Análisis de clustering

Se presentan varios análisis del conjunto de datos usando clustering para agrupar usuarios.

Objetivos

- Separar los consumidores en altos consumidores, consumidores promedio y bajos consumidores para realizar tratamientos especializados.
- Encontrar patrones horarios para cada grupo para posibles estudios sobre ampliación de ancho de banda en ciertos horarios o mayor velocidad para cierto tipo de usuarios.
- Estudiar consumo vs tiempo de actividad en cada grupo para ver si es más ventajoso proponer un plan por tiempo y no por consumo (ej. nauta hogar).
- Analizar si los bajos consumidores de datos prefieren otros servicios (se encuentran entre los consumidores promedio o altos de otros servicios).
- Evaluar y optimizar el número de clusters usando Coeficiente de Silueta para determinar la segmentación óptima de usuarios (K=2 a 10).

Resultados

Objetivo 1: Separación de consumidores en perfiles especializados

El clustering K-means fue exitoso en segregar usuarios en tres categorías bien diferenciadas según su consumo de datos:

Perfil	Usuarios	Consumo Total	% del Tráfico	Consumo/Usuario
Bajo	459	25,020,000 MB	29.51 %	∼54,500 MB
Normal	422	45,440,000 MB	53.59 %	∼107,700 MB
Alto	1	14,340,000 MB	16.91 %	14,340,000 MB

Hallazgos principales:

- El perfil **Normal** domina tanto en cantidad de usuarios (45 %) como en volumen de tráfico (54 %).
- Los usuarios de perfil **Bajo** representan el 47 % de la base pero solo generan el 30 % del tráfico.
- Un único usuario de perfil **Alto** genera casi el 17 % de todo el tráfico, sugiriendo una posible anomalía o cliente empresarial.
- Estos tres segmentos justifican el desarrollo de planes y tratamientos especializados para cada grupo.

Objetivo 2: Patrones horarios por grupo

El análisis del período 00:00–8:00 horas reveló patrones de actividad diferenciados y bien definidos:

Comportamiento horario general:

- **Hora 00:00** (Pico nocturno): 24,040,000 MB, dominado por Normal (62 %) y Bajo (38 %) – 508 usuarios activos.
- **Hora 03:00** (Pico principal): 23,540,000 MB con dominio absoluto del perfil Alto (60 % del consumo).
- **Horas 06:00–07:00** (Declive): Consumo decrece a 2,300,000 MB y 269,000 MB; usuarios Bajo se vuelven dominantes (59–81 %).

Patrones identificados por perfil:

1. **Perfil Bajo**: Actividad sostenida en madrugada (00:00–04:00). Mayor concentración 04:00–07:00 (59 % del consumo horario). Usuarios consistentes pero con bajo volumen.
2. **Perfil Normal**: Distribuido en dos picos: 00:00–03:00 con 62 % participación. Mantiene presencia significativa durante todo el periodo. Más estable en relación consumo/usuarios.
3. **Perfil Alto**: Concentración extrema en hora punta 03:00 (60 % del consumo total). Ausente en otras horas. Sugiere usuario empresarial o con patrón de uso muy específico.

Oportunidades identificadas:

- Se requiere mayor capacidad de infraestructura en horario 00:00–03:00.
- Es posible optimizar ancho de banda según perfiles: mayor capacidad para Normal, menor para Bajo en off-peak.
- El usuario Alto puede requerir línea dedicada especial en horas punta.

Objetivo 3: Análisis consumo vs tiempo de actividad

Se analizó la viabilidad de proponer planes por tiempo de conexión en lugar de consumo de datos:

Perfil	Duración Promedio	Consumo Promedio	Ratio (MB/hora)
Bajo	2.37 h	54,500 MB	23,000 MB/h
Normal	10.89 h	107,700 MB	9,900 MB/h
Alto	2.31 h	14,340,000 MB	6,200,000 MB/h (anomalía)

Conclusión: SÍ es viable un modelo de planes por tiempo

- **Para usuarios Bajo**: Presentan ratio más elevado pero consistente. Plan por tiempo sería muy beneficioso para evitar penalizaciones a quienes usan mucho en corto tiempo.

- **Para usuarios Normal**: Utilizan 10+ horas con consumo variable. Plan híbrido (horas + datos) sería óptimo.
- **Para usuario Alto**: Requiere plan empresarial especial.

Recomendaciones operacionales:

- Ofrecer planes diarios/mensuales por horas de conexión sería atractivo para usuarios como los del perfil Bajo.
- Combinar límites de tiempo con límites de datos mejora la experiencia del usuario Normal.

Objetivo 4: Preferencia de bajos consumidores por otros servicios

Se realizó un estudio estadístico exhaustivo para validar si los bajos consumidores de datos tienen preferencia por otros servicios (SMS y Voz), utilizando test de hipótesis con 95 % de confianza.

Metodología:

1. **Fuente de datos:** Base de datos muestra.xlsx con 10,000 registros de múltiples servicios.
2. **Clasificación de usuarios:**
   - SERVICE_CATEGORY=5: Datos (881 usuarios únicos).
   - SERVICE_CATEGORY=2: SMS (324 usuarios únicos).
   - SERVICE_CATEGORY=1: Voz (89 usuarios únicos).
3. **Segmentación:** Aplicación de K-means en cada servicio para clasificar usuarios en Bajo/Normal/Alto.
4. **Análisis:** Test binomial unilateral con Intervalo de Confianza de Wilson al 95 %.

