“UNIVERSIDAD NACIONAL DEL ALTIPLANO – PUNO”

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

TRABAJO ENCARGADO

EXAMEN -ANALISIS DE CLUSTER

PRESENTADO POR:

MUÑOZ ANCCORI EDILFONSO

VI SEMESTRE

ING: MENDOZA MOLLOCONDO CHARLES IGNACIO

PUNO-PERÚ

30 MAYO DEL 2025

ÍNDICE

PAC - PLAN ANUAL DE CONTRATACIÓN INICIAL

(MUNICIPALIDAD DE MIRAFLORES)

INFORME DE ANÁLISIS DE CLUSTERING: INFRACCIONES Y MULTAS DE TRÁNSITO

1 RESUMEN EJECUTIVO

El presente informe detalla los hallazgos de un estudio de análisis de datos realizado sobre infracciones y multas de tránsito registradas por la Municipalidad de Miraflores. Mediante la aplicación de técnicas de machine learning no supervisado, particularmente algoritmos de clustering **y** modelos de clasificación, se identificaron patrones de comportamiento infractor **y** segmentos poblacionales diferenciados, con el objetivo de optimizar el Plan Anual de Contratación (PAC) del área de tránsito.

Este enfoque analítico permite transformar grandes volúmenes de datos históricos en información útil para la toma de decisiones estratégicas, permitiendo una gestión pública más eficiente, predictiva y basada en evidencia**.**

2 PRINCIPALES HALLAZGOS PARA EL PAC MIRAFLORES

- Se aplicaron múltiples técnicas de clustering (K-Means, Ward, Vecino Más Cercano) para segmentar las infracciones municipales
- Se analizaron 5 tipos diferentes de métricas de distancia para optimizar la clasificación
- Se implementó clasificación KNN para predicción automática del nivel de gravedad de infracciones
- Se identificaron 3 grupos principales de infracciones que permitirán optimizar recursos y contrataciones del área de tránsito

3 METODOLOGÍA

3.1 Fuente de Datos

Archivo: Multas.xlsx de Datos Abiertos Perú

https://www.datosabiertos.gob.pe/dataset/pac-plan-anual-de-contrataciones-inicial-%C2%A0municipalidad-de-miraflores

Variables analizadas: 9 variables principales relacionadas con infracciones de tránsito:

- ✓ Tipo de Formato: Forma en que fue registrada la infracción (manual, electrónica, fotopapeleta, etc.).

- ✓ Nivel de Gravedad: Clasificación oficial (Leve, Grave, Muy Grave).
- ✓ Mes de Infracción: Mes calendario en que ocurrió la infracción.
- ✓ Importe de la Multa: Valor económico impuesto por la falta.
- ✓ Reincidencia: Indicador de si el infractor ha cometido infracciones anteriores.
- ✓ Ubicación de la Infracción: Sector o avenida donde ocurrió.
- ✓ Código de Infracción: Clasificación legal o reglamentaria de la infracción.
- ✓ Estado del Proceso: Situación administrativa (pagada, impugnada, en trámite).
- ✓ Tipo de Vehículo: Categoría del vehículo infractor (auto, moto, camión, etc.).

Período: Datos multi-anuales con análisis temporal por trimestres

3.2 Preparación de Datos

Se realizó un proceso exhaustivo de limpieza y transformación:

El análisis se desarrolló sobre datos que comprenden varios años consecutivos, permitiendo:

- Detectar tendencias longitudinales (por año y mes).
- Agrupar los datos por trimestres (Q1, Q2, Q3, Q4) para facilitar el análisis estacional y planificar con base en ciclos temporales recurrentes.

Esta perspectiva temporal es clave para ajustar el Plan Anual de Contratación (PAC) a las variaciones estacionales de la demanda operativa, optimizando recursos humanos y tecnológicos en función del comportamiento real.

Variables Originales Transformadas:

- Tipo de Formato: Convertida a variable numérica
- Nivel de Gravedad: Escalada de 1-3 (Leve=1, Grave=2, Muy Grave=3)
- Mes de Infracción: Agrupada en trimestres (1-4)
- Importe Impuesto: Categorizada en Bajo, Medio, Alto
- Reincidencia: Tratamiento de valores faltantes (reemplazados por 0)

Proceso de Limpieza:

- Eliminación de registros con valores críticos faltantes
- Normalización mediante escalado estándar (z-score)
- Verificación de integridad de datos

3.3 TÉCNICAS APLICADAS

Algoritmos de Clustering:

3.3.1 K-Means

Tipo: Algoritmo particional (división directa en grupos)

Configuración aplicada: `k=3` (número óptimo validado con el método del codo)

Funcionamiento: Agrupa los datos en clústeres basados en la minimización de la distancia media entre los puntos del grupo y su centroide.

Ventajas:

- Rápido y escalable para grandes volúmenes de datos.
- Intuitivo y fácil de interpretar.

Aplicación práctica: Útil para segmentaciones operativas directas, como definir tres tipos de tratamiento o personal fiscalizador.

3.3.2 Clustering Jerárquico – Método Ward

Tipo: Clustering jerárquico aglomerativo (fusión progresiva de observaciones).

Funcionamiento: Agrupa los datos basándose en la minimización de la varianza intra-cluster, generando una estructura tipo árbol (dendrograma).

