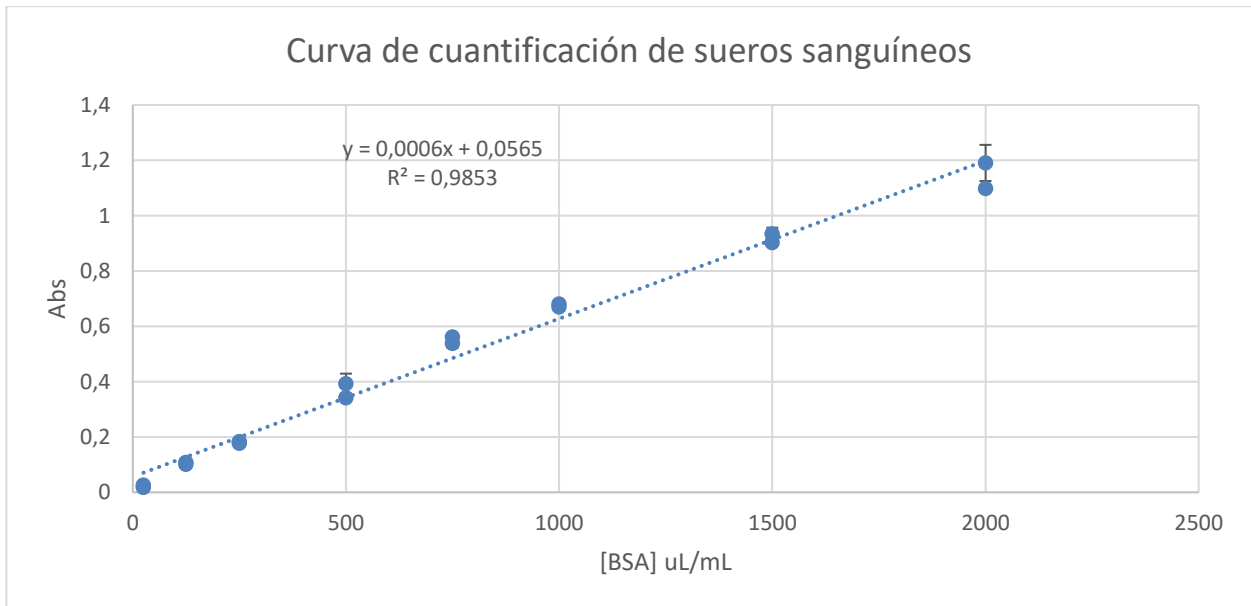


Figura 34. Curva de cuantificación para sueros sanguíneos originales en agua.

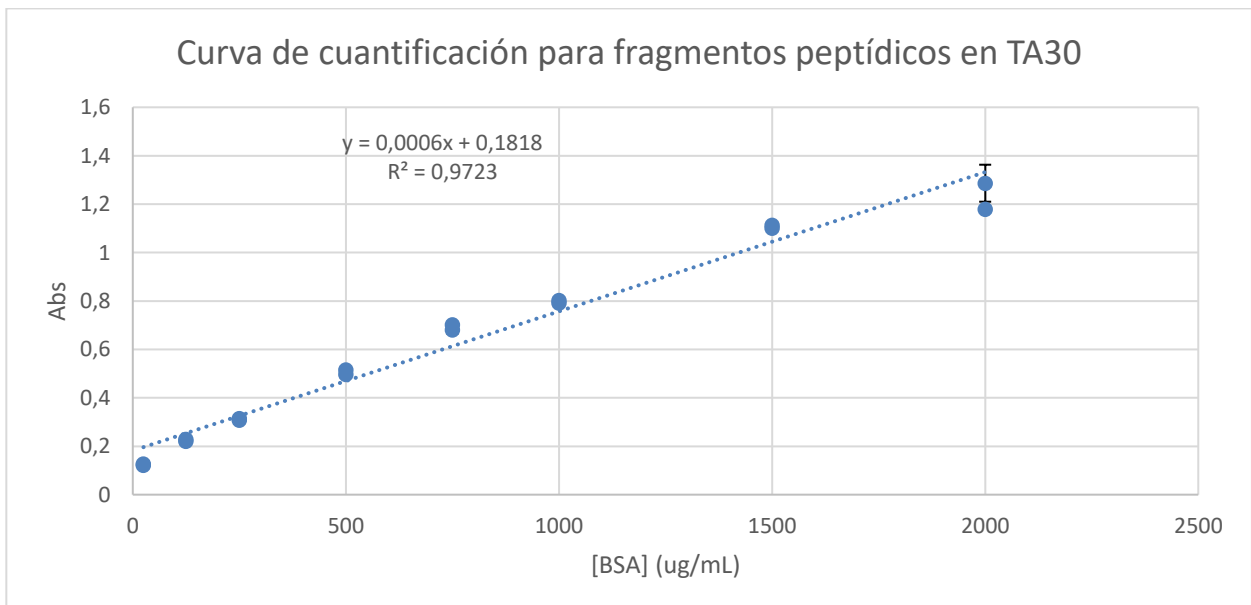


Figura 35. Curva de cuantificación de fragmentos peptídicos en TA30.

Ambas curvas muestran buenas correlaciones lineales. En cuanto a la sensibilidad, Simpson [119]. Por ejemplo, obtuvo una pendiente de 0,0016 entre 20 a 250 ug/mL con el mismo kit, también Kreuzsch [120] realizó curvas de cuantificación con un solvente diferente al agua, diluyó BCA en una solución de 10Mm de buffer de fosfato a pH 7,4 y obtuvo diferentes curvas de cuantificación con pendientes desde 0,2085 mg/mL hasta 0,000477 ug/mL. Por otro lado, las concentraciones encontradas para los S.S no están muy lejanos de los reportados en la literatura, el contenido promedio de proteínas en S.S en un ser humano es de 60000 - 80000 ug/mL [121], [122], [123], en este caso se encontró un rango de 42667,1 - 119612,3 ug/mL y el contenido de los fragmentos peptídicos en TA30 está en el rango de 3,365 – 135,396 ug/mL aunque el límite inferior no sería correcto ya que no entra en el modelo lineal de la curva de cuantificación, sin embargo, son mostrados en la **Tabla 10**; así entonces el verdadero rango cuantificado sería 24,911 – 135,396 ug/mL.

**Tabla 10.** Comparación de concentración de proteínas en sueros sanguíneos intactos y fragmentos peptídicos después de digestión enzimática disueltos en TA30.

Muestra	Etiqueta	Concentración de proteína en suero sanguíneo original [ug/mL]	Concentración de fragmentos peptídicos en TA30 [ug/mL]
115	sí	94235,9	17,709
135	no	79097	9,385
26	no	100265,9	68,297
18	no	119612,3	20,059
23	no	42667,1	137,396
30	no	87369	112,282
1	no	103313	26,677
7	no	82803,9	28,078
122	no	92550,9	17,810
19	sí	57037,4	41,434
27	no	93474,6	45,175
24	sí	73915,5	5,271
140	sí	79222,9	58,973
138	no	80863,6	35,746
22	no	93636,3	66,858
13	no	101410,9	75,369
119	no	71758	24,911
132	sí	65348,8	32,349
148	sí	94954	14,939
137	no	82320,6	3,365
37	no	99859,1	85,563
6	no	105341,2	8,063
14	no	114337,2	71,608
35	no	75234,4	53,106
16	no	75767,9	83,600
PROMEDIO		86655,9	45,761
MAXIMO		119612,3	137,396
MINIMO		42667,1	3,365

### 10.4 Etapa 3

Los datos en Python podían trabajarse tanto en anaconda si los datos se encontraban guardados en el computador o desde Google Colab si se encontraban en la nube como Google drive. El inicio del cuadernillo, que se muestra en el **anexo 3**, es la importación de librerías como se mostró en la sección de la metodología.

La particularidad de ver una imagen de un espectro de masas es que no se cuentan la cantidad de puntos por espectro, solo un valor de intensidad en un determinado rango de m/z. Pero cuando se exportan en formato .txt todos los datos se desglosan en dos columnas separadas por un espacio ( ' ') donde cada entrada o fila representa un punto que forma el espectro de masas. En este caso se encontró un valor de 194049 filas en todos los documentos de texto. Que fueron organizados en una lista (DataFrame) y se visualizan en la **Figura 36**.

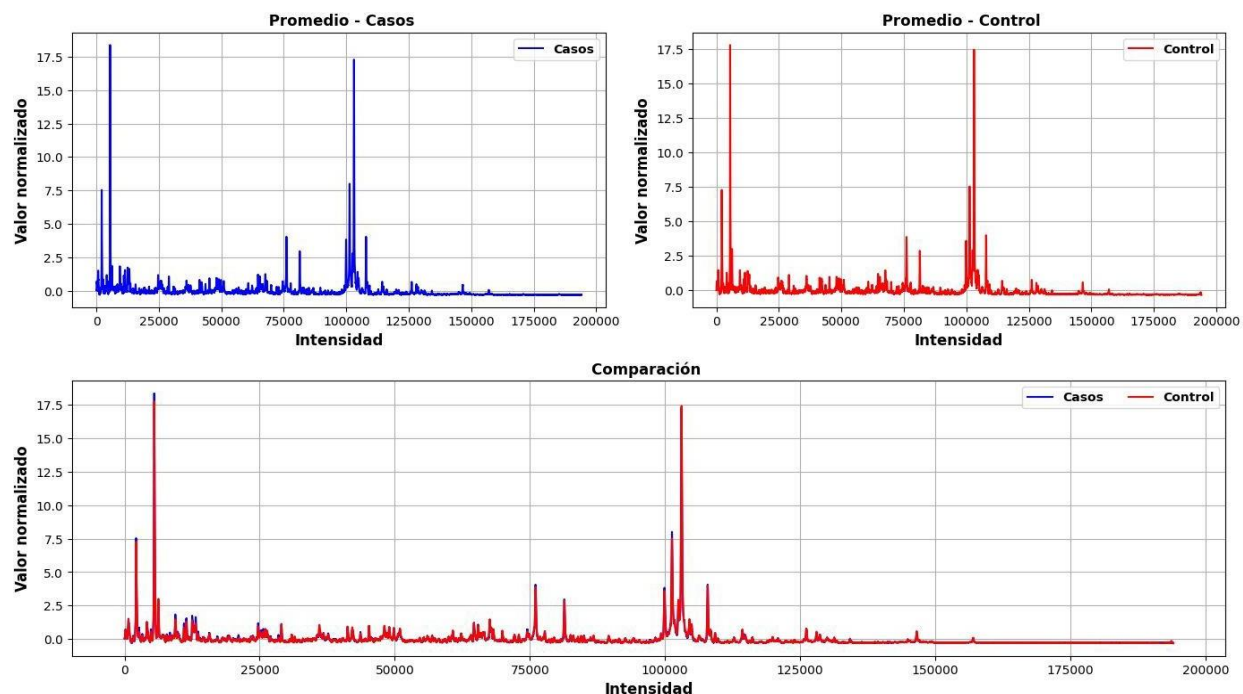
	0	1	2	3	4	5	6	7	8	9	...	154	155	156	157	158	159	160	161	162	163	
0	315	205	125	221	136	437	194	72	51	410	...	183	3023	30	276	175	135	302	124	273	290	
1	317	199	126	200	127	441	196	67	54	404	...	169	3082	31	254	173	116	272	129	273	244	
2	319	195	128	217	142	444	190	73	57	379	...	186	3132	31	261	179	105	280	134	247	239	
3	320	198	130	203	137	447	203	75	60	393	...	194	3130	32	296	176	119	242	139	215	250	
4	322	157	132	187	134	451	193	84	63	390	...	178	3146	33	291	204	121	251	144	231	233	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
194043	6	20	2	858	30	8	6	1	8	32	...	5	-6	23	17	12	1	4	14	19	17	
194044	5	2	2	901	34	8	18	0	8	34	...	7	-5	23	4	8	6	-9	13	19	29	
194045	5	18	1	870	33	8	19	-2	8	24	...	8	-2	23	7	13	7	-17	13	22	21	
194046	4	2	1	846	30	7	12	-4	8	31	...	14	0	23	14	21	2	-10	12	14	11	
194047	-28	-29	-59	-85	24	-56	5	1	-6	5	...	5	-27	20	3	0	0	-14	-40	5	7	

194048 rows × 164 columns

**Figura 36.** Visualización de la matriz de datos. 194049 puntos de (intensidad, m/z) por 164 espectros.

Con el fin de compararlos correctamente se promediaron los valores de intensidad tanto para los 67 casos como los 97 controles para, normalizarlos, es decir, restarles la media y dividirlos por la desviación estándar, para ajustar los valores a un punto de referencia como se visualiza en la **Figura 37**. Nótese que en el eje x se encuentran el número de filas mencionadas anteriormente. La intención de estos dos pasos es mejorar la calidad de los datos y así mismo, facilitar el análisis por modelos como el PCA y los concernientes al aprendizaje supervisado. Lo que se observa del espectro de comparación es que no se ven muchas diferencias en las alturas de intensidades al ser superpuestos.





