Article

MODELAMIENTO GEO-ESPACIAL Y TEMPORAL DEL ROBO EN BARRANQUILLA Y SOLEDAD.

Jorge Arteaga C [1] and Carlos De Oro [2]

[1] Estudiante de la Maestría en Estadística Aplicada. Departamento de Matemáticas y Estadística, Universidad del Norte, Barranquilla 080001, Colombia; jarteagaa@uninorte.edu.co (J.A.) [2]Profesor Asistente. Departamento de Matemáticas y Estadística, Universidad del Norte, Barranquilla 080001, Colombia; cdeoroaguado@uninorte.edu.co (C.O.)

Abstract: El robo es uno de los factores clave que afectan la convivencia en nuestra sociedad. Además, es uno de los delitos con mayor incidencia y afecta severamente el índice de percepción de seguridad que utiliza el DANE. Esto, sumado a la falta de herramientas para explorar el comportamiento del robo, pone a la Policía Nacional en una posición en la que no tienen información ni conocimientos para desplegar adecuadamente sus recursos. El objetivo principal del proyecto fue el desarrollo de un modelo predictivo base que pudiera determinar la cantidad de robos durante un espacio y tiempo específicos, y de este manera permitir que la policía enfoque sus esfuerzos para combatir el crimen de manera eficiente al tomar acciones basadas en hechos reales. Para conseguirlo, se entrenaron diferentes tipos de modelos con el fin de seleccionar el mejor. La lista incluye modelos clásicos para este tipo de problemas, como modelos de regresión poisson, pero también se incluyeron modelos de regresión quasi-poisson y binomial negativo, debido a la sobre-dispersión observada en la variable respuesta, y random Forests regressors. Luego de comparar los modelos, se observó que el modelo de regresión poisson tiene las mejoras métricas en comparativa con los demás modelos. Se construyó un dashboard para mostrar las predicciones del modelo final y se desplegó en Heroku.

Keywords: Modelamiento del crimen; predicción de la cantidad de robos; análisis geo-espacial; regresión poisson; sobre-dispersión.

1. Introducción

Según la página de la policía nacional, la visión de la institución es "el mantenimiento de la convivencia como condición necesaria, para el ejercicio de los derechos y libertades públicas y para asegurar que los habitantes de Colombia convivan en paz fundamentada en el código de ética policial" [11]. Lamentablemente, con la alta frecuencia de criminalidad, esta es una visión muy difícil de lograr. De los crímenes presentes en nuestro país, el robo a personas es uno de los más comunes y frecuentes. Los robos son la forma más habitual de violencia, según las estadísticas presentadas por la OCDE en su informe sobre bienestar en Latinoamérica [3]. También, se observa que los robos a personas son el segundo crimen más frecuente en la encuesta de convivencia y seguridad ciudadana más reciente (2020), detrás de los hurtos a vehículos [5].

Todo esto sumado, a la falta de herramientas para explorar el comportamiento del robo a personas en Colombia, hace que el trabajo de la policía nacional sea más difícil, y afecta en gran manera la percepción de seguridad de los colombianos. La proporción de la población que declaró que la delincuencia era la mayor amenaza para su seguridad personal fue del 55%, lo cual duplicaba el promedio de la OCDE (22%) en 2019 [7].

Por todas estas razones, el uso de herramientas que permitan predecir dónde pueden ocurrir los robos y con qué frecuencia ganan más relevancia. Se decidió trabajar en un modelo que nos permita estimar la cantidad de robos en un área geográfica determinada, dependiendo de cierta temporalidad. Las agrupaciones de los variables se realizaron de acuerdo a la granularidad de los datos, y dependieron de la calidad de éstos.

Citation: Arteaga, J.; De Oro, C. Modelamiento geo-espacial y temporal del robo en Barranquilla y Soledad.. *Journal Not Specified* **2022**, *1*, 0.

Received:
Accepted:
Published:

2. Metodología

En esta sección se definieron los conceptos relevantes para el proyecto, además de identificar el tipo de estudio que se realizó, las variables y área de estudio, fuentes de información y pre-procesamiento de los datos.

2.1. Definición de robo

La definición de robo es "tomar la propiedad o servicios de otra persona sin su consentimiento" [10]. Este término es muy usado en Colombia también para indicar el hurto a personas, cuando se menciona que a alguien "lo robaron". La definición de robo implica el uso de fuerza, o amenaza de ésta. En los datos extraídos de la policía nacional, este delito se tipifica como **Hurto a personas**, por lo cual usaremos ambas palabras indistintamente en lo que resta del documento.

2.2. Área geográfica de estudio

El área geográfica elegida comprende las dos ciudades más grandes del área metropolitana de Barranquilla. Barranquilla tiene una superficie de 154 km² y una población de 1.206.319 (en 2020) [9]. Su densidad de población es de 7.548 hab/km². Está ubicada en el lado noroeste del país. Barranquilla es la ciudad más grande y el segundo puerto de la región de la Costa Caribe Norte. La ciudad celebra una de las fiestas folclóricas y culturales más importantes de Colombia, el Carnaval de Barranquilla, que fue declarado Patrimonio Cultural de la Nación por el Congreso de Colombia en 2001 y reconocida por la UNESCO en 2003.

La otra ciudad es Soledad, la cual tiene una superficie de 67 km² y además es la ciudad con mayor crecimiento demográfico en Colombia, en 2005 tuvo una población de 455.734 y en 2020 creció hasta alcanzar los 683.486 habitantes [14]. Su densidad de población es de 6.802,93 hab/km².

2.3. Fuentes de información

2.3.1. Fuente de datos de la policía

Los datos relacionados con los robos (su ubicación, fecha y hora, así como otra información relevante) fueron obtenidos del Sistema de Información Estadístico, Delincuencial Contravencional y Operativo de la Policía Nacional – SIEDCO [6]. Los datos son accesibles a través de un formulario web, y se pueden descargar los consolidados por año en archivos *xls* (MS Excel). Los datos fueron manualmente procesados para eliminar la información del encabezado y pie de página, y exportados a formato *csv*.

Se descargaron los archivos correspondientes desde el año 2010 hasta el 2019. Dicha información se encuentra desagregada por variables de tiempo, modo y lugar. Cabe mencionar que cada archivo tiene información de todos los municipios de Colombia, por lo que se filtraron los datasets para solo tener información de las ciudades correspondientes.

2.3.2. Fuente de datos geográfica

Los datos geográficos se obtuvieron de diferentes fuentes. Los límites de los barrios en Barranquilla fueron tomados del Plan de Ordenamiento Territorial (POT) de Barranquilla 2012-2032 [13]. Los datos geográficos de Soledad se crearon a partir de *OpenStreetMap*, *Google Maps* y la página del Código Postal de Colombia. El procesamiento se realizó con QGIS 3.10. Las formas de los polígonos se exportaron como un archivo *geoJson*.

2.4. Tipo de estudio

El tipo de estudio del proyecto es descriptivo, porque realizamos un análisis exploratorio de los datos, en el cual se observan ciertas tendencias y patrones. Adicionalmente, el proyecto también posee una parte predictiva, ya que utilizamos los datos para tratar de predecir la cantidad de robos en un espacio y tiempos definidos.

2.5. Variables de estudio

Las variables independientes del proyecto se encuentran en la tabla 1 con su respectivo tipo y su descripción.

Table 1. Variables independientes del proyecto.