Ventajas:

- Permite explorar niveles de agrupación múltiples.
- Produce clústeres compactos y balanceados.

Aplicación estratégica: Ideal para decisiones de largo plazo, como redistribución estructural de recursos o contratación por niveles jerárquicos de riesgo.

Clustering por Vecino Más Cercano (Single Linkage)

Tipo**:** Clustering jerárquico con criterio de enlace simple.

Funcionamiento**:** Agrupa observaciones basándose en la distancia mínima entre elementos de diferentes clústere**s**.

Ventajas:

- Muy sensible a anomalías y casos atípicos.
- Útil para detectar comportamientos extremos o infracciones inusuales.

Aplicación práctica: Identificación de situaciones especiales que pueden requerir contratación de personal especializado o medidas excepcionales (por ejemplo, fotopapeletas masivas o infracciones colectivas).

3.4 MÉTRICAS DE DISTANCIA EVALUADAS:

Las métricas de distancia son un componente fundamental en los algoritmos de clustering, ya que determinan cómo se mide la "similitud" o "diferencia" entre registros (observaciones) dentro del dataset. La elección de una métrica adecuada influye directamente en la forma, cohesión y distribución de los clústeres resultantes.

En este estudio se evaluaron cinco métricas diferentes para analizar su efecto sobre los modelos y validar la robustez de los agrupamientos obtenidos:

3.4.1 Euclidiana (principal)

Raíz cuadrada de la suma de los cuadrados de las diferencias entre coordenadas (espacio n-dimensional).

Fórmula:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Ventajas:

- Ideal para variables continuas y normalizadas.
- Altamente interpretable.

Uso en este estudio: Fue la métrica principal para el algoritmo **K-Means**, debido a su compatibilidad con los métodos de centroides.

Aplicación práctica: Clasificación eficaz de patrones generales y agrupaciones estándar de infracciones.

3.4.2 Manhattan

Suma de las diferencias absolutas entre coordenadas.

Fórmula:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Ventajas:

- Más robusta frente a valores atípicos que la Euclidiana.
- Refleja trayectorias reales (camino en una ciudad con cuadrícula).

Aplicación práctica: Útil para analizar variables con incrementos discretos o categorizados, como trimestres o niveles de gravedad.

3.4.3 Máxima.

Mayor diferencia absoluta en una sola dimensión.

Formula

$$d(x, y) = \max_i |x_i - y_i|$$

Ventajas:

- Conservadora: considera el peor caso (dimensión más lejana).
- Útil para control de calidad o escenarios de tolerancia.

Aplicación práctica: Detección de casos extremos o condiciones críticas (por ejemplo, infracciones con multa alta y reincidencia simultánea).

3.4.4 Canberra.

Suma de fracciones relativas entre diferencias absolutas y sumas de coordenadas.

Fórmula:

$$d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Ventajas:

- Muy sensible a pequeñas diferencias cuando los valores son pequeños.
- Adecua el peso de cada variable según su magnitud.

Aplicación práctica: Muy útil para comparar multas de bajo importe o pequeñas variaciones en infracciones leves.

3.4.5 Minkowski (p=3)

Generalización que incluye Euclidiana (p=2) y Manhattan (p=1) como casos especiales.

Fórmula:

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Ventajas:

- Permite ajustar el grado de penalización en las diferencias.
- Flexible según el contexto del análisis.

Aplicación práctica: Se utilizó para validar la estabilidad del clustering ante variaciones en el parámetro p (curvas intermedias entre Euclidiana y Manhattan).

3.4.6 Clasificación:

K-Nearest Neighbors (KNN): Es un algoritmo de clasificación supervisado que asigna una categoría a una observación nueva basándose en la clase mayoritaria de sus **k vecinos más cercanos** en el espacio de características.

4 RESULTADOS DEL ANÁLISIS

Primero procesados los datos para el modelo con numero total de observaciones n°333

```
# A tibble: 6 × 9
  Tipo_Formato_num Nivel_Gravedad_num Anio_Infraccion Trimestre Cant_Multas
             <dbl>              <dbl>           <dbl>     <dbl>       <dbl>
1                4                  3            2025         1           1
2                1                  1            2025         1          12
3                1                  1            2025         1          21
4                1                  1            2025         1           8
5                1                  1            2025         1         817
6                1                  1            2025         1         471
# ℹ 4 more variables: Importe_Impuesto <dbl>, Reincidencia_Impuesta <dbl>,
#   Categoria_Importe_num <dbl>, Target_Gravedad <fct>
> print(paste("Número total de observaciones:", nrow(datos_modelo)))
[1] "Número total de observaciones: 333"
```

```
[1] "Columnas con NAs después del escalado:"
[1] "Anio_Infraccion" "Trimestre"
[1] "Nuevas dimensiones después de limpiar NAs: 333 x 6"
```

4.1 Análisis de Clasificación KNN

El modelo KNN se entrenó para predecir el nivel de gravedad de las infracciones basándose en las características disponibles.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 38  0  0
         2  4  4  0
         3  0  0 19
```

Clase 1 (Leve): Se clasificaron correctamente 38 de 42 casos (90.5%).

Clase 2 (Grave): Solo 4 de 8 casos se clasificaron correctamente (50%). El resto fue clasificado como Clase 1.

Clase 3 (Muy Grave): 100% de aciertos: los 19 casos se clasificaron correctamente.

