# Análisis del perfil proteómico de sueros sanguíneos de pacientes con preeclampsia empleando espectrometría de masas MALDI-TOF

Vanessa Zambrano Martínez

Trabajo de Grado para Optar al Título de Químico

Director

Enrique Mejía Ospino

Doctor en Ciencias Químicas

Codirectora

Yuly Andrea Prada Vargas

Doctora en Química

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Química

Química

Bucaramanga

2026

**Dedicatoria**

A Dios, por siempre ser tan bueno conmigo.

A mi mami Rosalba Diaz, quien me crio y, con sus valores, me formó en la mujer que soy.

A mi papá Alberto Martínez, que desde el cielo me acompaña siempre.

A Ronal de Freitas, por ser mi apoyo incondicional.

A mis amigos Harold, Bryan y Dani por siempre creer en mí.

**Agradecimientos**

A Dios, por acompañarme durante todo mi proceso académico, por concederme sabiduría, fortaleza y perseverancia, y por sostenerme en cada etapa de este camino.

A la Universidad Industrial de Santander, a la Escuela de Química y a la Facultad de Ciencias, por brindarme la formación académica y los espacios necesarios para el desarrollo de este trabajo de investigación.

Al doctor Enrique Mejía, director de la tesis, y a la doctora Yuly Prada, codirectora, por su acompañamiento, orientación académica, paciencia y apoyo constante durante el desarrollo del proyecto.

Al Grupo de Investigación LEAM, por facilitar el acceso a sus laboratorios, en especial al laboratorio de espectrometría de masas, y al grupo GIBIM por el apoyo brindado y por permitir el uso de sus instalaciones. De manera especial, agradezco a la magister Jennifer Ruiz por su valioso acompañamiento durante el proceso experimental.

Agradezco al Ministerio de Ciencias por la financiación otorgada y a la Fundación Cardiovascular de Colombia (FCV) por permitirme hacer parte del programa de Jóvenes Investigadores, así como a la doctora Paula Bautista por su acompañamiento durante este proceso.

A mis amigos Harold, Bryan y Daniela, por su apoyo constante y por creer en mí, y a Ronal, por su motivación y apoyo incondicional a lo largo de este camino.

A mi familia, especialmente a mi abuela Rosalba, quien ha sido un pilar fundamental en mi vida, y a mi mamá Faida y tíos, por su apoyo y acompañamiento.

Finalmente, me agradezco a mí misma por la perseverancia, la fortaleza y la determinación para culminar este proceso y alcanzar este logro.

## Tabla de Contenido

**Pág.**

## Lista de Tablas

**Pág.**

## Lista de Figuras

**Pág.**

## Lista de Apéndices

**Pág.**

## Glosario

**ACN:** acetonitrilo.

**ANN:** redes neuronales artificiales

**BR:** buffer Britton-Robinson.

**CD5L:** proteína CD5-like.

**FASP:** preparación de muestras asistida por filtros.

**HIF-1α:** factor inducible por hipoxia 1 alfa.

**MALDI:** desorción/ionización láser asistida por matriz.

**MALDI-TOF:** desorción/ionización láser asistida por matriz con analizador de tiempo de vuelo.

**MS:** espectrometría de masas.

**PCA:** análisis de componentes principales.

**PlGF:** factor de crecimiento placentario.

**ROS/RNS:** especies reactivas de oxígeno y nitrógeno.

**sEng:** endoglina soluble.

**sFLT1:** tirosina quinasa tipo Fms soluble.

**SVM:** máquina de vectores de soporte.

**TOF:** tiempo de vuelo.

## Resumen

**Título:** Análisis del perfil proteómico de sueros sanguíneos de pacientes con preeclampsia empleando espectrometría de masas MALDI-TOF[1*]

**Autor:** Vanessa Zambrano Martínez[2*3*]

**Palabras Clave:** Preeclampsia; proteómica, espectrometría de masas, MALDI-TOF, aprendizaje automático.

**Descripción:** La preeclampsia es un trastorno hipertensivo del embarazo que afecta aproximadamente entre el 5–8% de las gestaciones y constituye una de las principales causas de morbilidad y mortalidad materna y perinatal. Su complejidad radica en la falta de claridad sobre su etiología y en la ausencia de biomarcadores confiables para su detección temprana. En este contexto, la proteómica surge como una estrategia prometedora para identificar perfiles proteicos diferenciales asociados a la enfermedad.

En esta investigación, se examinaron muestras de suero de mujeres embarazadas con preeclampsia y de un grupo de controles sanos, proporcionadas por la Fundación Cardiovascular de Colombia. Para el análisis proteómico, se empleó el método FASP, el cual permitió la digestión de las proteínas y la obtención de péptidos de bajo peso molecular para su análisis mediante espectrometría de masas MALDI-TOF.

Los datos generados fueron analizados mediante Python en JupyterLab. El análisis de componentes principales mostró una separación parcial entre los diferentes grupos, sugiriendo variaciones en los perfiles proteómicos. A continuación, se desarrollaron modelos de clasificación supervisada. El modelo SVM logró una exactitud del 79%, mientras que la red neuronal mostró un rendimiento superior, alcanzando una exactitud del 90,70% y una mayor efectividad en la identificación de casos de preeclampsia. Estos hallazgos resaltan el potencial del análisis proteómico, junto con técnicas de aprendizaje automático, en la detección de patrones relacionados con la preeclampsia y su posible uso en el desarrollo de herramientas de diagnóstico.

---

[1*] Trabajo de Grado

[2**] Facultad de Ciencias. Escuela de Química. Director: Enrique Mejia Ospino. Ph.D. en Ciencias Químicas. Codirector: Yuly Andrea Prada. Ph.D. en Ciencias Químicas.

[3]

## Abstract

**Title:** Proteomic profile analysis of blood serum from patients with preeclampsia using MALDI-TOF mass spectrometry [4*]

**Author(s):** Vanessa Zambrano Martínez[5]

**Key Words:** Preeclampsia, proteomics, mass spectrometry, MALDI-TOF, machine learning.

**Description:** Preeclampsia is a hypertensive disorder of pregnancy that affects approximately 5–8% of pregnancies and is a leading cause of maternal and perinatal morbidity and mortality. Its complexity lies in the unclear etiology and the lack of reliable biomarkers for early detection. In this context, proteomics emerges as a promising strategy for identifying differential protein profiles associated with the disease.

In this study, serum samples from pregnant women with preeclampsia and healthy controls, provided by the Cardiovascular Foundation of Colombia, were analyzed. For the proteomic analysis, the FASP method was employed, enabling protein digestion and the generation of low-molecular-weight peptides for analysis by MALDI-TOF mass spectrometry.

The generated data were analyzed using Python in JupyterLab. Principal component analysis showed partial separation between the groups, suggesting variations in proteomic profiles. Subsequently, supervised classification models were developed. The SVM model achieved an accuracy of 79%, while the neural network showed superior performance, reaching 90.70% accuracy and greater effectiveness in identifying cases of preeclampsia.

These findings highlight the potential of proteomic analysis, combined with machine learning techniques, for detecting patterns associated with preeclampsia and its potential application in the development of diagnostic tools.

---

[4*] Degree Work

[5]Faculty of Science. School of Chemistry. Advisor: Enrique Mejia Ospino. Ph. D, Chemistry Science. Co-advisor: Yuly Andrea Prada. Ph. D, Chemistry Science.

## Introducción

Actualmente, la preeclampsia se cataloga como una de las enfermedades más extrañas existentes que se desarrollan en el embarazo humano. Esta patología afecta aproximadamente al 5-8% de las gestaciones y produce morbilidad, mortalidad materna y perinatal. Su complejidad radica en el desconocimiento de las causas que la producen y la existencia de múltiples teorías acerca de su etiología, lo que dificulta establecer diferencias exactas entre un embarazo normal y uno con preeclampsia.

Si bien, en los últimos 10 años se han dado a conocer principios sobre la fisiopatología de la preeclampsia, aún no existen herramientas de detección o biomarcadores precisos, que permitan un diagnóstico de la preeclampsia en las primeras etapas de la gestación (Nirupama et al., 2021).

Los estudios clínicos realizados para su detección temprana son costosos y limitados (Anand et al., 2016). Por consiguiente, lograr un diagnóstico clínico de preeclampsia, representa un progreso en la búsqueda de soluciones para el tratamiento de la enfermedad, de tal manera que las pacientes pueden acceder a un tratamiento oportuno y adecuado con las mínimas complicaciones para la madre o el feto (Pasyar et al., 2020).

A raíz de esa problemática, se han propuesto diferentes métodos de análisis para identificar nuevos biomarcadores que permitan predecir de manera temprana la preeclampsia y, de este modo, sugerir un posible tratamiento. Entre los métodos de análisis, se destaca el análisis proteómico, que consiste en la evaluación de los perfiles proteicos completos de muestras sanguíneas, fluidos fisiológicos y tejidos como la placenta y el cordón umbilical.

La adquisición y el análisis del perfil de proteínas y su asociación con la manifestación de diferentes afecciones clínicas se han convertido en una herramienta de *screening* y, en algunos

casos, en una herramienta complementaria para la identificación y confirmación del diagnóstico de enfermedades complejas.

Los análisis proteómicos requieren una preparación secuencial, exhaustiva y precisa de la muestra, en la que intervienen pasos de aislamiento de proteínas, separación electroforética y posterior identificación de su estructura primaria por técnicas analíticas robustas y de alta resolución, como la espectrometría de masas. Notablemente, la combinación de la proteómica diferencial basada en la espectrometría de masas y el uso de herramientas de machine learning (ML) en el tratamiento y análisis de datos bioquímicos se ha posicionado como una alternativa versátil para la elaboración de modelos de predicción de enfermedades multifactoriales, como la preeclampsia, en la que el origen de la enfermedad aún se desconoce y no ha sido atribuida a un único factor; estas estrategias analíticas proveen un soporte completo en la comprensión de los mecanismos de interacción biológicos, las implicaciones y rol de las proteínas en la progresión de las enfermedades, y en la búsqueda de nuevos marcadores moleculares para el desarrollo de otros métodos de diagnóstico que puedan ser implementados en el ámbito clínico.

En consecuencia, el presente documento de investigación muestra los hallazgos y resultados de un proyecto propuesto por la Fundación Cardiovascular de Colombia (FCV) junto con el Laboratorio de Espectroscopía Atómica y Molecular (LEAM), cuyo objetivo fue analizar el perfil de proteínas a partir de sueros sanguíneos de muestras de mujeres en embarazo cuyo desenlace fue normal y mujeres con preeclampsia, como un análisis complementario de un estudio multi-ómico de casos y controles llamado proyecto GenPE (Genética en Preeclampsia) el cual tuvo como objetivo el estudio, comprensión y prevención temprana de la preeclampsia en Colombia y realizado entre 2001 y 2012; en este proyecto se recolecto y preservó de material biológico (suero

sanguíneo, cordón umbilical y placenta) de aproximadamente 4500 mujeres gestantes en diferente regiones el país.

Para este trabajo, se analizaron muestras de suero sanguíneo de gestantes con preeclampsia (100) y de mujeres con embarazo normal (112) mediante espectrometría de masas MALDI-TOF, con el propósito de identificar un perfil proteómico diferencial. También se emplearon herramientas de análisis estadístico, como el análisis de componentes principales (PCA), para evaluar la heterogeneidad y la separación de las muestras en un análisis no supervisado, y posteriormente se aplicaron modelos de aprendizaje supervisado, incluyendo máquinas de soporte vectorial (SVM) y redes neuronales artificiales (ANN), con el fin de evaluar la utilidad de la información biológica de los perfiles de proteínas para desarrollar modelos que permitieran predecir y diferenciar casos de preeclampsia. Los resultados mostraron que el modelo SVM alcanzó una exactitud del 79%, mientras que el modelo ANN presentó un mejor desempeño con una exactitud del 90,70%. En consecuencia, los perfiles proteicos aportaron patrones relevantes para distinguir y clasificar casos y controles de preeclampsia, lo que constituye un conocimiento valioso y evidencia de que la búsqueda de nuevos biomarcadores aplicables al diagnóstico temprano de la preeclampsia aún no se ha concluido.

# 1. Objetivos

## 1.1 Objetivo General

Determinar el perfil proteómico de sueros sanguíneos de pacientes mujeres con preeclampsia mediante la técnica de espectrometría de masas con ionización/desorción asistida por matriz y analizador de tiempo de vuelo (MALDI-TOF).

## 1.2 Objetivos Específicos

- Analizar el perfil proteico en muestras de suero sanguíneo de pacientes con preeclampsia por espectrometría de masas MALDI-TOF
- Ajustar las condiciones instrumentales en el equipo de MALDI-TOF, como la afluencia del láser, el rango de masas detectado, los métodos de análisis (iones negativos o iones positivos), y seleccionar una matriz adecuada para los procesos de ionización/desorción de la muestra, así como la relación molar entre la matriz utilizada y el suero sanguíneo.
- Analizar por MALDI-TOF las fracciones proteicas del suero sanguíneo obtenidas por técnicas de exclusión de tamaño, como la extracción por membrana.
- Hallar diferencias significativas en los espectros de masas del suero sanguíneo “in silico” respecto de los de sus fracciones, a fin de establecer un perfil proteómico diferencial.

## 2. Estado del arte

En 1994, el científico Marc Wilkins definió el concepto de proteoma (Wasinger et al., 1995). Wilkins, usó el término para poder describir el complemento completo de las proteínas que se expresan en el genoma, las células y tejidos. Para la década de 1990, el análisis proteómico se aceleró drásticamente, en parte debido a la disponibilidad de técnicas de identificación de proteínas, lo que permitió convertir una tarea difícil en un procedimiento simple y escalable. En el año 2001, el bioquímico Matthias Mann redefine el concepto de proteómica como el análisis a gran escala de las funciones de los productos génicos, diferenciándola de disciplinas como la genómica y la bioinformática, que, a su vez, se complementan en el análisis proteómico. Mann resaltó que tanto los avances en espectrometría de masas (MS), como la técnica MALDI-MS (desorción/ionización láser asistida por matriz), la MS/MS con nanoelectrospray, y el TOF cuadripolar, permitieron la identificación de 250 proteínas con alta sensibilidad y rendimiento (Jain, 2001).

Con la comprensión de la importancia de la identificación de proteínas y marcadores clínicos presentes en la sangre, y de su relación con la abundancia de proteínas, se empezaron a considerar los análisis proteómicos una herramienta precisa para la detección de enfermedades mediante la comparación de proteomas (Noroña Calvachi, 2014).

El análisis proteómico se ha demostrado como un avance clave en la determinación de biomarcadores para diversas enfermedades, incluida la preeclampsia (Tomkiewicz & Darmochwał-Kolarz, 2024), al estudiar cambios en la expresión proteica que ocurren durante el embarazo y proporcionar información que permite la detección temprana de la enfermedad (Rybak-Krzyszkowska et al., 2023). La proteinuria y las apolipoproteínas (ApoB, ApoC-III y ApoE), por ejemplo, se expresan significativamente en la preeclampsia, debido a un estado

proinflamatorio, riesgo cardiovascular y disfunción endotelial como manifestaciones en el embarazo (Serrano et al., 2018), pero también está asociada a otras enfermedades como la diabetes mellitus y afecciones hipertensivas no relacionadas con la preeclampsia (Chaemsaithong et al., 2022).

En estudios clínicos, bajo las directrices internacionales como el Instituto Nacional Británico para la Excelencia en Salud, se ha aprobado e incluido la combinación de biomarcadores como PlGF, sFLT1 y sEng, junto con modelos predictivos, la cual ha alcanzado tasas de detección superiores al 75% y falsos positivos por debajo del 10%, lo cual sustenta su uso en la práctica clínica aun en entornos con recursos limitados (Rana et al., 2019).

Por otro lado, investigaciones realizadas en Latinoamérica han proporcionado información muy útil en la búsqueda de nuevos biomarcadores, como lo describen Valderrama-Aguirre et al. (2011), quienes identificaron SERPINA-1 como un biomarcador urinario capaz de detectar preeclampsia hasta 10 semanas antes de que se presenten los síntomas clínicos, lo que crea una oportunidad para una intervención temprana. Además, también se han realizado estudios en países como Bolivia y Cuba donde se han documentado alteraciones en la glicosilación de apolipoproteína E, sin cambios en su concentración, lo que sugiere modificaciones funcionales importantes (Noroña Calvachi, 2014). En la Tabla 1, se ilustran algunos resultados de investigaciones de los últimos 10 años en las que se realizaron análisis proteómicos en muestras de suero sanguíneo con preeclampsia. En resumen, los artículos de investigación con mayor relevancia que se encontraron en las bases de datos Scopus y ScienceDirect, y las tecnologías que implementadas se muestran en la Tabla 1.

**Tabla 1. Análisis proteómico en preeclampsia**

| Autor/año | Tipo de muestra | Tecnología | Resultado de la investigación |
|---|---|---|---|
| (Kolialexi et al., 2017) | Plasma | MALDI-TOF | Sobreexpresión de 12 proteínas, entre ellas Alfa-1-antitripsina (A1AT), molécula similar al antígeno CD5 (CD5L) queratina, citoesqueleto tipo I 9 (K1C9, en paciente con PE. |
| (Bahado-Singh et al., 2017) | Suero | MALDI-TOF | Identificación de Proteína de unión TATA (TBP) como posibles biomarcadores predictivos de pacientes con PE. |
| (Hua et al., 2018) | Placenta | MALDI-TOF | Hallazgo del polimorfismo de SNP (rs9393931) con el genotipo TT recesivo ubicado en el gen 3-UTR del gen ERVFRDE-1 asociado a pacientes con PE. |
| (Szabo et al., 2020) | Placenta | LC-MS | Identificación PP1, elevada en pacientes con PE. |

| | | | |
|---|---|---|---|
| (Sergeeva et al., 2020) | Orina | HPLC-MS/MS | Identificación de biomarcadores clave: α-1-antitripsina (SERPINA1), complemento C3, haptoglobina, ceruloplasmina y tripstatina. |
| (Chen et al., 2022) | Plasma | LC-MS/MS | Identificación de proteínas diferencialmente expresadas en preeclampsia temprana y tardía; biomarcadores clave: IGFBP4, ITIH2, ITIH3, ITIH4. |
| (Andresen et al., 2025) | Plasma | SomaScan | Cuantificación de ~7,000 proteínas; biomarcadores para preeclampsia de inicio tardío: FAAH2, IL17RC, SIGLEC6, HTRA1. |

En conjunto, estas investigaciones evidencian avances significativos en la caracterización proteómica de la preeclampsia; no obstante, aún se requieren más estudios que permitan trasladar estos hallazgos al ámbito clínico, así como fortalecer su análisis mediante herramientas de ciencia de datos (Rodríguez, 2025).

# 3. Marco teórico

## 3.1 Preeclampsia

En el embarazo la triada mortal, representa una de las principales causas de complicaciones de salud. La triada mortal está conformada por las hemorragias, infecciones y trastornos hipertensivos, dentro de esta última se encuentra la preeclampsia (Cunningham et al., 2022).

La preeclampsia es un trastorno multisistémico grave que puede afectar todos los órganos del cuerpo humano (American College of Obstetricians & Gynecologists, 2021). Se caracteriza por la presencia de hipertensión arterial (presión arterial sistólica >140 y/o diastólica >90mmHg) (Gallos et al., 2013) y proteinuria (300mg/24hr o 1+ por tirametría en muestra aislada de orina), después de la semana 20 de gestación (Sánchez, 2018). Es un problema obstétrico considerable y una fuente importante de morbilidad y mortalidad materna y neonatal (Gilbert et al., 2008; National High Blood Pressure Education Program, 2000).

La preeclampsia se debe a una anomalía de la implantación de la placenta, de mecanismo no bien conocido, que ocasiona isquemia placentaria; ésta, a su vez, provoca hipoxia fetal crónica y lesión endotelial difusa materna, lo que origina la hipertensión arterial y las complicaciones orgánicas. La gravedad de la preeclampsia depende de la intensidad de la hipertensión arterial o de la severidad de las lesiones orgánicas como la formación de edema pulmonar, insuficiencia renal, síndrome HELLP (hemolysis-elevated liver enzymes-low platelets count; hemólisis, que es la elevación de las enzimas hepáticas y disminución del recuento de plaquetas), alteraciones de la coagulación, hematoma retroplacentario o hematoma hepático, eclampsia y retraso de crecimiento o muerte fetal intrauterina (Sibai et al., 2005).

La preeclampsia puede clasificarse según se observa en la Tabla 2 (Cunningham et al., 2022).

**Tabla 2. Clasificación general preeclampsia**

| | **Clasificación Preeclampsia** | | | |
|---|---|---|---|---|
| | Inicio temprano | Inicio tardío | Inicio prematuro | Inicio del término |
| **Semanas** | <34 semanas | ≥34 semanas | <37 semanas | ≥37 semanas |

*Nota:* Adaptado de Cunningham et al. (2022).

Los factores de riesgos característicos en la aparición de preeclampsia son los siguientes: antecedentes de preeclampsia en embarazos anteriores, embarazos múltiples, hipertensión crónica, enfermedades renales, diabetes mellitus y trastornos autoinmunes como el lupus (Duckitt & Harrington, 2005). También se encuentran factores de riesgo moderados, como el primer embarazo, un intervalo superior a 10 años desde el último parto, un índice de masa corporal mayor a 30, antecedentes familiares de preeclampsia, edad materna de 35 años o más, y complicaciones en embarazos previos (Bartsch et al., 2016).

Actualmente, el tratamiento de la preeclampsia según la mayoría de los obstetras es el uso de la aspirina, ácido acetilsalicílico, en bajas dosis, el cual inhibe la ciclooxigenasa con propiedades antiinflamatorias y antiplaquetarias (Rolnik et al., 2022).

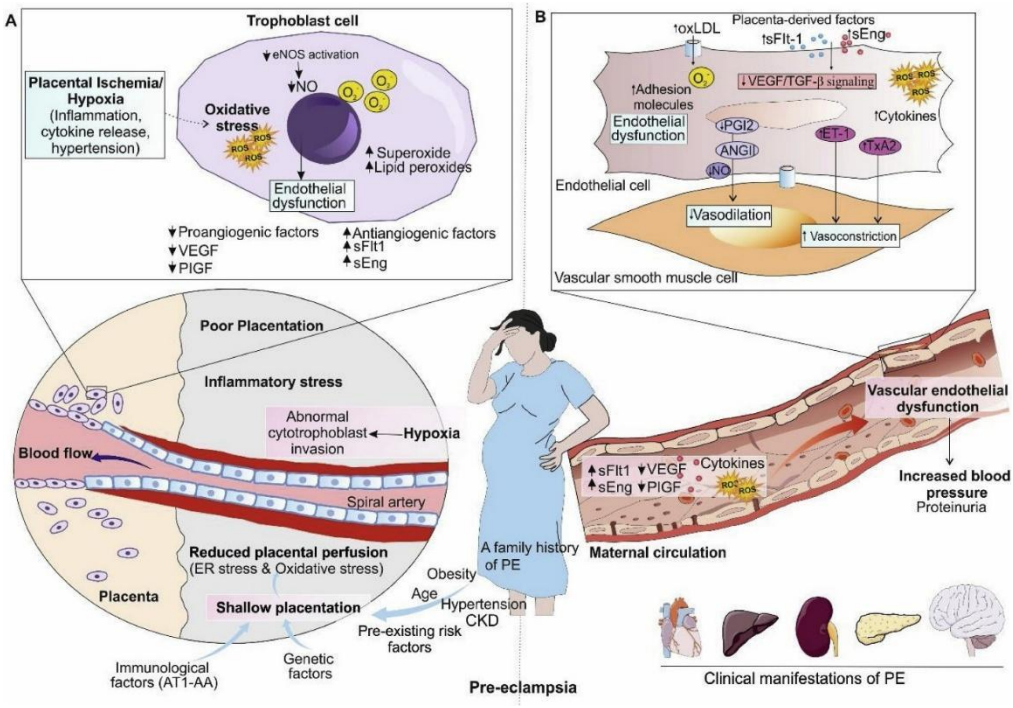
### *3.1.1 Fisiopatología*

Aunque la fisiopatología de la preeclampsia sigue sin estar clara, el proceso comienza en el primer trimestre de la gestación. Las principales características de la preeclampsia incluyen un defecto de placentación y la transformación incompleta de las arterias espirales, lo que conduce a hipoperfusión placentaria, hipoxia, estrés oxidativo y reducción del suministro nutricional al feto,

lo que resulta en restricción del crecimiento fetal. Como resultado, se producen una inflamación materna sistémica y daño endotelial por la liberación de factores derivados de la placenta en la circulación materna (Kolialexi et al., 2017; Redman & Sargent, 2009).

