Resumen Completo de Análisis de Clusters

Estudio de Perfiles de Consumo de Usuarios

ML Final Project

January 19, 2026

Contents

1 Introducción y Contexto **3**

2 Metodología de Clustering **3**
2.1 Algoritmo Utilizado . . . 3
2.2 Justificación de K=4 . . . 3

3 Descripción de los Clusters **3**
3.1 Cluster 1: Alto Regular . . . 3
3.2 Cluster 2: Bajo Extremo . . . 4
3.3 Cluster 3: Alto Extremo . . . 4
3.4 Cluster 4: Bajo Regular . . . 4

4 Análisis Estadístico de Perfiles **5**
4.1 Estadística Descriptiva . . . 5
4.2 Variabilidad (Coeficiente de Variación) . . . 5
4.2.1 Interpretación de CV . . . 5
4.3 Tests de Significancia Estadística . . . 5
4.3.1 ANOVA (Análisis de Varianza) . . . 5
4.3.2 Kruskal-Wallis (Test No-Paramétrico) . . . 6
4.4 Correlaciones Duración-Consumo . . . 6

5 Análisis de Patrones por Tipo de Servicio **7**
5.1 Distribución por Tipo de Consumo (Datos, SMS, Voz) . . . 7
5.1.1 Consumo de Datos . . . 7
5.1.2 Consumo de SMS . . . 7
5.1.3 Consumo de Voz . . . 7

6 Análisis Temporal: Influencia por Horario **8**
6.1 Dominancia Temporal . . . 8
6.2 Contribución Total al Consumo . . . 8
6.3 Matriz de Influencia Relativa (%) . . . 8
6.4 Patrones Horarios Específicos por Cluster . . . 8
6.4.1 Alto Regular - Variabilidad Media . . . 8
6.4.2 Bajo Extremo - Variabilidad Alta . . . 9
6.4.3 Alto Extremo - Variabilidad Extrema . . . 9
6.4.4 Bajo Regular - Variabilidad Moderada . . . 9

7 Hallazgos Clave **10**

8 Conclusiones **10**

1 Introducción y Contexto

Este documento presenta un resumen ejecutivo de todos los análisis estadísticos y descriptivos realizados sobre los clusters de usuarios identificados mediante algoritmos de aprendizaje no supervisado. El objetivo es comprender los patrones de consumo de datos, SMS y voz en diferentes segmentos de población de usuarios.

2 Metodología de Clustering

2.1 Algoritmo Utilizado

Se empleó el algoritmo **K-Means** con el siguiente procedimiento:

1. **Preparación de datos**: Normalización con StandardScaler
2. **Selección de K**: Análisis de Silhouette para k = 3, 4, 5, 6, 7, 8
3. **Resultado**: K = 4 seleccionado como óptimo

2.2 Justificación de K=4

Table 1: Comparación de Silhouette Score por número de clusters

K	Silhouette Score	Singletons	Interpretabilidad	Selección
3	0.4739	Alto	Media	
4	0.4901	Bajo	**Excelente**	✓
5	0.4902	Muy Alto	Baja	
6+	< 0.48	Extremo	Pobre	

K=4 fue seleccionado por:

- Silhouette Score máximo (0.4901)
- Menor número de singletons respecto a K=5
- Mejor interpretabilidad de perfiles
- Balance óptimo entre granularidad y estabilidad

3 Descripción de los Clusters

Los 4 clusters identificados representan perfiles de consumo distintos:

3.1 Cluster 1: Alto Regular

- **Tamaño**: 396 usuarios (44.9%)
- **Consumo promedio**: 111.8 MB
- **Rango**: 113 KB a 2,710 MB
- **Características**: Heavy users con consumo consistente y predecible
- **Color**: • Azul

3.2 Cluster 2: Bajo Extremo

- **Tamaño**: 444 usuarios (50.3%)
- **Consumo promedio**: 53.2 MB
- **Rango**: 60 bytes a 1,335 MB
- **Características**: Light users ocasionales con comportamiento variable
- **Color**: • Naranja

3.3 Cluster 3: Alto Extremo

- **Tamaño**: 1 usuario (0.1%)
- **Consumo total**: 14,335.7 GB
- **Características**: Power user con consumo extremadamente alto
- **Color**: • Rojo

3.4 Cluster 4: Bajo Regular

- **Tamaño**: 41 usuarios (4.6%)
- **Consumo promedio**: 62.3 MB
- **Rango**: 827 KB a 644 MB
- **Características**: Light users regulares con patrones predecibles
- **Color**: • Verde

4 Análisis Estadístico de Perfiles

4.1 Estadística Descriptiva

Table 2: Resumen estadístico por cluster

Cluster	n	Media (MB)	Mediana (MB)	Máx (MB)
Alto Regular	396	111.81	16.08	2,710.55
Bajo Extremo	444	53.23	6.17	1,335.74
Bajo Regular	41	62.26	27.85	644.84
Alto Extremo	1	14,335.65	14,335.65	14,335.65

