UNIVERSIDAD AUTÓNOMA DE YUCATÁN
FACULTAD DE MATEMÁTICAS

Análisis Multivariado
Actividad 5

Profesor: Dr. Jorge Armando Argáez Sosa

Alumno: Amir Oswaldo Canto Palomo

Actividad 5.

1. Introducción

Contexto y objetivo

En el presente estudio y análisis de conglomerados contamos con una base de datos con 25 Universidades que contienen el reconocimiento que según las personas encuestadas perciben.

Variables consideradas

Para el presente trabajo se consideraron las siguientes variables que contiene la base de datos:

1. Calificación promedio obtenida por los admitidos en el examen SAT;
2. Porcentaje de estudiantes en el percentil de al menos 90% en su aprovechamiento;
3. Porcentaje de estudiantes aceptados;
4. Razón de alumnos por profesor;
5. Gastos anuales por alumno;
6. Porcentaje de alumnos graduados.

Se nos ha solicitado agrupar estas Universidades con base a estas variables en 3 diferentes conglomerados:

- Excelentes
- Buenas
- Regulares.

Al finalizar este estudio tendremos agrupados en estos conglomerados las Universidades de nuestra base de datos.

2. Metodología

Comenzamos nuestro estudio utilizando Python para el análisis mediante las siguientes librerías:

1. Pandas
2. Numpy
3. Scipy
4. Sklearn.

Realizamos un pequeño análisis exploratorio de datos para entender cómo está estructurada nuestra base de datos:

	Calificacion en el SAT	% de estudiantes en el percentil del 90% de aprovechamiento	% de estudiantes aceptados	Razón alumnos por profesor	Gastos anuales estimados	Porcentaje de graduacion
count	25.000000	25.000000	25.000000	25.00000	25.000000	25.000000
mean	12.664400	76.480000	39.200000	12.72000	27.388000	86.720000
std	1.083598	19.433905	19.727308	4.06735	14.424883	9.057778
min	10.050000	28.000000	14.000000	6.00000	8.704000	67.000000
25%	12.400000	74.000000	24.000000	11.00000	15.140000	81.000000
50%	12.850000	81.000000	36.000000	12.00000	27.553000	90.000000
75%	13.400000	90.000000	50.000000	14.00000	34.870000	94.000000
max	14.150000	100.000000	90.000000	25.00000	63.575000	97.000000

Observamos los rangos min y max de las variables y encontramos que ciertos datos han cambiado de “escala” como los gastos anuales que deben ser estimados en miles así como el valor del SAT.

Encontramos que son 25 Universidades y todas las 6 variables son de tipo numéricas.

```
Institución                                                25
Calificacion en el SAT                                     25
% de estudiantes en el percentil del 90% de aprovechamiento    25
% de estudiantes aceptados                                 25
Razón alumnos por profesor                                 25
Gastos anuales estimados                                   25
Porcentaje de graduacion                                   25
dtype: int64
```

Saber que nuestras variables son de tipo numéricas nos da una pista para realizar nuestro análisis con alguna de las distancias disponibles:

- Euclidiana
- Manhattan
- Minkowski
- Mahalanobis

Analizamos las columnas de nuestra base de datos y hemos renombrado estas a las siguientes para un mejor y más simple manejo del DataFrame:

- Institucion
- Calificación
- Percent90
- PercentAceptados
- RazonAlumnProf
- GastoAnual
- PercentGraduados

Dadas las escalas en las variables que hemos analizado, hemos decidido y procedido a estandarizar las variables anteriormente mencionadas.

Por la naturaleza de los datos hemos decidido elegir algoritmos jerárquicos aglomerativos.

Conglomerados Jerárquicos aglomerativos

Hemos elegido la distancia Euclidiana para realizar nuestro análisis con los siguientes métodos de enlace:

Método: WARD
Estandarizado y sin Estandarizar

Método: "Complete linkage"
Estandarizado y sin Estandarizar

Método: "Single linkage"
Estandarizado y sin Estandarizar

Método: "Centroide"
Estandarizado y sin Estandarizar

K-Medias

También hemos desarrollado el algoritmo de k medias con 2 y 3 clusters.