Esta enfermedad placentaria consta de dos etapas, como se observa en la Figura (Ahmadian et al., 2020).

**Figura 1. Fisiopatología de la preeclampsia**




*Nota:* Tomado de (Ahmadian et al., 2020)

***3.1.1.1 Etapa 1: Placentación anormal, invasión de trofoblastos e interfase materno-fetal. (Figura 1A)***

En esta primera etapa la placentación es anómala. Durante las primeras semanas de gestación, los trofoblastos extravellosos no invaden adecuadamente las arterias espirales del útero, lo que impide su conversión en vasos de baja resistencia. Debido a este cambio vascular deficiente, se conduce a una hipoperfusión e isquemia placentaria, que activan mecanismos moleculares como

el factor inducible por hipoxia 1 alfa (HIF-1α) y promueven la sobreexpresión de sFLT1 y endoglina soluble (sEng) (Rodríguez G et al., 2012). En la isquemia se liberan factores antiangiogénicos (sFlt-1, sEng) y disminuyen los factores proangiogénicos (VEGF, PlGF) de la placenta, lo que afecta la circulación materna y conlleva disfunción endotelial materna (Rana et al., 2019). En la isquemia placentaria disminuye la defensa antioxidante y, por el contrario, aumentan los peróxidos lipídicos, las RNS y ROS ROS, lo que produce estrés oxidativo. El estrés oxidativo provoca inflamación, oxidación de LDL, daño en proteínas, ADN y lípidos, y activación de células inmunitarias, principalmente neutrófilos y monocitos. Tanto el desequilibrio angiogénico como el estrés oxidativo dan lugar a disfunción endotelial (Ahmadian et al., 2020).

Además, como describe Cruz-Martínez (2024), este problema en la placentación puede tener un origen inmunológico, donde la interacción entre el HLA-C paterno y los receptores KIR maternos de las células NK uterinas puede promover o rechazar la tolerancia del aloinjerto fetal.

#### *3.1.1.2 Etapa 2: Patogenia del síndrome materno*

Después de la liberación de los factores antiangiogénicos y proinflamatorios al torrente sanguíneo, inicia la segunda etapa, en esta se observa una marcada disfunción endotelial sistémica, en donde el exceso de sFLT1 bloquea la acción de VEGF y reduce la producción de óxido nítrico, mientras que la endoglina interfiere con TGF-β1. Esto provoca un estado que causa la contracción de los vasos, favorece la coagulación y produce inflamación, lo que clínicamente se manifiesta mediante hipertensión, presencia de proteínas en la orina, hinchazón y daño en varios órganos, como el hígado, el cerebro y los riñones (Cruz-Pavlovich et al., 2023; Rana et al., 2019). En esta fase también se evidencian espasmos vasculares y activación plaquetaria, lo que agrava complicaciones como el síndrome HELLP y la eclampsia. También se reconocen citocinas que provocan inflamación, tales como TNF-α, IL-6 e IL-1, así como un desequilibrio en la relación

Th1/Th2 de los linfocitos CD4, que aumentan la inflamación sin infección en respuesta a las micropartículas del sincitiotrofoblasto (Cruz-Martínez, 2024). Las investigaciones experimentales han demostrado que la extracción del tejido placentario es el único tratamiento eficaz para prevenir la progresión hacia una falla orgánica múltiple (Sánchez Trigueros & Robles Selva, 2023).

### *3.1.2 Biomarcadores y relevancia en el diagnóstico.*

La detección de biomarcadores relacionados con la angiogénesis ha supuesto un avance significativo en la identificación temprana y la evaluación del riesgo de preeclampsia (Rodríguez, 2025). Aunque muchas organizaciones, como la ACOG (American College of Obstetricians and Gynecologists), no sugieren su implementación como método estándar de evaluación, se ha demostrado que estudios realizados en múltiples centros han mostrado la eficacia de la relación sFLT1/PlGF para anticipar el desarrollo de la enfermedad, particularmente antes de las 34 semanas de gestación. Una proporción elevada de sFLT1/PlGF se asocia con resultados adversos y con la posible aparición de preeclampsia, mientras que niveles bajos permiten descartar la enfermedad con gran certeza, lo que minimiza las hospitalizaciones innecesarias (Rana et al., 2019)

### *3.1.3 Preeclampsia en Colombia*

En Colombia, las enfermedades hipertensivas en el embarazo, como la preeclampsia, se han convertido en la principal causa directa de muerte materna, representando el 23,8 % de las muertes ocurridas durante el embarazo, el parto o el posparto inmediato (HUSI, 2020). Este conjunto de condiciones hipertensivas relacionadas con el embarazo (EHE) abarca la hipertensión gestacional, la preeclampsia, la eclampsia y el síndrome HELLP, trastornos que afectan múltiples sistemas y suponen un grave peligro tanto para la madre como para el feto (Greace et al., 2022). Su efecto no solo se manifiesta en las estadísticas de mortalidad, sino también en los índices de morbilidad materna extrema, lo que indica que muchas mujeres embarazadas enfrentan

complicaciones que pueden ser mortales, aun cuando logran sobrevivir. Resulta fundamental la necesidad urgente de mejorar la capacitación del personal de salud, promover la detección temprana de emergencias obstétricas y asegurar un manejo oportuno, adecuado y eficiente, para prevenir resultados fatales, especialmente en áreas con escasa cobertura médica.

Asimismo, la información más reciente del Instituto Nacional de Salud evidencia que, aunque la mortalidad materna en Colombia ha presentado una tendencia general al descenso en los últimos años, persisten importantes desigualdades sociales y demográficas (Instituto Nacional de Salud, 2025). Para 2025, los mayores riesgos se concentran en mujeres pertenecientes a poblaciones indígenas, residentes en áreas rurales dispersas y afiliadas al régimen subsidiado, así como en aquellas sin acceso a controles prenatales. Estos hallazgos refuerzan la relación entre las condiciones de vulnerabilidad social y el riesgo materno, evidenciando la necesidad de fortalecer estrategias de atención diferencial, acceso oportuno a los servicios de salud y vigilancia en salud pública para reducir la morbimortalidad materna en el país (Figura 2).

**Figura 2. Mortalidad materna según entidad territorial de residencia, Colombia, semana epidemiológica 21 de 2024 – 2025**

| Entidad territorial de residencia | Promedio histórico 2021-2024 a SE 21 | Acumulado de casos a SE 21 | |
|---|---|---|---|
| | | 2024 | 2025 |
| **Colombia** | **115** | **82** | **72** |
| Antioquia | 10 | 6 | 8 |
| La Guajira | 9 | 6 | 7 |
| Bogotá D.C. | 11 | 8 | 6 |
| Cesar | 5 | 7 | 5 |
| Chocó | 6 | 7 | 4 |
| Córdoba | 5 | 5 | 4 |
| Cundinamarca | 5 | 4 | 4 |
| Santiago de Cali D.E. | 3 | 2 | 3 |
| Casanare | 1 | 1 | 3 |
| Nariño | 5 | 4 | 3 |
| Atlántico | 3 | 1 | 2 |
| Barranquilla D.E. | 3 | 3 | 2 |
| Norte de Santander | 5 | 4 | 2 |
| Risaralda | 3 | 1 | 2 |
| Santander | 2 | 2 | 2 |
| Tolima | 3 | 0 | 2 |
| Valle del Cauca | 3 | 1 | 2 |
| Amazonas | 1 | 0 | 1 |
| Bolívar | 5 | 3 | 1 |
| Boyacá | 1 | 2 | 1 |
| Caquetá | 1 | 0 | 1 |
| Cauca | 4 | 3 | 1 |
| Guainía | 0 | 0 | 1 |
| Huila | 3 | 5 | 1 |
| Meta | 3 | 2 | 1 |
| Quindío | 1 | 0 | 1 |
| Sucre | 2 | 2 | 1 |
| Vichada | 1 | 0 | 1 |
| Arauca | 1 | 0 | 0 |
| Buenaventura D.E. | 2 | 0 | 0 |
| Caldas | 1 | 1 | 0 |
| Cartagena de Indias D.E. | 3 | 0 | 0 |
| Guaviare | 0 | 0 | 0 |
| Magdalena | 5 | 0 | 0 |
| Putumayo | 1 | 1 | 0 |
| Santa Marta D.T. | 4 | 1 | 0 |
| Archipiélago de San Andrés y Providencia | 0 | 0 | 0 |
| Vaupés | 0 | 0 | 0 |

*Nota:* Tomado de Fuente: Instituto Nacional de Salud, 2025 basado en datos de DANE y Sivigila (2024–2025).

## 3.2 Proteómica

La proteómica es el análisis de todo el sistema proteico de una célula, tejido, u organismo, en condiciones específicas (Yu et al., 2010) por medio de la cual se puede conocer los procesos

biológicos a nivel de proteínas (Duong & Lee, 2023; Rozanova et al., 2021). En la actualidad, el análisis proteómico está dado según el desarrollo tecnológico e instrumental. La separación, identificación y cuantificación de proteínas, requieren métodos de fraccionamiento de proteínas complejas o de mezclas de péptidos, la separación de proteínas/péptidos por medio de electroforesis en gel o por diversas técnicas cromatográficas, y la respectiva identificación de proteínas por medio de espectrometría de masas y finalmente un análisis bioinformático donde se analizan los espectros de masas(Cotes et al., 2013; Faktor et al., 2021) .

### *3.2.1 Proteómica en preeclampsia*

La proteómica ha transformado la investigación en el ámbito biomédico, se ha convertido en una herramienta transformadora al facilitar el estudio masivo de proteínas en líquidos biológicos como el plasma, el suero y la orina (Starodubtseva et al., 2025). La proteómica es fundamental en el descubrimiento de biomarcadores que diferencia entre mujeres sanas y aquellas que padecen la enfermedad (Navajas et al., 2021). De acuerdo con Starodubtseva et al., 2025, se han hallado más de 560 proteínas que presentan alteraciones, de las cuales 122 han sido confirmadas en diversos estudios, lo que subraya su importancia diagnóstica. También, se han diseñado paneles específicos según el trimestre del embarazo, lo que permite hacer predicciones más exactas en función del período gestacional. En particular, proteínas como sFLT1, PlGF, SERPINA-1, fibronectina y factores del complemento han mostrado vínculos sólidos con la fisiopatología de la enfermedad, que incluye inflamación, estrés oxidativo y disfunción endotelial (Jacobo-Baca et al., 2022; Zhao et al., 2025).

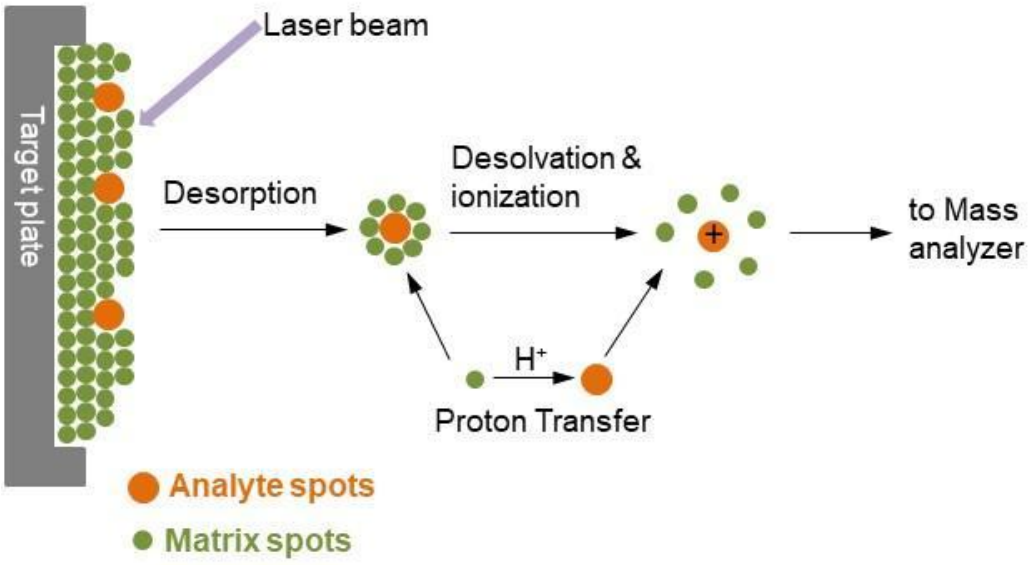
La proteómica proporciona una plataforma robusta para mejorar el diagnóstico, pronóstico y comprensión integral de esta afección obstétrica, sobre todo en contextos donde los métodos tradicionales son insuficientes (Faca et al., 2007; Rasanen et al., 2010).

### 3.3 Espectrometría de Masas MALDI-TOF

La espectrometría de masas (MS) se basa en la ionización de las moléculas de la muestra en fase gaseosa, así como en la separación y detección de los iones resultantes según la relación masa-carga (m/z). Según el tipo de ionización, las moléculas de la muestra se fragmentan idealmente, y así se obtienen los iones de producto. La abundancia de los iones con respecto a la relación m/z se observa en un espectro de masas. Esta información es útil, porque permite identificar la estructura de la molécula, por medio de la masa de los fragmentos (Smith, 2013).

En espectrometría de masas MALDI-TOF, la fuente de iones es la ionización/desorción láser asistida por matriz (MALDI) (Figura 3), y el analizador de masas es un analizador de tiempo de vuelo (TOF). La técnica MALDI implica una ionización suave que proporciona iones moleculares intactos del analito en fase gaseosa, es decir, sin fragmentar las moléculas (Gross, 2011).

**Figura 3. Ionización de analitos por MALDI**




*Nota:* Tomado de (Creative Proteomics, 2021).

La función de la matriz es principalmente diluir y aislar moléculas de analito unas de otra. Esta separación ocurre, mientras se evapora el solvente y se forma simultáneamente una solución sólida (Karas et al., 2000). Después de la irradiación con láser, la matriz funciona como un mediador para la absorción de la energía. En la Tabla 3, se observan las matrices de MALDI más comunes para el análisis de proteínas y péptidos (Gross, 2011a).

**Tabla 3. Matrices MALDI**

| Compuesto | Acrónimo |
|---|---|
| **Ácido nicotínico** | N/A |
| **Ácido 2,5-dihidroxibenzoico** | DHB |
| **Ácido α-ciano-4-hidroxicinámico** | α-CHC, α-CHCA, 4-HCCA, CHCA, HCCA |
| **Ácido 4-cloro-α-ciano-cinámico** | CICCA |
| **Ácido 3,5-dimetoxi-4-hidroxicinámico** | SA |
| **Ácido 2- (4-hidroxifenilazo) benzoico** | HABA |

*Nota:* Adaptado de (Gross, 2011)

El analizador de tiempo de vuelo TOF, se basa en que los iones de diferentes m/z se dispersan en el tiempo durante su vuelo, a lo largo de una trayectoria sin campo de longitud conocida. Cuando los iones comienzan su viaje al mismo tiempo o en un breve intervalo de tiempo, los más ligeros llegan más temprano al detector que los más pesados (Creative Proteomics, n.d.; Gross, 2011).

## 3.4 Método de preparación de muestra asistido por un filtro (*FASP: Filter-Aided Sample Preparation*)

En el año 2009, Jacek R. Wiśniewski reportó el método preparativo conocido como FASP, que se ha consolidado como un procedimiento eficaz y versátil para la preparación de muestras proteicas en estudios proteómicos bottom-up. Mediante este enfoque se puede trabajar con proteínas disueltas en detergentes como el dodecil sulfato de sodio (SDS), lo que simplifica su manejo y favorece condiciones ideales para la digestión enzimática (Wiśniewski, 2018).

### *3.4.1 Principio método FASP*

El principio del método FASP se basa en el uso de unidades de ultrafiltración centrífugas que actúan como reactores químicos cerrados. En estos amicones de ultrafiltración ocurren los siguientes pasos:

- Desnaturalización y eliminación de detergentes**:** las cuales se llevan a cabo mediante la aplicación de urea 8 M, que disocia micelas y mantiene las proteínas en solución.
- Modificación química: los residuos de cisteína suelen alquilarse con iodoacetamida para prevenir la formación de puentes disulfuro.
- Digestión enzimática: Las proteasas como tripsina, Lys-C y Arg-C, se emplean por su alta eficiencia en la generación de péptidos detectables por espectrometría de masas.
- Ultrafiltración: los péptidos fragmentos generados en la digestión enzimática atraviesan la membrana, mientras que las macromoléculas indeseadas quedan

retenidas, lo que asegura fragmentos de alta pureza (Nel et al., 2015; Wiśniewski, 2018).

## 3.5 Modelos estadísticos

Según McCullagh (2002), un modelo estadístico puede definirse como una representación matemática que describe la distribución de probabilidad de un conjunto de datos observables, condicionada por un conjunto de parámetros. Esta definición enfatiza que los modelos estadísticos no son simplemente herramientas de ajuste, sino estructuras formales que permiten inferencias sobre procesos subyacentes. Los modelos estadísticos no solo sirven para ajustar datos, sino que también ofrecen una base teórica para evaluar hipótesis, estimar parámetros y realizar predicciones (Davison, 2003).

### *3.5.1 Análisis de componentes principales (PCA)*

El Análisis de Componentes Principales (PCA) es un método estadístico multivariado que se utiliza para simplificar datos complejos a través de la reducción de su dimensionalidad, que puede ser empleado para extraer información de un espacio con alta dimensión al proyectarla en un subespacio de dimensión más baja, al mismo tiempo que se intenta conservar el máximo de varianza posible. Además, en esta técnica se convierte variables que pueden estar relacionadas en un conjunto de componentes principales que no están correlacionados, lo que facilita la identificación de patrones y relaciones importantes entre diversas variables (DataCamp, 2024; Souza, 2025)

## 3.6 Herramientas de análisis basadas en aprendizaje automático

El aprendizaje automático o Machine Learning, se define como la disciplina que investiga los algoritmos informáticos que se perfeccionan a partir de la experiencia. En términos generales, se trata de la creación de programas informáticos que se optimizan en función de las medidas de

evaluación mediante el uso de datos (Aracena et al., 2022). Este enfoque encuentra sus raíces en el trabajo de Alan Turing (TURING, 1950), quien propuso que una máquina digital también podría simular el comportamiento humano si estuviera adecuadamente programada. Su conocido "Juego de Imitación", también llamado prueba de *Turing*, sentó las bases filosóficas para la posibilidad de que las máquinas aprendieran y tomaran decisiones similares a las humanas.

Cuando se investiga el aprendizaje automático, se hace referencia a la creación de técnicas que pueden identificar automáticamente patrones en la información y emplearlos para anticipar eventos futuros o tomar decisiones en situaciones de incertidumbre. El aprendizaje automático se basa en la teoría de la probabilidad como herramienta esencial para extraer conclusiones (Murphy, 2012).

El aprendizaje automático puede clasificarse en tres tipos: **aprendizaje supervisado**, en el que los modelos se entrenan con datos etiquetados para tareas como la clasificación o la regresión; **aprendizaje no supervisado**, cuyo objetivo es descubrir estructuras ocultas en datos sin etiquetas; y *aprendizaje por refuerzo,* en el que los algoritmos adquieren conocimientos mediante la interacción con su entorno, recibiendo recompensas o castigos según sus acciones(Murphy, 2012).

### *3.6.1 Entorno computacional*

Jupyter Notebook es un entorno interactivo que combina código ejecutable, texto informativo, ecuaciones matemáticas, gráficos y resultados, todo en una única plataforma dinámica. Este sistema emplea archivos *.ipynb*, que permiten conservar y compartir todo el progreso del análisis. En JupyterLab, los notebooks se integran de manera versátil con consolas de programación, editores de texto, visualizadores de datos y terminales. También permite ejecutar código de notebooks simultáneamente y mostrar las salidas gráficas en diferentes pestañas, lo que facilita la exploración de datos y la documentación del proceso de análisis (JupyterLab, 2025).

Este entorno favorece la integración con Python, un lenguaje de programación muy popular en el ámbito del análisis de datos, la biomedicina y el aprendizaje automático.

En el entorno de *Jupyter Notebook*, existen herramientas que permiten cargar, procesar, visualizar y modelar datos sin necesidad de desarrollar código desde cero; se llaman librerías de Python, fundamentales en el análisis de datos. Las librerías más utilizadas se presentan en la Tabla 4.

**Tabla 4. Librerías de Python empleada para formular los modelos predictivos**

| Librería | Caracteristica |
|---|---|
| **Pandas** | Manipulación y análisis de datos tabulares con estructuras como DataFrame (Pandas, 2025). |
| **Numpy** | Computación numérica con arrays multidimensionales y operaciones matemáticas avanzadas (NumPy, 2025). |
| **matplotlib.pyplot** | Visualización de datos mediante gráficos estáticos (líneas, barras, histogramas) (Matplotlib, 2025) |
| **matplotlib inline** | Comando mágico de Jupyter para mostrar gráficos directamente en el notebook (IPython, 2025). |
| **sklearn.preprocessing** | Normalización y transformación de variables para modelos (Scikit-learn, 2009c). |
| **sklearn.decomposition.PCA** | Reducción de dimensionalidad mediante Análisis de Componentes Principales(Scikit-learn, 2009a). |
| **sklearn.metrics** | Evaluación de modelos predictivos con métricas como MAE, MSE, $R^2$ y varianza explicada (Scikit-learn, 2009b). |

### *3.6.2 Máquinas de vectores de soporte (SVM)*

Una máquina de vectores de soporte (SVM) es un algoritmo de aprendizaje supervisado que clasifica los datos mediante la identificación del hiperplano óptimo que maximiza la distancia entre las clases en un espacio N-dimensional. Esta técnica se basa en los vectores de soporte, que son los datos más cercanos a la frontera de decisión y definen el margen de separación. Cuando los datos no son linealmente separables, se emplean funciones kernel para transformarlos a espacios de mayor dimensión, lo que permite una separación más eficaz (IBM, 2023).

### *3.6.3 Redes neuronales*

Las redes neuronales artificiales (ANN) son modelos computacionales inspirados en la estructura y funcionamiento del cerebro humano. Su objetivo principal es reconocer patrones complejos y tomar decisiones a partir de datos, mediante un sistema de procesamiento distribuido y adaptativo (IBM, 2025).

### *3.6.4 Métricas de evaluación del desempeño*

La utilidad de un modelo depende en gran medida de su capacidad para realizar predicciones precisas y confiables. Por tal razón, es importante evaluar los modelos mediante métricas de desempeño, que se calculan con base en la matriz de confusión (Chukwura & Chukwura Obi, 2023).

#### 3.6.4.1 Matriz de confusión

La matriz de confusión es una tabla que muestra los resultados de las predicciones de un modelo frente a los valores observados. En un problema binario, la tabla de confusión se organiza como se ilustra en la Tabla 5.

**Tabla 5. Ilustración esquemática de una matriz de confusión**

| | | Valores reales | |
|---|---|---|---|
| | | P | N |
| Valores predichos | P | Verdadero positivo (VP) | Falso positivo (FP) |
| | N | Falso Negativo (FN) | Verdadero negativo (VN) |

De acuerdo con esta matriz, es posible calcular métricas importantes como precisión, sensibilidad o *recall*, especificidad, exactitud, F1-Score.

#### 3.6.4.2 Precisión

La precisión se define como la proporción de verdaderos positivos (TP) respecto de la suma total de positivos observados. Esta suma abarca tanto los verdaderos positivos (TP) como los falsos positivos (FP), que equivalen a los verdaderos negativos clasificados incorrectamente como positivos, de acuerdo con la ecuación 1.

$$Precisión = \frac{TP}{TP + FP} \qquad \textbf{(ecuación 1)}$$

#### 3.6.4.3 Sensibilidad

La sensibilidad o recall, evalúa la capacidad de un modelo para detectar adecuadamente los verdaderos positivos (TP). El total de positivos abarca tanto los verdaderos positivos (TP) como los falsos negativos (FN), es decir los positivos que fueron clasificados incorrectamente como negativos. Una alta sensibilidad indica que el modelo detecta la mayoría de los positivos reales, con pocos falsos negativos.

$$Sensibilidad = \frac{TP}{TP + FN} \qquad \textbf{( 2 )}$$

#### 3.6.4.4 Especificidad

La especificidad, evalúa la proporción de verdaderos negativos que un modelo identifica como negativos. Es esencial para valorar la efectividad del modelo en reconocer adecuadamente los casos que no son positivos.

$$Especificidad = \frac{TN}{TN + FP} \quad \textbf{( 3 )}$$

#### 3.6.4.5 Exactitud

La exactitud, se describe como la relación entre las clasificaciones acertadas y el total de conjunto de datos de prueba. Es decir, la exactitud mide la proporción de predicciones acertadas que un modelo realiza respecto de todas las posibles.

