PROYECTO DE FIN DE MÁSTER

ÁREA: MÁSTER EN ANALITICA WEB Y BIG DATA (DATA SCIENCE)

PROFESOR: ALEJANDRO PADILLA

INTRODUCCIÓN AL BRIEFING DEL PROYECTO:

El presente trabajo se basa en la información contenida en el site Birrapedia.com, donde se encuentra recolectada la información de más de 15 mil cervezas. Por lo que utilizaremos su contenido para obtener insights sobre los datos y poder inferencias a partir de los mismos.

A lo largo del trabajo se deberá pasar por diferentes fases de desarrollo, desde la adquisición. limpieza y transformación de datos, un análisis exploratorio de los datos (EDA por sus siglas en inglés), así como la generación de modelos de Machine Learning con su correspondiente análisis y comparación de resultados y finalmente la comunicación del viaje a través de los datos y sus insights.

Veremos que en el sitio web podremos encontrar la información de cada cerveza contenida en una URL en donde encontraremos (en la mayoría de los casos) la siguiente información sobre cada una de ellas.

Nombre	Tipo	Descripción
id	STRING	Identificador de la cerveza o producto en el sitio web.
sku	STRING	Identificador del producto en el mercado.
name	STRING	Nombre de la cerveza o producto.
type	STRING	Tipo de cerveza o producto.
alc	STRING	Graduación de alcohol contiene valores de 0 a 100°.
ibu	STRING	Unidades internacionales de amargor.
price	STRING	Precio de la cerveza o producto
currency	STRING	Moneda en la que está valuada la cerveza

price_estimate	FLOAT	Precio mediano de la cerveza o similares
rating_count	STRING	Conteo de reseñas de la cerveza o producto
rating_value	STRING	Valor promedio de las reseñas (valores entre 0 y 5)
rating_best	STRING	Valor máximo de las reseñas (valores entre 0 y 5)
section	STRING	Sección del sitio web donde está alojado el producto.
country	STRING	País de producción de la cerveza.
brewery	STRING	Nombre de la cervecería que la produce.
url	STRING	URL donde están alojados los datos de la cerveza en el site.
last_modified	STRING	Fecha de la última modificación de la cerveza o producto.

OBJETIVO DEL TRABAJO:

El objetivo del trabajo final de máster consistirá en utilizar lo aprendido a lo largo del mismo y aplicarlo a un caso práctico. Tu trabajo consistirá en enfrentarse al ciclo de desarrollo de un proyecto de Data Science iniciado por la extracción de datos desde un sitio web (si así lo decides), pasando a la limpieza y transformación de éstos para poder hacerlos de utilidad, generar un análisis descriptivo del comportamiento y distribución de los datos, de modo que seas capaz de evaluar la generación de atributos nuevos, ajustes un modelo a los datos y expongas el viaje realizado.

TRABAJO A REALIZAR:

El trabajo se compone de tres partes identificadas como:

1. Presentación con una extensión máxima de 30 diapositivas. Muy importante la conceptualización de la visualización y los mensajes aportados.
2. Documentación de soporte de la presentación en **formato PDF** construido a partir de un Notebook de Jupyter ocultando las celdas de código que no sean relevantes por lo que su extensión no deberá superar las 20 páginas. La estructura de dicho documento será definida por el alumno. Aquí estará recolectado la mayor parte del realizado.

3. Un fichero **ipynb** con el código completo del trabajo realizado. En el que sea capaz de ejecutar todas las celdas contenidas en el mismo.

A continuación, se detallarán los aspectos que se deberán abordar en el proyecto:

1. Extracción de datos usando web *scraping* (**Opcional** para obtener crédito extra que ayude a alcanzar el 100% de créditos)
   i. Se usará el paquete `requests` de python para la extracción del código HTML/XML del sitio web.
   ii. Así mismo el paquete `BeautifulSoup` para la extracción del contenido específico del punto anterior.
   iii. Dado que son más de 16 mil enlaces a consultar, se recomienda hacer una consulta cada segundo por enlace y almacenar los resultados en un fichero cada mil enlaces visitados. En caso de usar Google Colab habrá que montar drive para poder almacenar los datos en formato JSON o CSV.

```
# Habilita la conexión con Google Drive para almacenar y
extraer ficheros
from google.colab import drive
drive.mount('/content/gdrive')

# Almacena un diccionario o lista con diccionarios a un
fichero tipo JSON
with open('/content/gdrive/My Drive/TFM/birrapedia.json','a')
as fp:
      json.dump(beer_list,fp,ensure_ascii=False,indent=4)

# Almacena un DataFrame como fichero CSV
beers = pd.DataFrame(beer_list)
beers.to_csv('/content/gdrive/My Drive/TFM/birrapedia.csv')
```

   iv. Es recomendable usar las funciones `try`, `except` y `finally` para un mejor flujo en la extracción de datos y evitar hacer la extracción del mismo dato múltiples veces.
   v. El id de la cerveza se encuentra ubicado en la última sección de la URL, por lo que habrá que quedarse con ésa última sección para ello.
   vi. Algunas de las cervezas contienen precios de cervezas similares o la misma cerveza pero de distintas fuentes. Éste dato se encuentra almacenado en `<span`

`class='colorRojo'>` para cada una de las distintas fuentes por lo que habrá que quedarse con un precio medio o mediano de preferencia dado que alguna de las fuentes puede contener el precio de un pack y no de una única cerveza.

vii. Además, los precios vendrán en euros, lo que significa que vendrán con **coma** decimal y no con **punto** decimal, por lo que habrá que transformar el dato a un punto flotante que pueda operar Python.

viii. El resto de los datos se podrá conseguir buscando su etiqueta y clase o lista de clases.

