Trabajo Practico N° 1
Minería de datos

Objetivo

El objetivo de este trabajo practico es integrar los conocimientos adquiridos en las unidades 2 y 3 en un problema real asociado a los cultivos.

Actividades

1. Descargar un conjunto de datos, dxCrop_Soil.csv[1], para realizar el trabajo práctico. Realizar un análisis exploratorio de datos: visualizar distribuciones, valores faltantes, correlaciones, etc. Limpiar el conjunto de datos (manejar valores faltantes, eliminar outliers) si es necesario. Codificar variables categóricas (si es necesario). Normalizar o estandarizar las características. La variable tipoCultivo es nuestro objetivo.

2. Realizar PCA y determinar el número de componentes principales considerando alguno de los 3 criterios datos en la práctica. Graficar la varianza acumulada y las componentes de PCA en un gráfico 2 o 3D con sus respectivas clases.

3. Aplicar Isomap y analizar los resultados obtenidos variando el número de vecinos y componentes. Realizar un gráfico en 2D de utilizando dos componentes.

4. Aplicar t-SNE y analizar los resultados obtenidos variando el número de iteraciones, componentes y perplejidad. Realizar un gráfico en 2D de utilizando dos componentes.

   Ahora, realizar un subconjunto de los datos considerando solamente estos tipos de cultivos: Maíz, Cebada, Trigo

5. Realizar nuevamente PCA y graficar las componentes de PCA en un gráfico 2D con sus respectivas clases.

6. Aplicar K-means y analizar los resultados obtenidos variando el número de clusters y obtener el número óptimo de clusters mediante GAP. Realizar un gráfico en 3D de utilizando tres atributos de los datos y donde los colores estén asociados a los clusters.

7. Aplicar clustering jerárquico y determinar cuál número sería el que mejor represente los datos. Utilizar el score de Silhouette y calcular el número óptimo de cluster por medio de GAP.

Presentación

La entrega es por grupos de dos estudiantes y se entregan un archivo por grupo.

Cualquier integrante del grupo puede hacer la entrega mediante el campus de la materia.

[1] https://www.kaggle.com/datasets/shankarpriya2913/crop-and-soil-dataset

Trabajo Practico N° 1
Minería de datos

El informe deberá tener una cabecera en la que se indique: año, materia, integrantes. Además, deberá contar con una sección de conclusiones al final del mismo.

El formato del informe deberá ser en formato ipynb y no debe contener las definiciones teóricas ni el significado de los parámetros de los métodos dados en clase.

Las gráficas mostradas en el informe deben contener una explicación de lo observado y si es coherente que su hipótesis previa, por tanto, la cantidad de graficas debe estar acotadas y ser representativas.

Las entregas fuera del plazo establecido no serán consideradas salvo excepciones previamente justificadas por el grupo.