Hipótesis planteada: "60 % o más de los bajos consumidores de datos prefieren otro servicio" (son altos/normales en SMS o Voz).

Estudio 1: SMS - Preferencia de Mensajes

Datos del cruce:

- Bajos consumidores de DATOS identificados: **152 usuarios**.
- Usuarios que son ALTO/NORMAL en SMS: **14 usuarios (9.21 %)**.
- Usuarios que son BAJO en SMS: **138 usuarios (90.79 %)**.

Resultados estadísticos (95 % confianza):

- p-value: 1.000000 (altamente no significativo).
- Intervalo de Confianza: [5.57 %, 14.87 %].

Conclusión SMS: HIPÓTESIS RECHAZADA.

Estudio 2: Voz - Preferencia de Llamadas

Datos del cruce:

- Bajos consumidores de DATOS con datos de VOZ: **43 usuarios**.
- Usuarios que son ALTO/NORMAL en VOZ: **9 usuarios (20.93 %)**.
- Usuarios que son BAJO en VOZ: **34 usuarios (79.07 %)**.

Resultados estadísticos (95 % confianza):

- p-value: 1.000000 (altamente no significativo).
- Intervalo de Confianza: [11.42 %, 35.21 %].

Conclusión VOZ: HIPÓTESIS RECHAZADA.

CONCLUSIÓN GENERAL DEL OBJETIVO 4:
La hipótesis es FALSA en ambos servicios.
Los bajos consumidores de datos **NO prefieren otros servicios**. Por el contrario:

- Solo 9.21 % de bajos consumidores de datos son altos/normales en SMS.
- Solo 20.93 % de bajos consumidores de datos son altos/normales en Voz.
- Ambas proporciones están **lejos del 60 % hipotético**.

Interpretación: Existe un perfil de usuario claramente diferenciado:

- Usuarios "frugales" que consumen poco en TODOS los servicios (datos, SMS, voz).
- No hay compensación de consumo entre servicios.
- Los bajos consumidores de datos mantienen un patrón consistente de bajo uso en toda la plataforma.

Implicaciones para la estrategia comercial:

- Los bajos consumidores de datos NO son un segmento que busque alternativas en otros servicios.
- Estrategias de upsell basadas en ofrecer SMS/Voz a estos usuarios tendrían bajo potencial.
- Estos usuarios pueden estar orientados a aplicaciones OTT (WhatsApp, redes sociales) en lugar de servicios tradicionales.
- Se recomienda analizar patrones de uso de datos más profundamente (tipo de aplicación: streaming, redes sociales, etc.).

Objetivo 5: Evaluación y Optimización del Número de Clusters

Se realizó un análisis exhaustivo para determinar el número óptimo de clusters (K) usando el **Coeficiente de Silueta**, una métrica que combina cohesión (cercanía dentro del cluster) y separación (distancia entre clusters).

Metodología:
El Coeficiente de Silueta se define como:

$$\text{Silueta} = \frac{d_{\text{inter}} - d_{\text{intra}}}{\max(d_{\text{inter}}, d_{\text{intra}})}$$

Donde:

- d_{intra}: distancia promedio dentro del cluster (menor es mejor)
- d_{inter}: distancia promedio al cluster más cercano (mayor es mejor)
- Rango: -1 a +1 (valores >0.5 indican clustering bien definido)

Resultados del análisis de K = 2 a 10:

K	Silueta	Evaluación
2	0.4488	Clustering débil
3	0.4586	Clustering débil (K actual)
4	0.4901	Excelente
5	**0.4902**	**Óptimo – Mejor silueta**
6	0.4197	Comienza a degradarse
7–10	<0.43	Degradación progresiva

Comparativa K=4 vs K=5 (decisión crítica):

Métrica	K=4	K=5	Decisión
Silueta	0.4901	0.4902	Prácticamente idéntica (Δ=0.0001)
Distribución	Balanceada	Balanceada	K=4 gana por singletons
Singletons	1 (0.1 %)	2 (0.2 %)	K=4 es mejor
Interpretabilidad	Alta	Media	K=4 es más simple
Selección	**ELEGIDO**	Descartado	Mejor opción

Conclusión del análisis:
K=4 es el número óptimo de clusters

- Silueta es **6.9 % mejor que K=3**
- K=4 y K=5 tienen silueta prácticamente idéntica (Δ=0.0001: 0.4901 vs 0.4902)
- **K=4 tiene mejor distribución:** 1 singleton vs 2 en K=5

- K=4 es más simple operacionalmente

Nuevos perfiles optimizados (K=4):

Perfil	Usuarios	% del Total	Características
Cluster 0	396	44.9 %	Usuarios altos regulares
Cluster 1	444	50.3 %	Usuarios bajos extremos
Cluster 2	41	4.6 %	Usuarios bajos regulares
Cluster 3	1	0.1 %	Usuarios altos extremos

Beneficios operacionales de K=4:

1. **Mejor balance:** 1 singleton vs 2 en K=5.
2. **Silueta prácticamente idéntica:** 0.4901 vs 0.4902 (diferencia de 0.0001).
3. **Calidad mejorada:** 6.9 % superior a K=3.
4. **Simplicidad:** 4 clusters más fáciles de gestionar operacionalmente.

Recomendación:

Se recomienda adoptar **K=4 clusters** para futuras segmentaciones de usuarios, reemplazando el K=3 anterior. Aunque K=5 tiene silueta marginalmente mejor (0.0001 de diferencia), K=4 ofrece mejor distribución con menos singletons y es más simple de interpretar operacionalmente.