

---

## 1. INDICACIONES

---

- En esta actividad se evalúa si el estudiante *Criterio 3.2: Aplica modelos de aprendizaje supervisado en escenarios del mundo real, ajustando los modelos para maximizar su precisión y eficiencia mediante técnicas de ajuste de hiperparámetros y regularización.*

---

## 2. DESCRIPCIÓN

---

A partir del trabajo realizado en el reto anterior, el equipo de investigación de la PUCE ha solicitado continuar con el análisis del conjunto de datos de viviendas para abordar un nuevo objetivo: desarrollar modelos predictivos que permitan anticipar el *estatus entomológico* de una vivienda, es decir, si presenta o no indicios de la presencia del vector del mal de Chagas.

Este nuevo enfoque requiere aplicar modelos de **aprendizaje supervisado**, en donde se utilizará la variable 'Status entomológico' como etiqueta, a partir de las demás variables del entorno, materiales de construcción, servicios básicos, animales presentes, prácticas higiénicas y condiciones ambientales.

### Pregunta esencial

- ¿Qué características de una vivienda permiten anticipar la presencia del vector del Chagas, y cómo pueden los modelos supervisados apoyar acciones preventivas y decisiones comunitarias?

### Reto

Tu misión es desarrollar un modelo supervisado que prediga el estatus entomológico de una vivienda, contribuyendo a priorizar acciones de intervención, fumigación o concienciación en comunidades vulnerables. Deberás:

- Preprocesar el conjunto de datos del proyecto Chagas.
- Entrenar y evaluar al menos tres modelos supervisados distintos.
- Comunicar los hallazgos de forma técnica y también comprensible para equipos de salud pública.

Tu objetivo final será entregar:

- Un `Jupyter Notebook` documentado, que contenga el desarrollo completo del modelo y visualizaciones.
- Un informe técnico de divulgación para equipos de intervención.

### Preguntas guía

- ¿Qué variables son más relevantes para la predicción del estatus entomológico?
- ¿Qué tipo de codificación es adecuada para las variables categóricas?
- ¿Qué algoritmos ofrecen mejor desempeño y explicabilidad en este contexto?
- ¿Cómo se puede interpretar el modelo para guiar decisiones prácticas?
- ¿Qué recomendaciones concretas se derivan del modelo generado?

### Actividades guía

1. Revisar y seleccionar las variables predictoras que puedan influir en la presencia del vector.
2. Preprocesar el dataset:
  - Limpieza y codificación de variables.
  - Análisis de balanceo de clases.
  - Escalado y selección de características.
3. Dividir los datos en conjunto de entrenamiento y prueba.
4. Entrenar al menos tres modelos (p. ej. regresión logística, árboles de decisión, Random Forest, redes neuronales).
5. Evaluar los modelos con métricas de clasificación (accuracy, precision, recall, F1-score, AUC).
6. Optimizar el modelo con mejor desempeño utilizando técnicas como GridSearchCV.
7. Interpretar los resultados, destacando las variables más influyentes.
8. Generar visualizaciones como:
  - Matriz de confusión.
  - Importancia de variables.
  - Curvas ROC.
9. Redactar un informe claro, accesible y útil para equipos comunitarios.

### Recursos

- Dataset del proyecto PUCE sobre el mal de Chagas.
- Bibliotecas: Scikit-learn, Pandas, Seaborn, Matplotlib, Plotly.
- Cuadernos y ejemplos proporcionados en clase sobre aprendizaje supervisado.

---

## PRODUCTO

---

### 2.1 Jupyter Notebook

El *Jupyter Notebook* debe ser técnico, bien documentado y fácilmente replicable, permitiendo que otros equipos puedan usar el mismo enfoque para analizar datos similares en nuevas comunidades. Debe estar generado en el [formato base](#) dado en clases.

- 1. Título y descripción inicial:** Título claro del proyecto, autores y una breve introducción al problema del Chagas y los objetivos del modelo predictivo.
- 2. Carga y descripción de datos:** Lectura del dataset, identificación de la variable objetivo ('Status entomológico') y descripción inicial de variables relevantes (estructura de la vivienda, tipo de agua, animales presentes, etc.).
- 3. Preprocesamiento de datos:** Limpieza de datos, codificación de variables categóricas, imputación de valores faltantes, escalado de variables numéricas y análisis de balanceo de clases. Justificación de cada decisión.
- 4. Entrenamiento de modelos supervisados:** Implementación de al menos tres modelos distintos (como regresión logística, Random Forest, y redes neuronales). Visualización de curvas de aprendizaje o métricas clave.
- 5. Evaluación de modelos:** Evaluación con métricas como accuracy, precisión, recall, F1-score y AUC. Comparación en una tabla resumen y representación gráfica (matriz de confusión, curvas ROC).
- 6. Optimización de modelos:** Ajuste de hiperparámetros del mejor modelo con técnicas como Grid Search o Random Search. Justificación de los parámetros elegidos.
- 7. Interpretación de resultados:** Análisis de las variables más influyentes y discusión sobre cómo estas se relacionan con el riesgo de presencia del vector.
- 8. Conclusión técnica:** Resumen de hallazgos, limitaciones del análisis y sugerencias para aplicaciones futuras del modelo.

