Clasificación de crímenes por zonas en la ciudad de Nueva York utilizando técnicas de aprendizaje automático no supervisado

Juan David Ceballos Sánchez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Efraín Alberto Oviedo Carrascal, Magíster (MSc) en Tecnologías de la Información y la Comunicación

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	(Ceballos Sánchez, 2023)
Referencia **Estilo APA 7 (2020)**	Ceballos Sánchez, Juan David (2023). *Clasificación de crímenes por zonas en la ciudad de Nueva York utilizando técnicas de Aprendizaje Automático No supervisado* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Elija un elemento.

Centro de Documentación de Ingeniería (CENDOI)

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes

Decano: Julio César Saldarriaga Molina

Jefe de Departamento: Diego José Luis Botia Valderrama

Dedicatoria

Se dedica este trabajo inicialmente a Dios por la vida y la oportunidad de haber llegado hasta este punto tan importante, a mi familia por el apoyo que me ha brindado en la formación de valores especialmente la perseverancia y el hacer las cosas con amor.

Agradecimientos

Agradezco a todos mis compañeros de estudio con los que compartí las clases y entregas, los espacios universitarios; agradezco a la Universidad y su cuerpo de profesores especialmente al profesor Efraín por su guianza y paciencia en este largo proceso.

Tabla de contenido

Lista de figuras

Siglas, acrónimos y abreviaturas

APA	American Psychological Association
Esp.	Especialista
KMeans	K-medias
KModes	K-modas
MSc	Magister Science
PCA	Análisis de componentes principales
UdeA	Universidad de Antioquia

Resumen

Este documento presenta los resultados de un proyecto de aprendizaje automático para analizar la criminalidad en la ciudad de Nueva York a partir de denuncias de distintos tipos de delitos registrados en un histórico de 2016 a 2019. El objetivo principal fue implementar modelos de agrupamiento que permitieran categorizar y agrupar los distintos tipos de crímenes en tres de los cinco distritos de la ciudad: Brooklyn, Bronx y Manhattan, con el fin encontrar patrones, entender y generar conocimiento y estrategias que permitan a las agencias encargadas de la ley y otros entes de seguridad desplegar sus recursos de manera más eficiente.

Para lograr lo planteado se implementaron los algoritmos KMeans y KModes y, a pesar de la complejidad del problema y la estructura de los datos, se encontró un patrón marcado en cada modelo por medio del cual se puede interpretar y caracterizar cómo son los principales tipos de crímenes en los tres distritos.

Palabras clave: aprendizaje automático, aprendizaje automático no supervisado, big data, crímenes, clúster, variables categóricas, visualización de datos.

Abstract

This document presents the results of a machine learning project aimed at analyzing crime in the city of New York based on reported incidents of various types of crimes recorded from 2016 to 2019. The main objective was to implement clustering models that would allow for categorizing and grouping the different types of crimes in three out of the five boroughs of the city: Brooklyn, Bronx, and Manhattan. The goal was to identify patterns, gain understanding, and generate knowledge and strategies that would enable law enforcement agencies and other security entities to deploy their resources more efficiently.

To achieve this, the KMeans and KModes algorithms were implemented. Despite the complexity of the problem and the structure of the data, a distinct pattern was found in each model, which can be interpreted and characterized to understand the main types of crimes in the three boroughs.

*Keywords***:** big data, categorical data, crime data, clustering, data visualization, machine learning, unsupervised machine learning.

Introducción

La ciudad de Nueva York es una de las más grandes y complejas del mundo, posee un encanto arquitectónico y financiero, sin embargo, tiene una particularidad y es que "a lo largo de los últimos doscientos años, (...) ha sido escenario de miles de delitos y crímenes de todo tipo: asesinatos fríamente calculados, robos espectaculares, matanzas mafiosas, atentados devastadores, casos memorables que contribuyeron a mejorar la metodología policial, psicópatas desatados…" (Whalen, Mladinich, & Messing, 2019).

Teniendo en cuenta que la criminalidad es un problema que afecta no solo a Nueva York, sino a muchas de las grandes ciudades del mundo, se requieren medidas efectivas para combatirla. En este contexto, este trabajo tiene como objetivo analizar los datos de criminalidad en dicha ciudad, para un conjunto de datos que recopila información entre los años 2016 a 2019, a través de la implementación de técnicas de aprendizaje automático que es una rama de la inteligencia artificial que se enfoca en el estudio y el desarrollo de algoritmos y sistemas que permiten a las computadoras aprender y mejorar su desempeño a partir de la experiencia adquirida con los datos (Alpaydin, 2016). Lo anterior, aplicado al caso de clustering, quiere decir que en cada iteración el modelo aplicado intenta mejorar sus resultados sin empeorar los que ya ha logrado, o sea, continúa generando agrupaciones de modo que se espera que los grupos creados sean óptimos. Para el tema de interés se utilizan las técnicas de KMeans y KModes para intentar lograr un perfilamiento de los tipos de crímenes en tres de los cinco distritos de la ciudad de Nueva York.

Cabe destacar que el trabajo incluye principalmente variables categóricas y el uso de la técnica KMeans funciona para variables continuas. Adicionalmente, si bien son algoritmos que trabajan con medidas de tendencia central, al compararlos se debe tener cuidado en las interpretaciones y comparaciones entre ellos ya que, por ejemplo, según los autores (Hamzah, Kek, & & Saharan, 2017)[1], el KMeans utiliza la suma de cuadrados y el cálculo del valor medio para determinar los centroides y realizar el agrupamiento. En cambio, el KModes utiliza la distancia de coincidencia simple y el valor de moda como centroides. Además, el modelo KMeans brinda mejores resultados que el KModes en conjuntos de datos numéricos. Sin embargo, es importante considerar las

[1] Consultado en: https://penerbit.uthm.edu.my/ojs/index.php/JST/article/view/2038

características y la naturaleza de los datos antes de seleccionar el algoritmo de agrupamiento más adecuado para un determinado problema

Más allá del resultado, este estudio muestra un intento por lograr la agrupación de crimines con ambas técnicas y contribuir a futuros estudios en el área tanto de Aprendizaje Automático como de seguridad.

1. Planteamiento del problema

Esta monografía aborda el desafío de identificar patrones y generar información relevante a partir de los datos históricos de denuncias proporcionados por el Departamento de Policía de la ciudad de Nueva York para diversos organismos encargados de la seguridad y el cumplimiento de la ley en la ciudad. Dada la naturaleza cosmopolita de la ciudad y la amplia fuente de datos disponible, se registra un alto volumen de denuncias que abarcan desde infracciones menores de tránsito hasta delitos graves como homicidios y secuestros.

Uno de los objetivos principales es lograr la caracterización de diferentes grupos dentro de cada distrito (Brooklyn, Bronx y Manhattan). Esto permitirá identificar similitudes y diferencias entre ellos, lo cual resultará en estrategias de seguridad más efectivas, adaptadas a las necesidades específicas de cada grupo.

1.1 **Antecedentes**

Teniendo en cuenta el enfoque experimental de este trabajo y que la ciencia de datos es un campo que ha tomado un auge importante en las últimas dos décadas, se encuentra en repositorios de GitHub y en la comunidad en línea Kaggle, algunas referencias de ejercicios que se han desarrollado basados en el mismo conjunto de datos que se ha utilizado para la investigación. Esto es algo positivo puesto que cada uno de los ejercicios y planteamientos desarrollados aporta a la ciencia y hace que investigaciones más recientes, como esta, tengan un punto de partida y también un punto de referencia para comparar resultados. Por ejemplo, un ejercicio realizado en Kaggle (Feder, 2018) que propone un análisis de tipo exploratorio de datos, se concluye que el Distrito donde mayores denuncias se han registrado es en Brooklyn. En los resultados de esta investigación se encuentra que, efectivamente, lo anterior es cierto, incluso se logra una categorización en el Distrito de Brooklyn donde se logran clasificar algunos de los barrios de este distrito donde hay mayor y menor incidencias criminales siendo que hay características como la raza y la hora del día que influyen en el total de incidentes reportados.

En otro estudio realizado y alojado en RPubs, el autor (Zhou, 2019) aplica técnicas de clustering para responder a la pregunta de dónde deberían estar localizadas las estaciones de policía basado en los hechos reportados. Si bien en este estudio se aplican varias técnicas de aprendizaje no supervisado como KMeans, DBSCAN y otros, el estudio se basa mayormente en frecuencia de crímenes por locación en el mapa (georreferencia) de la ciudad y en encontrar la probabilidad de que dado un lugar ocurra una denuncia en particular. Con lo anterior, el estudio concluye que para el conjunto de datos podrían generarse grupos de 20 clústeres y así, alrededor de los mismos, basar las estaciones de policía. El estudio logra identificar ciertos patrones principalmente con el uso del algoritmo DBSCAN, pero no hay un detalle profundo en esto. Se toma también como referencia para la construcción de mapas de calor y de los clústeres dentro del mapa de la ciudad.

Otra referencia (Reddy, 2019) que analiza temas de seguridad y crimen en la ciudad para el año 2021, es un ejercicio de big data que realiza un análisis descriptivo del conjunto de datos para aplicar estrategias de limpieza y tratamiento datos (problemas que también fueron encontrados con el conjunto de datos de 2006 a 2019). Los esfuerzos del ejercicio para 2021 se centraron en saber qué tan confiables y limpios se encuentran los datos para ser usados en análisis posteriores y se concluye que parte de los datos contienen, no sólo datos vacíos que producen ruido en los modelos

de aprendizaje automático, sino que también intentan ligar su ejercicio de análisis a conjuntos de datos similares tales como: Arrestos en la ciudad de Nueva York, Incidentes relacionados a Tiroteos en la ciudad, Crímenes de Odio en la ciudad, Choques de Tránsito en la ciudad que son conjuntos de datos abiertos que se encuentran también disponibles en NYC Open Data (https://opendata.cityofnewyork.us/) que es de donde se obtiene el conjunto de datos de base para éste trabajo.

