



**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

---

**Facultad de Estudios Superiores  
Acatlán**

## **Practica 1:**

*Clúster jerárquico*

**Análisis multivariado y regresiones**

**Integrantes:**

Gustavo Adolfo Alvarez Hernández

**Profesor:**

Mahil Herrera Maldonado



**FES Acatlán, 2 de noviembre 2022**

## Contenido

Portada .....	1
Objetivo .....	3
Introducción .....	3
Problemática .....	3
Técnicas estadísticas utilizadas.....	3
Cluster Jerárquico Aglomerativo.....	3
Percentiles .....	4
Análisis descriptivo .....	4
Por materia .....	4
Por orden de preferencia .....	5
Análisis utilizando la técnica.....	5
Dendrograma.....	5
Análisis de los clusters.....	6
Cluster 0.....	6
Cluster 2.....	7
Cluster 3.....	7
Análisis 3 principales clusters .....	8
Conclusiones.....	9
Referencias .....	9
Anexo .....	9
Errores de llenado.....	9
Formato de los datos. ....	10
Votos perdidos.....	10

## Objetivo

Se busca implementar el algoritmo de cluster jerárquico sobre los datos de preferencia de optativas que se realizó en la carrera de actuaria, con el fin de encontrar al grupo que contenga a la población más representativa y así tomarla en cuenta para la elección de las 5 optativas que se ofertaran en el siguiente semestre.

## Introducción

En la presente práctica se irá desarrollando la problemática de ofertar optativas en la FES Acatlán, la cual se tienen los datos recabados en una encuesta, la cual de diversas transformaciones nos permite aplicar el cluster jerárquico; posteriormente realizar un análisis de los grupos generados, ponderarles su importancia y así obtener las 5 materias más deseadas.

También se explicará grosso modo las complejidades de la implementación y las transformaciones necesarias para que los datos sean adecuados para el modelaje.

## Problemática

Se presenta una encuesta realizada por alumnos de actuaria de la FES Acatlán, en la cual el alumnado tiene la oportunidad de dar a conocer las 5 materias optativas que le gustaría estuvieran disponibles para el siguiente semestre. Dado que los recursos para contratar profesores como los espacios para impartir la clase son limitados, no es posible que se impartan todas las materias, por lo cual por medio de segmentación de los alumnos encuestados se busca proponer 5 materias de tal manera que la comunidad este de acuerdo con las ofertadas.

En cuanto al formato de los datos, vienen dados por el alumno, clave y materia que les agrada, así como un indicador del 1 al 5 para marcar su preferencia por esa materia, donde 1 le interesa mucho y 5 representa la preferencia mínima pero que aun así le interesa la materia.

## Técnicas estadísticas utilizadas

Además del cálculo de estadísticos relevantes como puede ser la moda, media y mediana, se abordó el algoritmo de clustering también analizamos los grupos por medio de sus percentiles.

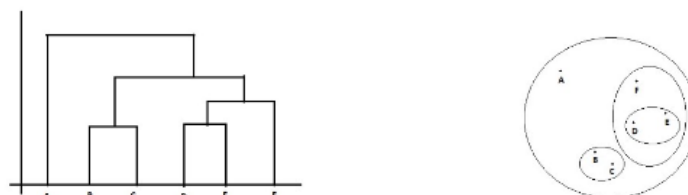
### Cluster Jerárquico Aglomerativo

Es un tipo de implementación del algoritmo no supervisado que se basa en considerar a cada dato como un cluster, se procede a encontrar la distancia entre cada cluster y se agrupan los clusters que se encuentren más cercanos y se genera uno nuevo.

Todo este proceso se lleva a cabo hasta que solo haya un cluster y se da a través de la matriz de proximidad y en cada iteración iría actualizándola con el nuevo cluster. El método para usar puede variar y determinan resultados distintos como pueden ser:

1. Los dos puntos más lejanos entre clusters
2. Los dos puntos más cercanos entre clusters
3. Distancia entre centroides
4. Método Ward

Por último, podemos visualizar los distintos grupos que se realizaron a través de un dendrograma.

