ANEXO D

Metodología analítica CRISP

1. Entendimiento del negocio

1.1. Estado actual

En Colombia para el hurto a personas se registraron en promedio 830 casos diarios, lo que quiere decir que cada hora 35 personas están siendo víctimas de un robo en las calles del país. Las horas con mayor reporte de hurtos están ubicadas entre las 6:00 a.m. y las 11:59 a.m. donde se reportaron hasta 30% más casos el año pasado que en el promedio general [1]. En la capital Bogotá D.C. se dieron a conocer cifras de hurtos del año 2020, el concejal Emel Rojas de Colombia Justa Libres reveló que según las cifras oficiales en los primeros días del mes de enero, se presentaron 7.328 hurtos a personas en la ciudad, es decir un promedio de 271 robos por día [2].

1.2. Objetivos de negocio

- En cuanto al proyecto de grado RMFAR, crear una fuente de datos potencial que sirva como base para crear reglas de asociación y utilizarlas para crear recomendaciones para mitigar el riesgo de delito de hurto.
- Analizar y experimentar mediante las variables más importantes para identificar las localidades más interesantes, los periodos de tiempo más importantes y otros aspectos de interés.

1.3. Criterios de éxito del negocio

- Creación de una base sólida de al menos 5000 registros para la generación de reglas de asociación.

1.4. Objetivos de la minería

- Elaborar un clasificador de machine learning para categorizar los tipos de delito en delito de hurto a personas y otro tipo de delito.
- Estimar la probabilidad de que un delito sea de tipo hurto a personas u otro tipo de delito.

1.5. Criterios de éxito de los objetivos de minería

- Obtener un rendimiento considerable y una precisión por encima del 60%.

2. Entendimiento de los datos

2.1. Recolección de los datos

- (FIP) Base de datos de homicidios en Bogotá D.C. desde el año 2016
- (SIEDCO) Base de datos de delitos de todo tipo para Bogotá D.C. desde el año 2016.
- (UAECD) Base de datos de Sectores Catastrales en Bogotá D.C.
- (SDM) Base de datos de índices de seguridad nocturna en Bogotá D.C.

2.2. Descripción de los datos

- (FIP)

Variable	Tipo de dto	Tipo de atributo	Descripción
year	número	Escalar discreta	Año en el que ocurrio el homicidio
fecha	fecha	Escalar discreta	Fecha en la que ocurrio el homicidio
dia	número	Escalar discreta	Dia en el que ocurrio el homicidio
mes	número	Escalar discreta	Mes en el que ocurrio el homicidio
hora	número	Escalar discreta	Hora en la que ocurrio el homicidio
hora_grupo	número	Categorica Ordinal	Grupo de hora (1,2,3 o 4) en el que ocurrio el homicidio
zona	categorica	Categorica Nominal	Zona en la que ocurrio el homicidio
barrio	categorica	Categorica Nominal	Barrio en el que ocurrio el homicidio
clase_sitio	categorica	Categorica Nominal	Clase del sitio en el que ocurrio el homicidio
arma_empleada	categorica	Categorica Nominal	Arma empleada durante el homicio
movil_agresor	categorica	Categorica Nominal	En que se movilizaba el agresor del homicidio
movil_victima	categorica	Categorica Nominal	En que se moviliaba la victima del homicio
edad	número	Escalar discreta	Edad de la victima del homicidio
sexo	categorica	Categorica Nominal	Sexo de la victima del homicio
clase_empleado	categorica	Categorica Nominal	Clase de empleado de la victima del homicidio
escolaridad	categorica	Categorica Nominal	Nivel escolar de la victima del homicidio
profesion	categorica	Categorica Nominal	Profesión de la victima del homicidio
estado_civil	categorica	Categorica Nominal	Estado civil de la victima del homicidio
pais_nace	categorica	Categorica Nominal	Pais de nacimiento de la victima del homcidio

Tabla 1: Columnas de la base de datos de homicidios desde el año 2016 (FIP)

- (SIEDCO)

Variable	Tipo de dto	Tipo de atributo	Descripción
fecha	fecha	Escalar discreta	Fecha en la que ocurrio el delito
barrios	fecha	Categorica Nominal	Barrio en el que ocurrio el delito
zona	categorica	Categorica Nominal	Zona en la que ocurrio el delito
clase sitio	categorica	Categorica Nominal	Clase del sitio en el que ocurrio el delito
armas medios	categorica	Categorica Nominal	Armas empleadas por el agresor durante el delito
genero	categorica	Categorica Nominal	Genero de la victima del delito
delitos	categorica	Categorica Nominal	Tipo de delito

Tabla 2: Columnas de la base de datos de hurtos desde el año 2016 (SIEDCO)

- (UAECD)

Variable	Tipo de dto	Tipo de atributo	Descripción
scacodigo	fecha	Escalar discreta	Codigo del SCA
scanombre	fecha	Categorica Nominal	Nombre del SCA
objectid	categorica	Categorica Nominal	ID del SCA
geometry	object	Geojson	Geojson del poligono del SCA

Tabla 3: Columnas de la base de datos de SCA (UAECD)

- (SDM)

Variable	Tipo de dto	Tipo de atributo	Descripción
upInombre	número	Categorica Nominal	Nombre de la UPZ
upIcodigo	número	Escalar discreta	Codigo de UPZ
fecha	número	Escalar discreta	Timestamp de la toma de los indicadores
t_que_veo	número	Escalar discreta	Total ¿Qué veo?
t_quienes	número	Escalar discreta	Total ¿Quién me ve?
t_transpor	número	Escalar discreta	Total de transporte
t_sendero	número	Escalar discreta	Total de sendero
t_presenci	número	Escalar discreta	Total de seguridad
t_presen_1	número	Escalar discreta	Total de diversidad de genero
t_personas	número	Escalar discreta	Total de personas en la UPZ
loccodigo	número	Categorica Nominal	Codigo de la localidad a la que pertenece la UPZ

Tabla 4: Columnas de la base de datos de UPZ (SDM)

2.3. Exploración de los datos

Para el dataset de FIP de delitos de homicidio desde el año 2010 hasta 2018 en Bogotá D.C. se realizó la exploración de los datos (ver Figura 12) en la cual se encuentran aspectos importantes que dan un amplio panorama del histórico de los delitos de homicidios en la ciudad de Bogotá D.C., los supuestos más importantes son:

- La localidad con más homicidios es Ciudad Bolívar con 2160 casos de homicidios.