2. Justificación

Existe una motivación por abordar este tipo de temática de carácter socioeconómico ya que la seguridad es una problemática no sólo de una ciudad en particular sino algo que afecta el mundo entero y no es un suceso coyuntural, sino estructural intrínseco al ser humano. Así, desde un enfoque económico, "una persona comete un delito si la utilidad esperada para él excede la utilidad que podría obtener usando su tiempo y otros recursos en otras actividades" (Becker, 1968). Lo anterior, es una concepción fundamentada en un enfoque económico neoclásico en la que los individuos son racionales y basan sus decisiones dependiendo de la utilidad (ganancia) esperada de sus acciones. Fue esta investigación piedra angular en materia y aunque es interesante poder abordar la temática en términos de dinero y ver cuánto les cuesta el crimen a diferentes instituciones, éste estudio es más de tipo experimental para probar y entrenar modelos basados en el conjunto de datos disponibles.

Generalmente estos problemas están a cargo de intervenciones político-económicas donde el enfoque es más teórico que práctico y donde se basan los esfuerzos en el uso de técnicas estadísticas como la espacial o a discreción de decisiones políticas. Al aplicar un enfoque empírico basado en experimentos y en evidencia a través del uso de distintos algoritmos que provean patrones que lleven a generar información a través de los datos, el problema puede ser resuelto con mayor precisión y, entraría, además, el papel crítico por parte del autor en plantear estrategias basadas en evidencia que respalden tales sugerencias. De este modo, se pretende impactar la capacidad de respuesta y la eficacia de las agencias encargadas de la seguridad y los cuerpos policiales en la prevención y disminución de la criminalidad en la ciudad de Nueva York.

El uso de los algoritmos KMeans (Morissette & Chartier, 2013), que ha demostrado ser simple y elegante en la manera de particionar los datos, y KModes (Dewia & Dwidsamara, 2021), que fue introducido por primera vez por Huang en 1998 como un método de agrupamiento modificado a partir del método k-means. K-means es un algoritmo no supervisado simple, cuyos resultados son bastante buenos para problemas generales de agrupamiento. Sin embargo, k-means generalmente trabaja con atributos o datos con valores numéricos, no valores categóricos. Por lo tanto, se modificó el k-means para poder ser utilizado con datos categóricos. Mediante el uso de estos algoritmos se aborda el problema de análisis de la criminalidad ya que son modelos que han sido ampliamente utilizados en el campo del aprendizaje automático y la minería de datos,

demostrando su eficacia en la identificación de patrones y agrupamiento de datos. Queda a disposición de los entes encargados de la seguridad y ley en la ciudad de Nueva York el utilizar los resultados arrojados por estos modelos y basar su toma de decisiones en ellos de modo que su actuar pueda ser más rápido y preciso

En un contexto en constante evolución tecnológica, donde los desafíos actuales difieren significativamente de los del pasado, este trabajo ofrece un enfoque innovador al presentar alternativas como el uso de mapas de calor que se encuentran principalmente en los notebooks almacenados en el repositorio que acompaña la investigación. Este mapa permite comparar y analizar la relación y comportamiento entre diversos grupos, ubicaciones y eventos a través de distintas variables. Este enfoque proporciona una nueva perspectiva para comprender y abordar de manera más efectiva los fenómenos estudiados.

Queda fuera del alcance de este estudio la creación de gráficos utilizando el croquis de la ciudad de Nueva York para identificar visualmente las zonas con mayor incidencia de delitos. Sin embargo, se sugiere que esta línea de investigación pueda ser explorada en futuros trabajos y contribuir a la literatura existente. La identificación de las áreas con mayor frecuencia de delitos proporcionaría un valioso apoyo a las autoridades en la asignación de recursos y la implementación de estrategias de seguridad, permitiéndoles focalizar sus esfuerzos en aquellas zonas que requieren mayor atención y ayuda.

3. Objetivos

3.1 **Objetivo general**

Identificar patrones y agrupar delitos en un conjunto de datos históricos de denuncias de la policía de la ciudad de Nueva York aplicando técnicas de aprendizaje automático no supervisado como herramienta de apoyo a la toma de decisiones en cuanto a la prevención y disminución de la criminalidad en la ciudad.

3.2 **Objetivos específicos**

- Implementar el uso de diferentes tipos de algoritmos como KMeans y KModes de aprendizaje automático para ver su comportamiento, comparar resultados y lograr la categorización de los crímenes en la ciudad de Nueva York y a través de un perfilamiento en los hechos.
- A través de los resultados obtenidos, proponer una estrategia en la cual se pueda impactar la prevención y, por ende. la disminución de los hechos en la ciudad.

4. Marco teórico

El agrupamiento en el aprendizaje automático no supervisado es una técnica de análisis de datos que se utiliza para identificar patrones y estructuras ocultas en un conjunto de datos. El objetivo es dividir el conjunto de datos en grupos o clústeres que sean similares entre sí, pero diferentes de otros grupos. Existen diferentes (Müller & Guido, 2017) enfoques para el agrupamiento, como el agrupamiento jerárquico, el agrupamiento basado en densidad y el agrupamiento basado en prototipos. Cada uno de estos enfoques tiene sus propias ventajas y desventajas, y la elección del método adecuado dependerá del conjunto de datos y del objetivo del análisis. KMeans y KModes son métodos comunes de agrupamiento no supervisado utilizados para alcanzar este objetivo.

KMeans es un algoritmo de agrupamiento que se basa en prototipos para dividir un conjunto de datos en k clústeres predefinidos. Su objetivo es minimizar la distancia euclidiana entre los puntos de datos y los centroides de los clústeres. Aunque la mayoría de los algoritmos de agrupamiento utilizan la distancia euclidiana, también existen otras medidas como la distancia Manhattan y la distancia coseno, esta última comúnmente utilizada en el procesamiento de texto. Por otro lado, el algoritmo KModes es una extensión del algoritmo KMeans (López, 2007), diseñado especialmente para manejar conjuntos de datos categóricos, como las descripciones de los crímenes y sus características asociadas. En lugar de utilizar la distancia euclidiana, KModes utiliza una medida de disimilitud que busca encontrar modos o categorías donde las quejas sean lo más similares posible en términos de sus atributos. La representación de cada grupo se basa en las frecuencias de las características presentes en las denuncias. Este enfoque adaptado del algoritmo KMeans permite agrupar las denuncias de crímenes en categorías distintas, facilitando la identificación de patrones similares. Al utilizar la medida de disimilitud, se busca encontrar grupos de quejas con características similares, pero no necesariamente cercanas en términos de distancia euclidiana.

En general, el proceso de agrupamiento consta de tres etapas: selección de características, definición de la medida de similitud y aplicación del algoritmo de agrupamiento. La selección de características implica identificar las variables más relevantes para el análisis. La medida de similitud se utiliza para calcular la distancia entre los puntos de datos y determinar la similitud entre ellos. Finalmente, el algoritmo de agrupamiento se aplica para dividir el conjunto de datos en

grupos. Este proceso (Bishop, 2006) puede ser muy útil para descubrir patrones ocultos en grandes conjuntos de datos, lo que puede llevar a una mejor comprensión de los datos y a la identificación de posibles relaciones entre las variables.

Es relevante destacar que tanto KMeans como KModes son algoritmos de agrupamiento que pueden ser comparados utilizando índices de evaluación, como el índice de Calinski-Harabasz y el coeficiente de silueta. El índice de Calinski-Harabasz es una medida utilizada para evaluar la calidad de los resultados de agrupamiento y se basa en la relación entre la dispersión interna de los datos dentro de cada clúster y la dispersión entre los clústeres (Calinski & Harabasz, 1974). Por otro lado, el coeficiente de silueta es una medida empleada para evaluar la calidad de un agrupamiento y asigna un valor entre -1 y 1 a cada muestra, en función de su similitud con su propio grupo en comparación con otros grupos cercanos (Rousseeuw, 1987).

Estos índices permiten evaluar la calidad de los resultados de agrupamiento generados por ambos algoritmos, teniendo en cuenta la separación y cohesión de los clústeres. Sin embargo, es importante tener en cuenta que tanto KMeans como KModes son métodos heurísticos, lo que significa que su convergencia a los resultados óptimos no está garantizada.

5. Metodología

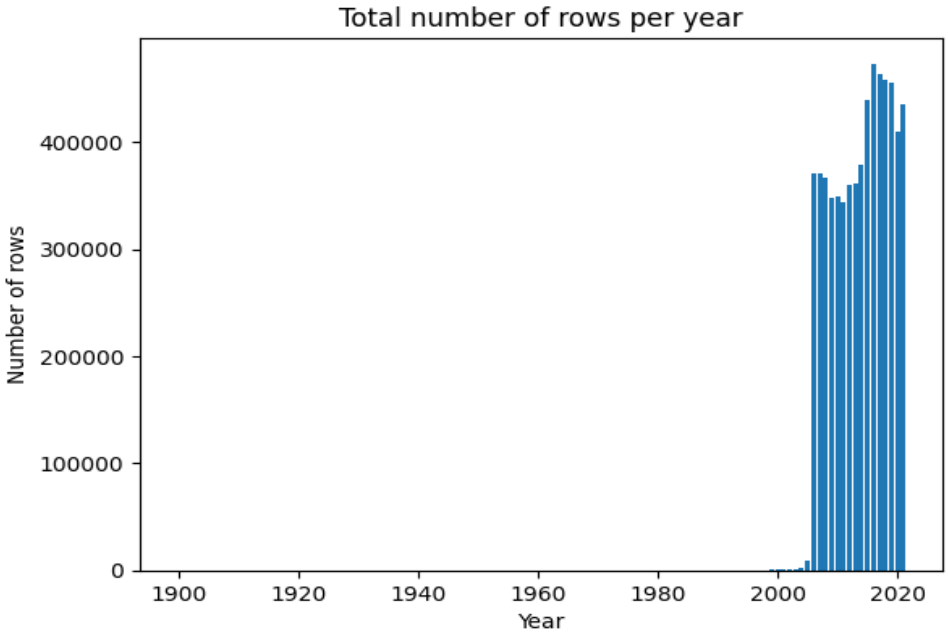
En este apartado, se presentará el diseño de investigación, los métodos empleados y los pasos seguidos para lograr los objetivos planteados. Esta sección proporciona una guía detallada de cómo se recopiló y analizó la información, permitiendo a los lectores comprender la rigurosidad y validez del estudio

5.1 Origen de los datos

La metodología que emplea esta monografía es cuantitativa con un enfoque de carácter experimental, se trabaja con una base de datos compuesta por denuncias hechas por los ciudadanos antes el departamento de Policía de la ciudad de Nueva York. Se destaca, entre las variables, atributos de tipo categórico que son aquellas que representan características o cualidades no numéricas, sino categorías o grupos discretos. Estas variables son utilizadas para clasificar elementos en diferentes categorías, en este caso en particular, como género del perpetrador y la víctima, el rango de edad, las zonas y barrios, tipo de crimen entre otras variables que serán descritas más adelante. Estas variables capturan información cualitativa y permiten analizar la relación o diferencia entre las categorías (Ghasemi & Zahediasl, 2012). Los datos se obtuvieron del conjunto de datos históricos de denuncias del Departamento de Policía de la ciudad de Nueva York (NYPD), están abiertos y están disponibles en el portal NYC Open Data (https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i). Si bien existe una API para la descarga de los datos ésta no se incluye en el proceso de este trabajo ya que, al tener una extensión de más de 7 millones de registros, el peso de la base es de 2.4G lo que hace realmente lento el incorporar los datos por esta vía.

