#### 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**

**Bryan Escobar**

**Juan Roman**

**23/10/2025**

* **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:**

1. **Cargar datasets locales con información de las ciudades sobre las categorías críticas (salud, educación, seguridad, medio ambiente).**
2. **Procesar y normalizar los datos automáticamente, identificando columnas relevantes y eliminando información sensible o duplicada.**
3. **Analizar y resumir los datos para generar estadísticas por ciudad y categoría, incluyendo:**
4. **Número total de reportes por ciudad y categoría.**
5. **Distribución de nivel de urgencia.**
6. **Combinaciones relevantes entre categoría, ciudad y urgencia.**
7. **Responder preguntas de los usuarios en lenguaje natural sobre los datos.**
8. **Generar informes automáticos basados en los datos, listos para análisis ejecutivo o toma de decisiones.**
9. **Interfaz amigable tipo chatbot, con dos vistas principales:**
10. **Vista de chat para interactuar con el bot y consultar información.**
11. **Vista de análisis textual para visualizar resúmenes completos de los datos cargados.**

* **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):**

1. **Secundaria: Datos obtenidos de fuentes ya existentes, como bases de datos gubernamentales, estudios previos, reportes de ONGs, portales estadísticos o investigaciones académicas.**
2. **Estructurada: Datos organizados en tablas o bases de datos, como registros de incidentes, reportes de salud o educación con campos específicos (ciudad, categoría, urgencia, fecha).**
3. **Interna: Datos generados y gestionados dentro de la organización o comunidad, como registros internos de municipalidades, escuelas o centros de salud.**

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Proceso** | **Método seleccionado** | **Justificación** | **Herramientas usadas** |
| **Extracción** | **Carga de archivos CSV proporcionados por la comunidad o instituciones locales** | **Permite recolectar datos directos y estructurados sobre comentarios, categorías y ciudades, asegurando información confiable y específica del contexto.** | **Python (FastAPI, Pandas), formulario de carga en frontend** |
| **Transformación** | |  | | --- | |  |  |  | | --- | | **Limpieza y normalización de columnas, eliminación de duplicados y columnas sensibles** | | **Necesario para garantizar la consistencia de los datos y proteger información privada; homogeneiza nombres de columnas y prepara datos para análisis.** | **Python (Pandas, unicodedata)** |
| **Cargue** | |  | | --- | |  |  |  | | --- | | **Guardado de datos limpios en CSV y JSON de resumen** | | **Facilita que el chatbot y los endpoints consuman datos ya procesados, asegurando que el análisis sea rápido y reutilizable.** | **Python (Pandas, JSON), almacenamiento local (data/)** |
| **Análisis estadístico** | **Generación de conteos, porcentajes, combinaciones por ciudad/categoría y jerarquías** | **Permite entender la distribución de reportes, identificar tendencias y destacar las categorías con más incidencias en cada ciudad, base para el chatbot** | **Python (Pandas), API de IA (Gemini)** |

| **Algoritmo aplicado** | **Qué algoritmo se usó** | **Qué librerías** | **Razón de uso** |
| --- | --- | --- | --- |
| **Clasificación de texto / Reconocimiento de patrones** | Búsqueda por coincidencia de palabras clave en comentarios y categorías | Python: pandas, unicodedata | Permite categorizar los comentarios según palabras clave relacionadas con salud, educación, seguridad o medio ambiente. |
| **Resumen estadístico** | Conteo, porcentajes, jerarquía ciudad → categoría → urgencia | Python: pandas, json | Para generar resúmenes cuantitativos y estructurados del dataset, base para análisis del chatbot. |
| **Chatbot basado en IA** | Modelo de lenguaje GPT / LLaMA vía IBM Granite | requests para API, JSON | Para responder preguntas del usuario usando el resumen estadístico y generar informes en lenguaje natural. |
| **Protección de datos** | Filtrado de columnas sensibles / normalización | Python: pandas, unicodedata | Para garantizar que la información sensible o no relevante no sea procesada ni mostrada al usuario. |

| **Preparación, validación y ajuste del modelo analítico** | **Detalle** |
| --- | --- |
| **Tipo de modelo utilizado** | **Modelo de lenguaje basado en IA (GPT / LLaMA) para generar respuestas en lenguaje natural, combinado con análisis estadístico descriptivo del dataset (conteos, porcentajes, jerarquías por ciudad, categoría y urgencia).** |
| **Calibración o ajustes realizados al modelo** | **Normalización y limpieza de los datos del CSV (eliminación de duplicados, manejo de valores faltantes, unificación de nombres de columnas).**  **Filtrado de columnas sensibles para proteger la privacidad.**  **Ajuste de prompts al modelo de IA para obtener respuestas coherentes y contextuales usando los resúmenes estadísticos.**  **Validación de la correspondencia entre la pregunta del usuario y los datos disponibles para evitar respuestas inventadas.** |
| **Interpretación de resultados** | **La IA interpreta los resúmenes generados para dar insights claros por ciudad y categoría.**  **Se identifican tendencias, categorías con mayor número de reportes y niveles de urgencia predominantes.**  **La salida es en lenguaje natural, lo que permite a usuarios no técnicos comprender el estado de cada categoría en cada ciudad y posibles áreas de intervención.** |

* **4. Análisis y hallazgos relevantes**
* **Hallazgos : Resalta solo los hallazgos relevantes: tendencias, problemas detectados, patrones importantes. Incluye gráficas o tablas de ser necesarias**
* **5. Visualizaciones**
* **Gráficos que sustenten los hallazgos (de barras, líneas, pastel, mapas de calor).**
* **6. Conclusiones**

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

**Firma candidato (Digital)**

**Nombres Apellidos: Bryan Andrei Escobar Valencia**

**Documento Identidad: 1126664332**

**Numero celular:3224572659**

**Correo electrónico: lobo638yt** **@gmail.com**

**Firma candidato (Digital)**

**Nombres Apellidos:**

**Documento Identidad:**

**Numero celular:**

**Correo electrónico:**