Para nuestra comparativa de conglomerados y obtención del Cluster correcto hemos usado diferentes métodos. Para el caso del algoritmo jerárquico usamos la prueba de la silueta para las k medias usamos el método del codo (elbow method).

y para la comparación de conglomerados usamos el Adjusted Index Score (ARI).

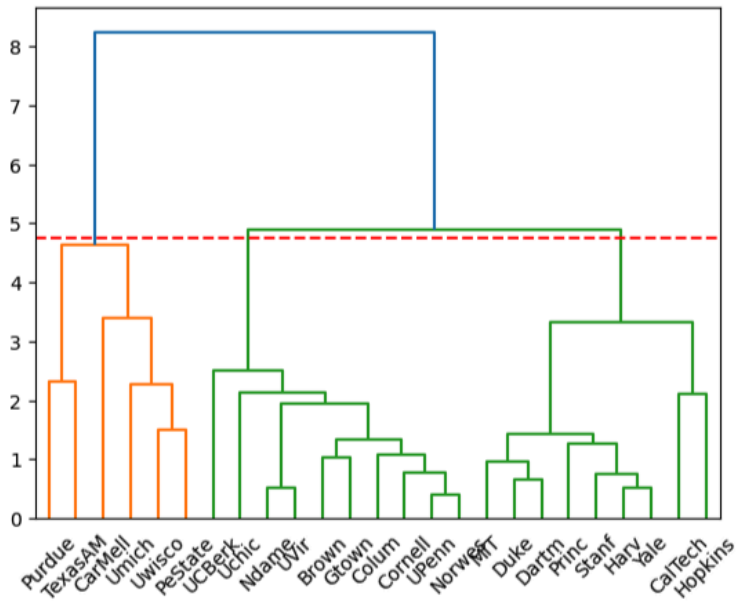
3. Resultados

Se evaluaron diferentes métodos de enlace jerárquico (Ward, Complete, Average) para determinar cuál produce la estructura de clusters más robusta. Ward fue seleccionado como método principal por dos razones: (1) obtuvo el mayor coeficiente de silueta (0.322 vs. 0.301 de Complete para k=3), indicando mejor separación entre grupos, y (2) es el método más utilizado en la literatura de clasificación institucional debido a su criterio de minimización de varianza intra-cluster, que favorece grupos homogéneos. Complete Linkage fue utilizado como método de validación.
La comparación entre ambos dendrogramas mostró alta consistencia (ARI=0.792), confirmando que la estructura de tres clusters no depende críticamente del método de enlace específico.

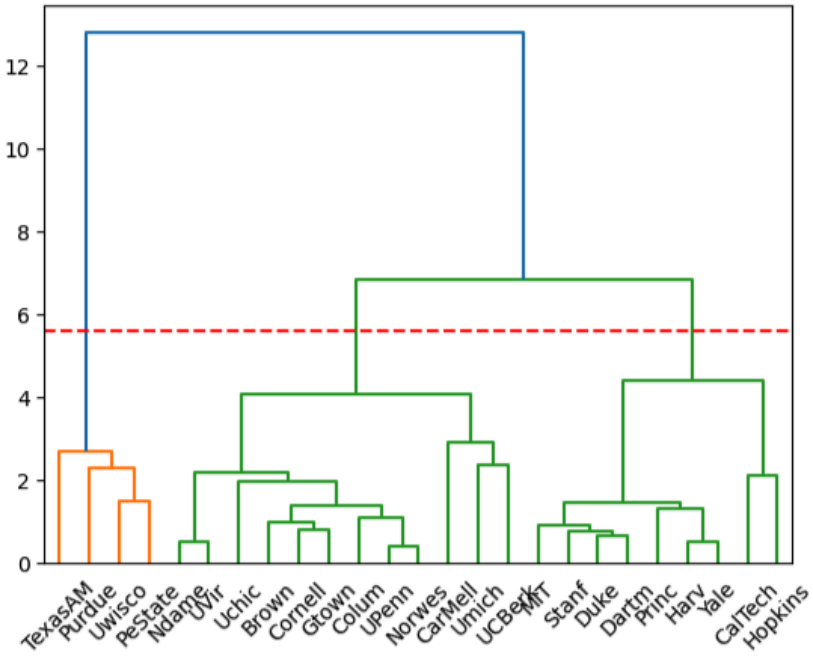
Ambos dendrogramas revelaron una característica común: CalTech se fusiona a una distancia considerablemente mayor que el resto de instituciones, (máximo SAT, máximos gastos, máximo %Percentil90). Esta separación confirma la heterogeneidad dentro del grupo de universidades excelentes. Los dendrogramas con k=3 identificaron consistentemente tres grupos interpretables:

1. Universidades públicas grandes con recursos limitados
2. Instituciones privadas de alta calidad con selectividad moderada
3. Universidades élite con máxima selectividad y recursos.

Complete Linkage



Ward

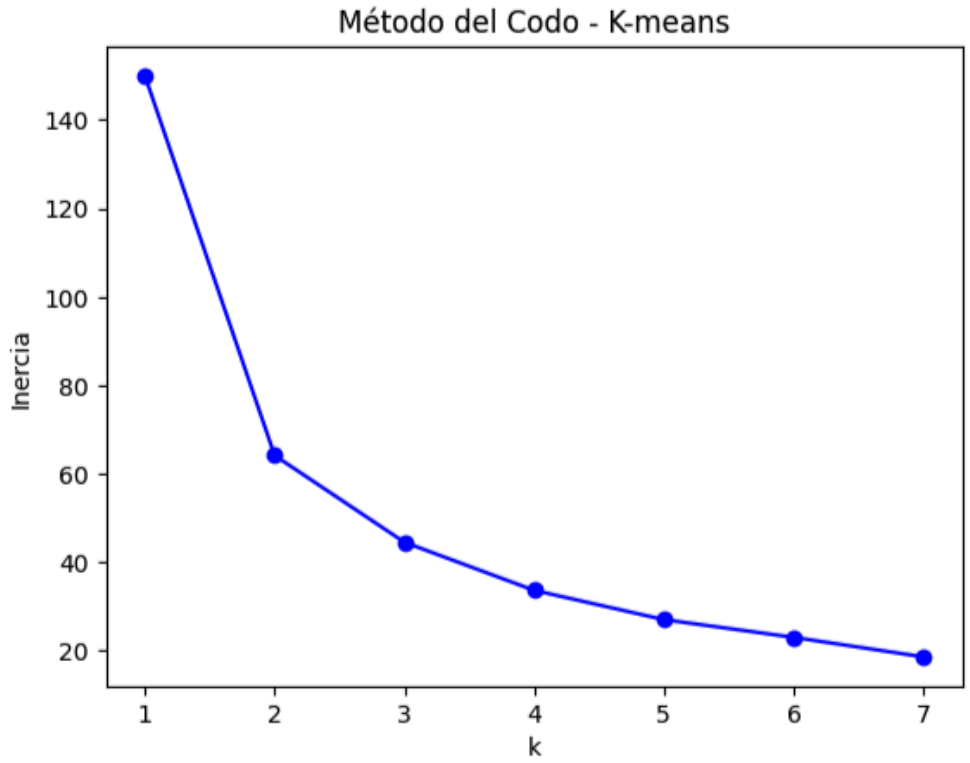


Determinación del número óptimo de clusters

Se evaluó el número óptimo de clusters mediante dos enfoques complementarios:

Método del Codo (K-means):

El análisis de inercia en función de k mostró un cambio pronunciado de pendiente en k=2, sugiriendo este valor como punto óptimo.
Para valores superiores (k=3, 4, 5), la reducción en inercia fue marginal, indicando que clusters adicionales no aportan mejora sustancial en la compacidad de los grupos.