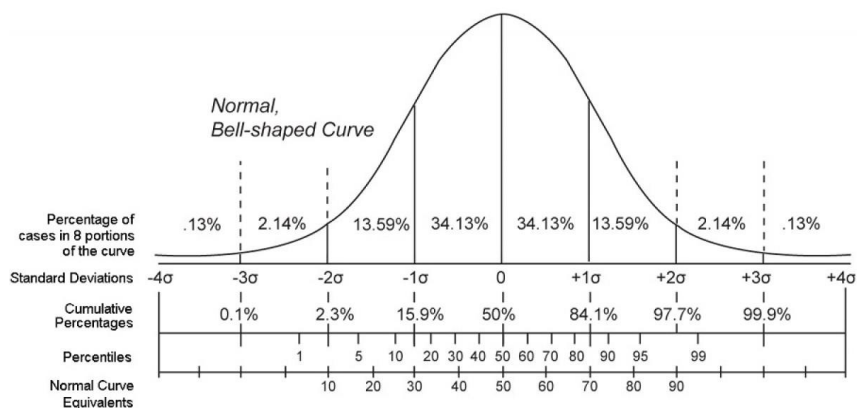


Dendrogram representation

## Percentiles

Los percentiles es un estadístico el cual arroja el valor a partir del cual se concentra cierto porcentaje de la muestra. Esto ordenando los valores de menor a mayor y se va buscando hasta que valor se tiene ese k porcentaje de valores con respecto al tamaño de la muestra.

Además, en la distribución normal existe cierta relación con su desviación estándar, como se muestra en el siguiente diagrama.



## Análisis descriptivo

Tenemos el listado que contiene 608 votos que contiene el identificador del usuario, el orden de preferencia y la materia en cuestión, por lo cual analizaremos lo siguiente:

### Por materia

Contamos con 25 materias distintas que fueron votadas con diferentes ordenes de preferencia dados por el alumno, siendo el top 5 más votado:

- |                            |    |
|----------------------------|----|
| 1. Modelos y Simulación    | 85 |
| 2. Análisis de Regresión   | 64 |
| 3. Evaluación de proyectos | 56 |
| 4. Derivados               | 54 |

## 5. Muestreo 43

Y en su contraparte, tenemos el top materias menos votadas, que son:

- |                             |   |
|-----------------------------|---|
| 1. Análisis econométrico    | 1 |
| 2. Modelos microeconómicos  | 1 |
| 3. Legislación de seguros   | 2 |
| 4. Seguros de personas      | 3 |
| 5. Procesos estocásticos II | 3 |

### Por orden de preferencia

Existen opciones desde el 1 al 5, que establecen la importancia de la materia para determinado alumno, esta se distribuye de la siguiente manera:

- |             |     |
|-------------|-----|
| 1. Opción 5 | 124 |
| 2. Opción 4 | 124 |
| 3. Opción 3 | 122 |
| 4. Opción 2 | 120 |
| 5. Opción 1 | 118 |

Y podemos observar cómo se distribuye el voto de máxima prioridad en los datos, o sea las materias que se asignaron como:

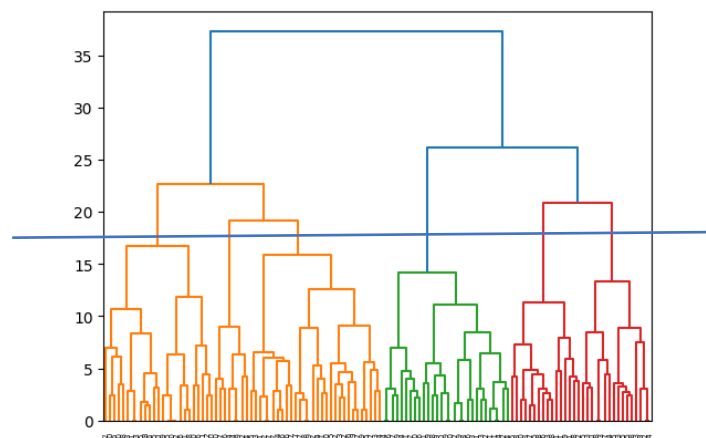
- |                                    |    |
|------------------------------------|----|
| 1. Evaluación de proyectos         | 14 |
| 2. Muestreo                        | 13 |
| 3. Investigación de Operaciones II | 10 |
| 4. Derivados                       | 10 |
| 5. Modelos y Simulación            | 9  |

### Análisis utilizando la técnica

Una vez que tenemos nuestros datos presentados de una manera apta para el modelaje del cluster (se detalla el anexo), procedemos a generar el dendrograma para determinar cuántos grupos tenemos.

