Laboratorio #5 – Aprendizaje no Supervisado

Descripción:

Para este laboratorio aplicaremos métodos de aprendizaje no supervisado para determinar la asignación de clusters de cada fila del dataset, y así verificar cual es el desempeño de cada mecanismo en ambientes de incertidumbre.

Utilizando el archivo proporcionado en la carpeta adjunta realice lo siguiente:

1) Valide si la columna status_id vale la pena mantenerla en el dataset.
2) Realice un análisis estadístico de cada variable dentro del dataset Live.csv las cuales apliquen (todas excepto status_id, status_published), esto es:
    a. Gráfica del histograma y distribución de las variables numéricas continuas, promedio, mediana, varianza, desviación estándar, rango y describe.
    b. Distribución de frecuencias para las variables que se consideren discretas o categóricas.
3) Muestre una gráfica de serie temporal (debe ordenar las fechas) para cada tipo de entidad:
    a. Num_reactions
    b. Num_shares
    c. Num_likes
    d. Num_loves
    e. Num_haha
    f. Num_wows.
    g. Num_sads.
    h. Num_angrys.

4) Separe el dataset en X e y donde y será la columna status_type.

5) Aplique el procedimiento de ingeniería de características para preparar el dataset para aplicar clustering sobre el, esto es:
    a. Convertir las variables categóricas a numéricas (incluyendo la y).
    b. Aplicar Feature Scaling (recomendamos MinMaxScaler).

6) K-Means: Utilice el enfoque con las variables que quedaron como X para encontrar el numero de clusters más adecuado para agrupar los datos, recuerde que puede utilizar el metodo del codo para determinar cual es la cantidad de clusters adecuada, sin embargo en esta ocación debemos considerar cual es la cantidad de etiquetas colocadas correctamente a cada fila X y determinar el accuracy de la asignación.

7) Aplique el método de PCA para realizar la clusterisación con el mismo criterio que se menciona en el inciso anterior.

8) Investigue en que consiste el método t-sne y apliquelo para realizar la agrupación de los datos.

9) Investigue como funciona el método de clustering jerarquico y apliquelo a este conjunto de datos.

10) Finalmente compare cual es el método que obtiene mejor accuracy en la asignación de clusters para los datos.