La base de datos contiene variables como la fecha del hecho; la fecha de la denuncia, la hora del hecho, las coordenadas del lugar, el nombre del distrito, qué tipo de crimen se efectuó, si el crimen fue completado, interrumpido o fallido, entre otros. Los datos cubren el período desde 2006 hasta el 2019 sin embargo como se observa en la Figura 1 existe ruido que debe ser tratado para efectos de calidad en la modelación.

Figura 1. Distribución de las denuncias a lo largo del tiempo



Recientemente el departamento de policía de la ciudad ha creado una estrategia por medio de la cual la ciudadanía puede registrar denuncias menores relacionadas a la pérdida de billeteras o celulares a través del portal web: nyc.gov.

Este conjunto de datos incluye todos los delitos graves, delitos menores y violaciones denunciados al Departamento de Policía de la Ciudad de Nueva York (NYPD) desde 2006 hasta finales del 2019.

La volumetría del conjunto de datos es de un total de 7’825.499 de registros en el que cada uno es una denuncia interpuesta ante el Departamento de Policía. Adicional, presenta un total de 35 columnas o atributos dentro de las cuales se encuentran:

1. CMPLNT_NUM: Identificación persistente generada aleatoriamente para cada queja (int64). Renombrada por: "num_denuncia".

2. CMPLNT_FR_DT: Fecha exacta de ocurrencia del evento informado (o fecha de inicio de ocurrencia, si existe CMPLNT_TO_DT) (datetime). Renombrada por: "fecha_inicio_suceso".

3. CMPLNT_FR_TM: Hora exacta de ocurrencia del evento informado (o hora de inicio de ocurrencia, si existe CMPLNT_TO_TM) (object). Renombrada por: "hora_inicio_suceso".

4. CMPLNT_TO_DT: Fecha final de ocurrencia del evento informado, si se desconoce la hora exacta de ocurrencia (datetime). Renombrada por: "fecha_final_suceso".

5.CMPLNT_TO_TM: Hora final de ocurrencia del evento informado, si se desconoce la hora exacta de ocurrencia (object). Renombrada por: "hora_final_suceso".

6. ADDR_PCT_CD: El recinto en el que ocurrió el hecho (int64). Renombrada por: "direccion_suceso".

7. RPT_DT: Fecha en que se informó el evento a la policía (datetime). Renombrada por: "fecha_denuncia".

8. KY_CD: Código de clasificación de delitos de tres dígitos (int64). Renombrada por: "codigo_clasificacion".

9. OFNS_DESC: Descripción del delito correspondiente al código clave (object). Renombrada por: "descripcion_suceso".

10. PD_CD: Código de clasificación interna de tres dígitos (más granular que el código clave) (int64). Renombrada por: "codigo_clasificacion_granular".

11. PD_DESC: Descripción de la clasificación interna correspondiente al código PD (más granular que la Descripción del delito) (object). Renombrada por: "descripcion_codigo_clasificacion_granular".

12. CRM_ATPT_CPTD_CD: Indicador de si el delito se completó o intentó con éxito, pero fracasó o se interrumpió prematuramente (object). Renombrada por: "delito_completado_interrumpido".

13. LAW_CAT_CD: Nivel de ofensa: delito grave, delito menor, violación (objetc). Renombrado por: "nivel_ofensa".

14. BORO_NM: El nombre del distrito en el que ocurrió el incidente (object). Renombrada por: "distrito".

15. LOC_OF_OCCUR_DESC: Ubicación específica de la ocurrencia en o alrededor de las instalaciones; dentro, enfrente de, delante de, detrás de (object). Renombrada por: "ubicacion_especifica".

16. PREM_TYP_DESC: Descripción específica de las instalaciones; tienda de abarrotes, residencia, calle, etc. (object). Renombrada por: "descripcion_ubicacion".

17. JURIS_DESC: Descripción del código de jurisdicción (object). Renombrada por: "descripcion_codigo_jurisdiccion".

18. JURISDICTION_CODE: Jurisdicción responsable del incidente. Ya sea interno, como Policía (0), Tránsito (1) y Vivienda (2); o externas (3), como Corrección, Autoridad Portuaria, etc. (int64). Renombrada por: "jurisdiccion_encargada".

19. PARKS_NM: Nombre del parque, área de juegos o espacio verde de la ciudad de Nueva York en el que ocurrió, si corresponde (los parques estatales no están incluidos) (object). Renombrada por: "nombre_area_suceso".

20. HADEVELOPT: Nombre de la urbanización de viviendas de la Autoridad de la Vivienda de la Ciudad de Nueva York en la que ocurrió, si corresponde (object). Renombrada por: "nombre_urbanizacion".

21. HOUSING_PSA: Código de nivel de desarrollo (int64). Renombrada por: "codigo_nivel_desarrollo".

22. X_COORD_CD: Coordenada X para el sistema de coordenadas planas del estado de Nueva York, zona de Long Island, NAD 83, unidades pies (FIPS 3104) (float64). Renombrada por: "coordenada_x".

23. Y_COORD_CD: Coordenada Y para el sistema de coordenadas planas del estado de Nueva York, zona de Long Island, NAD 83, unidades pies (FIPS 3104) (float64). Renombrada por: "coordenada_y".

24. SUSP_AGE_GROUP: Grupo de edad del sospechoso (object). Renombrada por: "grupo_edad_sospechoso".

25. SUSP_RACE: Descripción de la raza (etnia) del sospechoso (object). Renombrada por: "raza_sospechoso".

26. SUSP_SEX: Descripción del sexo del sospechoso (object). Renombrado por: "sexo_sospechoso".

27. TRANSIT_DISTRICT: Distrito de tránsito en el que ocurrió la infracción (int64. Renombrado por: "distrito_transito".

28. Latitude: Coordenada de latitud de bloque medio para el sistema de coordenadas global, WGS 1984, grados decimales (EPSG 4326) (float64). Renombrado por: "latitud".

29. Longitude: Coordenada de longitud de bloque medio para el sistema de coordenadas global, WGS 1984, grados decimales (EPSG 4326) (float64). Renombrado por: "longitud".

30. Lat_Lon: Punto de ubicación geoespacial (latitud y longitud combinadas) (object). Renombrado por: "geoespacial".

31. PATROL_BORO: El nombre del distrito de patrulla en el que ocurrió el incidente (object). Renombrado por: "nombre_patrulla_suceso".

32. STATION_NAME: Nombre de la estación de tránsito (object). Renombrado por: "nombre_estacion_transito".

33. VIC_AGE_GROUP: Grupo de edad de la víctima (object). Renombrado por: "grupo_edad_victima".

34. VIC_RACE: Raza (etnia) de la víctima (object). Renombrado por: "raza_victima".

35. VIC_SEX: Sexo de la víctima (object). Renombrado por: "sexo_victima".

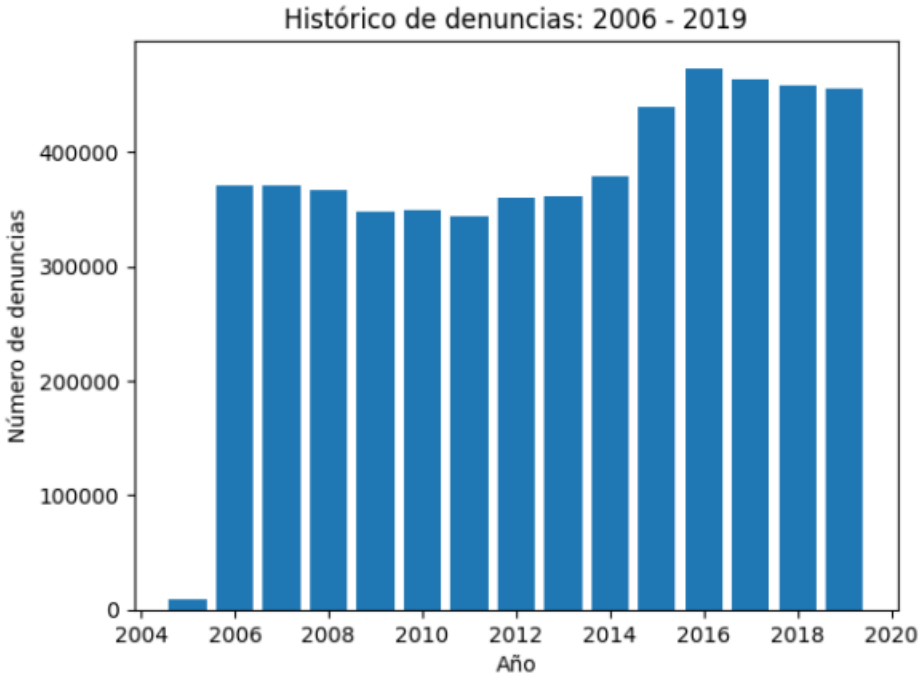
5.2 Limpieza de datos

Se realizó una limpieza exhaustiva de los datos tratando principalmente datos faltantes y atípicos. Algo que se encontró fue que no sólo hay registros mal datados en cuanto a fechas (ver figura 4), sino que también por la naturaleza del problema debería haber mayor precisión al momento de registrar los sucesos ya que esto permitirá también una mejor aproximación con los resultados. Adicional al tratamiento de datos, se llevó a cabo la transformación de variables categóricas a variables numéricas utilizando técnicas de codificación adecuadas para poder suministrar estos datos a los modelos KMeans y KModes.