4.2 Variabilidad (Coeficiente de Variación)

El Coeficiente de Variación (CV) mide la predictibilidad del comportamiento:

Table 3: Coeficiente de Variación (CV) por cluster

Cluster	CV Consumo	CV Duración	CV Sesiones
Alto Regular	263.8%	25.0%	55.8%
Bajo Extremo	236.6%	89.7%	83.3%
Bajo Regular	185.3%	61.8%	62.8%

Nota: CV Alto Extremo no calculable (n=1)

4.2.1 Interpretación de CV

- **CV ¡ 50%**: Comportamiento muy predecible
- **50% ≤ CV ¡ 100%**: Comportamiento moderadamente predecible
- **CV ≥ 100%**: Comportamiento muy variable e impredecible

Hallazgos:

- Bajo Regular es el más predecible (CV=185.3% en consumo)
- Alto Regular muestra duración muy predecible (CV=25.0%)
- Bajo Extremo es el menos predecible (CV=89.7% en duración)

4.3 Tests de Significancia Estadística

Se realizaron tests para verificar si las diferencias entre clusters son estadísticamente significativas:

4.3.1 ANOVA (Análisis de Varianza)

Table 4: Resultados ANOVA

Variable	F-statistic	p-value	Significancia
Consumo Total	1,422.61	< 0.0001	***
Duración	742.63	< 0.0001	***

Ambos tests son **altamente significativos** ($p < 0.0001$), indicando que los clusters difieren genuinamente en consumo y duración.

4.3.2 Kruskal-Wallis (Test No-Paramétrico)

$$H = 54.25$$
$$p < 0.0001 \ (***)$$
Conclusión : Diferencias significativas sin asumir normalidad

4.4 Correlaciones Duración-Consumo

Análisis de la relación entre duración de sesiones y consumo de datos:

Table 5: Correlaciones por cluster (Pearson y Spearman)

Cluster	n	Pearson r	p-value	Spearman rho	p-value
Bajo Extremo	444	0.235	0.000001	0.463	0.000000
Bajo Regular	41	0.009	0.954	-0.067	0.677
Alto Regular	396	0.034	0.498	0.168	0.001

Interpretación:

- Bajo Extremo: Correlación moderada (r=0.235, significativa), lo que indica que mayor duración se asocia con mayor consumo en usuarios ocasionales
- Alto Regular: Correlación débil (r=0.034, no significativa en Pearson), pero significativa en Spearman (rho=0.168), sugeriendo relación no lineal
- Bajo Regular: Sin correlación (r=0.009), duración no predice consumo

5 Análisis de Patrones por Tipo de Servicio

5.1 Distribución por Tipo de Consumo (Datos, SMS, Voz)

5.1.1 Consumo de Datos

Table 6: Perfil de consumo de datos por cluster

Cluster	Promedio (MB)	Mediana (MB)	Desviación (MB)
Alto Regular	72.45	8.53	259.93
Bajo Extremo	36.18	3.82	85.74
Bajo Regular	41.70	17.62	80.24

Hallazgo: Alto Regular consume 2x más datos que los demás clusters.

5.1.2 Consumo de SMS

Table 7: Distribución de SMS por cluster

Cluster	Usuarios SMS	Porcentaje
Alto Regular	177	44.7%
Bajo Extremo	245	55.2%
Bajo Regular	32	78.0%

Hallazgo: Bajo Regular es más propenso a usar SMS (78% vs 45% en Alto Regular).

5.1.3 Consumo de Voz

Table 8: Distribución de Voz por cluster

Cluster	Usuarios Voz	Porcentaje
Alto Regular	148	37.4%
Bajo Extremo	134	30.2%
Bajo Regular	18	43.9%

Hallazgo: Distribución relativamente balanceada entre clusters, con Bajo Regular ligeramente más propenso a usar voz.

6 Análisis Temporal: Influencia por Horario

6.1 Dominancia Temporal

Análisis de qué cluster domina (mayor consumo) en cada hora del período analizado:

Table 9: Dominancia de clusters por hora del día

Hora UTC	Cluster Dominante	Consumo (GB)	% Dominancia	Estabilidad
00:00	Alto Regular	14.66	61.0%	Alta
01:00	Alto Regular	6.49	57.3%	Alta
02:00	Alto Regular	8.86	66.8%	Muy Alta
03:00	Alto Extremo	14.11	60.0%	Baja
04:00	Alto Regular	3.43	59.8%	Alta
05:00	Alto Regular	2.31	53.3%	Media
06:00	Bajo Extremo	1.36	59.0%	Media
07:00	Alto Regular	0.11	42.7%	Baja

