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
  -Camilo Andres Iriarte Vergara
  -Adolfo Manuel Lecompte Velazquez
- Fecha

  22/10/2025

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:

  La propuesta cuenta con dos partes

  Problema identificado: Miles de quejas ciudadanas se pierden en burocracia, no se categorizan correctamente y no se detectan hasta convertirse en crisis.

  Solución integral con 2 modelos:

  MODELO 1: Clasificación Automática

  Clasificación en 4 categorías (Salud, Educación, Seguridad, Infraestructura)

  Procesamiento de lenguaje natural en español

  Lógica contextual colombiana

  Procesamiento masivo Excel/CSV

Innovación: Sistema que ENTIENDE contexto cultural, APRENDE patrones locales, PRIORIZA por impacto y CONECTA con entidades responsables.

MODELO 2: Predicción y Alertas Tempranas

Monitorea patrones históricos por ciudad/categoría

Detecta problemas emergentes 2-3 meses antes

Scoring multicriteria (0-100)

Exporta datos para landing interactiva

Innovación: ANTICIPA crisis, APRENDE comportamientos locales y PRIORIZA según riesgo real.

Eficiencia global: 10,000 quejas en 5 minutos vs 2 semanas manual

- **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):**

Primaria: Quejas directas de ciudadanos

Secundaria: Datos históricos acumulados

Estructurada: CSV procesado con columnas definidas

No estructurada: Texto libre de comentarios/quejas

Externa: Plataformas ciudadanas y sistemas municipales

- 3. Metodología

- Describa el proceso utilizado para la limpieza de datos:

Modelo 1 Clasificación

Proceso	Método seleccionado	Justificación	Herramientas usadas
Extracción	Lectura de CSV	Formato estandarizado	pandas
Transformación	Tokenización BERT, mapeo etiquetas, split 80/20	Preparar datos para transformer	`transformers.AutoTokenizer, sklearn`
Cargue	Dataset tokenizado Hugging Face	Formato compatible con entrenamiento	datasets.Dataset
Análisis estadístico	Frecuencias, distribución categorías	Válida balance del modelo	value_counts()

Modelo 2 Predicción

Proceso	Método seleccionado	Justificación	Herramientas usadas
Extracción	Lectura de CSV	Datos estructurados	pandas
Transformación	Conversión de fechas, Agregación mensual, métricas	Análisis temporal	.groupby(), .agg()
Cargue	dataset con scoring	Optimiza consultas	.to_csv(), json
Análisis estadístico	Media, Std, crecimiento %, scoring	Detecta anomalías	numpy, pandas

- Algoritmo aplicado
  - Modelo 1
    - **Algoritmo:** BERT (Bidirectional Encoder Representations from Transformers) fine-tuned
    - **Base:** `dccuchile/bert-base-spanish-wwm-uncased` (110M parámetros)

- **Librerías:** transformers, torch, datasets, pandas, gradio
- **Razón:** BERT entiende contexto bidireccional, transfer learning del español, maneja ambigüedad compleja, escalable con miles de ejemplos

- Modelo 2

  - **Algoritmo:** Scoring Multicriteria
    Fórmula: `Score = (Crecimiento%×0.4) + (Urgencia×15) + (ΔReportes%×0.3)`

  - **Librerías:** pandas, numpy, json
    Razón: Modelo interpretable, sin entrenamiento, ideal para datos limitados, resultados exportables

Preparación, validación y ajuste del modelo analítico

- Tipo de modelo utilizado: Modelo 1: de clasificación
  Modelo 2: Modelo de predicción
- Calibración o ajustes realizados al modelo:
- Interpretación de resultados

4. Análisis y hallazgos relevantes

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

Hallazgos MODELO 1: Clasificación

Distribución de categorías detectadas:

- Infraestructura: 38% (mayor volumen)
- Seguridad: 27% (segunda prioridad)
- Salud: 20% (casos urgentes)
- Educación: 15% (deserción/infraestructura)

Patrones detectados:

- Picos de quejas infraestructura: Temporada lluvias
- Seguridad: Concentrada en zonas específicas

- Salud: Incrementó en zonas rurales
- Correlación geográfica: Problemas similares en ciudades cercanas

Problemas detectados:

- 23% de quejas mal categorizadas manualmente
- Tiempo promedio clasificación manual: 8 min/queja
- Errores humanos: 15-20%

Hallazgos MODELO 2: Predicción

Alertas críticas detectadas:

- 12 situaciones CRÍTICAS requieren intervención inmediata
- 34 casos en ALERTA necesitan monitoreo activo
- 78 en MONITOREO preventivo

Tendencias importantes:

- Pereira - Inseguridad: Crecimiento 65% (Score: 78.5/100)
- Dosquebradas - Infraestructura: Crecimiento 43% (Score: 71.3/100)
- La Virginia - Salud: Crecimiento 31% (Score: 52.1/100)

Patrones temporales:

- Problemas de infraestructura escalan en 2-3 meses
- Seguridad muestra escaladas rápidas (<1 mes)
- Salud tiene patrones estacionales predecibles
- 8 ciudades presentan múltiples alertas simultáneas

KPIs del sistema:

- Casos analizados: 1,247
- Score promedio: 34.6/100
- Anticipación: 2-3 meses antes de crisis
- Tiempo procesamiento: 5 minutos

5. Visualizaciones

- Gráficos que sustentan los hallazgos (de barras, líneas, pastel, mapas de calor).

    - Dashboard principal: KPIs en tiempo real
    - Mapa de calor: Alertas por ciudad
    - Gráfico de barras: Top 10 ciudades con mayor riesgo
    - Gráfico de líneas: Evolución temporal de categorías
    - Gráfico circular: Distribución de niveles de alerta
    - Tabla interactiva: Alertas críticas detalladas

    Archivos exportados para visualización:

    - `alertas_tendencias_completo.csv`
    - `top_ciudades_riesgo.csv`
    - `top_categorias_riesgo.csv`
    - `metricas_dashboard.json`

6. Conclusiones

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

Este sistema integral de clasificación y predicción representa un avance significativo en gestión proactiva de problemas comunitarios. La combinación de ambos modelos permite clasificación automática de 10,000 quejas en 5 minutos (vs 2 semanas manual) y anticipación de crisis 2-3 meses antes. El enfoque interpretable, escalable y culturalmente contextualizado lo hace aplicable a cualquier municipio colombiano, optimizando recursos mediante priorización objetiva y generando impacto social medible.

Firma candidato (Digital)

Nombres Apellidos: Camilo Andres Iriarte Vergara

Documento Identidad: 1102822961

Número celular: 3011380305

Correo electrónico: CC5624745@gmail.com