MINERÍA DE DATOS

PROYECTO:
“MODELO PREDICTIVO DE CLASIFICACIÓN DE RIESGO DE PREECLAMPSIA EN GESTANTES”

Integrantes:

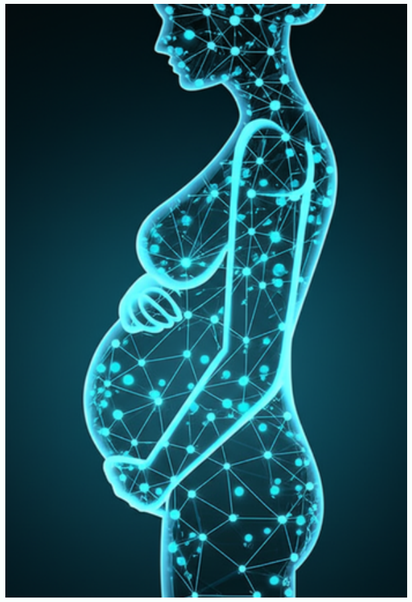
- Quispe Mamani Deyvis
- Pedraza Perez Joshua Josue
- Hidalgo Jauregui Karla Monica

Índice de
CONTENIDOS

I. PERFIL DEL PROYECTO

1. TITULO

Modelo Predictivo de Clasificación de Riesgo de Preeclampsia en Gestantes



2. PROBLEMA

La preeclampsia constituye una de las principales complicaciones durante el embarazo, afectando tanto la salud materna como neonatal. Tradicionalmente, su diagnóstico se ha basado en criterios clínicos y pruebas bioquímicas, pero recientes investigaciones destacan la necesidad de métodos más eficientes y predictivos.

3. PROPÓSITO

01 Desarrollar un modelo de clasificación que sirva como sistema de alerta temprana para gestantes en riesgo, reduciendo complicaciones en salud materna y neonatal.

02 Apoyar a profesionales de la salud en la toma de decisiones clínicas preventivas.

03 Reducir la incidencia de complicaciones maternas y perinatales mediante detección temprana y monitoreo proactivo .

4. JUSTIFICACIÓN

La detección temprana de la preeclampsia tiene un impacto directo en la salud materna e infantil, reduciendo riesgos como parto prematuro, daño multiorgánico o mortalidad materna.

El uso de técnicas de minería de datos e inteligencia artificial permite aprovechar información clínica (presión arterial, IMC, antecedentes, biomarcadores, estilo de vida, etc.) para generar modelos predictivos más precisos que los métodos tradicionales .

Este proyecto responde a la necesidad de soluciones basadas en datos en el sector salud, alineado con los objetivos de innovación en medicina preventiva y con los principios de desarrollo sostenible y bienestar social.

5. OBJETIVOS

Construir un modelo que prediga si una gestante estará en RIESGO DE PREECLAMPSIA (RIESGO = 1, NO RIESGO = 0) usando características clínicas, demográficas y de estilo de vida de la paciente, sin necesidad de esperar complicaciones avanzadas.

El modelo servirá como sistema de alerta temprana para apoyar decisiones médicas preventivas.

