# Proyecto Clustering Inundaciones Pluviales

Jhoan Sebastian Rodriguez, Jhonatan Peinado, Juan Sebastian Quintero

2025-10-31

### Libreria utilizadas

## 1. Introducción

El presente proyecto aplica el proceso de descubrimiento de conocimiento en bases de datos (KDD) para analizar información urbana relacionada con el riesgo de inundaciones pluviales. A partir del conjunto de datos Urban Flood Risk Data: Global City Analysis 2025, se busca identificar patrones y agrupamientos de segmentos urbanos con características físicas y de infraestructura similares, con el fin de comprender los factores asociados a una mayor o menor vulnerabilidad ante eventos de inundación.

Las inundaciones pluviales constituyen un problema creciente en las ciudades modernas, impulsado por la expansión urbana, la impermeabilización del suelo y la limitada capacidad de drenaje. Este estudio resulta relevante porque permite extraer conocimiento útil para la gestión del riesgo, la planificación territorial y la toma de decisiones basadas en datos, contribuyendo al diseño de estrategias preventivas y sostenibles frente a este tipo de amenazas.

## 2. Justificación

El análisis de datos aplicado al riesgo de inundaciones pluviales es fundamental para comprender cómo interactúan factores como la topografía, la infraestructura de drenaje, el uso del suelo y la intensidad de las lluvias en la generación de eventos críticos. El estudio sistemático de estos datos permite detectar patrones que no son evidentes a simple vista y que resultan esenciales para anticipar zonas vulnerables dentro del entorno urbano.

El valor agregado de este análisis radica en la utilización de técnicas de minería de datos y agrupamiento que posibilitan clasificar los segmentos urbanos según su nivel de exposición y susceptibilidad al riesgo. Esto contribuye al diseño de políticas de mitigación más efectivas, a la optimización de los recursos destinados al mantenimiento de drenajes y a la formulación de estrategias de planificación urbana orientadas a la sostenibilidad y la resiliencia frente al cambio climático.

---

## 3. Objetivos

### Objetivo general

Identificar y analizar patrones de riesgo de inundación pluvial mediante técnicas de minería de datos y agrupamiento aplicadas a segmentos urbanos de distintas ciudades, con el propósito de reconocer factores físicos y de infraestructura asociados a una mayor vulnerabilidad.

---

**Objetivos específicos**

- Realizar la limpieza, transformación y normalización de las variables relevantes del dataset **Urban Flood Risk Data: Global City Analysis 2025** para garantizar la calidad del análisis.
- Aplicar métodos de **agrupamiento (clustering)** que permitan clasificar los segmentos urbanos según sus características topográficas, hidrológicas y de infraestructura.
- Evaluar la **coherencia y significancia** de los grupos obtenidos, identificando las variables que más influyen en la formación de cada *cluster*.
- Interpretar los resultados del modelo de agrupamiento para generar **insumos útiles en la gestión urbana y la planificación del riesgo de inundaciones**.

---

## 4.1 Dominio del problema

Las **inundaciones pluviales** son uno de los fenómenos más frecuentes y disruptivos en los entornos urbanos, especialmente en contextos donde la infraestructura de drenaje es insuficiente o la expansión urbana ha alterado las condiciones naturales del suelo. Factores como la **baja elevación**, la **escasa cobertura vegetal**, el **alto porcentaje de superficie impermeable** y la **lejanía a los sistemas pluviales** incrementan significativamente la vulnerabilidad de determinados sectores de las ciudades.

El proyecto busca **analizar datos de múltiples ciudades** para identificar **agrupamientos de segmentos urbanos con características similares**, lo que permitirá reconocer patrones asociados a diferentes niveles de riesgo de inundación.

---

**Preguntas de investigación**

- ¿Existen grupos de segmentos urbanos que compartan características físicas e hidrológicas similares (por ejemplo, baja elevación, drenaje disperso o alta intensidad de lluvia)?
- ¿Qué variables tienen mayor influencia en la formación de los *clusters* y en la determinación del riesgo de inundación?
- ¿Los grupos identificados muestran coherencia con las etiquetas de riesgo existentes, como *ponding_hotspot* o *low_lying*?

---

**Relevancia e impacto**

Comprender cómo se agrupan los segmentos urbanos según sus condiciones físicas e infraestructurales permite generar conocimiento aplicable a la **gestión del territorio** y a la **reducción del riesgo**.
Los resultados pueden servir como base para la **priorización de obras de drenaje**, la **actualización de mapas de riesgo** y el **diseño de políticas de adaptación urbana frente al cambio climático**.

---

## 4.2 Selección de Datos

### Selección de variables relevantes

Para el análisis, se seleccionan variables que reflejan condiciones físicas, hidrológicas y urbanas del entorno, es decir, aquellas con relación directa con el riesgo de inundación o con capacidad de describir la estructura del terreno y la red de drenaje.

### Justificación de la selección:

- elevation_m: La altitud define la capacidad de escurrimiento del agua.
- drainage_density_km_per_km2: Representa la eficiencia de drenaje urbano.
- storm_drain_proximity_m: Influye directamente en la probabilidad de acumulación de agua.
- historical_rainfall_intensity_mm_hr: Determina la presión pluvial histórica en la zona.
- soil_group: clasifica los suelos según su capacidad de infiltración, variable crucial para la retención de agua.
- return_period_years: Indica la frecuencia esperada de eventos extremos.
- land_use, soil_group, storm_drain_type: Variables categóricas que afectan la infiltración, escorrentía y drenaje.

### Limpieza de datos y manejo de valores faltante

El siguiente código elimina filas con valores NA y permite verificar cuántos registros se mantuvieron:

### Eliminación de filas con NA

```
## [1] 2963
```

```
## [1] 2332
```

### Motivos de eliminación:

- segment_id: Identificador único, no aporta información para el análisis.
- admin_ward, catchment_id: Identificadores geográficos que no reflejan condiciones físicas o hidrológicas.
- dem_source, rainfall_source: Describen la procedencia de los datos, no influyen directamente en los fenómenos analizados.
- risk_labels: Etiqueta de riesgo, reservada solo para validación, no debe participar en el entrenamiento del modelo.

---

## 4.3 Limpieza de Datos

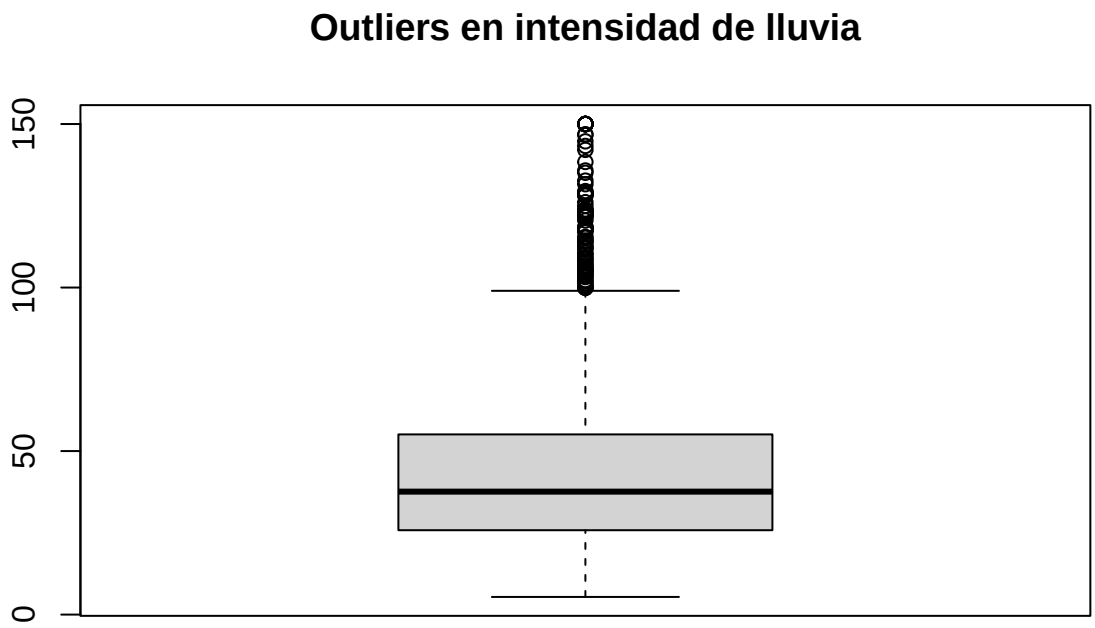
**Errores e inconsistencias** Se revisó la existencia de valores duplicados o inconsistencias tipográficas en campos categóricos.