**Figura 37.** Datos de intensidades normalizadas de espectros de casos y controles de PE.

El PCA es una técnica para reducir la dimensionalidad de los datos y así encontrar el menor número de variables ahora llamadas componentes principales que expliquen la mayor parte de las anomalías o diferencias en los espectros. La dimensionalidad es un factor que siempre debe estudiarse cuando el conjunto de datos es grande, así como los grados de libertad y esto en parte tiene que ver que visualmente se pueden graficar hasta 3 dimensiones. En este caso el tener dos clases de datos (sí y no) se pueden comparar en dos dimensiones combinando las variables más importantes.

Al leer los datos en el cuadernillo lo que se forma es una matriz que contiene  $n$  filas (194049) y  $p$  columnas (164), entonces el tamaño de la matriz es  $n \times p$ . El PCA busca una transformación de cada vector de esa matriz para formar otra matriz de nuevos vectores, pero en otro espacio dimensional, estos son conocidos como *eigenvectores* que son múltiplos de los vectores originales y se organizan para que los primeros sean los que mayor varianza aporten en la reducción de la dimensionalidad, es decir, se deben escoger los primeros *eigenvectores* o primeros componentes principales. Para poder elegirlos se realizó un “diagrama de codo” que ayuda a identificar hasta cual componente vale la pena reducir la dimensionalidad, esto se observa cuando haya un cambio brusco en la varianza. En la **Figura 38** se demuestra dicho diagrama para los 8 componentes principales y efectivamente los 2 primeros componentes son los que mayor varianza aportan al momento de diferenciar entre las dos clases de datos, ver la **Figura 39**. Algo por recalcar es que el primer componente no llega ni a la mitad de la varianza explicada por lo que se puede decir que una clasificación entre casos y controles para la PE, utilizando datos de MALDI-TOF-MS por métodos de aprendizaje no supervisado de PCA no es lo suficientemente favorable. Se encontró una suma de varianza total para 8 componentes de 0,81.

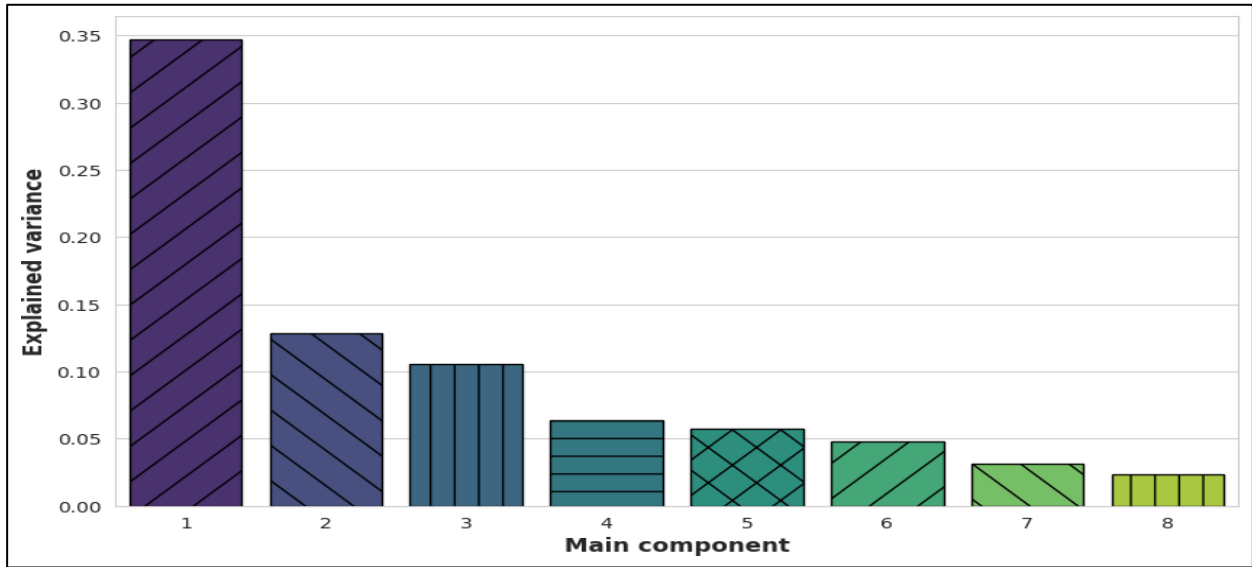


Figura 38. Diagrama de Codo para la elección de componentes principales.

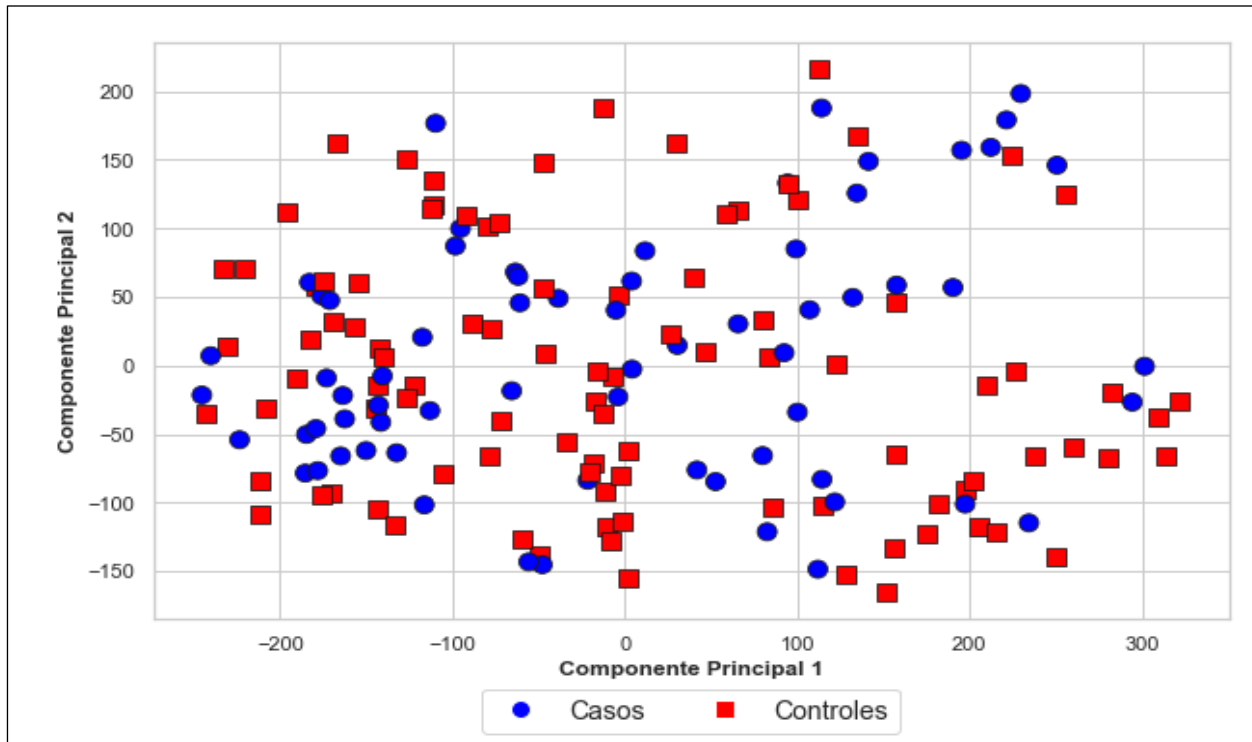


Figura 39. Diagrama de PCA para reducción de dimensionalidad a 2 componentes principales.

Por otro lado, existe el diagrama de pares que permitió visualizar la combinación de todos los componentes principales y poder ver con mayor claridad si los datos se agrupan, se separan o forman patrones, este se muestra en la **Figura 40**. La forma de las curvas es asimétrica en la mayoría de las combinaciones de componentes. Entre más anchas son las curvas significa que los datos están más dispersos, por ejemplo, la relación de todos los componentes con el componente 7 es el único que tiene las curvas más simétricas, es decir esa variable no aporta a la clasificación de las clases. Lo ideal es que en al menos un punto de la diagonal ambas curvas estuvieran alejadas a lo largo del eje x. La discrepancia en las alturas es debido al desbalance de datos entre casos y controles. Y en general las curvas presentan anchura de la cola positiva.

La literatura reporta el uso de esta herramienta discriminadora en varios campos de la investigación, un ejemplo encaminado a la identificación de perfil proteómico también a partir de espectros de masas, como el trabajo de Elbehiry (2023) [124]; se vio una clasificación de tres proteínas diferentes cuando graficaron los 3 componentes principales, además mostraron la gráfica de cuanto aporta cada componente a la varianza explicada. El primer componente tiene un valor de 26% y los demás van disminuyendo gradualmente hasta 10 componentes principales.

Por otro lado, el aporte de Correnti [125], un estudio similar a este, donde su objetivo era diagnosticar la infertilidad masculina a partir de MALDI-TOF-MS encontraron diferencias significativas entre pacientes fértiles e infértiles y se puede visualizar con el diagrama de componentes principales, lo importante para destacar en su trabajo es que las señales de masas fueron tratados por técnicas separativas, ellos ensayaron varios sorbentes como C8 y C18 antes de realizar el PCA, así pudieron detectar qué posibles péptidos hacían las diferencias en la separación. Adicionalmente, resaltan que la homogeneidad de los datos es algo importante para realizar el estudio discriminante.