#### Dendrograma

Utilizando el método Ward que toma en cuenta la distancia de todos los puntos y usando la métrica euclidiana se genera el siguiente diagrama:



Como podemos observar, queda un poco a interpretación en cuantos grupos se van a dividir, lo que se tomo en cuenta fueron los clusters más grandes que están por debajo de la línea azul, puesto que considere que se encuentran lo suficientemente divididos los alumnos en 5 grupos.

### Análisis de los clusters

Una vez que se determino que sean 5 grupos, utilizamos el algoritmo y le asignamos su nueva etiqueta al alumnado y con fin de saber que grupos son mas relevante tomamos las siguientes estadísticas:

Cluster	Proporción	Alumnos con 5 votos	Alumnos con preferencia 1
0	30.64%	73%	57%
1	20.16%	51%	44%
2	23.38%	83%	62%
3	12.90%	91%	93%
4	12.80%	92%	75%

Marcamos en color verde, amarillo y rojo a los clusters mas significativos siguiendo ese orden, esto se da a que el cluster contiene un buen porcentaje del alumnado, esta en un 83% lleno los votos de sus alumnos (esto es, están sus 5 materias) y en más de la mitad contiene la primera opción.

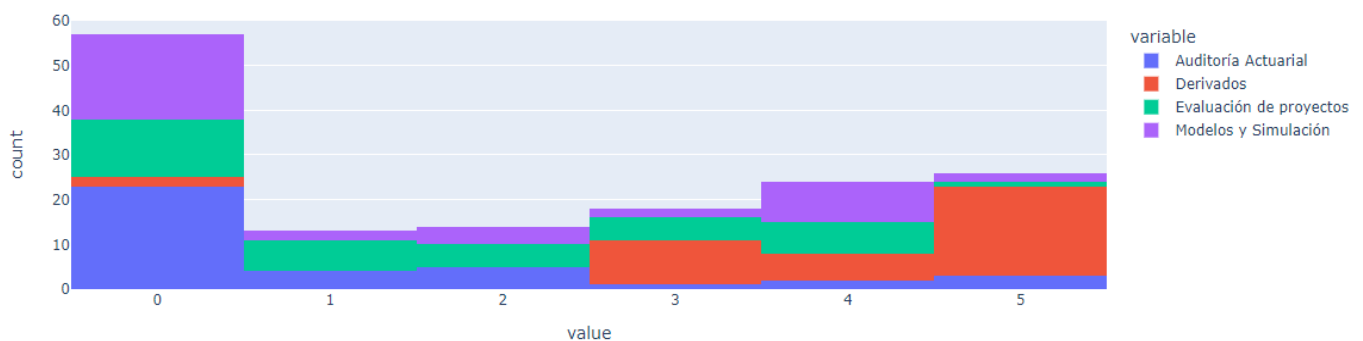
Por su parte el cluster 3, tiene pocos alumnos, pero la calidad de sus datos son bastantes buenos y por último el cluster 0 contiene a la mayor cantidad de alumnos, pero esta muy lleno de ceros, por lo cual no es tan significativo.

El análisis que se hará es ver el percentil 50 o 60, para observar que materias en ese percentil es mayor a 0, esto indica que al menos al 40% del cluster tiene preferencia por esta materia.

### Cluster 0

Utilizando el método de los percentiles tenemos que las materias más aceptadas son:

- Auditoria Actuarial
- Derivados
- Evaluación de proyectos
- Modelos y simulación