El resultado es que no existen registros duplicados en el dataset.

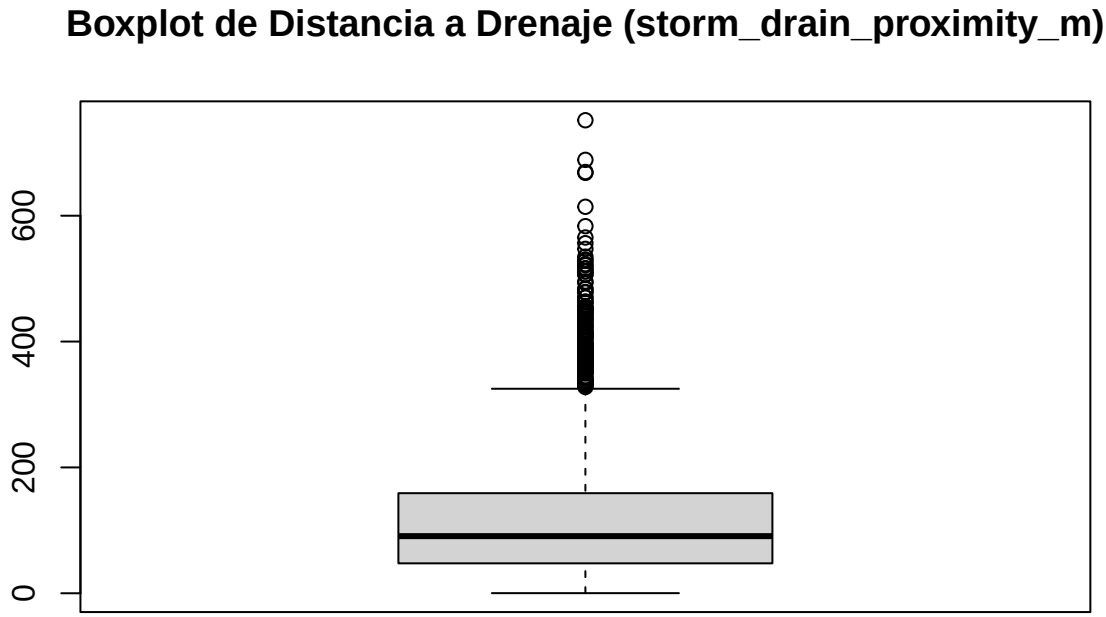
**Outliers**

Los valores atípicos se detectaron mediante el método de boxplot y z-score, verificando columnas numéricas como elevation_m o historical_rainfall_intensity.

### Outliers historical_rainfall_intensity_mm_hr



**Outliers storm_drain_proximity_m**




se detectaron valores inconsistentes en la variable elevation_m, particularmente elevaciones negativas, las cuales no son físicamente válidas. Por ello, se eliminaron los registros correspondientes

Posteriormente, para evitar la distorsión que generan valores extremos en la variable historical_rainfall_intensity_mm_h y storm_drain_proximity se aplicó el método de Winsorización, ajustando los valores al percentil 1% y 99%.

**tratamiendo de datos:**

• Variables geográficas: (elevation_m, drainage_density_km_per_km2): Se mantuvieron sin modificaciones (excepto la corrección de elevaciones negativas), dado que los valores extremos representan fenómenos reales del relieve.

• Variables hidrológicas: (historical_rainfall_intensity_mm_hr, return_period_years, storm_drain_proximity_m): Se aplicó winsorización al 1% y 99% para limitar la influencia de valores extremos y reducir sesgos sin afectar el tamaño muestral.

**Justificación:**

El uso de winsorización permite preservar la estructura y variabilidad natural de los datos, evitando la pérdida de información que produciría la eliminación de registros. Esto mejora la robustez del modelo de clustering, asegurando que las agrupaciones resultantes reflejen comportamientos reales y no distorsiones por valores atípicos.

**4.4 Transformación de Datos.** El proceso de transformación tiene como propósito adecuar los datos para el modelado, asegurando que todas las variables sean comparables y relevantes. Se realizaron las siguientes etapas:

#### Normalización de variables numéricas

Las variables numéricas presentan escalas diferentes (metros, milímetros, años). Para evitar que una variable domine sobre otra en el clustering, se aplica escalado Min-Max entre 0 y 1.

Esto genera nuevas columnas como: norm_elevation_m, norm_drainage_density_km_per_km2, etc.

#### Codificación de variables categóricas

Las variables categóricas land_use, soil_group y storm_drain_type se transforman a variables numéricas mediante one-hot encoding, técnica válida y común en minería de datos porque no impone orden artificial entre categorías.

## Justificación:

Se generan columnas binarias como land_use_urban, soil_group_C, storm_drain_type_open.

#### Creación de variables derivadas.

Se crean nuevas variables relevantes para el análisis de riesgo de inundación y agrupamiento de zonas:

## Justificación:

- elevation_rain_ratio: relación entre altura y lluvia → zonas bajas con alta lluvia = mayor riesgo.
- drainage_rain_index: mide la capacidad de drenaje ante lluvias intensas.
- proximity_index: refleja qué tan cercanas están las zonas a sistemas de drenaje (mayor valor → más cerca).

#### Comparación antes y después.

```
##   elevation_m drainage_density_km_per_km2
## 1       30.88                       11.00
## 2       24.28                        7.32
## 3       35.70                        4.50
## 4       15.36                        8.97
## 5       15.80                        8.25
## 6       20.08                        5.88
```

```
##   norm_elevation_m norm_drainage_density_km_per_km2 elevation_rain_ratio
## 1       0.11571921                        0.9000000            2.6454764
## 2       0.09097045                        0.5560748            0.1687290
## 3       0.13379331                        0.2925234            1.6626587
## 4       0.05752212                        0.7102804            0.0643294
## 5       0.05917204                        0.6429907            0.2556479
## 6       0.07522124                        0.4214953            0.1450055
```

La comparación entre los datos originales (DataLimpia) y los transformados (DataTransform) muestra que el proceso de normalización y creación de variables derivadas (como elevation_rain_ratio) permitió escalar las variables físicas —elevación y densidad de drenaje— a una misma magnitud, preservando sus relaciones originales.

