

Guía para de Proyecto 2. Resultados Iniciales

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

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. Haga una revisión bibliográfica en la que investigue qué algoritmos de aprendizaje automático se han usado para resolver problemas similares. Puede investigar sobre arquitecturas y modelos diferentes de algún algoritmo en específico. Use el formato APA para referenciar los artículos, de revistas indexadas, que fueron consultados.
2. En base a la investigación revisada, seleccione los algoritmos y/o modelos que usará para resolver su problema.
3. Construya varios modelos con los algoritmos que determinó que serían los más útiles. Si es necesario tunear parámetros hágalo.
4. Pruebe los modelos. Determine su eficiencia. Recuerde que debe usar métricas adecuadas dependiendo de si su problema es de clasificación o de regresión.
5. Discuta sobre los resultados obtenidos con la aplicación de los modelos. Lleve a cabo comparaciones que le permitan seleccionar el o los modelos que usará para resolver su problema.
6. Incluya al menos 3 visualizaciones estáticas que ilustren de una forma gráfica los resultados obtenidos.

EVALUACIÓN

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

- **(20 puntos) Investigación de algoritmos:** Investiga y consulta bibliografía sobre varios algoritmos de aprendizaje de máquinas que le pueden ser útiles para resolver el problema planteado. Se referencian fuentes bibliográficas indexadas.
- **(10 puntos). Selección de algoritmos a probar:** Se explican las razones por las cuales se seleccionaron los algoritmos.
- **(10 puntos). Modelos:** Se elaboran varios modelos con los algoritmos seleccionados.
- **(20 puntos). Eficiencia de los modelos:** Se muestra la eficiencia de los modelos.
- **(30 puntos). Discusión y Visualizaciones estáticas:**
 - Discute sobre el desempeño de los modelos y determina cual es el que mejor funciona para el problema a resolver.
 - Elabora visualizaciones estáticas que ayudan a explicar de forma gráfica los resultados obtenidos.
- **(10 puntos). Referencias y Bibliografía:**
 - Construye la bibliografía consultada, siguiendo las normas APA.
 - Pone referencias consultadas en el texto.
 - La bibliografía referenciada está indexada por lo que es confiable

MATERIAL A ENTREGAR

- Archivo .pdf con la investigación y discusión de los resultados de los modelos.
- Script de R (.r o .rmd) o de Python que utilizó para responder las preguntas con el código utilizado.
- Link de Google Drive donde trabajó el grupo el informe
- Link del repositorio usado para versionar el código.
- Presentación de Power Point a usar para presentar resultados.

FECHAS DE ENTREGA

- **PRESENTACIÓN Y DOCUMENTO FINAL: 23 de octubre de 2025 durante el período de clase**

REFERENCIAS

- <https://www.kaggle.com/general/33266>
- <https://medium.com/analytics-vidhya/how-to-use-google-colab-with-github-via-google-drive-68efb23a42d>
- Para manejar referencias puede usar Mendeley:
 - o <https://www.mendeley.com/>
- Puede consultar las fuentes de revistas indexadas que proveen los recursos digitales de la biblioteca:
 - o <https://bibliotecas.uvg.edu.gt/recursos/>
 - o (IEEE Computer Society) <https://www.computer.org/csdl/home>
 - Este recurso debe ser accedido solo desde la universidad.