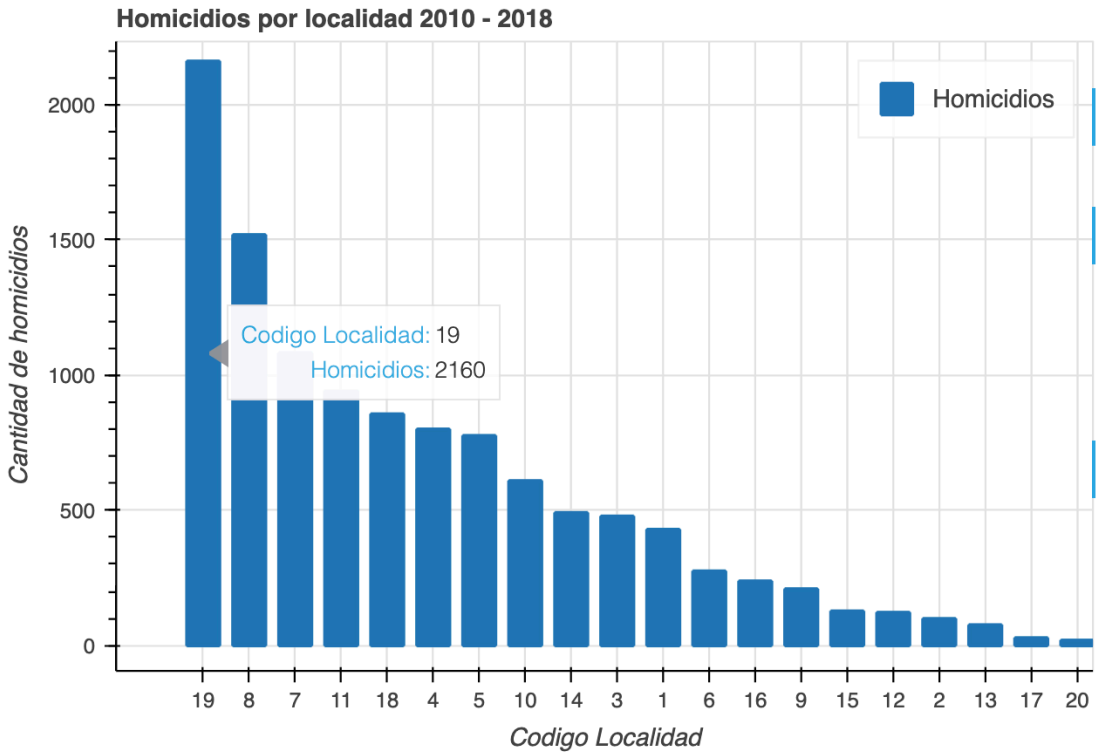


Figura 1: Homicidios por localidad

- Los casos de homicidio han tenido una tendencia a disminuir a partir del año 2015.

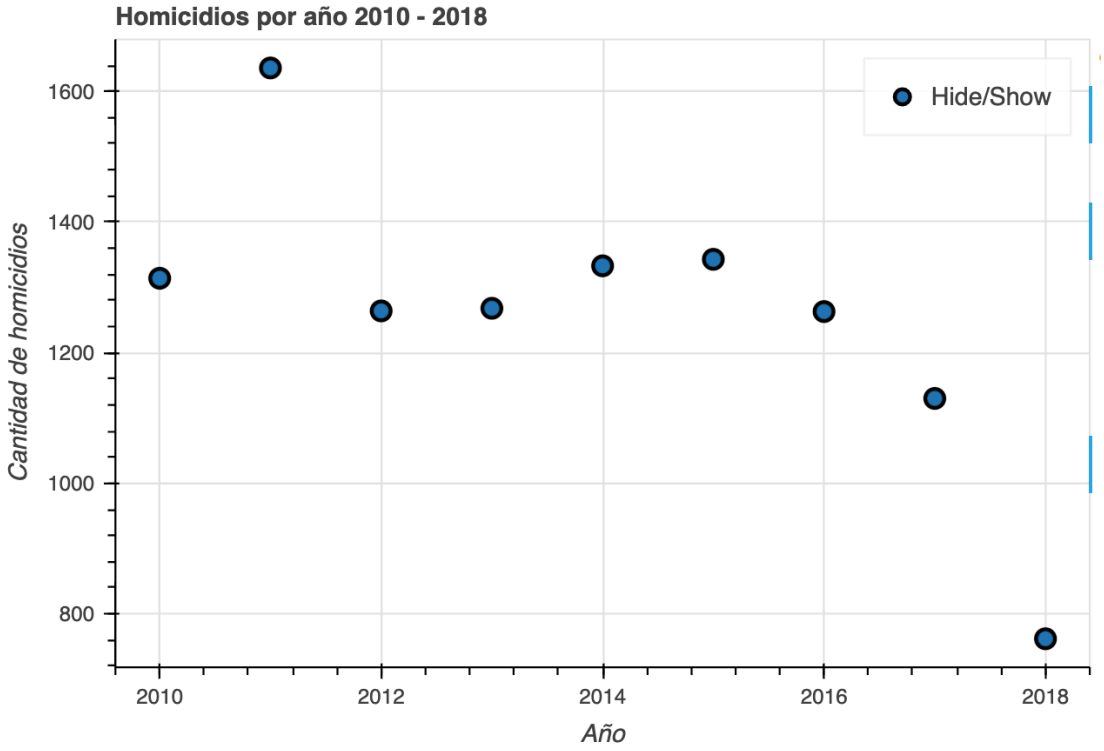


Figura 2: Homicidios por año

- Los homicidios en víctimas de género masculino es 10 veces los homicidios en víctimas de género femenino.