Para el conjunto de datos utilizado y teniendo en cuenta la cantidad de categorías en cada una de las características, se procedió a realizar un agrupamiento en las variables tipo de crimen y lugar del suceso. Según la literatura (Kuhn & Johnson, 2013) no hay un número óptimo de etiquetas para todas las situaciones, pero se recomienda no tener más de 5-10 etiquetas por variable categórica para mantener la capacidad explicativa y la facilidad de interpretación de los resultados. Se sugiere, además, evitar las variables categóricas con un gran número de etiquetas debido a que pueden causar problemas de desbordamiento y reducir la capacidad de interpretación de los modelos.

Por la extensión de los registros en el conjunto de datos y, teniendo en cuenta el ruido que se encontró en los mismos, se realiza una primera iteración utilizando le modelo KMeans con los datos de 2006 a 2019 y, luego, el trabajó se concentra en el rango de 2016 a 2019 haciendo una separación por los distritos de Manhattan, Brooklyn y Bronx. La idea es poder validar los resultados entre modelos para saber cuál logra una mejor agrupación de datos y centroides diferentes.

Figura 2. Distribución de los datos con tratamiento



La Figura 2 muestra como es la distribución de las denuncias en la ciudad para el rango 2006 a 2019 una vez se ha eliminado el ruido en las fechas.

5.3 Métodos de análisis

Para el detalle práctico de la investigación llevada a cabo, se invita al lector a visitar el repositorio de GitHub: https://github.com/JuanDa-Machine/nypd-complaint-clustering-crime, donde se pueden encontrar los diferentes cuadernos con los códigos utilizados.

Para iniciar el análisis aplicando los modelos de aprendizaje automático, en primer lugar, se utilizó un modelo base de agrupamiento KMeans con los datos completos de 2006 a 2019 con la idea de tener una primera noción o el punto de partida a partir del cual se pueden medir las mejoras en los siguientes modelos utilizados (Géron, 2019). Esta iteración se llevó a cabo con un 15% de los datos ya que, si se consideran todos los registros, la demanda de recursos computacionales incrementa y es más complejo obtener un resultado. Se utilizó el coeficiente de la Silueta para validar la calidad de los resultados.

Una vez esto, se procede a realizar un tratamiento extra de los datos a través de la generación de nueva información con base en las características ya existentes, es decir, a partir de la variable fecha se extrae información como el nombre del día de la semana, la época del año en términos de las estaciones, el nombre del mes del año, entre otras. Posterior a esto se realiza un análisis exploratorio de los datos con un enfoque estadístico en variables categóricas en donde prevalece el uso de tablas pivotes que resumen la información por distrito y, además, el uso de tablas de contingencia que son utilizadas cuando se quiere comparar variables categóricas como el tipo de crimen y el lugar del suceso.

En las siguientes iteraciones se aplicó el método del codo y la técnica de KMeans de KMeansDask (Team, Dask: Library for dynamic task scheduling, 2016). El método del codo es una técnica ampliamente utilizada en análisis de agrupamiento y, como se verá más adelante, se aplica tanto para KMeans como para KModes. Su análisis se basa en la varianza explicada por diferentes números de grupos (K). Mediante la identificación de una inflexión en un gráfico, se determina el número de clústeres que maximiza la varianza en el conjunto de datos (Humaira & Rasyidah, 2018).

Continuando con las iteraciones, se aplica reducción de dimensionalidad antes de volver a utilizar la técnica de KMeans extrayendo las características más relevantes dentro del conjunto de datos. Esto debido a que, aplicando PCA, las agrupaciones generadas podrían ser mejores en términos de distinción de grupos. Finalmente, aunque el problema de la cantidad de datos fue resuelto abordando el tema no a nivel ciudad sino a nivel distrito, se podría ganar en términos de eficiencia computacional. Un estudio que ha trabajado en lo anterior (Ding & He, 2004), mostró que a medida que la reducción es aplicada, los resultados para la técnica KMeans mejoran de manera sistemática y significativa.

La última técnica utilizada fue KModes y con su uso, como menciona (López, 2007), en su estudio, se podría esperar mejores resultados por el contenido de variables categóricas. No se aplica reducción de dimensionalidad dado que acá las variables son categóricas y lo que se obtiene con PCA es una variable continua. Como se menciona en la sección 4 del marco teórico, se recurre al índice de Calinski-Harabasz y al coeficiente de la Silueta para comparar los resultados de los modelos utilizados.

5.4 Entrenamiento del modelo

Una vez la data fue tratada y procesada y se logró la obtención del conjunto de datos de entre 2016 y 2019 que contenía información para los distritos de Manhattan, Brooklyn y Bronx, se procede a utilizar los métodos ya mencionados KMeans, PCA – Kmeans y KModes.

5.4.1 KMeans

Como se menciona en la sección 5.1, se aplicó el método del codo para conocer un posible número K óptimo de clústeres y, con la implementación de KMeansDask, se validan los resultados de dicho método. Para esto, se probaron diferentes valores de número de clústeres (K), específicamente [2, 3, 5, 7, 9] a modo de utilizar el K como parámetro ya que en estos marcos de trabajo (framework) no es posible cambiar las distancias para hacer búsqueda de los mejores hiper parámetros. Los datos se transformaron utilizando la función get_dummies para convertir variables categóricas en variables dummy. Una vez se obtuvieron las etiquetas asignadas a cada muestra con el uso del modelo KMeans, se calcula el índice de Calinski-Harabasz y el coeficiente de silueta.

5.4.2 PCA – Kmeans

El proceso consistió en la transformación de los datos originales en variables categóricas codificadas y la creación de una matriz Dask para un procesamiento eficiente. A continuación, se aplicó el algoritmo PCA para reducir la dimensionalidad de los datos y se ajustó el modelo utilizando la función fit(). Posteriormente, se entrenó el algoritmo KMeans con diferentes números de clústeres utilizando los datos transformados por PCA. Se calcularon las métricas de evaluación,

como el índice de Calinski-Harabasz y el coeficiente de silueta, para analizar el rendimiento del modelo en cada número de clústeres.

Este proceso de entrenamiento del modelo, basado en PCA y KMeans, permitirá realizar un análisis de agrupamiento en el conjunto de datos, proporcionando información relevante para el desarrollo de la investigación.

5.4.3 KModes

Se aplicó el codificador ordinal a los datos de variables de tipo categórico. A continuación, se realizó una búsqueda del número óptimo de clústeres utilizando el método del codo. Se iteró sobre diferentes valores de n_clusters (2, 3, 4, 5) y se ajustó el modelo KModes con el método de inicialización 'Cao'. Se registró el costo obtenido en cada iteración y se generó un gráfico para visualizar la relación entre el número de clústeres y el costo. Después de determinar el número óptimo de clústeres, se procedió al entrenamiento del modelo KModes con el número de clústeres seleccionado. Se utilizó una muestra del 50% de los datos codificados para entrenar el modelo. Se calcularon las métricas de evaluación, como el índice de Calinski-Harabasz y el coeficiente de silueta, para analizar el rendimiento del modelo en cada número de clústeres.

6. Resultados

Con la investigación realizada y a la luz de los resultados que podrían generarse entre uno y otro modelo, este documento plasma el camino recorrido en un intento por solucionar el problema a través de agrupamientos utilizando KMeans inicialmente, sin embargo, a pesar de que es el modelo más estándar y tradicional podría no ser el mejor en cuanto al tratamiento de variables categóricas. Esto se explica principalmente por el hecho de que KMeans calcula distancias entre puntos de datos basadas en valores numéricos, mientras que KModes aborda esta limitación utilizando una medida de disimilitud o similitud entre variables categóricas utilizando un criterio de coincidencia enfocándose en encontrar modas y frecuencias dentro de cada grupo, en lugar de calcular distancias o promedios.

Lo anterior es apoyado en (San, Huynh, & Nakamori, 2004) quienes en su estudio explican cómo KModes supera a KMeans en cuanto al agrupamiento de variables categóricas se trata: "(…) como hemos visto, al aplicar el método KMeans a objetos categóricos, nos enfrentamos a dos problemas principales: la formación de centroides de clúster y el cálculo de la disimilitud entre objetos y centroides de clúster. Estos problemas se han resuelto por completo en el algoritmo KModes mediante el uso de la medida de disimilitud de coincidencia simple para datos categóricos en lugar de la medida de distancia euclidiana y reemplazando las medias de los clústeres por las modas. Estas modificaciones también cumplen con la condición de minimización, como se mostró en (Huang, 1998)". Así pues, considerando los resultados obtenidos a través del proceso de iteraciones con diferentes modelos, y, como se verá más adelante, a pesar de que se obtiene un patrón en cada modelo y los resultados difieren entre cada algoritmo, se sugiere, para investigaciones futuras, el uso de modelos de agrupamiento enfocados en el tratamiento de variables categóricas como: KModes-TOPSIS, Fuzzy C-Modes, ROCK y clúster jerárquico.

Esta sección presenta los resultados obtenidos para cada iteración y modelo utilizado a través de figuras junto con una breve descripción de estos. Los resultados serán analizados y discutidos en la sección siguiente para determinar el mejor modelo.

