UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERÍA

E.P. INGENIERÍA DE SISTEMAS

Proyecto:

"Modelo Predictivo de Clasificación de Riesgo de Preeclampsia en Gestantes"

Integrantes:
Quispe Mamani Deyvis
Pedraza Perez Joshua Josue
Hidalgo Jauregui Karla Monica

Curso:
Minería de Datos

Docente:
Sullon Macalupu Abel Angel

Lima – Perú
2025

Entregable: Perfil del Proyecto y Dataset Preprocesado

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 [1].

3. Propósito

- Desarrollar un modelo de clasificación que sirva como sistema de alerta temprana para gestantes en riesgo, reduciendo complicaciones en salud materna y neonatal.
- Apoyar a profesionales de la salud en la toma de decisiones clínicas preventivas.
- Reducir la incidencia de complicaciones maternas y perinatales mediante detección temprana y monitoreo proactivo [6].

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 [3][2].

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

Objetivo General

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.

Objetivos Específicos

- Recolectar y comprender los datos clínicos disponibles en el dataset “Preeclampsia.csv”.
- Realizar un diagnóstico de calidad de los datos (valores faltantes, duplicados, outliers, consistencia).
- Aplicar técnicas de preprocesamiento (limpieza, codificación, escalado, reducción si es necesario).

- Implementar un análisis exploratorio para identificar patrones y correlaciones relevantes.
- Definir métricas de éxito para el modelo predictivo (Accuracy ≥ 0.80, AUC ≥ 0.85, Recall positivo alto).
- Preparar un dataset final listo para la fase de modelado.

6. Alcance del Proyecto

- Fase 1. Comprensión del negocio
- Fase 2: Comprensión de los datos
- Fase 3. Preparacion de los datos

7. Metodología de Minería de Datos

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: Cronograma, presupuesto y financiamiento

a. **Cronograma:** 4 semanas
   - Semana 1: Perfil del proyecto
   - Semana 2: Comprensión de datos
   - Semana 3: Preprocesamiento y limpieza
   - Semana 4: Documentación y dataset final

b. **Presupuesto:** no aplica (uso de Python, Jupyter, librerías open source).

c. **Financiamiento:** recursos propios del equipo.

II. Aplicación de la Metodología CRISP-DM

Para esta unidad, solo se entregará las fases 1, 2 y 3.

1. Comprensión del negocio

En el contexto peruano, los embarazos de riesgo han sido abordados con técnicas de machine learning, evidenciando la aplicabilidad de estos modelos en hospitales locales [5]. Esto refuerza la pertinencia de investigar con datasets clínicos y variables sociodemográficas relevantes.

Objetivo de negocio: Detectar de manera temprana el riesgo de preeclampsia en mujeres embarazadas mediante el análisis de datos clínicos, con el fin de apoyar la toma de decisiones médicas.

- **Problema:** preeclampsia como riesgo crítico en gestantes.
- **Objetivo del modelo:** clasificar *Riesgo = 1* vs *No riesgo = 0*.
- **Criterios de éxito:**
  a. Métricas del modelo: Accuracy ≥ 0.80, AUC ≥ 0.85 (binario) o AUC-macro ≥ 0.80 (multiclase).
  b. Recall alto para la clase positiva (riesgo de preeclampsia).
  c. Dataset preprocesado y listo para modelar.

- **Restricciones y consideraciones:**
  a. Los datos corresponden a registros clínicos simulados o recolectados previamente (no intervención directa en pacientes).
  b. El uso es académico e investigativo, no reemplaza el diagnóstico médico.
  c. Validación clínica externa queda fuera del alcance de esta fase.

2. Comprensión y preparación de los datos

2.1 Comprensión de los Datos

Los datos biomédicos y clínicos son fundamentales para entrenar modelos de predicción, y en estudios recientes se han considerado incluso señales no invasivas como las características de la retina [4].

- **Fuente de datos:** Archivo Preeclampsia.csv, proporcionado en el curso, con 1,800 registros y 25 variables.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from pathlib import Path

# 1. Recolección de datos
# Cargar el archivo CSV en un DataFrame de pandas
df = pd.read_csv('/content/Preeclampsia.csv')
df.head()
```

- **Descripción general:**

  a. Variables numéricas: presión arterial, talla, peso, IMC, biomarcadores, triglicéridos, creatinina, etc.

  b. Variables categóricas/binarias: hipertensión (0/1), diabetes (0/1), dieta, actividad física, sueño, ocupación, antecedentes familiares, técnicas de reproducción asistida.

