1. Objetivo del laboratorio

*Desarrollar de forma autónoma **tres Notebooks** que permitan explicar distintas hipótesis a partir de varios datasets de entrada, mediante la preparación y visualización de estos.*

2. Elementos a utilizar:

- Lenguaje Python
- Librería numérica *NumPy*, pandas, scikit-learn y gráfica *Matplotlib*
- Entorno Anaconda
- Editor Jupyter

3. Práctica 1 (AirBnB es un indicar de cómo se transforma la ciudad)

Objetivo

El problema del alquiler de la vivienda ya forma parte del contexto de la mayoría de grandes ciudades del mundo. Hay varios factores que se enumeran como parte del problema: el turismo masivo, la distribución de las ciudades o el auge de los apartamentos turísticos como es el caso de Airbnb. Lo que si queda claro es que los datos obtenidos del portal sirven para obtener una radiografía más o menos valida de la ciudad. A través del dataset proporcionado en Moodle en el que se miden las estancias en los últimos años en la ciudad de New York. Hay varias preguntas que nos hacemos.

1.- (1 punto) El barrio de Williamsburg es uno de los barrios de moda hoy en día. Desde 2005 ha pasado de ser un distrito donde principalmente se encontraban fabricas a acoger una gran cantidad de gente joven y nuevos negocios. ¿Teniendo en cuenta los datos proporcionados que posición en términos de popularidad/calidad se puede decir que ostenta el barrio con respecto a los demás que forma New York?

Lo primero que tendremos que hacer es cargar todo el archivo csv en un DataFrame para poder manipularlos. Habrá que comprobar si existen datos redundantes o anómalos.

Para medir la popularidad/calidad del barrio, tendremos primero en cuenta el número de apartamentos de calidad según los usuarios. Para ello, lo primero será clasificar los apartamentos como “Muy Malos”, “Malos”, “Regulares”, “Buenos” y “Muy Buenos”. Para ello primero haremos un par de transformaciones de los datos. La columna de reseñas por mes, se dividirán en 3 rangos iguales, después dependiendo del valor concreto que tenga cada apartamento se le asignarán las etiquetas: “Baja”, “Media” y ”Alta”. Para la columna que indica el número de días que está disponible asignaremos las etiquetas de menor valor a mayor de la siguiente manera: “Poco disponible”, “Disponible normalmente” y “Altamente disponible”. Finalmente crearemos una columna para clasificar los apartamentos, teniendo en cuenta las siguientes reglas:

- Si Reseñas es Baja y Disponibilidad es “Poco disponible” -> “Bueno”
- Si Reseñas es Media y Disponibilidad es “Poco disponible” -> “Bueno”
- Si Reseñas es Alta y Disponibilidad es “Poco disponible” -> “Muy Bueno”
- Si Reseñas es Baja y Disponibilidad es “Disponible normalmente” -> “Regular”
- Si Reseñas es Media y Disponibilidad es “Disponible normalmente” -> “Regular”
- Si Reseñas es Alta y Disponibilidad es “Disponible normalmente” -> “Bueno”
- Si Reseñas es Baja y Disponibilidad es “Altamente disponible” -> “Malo”
- Si Reseñas es Media y Disponibilidad es “Altamente disponible” -> “Malo”
- Si Reseñas es alta y Disponibilidad es “Altamente disponible” -> “Muy malo”

2.- (0,5 puntos) Explica visualmente como se distribuyen los 5 barrios más populares (esto serán aquellos cuyos apartamentos estén más solicitados en números absolutos). Para ello utiliza un diagrama de burbuja donde el eje de las X indica el barrio y el eje Y la calidad de los apartamentos. Usa todos los diagramas que necesites para llegar a esta conclusión final.

3.- (1 punto) Una vez obtenida la información del apartado anterior, queremos tratar de entender las diferencias de precio entre alquilar un apartamento entero y una habitación privada. Usa los diagramas de cajas donde cada uno de los 5 barrios más populares está representado por una caja y haz una interpretación de los resultados.

4.- (1 punto) Por último, se intuye que la tendencia a dejar reseñas en las apps que prestan servicios, ha aumentado en los últimos años. Decide que diagrama es más útil para este caso. Dibújalo y realiza una interpretación del mismo. Tomaremos la fecha de la última reseña como el dato útil para realizar dicho caso.

4. Práctica 2 (el perfil de los clientes de un banco)

Objetivo

La sucursal del Banco Santander situada en el campus de la Universidad Francisco de Vitoria nos solicita hacer varios estudios de sus clientes. Para ello haremos uso de un dataset proporcionado por la propia sucursal que se puede encontrar en Moodle.