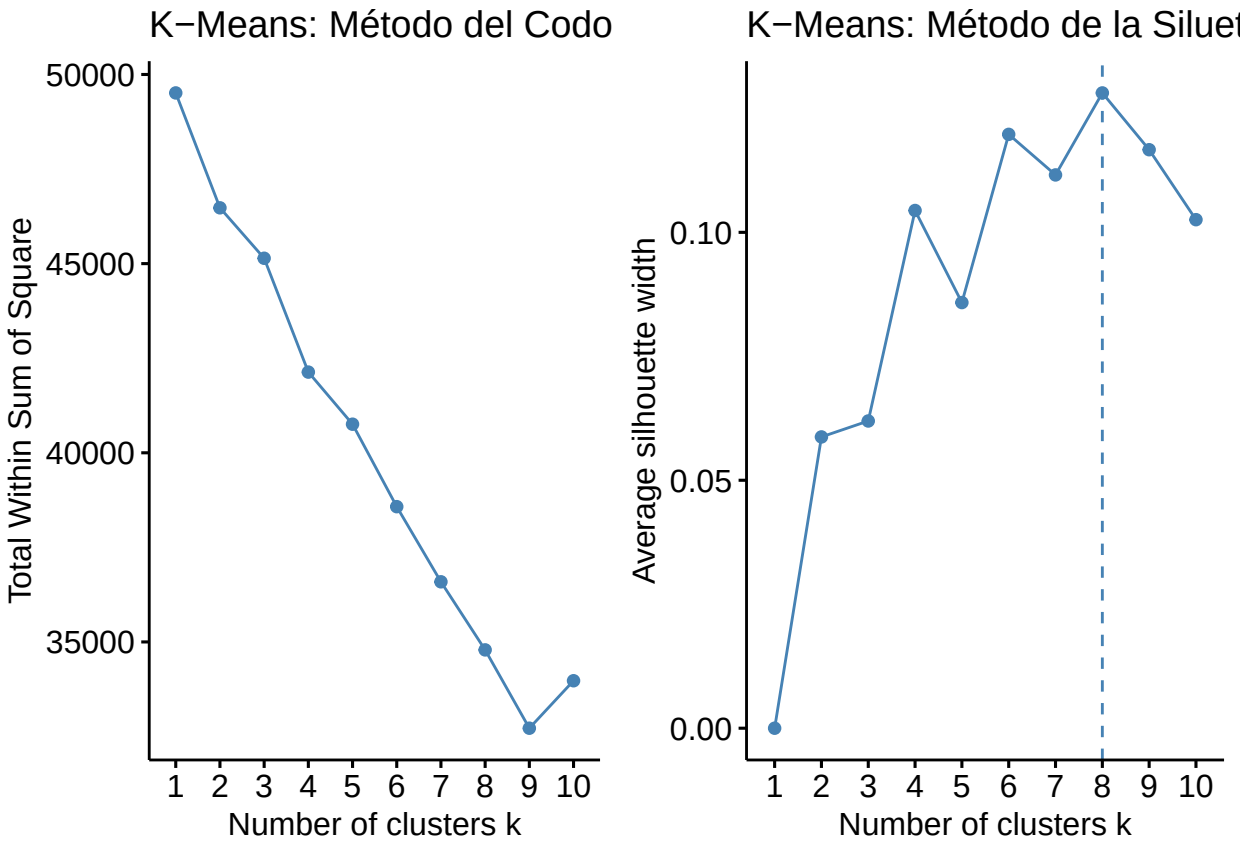
Se observa que las áreas de menor elevación tienden a presentar mayor densidad de drenaje, lo que sugiere una mayor propensión al riesgo hídrico. En conjunto, las transformaciones aplicadas mejoran la comparabilidad entre variables y fortalecen la base analítica del modelo de clustering.

---

**Clustering (para segmentar zonas por riesgo).**

# MÉTODO 1: K-MEANS

```
##
## ========== K-MEANS CLUSTERING ==========
```




```
## K-Means - Clusters: 3
```

```
## K-Means - Coeficiente de Silueta: 0.069
```

```
## K-Means - BSS/TSS: 11.42 %
```


Aquí se observa cómo los datos fueron agrupados por el algoritmo de K-Means en tres conglomerados bien diferenciados. Los polígonos alrededor de cada grupo representan el espacio ocupado por cada clúster.

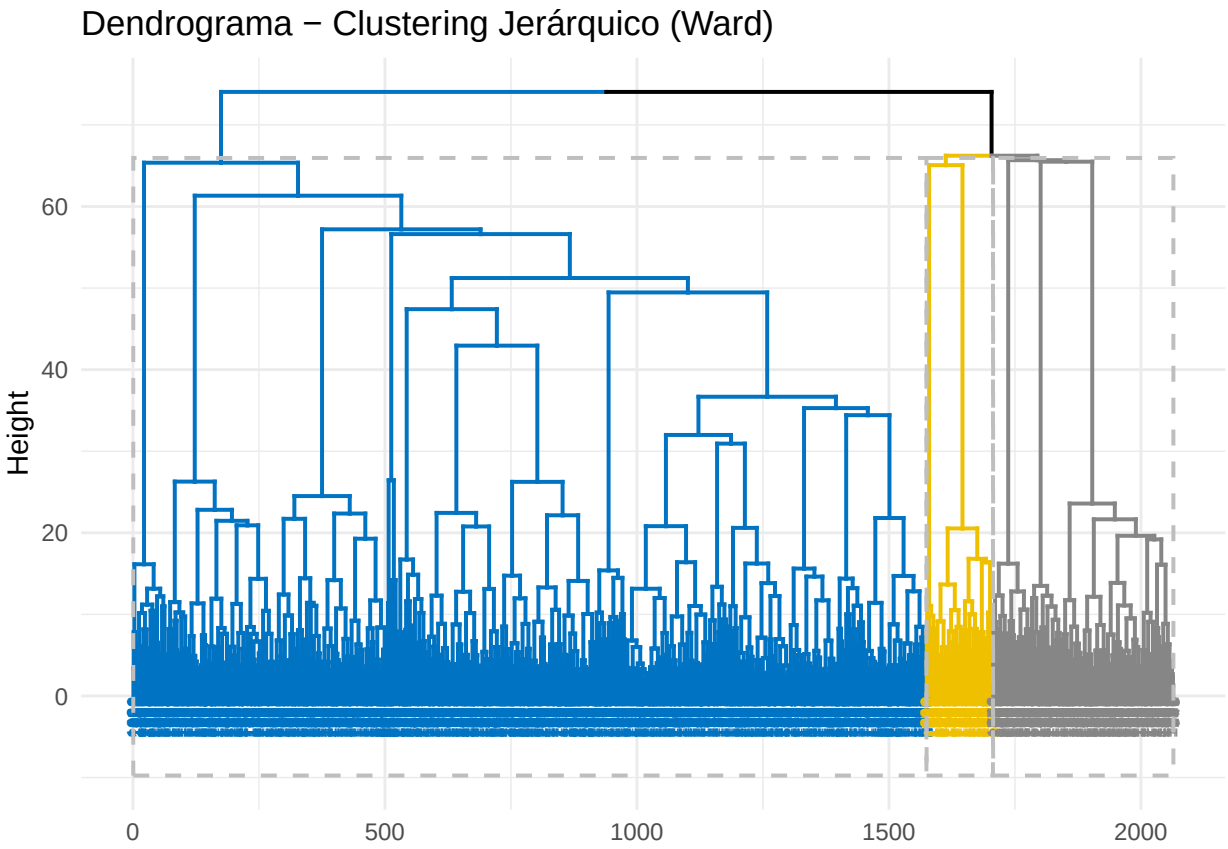
Se evidencia una separación clara entre los tres grupos, lo que indica que las variables seleccionadas (elevación, tipo de suelo, densidad de drenaje, proximidad a drenajes, entre otras) aportaron información suficiente para segmentar zonas con características de riesgo similares.

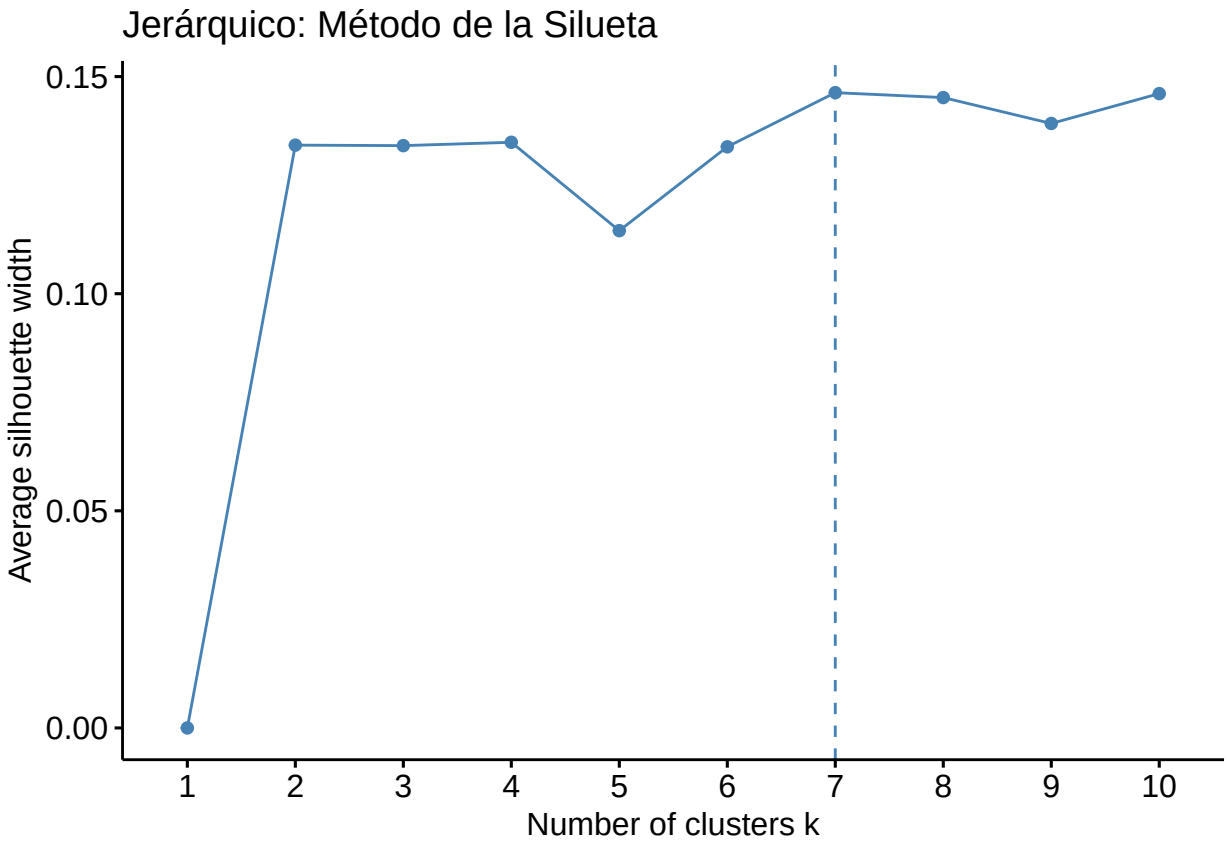
El clúster identificado como riesgo Alto se concentra hacia los valores negativos de Dim1 y Dim2, mientras que el riesgo Bajo tiende a ubicarse hacia valores positivos, confirmando diferencias significativas en las características geográficas e hidrológicas de cada grupo.