Variable	Tipo de variable	Descripción
Fecha	String con formato (DD/MM/AAAA)	Fecha del incidente
Departamento	Categórica	Departamento donde sucedió el incidente
Municipio	Categórica	Municipio donde sucedió el incidente
Día	Categórica	Día de la semana cuando ocurrió el incidente
Hora	String con formato 24H (HH:MM:SS)	Hora en la cual ocurrió el incidente
Barrio	Categórica	Barrio donde ocurrió el incidente
Zona	Categórica	Zona (Urbana o rural) donde ocurrió el incidente
Clase de sitio	Categórica	Tipo de sitio donde ocurrió el incidente
Arma empleada	Categórica	Arma empleada para cometer el incidente
Móvil Agresor	Categórica	Móvil usado por el agresor para cometer el incidente
Móvil Victima	Categórica	Móvil usado por la víctima al momento del incidente
Edad	Numérica	Edad de la víctima
Sexo	Categórica	Sexo reportado de la víctima
Estado civil	Categórica	Estado civil de la víctima
País de nacimiento	Categórica	País de nacimiento de la víctima
Clase de empleado	Categórica	Tipo de empleo de la víctima
Profesión	Categórica	Profesión de la víctima
Escolaridad	Categórica	Escolaridad de la víctima
Código DANE	Categórica	Código DANE de la ciudad

La variable respuesta de nuestro estudio fue **la cantidad de robos ocurrida en el incidente reportado**. Dicha variable es un entero positivo.

2.6. Pre-procesamiento de los datos

El paso inicial que se realizó antes, fue filtrar el dataframe para solo tener incidentes de Barranquilla y Soledad. Después de este primer filtro, tenemos 68.966 incidentes en los 10 años para ambas ciudades. Se unió el dataframe de la información de la policía con el dataframe geográfico (que hace el match de los barrios con su GeoID), y este nuevo dataframe quedó con 68.668 incidentes.

El paso siguiente fue eliminar las columnas que tienen información redundante o no relevante para el análisis:

- **Departamento:** Después de filtrar, solo quedó la opción *ATLÁNTICO*.
- **Zona:** Solo tiene dos opciones, Urbana y Rural, y urbana tiene el 99.17% de los datos.
- **Código DANE:** Este código indica la ciudad, por lo que es información redundante.

La columna *Clase de sitio* poseía muchas opciones, por lo cuál se decidió agrupar las diferentes opciones en unas cuantas categorías, y renombrar la columna como *Categoría de sitio*.

- **VÍAS PUBLICAS Y LOTES.** "LOTE BALDIO", "AREA RURAL", "VIAS PUBLICAS".

- **VIVIENDAS.**
- **SERVICIO Y ENTIDADES PUBLICAS.** "FUERZA PUBLICA", "CARCELES", "AEROPUERTO", "ESTACION TRANSPORTE PUBLICO", "INSTITUTOS DE SALUD", "ENTIDAD PUBLICA", "FUERZA PUBLICA", "SECTOR MARITIMO Y FLUVIAL".
- **OTRO.** "OTRO", "NO REPORTADO", "NO DEFINIDO".
- **LOCALES COMERCIALES.** El resto de opciones.

También se agruparon categorías similares en las columnas de *Móvil agresor*, *Móvil Victima* y *Arma empleada*. La forma en la que se agruparon estas variables se encuentra en el apéndice A. Después de agrupar, se realizó un Data QA en los datos, para confirmar el tipo de datos, y eliminar o imputar datos faltantes. La proporción de datos faltantes para cada columna de datos se puede observar en la tabla 2.

Table 2. Porcentaje de datos faltantes por columna.

Variable	Porcentaje de datos faltantes
Fecha	0.00%
Municipio	0.00%
Día	0.00%
Hora	0.00%
Arma empleada	0.31%
Móvil Agresor	2.76%
Móvil Victima	1.14%
Edad	1.10%
Sexo	0.04%
Estado civil	1.08%
País de nacimiento	0.58%
Clase de empleado	0.04%
Profesión	57.01%
Escolaridad	1.08%
Cantidad	0.00%
GEOID	0.00%
Barrio	0.00%
Categoria de sitio	0.01%

Se puede observar que la mayoría de las columnas no tienen datos faltantes. La columna con mayor porcentaje es *Profesión* con un 57.01% de los datos. Si se imputan los datos de esta columna se agregaría mucho ruido a los datos, por lo cual se eliminó del dataset. Para las columnas restantes, se revisó cada caso específico, para mirar si eliminaban esos incidentes o se imputaban los datos faltantes:

- **Categoria de sitio:** Imputados como "OTRO".
- **Arma empleada:** Imputados como "NO REPORTADO".
- **Móvil Agresor:** Imputados como "NO REPORTADO".
- **Móvil Victima:** Imputados como "NO REPORTADO".
- **Edad:** No imputado, no se usará en los modelos.
- **Sexo:** Imputado como "NO REPORTA".
- **Estado civil:** Imputados como "NO REPORTADO".
- **País de nacimiento:** Imputados como "NO REPORTADO".
- **Clase de empleado.** Imputado como "NO REPORTA".
- **Escolaridad:** Imputados como "NO REPORTADO".

2.7. Feature engineering

Según D. Guo y J. Wu, los datos sobre delitos se pueden dividir en dos categorías principales en función de la representación espacial: datos puntuales (con coordenadas o direcciones de incidentes delictivos) y datos de área (recuentos de incidentes agregados a límites espaciales predefinidos)[2]. Los datos puntuales se pueden convertir (agregar) a datos de área. Sin embargo, el caso contrario no es posible.

Como los datos de la policía ya venían agregados por barrio, se decidió agregar la información disponible para la construcción de un modelo que estimara los robos dependiendo de las variables agregadoras. Entre estas variables, se encontraron tres grupos, la dimensión espacial, que contiene la ciudad, el barrio y la categoría de sitio. La dimensión temporal, la cual tiene fecha y hora específica, y la dimensión de las variables relacionadas con el crimen, tales como el arma usada, características de la víctima, entre otras.

Antes de agregar los datos, se decidió crear otras variables temporales relacionadas con eventos específicos, ya que se ven reportes de prensa sobre incrementos de la actividad delincuencial en estas fechas. Las fechas elegidas fueron carnavales, el día de la madre y las quincena. Para mayor simplicidad, los días de quincena se definieron como los días 14, 15, 16, 28, 29, 30 y 31 de cada mes. Para los otros días especiales, se realizó un web-scrapping manual para conseguir las fechas de cada año, y se agregaron al dataframe.

Con estas nuevas variables ya disponibles en el dataframe, se procedió a agregar la cantidad de robos. Sin embargo, se observó una muy alta granularidad, lo que conllevo a obtener un resultado concentrado en un solo valor (1) en las múltiples combinaciones de las variables. Luego de prueba y error, se concluyó que lo mejor era evitar las variables relacionadas con el crimen (como arma empleada y categoría de sitio), y eliminar algunas variables temporales (las recién añadidas, y el mes), y agrupar aún más otras variables. La variable hora se agrupó en una nueva variable llamada "Timeframe", el proceso se muestra en el apéndice A.3.

Al final, las variables seleccionadas para agregar la cantidad de robos fueron las siguientes: "Municipio", "Barrio", "Year", "WeekDay" y "Timeframe". El histograma para la cantidad de robos en este dataset agregado se muestra en la figura 1.