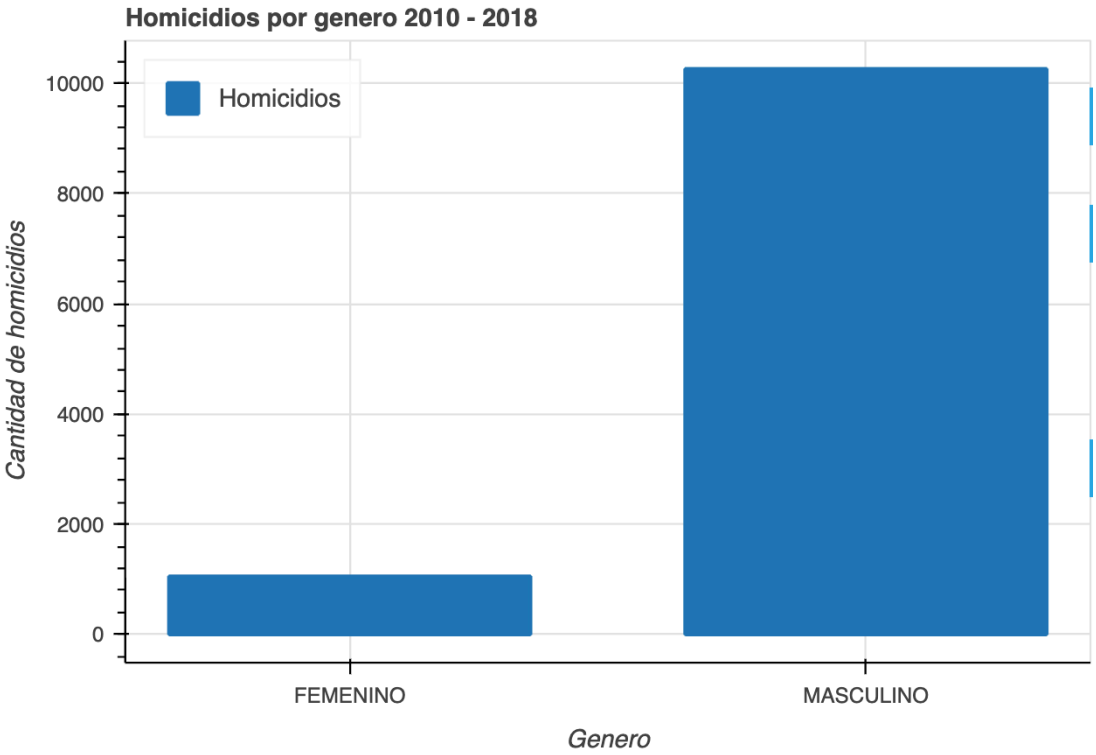


Figura 3: Homicidios por genero

- Las edades más frecuentes de las víctimas de homicidio están entre los 18 y los 40 años.

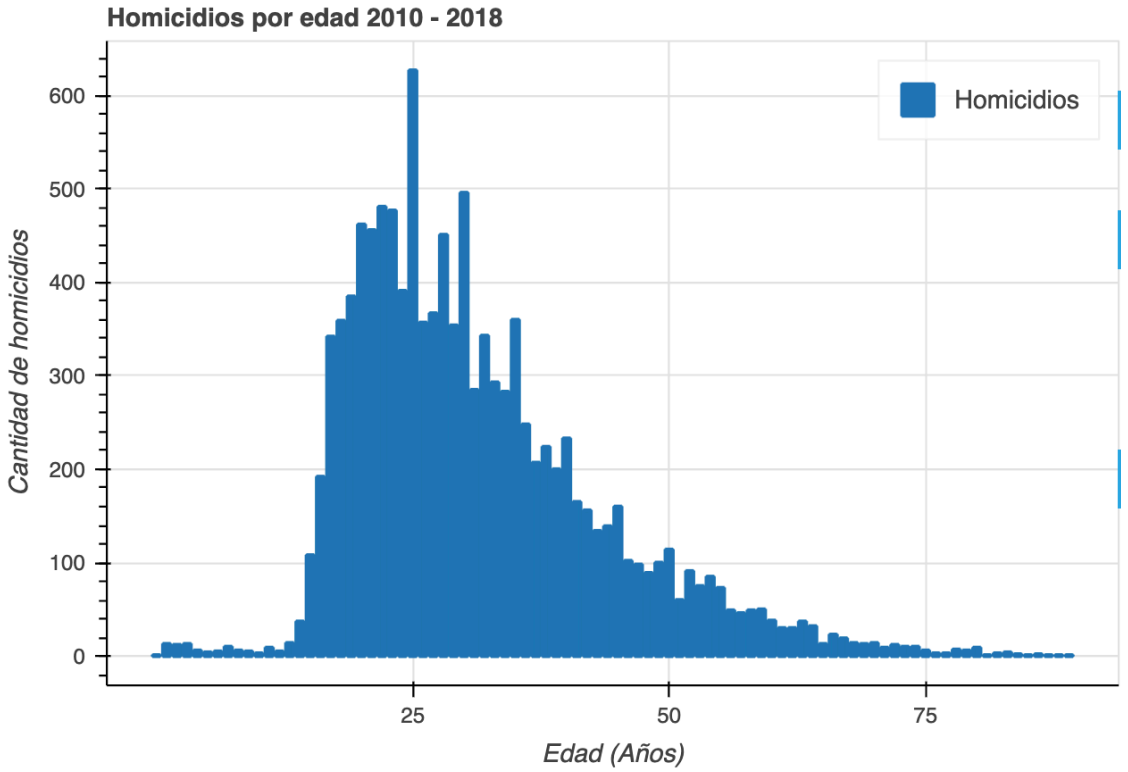


Figura 4: Homicidios por edad

- Las horas en las que más se presentan homicidios es en la noche y en la madrugada.