2. Exploración, tratamiento y análisis de los datos.

   i. La intención es poder describir la distribución, comportamiento y relación entre las distintas variables.

   ii. Tratar de obtener insights relevantes a la información que se presenta. Por ejemplo: ¿qué tipo de cerveza es la más producida en España?, ¿qué países presentan similitud en preferencias de cerveza?, ¿qué país o países consumen un mayor nivel de alcohol en la cerveza?, etc.

   iii. Si se quisiera abrir una cervecería en un una determinada región, ¿qué tipo cerveza o cervezas serían la mejor opción para abrir mercado? ¿qué amargor le gusta más a la población de dicha región? ¿existe algún tipo de cerveza se que haya puesto de moda en el último año? ¿Cómo se ha ido aceptando o rechazando el nivel de amargor en la región en cuestión?

3. Generación de modelos de clasificación para determinar el tipo de cerveza según los datos recabados.

   i. Generar al menos 3 modelos distintos de ML para la clasificación del tipo de cerveza.

   ii. Comparar los distintos modelos y determinar cuál performa mejor.

   iii. Determinar cuáles son las variables más importantes para determinar el tipo de cerveza y con qué nivel de relevancia.

4. Conclusiones.

ACCESO A LOS DATOS:

Los datos vendrán del sitio web https://birrapedia.com/. Para ello el alumno tendrá que elegir una de las dos opciones siguientes:

a) Se realizará un *scraping* del sitio, realizando como primer parada el fichero robots.txt alojado en https://birrapedia.com/robots.txt, de donde obtendremos los enlaces a *scrapear* de cada cerveza.
b) Se podrá descargar un fichero CSV con los datos desde el enlace https://github.com/AlexPJ/SBS-AlejandroPadilla/blob/master/PFM2020/birrapedia.csv.

Aquellos que decidan obtener los datos a partir de *scraping* recibirán hasta un 15% de crédito de ayuda para alcanzar el crédito máximo de un 100%.

MATRIZ DE VALORACIÓN:

La calificación que obtendrá el alumno o grupo será producto de la calificación ponderada de los siguientes aspectos:

Contenido de la presentación:	50%
Documentación de soporte:	30%
Diseño de la presentación:	10%
Acciones propuestas realizables:	10%
AYUDA:	
Obtención de datos desde scraping	15% max.

Por contenido de la presentación se entiende como la profundidad del análisis abordado y el nivel de conocimientos desplegados para la realización del trabajo en el PowerPoint/Google presentaciones. Así mismo se valorará positivamente la capacidad de síntesis y resumen que demuestra un claro manejo de la materia por parte del alumno.

DISEÑO DE LA PRESENTACIÓN:

De igual forma que el contenido es relevante, muchas veces resulta imprescindible que el mensaje sea claro, que se entienda y que a través de las presentaciones quede claro lo que se quiere transmitir. Por ello resultará necesario que las diapositivas no se encuentren excesivamente cargadas, que sean claras, con colores agradables para que inviten a leerlas y que se cuide mucho la tipografía utilizada, así como los tamaños y cuidar todos los aspectos que son fundamentales de la presentación.

DOCUMENTACIÓN DE SOPORTE:

Junto con la presentación, se deberá entregar un **Notebook de Jupyter** de soporte que justifique la presentación elaborada. El índice y/o estructura será definido por el alumno o el grupo, aunque se valorará positivamente el orden y armonía del documento, así como la organización del contenido.

PLAZOS Y MÉTODOS DE ENTREGA:

La fecha límite de entrega será:

- **22 de febrero 2019 en Primera Convocatoria**
- **15 de marzo de 2019 en Segunda Convocatoria** hasta las 23:59hs.

Pasada esa fecha no se aceptará ningún trabajo sin excepción alguna. El formato de entrega de los documentos será a través de un enlace en Drive, Dropbox o Wetransfer, y que deberá contener en un fichero WINZIP o WINRAR todos los documentos. La dirección de correo electrónico a la que se deberá enviar el trabajo será: alejandro.padilla@bmind.es.

Así mismo el alumno será responsable de haber verificado previamente que el enlace efectivamente funciona y que se encuentra la información que se quiere presentar. El asunto del correo deberá seguir la siguiente estructura:

Apellidos, NOMBRE – TFM SBS