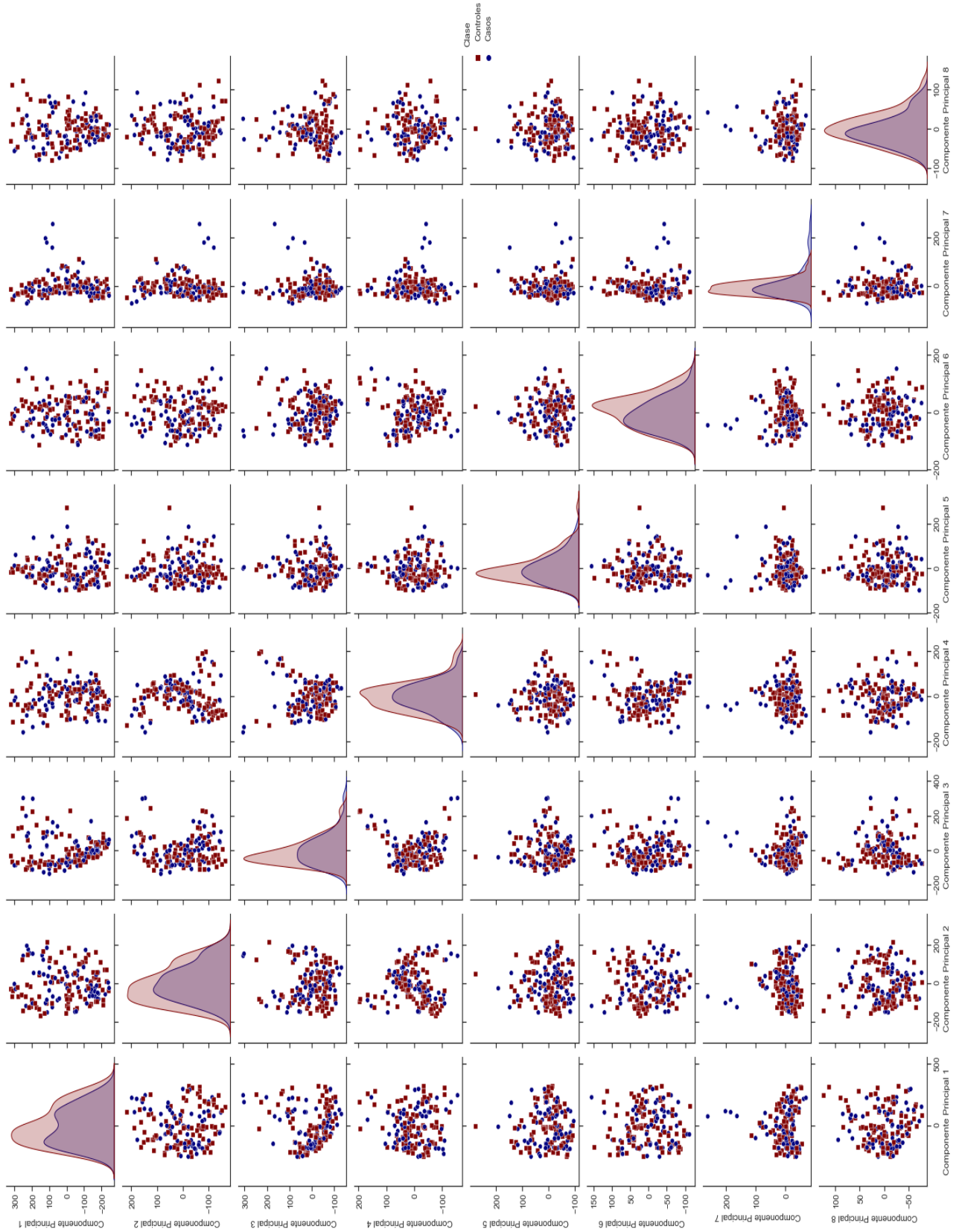
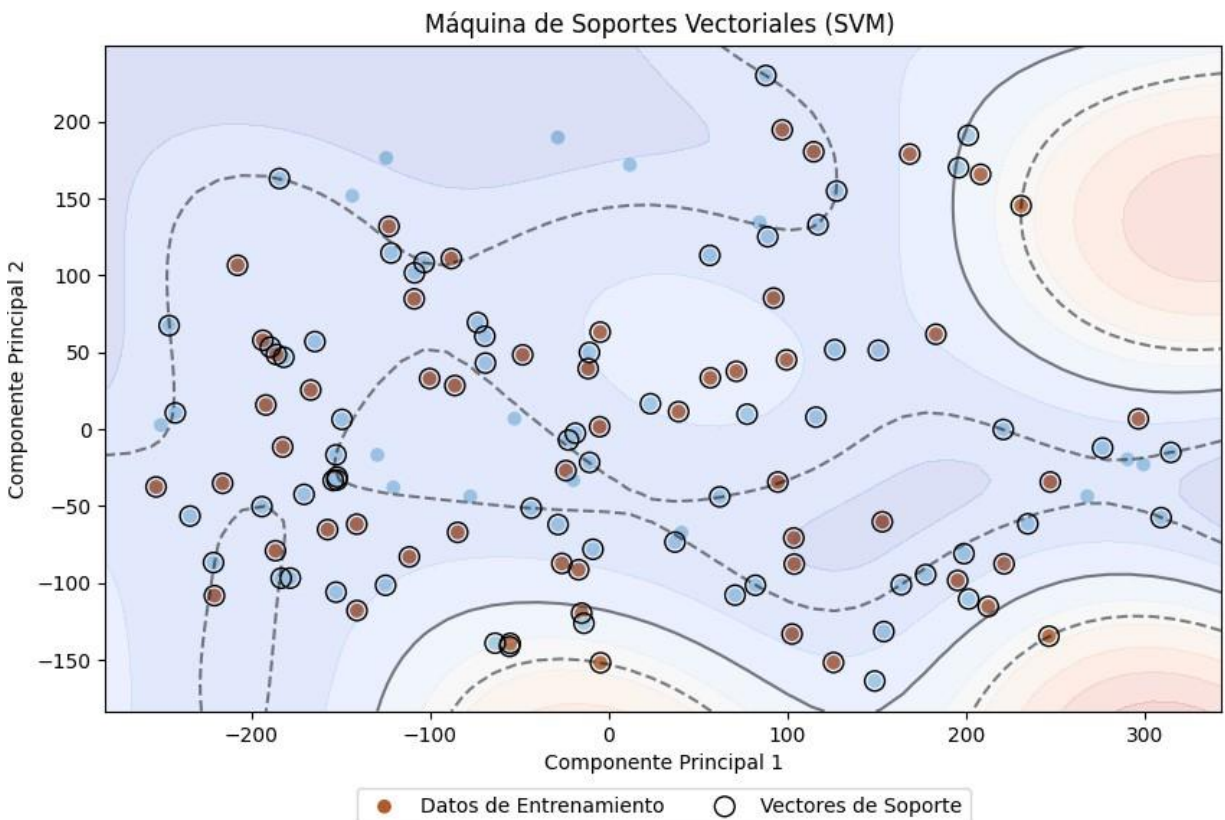


Figura 40. Diagrama de pares para relacionar los 8 componentes principales.

A continuación, se ensayaron modelos supervisados como SVM y RL utilizando la librería *Scikit-learn* y la función de división de entrenamiento y prueba, *train* y *test* respectivamente. En el entrenamiento no se ven como tal, valores que nos permitan ver y analizar ese proceso, solamente se encuentran y escogen los hiperparámetros, es por eso por lo que el conjunto de prueba permite evaluar el rendimiento de los modelos, evitar que se sobreajusten los datos y validar de cierta forma para generalizar la predicción de casos y controles.

El modelo SVM consiste en encontrar un hiperplano o varios, que se ajusten a la clasificación de las clases a partir de los hiperparámetros. Para SVM son: la regularización y el tipo de kernel, este último se refiere al grado del polinomio del hiperplano que a medida que aumenta es más difícil de visualizar. En nuestro caso el PCA nos dio una visión de cómo se encuentran distribuidos los datos, por lo que un kernel lineal y polinómico (plano) no serían suficientes para separarlos. Por eso se ensayó un kernel en base radial o Gaussiano (RBF) el cual, permite tener una clasificación más flexible como se observa en la **Figura 41**, donde se nota el hiperplano encontrado en datos de entrenamiento por el código en dos dimensiones utilizando las técnicas de PCA. El entrenamiento se hace aleatorio y repetidamente para encontrar el mayor valor aleatorio en el que los datos son mejor entrenados, y luego este se utiliza en el conjunto de datos de prueba para que encontremos el mejor valor de exactitud al momento de clasificar entre las clases. El mejor valor aleatorio fue 115313 con el cual, se encontró un valor de exactitud de 0,87 que significa que el 87% de las predicciones realizadas por el modelo SVM fueron correctas en el conjunto de prueba.



**Figura 41.** Ilustración de hiperplanos del modelo SVM en conjunto de entrenamiento.

La RL se especializa en la clasificación binaria estudiando la probabilidad de que una observación, en este caso las intensidades, se clasifique en alguno de dos grupos (caso y control), por lo cual se debe codificar a valores 1 y 0 como se hizo respectivamente. Para todos los modelos supervisados se buscó el mejor valor aleatorio en el entrenamiento y aplicarlo para encontrar la mejor exactitud en el conjunto de prueba. Para RL fue de 0,78. Métrica que se representa con la **Ecuación 16**.

En cuanto a RF, matemáticamente este modelo se enfoca en problemas donde no hay un patrón de clasificación y más bien se toma el trabajo de identificar la más mínima diferencia entre el conjunto de datos y así generar una “rama” de clasificación con base a arboles de decisión. Por lo que forma varios árboles con un valor de predicción que luego son ensamblados para elegir el mejor valor de predicción. Es una poderosa técnica de modelado, pero siempre se debe tener cuidado con el sobre ajuste ya que tiende a apegarse tanto a la clasificación que se obtuvo en los datos de entrenamiento, que cuando se aplica a datos de prueba no lo hace correctamente. En nuestro caso se encontró un valor de exactitud de 0,75.

El ensamblaje de los árboles que se mencionó anteriormente se conoce como técnica de *boosting*, a partir de esto nace otro modelo, XGBOOST, el cual, toma de referencia un árbol de decisión simple, pero lo va ajustando de acuerdo con dos funciones: una de objetivo que busca evitar el sobreajuste y otra de pérdida que mide la diferencia entre predicción y verdad y busca minimizarla para que el valor de exactitud sea la más generalizada en el modelo. Se forma un gradiente cuando se realiza una derivada a la función objetivo respecto a las predicciones. En nuestro caso el valor de exactitud para XGBOOST fue de 0,78.

En resumen, la **Tabla 11** muestra los valores de la métrica global, exactitud, representada por la **Ecuación 8** para cada modelo aplicado a los datos de prueba. Pero, debido a que la exactitud puede enmascarar las predicciones por clase cuando por lo general se cuenta, con un conjunto desbalanceado en etiquetas, como es nuestro caso, es pertinente evaluar otros parámetros. En la **Tabla 12** otras métricas de clasificación por clase que dan más información de cómo fue la predicción de cada modelo ante verdaderos y falsos casos y controles, de acuerdo con la **Ecuación 11**, **Ecuación 12** y **Ecuación 13** que fueron señaladas anteriormente. La representación gráfica de dicha tabla son las matrices de confusión que se ilustran en la **Figura 42**.

$$Exactitud = \frac{\text{número de predicciones correctas}}{\text{número de predicciones totales}}$$

**Ecuación 16.** Exactitud. Métrica global de evaluación de los modelos supervisados.

**Tabla 11.** Comparación de exactitud de los modelos predictivos supervisados.

Modelo	Valor de exactitud
SVM	0,88
RL	0,78
RF	0,75
XGBOOST	0,78

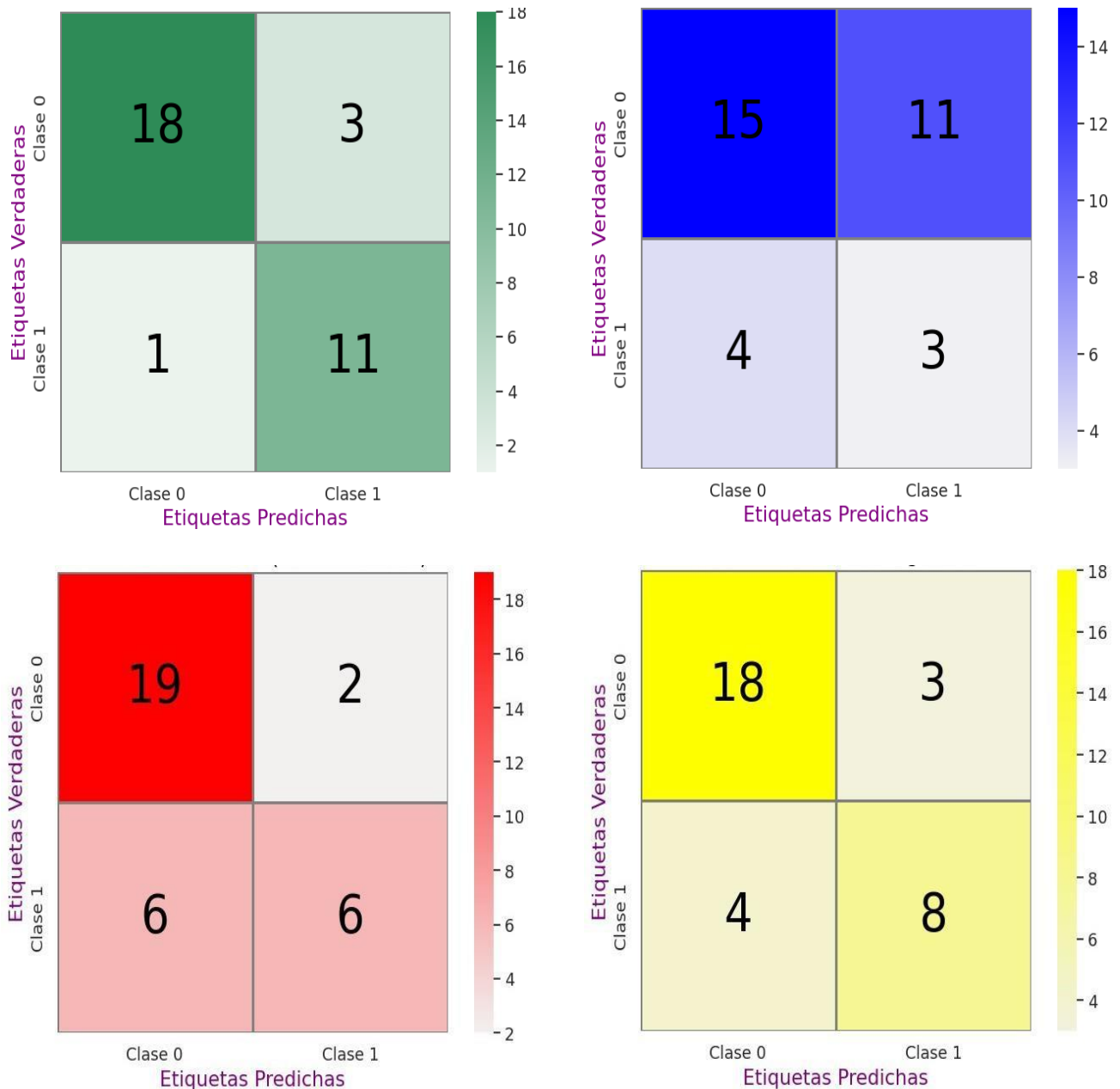
**Tabla 12.** Reporte de clasificación por clases de modelos supervisados.

