README – Proyecto Final de Ciencia de Datos

Modelo Geoespacial para Segmentación del Mercado Mexicano en Estados Unidos

1. Descripción General

Este proyecto implementa una aplicación interactiva en Streamlit que integra datos demográficos del U.S. Census Bureau (ACS 5-Year), información geoespacial a nivel ZIP code y un modelo de Machine Learning (K-Means) para segmentar el mercado hispano en Estados Unidos. El objetivo principal es identificar zonas prioritarias para penetración comercial, marketing y estrategia de expansión de productos con enfoque en hogares mexicanos.

2. Objetivo del Proyecto

Crear un prototipo funcional de un modelo Go-to-Market geoespacial, capaz de:

- Integrar datos públicos del censo estadounidense.

- Unificar información con bases de ZIP codes y coordenadas geográficas.

- Identificar segmentos homogéneos de ZIP codes según ingreso, población mexicana y densidad.

- Visualizar estos segmentos en un mapa interactivo.

- Apoyar decisiones estratégicas para productos dirigidos al mercado latino en EE.UU.

3. Datos Utilizados

a) U.S. Census Bureau – ACS 5-Year API

Variables:

- Población mexicana

- Población total

- Ingreso medio del hogar

- ZIP code tabulation area

b) GeoParquet (WKB geometry)

Contiene coordenadas lat/lon por ZIP code.

c) ZIP Locale Detail CSV

Incluye ciudad, estado y district name.

4. Procesamiento y Limpieza

El pipeline realiza:

- Conversión de columnas numéricas
- Normalización de ZIP codes
- Eliminación de errores del Census
- Validación lógica de datos
- Cálculo de porcentaje de población mexicana
- Unión con dimensión de ciudades y estados
- Eliminación de outliers extremos

5. Exploración de Datos (EDA)

Incluye:

- Histogramas de población e ingreso
- Boxplots para detección de valores atípicos
- Bubble chart ingreso vs porcentaje mexicano
- Filtros por ZIP, estado, ciudad
- Tabla de mercados principales

6. Modelo de Machine Learning: K-Means

K-Means fue seleccionado debido a:

- Su eficiencia y escalabilidad para más de 30,000 observaciones.

- Su buen desempeño con datos numéricos continuos.

- La fácil interpretación de los segmentos mediante centroides.

- La posibilidad de seleccionar automáticamente el número óptimo de clusters mediante Silhouette Score.

7. Visualización Geoespacial

Mediante PyDeck y Mapbox, la app despliega:

- HexagonLayer para densidad de ZIP codes

- Voronoi Polygons por cluster

Incluye tooltips interactivos, zoom, rotación y estadísticas por cluster.

8. Ejecución del Proyecto

Local:

pip install -r requirements.txt

streamlit run app.py

10. Conclusiones

- Se desarrolló una herramienta analítica eficiente y replicable.

- K-Means permitió agrupar ZIP codes en segmentos homogéneos.

- La visualización voronoi facilita interpretar territorios comerciales.

- Este enfoque puede ampliarse incorporando datos de ventas y modelos predictivos futuros.

Autor

Héctor Arturo Argente Amaya