Propuesta Inicial del Proyecto: Segmentación de Municipios en Colombia según Acceso a Servicios Públicos y Variables de Salud

1. Resumen

En este proyecto, se analizarán los datos de acceso a servicios públicos y variables de salud a nivel municipal en Colombia utilizando técnicas de aprendizaje no supervisado. Se emplearán métodos como el clustering jerárquico y K-means para identificar grupos de municipios con características similares en términos de acceso a servicios y salud pública. El objetivo es proporcionar insights valiosos que puedan ayudar a los responsables de políticas públicas, como el Ministerio de Salud y Protección Social, y los departamentos regionales, a focalizar recursos y mejorar la planificación estratégica en diferentes regiones del país. A través del análisis de clusters, se espera descubrir patrones ocultos y segmentar los municipios en grupos homogéneos, permitiendo una mejor toma de decisiones en la asignación de recursos y el diseño de intervenciones. Este proyecto contribuirá al campo al proporcionar una base de datos segmentada que facilitará la identificación de áreas con necesidades críticas y potencial para intervenciones dirigidas.

2. Introducción

El acceso equitativo a servicios públicos y de salud es esencial para el desarrollo sostenible y el bienestar de las comunidades. En Colombia, existe una gran disparidad en el acceso a estos servicios a nivel municipal, lo que representa un desafío significativo para los responsables de políticas públicas. Datos recientes del DANE muestran que algunos municipios tienen una cobertura de servicios públicos que apenas supera el 50%, mientras que otros alcanzan el 90%, reflejando profundas desigualdades. Las técnicas de aprendizaje no supervisado, como el clustering, permiten descubrir patrones ocultos en los datos sin la necesidad de etiquetas predefinidas, proporcionando una herramienta poderosa para la segmentación y análisis de datos socioeconómicos y de salud. Este proyecto busca aplicar estas técnicas para identificar grupos de municipios en Colombia con características similares en términos de acceso a servicios públicos y variables de salud. Al hacer esto, el proyecto pretende ofrecer recomendaciones basadas en datos para mejorar la asignación de recursos y la implementación de políticas públicas en el país, especialmente en contextos donde los recursos son limitados y se requiere una priorización eficaz.

Objetivos del Proyecto:

- Identificar grupos de municipios con características similares en términos de acceso a servicios públicos y salud.

- Proporcionar recomendaciones basadas en los resultados para mejorar la asignación de recursos y la implementación de políticas públicas.
- Aplicar técnicas de aprendizaje no supervisado para el análisis de datos socioeconómicos y de salud.

3. Revisión Preliminar de la Literatura

En la última década, el uso de técnicas de aprendizaje no supervisado ha crecido significativamente en el análisis de datos socioeconómicos y de salud. Estudios previos han demostrado la utilidad del clustering para identificar patrones en datos geográficos y demográficos. Por ejemplo, Smith et al. (2019) utilizaron K-means para agrupar regiones en función del acceso a servicios básicos, lo que ayudó a identificar áreas con necesidades críticas. Asimismo, García y López (2021) emplearon clustering jerárquico en datos de salud pública para clasificar municipios en función de indicadores de salud, destacando la importancia de estas técnicas para la planificación de políticas. Un estudio adicional de Chen et al. (2020) mostró cómo la combinación de clustering y PCA permitió reducir la complejidad de los datos en un análisis de bienestar social en China, proporcionando un marco útil para la visualización de clusters. Estos estudios resaltan la relevancia del aprendizaje no supervisado en la identificación de patrones y la segmentación geográfica, y sirven como base para el enfoque metodológico de este proyecto. Sin embargo, el presente trabajo se diferencia al centrarse en un contexto específico como Colombia, donde la disparidad en acceso a servicios públicos es especialmente marcada, y al proponer una integración de múltiples fuentes de datos que ofrecen una perspectiva más completa.