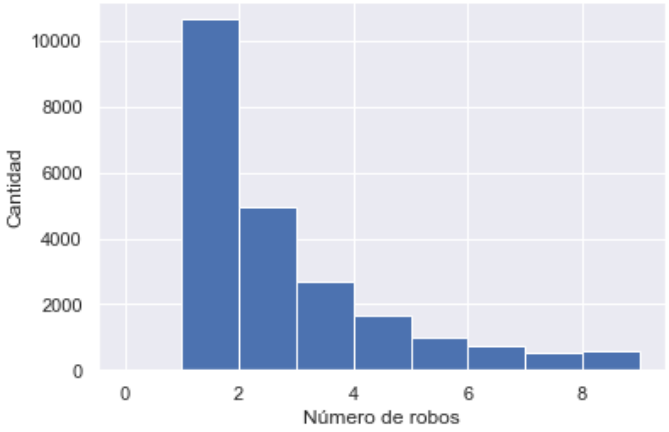

Figure 1. Histograma de la cantidad de robos del dataset agregado.

3. Análisis estadístico

Esta sección contiene el apartado técnico y estadístico del proyecto, desde el análisis exploratorio de los datos, hasta la selección de los modelos y su base matemática.

3.1. Análisis exploratorio de los datos

El análisis exploratorio se realizó sobre los datos desagregados. En la tabla 3 se puede ver la distribución histórica por sexo de los incidentes. Se puede observar que en su mayoría, las víctimas se identificaron como hombres, con un 64.74% de los incidentes, seguido de las mujeres con un 35.21% de los incidentes. Treinta y dos personas no reportaron su sexo, lo cual corresponde a un 0.05%. Es clara la tendencia de los hombres a sufrir más robos que las mujeres.

Table 3. Porcentaje de incidentes por sexo reportado.

Sexo reportado	Cantidad de incidentes (n=68668)	Porcentaje de incidentes
MASCULINO	44456	64.74%
FEMENINO	24180	35.21%
NO REPORTA	32	0.05%

Para la clase de empleado, se observa que el 57.43% de las víctimas reportadas son empleados particulares, seguido por los independientes con un 17.46% de los casos. Luego le siguen los estudiantes, amas de casas y comerciantes con un 6.23%, 3.56% y 3.36% respectivamente. La información se puede ver en la tabla 4

Table 4. Porcentaje de incidentes total para el top 5 de las clases de empleados

Clase de empleo reportado	Cantidad de incidentes (n=68668)	Porcentaje de incidentes
EMPLEADO PARTICULAR	39436	57.43%
INDEPENDIENTE	11994	17.46%
ESTUDIANTE	4280	6.23%
AMA DE CASA	2445	3.56%
COMERCIANTE	2308	3.36%

Para el caso del móvil del agresor en los incidentes, el móvil más frecuente fue "a pie", con un 59.21% de los incidentes, seguido de conductor de motocicleta y pasajero de motocicleta con un 16.01% y 13.02% respectivamente. En la tabla 5 se pueden observar los móviles restantes.

Table 5. Porcentaje de incidentes total por tipo de móvil agresor.

Móvil agresor reportado	Cantidad de incidentes (n=68668)	Porcentaje de incidentes
A PIE	40661	59.21%
CONDUCTOR MOTOCICLETA	10994	16.01%
PASAJERO MOTOCICLETA	8941	13.02%
PASAJERO BUS	1897	2.76%
NO REPORTADO	1895	2.75%
PASAJERO TAXI	1308	1.92%
CONDUCTOR VEHÍCULO	1091	1.59%
CONDUCTOR TAXI	1086	1.58%
BICICLETA	387	0.56%
PASAJERO VEHÍCULO	298	0.45%
CONDUCTOR BUS	60	0.08%
OTROS	50	0.07%

En el caso del móvil de la víctima del robo, se ven concentraciones más acentuadas. El 71.86% iba "a pie", seguido de un 12.32% siendo el conductor de un vehículo. El tercer lugar lo tiene "conductor motocicleta" con un 4.77% de los incidentes. El resto de opciones se pueden ver en la tabla 6. Se ve una clara relación en el caso de los dos móviles, donde lo más común es que el robo suceda a pie.

Table 6. Porcentaje de incidentes total por tipo de móvil víctima.

Móvil víctima reportado	Cantidad de incidentes (n=68668)	Porcentaje de incidentes
A PIE	49348	71.86%
CONDUCTOR VEHÍCULO	8463	12.32%
CONDUCTOR MOTOCICLETA	3277	4.77%
PASAJERO BUS	1985	2.90%
CONDUCTOR TAXI	1263	1.83%
PASAJERO TAXI	1128	1.65%
NO REPORTADO	787	1.15%
PASAJERO MOTOCICLETA	777	1.13%
CONDUCTOR BUS	575	0.84%
PASAJERO VEHÍCULO	553	0.80%
BICICLETA	441	0.65%
OTROS	71	0.10%

Agrupando por país de nacimiento, se tiene que en el 98.5% de los casos las víctimas fueron colombianas. Luego, es seguido por Venezuela con el 0.59% de los casos, y Estados Unidos con el 0.055% de los casos. Como se puede observar, los casos extranjeros solo contribuyen con el 1.5% de los casos en total. El top 5 de nacionalidad se puede ver en la tabla 7.

Table 7. Top 5 Nacionalidad de la víctima por porcentaje de incidentes total.

País de nacimiento	Cantidad de incidentes (n=68668)	Porcentaje de incidentes
COLOMBIA	67641	98.5044%
VENEZUELA	407	0.5927%
NO REPORTADO	398	0.5796%
ESTADOS UNIDOS	38	0.0553%
ESPAÑA	23	0.0335%
MÉXICO	13	0.0189%

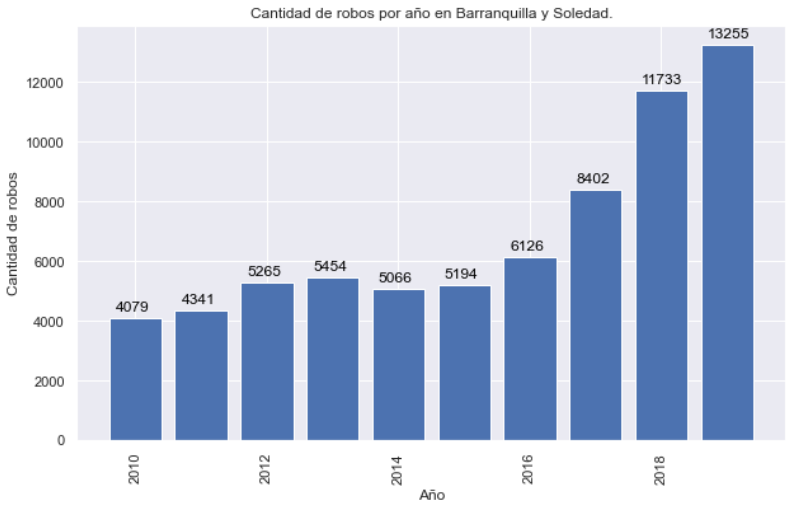

Figure 2. Robos ocurridos en Barranquilla y Soledad por año.

Otra forma de analizar la información es agregando temporalmente los incidentes, agrupando por fechas (o cualquier otra escala temporal). En la figura 2, se puede observar como el número de robos va incrementando año tras año. Se ve una tendencia a la baja

entre 2013 y 2015, pero luego la tendencia cambia rápidamente, aumentando casi al doble entre 2016 y 2018.