MODELO	METRICAS DE EVALUACION POR CLASE							
	Precisión		Sensibilidad		Puntuación		Soporte	
CLASE	0	1	0	1	0	1	0	1
SVM	0,95	0,79	0,86	0,92	0,90	0,85	21	12
RL	0,79	0,21	0,58	0,43	0,67	0,29	26	7
RF	0,76	0,75	0,90	0,50	0,83	0,60	21	12
XGBOOST	0,82	0,75	0,86	0,67	0,84	0,70	21	12

Verdaderos Controles Verdaderos Negativos Verdaderos 0	Falsos Casos Falsos positivos Falsos 1
Falsos Controles Falsos Negativos Falsos 0	Verdaderos Casos Verdaderos positivos Verdaderos 1

Recordatorio de la **Figura 13**

Recordando la **Figura 13** se pueden interpretar las matrices de confusión y por ende las métricas de evaluación. Según las ecuaciones de las métricas, las palabras positivo y negativo se podrían igualar a casos y controles de PE, o en su efecto nomenclatura que se asumió para que los modelos entendieran que se trata de una clasificación binaria (1) y (0) respectivamente. En principio, el conjunto de prueba se eligió aleatoriamente entre 164 espectros, con un peso del 20% que representan 33 espectros entre los cuales hay una porción de casos (1) y controles (0) y estarán distribuidos aleatoriamente en las 4 secciones de la matriz de confusión.



Verde (SVM), azul (RL), rojo (RF) y amarillo (XGBOOST)

**Figura 42.** Matrices de confusión para modelo supervisado.

Para leer las matrices se debe tener en cuenta que cada fila y columna esta nombrada con una clase, en nuestro caso es 0 o 1. La diagonal principal indica el número de elementos que en cada clase fueron predichos correctamente por el modelo y la diagonal inversa indica el número de elementos e los que el modelo falló. Entonces la idealidad consiste en que la diagonal inversa tuviera valor de 0. Debido a que es muy difícil que la idealidad se dé, entonces la métrica de precisión indica cuantos de los casos predichos como positivos son realmente casos. Asimismo, la sensibilidad expresa la cantidad de casos reales que se identificaron



correctamente. Este análisis también se aplica y concuerda con los valores mostrados en las matrices de confusión.

Por ejemplo, en la **Ecuación 17** la precisión con la que el modelo SVM predice realmente los casos respecto a todos los predichos como casos es del 79 %, de igual forma en la **Ecuación 18** para los controles el modelo SVM predice realmente los controles respecto a todos los controles predichos es del 95%.

$$\textit{Precision SVM Casos (1)} = \frac{V(1)}{V(1) + F(1)} = \frac{11}{11 + 3} = \frac{11}{14} = 0,79$$

**Ecuación 17.** Ejemplo de cálculo de la métrica precisión para casos en el modelo SVM.

$$\textit{Precision SVM Controles (0)} = \frac{V(0)}{V(0) + F(0)} = \frac{18}{18 + 1} = \frac{18}{19} = 0,95$$

**Ecuación 18.** Ejemplo de cálculo de métrica precisión para controles en el modelo SVM.

La diagonal principal de la matriz del modelo SVM (verde) representa la cantidad de espectros predichos correctamente de casos y controles por el modelo (verdaderos). En cambio, la diagonal inversa muestra la cantidad de espectros que fueron predichos falsamente. Esto se puede comprobar con los valores de soporte, donde si sumamos las filas de la matriz dan 21 controles (0) y 12 casos (1). En este contexto, por ejemplo, la sensibilidad ante los casos que tiene el modelo RL es del 43 % para casos y del 58% para los controles. Esto se ve numéricamente en la **Ecuación 19** y **Ecuación 20**.

$$\textit{Sensibilidad RL Casos (1)} = \frac{V(1)}{V(1) + F(0)} = \frac{3}{3 + 4} = \frac{3}{7} = 0,43$$

**Ecuación 19.** Ejemplo de cálculo de métrica sensibilidad para casos en el modelo RL.

$$\textit{Sensibilidad RL Controles (0)} = \frac{V(0)}{V(0) + F(1)} = \frac{15}{15 + 11} = \frac{15}{26} = 0,58$$

**Ecuación 20.** Ejemplo de cálculo de métrica sensibilidad para controles en el modelo RL.

Por último, la puntuación F1, es una métrica de combinación armónica entre la precisión y sensibilidad cuando se quiere dar la misma importancia a las dos métricas de clasificación. Que sea tanto preciso y sensible a la hora de predecir cuando una mujer padezca PE o no, ya que se trata de la salud de 2 seres vivos. La **Ecuación 21** y **Ecuación 22** muestran el cálculo para el modelo RF.

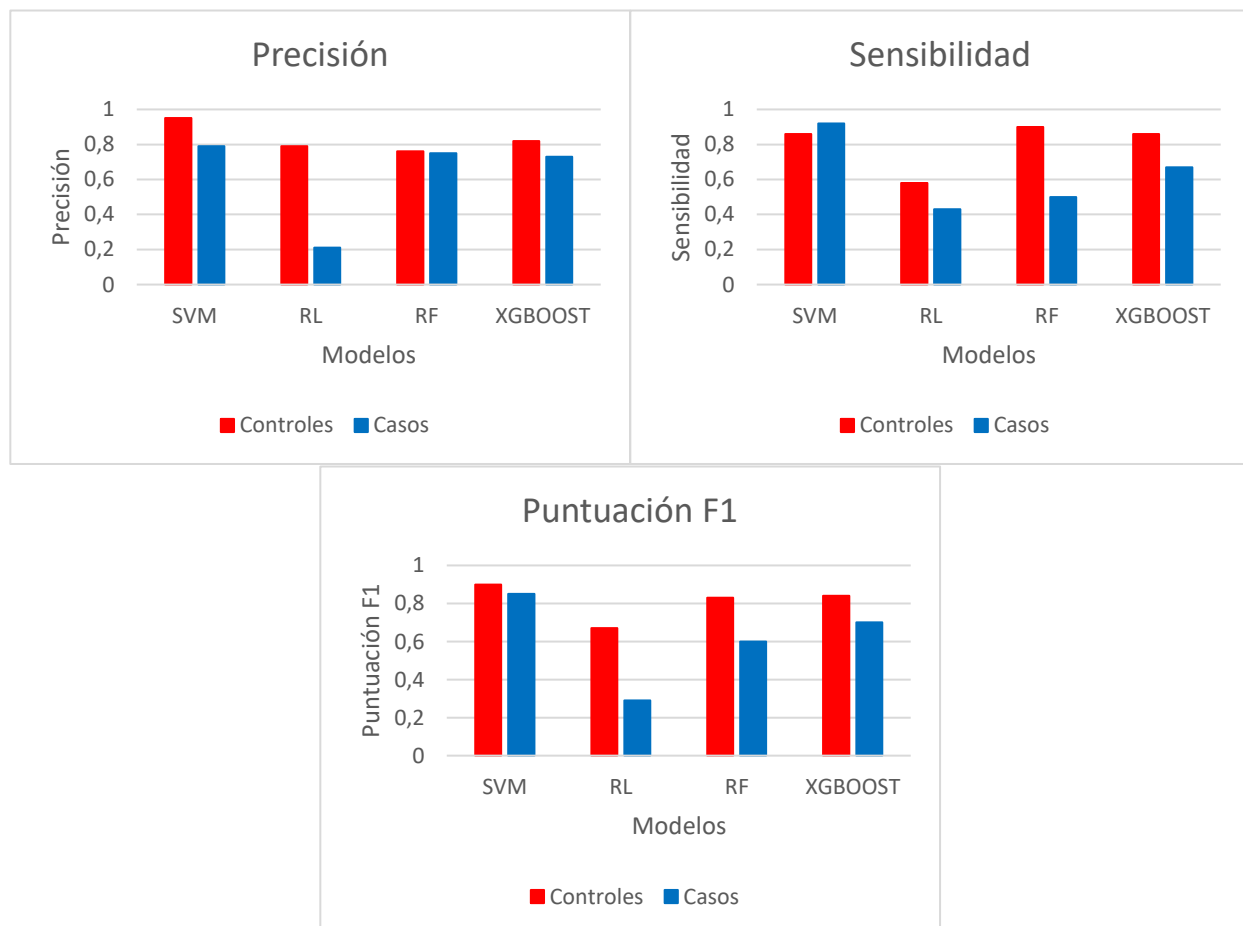
$$Puntuación\ F1\ RF\ Casos(1) = \frac{2\ Precisión\ (1) \times\ Sensibilidad\ (1)}{Precisión\ (1) +\ Sensibilidad\ (1)} = \frac{2 \times 0,75 \times 0,50}{0,75 + 0,50} = 0,60$$

**Ecuación 21.** Ejemplo de cálculo de puntuación F1 para casos del modelo RF.

$$Puntuación\ F1\ RF\ Controles(0) = \frac{2\ Precisión\ (0) \times\ Sensibilidad\ (0)}{Precisión\ (0) +\ Sensibilidad\ (0)} = \frac{2 \times 0,76 \times 0,90}{0,76 + 0,90} = 0,83$$

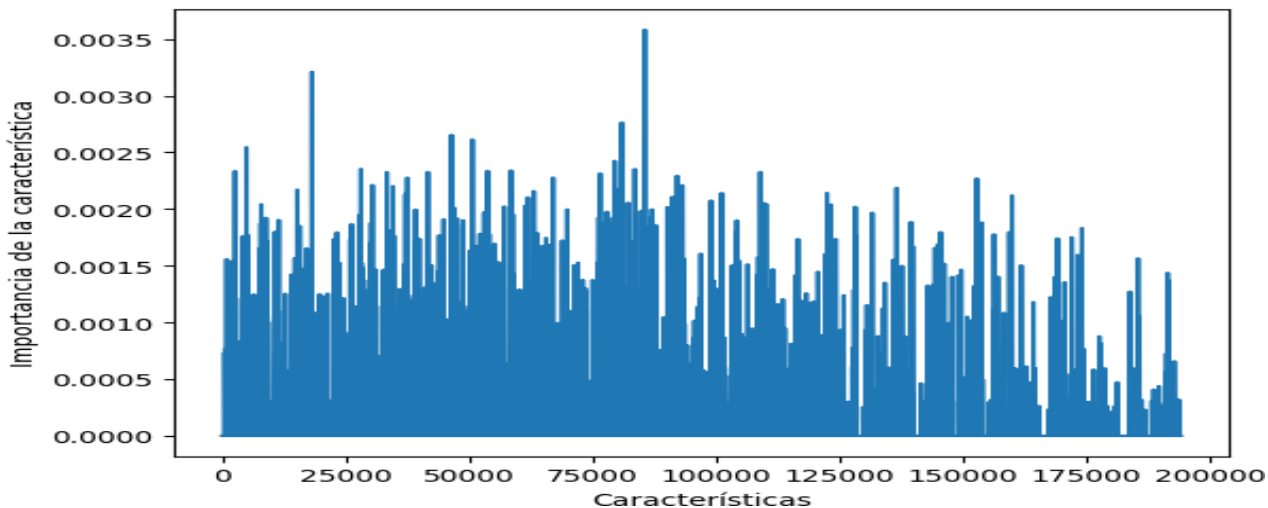
**Ecuación 22.** Ejemplo de cálculo de puntuación f1 para controles del modelo RF.