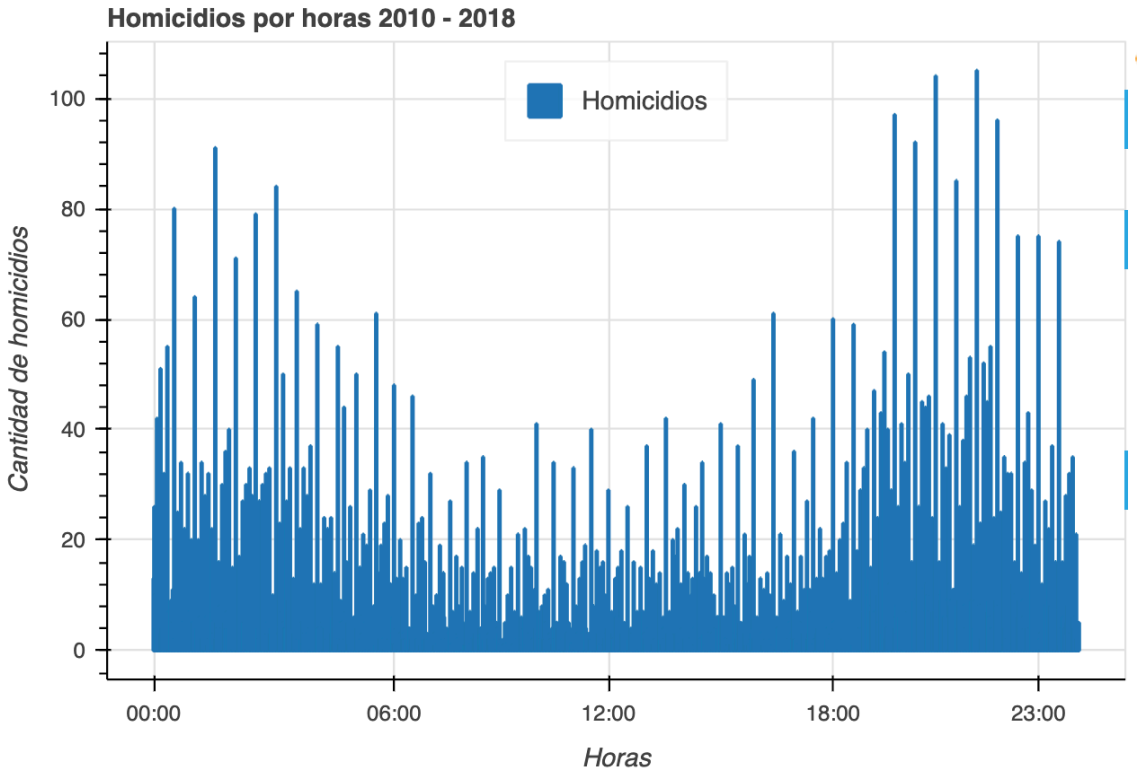


Figura 5: Homicidios por horas

Para el dataset de SIEDCO de delitos de hurto desde el año 2016 hasta el 2020 en Bogotá D.C. se realizó la exploración de datos (ver Figura 13) en la que se encontraron las siguientes premisas.

- La localidad con más delitos de hurto de todo tipo es la localidad de Suba con 53852 casos, seguido de Kennedy y Engativá.

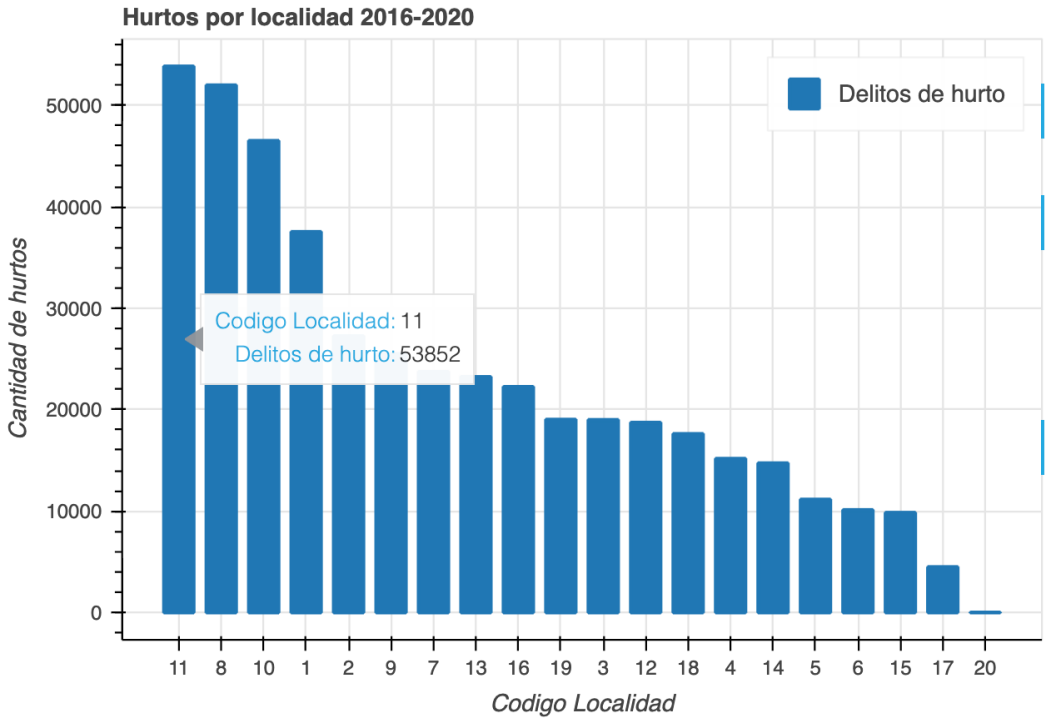


Figura 6: Hurtos por localidad

- Los delitos de hurto han venido en aumento en los últimos años.

