#### EVIDENCIA DE PRODUCTO PARA CERTIFICAR

#### POR COMPETENCIAS LA NORMA 220501114

#### Informe de analítica de datos

**1. Portada**

* **Nombre del proyecto: Informe de analítica de datos, evidencia producto norma 22050114.**

**Sistematizar datos masivos de acuerdo con métodos de analítica y herramientas tecnológicas**

* **Aprendices: Emilia Gallo Alzate, Maria Ximena Marin Delgado**
* **Fecha: 22 de Octubre de 2025**
* **2. Introducción**
* **Objetivo del proyecto:** Desarrollar una solución con apoyo de Inteligencia Artificial que ayude a resolver un problema real en una comunidad local en los campos de educación, salud, medio ambiente o seguridad, utilizando técnicas de clasificación y resumen de datos, metodologías ágiles y principios éticos de la IA.
* **Propuesta de solución: Se desarrolló una aplicación de analítica de datos impulsada por IA que procesa información proveniente de comunidades locales (encuestas, reportes y comentarios). El sistema aplica limpieza de datos, análisis de sentimientos y clustering para detectar patrones de urgencia y generar reportes automatizados con IBM Granite y modelos de lenguaje. Esto permite identificar zonas críticas y necesidades prioritarias de atención para los entes gubernamentales.**
* **Fuentes de datos:** La ruta de habilitación para el reto SenaSoft en la categoría Inteligencia Artificial está definida dentro de los parámetros formativos con los que cuenta IBM SkillsBuild, dando una base que permita obtener los conocimientos básicos para la aplicación en resolución de problemas específicos con apoyo y uso de IA:

<https://skills.yourlearning.ibm.com/activity/PLAN-D40AB1C86960?ngoid=0302&utm_campaign=open-SENASOFT2025>

* **Tipos de Fuentes de datos ( Primaria, secundaria / Estructurada, No estructurada / Interna, externa):**
* **Secundaria (proporcionada por IBM)**
* **Estructurada (CSV)**
* **Externa (no pertenece a una base institucional)**

**Para conocer más sobre el proceso de desarrollo de este proyecto visítanos en:** <https://mari0724.github.io/senasoft/>

* **3. Metodología**
* **Describa el proceso utilizado para la limpieza de datos:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Proceso** | **Método seleccionado** | **Justificación** | **Herramientas usadas** |
| **Extracción** | **Lectura y carga de datos con pandas y numpy desde CSV, controlando codificación y delimitador.** | **Facilita la manipulación eficiente de datos estructurados.** | **Python, pandas, numpy, python-dateutil** |
| **Transformación** | **Limpieza y normalización de texto: minúsculas, eliminación de acentos y símbolos. Conversión de tipos y normalización de columnas.** | **Garantiza consistencia y precisión antes del análisis.** | **regex, unicodedata, nltk, scikit-learn, pandas** |
| **Cargue** | |  | | --- | | **Exportación del dataset limpio y unificado a formato CSV para posterior análisis y visualización.** |  |  | | --- | |  | | |  | | --- | | **Permite interoperabilidad con librerías analíticas y gráficas.** |  |  | | --- | |  | | **pandas, csv, python-dotenv** |
| **Análisis estadístico** | |  | | --- | | **Aplicación de conteos, correlaciones, métricas de clustering y evaluación de rendimiento.** |  |  | | --- | |  | | |  | | --- | | **Detecta patrones, relaciones y tendencias dentro de los datos.** |  |  | | --- | |  | | **numpy, scikit-learn, matplotlib, seaborn, torch, tqdm** |

* **Algoritmo aplicado**
  + **Algoritmo: K-Means Clustering + Análisis de Sentimiento (Transformers)**
* **Librerías: scikit-learn, transformers, pandas, nltk, sentence-transformers**
* **Razón de uso:**

**K-Means permitió agrupar las comunidades según nivel de urgencia y similitud textual, mientras que el análisis de sentimiento clasificó los comentarios como positivos, negativos o neutros, permitiendo contextualizar los resultados.**

**Preparación, validación y ajuste del modelo analítico**

* **Tipo de modelo utilizado: Modelo no supervisado (K-Means) combinado con análisis semántico supervisado (sentiment analysis).**
* **Calibración o ajustes realizados al modelo:** Se evaluó la calidad del clustering mediante métricas como:
* Silhouette: **0.56**
* Davies-Bouldin: **1.08**
* Calinski-Harabasz: **3150.20**  
  Estos valores indicaron una buena separación y cohesión entre grupos.
* **Interpretación de resultados**

Se identificaron cuatro grupos principales:

* Zona crítica
* Zona invisible
* Zona puntual
* Zona estable

Cada grupo representó un nivel diferente de urgencia social y atención prioritaria.

* **4. Análisis y hallazgos relevantes**
* **Hallazgos : Resalta solo los hallazgos relevantes: tendencias, problemas detectados, patrones importantes. Incluye gráficas o tablas de ser necesarias**

Los datos revelaron que:

Las zonas rurales con bajo acceso a internet y sin atención previa del gobierno se concentraron en la Zona crítica.

Comunidades con comentarios negativos o de frustración coincidieron con altos niveles de urgencia.

Las áreas con apoyo gubernamental previo mostraron mayor estabilidad en los indicadores.

Estos hallazgos permitieron priorizar recursos y detectar comunidades invisibilizadas dentro del conjunto de datos.

* **5. Visualizaciones**
* **Gráficos que sustenten los hallazgos (de barras, líneas, pastel, mapas de calor).**

**El dashboard desarrollado, “ComuniMind – Panel de Impacto Social Comunitario”, integra cinco gráficos que resumen los principales hallazgos del análisis de datos:**

1. **Impacto social promedio por ciudad (barras horizontales): identifica los municipios con mayor necesidad de intervención.**
2. **Categorías de problema más urgentes (barras verticales): muestra los temas con mayor prioridad, como salud y educación.**
3. **Acceso a internet vs nivel de urgencia (barras apiladas): evidencia cómo la brecha digital influye en la urgencia social.**
4. **Distribución de reportes por género (gráfico circular): refleja la participación y afectación de los distintos grupos poblacionales.**
5. **Sentimiento promedio por ciudad (barras coloreadas): resume la percepción emocional de las comunidades.**

**Estas visualizaciones, generadas con Python y Matplotlib, facilitan la interpretación de los resultados y permiten a la ONG enfocar sus recursos en las zonas y problemáticas de mayor impacto social.**

* **6. Conclusiones**

**Resumen breve 3-4 líneas, centrado en hallazgos y recomendaciones estratégicas.**

**Un letrero de color blanco

El contenido generado por IA puede ser incorrecto.**La solución de IA permitió identificar y priorizar necesidades sociales, destacando que el acceso a internet y la atención gubernamental son factores clave del bienestar comunitario. Su análisis ético mediante clustering y sentimiento generó información útil para decisiones más sostenibles.

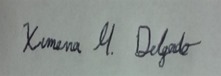
**Firma candidato (Digital)**

**Nombres Apellidos: Emilia Gallo Alzate**

**Documento Identidad: 1011396062**

**Numero celular:3127131921**

**Correo electrónico:** [emiliagalloalzate85@gmail.com](mailto:emiliagalloalzate85@gmail.com)



**Firma candidato (Digital)**

**Nombres Apellidos: Maria Ximena Marin Delgado**

**Documento Identidad: 1020112289**

**Numero celular:3178523898**

**Correo electrónico:** [ximenadelgadom07@gmail.com](mailto:ximenadelgadom07@gmail.com)