

Desafío 4 - Proyecto integrador

Introducción

En el proyecto integrador final, al igual que en el Desafío 3, pueden elegir con libertad con qué dataset trabajar, teniendo en cuenta los mismos requisitos planteados con anterioridad.

Escenario

Están trabajando como data scientists para una firma que se está expandiendo rápidamente. Para consolidar su posición como analistas en la compañía, deciden presentar un tópico innovador y poco habitual al directorio. Su propuesta tiene que constituir un problema que se pueda abordar con machine learning, y apuntar a conocer un sector desconocido o poco explotado hasta el momento. **Recuerden que tienen que mostrar cómo los datos aportan valor al problema que se plantearon.** Cualquier pregunta o problema vale, siempre y cuando esté bien fundamentada y encuadrada como algunos de los problemas de machine-learning que vimos en el curso (ya sea una regresión, una clasificación, o un problema de clustering).

Resumen del proyecto

Su trabajo consistirá en:

- Recolectar datos de su elección
- El dataset **debe** contener **al menos 1000** observaciones limpias para trabajar.
- Identificar los principales predictores de la variable objetivo. Hacer los preprocesamientos necesarios y realizar un análisis de datos de su base para comprender la relación entre variables.
- Proponer un conjunto de modelos acordes al problema.
- Testear, validar y describir los modelos generados.
 - ¿Cuáles son los factores que predicen la variable en estudio?
 - ¿Cuál es la *performance* del modelo?
 - ¿Cuál es el modelo que tuvo un mejor desempeño?
- **Realizar una pre-entrega el 17 de diciembre** que contenga los lineamientos generales del proyecto. Debe ser un documento de no más de 3 carillas que contenga una breve descripción de los siguientes puntos:
 - Tema de investigación
 - Antecedentes sobre el tema
 - Aporte esperado
 - Disponibilidad de datos e infraestructura
 - Plan de trabajo y Cronograma tentativo
- **Para el martes 15 de febrero, entregar un reporte técnico** para los responsables del área de Data Science de la empresa detallando los hallazgos (jupyter notebook). Aparte debe entregarse un archivo *pickle* que debe contener el mejor rendimiento.
- **Para el martes 15 de febrero, confeccionar una presentación** que se expondrá ante el

CEO de la empresa detallando los hallazgos. El reporte deberá tener un carácter no técnico.

Requisitos

- Se puede utilizar cualquiera de los modelos vistos en los módulos anteriores, pero sí o sí hay que evaluar los modelos vistos en el módulo 7 (modelos de ensambles).
- Se debe realizar un tuneo de hiperparámetros utilizando alguna de las herramientas y estrategias vistas en clase.
- Se debe generar por lo menos un pipeline que contenga alguno de los pasos de preprocesamiento del dataset elegido.
- Se debe realizar un análisis de la importancia de las features, por lo menos, del modelo que haya tenido el mejor desempeño.
- Se debe utilizar la librería pickle para serializar el modelo que haya tenido el mejor desempeño para luego ser entregado a la persona que se encarga de hacer la puesta en producción de los modelos.

Características del entregable técnico:

Jupyter Notebook que contiene el reporte técnico (código, análisis, visualizaciones, conclusiones). El mismo debe tener la forma de un reporte con los siguientes contenidos:

- una *introducción* en la que se plantea el problema.
- un apartado en el que describen sucintamente las *técnicas a utilizar y las características del/los dataset/s* utilizados.
- uno o más apartados en los que desarrolla el *análisis, visualizaciones, resultados* de los modelos, etc.
- un párrafo en el que se resumen los *principales hallazgos*, conclusiones y se realizan recomendaciones para los interesados (si corresponde).

Características de la presentación:

Una presentación (en formato ppt, pdf o similar) destinada a un público no técnico en la que se resume el problema, los métodos usados y los principales resultados.

Fecha de entrega

Los materiales deben ser entregados el **martes 15 de febrero** de 2022, antes de la clase de ese día.

Dataset

- Van a utilizar un dataset elegido por ustedes, sea tanto existente o construido a partir de los

datos recolectados de la web o consumiendo una API por ustedes mismos.

- Pueden buscar links a bases de datos en el github de Digital House:
 - https://github.com/Digital-House-DATA/ds_blend_2021_img/wiki/Public-datasets

¿Cómo empezar? Sugerencias:

En términos generales, recuerden las siguientes sugerencias:

- Escribir un pseudocódigo antes de empezar a codear. Suele ser muy útil para darle un esquema y una lógica generales al análisis.
- Leer la documentación de cualquier tecnología o herramienta de análisis que uses. A veces no hay tutoriales para todo y los documentos y las ayudas son fundamentales para entender el funcionamiento de las herramientas utilizadas.
- Documentar todos los pasos, transformaciones, comandos y análisis que realices.

Recursos útiles

- [Algunos consejos sobre cómo escribir para no especialistas](#)
- [Documentación de la librería Requests](#)
- [Documentación de la librería BeautifulSoup](#)