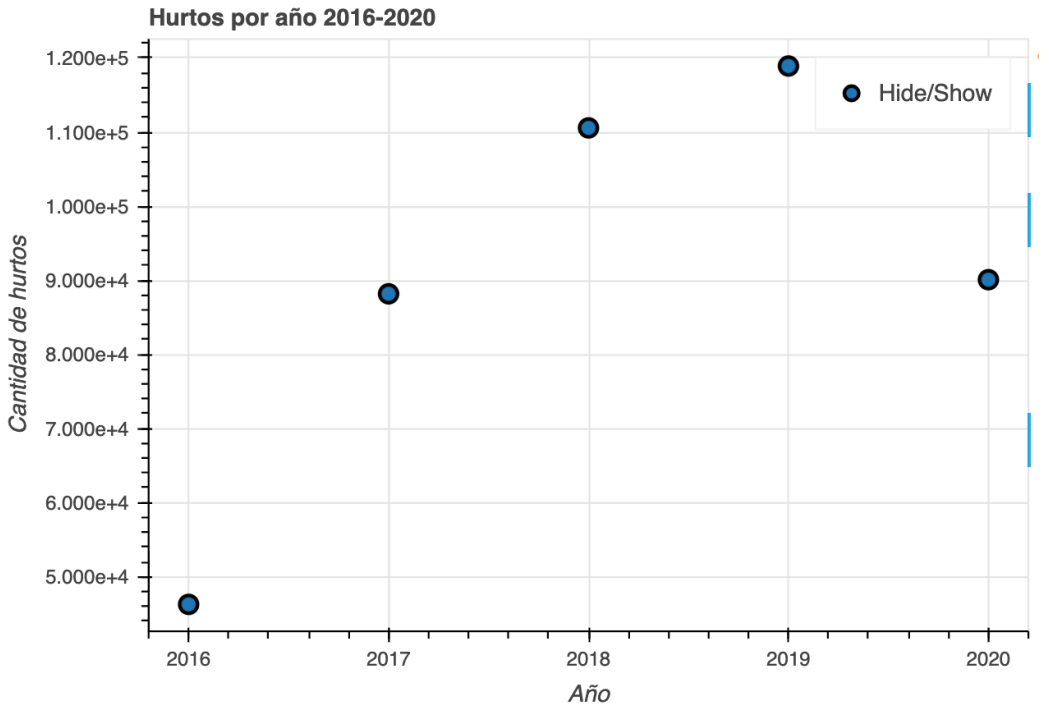


Figura 7: Hurtos por año

Para el dataset de SIEDCO de delitos de hurto en Suba en el año 2020, se realizó la exploración de datos (ver Figura 14) en la que se encontraron las siguientes premisas:

- Hay 9 tipos de delitos y el delito de hurto a personas es el más frecuente.

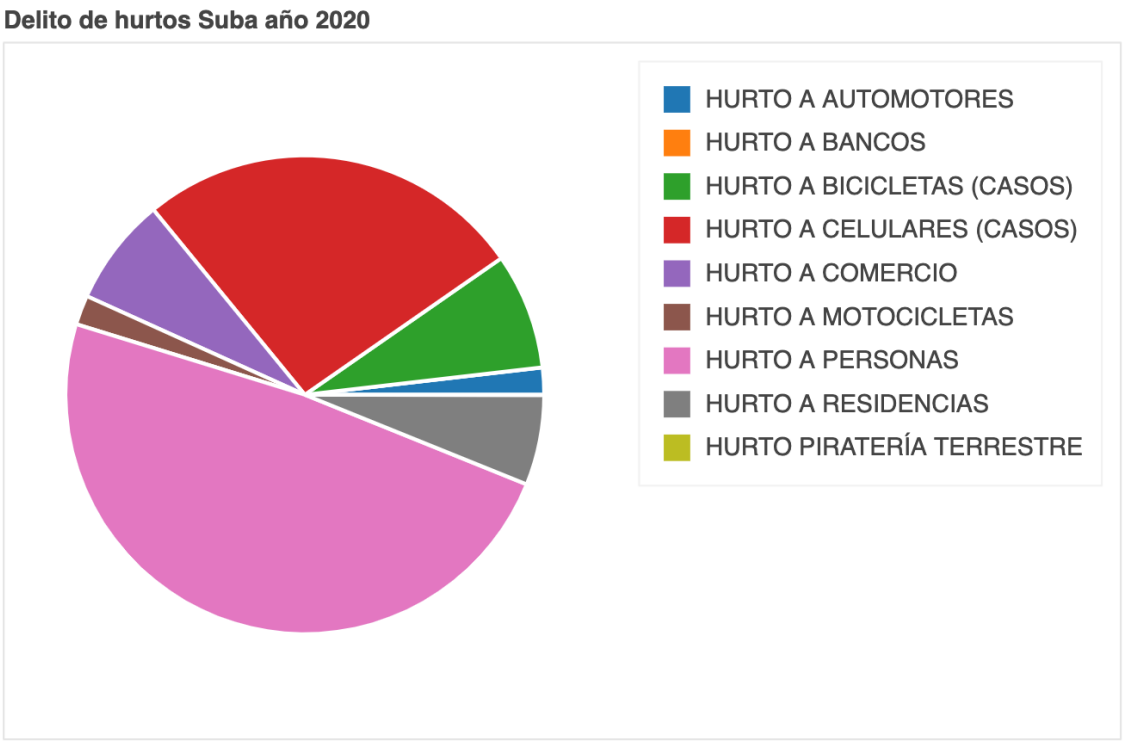


Figura 8: Tipos de hurto

- Las épocas en los que más delitos de hurto se presentan es a inicio de año, seguido de mitad de año y fin de año.

