Desafío 3 - Construyendo un clasificador

Introducción

En este módulo vimos algunas técnicas de clasificación, como los algoritmos KNN, regresión logística y Naive Bayes (en sus diversas implementaciones). En este tercer desafío, vamos a poner a prueba la implementación de estos contenidos.

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 de clasificación y apuntar a conocer un sector desconocido o poco explotado hasta el momento.

Por ejemplo, podrían analizar distintas alternativas de problemas de clasificación, con el propósito de identificar si ciertos correos son *spam* o no, si un diagnóstico es benigno o maligno, etc., o cualquier otro problema que les interese. Cualquier pregunta o problema vale siempre y cuando esté bien fundamentada y encuadrada como un problema de clasificación.

Resumen del proyecto

En este desafío se pondrán en práctica habilidades importantes como la construcción de un clasificador binario (o multiclase) usando uno o varios algoritmos de clasificación.

Su trabajo consistirá en:

- Recolectar datos de su elección, en un dataset que **debe** contener **al menos 1000** observaciones limpias para trabajar.
- Identificar los principales predictores de la variable objetivo.
- Hacer los preprocesamientos necesarios.
- Testear, validar y describir los modelos generados.
    - ¿Cuáles son los factores que predicen la variable en estudio?
    - ¿Cuál es la *performance* del modelo?
- Escribir un reporte técnico para los responsables del área de Data Science de la empresa detallando el trabajo realizado y los hallazgos (Jupyter Notebook/Lab).
- Confeccionar una presentación para exponer ante el CEO de la empresa detallando los hallazgos. La presentación deberá tener un carácter no técnico.

Requisitos y Material a entregar

En este desafío hay dos entregables básicos:

1. Jupyter Notebook/Lab que contiene el reporte técnico (código, análisis, visualizaciones, conclusiones) destinado al área de Data Science de la empresa. El mismo **debe tener la forma de un reporte** orientado a una audiencia de pares, 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, preprocesamientos, 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).

2. Una exposición dirigida al CEO de la compañía, de no más de 15 minutos del trabajo realizado, consistente en una presentación acompañada con algunos slides no técnicos (ppt o Google Slides).

   La presentación debe constar de:

   - Una introducción (planteo del problema, la pregunta, el objetivo del trabajo)
   - Un desarrollo de los análisis realizados, exponiendo los métodos utilizados.
   - Una exposición de los principales resultados y conclusiones.

   Esta presentación al CEO se expondrá durante la clase del 26/11.

Fecha de entrega

Los materiales (notebooks y ppt) deben ser entregados el día **viernes 26/11/2021, antes de comenzar la clase**.

Dataset

- Van a utilizar un dataset elegido por ustedes.
- Pueden buscar bases de datos en:
  - https://data.buenosaires.gob.ar/
  - https://toolbox.google.com/datasetsearch
  - https://www.kaggle.com/
  - https://github.com/
  - https://registry.opendata.aws/
  - https://archive.ics.uci.edu/ml/index.php

¿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 usen. 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