Informe Analítico Bajo la Metodología

ASUM-DM

Cristian Alejandro Alarcón Osorio

Sarah Katalina Gonzales González

José Daniel Muñoz Velandia

3066474

SENA

Centro de Biotecnología Agropecuaria

Análisis y Desarrollo de software

2025

Análisis general del proyecto

El Servicio Nacional de Aprendizaje (SENA) cuenta con una amplia red de Centros de Aprendizaje distribuidos a lo largo del territorio colombiano. La correcta georreferenciación de estos centros es fundamental para la planeación estratégica, la toma de decisiones institucionales y el análisis territorial de la oferta educativa. El presente proyecto busca analizar, depurar y preparar un *dataset* de georreferenciación de los Centros de Aprendizaje del SENA, con el fin de garantizar su calidad y permitir posteriores análisis avanzados como segmentación geográfica, *clustering* y modelado analítico.

1.2 Objetivo general

Analizar y preparar un conjunto de datos de georreferenciación de los Centros de Aprendizaje del SENA mediante la metodología ASUM-DM, garantizando la calidad de los datos y sentando las bases para modelos analíticos posteriores.

1.3 Objetivos específicos

- Comprender la estructura y características del *dataset*.
- Identificar patrones, anomalías y posibles inconsistencias en los datos geográficos.
- Realizar un proceso de limpieza y transformación de los datos.
- Preparar un *dataset* confiable para etapas de modelado y optimización.
- Utilizar un Modelo de IA o analítica de datos.

1.4 Metodología

El proyecto se desarrolla siguiendo la metodología ASUM-DM, abordando las fases de:

- Comprensión del Negocio
- Comprensión de los Datos
- Preparación de los Datos

1.5 Roles del Equipo de Trabajo

- **Cristián:** Análisis técnico, EDA, limpieza de datos
- **Daniel:** Modelado inicial y optimización del modelo.
- **Sarah:** Diseño de la comunicación final, Diseño del *Dashboard* y creación del informe.

Comprensión de los Datos (EDA)

2.1 Descripción del dataset

El *dataset* analizado corresponde a la Georreferenciación de Centros de Aprendizaje del SENA, e incluye información relacionada con:

- Códigos de centro y regional.
- Nombres de regionales, departamentos y municipios.
- Coordenadas geográficas (latitud y longitud).

La carga del *dataset* permitió verificar que los datos se importaron correctamente y que las columnas clave estaban disponibles para el análisis.

2.2 Estructura y tipos de datos

El análisis inicial de la estructura del *dataset* permitió identificar:

- Variables numéricas: LATITUD y LONGITUD.
- Variables categóricas: regionales, departamentos, municipios y códigos administrativos.
- Posibles inconsistencias en tipos de datos, especialmente en variables de código almacenadas como texto.

Este paso fue clave para definir las estrategias de limpieza y transformación posteriores.

2.3 Valores faltantes

El análisis de valores nulos indicó que el *dataset* no presenta valores faltantes críticos, lo cual sugiere una buena calidad inicial de los datos y reduce la necesidad de imputaciones complejas.

2.4 Análisis de distribuciones

Se analizaron las distribuciones de las variables LATITUD y LONGITUD mediante histogramas.

Conclusiones principales:

- La mayoría de los valores se concentran dentro de los rangos esperados para el territorio colombiano.
- No se observan dispersiones extremas ni concentraciones anómalas significativas.
- La distribución es coherente con una cobertura geográfica nacional.

2.5 Análisis de correlaciones

Se evaluó la correlación entre las variables numéricas del *dataset*.

Hallazgos:

- La correlación entre latitud y longitud es baja o moderada.
- Este comportamiento es esperado, ya que ambas variables representan dimensiones espaciales independientes y no una relación causal directa.

2.6 Identificación de *outliers*

Mediante diagramas de caja (*boxplots*) se identificaron posibles valores atípicos en las coordenadas geográficas.

Conclusión:

- Se detectaron algunos puntos potencialmente atípicos que podrían corresponder a errores de digitación o registros fuera del territorio nacional.
- Estos casos requieren revisión antes de análisis geográficos avanzados.

2.7 Detección de anomalías geográficas

Se realizó una validación de rangos geográficos considerando los límites aproximados del territorio colombiano:

- Latitud: entre -5 y 14.
- Longitud: entre -82 y -65.

Resultados:

- No se identificaron registros con coordenadas fuera de estos rangos.

Preparación de los Datos

3.1 Copia de seguridad del dataset

Se creó una copia del *dataset* original para garantizar la trazabilidad y evitar la pérdida de información durante el proceso de limpieza.

3.2 Normalización de nombres de columnas

Se eliminaron espacios innecesarios en los nombres de las columnas para asegurar consistencia y facilitar su manipulación programática.

3.3 Corrección de tipos de datos

Se realizaron las siguientes transformaciones:

- Conversión del CODIGO_CENTRO a tipo entero, eliminando caracteres no numéricos.
- Conversión de códigos administrativos y nombres de regionales, departamentos y municipios a tipo texto.