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{( 4 )}$$

#### 3.6.4.6 F1-Score

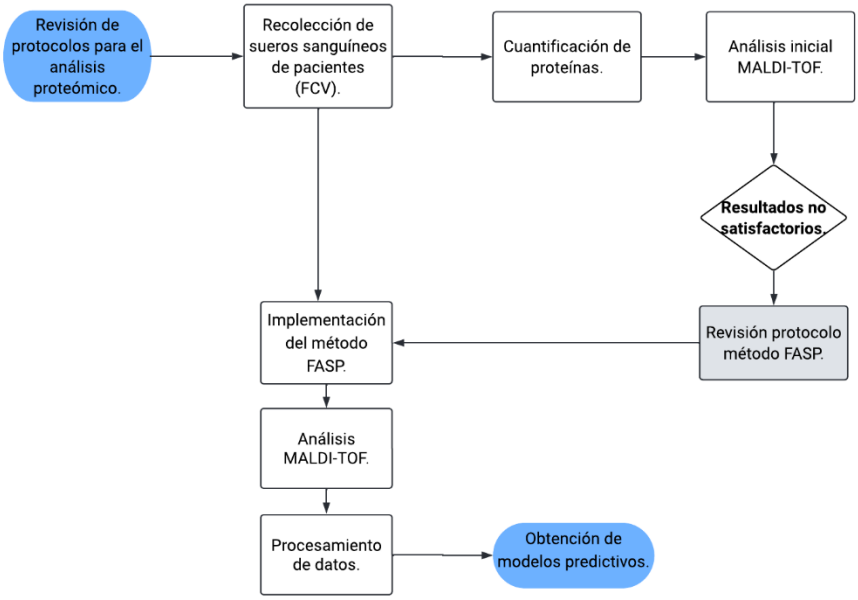
El F1-Score, o F-Score o medida F, es la media armónica de la precisión y la sensibilidad, y permite estimar la capacidad de clasificación de una prueba; es decir, el F1-Score nos permite evaluar el desempeño de un modelo (Molina Arias, 2024).

$$F1 - Score = 2 \times \frac{(Precision \times Sensibilidad)}{(Precision + Sensibilidad)} \quad \textbf{( 5 )}$$

## 4. Metodología

Con el propósito de realizar el análisis proteómico discriminatorio de sueros sanguíneos de pacientes con preeclampsia y mujeres con parto normal empleando la espectrometría de masas MALDI-TOF, se estableció una secuencia metodológica estructurada, desde la revisión de protocolos en artículos de investigación, hasta la implementación del método eficiente de preparación para el análisis MALDI-TOF y finalmente el análisis de los espectros de masas adquiridos con la información del perfil proteómico de cada grupo empleando el algoritmo de VSM y redes neuronales. El siguiente diagrama de flujos (Figura 4) representa la metodología general utilizada:

**Figura 4. Diagrama de flujo del procedimiento experimental para el análisis proteómico de sueros sanguíneos mediante MALDI-TOF**



## 4.1 Materiales y reactivos

**Tabla 6. Materiales y reactivos usados en el desarrollo del proyecto de investigación**

| Reactivo/Material | Grado/Pureza | Proveedor |
|---|---|---|
| **Agua desionizada** | 18,2 MΩ.cm | Milli-Q |
| **Urea** | Grado analítico | Merck Millipore |
| **Tris-HCl** | Grado analítico | Merck Millipore |
| **Acetonitrilo (ACN)** | Grado HPLC | Merck Millipore |
| **Cloroformo** | Grado HPLC | Merck Millipore |
| **Ácido tricloroacético (TFA)** | Grado analítico | Merck Millipore |
| **Tripsina** | Grado secuenciación | Thermo Fisher Bioreagents |
| **Ditiotreitol (DTT)** | Grado análisis | Thermo Fisher Bioreagents |
| **Iodoacetamida (IAA)** | Grado análisis | Thermo Fisher Bioreagents |
| **Filtros Amicon de 3 kDa** | Porcentaje de retencion >95% | Merck Millipore |
| **Buffer fosfato salino (PBS)** | Reactivos de alta pureza | No especificado |
| **Acetona fría** | Grado analítico >98% | Merck Millipore |
| **Buffer Britton-Robinson (BR)** | Reactivos de alta pureza | Biorad |
| **Matriz HCCA** | Grado analítico espectromtria de masas MALDI | Sigma-Aldrich |
| **Solución de acrilamida/bis acrilamida 30%** | Grado elcetroforesis .98% | Biorad |
| **Dodecilsulfato de sodio (SDS)** | Reactivos de alta pureza | No especificado |
| **Buffer Tris HCl 1 M, pH 8,8** | Reactivos de alta pureza | No reportado |
| **Buffer Tris HCl 0,5 M, pH 6,8** | Reactivos de alta pureza | No reportado |

| | | |
|---|---|---|
| **Buffer de carga (Laemmli 2x)** | Reactivos de alta pureza | No reportado |

### 4.2 Recolección y selección de muestra

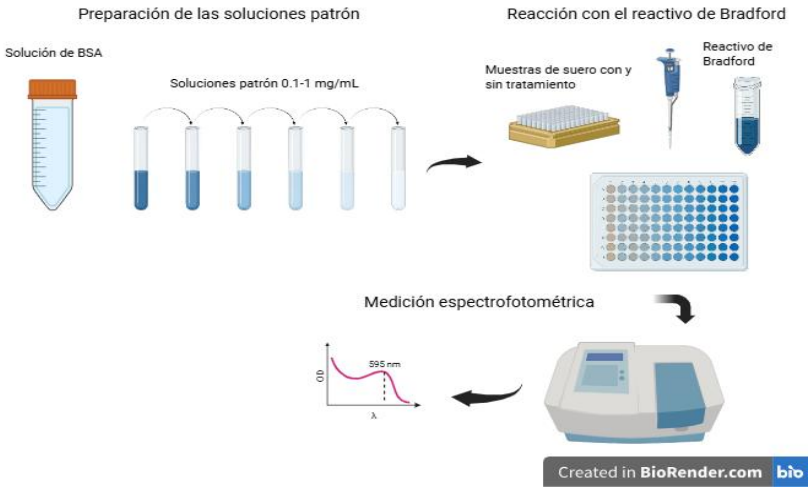
Las muestras de suero sanguíneo fueron suministradas por la Fundación Cardiovascular (FCV) a través del biobanco del proyecto GenPE, donde han permanecido almacenadas desde 2004 hasta la fecha. Las muestras fueron colectadas durante los años 2004-2012 de mujeres embarazadas que padecieron preeclampsia y de mujeres embarazadas en diferentes zonas demográficas del país, bajo las normativas éticas correspondientes, y han permanecido almacenadas a -85 °C para mantener su integridad. Para efectos de este proyecto, las muestras se seleccionaron de manera aleatorizada desde el biobanco y se procesaron para efecto del presente trabajo en el 2022. El estudio incluyó un total de 212 muestras, de las cuales 100 correspondieron a pacientes con preeclampsia y 112 a gestantes sin la enfermedad (grupo de control).

### 4.3 Determinación de concentración de proteínas

La cuantificación de proteínas totales en las muestras de suero se realizó mediante el método colorimétrico de Bradford, siguiendo el protocolo de Merck Millipore (2020), que se basa en la interacción del colorante Coomassie Brilliant Blue G-250 con residuos básicos y aromáticos de las proteínas, principalmente arginina, lisina, histidina y fenilalanina. En condiciones ácidas, el colorante presenta una forma catiónica de color rojo-amarillo, que, al unirse a las proteínas, se estabiliza en su forma aniónica de color azul, desplazando el máximo de absorción desde 465 nm hasta aproximadamente 595 nm. La intensidad del color desarrollado es proporcional a la concentración de proteína en la muestra.

El esquema del procedimiento de cuantificación de proteínas se muestra en la Figura 5.

**Figura 5. Determinación de concentración de proteínas**




*Nota:* Elaboración propia en BioRender.

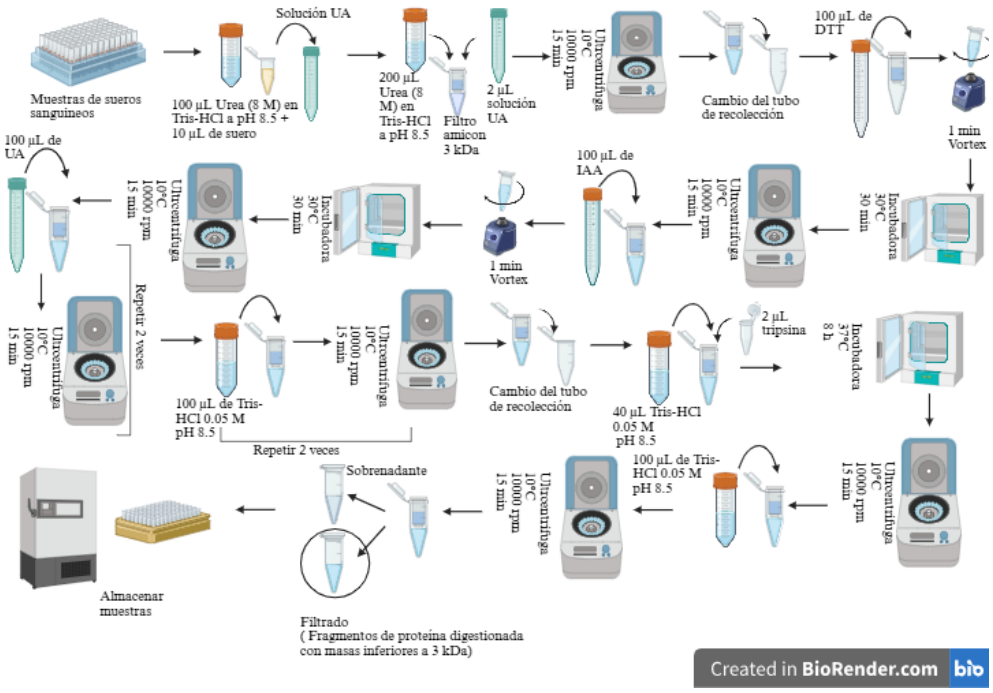
Se preparó una curva de calibración utilizando albúmina sérica bovina (BSA) como patrón, en un rango de concentraciones previamente establecido. Las muestras de suero y las previamente tratadas fueron diluidas y analizadas en duplicado. La medición de la absorbancia se realizó en un espectrofotómetro UV-Vis a 595 nm, utilizando como blanco la mezcla de reactivos sin proteína.

### 4.4 Implementación del método FASP para preparación de muestras

Como un nuevo procedimiento para el procesamiento del suero sanguíneo se empleó la técnica de preparación de muestras asistida por filtración (FASP) (Wiśniewski, 2018), utilizando unidades de ultrafiltración Amicon de 3 kDa como reactores proteómicos. La configuración del reactor (amicon) facilitó la eliminación eficiente de reactivos mediante centrifugaciones sucesivas, lo que garantizó un entorno óptimo para la digestión. Este método permitió la adición sucesiva de los reactivos requeridos en el proceso de digestión triptica, como la reducción con DTT y una alquilación con la iodacetamida de los residuos de cisteína directamente sobre la membrana del filtro. La digestión de las proteínas se llevó a cabo mediante la adición de tripsina en una relación

enzima-sustrato de 1:100 y la incubación a 37 °C durante 8 horas. Finalizada la proteólisis, los péptidos resultantes con masas inferiores al límite de corte del filtro (3 kDa) se recuperaron por centrifugación en el tubo de recolección, mientras que los fragmentos de mayor tamaño y la enzima permanecieron retenidos en la membrana. El procedimiento detallado de las etapas de lavado e incubación, así como los parámetros de centrifugación, se presenta en la Figura 6.

**Figura 6. Metodología método FASP**



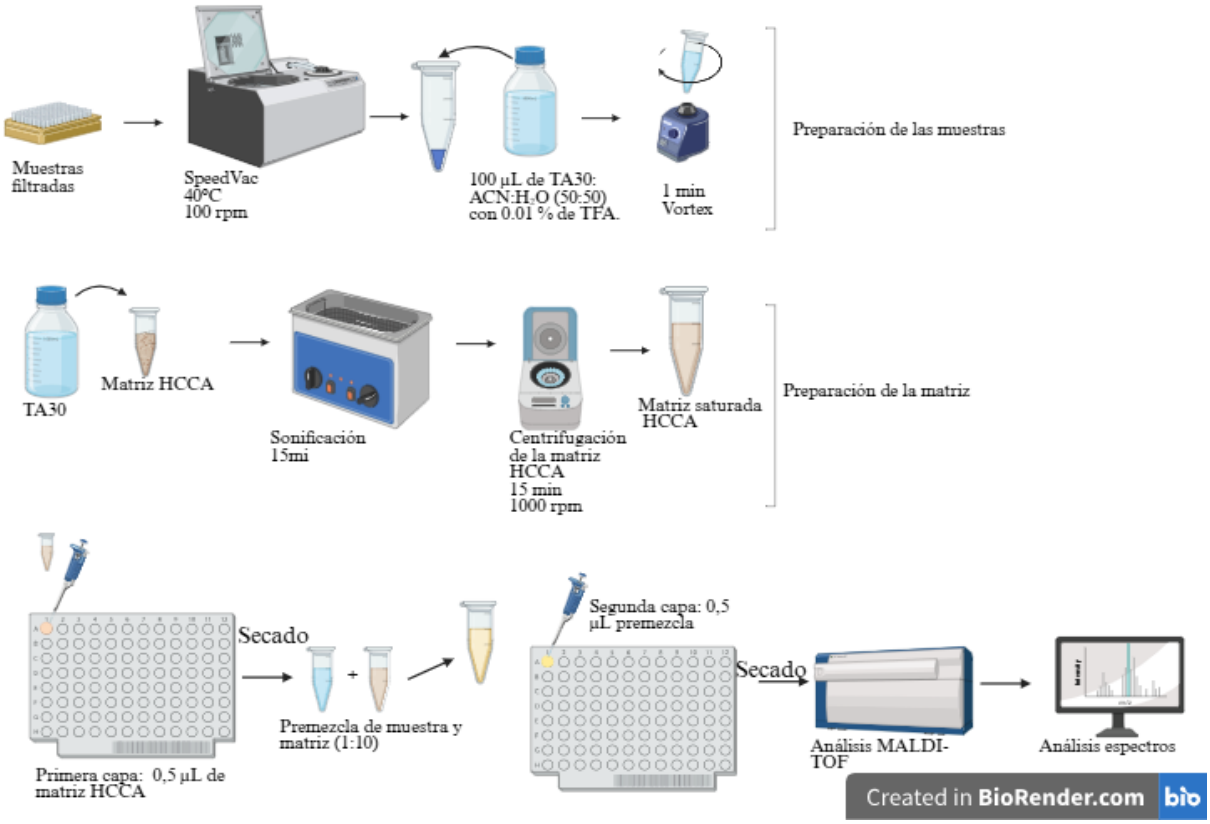

*Nota:* Elaboración propia en BioRender.

## 4.5 Análisis MALDI-TOF/MS

Los péptidos recuperados del proceso anterior fueron concentrados mediante un sistema de vacío (SpeedVac) a 40 °C y resuspendidos en una mezcla de ACN$_2$O (50:50) con 0,01 % de TFA. Para el análisis, se empleó la técnica de cristalización de doble capa en el portamuestras, utilizando

como matriz una solución saturada de HCCA. El proceso de preparación de la matriz y de las muestras, junto con la técnica de depósito sobre la portamuestra, se ilustran en la Figura 7. Adicionalmente, se realizó la adquisición de espectros sin digestión previa de las muestras, con el fin de evaluar directamente el perfil proteico y comparar su comportamiento frente a las muestras digeridas.

**Figura 7. Metodología análisis MALDI-TOF/MS**




*Nota:* Elaboración propia en BioRender.

La adquisición de espectros se realizó en un espectrómetro Ultraflex II MALDI-TOF-TOF, operado en modo positivo, con un voltaje de aceleración de 89 kV. Se empleó el método de adquisición PR-700-3500_Da.par, optimizado para el rango de masa de los fragmentos obtenidos. Finalmente, los espectros fueron procesados y suavizados con el software flexAnalysis 3.3.

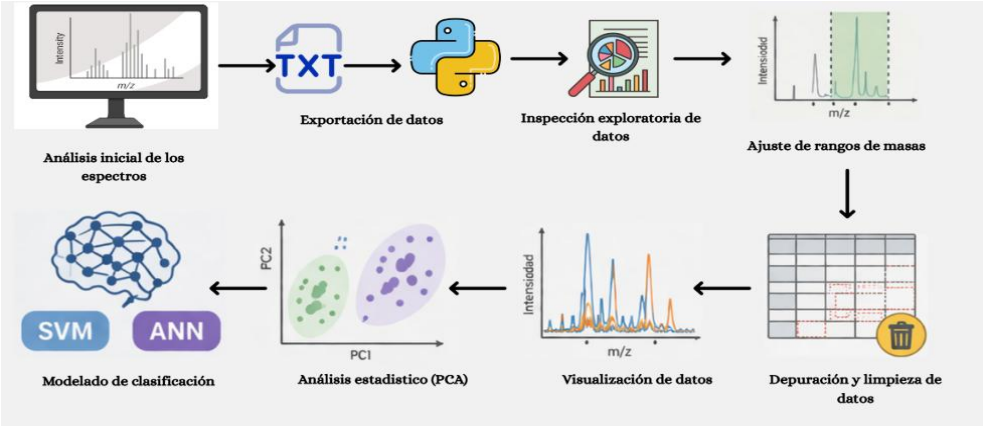
### 4.6 Procesamiento de datos

El tratamiento de los datos abarcó la visualización preliminar de los espectros en FlexAnalysis, su posterior transferencia a Python y una fase de revisión y corrección de los datos. A lo largo de esta etapa, se modificaron los intervalos de masas y se descartaron los espectros que contenían información insuficiente.

Posteriormente, se generaron representaciones gráficas y se aplicó el Análisis de Componentes Principales (PCA) para explorar la variabilidad de los datos. Finalmente, se implementaron modelos de clasificación supervisada, específicamente Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Artificiales (ANN), con el fin de evaluar la capacidad de discriminación entre las muestras.

El flujo general del análisis y procesamiento de datos se resume en la Figura 8.

**Figura 8. Análisis y procesamiento de datos**




*Nota:* Elaboración propia en Canva.

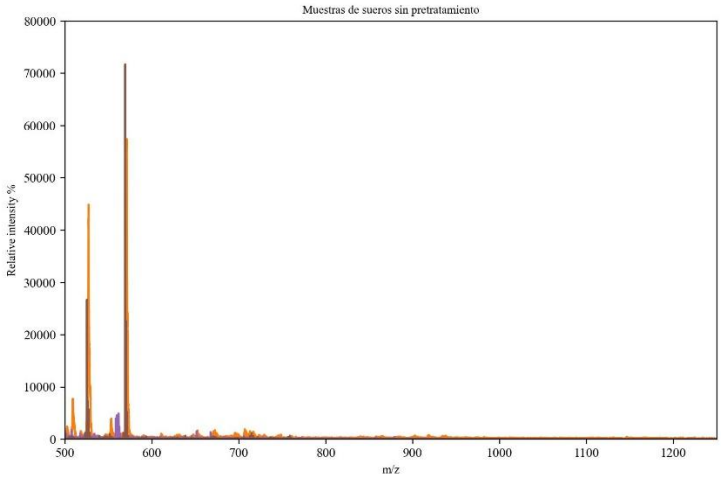
## 5. Resultados y análisis

### 5.1 Adquisición de espectros sin digestión previa

Se realizó la adquisición de espectros directamente a partir de las muestras sin digestión previa con el fin de evaluar el perfil proteico en estado nativo. Sin embargo, los espectros obtenidos, al ser superpuestos (Figura 9), evidenciaron la presencia predominante de señales características de la matriz utilizada (HCCA), las cuales coinciden con el espectro de la matriz presentado en el Apéndice A. Esto indica que no se logró una adecuada ionización diferencial de los componentes proteicos de la muestra, generando señales poco representativas y de baja utilidad analítica.

Este resultado justifica la necesidad de implementar estrategias de preparación de muestra que permitan mejorar la resolución y detección de péptidos, ya que el análisis directo sin un proceso adecuado limita la identificación de proteínas de interés, probablemente debido a la complejidad de la muestra y a efectos de supresión iónica. En consecuencia, se confirma la importancia de realizar la digestión proteica mediante el método FASP, el cual facilita la obtención de péptidos más adecuados para su análisis por espectrometría de masas.

**Figura 9. Espectros de masas de muestras de suero sanguíneo sin pretratamiento.**




## 5.2 Determinación de concentración de proteínas

Con el fin de descartar que la ausencia de señales peptídicas estuviera asociada a una baja concentración proteica, se seleccionó un subconjunto representativo de aproximadamente 12 muestras de suero para su cuantificación. Los resultados, obtenidos a partir de la curva de calibración (0–1 mg/mL) presentada en el Apéndice B, mostraron concentraciones en el rango de 90 a 197 µg/mL.

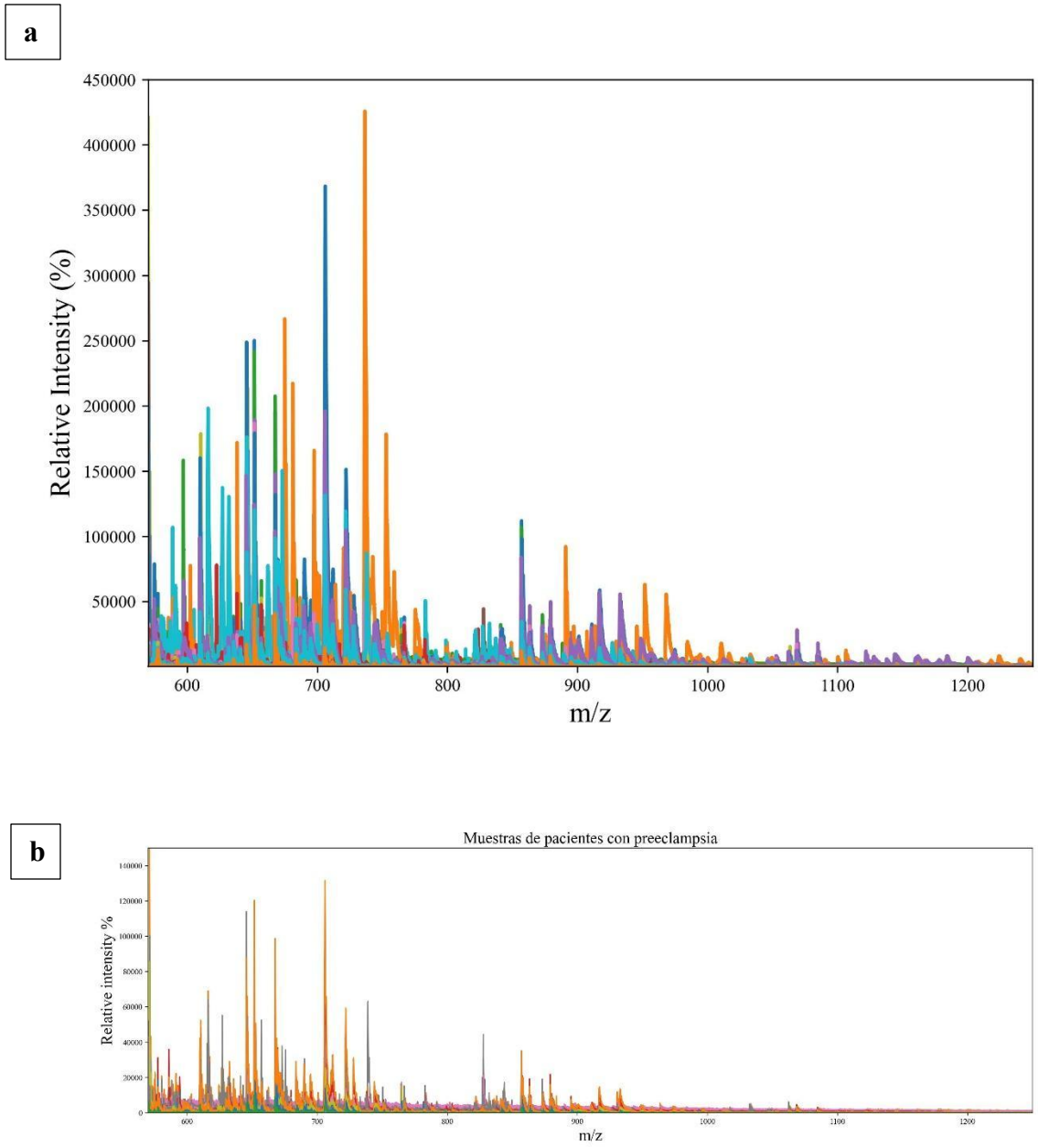
Estos valores obtenidos permiten descartar la baja concentración de proteínas como causa de la ausencia de señales detectables. Además, este rango de concentración indica que las muestras contienen suficiente material proteico para llevar a cabo de manera adecuada la digestión enzimática, permitiendo establecer proporciones apropiadas entre proteína y enzima. La falta de respuesta analítica en los espectros sin pretratamiento podría estar asociada a factores como la complejidad de la matriz biológica o efectos de supresión iónica (Van Belkum et al., 2017).

