Taller 8 PR: *Clustering*

Facultad de Ingeniería

Departamento de Electrónica

Nota: fecha máxima de entrega del informe domingo 30 de octubre de 2022 a las 11:59 p.m. Por cada minuto de retraso en la entrega se descontará una (1) décima.

Objetivo:

- Estudiar el modelo de conectividad y el modelo de centroide. Para ello, se utilizará el agrupamiento jerárquico tipo aglomerativo y el algoritmo k-means. Además, se determinará el número "óptimo" de grupos, se interpretarán los resultados obtenidos y su utilidad como estrategia en el deporte analizado.

1. (5.0 p.t) Utilice el conjunto de datos "*Basketball data set*" disponible en: https://sci2s.ugr.es/keel/dataset.php?cod=1293sub1, para estudiar el aprendizaje automático no supervisado empleando un modelo de conectividad (agrupamiento jerárquico) y un modelo de centroide (algoritmo k–means), y así, realizar un análisis exploratorio de los datos.

Con base en lo anterior, realice:

a.) (1.0 p.t) Estadística descriptiva de los datos:

   i.) (0.1 p.t) Describa estadísticamente el conjunto de observaciones.

   ii.) (0.2 p.t) Obtenga los histogramas de las variables de entrada y analice si las observaciones provienen de una población con una distribución normal (Gaussiana). Si considera pertinente realizar una prueba adicional de normalidad, i.e., *Kolmogorov–Smirnov*, *pp–plot*, *qq–plot*, etc., siéntase libre de hacerlo.

   iii.) (0.3 p.t) Obtenga los diagramas de dispersión.

   iv.) (0.4 p.t) Examine la dependencia entre las variables de entrada con base en el criterio que considere idóneo (i.e., matriz de covarianza, coeficiente de correlación de *Pearson*, etc.).

b.) (2.0 p.t) Agrupamiento jerárquico (modelo de conectividad):

   i.) (0.5 p.t) Utilice el algoritmo de agrupamiento jerárquico aglomerativo para agrupar los datos.

   ii.) (0.5 p.t) Grafique el dendrograma y estime el número "óptimo" de grupos a través de la técnica vista en clase basada en la distancia máxima en el dendrograma.

   iii.) (0.5 p.t) Utilice la técnica del codo (*Elbow Method*) para tener un criterio adicional del número "óptimo" de grupos. El código de este ítem deberá ser propio.

   iv.) (0.5 p.t) Seleccione el número "óptimo" de grupos (k). Justifique su respuesta.

c.) (1.2 p.t) Agrupamiento con k–means (modelo de centroide):

i.) (0.5 p.t) Utilice el algoritmo k–means para agrupar los datos, teniendo como referencia el k seleccionado en el ítem previo.

ii.) (0.2 p.t) Grafique los clústeres en 3D (escoja las variables que considere pertinentes).

iii.) (0.5 p.t) Grafique el coeficiente de silueta e interprete los resultados.

d.) (0.4 p.t) Defina ¿cuáles podrían ser las k–etiquetas?

e.) (0.4 p.t) Concluya sobre los resultados obtenidos.