De acuerdo con la explicación de la puntuación F1, SVM fue el mejor modelo capaz de tener esa armonía entre la precisión y sensibilidad tanto a la hora de diferenciar casos como de controles con un valor de 0,85 y 0,90 respectivamente. Las gráficas de barras de la **Figura 43** muestra que la RL fue el modelo que menor precisión y sensibilidad demostró al momento de clasificar casos de PE y el de menor sensibilidad en cuanto a catalogar controles. También se infiere que RF fue el modelo con mayor sensibilidad para predecir correctamente controles de PE. En general SVM obtuvo un mejor desempeño predictivo con un valor del 88 % en términos de exactitud y precisión con un valor de 95% de controles y 79% de casos, pero en sensibilidad de controles fue superado por el algoritmo de RF con una diferencia del 4%.



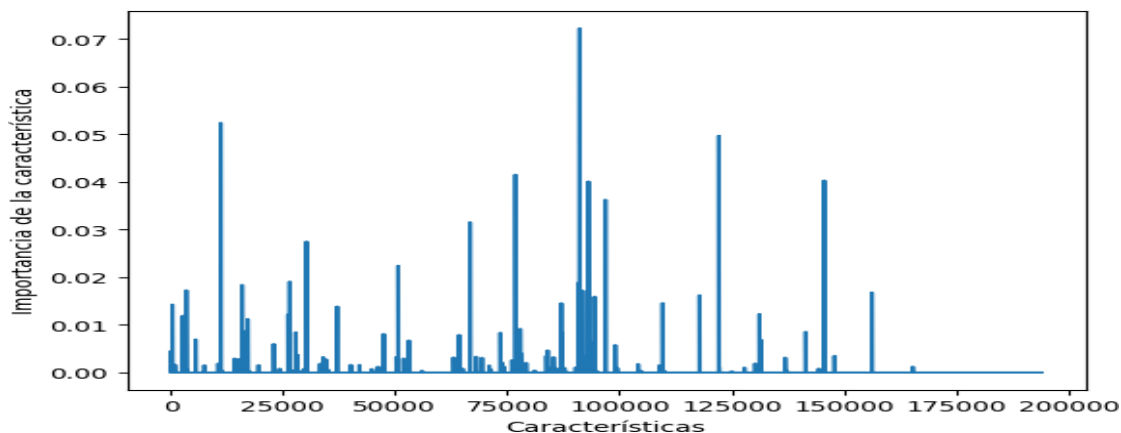
**Figura 43.** Gráfico de barras de comparación de métricas por clase de cada modelo supervisado.

Como adición, aprovechando que se utilizaron modelos basados en arboles de decisión, RF y XGBOOST, estos tienen la particularidad de poder identificar las características más importantes que se calculan para RF como la reducción de impurezas o la medida de que tan mezcladas se encuentran las clases ya que a partir de estos los árboles de decisión clasifican los casos y controles. En cambio, para XGBOOST la importancia de la característica es la ganancia al realizar la clasificación de los árboles de decisión.



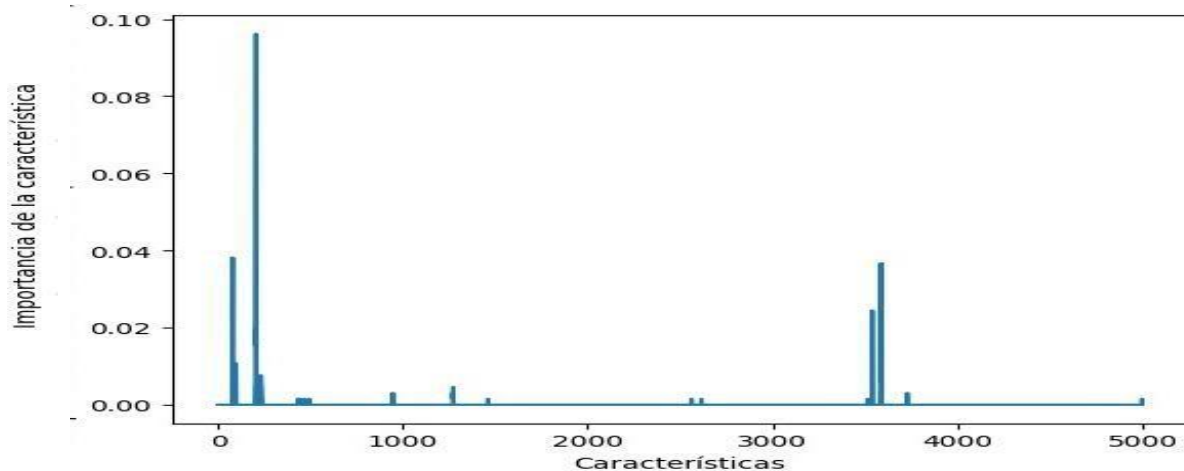
**Figura 44.** Selección de características a partir del modelo RF.

Al comparar la **Figura 44** con la **Figura 45**, se observa que RF exhibe más características, sugiriendo posiblemente una mayor permisibilidad al tomar decisiones en comparación con XGBOOST. La dependencia de la naturaleza de los datos queda descartada, dado que, a pesar de tratarse de una clasificación binaria, todas las muestras de S.S fueron analizadas por MALDI-TOF-MS bajo las mismas condiciones. Al examinar las etiquetas de los ejes X de estas figuras, se constata que las características no exceden el número de entradas de intensidades del espectro. También, podría atribuirse para XGBOOST que en el eje Y de la importancia de las características, estas poseen una relevancia numérica mayor que las encontradas en RF. Incluso con dicha importancia y la menor cantidad de características, no se descarta que pudiera deberse a señales de ruido en los espectros, que han sido interpretados como importantes al diferenciar clases.



**Figura 45.** Selección de características a partir del modelo XGBOOST.

Por otro lado, en un rango de 0 a 5000 se eligieron las características más importantes en la permutación de SVM, donde se obtuvo un perfil de características, ver **Figura 46**, que podrían relacionarse con bases de datos para identificar posibles biomarcadores proteicos, es pertinente aclarar que esas señales no corresponden a secuencias de péptidos o metabolitos, esas señales extraídas son intensidades relativas que hacen que se diferencien espectros de pacientes con PE de pacientes sin PE, por lo que no sería correcto clasificar este trabajo como uno más en el campo de la proteómica o metabolómica, por supuesto se extraen herramientas de ambas ramas pero realmente es un modelo a partir de espectros de masas a los cuales se les aplicó aprendizaje automatizado con el fin de dar un resultado de predictibilidad ante una enfermedad que en la actualidad aún no tiene una explicación de origen y que muestra efectividad del 88%. El siguiente paso es realizar separaciones cromatográficas, electroforéticas e inversión en bases de datos para la validación con expertos en el campo clínico.



**Figura 46.** Huella peptídica de intensidades relativas para los pacientes de casos PE.

Algunos autores como Pusch [126]. y colaboradores han visto en esta selección de características una herramienta de diagnóstico también para cáncer de próstata, donde los picos de diagnóstico se obtienen a partir de árboles de decisión y que, se pueden correlacionar con biomarcadores proteómicos que den un indicio de la enfermedad. Por otro lado, trabajos como el de Liu Mengyan y colaboradores [127], que tuvieron como objetivo analizar datos clínicos de mujeres con PE, también resalta la implementación de SVM con una exactitud de 0,68 y una precisión del 0,51 para casos. Sin embargo, el modelo RF para ellos fue el de mejor precisión de casos con un valor de 0,86. También Aljameel [81] resalta la importancia de la cantidad y calidad de los datos espectrométricos, ya que, entre más datos de entrenamiento, los modelos aprenderán de una manera más generalizada, robusta y con mayor capacidad predictiva.

Es posible pensar en el estudio de muestras intactas, Sarkar [128], por ejemplo, realizó una comparación con estudios que emplearon muestras proteómicas intactas y obtuvieron buenos resultados, especialmente en la clasificación binaria. Estos estudios resaltan la versatilidad de las técnicas de análisis de masas, destacando la efectividad tanto en muestras digeridas como en muestras proteómicas intactas, lo que respalda la idoneidad de

la metodología utilizada en este estudio y una posible propuesta a realizar estudios parecidos a este, pero con sueros intactos.

Por último, este trabajo, más allá de mostrar un prototipo de modelo diagnóstico para la PE, muestra cómo se pueden enlazar varias ciencias como la medicina, química y ciencia de datos para tratar una cantidad considerable de datos de manera más eficiente. La comparación con otros trabajos demostró que estas técnicas han sido antes utilizadas, pero gracias a la creatividad que es concerniente a la humanidad se pueden obtener muchas investigaciones que trasciendan las habilidades de estos.

## **11. ANEXOS**

En versión física se encuentran en el CD al final de las páginas, y en versión digital se pueden ver en el siguiente enlace de Google drive.

[https://drive.google.com/drive/folders/1XRC4IUSeKu\\_OK6pif\\_gJLgUmM5U5bXXA?usp=sharing](https://drive.google.com/drive/folders/1XRC4IUSeKu_OK6pif_gJLgUmM5U5bXXA?usp=sharing)

## 12. CONCLUSIONES

En el desarrollo de este trabajo se pudo llevar a cabo la digestión asistida por la metodología FASP de 164 muestras de suero sanguíneo de pacientes casos y control de Preeclampsia, el resultado de esta primera etapa fue el conjunto de 164 muestras de fragmentos peptídicos secados al vacío. Al tener listos los fragmentos, se observó que el análisis de muestras por MALDI-TOF-MS depende de múltiples factores. Adicionalmente, tratándose muestras biológicas que, aunque comparten una misma naturaleza, provienen de diferentes sujetos, se llevó a cabo un diseño experimental factorial mixto 3x3x2 para evaluar los 3 factores (la dilución, el tipo de matriz y el tipo de sembrado en la placa) que, se consideró, eran los más influyentes en la deposición de las muestras en MALDI-TOF-MS, y que se podrían relacionar con una variable respuesta que fue encontrar un espectro de calidad con el mayor número de señales a una determinada altura o intensidad de abundancia relativa. También se evaluó otro factor como el volumen de redisolución de los fragmentos peptídicos, pero no se incluyó en el diseño de experimentos debido a que trabajar con otros volúmenes representaba inconsistencias en la continuidad del procedimiento, por lo cual se adoptó un nivel fijo de 50 uL para ese factor. El resultado del diseño experimental indicó que las condiciones que se utilizaron fueron: dilución de 1:10, tipo de matriz de HCCA y tipo de sembrado de doble capa. Esto indica el cumplimiento satisfactorio del primer objetivo específico que era ajustar las condiciones para la obtención de espectros de masas MALDI-TOF y permitió seguir desarrollando la investigación.

La segunda etapa consistió en adquirir los espectros de masas en el que se aplicaron las condiciones encontradas en el diseño experimental para las 164 muestras de fragmentos peptídicos, el resultado principal fue la construcción de una carpeta con todos los datos de espectros de masas en formato .txt, cumpliendo así con el segundo objetivo específico propuesto al principio de la investigación. También se logró adquirir experiencia en el manejo del equipo MALDI-TOF.