### 5.3 Perfil de proteínas MALDI-TOF/MS

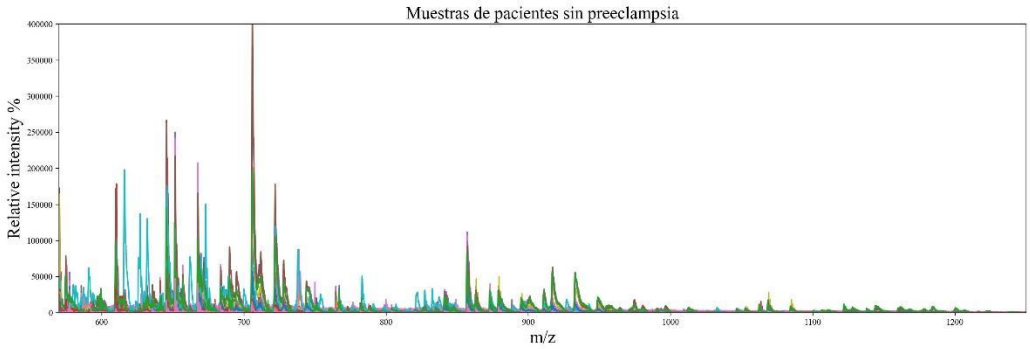
Los espectros se adquirieron utilizando el método *PR-700-3500_Da.par* incorporado en el equipo y fueron procesados mediante el software flexAnalysis versión 3.3 (Bruker Daltonics GmbH, Alemania). La huella peptídica proveniente de las proteínas en las 212 muestras, se analizaron en un intervalo señales con relación m/z de 400 a 3500 Da. A continuación, se observan los diferentes espectros obtenidos y sus representaciones gráficas.

La Figura 10A muestra el espectro promedio de todas las muestras analizadas (sin distinción entre positivas y negativas), lo que permite observar el patrón global de señales y la densidad de picos en el rango estudiado. En la Figura 10B se presentan los espectros totales del subconjunto de muestras positivas (preeclampsia), mientras que la Figura 10C corresponde al espectro promedio del subconjunto de muestras negativas (control). Estas figuras permiten comparar visualmente las regiones con mayor intensidad y detectar picos con comportamiento diferencial entre grupos.

**Figura 10. Espectros de muestras totales.**



c




Los espectros obtenidos evidencian un patrón general similar entre las muestras, tanto en el conjunto total como en los grupos de pacientes con y sin preeclampsia, lo que propone la presencia de una huella peptídica compartida característica del suero sanguíneo (Wen et al., 2013). En todos los casos, las señales corresponden a péptidos ionizados y se concentran principalmente en el rango de m/z entre 500 y 800, donde se observa la mayor densidad de picos y las intensidades más elevadas, consistente con la detección de péptidos de bajo peso molecular generados tras la digestión.

A partir de valores de m/z cercanos a 900–1000, se evidencia una disminución progresiva tanto en la intensidad de las señales como en la cantidad de picos detectables, lo que sugiere una menor abundancia o eficiencia de ionización de péptidos de mayor masa. Este comportamiento se mantiene de forma consistente en los tres conjuntos analizados.

Los perfiles espectrales presentan similitudes en cuanto a la distribución de los picos, sin embargo, se observan diferencias notables en las intensidades relativas entre los grupos. En particular, las muestras correspondientes a pacientes sin preeclampsia muestran picos de mayor intensidad y variabilidad, mientras que las muestras de pacientes con preeclampsia presentan intensidades relativamente menores y perfiles más homogéneos. Estas diferencias podrían estar

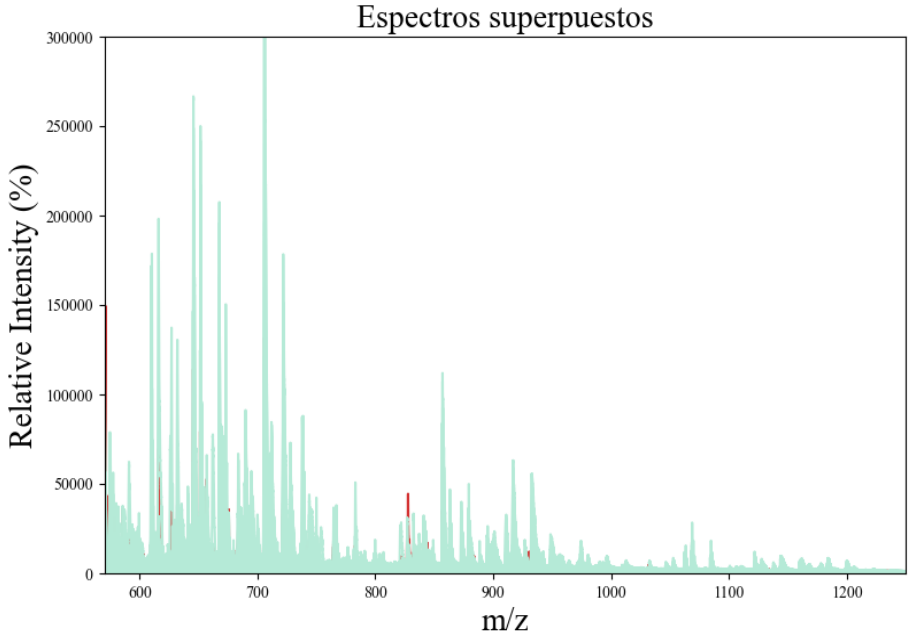
asociadas a variaciones en la abundancia relativa de péptidos derivados de proteínas específicas relacionadas con la preeclampsia (Park et al., 2011).

En el análisis total se identifican picos recurrentes en múltiples muestras, lo que sugiere la presencia de péptidos comunes del suero, mientras que señales de menor intensidad podrían corresponder a variaciones individuales o a péptidos de baja abundancia. De manera general, estos resultados indican que, aunque existe una huella peptídica base compartida, las variaciones en la intensidad de las señales podrían reflejar diferencias biológicas relevantes entre los grupos.

La Figura 11 contiene la superposición de los espectros promedio de muestras positivas y negativas, facilitando la identificación visual de picos que difieren en intensidad entre ambos grupos. Los espectros de color **rojo** corresponden a las muestras positivas (preeclampsia) y los espectros con color **azul** a las muestras negativas (sin preeclampsia).

Los espectros superpuestos permiten visualizar con mayor claridad que las señales correspondientes a las muestras sin preeclampsia presentan mayores intensidades relativas, lo que genera un efecto de predominio visual sobre las muestras con preeclampsia, cuyas señales tienden a verse atenuadas u opacadas en la superposición. A pesar de que ambos grupos comparten un patrón general de distribución en el rango de m/z, este comportamiento sugiere diferencias en la intensidad de las señales, lo que podría estar asociado a una mayor abundancia relativa de ciertos péptidos en el grupo control.

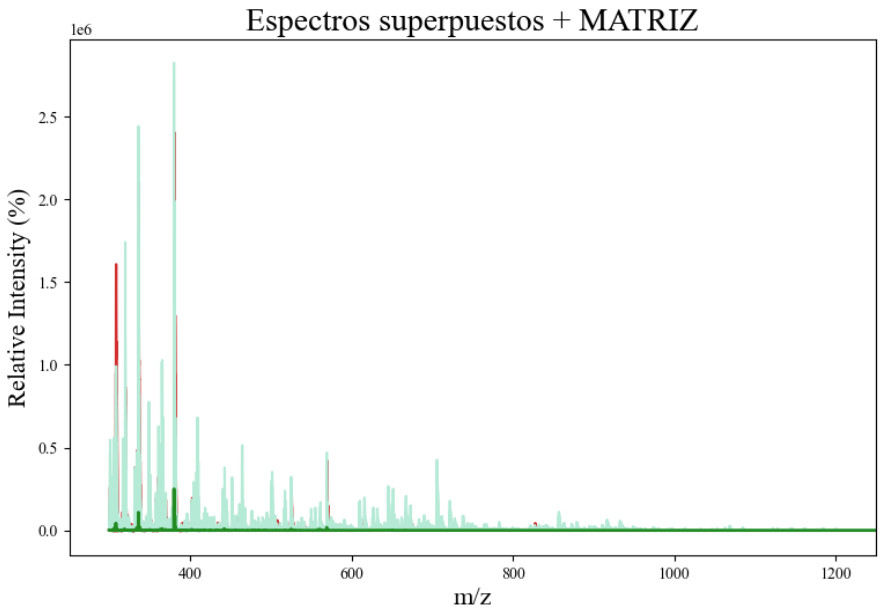
**Figura 11. Espectro de muestras con preeclampsia y sin preeclampsia superpuestos**




En la Figura 12 se agrega además el espectro de la matriz (color **verde**) a la superposición (muestras positivas + negativas + matriz), con el propósito de identificar picos atribuibles exclusivamente a la matriz y evaluar su posible interferencia en los perfiles biológicos.

En la comparación de los espectros con el perfil de la matriz, se observa que, aunque existen señales características de HCCA en la región de bajo m/z, la cantidad y diversidad de picos observados en las muestras de suero es considerablemente mayor, este resultado demuestra que las señales registradas no corresponden exclusivamente a interferencias de la matriz, sino que reflejan de manera significativa la huella peptídica propia del suero sanguíneo, así, poder diferenciar claramente entre las señales aportadas por la matriz y los péptidos ionizados de las muestras analizadas.

**Figura 12. Espectro de muestras con preeclampsia, sin preeclampsia y matriz superpuestos**




Aunque se evidencian diferencias en las intensidades de las señales, con tendencia a valores mayores en el grupo control, la variabilidad intragrupo y la coincidencia de picos limitan una discriminación clara basada únicamente en la inspección visual. Por ello, fue necesario complementar el análisis con métodos multivariados (PCA) y modelos de predicción.

## 5.4 Análisis de datos empleando Machine learning

El análisis y el procesamiento estadístico de los datos obtenidos a través de la espectrometría de masas se llevó a cabo utilizando el lenguaje de programación Python, y se utilizó la herramienta interactiva Jupyter Notebook como el entorno de desarrollo para realizar depuraciones y exploraciones iniciales.

Primero, se realizó un procedimiento de depuración de datos, que abarcó la detección y el tratamiento de valores nulos o erróneos, para asegurar la calidad y la fiabilidad de los datos. Luego,

se organizaron y estructuraron los datos en *dataframes* utilizando la biblioteca *pandas*, lo que facilitó su manipulación y análisis.

Para redactar y ejecutar scripts adicionales, así como para la administración completa del proyecto, se utilizó el entorno de desarrollo integrado Visual Studio Code, aprovechando sus capacidades de control de versiones y depuración.

El código completo utilizado para la depuración, la organización y el análisis de los datos está disponible públicamente en el repositorio del proyecto en GitHub y en el Apéndice C del presente documento, donde también se incluyen los análisis exploratorios, como el PCA, y la implementación de los modelos predictivos.

### 5.5 Análisis exploratorio y modelos predictivos

Antes de proceder con la etapa de clasificación, se llevó a cabo un Análisis de Componentes Principales (PCA) como técnica de análisis exploratorio (Bro & Smilde, 2014). El objetivo del análisis es reducir la dimensionalidad de los datos espectrales obtenidos mediante MALDI-TOF, así como evaluar de manera visual la existencia de patrones, agrupamientos o tendencias naturales entre las muestras correspondientes a pacientes con preeclampsia y el grupo control.

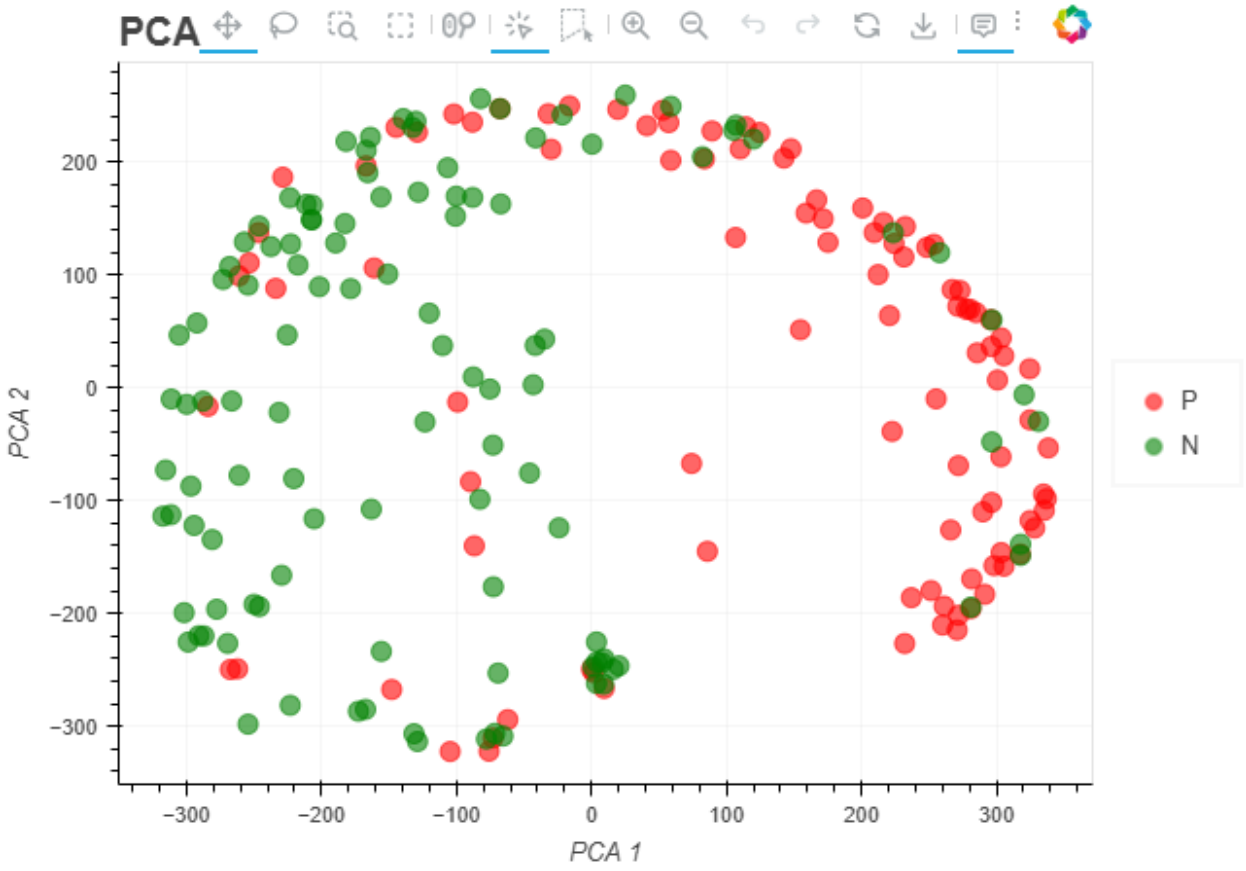
Posteriormente, con la información obtenida en el análisis preliminar, se llevaron a cabo la implementación y comparación de dos técnicas de aprendizaje supervisado: Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Artificiales (ANN). Estos modelos fueron preparados para clasificar las muestras, con el objetivo de medir su rendimiento con relación a indicadores de medición como la sensibilidad y la especificidad.

Para la formación y evaluación de los modelos, se dividió el conjunto de datos, destinando un 80% para el entrenamiento y un 20% para las pruebas, asegurando que el modelo fuese evaluado con datos que no se habían utilizado en la fase de entrenamiento.

### 5.6. Análisis de componentes principales (PCA)

Se realizó un análisis de componentes principales (PCA) de los datos proteómicos, con el objetivo de reducir la dimensionalidad y explorar la estructura interna de los datos de las muestras. Inicialmente, los datos fueron estandarizados para garantizar que todas las variables contribuyeran de manera equitativa al análisis. A partir de este procedimiento se obtuvieron nueve componentes principales que representan combinaciones lineales de las variables originales. Como se observa en la figura 13.

**Figura 13. Grafica de análisis de componentes principales (PCA)**




Este resultado en el análisis de PCA no muestra una separación clásica de los grupos, lo que sugiere una distribución no lineal (semicírculo) de tipo herradura, lo que evidencia una transición progresiva entre los grupos con solapamiento en algunos puntos, lo cual representa una

variación no específica entre casos y controles que no permite la separación de los clústeres (Shah et al., 2024). En el caso de la preeclampsia, este patrón puede atribuirse a la heterogeneidad de los datos al ser un trastorno multifactorial y de severidad variable, en donde la expresión se proteínas evidencian un continuo de estado fisiopatológicos sin separación binaria entre grupos.

En la Tabla 7 se presentan las varianzas explicadas y acumuladas de los PC. La primera componente principal (PC1) explicó el 29,07% de la variabilidad, lo que indica que captura una fracción baja de la variabilidad fenotípica en los datos. La segunda componente (PC2) presentó una varianza del 20,45%, lo que aporta información adicional.

**Tabla 7. Varianza explicada y acumulada de los componentes principales (PCA)**

| PC | Varianza Explicada | Varianza Acumulada |
|---|---|---|
| PC1 | 0.290711 | 0.290711 |
| PC2 | 0.204477 | 0.495188 |
| PC3 | 0.100954 | 0.596143 |
| PC4 | 0.071295 | 0.667438 |
| PC5 | 0.055750 | 0.723188 |
| PC6 | 0.035079 | 0.758267 |
| PC7 | 0.029850 | 0.788117 |
| PC8 | 0.024625 | 0.812741 |
| PC9 | 0.021847 | 0.834589 |

En total, las dos componentes juntas logran una varianza acumulada de 49.51, lo que indica que cerca de la mitad de la información completa del conjunto de datos puede representarse con estas dos dimensiones. Al incluir las nueve componentes principales, la varianza acumulada

aumenta al 83.46%, lo que muestra que se retiene una parte considerable de la información original tras el proceso de reducción de dimensiones.

La proyección de las muestras en el espacio de PC1 y PC2 (Figura 13) evidenció un comportamiento multidimensional entre pacientes con preeclampsia (P) y el grupo de control (N), que podría ser el resultado de múltiples procesos biológicos simultáneos, en los que se afecta la expresión de proteínas y no predomina un solo eje biológico (Assani et al., 2026; Le et al., 2019).

Se observó también que las muestras de pacientes con preeclampsia tienden a ubicarse en valores positivos de PC1, mientras que el grupo de control se posiciona mayormente en valores negativos. Sin embargo, hay solapamiento entre ambos grupos, lo que sugiere que la distinción no es del todo clara

### *5.6.1 Máquinas de vectores de soporte (SVM)*

Se implementó el modelo de Máquinas de Soporte Vectorial (SVM) para la clasificación de las muestras de pacientes con preeclampsia (P) y del grupo control (N). Para ello, la variable dependiente fue previamente codificada en valores numéricos, donde el grupo control (N) fue representado como (0.0) y los pacientes con preeclampsia (P) como (1.0).

El modelo se estableció con un **kernel de tipo radial** (RBF), el cual es apto para identificar conexiones no lineales entre las variables, de acuerdo con la estructura que se notó en el análisis exploratorio. Además, se utilizó un parámetro de regularización **C = 40**, que permite el equilibrio entre la correcta clasificación de los casos y la complejidad del modelo, mientras que el parámetro gamma se definió como ***scale***.

Para el entrenamiento y evaluación del modelo, el conjunto de datos fue dividido en un 80% para entrenamiento y un 20% para prueba. La distribución de las muestras por clase en cada

conjunto se presenta en la Tabla 8, donde se observa el balance entre pacientes con preeclampsia y el grupo control.

**Tabla 8. Distribución de muestras en los conjuntos de entrenamiento y prueba.**

| Conjunto | Control (N) | Preeclampsia (P) | Total |
|---|---|---|---|
| **Entrenamiento** | 89 | 80 | 169 |
| **Prueba** | 23 | 20 | 43 |
| **Total** | 112 | 100 | 212 |

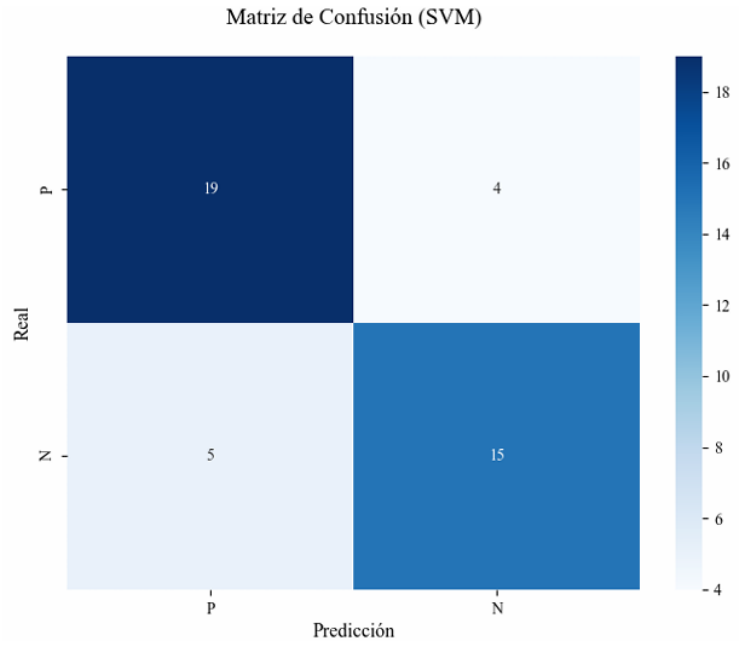
El desempeño del modelo se evaluó mediante métricas de clasificación, cuyos resultados se presentan en la Tabla 9. El modelo alcanzó una exactitud del 79%, lo que evidencia una adecuada capacidad de clasificación global. Para el grupo de control (N), se obtuvo una precisión de 0.79, un recall de 0.83 y un F1-score de 0.81. Por su parte, para los pacientes con preeclampsia (P), el modelo presentó una precisión de 0.79, un recall de 0.75 y un F1-score de 0.77.

**Tabla 9. Métricas de desempeño del modelo SVM**

| Clase | Precisión | Recall | F1-score | Soporte |
|---|---|---|---|---|
| **0.0** | 0.79 | 0.83 | 0.81 | 23 |
| **1.0** | 0.79 | 0.75 | 0.77 | 20 |

Adicionalmente, la matriz de confusión de la Figura 14, permitió analizar con mayor detalle el comportamiento del modelo. Se identificaron 19 verdaderos positivos y 15 verdaderos negativos, así como 4 falsos positivos y 5 falsos negativos.

**Figura 14. Matriz de confusión modelo SVM**




El algoritmo SVM mostró un desempeño global adecuado, con una exactitud del 79,1 % y una sensibilidad para la clasificación de preeclampsia del 73,1 %, mientras que la especificidad alcanzó el 88,2 %, indicando una alta capacidad para identificar correctamente a los controles sanos. Asimismo, la precisión del modelo fue elevada (90,5 %), lo que sugiere que las predicciones positivas son confiables. Sin embargo, la presencia de falsos negativos (4) evidencia limitaciones en la clasificación de todos los casos afectados, lo cual podría estar asociado a la heterogeneidad biológica de la enfermedad y al solapamiento observado en el análisis de componentes principales. En conjunto, estos resultados respaldan el potencial del perfil proteómico sérico como herramienta diagnóstica, aunque subrayan la necesidad de optimizar los modelos para aumentar la sensibilidad.

### *5.6.2 Redes neuronales*

El modelo de red neuronal artificial (ANN) se implementó con el objetivo de clasificar las muestras de pacientes con preeclampsia (P) y del grupo control (N). Tal como se describió anteriormente para el modelo SVM, la variable dependiente fue codificada en valores numéricos,

donde el grupo control (N) se representó como (0.0) y los pacientes con preeclampsia (P) como (1.0). La configuración de modelo se estableció mediante una arquitectura de dos capas ocultas, cada una con 20 neuronas **(hidden_layer_sizes = (20, 20))**, lo que permite capturar relaciones complejas en los datos. Se utilizó una función de activación *ReLU* y una tasa de aprendizaje inicial de 0.01 **(***learning_rate_init***)**. En la optimización, se empleó el algoritmo ***LBFGS***, adecuado para conjuntos de datos de tamaño mediano. De igual forma, se estableció un máximo de 5000 iteraciones, para así garantizar la convergencia del modelo.