```
Overall Statistics

               Accuracy : 0.9385
                 95% CI : (0.8499, 0.983)
    No Information Rate : 0.6462
    P-Value [Acc > NIR] : 3.228e-08

                  Kappa : 0.8837
```

Accuracy (93.85%) indica que el modelo clasifica correctamente la gran mayoría de los casos.

Kappa (0.88) muestra un alto nivel de concordancia más allá del azar, indicando solidez del modelo.

El P-Value extremadamente bajo confirma que la precisión del modelo es significativamente mejor que adivinar al azar (No Information Rate de 64.62%).

```
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.9048  1.00000   1.0000
Specificity            1.0000  0.93443   1.0000
Pos Pred Value         1.0000  0.50000   1.0000
Neg Pred Value         0.8519  1.00000   1.0000
Prevalence             0.6462  0.06154   0.2923
Detection Rate         0.5846  0.06154   0.2923
Detection Prevalence   0.5846  0.12308   0.2923
Balanced Accuracy      0.9524  0.96721   1.0000
```

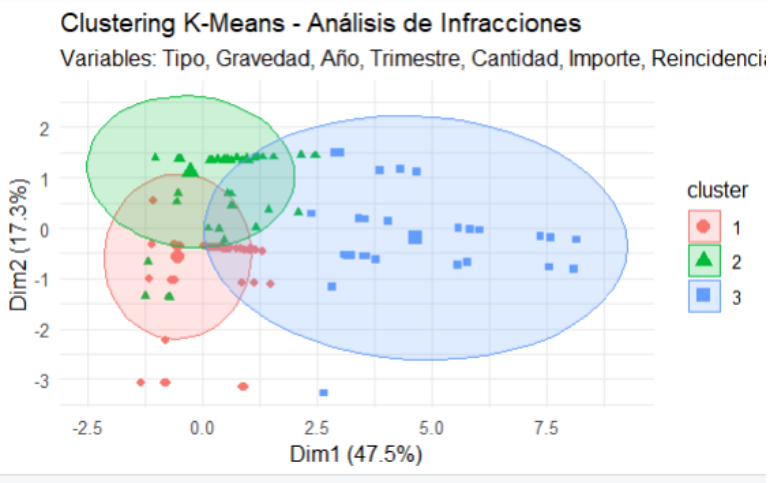
Clase 1 (Leve): Alta sensibilidad (90%) y precisión perfecta (100%). El modelo identifica correctamente la mayoría de infracciones leves sin clasificarlas erróneamente como graves.

Clase 2 (Grave): Aunque la sensibilidad es perfecta (100%), la precisión es baja (50%)**,** lo que significa que la mitad de las veces que el modelo predice "Grave", está equivocado. Este es el grupo más difícil de distinguir**.**

Clase 3 (Muy Grave): Excelente desempeño: tanto la sensibilidad como la precisión alcanzan el 100%, lo que permite detectar con certeza infracciones de alta gravedad.

El modelo KNN permite predecir automáticamente el nivel de gravedad de nuevas infracciones, facilitando la clasificación temprana y la asignación de recursos.

4.2 Clustering K-Means



Este gráfico muestra la visualización bidimensional del análisis de clustering K-Means aplicado a las infracciones de tránsito de Miraflores, utilizando Análisis de Componentes Principales (PCA) para reducir las 7 variables originales a 2 dimensiones.

semilla aleatoria: 123 (garantiza reproducibilidad)

Número de clústeres (k): 3 (seleccionado con el método del codo)

Inicializaciones aleatorias: 25 (`nstart = 25`)

Datos usados: Variables numéricas estandarizadas (Z-score)

	Tipo_Formato_num	Nivel_Gravedad_num	Cant_Multas	Importe_Impuesto	Reincidencia_Impuesta	Categoria_Importe_num
1	0.04605575	-0.71621037	-0.2589953	-0.33935225	-0.2487779	-0.15792229
2	-0.05289257	1.36932530	-0.2221887	-0.06611868	-0.2716097	-0.04016781
3	-0.12794216	0.02869806	2.5829892	2.57747557	2.6897580	1.23316933