```
print('\n*** Info general ***')
print(df.info())
print('\n*** Primeras filas ***')
display(df.head())
```

```
*** Info general ***
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1800 entries, 0 to 1799
Data columns (total 25 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   age          1800 non-null   int64  
 1   gest_age     1800 non-null   int64  
 2   height       1800 non-null   int64  
 3   weight       1800 non-null   int64  
 4   bmi          1800 non-null   float64
 5   sysbp        1800 non-null   int64  
 6   diabp        1800 non-null   int64  
 7   hb           1800 non-null   float64
 8   pcv          1800 non-null   float64
 9   tsh          1800 non-null   float64
 10  platelet     1800 non-null   int64  
 11  creatinine   1800 non-null   float64
 12  plgf:sflt    1800 non-null   float64
 13  SEng         1800 non-null   float64
 14  cysC         1800 non-null   float64
 15  pp_13        1800 non-null   float64
 16  glycerides   1800 non-null   float64
 17  htn          1800 non-null   int64  
 18  diabetes     1800 non-null   int64  
 19  fam_htn      1800 non-null   int64  
 20  sp_art       1800 non-null   int64  
 21  occupation   1800 non-null   int64  
 22  diet         1800 non-null   int64  
 23  activity     1800 non-null   int64  
 24  sleep        1800 non-null   int64  
dtypes: float64(10), int64(15)
memory usage: 351.7 KB
None
```

```
*** Estructura del dataset ***
Filas: 1800
Columnas: 25
Total de registros: 1800

*** Tipos de variables ***
int64      15
float64    10
Name: count, dtype: int64

Columnas de tipo fecha: No se encontraron columnas de fecha
Columnas de tipo texto: No se encontraron columnas de texto
```

- **Diagnóstico de calidad:**

  a. No se detectaron valores nulos.
  b. No se encontraron registros duplicados.
  c. Se observaron inconsistencias en las variables de presión arterial: los campos sistólica y diastólica parecen estar cruzados (sistólica con valores de diastólica y viceversa). Este hallazgo debe ser corregido en la etapa de preprocesamiento.
  d. Escalas y rangos: biomarcadores y medidas clínicas presentan variabilidad normal, sin valores atípicos extremos no clínicamente plausibles.

2.2 Preparación de los Datos

- **Proceso de limpieza aplicado**

  1. Eliminación de duplicados

- Se verificó que no existían registros duplicados.

2. Manejo de valores faltantes
   - El dataset no contiene valores nulos, por lo que no fue necesario aplicar imputación.

- **Transformaciones realizadas**

1. Codificación de variables categóricas/binarias
   - Variables binarias (hipertension, diabetes, ant_fam_hiper, tec_repro_asistida, dieta, actividad_fisica, sueno) se mantuvieron en formato numérico (0/1).
   - Variable ocupacion fue tratada con One Hot Encoding para su uso en modelos de clasificación.

2. Escalado de variables numéricas
   - Variables continuas (edad, talla, peso, imc, presión arterial, biomarcadores) fueron estandarizadas con StandardScaler para garantizar homogeneidad de rangos en algoritmos sensibles a escala.

3. Generación de variable objetivo
   - Se plantearon dos enfoques:
   - Enfoque A (variable clínica directa): usar la variable binaria hipertension como proxy inicial de riesgo de preeclampsia.
   - Enfoque B (regla compuesta): creación de una variable Riesgo según criterios clínicos (≥140 mmHg sistólica o ≥90 mmHg diastólica, IMC ≥30, diabetes), con categorías Bajo, Medio, Alto.
   - En este entregable se documentan ambos enfoques, dejando al modelado la selección más adecuada.

4. Selección de variables finales
   - Se excluyeron atributos redundantes o no informativos.
   - Dataset final:
   - Variables predictoras (24 columnas, numéricas y categóricas transformadas).
   - Variable objetivo (Riesgo).

3. Planificación para modelado

Dataset listo para aplicar algoritmos supervisados, se evalúa lo siguiente:

- Árboles de decisión (modelo base, interpretable).
- Random Forest o XGBoost (más potentes)
- Regresión logística (modelo base, interpretable)

El modelo debe priorizar métricas robustas como precisión y recall, especialmente en la clase positiva de riesgo, con el fin de reducir falsos negativos que podrían comprometer la salud materna [2][1].

III. Dataset final entregado

- Conjunto de datos: Limpio, transformado y documentado en formato .csv y .ipynb.
- Descripción de variables y tratamiento aplicado:
    - Eliminación de duplicados y corrección de inconsistencias.
    - Imputación de valores faltantes.
    - Codificación de variables categóricas y binarias.
    - Escalado de variables numéricas.
    - Conservación de variables relevantes para modelado supervisado.

Investigaciones previas han demostrado que el preprocesamiento es clave para garantizar el rendimiento de modelos predictivos en contextos médicos [6].

IV. 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.

V. 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