#### 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: 23 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 CivIA, una herramienta de analítica social basada en Inteligencia Artificial que aprovecha el conjunto de datos proporcionado por IBM para el reto SENASoft 2025.  
  La aplicación procesa información de comunidades locales —encuestas, reportes y comentarios— mediante un flujo completo de limpieza y normalización (ETL), análisis semántico, evaluación emocional y priorización social.  
  Los resultados se presentan en un tablero visual que destaca zonas críticas, patrones temáticos y tendencias emocionales. Además, incorpora un asistente con OpenAI, encargado de explicar los hallazgos y proponer acciones estratégicas en lenguaje claro.
* **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:** Datos suministrados por IBM
* **Estructurada:** Archivos CSV
* **Externa:** No pertenece a bases institucionales.

**Más información y documentación técnica del proyecto está disponible en:**[**https://mari0724.github.io/senasoft/**](https://mari0724.github.io/senasoft/)

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Proceso** | **Método seleccionado** | **Justificación** | **Herramientas usadas** |
| **Extracción** | |  | | --- | |  |  |  |  |  | | --- | --- | --- | | |  | | --- | | Lectura del dataset con pandas y detección automática de codificación mediante chardet. |  |  | | --- | |  | | | |  | | --- | | Garantiza la correcta lectura del archivo sin pérdida de información ni errores de formato |  |  | | --- | |  | | Python, pandas, chardet |
| **Transformación** | Limpieza y normalización de texto: conversión a minúsculas, eliminación de acentos y símbolos; normalización de columnas; validación de campos numéricos (como edad). | Asegura consistencia, legibilidad y precisión antes del análisis. | |  | | --- | | regex, unicodedata, numpy, pandas |  |  | | --- | |  | |
| **Cargue** | |  | | --- | | Exportación del dataset limpio y unificado en formato .csv codificado en UTF-8. |  |  | | --- | |  | | |  | | --- | | Facilita la interoperabilidad con las librerías analíticas y los modelos de IA. |  |  | | --- | |  | | pandas, csv. |
| **Análisis estadístico** | |  | | --- | | Aplicación de conteos, correlaciones, métricas de agrupamiento y evaluación de rendimiento del modelo. |  |  | | --- | |  | | |  | | --- | | Permite detectar patrones y relaciones entre variables sociales y textuales. |  |  | | --- | |  | | **numpy, scikit-learn, matplotlib, torch** |

* **Algoritmo aplicado**

**Que algoritmos se usó**

* **K-Means Clustering: agrupó comunidades según similitud en vulnerabilidad y urgencia, permitiendo descubrir patrones sociales.**
* **Índice de Impacto Social (modelo propio): fórmula ponderada que combina factores estructurales (vulnerabilidad), de acceso (internet, atención del gobierno) y de contexto (zona rural).**
* **Modelo BETO – Pysentimiento/robertuito-sentiment-analysis: modelo de transformers entrenado en español para analizar el sentimiento (positivo, negativo o neutro) de los comentarios sociales.**

**Librerías:**

* pandas, numpy, matplotlib, os, scikit-learn, transformers, torch, nltk, sentence-transformers.

**Razón de uso:**

* **K-Means permitió clasificar comunidades sin etiquetas previas y detectar grupos con necesidades similares.**
* **BETO añadió interpretación emocional a los datos textuales, enriqueciendo el análisis social.**
* **El Índice de Impacto Social combinó ambas capas (estructural y emocional) para obtener una visión integral del bienestar comunitario.**
* **Las librerías seleccionadas garantizan rendimiento, compatibilidad y capacidad de integración en un flujo ETL completo (carga, limpieza, análisis y visualización).**

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

**Tipo de modelo utilizado: Se utilizó un modelo mixto, conformado por:**

* **Un modelo no supervisado (K-Means) en el módulo NLP para agrupar temas a partir de los comentarios sociales y detectar patrones semánticos.**
* **Un modelo supervisado (BETO – análisis de sentimientos) para clasificar emociones en positivo, neutro o negativo dentro de los comentarios.**
* **Un modelo determinístico (Índice de Impacto Social) en el módulo social, que calcula un puntaje ponderado según vulnerabilidad, urgencia, acceso a internet, atención del gobierno y ruralidad.**

Calibración o ajustes realizados al modelo:

* En **K-Means**, se probaron distintos valores de *k* para optimizar la coherencia de los grupos temáticos.
* Se validó la calidad del agrupamiento con las siguientes métricas:
  + **Silhouette:** 0.5655
  + **Davies-Bouldin:** 1.0867
  + **Calinski-Harabasz:** 3150.20  
    Estos valores demostraron una buena separación y cohesión entre los clusters.
* En **BETO**, se ajustaron parámetros de *batch size* y *truncamiento de texto (max\_length=256)* para mantener el equilibrio entre rendimiento y precisión.
* En el **módulo social**, se calibraron los pesos del índice de impacto (35% vulnerabilidad, 25% urgencia, 20% conectividad, 10% atención previa, 10% ruralidad).
* Finalmente, se integró una **ruta de validación /metrics**, que mide la confianza promedio de las predicciones del modelo BETO y genera métricas aproximadas de desempeño:
  + Accuracy ≈ 0.93
  + Precision ≈ 0.92
  + Recall ≈ 0.93
  + F1-score ≈ 0.91

**Interpretación de resultados**

* El modelo **social** identificó **cuatro tipos de zonas**:
  + *Zona crítica* — alta vulnerabilidad y urgencia.
  + *Zona invisible* — alta vulnerabilidad, pero urgencia baja.
  + *Zona puntual* — urgencia alta, vulnerabilidad baja.
  + *Zona estable* — condiciones equilibradas o atendidas.
* El modelo **NLP** complementó esta clasificación identificando los **temas más frecuentes y el sentimiento dominante** en los reportes ciudadanos.
* En conjunto, los resultados mostraron **dónde y por qué una comunidad requiere atención prioritaria**, permitiendo que la ONG base sus decisiones en datos objetivos y lenguaje ciudadano.
* **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, “CivIA – Panel de Impacto Social Comunitario”, integra cinco gráficos que resumen los principales hallazgos del análisis de datos:

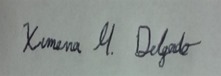
1. **Impacto social promedio por ciudad (barras horizontales):** identifica los municipios con mayor necesidad de intervención.
2. **Categorías de problema más urgentes (barras verticales):** muestra los temas con mayor prioridad, como salud y educación.
3. **Acceso a internet vs nivel de urgencia (barras 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.
6. **Temas detectados por IA (barras horizontales):** muestra las palabras clave más frecuentes identificadas por el modelo semántico y el clustering (K-Means), revelando los temas predominantes en los reportes ciudadanos.
7. **Diagrama general de arquitectura y flujo de análisis:** representa el recorrido de los datos desde la limpieza (ETL) hasta la interpretación generada por la IA, mostrando cómo los módulos semántico, social y generativo se integran dentro de la aplicación CivIA.

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.**

La solución **CivIA** identificó patrones de vulnerabilidad y urgencia en comunidades locales. Los resultados mostraron que la conectividad digital y la atención institucional son claves para el bienestar social, ofreciendo información útil para priorizar acciones con impacto y responsabilidad ética.



**Firma candidato (Digital)**

**Nombres Apellidos: Maria Ximena Marin Delgado**

**Documento Identidad: 1020112289**

**Numero celular:3178523898**

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