

Clase 6: Intro Data Science & ML

Teoría DS

1. ¿Cómo definiría machine learning?
2. Nombrar 3 situaciones que se pueden resolver con Data Science y Machine Learning.
3. ¿Qué es un set de entrenamiento etiquetado?
4. ¿Qué es el entrenamiento incremental? Nombre 2 ventajas.
5. ¿Cuales con los dos tipos de tareas más comunes donde se aplica aprendizaje supervisado?
6. ¿Qué tipo de algoritmo se basan en una medida de similaridad para hacer predicciones?
7. Nombre 3 tipos de tarea donde se aplique aprendizaje no supervisado.
8. ¿Qué tipo de algoritmo utilizaría para entrenar un robot a caminar en varios tipos de superficies?
9. ¿Qué tipo de algoritmo utilizaría para segmentar clientes en múltiples grupos? ¿Por qué?
10. Dar un ejemplo donde aplicar:
 - a. Clasificación
 - b. Regresión
 - c. Clusterización
 - d. Reglas de asociación
11. Si un modelo tiene buena performance en el set de entrenamiento, pero baja frente a nuevas observaciones o instancias. ¿Que está sucediendo? Nombre dos soluciones posibles.

Python & Pandas

1. Crear un *dataframe* con los datos del archivo TSV orders.tsv
2. Mostrar las primeras 5 observaciones
3. Mostrar las últimas 10 observaciones
4. ¿Cuántas observaciones contiene?
5. ¿Cuántos atributos contiene?
6. Listar el nombre de cada atributo
7. Obtener el nombre del segundo atributo
8. Listar el tipo de datos de cada columna
9. Obtener el tipo de datos solamente del atributo *item_price*
10. Listar los valores en *items_name*
11. Listar catálogo de ítems vendidos
12. ¿Cuántos ítems distintos se vendieron?
13. ¿Cuál es la fila (observación) en la posición 190? (Mostrar como una serie de pandas)
14. Obtener el precio del ítem en la posición 210
15. ¿Cuál es el ítem que más cantidad de veces aparece en el dataSet?
16. ¿Cuántos ítems se compraron en total?
17. ¿Cuál es el ítem más vendido?
18. Convertir el precio de los ítems de string un float realizando los pasos necesarios

19. Confirmar que la transformación se realizó correctamente consultando el tipo de datos de *item_price* y los primeros registros.
20. Obtener estadísticas generales (cantidad de registro, máximos, promedios, etc) de las observaciones para los atributos numéricos
21. Obtener estadísticas generales de las observaciones para *item_name*
22. Obtener estadísticas generales de las observaciones para todos los atributos
23. Obtener el precio promedio redondeado a dos decimales
24. ¿Cuánto fue la ganancia bruta total?
25. Agregar una columna que por cada combinación de *item-orden*, muestre la ganancia bruta de cada ítem en dicha orden.
26. ¿Cuántas órdenes se realizaron?
27. ¿Cuanto es el la cantidad promedio de ítems de una orden?
28. Imprimir el listado de ítems con precio superior a 10.
29. Generar el catalogo de items, es decir un nuevo *dataframe* con el nombre de cada ítem y su precio, ordenados por precio, y luego por nombre descendente
Hint: investigen la funcion drop_duplicates.
30. Defina *item_name* como índice del dataset
31. Seleccione del catálogo los ítems Steak Soft Tacos, Chicken Bowl y Veggie Burrito.
32. Seleccione los ítems entre el rango definido por la posición 3 y 7.
33. Mediante *iloc* seleccione los primeros 5 registros del catálogo
34. Mediante *iloc* seleccione del catálogo todos los registros a partir del número 40
35. Seleccione los últimos 5 registros del catálogo
36. Seleccione todos los registros a partir de la posición 25 sin incluir los últimos 15.

Filtros:

37. ¿Cuántas pedidos tuvieron Veggie Salad Bowl?
38. ¿Cuántas personas ordenaron más de una Canned Soda?
39. Obtener el listado de productos que su precio sea menor a 1 o mayor a 11
40. De cada producto, ¿Cuál es el máximo, mínimo y promedio vendidos? Ordenar la lista por nombre de producto.