6. ALCANCE DEL PROYECTO

01 PRIMERA FASE

Comprensión del negocio

02 SEGUNDA FASE

Comprensión de los datos

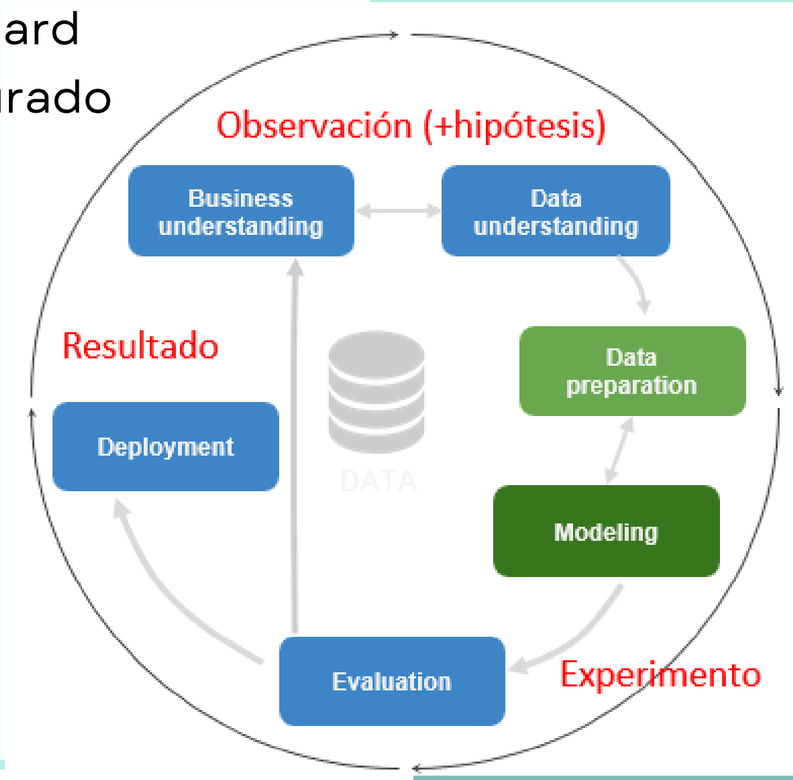
03 TERCERA FASE

Preparación de datos

7. METODOLOGÍA

Se sigue la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que establece un flujo estructurado para el desarrollo de proyectos de minería de datos.
Las fases son:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado (planificado)
- Evaluación (planificado)
- Despliegue (planificado)



8. ADMINISTRACIÓN:

01 CRONOGRAMA

4 semanas

·Semana 1: Perfil del proyecto

·Semana 2: Comprensión de datos

·Semana 3: Preprocesamiento y limpieza

·Semana 4: Documentación y dataset final

02 PRESUPUESTO

No aplica (uso de Python, Jupyter, librerías open source).

03 FINANCIAMIENTO

Recursos propios del equipo

II. APLICACIÓN DE LA METODOLOGÍA CRISP-DM

1. COMPRENSION DEL NEGOCIO

Construir un modelo de clasificación predictivo para identificar si una gestante estará en RIESGO DE PREECLAMPSIA (RIESGO = 1, NO RIESGO = 0) usando características clínicas, demográficas y de estilo de vida de la paciente, sin necesidad de esperar complicaciones avanzadas. El modelo servirá como sistema de alerta temprana para apoyar decisiones médicas preventivas.

Se han definido métricas de éxito tanto técnicas como de negocio, las cuales serán validadas con especialistas en salud para asegurar la relevancia del proyecto.

- Métrica Técnica: Precisión (Accuracy) ≥ 0.80, AUC ≥ 0.85, y Recall alto en la clase positiva (riesgo), minimizando falsos negativos.
- Métrica de Negocio:
  - Contribuir a la detección temprana y monitoreo de gestantes en riesgo de preeclampsia.
  - Generar evidencia cuantitativa que permita a instituciones de salud diseñar estrategias preventivas que reduzcan la morbilidad y mortalidad materna y neonatal.

2.COMPRENSIÓN Y PREPARACIÓN DE LOS DATOS

Los datos biomédicos y clínicos son fundamentales para entrenar modelos de predicción, y en estudios recientes.

Fuente de datos: Archivo Preeclampsia.csv, cuenta con 1,800 registros y 25 variables.

3. VARIABLES

01 VARIABLES NUMÉRICAS CONTINUAS

age → Numérica continua (edad en años).
gest_age → Numérica continua (edad gestacional en semanas).
height → Numérica continua (altura en cm).
weight → Numérica continua (peso en kg).
bmi → Numérica continua (índice de masa corporal).
sysbp → Numérica continua (presión arterial sistólica en mmHg).
diabp → Numérica continua (presión arterial diastólica en mmHg).
hb → Numérica continua (hemoglobina).
pcv → Numérica continua (packed cell volume, hematocrito).
tsh → Numérica continua (hormona tiroidea).
platelet → Numérica continua (conteo de plaquetas).
creatinine → Numérica continua (nivel de creatinina en sangre).
plgf:sflt → Numérica continua (relación PLGF/sFlt, biomarcador).
SEng → Numérica continua (soluble endoglina, biomarcador).
cysC → Numérica continua (cistatina C, biomarcador renal).
pp_13 → Numérica continua (proteína placentaria 13).
glycerides → Numérica continua (triglicéridos).

