Guía del Proyecto 2.
Análisis Exploratorio

INTRODUCCIÓN

La Ciencia de Datos permite abordar una gran variedad de problemas complejos. Con el fin de que los estudiantes desarrollen habilidades prácticas, este año se proponen diferentes retos de proyecto. La actividad se realizará en grupos de 4 integrantes, quienes deberán seleccionar uno de los siguientes desafíos:

#	Reto	Tema
1	**Jigsaw - Agile Community Rules Classification**	**Procesamiento del Lenguaje Natural**
2	**MITSUI&CO. Commodity Prediction Challenge**	**Series de tiempo**
3	**MAP - Charting Student Math Misunderstandings**	**Procesamiento del Lenguaje Natural**
4	**RSNA Intracranial Aneurysm Detection**	**Visión Artificial**
5	NeurIPS - Desafío de datos Ariel 2024	Visión Artificial
6	Clasificación degenerativa de la columna lumbar RSNA 2024	Visión Artificial
7	Predecir qué respuesta preferirá un usuario en esta batalla cara a cara con datos del Chatbot Arena	Procesamiento del Lenguaje Natural
8	Predicción de compradores recurrentes: cuestionar la línea base	Negocios
9	CommonLit - Evaluar resúmenes de estudiantes	Procesamiento del Lenguaje Natural
10	Detección de estructuras microvasculares en tejidos de riñón humano sanos.	Visión Artificial
11	Detectar y clasificar lesiones abdominales traumáticas	Visión Artificial
12	Google: reconocimiento de deletreo manual del lenguaje de señas estadounidense	Visión Artificial
13	Desafío Ojos en el Terreno del CGIAR. Detección de enfermedades en plantas	Visión Artificial
14	Identificación de Especies de Mosquitos	Visión Artificial
15	Hackeando el cuerpo humano	Visión Artificial
16	Predicción de argumentos efectivos	Procesamiento del Lenguaje Natural
17	Detección de fracturas de las vértebras cervicales en radiografías	Visión Artificial
18	Detección de pases en videos de jugadas de football de la liga alemana	Visión Artificial
19	Clasificación de los orígenes de un coágulo de sangre en un accidente cerebrovascular	Visión Artificial
20	Identificación de menciones de entidades biomédicas en resúmenes de artículos de investigación	Procesamiento del Lenguaje Natural
21	Identificación de relaciones de entidades biomédicas en resúmenes de artículos de investigación	Procesamiento del Lenguaje Natural

Nota: Los primeros 4 retos son competencias activas.

INSTRUCCIONES

Sobre los Grupos:

- Los estudiantes deben inscribirse en los grupos de Canvas para el Proyecto 2. **Si no se inscriben, no recibirán calificación**.

Sobre los Retos:

- Cada grupo debe seleccionar un reto **diferente**. En caso de repetición, se evaluará únicamente la entrega del primer grupo inscrito.

Sobre el proyecto:

- Pueden trabajar en **Google Colab** o **Kaggle**, pero es obligatorio versionar el código en **GitHub**, ya que se evaluarán las contribuciones individuales.

ACTIVIDADES

1. **Planteamiento inicial del problema:**
    a. **Situación problemática:** describir el contexto o situación que origina el reto seleccionado.
    b. **Problema científico:** enunciar con claridad el problema a resolver derivado de la situación planteada.
    c. **Objetivos:** redactar al menos un objetivo general y dos objetivos específicos, que sean medibles y alcanzables durante el desarrollo del proyecto.

    Ejemplo (para un reto ficticio de Visión por Computadora):

    a. *Situación problemática:* En los hospitales públicos se acumulan grandes cantidades de radiografías que requieren diagnóstico rápido. El retraso en su revisión puede poner en riesgo la atención médica oportuna.
    b. *Problema científico:* ¿Es posible entrenar un modelo de visión por computadora que identifique de manera confiable fracturas en radiografías de tórax, reduciendo los tiempos de diagnóstico?
    c. *Objetivo general:* Desarrollar un modelo de aprendizaje automático que detecte fracturas en radiografías de tórax.
    d. *Objetivos específicos:*
        i. Realizar un análisis exploratorio de los datos radiológicos para identificar variables relevantes.
        ii. Implementar técnicas de preprocesamiento de imágenes para mejorar la calidad de los datos de entrada.

iii. Evaluar modelos de clasificación y comparar su desempeño en la detección de fracturas.

Ejemplo (para un reto de Procesamiento de Lenguaje Natural):

1. *Situación problemática:* Las plataformas de redes sociales enfrentan el reto de identificar mensajes con lenguaje tóxico o discriminatorio que afectan la convivencia en línea.
2. *Problema científico:* ¿Se pueden detectar patrones lingüísticos que permitan a un modelo identificar automáticamente mensajes tóxicos en comunidades digitales?
3. *Objetivo general:* Desarrollar un modelo de PLN que clasifique mensajes según su nivel de toxicidad.
4. *Objetivos específicos:*

    iv. Analizar un corpus de texto para identificar características lingüísticas asociadas con mensajes tóxicos.

    v. Implementar técnicas de limpieza y normalización de texto para mejorar la calidad del dataset.

    vi. Evaluar diferentes modelos de clasificación y métricas de desempeño para determinar la mejor opción.