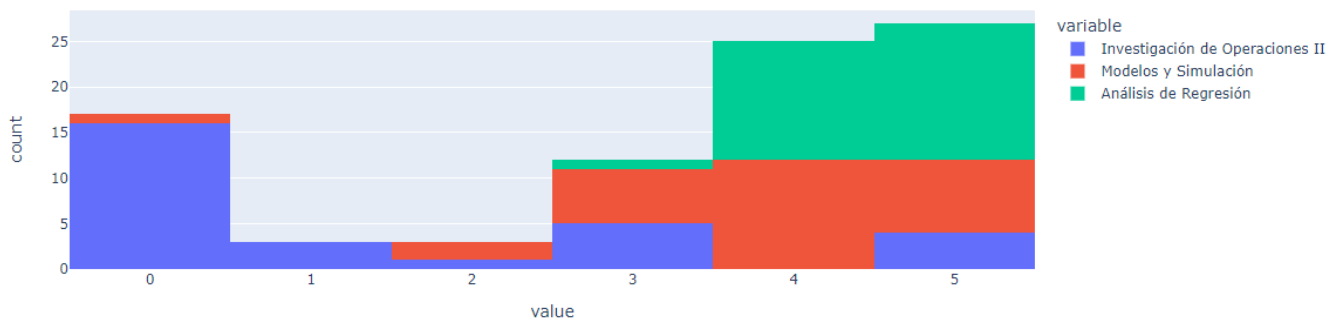


Y se distribuyen así sus votos, hay que hacer hincapié en que Derivados es muy poco rechazado y evaluación de proyectos es la que más 1's tiene.

### Cluster 2

Por su parte en este cluster tenemos menor cantidad de materias que cumplan con el filtro, siendo:

- Investigación de operaciones II
- Modelos y simulación
- Análisis de regresión

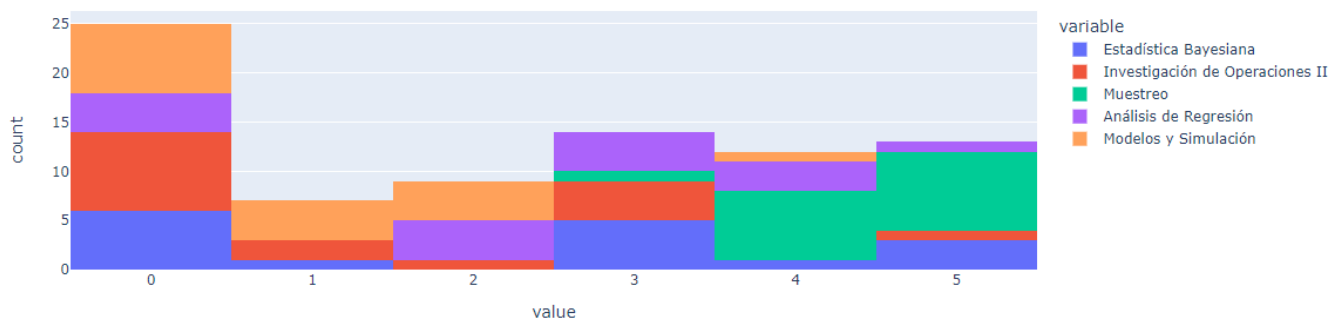


En este cluster Análisis y Modelos es altamente aceptado al menos como una de sus opciones no tan populares e investigación tiene varios unos, pero no es del todo aceptada en general.

### Cluster 3

Por último, tenemos a la mayor cantidad materias deseadas por este grupo, que son:

- Estadística bayesiana
- Investigación de operaciones II
- Muestreo
- Análisis de regresión
- Modelos y simulación