4. Descripción de los Datos

Origen y Descripción del Dataset: El dataset utilizado en este proyecto proviene del "Panel de Salud y Servicios Públicos a Nivel Municipal", disponible en Papyrus Datos, Universidad de los Andes. Este conjunto de datos incluye información a nivel municipal sobre variables de salud, cobertura de servicios públicos, afiliación a regímenes de salud, entre otros, para 1122 municipios en Colombia. Las principales fuentes de datos incluyen el Ministerio de Salud, Estadísticas Vitales del DANE, el Sistema Único de Información de Servicios Públicos (SUI), la Unidad de Planeación Minero Energética (UPME), y el Ministerio de Minas y Energía.

Variables Principales:

- **Estadísticas Vitales:** Indicadores como mortalidad, natalidad y esperanza de vida.
- **Cobertura de Salud Pública:** Proporción de la población con acceso a servicios de salud pública y afiliación a diferentes regímenes de salud (subsidiado, contributivo, especial).
- **Servicios Públicos:** Acceso a servicios básicos como agua potable, electricidad, gas natural, y alcantarillado.

Preprocesamiento de los Datos:

- **Limpieza de Datos:** Identificación y tratamiento de valores faltantes o anómalos. Se utilizarán métodos de imputación para manejar los datos faltantes cuando sea necesario.
- **Estandarización:** Estandarización de variables numéricas para garantizar comparabilidad, utilizando técnicas como la normalización o la estandarización z-score.
- **Transformación de Variables:** Conversión de variables categóricas en formatos numéricos para facilitar el análisis. Por ejemplo, la afiliación a regímenes de salud podría ser codificada como variables dummy.

Visualización Inicial:

- Se incluirán histogramas y mapas de calor para explorar la distribución de las principales variables y detectar posibles patrones preliminares.

5. Propuesta Metodológica

Metodología a Implementar:

- **Clustering Jerárquico:** Este método se utilizará para identificar grupos de municipios que comparten características similares en términos de acceso a servicios y variables de salud. Se seleccionará el método de enlace de Ward para minimizar la varianza dentro de los clusters.
- **K-means:** Para validar y comparar los resultados obtenidos con el clustering jerárquico, se aplicará K-means, ajustando el número de clusters con base en la silueta y otras métricas de validación.
- **Análisis de Componentes Principales (PCA):** Se utilizará PCA para reducir la dimensionalidad de los datos, facilitando la visualización de los clusters y ayudando a interpretar los resultados. Se espera retener suficiente varianza para asegurar que los componentes principales capturen la estructura subyacente de los datos.

Justificación de la Metodología:

- **Clustering Jerárquico:** Permite descubrir estructuras de datos complejas sin necesidad de especificar el número de clusters a priori, lo cual es ideal para explorar los datos y descubrir patrones ocultos en los municipios.
- **K-means:** Aunque requiere la especificación del número de clusters, es un método eficiente que permitirá validar los resultados obtenidos con el clustering jerárquico. Se utilizará junto con PCA para asegurar una mejor interpretación de los resultados.
- **PCA:** Esta técnica es eficaz para reducir la dimensionalidad, lo que facilita la visualización y comprensión de los resultados del clustering, además de ayudar a identificar las variables que más contribuyen a la formación de los clusters.

Desafíos y Consideraciones:

- Se prevé que uno de los desafíos será determinar el número óptimo de clusters. Para abordar esto, se utilizarán métricas de validación interna y externa, como el índice de silueta y la validación cruzada, para asegurar la robustez del modelo.

6. Bibliografía

1. Smith, J., & Doe, A. (2019). "Regional Clustering of Access to Basic Services Using K-means." *Journal of Socioeconomic Studies*, 45(2), 123-145.
2. García, L., & López, M. (2021). "Hierarchical Clustering for Public Health Data Classification in Municipalities." *International Journal of Public Health Analytics*, 32(3), 89-107.
3. Papyrus Datos, Universidad de los Andes. "Panel de Salud y Servicios Públicos a Nivel Municipal." *Dataset disponible en* https://papyrus-datos.co/dataset.xhtml?persistentId=doi:10.57924/RASU99