**Desafío 4 - Construyendo un clasificador**

**Introducción**

En este módulo veremos temas relacionados técnicas de clasificación, incluyendo webscraping y regresión logística. Vamos a poner a prueba la implementación de algunos 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. El mismo tiene que constituir un problema de clasificación y apuntar a conocer un sector desconocido o poco aprovechado.

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/ problema vale siempre y cuando esté bien fundamentado y encuadrado como un problema de clasificación.  
  
**Resumen del proyecto**

En este desafío se pondrán en práctica habilidades importantes como la construcción de un clasificador binario usando uno o varios algoritmos de clasificación (KNN, regresión logística, etc.).  
  
Su trabajo consistirá en:

* Recolectar datos de tu elección (recuerda que puede ser el primer paso del desafío integrador!)
* El dataset DEBE contener AL MENOS 1000 entradas/filas
* Identificar los principales predictores de la variable objetivo; si la variable dependiente es continua o policotómica deberán dicotomizarla.
* 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 los hallazgos.
* Confeccionar una presentación para exponer ante el CEO de la empresa detallando los hallazgos. El reporte deberá tener un carácter no técnico.
* **BONUS** 
  + Armar un dataset a partir de datos scrapeados. Por ejemplo, se podría analizar la vinculación entre los salarios, los diferentes puestos de trabajo y las búsquedas de empleo en [USAJob](https://www.usajobs.gov/) o en [Bumeran](http://www.bumeran.com.ar/). También podrían tratar de construir clasificador sobre la probabilidad de que una película gane un Oscar en función de los actores y directores que participan ([DataBase Acadademy Awards](http://awardsdatabase.oscars.org/)).
  + Utilizar técnicas de regularización vistas (Ridge o LASSO), buscando sus hiperparámetros a través del método Gridsearch.

**Requisitos**

* Producir un reporte técnico en Jupyter Notebook con los análisis de regresión para una audiencia de pares (data scientists).
* Una presentación (no técnica) para el CEO de la empresa
* Tip: Hay un muy buen ejemplo de reporte no técnico (que podría servir como base para una presentación) [acá](https://www.dlsweb.rmit.edu.au/lsu/content/2_assessmenttasks/assess_tuts/reports_ll/report.pdf).

**Material a entregar**

En este desafío hay dos entregables básicos:

* 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)
  + Este entregable NO debe ser el notebook con el “código básico” que se presenta como orientador.
* 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 en la clase 34:

* Ma-Mi-Vi: 17 de octubre de 2017
* Ma-Mi-Sa: 19 de octubre de 2017
* Lu-Mi-Vi: 20 de octubre de 2017

**Código básico**

Se adjunta un notebook con algunos comentarios orientadores. El mismo tiene un carácter de apoyo y no constituye un entregable.

**Dataset**

* Van a utilizar un dataset elegido por ustedes, sea tanto existente o construido a partir de los datos recolectados de la web por ustedes mismos.
* Para extraer los datos pueden usar la librería BeatifulSoup.

**¿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](http://programmers.stackexchange.com/questions/11523/explaining-technical-things-to-non-technical-people)
* [Documentación de la librería BeatifulSoup](http://www.crummy.com/software/BeautifulSoup/)