Estas acciones garantizan coherencia semántica y evitan errores en análisis posteriores.

3.4 Eliminación de duplicados

Se identificaron y eliminaron registros duplicados, asegurando que cada centro de aprendizaje estuviera representado una sola vez en el *dataset* final.

3.5 Codificación de variables categóricas

Se aplicó *One-Hot Encoding* a las variables categóricas, generando variables binarias que permiten su uso en algoritmos de modelado y *clustering*.

3.6 Normalización de variables numéricas

Las variables LATITUD y LONGITUD fueron estandarizadas mediante *StandardScaler*, generando nuevas variables normalizadas:

- LATITUD_norm
- LONGITUD_norm

Este paso es fundamental para algoritmos sensibles a la escala de los datos.

3.7 Ingeniería de características

Se creó una nueva variable denominada *distancia_bog*, que representa la distancia euclidiana aproximada de cada centro de aprendizaje respecto a la ciudad de Bogotá. Esta variable permite:

- Analizar la centralidad geográfica.
- Facilitar análisis de *clustering* territorial.

3.8 Dataset final

El *dataset* limpio y transformado fue exportado como *dataset_limpio.csv*, quedando listo para las fases de modelado, optimización y evaluación.

Modelado: Clustering con K-Means

4.1 Objetivo del modelado

El objetivo del modelado fue aplicar un algoritmo de *clustering* no supervisado (*K-Means*) sobre el *dataset* limpio, con el fin de identificar patrones y segmentaciones geográficas entre los Centros de Aprendizaje del SENA, sin necesidad de una variable objetivo.

4.2 Selección de variables

Para el proceso de *clustering* se seleccionaron las siguientes variables numéricas:

- LATITUD
- LONGITUD
- *distancia_bog* (distancia euclidiana aproximada respecto a Bogotá)

Estas variables permiten capturar tanto la ubicación geográfica como la relación espacial con el principal centro urbano del país.

4.3 Escalado de datos

Dado que *K-Means* es sensible a la escala de los datos, se aplicó *StandardScaler* para estandarizar las variables seleccionadas, asegurando que todas contribuyen de manera equilibrada al cálculo de distancias.

4.4 Determinación del número óptimo de clusters

Se utilizó el Método del Codo (*Elbow Method*) evaluando valores de *K* entre 1 y 10. El análisis de la curva de inercia mostró un punto de inflexión claro en $K = 4$, lo que sugiere que este número de *clusters* ofrece un buen equilibrio entre complejidad del modelo y capacidad de agrupamiento.

4.5 Entrenamiento del modelo final

Con base en el análisis previo, se entrenó el modelo *K-Means* con los siguientes parámetros:

- Número de *clusters*: 4
- Escalado: *StandardScaler*
- Variables: latitud, longitud y distancia a Bogotá

A cada registro del *dataset* se le asignó una etiqueta de *cluster*, generando una nueva variable denominada *cluster*.

Evaluación y Optimización del Modelo

5.1 Evaluación mediante *Silhouette Score*

La calidad del *clustering* se evaluó utilizando el *Silhouette Score*, métrica que mide qué tan bien definidos están los *clusters*. Valores cercanos a 1 indican una buena separación entre grupos. El modelo base (*StandardScaler*, $K = 4$, tres variables) obtuvo el mejor *Silhouette Score*, confirmando una segmentación adecuada.

5.2 Comparación de escalados

Se probó un escalado alternativo mediante *MinMaxScaler* manteniendo el mismo número de *clusters*.

Resultado:

- El *Silhouette Score* obtenido con *MinMaxScaler* fue ligeramente inferior al del modelo base.
- Se decidió mantener *StandardScaler* como método de escalado definitivo.

5.3 Comparación de diferentes valores de K

Se evaluaron configuraciones adicionales con distintos números de *clusters*:

- $K = 3$
- $K = 4$
- $K = 5$

El análisis mostró que $K = 4$ produce el mejor desempeño según el *Silhouette Score*, reforzando la elección inicial.

5.4 Reducción de variables

Se realizó una prueba adicional utilizando únicamente las variables geográficas (latitud y longitud).

Conclusión:

- El *Silhouette Score* disminuyó al eliminar la variable *distancia_bog*.
- Esto indica que dicha variable aporta información relevante para la correcta definición de los *clusters*.

Por lo tanto, se decidió conservar las tres variables originales en el modelo final.

Resultados del Clustering

6.1 Interpretación de los clusters

El análisis de las medias de cada *cluster* permite observar diferencias claras en términos de ubicación geográfica y distancia respecto a Bogotá, lo que sugiere una segmentación territorial coherente de los Centros de Aprendizaje del SENA. Cada *cluster* representa un grupo de centros con características espaciales similares, lo cual puede ser útil para:

- Planeación regional
- Asignación de recursos
- Análisis de cobertura territorial

6.2 Dataset final con clusters

El *dataset* final fue almacenado como *dataset_con_clusters.csv*, incluyendo la variable *cluster* asignada a cada centro, quedando listo para análisis posteriores o visualización geoespacial.