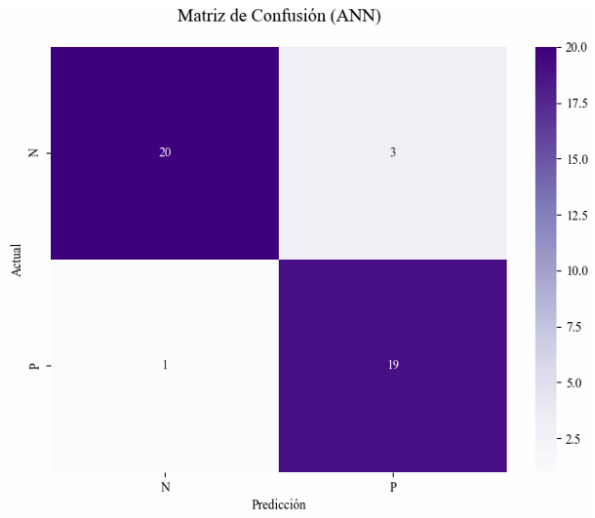
En el entrenamiento y la prueba, se utilizó la misma división de datos descrita, la cual corresponde a un 80% para entrenamiento y un 20% para prueba, manteniendo la distribución de clases presentada en la Tabla 8.

**Tabla 10. Métricas de desempeño del modelo ANN**

| Clase | Precisión | Recall | F1-score | Soporte |
|---|---|---|---|---|
| 0.0 | 0.95 | 0.87 | 0.91 | 23 |
| 1.0 | 0.86 | 0.95 | 0.90 | 20 |

El desempeño de las métricas del modelo se presenta en la Tabla 10, donde se observa que la red neuronal obtuvo una exactitud del 90,70%. Cuya matriz de confusión, representada en la Figura 15, demuestra que el modelo clasificó correctamente 20 muestras del grupo de control y 22 de preeclampsia, con 3 falsos positivos y 1 falso negativo, lo que revela un alto desempeño en la clasificación de ambas clases.

**Figura 15. Matriz de confusión modelo ANN**




A diferencia del modelo SVM descrito, el modelo basado en ANN mostró un desempeño robusto en la clasificación de muestras, con una exactitud global del 90,70 %. Destaca su alta sensibilidad (95%), lo que indica una notable capacidad para identificar correctamente a las pacientes afectadas. Adicionalmente, obtuvo una precisión del 86 % y un *F1-score* de 0,90 para esta clase. En contraste, la clasificación de controles alcanzó una precisión del 95% y un *recall* del 87%. Además el modelo presentó un menor número de falsos positivos que el SVM, un valor de sensibilidad más alto sugiere un mejor desempeño en contextos clínicos en los que la detección temprana de la preeclampsia es prioritaria. Estos resultados reflejan la capacidad del ANN para capturar relaciones no lineales en los perfiles proteómicos séricos, de manera más efectiva dada las relaciones complejas presentes en los datos proteómicos.

## 6. Conclusiones

La combinación de la espectrometría de masas MALDI-TOF y el machine learning permitió analizar patrones significativos en los perfiles de proteínas de gestantes con preeclampsia y de mujeres con un embarazo normal, lo que confirma el potencial y la versatilidad de esta técnica para el análisis de muestras biológicas complejas en estudios de casos y controles, lo cual contribuyó al desarrollo de estrategias para la identificación de patrones asociados a esta enfermedad en el contexto nacional dentro de estudios multiómicos.

Se implementó el método preparativo asistido por un filtro, FASP, el cual permitió la obtención de los fragmentos de proteínas de bajo peso molecular mediante un proceso de digestión enzimática asistido por filtración en membrana, el cual mejoró significativamente la calidad de los espectros obtenidos y permitió el análisis de un perfil proteómico asociado a la preeclampsia.

Aunque el análisis de los datos mediante el método estadístico (PCA) no logró una separación clara entre grupos, que permitiera clasificar los casos y controles; sí logro corroborar la complejidad multifactorial que condiciona la predicción temprana de la preeclampsia.

Por otra parte, los modelos de aprendizaje automático como SVM y ANN mostraron ser una herramienta eficiente en la formulación de modelos predictivos basados los perfiles proteómicos entre pacientes con preeclampsia y el grupo de control.

En términos de desempeño, el modelo de Máquinas de Soporte Vectorial alcanzó una exactitud del 79%, mientras que el modelo de Red Neuronal Artificial presentó un mejor comportamiento, con una exactitud del 90.70% y una mayor capacidad para detectar casos de preeclampsia, con un recall de 0,95.

## 7. Recomendaciones

Se recomienda la implementación continua del método FASP como estrategia de preparación de muestras, dada su efectividad al garantizar la integridad, limpieza y concentración de los fragmentos proteicos, un factor clave en la obtención de espectros de calidad en el análisis por MALDI-TOF/MS. Adicionalmente, se propone ampliar el tamaño de las muestras en una nueva cohorte, lo que permitiría enriquecer y mejorar la robustez de los modelos de aprendizaje automático y optimizar su capacidad de predicción.

## Referencias Bibliográficas

Ahmadian, E., Rahbar Saadat, Y., Hosseiniyan Khatibi, S. M., Nariman-Saleh-Fam, Z., Bastami, M., Zununi Vahed, F., Ardalan, M., & Zununi Vahed, S. (2020). Pre-Eclampsia: Microbiota possibly playing a role. In *Pharmacological Research* (Vol. 155, p. 104692). Academic Press. https://doi.org/10.1016/j.phrs.2020.104692

American College of Obstetricians & Gynecologists. (2021). *Infographic: Preeclampsia and Pregnancy | ACOG*. https://www.acog.org/womens-health/infographics/preeclampsia-and-pregnancy

Anand, S., Young, S. A., Esplin, M. S., Peaden, B., Tolley, H. D., Porter, T. F., Varner, M. W., D'Alton, M. E., Jackson, B. J., & Graves, S. W. (2016). Detection and confirmation of serum lipid biomarkers for preeclampsia using direct infusion mass spectrometry. *Journal of Lipid Research*, *57*(4), 687–696. https://doi.org/10.1194/jlr.P064451

Andresen, I. J., Romero, R., Westerberg, A. C., Than, N. G., Gomez-Lopez, N., Bhatti, G., Ahmodu, O., Gudicha, D. W., Meyyazhagan, A., Awonuga, A., Chaiworapongsa, T., Bryant, D. R., Michelsen, T. M., & Tarca, A. L. (2025). Large-scale proteomics reveals new candidate biomarkers for late-onset preeclampsia. *Hypertension (Dallas, Tex. : 1979)*, *83*(2), e25189. https://doi.org/10.1161/HYPERTENSIONAHA.125.25189

Aracena, C., Villena, F., Arias, F., & Dunstan, J. (2022). Aplicaciones de aprendizaje automático en salud. *Revista Médica Clínica Las Condes*, *33*(6), 568–575. https://doi.org/10.1016/J.RMCLC.2022.10.001

Assani, A. D., Boldeanu, L., Novac, M. B., Assani, M. Z., Siloși, I., Boldeanu, M. V., Dijmărescu, A. L., Manolea, M. M., Dinescu, V. C., & Văduva, C. C. (2026). Angiogenic Imbalance in Preeclampsia: Profiling VEGF A, sFlt1, PlGF, and sFlt1/PlGF Ratios. *International Journal*

*of Molecular Sciences 2026, Vol. 27, Page 2438*, *27*(5), 2438. https://doi.org/10.3390/IJMS27052438

Bahado-Singh, R., Poon, L. C., Yilmaz, A., Syngelaki, A., Turkoglu, O., Kumar, P., Kirma, J., Allos, M., Accurti, V., Li, J., Zhao, P., Graham, S. F., Cool, D. R., & Nicolaides, K. (2017). Integrated Proteomic and Metabolomic prediction of Term Preeclampsia. *Scientific Reports*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-15882-9

Bartsch, E., Medcalf, K. E., Park, A. L., Ray, J. G., Al-Rubaie, Z. T. A., Askie, L. M., Berger, H., Blake, J., Graves, L., Kingdom, J. C., Lebovic, G., Lord, S. J., Maguire, J. L., Mamdani, M. M., Meloche, J., Urquia, M. L., & Van Wagner, V. (2016). Clinical risk factors for pre-eclampsia determined in early pregnancy: systematic review and meta-analysis of large cohort studies. *BMJ (Clinical Research Ed.)*, *353*. https://doi.org/10.1136/BMJ.I1753

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812–2831. https://doi.org/10.1039/C3AY41907J

Chaemsaithong, P., Sahota, D. S., & Poon, L. C. (2022). First trimester preeclampsia screening and prediction. *American Journal of Obstetrics and Gynecology*, *226*(2), S1071-S1097.e2. https://doi.org/10.1016/j.ajog.2020.07.020

Chen, H., Aneman, I., Nikolic, V., Karadzov Orlic, N., Mikovic, Z., Stefanovic, M., Cakic, Z., Jovanovic, H., Town, S. E. L., Padula, M. P., & McClements, L. (2022). Maternal plasma proteome profiling of biomarkers and pathogenic mechanisms of early-onset and late-onset preeclampsia. *Scientific Reports 2022 12:1*, *12*(1), 19099-. https://doi.org/10.1038/s41598-022-20658-x

Chukwura, J., & Chukwura Obi, J. (2023). A comparative study of several classification metrics and their performances on data. *Https://Wjaets.Com/Sites/Default/Files/WJAETS-2023-0054.Pdf*, *8*(1), 308–314. https://doi.org/10.30574/WJAETS.2023.8.1.0054

Cotes, E. J., Sánchez, L. M., Grisales, N. V., Vélez, C. A., & Trujillo, I. O. (2013). Preeclampsia: la evolución diagnóstica desde la genómica y la proteómica. *Revista Chilena de Obstetricia y Ginecología*, *78*(2), 148–153. https://doi.org/10.4067/S0717-75262013000200014

Creative Proteomics. (n.d.). *MALDI-TOF Mass Spectrometry - Creative Proteomics*. Retrieved April 20, 2026, from https://www.creative-proteomics.com/technology/maldi-tof-mass-spectrometry.htm

Creative Proteomics. (2021). *MALDI-TOF Mass Spectrometry - Creative Proteomics*. https://www.creative-proteomics.com/technology/maldi-tof-mass-spectrometry.htm

Cruz-Martínez, F. (2024). Fisiopatología de la preeclamsia placentaria. *Arch Med Urgen Mex*, *16*(1), 37–44. https://doi.org/10.35366/115761

Cruz-Pavlovich, S., Salmeron-Salcedo, F., Ponce-Rivera, C. S., & Luna-Flores, M. (2023). ARTÍCULO DE REVISIÓN PREECLAMPSIA: REVISIÓN. *Preeclampsia: Revisión. Artículo de Revisión. Revista Homeostasis*, *2023*(5).

Cunningham, F. G., Leveno, K. J., Dashe, J. S., Hoffman, B. L., Spong, C. Y., & Casey, B. M. (2022). Preeclampsia Syndrome. In *Williams Obstetrics, 26e*. McGraw Hill. obgyn.mhmedical.com/content.aspx?aid=1190762794

DataCamp. (2024). *Tutorial de Análisis de Componentes Principales (ACP) en Python | DataCamp*. https://www.datacamp.com/es/tutorial/principal-component-analysis-in-python

Davison, A. C. (2003). Statistical Models. *Statistical Models*. https://doi.org/10.1017/CBO9780511815850

Duckitt, K., & Harrington, D. (2005). Risk factors for pre-eclampsia at antenatal booking: systematic review of controlled studies. *BMJ : British Medical Journal*, *330*(7491), 565. https://doi.org/10.1136/BMJ.38380.674340.E0

Duong, V. A., & Lee, H. (2023). Bottom-Up Proteomics: Advancements in Sample Preparation. *International Journal of Molecular Sciences*, *24*(6). https://doi.org/10.3390/IJMS24065350

Faca, V., Pitteri, S. J., Newcomb, L., Glukhova, V., Phanstiel, D., Krasnoselsky, A., Zhang, Q., Struthers, J., Wang, H., Eng, J., Fitzgibbon, M., McIntosh, M., & Hanash, S. (2007). Contribution of Protein Fractionation to Depth of Analysis of the Serum and Plasma Proteomes. *Journal of Proteome Research*, *6*(9), 3558–3565. https://doi.org/10.1021/PR070233Q

Faktor, J., R. Goodlett, D., & Dapic, I. (2021). Trends in Sample Preparation for Proteome Analysis. *Mass Spectrometry in Life Sciences and Clinical Laboratory*. https://doi.org/10.5772/INTECHOPEN.95962

Gallos, I., Sivakumar, K., Kilby, M., Coomarasamy, A., Thangaratinam, S., & Vatish, M. (2013). Pre-eclampsia is associated with, and preceded by, hypertriglyceridaemia: a meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology*, *120*(11), 1321–1332. https://doi.org/10.1111/1471-0528.12375

Gilbert, J. S., Ryan, M. J., Lamarca, B. B., Sedeek, M., Murphy, S. R., & Granger, J. P. (2008). Pathophysiology of hypertension during preeclampsia: Linking placental ischemia with endothelial dysfunction. In *American Journal of Physiology - Heart and Circulatory Physiology* (Vol. 294, Number 2). Am J Physiol Heart Circ Physiol. https://doi.org/10.1152/ajpheart.01113.2007

Greace, C., Ávila, A., Subdirector, M., Marcela, D., Acero, W., María, A., Bedoya, G., Alejandra, G., Mellizo, Á., Narváez, N. S., Patricia, H., & Suspes, S. (2022). *Informe de mortalidad materna, Colombia, 2022 INSTITUTO NACIONAL DE SALUD Elaborado por: Revisado por: Aprobado por*.

Gross, J. H. (2011). Matrix-Assisted Laser Desorption/Ionization. In *Mass Spectrometry: A Textbook* (pp. 507–559). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10711-5_11

Hua, Y., Wang, J., Yuan, D. L., Qi, Y., Tang, Z., Zhu, X., & Jiang, S. W. (2018). A tag SNP in syncytin-2 3-UTR significantly correlates with the risk of severe preeclampsia. *Clinica Chimica Acta*, *483*, 265–270. https://doi.org/10.1016/j.cca.2018.05.013

HUSI. (2020). *Preeclampsia, la principal causa de mortalidad materna en Colombia - HUSI en los medios - HUSI*. https://www.husi.org.co/el-husi-hoy/husi-en-los-medios/-/asset_publisher/rMVQOyye5rdo/content/preeclampsia-la-principal-causa-de-mortalidad-materna-en-colombia

IBM. (2023). *¿Qué es Support Vector Machine? | IBM*. https://www.ibm.com/mx-es/think/topics/support-vector-machine

IBM. (2025). *¿Qué es una red neuronal?* . https://www.ibm.com/think/topics/neural-networks

Instituto Nacional de Salud (INS). (2025). *Semana epidemiológica 18 al 24 de mayo de 2025*. https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2025_Boletin_epidemiologico_semana_21.pdf

IPython. (2025). *Built-in magic commands — IPython 9.4.0 documentation*. https://ipython.readthedocs.io/en/stable/interactive/magics.html

Jacobo-Baca, G., Salazar-Ybarra, R. A., Torres-de-la-Cruz, V., Guzmán-López, S., Elizondo-Omaña, R. E., Guzmán-López, A., Vázquez-Barragán, M. Á., & Martínez-de-Villarreal, L. E. (2022). Proteomic profile of preeclampsia in the first trimester of pregnancy. *The Journal of Maternal-Fetal & Neonatal Medicine : The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*, *35*(18), 3446–3452. https://doi.org/10.1080/14767058.2020.1820980

Jain, K. K. (2001). Proteomics: new technologies and their applications. *Drug Discovery Today*, *6*(9), 457–459. https://doi.org/10.1016/S1359-6446(01)01785-8

JupyterLab. (2025). *Get Started — JupyterLab 4.4.4 documentation*. https://jupyterlab.readthedocs.io/en/stable/getting_started/overview.html

Karas, M., Glückmann, M., & Schäfer, J. (2000). Ionization in matrix-assisted laser desorption/ionization: Singly charged molecular ions are the lucky survivors. *Journal of Mass Spectrometry*, *35*(1), 1–12. https://doi.org/10.1002/(SICI)1096-9888(200001)35:1<1::AID-JMS904>3.0.CO;2-0

Kolialexi, A., Tsangaris, G. T., Sifakis, S., Gourgiotis, D., Katsafadou, A., Lykoudi, A., Marmarinos, A., Mavreli, D., Pergialiotis, V., Fexi, D., Mavrou, A., Papaioanou, G. K., & Papantoniou, N. (2017). Plasma biomarkers for the identification of women at risk for early-onset preeclampsia. *Expert Review of Proteomics*, *14*(3), 269–276. https://doi.org/10.1080/14789450.2017.1291345

Le, Y., Ye, J., & Lin, J. (2019). Expectant management of early-onset severe preeclampsia: a principal component analysis. *Annals of Translational Medicine*, *7*(20), 519–519. https://doi.org/10.21037/ATM.2019.10.11

Matplotlib. (2025). *Matplotlib documentation — Matplotlib 3.10.3 documentation*. https://matplotlib.org/stable/index.html

McCullagh, P. (2002). What is a statistical model? *Https://Doi.Org/10.1214/Aos/1035844977*, *30*(5), 1225–1310. https://doi.org/10.1214/AOS/1035844977

Merck Millipore. (2020). *Proteínas (según el método de Bradford)*. 1–3. www.sigmaaldrich.com

Molina Arias, M. (2024). Un intruso de otro mundo: F1-score. *Revista Electrónica AnestesiaR, ISSN-e 1989-4090, Vol. 16, Nº. 4, 2024*, *16*(4), 3. https://dialnet.unirioja.es/servlet/articulo?codigo=9532114&info=resumen&idioma=SPA

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* (MIT Press).

National High Blood Pressure Education Program. (2000). *WORKING GROUP REPORT ON HIGH BLOOD PRESSURE IN PREGNANCY National High Blood Pressure Education Program*.

Navajas, R., Corrales, F., & Paradela, A. (2021). Quantitative proteomics-based analyses performed on pre-eclampsia samples in the 2004–2020 period: a systematic review. *Clinical Proteomics*, *18*(1), 1–12. https://doi.org/10.1186/S12014-021-09313-1/TABLES/1

Nel, A. J. M., Garnett, S., Blackburn, J. M., & Soares, N. C. (2015). Comparative reevaluation of FASP and enhanced FASP methods by LC - MS/MS. *Journal of Proteome Research*, *14*(3), 1637–1642. https://doi.org/10.1021/PR501266C/SUPPL_FILE/PR501266C_SI_001.XLSX

Nirupama, R., Divyashree, S., Janhavi, P., Muthukumar, S. P., & Ravindra, P. V. (2021). Preeclampsia: Pathophysiology and management. In *Journal of Gynecology Obstetrics and Human Reproduction* (Vol. 50, Number 2, p. 101975). Elsevier Masson s.r.l. https://doi.org/10.1016/j.jogoh.2020.101975

Noroña Calvachi, C. D. (2014). Preeclampsia: la Era de los Marcadores Bioquímicos. *Revista Científica Ciencia Médica*, *17*, 32–38.

NumPy. (2025). *NumPy documentation — NumPy v2.3 Manual*. https://numpy.org/doc/stable/

Pandas. (2025). *pandas documentation — pandas 2.3.1 documentation*. https://pandas.pydata.org/docs/

Park, J., Cha, D. H., Lee, S. J., Kim, Y. N., Kim, Y. H., & Kim, K. P. (2011). Discovery of the serum biomarker proteins in severe preeclampsia by proteomic analysis. *Experimental & Molecular Medicine*, *43*(7), 427. https://doi.org/10.3858/EMM.2011.43.7.047

Pasyar, S., Wilson, L. M., Pudwell, J., Peng, Y. P., & Smith, G. N. (2020). Investigating the diagnostic capacity of uric acid in the occurrence of preeclampsia. *Pregnancy Hypertension*, *19*, 106–111. https://doi.org/10.1016/j.preghy.2019.12.010

Rana, S., Lemoine, E., Granger, J., & Karumanchi, S. A. (2019). Preeclampsia: Pathophysiology, Challenges, and Perspectives. *Circulation Research*, *124*(7), 1094–1112. https://doi.org/10.1161/CIRCRESAHA.118.313276/ASSET/093CF6BB-9496-4C85-9583-5C41E1DFFFEA/ASSETS/IMAGES/LARGE/1094FIG03.JPG

Rasanen, J., Girsen, A., Lu, X., Lapidus, J. A., Standley, M., Reddy, A., Dasari, S., Thomas, A., Jacob, T., Pouta, A., Surcel, H. M., Tolosa, J. E., Gravett, M. G., & Nagalla, S. R. (2010). Comprehensive maternal serum proteomic profiles of preclinical and clinical preeclampsia. *Journal of Proteome Research*, *9*(8), 4274–4281. https://doi.org/10.1021/PR100198M

Redman, C. W. G., & Sargent, I. L. (2009). Placental Stress and Pre-eclampsia: A Revised View. *Placenta*, *30*, 38–42. https://doi.org/https://doi.org/10.1016/j.placenta.2008.11.021

Rodríguez, A. F. A. (2025). Biomarkers: A promising alternative for the early detection of the risk of preeclampsia. *Ginecologia y Obstetricia de Mexico*, *93*(5), 184–191. https://doi.org/10.24245/GOM.V93I5.153

Rodríguez G, M., Egaña, G., Márquez, R., Bachmann, M., & Soto, A. (2012). Preeclampsia: mediadores moleculares del daño placentario. *Revista Chilena de Obstetricia y Ginecología*, *77*(1), 72–78. https://doi.org/10.4067/S0717-75262012000100014

Rolnik, D. L., Nicolaides, K. H., & Poon, L. C. (2022). Prevention of preeclampsia with aspirin. *American Journal of Obstetrics and Gynecology*, *226*(2), S1108–S1119. https://doi.org/10.1016/J.AJOG.2020.08.045

Rozanova, S., Barkovits, K., Nikolov, M., Schmidt, C., Urlaub, H., & Marcus, K. (2021). Quantitative Mass Spectrometry-Based Proteomics: An Overview. *Methods in Molecular Biology (Clifton, N.J.)*, *2228*, 85–116. https://doi.org/10.1007/978-1-0716-1024-4_8

Rybak-Krzyszkowska, M., Staniczek, J., Kondracka, A., Bogusławska, J., Kwiatkowski, S., Góra, T., Strus, M., & Górczewski, W. (2023). From Biomarkers to the Molecular Mechanism of Preeclampsia-A Comprehensive Literature Review. *International Journal of Molecular Sciences*, *24*(17). https://doi.org/10.3390/IJMS241713252

Sánchez, K. (2018). Preeclampsia. In *REVISTA MEDICA SINERGIA* (Vol. 3, Number 3).