Coeficiente de Silueta (Clustering Jerárquico):

- k=2: Silueta = 0.550 (buena separación)
- k=3: Silueta = 0.322 (separación débil)
- k=4: Silueta = 0.323 (sin mejora)
- k=5: Silueta = 0.313 (decrece)

Ambos métodos coinciden en que k=2 es el número óptimo según la estructura natural de los datos. Sin embargo, se procedió con k=3 según el objetivo del análisis (clasificar en Excelentes, Buenas y Regulares), logrando agrupaciones válidas y consistentes (ARI=1.0 entre Ward y K-means) a pesar de la menor separación.

```
Complete k=2: Silueta = 0.521
Complete k=3: Silueta = 0.301
Complete k=4: Silueta = 0.264
Complete k=5: Silueta = 0.268
Ward k=2: Silueta = 0.550
Ward k=3: Silueta = 0.322
Ward k=4: Silueta = 0.323
Ward k=5: Silueta = 0.313
```

Asignación de universidades

Si agrupamos en 3 conglomerados obtenemos la siguiente lista de Universidades:

CLUSTER 1: REGULARES (6 universidades)

1. - Carnegie Mellon
2. - University of Michigan
3. - University of Wisconsin
4. - Penn State
5. - Purdue
6. - Texas A&M

CLUSTER 2: BUENAS (10 universidades)

1. - Brown
2. - University of Chicago
3. - University of Pennsylvania
4. - Cornell
5. - Northwestern
6. - Columbia
7. - Notre Dame
8. - University of Virginia
9. - Georgetown
10. - UC Berkeley

CLUSTER 3: EXCELENTES (9 universidades)

1. - Harvard
2. - Princeton
3. - Yale
4. - Stanford
5. - MIT
6. - Duke
7. - CalTech
8. - Dartmouth
9. - Johns Hopkins

Ahora si agrupamos por 2 conglomerados.

CLUSTER 1: REGULARES (4 universidades)

1. University of Wisconsin

2. Penn State
3. Purdue
4. Texas A&M

CLUSTER 2: ÉLITE Y BUENAS (21 universidades)

1. Harvard
2. Princeton
3. Yale
4. Stanford
5. MIT
6. Duke
7. CalTech
8. Dartmouth
9. Brown
10. Johns Hopkins
11. University of Chicago
12. University of Pennsylvania
13. Cornell
14. Northwestern
15. Columbia
16. Notre Dame
17. University of Virginia
18. Georgetown
19. Carnegie Mellon
20. University of Michigan
21. - UC Berkeley

Validaciones.

Comparación de métodos

Se implementó una estrategia de validación múltiple para asegurar la robustez de los resultados.

Comparación entre métodos de clustering:

(1) Ward vs. K-means (k=3): ARI = 1.000
La coincidencia perfecta entre el clustering jerárquico (Ward) y el método no jerárquico (K-means) indica que ambos algoritmos, con fundamentos matemáticos diferentes, identificaron exactamente la misma estructura de agrupamiento. Este resultado, aunque poco común en la práctica, demuestra alta robustez y validez de los tres clusters identificados.

(2) Ward vs. Complete Linkage (k=3): ARI = 0.792
La comparación entre métodos de enlace mostró alta consistencia (79.2% de acuerdo ajustado por azar), confirmando que la estructura de grupos no depende críticamente del método de enlace específico utilizado.

Datos Estandarizados

Se realizaron análisis comparativos con datos estandarizados y no estandarizados para evaluar el impacto del preprocesamiento. Con k=3, el ARI entre resultados estandarizados y no estandarizados fue de 0.209 (Ward) y 0.396 (Complete), indicando diferencias sustanciales en las asignaciones.

```
Ward k=3 (std vs no-std): 0.209
Complete k=3 (std vs no-std): 0.396
```

Específicamente, 13 de 25 universidades (52%) cambiaron de cluster al no estandarizar. Sin estandarización, la variable Gastos Anuales (rango 54.9 miles de dólares) domina sobre variables con rangos menores como SAT (rango 4.1 puntos), resultando en una clasificación sesgada hacia recursos económicos en lugar de un balance entre calidad académica, selectividad y recursos. Esto confirma que la estandarización es esencial para otorgar igual peso a todas las dimensiones de calidad universitaria.

