Ciencia de Datos e Inteligencia de Negocios

Tarea 3. Medidas de Similitud

Habiendo hecho un estudio preliminar de la calidad de los datos. Normalmente conviene definir el tipo de dato de cada variable. La clasificación de los datos, normalmente se hace en las siguientes categorías: Variables Binarias o de doble estado, variables cualitativas o multi estado, variables cuantitativas y variables genéticas.

En clase se realizó un resumen de las medidas de similitud más usadas en el análisis de datos. Enseguida se muestran dos tablas de medidas o índices de similitud, que han sido propuestas para cada tipo de dato.

La realización de esta tarea considera la elaboración de un reporte de las siguientes actividades:

1. Del archivo con información de la contaminación de zona metropolitana, se tomarán las mediciones de al menos 4 estaciones que miden la calidad del aire. Cada estación mide la cantidad de las siguientes sustancias: “CO”, “NO2”, “O3”, “PM10”, “SO2”, donde la frecuencia de medición es cada hora durante todo el año 2015. Pero para esta tarea solo nos enfocaremos en estudiar los contaminantes “CO” y “PM10”.
2. Obtener un reporte de la calidad de los datos de al menos una tabla, para determinar la información básica de la base de datos.
3. Con la ayuda del cálculo de las distancias, determine las estaciones que tienen un comportamiento similar durante todo el año. Es decir, se requiere comparar el “CO” de todas las estaciones durante todo el año para determinar si hay unas estaciones que se comportaron de forma similar.
4. Se debe de hacer la comparación de todas las estaciones ahora considerando solamente el contaminante “PM10” de cada estación y así sucesivamente.
5. Incluya en una tabla las medidas de similitud o distancia en cada una de las comparaciones.
6. El reporte debe de contener lo siguiente:
    a. Portada que incluye, el nombre de la tarea y del alumno como mínimo.
    b. En el desarrollo del reporte se debe de incluir el índice de similitud que se haya encontrado, su fórmula y un pequeño ejemplo de cómo se aplica en una serie de datos de prueba.
7. En cada figura, tabla, o resultado que se incluya, se debe de incluir el código que se utilizó para obtenerlo. Se recomienda incluir este código con un tipo de fuente distinta para que se diferencie del texto del reporte.

8. Todas las figuras (si las hubiera) deben de estar comentadas o descritas.
9. El reporte será entregado en digital (Word, PDF) y será subido en la liga de Moodle disponible.