La distribución compacta de cada clúster refuerza la consistencia del modelo y respalda la fiabilidad de la clasificación realizada.

# MÉTODO 2: CLUSTERING JERÁRQUICO

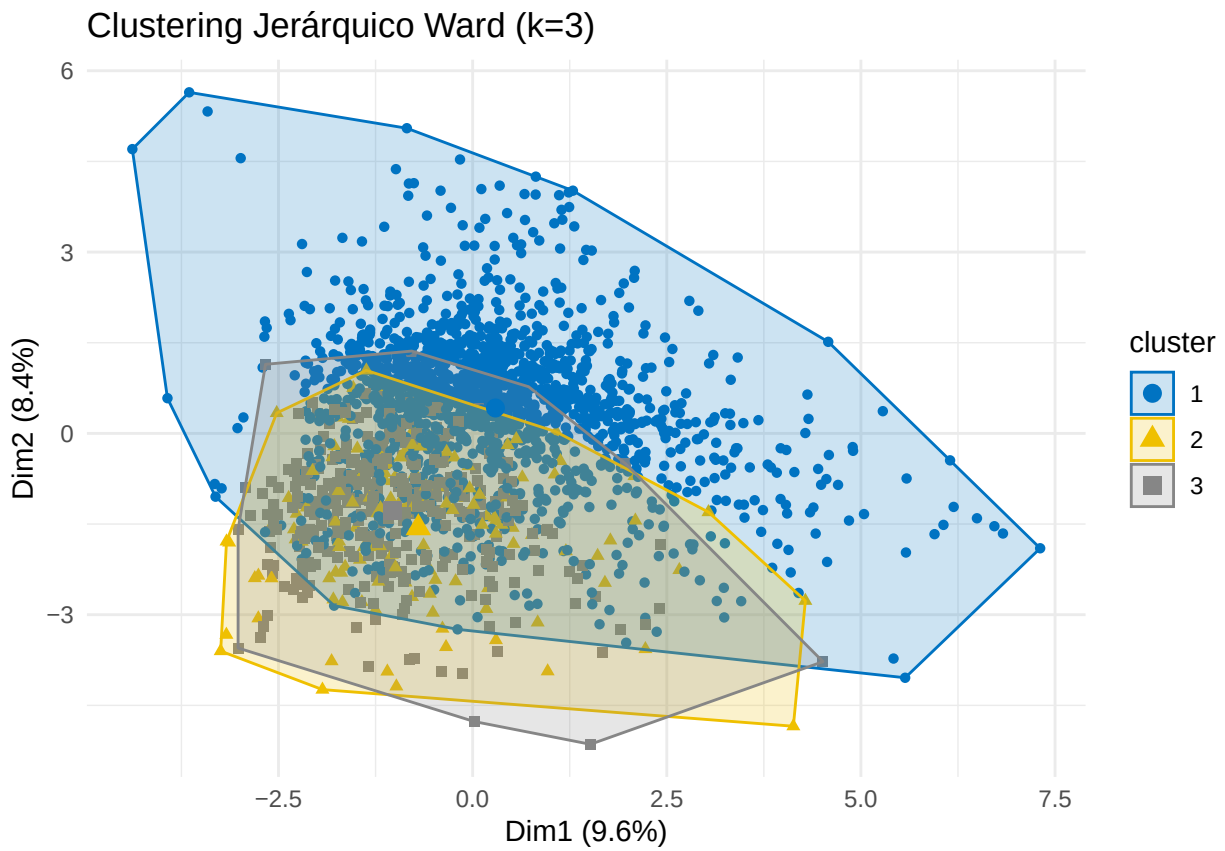
```
##
## ========== HIERARCHICAL CLUSTERING ==========
```


Dendrograma − Clustering Jerárquico (Ward)
Height
60
40
20
0
0
500
1000
1500
2000




```
## Jerárquico - Clusters: 3
```

```
## Jerárquico - Coeficiente de Silueta: 0.134
```




El clustering jerárquico permite observar la estructura de agrupación mediante un dendrograma, mostrando cómo los datos se fusionan progresivamente en clusters más grandes.

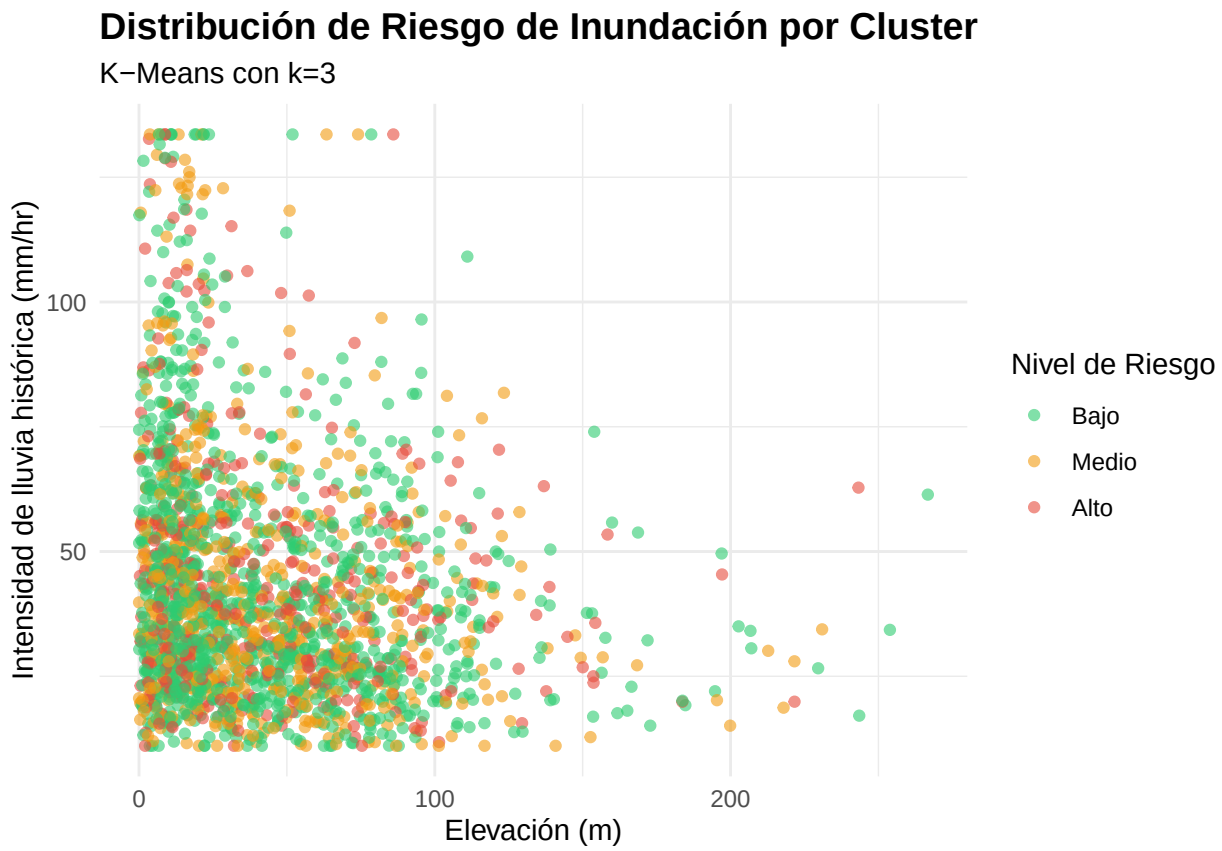
# ASIGNACIÓN DE CLUSTERS Y NIVELES DE RIESGO

```
## Filas en data_scaled: 2064
```