02 BINARIAS (DICOTÓMICAS, 0 = NO, 1 = SÍ)

htn → Binaria (antecedente de hipertensión).
diabetes → Binaria (antecedente de diabetes).
fam_htn → Binaria (antecedente familiar de hipertensión).
sp_art → Binaria (uso de técnicas de reproducción asistida).
diet → Binaria (dieta saludable/no saludable).
activity → Binaria (actividad física adecuada/inadecuada).
sleep → Binaria (calidad del sueño adecuada/inadecuada).

03 CATEGÓRICA NOMINAL

occupation → Categórica nominal (tipo de ocupación)

3.PREPARACION DE LOS DATOS

01 VERIFICACION DE DUPLICADOS

```
# A2) Nulos y duplicados
nulls_limpio = int(df_limpio.isnull().sum().sum())
dups_limpio  = int(df_limpio.duplicated().sum())
print(f"\nNulos (limpio): {nulls_limpio} | Duplicados (limpio): {dups_limpio}")
```

02 VERIFICACION DEVALORES FALTANTES

```
print("\n*** Valores faltantes por columna ***")
missing = df.isnull().sum()
missing_pct = (missing / len(df)) * 100
missing_report = pd.DataFrame({"Faltantes": missing, "Porcentaje": missing_pct})
print(missing_report)
```

03 CODIFICACION DE VARIABLES

Variables categóricas: *dieta*, *actividad_fisica*, *sueno*, *ocupacion*.
Se utilizó **One-Hot Encoding**, generando columnas binarias para cada categoría

Variable objetivo: Riesgo
Se aplicó codificación binaria:

- Riesgo = 1
- No riesgo = 0

El objetivo es un problema de clasificación binaria, por lo que esta representación es directa, eficiente y compatible con la mayoría de algoritmos de Machine Learning.


Open in Colab

CONCLUSIONES

- Se formuló un proyecto de minería de datos en salud con base en CRISP-DM.
- El dataset fue limpiado, transformado y documentado para su aplicación en modelos de clasificación.
- El proyecto queda listo para pasar a la fase de modelado con métricas enfocadas en la detección temprana de riesgo de preeclampsia.

REFERENCIAS

·[1] Feng, W., & Luo, Y. (2024). Preeclampsia and its prediction: Traditional versus contemporary predictive methods. The Journal of Maternal-Fetal & Neonatal Medicine, 37(1). https://doi.org/10.1080/14767058.2024.2388171

·[2] Kaur, M., Girija, R., Singh, M., Gupta, T., & Marwaha, S. (2025). Maternal risk prediction by early detection of pre-eclampsia and high-risk pregnancies using machine learning. In 2025 International Conference on Emerging Systems and Intelligent Computing (ESIC) (pp. 238–244). IEEE. https://doi.org/10.1109/ESIC64052.2025.10962644

·[3] Purwanti, E., Preswari, I. S., & Ernawati, E. (2019). Early risk detection of pre-eclampsia for pregnant women using artificial neural network. International Journal of Online and Biomedical Engineering (iJOE), 15(2), 71–80. https://doi.org/10.3991/ijoe.v15i02.9680

·[4] Wu, Y., Shen, L., Zhao, L., Zhang, X., & Chen, H. (2025). Noninvasive early prediction of preeclampsia in pregnancy using retinal vascular features. npj Digital Medicine, 8, 188. https://doi.org/10.1038/s41746-025-01582-6

·[5] Roque, A., Huamanzana, J., & Mauricio, D. (2024). Forecasting risk pregnancies in Peru using machine learning. In 2024 10th International Conference on Optimization and Applications (ICOA) (pp. 1–6). IEEE. https://doi.org/10.1109/ICOA62581.2024.10753982

·[6] Shabani, F., Jodeiri, A., Mohammad-Alizadeh-Charandabi, S., & Abedini, A. (2025). Developing and validating an artificial intelligence-based application for predicting some pregnancy outcomes: A multi-phase study protocol. Reproductive Health, 22, 99. https://doi.org/10.1186/s12978-025-02048-4

GRACIAS