El análisis no supervisado produjo resultados poco prometedores porque al reducir la dimensionalidad con el método de PCA a 8 componentes principales no se vio una separación entre las clases caso y control de PE. De otro lado, el análisis supervisado mejoró considerablemente la exactitud para el modelo de SVM (88%) que fue el elegido como el modelo que mejor desempeño mostró al momento de predecir o diferenciar entre un caso y control. Sin embargo, otros tres modelos también fueron analizados (RL, RF y XGBOOST) y evaluados bajo métricas de clasificación por clases como la precisión, sensibilidad y puntuación F1 estimados por matrices de confusión. Así se concluye que se ha cumplido el tercer objetivo con la etapa 3 y, por tanto, con el objetivo general de la investigación.

De forma adicional, y mostrando un potencial superior de la indagación, se obtuvo un resultado que indica una selección de características más importantes elegidas bajo los modelos basados en arboles de decisión que podrían relacionarse con posibles biomarcadores moleculares en la enfermedad de PE. De otro lado, ante la duda de si se perdía muestra proteica por el uso del filtro en la metodología FASP se realizó una comparación en la concentración de proteínas de sueros sanguíneos intactos y la redisolución de fragmentos, encontrando que se perdió aproximadamente el 99,95% del material proteico. Sin embargo, el 0,05 % restante fue suficiente para

el análisis y detección de fragmentos peptídicos por MALDI-TOF-MS.

En definitiva, la investigación permite centrar el foco de los estudios químicos en la exploración de nuevas áreas como lo es la ciencia de datos y la programación, que aun sin ser una rama de los estudios de pregrado en química de la Universidad del Cauca, demuestran que hoy en día en campos tan importantes como la medicina y la bioquímica son indispensables por la cantidad de datos que se generan en los laboratorios y no son aprovechados de la mejor manera posible. Esto deja en evidencia la importancia de preparar a los químicos de la actualidad de una forma más contundente para que se puedan enfrentar de una mejor forma a la interdisciplinariedad inherente a la investigación de hoy en día.

Aunque se da un resultado fijo de exactitud, el amplio rango de aplicación de la ciencia de datos puede favorecer el refinamiento de este modelo, ya sea por la inclusión de más datos o la mejora en la calidad de estos, o incluso puede facilitar el inicio de investigaciones derivadas que puedan complementar la aquí presentada.

### 13. BIBLIOGRAFIA

- [1] S. A. Karrar y P. L. Hong, "Preeclampsia," in StatPearls [Internet], StatPearls Publishing, 2023.
- [2] J. P. Calvo, Y. P. Rodríguez, L. Q. Figueroa, "Actualización en preeclampsia. Revista médica sinergia", 5(01), 345, 2020.
- [3] J. R. Wiśniewski, "Filter aided sample preparation—a tutorial," *Anal Chim Acta*, vol. 1090, pp. 23–30, 2019.
- [4] M. Feucherolles, M. Nenning, S. Becker, et al., "Combination of MALDI-TOF mass spectrometry and machine learning for rapid antimicrobial resistance screening: The case of *Campylobacter* spp.," *Front Microbiol*, vol. 12, p. 804484, 2022.
- [5] M. A. Farrán, S. L. Cabanillas, y J. R. Cabero, "Machine learning aplicado a la química," Canal UNED, Ciencias en radio 3, Jun. 25, 2019.
- [6] O. M. de la Salud. OMS, "Recomendaciones de la OMS sobre el tratamiento farmacológico de la hipertensión arterial leve o moderada en el embarazo", recuperado de [\[https://iris.paho.org/bitstream/handle/10665.2/56658/9789275326350\\_spa.pdf?sequence=1&isAllowed=y\]](https://iris.paho.org/bitstream/handle/10665.2/56658/9789275326350_spa.pdf?sequence=1&isAllowed=y).
- [7] O. P. de la S. OPS, "Día de Concientización sobre la Preeclampsia", recuperado de [\[https://www.paho.org/es/noticias/1-8-2019-dia-concientizacion-sobre-preeclampsia#:~:text=A%20nivel%20mundial%2C%20la%20preeclampsia,son%20provocados%20por%20problemas%20hipertensivos\]](https://www.paho.org/es/noticias/1-8-2019-dia-concientizacion-sobre-preeclampsia#:~:text=A%20nivel%20mundial%2C%20la%20preeclampsia,son%20provocados%20por%20problemas%20hipertensivos), 2019.
- [8] I. N. de S. INS and Minsalud, "Boletín Epidemiológico Semanal," Bogotá, May 2021; recuperado de [\[chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2021\\_Boletin\\_epidemiologico\\_semana\\_8.pdf\]](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2021_Boletin_epidemiologico_semana_8.pdf), 2021.
- [9] N. C. Serrano, E. Guio-Machena, D.C. Quintero, et al., "Lipid profile, plasma apolipoproteins, and pre-eclampsia risk in the GenPE case-control study.," *Atherosclerosis*, vol. 276, pp. 189–194, Sep. 2018, doi: 10.1016/j.atherosclerosis.2018.05.051.
- [10] J. Williams y G. Cunningham, "Hipertensión en el embarazo," in *Obstetricia de Williams*, 23rd ed., Ciudad de México: McGraw-Hill Interamericana, pp. 710–756., 2011.
- [11] T. Sundsten, y H. Ortsäter, "Proteomics in diabetes research. *Molecular and Cellular Endocrinology*", 297(1-2), 93-103, 2009.
- [12] S. Barceló-Batllori, y R. Gomis, "Proteomics in obesity research," *PROTEOMICS—Clinical Applications*, 3(2), 263-278, 2009.
- [13] E. Drummond, y T. Wisniewski, "Using proteomics to understand Alzheimer's disease pathogenesis," *Exon Publications*, 37-51, 2019.
- [14] P. Sousa, L. Silva, C. Luís, J. S. Câmara, y R. Perestrelo, "MALDI-TOF MS: A Promising Analytical Approach to



Cancer Diagnostics and Monitoring,” *Separations*, vol. 10, no. 8, p. 453, Aug. 2023, doi: 10.3390/separations10080453.

- [15] NIH (Institutos Nacionales de Salud), “Información sobre el embarazo,” recuperado de [ <https://espanol.nichd.nih.gov/salud/temas/pregnancy/informacion>]. Mar. 2020.
- [16] D. Lang, “Hitos del desarrollo-Hitos del desarrollo,” en *PROBLEMAS DE CRIANZA Y DIVERSIDAD FAMILIAR (LANG)*, Iowa State University, recuperado de [ [https://espanol.libretexts.org/Ciencias\\_Sociales/Libro%3A\\_Problemas\\_de\\_crianza\\_y\\_diversidad\\_familiar\\_\(Lang\)/07%3A\\_A\\_Hitos\\_del\\_desarrollo](https://espanol.libretexts.org/Ciencias_Sociales/Libro%3A_Problemas_de_crianza_y_diversidad_familiar_(Lang)/07%3A_A_Hitos_del_desarrollo)].
- [17] J. E. Hall, “Guyton y Hall. Tratado de fisiología médica. Elsevier Health Sciences,” 2011.
- [18] M. D. Lindheimer, R. N. Taylor, J. M. Roberts, F. G. Cunningham, and L. Chesley, “Introduction, History, Controversies, and Definitions,” in *Chesley’s Hypertensive Disorders in Pregnancy*, 4th ed., Elsevier, 2015, pp. 1–24. doi: 10.1016/B978-0-12-407866-6.00001-8.
- [19] L. C. Chesley, “A short history of eclampsia,” *Obstetrics & Gynecology*, vol. 43, no. 4, pp. 599–602, 1974.
- [20] Meiat, J. *Hippocratis Magni Coacae Praenotiones: Opus Admirabile in tres libros distributum; cum rerum commemorabilium indice amplissimo*, 1621.
- [21] J. Williams, y G. Cunningham, “Hipertensión en el embarazo,” in *Obstetricia de Williams*, 23rd ed., Ciudad de México: McGraw-Hill Interamericana, pp. 706–708, 2011.
- [22] L. V. Beltrán Chaparro, P. Benavides, J. A. López Rios, and W. Onatra Herrera, “Estados hipertensivos en el embarazo: revisión,” *Revista UDCA Actualidad & Divulgación Científica*, vol. 17, no. 2, pp. 311–323, 2014.
- [23] Real Academia Nacional De Medicina De España, “Diccionario de Términos Médicos,” recuperado de [ <https://dtme.ranm.es/busador.aspx>].
- [24] S. E. Sánchez, “Actualización en la epidemiología de la preeclampsia: update,” *Revista Peruana de Ginecología y Obstetricia*, vol. 60, no. 4, pp. 309–320, 2014.
- [25] L. Badimón y J. Martínez-González, “Disfunción endotelial,” *Revista Española de Cardiología Suplementos*, vol. 6, no. 1, pp. 21A-30A, Jan. 2006, doi: 10.1016/S1131-3587(06)74817-8.
- [26] C. W. G. Redman y I. L. Sargent, “Placental Stress and Pre-eclampsia: A Revised View,” *Placenta*, vol. 30, pp. 38–42, Mar. 2009, doi: 10.1016/j.placenta.2008.11.021.
- [27] A. Negre-Salvayre, A. Swiader, R. Salvayre, y P. Guerby, “Oxidative stress, lipid peroxidation and premature placental senescence in preeclampsia,” *Arch Biochem Biophys*, vol. 730, p. 109416, Nov. 2022, doi: 10.1016/j.abb.2022.109416.
- [28] M. A. Opichka, M. W. Rappelt, D. D. Gutterman, J. L. Grobe, and J. J. McIntosh, “Vascular Dysfunction in Preeclampsia,” *Cells*, vol. 10, no. 11, p. 3055, Nov. 2021, doi: 10.3390/cells10113055.
- [29] C. A. Hubel, J. M. Roberts, R. N. Taylor, et al., “Lipid peroxidation in pregnancy: new perspectives on preeclampsia,” *American journal of obstetrics and gynecology*, 161(4), 1025-1034, 1989.
- [30] C. A. Labarrere y W. P. Faulk, “Anchoring villi in human placental basal plate: Lymphocytes, macrophages and

coagulation," *Placenta*, vol. 12, no. 2, pp. 173–182, Mar. 1991, doi: 10.1016/0143-4004(91)90021-7.

[31] V. NEWMAN and J. FULLERTON, "Role of nutrition in the prevention of preeclampsia \*1Review of the literature," *J Nurse Midwifery*, vol. 35, no. 5, pp. 282–291, Sep. 1990, doi: 10.1016/0091-2182(90)90081-F.

[32] A. C. Ariza, N. A. Bobadilla, and A. Halhali, "[Endothelin 1 and angiotensin II in preeclampsia].," *Rev Invest Clin*, vol. 59, no. 1, pp. 48–56, 2007.

[33] L. C. Chesley, "Diagnosis of preeclampsia.," *Obstetrics and gynecology*, vol. 65, no. 3, pp. 423–425, 1985.

[34] P. August and B. M. Sibai, "Preeclampsia: Clinical features and diagnosis," *UpToDate* Accessed December, vol. 22, 2017.

[35] L. K. Wagner, "Diagnosis and management of preeclampsia," *Am Fam Physician*, vol. 70, no. 12, pp. 2317–2324, 2004.

[36] J. M. Sánchez-Romero and J. M. González-Buitrago, "Biobancos, laboratorios clínicos e investigación biomédica," *Revista del Laboratorio Clínico*, vol. 3, no. 4, pp. 201–205, Oct. 2010, doi: 10.1016/j.labcli.2010.09.001.

[37] B. D. Vacutainer, "Sistema de extracción de sangre al vacío," recuperado de [chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://static.bd.com/documents/eifu/VDP40161\_ES.pdf]

[38] E. M. Keohane, "Blood specimen collection," *Rodak's Hematology: Clinical Principles and Applications*, p. 19, 2015.