En la figura 3, se observa como dependiendo del mes, la cantidad de robos va cambiando ligeramente (Parte (a)). Los meses con mayor incidencia son Mayo y Enero, y el mes con menor incidencia es Abril. Si miramos el otro gráfico (Parte (b)), se observa una diferencia más pronunciada, el conteo de robos sube a medida que la semana va transcurriendo hasta llegar a su pico el viernes, luego empieza a decrecer, hasta llegar al domingo que es el día de la semana con menos incidentes.

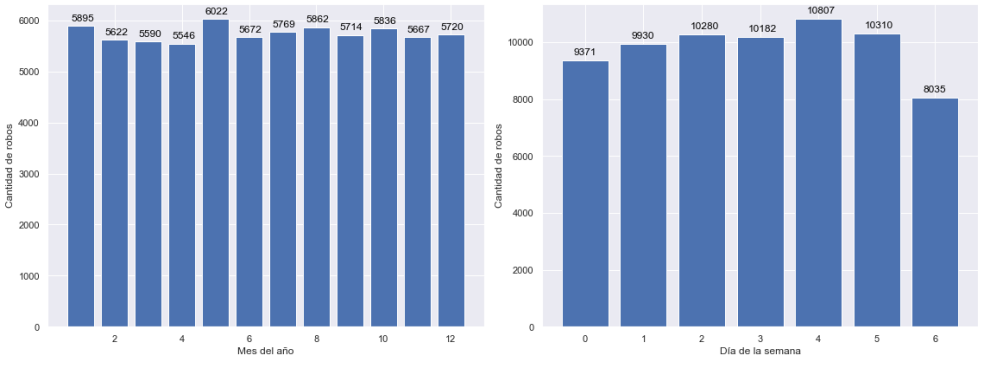

Figure 3. Robos agrupados por mes y día de la semana en Barranquilla y Soledad (sin importar el año). **(a)** Agrupado por mes. **(b)** Agrupado por día de la semana.

En la figura 4, se ve que la hora con mayor incidencia de robos es a media noche, luego la tendencia baja abruptamente y vuelve a subir a lo largo de la mañana para alcanzar aproximadamente el mismo punto a las 10-11AM. Luego de estas horas, la incidencia empieza a bajar paulatinamente.

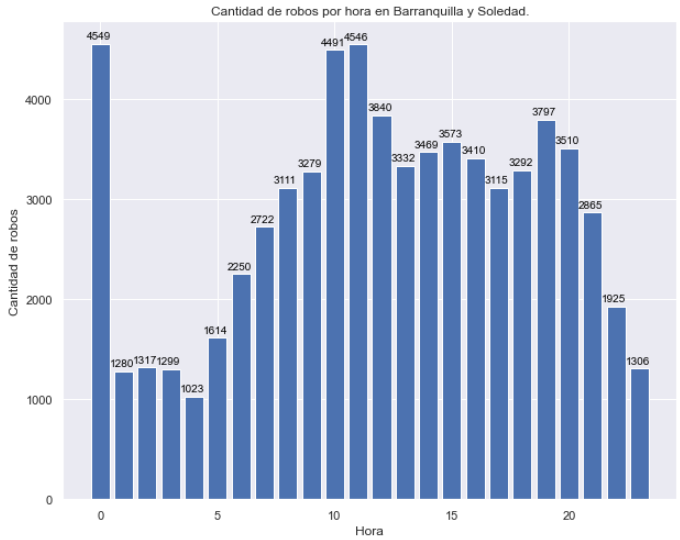

Figure 4. Robos ocurridos en Barranquilla y Soledad por hora.

En la figura 5 se puede observar una serie de tiempo más fina, al agrupar los datos por mes-año y contar los robos ocurridos. Hay ciertas variaciones a lo largo de los meses del mismo año, pero la tendencia general es ascendente. Los incidentes se disparan en febrero de 2017 y siguen creciendo rápidamente hasta marzo de 2019, donde hay una pequeña caída, para luego seguir subiendo.

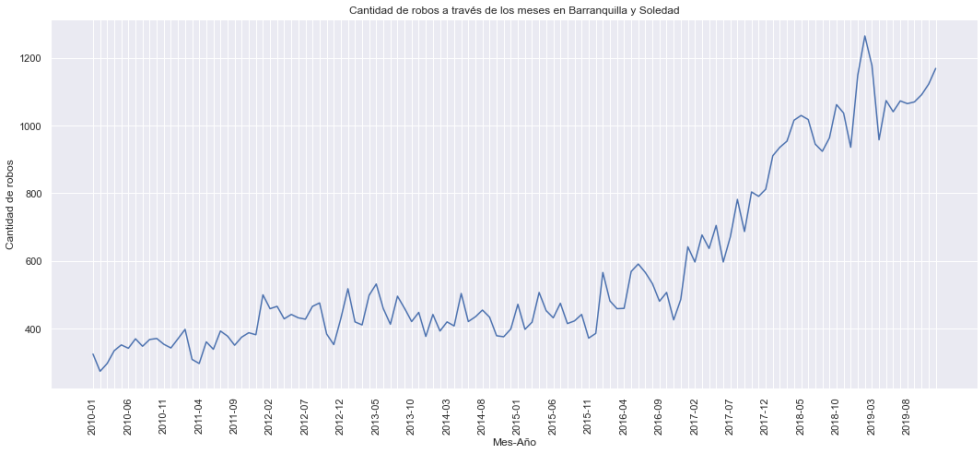

Figure 5. Robos ocurridos en Barranquilla y Soledad a través de los meses y años.

La tabla 8 muestra el porcentaje de incidencias total por arma empleada para cometer el robo. Se observa que el 45.99% de los incidentes fueron cometidos con armas de fuego, seguido de un 29.4% sin uso de armas. El tercer lugar es para las armas blancas/corto-punzantes con un 15.31%.

Table 8. Porcentaje de incidentes total por arma empleada.

Arma reportado	Cantidad de incidentes (n=68668)	Porcentaje de incidentes
ARMA DE FUEGO	31585	45.99%
SIN EMPLEO DE ARMAS	20195	29.40%
ARMA BLANCA / CORTO-PUNZANTE	10555	15.37%
CONTUNDENTES	3923	5.72%
NO REPORTADO	2012	2.94%
ESCOPOLAMINA	398	0.58%

En la figura 6, se puede ver la progresión de la cantidad de robos según el arma empleada. En general, se observan las mismas tendencias para las armas mostradas, exceptuando el segundo y tercer trimestre de 2014, donde "SIN EMPLEO DE ARMAS" superó a "ARMA DE FUEGO" como arma preferida para robar.