6.2 Contribución Total al Consumo

Table 10: Aportación de cada cluster al consumo total

Cluster	Consumo Total (GB)	Porcentaje
Alto Regular	44.28	52.2%
Bajo Extremo	23.63	27.9%
Alto Extremo	14.34	16.9%
Bajo Regular	2.55	3.0%
TOTAL	**84.80 GB**	**100.0%**

6.3 Matriz de Influencia Relativa (%)

Porcentaje de consumo de cada cluster en cada hora:

Table 11: Influencia relativa de clusters por hora UTC (%)

Hora UTC	Alto Regular	Bajo Extremo	Alto Extremo	Bajo Regular
00:00	61.0%	36.1%	0.0%	3.0%
01:00	57.3%	41.8%	0.0%	0.9%
02:00	66.8%	29.1%	1.7%	2.3%
03:00	32.8%	6.0%	60.0%	1.3%
04:00	59.8%	32.9%	0.0%	7.3%
05:00	53.3%	37.1%	0.0%	9.6%
06:00	30.0%	59.0%	0.0%	11.0%
07:00	42.7%	40.2%	0.0%	17.1%

6.4 Patrones Horarios Específicos por Cluster

6.4.1 Alto Regular - Variabilidad Media

- **Patrón**: Consistente con picos matutinos
- **Máximo**: 00:00 UTC (14.7 GB)

- **Mínimo**: 07:00 UTC (0.11 GB)
- **Coeficiente de Variación**: ≈ 50% (Predecible)
- **Interpretación**: Heavy users con consumo regulado y predecible
- **Recomendación Operacional**: Provisionar recursos en 00:00-02:00 UTC; optimizar para streaming

6.4.2 Bajo Extremo - Variabilidad Alta

- **Patrón**: Crece hacia final del período
- **Máximo**: 01:00 UTC (9.3 GB)
- **Mínimo**: 03:00 UTC (1.4 GB)
- **Coeficiente de Variación**: ≈ 100% (Impredecible)
- **Interpretación**: Usuarios ocasionales con comportamiento variable
- **Recomendación Operacional**: Paquetes de datos flexibles; monitoreo proactivo

6.4.3 Alto Extremo - Variabilidad Extrema

- **Patrón**: Un único pico en 03:00 UTC
- **Máximo**: 03:00 UTC (14.1 GB ≈ 60% del consumo en esa hora)
- **Resto del período**: Casi cero
- **Coeficiente de Variación**: Máximo (Muy impredecible)
- **Interpretación**: Power user (1 usuario) con actividad extremadamente localizada
- **Recomendación Operacional**: QoS separado; traffic shaping; tarifa especial

6.4.4 Bajo Regular - Variabilidad Moderada

- **Patrón**: Crecimiento progresivo
- **Máximo**: 07:00 UTC (0.44 GB)
- **Mínimo**: 03:00 UTC (0.03 GB)
- **Coeficiente de Variación**: ≈ 70% (Variable pero predecible)
- **Interpretación**: Usuarios light que se activan gradualmente
- **Recomendación Operacional**: Fomentar crecimiento; planes graduales; campañas en horas pico

7 Hallazgos Clave

1. **Dominancia de Alto Regular**: Este cluster aporta más del 52% del consumo total y domina el 75% del período (6 de 8 horas), siendo crítico para la operación de red.

2. **Anomalía en 03:00 UTC**: Alto Extremo (1 solo usuario) domina completamente en 03:00 UTC, consumiendo el 60% del tráfico en esa hora. Requiere atención especial.

3. **Transición en 06:00 UTC**: Bajo Extremo supera a Alto Regular por primera vez en 06:00 UTC, indicando cambio de patrón hacia final del período.

4. **Predictibilidad**: Alto Regular es más predecible (CV bajo en duración), mientras que Bajo Extremo es altamente impredecible (CV alto).

5. **Correlación Débil Duración-Consumo**: En Alto Regular, la duración no es buen predictor de consumo, sugiriendo que otros factores (tipos de aplicación, calidad de video, etc.) son más importantes.

8 Conclusiones

Los análisis estadísticos realizados validan la existencia de 4 clusters distintivos y estables de usuarios, cada uno con características únicas de consumo y patrones temporales específicos. Los clusters no son artefactos aleatorios, sino segmentos reales diferenciados estadísticamente (ANOVA p¡0.0001, Kruskal-Wallis p¡0.0001).

El cluster **Alto Regular** es el segmento más importante estratégicamente, seguido por **Bajo Extremo**. La presencia de un único usuario en **Alto Extremo** requiere monitoreo especial.

Las recomendaciones operacionales deben adaptarse al patrón temporal específico de cada cluster, con provisioning diferenciado y QoS segregado para optimizar tanto la experiencia de usuario como la eficiencia operacional.