[39] J. N. Perry, A. Jasim, A. Hojat, and W. H. Yong, "Procurement, Storage, and Use of Blood in Biobanks," in *Biobanking Methods and Protocols*, vol. 1897, Humana Press, 2019, pp. 89–97. doi: 10.1007/978-1-4939-8935-5\_9.

[40] E. M. Keohane, C. N. Otto, y J. M. Walenga, "Rodak's hematology-e-book: clinical principles and applications," Elsevier Health Sciences, 2019.

[41] Z. Zhang, S. Wu, D. L. Stenoién, y L. Paša-Tolić, "High-Throughput Proteomics," *Annual Review of Analytical Chemistry*, vol. 7, no. 1, pp. 427–454, Jun. 2014, doi: 10.1146/annurev-anchem-071213-020216

[42] M. Carrera, "Proteómica bottom-up y top-down de organismos poco secuenciados. secuenciación de novo de péptidos biomarcadores para la identificación de especies de la familia merlucciidae ," *Universidad de Vigo, instituto de investigaciones marinas*, 2008.

[43] D. J. Swiner, S. Jackson, B. J. Burris, y A. K. Badu-Tawiah, "Applications of Mass Spectrometry for Clinical Diagnostics: The Influence of Turnaround Time," *Anal Chem*, vol. 92, no. 1, pp. 183–202, Jan. 2020, doi: 10.1021/acs.analchem.9b04901.

[44] T. Cabras, E. Pisano, C. Montaldo, et al., "Significant Modifications of the Salivary Proteome Potentially Associated with Complications of Down Syndrome Revealed by Top-down Proteomics," *Molecular & Cellular Proteomics*, vol. 12, no. 7, pp. 1844–1852, Jul. 2013, doi: 10.1074/mcp.M112.026708.

[45] I. Messana, T. Cabras, R. Inzitari, et al., "Characterization of the Human Salivary Basic Proline-Rich Protein Complex by a Proteomic Approach," *J Proteome Res*, vol. 3, no. 4, pp. 792–800, Aug. 2004, doi: 10.1021/pr049953c.

[46] M. Palmblad, A. Tiss y R. Cramer, "Mass spectrometry in clinical proteomics – from the present to the future,"

Proteomics Clin Appl, vol. 3, no. 1, pp. 6–17, Jan. 2009, doi: 10.1002/prca.200800090.

[47] E. F. Petricoin, D. A. Fishman, T. P. Conrads, T. D. Veenstra, y L. A. Liotta, “Lessons from Kitty Hawk: From feasibility to routine clinical use for the field of proteomic pattern diagnostics,” *Proteomics*, vol. 4, no. 8, pp. 2357–2360, Aug. 2004, doi: 10.1002/pmic.200400865.

[48] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, y C. M. Whitehouse, “Electrospray Ionization for Mass Spectrometry of Large Biomolecules,” *Science* (1979), vol. 246, no. 4926, pp. 64–71, Oct. 1989, doi: 10.1126/science.2675315.

[49] Michael. Karas y Franz. Hillenkamp, “Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons,” *Anal Chem*, vol. 60, no. 20, pp. 2299–2301, Oct. 1988, doi: 10.1021/ac00171a028.

[50] S. A. Pshenichnyuk y N. L. Asfandiarov, “The Role of Free Electrons in Matrix-Assisted Laser Desorption/Ionization: Electron Capture by Molecules of  $\alpha$ -Cyano-4-Hydroxycinnamic Acid,” *European Journal of Mass Spectrometry*, vol. 10, no. 4, pp. 477–486, Aug. 2004, doi: 10.1255/ejms.650.

[51] J. Gobom, M. Schuerenberg, M. Mueller, D. Theiss, H. Lehrach, y E. Nordhoff, “ $\alpha$ -Cyano-4- hydroxycinnamic Acid Affinity Sample Preparation. A Protocol for MALDI-MS Peptide Analysis in Proteomics,” *Anal Chem*, vol. 73, no. 3, pp. 434–438, Feb. 2001, doi: 10.1021/ac001241s.

[52] S. L. Cohen y B. T. Chait, “Influence of Matrix Solution Conditions on the MALDI-MS Analysis of Peptides and Proteins,” *Anal Chem*, vol. 68, no. 1, pp. 31–37, Jan. 1996, doi: 10.1021/ac9507956.

[53] R. C. Beavis y B. T. Chait, “[22] Matrix-assisted laser desorption ionization mass-spectrometry of proteins,” 1996, pp. 519–551. doi: 10.1016/S0076-6879(96)70024-1.

[54] R. C. Beavis, T. Chaudhary, y B. T. Chait, “ $\alpha$ -Cyano-4-hydroxycinnamic acid as a matrix for matrixassisted laser desorption mass spectrometry,” *Organic Mass Spectrometry*, vol. 27, no. 2, pp. 156–158, Feb. 1992, doi: 10.1002/oms.1210270217.

[55] K. Strupat, M. Karas, y F. Hillenkamp, “2,5-Dihydroxybenzoic acid: a new matrix for laser desorption—ionization mass spectrometry,” *Int J Mass Spectrom Ion Process*, vol. 111, pp. 89–102, Dec. 1991, doi: 10.1016/0168-1176(91)85050-V.

[56] T. Y. Hou, C. Chiang-Ni, y S. H. Teng, “Current status of MALDI-TOF mass spectrometry in clinical microbiology,” *Journal of food and drug analysis*, 27(2), 404-414, 2019.

[57] D.A. Skoog, F.J. Holler, S.R.Crouch, y D.M. West, “Espectrometría de masas ,” in *Fundamentos de Química Analítica*, 9th ed., CENGAGE Learning, 2014.

[58] J. Beynon, J. Herbert, y L. Brown, “Espectrometría de masas,” *Enciclopedia Británica*. 2023.

[59] M. L. Vestal, “Methods of Ion Generation,” *Chem Rev*, vol. 101, no. 2, pp. 361–376, Feb. 2001, doi: 10.1021/cr990104w.

[60] M. L. Gross y R. M. Caprioli, “The encyclopedia of mass spectrometry,” vol. 1. Elsevier, 2003.

[61] V. Gomis Yagües, “Tema 5. Espectrometría de masas,” *Técnicas Instrumentales en el Análisis Industrial*, 2008.

- [62] F. Hillenkamp y J. Peter, "MALDI MS A Practical guide to Instrumentation Methods and Applications," vol. 31 48149, Institute for Medical Physics and Biophysics University of Münster Robert-Koch-Str., Ed., Münster Germany: WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2007.
- [63] V.-A. Duong y H. Lee, "Bottom-Up Proteomics: Advancements in Sample Preparation," *Int J Mol Sci*, vol. 24, no. 6, p. 5350, Mar. 2023, doi: 10.3390/ijms24065350.
- [64] J. A. Paulo, "Sample preparation for proteomic analysis using a GeLC-MS/MS strategy," *J Biol Methods*, vol. 3, no. 3, p. e45, Jul. 2016, doi: 10.14440/jbm.2016.106.
- [65] J. R. Wiśniewski, "Filter-Aided Sample Preparation," 2017, pp. 15–27. doi: 10.1016/bs.mie.2016.09.013.
- [66] L. E. Prestagiacomo, C. Gabriele, P. Morelli, et al., "Proteomic Profile of EPS-Urine through FASP Digestion and Data-Independent Analysis," *Journal of Visualized Experiments*, no. 171, May 2021, doi: 10.3791/62512.
- [67] C. Delles, E. Carrick, D. Graham, y S. A. Nicklin, "Utilizing proteomics to understand and define hypertension: where are we and where do we go?," *Expert Rev Proteomics*, vol. 15, no. 7, pp. 581–592, Jul. 2018, doi: 10.1080/14789450.2018.1493927.
- [68] D. Pellerin, H. Gagnon, J. Dubé, y F. Corbin, "Amicon-adapted enhanced FASP: an in-solution digestion- based alternative sample preparation method to FASP," *F1000Res*, vol. 4, p. 140, Sep. 2015, doi: 10.12688/f1000research.6529.2.
- [69] Y. Yu, S. Bekele, and R. Pieper, "Quick 96FASP for high throughput quantitative proteome analysis," *J Proteomics*, vol. 166, pp. 1–7, Aug. 2017, doi: 10.1016/j.jprot.2017.06.019.
- [70] Technical Bulletin, "Trypsin Gold, Mass Spectrometry Grade Instructions for Use of Product," vol. V5280, recuperado de [<https://worldwide.promega.com/resources/protocols/technical-bulletins/101/trypsin-gold-mass-spectrometry-grade-protocol/>].
- [71] NIH Instituto Nacional de Investigación del Genoma Humano, "bioinformática," recuperado de [<https://www.genome.gov/es/genetics-glossary/Bioinformatica#:~:text=La%20bioinform%C3%A1tica%2C%20en%20relaci%C3%B3n%20con%20la%20gen%C3%A9tica%20y,ADN%20y%20amino%C3%A1cidos%20o%20anotaciones%20sobre%20esas%20secuencias>]
- [72] A. D. Baxevanis y A. Bateman, "The importance of biological databases in biological discovery," *Curr Protoc Bioinformatics*, vol. 50, no. 1, p. 1, 2015.
- [73] J. Garcia, J. Molina, y A. Berlanga, "introduccion - etapas en los procesos de big data," en *Ciencia de Datos- Tecnicas Analiticas y aprendizaje estadistico en un enfoque practico*, Bogotá: Alfaomega, Publicaciones Altaria, 2018.
- [74] J. Díaz-Ramírez, "Aprendizaje Automático y Aprendizaje Profundo," *Ingeniare. Revista chilena de ingeniería*, vol. 29, no. 2, pp. 180–181, 2021.
- [75] N. P. Méndez y J. P. Rubier, "Ciencia de datos: una revisión del estado del arte," *UCE Ciencia. Revista de postgrado*, vol. 6, no. 3, 2018.
- [76] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, y A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised and unsupervised learning for data science*, pp.

3–21, 2020, doi: 10.1007/978-3-030-22475-2.

- [77] J. Hernández Orallo, "Introducción a la Minería de Datos," 2004.
- [78] E. M. Rojas, "Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo," *Revista Ibérica de Sistemas e Tecnologías de Informação*, no. E28, pp. 586–599, 2020.
- [79] Gudta Sakshi, "¿Cuál es el mejor lenguaje para el aprendizaje automático?," recuperado de [<https://www.springboard.com/blog/data-science/best-language-for-machine-learning/>].
- [80] Delua Juliana, "Aprendizaje supervisado versus no supervisado: ¿cuál es la diferencia?," [<https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>].
- [81] S. S. Aljameel et al., "Prediction of Preeclampsia Using Machine Learning and Deep Learning Models: A Review," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 32, Feb. 2023, doi: 10.3390/bdcc7010032.
- [82] A. Thomas, G. D. Tourassi, A. S. Elmaghraby, R. Valdes, and S. A. Jortani, "Data mining in proteomic mass spectrometry," *Clin Proteomics*, vol. 2, no. 1–2, pp. 13–32, Mar. 2006, doi: 10.1385/CP:2:1:13.
- [83] K. R. Coombes, J. M. Koomen, K. A. Baggerly, J. S. Morris, and R. Kobayashi, "Understanding the characteristics of mass spectrometry data through the use of simulation," *Cancer Inform*, vol. 1, p. 117693510500100100, 2005, doi: <https://doi.org/10.1177/1176935105001001>.
- [84] C. M. Bishop, "Pattern recognition and machine learning," Springer google schola, 2, 645-678, 2006.
- [85] Builtin. "PCA in Python," recuperado de [<https://builtin.com/machine-learning/pca-in-python>], (s.f.).
- [86] J. M. Pérez, y P. P. Martín, "Regresión logística". *Medicina de Familia. SEMERGEN*, 50(1), 102086, 2024.
- [87] E. C. León, "Introducción a las máquinas de vector soporte (SVM) en aprendizaje supervisado," Trabajo de Fin de Grado en Matemáticas, Universidad de Zaragoza, 2017.
- [88] A. Pal, "Gradient boosting trees for classification: A beginner's guide". Medium. Recuperado de [<https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea>], (2020).
- [89] T. R. Transue, J. M. Krahn, S. A. Gabel, E. F. DeRose, y R. E. London, "X-ray and NMR Characterization of Covalent Complexes of Trypsin, Borate, and Alcohols," *Biochemistry*, vol. 43, no. 10, pp. 2829–2839, Mar. 2004, doi: 10.1021/bi035782y.
- [90] A. Bujacz, "Structures of bovine, equine and leporine serum albumin," *Acta Crystallogr D Biol Crystallogr*, vol. 68, no. 10, pp. 1278–1289, Oct. 2012, doi: 10.1107/S0907444912027047.
- [91] Pierce, "Protein concentrators PES, 3K or 5 K MWCO, 0,5-100 mL" Catalogo número 88512, recuperado de [<https://www.thermofisher.com/order/catalog/product/88512?SID=srch-srp-88512>]
- [92] C. Barajas-Solano, B. Muñoz, E. Chicano-Gálvez, P. Escobar, y E. Mejía-Ospino, "Discriminator for Cutaneous Leishmaniasis Using MALDI-MSI in a Murine Model," *Journal of the American Society for Mass*