6.1 KMeans

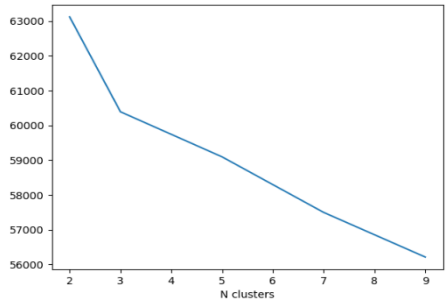
Como se observa en la Figura 3, el método del codo indica que el número óptimo de clústeres es 3, donde se produce la primera inflexión en el gráfico. Este resultado se valida mediante el coeficiente de la silueta, el cual, aunque cercano a 0, indica que las muestras están en o cerca del límite de decisión entre dos clústeres vecinos, lo que implica cierta superposición entre los valores. Además, se utiliza el índice de Calinski-Harabasz en estas comparaciones, y aunque es mayor para K = 2, es con K = 3 donde tanto el método del codo como el coeficiente de la silueta convergen. Se ve que hay cercanías en los resultados entre distritos puesto que en todos ellos para K=3 el coeficiente de la silueta es 0.10.

Figura 3. Resultados obtenidos por Distrito utilizando la técnica de KMeans

KMEANS

MANHATTAN

Método del codo

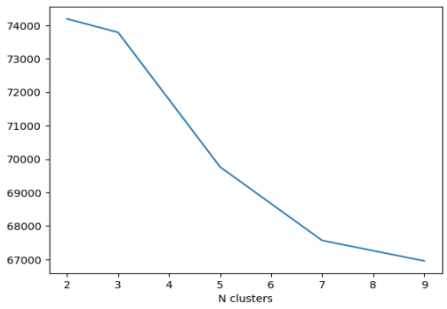


Iteración utilizando Kmeans:

k	Calinski-Harabasz Index	Silhouette Coefficient
2	2.746.653.442	0.103252
3	2.418.787.910	0.106804
5	1.716.989.560	0.085804
7	1.316.076.073	0.075724
9	1.007.873.576	0.084520

BROOKLYN

Método del codo

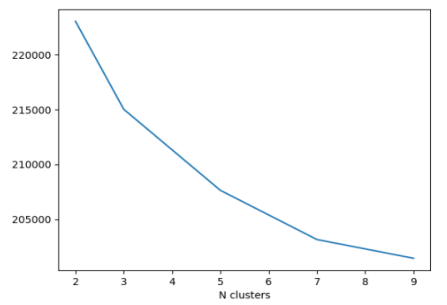


Iteración utilizando Kmeans:

k	Calinski-Harabasz Index	Silhouette Coefficient
2	3.214.809.666	0.101131
3	2.627.229.091	0.103900
5	1.961.062.258	0.087516
7	1.368.699.842	0.075553
9	1.249.444.090	0.064177

BRONX

Método del codo



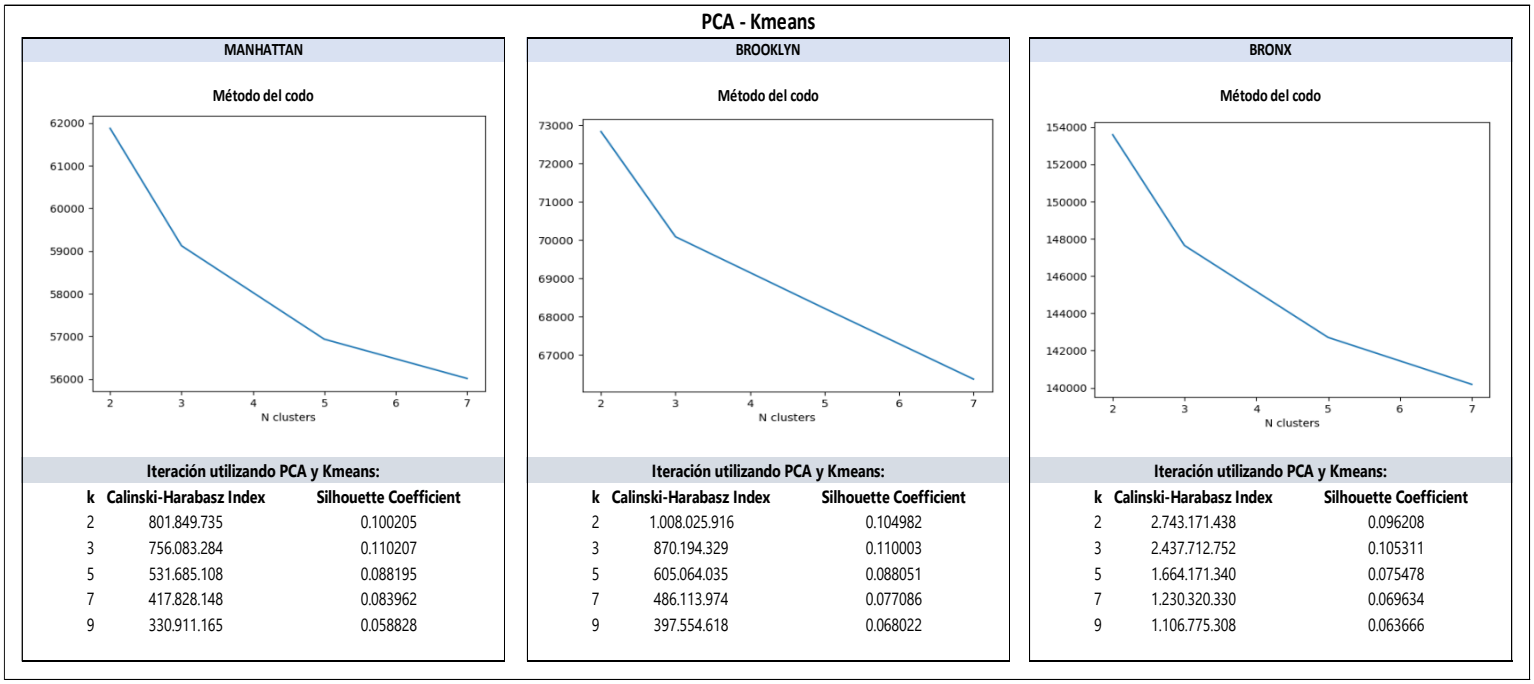
Iteración utilizando Kmeans:

k	Calinski-Harabasz Index	Silhouette Coefficient
2	8.880.606.503	0.093576
3	7.835.419.454	0.102561
5	5.382.366.464	0.074459
7	4.221.401.989	0.064865
9	3.526.144.750	0.068789

6.2 PCA – KMeans

En la Figura 4, se muestran los resultados obtenidos aplicando reducción de dimensionalidad y luego la técnica de KMeans. El método del codo indica que el número óptimo de clústeres es 3. El resultado se valida mediante el coeficiente de la silueta, el cual, pasa de 0.10 con K igual a 2 y a 3, a 0.11 en Manhattan y en Brooklyn con K = 3. Para Bronx, a diferencia de los anteriores distritos, al aplicar PCA antes de la técnica KMeans, los resultados con K igual a 2 y a 3 permanecen en 0.09 y 0.10 respectivamente.

Figura 4. Resultados obtenidos por Distrito utilizando reducción de dimensionalidad PCA y KMeans



Iteración utilizando PCA y Kmeans:

k	Calinski-Harabasz Index	Silhouette Coefficient
2	801.849.735	0.100205
3	756.083.284	0.110207
5	531.685.108	0.088195
7	417.828.148	0.083962
9	330.911.165	0.058828

Iteración utilizando PCA y Kmeans:

k	Calinski-Harabasz Index	Silhouette Coefficient
2	1.008.025.916	0.104982
3	870.194.329	0.110003
5	605.064.035	0.088051
7	486.113.974	0.077086
9	397.554.618	0.068022

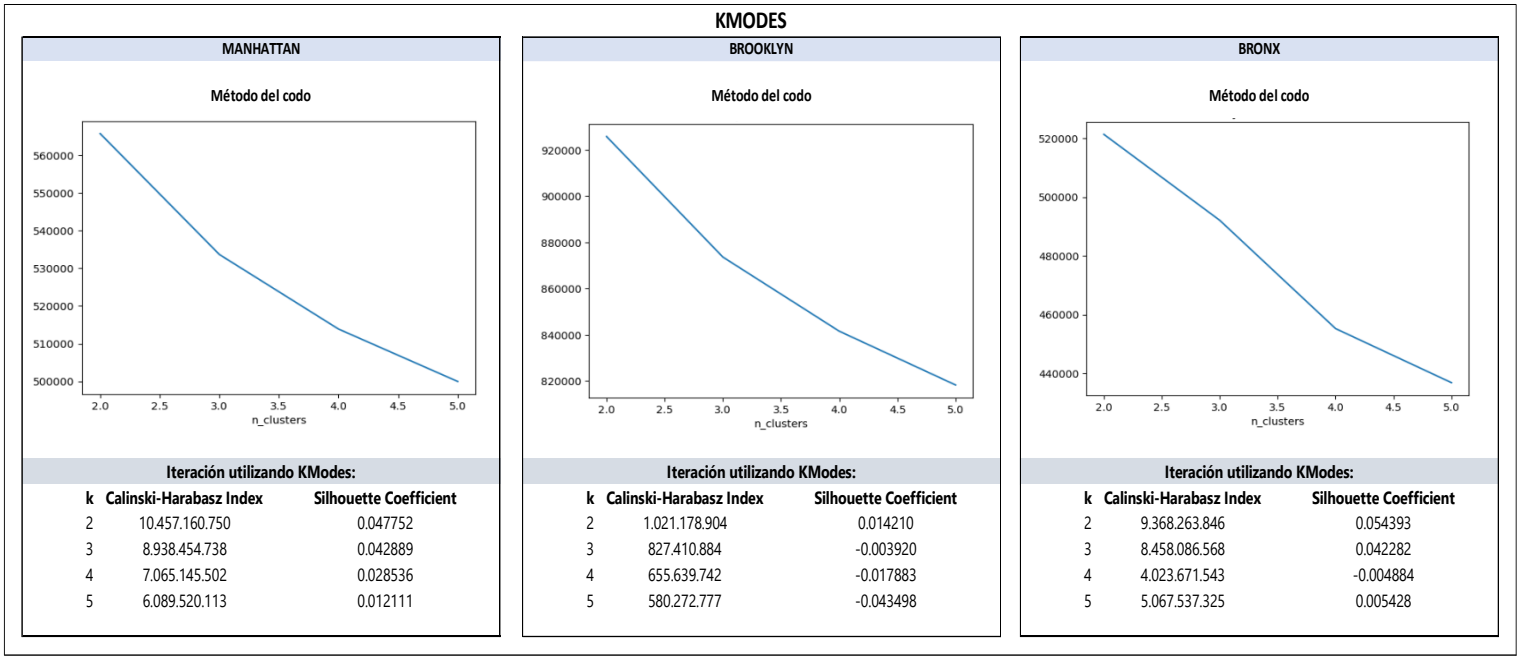
Iteración utilizando PCA y Kmeans:

k	Calinski-Harabasz Index	Silhouette Coefficient
2	2.743.171.438	0.096208
3	2.437.712.752	0.105311
5	1.664.171.340	0.075478
7	1.230.320.330	0.069634
9	1.106.775.308	0.063666

6.3 KModes

Los resultados para esta técnica mostraron que, en los tres conjuntos de datos, el índice de Calinski-Harabasz tendió a disminuir a medida que aumentaba el número de clústeres. El coeficiente de la silueta también presentó valores bajos o cercanos a cero, indicando superposición entre los clústeres en los datos. Sin embargo, se observó una mejora marginal en el índice de Calinski-Harabasz y el coeficiente de la silueta cuando se aumentó de K igual a 2 a K igual a 3. Lo anterior se detalla en la Figura 5 a continuación.