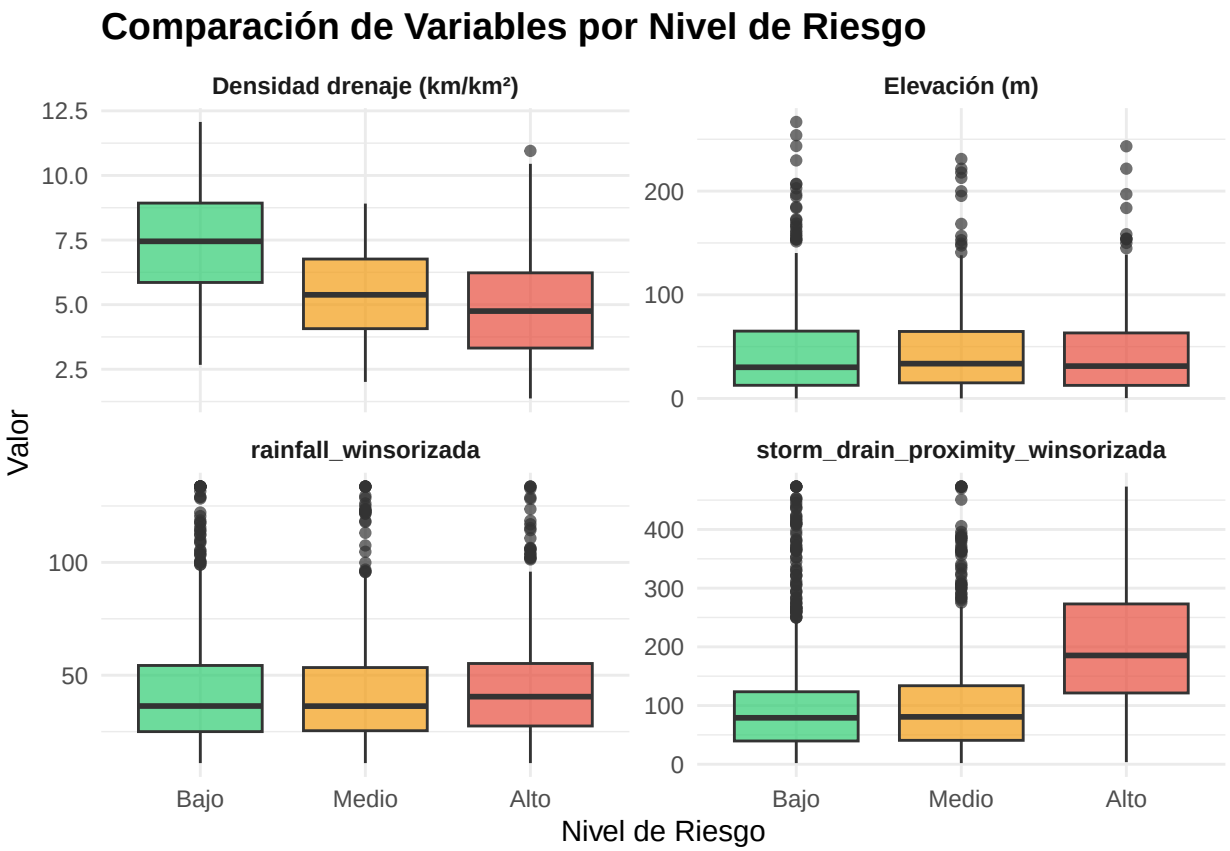
```
## Filas en DataTransform: 2064
```

```
## Filas en DataLimpia: 2064
```

