***Resumen del proyecto***

¡Aplica Procesamiento del Lenguaje Natural, Sistemas de Recomendación y Series de Tiempo para resolver problemas de relevancia contemporánea!

La realización y entrega del Proyecto es individual.

**Entregables**

Un Notebook de Jupyter con la resolución de la consigna. El Notebook debe poder ejecutarse sin errores. En el notebook debe estar el link al repositorio (por ejemplo, de GitHub) donde se pueda encontrar el proyecto entregado.

**Referencias**

Apóyate en las bitácoras, los notebooks trabajados y las presentaciones vistas en clase para resolver tu proyecto. También será de mucha utilidad la documentación de las librerías de Python. No dudes en consultar comunidades online como Stack Overflow y, por supuesto, buscar en la web (*googlear).*

**Consigna**

Elige **una de las tres opciones** de aplicación para elaborar tu proyecto. El objetivo es que apliques las herramientas aprendidas en el dominio que hayas seleccionado. Verás que para cualquiera de las opciones, el trabajo se organiza en tres partes:

* **Parte A - Exploración de Datos.**Todo proyecto de Ciencia de Datos empieza con un Análisis Exploratorio de Datos. Y todo Análisis Exploratorio de Datos debe responder preguntas.
* **Parte B - Modelo de Machine Learning.**En esta sección deberás aplicar las técnicas de Machine Learning aprendidas para crear un modelo predictivo a partir del dataset provisto.
* **Parte C - Investigación.**Las preguntas y cosas para probar nunca se agotan. El objetivo de esta sección es que sugieras cómo continuarías el proyecto, con el fin de mejorar el modelo o responder una pregunta que consideres interesante. En todos los notebooks dejamos algunas sugerencias, pero puedes proponer otras.

Aquí debajo encontrarás el notebook con las consignas y el dataset para cada caso. Ten en cuenta que este proyecto se monta sobre los proyectos anteriores, y que ya esperamos cierto grado de independencia de parte tuya. Por eso, las consignas y recomendaciones de los notebooks no son exhaustivas. Es decir, si hay un paso en el flujo de trabajo que no está mencionado en el notebook que te dejamos, no implica que no haya que hacerlo.**Recuerda que para ver lo que sí tiene que estar sin falta en tu entregable debes consultar el Checklist.**

**Opción 1: Sistemas de recomendación**

*Implementa un Sistema de Recomendación para videojuegos de la plataforma*[*Steam*](https://store.steampowered.com/?l=spanish)*.*

* **Dataset.** En [este repositorio](https://github.com/kang205/SASRec) puedes encontrar el dataset, junto con información sobre sus creadores (¡a quienes les agradecemos fuertemente!). Descarga los dos archivos (*reviews* y *game\_info*) con anticipación. ¡Ten en cuenta que el primero ocupa bastante lugar!
* **Notebook.** [Aquí](https://colab.research.google.com/drive/1byeX_htRp3NqV08X3eLlWIp02cnPb_WP?usp=sharing) te dejamos un notebook con algunas recomendaciones y líneas de código para que empieces a explorar los datos.

**Opción 2: Procesamiento de Lenguaje Natural**

*Implementa un modelo para reconocer el puntaje asignado a un ítem de Amazon a partir de la crítica que hace un/a usuario/a.*

* **Dataset.** [Aquí](https://drive.google.com/uc?export=download&id=11XnXB7Ubgf3t6gotXGlM4FCwPOMHhDLX) puedes encontrar la descarga del dataset e información sobre el mismo [aquí](https://registry.opendata.aws/amazon-reviews-ml/). Es importante que tengas en cuenta la [licencia](https://docs.opendata.aws/amazon-reviews-ml/license.txt) de este dataset.
* **Notebook.** [Aquí](https://colab.research.google.com/drive/1oK2UQA57i6GpzRjqmW2wjOQ4l1MyEyia?usp=sharing) te dejamos un notebook con algunas recomendaciones para tu trabajo.

**Opción 3: Series de Tiempo**

*Implementa un modelo para predecir el flujo vehicular en una autopista de la Ciudad de Buenos Aires, Argentina.*

* **Dataset.** [Aquí](https://data.buenosaires.gob.ar/dataset/flujo-vehicular-por-unidades-peaje-ausa) puedes descargar el dataset. Deberás descargar - al menos para comenzar - los años 2017, 2018 y 2019. ¡Agradecemos a todos/as los que hacen Datos Abiertos!
* **Notebook.** [Aquí](https://colab.research.google.com/drive/1aQcWJmd6Unbtxoy6XOI5feSfo3ZkI1JO?usp=sharing) te dejamos un notebook con algunas recomendaciones y líneas de código para que empieces a explorar los datos.

**Checklist de evaluación**

Sugerencias para desarrollar el proyecto:

* La resolución del proyecto te puede enfrentar a desafíos que no trabajamos durante los encuentros. Es importante que desarrolles la capacidad de resolverlos. Para ello, consulta las bitácoras, las referencias, los notebooks vistos en clase, la documentación de la librería y, sobre todo, *googlea*.
* Los pasos deben estar correctamente justificados.
* Las preguntas que se respondan deben estar correctamente explicitadas.
* Imagina que este proyecto lo usarías para presentar en una entrevista de trabajo, o que lo debes presentar en tu trabajo. Presta mucha atención a la redacción, presentación de gráficos, etc.

Antes de subir tu proyecto a la plataforma Acámica para que sea evaluado, verifica que el Notebook se ejecute sin errores. Además, asegúrate de cumplir con las siguientes condiciones (son las que los/as evaluadores/as tendrán en consideración al momento de corregir tu trabajo):

**Parte A - Exploración de Datos**

1. El Análisis Exploratorio de Datos debe servir para comprender el dataset y todo el flujo de trabajo que le siga.
2. Debes responder al menos una pregunta original con este dataset. La pregunta debe estar correctamente explicitada.

**Parte B - Modelo de Machine Learning**

1. Debes evaluar correctamente el modelo que realices. Esto implica un correcto manejo de datos de Train y Test, elegir una métrica apropiada y justificar su elección, y comparar los resultados contra un modelo benchmark.
2. Puedes aplicar más de una de las técnicas vistas para crear tus modelos. Pero ten en cuenta que es preferible un modelo bien hecho (apropiada transformación de datos, optimización de hiperparámetros y análisis de sus resultados) que muchos modelos a medias. En caso de entrenar más de un modelo, debes comparar sus resultados y justificar cuál elegirías.
3. Si el modelo lo permite, debes explorar qué información utiliza para predecir e interpretar ese resultado. ¿Coincide con lo que esperabas a partir de tu experiencia con el dataset?

**Parte C - Investigación**

1. Debes explicar qué te gustaría probar, por qué y cómo lo harías. Si tienes referencias (por ejemplo, un artículo que hayas encontrado, capítulo de libro, etc.), debes mencionarlas.
2. Debes comentar también qué resultados esperas encontrar. Por ejemplo, puedes implementar una prueba rápida y mostrar resultados preliminares, para ver si estás correctamente orientado.