Análisis de Ward con estandarización y sin estandarización:

```
Universidades que cambian de cluster (War
   Universidad  Ward_std  Ward_no_std
6      CalTech         3            2
8        Brown         2            3
9      Hopkins         3            2
11       UPenn         2            3
12     Cornell         2            3
13      Norwes         2            3
14       Colum         2            3
15       Ndame         2            3
16        UVir         2            3
17       Gtown         2            3
18     CarMell         2            1
19       Umich         2            1
20      UCBerk         2            3
```

Análisis de Importancia de Variables

Análisis de Importancia de Variables Se evaluó la importancia de cada variable mediante clustering eliminando una variable a la vez y comparando con el análisis completo mediante ARI.

Resultados:

- Porcentaje de Aceptados: ARI = 1.000 (variable PRESCINDIBLE)
- Percentil 90: ARI = 0.215 (variable CRÍTICA)
- Calificación SAT: ARI = 0.256 (variable CRÍTICA)
- Gastos Anuales: ARI = 0.540 (variable CRÍTICA)
- Razón Alumnos/Profesor: ARI = 0.542 (variable CRÍTICA)
- Porcentaje Graduados: ARI = 0.752 (variable moderadamente importante)

El análisis reveló que cinco de las seis variables son necesarias para la clasificación. La única variable prescindible fue Porcentaje de Estudiantes Aceptados, lo que sugiere que la selectividad es una consecuencia de la calidad académica y recursos disponibles, no un factor independiente que distingue instituciones.

4. Conclusiones

El análisis de conglomerados reveló que la estructura natural de los datos favorece k=2 grupos (silueta=0.55), separando claramente universidades regulares (n=4) de las demás instituciones

(n=21). Sin embargo, la clasificación en k=3 grupos solicitada (Excelentes, Buenas, Regulares) es válida, con una silueta de 0.32 que indica separación débil pero aceptable.

Las validaciones múltiples demostraron alta robustez en las asignaciones. El clustering jerárquico (Ward) y K-means coincidieron perfectamente (ARI=1.0), confirmando la consistencia de los grupos identificados. Adicionalmente, la comparación entre métodos de enlace (Ward vs. Complete) mostró alta concordancia (ARI=0.79).

Respecto a las preguntas de investigación:

¿Cambian los resultados con variables estandarizadas vs. no estandarizadas?

Sí, significativamente. Con k=3, el ARI entre resultados estandarizados y no estandarizados fue de 0.21, indicando asignaciones muy diferentes. Sin estandarización, la variable Gastos (rango 54.9) domina sobre SAT (rango 4.1), resultando en 13 universidades (52%) cambiando de cluster. Esto confirma que la estandarización es esencial para balancear la contribución de todas las

dimensiones.

¿Todas las variables son necesarias?

No. El porcentaje de estudiantes aceptados resultó completamente prescindible (ARI=1.0 al eliminarla), mientras que las demás variables mostraron ser críticas:

Calificación SAT (ARI=0.26), %Percentil 90 (ARI=0.22), Gastos anuales (ARI=0.54), y Razón alumnos/profesor (ARI=0.54). La redundancia de %Aceptados sugiere que la selectividad es una consecuencia de la calidad académica y recursos, no un factor independiente.

En síntesis, aunque los datos sugieren naturalmente dos grupos, la clasificación en tres categorías es válida y consistente entre métodos. El análisis demuestra que cinco de las seis variables son necesarias para una clasificación robusta, siendo fundamental la estandarización para resultados confiables.