Ejercicios Repaso para Examen Final

Teoría

1. ¿Cuál es la finalidad de una normalización de datos?
2. Indique 5 ejemplos de la aplicación de KNN
3. Indique 5 ejemplos de la aplicación de KMeans
4. ¿Describa 2 Ejemplos donde se aplique una Regla de Asociación?
5. ¿Cuáles son las diferencias entre una Regresión Lineal y Polinomial?
6. ¿Qué casos se utiliza como métrica de similaridad una distancia Jackard y coseno?
7. ¿Cuál es la finalidad de una Matriz de Confusión?
8. ¿Dentro del Aprendizaje Automático, a que se denomina fiting?
9. ¿En qué caso aplicaría SVM?
10. ¿A qué se denomina Falso Positivo y Falso Negativo?
11. ¿Para qué se usa una matriz de correlación?

Implementación

1. Caso de Estudio: Deserción Estudiantil

Usted ha sido llamado para hacer un análisis de deserción estudiantil en alumnos de los primeros ciclos de Ingeniería por tanto se solicita:

- Proponga un dataset indicando los atributos y tipos de datos necesarios.
- Indique 5 preguntas que se resolverían con Reglas de Asociación, Kmeans, Árbol de Decisión, Regresión Polinomial, SVM.
- Ejemplifique un caso de matriz de confusión.

2. Caso de Estudio "Análisis de Datos de Netflix"

Introducción:

Netflix tiene 86,74 millones de clientes en todo el mundo de streaming. Tener esta gran base de usuarios permite a Netflix reunir una enorme cantidad de datos y poder tomar mejores decisiones haciendo que los usuarios estén cada vez más satisfechos con su servicio.

Presentamos un listado de captura de datos basado en eventos de los usuarios de Netflix:

- Al hacer una pausa, rebobinar o avanzar.
- La fecha y hora de visualización del contenido
- Frecuencia de visualización de contenido, por ejemplo, Netflix ha encontrado que los usuarios ven series de televisión durante la semana y películas durante el fin de semana.
- Localización de acceso (dirección IP)
- Tipo de dispositivo utilizado para acceder al contenido
- Las calificaciones dadas (alrededor de 4 millones por día)
- Búsquedas de contenido (alrededor de 3 millones por día)

- Comportamiento de navegación y desplazamiento de usuario
- Netflix también examina los datos dentro de las películas. Tomando varias "capturas de pantalla" en tiempo real para determinar "en el momento" características especiales sobre video que pueden ser: volumen, colores, y el paisaje que ayudan a Netflix averiguar los gustos y preferencias de los usuarios.
- De esto Netflix cuenta con algoritmos de personalización que tienen como objetivo predecir con exactitud lo que los usuarios van a ver a continuación por lo que tiene capacidades de realizar recomendaciones de películas.

A través de su análisis, Netflix puede saber cuánto contenido los usuarios necesitan para ver con el fin de tener menos probabilidades de cancelar. Por ejemplo, tal vez ellos saben "Si podemos conseguir que cada usuario vea por lo menos 15 horas de contenido cada mes, es 75% menos probable que cancelen. Si caen por debajo de 5 horas, hay una probabilidad del 95% de que cancelarán ".

Fuente: Recorte Kissmetrics.com

Se solicita:

1. **Generación de Datos**

   (4 puntos) Implementar según formato adjunto una función para generar de 100 a 200 registros de accesos que representa el fin de cada visualización (no necesariamente película concluida).

	Ejemplo
Autogenerado	725975413462163456
Fecha	2016-11-28
Hora Actual	09:50
IP	190.50.17.8
Hora Inicio	09:10
Hora Fin	09:50
Nombre usuario	HugoRamirez
Correo	HugoRamirez@upc.edu.pe
Plan	Básico
Película	Vengador Anónimo
Genero	Action, Adventure
Duración de película	90 min
Tipo Usuario	Premium
País	PE
Idioma de Película	En
Tipo de Dispositivo	Tablet

Calificación (1-5)	4

- Para la elección de películas, hacer uso del archivo “movies.csv”

2. Serialización de Datos

   (2 Puntos) Insertar en un solo Archivo **“Datasets.csv”** los registros generados.

3. Manipulación de Datos

   (2 Puntos c/u) De los datos insertados, realizar los siguientes reportes, hacer uso de funciones

   a. Cantidad de usuarios activos, con parámetros (Fecha inicio, Fecha Fin, Tipo Usuario)
   b. Top 100 Películas más vistas por continentes
   c. Promedio de géneros más vistos por países
   d. Número de películas no concluidas clasificado por Idioma. (Hora Fin- Hora Inicio < Duración de Película)
      - Por ejemplo:
        9:50-9:10= 40min, es menor a la duración 90

   e. Cantidad de accesos por tipo de Dispositivo (TV, Tablet, Desktop.
   f. Promedio acceso por Tipo de Plan (Básico, Estándar, Premium)
   g. Top 5 categorías de películas con mayor cantidad de visualizaciones.

3. Caso de Estudio: Mesa Redonda

Descripción:

La siguiente semana es la conferencia anual de Data Science y asistirán 8 expositores de distintas marcas de software. Las veces pasadas que se ha organizado la conferencia ha habido mucha tensión entre expositores, discusiones y todo tipo de conflictos. Este año, se ha determinado que se utilizará un algoritmo para ubicar a los expositores de manera que se reduzcan los conflictos. Para esto, se digitalizó la información de las conferencias pasadas, se determinó un indicador de conflictos (IDC) en función a la frecuencia y gravedad de las discusiones entre expositores, y se recopiló los datos en un archivo expositores.txt y tiene el siguiente formato:

El IDC de Microsoft aumenta en 95 cuando se sienta junto a IBM.
El IDC de Microsoft disminuye en 36 cuando se sienta junto a WEKA.
El IDC de Microsoft disminuye en 40 cuando se sienta junto a Teradata.
El IDC de Microsoft aumenta en 3 cuando se sienta junto a Python.
El IDC de Microsoft aumenta en 3 cuando se sienta junto a R.