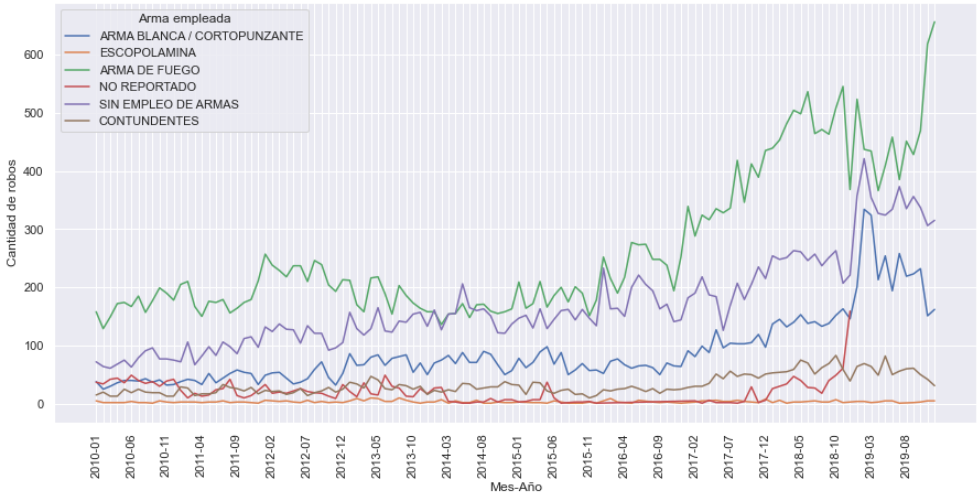

Figure 6. Robos ocurridos en Barranquilla y Soledad a través de los meses y años según el tipo de arma empleada.

La tabla 9 muestra el porcentaje total de incidentes por barrio. El primer lugar es para el barrio "El Centro" con un 6.55% de los casos, seguido del "Prado", "Ciudadela 20 de Julio", "Rebolo" y "Boston" con un 2.90%, 2.57%, 2.54% y 2.22% respectivamente. De estos barrios, algunos tienen mala percepción entre los habitantes, como lo son "El Centro", "Rebolo" y "El Bosque".

Table 9. Porcentaje de incidentes total por barrio (Top 10).

Barrio	Cantidad de incidentes (n=68668)	Porcentaje de incidentes
CENTRO	4508	6.56%
EL PRADO	1995	2.90%
CIUDADELA 20 DE JULIO	1771	2.57%
REBOLO	1750	2.54%
BOSTON	1527	2.22%
LAS TORRES	1344	1.95%
ALTO PRADO	1245	1.81%
SIMON BOLIVAR	1162	1.69%
EL BOSQUE	1126	1.63%
DELICIAS	1066	1.55%

En la figura 7 se puede observar los cambios en las tendencias de los barrios con mayor incidencia de robo a lo largo del mes-año. Si bien, "El Centro" usualmente lleva siempre la mayor cantidad de robos, el año 2017 hubo un crecimiento muy grande de la incidencia en el barrio "Las Torres", los que hizo que este año superara "El Centro". Luego de este año, volvió a sus niveles normales. Sería bueno revisar las noticias de ese año para detectar si hubo o hubieron eventos masivos de delincuencia.

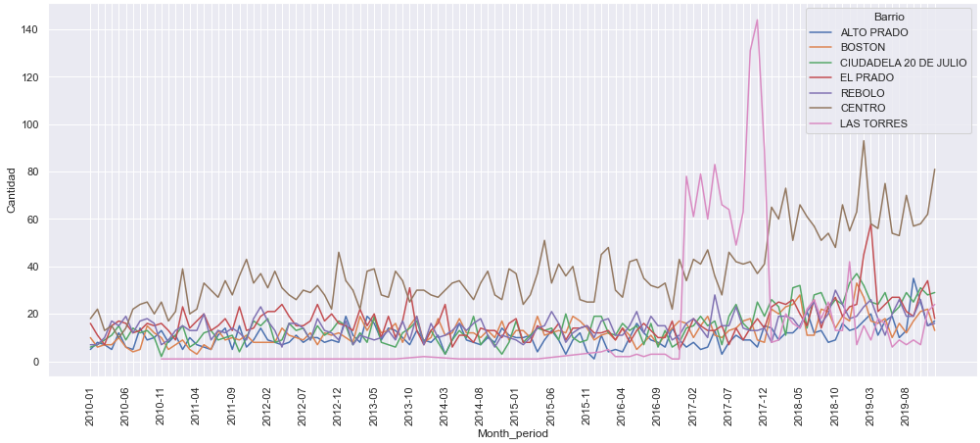

Figure 7. Robos ocurridos en Barranquilla y Soledad a través de los meses y años según el barrio.

En la figura 8, se observa el histograma para la edad de las víctimas de los robos. El 90% de las víctimas tiene entre 19 y 59 años de edad, se observa que el 50% está comprendido entre los 26 y 43 años de edad. La edad promedio de la víctima es de 35.4 años de edad.

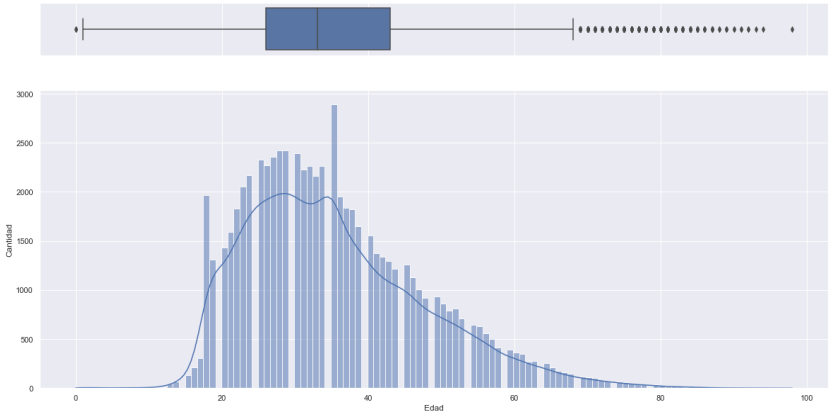

Figure 8. Histograma para la edad de la víctimas.

Revisando los días especiales, los análisis preliminares sugieren que no hay una diferencia tan grande al comparar el promedio diario de robos de un día especial (cualquiera de los tres) con un día normal, y que no hay patrones específicos.

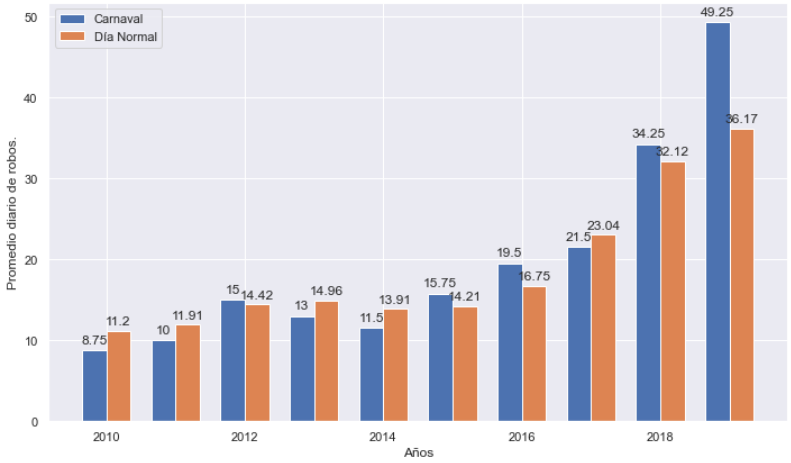

Figure 9. Promedio diario de robos en días de carnaval/días normales por año.