### 2.2 Informe Divulgativo

El informe divulgativo debe estar dirigido a un público no técnico, como responsables comunitarios, autoridades sanitarias y organizaciones sociales. Debe estar hecho en  $\text{\LaTeX}$ , siguiendo el [formato de tareas PUCE](#), con lenguaje claro y visualizaciones interpretables.

- 1. Introducción:** Contextualización del problema del Chagas y de la necesidad de modelos predictivos. Objetivo del análisis y relevancia para la salud pública.
- 2. Metodología:** Breve explicación de los modelos supervisados utilizados, sin entrar en detalles técnicos, y por qué se eligieron para este problema.

- 3. Resultados principales:** Descripción del modelo seleccionado, sus predicciones más relevantes, y cómo estas permiten identificar viviendas en riesgo. Gráficos comprensibles (ej. semáforos de riesgo, mapas conceptuales, barras de importancia de variables).
- 4. Implicaciones prácticas:** Cómo los resultados pueden apoyar decisiones de intervención (como priorización de fumigaciones, campañas informativas o mejoras estructurales).
- 5. Recomendaciones:** Acciones sugeridas con base en los datos: reforzar ciertos servicios, enfocarse en tipos específicos de vivienda, entre otros.
- 6. Conclusión:** Relevancia del uso de datos e inteligencia artificial para combatir el mal de Chagas, y la importancia de continuar recolectando y analizando información con enfoque preventivo.

---

### 3. RÚBRICA DE EVALUACIÓN

---

#### Jupyter Notebook (30 puntos totales)

##### 1. Título y descripción inicial (2 puntos)

- Contextualiza el problema del Chagas y plantea claramente los objetivos del modelo predictivo (1 punto).
- Describe brevemente las técnicas de aprendizaje supervisado aplicadas (1 punto).

##### 2. Carga y descripción de datos (4 puntos)

- Presenta las variables relevantes con un análisis exploratorio básico (3 puntos).
- Incluye representaciones gráficas adecuadas de distribuciones y relaciones (1 punto).

##### 3. Preprocesamiento de datos (5 puntos)

- Limpieza e imputación adecuadas (2 puntos).
- Codificación y normalización bien implementadas (2 puntos).
- Justificación de las decisiones de preprocesamiento (1 punto).

##### 4. Entrenamiento de modelos supervisados (6 puntos)

- Entrenamiento correcto de al menos tres modelos supervisados distintos (3 puntos).
- Visualización y explicación de los resultados por modelo (3 puntos).

##### 5. Evaluación de modelos (5 puntos)

- Uso adecuado de métricas de clasificación (accuracy, recall, F1, AUC) (3 puntos).
- Inclusión de gráficas como matrices de confusión o curvas ROC (2 puntos).

**6. Optimización de modelos (3 puntos)**

- Aplicación de técnicas de ajuste de hiperparámetros (2 puntos).
- Justificación y análisis de los cambios en rendimiento (1 punto).

**7. Interpretación de resultados (2 puntos)**

- Identifica y analiza variables clave para la predicción del estatus entomológico (1.5 puntos).
- Relación de los hallazgos con el problema sanitario (0.5 puntos).

**8. Conclusión técnica (3 puntos)**

- Resumen técnico claro de hallazgos y desempeño de modelos (2 puntos).
- Limitaciones identificadas y posibles mejoras futuras (1 punto).

**Informe Divulgativo (20 puntos totales)****1. Introducción (3 puntos)**

- Presenta el contexto del mal de Chagas y el propósito del análisis (2 puntos).
- Define claramente los objetivos del informe (1 punto).

**2. Metodología (3 puntos)**

- Describe de forma sencilla las técnicas utilizadas (2 puntos).
- Explica cómo estas técnicas permiten abordar el problema (1 punto).

**3. Resultados principales (6 puntos)**

- Presenta los hallazgos de forma comprensible y contextualizada (3 puntos).
- Utiliza visualizaciones claras e interpretables por público no técnico (3 puntos).

**4. Implicaciones prácticas (4 puntos)**

- Relaciona resultados con acciones concretas de intervención (2.5 puntos).
- Reflexiona sobre el impacto en salud pública y priorización de zonas (1.5 puntos).

**5. Recomendaciones (2 puntos)**

- Plantea acciones claras y realistas con base en el modelo (2 puntos).

**6. Conclusión (2 puntos)**

- Resume la importancia del uso de modelos para combatir el Chagas (1.5 puntos).
- Muestra conciencia sobre los límites y el valor del análisis (0.5 puntos).