El IDC de Microsoft disminuye en 18 cuando se sienta junto a Scala.
El IDC de Microsoft disminuye en 39 cuando se sienta junto a RapidMiner.

Idea Extraída de Tasa Challenge.

Esto se puede representar de la siguiente manera:

	1	2	3	4	5	6	7	8
1	0	95	36	40	3	3	18	39
2	..	0	..	..	..	..	..	..
3	..	..	0	..	..	..	..	..
4	..	..	..	0	..	..	..	..
5	..	..	..	..	0	..	..	..
6	..	..	..	..	..	0	..	..
7	..	..	..	..	..	..	0	..
8	..	..	..	..	..	..	..	0

Por lo tanto, necesitamos recrear estos escenarios de tal manera que se pueda determinar la ubicación óptima de los expositores, asumiendo que estarán sentados en una mesa redonda, así como el valor minimizado del IDC que fue hallado.

Se solicita:

- **(2 puntos)** Leer el archivo, pues la matriz se forma desde el procesamiento de texto, por ejemplo:

El IDC de Microsoft aumenta en 95 cuando se sienta junto a IBM ======> 1, 95,2

El IDC de Microsoft disminuye en 36 cuando se sienta junto a Weka ======> 1,-36,3

	1	2	3
1	0	95	-36
2	95	0	...
3	-36	...	0

- **(2 puntos)** Calcular cual es el representante más y menos conflictivo.
- **(4 puntos)** Implementar una función que devuelva el orden de los representantes donde se pueda minimizar los conflictos, es posible que tenga múltiples soluciones.
- **(2 puntos)** Validar su algoritmo de optimización generando 10 matrices de 8x8 con IDC aleatorios entre -100 y 100, tome en cuenta que el valor fila-columna tiene que ser igual a columna-fila, además que fila=columna es 0. Cada matriz guardar en un archivo tipo csv, con nombre “Dataset”<nro>”.csv”

4. Caso de Estudio: Venta de Juegos de Consola

Descripción

Ha sido contratado para formar parte del equipo interno de Data Science de una empresa de diseño de videojuegos. La empresa diseña juegos para computadoras, pero está considerando entrar en el negocio de los juegos de consola.

Nota: las consolas se refieren a los dispositivos conectados a televisores. Por ejemplo, Play Station. Sin embargo, los ejecutivos de la empresa han notado que otras empresas rivales de diseño de juegos de consola han sufrido pérdidas crecientes en los últimos dos años. Es por eso por lo que quieren que usted investigue el estado de la industria para ayudarles a tomar la decisión de entrar en este negocio o no.

En el dataset adjunto al caso de estudio contiene más de 16.000 juegos de consola vendidos entre 1980 y 2015. Las ventas se desglosan en 4 regiones y se muestran en millones de dólares.

Dataset:

Puesto
Nombre
Plataforma
Año
Género
Compañía
Ventas Norte America
Ventas Europa
Ventas Japon
Otras_Ventas

Se plantea las siguientes preguntas:

1. **(2 puntos)** ¿Cómo han fluctuado los totalizados de las ventas de juegos de consola en los diversos géneros a lo largo de los años?

   Tip: Usar una línea de tiempo

2. **(2 puntos)** Compare las distintas plataformas en términos de ventas acumuladas a lo largo del tiempo, ¿Quién es el líder?
   Tip: intente usar un gráfico de barra y retire cualquier plataforma con ventas insignificantes para ahorrar espacio visual. Agregue detalles colocando género en color.
3. **(2 puntos)** Compare las diferentes compañías entre sí en términos de ventas globales acumuladas desde su inicio ¿Quién es el líder?

   Tenga en cuenta que en este conjunto de datos el número de Registros NO es el número total de unidades vendidas. Puesto que cada fila representa un título único (nombre del

juego de consola), entonces el número de registros representa el número total de títulos UNICOS vendidos. Trate de usar esta información para colorear su mapa de árbol.

4. **(2 puntos)** ¿Cómo se comparan las consolas de nueva generación en términos de ventas globales totales para 2014 y 2015? Las nuevas plataformas en este conjunto de datos son PS4, XOne y WiiU.
   Tip: Usar un gráfico pastel.
5. **(2 puntos)** ¿Cuáles son los 10 mejores juegos con las mayores ventas globales?
   Tip: usar un gráfico de barras.

5. Caso de Estudio: Predicción de precios de venta de viviendas

Fuente: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Descripción:

Usted es un científico de datos que ayuda a la industria inmobiliaria a predecir los precios de las viviendas. Durante este estudio de caso, usted se familiarizará con los Algoritmos Básicos de Machine Learning: Regresión Lineal y Algoritmo de Clasificación.

Su análisis debe ser capaz de responder a las siguientes solicitudes:

1- **(2 puntos) Prepare** su conjunto de datos (sin valores faltantes, con el tipo de datos correcto) para hacer predicción de precios de vivienda
2- **(1 punto)** Realice un **muestreo** del 75% para el entrenamiento, 15% para la validación y 5% para la prueba.
3- **(2.5 puntos)** Realice la **Regresión Lineal**, crear su propia función e interprete los resultados.
4- **(2.5 puntos)** Realice un **Árbol de Clasificación**, elija las variables e interprete los resultados.
5- **(1 punto)** Implemente una función para el cálculo de la **Distancia Euclidiana** de un punto determinado.
6- **(1 punto)** Implemente una función para el cálculo de la **Distancia Manhatan** de un punto determinado.