



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ALMACÉN Y MINERÍA DE DATOS

PRÁCTICA 13

AGRUPAMIENTO

- **CALDERÓN FERNÁNDEZ GABRIEL**
- **FLORES GONZÁLEZ LUIS BRANDON**
 - **SANTAELLA MARÍN HÉCTOR**

1) Detecte los grupos más representativos del censo.

Edad

El grupo más representativo para este serán los que se encuentran entre las edades de 20 a 49 en el cluster número 3.

Tipo Empleado

El grupo más representativo sería del 1 al 5, dentro de la categoría Private. Claramente en este rango se encuentra el mayor número de personas por una gran diferencia en comparación a las demás categorías.

Nivel Educacion

Dentro de este apartado no hay mucha diferencia, pero se puede notar que dentro de la categoría HS-grad en los clusters 2, 3, 4 hay un cúmulo de gente mayor a los demás agrupamientos.

Estado Civil

La mayor cantidad de personas se encuentra en Married-civ-spouse del cluster 1 al 4, es más probable encontrar gente dentro de este rango.

Ocupacion

La ocupación igual no es muy notoria las agrupaciones, ya que la diferencia entre ellas no es demasiada. Pero el mayor cúmulo se registra dentro de Craft-repair del cluster 1, 2 y 3.

Raza

En este apartado es muy sencillo ver la diferencia de las agrupaciones. La raza blanca es el grupo más representativo en todos los clusters.

Sexo

El grupo más representativo es el de los hombres en todos los clusters.

Ganancias

La ganancia más común es 15024 dentro de los grupos 2, 3 y 5.

Perdidas

Las mayores pérdidas están registradas en 1902, en los clusters 2, 3 y 4.

Horas Trabajadas

El grupo más representativo es dentro de 40, para todos los clusters.

Pais Origen

Aquí tenemos la mayoría de gente dentro de la categoría de Estados Unidos dentro del cluster 3.

Ingresos

La mayor gente gana ingresos dentro del cluster 2 y 3, dentro de ambos intervalos de ingresos. Pero $\leq 50k$ es el mayor ingreso en el cluster 2 y 3.

2)Indique cuáles son los valores atípicos encontrados bajo esta configuración. Justifique. Indique cuáles son los valores atípicos encontrados bajo esta configuración. Justifique.

Edad

Tiene valores atípicos ya que nos encontramos que en el rango de las edades 60 a 90 no existen datos en mayor parte.

Tipo Empleado

Tiene valores atípicos ya que nos encontramos que en la columna "Without-pay" a excepción del primer cluster no contamos con suficientes datos.

Nivel Educacion

Bajo la configuración dada nos podemos encontrar que en los cluster 1, 2, 3, 5 ya no existen datos, las columnas con mayor datos atípicos son "1st-4th" y "Preschool".

Grado Educacion

Tiene valores atípicos ya que nos encontramos que en los cluster, no hay datos correspondiente para las columnas 1, 2, 3, pero donde se encuentra los valores atípicos son en la columna 1.

Estado Civil

En la columna "Married-AF-spouse" en los cluster 1, 2, 4, no se encuentran valores.

Ocupacion

La mayoría de los valores atípicos se encuentran en la columna "Armed-Forces" en los cluster 1, 2 y 5.

Raza

En el cluster 5 es donde se encuentran los valores atípicos en la columna “Amer-Indian-Eskimo” y “other”.

Ganancias

En este atributo es donde se encuentran más valores atípicos en los cluster y en cada una de las columnas de ganancias. En la columna 0 de ganancias no existen valores atípicos.

Perdidas

Perdidas

En este atributo es donde se encuentran más valores atípicos en los cluster y en cada una de las columnas de pérdidas. En la columna 0 de pérdidas no existen valores atípicos.

Horas Trabajadas

Valores atípicos se encuentran en todos los clusters solo que difiere la columna, en algunas columnas en particular no existen valores atípicos.

Pais Origen

Valores atípicos se encuentran en todos los clusters solo que difiere la columna, en algunas columnas en particular no existen valores atípicos.