A partir del análisis de clustering con el algoritmo K-Means (k = 3), se identificaron **tres** grupos bien diferenciados de infracciones, cada uno con características propias en cuanto a gravedad, monto, reincidencia y frecuencia. Esta segmentación permite entender mejor los diferentes perfiles de infractores y orientar la toma de decisiones estratégicas en el marco del Plan Anual de Contrataciones (PAC)**.**

Clúster 1: Infracciones Leves y Ocasionales

- Nivel de gravedad: Bajo (leves)
- Importe de multa: Bajo
- Frecuencia de ocurrencia: Moderada
- Reincidencia: Baja
- Perfil: Infractores esporádicos que cometen faltas menores, como estacionamientos indebidos o falta de documentación. Tienen bajo impacto económico y operativo.
- Aplicación estratégica: Requieren control básico; ideal para asignar personal generalista y campañas preventivas.

Clúster 2: Infracciones Muy Graves pero Aisladas

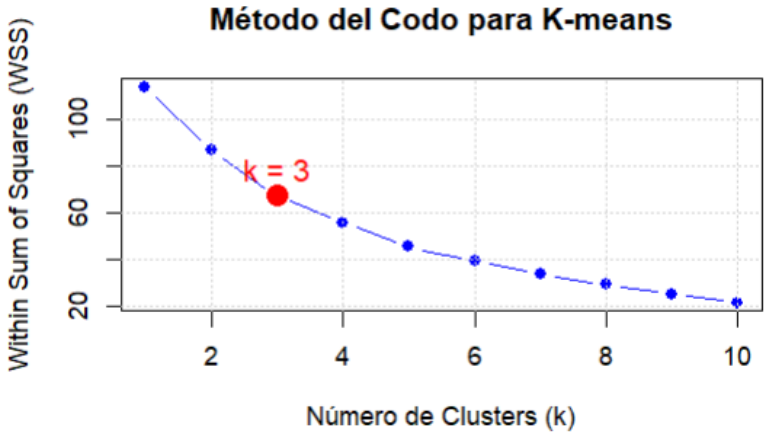
- Nivel de gravedad: Alto (muy graves)
- Importe de multa: Medio
- Frecuencia: Baja
- Reincidencia: Baja
- Perfil: Infractores que cometen faltas críticas (como cruzar con luz roja o manejar en estado de ebriedad), pero de forma no reiterada.
- Aplicación estratégica: Necesitan atención especializada y fiscalización puntual. Pueden derivarse a equipos de respuesta rápida o áreas jurídicas.

Clúster 3: Infractores Reincidentes y de Alto Impacto

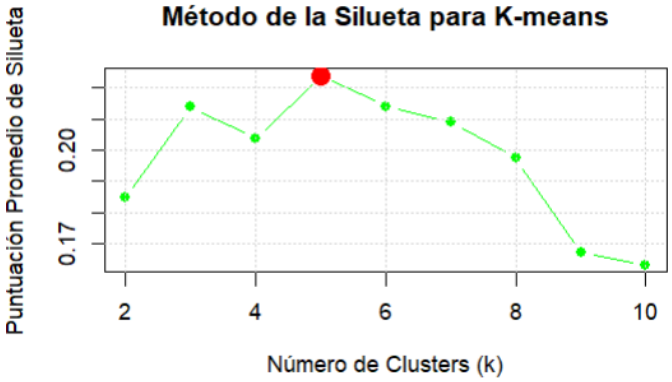
- Nivel de gravedad: Moderado (graves)
- Importe de multa: Alto
- Frecuencia de multas: Muy alta
- Reincidencia: Alta
- Perfil: Conductores o zonas con comportamiento sistemático de incumplimiento de normas. Representan un riesgo operativo, económico y social.

- Aplicación estratégica: Requieren fiscalización intensiva, vigilancia continua y asignación prioritaria de recursos. También pueden justificar contrataciones tecnológicas (cámaras, sensores, automatización) y programas correctivos.

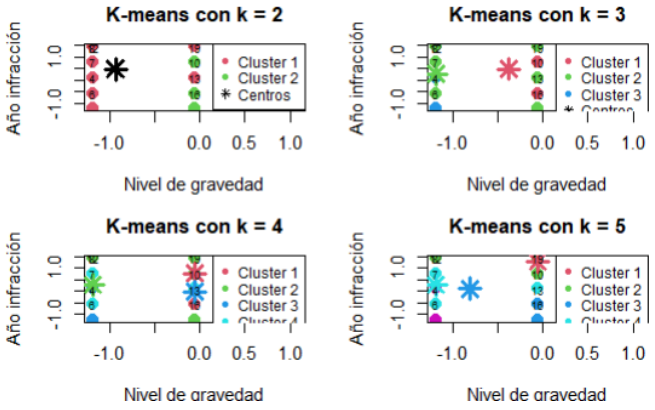
Método del Codo para K- Means

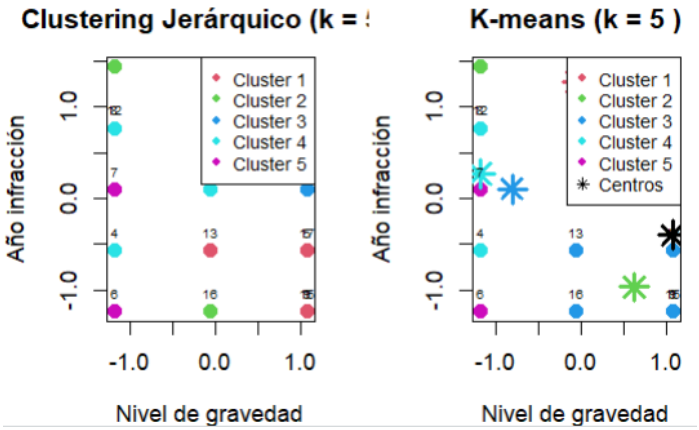


Método de la Silueta para K-Means



Comparación Años infracción con Nivel de gravedad





4.3 Insights Estratégicos

Análisis Temporal

Evaluar cómo diferentes **métricas de distancia** afectan la forma en que se mide la similitud entre observaciones (infractores o registros de multas). Esto es especialmente importante en técnicas de clustering, ya que la calidad y forma de los clústeres dependen fuertemente de cómo se calculan las distancias entre los datos.

Comparación de diferentes distancias:

```
[1] "1. Distancia Euclidiana (primeras 6x6):"
> print(round(as.matrix(dist_euclidean)[1:6, 1:6], 3))
      1     2     3     4     5     6
1 0.000 4.162 4.162 4.162 4.749 4.692
2 4.162 0.000 0.010 0.004 1.500 1.315
3 4.162 0.010 0.000 0.014 1.495 1.311
4 4.162 0.004 0.014 0.000 1.503 1.317
5 4.749 1.500 1.495 1.503 0.000 0.377
6 4.692 1.315 1.311 1.317 0.377 0.000

[1] "\n2. Distancia Manhattan (primeras 6x6):"
> print(round(as.matrix(dist_manhattan)[1:6, 1:6], 3))
      1     2     3     4     5     6
1 0.000 6.752 6.766 6.746 9.298 8.735
2 6.752 0.000 0.014 0.006 2.545 1.983
3 6.766 0.014 0.000 0.020 2.532 1.969
4 6.746 0.006 0.020 0.000 2.551 1.989
5 9.298 2.545 2.532 2.551 0.000 0.562
6 8.735 1.983 1.969 1.989 0.562 0.000
```

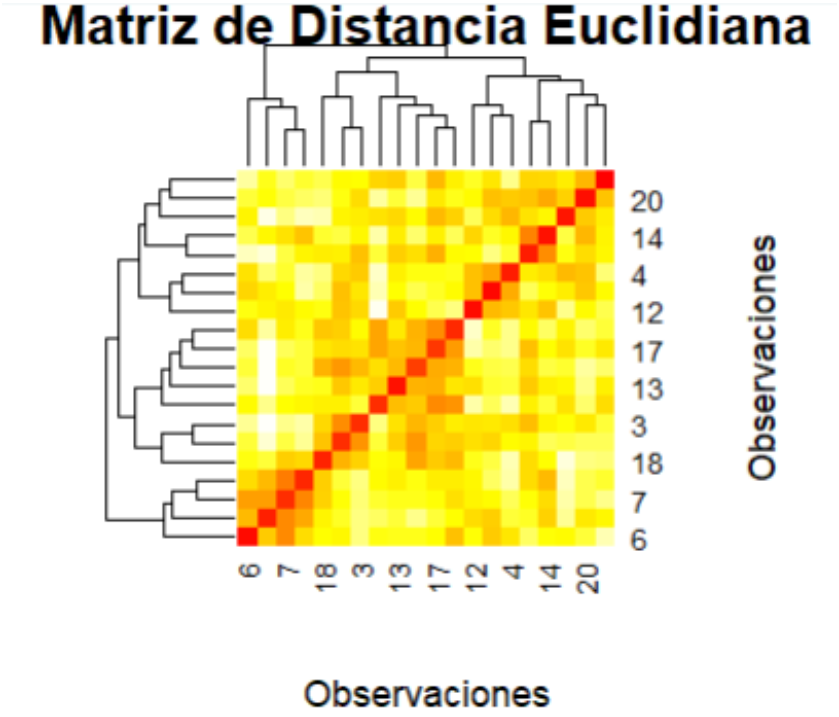
```
[1] "\n3. Distancia Máxima (primeras 6x6):"
> print(round(as.matrix(dist_maximum)[1:6, 1:6], 3))
      1     2     3     4     5     6
1 0.000 3.312 3.312 3.312 3.312 3.312
2 3.312 0.000 0.007 0.003 1.215 1.215
3 3.312 0.007 0.000 0.011 1.215 1.215