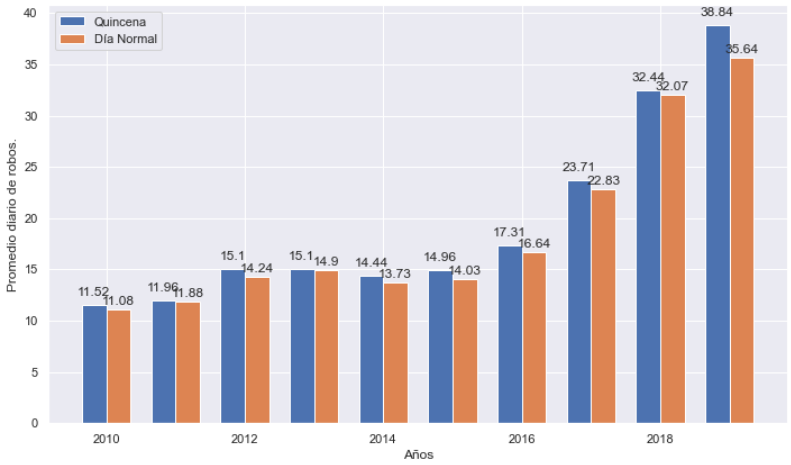

Figure 10. Promedio diario de robos en días de quincena/días normales por año.

Por ejemplo, se ve en la figura 9 que el promedio diario de robos es parecido los días de carnaval y los días normales los primeros años reportados, con pequeñas diferencias inter-anuales. Sin embargo, se observa que los dos últimos años, los días de carnaval

empiezan a tener diferencias cada vez más grandes. El mismo comportamiento se observa los días de quincena, aunque en este caso la tendencia es que en general en quincena el promedio de robos sea mayor que en un día normal.

Por último, para el día de las madres no hay patrones claros, y depende particularmente del año para indicar si el día de las madres tuvo un promedio de robos más alto que un día normal.

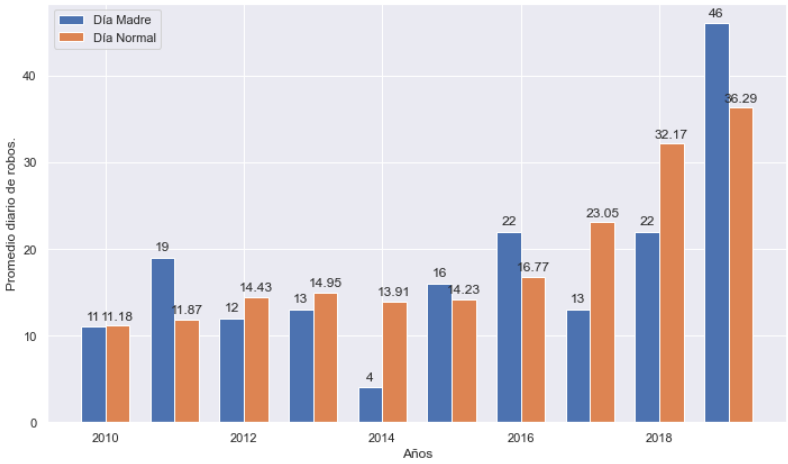

Figure 11. Promedio diario de robos en día de madres/días normales por año.

3.2. Modelo de regresión poisson

La necesidad de contar cosas siempre ha estado presente a lo largo de la historia, por lo que este tipo de problemas surgen frecuentemente. En este caso particular, se cuentan los robos ocurridos por año en cierto barrio, cierto día de la semana, a cierta franja horaria.

La distribución más usada para modelar conteos es la distribución de Poisson [12], la cual tiene una función de probabilidad dada por:

$$\mathcal{P}(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} \tag{1}$$

para $y = 0, 1, 2, \ldots$, con conteos esperados $\mu > 0$. Para la distribución poisson, se da que la media y la varianza son la misma y su valor es igual al parámetro μ.

La función de enlace más común para los modelos lineales generalizados de Poisson es la función de enlace logarítmica (la cual es la función de enlace canónica) [12]. Esta función de enlace asegura que $\mu > 0$ y hace que los parámetros de la regresión puedan ser interpretados al tener efectos multiplicativos [12]. Usando la función logarítmica en R, la forma general de un GLM Poisson es la siguiente:

$$\begin{cases} y \sim \text{Pois}(\mu) \\ \log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \end{cases} \tag{2}$$

Con esto en mente, se entrenó el modelo GLM de Poisson con las variables "Barrio", "WeekDay" y "Timeframe" en R usando el paquete glm(). El modelo cuenta la cantidad de robos en un año para todos los barrios en un día de la semana y franja horaria específica.

3.3. Sobre-dispersión y GLM quasi-poisson/binomial negativo.

Como se había mencionado anteriormente, una distribución de Poisson debe cumplir que la varianza y la media sean iguales. Es decir, var[y]=μ. Sin embargo, en la práctica la varianza aparente usualmente excede la media observada [12]. Esto es lo que se conoce como sobre-dispersión. La sub-dispersión también ocurre, pero es menos frecuente [12].

Para saber si nuestra variable respuesta presenta sobre-dispersión, una primera aproximación es calcular la media y la varianza, y comprobar si estos valores son parecidos.

$$\bar{x} = 2.979 \tag{3}$$

$$\text{var} = 18.656 \tag{4}$$

Como se observa, la varianza es mucho más grande que la media observada en la variable respuesta. Esto es un indicio de sobre-dispersión. Para detectar con mayor precisión la sobre-dispersión, se puede calcular el coeficiente deviance/gl, o el coeficiente pearson-chi2/gl con los datos arrojados por el modelo de Poisson. Los cálculos están a continuación:

$$\text{deviance/gl} = \frac{25917}{22853} = 1.134 \tag{5}$$

$$\text{pearson-chi2/gl} = \frac{31826.23}{22853} = 1.392 \tag{6}$$

En ambos casos el coeficiente es mayor a 1, lo que indica sobre-dispersión. Una manera de modelar la sobre-dispersión es a través de un modelo jerárquico. En vez de asumir que la variable respuesta tiene una distribución Poisson, se puede añadir una segunda capa de variabilidad permitiendo que el parámetro μ sea una variable aleatoria [12].

Dos modelos que cumplen con estas condiciones son los GLMs con distribución binomial negativa, y con distribución quasi-poisson [12]. Dichos modelos producen sobre-dispersión relativa a la distribución Poisson pero cada modelo asume relaciones diferentes entre la media y la varianza [12]. Los modelos quasi-poisson asumen una función lineal para la varianza $(V(\mu) = \phi\mu)$, mientras que los modelos binomiales negativo usan una función cuadrática $\left(V(\mu) = \mu + \mu^2/k\right)$ [12]. Se entrenaron ambos modelos, y se consolidaron todos los coeficientes en la tabla 10.

Table 10. Consolidado de los coeficientes para identificar sobre-dispersión.

Modelo entrenado	deviance/gl	pearson-chi2/gl
Poisson	1.134	1.392
Binomial negativo	0.726	0.859
Quasi-poisson	1.139	1.389

Se puede observar que el modelo quasi-poisson tiene valores muy parecidos a los del modelo poisson, sin embargo, el modelo binomial si tiene valores menores a cero en los coeficientes, por lo cual no presenta sobre-dispersión.

3.4. Random Forests regressors

Por último, además de entrenar los tres modelos anteriores, por la naturaleza de las variables independientes (todas son factores), se decidió entrenar un random Forest regressor, ya que estos modelos tienen buenos resultados para este tipo de variables.

4. Resultados

Se presentan los resultados obtenidos de la comparación para los cuatro modelos entrenados, y adicionalmente se presenta el modelo final en un dashboard accesible desde internet. En el caso del modelo de Poisson, también se entrenó una variante adicional en Python con el módulo *scikit-learn* utilizando cross-validation y gridSearch para escoger el parámetro de regularización. Se decidió entrenar en un 80% del dataset, y dejar un 20% restante para el test. La semilla usada fue 30.