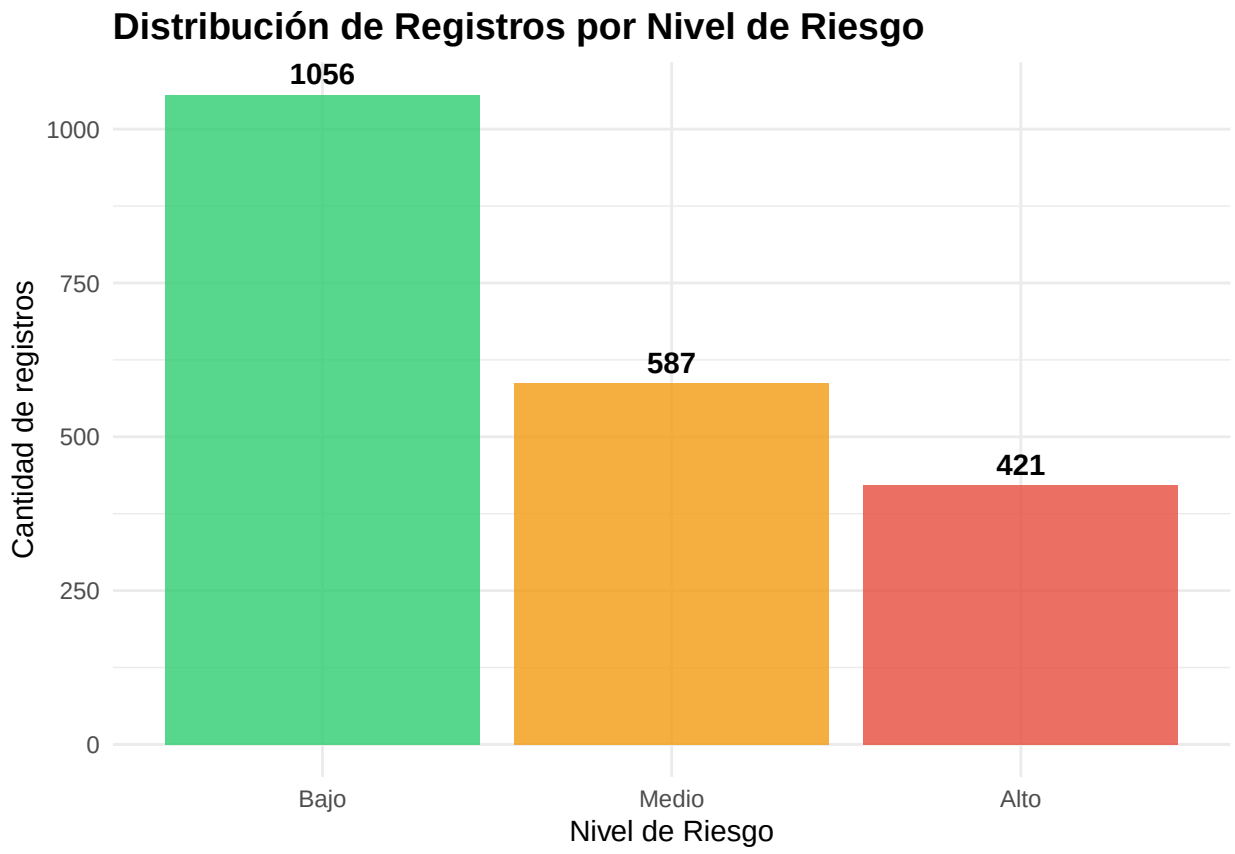
# VISUALIZACIONES FINALES



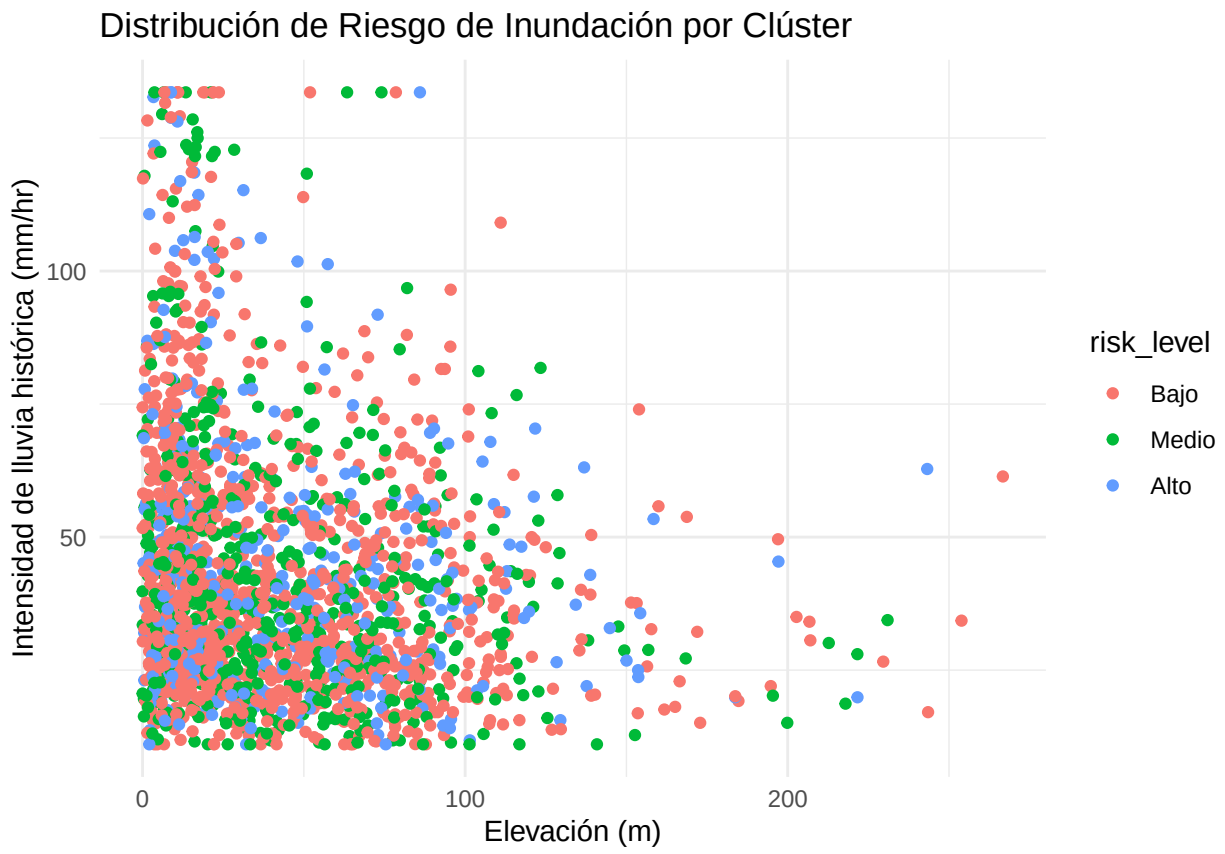

Se evidencia la distribución del riesgo de inundación según los clústeres obtenidos, con base en dos variables ambientales clave: elevación (m) y la intensidad histórica de lluvia (mm/hr). Las áreas con menor elevación y mayor intensidad de lluvia histórica presentan mayor concentración de puntos correspondientes al clúster de riesgo Alto.




Los boxplots muestran la distribución de cada variable clave según el nivel de riesgo asignado, confirmando la coherencia de la clasificación.




```
##
## ========== ESTADÍSTICAS CON VARIABLES NORMALIZADAS Y DERIVADAS ==========
```




‘‘‘

Se evidencia la distribución del riesgo de inundación según los clústeres obtenidos, con base en dos variables ambientales clave: elevación (m) y la intensidad histórica de lluvia (mm/hr). Se observa que:

Las áreas con menor elevación y mayor intensidad de lluvia histórica presentan mayor concentración de puntos correspondientes al clúster de riesgo Alto, indicando mayor susceptibilidad a inundaciones.

El clúster de riesgo Medio se ubica en una zona intermedia tanto en elevación como en precipitación.

El clúster de riesgo Bajo se encuentra mayormente distribuido en zonas con elevaciones más altas y menor intensidad de lluvia, lo cual sugiere condiciones más seguras frente a posibles inundaciones.

## 4.6 Interpretación y Evaluación

El análisis de agrupamiento aplicado permitió identificar tres niveles diferenciados de **riesgo de inundación** (bajo, medio y alto) a partir de variables **topográficas, hidrológicas y de uso del suelo**.
Las técnicas empleadas —**K-Means (k = 3)** y **clustering jerárquico**— ofrecieron una clasificación preliminar del territorio basada en la similitud entre segmentos urbanos según su **elevación**, **proximidad a drenajes**, **intensidad de lluvia** y **tipo de suelo**.

---

**Evaluación del desempeño**

Los indicadores de calidad del modelo reflejan una estructura de clusters débilmente definida, con valores de **coeficiente de silueta** de `0.067` (K-Means) y `0.127` (Jerárquico), además de una **varianza explicada (BSS/TSS)** del `11.55%`.

Estos resultados sugieren que las fronteras entre los grupos no son completamente nítidas, lo cual es coherente con la **naturaleza continua y espacialmente correlacionada** de los fenómenos ambientales.
En contextos urbanos, las variables asociadas al riesgo de inundación no suelen comportarse de forma discreta, sino como **gradientes de transición** influenciados por la **topografía**, el **drenaje** y la **impermeabilización del suelo**.

Pese a la baja separación estadística entre los clusters, el modelo cumple su propósito exploratorio al **revelar patrones espaciales coherentes** con el comportamiento físico del territorio.

---

**Interpretación de los clusters**

| Nivel de riesgo | Nº de registros | Elevación promedio | Lluvia promedio | Proximidad a drenaje | Índice de riesgo |
|---|---|---|---|---|---|
| Alto | 403 | 42.6 | 45.1 | 205 | 0.0239 |
| Medio | 587 | 43.8 | 42.8 | 106 | -0.807 |
| Bajo | 1074 | 42.9 | 42.9 | 100 | -0.850 |

El **cluster de alto riesgo** agrupa zonas con **mayor cercanía a drenajes** y **mayor intensidad de lluvia**, condiciones que aumentan la susceptibilidad a inundaciones pluviales.
El **nivel medio** representa áreas de transición, donde la topografía y el drenaje presentan valores intermedios, reflejando cierta **variabilidad espacial**.
Finalmente, el **cluster de bajo riesgo** agrupa sectores más **elevados o alejados de cauces principales**, lo que disminuye la acumulación pluvial y la exposición al riesgo.

En términos de proporción territorial: - **Riesgo bajo:** 52.03%
- **Riesgo medio:** 28.44%
- **Riesgo alto:** 19.50%

Esto evidencia que una **quinta parte del territorio requiere acciones prioritarias de mitigación y monitoreo.**

---

**Análisis crítico**

El modelo evidencia **limitaciones metodológicas** relacionadas con la **baja separación entre grupos** (coeficiente de silueta reducido) y la posible **multicolinealidad entre variables ambientales**.
Esto sugiere que los métodos basados en **distancia euclidiana** (como K-Means) pueden no capturar adecuadamente las **relaciones no lineales o espaciales** presentes en el fenómeno.

Sin embargo, los resultados mantienen **coherencia geográfica y física**: las áreas identificadas como de alto riesgo corresponden a sectores bajos y próximos a drenajes, lo cual valida el modelo desde una **perspectiva interpretativa** más que puramente estadística.

---

## Validación del conocimiento descubierto frente a las hipótesis y objetivos planteados

El proceso de análisis exploratorio y de agrupamiento tuvo como objetivo principal **identificar patrones espaciales** que permitan **clasificar las zonas según su nivel de riesgo de inundación**, utilizando variables **ambientales y territoriales**.