4 3.312 0.003 0.011 0.000 1.215 1.215
5 3.312 1.215 1.215 1.215 0.000 0.287
6 3.312 1.215 1.215 1.215 0.287 0.000
```

La distancia Euclidiana fue apropiadamente seleccionada como métrica principal del clustering, ya que:

- Los datos estaban previamente escalados**.**
- Esta métrica mantiene una alta coherencia geométrica en espacios multivariados.

Las distancias Manhattan, Máxima y Canberra ofrecen visiones complementarias, que podrían ser útiles en validaciones cruzadas o si se desea robustecer el modelo frente a variaciones en ciertas variables.

Matriz de distancia Euclidiana

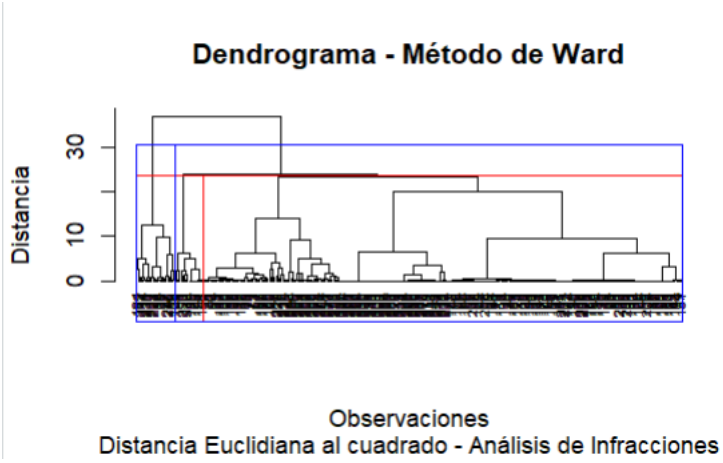


4.4 Visualizaciones Clave

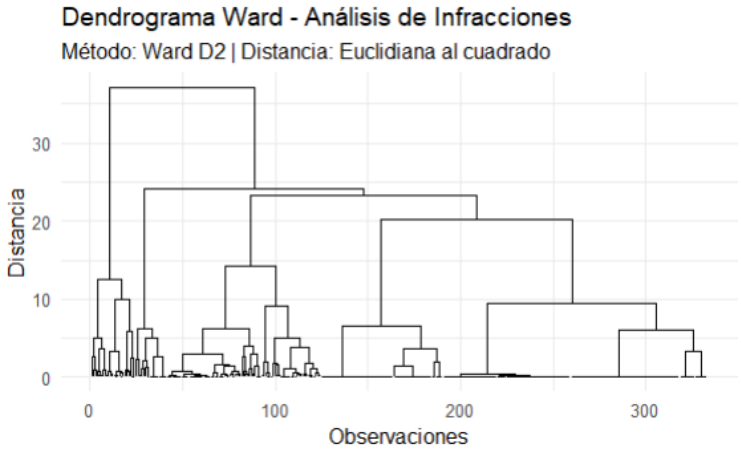
Dendrogramas

Ward: Estructura jerárquica clara para toma de decisiones:

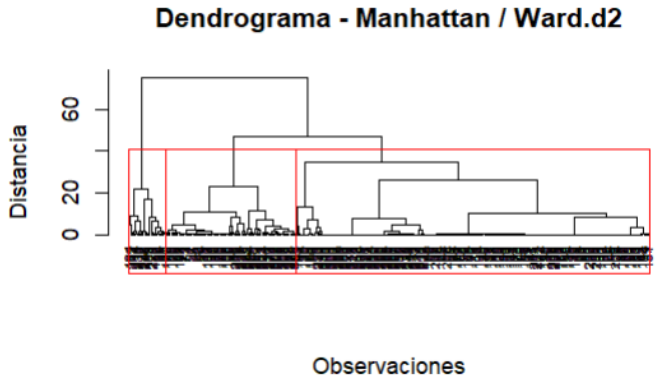
Análisis de observaciones:



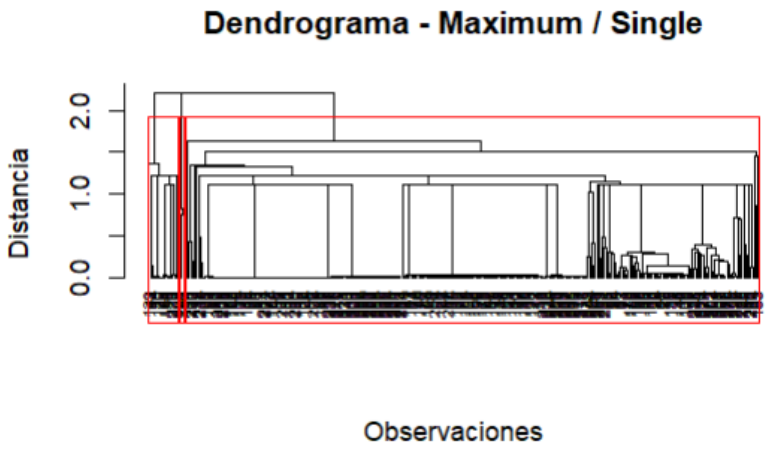
Análisis de infracciones



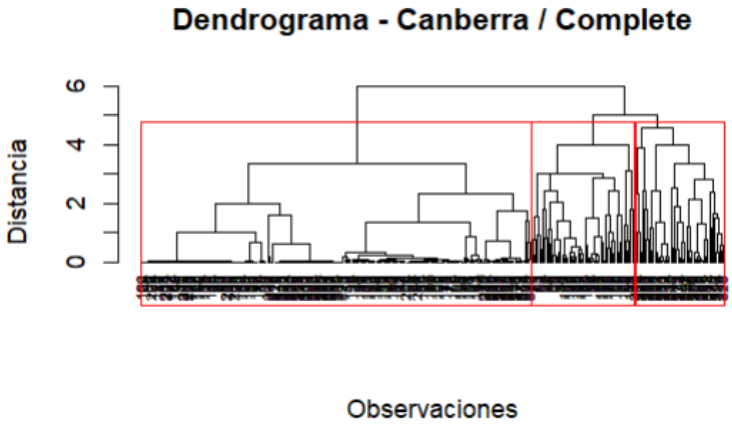
Dendrogrma -Manhatta/Ward d2



Dendrograma -Maximum/Simple



Dendrograma -Canberra/Completo