Sánchez Trigueros, M. J., & Robles Selva, J. A. (2023). *Preeclampsia: Etiopatogenia y fisiopatología*. https://www.revista-portalesmedicos.com/revista-medica/preeclampsia-etiopatogenia-y-fisiopatologia/

Scikit-learn. (2009a). *2.5. Decomposing signals in components (matrix factorization problems) — scikit-learn 1.7.0 documentation*. https://scikit-learn.org/stable/modules/decomposition.html#pca

Scikit-learn. (2009b). *3.4. Metrics and scoring: quantifying the quality of predictions — scikit-learn 1.7.0 documentation*. https://scikit-learn.org/stable/modules/model_evaluation.html

Scikit-learn. (2009c). *7.3. Preprocessing data — scikit-learn 1.7.0 documentation*. https://scikit-learn.org/stable/modules/preprocessing.html

Sergeeva, V. A., Zakharova, N. V., Bugrova, A. E., Starodubtseva, N. L., Indeykina, M. I., Kononikhin, A. S., Frankevich, V. E., & Nikolaev, E. N. (2020). The high-resolution mass spectrometry study of the protein composition of amyloid-like urine aggregates associated with preeclampsia. *European Journal of Mass Spectrometry*, *26*(2), 158–161. https://doi.org/10.1177/1469066719860076

Serrano, N. C., Guio-Mahecha, E., Quintero-Lesmes, D. C., Becerra-Bayona, S., Paez, M. C., Beltran, M., Herrera, V. M., Leon, L. J., Williams, D., & Casas, J. P. (2018). Lipid profile, plasma apolipoproteins, and pre-eclampsia risk in the GenPE case-control study. *Atherosclerosis*, *276*, 189–194. https://doi.org/10.1016/J.ATHEROSCLEROSIS.2018.05.051

Shah, N., Meng, Q., Zou, Z., & Zhang, X. (2024). Systematic analysis on the horse-shoe-like effect in PCA plots of scRNA-seq data. *Bioinformatics Advances*, *4*(1). https://doi.org/10.1093/BIOADV/VBAE109

Sibai, B., Dekker, G., & Kupferminc, M. (2005). Pre-eclampsia. *The Lancet*, *365*(9461), 785–799. https://doi.org/https://doi.org/10.1016/S0140-6736(05)17987-2

Smith, R. W. (2013). Mass Spectrometry. In J. A. Siegel, P. J. Saukko, & M. M. Houck (Eds.), *Encyclopedia of Forensic Sciences (Second Edition)* (Second Edition, pp. 603–608). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-382165-2.00250-6

Souza, T. (2025). Principal Component Analysis (PCA). *Advanced Statistical Analysis for Soil Scientists*, 43–56. https://doi.org/10.1007/978-3-031-88161-9_4

Starodubtseva, N., Poluektova, A., Tokareva, A., Kukaev, E., Avdeeva, A., Rimskaya, E., & Khodzayeva, Z. (2025). Proteome-Based Maternal Plasma and Serum Biomarkers for Preeclampsia: A Systematic Review and Meta-Analysis. *Life*, *15*(5), 776. https://doi.org/10.3390/LIFE15050776/S1

Szabo, S., Karaszi, K., Romero, R., Toth, E., Szilagyi, A., Gelencser, Z., Xu, Y., Balogh, A., Szalai, G., Hupuczi, P., Hargitai, B., Krenacs, T., Hunyadi-Gulyas, E., Darula, Z., Kekesi, K. A., Tarca, A. L., Erez, O., Juhasz, G., Kovalszky, I., … Than, N. G. (2020). Proteomic identification of Placental Protein 1 (PP1), PP8, and PP22 and characterization of their placental expression in healthy pregnancies and in preeclampsia. *Placenta*, *99*, 197–207. https://doi.org/10.1016/J.PLACENTA.2020.05.013

Tomkiewicz, J., & Darmochwał-Kolarz, D. A. (2024). Biomarkers for Early Prediction and Management of Preeclampsia: A Comprehensive Review. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, *30*, e944104-1. https://doi.org/10.12659/MSM.944104

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/MIND/LIX.236.433

Valderrama-Aguirre, A., Gallo, D., & Cifuentes B, R. (2011). ¿Cuáles son los avances de la genómica y la proteómica en el tamizaje y/o predicción de la preeclampsia? *Revista Colombiana de Obstetricia y GinecologÃa*, *62*, 64–70. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0034-74342011000100008&nrm=iso

Van Belkum, A., Welker, M., Pincus, D., Charrier, J. P., & Girard, V. (2017). Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry in Clinical Microbiology: What Are the Current Issues? *Annals of Laboratory Medicine*, *37*(6), 475. https://doi.org/10.3343/ALM.2017.37.6.475

Wasinger, V. C., Cordwell, S. J., Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., & Humphery-Smith, I. (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *ELECTROPHORESIS*, *16*(1), 1090–1094. https://doi.org/10.1002/elps.11501601185

Wen, Q., Liu, L. Y., Yang, T., Alev, C., Wu, S., Stevenson, D. K., Sheng, G., Butte, A. J., & Ling, X. B. (2013). Peptidomic Identification of Serum Peptides Diagnosing Preeclampsia. *PLOS ONE*, *8*(6), e65571. https://doi.org/10.1371/JOURNAL.PONE.0065571

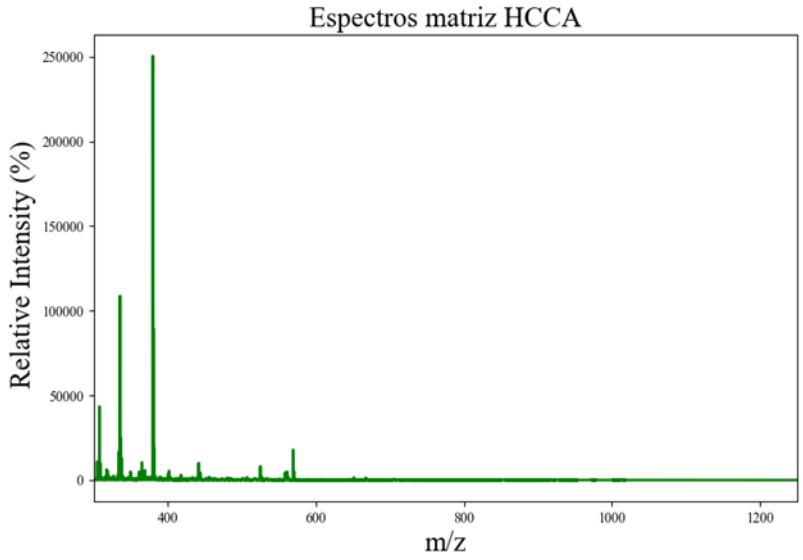
Wiśniewski, J. R. (2018). Filter-aided sample preparation for proteome analysis. *Methods in Molecular Biology*, *1841*, 3–10. https://doi.org/10.1007/978-1-4939-8695-8_1,

Yu, L. R., Stewart, N. A., & Veenstra, T. D. (2010). Proteomics. The Deciphering of the Functional Genome. In *Essentials of Genomic and Personalized Medicine* (pp. 89–96). Elsevier Inc. https://doi.org/10.1016/B978-0-12-374934-5.00008-8

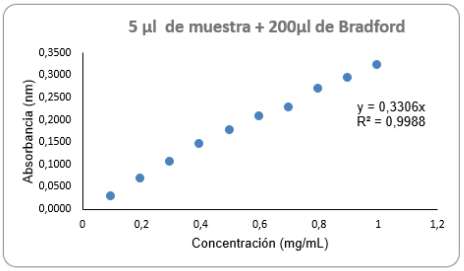
Zhao, Q., Li, J., Diao, Z., Zhang, X., Feng, S., Hou, G., Xu, W., Zhao, Z., Qiu, Z., Yang, W., Zhou, S., Tian, P., Zhang, Q., Chen, W., Li, H., Xiao, G., Qin, J., Hu, L., Li, Z., … Zhang, R. (2025). Early prediction of preeclampsia from clinical, multi-omics and laboratory data using random forest model. *BMC Pregnancy and Childbirth*, *25*(1), 1–15. https://doi.org/10.1186/S12884-025-07582-4/TABLES/2

# Apéndices

**Apéndice A.** Espectro matriz HCCA




**Apéndice B.** Determinación concentración de proteínas




| Muestras | Absorbancia | mg/mL | Concentración ug | d | Concentración mg | Concentración Promedio mg/mL |
|---|---|---|---|---|---|---|
| Muestra 1 | 0,4173 | 1,275 | 1275,457 | 100 | 127,546 | |
| Muestra 2 | 0,6403 | 1,966 | 1965,645 | 100 | 196,565 | |
| Muestra 3 | 0,5662 | 1,736 | 1736,305 | 100 | 173,630 | 165,9 |
| Muestra 4 | 0,3968 | 1,212 | 1212,009 | 100 | 121,201 | |
| Muestra 5 | 0,3634 | 1,109 | 1108,635 | 100 | 110,864 | |
| Muestra 6 | 0,4246 | 1,298 | 1298,050 | 100 | 129,805 | 120,3 |
| Muestra 7 | 0,3172 | 0,966 | 965,645 | 100 | 96,565 | |
| Muestra 8 | 0,487 | 1,491 | 1491,179 | 100 | 149,118 | |
| Muestra 9 | 0,496 | 1,519 | 1519,034 | 100 | 151,903 | 132,5 |
| Muestra 10 | 0,3555 | 1,084 | 1084,184 | 100 | 108,418 | |
| Muestra 11 | 0,358 | 1,092 | 1091,922 | 100 | 109,192 | |
| Muestra 12 | 0,3361 | 1,024 | 1024,141 | 100 | 102,414 | 106,7 |

**Apéndice C.** Análisis y procesamiento de datos

# Análisis del perfil proteómico de sueros sanguíneos de pacientes con preeclampsia

Vanessa Zambrano Martínez

Escuela de Química

Universidad Industrial de Santander

In [1]:
```
#Cargamos las librerias que vamos a requeririr para hacer el desarrollo
import pandas as pd
import numpy as np
import glob
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.decomposition import PCA
from sklearn import preprocessing
from sklearn.metrics import explained_variance_score, mean_absolute_error, mean_squ
from sklearn.decomposition import PCA
```

## Leer y pegar todos los espectros de masas, luego separar las intensidades y las relaciones m/z

In [2]:
```
import glob
fn = []
all_ms = pd.DataFrame()
#nombres=['MF']
for f in glob.glob("./all-ms/*.txt"):
    df = pd.read_csv(f, header=None, delimiter=' ' )
    all_ms = pd.concat([all_ms, df], axis=1)
    fn.append(f) # Guardamos parte del nombre del archivo como una etiqueta de cada
Datamz = all_ms[0] # relaciones m/z
Data = all_ms[1] # Intensidades
```

## Al inspeccionar nos damos cuenta que no todos los espectros han sido tomados en el mismo rango de relaciones m/z. Debemos eliminar los datos faltantes y dejar todos los espectros con el mismo número de datos

In [3]:
```
f
```

Out[3]:
```
'./all-ms\\T9R-D7_0_D7_1.txt'
```

In [4]:
```
Data.columns=range(Data.shape[1]) # Aquí le ponemos indices ordenados a las columns
#Datamz.columns=range(Datamz.shape[1])
#Datamz=Data[[0]]
Datamz.tail(3590) # Inspeccionamos los datos de la relación m/z al final para saber
```

Out[4]:

| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
|---|---|---|---|---|---|---|---|---|---|
| **252922** | 3325.897 | 3325.897 | 3325.897 | 3325.897 | 3325.897 | 3325.897 | 3325.897 | 3325.897 | 33 |
| **252923** | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 33 |
| **252924** | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 33 |
| **252925** | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 33 |
| **252926** | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 33 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | |
| **256507** | 3392.288 | 3392.288 | 3392.288 | 3392.288 | 3392.288 | 3392.288 | 3392.288 | 3392.288 | 33 |
| **256508** | 3392.307 | 3392.307 | 3392.307 | 3392.307 | 3392.307 | 3392.307 | 3392.307 | 3392.307 | 33 |
| **256509** | 3392.325 | 3392.325 | 3392.325 | 3392.325 | 3392.325 | 3392.325 | 3392.325 | 3392.325 | 33 |
| **256510** | 3392.344 | 3392.344 | 3392.344 | 3392.344 | 3392.344 | 3392.344 | 3392.344 | 3392.344 | 33 |
| **256511** | 3392.362 | 3392.362 | 3392.362 | 3392.362 | 3392.362 | 3392.362 | 3392.362 | 3392.362 | 33 |

3590 rows × 215 columns

## m/z

In [5]:
```
pd.set_option('display.max_columns', None)  # Mostrar todas las columnas
Datamz[252926:252929]
```

Out[5]:

| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
|---|---|---|---|---|---|---|---|---|---|
| **252926** | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 33 |
| **252927** | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 33 |
| **252928** | 3326.008 | 3326.008 | 3326.008 | 3326.008 | 3326.008 | 3326.008 | 3326.008 | 3326.008 | 33 |

## Se crea una nueva tabla de relaciones m/z que solo contenga las filas hasta donde todos los

## espectros tengan señal

In [6]:
```
Dsna=Datamz[0:252928]
Dsna
```

Out[6]:

| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 299.901 | 299.901 | 299.901 | 299.901 | 299.901 | 299.901 | 299.901 | 299.901 | 2 |
| **1** | 299.907 | 299.907 | 299.907 | 299.907 | 299.907 | 299.907 | 299.907 | 299.907 | 2 |
| **2** | 299.912 | 299.912 | 299.912 | 299.912 | 299.912 | 299.912 | 299.912 | 299.912 | 2 |
| **3** | 299.918 | 299.918 | 299.918 | 299.918 | 299.918 | 299.918 | 299.918 | 299.918 | 2 |
| **4** | 299.923 | 299.923 | 299.923 | 299.923 | 299.923 | 299.923 | 299.923 | 299.923 | 2 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | |
| **252923** | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 3325.915 | 33 |
| **252924** | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 3325.934 | 33 |
| **252925** | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 3325.952 | 33 |
| **252926** | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 3325.971 | 33 |
| **252927** | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 3325.989 | 33 |

252928 rows × 215 columns

## Verificamos que no hallan quedado casillas vacías

In [7]:
```
D=Dsna.isnull().any() # En la tabla D quedan la información booleana (False or True
```

In [8]:
```
df = D[D[0]==True]
df
```

Out[8]:
```
Series([], dtype: bool)
```

## Ya seguros de que hasta la fila 32255 no hay espectros sin datos de intensidad, volvemos a construir la tabla de intensidades hasta esa fila

In [9]:
```
Dataint=Data[0:252928]
Dataint
```

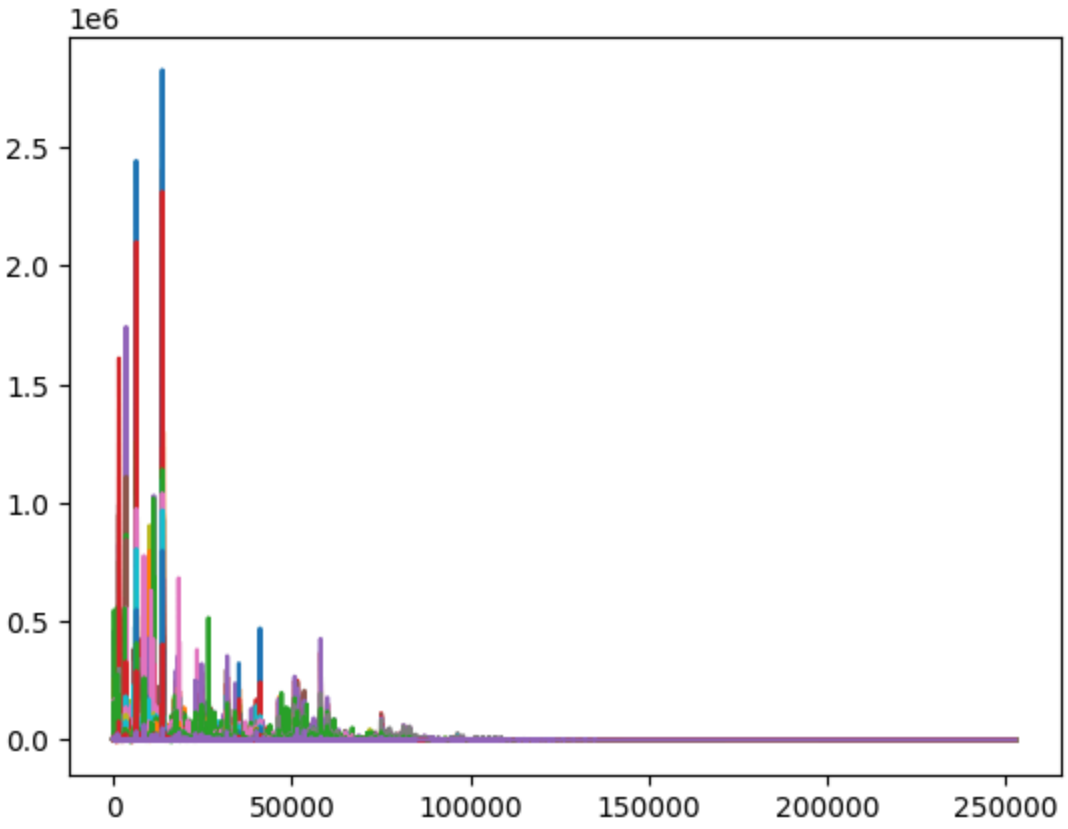
Out[9]:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 267 | 411 | 94 | 390 | 355 | 254 | 278 | 334 | 160 | 790 | 1496 | 21 | 299 | 403 | 8779 | 51 |
| **1** | 323 | 381 | 161 | 456 | 349 | 257 | 332 | 350 | 126 | 737 | 1515 | 15 | 286 | 411 | 8803 | 30 |
| **2** | 291 | 362 | 161 | 444 | 353 | 272 | 352 | 398 | 124 | 839 | 1603 | 14 | 291 | 451 | 9012 | 59 |
| **3** | 365 | 423 | 156 | 442 | 392 | 321 | 386 | 389 | 129 | 818 | 1527 | 11 | 300 | 408 | 9000 | 56 |
| **4** | 357 | 402 | 177 | 422 | 367 | 319 | 390 | 379 | 115 | 719 | 1567 | 15 | 258 | 444 | 9056 | 52 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **252923** | 4 | 6 | 5 | 0 | 2 | 12 | 11 | 2 | 0 | 4 | 0 | 2 | 2 | 0 | 0 | 5 |
| **252924** | 7 | 9 | 0 | 6 | 1 | 7 | 8 | 3 | 0 | 4 | 0 | 1 | 2 | 1 | 1 | 2 |
| **252925** | -1 | 3 | 0 | 7 | 6 | 9 | 15 | 5 | 0 | 5 | 1 | 0 | 0 | 1 | 0 | 7 |
| **252926** | -1 | 2 | 1 | 10 | 2 | 7 | 12 | 8 | 3 | 3 | 1 | 0 | 2 | 4 | 0 | 1 |
| **252927** | 0 | 1 | 1 | 10 | 8 | 2 | 12 | 6 | 1 | -1 | 1 | 0 | 2 | 1 | 1 | 2 |

252928 rows × 215 columns

## Visualización de todos los espectros, se encontro que algunos no se les realizó línea base

In [10]:
```
plt.plot(Dataint);
```




## Debemos encontrar esos espectros y eliminarlos de la tabla de datos, hacemos una pequeña rutina

In [11]:
```
len(Dataint.columns)
```

Out[11]: 215

## Eliminamos los espectros y inspeccionamos con la visualización, observamos que no aparecen esos espectros y quedamos contentos

In [12]:
```
ET = pd.read_excel('etiquetas6.xlsx', header=None)
Y=ET[[1]]
Y=Y.set_axis(['Clase'], axis=1)
Y
```

Out[12]:

| | Clase |
|---|---|
| **0** | P |
| **1** | P |
| **2** | P |
| **3** | P |
| **4** | P |
| **...** | ... |
| **210** | N |
| **211** | N |
| **212** | N |
| **213** | N |
| **214** | N |

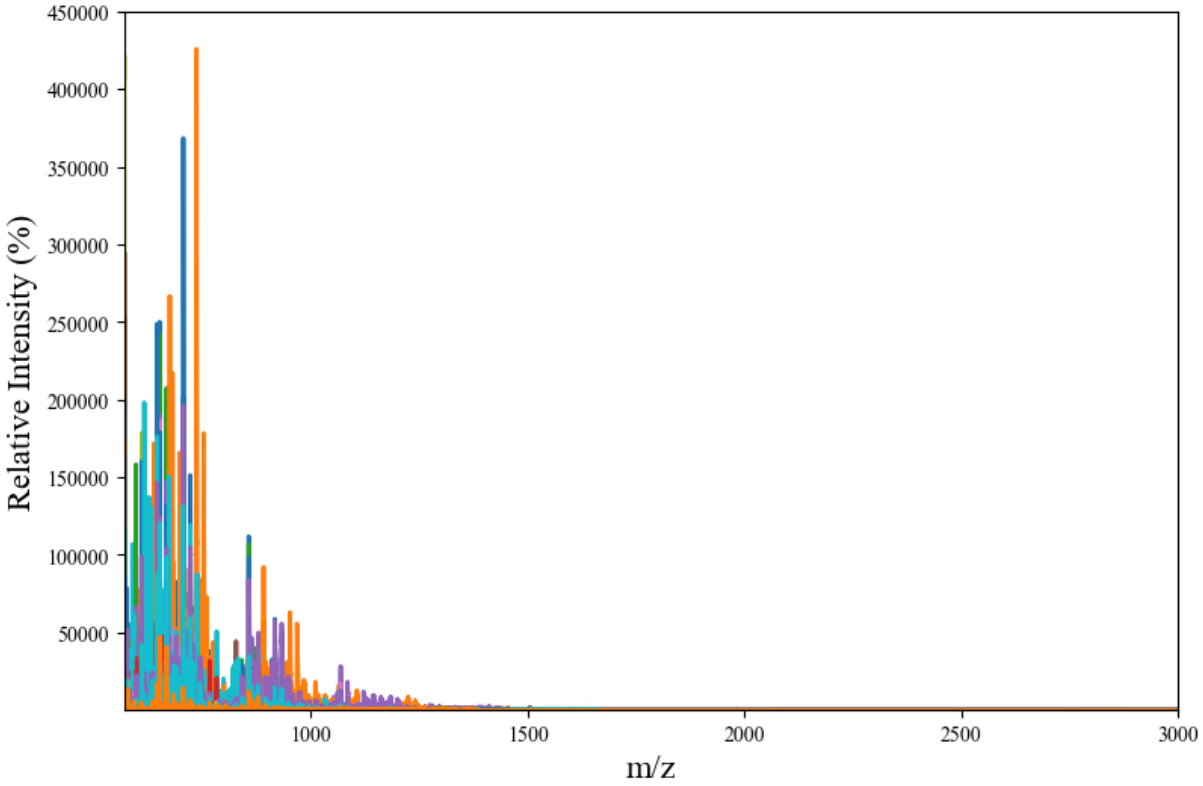
215 rows × 1 columns

In [13]:
```
#disminuir uso de memoria
import gc
gc.collect()
```

Out[13]: 17008

In [14]:
```
Dsna.columns=range(Dsna.shape[1]) # Aquí le ponemos indices ordenados a las columns
Dataintn=Dataint.drop([1, 2, 4], axis=1) # Eliminamos estas muestras
#Dsnan = Dsna.drop([520, 458], axis=1) # Elminamos de los datos de la relación m/z
Y1 = Y.drop([1, 2, 4], axis=0) # Eliminamos de las etiquetas los espectros defectuo
# Configuración de la visualización
plt.rcParams["font.family"] = 'Times New Roman'
fig, ax = plt.subplots(figsize=(9, 6))
# Graficamos las intensidades de los espectros no defectuosos
for i in range(Dataintn.shape[1]):
    ax.plot(Dsna.iloc[:, i], Dataintn.iloc[:, i], lw=2, label=f'Muestra {i+1}')
# Configuramos los límites de los ejes
ax.set_xlim([570, 3000])
ax.set_ylim(500, 450000)  # Descomenta si necesitas establecer límites en el eje y
# Etiquetas de los ejes
ax.set_xlabel('m/z', fontsize=16)
ax.set_ylabel('Relative Intensity (%)', fontsize=16)
# Guardamos la figura
fig.savefig("espectros1.jpg", dpi=300)
plt.show()
#plt.plot(Dataintn);
#from matplotlib.ticker import MultipleLocator, AutoMinorLocator
#plt.rcParams.update({'font.size': 18, 'font.family': 'STIXGeneral', 'mathtext.font
#plt.rcParams["font.family"] = 'Times New Roman'
#fig, ax = plt.subplots(figsize=(9,6))
#ax.plot(Dsna, Dataintn, lw=2)
```

```
#ax.set_xlim([570, 8000])
#ax.set_ylim(500, 5000)
#ax.set_xlabel('m/z',  fontsize=16)
#ax.set_ylabel('Relative Intensity (%)', fontsize=16)
#fig.savefig("todosesp.jpg", dpi=300)
#plt.show()
```

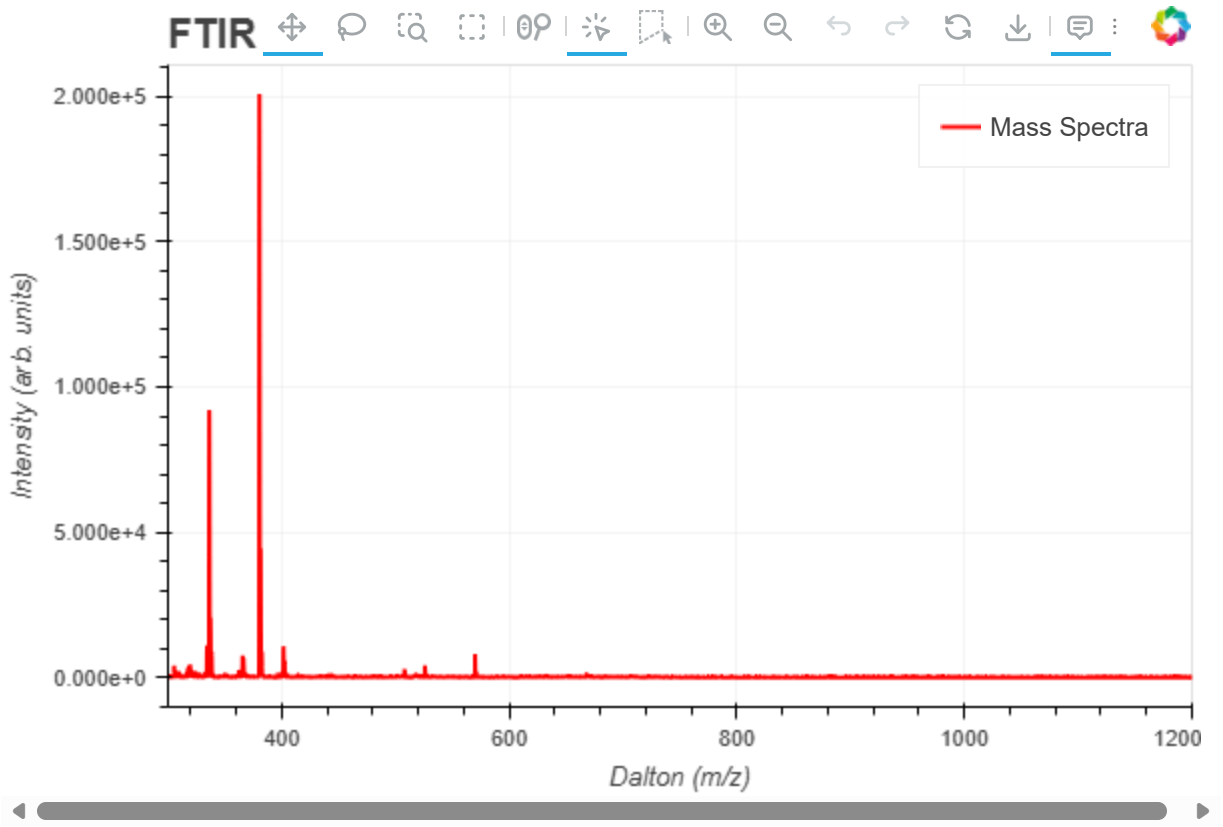



In [15]:

```
from bokeh.io import output_notebook, show
from bokeh.plotting import figure

# Configuración para mostrar gráficos en el cuaderno de Jupyter
output_notebook()

# Herramientas para la interacción
TOOLS = "hover,crosshair,pan,wheel_zoom,zoom_in,zoom_out,box_zoom,undo,redo,reset,t

# Crear un gráfico
p = figure(title='FTIR', width=600, height=400, x_range=(300, 1200),
           x_axis_label='Dalton (m/z)', y_axis_label='Intensity (arb. units)', tool
p.grid.grid_line_alpha = 0.3

# Asegúrate de que las columnas que estás graficando existan y tengan la misma long
if Dsna.shape[1] > 1 and Dataintn.shape[1] > 1:
    p.line(Dsna.iloc[:, 1], Dataintn.iloc[:, 1], color='red', legend_label='Mass Sp
else:
    print("Error: Asegúrate de que Dsna y Dataintn tengan al menos dos columnas.")

# Configuración adicional del gráfico
p.legend.location = "top_right"
p.grid.visible = True
p.title.align = "center"
```

```
p.title.text_font_size = "20px"

# Mostrar el gráfico
show(p)
```

Loading BokehJS ...