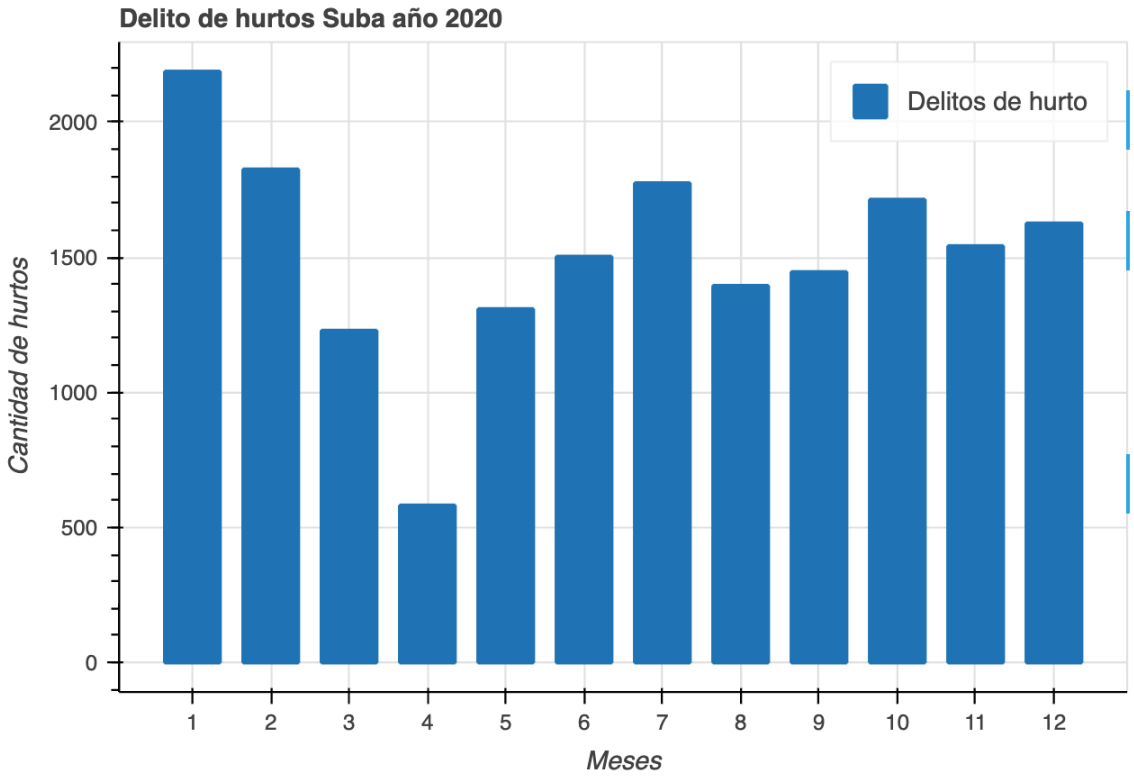


Figura 9: Hurtos por meses

- Los hombres son a los que más delitos de hurto les cometen, el doble que a las mujeres.

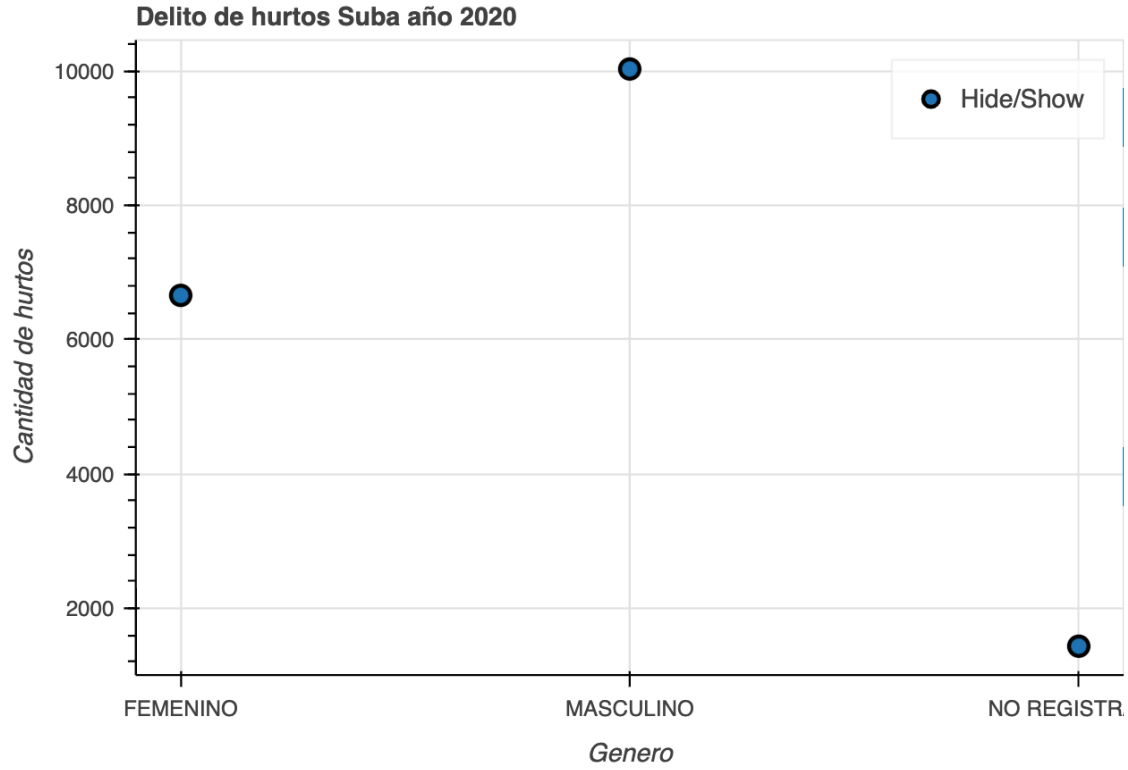


Figura 10: Hurtos por genero

- Las vías públicas y callejones es el lugar en donde más delitos de hurto se cometen.

- Las habitaciones, centros comerciales y estaciones de transporte público son los siguientes lugares después de las vías públicas y callejones en donde más delitos de hurto se cometen.

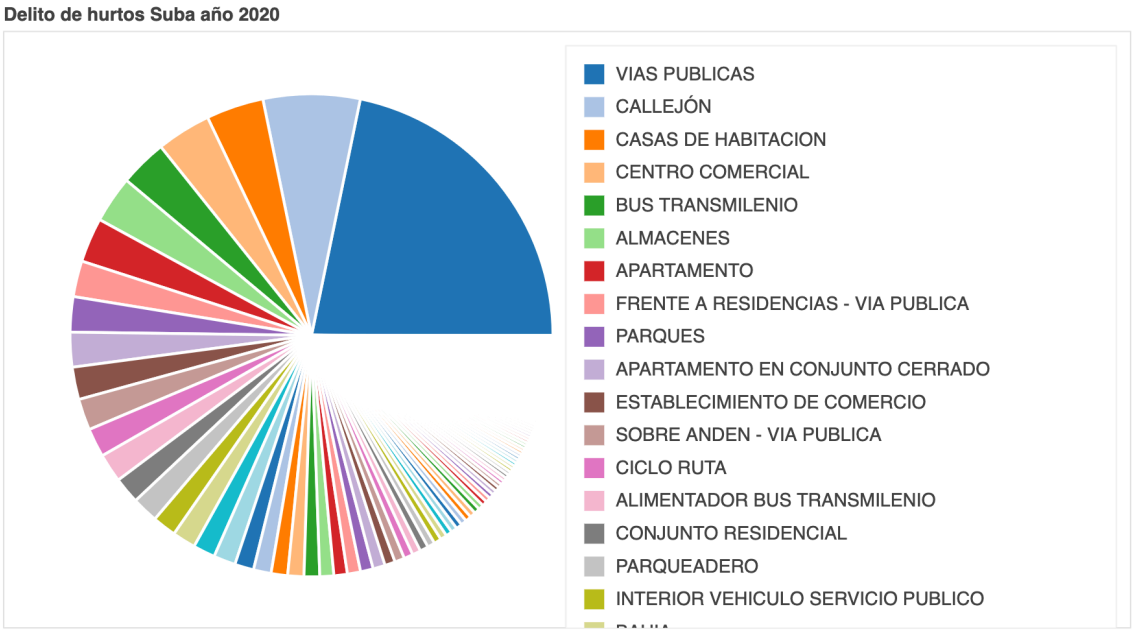


Figura 1: Sitios y lugares de hurto

2.4. Verificar la calidad de los datos

Para verificar la calidad de los datos se revisó cada una de las dimensiones sugeridas por DAMA (Data Management Association) enunciadas a continuación:

- Completitud: Los datos no están completos hay algunos registros que tienen NaN o valores por defecto como por ejemplo el género de la víctima.
- Unicidad: Hay un id único para cada registro de delito por lo tanto no se presentan problemas en esta dimensión.
- Exactitud: Los datos de los delitos son acordes a los periodos de tiempo mencionados.
- Consistencia: Existen algunos datos incorrectos que deben ser tratados en la preparación de los datos.

3. Preparación de los datos

3.1. Selección de los datos

El dataset utilizado para la construcción del modelo es el dataset de delitos de hurto para la localidad de suba en el año 2020 proveído por SIEDCO. Contiene 18123 registros de 9 tipos de delitos de hurto. Sin embargos los otros datasets (FIP, UAECD, SDM) mencionados anteriormente se utilizan para realizar un análisis, exploración y construcción de nuevas variables que permitan elaborar una base más sólida que será insumo del modelo de recomendaciones RMFAR y del modelo analítico adicional que será utilizado como punto de comparación.

3.2. Limpieza de los datos

Las siguientes variables fueron eliminadas del análisis, puesto que no tienen ningún aporta en el objetivo proyecto ni al modelo: municipio, zona y armas medios. Los delitos de HURTO PIRATERÍA TERRESTRE (4 registros) y HURTO A BANCOS (2 registros) fueron eliminados debido a la poca cantidad de registros que pueden ser detractores y afectar el modelo. Por otro

lado, para la variable GENERO, existen unos registros que tienen como valor NO REGISTRADO debido al objetivo principal del proyecto el cual es generar recomendaciones la opción más viable para llenar estos valores NaN o defecto fue utilizar el GENERO más frecuente el cual es el MASCULINO.

3.3. Construcción de los datos

Debido a la poca cantidad de variables disponibles en el dataset de delitos de hurto de SIEDCO, surgió la necesidad de crear nuevas variables a partir de variables existentes y de otras fuentes que permitan extraer nueva información. A partir de la fecha se crean 4 nuevas variables categóricas:

- HORA_INT: Intervalo o grupo de hora, los valores que puede tomar son: MAÑANA (Si la hora esta entre las 00:00 y las 12:00), MEDIO DIA (Si la hora esta entre las 12:00 y las 13:00), TARDE (Si la hora esta entre las 13:00 y las 18:00) y NOCHE (si a hora esta entre las 19:00 y las 23:59).
- DIA_SEMANA_INT: Intervalo o grupo del día de la semana, los valores que puede tomar son: FIN DE SEMANA (si el día es viernes, sábado o domingo) y ENTRE SEMANA (Si el día es lunes, martes, miércoles o jueves).
- DIA_MES_INT: Intervalo o grupo del día del mes, los valores que puede tomar son: QUINCENA (si el día esta entre 14 y el 18 o entre el 28 y el 2) y NO QUINCENA (el resto de los días).
- MES_INT: Intervalo o grupo del mes del año, los valores que puede tomar son: FIN DE AÑO (si el mes es noviembre o diciembre), INICIO DE AÑO (si el mes es enero o febrero), PRIMA DE MITAD DE AÑO (si el mes es junio o julio) y MES COMÚN (si el mes es cualquier otro mes que no entra en ningún grupo).

Para de crear más variables continuas que puedan aportar más información sobre un delito de hurto, se optó por crear variables espaciales o geográficas relacionadas al lugar en donde se cometió el hurto. Para esto se combinaron el dataset de SIEDCO con el dataset de UAECD, mediante la relación de barrio con sector catastral, un SCA puede estar conformado por 1 o más barrios. Posteriormente mediante el polígono de cada SCA se calculó el centroide, posteriormente, se calculó la distancia del centroide a cada punto que conforma el polígono para obtener la mediana de la distancia.

La mediana de la distancia es utilizada como parámetro de radio de búsqueda en el API de Places de Google Maps según la documentación [3], mediante el cual se pueden consultar todos los lugares de algún tipo en específico cerca de una coordenada (compuesta de latitud y longitud) en un radio especificado por parámetro a través de peticiones GET. Los criterios o parámetros de búsqueda fueron los siguientes:

- Location: Centroide del polígono del SCA conformado por latitud y longitud.
- Radius: Mediana de la distancia desde el centroide a los puntos del polígono del SCA.
- Type: Tipo de lugar se realizaron 5 consultas por cada tipo (bar, bank, store, atm y park).

Con las consultas realizadas se construyeron 5 nuevas variables:

- DENSIDAD_NOCTURNA: Número de bares cercanos.
- DENSIDAD_BANCARIA: Número de bancos cercanos.
- DENSIDAD_TIENDAS: Número de tiendas cercanas.
- DENSIDAD_CAJEROS: Número de cajeros cercanos.
- DENSIDAD_PARQUES: Número de parques cercanos.

Por otro lado, a partir del dataset de SDM que contiene los índices de seguridad nocturna para las UPZ, se calculó el área de la UPZ y el área de cada SCA mediante los respectivos polígonos, para luego calcular la proporción del SCA con relación a la UPZ que pertenece y así dividir cada índice de seguridad en los SCA que componen la UPZ.

Para el problema que ataca esta metodología analítica de clasificación, se optó por agrupar los delitos de: HURTO A BICICLETAS, CELULARES. COMERCIO, RESIDENCIAS. MOTOCICLETAS Y AUTOMOTORES, como HURTO A PATRIMONIO. Y así tener 2 categorías para delitos HURTO A PERSONAS y HURTO a PATRIMONIO. Esto debido a que como la cantidad de delitos de HURTO A PERSONAS es bastante grande en comparación a los demás al agruparlas como patrimonio la cantidad queda proporcional a un 55% para personas y 45% para patrimonio, lo cual permite entender de una mejor manera explicativa las decisiones y reglas del modelo. Además, esta desproporción a nivel de delitos hace que las reglas para este tipo de delitos no sean generadas o tengan medidas de interés muy bajas lo que hace que se descarten.