4.1. Métricas

La tabla 11 muestra las métricas usadas para evaluar los modelos. Se calcularon el MAE (Mean Absolut Error), RMSE (Root Mean Squared Error), y el ajuste del modelo

(R-cuadrado). Si bien, las mejores métricas se observan para el random Forest regressor, con un MAE de 1.28 y RMSE de 2.90 en el train set, se observa que el ajuste del modelo es muy alto en el train (0.67) en comparación con el test (0.39). Esto sumado a que la diferencia entre las métricas del train y el test son mayores que en los otros modelos, arroja fuertes indicios de over-fitting.

Table 11. Métricas de predicción para los modelos entrenados.

Modelo entrenado	**Modulo usado**	**MAE**	**RSME**	R^2
Poisson (train)	R-glm()	2.166485	4.518292	0.3227721
Poisson (test)	R-glm()	2.144079	4.549835	0.2969717
Binomial negativo (train)	R-glm()	2.157172	4.522903	0.3235474
Binomial negativo (test)	R-glm()	2.134694	4.552695	0.2995746
Quasi-poisson (train)	R-glm()	2.166485	4.518292	0.3227721
Quasi-poisson (test)	R-glm()	2.144079	4.549835	0.2969717
Random forest regressor (train)	Python - sklearn	1.280827	2.906988	0.673685
Random forest regressor (test)	Python - sklearn	1.5570119	3.695990	0.395518
Poisson con CV (train)	Python - sklearn	1.60117	3.29834	0.338394
Poisson con CV (test)	Python - sklearn	1.6381914	3.745823	0.319877

Para evitar el over-fitting, se pueden tomar diferentes caminos, ya sea usar cross-validation o tunear diferentes parámetros del modelo, pero por poder computacional y limitaciones de tiempo se decidió no implementar estas opciones. Comparando los modelos restantes, el siguiente modelo con mejores métricas es el de Poisson entrenado en *scikit-learn* con cross-validation y gridSearch. Este fue el modelo seleccionado, y puede explicar aproximadamente un 31% de la variabilidad de los datos, con un MAE de 1.161 y RMSE de 3.74 aproximadamente.

4.2. Dashboard

Una vez seleccionado el modelo final, se decidió disponibilizar el modelo en un dashboard, usando la capa gratis que ofrece el servicio web **Heroku**. Para el dashboard, se pensó en una barra lateral donde se pudieran escoger las variables temporales (día de la semana y franja horaria), y al cambiarlas el dashboard realiza las predicciones del conteo de robos en todos los barrios. Los conteos se muestran en un mapa cloroplético, ya que este tipo de gráficos permite observar patrones espaciales [1] [4].

El dashboard se construyó usando los módulos de **Dash** de *plotly*. El modelo final se reentrenó con todos los datos, y se exportó en un archivo binario *.pickle* para poder realizar las predicciones bajo demanda. El modelo final está disponible en el siguiente link: https://dashboard-thefts-baq.herokuapp.com/dash/. En la figura 12 se muestra la versión final. Adicionalmente, se adjunta la estructura y el link del repositorio de los datos en el apéndice B.

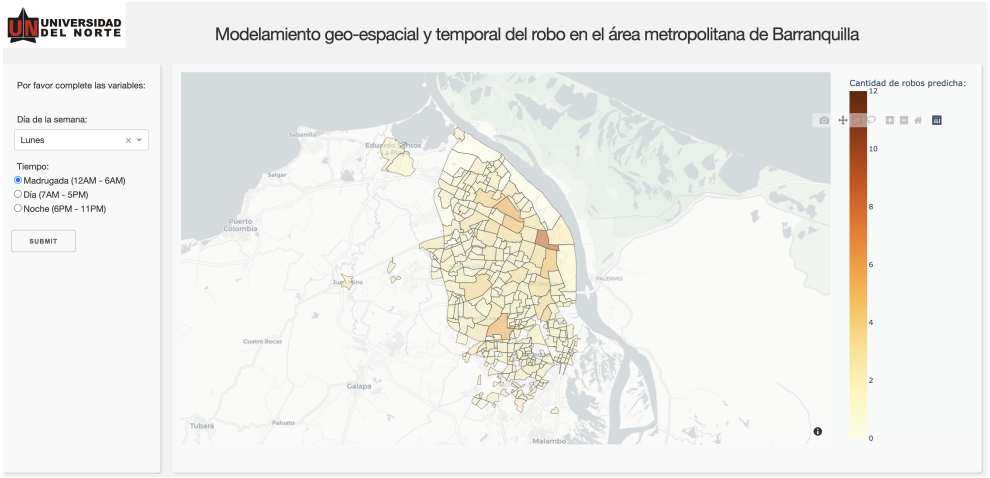

Figure 12. Versión final del dashboard desplegado en Heroku.

5. Discusión y conclusiones

Como punto de partida, el estudio reveló un aumento importante inter-anual en la cantidad de robos en las ciudades de Barranquilla y Soledad. Este aumento es mucho más acentuado desde 2016, y la cifra de robos llega a duplicarse en 2018 (comparando con 2016). Tocaría investigar con las autoridades competentes si efectivamente hubo un aumento de la delincuencia en esos años (analizando más datos), o si simplemente antes había problema de sub-reportes, es decir, que la gente sufría de los robos, pero no los denunciaba.

Comparando a una escala mensual, no se observa mucha variación en la cantidad de robos mes a mes. Sin embargo, si se toma una escala semanal si se observan cambios día a día, siendo el viernes el día con mayor incidencia de robos, y el domingo el día con menor incidencia. Se observa también un patrón si se compara hora a hora, la hora con mayor incidencia de robos es a media noche, luego la tendencia baja abruptamente y vuelve a subir a lo largo de la mañana para alcanzar aproximadamente el mismo punto a las 10-11AM. Luego de estas horas, la incidencia empieza a bajar paulatinamente.

Al observar las características relacionadas con los robos, también se tienen resultados interesantes. Del lado de las víctimas, el 64.74% de éstas se identificó como hombre, seguido de un 35.21% identificadas como mujer. Usualmente las víctimas están caminando cuando sufren el robo (71.86% de los casos), tienen entre 19 a 59 años de edad (el 90%) y el tipo de arma más común son las armas de fuego con un 45.99% de los casos.

A nivel geo-espacial, se observó que la mayoría de barrios con mayor incidencia de robos (tales como *El Centro*, *Rebolo*, *Simón Bolívar* y *El Bosque*) coinciden con la percepción popular de zonas con menor índice de seguridad. Otro punto a destacar es que se observan mayor incidencia de robos para los días de carnavales y quincenas en los últimos años reportados (2018, 2019). En el caso del día de la madre, no hay datos concluyentes para afirmar que es un día con mayor peligrosidad, y depende en parte del año reportado.

Pasando al pre-procesamiento de los datos, lamentablemente los datos ya venían agregados a nivel de barrio, por lo cual no se pudieron implementar métodos de distancias, auto-correlación espacial ni estimaciones de densidad. También se observó que al tratar de agregar los datos incluyendo variables relacionadas con el incidente, y features temporales, la variable respuesta tenía una acumulación muy fuerte en un único valor (un robo por cada combinación).

