1. Indicaciones

- En esta actividad se evalúa si el estudiante *(Criterio 3.1: Aplica modelos de aprendizaje no supervisado en casos prácticos complejos, analizando los resultados y proponiendo mejoras basadas en métricas de rendimiento.*

2. Descripción

Desde hace más de 15 años, la PUCE ha liderado estudios sobre el mal de Chagas en Ecuador, recolectando datos sobre viviendas, condiciones sanitarias, materiales de construcción, presencia de animales, prácticas domésticas y características del entorno natural. Este valioso conjunto de datos permite analizar los factores que podrían incidir en la presencia del vector y el riesgo de transmisión.

Tu rol como analista de datos será identificar patrones ocultos que permitan generar *perfiles de viviendas con características de riesgo similares*, usando técnicas de clustering y reducción de dimensiones.

Pregunta esencial

- ¿Qué perfiles de viviendas presentan condiciones que favorecen la presencia del vector del Chagas y cómo podemos visualizarlos de manera clara para apoyar intervenciones de salud pública?

Reto

Formas parte del equipo de ciencia de datos del proyecto PUCE-Chagas. Tu reto es descubrir agrupaciones de viviendas que compartan características estructurales, sanitarias y ambientales relevantes para la presencia del vector. Para ello, aplicarás técnicas de clustering y reducción de dimensiones para producir visualizaciones e informes comprensibles que puedan ser utilizados por profesionales de salud pública y líderes comunitarios. Tu objetivo final será entregar:

- Un Jupyter Notebook replicable, que incluya código y visualizaciones técnicas para explorar y analizar los datos.
- Un informe divulgativo para instituciones públicas y comunidades, que presente los hallazgos clave de manera clara y accesible, con visualizaciones comprensibles.

Preguntas guía

- ¿Qué variables son más relevantes para caracterizar condiciones de riesgo?
- ¿Cómo transformar las variables para análisis cuantitativo?

- ¿Qué tipos de agrupaciones emergen del análisis?
- ¿Qué relaciones se observan entre materiales de construcción, acceso a agua, presencia de animales y riesgo entomológico?
- ¿Cómo pueden visualizarse y comunicarse estos patrones de forma accesible para la acción comunitaria?

Actividades guía

1. Investigar técnicas de clustering (K-Means, DBSCAN) y reducción de dimensiones (PCA, UMAP, t-SNE).
2. Explorar el conjunto de datos:
    - Identificar variables relevantes para análisis de riesgo (presencia de animales, materiales, cercanía a basura, palmas, tipo de agua, etc.). Tomar de 10 a 20 variables.
    - Transformar variables categóricas (e.g. tipo de techo, material del piso) y normalizar variables cuantitativas (e.g. distancias).
3. Aplicar técnicas de clustering para agrupar viviendas por condiciones de riesgo.
4. Reducir dimensiones para facilitar la visualización de los datos en 2D o 3D.
5. Comparar diferentes algoritmos de clustering y justificar la elección.
6. Elaborar gráficos que muestren claramente los perfiles de riesgo.
7. Redactar un informe final, estructurando los hallazgos en:
    - Introducción al problema.
    - Metodología aplicada.
    - Resultados obtenidos y su interpretación.
    - Recomendaciones basadas en los hallazgos.

Recursos

- Conjunto de datos del proyecto PUCE-Chagas (acceso en el aula virtual).
- Tutoriales y notebooks de clustering y reducción de dimensiones compartidos en clase.
- Documentación de Scikit-learn, Pandas y Matplotlib.
- Tutoriales y ejemplos sobre clustering y reducción de dimensiones dados en clases.

PRODUCTO

PRODUCTO

2.1 Jupyter Notebook

El *Jupyter Notebook* debe ser técnico, bien documentado y fácilmente replicable, permitiendo que otros equipos de análisis puedan aplicar el mismo proceso a nuevas comunidades. Debe estar generado en el formato base dado en clases y contener:

1. **Título y descripción inicial:** Presentación del equipo, contexto del problema, objetivos del análisis y justificación del enfoque metodológico.
2. **Carga y descripción de datos:** Lectura del conjunto de datos, descripción de las variables más relevantes (materiales de la vivienda, presencia de animales, condiciones sanitarias, proximidad de residuos, etc.) y análisis exploratorio inicial.
3. **Preprocesamiento de datos:** Transformación y limpieza de los datos, manejo de valores faltantes, codificación de variables categóricas, estandarización de variables numéricas y justificación de las decisiones tomadas.
4. **Análisis con clustering:** Aplicación de al menos dos algoritmos de agrupamiento (como `K-Means` y `DBSCAN`) sobre subconjuntos de variables seleccionadas. Visualización técnica de los clusters y análisis de cohesión y separación.
5. **Reducción de dimensiones:** Uso de técnicas como `PCA`, `t-SNE` o `UMAP` para representar los datos en espacios de dos o tres dimensiones. Gráficos que ilustren visualmente los clusters y sus diferencias.
6. **Interpretación de resultados:** Análisis de los perfiles de viviendas emergentes a partir de los agrupamientos. Relación con variables de riesgo entomológico y condiciones estructurales y sanitarias.
7. **Conclusión técnica:** Resumen de hallazgos, reflexión sobre la calidad del agrupamiento y limitaciones del análisis. Propuestas para futuras mejoras o extensiones del trabajo.