Figura 5. Resultados obtenidos por Distrito utilizando la técnica de KModes



6.4 Análisis de resultados

Dado que las métricas utilizadas muestran mejor aproximación con K = 3, se utiliza este número de clústeres para proceder a hacer el análisis de cada uno de los centroides generados en cada técnica. Posterior a esto, y con el ánimo ver la veracidad de los resultados en términos del problema inicial, se hace la comparativa entre ellos para determinar cuál sería el mejor modelo y

con él poder generar estrategias en torno a la seguridad y los hechos ocurridos entre 2016 y 2019 para cada uno de los distritos analizados.

6.4.1 Centroides en KMeans

Al analizar los resultados obtenidos con la técnica KMeans en los tres distritos y considerando la información proporcionada, se pueden identificar algunas similitudes comunes entre ellos:

- Crimen: en los tres distritos, el crimen más común es la extorsión, que ocupa una proporción significativa en los tres centroides. Esto sugiere que la extorsión es un problema común en estos distritos.
- Lugar: los tres distritos presentan una alta incidencia de crímenes en espacios públicos, lo que indica que estos lugares pueden ser considerados como puntos críticos en términos de seguridad.
- Temporalidad: aunque varía ligeramente entre los centroides, en general, los crímenes tienden a ocurrir con mayor frecuencia los viernes.
- Características demográficas: si bien hay diferencias entre los centroides, hay una presencia significativa de sospechosos de raza desconocida en todos los distritos. Además, en algunos centroides, se observa una alta proporción de sospechosos de raza negra. Estos datos demográficos resaltan la importancia de considerar la diversidad racial en el análisis de la criminalidad en estos distritos.

Lo anterior se detalla en la siguiente figura 6.

Figura 6. Centroides obtenidos con la técnica KMeans

DISTRITO		MANHATTAN			BROOKLYN			BRONX		
CENTROIDE		0	1	2	0	1	2	0	1	2
CRIMEN	TIPO	Extorsión (48.5%)	Extorsión (48.6%)	Extorsión (47.7%)	Extorsión (55.9%)	Sexual (25.1%)	Extorsión (25%)	Delitos graves (35.4%)	Extorsión (56.6%)	Extorsión (36%)
	LUGAR	Espacio público (55.9%)	Espacio público (55.8%)	Espacio público (56.5%)	Espacio público (75%)	Espacio público (74.9%)	Espacio público (74.1%)	Espacio público (85.3%)	Espacio público (82.8%)	Espacio público (58.1%)
TEMPORALIDAD	DÍA	Viernes (15.6%)	Viernes (15.8%)	Viernes (15.4%)	Viernes (15.8%)	Sábado (15%)	Viernes (14.9%)	Domingo (14.7%)	Viernes (15%)	Viernes (16.1%)
	MES	Diciembre (16.4%)	Diciembre (16.2%)	Diciembre (15.6%)	Diciembre (16.1%)	Diciembre (15.4%)	Diciembre (12.5%)	Diciembre (15.7%)	Diciembre (16.6%)	Diciembre (14.9%)
	ESTACIÓN	Verano (26.3%)	Verano (26.7%)	Verano (27.2%)	Verano (26.9%)	Verano (26.1%)	Verano (26.9%)	Verano (26.9%)	Verano (26.6%)	Primavera (25.9%)
	HORA	Tarde (37.3%)	Tarde (37%)	Tarde (36.6%)	Tarde (32.3%)	Tarde (33%)	Tarde (33.7%)	Tarde (32.3%)	Noche (33.2%)	Tarde (36.3%)
SOSPECHOSO	RAZA	Desconocida (45%)	Desconocida (44.9%)	Desconocida (44.3%)	Desconocida (98.4%)	Blanca (85.6%)	Negra (97.5%)	Negra (59.4%)	Desconocida (97.6%)	Negra (41.7%)
	EDAD (años)	Desconocida (63.8%)	Desconocida (63.1%)	Desconocida (63.7%)	Desconocida (99.7%)	25-44 (46.2%)	25-44 (41.9%)	25-44 (43.6%)	Desconocida (99.5%)	Desconocida (52.3%)
	SEXO	Masculino (45.9%)	Masculino (45.7%)	Masculino (46%)	Desconocido (94%)	Masculino (77.8%)	Masculino (74.1%)	Masculino (71.9%)	Desconocido (91%)	Masculino (60%)
VÍCTIMA	RAZA	Desconocida (40.2%)	Desconocida (40.4%)	Desconocida (40.2%)	Blanca (32.9%)	Blanca (51%)	Negra (54.9%)	Negra (49.4%)	Negra (45.1%)	Desconocida (99.6%)
	EDAD (años)	Desconocida (51.1%)	Desconocida (51.5%)	Desconocida (51.8%)	Desconocida (39.2%)	Desconocida (42.8%)	Desconocida (41.6%)	25-44 (48.5%)	25-44 (46.5%)	Desconocida (99.9%)
	SEXO	Transgénero (35.3%)	Transgénero (36%)	Transgénero (35.2%)	Masculino (42.4%)	Femenino (42%)	Femenino (46.6%)	Femenino (64.2%)	Masculino (55.1%)	Transgénero (95.4%)

6.4.2 Centroides en PCA – KMeans

- Crimen: en los tres distritos, el tipo de crimen más común en todos los centroides sigue siendo la extorsión. Sin embargo, las proporciones específicas han cambiado ligeramente en comparación con los resultados anteriores.
- Lugar: la mayoría de los delitos en los tres distritos continúan teniendo lugar en espacios públicos. En general, los centroides muestran una alta proporción de crímenes en lugares de acceso público, aunque las proporciones específicas pueden variar entre los distritos.
- Temporalidad: los viernes siguen siendo el día con mayor frecuencia de crímenes en los tres distritos, como se observó en los resultados anteriores. Esto implica que los viernes pueden ser considerados días críticos en términos de seguridad en estos distritos.
- Características demográficas: en los centroides generados después de aplicar PCA y KMeans, la proporción de sospechosos y víctimas de raza desconocida sigue siendo alta en todos los distritos. Sin embargo, también se observa una alta proporción de sospechosos masculinos y víctimas mujeres de raza negra y transgéneros en algunos centroides.

La figura 7 muestra en detalle lo anterior.

Figura 7. Centroides obtenidos con la técnica PCA - KMeans

DISTRITO		MANHATTAN			BROOKLYN			BRONX		
CENTROIDE		0	1	2	0	1	2	0	1	2
CRIMEN	TIPO	Extorsión (48.6%)	Extorsión (47.7%)	Extorsión (48.5%)	Extorsión (37.1%)	Extorsión (38%)	Extorsión (37.7%)	Extorsión (31.4%)	Extorsión (32%)	Extorsión (31.5%)
	LUGAR	Espacio público (56%)	Espacio público (56.3%)	Espacio público (56%)	Espacio público (75%)	Espacio público (75%)	Espacio público (74.2%)	Espacio público (77.6%)	Espacio público (77.7%)	Espacio público (78.1%)
TEMPORALIDAD	DÍA	Viernes (15.7%)	Viernes (15.5%)	Viernes (15.6%)	Viernes (15.3%)	Viernes (15.7%)	Martes (14.9%)	Viernes (15%)	Viernes (15.5%)	Viernes (15.4%)
	MES	Diciembre (16.1%)	Diciembre (15.7%)	Diciembre (16.4%)	Diciembre (16.2%)	Diciembre (16%)	Diciembre (16%)	Diciembre (15.8%)	Diciembre (15.9%)	Diciembre (15.6%)
	ESTACIÓN	Verano (26.4%)	Verano (27.4%)	Verano (26.4%)	Verano (26.9%)	Verano (26.9%)	Verano (27%)	Verano (26.2%)	Verano (26.6%)	Verano (26.9%)
	HORA	Tarde (37%)	Tarde (36.6%)	Tarde (37.2%)	Tarde (32.9%)	Tarde (32.8%)	Tarde (33%)	Tarde (33.7%)	Tarde (33.3%)	Tarde (32.2%)
SOSPECHOSO	RAZA	Desconocida (45%)	Desconocida (44.2%)	Desconocida (44.9%)	Desconocida (44.6%)	Desconocida (45%)	Desconocida (44.3%)	Desconocida (38.5%)	Desconocida (38.9%)	Desconocida (38.8%)
	EDAD (años)	Desconocida (63.1%)	Desconocida (63.7%)	Desconocida (63.8%)	Desconocida (63.7%)	Desconocida (63.9%)	Desconocida (63.6%)	Desconocida (59%)	Desconocida (59.1%)	Desconocida (59.6%)
	SEXO	Masculino (45.7%)	Masculino (46%)	Masculino (45.9%)	Masculino (42.1%)	Masculino (44.1%)	Masculino (44.7%)	Masculino (49.3%)	Masculino (49.2%)	Masculino (49.1%)
VÍCTIMA	RAZA	Desconocida (40.2%)	Desconocida (40.4%)	Desconocida (40.2%)	Negra (35.5%)	Negra (35.9%)	Negra (35.9%)	Negra (35.2%)	Negra (35.8%)	Negra (35.8%)
	EDAD (años)	Desconocida (51.3%)	Desconocida (51.9%)	Desconocida (51.2%)	Desconocida (41.4%)	Desconocida (39.9%)	Desconocida (41%)	Desconocida (41%)	Desconocida (40.2%)	Desconocida (40.1%)
	SEXO	Transgénero (35.9%)	Transgénero (35.3%)	Transgénero (35.3%)	Femenino (40.4%)	Femenino (41.3%)	Femenino (40.9%)	Femenino (42.7%)	Femenino (42.7%)	Femenino (42.7%)