Ejemplo (para un reto de Negocios):

1. *Situación problemática:* Una empresa de comercio electrónico observa que muchos clientes compran una sola vez y no regresan, lo que reduce sus ingresos recurrentes.
2. *Problema científico:* ¿Es posible predecir qué clientes tienen mayor probabilidad de ser compradores recurrentes y diseñar estrategias de retención personalizadas?
3. *Objetivo general:* Construir un modelo de predicción que identifique clientes con alta probabilidad de recompra.
4. *Objetivos específicos:*

    vii. Realizar un análisis exploratorio de las variables de comportamiento de los clientes (frecuencia, monto, tipo de productos).

    viii. Aplicar técnicas de segmentación de clientes para identificar perfiles con mayor propensión a la recompra.

ix. Evaluar diferentes algoritmos de clasificación y recomendar estrategias de retención basadas en los hallazgos.

2. **Investigación preliminar:**
    a. Para problemas médicos: describir la enfermedad, síntomas y métodos de diagnóstico (con énfasis en diagnóstico por imágenes).
    b. Para problemas de PLN: investigar técnicas comunes para detección de patrones en texto.
    c. Para problemas de negocios: investigar estrategias de retención de clientes y optimización de ofertas.
3. **Análisis inicial del problema y los datos disponibles.**
4. **Preprocesamiento de datos:** describir y documentar las tareas de limpieza realizadas.
5. **Análisis exploratorio de datos (EDA):**
    a. Describir cuántas variables y observaciones hay en el dataset y sus tipos de datos.
    b. Resumir variables numéricas (medidas de tendencia central y dispersión).
    c. Elaborar tablas de frecuencia para variables categóricas.
    d. Realizar cruces entre variables clave para detectar patrones relevantes.
    e. Generar gráficos exploratorios (histogramas, diagramas de cajas, dispersión, barras, etc.) para facilitar la interpretación.
6. **Conclusiones:** redactar un resumen con los principales hallazgos y posibles implicaciones para etapas posteriores del proyecto.

EVALUACIÓN

NOTA: La evaluación de cada integrante del grupo será de acuerdo con sus contribuciones al trabajo grupal

- **(10 puntos) Situación Problemática**: Describe la situación problemática que origina el problema.
- **(10 puntos). Problema científico:** Enuncia adecuadamente el problema derivado de la situación planteada.
- **(10 puntos). Objetivos:** Plantea un objetivo general y al menos dos específicos, que sean medibles y alcanzables.
- **(20 puntos). Descripción de los datos:** Presenta una descripción clara de las variables, observaciones y operaciones de limpieza aplicadas.
- **(30 puntos). Análisis Exploratorio:**
    - Uso correcto de estadística descriptiva en variables cuantitativas.
    - Gráficos exploratorios adecuados.

- Identificación de datos atípicos, valores faltantes y correlaciones entre variables.
- Análisis de variables categóricas con tablas y gráficos apropiados.
- Explicación clara de procedimientos y hallazgos.

- **(20 puntos). Hallazgos y conclusiones:**
  - Resumen de hallazgos y sugerencias para fases posteriores.

RUBRICA

Criterio	Excelente (100%)	Bueno (75%)	Aceptable (50%)	Insuficiente (25% o menos)
Situación problemática (10 pts)	Contexto claramente descrito, con relevancia bien justificada.	Contexto descrito con algunos detalles faltantes.	Descripción general, poco precisa o sin justificación.	No se describe la situación problemática.
Problema científico (10 pts)	Pregunta de investigación clara, específica y bien formulada.	Pregunta adecuada pero poco precisa.	Pregunta poco clara o demasiado general.	No se formula el problema científico.
Objetivos (10 pts)	Objetivo general y ≥2 específicos claros, medibles y alcanzables.	Objetivo general claro, objetivos específicos poco precisos.	Objetivos vagos o difíciles de medir.	No se presentan objetivos.
Descripción de los datos (20 pts)	Variables, observaciones y limpieza descritas con detalle y claridad.	Buena descripción, pero faltan algunos aspectos.	Descripción incompleta o superficial.	No describe los datos.
Análisis Exploratorio (30 pts)	Uso completo de estadística descriptiva, gráficos variados y pertinentes, análisis de correlaciones, outliers y valores faltantes.	Análisis adecuado pero incompleto en algún aspecto (ej. gráficos o correlaciones).	Análisis limitado, solo con estadísticas básicas y pocos gráficos.	No se realiza análisis exploratorio.
Hallazgos y conclusiones (20	Resumen claro y profundo de hallazgos,	Resumen adecuado, pero sin suficiente	Conclusiones vagas o poco	No presenta

Criterio	Excelente (100%)	Bueno (75%)	Aceptable (50%)	Insuficiente (25% o menos)
pts)	con implicaciones bien justificadas.	profundidad.	relacionadas con el análisis.	conclusiones.

MATERIALES A ENTREGAR

- Informe en formato **PDF** con el análisis exploratorio.
- Link al repositorio de **GitHub** (y a Kaggle o Colab en caso de usarse).
- Presentación en **PowerPoint** con los resultados principales.

FECHAS DE ENTREGA

- **PRESENTACIÓN Y DOCUMENTO FINAL COMPLETO: 11 de septiembre de 2025**

REFERENCIAS

- https://www.kaggle.com/general/33266
- https://medium.com/analytics-vidhya/how-to-use-google-colab-with-github-via-google-drive-68efb23a42d