- Spectrometry, 33(6), 952-960, 2022.
- [93] Pierce, "BCA protein Assay Kit," catálogos números 23225 y23227, Recuperado de [https://www.thermofisher.com/order/catalog/product/15045?gclid=Cj0KCQjw\_qexBhCoARIsAFgBleupLGptaX5J\_2FJqLSKwuosbSwd\_-KoPv3s4JBIP97kBGwa4Qc2sv8aAog4EALw\_wcB&ef\_id=Cj0KCQjw\_qexBhCoARIsAFgBleupLGptaX5J\_2FJqLSKwuosbSwd\_-KoPv3s4JBIP97kBGwa4Qc2sv8aAog4EALw\_wcB:G:s&s\_kwid=AL13652!3!562127230705!!!g!!!15287362487!128563401494&cid=bid\_clb\_ccp\_r01\_co\_cp0000\_pjt0000\_bid00000\_0se\_gaw\_dy\_pur\_con&gad\_source=1]
- [94] E. P. Dimagno, D. Corle, J. F. O'brien, I. J. Masnyk, V. L. W. Go, y R. Aamodt, "Effect of Long-Term Freezer Storage, Thawing, and Refreezing on Selected Constituents of Serum," *Mayo Clin Proc*, vol. 64, no. 10, pp. 1226–1234, Oct. 1989, doi: 10.1016/S0025-6196(12)61285-3.
- [95] S. Cuhadar, M. Koseoglu, A. Atay, y A. Dirican, "The effect of storage time and freeze-thaw cycles on the stability of serum samples," *Biochem Med (Zagreb)*, pp. 70–77, 2013, doi: 10.11613/BM.2013.009.
- [96] R. E. Gislefoss, M. Lauritzen, H. Langseth, y L. Mørkrid, "Effect of multiple freeze-thaw cycles on selected biochemical serum components," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 55, no. 7, Jan. 2017, doi: 10.1515/cclm-2016-0892.
- [97] M. Montelongo, "Act. 4.3. Flujo sanguíneo y gasto cardíaco," Sep. 2021. Accessed: Jan. 04, 2024. Recuperado de [https://www.studocu.com/es-mx/document/universidad-de-guadalajara/fisiologia-y-fundamentos-de-fisiopatologia/act-43-flujo-sanguineo-y-gasto-cardiaco/20248254].
- [98] A. Schmudlach, J. Felton, C. Cipolla, L. Sun, R. T. Kennedy, y N. J. Dovichi, "Sample preparation protocol for bottom-up proteomic analysis of the secretome of the islets of Langerhans," *Analyst*, vol. 141, no. 5, pp. 1700–1706, 2016, doi: 10.1039/C5AN02265G.
- [99] D. R. Canchi, D. Paschek, y A. E. García, "Equilibrium Study of Protein Denaturation by Urea," *J Am Chem Soc*, vol. 132, no. 7, pp. 2338–2344, Feb. 2010, doi: 10.1021/ja909348c.
- [100] A. Arsiccio, P. Ganguly, y J.-E. Shea, "A Transfer Free Energy Based Implicit Solvent Model for Protein Simulations in Solvent Mixtures: Urea-Induced Denaturation as a Case Study," *J Phys Chem B*, vol. 126, no. 24, pp. 4472–4482, Jun. 2022, doi: 10.1021/acs.jpcc.2c00889.
- [101] J. R. Wiśniewski y M. Mann, "Consecutive Proteolytic Digestion in an Enzyme Reactor Increases Depth of Proteomic and Phosphoproteomic Analysis," *Anal Chem*, vol. 84, no. 6, pp. 2631–2637, Mar. 2012, doi: 10.1021/ac300006b.
- [102] J. R. Wiśniewski, "Quantitative Evaluation of Filter Aided Sample Preparation (FASP) and Multienzyme Digestion FASP Protocols," *Anal Chem*, vol. 88, no. 10, pp. 5438–5443, May 2016, doi: 10.1021/acs.analchem.6b00859.
- [103] M. Trivedi, J. Laurence, y T. Siahaan, "The Role of Thiols and Disulfides on Protein Stability," *Curr Protein Pept Sci*, vol. 10, no. 6, pp. 614–625, Dec. 2009, doi: 10.2174/138920309789630534.
- [104] J. Garcia, A. Herrera, y M. Castillo, "Ditiotreitol (DTT)." Instituto LINCON. Accessed: Jan. 04, 2024. [Online].

Available: <https://licon.com.mx/ditiotreititol-dtt/>

- [105] M. R. Á. Santos, "evaluación del efecto de la cicloheximida sobre la memoria emocional y la expresión de proteínas en el hipocampo dorsal de ratas wistar expuestas al laberinto en cruz elevado".
- [106] J. R. Wiśniewski, "Filter-Aided Sample Preparation," 2017, pp. 15–27. doi: 10.1016/bs.mie.2016.09.013.
- [107] H. K. Hustoft, H. Malerod, S. R. Wilson, L. Reubsaet, E. Lundanes, y T. Greibrokk, "A critical review of trypsin digestion for LC-MS based proteomics," *Integrative Proteomics*, vol. 73, 2012.
- [108] B. R. Fonslow et al., "Digestion and depletion of abundant proteins improves proteomic coverage," *Nat Methods*, vol. 10, no. 1, pp. 54–56, Jan. 2013, doi: 10.1038/nmeth.2250.
- [109] J. R. Wiśniewski y G. Prus, "Homogenous Phase Enrichment of Cysteine-Containing Peptides for Improved Proteome Coverage," *Anal Chem*, vol. 87, no. 13, pp. 6861–6867, Jul. 2015, doi: 10.1021/acs.analchem.5b01215.
- [110] A. A. Patil, C.-K. Chiang, C.-H. Wen, y W.-P. Peng, "Forced dried droplet method for MALDI sample preparation," *Anal Chim Acta*, vol. 1031, pp. 128–133, Nov. 2018, doi: 10.1016/j.aca.2018.05.056.
- [111] R. Bhardwaj, X. Fang, P. Somasundaran, y D. Attinger, "Self-Assembly of Colloidal Particles from Evaporating Droplets: Role of DLVO Interactions and Proposition of a Phase Diagram," *Langmuir*, vol. 26, no. 11, pp. 7833–7842, Jun. 2010, doi: 10.1021/la9047227.
- [112] S. K. Wilson y H.-M. D'Ambrosio, "Evaporation of Sessile Droplets," *Annu Rev Fluid Mech*, vol. 55, no. 1, pp. 481–509, Jan. 2023, doi: 10.1146/annurev-fluid-031822-013213.
- [113] M. J. Roth et al., "Thin-Layer Matrix Sublimation with Vapor-Sorption Induced Co-Crystallization for Sensitive and Reproducible SAMDI-TOF MS Analysis of Protein Biosensors," *J Am Soc Mass Spectrom*, vol. 23, no. 10, pp. 1661–1669, Oct. 2012, doi: 10.1007/s13361-012-0442-7.
- [114] R. D. Deegan, "Pattern formation in drying drops," *Phys Rev E*, vol. 61, no. 1, pp. 475–485, Jan. 2000, doi: 10.1103/PhysRevE.61.475.
- [115] T. W. Jaskolla, M. Karas, U. Roth, K. Steinert, C. Menzel, and K. Reihls, "Comparison between vacuum sublimed matrices and conventional dried droplet preparation in MALDI-TOF mass spectrometry," *J Am Soc Mass Spectrom*, vol. 20, no. 6, pp. 1104–1114, Jun. 2009, doi: 10.1016/j.jasms.2009.02.010.
- [116] M. Karas, y R. Krüger, "Ion formation in MALDI: the cluster ionization mechanism," *Chemical reviews*, 103(2), 427-440, 2003.
- [117] B. A. Otieno, C. E. Krause, y J. F. Rusling, "Bioconjugation of Antibodies and Enzyme Labels onto Magnetic Beads," 2016, pp. 135–150. doi: 10.1016/bs.mie.2015.10.005.
- [118] Fisher Scientific S.L., "Thermo Scientific™ Pierce™ BCA Protein Assay Kits," <https://www.fishersci.es/shop/products/pierce-bca-protein-assay-kits/p-200000333>.
- [119] R. J. Simpson, "Quantifying Protein by Bicinchoninic Acid," *Cold Spring Harb Protoc*, vol. 2008, no. 8, p. pdb.prot4722, Aug. 2008, doi: 10.1101/pdb.prot4722.
- [120] S. Kreuzsch, S. Schwedler, B. Tautkus, G. A. Cumme, y A. Horn, "UV measurements in microplates suitable for

high-throughput protein determination," *Anal Biochem*, vol. 313, no. 2, pp. 208–215, Feb. 2003, doi: 10.1016/S0003-2697(02)00460-8.

[121] N. A. Tashtush, E. M. Altamimi, y A. J. Aleshawi, "Ménétrier's disease in a 5-year-old girl without cytomegalovirus infection: Case report," *Paedia Open A Open J*, pp. 1–5, 2019.

[122] J. D. W. Choi et al., "Is preoperative hypoalbuminaemia or hypoproteinaemia a reliable marker for anastomotic leakage risk in patients undergoing elective colorectal surgery in an Enhanced Recovery after Surgery (ERAS) program?," 2023.

[123] G. Kok, C. D. M. van Karnebeek, yS. A. Fuchs, "Response to Shen and Zou," *Genetics in Medicine*, vol. 23, no. 3, pp. 589–590, Mar. 2021, doi: 10.1038/s41436-020-01014-8.

[124] A. Elbehiry et al., "Using Protein Fingerprinting for Identifying and Discriminating Methicillin Resistant *Staphylococcus aureus* Isolates from Inpatient and Outpatient Clinics," *Diagnostics*, vol. 13, no. 17, p. 2825, Aug. 2023, doi: 10.3390/diagnostics13172825.

[125] S. Correnti et al., "Revealing the Hidden Diagnostic Clues of Male Infertility from Human Seminal Plasma by Dispersive Solid Phase Extraction and MALDI-TOF MS," *Int J Mol Sci*, vol. 23, no. 18, p. 10786, Sep. 2022, doi: 10.3390/ijms231810786.

[126] W. Pusch, M. T. Flocco, S.-M. Leung, H. Thiele, and M. Kostrzewa, "Mass spectrometry-based clinical proteomics," *Pharmacogenomics*, vol. 4, no. 4, pp. 463–476, Jul. 2003, doi: 10.1517/phgs.4.4.463.22753.

[127] M. Liu et al., "Development of a prediction model on preeclampsia using machine learning-based method: a retrospective cohort study in China," *Front Physiol*, vol. 13, Aug. 2022, doi: 10.3389/fphys.2022.896969.

[128] S. Sarkar et al., "Effect of Tryptic Digestion on Sensitivity and Specificity in MALDI-TOF-Based Molecular Diagnostics through Machine Learning," *Sensors*, vol. 23, no. 19, p. 8042, Sep. 2023, doi: 10.3390/s23198042.