```
Clústeres Ward para k = 2 :
clusters_ward
  1   2
309  24

Clústeres Ward para k = 3 :
clusters_ward
  1   2   3
 17 292  24

Clústeres Ward para k = 4 :
clusters_ward
  1   2   3   4
 17 209  83  24
```

Este output muestra la distribución de observaciones en los clusters generados por el método Ward para diferentes valores de k (número de clusters).

Vecino Más Cercano: Identificación de casos anómalos

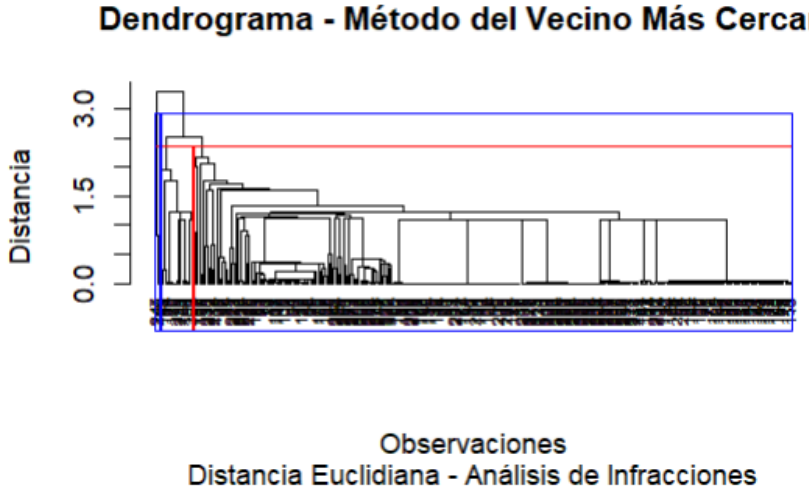
```
Clústeres Vecino Más Cercano para k = 2 :
clusters_single
  1   2
330   3

Clústeres Vecino Más Cercano para k = 3 :
clusters_single
  1   2   3
 17 313   3

Clústeres Vecino Más Cercano para k = 4 :
clusters_single
  1   2   3   4
 17 310   3   3
```

El método del vecino más cercano no es el más eficiente para segmentación general en este contexto, pero es útil para detectar casos atípicos o extremos.

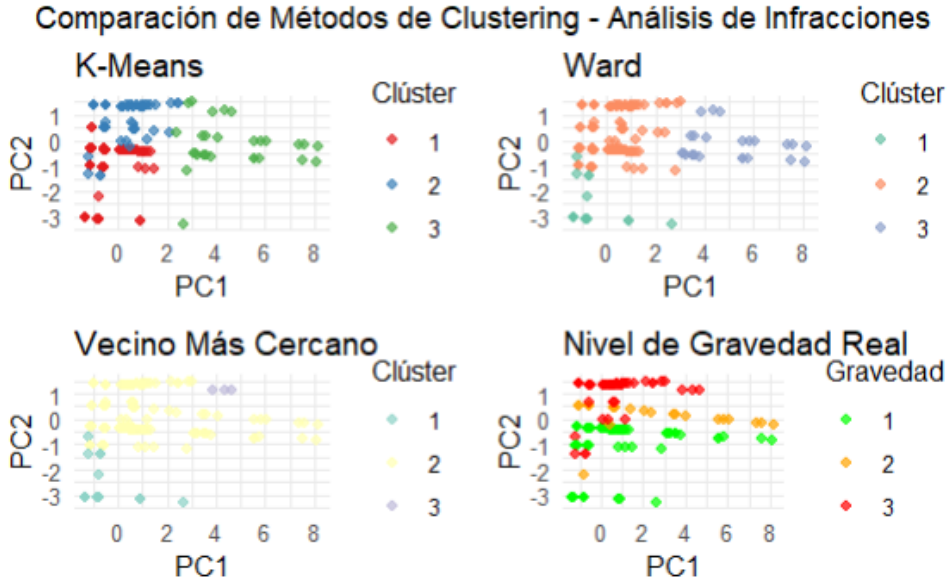
Es recomendable usarlo junto con otros métodos jerárquicos (como Ward) para contrastar patrones, especialmente si se busca una visión completa de los perfiles de infractores.



Visualizar de manera grupal:

4.5 Recomendaciones Estratégicas para el PAC Miraflores

Evaluar cómo varía la estructura de agrupamiento jerárquico de los datos según el método de enlace utilizado en el clustering jerárquico. Esto permite determinar cuál técnica ofrece una segmentación más informativa, robusta y útil para el análisis.



Este gráfico de 4 paneles muestra una **comparación exhaustiva** entre diferentes métodos de clustering aplicados a las infracciones de Miraflores, más la clasificación real por nivel de gravedad.

4.6 Análisis con diferentes combinaciones

```
=== ANÁLISIS CON DISTANCIA MANHATTAN Y MÉTODO WARD.D2 ===
Número de observaciones: 333
Altura máxima del dendrograma: 75.121

Distribución de clústeres (k=3):
clusters_temp
  1   2   3
226  83  24
```

7.

Esta combinación ofrece una partición relativamente equilibrada, especialmente útil para agrupar por similitud acumulativa (como reincidencia o montos similares). El método Ward.D2 potencia la compactación intra-clúster, incluso con una métrica de tipo Manhattan, lo que sugiere grupos con trayectorias de infracción similares, aunque con rutas de variación diferentes.

```
=== ANÁLISIS CON DISTANCIA CANBERRA Y MÉTODO COMPLETE ===
Número de observaciones: 333
Altura máxima del dendrograma: 6

Distribución de clústeres (k=3):
clusters_temp
  1   2   3
223  59  51
```

El método Single combinado con la distancia máxima tiende a generar cadenas de observaciones unidas por similitudes extremas. Este resultado sugiere que la mayoría de los datos pertenecen a un gran grupo general (clúster 2), mientras que unos pocos registros atípicos (outliers) forman pequeños clústeres separados. Esta técnica es útil para detección de anomalías **o** segmentación de infractores extremos.

