Ciencia de Datos
Tarea 1

Profesor: Diego Ramírez
Ayudante: Abner Astete

22 de abril de 2023

1. Resumen

La presente tarea tiene por objetivo evaluar la aplicación de preprocesamiento y clustering sobre un conjunto de datos.

2. Análisis de Datos

En primera instancia utilizar el dataset `Pokemon`, con todos sus datos.

2.1. Perfilado (15 pts)

Para este punto debe obtener un resumen del dataset indicando claramente:

1. Cantidad de atributos.
2. Cantidad de registros.

Además, entregar la descripción completa de cada atributo dataset incluyendo:

1. Significado
2. Tipo de dato
3. Valores faltantes
4. Mínimo (si es numérico)
5. Máximo (si es numérico)
6. Desviación estándar (si es numérico)

2.2. Preprocesamiento (50 ptos)

Para este punto es libre de utilizar WEKA y/o Python para preprocesar los datos. Debe aplicar:

Normalización y Discretización (25 ptos)

1. Normalización de todas las variables numéricas. Indicando claramente que tipo de reescalamiento o normalización aplicó y porque.
2. Discretización de todas las variables numéricas:
   - Obtener intervalos de igual amplitud. Exponga el objetivo de este preprocesamiento.
   - Obtener intervalos de igual frecuencia. Exponga el objetivo de este preprocesamiento.
   - Elija alguna variable categórica del dataset y conviertala a numérica. Porque aplicó la transformación a esta columna?

Valores Nulos y Outliers (25 ptos)

Para el caso de los valores nulos, utilice el dataset `pokemon_with_nulls`.

1. Apliqué imputación a los datos para rellenar los nulos. Defina claramente cual imputación aplicó y porque.
2. Compare los resultados de la imputación sobre el dataset `pokemon_with_nulls` y el dataset original `pokemon`. Concluya que tan efectiva fue la imputación de los datos.

Para los valores atípicos utilice el dataset `pokemon`

1. Obtenga mediante candidatos a Outliers aplicando los métodos vistos en clases.
2. Elija uno de sus resultados del punto anterior, justificando porque eligió ese método, exponga los resultados obtenidos y sus conclusiones al respecto.

2.3. Clustering (35 ptos)

Descubra, mediante un análisis no supervisado, los grupos (clusters) en el dataset `pokemon`. Considere utilizar los datos numéricos normalizados vs no normalizados. Explique qué consideración toma respecto a la cantidad de clusters que pretende utilizar. Realice este procedimiento con el dataset completo, pero posteriormente considere retirar atributos irrelevantes y outliers. Interprete los resultados obtenidos y cuál puede ser el uso de estos grupos obtenidos.

Compare los resultados obtenidos usando el algoritmo *K-Means* y *DBSCAN*. Con cual de los 2 agrupamiento se quedaría y porque?

3. Datasets

Los datasets necesarios se encuentran disponibles en
`https://github.com/diegoalrv/demonic.soda/tree/master/datasets`

4. Entregable

Se deberá realizar en grupos de máximo 2 personas. Se debe enviar el documento de informe en formato PDF con plazo máximo hasta el 8 de Mayo del 2023. Esta permitido el uso de cualquier herramienta digital, el puntaje asignado principalmente es por sus justificaciones y conclusiones.

El documento debe ser en formato presentación (PPT), mostrando el procedimiento aplicado, los resultados obtenidos y las conclusiones al respecto. Debe entregar un documento **ordenado y legible**.