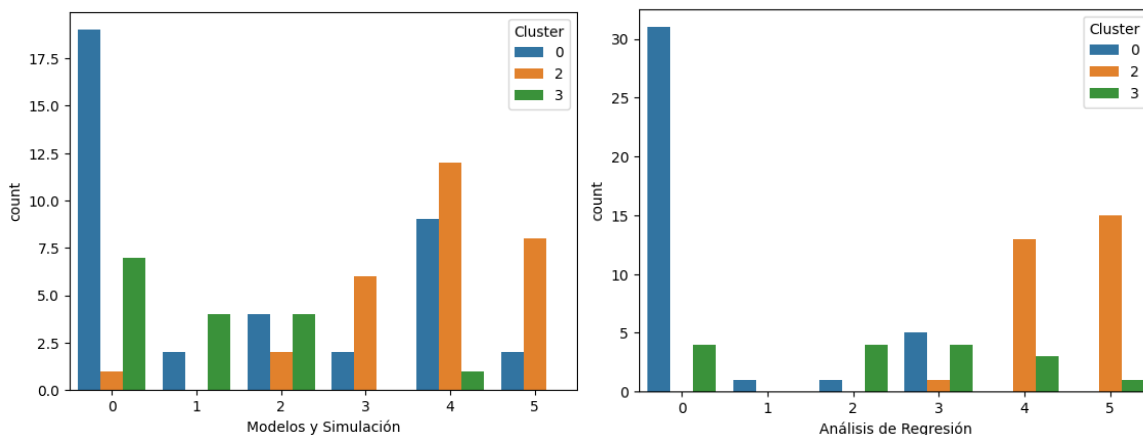


Este es el mas diverso y con mayor dispersión en los valores y deja en claro que Modelos es una materia bastante deseada, en los 3 grupos.

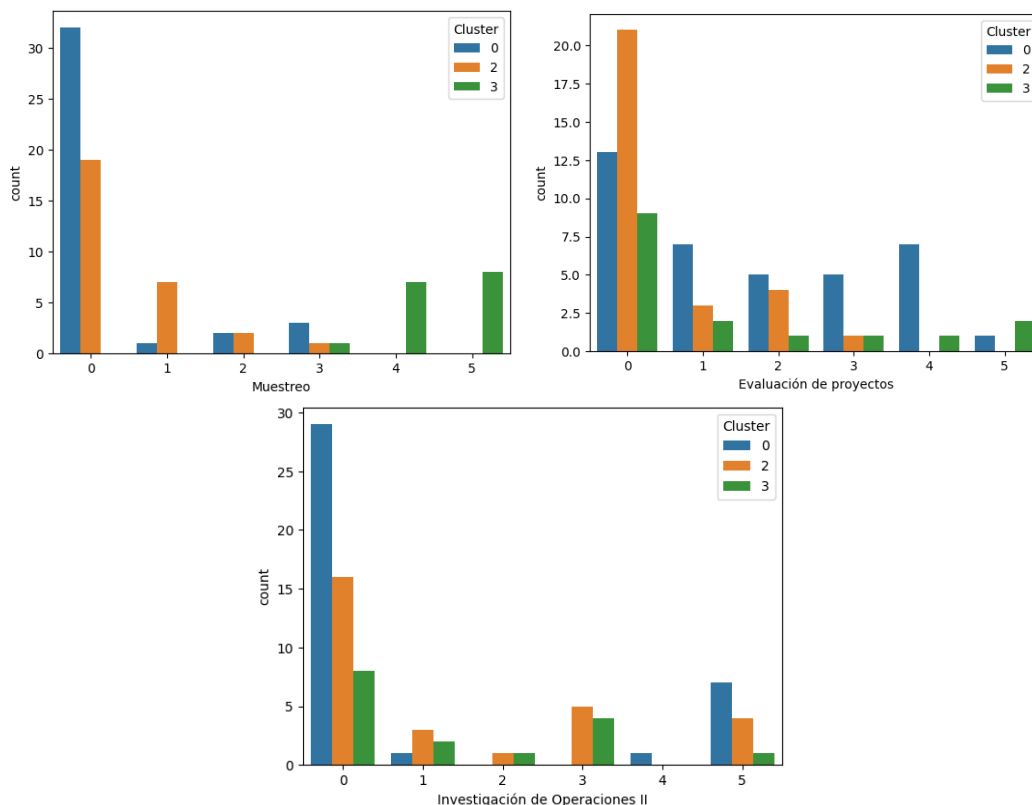
### Análisis 3 principales clusters

Para incorporar al final estos 3 análisis separados, los graficaremos y tomaremos en cuenta como se distribuyen en cada una de esas materias predilectas, dando preferencia al etiquetado de colores que hicimos en la tabla.

Mostrando algunas gráficas interesantes para la elección de materias, nos queda que:



Como era de esperarse, hay muchos votos mayores a 0 en Modelos y simulación y por tanto se le considera una materia por excelencia a publicarse, también análisis de regresión que, aunque tiene oposición por el cluster 0, quedamos que su participación era la menos relevante. Y así realizando este análisis de la grafica se seleccionan las demás materias, se muestran las graficas de las 3 faltantes. Como comentario se vuelve un poco más complejo escoger en algunos casos.