Visualización de Resultados y Dashboard

La fase final del análisis se enfoca en la comunicación efectiva de los resultados de clustering ($K=4$) a través de un dashboard interactivo en Power BI. Cada visualización tiene un propósito específico para validar, contextualizar y dimensionar la segmentación territorial de los Centros de Aprendizaje del SENA.

7.1 Métricas Clave (KPIs)

Justificación: Las Tarjetas KPI son esenciales para establecer un **punto de referencia** cuantitativo y permitir al usuario obtener un resumen numérico inmediato del universo analizado (o del subconjunto filtrado).

	Métrica y Función	**Objetivo de Análisis**
Total de Centros Analizados	CODIGO_CENTRO (Recuento Distinto)	Muestra el tamaño total de la muestra analizada (N).
Número de Clústeres Definidos	cluster (Recuento Distinto)	Confirma el valor de K seleccionado en la fase de modelado ($K=4$).
Distancia Promedio Total (km)	distancia_bog (Promedio)	Establece el valor central de la variable de ingeniería de características.

7.2 Gráfico de Dispersión (Scatter Plot)

Justificación: Este es el gráfico de **validación visual** del modelo K-Means. Permite demostrar que los clústeres no se generaron al azar, sino que se separan lógicamente en función de las variables de entrada.

Distribución del Clustering por Ubicación y Distancia

Composición (Eje)	Campo y Función	Objetivo del Análisis
Eje X	distancia_bog (Promedio)	Muestra el alejamiento o cercanía de Bogotá.
Eje Y	LATITUD (Promedio)	Muestra la distribución geográfica Norte-Sur.
Leyenda (Color)	cluster	El color identifica visualmente los límites de cada grupo.
Tamaño	CODIGO_CENTRO (Recuento)	Muestra la concentración de centros en cada punto geográfico.

7.3 Visual de Mapa (Map Visual)

Justificación: Proporciona el **contexto geográfico** y la visualización espacial de los resultados, permitiendo confirmar si los grupos de clústeres forman regiones geográficas contiguas.

Ubicación de Centros y Pertenencia a Clúster

Composición	Campo y Función	Objetivo del Análisis
Latitud/Longitud	LATITUD, LONGITUD (No resumir)	Posiciona cada centro en el territorio colombiano.
Leyenda (Color)	cluster	Muestra la distribución geográfica del resultado del clustering.

7.4 Gráfico de Columnas Apiladas

Justificación: Este gráfico evalúa la **composición interna** de cada clúster, identificando qué entidades administrativas (Regionales) son las que más contribuyen a la definición de cada grupo.

Composición Regional de los Clústeres (Distribución %)

Composición	Campo y Función	Objetivo del Análisis
Eje X	cluster	Las categorías principales (Clúster 0, 1, 2, 3).
Eje Y	CODIGO_CENTRO (Recuento)	El total de la columna, que representa el **100%** de cada clúster.
Leyenda (Color)	NOMBRE_REGIONAL	Los colores dentro de la columna muestran la contribución proporcional de cada Regional.

7.5 Gráfico de Columnas Agrupadas y de líneas

Justificación:

Responder preguntas como:

- **¿La regional con más centros (columna más alta) es también la más cercana a Bogotá (línea más baja)?**
    - Si la columna es alta y la línea es baja, esa regional tiene muchos centros y está muy cerca de Bogotá.
- **¿Existen regiones con pocos centros (columna baja) pero que están muy lejos (línea alta)?**
    - Estas podrían ser candidatas a requerir más atención o inversión, ya que están dispersas geográficamente.

Composición Regional de los Clústeres (Distribución %)

Composición	Campo y Función	Objetivo del Análisis
Eje X (Compartido)	NOMBRE_REGIONAL	La categoría principal que vamos a analizar.
Columna (Eje Y)	CODIGO_CENTRO	**Contar** cuántos centros tiene cada regional.
Línea (Eje Y)	distancia_bog	Mostrar la **distancia promedio** a Bogotá.

El **diseño del dashboard**, basado en estos cinco visuales clave, permite comprender rápidamente la segmentación geográfica de sus centros, **validar** la coherencia del modelo K-Means, y tomar decisiones informadas sobre la asignación de recursos y el análisis de cobertura territorial.

Conclusiones Generales

- El *dataset* presenta una buena calidad inicial y permitió un proceso de análisis y limpieza eficiente.
- El uso de técnicas de EDA facilitó la identificación de anomalías y patrones espaciales.
- El modelo *K-Means* con 4 *clusters*, *StandardScaler* y tres variables demostró ser la mejor configuración.
- La variable distancia a Bogotá aporta información relevante al proceso de segmentación.
- Los resultados obtenidos permiten comprender mejor la distribución geográfica de los Centros de Aprendizaje del SENA y sirven como base para análisis estratégicos futuros.

Webgrafías:

Dataset :Georeferenciación Centros de Aprendizaje SENA | Datos Abiertos Colombia