## PCA

In [16]:
```
# Paso 1: Obtener los índices válidos de Dataintn
indices_validos = Dataintn.index
# Paso 2: Filtrar Y1 para mantener solo las filas con índices que existen en Datain
Y1_filtrado = Y1[Y1.index.isin(indices_validos)].reset_index(drop=True)
```

In [17]:
```
Y1_filtrado
```

Out[17]:

| | Clase |
|---|---|
| **0** | P |
| **1** | P |
| **2** | N |
| **3** | N |
| **4** | N |
| **...** | ... |
| **207** | N |
| **208** | N |
| **209** | N |
| **210** | N |
| **211** | N |

212 rows × 1 columns

In [18]:
```
#normalizamos los datos
from sklearn.preprocessing import StandardScaler, Normalizer
scaler=StandardScaler()
#norma=Normalizer() NO ES NECESARIO

scaler.fit(Dataintn) # calculo la media para poder hacer la transformacion
x_scaled=scaler.transform(Dataintn)# Ahora si, escalo los datos y los normalizo
#x_norm = preprocessing.normalize(data_signals, norm='l1')# Normaliza los datos
#x_nsd = savgol_filter(x_norm, 17, polyorder=2, deriv=2) # Realiza segunda derivada
```

In [19]:
```
#NM=len(data.columns) # Número de muestras
#suma = np.empty((NM, 1)) #creamos un vector vacío para luego ir llenandolo con el
#k=0 #inicializamos el contador
#for k in np.arange(NM): #lo ponemos a correr hasta 40, que es el numero de espectr
#    suma[k]=sum(val[:, k]) #desarrollamos la suma a lo largo de cada columna, cada
#    k +=1 #variamos el contador en 1

valt=Dataintn.T
#valtn=valt/suma

valtn = preprocessing.normalize(valt, norm='l1')# Normaliza los datos
# Iniciamos el desarrollo del PCA en este caso con 9 componentes
pca=PCA(n_components=9) # Otra opción es hacer pca hasta obtener un mínimo explicad
pca1=pca.fit(x_scaled.T) # obtener los componentes principales
datos_pca=pca.transform(x_scaled.T) # convertimos nuestros datos con las nuevas dim

#out = pca.fit_transform(valtn) # otra opción

# Esta celda es para observar la varianza explicada con 5 componentes, se podría va
print("shape of datos_pca", datos_pca.shape)
expl = pca.explained_variance_ratio_
```

```
print(expl)
print('suma:',sum(expl[0:9]))
#Vemos que con 5 componentes tenemos algo mas del 85% de varianza explicada
datos_pca1 = pd.DataFrame(datos_pca) # Convierte los datos pca en un DataFrame
datos_pca1=pd.concat([datos_pca1, Y1_filtrado] ,axis=1) # Se agrega la columna del
```

```
shape of datos_pca (212, 9)
[0.29071117 0.20447723 0.10095436 0.07129545 0.05574986 0.03507871
 0.02984988 0.02462483 0.02184711]
suma: 0.8345885999287804
```

In [20]:
```
# Varianza explicada
expl = pca.explained_variance_ratio_

# Crear tabla
tabla_pca = pd.DataFrame({
    "PC": [f"PC{i+1}" for i in range(len(expl))],
    "Varianza Explicada": expl,
    "Varianza Acumulada": np.cumsum(expl)
})

# Redondear (como en tu ejemplo)
tabla_pca = tabla_pca.round(6)

tabla_pca
```

Out[20]:

| | PC | Varianza Explicada | Varianza Acumulada |
|---|---|---|---|
| **0** | PC1 | 0.290711 | 0.290711 |
| **1** | PC2 | 0.204477 | 0.495188 |
| **2** | PC3 | 0.100954 | 0.596143 |
| **3** | PC4 | 0.071295 | 0.667438 |
| **4** | PC5 | 0.055750 | 0.723188 |
| **5** | PC6 | 0.035079 | 0.758267 |
| **6** | PC7 | 0.029850 | 0.788117 |
| **7** | PC8 | 0.024625 | 0.812741 |
| **8** | PC9 | 0.021847 | 0.834589 |

In [21]:
```
datos_pca1
```

Out[21]:

| | 0 | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|---|
| **0** | -144.665623 | 230.516980 | -39.466869 | 39.501586 | 40.213775 | -2.549713 | 18.50412 |
| **1** | 19.540504 | 246.431348 | -37.452695 | -21.698920 | -23.187857 | -12.762324 | -20.06652 |
| **2** | -130.036441 | 235.987843 | -99.907257 | 78.217042 | 80.050157 | -36.388989 | -34.81279 |
| **3** | -181.710379 | 218.071344 | -123.337045 | 83.466008 | 93.602259 | -30.677908 | -39.63004 |
| **4** | 58.928823 | 249.238140 | -56.672367 | -26.207643 | -41.017102 | -46.368744 | -29.81536 |
| **...** | ... | ... | ... | ... | ... | ... | |
| **207** | -266.408671 | -11.863760 | 156.467518 | -15.424923 | -42.874117 | 47.918201 | -35.59692 |
| **208** | -65.532194 | -308.598974 | -211.877477 | -203.453569 | 114.785775 | -143.088794 | -7.03931 |
| **209** | -24.017537 | -124.109608 | -81.372175 | -252.574843 | 93.860369 | 176.983496 | -1.39588 |
| **210** | -287.787319 | -12.032986 | 120.105243 | -2.814530 | 0.066547 | 58.369634 | -25.35145 |
| **211** | -155.581677 | -233.664868 | -186.453973 | -206.903880 | 169.879423 | -51.964106 | 75.87805 |

212 rows × 10 columns

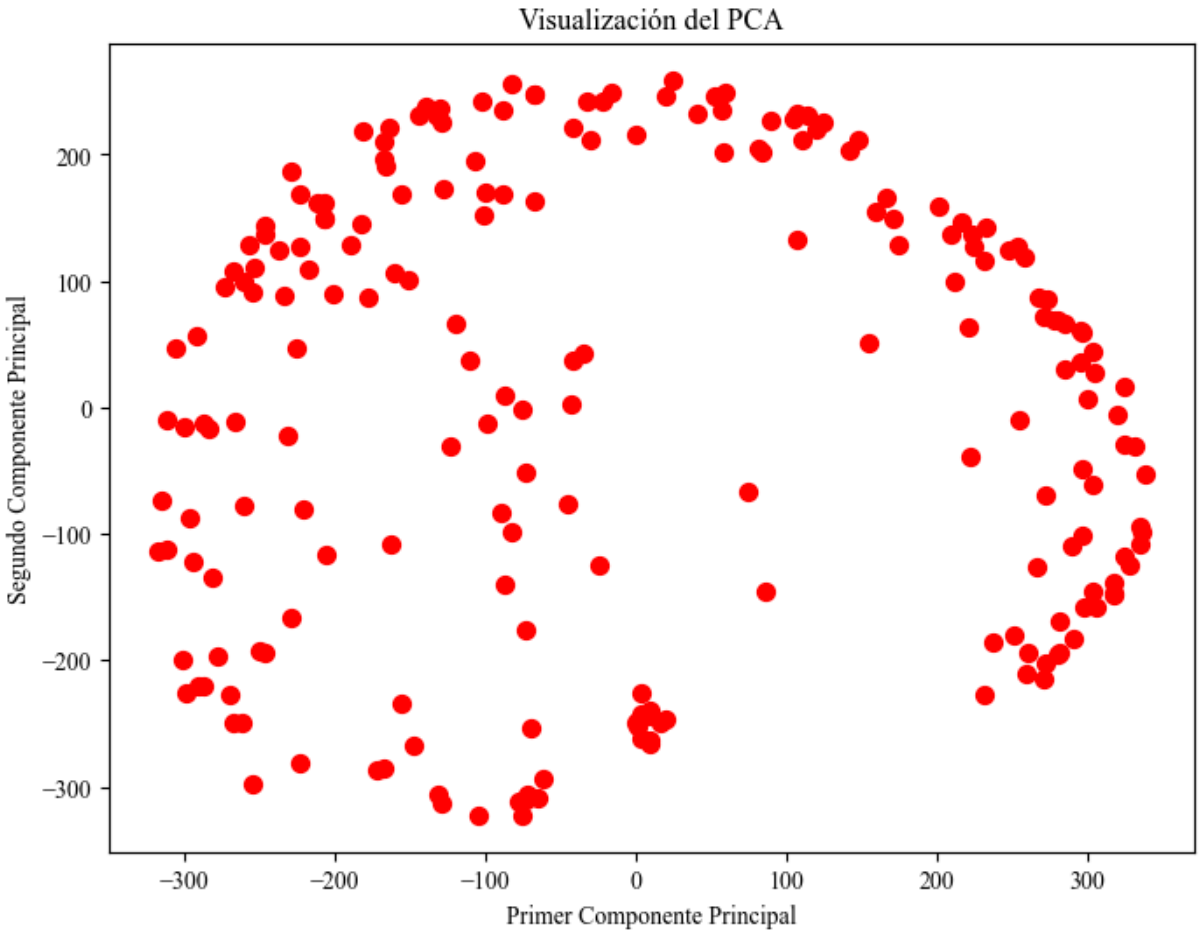
In [22]:
```
scores_df = pd.DataFrame(datos_pca, columns = [f"PC{i+1}" for i in range(datos_pca.
scores_df.head()
```

Out[22]:

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | |
|---|---|---|---|---|---|---|---|---|
| **0** | -144.665623 | 230.516980 | -39.466869 | 39.501586 | 40.213775 | -2.549713 | 18.504127 | -1 |
| **1** | 19.540504 | 246.431348 | -37.452695 | -21.698920 | -23.187857 | -12.762324 | -20.066521 | 0 |
| **2** | -130.036441 | 235.987843 | -99.907257 | 78.217042 | 80.050157 | -36.388989 | -34.812793 | -21 |
| **3** | -181.710379 | 218.071344 | -123.337045 | 83.466008 | 93.602259 | -30.677908 | -39.630042 | -20 |
| **4** | 58.928823 | 249.238140 | -56.672367 | -26.207643 | -41.017102 | -46.368744 | -29.815362 | 0 |

## Graficamos Resultados de PCA

In [23]:
```
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))
plt.scatter(datos_pca1.iloc[:, 0], datos_pca1.iloc[:, 1], c='red', s=50)
plt.xlabel('Primer Componente Principal')
plt.ylabel('Segundo Componente Principal')
plt.title('Visualización del PCA')
plt.savefig('PCA.png');
plt.show()
```




## VERIFICACIÓN DE DATOS NULOS PCA

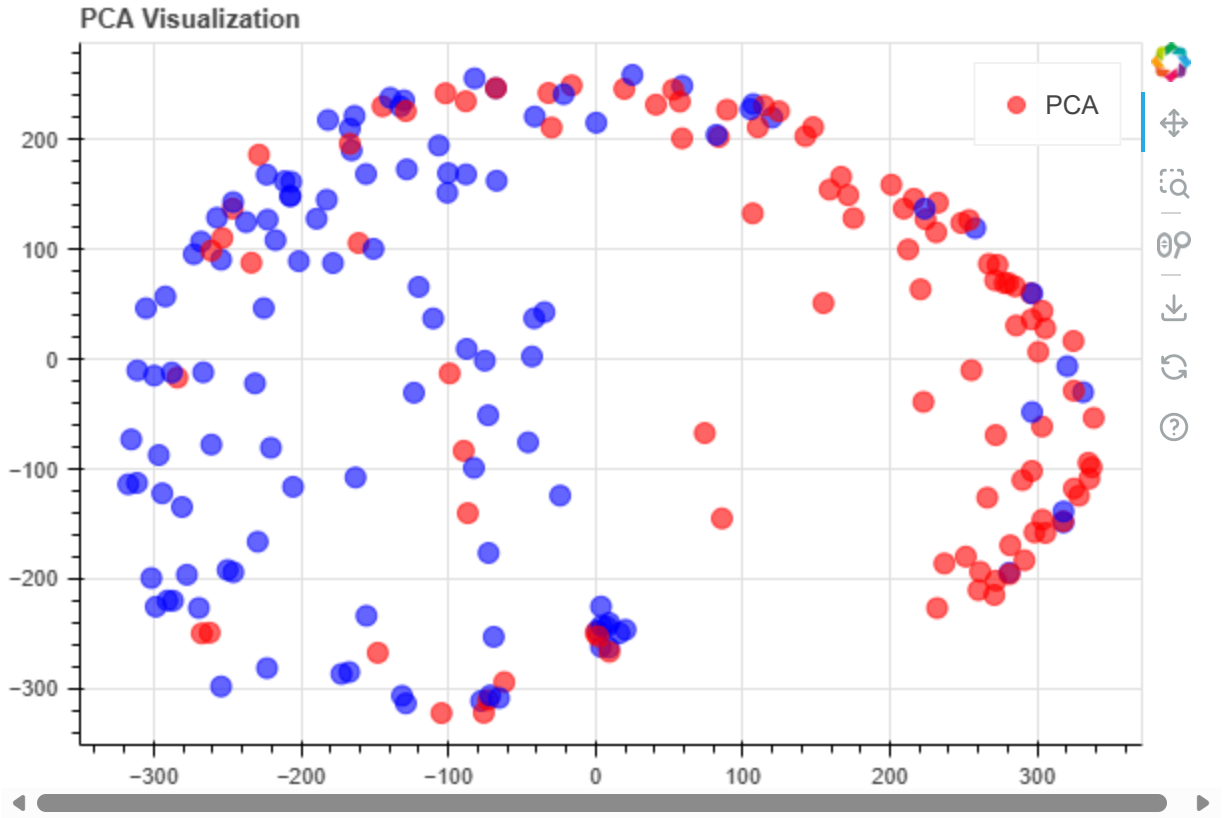
In [24]:
```
print(datos_pca1.isnull().sum())
```

```
0        0
1        0
2        0
3        0
4        0
5        0
6        0
7        0
8        0
Clase    0
dtype: int64
```

In [25]:
```
from bokeh.plotting import figure, show
import random
# Define los colores para cada etiqueta
colors = ['blue' if label == 'N' else 'red' for label in datos_pca1['Clase']]

# Crear figura de Bokeh
p = figure(title='PCA Visualization', width=600, height=400)

# Verifica longitudes
print(len(datos_pca1.iloc[:, 0]), len(datos_pca1.iloc[:, 1]), len(colors))
```

```
# Asegura que todos tengan la misma longitud
if len(colors) != len(datos_pca1):
    while len(colors) < len(datos_pca1):
        colors.append(random.choice(colors))
    colors = colors[:len(datos_pca1)]  # En caso de que sobre

# Gráfica con scatter
p.scatter(datos_pca1.iloc[:, 0], datos_pca1.iloc[:, 1], color=colors, legend_label=

# Mostrar gráfico
show(p)
```

```
212 212 212
```

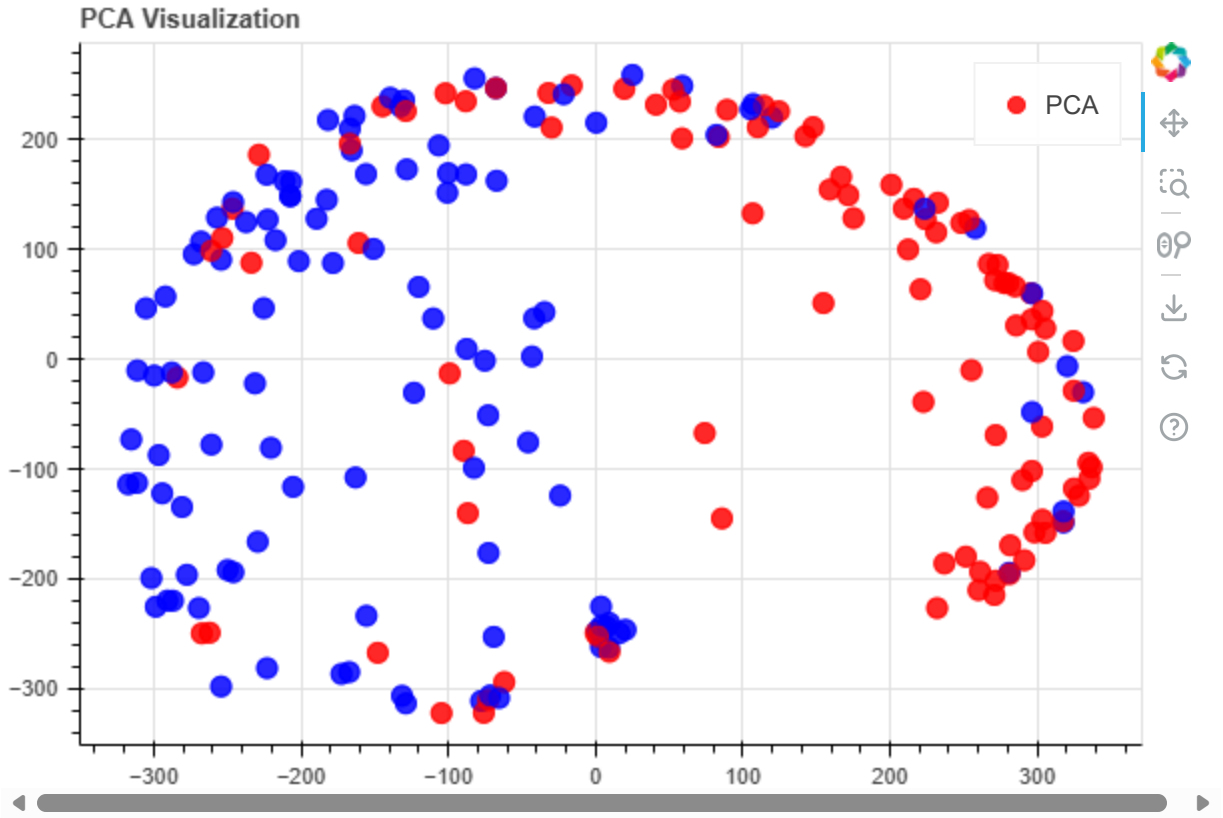



In [26]:

```
import random

# Verificar longitudes
print(len(datos_pca1.iloc[:, 0]), len(datos_pca1.iloc[:, 1]), len(colors))

# Asegura que todos tengan la misma longitud
if len(colors) != len(datos_pca1):
    # Si colors es más corto, lo extendemos
    while len(colors) < len(datos_pca1):
        colors.append(random.choice(colors))  # Elige un color aleatorio de la list

    # Recorta colors si es más largo
    colors = colors[:len(datos_pca1)]

# Gráfica con scatter
p.scatter(datos_pca1.iloc[:, 0], datos_pca1.iloc[:, 1], color=colors, legend_label=
```

```
# Muestra la figura
show(p)
```

```
212 212 212
```




In [27]:
```
print(datos_pca1.shape)  # filas, columnas
print(datos_pca1.columns)  # nombres actuales
```

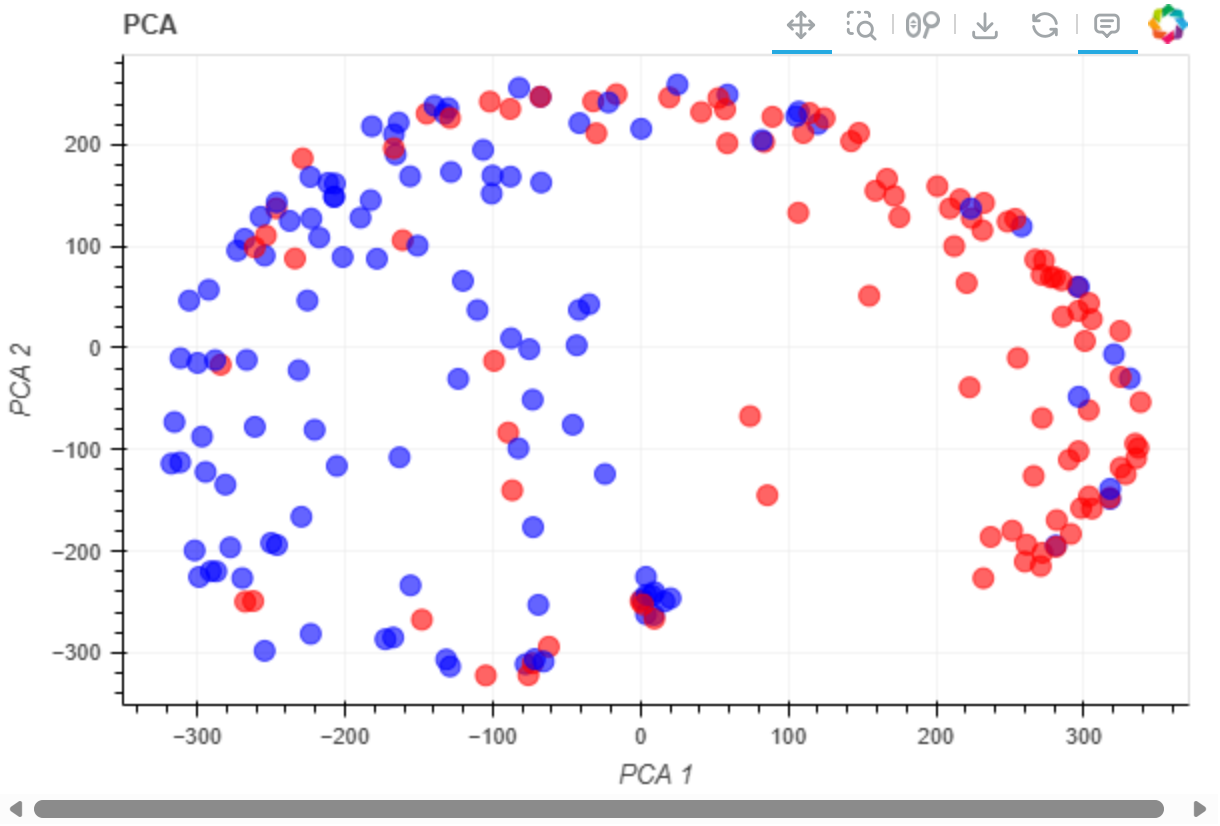
```
(212, 10)
Index([0, 1, 2, 3, 4, 5, 6, 7, 8, 'Clase'], dtype='object')
```