2.2 Informe Divulgativo

El informe divulgativo debe estar dirigido a autoridades de salud pública, responsables comunitarios y actores sociales involucrados en la erradicación del mal de Chagas. Deberá elaborarse en LaTeX, siguiendo el formato de tareas PUCE, e incluir visualizaciones claras y recomendaciones prácticas.

1. **Introducción:** Contextualización del problema del Chagas en Ecuador, objetivo del estudio y relevancia del análisis de datos para la prevención.

2. **Metodología:** Explicación clara de las técnicas utilizadas (clustering y reducción de dimensiones), su propósito y valor en el análisis de datos complejos.
3. **Resultados principales:** Descripción de los perfiles de viviendas identificados, gráficos comprensibles, mapas o representaciones visuales que permitan a los tomadores de decisiones entender rápidamente los hallazgos.
4. **Implicaciones prácticas:** Discusión de cómo los resultados pueden guiar intervenciones focalizadas, campañas de prevención, mejora de servicios y priorización territorial.
5. **Recomendaciones:** Sugerencias prácticas basadas en los hallazgos, orientadas a mejorar condiciones habitacionales, sanitarias y de control del vector.
6. **Conclusión:** Cierre que resuma la importancia del uso de datos para combatir enfermedades como el Chagas, con énfasis en el trabajo colaborativo entre ciencia, salud y comunidad.

3. Rúbrica de evaluación

Jupyter Notebook (30 puntos totales)

1. **Título y descripción inicial (2 puntos)**
   - Introducción que contextualiza el problema y los objetivos (1 punto).
   - Breve descripción de las técnicas utilizadas (1 punto).
2. **Carga y descripción de datos (4 puntos)**
   - Descripción de las variables relevantes, análisis exploratorio básico (3 puntos).
   - Identificación y representación gráfica adecuada de distribuciones (1 punto).
3. **Preprocesamiento de datos (5 puntos)**
   - Limpieza de datos correctamente implementada (3 puntos).
   - Justificación clara de las decisiones tomadas en el preprocesamiento (2 puntos).
4. **Análisis con clustering (8 puntos)**
   - Implementación adecuada de al menos dos algoritmos de clustering (4 puntos).
   - Visualizaciones técnicas que representen los resultados (2 puntos).
   - Explicación de la calidad de los clusters obtenidos (2 puntos).
5. **Reducción de dimensiones (4 puntos)**
   - Implementación de una técnica de reducción de dimensiones (2 puntos).
   - Visualizaciones técnicas en 2D o 3D que resalten patrones clave (2 puntos).
6. **Interpretación de resultados (2 puntos)**

- Discusión clara de los clusters formados y los patrones identificados (1.5 puntos).
- Relación de los resultados con el problema inicial (0.5 puntos).

7. **Conclusión técnica (2 puntos)**
   - Resumen conciso de los hallazgos técnicos (1.5 puntos).
   - Identificación de limitaciones y posibles mejoras (0.5 puntos).
8. **Legibilidad y simpleza del código (3 puntos)**
   - Código bien comentado, organizado en secciones claras (1.5 puntos).
   - Uso de celdas Markdown para explicar los pasos (1.5 puntos).

Informe Divulgativo (20 puntos totales)

1. **Introducción (3 puntos)**
   - Contextualiza el problema y la relevancia del análisis (2 puntos).
   - Explica claramente los objetivos de la investigación (1 punto).
2. **Metodología (2 puntos)**
   - Breve descripción de las técnicas utilizadas (1.5 puntos).
   - Explicación de cómo estas técnicas ayudan a entender los datos (0.5 puntos).
3. **Resultados principales (5 puntos)**
   - Explicación de los clusters encontrados y sus características (2.5 puntos).
   - Uso de gráficos comprensibles y relevantes para el público objetivo (2.5 puntos).
4. **Implicaciones prácticas (3 puntos)**
   - Relación de los resultados con acciones concretas (2 puntos).
   - Interpretación de patrones significativos y su importancia (1 punto).
5. **Recomendaciones (2 puntos)**
   - Propuestas claras y aplicables basadas en los resultados del análisis (2 puntos).
6. **Conclusión (3 puntos)**
   - Resumen breve y relevante de los hallazgos (2 puntos).
   - Importancia de los resultados para la resolución del problema (1 punto).
7. **Conclusión (3 puntos)**
   - Diseño profesional y organizado con secciones bien definidas (1.5 puntos).
   - Uso de gráficos y tablas adecuadamente formateados (1.5 puntos).