3.4. Integración de los datos

La integración de los datos se realizó combinando los datasets de SIEDCO y SDM. Para tener una base sólida y completa con variables que permitan obtener buenas reglas de comportamientos comunes en los delitos de hurto en la localidad 11 de Suba en Bogotá D.C. Teniendo así las variables que serán descritas en la sección de formato de los datos.

3.5. Formato de los datos

Teniendo en cuenta las etapas anteriores la variedad de datos ahora es mucho más amplia se cuentan con datos categóricos de tipo Object, datos continuos de tipo float64 y datos discretos de tipo int64 como se observa en la Tabla 8.

```
df_base_delitos[features].dtypes
```

```
HORA_INT              object
DIA_SEMANA_INT        object
DIA_MES_INT           object
MES                    int64
ACTIVIDAD             object
SCA_AREA             float64
DENSIDAD_NOCTURNA      int64
DENSIDAD_BANCARIA      int64
DENSIDAD_TIENDAS       int64
DENSIDAD_PARQUES       int64
ILUMINACION          float64
PERSONAS             float64
SEGURIDAD            float64
SENDERO              float64
TRANSPORTE           float64
GENERO                object
DELITOS               object
dtype: object
```

4. Modelado

4.1. Técnica de modelado

La fase de exploración analítica incluye la selección de un modelo aparte del modelo principal de este proyecto RMFAR, el modelo seleccionado es un Árbol de Decisión (DT) debido a que es un clasificador que ha mostrado buenos resultados y además utiliza reglas o decisiones para subdividir la mayor proporción de individuos en cada uno de los grupos.

4.2. Generación plan de prueba

La variable objetivo es la clasificación del DELITO, inicialmente 9 tipos de hurto y con la limpieza se tomaron en cuenta 2 tipos de hurto. Para la experimentación con el modelo se realizaron 3 casos, variando la proporción de los datasets:

- Entrenamiento 70% y pruebas 30%
- Entrenamiento 80% y pruebas 20%
- Ajuste de hiperparámetros (criterio, max_depth, max_features, min_samples_leaf)

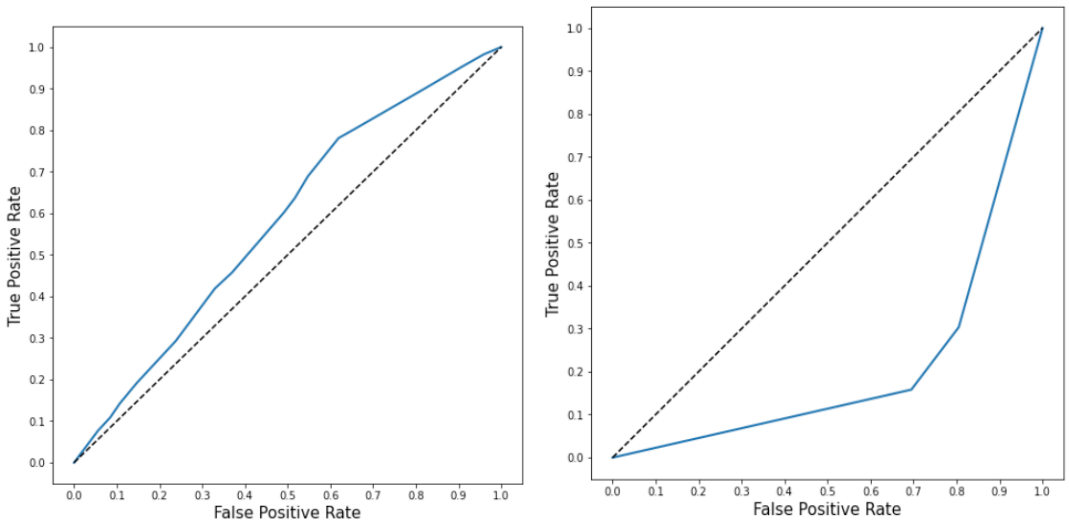
Como métricas se define la matriz de confusión para obtener la precisión y exactitud del modelo.

4.3. Construcción del modelo

La división del dataset se realiza con 70% para entrenamiento y 30% para pruebas y 80% para entrenamiento y 20% para pruebas.

5. Evaluación

Para la evaluación de un problema de clasificación se utilizan con mucha frecuencia las matrices de confusión, que permiten ver la precisión y la exhaustividad (precisión and recall). Se puede aumentar la precisión disminuyendo la exhaustividad y viceversa, cambiando simplemente el punto de corte (disminuyéndolo para aumentar precisión o aumentándolo para aumentar exhaustividad) Para observar ese balance se utiliza la curva COR (o ROC por sus siglas en inglés). Al graficar la curva de todas las posibles combinaciones de falsos positivos (que hablan de la precisión) contra los verdaderos positivos (que hablan de la exhaustividad) podemos darnos una idea de cómo funcionaría el modelo en diferentes puntos de corte. El área bajo la curva dibujada es un indicador adicional de la calidad del modelo. En la figura 15 se puede observar que el modelo de árbol de decisión (DT) con ajuste de hiperparámetros obtiene mejores resultados que el modelo de DT ingenuo.



6. Referencias

[1] "Reloj de la Criminalidad: horas y días en los que más delitos se cometen en Colombia." https://www.asuntoslegales.com.co/actualidad/reloj-de-la-criminalidad-horas-y-dias-en-los-que-mas-delitos-se-cometen-en-colombia-3038027 (accessed May 17, 2021).

[2] "Más de 7 mil hurtos se han registrado en Bogotá en el 2020 | RCN Radio." https://www.rcnradio.com/bogota/mas-de-7-mil-hurtos-se-han-registrado-en-bogota-en-el-2020 (accessed May 17, 2021).

[3] "Place Search | Places API | Google Developers." https://developers.google.com/maps/documentation/places/web-service/search#PlaceSearchRequests (accessed May 18, 2021).