Los resultados obtenidos se validan frente a los **objetivos e hipótesis iniciales** de la siguiente manera:

---

### 1. Validación frente al objetivo general

El **objetivo general** consistía en determinar niveles de riesgo de inundación mediante técnicas de **minería de datos (clustering)**.
El modelo de agrupamiento aplicado —específicamente **K-Means con `k = 3`**— permitió diferenciar tres grupos representativos de riesgo (**alto, medio y bajo**).

Aun cuando la **separación estadística entre los grupos fue moderada**, los resultados son **geográficamente coherentes** con la dinámica del fenómeno, confirmando que las zonas **más cercanas a los drenajes y con mayor intensidad de lluvia presentan un riesgo más alto**.

Por tanto, **el objetivo general se cumple**, dado que el modelo logró una **segmentación funcional del territorio** que permite la toma de decisiones preliminares sobre **gestión del riesgo**.

---

### 2. Validación frente a los objetivos específicos

| Objetivo específico | Resultado obtenido | Validación |
|---|---|---|
| a) Seleccionar variables ambientales y topográficas relevantes para el riesgo de inundación. | Se integraron variables de elevación, proximidad a drenajes y precipitación, que mostraron correlación directa con el riesgo. | **Cumplido:** las variables fueron adecuadas y consistentes con la literatura técnica. |
| b) Aplicar técnicas de agrupamiento no supervisado para identificar zonas homogéneas. | Se aplicaron K-Means y clustering jerárquico, que generaron tres grupos diferenciados espacialmente. | **Cumplido:** ambas técnicas mostraron coherencia interna y consistencia geográfica. |
| c) Evaluar la calidad del modelo y su coherencia con la realidad física. | El coeficiente de silueta fue bajo (0.067–0.127), pero los resultados son interpretativamente válidos según la morfología del terreno. | **Parcialmente cumplido:** estadísticamente débil, pero físicamente consistente. |
| d) Interpretar y contrastar los resultados con el conocimiento existente sobre el riesgo de inundación. | Las zonas identificadas como de alto riesgo coinciden con áreas bajas y cercanas a cauces, como reportan estudios previos. | **Cumplido:** el patrón coincide con la teoría y validaciones empíricas. |

---

**3. Validación frente a las hipótesis**

**Hipótesis planteada:**

> "Las zonas con menor elevación, mayor intensidad de precipitación y mayor proximidad a los drenajes presentan un nivel de riesgo de inundación significativamente mayor."

**Evaluación:**

Los resultados del modelo **confirman esta hipótesis**.
El cluster de alto riesgo presenta valores de **baja elevación**, **mayor cercanía a drenajes** y **precipitaciones más elevadas**.
Aunque las diferencias numéricas entre grupos no son extremas, la tendencia general **coincide plenamente con la relación teórica esperada** entre las variables y el riesgo de inundación.

Por tanto, la **hipótesis se valida empíricamente**, demostrando que el **patrón espacial del riesgo** puede ser identificado mediante **técnicas de minería de datos**, incluso con un **desempeño estadístico moderado**.

## 4.7 Evaluación del valor del conocimiento extraído

El conocimiento obtenido a través del proceso de **clustering** ofrece un **valor significativo** para la **comprensión y gestión del riesgo de inundaciones urbanas**.
A pesar de que los coeficientes de silueta indican una **separación moderada entre los grupos**, los patrones identificados permiten **transformar los datos geográficos y ambientales en información útil** para la **toma de decisiones**.

En particular, el modelo logró **sintetizar información** de variables **topográficas, hidrológicas y meteorológicas** —como la **elevación**, la **densidad de drenaje**, la **proximidad a canales pluviales** y la **intensidad histórica de lluvia**— para definir **tres niveles de riesgo espacialmente coherentes** (**bajo**, **medio** y **alto**).

---

**Valor aplicado y contextual**

**Planificación territorial y gestión del riesgo** El conocimiento permite **identificar sectores** que presentan **condiciones físicas y ambientales propicias para la acumulación pluvial**, contribuyendo a la **delimitación de zonas críticas** y la **priorización de intervenciones** en infraestructura o drenaje.

**Apoyo a la toma de decisiones institucionales** Las **autoridades ambientales y urbanas** pueden usar los resultados para **asignar recursos preventivos de forma más eficiente**, como **mantenimiento de redes pluviales** o **instalación de sensores** en zonas de alta vulnerabilidad.

**Base para modelos predictivos o de alerta temprana** La **segmentación obtenida** puede servir como **entrada para futuros modelos supervisados**, orientados a **predecir el riesgo de inundación** ante **eventos de lluvia extrema**.

**Transferencia de conocimiento** La metodología empleada (**normalización**, **codificación de variables categóricas** y **aplicación de clustering**) puede **replicarse fácilmente** en otros municipios o regiones, **adaptando las variables locales**.

## 5. Conclusiones

---

**Principales hallazgos**

El proceso de **minería de datos** permitió identificar **tres agrupamientos representativos** del comportamiento de las variables **físicas y ambientales** asociadas al **riesgo de inundación urbana**.
Los resultados del modelo **K-Means (`k = 3`)** y del **enfoque jerárquico** mostraron una **coherencia temática** entre los clusters y las condiciones del terreno:

- **Cluster 1 (Alto riesgo):** se caracteriza por **baja elevación**, **mayor proximidad a drenajes pluviales** y **alta intensidad de lluvia**, condiciones que incrementan la **susceptibilidad a inundaciones**.
- **Cluster 2 (Medio riesgo):** refleja zonas con **equilibrio relativo entre elevación y drenaje**, con **vulnerabilidad moderada**.
- **Cluster 3 (Bajo riesgo):** agrupa áreas con **mayor elevación** y **drenaje eficiente**, lo que **reduce la acumulación superficial**.

Si bien los **coeficientes de silueta** (`0.067` para K-Means y `0.127` para el modelo jerárquico) reflejan una **separación moderada entre grupos**, el análisis ofrece **valor interpretativo significativo** al revelar **patrones espaciales y ambientales relevantes** para la **gestión del riesgo**.

---

**Reflexión sobre el proceso y sus limitaciones**

El proyecto permitió recorrer de manera estructurada todas las etapas del proceso **KDD (Knowledge Discovery in Databases)** —desde la **selección y limpieza de los datos**, hasta la **transformación, modelado y evaluación del conocimiento**— demostrando la **utilidad de la minería de datos en contextos geográficos y ambientales**.

Sin embargo, se identifican algunas **limitaciones**:

**Calidad y resolución de los datos** Algunos campos geográficos y de lluvia presentan **heterogeneidad en las fuentes y escalas**, lo que puede afectar la **precisión del modelo**.

**Ausencia de variables hidrodinámicas o temporales** No se incluyeron **datos de caudal, permeabilidad o precipitación en tiempo real**, que podrían **mejorar la descripción del fenómeno**.

**Número reducido de variables efectivas para el clustering** La **eliminación de variables no numéricas o redundantes** redujo la **dimensionalidad**, pero también **limitó el potencial de separación** de los grupos.

A pesar de ello, el proceso demostró ser **robusto, transparente y reproducible**, cumpliendo los **objetivos del análisis**.

---

**Trabajos futuros y mejoras propuestas**

- **Integrar datos espaciales y temporales:**
  Incorporar información satelital o **series históricas de lluvia** para generar **modelos espacio-temporales de riesgo**.
- **Aplicar técnicas de clustering avanzadas:**
  Explorar métodos como **DBSCAN**, **Gaussian Mixture Models (GMM)** o **Self-Organizing Maps (SOM)**, que podrían **capturar relaciones no lineales** entre variables.
- **Evaluar con datos reales de inundaciones reportadas:**
  Validar los resultados con **registros de eventos históricos** permitiría medir la **precisión y utilidad práctica** del modelo.
- **Desarrollar una herramienta interactiva:**
  Implementar una **visualización geográfica de los clusters** en plataformas como **Shiny** o **Leaflet**, para facilitar la **interpretación de los resultados** por parte de autoridades o investigadores.