In [28]:
```
datos_pca_df = pd.DataFrame({
    'PC1': datos_pca1[0],
    'PC2': datos_pca1[1],
    'Clase': Y1_filtrado['Clase'].values  # me aseguro de que esté alineado
})
```

In [29]:
```
from bokeh.plotting import figure
from bokeh.models import ColumnDataSource, HoverTool
from bokeh.io import show

TOOLS = "hover,save,pan,box_zoom,reset,wheel_zoom"

colormap = {'N': 'blue', 'P':'red' }
colors = [colormap[x] for x in Y1_filtrado['Clase']]

#datos_pca1 = pd.DataFrame(data={'PCA1': X_pca[:,0], 'PCA2': X_pca[:,1]})

p = figure(title='PCA', width=600, height=400,
           x_axis_label='PCA 1', y_axis_label='PCA 2', toolbar_location="above", to
p.grid.grid_line_alpha=0.3
p.circle(datos_pca1[0], datos_pca1[1], color=colors, size=10, alpha=0.6)
```

```
show(p)
```

```
BokehDeprecationWarning: 'circle() method with size value' was deprecated in Bokeh
3.4.0 and will be removed, use 'scatter(size=...) instead' instead.
```

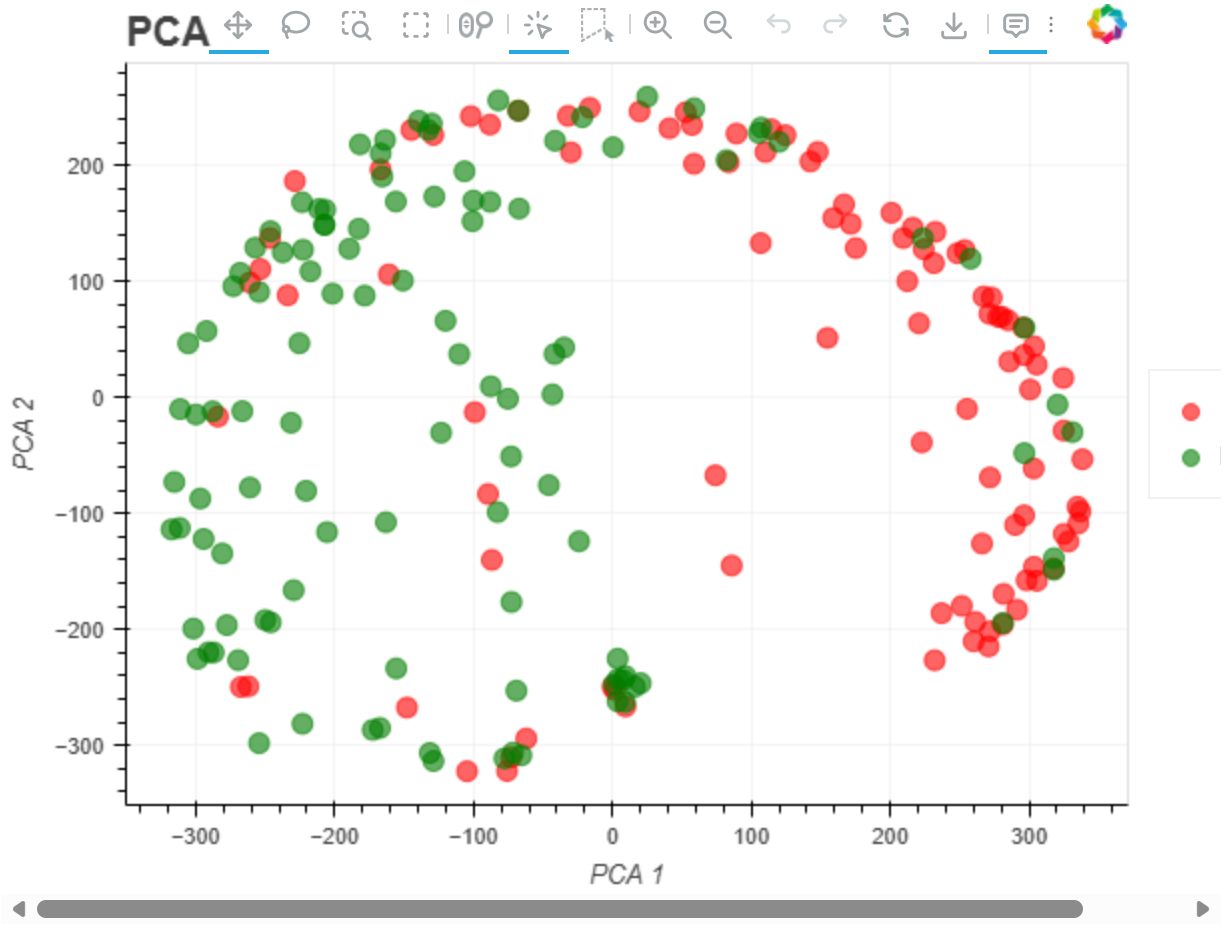



In [30]:
```
from bokeh.models import Legend
from bokeh.io import output_notebook, show
from bokeh.plotting import figure

output_notebook()

colormap = {'P': 'red', 'N': 'green'}
colors = [colormap[x] for x in Y1_filtrado['Clase']]

TOOLS = "hover,crosshair,pan,wheel_zoom,zoom_in,zoom_out,box_zoom,undo,redo,reset,t

p = figure(title='PCA', width=650, height=450,  # Aumenté el tamaño para acomodar l
           x_axis_label='PCA 1', y_axis_label='PCA 2',
           toolbar_location="above", tools=TOOLS)

# Crear renderizadores separados para cada clase
renderers = []
for label, color in colormap.items():
    class_mask = Y1_filtrado['Clase'] == label
    r = p.circle(datos_pca1[0][class_mask], datos_pca1[1][class_mask],
                 color=color, size=10, alpha=0.6)
    renderers.append((label, [r]))

# Crear leyenda fuera del gráfico
legend = Legend(items=renderers, location="center")
p.add_layout(legend, 'right')  # Posición a la derecha del gráfico
```

```
# Configuraciones adicionales
p.grid.grid_line_alpha = 0.3
p.title.align = "center"
p.title.text_font_size = "20px"
plt.savefig('GraficaPCA.png');
show(p)
```

BokehJS 3.7.3 successfully loaded.

```
BokehDeprecationWarning: 'circle() method with size value' was deprecated in Bokeh
3.4.0 and will be removed, use 'scatter(size=...) instead' instead.
BokehDeprecationWarning: 'circle() method with size value' was deprecated in Bokeh
3.4.0 and will be removed, use 'scatter(size=...) instead' instead.
```




```
<Figure size 640x480 with 0 Axes>
```

## Codificamos la variable dependiente

In [31]:
```
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import OneHotEncoder

encoder = OrdinalEncoder()
encoder.fit(Y1[['Clase']])
Y1_code = encoder.transform (Y1[['Clase']])
Y1_code = pd.DataFrame(Y1_code)
Y1_code
```

Out[31]:

| | 0 |
|---|---|
| **0** | 1.0 |
| **1** | 1.0 |
| **2** | 0.0 |
| **3** | 0.0 |
| **4** | 0.0 |
| **...** | ... |
| **207** | 0.0 |
| **208** | 0.0 |
| **209** | 0.0 |
| **210** | 0.0 |
| **211** | 0.0 |

212 rows × 1 columns

## Segundo Modelo: Vectores Suporte, SVM

In [32]:

```
import numpy as np
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# División de datos
X_tr2, X_te2, y_tr2, y_te2 = train_test_split(x_scaled.T, Y1_code.values.ravel(), t

# Creación y entrenamiento del modelo SVM
svc = SVC(C=40, kernel='rbf', random_state=123)
svc.fit(X_tr2, y_tr2)

# Predicciones en el conjunto de prueba
predicciones = svc.predict(X_te2)

# Cálculo de la precisión
accuracy = accuracy_score(y_true=y_te2, y_pred=predicciones, normalize=True)
print(f"\nEl accuracy de test es: {100 * accuracy:.2f}%")

# Función para hacer el segundo modelo
def opt_svc(X, y, x_test, rs):
    svc = SVC(C=40, kernel='rbf', random_state=123)
    svc.fit(X, y)
    y_pred = svc.predict(x_test)
    ex = accuracy_score(y_true=y_test, y_pred=y_pred, normalize=True)
    return y_pred, ex

# Ejecución de 99 modelos
```

```
exs = []
yps = []
rss = []
arr_rs = np.arange(1, 100)

for rs in arr_rs:
    x_train, x_test, y_train, y_test = train_test_split(x_scaled.T, Y1_code.values.
    y_pred, ex = opt_svc(x_train, y_train, x_test, rs)
    rss.append(rs)
    yps.append(y_pred)
    exs.append(ex)

# Máxima precisión obtenida
max_accuracy = np.max(exs)
print(f"La máxima precisión obtenida es: {100 * max_accuracy:.2f}%")
```

```
El accuracy de test es: 79.07%
La máxima precisión obtenida es: 88.37%
```

# OTRA MANERA DE REALIZAR EL MODELO

In [33]:
```
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import accuracy_score

#X_tr1, X_te1, y_tr1, y_te1 = train_test_split(datos_pca, Y1_code, train_size   = 0
X_tr2, X_te2, y_tr2, y_te2 = train_test_split(x_scaled.T, Y1_code.values.ravel(), t

# Creación del modelo SVM lineal
# ==============================================================================
#modelo1 = SVC(C = 60, kernel = 'linear', random_state=123)
#modelo1.fit(X_tr1, y_tr1)
svc = SVC(C = 40, kernel = 'rbf', random_state=123)
svc.fit(X_tr2, y_tr2)
```

Out[33]:

SVC

▶ Parameters

In [49]:
```
# Predicciones test
# ==============================================================================
predicciones = svc.predict(X_te2)
predicciones
```

Out[49]:
```
array([1., 1., 0., 1., 1., 1., 1., 0., 1., 0., 0., 1., 1., 0., 0., 1., 0.,
       1., 0., 0., 0., 0., 0., 0., 1., 0., 0., 1., 0., 1., 1., 1., 1., 0.,
       1., 0., 0., 1., 0., 0., 0., 0., 0.])
```

In [50]:
```
from sklearn.metrics import classification_report

print(classification_report(y_te2, predicciones))
```

```
              precision    recall  f1-score   support

         0.0       0.79      0.83      0.81        23
         1.0       0.79      0.75      0.77        20

    accuracy                           0.79        43
   macro avg       0.79      0.79      0.79        43
weighted avg       0.79      0.79      0.79        43
```

In [35]:
```python
# Accuracy de test del modelo
# ================================================================================
accuracy = accuracy_score(y_true = y_te2, y_pred = predicciones, normalize = True)
print("")
print(f"El accuracy de test es: {100*accuracy}%")
```

```
El accuracy de test es: 79.06976744186046%
```

## Definimos una función para hacer el segundo modelo y obtener todos los parametros

In [36]:
```python
def opt_svc(X, y, xt, rs):

    # Definimos PLS y el número de componentes
    svc = SVC(C = 40, kernel = 'rbf', random_state=123)
    svc.fit(X, y)
    y_pred = svc.predict(x_test)
    # Calculamos métricas
    ex = accuracy_score(y_true = y_test, y_pred = y_pred, normalize = True)

    return (y_pred, ex)
```

## Ejecutamos cien modelos para encontrar el mejor ramdon state

In [37]:
```python
# Probamos con 30 componentes
exs = []
yps = []
rss = []
arr_rs = np.arange(1, 100)

for rs in arr_rs:
    x_train, x_test, y_train, y_test = train_test_split(x_scaled.T, Y1_code.values.
    y_pred, ex = opt_svc(x_train, y_train, x_test, rs)
    rss.append(rs)
    yps.append(y_pred)
    exs.append(ex)
```

## Exactitud del mejor random state

In [38]:
```
np.max(exs)
```

Out[38]:
```
np.float64(0.8837209302325582)
```

In [48]:
```
from sklearn.model_selection import train_test_split
import numpy as np

# División de datos
x_train, x_test, y_train, y_test = train_test_split(
    x_scaled.T,
    Y1_code.values.ravel(),
    train_size=0.8,
    random_state=12,
    shuffle=True
)

# Conteo de clases en entrenamiento
print("Entrenamiento:")
print("Control (N):", sum(y_train == 0))
print("Preeclampsia (P):", sum(y_train == 1))

# Conteo de clases en prueba
print("\nPrueba:")
print("Control (N):", sum(y_test == 0))
print("Preeclampsia (P):", sum(y_test == 1))
```

```
Entrenamiento:
Control (N): 89
Preeclampsia (P): 80

Prueba:
Control (N): 23
Preeclampsia (P): 20
```

## MATRIZ DE CONFUSIÓN

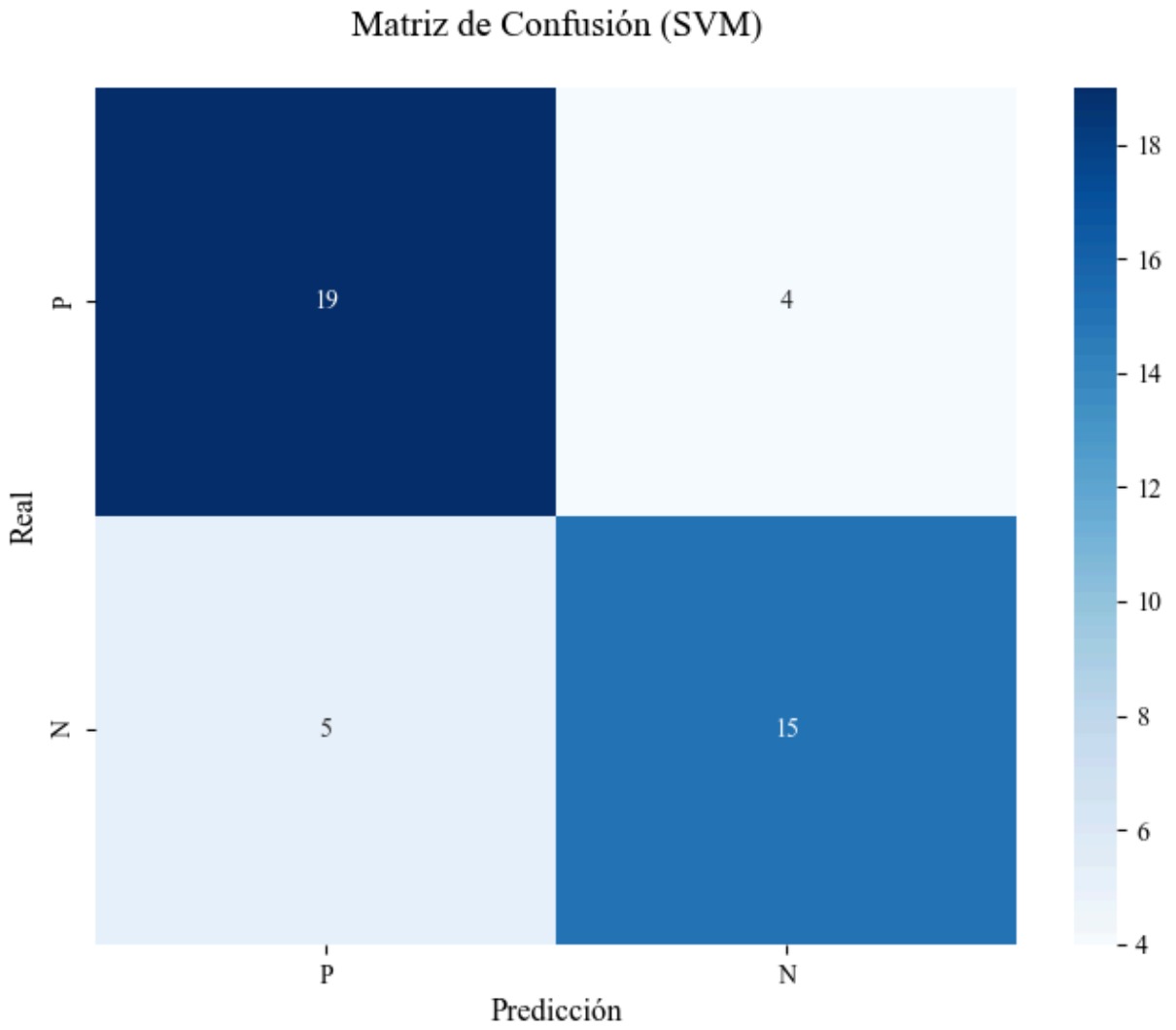
In [39]:
```
from sklearn.metrics import confusion_matrix
# Calcular la matriz de confusión
cm = confusion_matrix(y_te2, predicciones)
print("Matriz de Confusión:")
print(cm)
```

```
Matriz de Confusión:
[[19  4]
 [ 5 15]]
```

In [40]:
```
import seaborn as sns
# Visualizar la matriz de confusión
plt.figure(figsize=(8, 6))
sns.heatmap(cm,
            annot=True,          # Muestra los valores en cada celda
            fmt='d',             # Formato: números enteros
            cmap='Blues',        # Mapa de colores
            xticklabels=['P', 'N'],  # Etiquetas personalizadas en el eje X
            yticklabels=['P', 'N']   # Etiquetas personalizadas en el eje Y
```

```
            )
plt.ylabel('Real', fontsize=12)
plt.xlabel('Predicción', fontsize=12)
plt.title('Matriz de Confusión (SVM)', fontsize=14, pad=20)
plt.savefig('matriz_confusion_svm.png', bbox_inches='tight', dpi=300)
print("Matriz guardada como 'matriz_confusion_svm.png'")
plt.show()
```

```
Matriz guardada como 'matriz_confusion_svm.png'
```




## Tercer Modelo: Red Neuronal Artificial, ANN

In [41]:

```
from sklearn.neural_network import MLPClassifier
x_train, x_test, y_train, y_test = train_test_split(x_scaled.T, Y1_code.values.rave
modelo_3 = MLPClassifier(
                hidden_layer_sizes=(20, 20),
                learning_rate_init=0.01,
                solver = 'lbfgs',
                max_iter = 5000,
                random_state = 123
            )
modelo_3.fit(X=x_train, y=y_train)
```

Out[41]:

MLPClassifier

▶ Parameters

In [42]:
```
y_pred = modelo_3.predict(x_test)
score = modelo_3.score(x_test, y_test)
score
```

Out[42]: 0.9069767441860465

In [43]:
```
from sklearn.metrics import confusion_matrix, precision_score, recall_score, f1_sco

# Confusion matrix
confusion_matrix(y_test, y_pred)
```

Out[43]:
```
array([[20,  3],
       [ 1, 19]])
```

In [44]:
```
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

         0.0       0.95      0.87      0.91        23
         1.0       0.86      0.95      0.90        20

    accuracy                           0.91        43
   macro avg       0.91      0.91      0.91        43
weighted avg       0.91      0.91      0.91        43
```

## OTRA FORMA

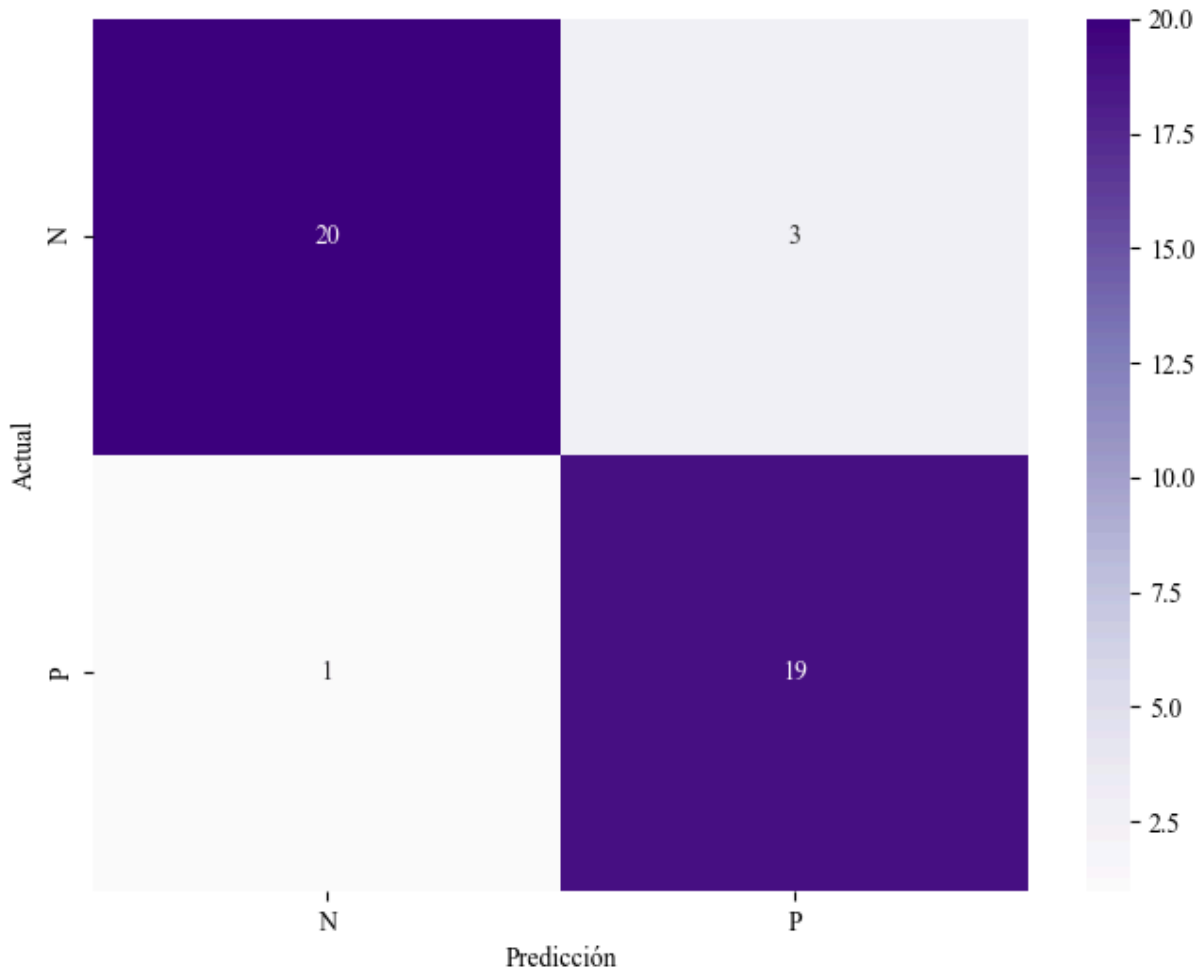
In [45]:
```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import confusion_matrix, classification_report

# Supongamos que x_scaled y Y1_code ya están definidos
x_train, x_test, y_train, y_test = train_test_split(
    x_scaled.T,
    Y1_code.values.ravel(),
    train_size=0.8,
    random_state=rss[np.argmax(exs)],
    shuffle=True
)

# Creación del modelo de Red Neuronal
modelo_3 = MLPClassifier(
    hidden_layer_sizes=(20, 20),
    learning_rate_init=0.01,
    solver='lbfgs',
    max_iter=5000,
```

```
    random_state=123
)

# Entrenamiento del modelo
modelo_3.fit(X=x_train, y=y_train)

# Predicciones
y_pred = modelo_3.predict(x_test)

# Precisión del modelo
score = modelo_3.score(x_test, y_test)
print(f"El accuracy del modelo es: {100 * score:.2f}%")

# Matriz de confusión
cm = confusion_matrix(y_test, y_pred)
print("Matriz de Confusión:")
print(cm)

# Informe de clasificación
print(classification_report(y_test, y_pred))
```

```
El accuracy del modelo es: 90.70%
Matriz de Confusión:
[[20  3]
 [ 1 19]]
              precision    recall  f1-score   support

         0.0       0.95      0.87      0.91        23
         1.0       0.86      0.95      0.90        20

    accuracy                           0.91        43
   macro avg       0.91      0.91      0.91        43
weighted avg       0.91      0.91      0.91        43
```

In [46]:
```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix

# Calcular la matriz de confusión
cm = confusion_matrix(y_test, y_pred)

# Visualizar la matriz de confusión
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Purples',
            xticklabels=['N', 'P'], yticklabels=['N', 'P'])
plt.ylabel('Actual')
plt.xlabel('Predicción')
plt.title('Matriz de Confusión (ANN)', fontsize=14, pad=20)
plt.savefig('matriz_confusion_ANN.png', bbox_inches='tight', dpi=300)
print("Matriz guardada como 'matriz_confusion_ANN.png'")
plt.show()
```

```
Matriz guardada como 'matriz_confusion_ANN.png'
```


Matriz de Confusión (ANN)
20
3
1
19
N
P
Actual
N
P
Predicción
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5