ITESO

DEPARTAMENTO DE MATEMÁTICAS Y FÍSICA

CIENCIA DE DATOS E INTELIGENCIA DE NEGOCIOS

PROYECTO DE APLICACIÓN: MANEJO DE DATOS, SIMILITUD y CLUSTERING.

ESTE PROYECTO TIENE COMO FINALIDAD LA EVALUACIÓN DE LOS CONOCIMIENTOS ADQUIRIDOS DE MANEJO DE DATOS, ANALISIS EXPLORATORIO DE DATOS, INDICES DE SIMILITUD Y ALGORITMOS DE CLUSTERING.

INTRODUCCION:

Actualmente, con los avances en las tecnologías de la información, la generación de datos de diversos tipos en un solo día tiene volúmenes muy altos y con tendencia creciente. Con el fin del aprovechamiento de la información valiosa que pueda estar oculta en los datos que se generan, se requieren de tener conocimientos básicos de manejo de información y de análisis exploratorio de datos.

De forma general, a menos que la persona sea una experta en el fenómeno en el cual se están generando los datos, el ingeniero que se disponga al análisis de los datos generados debe de realizar un análisis exploratorio para rescatar las características básicas que poseen los datos. Además de realizar un agrupamiento de los mismos datos en base a una característica de interés.

OBJETIVO:

El objetivo de este proyecto de aplicación se puede separar en tres fases:

1.- La limpieza y extracción de la información estadística básica que tienen los datos que se están analizando.

2.- Realización de un agrupamiento (“Clustering”) de los datos en base a una característica de interés.

3.- La obtención o formulación de conclusiones sobre el fenómeno del cual provienen los datos en base a los resultados de los análisis anteriores.

Para la realización de la práctica se contemplan las siguientes actividades:

1.- Obtención de una base de datos que fuera generada por un fenómeno de interés (La orientación o tema de los datos será especificada para cada equipo por el profesor).

2.- Aplicar el estudio de calidad de los datos para determinar el tipo de datos, categorías e información estadística básica.

3.- Realizar una limpieza de datos y obtener un análisis exploratorio de datos (EDA) donde se muestren gráficas y conclusiones acerca del análisis. Al menos obtener 5 insights.

3.- En base al estudio anterior, realizar un análisis de similitud entre variables y entre muestras disponibles en su base de datos.

4.- Crear agrupamientos o "clusters" basados en el algoritmo "hierarchical clustering" ó "Kmeans", y presentar sus resultados (si los datos lo permiten).

6.- Basados en los análisis anteriores, formular conclusiones sobre la información importante que se haya logrado encontrar de los datos analizados.

NOTA 1: La asignación de las bases de datos por equipo se encontrará en el archivo 'Equipos_Proyecto.xlsx'.

NOTA 2: Cada equipo debe de verificar la base de datos que le corresponde y si hubiera algún inconveniente **(no se puede descargar, son muy pocos datos, son muchos datos, la base de datos está dañada, etc.)**, se debe notificar con el profesor para realizar un cambio de la base de datos.

NOTA 3: A todas las bases de datos es posible aplicar algoritmos de clustering, pero depende de la naturaleza de los datos o el contexto del problema, los resultados obtenidos son válidos o no.

Entregables:

- **Reporte** que valide la realización de las actividades anteriores.
- **Presentación** del proyecto donde se explique el desarrollo hecho para lograr los resultados obtenidos.
- **Código** del desarrollo del proyecto (se deberá enviar un archivo comprimido al link abierto en la plataforma de CANVAS, el archivo comprimido deberá incluir la carpeta de "Code" donde deberá estar el código del proyecto y "Data" donde se deberá tener el dataset asignado al equipo).

Especificaciones de evaluación del reporte:

- El reporte del trabajo realizado se entregará en digital en un documento elaborado en computadora (no fotografías de hojas ó cuadernos), e incluirá como mínimo:
  - ✓ Nombre y apellidos de alumno de los integrantes del equipo.

- ✓ Es obligatorio que incluyan el código que generaron para la realización para este proyecto (si adjuntan los archivos .py del código no es necesario que el código aparezca en su reporte; es suficiente que lo mencionen en alguna parte de su reporte y que indiquen el orden o forma hicieron la implementación de su algoritmo). La recomendación es que agreguen los códigos como apéndices en su reporte.
- ✓ Es necesario que exista un capítulo que describa la solución que se propuso para el problema.
- ✓ Es necesario adjuntar también los archivos de datos históricos que hayan utilizado.
- ✓ Por la asignación de datos a cada equipo, se espera que se obtengan resultados distintos en cada reporte.
- ✓ Incluyan figuras para visualizar los resultados.
- ✓ **NO** es necesario imprimir el reporte a color, es posible identificar las figuras por el tipo de línea que se utiliza.
- ✓ Es obligatorio tener una sección de conclusiones.
- ✓ Se realizará una revisión de las faltas de ortografía.