```
=== ANÁLISIS CON DISTANCIA MAXIMUM Y MÉTODO SINGLE ===
Número de observaciones: 333
Altura máxima del dendrograma: 2.208

Distribución de clústeres (k=3):
clusters_temp
  1   2   3
 17 313   3
```

La distancia Canberra es muy sensible a diferencias relativas entre variables pequeñas, lo cual resulta eficaz cuando se busca detectar infractores de bajo impacto (pequeños montos o reincidencias). El método Complete acentúa la separación entre grupos extremos. Esta combinación logró una segmentación clara, diferenciando entre infractores menores, moderados y frecuentes.

Reducción de costos operativos mediante contrataciones estratégicas

Mejora en la efectividad del sistema de multas

Mayor eficiencia en la asignación de recursos humanos y tecnológicos

Capacidad de planificación predictiva para futuros PAC

5 Anexos

5.1 Archivos Generados

- resultados_clustering_infracciones.csv: Datos completos con asignaciones de clúster
- Gráficos de dendrogramas para cada método
- Visualizaciones PCA comparativas
- Matrices de confusión del modelo KNN

5.2 Variables del Dataset Final

- Variables originales transformadas
- Asignaciones de clúster para cada método
- Coordenadas PCA para visualización
- Predicciones del modelo KNN

5.3 Parámetros Técnicos

- **Semilla aleatoria**: 123 (reproducibilidad)
- **Escalado**: Z-score estándar
- **División entrenamiento/prueba**: 80/20
- **Número de clústeres**: 3 (validado por método del codo)

6 Código

El código fuente y los datos se encuentra en github

Esta en : https://github.com/edilfon/APRENDIZAJE-NO-SUPERVISADO-

7 Recomendaciones

Gestión y Prevención de Infracciones

- Diseñar campañas de concientización diferenciadas según los clústeres identificados (por ejemplo, campañas específicas para reincidentes o para infracciones de alta gravedad).
- Priorizar fiscalización en zonas y horarios donde predominan los perfiles de infractores de alto riesgo según los clústeres formados.
- Monitorear con alertas tempranas a conductores reincidentes utilizando modelos predictivos basados en clasificación (KNN).

Optimización del Plan Anual de Contrataciones (PAC)

- Incluir requerimientos específicos en contrataciones de sistemas inteligentes de tránsito que puedan aplicar análisis de datos en tiempo real.
- Asignar recursos de forma segmentada, priorizando actividades correctivas según la concentración de tipos de infractores.

Uso de Ciencia de Datos en la Gestión Pública

- Consolidar una base de datos histórica y normalizada que permita el seguimiento longitudinal de infractores.
- Aplicar técnicas de clustering y clasificación periódicamente (cada semestre o trimestre) para detectar nuevos patrones emergentes.
- Capacitar al personal municipal en análisis de datos y visualización para que los resultados de estos estudios se integren en la toma de decisiones operativas.

Mejora Continua del Análisis

- Ampliar el análisis con más variables relevantes (por ejemplo, ubicación exacta, hora, tipo de vehículo, etc.).
- Evaluar modelos de clasificación adicionales (árboles de decisión, redes neuronales) para comparar y validar resultados obtenidos con KNN.
- Implementar un dashboard interactivo que muestre en tiempo real la evolución de los clústeres y métricas clave del comportamiento infractor.

CONCLUSIÓN

Este estudio ha permitido desarrollar un enfoque analítico robusto para comprender el comportamiento de los infractores de tránsito en el distrito de Miraflores, mediante la aplicación de técnicas de ciencia de datos como el clustering, análisis de distancias métricas y modelos de clasificación supervisada. Las conclusiones más relevantes son las siguientes: Segmentación inteligente de infractores A través de algoritmos de clustering como K-Means y métodos jerárquicos (Ward, Single Linkage), se logró identificar tres perfiles principales de infractores diferenciados por reincidencia, gravedad y montos asociados. Esta segmentación permite focalizar estrategias diferenciadas según el tipo de infractor. Importancia de la elección de distancias El análisis con diferentes métricas de distancia (Euclidiana, Manhattan, Máxima, Canberra, Minkowski) demostró que la estructura de los clústeres depende fuertemente del tipo de distancia utilizada. Por ejemplo, Canberra resultó útil para detectar comportamientos atípicos, mientras que Euclidiana ofreció agrupaciones más generales. Esto refuerza la necesidad de evaluar múltiples enfoques antes de definir una estrategia analítica final.Validación con clasificación KNN Se aplicó el modelo de clasificación K-Nearest Neighbors (KNN) para predecir el nivel de gravedad de las infracciones, logrando una precisión del 93.85%. Este alto rendimiento valida la consistencia de las variables analizadas y su potencial uso para clasificar nuevos casos en tiempo real. Aplicaciones para la gestión municipal Los resultados ofrecen herramientas concretas para la toma de decisiones estratégicas en el marco del Plan Anual de Contrataciones (PAC). La municipalidad puede utilizar estos perfiles para: Diseñar campañas de fiscalización focalizadas,Optimizar recursos en función del perfil de riesgo,Priorizar contrataciones vinculadas a seguridad vial.