## Conclusiones

Después de todos los análisis elaborados, se concluye que las materias a ofertar deberían de ser:

- Modelos y simulación
- Análisis de regresiones
- Investigación de operaciones
- Muestreo
- Evaluación de proyectos

Las cuales pertenecen al top 5 de votaciones en general exceptuando a Investigación de operaciones, pero si en el top de opciones 1, por lo cual los resultados suenan coherentes a lo que se obtendría con un solo conteo.

Entonces, ¿vale la pena todo el desarrollo? Sí porque por medio de esta segmentación podemos tomar perfiles de gustos distintos y considerarlos lo máximo posible, además de que tienen gustos compartidos por lo cual se puede medianamente complacer a todos, que es al final lo que se busca.

Por último, se tendría mejor segmentación si la matriz que se genera no fuera tan rala, esto es que tenemos muchos campos con valores 0, por lo cual complica el algoritmo que al final mide distancias, por lo cual fue adecuado el solo hacer columnas al top materias.

## Referencias

Patlolla, C. R. (10 de Diciembre de 2018). *Towards Data Science*. Recuperado el 2 de Noviembre de 2022, de <https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec>

Sharma, P. (27 de Mayo de 2019). *Analytics Vidhya*. Recuperado el 2 de Noviembre de 2022, de <https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/>

## Anexo

En cuanto a los datos obtenidos de la encuesta, presentaban varios problemas para la implementación del algoritmo que se divide principalmente en estas categorías:

### Errores de llenado

Los datos en el campo de “Nombre de la materia”, llegaban a ser inconsistentes puesto que a una misma clave de materia encontrabas nombres distintos o errores de escritura en el mismo como puede ser, por poner un ejemplo:

- Análisis de regresión
- Análisis regresión
- “Análisis de regresiones “

Por lo cual, se tuvo que filtrar por clave e ir estandarizando el nombre teniendo cuidado con los espacios, esta tarea fue hecha manualmente a través de la tabla en Excel e ir aplicando el filtro por clave, hasta unificar todas las claves.

### Formato de los datos.

Ya previamente se describió que cada fila de los datos traía el identificador del alumno, el orden de preferencia y los datos asociados a la materia como se muestra a continuación:

	Individuo	Orden de preferencia	Clave	Nombre de la Materia
588	113	2	2055.0	Series de Tiempo
342	66	2	2041.0	Muestreo
583	103	1	2055.0	Series de Tiempo
74	74	2	2032.0	Análisis de Regresión
95	101	3	2032.0	Análisis de Regresión

Pero nosotros necesitamos condensar los datos del alumno y que cada fila represente a un solo alumno con sus preferencias, por lo cual se plantea volver una columna para cada materia y marcar con 0 si no la escogió o con el valor de preferencia.

El inconveniente en este caso es que son 25 materias y los alumnos tienen 5 votos, por lo cual tendrían 20 campos nulos, lo cual entorpecería al modelo debido a la maldición de la multidimensionalidad, por lo cual solo se tomó las 10 materias más populares y se generó el siguiente formato:

	Series de Tiempo	Estadística Bayesiana	Investigación de Operaciones II	Análisis Multivariado	Auditoría Actuarial	Muestreo	Derivados	Evaluación de proyectos	Análisis de Regresión	Modelos y Simulación
Individuo										
1	0	0	0	4	2	0	0	3	0	0
2	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	2	0	4
4	0	0	0	0	0	0	3	2	0	0
5	0	0	0	0	1	0	3	0	0	5

Y así podemos saber por ejemplo que el alumno 1, tiene como 4ta opción a Análisis Multivariado y en segundo a Auditoría Actuarial.

### Votos perdidos

Este siendo la ultima dificultad, puesto que pueden existir los casos en el que alumno voto a alguna de las materias que no están el top, por lo cual no se tendrá en los nuevos datos; Un claro ejemplo es el alumno 1 que no tiene su voto 1 y 5.

Ante esta problemática no podemos votar por cualquiera, por lo cual se dejan así los datos, pero para ello se hizo el análisis de cluster y se verificaba el porcentaje de alumnos con todos sus votos y cuantos alumnos que votaron con un 1, se tienen y así se ve la importancia de los grupos.