En otros artículos, se ha reconocido la importancia de añadir variables exógenas relacionadas con el ambiente socio-económico de los barrios afectados para entender el contexto y los procesos que influencian la variación espacio-temporal del crimen[8]. Desafortunadamente, acceder a datos históricos relacionados con dichas variables fue sumamente difícil. Para próximos trabajos, se recomendaría buscar en otras bases de datos o establecer contactos con las autoridades competentes.

En el caso de los modelos, se observó que aunque el modelo binomial negativo soluciona el problema de sobre-dispersión presentado en los datos, esto no representa mayor mejora en las métricas de predicción. Al final, se selecciona el modelo de regresión de Poisson entrenado con gridSearch y cross-validation en *scikit-learn*, ya que tiene las mejores métricas y no se ven indicios de over-fitting, como en el caso del random forest regressor.

Se recomienda para próximos trabajos hacer un trabajo de fine-tunning con cross-validation en este último modelo, ya que se pueden obtener mejores resultados y evitar el over-fitting. También se recomienda probar con modelos de árboles que usen gradient Boosting, tales como xGBoost.

Acknowledgments: Además de agradecer a mi tutor de proyecto, Carlos De Oro Aguado, quiero también agradecer a los profesores Karen Floréz Lozano y Manuel Mendoza Becerra por su apoyo y guía en el desarrollo de este proyecto de grado.

Por último pero no menos importante, también quiero agradecer a mi familia por apoyarme a lo largo de estos dos años de carrera.

Abreviaciones

Las siguientes abreviaciones fueron usadas en este documento:

DANE	Departamento Nacional de Estadística
GLM	Generalized Linear Model
OCDE	Organización para la Cooperación y Desarrollo Económico
UNESCO	United Nations Educational, Scientific and Cultural Organization

Appendix A. Pre-procesamiento de los datos

Adicional a la variable *Clase de sitio*, otras variables también se agruparon para tener menos categorías. Las agrupaciones que se realizaron se muestran a continuación.

Appendix A.1. Móvil Agresor y Móvil Víctima

- **OTROS.** "PASAJERO METRO", "PASAJERO BARCO", "TRIPULANTE AERONAVE", "PASAJERO AERONAVE".

Appendix A.2. Arma Empleada

- **ARMA BLANCA / CORTOPUNZANTE.** "ARMAS BLANCAS", "JERINGA", "CORTANTES".
- **CONTUNDENTES.** "LLAVE MAESTRA".
- **SIN EMPLEO DE ARMAS.** "PERRO".

Appendix A.3. Timeframe

En la sección de feature engineering, se decidió agregar ciertas variables para evitar la alta granularidad de los robos. La variable hora se agrupó tomando el ejemplo de D. Guo y J. Wo. en su pre-procesamiento[2]. Se crearon tres grupos:

- **Madrugada:** De 12AM hasta las 6AM (inclusive).
- **Día:** De 7AM hasta las 5PM (inclusive).
- **Noche:** De 6PM hasta las 11PM (inclusive).

Appendix B. Repositorio de los datos

El repositorio con los datos, notebooks y código fuente del dashboard puede ser descargado del siguiente link: https://github.com/jorarcar/theft_project. La estructura del respositorio es la siguiente:

DS4A-data-source-master: Carpeta donde se encuentran los archivos *geoJson* y los datos originales descargados de la policía nacional.

dashboard-thefts-local: Carpeta donde está el código del dashboard.
images: Carpeta donde se guardan las imágenes usadas para el documento final.
Data_QA_EDA_Thefts.ipynb: Jupyter-notebook donde se realizó el Data QA, transformación, EDA, feature engineering y entrenamiento de modelos.
Theft_modeling.Rmd: R-markdown notebook donde se realizó parte del entrenamiento de los modelos.
datos_merged.csv: CSV que agrupa todos los datos de la policía nacional y datos geográficos después de agrupar las categorías.
df_merged_clean.csv: CSV que tiene los datos limpios y listos para el EDA.
df_model_dias_esp.csv: df_merged_clean con los datos de los días especiales (Carnavales, Día de la Madre, Quincena).

Referencias

1. Chainey S, Tompson L, Uhlig S. The utility of hotspot mapping for predicting spatial patterns of crime. Secur J 21:4–28.
2. D. Guo, J. Wu, Understanding Spatiotemporal Patterns of Multiple Crime Types with a Geovisual Analytics Approach. In *Crime Modeling and Mapping Using Geospatial Technologies*; Leitner, M, (ed.); Springer Science+Business Media Dordrecht 2013; pp. 368, 374.
3. "En los países analizados sobre los que hay datos disponibles, los robos son la forma más habitual de violencia", en ¿Cómo va la vida en América Latina?: Medición del bienestar para la formulación de políticas públicas, OECD Publishing, Paris, Disponible online: https://doi.org/10.1787/e95f39a2-es (Revisado el 2 de Junio de 2022).
4. Eck JE, Chainey S, Cameron JG, Leitner M, Wilson RE (2005) Mapping crime: understanding hotspots. In: NIJ special report. Disponible online en: https://www.ncjrs.gov/pdffiles1/nij/209393.pdf
5. Encuesta de Convivencia y Seguridad Ciudadana (ECSC). Disponible online: https://www.dane.gov.co/index.php/estadisticas-por-tema/seguridad-y-defensa/encuesta-de-convivencia-y-seguridad-ciudadana-ecsc (Revisado el 2 de Junio de 2022).
6. Estadística delictiva, SIEDCO. Disponible online: https://www.policia.gov.co/grupo-informaci%C3%B3n-criminalidad/estadistica-delictiva (Revisado el 3 de Enero de 2022).
7. Gallup World Poll, Disponible online: https://www.gallup.com/analytics/232838/world-poll.aspx (Revisado el 2 de Junio de 2022.)
8. Hagenauer J, Helbich M, Leitner M, Visualization of crime trajectories with self-organizing maps: a case study on evaluating the impact of hurricanes on spatio-temporal crime hotspots. In *Proceedings of the 25th conference of the International Cartographic Association, Paris*.
9. Información capital, DANE. Disponible online: https://www.dane.gov.co/files/varios/informacion-capital-DANE-2019.pdf (Revisado el 2 de Junio de 2022). pp. 40.
10. Kaplan, J. Weisberg, R. Binder, G. Criminal Law - Cases and Materials (ed. 7) Wolters Kluwer. Law & Business. ISBN 978-1-4548-0698-1.
11. Misión, visión, mega, valores, principios y funciones. Disponible online: https://www.policia.gov.co/mision-vision-mega-principios-valores-funciones (Revisado el 2 de Junio de 2022).
12. P. K. Dunn, G. K. Smyth, Generalized Linear Models with Examples in R. Chapter 10, Models for Counts: Poisson and Negative Binomial GLMs. Publisher: Springer Science+Business Media, LLC, part of Springer Nature 2018; pp. 371-372, 397,399,402.
13. Plan de ordenamiento territorial de la ciudad de Barranquilla. Disponible online en: https://www.barranquilla.gov.co/transparencia/planeacion/politicas-lineamientos-y-manuales/planes-estrategicos/plan-de-ordenamiento-territorial (Revisado el 30 de Junio de 2019.)
14. Proyección municipios 2005-2020, DANE. Disponible online: http://www.dane.gov.co/files/investigaciones/poblacion/proyepobla06_20/ProyeccionMunicipios2005_2020.xls (Revisado el 2 de Junio de 2022).