6.4.3 Centroides en KModes

En los centroides generados por KModes se muestra en detalle en la Figura 8 y a modo general se identifica lo siguiente:

- El crimen más común en todos los distritos sigue siendo la extorsión. Sin embargo, las proporciones específicas de cada tipo de crimen varían en los distintos centroides.
- Lugar: ocurren principalmente en espacios públicos, variando la proporción de crímenes entre cada centroide.
- Temporalidad: los días de la semana con mayor frecuencia de crímenes difieren en cada distrito, sin embargo, en al menos un centroide de cada distrito, los viernes ocurren hechos relacionados a la extorsión en lugares públicos.
- Características demográficas: se observa una alta proporción de sospechosos y víctimas con datos desconocidos en términos de raza y edad. Además, se destacan diferencias en la proporción de sospechosos y víctimas según el sexo y la raza en algunos centroides.

Figura 8. Centroides obtenidos con la técnica KModes

DISTRITO		MANHATTAN			BROOKLYN			BRONX		
CENTROIDE		0	1	2	0	1	2	0	1	2
CRIMEN	TIPO	Extorsión (54.8%)	Extorsión (57.8%)	Sexual (42.8%)	Extorsión (30.8%)	Extorsión (53%)	Incidentes menores (41.9%)	Extorsión (50%)	Delitos graves (42.5%)	Incidentes menores (30.1%)
	LUGAR	Espacio público (43.7%)	Espacio público (67.8%)	Espacio público (69.3%)	Espacio público (74.7%)	Espacio público (81.1%)	Espacio público (63.1%)	Espacio público (80%)	Espacio público (85.3%)	Espacio público (64.5%)
TEMPORALIDAD	DÍA	Viernes (19.7%)	Miércoles (19.7%)	Lunes (21.7%)	Jueves (18.4%)	Viernes (20.9%)	Sábado (22.1%)	Viernes (19.2%)	Sábado (20%)	Martes (20.1%)
	MES	Diciembre (21.9%)	Julio (16.1%)	Mayo (18.3%)	Diciembre (22.2%)	Julio (15.9%)	Junio (20.3%)	Diciembre (23%)	Agosto (16.1%)	Abril (15.5%)
	ESTACIÓN	Otoño (33.1%)	Verano (42.1%)	Primavera (45.3%)	Otoño (34.4%)	Verano (41.4%)	Primavera (45.7%)	Otoño (34.8%)	Verano (40.7%)	Primavera (39.8%)
	HORA	Tarde (46.1%)	Noche (40%)	Mañana(32.7%)	Tarde (43.1%)	Noche (42.2%)	Mañana(29.1%)	Tarde (37.2%)	Noche (42.4%)	Tarde (41.8%)
SOSPECHOSO	RAZA	Desconocida (38.8%)	Desconocida (71.8%)	Blanca (47.7%)	Negra (63%)	Desconocida (74.4%)	Desconocida (76%)	Desconocida (72.7%)	Negra (64%)	Blanca (38.6%)
	EDAD (años)	Desconocida (59.9%)	Desconocida (81.8%)	Desconocida (42.5%)	Desconocida (45.4%)	Desconocida (83%)	Desconocida (75.5%)	Desconocida (87.1%)	25-44 (57.3%)	Desconocida (54%)
	SEXO	Masculino (53.2%)	Desconocido (70.1%)	Masculino (75.6%)	Masculino (68.5%)	Desconocido (71.4%)	Desconocido (63.5%)	Desconocido (67.2%)	Masculino (77%)	Masculino (66.9%)
VÍCTIMA	RAZA	Desconocida (74.5%)	Blanca (58.4%)	Blanca (56.5%)	Negra (51.7%)	Blanca (55.8%)	Desconocida (81.3%)	Negra (50.7%)	Blanca (48.7%)	Desconocida (81.1%)
	EDAD (años)	Desconocida (84.8%)	25-44 (55.5%)	25-44 (53.7%)	Desconocida (41.2%)	25-44 (58.3%)	Desconocida (86.3%)	25-44 (44.3%)	25-44 (50.2%)	Desconocida (88.8%)
	SEXO	Transgénero (68.6%)	Femenino (61.1%)	Masculino (61.5%)	Femenino (56.8%)	Masculino (65.5%)	Transgénero (75.2%)	Femenino (48.5%)	Femenino (61%)	Transgénero (75.1%)

7. Discusión

Este trabajo logra aplicar las técnicas propuestas, resaltando patrones en donde se identifica que la extorsión, el robo y el fraude en lugares públicos son los crímenes más frecuentes en los tres distritos de estudio. Al analizar los resultados, los hechos denunciados ocurren principalmente a mujeres y personas transgénero, llama la atención que, en la gran mayoría de los casos, no hay un reconocimiento de los sospechosos, lo que podría indicar las víctimas están distraídas u ocurren de raponazo.

Basados en el rendimiento de cada modelo, KMeans es quien muestra mejor coeficiente de silueta en todos los K, sin embargo, cuando se aplica reducción de dimensionalidad, específicamente en K = 3, se observa que los resultados tienen una leve mejora. Comparando estos dos enfoques con KModes, se evidencia que KMeans, con y sin reducción de dimensionalidad, es superior.

PCA – Kmeans con K=3 Muestra un desempeño ligeramente superior en términos de las métricas de evaluación, como el índice de Calinski-Harabasz y el coeficiente de la silueta, en comparación con KModes y KMeans. Esto indica que, el estudio referenciado (Ding & He, 2004) es un método preciso y puede aplicarse en ejercicios de agrupamiento puesto que logra una mejor cohesión intra-cluster y una mayor separación inter-cluster en los datos. Con esta técnica se debe tener particular cuidado ya que los resultados, al estar en otra dimensión, tienden a perder interpretabilidad.

Al considerar la interpretabilidad de los resultados, KModes muestra una ventaja significativa, pues, al utilizarlo, se logra una mejor comprensión y representación de los datos a través de la identificación de centroides, que corresponden a perfiles de características específicas. Estos centroides permiten una interpretación más clara y directa de los patrones de criminalidad presentes en cada distrito.

Considerando tanto la calidad de los resultados en términos de métricas como la interpretabilidad de los centroides, este estudio se decanta por la técnica de KModes, es la más adecuada para el análisis de los datos de crímenes. Esta técnica permite obtener una visión más clara y detallada de los diferentes perfiles de criminalidad presentes en cada distrito. Por ejemplo, en el tema de la hora en que ocurren los sucesos, fue esta técnica la que logró mejor entendimiento pues en los ejercicios anteriores a éste con KMeans, con y sin PCA, se ve que la gran mayoría de

los sucesos ocurren en horas de la tarde, lo que podría ser cierto, pero para el desarrollo de estrategias podría no ser la mejor, puesto que las horas más peligrosas del día son las comprendidas entre las 7 p.m y las 5 a.m.

Llama la atención el hecho de que los resultados de los modelos con KMeans muestren discrepancia en la estación y el mes del año, es decir, la primavera en la ciudad va de marzo a mayo, el verano es entre junio y agosto, otoño se encuentra comprendido en los meses de septiembre y noviembre y, finalmente, invierno está entre los meses de diciembre y febrero. En cuanto a los resultados de KModes, también presentan discrepancia en dos ocasiones, sin embargo, esta es menor puesto que la cercanía de la estación con el mes podría llegar a aceptarse. Los centroides 0 de Manhattan y Brooklyn: otoño-diciembre, y el centroide 2 de Brooklyn: junio-primavera. Para KMeans, en cambio, se muestra la relación verano y diciembre, lo que difiere y podría considerarse no preciso en términos de los resultados. Por otro lado, KModes al trabajar con modas y frecuencias, logra una diversidad mayor en sus centroides lo que a la larga podría llevar a interpretar mejor el problema y sus posibles resultados.

A continuación, basado en la anterior aproximación, se presentan estrategias de como la policía podría actuar para que los hechos de criminalidad sean controlados e incluso puedan disminuir. Estas recomendaciones son propuestas a partir de la información que arrojan los resultados de la técnica KModes en los tres distritos analizados.

7.1 Manhattan

- Centroide 0: los transgéneros son extorsionados los viernes en la tarde: estos hechos ocurren a manos de hombres con edad y raza, o desconocida o negra, pero se sabe que son hombres. Las personas afectadas no reportan la edad ni su raza, y son vulnerados en lugares públicos en otoño y diciembre.

  Estos resultados son similares a KMeans en cuanto a los días de la semana, el tipo de hecho, los lugares, los sospechosos y las víctimas.

Teniendo en cuenta lo anterior, una estrategia para reforzar la seguridad en los lugares públicos podría no sólo ser mayor presencial de la policía sino cámaras de monitoreo que puedan identificar perfiles de personas transgénero que son las más vulnerables.

- Centroide 1: durante las noches, los hombres más jóvenes de raza blanca en Manhattan son víctimas de delitos como fraude, robos o hurtos. Estos incidentes son cometidos por individuos cuyas características de edad, sexo y raza son desconocidas. Se observa que las víctimas sufren violencia principalmente los miércoles, y estos hechos ocurren en lugares públicos durante el verano, especialmente en el mes de julio.

