EJERCICIO 2

ESCENARIO

El Data Mining Educacional, o "Educational Data Mining" es un nuevo campo emergente que pretende descubrir conocimiento a partir de datos que se originan en ambientes educativos.

En este caso de estudio utilizaremos RapidMiner para agrupar estudiantes de acuerdo a su rendimiento académico, para lograr equipos de alumnos más efectivos, que permitan alcanzar mejores rendimientos de aprendizaje de sus integrantes.

El objetivo es crear grupos de estudiantes de acuerdo a sus características personales. Luego estos grupos así formados podrán ser utilizados por los profesores para construir sistemas de aprendizaje más personalizados, para promover el aprendizaje efectivo en equipos o incluso para proveer contenidos adaptivos.

Datos

El dataset se obtuvo de la Facultad de Ciencias Organizacionales, Universidad de Belgrado, Serbia. Contiene 366 registros de estudiantes graduados y su rendimiento académico.

Se proveen los archivos "ClusteringStudents" (ejemplos) y ClusteringStudents.aml (descripción xml)

Atributos en el dataset:

- **Sexo:** femenino / masculino, binomial
- **Región:** De donde proviene el estudiante, nominal
- **Puntaje en la prueba de admisión:** Valores obtenidos en el examen de ingreso. Valores 40-100, tipo real.
- **Calificaciones en el primer año** Notas en cada uno de los 11 exámenes del primer año de estudios. Valores: 6 a 10, tipo: entero
- **Calificacion promedio:** Promedio de calificaciones del estudiante luego de la graduación, valores: 6 – 10, tipo: continuo
- **Rendimiento académico del estudiante:** Rendimiento al finalizar los estudios, valores: "Malo", "Bueno", y "Excelente"; tipo: polinomial

Rendimiento académico (RA): el atributo se ha obtenido discretizando la calificación promedio, de la siguiente manera:

- Si la calificación es 8 o menor, el RA es "Malo"
- Si la calificación está entre 8 y 9, el RA es "Bueno",
- Y si es mayor a 9, el RA es "Excelente"

Si bien este atributo no es necesario para los procesos de clustering, sí es conveniente disponer del mismo para las tareas de visualización.

Preparación de los Datos

1. Importar el dataset, editar los nombres y tipos de atributos.

Aspectos a tener en cuenta:

2. El atributo "Rendimiento Académico" no se utilizará para la generación los modelos de clustering.

3. Deseamos realizar los modelos de clustering basados solamente en el rendimiento de los alumnos en la facultad, no en sus datos personales.
4. Revisar outliers, faltantes y otros problemas en los datos. Analizar necesidades de normalización. Incorporar y documentar las acciones tomadas.

Modelos

Aplica los siguientes algoritmos para generar modelos

- k-means
- DBSCAN
- Aglomerative clustering
- Top-down Clustering

Documenta las configuraciones de parámetros.

Evaluación

- Analiza el rendimiento de generación de clusters con cada método, en forma análoga al TA1
- Visualiza los clusters e interpreta los resultados. Captura los gráficos y explica la interpretación.