1.- (1 punto) Para preparar los datos vamos a crear un DataFrame que luego guardaremos en un csv donde se almacenará sólo la información necesaria. Esta es: age (edad), education (nivel de estudios), balance (saldo) y duration (días como cliente). Explica si el nivel de estudios está directamente relacionado con el saldo de cada cliente en la cuenta. Para ello estableceremos 3 rangos numéricos: los que tienen deudas, la gente que tiene unos ahorros normales (en positivo, pero menos de 10.000 euros) y los que disponen de suficientes ahorros como para ofrecerles paquetes de inversión (aquellos en positivo con más de 10.000 euros). ¿Qué diagrama has usado y por qué? ¿Cuál es el grupo que más destaca? Aporta toda la información que creas necesaria que puedes obtener de la gráfica.

2.- (0,5 punto) Otro dato interesante sería conocer cuál ha sido el grupo de edad que tiene más clientes. Así sabremos si los alumnos usan la cuenta creada a través de la Universidad o no. Para ello haz una transformación donde los alumnos se considerarán personas de menos de 30 años. Trabajadores jóvenes serán personas de 30 a 45 años, de 46 a 65 años serán trabajadores veteranos y el resto serán clientes no vinculados a la Universidad. Obten un gráfico donde podamos ver como se distribuye cada clase. Interpreta los datos.

3.- (1 punto) Por último queremos saber cómo se distribuyen y cuáles son las frecuencias respecto a los días que el cliente lleva en la empresa. Transforma los datos de manera que pasemos de días a años (tendremos decimales) y busca la representación más útil (sólo una). ¿Qué nos dicen los datos?

5. Práctica 3 (Principal Component Analysis)

Objetivo

Existen casos en que las variables no se pueden representar visualmente debido a que necesitaríamos varias dimensiones para ello. Para evitar esto, existe una metodología en la cual, un set de datos multidimensional, podemos transformarlo para poder explicar gran parte de la información en 2 o 3 dimensiones. Dicha metodología se conoce con el nombre de Principal Component Analysis (PCA). Vamos a aplicarlo a un set de datos que está colgado en Moodle y vamos a dar una serie de explicaciones de que ocurre.

1.- (0,5 puntos) Lo primero que habrá que hacer será estandarizar los datos para que las diferencias de rango no supongan un problema a la hora de procesar la información. Usa para ello el método StandardScaler de la librería scikit-learn.

2.- (2 puntos) El segundo paso será a partir de los datos anteriores, obtener los autovalores (eigenvalues) y los autovectores (eigenvectors) que nos permitan explicar cuántos componentes necesitamos para representar los datos iniciales. Para ello primer habrá que obtener la matriz de covarianza mediante el método cov de numpy y después aplicarle a dicha matriz el método linalg.eig también de numpy. Obten un DataFrame con el porcentaje de varianza y el acumulado por cada componente. Explica que quieren decir estos datos. ¿Cuánto información perdemos con 2 componentes? ¿Cuánta información representamos con 3 componentes?

3.- (1 punto) Por último queremos representar gráficamente los individuos de nuestro dataset, pero usando los valores de las componentes principales obtenidas. Obtén un diagrama de dispersión en 2 dimensiones y comenta que has interpretado en él. Es necesario que el diagrama contenga toda la información necesaria. Habrá que interpreta que información proporciona el eje X y el eje Y. Por último, elegir al menos 4 individuos y explicar qué pasa con ellos.

4.- (0,5 puntos) Realiza los mismos pasos que en los pasos anteriores usando la librería scikit-learn. Compara los resultados y coméntalos.

6. Forma de entrega del laboratorio:

La entrega consistirá en un fichero comprimido RAR con nombre LAB01-GRUPOxx.RAR subido a la tarea **LAB1** que **contenga únicamente**

1. **Por cada práctica** un notebook de Jupyter (archivos con extensión **.ipynb**).
2. La **memoria del laboratorio** que se irá construyendo en el Notebook de manera que se explique todo lo que se hace.

Las entregas que no se ajusten exactamente a esta norma NO SERÁN EVALUADAS.

7. Rúbrica de la Práctica:

1. IMPLEMENTACIÓN: Multiplica la nota del trabajo por 0/1

Siendo una práctica de Data Mining, todos los aspectos de programación se dan por supuesto. La implementación será:

- Original: Código fuente no copiado de internet. Grupos con igual código fuente serán suspendidos
- Correcta: El programa funciona y ejecuta correctamente todo lo planteado en los apartados de cada práctica.
- Comentada: Inclusión (**obligatoria**) de comentarios.
- En las gráficas que se realicen proporciona todos los datos que creas necesarios.

2. MEMORIA DEL LABORATORIO

Obligatorio redacción clara y correcta ortográfica/gramaticalmente. Cada paso que se haga tiene que estar justificado.