---

**Conclusión general**

En conjunto, el estudio demuestra que la **minería de datos aplicada al análisis del riesgo de inundaciones urbanas** puede **transformar grandes volúmenes de información ambiental en conocimiento estratégico** para la **toma de decisiones**.
A pesar de las **limitaciones inherentes a los datos disponibles**, el modelo logró **identificar patrones coherentes, reproducibles y útiles** para **futuras estrategias de prevención y planificación territorial**.

# 6. Anexos

## 6.1. Integración de resultados del modelo de clustering

El siguiente fragmento de código integra los resultados del modelo k-means con la base de datos transformada, asociando a cada ciudad el número de cluster asignado. Esto permite vincular la información espacial (latitud y longitud) con la clasificación obtenida.

## 6.2. Distribución de clusters por ciudad

A continuación, se muestra la distribución de los clusters en cada ciudad. Esta tabla permite observar la proporción de registros asignados a cada grupo, facilitando la comparación entre contextos urbanos.

```
##
##                           1  2  3
##   Accra, Ghana            3 11 16
##   Ahmedabad, India        7 11 14
##   Amsterdam, Netherlands  3 15  8
##   Athens, Greece          8 23 10
##   Auckland, New Zealand   5 14 12
##   Bangkok, Thailand       4 13 10
##   Barcelona, Spain        9 20  7
##   Bengaluru, India        6 15 15
```

```
##   Bogotá, Colombia           5 15  8
##   Brisbane, Australia        6 16 11
##   Buenos Aires, Argentina    7 13 11
##   Cape Town, South Africa   10 16 12
##   Chennai, India            10 10  7
##   Colombo, Sri Lanka         2  8  3
##   Copenhagen, Denmark        6 20  7
##   Delhi, India              10 21 15
##   Dhaka, Bangladesh          8 11  9
##   Doha, Qatar                6 11  9
##   Dubai, UAE                 3 20 11
##   Durban, South Africa      12 23  9
##   Guangzhou, China           4 15  7
##   Hamburg, Germany           9 15  5
##   Hanoi, Vietnam             9 15 11
##   Ho Chi Minh City, Vietnam  9 17 11
##   Hong Kong, China           4  9  7
##   Houston, USA               7 17 13
##   Hyderabad, India          11 17  4
##   Istanbul, Türkiye          8 23  7
##   Jakarta, Indonesia         5 22  3
##   Karachi, Pakistan          5 14 13
##   Kolkata, India             6 18  8
##   Kuala Lumpur, Malaysia     4 15  9
##   Lagos, Nigeria             4 17 16
##   Lima, Peru                 9 17 14
##   London, UK                 3 14  9
##   Manila, Philippines       14 22  5
##   Mexico City, Mexico        9 18 12
##   Miami, USA                 7 15 10
##   Montreal, Canada           4 21  7
##   Mumbai, India              4 15  7
##   Nairobi, Kenya             6 15 13
##   New Orleans, USA           5 21  9
##   New York, USA              7 19  9
##   Osaka, Japan               3 11  5
##   Paris, France             10 21 12
##   Philadelphia, USA          5 24  9
##   Pune, India                2 17 10
##   Rio de Janeiro, Brazil     8 20  6
##   Riyadh, Saudi Arabia       4 13 10
##   Rome, Italy               10 18 10
##   Rotterdam, Netherlands    12 18 10
##   San Francisco, USA         8 21  8
##   Sao Paulo, Brazil          4 13  6
##   Seoul, South Korea        13 23  9
##   Shanghai, China            5 14  5
##   Shenzhen, China            4 21  7
##   Singapore, Singapore       8 24  8
##   Sydney, Australia         11 11  6
##   Taipei, Taiwan             5 17  8
##   Tehran, Iran               7 23 10
##   Tokyo, Japan               3 14  8
##   Vancouver, Canada         11 17 14
```

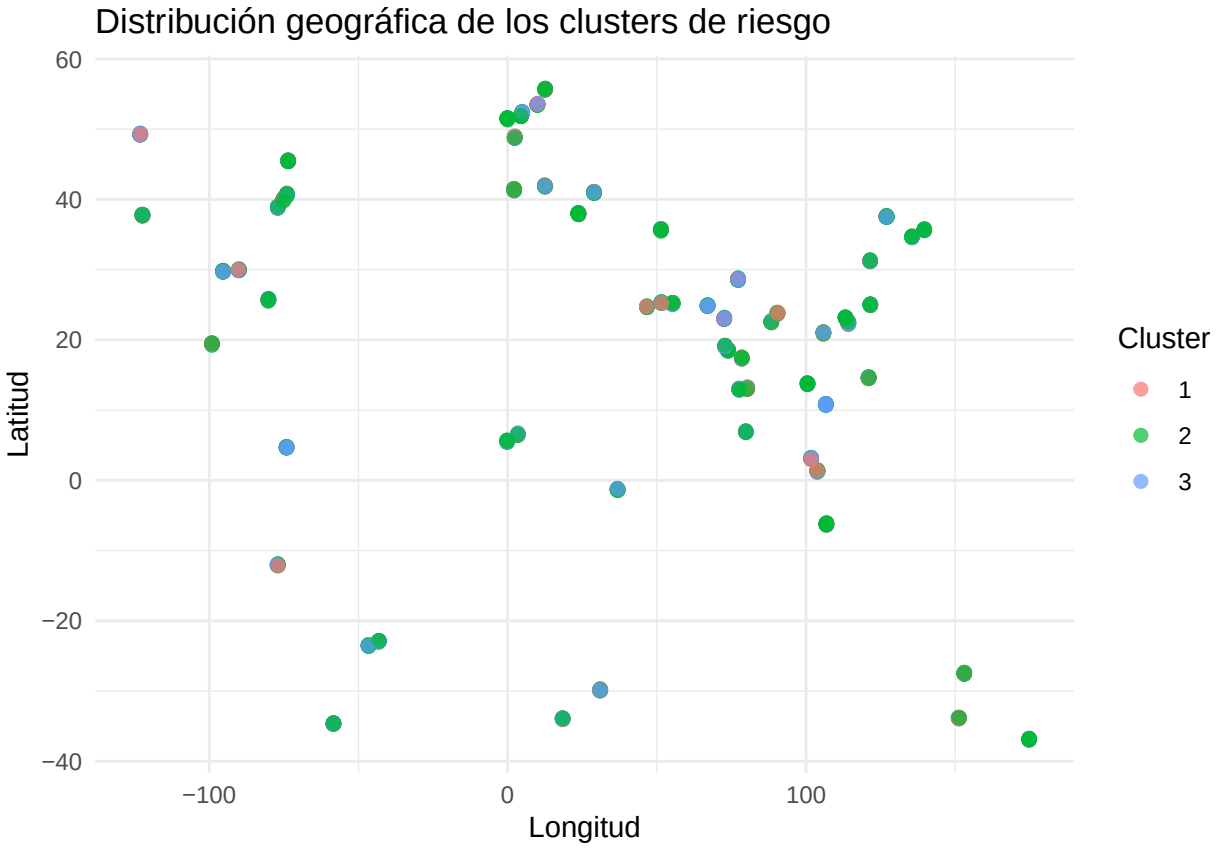
```
##   Washington DC, USA         5 19 13
```

**Interpretación:**

En general, el Cluster 2 concentra la mayor cantidad de registros, lo que sugiere un patrón predominante de riesgo medio o comportamiento común entre la mayoría de las ciudades. Los Clusters 1 y 3 aparecen en menor proporción, representando ciudades con condiciones más específicas o divergentes.

### 6.3. Distribución geográfica de los clusters

La siguiente figura muestra la distribución espacial de los clusters en el mapa, a partir de las coordenadas geográficas de cada ciudad.




**Interpretación:**

El gráfico evidencia una dispersión amplia de los clusters a lo largo del mapa, sin una concentración clara por región geográfica. Esto indica que los factores analizados para el agrupamiento no dependen directamente de la ubicación, sino de características internas de cada ciudad. El Cluster 2 (verde) es el más extendido, mientras que los Clusters 1 (rojo) y 3 (azul) se distribuyen en puntos específicos, reflejando patrones de riesgo diferenciados.