  De acuerdo con esto, y como se ha mencionado en el anterior agrupamiento, una estrategia no sólo es mayor despliegue de fuerza pública en los lugares públicos sino la alerta de que los hechos le ocurren más a jóvenes de raza blanca para que estas personas tengan mayor cuidado cuando transiten o estén en dichos lugares.

- Centroide 2: los lugares públicos de Manhattan no son tan seguros durante la primavera, en mayo. Hombres blancos de edad de entre 25 a 44 años son acosados o violados en horas de la mañana por hombres de edad desconocida y raza negra.

  Dado que no sólo hechos relacionados a la extorsión ocurren en lugares públicos, sino que también hay acoso y violaciones principalmente a hombres, la instalación de zonas de monitoreos y mayor control de la policía realizando requisas puede ejercer presión para ahuyentar los perpetradores de estos lugares en el mes de mayo.

7.2 Brooklyn

- Centroide 0: mujeres en Brooklyn corren peligro de ser víctimas de fraude o robo en las horas de la tarde de los jueves, en otoño o diciembre. Estos hechos ocurren a manos de hombres de raza negra de edad desconocida.

Estos resultados se asemejan al grupo 3 de KMeans, donde se mostró el mismo patrón con la diferencia de que no era jueves sino viernes; incluso, se podría usar esta información para soportar y completar un perfil en KMeans con reducción de dimensionalidad donde los resultados muestran sexo desconocido.

La estrategia propuesta se centra en la protección de las mujeres a través de un despliegue estratégico de fuerzas policiales en áreas públicas, utilizando un enfoque psicológico que no implica necesariamente la requisa indiscriminada de hombres, pero sí garantizando la presencia de agentes uniformados en lugares donde se congreguen grupos de mujeres.

- Centroide 1: en las noches de los viernes en verano, en el mes de julio, es cuando los hombres de raza blanca, que tienen entre 25 y 44 años, corren peligro de ser víctimas de extorsión en lugares públicos a manos de sospechosos de los cuales no se tiene registro de edad, raza o sexo.

  Este clúster, en particular, es similar al clúster 1 de KMeans sin reducción de dimensionalidad. Cabe señalar que es verano la temporada del año donde más crímenes se cometen, teniendo en cuenta el lugar de los hechos, es importante que las personas allí anden acompañadas y con precaución, además, la policía, durante esta temporada en particular, debería contar con centros móviles de atención inmediata.

- Centroide 2: las riñas y problemas de orden público en Brooklyn ocurren los sábados en horas de la mañana y están implicadas personas transgénero quienes son victimizadas por personas de raza, sexo y edad desconocida. Son incidentes menores y ocurren en primavera (o junio) en lugares públicos.

  Este clúster, en particular, difiere de todos los demás mencionados hasta ahora, pues el crimen no está relacionado a extorsión o temas sexuales, por ende, a pesar de que ocurren también principalmente en lugares públicos las estrategias antes mencionadas

podrían tener cabida acá con la particularidad de que son delitos menores y que el llamado debe ser más desde la tolerancia y cultura ciudadana.

7.3 Bronx

- Centroide 0: mujeres blancas de entre 25 a 44 años son víctimas de extorsión en lugares públicos. Estos hechos ocurren los viernes en la tarde en otoño o diciembre en los lugares públicos. Los perpetradores no ha sido identificados puesto que su sexo, raza y edad es desconocida.

  En general, a medida que el ejercicio continuó y con el análisis de los resultados, es destacable el hecho de que si bien muchos de los centroides arrojan sexo del sospechoso desconocido, por la gravedad de los hechos y el tipo de crimen es posible afirmar que estos hechos vienen marcados por presencia de hombres, quienes son los que protagonizan los crímenes, es decir, para temas relacionados a extorsión y delitos mayores, por el tipo de víctima que generalmente es femenino en su mayoría, transgéneros y hombres de 25 a 44 años de edad, es poco probable que una mujer sea la que realice dichos enfrentamientos y no logre ser identificada por sus rasgos de pelo largo o estatura que es generalmente más baja a la de los hombres.

- Centroide 1: mujeres blancas de entre 25 a 44 años son víctimas de delitos mayores, secuestro y homicidio. Esto ocurren en la noche de los sábados en lugares públicos durante el verano o el mes de agosto. Los criminales son hombres de edad de entre 25 a 44 años y de raza negra.

  Es importante hacer presencia y fomentar la seguridad para casos tan severos durante los sábados en la noche que es cuando generalmente las personas comparten su tiempo con amigos o familiares en discotecas o restaurantes, pero en el camino son víctimas de hombres que han ido más allá en su delinquir. Con la idea de generar impacto en la mitigación de estos crímenes es importante que la ley sea severa con las penas y castigos

para que no haya un incentivo sobre los hombres en arremeter contra las mujeres de esta manera.

- Centroide 2: hombres negros de edad desconocida protagonizan riñas e infracciones o accidentes de tránsito donde también hay transgéneros involucrados como víctimas. Estos hechos curren en las tardes de martes de primavera o abril en los lugares públicos de Bronx.

  Si bien no han sido graves los hechos, el llamado, como se hizo en el clúster 3 de Brooklyn, es a la tolerancia y a hacer de los lugares públicos un espacio de tránsito tranquilo y propicio para vivir en comunidad. Los hombres, de nuevo, son protagonistas como perpetradores y cabe la pena hacerse la pregunta de si para los hombres las penas de cárcel, las fianzas y castigos de servicio social deberían ser mayores a las de los demás géneros.

8. Conclusiones

En el contexto del procesamiento de grandes volúmenes de datos (big data), el uso de algoritmos de agrupamiento es una estrategia recomendada para encontrar patrones y realizar estudios descriptivos. Además, se sugiere emplear técnicas de reducción de dimensionalidad para obtener beneficios en términos computacionales y para identificar las características más relevantes en la resolución del problema. No se recomienda aplicar reducción de dimensionalidad en combinación con KModes, ya que esta técnica no es adecuada para variables categóricas (discretas) puesto que pasarían a ser continuas y se perdería contexto del ejercicio.

En este estudio se compararon los algoritmos KMeans y KModes y cabe destacarse que no se encontraron diferencias significativas en cuanto a las métricas de los modelos para la solución del problema puesto que se involucran variables categóricas. Sin embargo, al evaluar los centroides de los clústeres, se observó que KModes proporciona una mejor separabilidad y facilita la interpretación de los resultados, lo que a su vez se traduce en planes de acción más efectivos para las autoridades policiales.

Con la idea de obtener mejores resultados en un trabajo a futuro, se recomienda que el número de etiquetas se adapte mejor al contexto y al propósito de análisis de cada variable, es decir, para este trabajo tanto tipo de crimen como lugar del suceso contenían 70 categorías cada una, que fueron agrupadas en 6 y 5 respectivamente. Con lo que recomienda la literatura (Kuhn & Johnson, 2013) estas características podrían llevarse a un límite de 10 y, aunque existiría un desbalanceo de clases, también es cierto que se ganaría en términos de una mejor comprensión de los resultados.

El presente estudio revela que los delitos más frecuentes en la ciudad están relacionados con extorsión, como fraude, robos y hurtos, y concierto para delinquir, en Manhattan, Brooklyn y Bronx. Estos actos delictivos tienen lugar mayormente en espacios públicos, como calles, zonas residenciales, parques, centros públicos y templos religiosos, donde la presencia policial puede resultar menos costosa. Además, se identificó que las principales víctimas son mujeres y personas transgénero. Por lo tanto, es fundamental que la policía realice requisas y controles principalmente en hombres durante el horario comprendido entre las 7 pm y las 5 am.

Dado que es común encontrar datos incompletos o no identificados en casos relacionados con la criminalidad, se recomienda que las autoridades policiales enfaticen la importancia de identificar y denunciar los hechos de manera integral a la ciudadanía. Asimismo, se sugiere implementar tecnologías de vigilancia para complementar los esfuerzos de las autoridades en la lucha contra la criminalidad en la ciudad, lo cual permitiría validar las denuncias y recopilar información adicional para mejorar la resolución de los casos.

Referencias

Alpaydin, E. (2016). *Machine learning: the new AI. MIT press.*

Becker, G. (1968). *Crimen y castigo: un enfoque económico.*

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 1-27.

Dewia, N. P., & Dwidsamara, I. B. (2021). Implementation of K-Modes Algorithm for Clustering of Stress Causes in University Students. *Jurnal Elektronik Ilmu Komputer Udayana*.

Feder, F. (2018). *NYC Crime Complaints Guided EDA with Shiny App*. Retrieved from Kaggle: https://www.kaggle.com/code/spoons/nyc-crime-complaints-guided-eda-with-shiny-app

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media.

Hamzah, N. A., Kek, S. L., & & Saharan, S. (2017). *The performance of K-means and K-modes clustering to identify cluster in numerical data. Journal of Science and Technology.*

Hastie, T. T. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer Science & Business Media.

Humaira, H., & Rasyidah, R. (2018). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA).* Padang, Indonesia.

López, E. S. (2007). *Algoritmos de Agrupamiento Global para Datos Mezclados (Global Clustering Algorithms for Mixed Data).*

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, 15-24.

Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media.

Reddy, S. (2019). *BigData-Nyu-NYPD-Complaint-Data-HistoricPublic-Safety*. Retrieved from GitHub: https://github.com/sakethreddy997/BigData-Nyu-NYPD-Complaint-Data-HistoricPublic-Safety

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 53-65.

San, J. M., Huynh, V. N., & Nakamori, Y. (2004). *An alternative extension of the k-means algorithm for clustering categorical data.* athematics and Statistics Department at the Co-Operative Degree College, Sagaing, Myanmar, and the Japan Advanced Institute of Science and Technology.

Team, D. D. (2016). *Dask*. Retrieved from Library for dynamic task scheduling: https://ml.dask.org/modules/generated/dask_ml.cluster.KMeans.html

Whalen, J. B., Mladinich, R., & Messing, P. (2019). *El crimen en Nueva York: Los casos más famosos de la historia de la ciudad.* RBA Libros.

Zhou, K. (2019). *NYCrimes*. Retrieved from RPubs: https://rpubs.com/Keezo0227/NYCrimes