PREDICCIÓN DE TIPOS DE DELITOS EN LA CIUDAD AUTÓNOMA DE BUENOS AIRES

ABSTRACT

Tomando como base la información de delitos de los distintos barrios de la ciudad de Buenos Aires y utilizando métodos de clustering – aprendizaje no supervisado - el presente trabajo estudia las features necesarias para la predicción de los tipos de delitos con el fin de determinar la gravedad de los mismos.

KEY WORDS

Delitos, gravedad, features, clustering, K-Means, PCA.

INTRODUCCIÓN Y OBJETIVOS

La ciudad de Buenos Aires cuenta con un Centro de Monitoreo Urbano desde donde es posible realizar la detección y almacenamiento de la información de los delitos. Este centro obtiene la información desde aproximadamente 9000 cámaras de calle, 8000 cámaras instaladas en colectivos, un Anillo Digital para detección de patentes, los patrullajes con 2200 vehículos y las llamadas recibidas al 911. Dicha información, de los años 2019 y 2020 es la base del presente trabajo de investigación mediante el cual plantearemos como principal objetivo encontrar una relación entre las distintas features de cada una de las muestras a fin de determinar la gravedad de los delitos. Es decir, poder a futuro predecir de qué tipo de delito se trata según las variables a estudiar.

DESCRIPCIÓN DEL DATASET

Para la investigación realizada se utilizaron 3 datasets diferentes que posteriormente fueron unidos para la obtención de los distintos DataFrames. El primero de los datasets llamado “delitos_2020” contiene la información que respecta a los delitos ocurridos en CABA durante el año 2020 y aporta en sus columnas datos como fecha de ocurrencia, franja horaria, tipo de delito, subtipo, uso de armas, barrio, comuna, latitud, longitud y cantidad de víctimas en caso de que el tipo fuera homicidio. Por otro lado, el dataset llamado “delitos_2019” al igual que el anterior contiene la información sobre los delitos ocurridos en el año 2019 en CABA, y aporta en sus columnas distintos datos.

Como parte de la limpieza de datos, debimos realizar la modificación de los tipos de datos que no correspondieran, la normalización de los mismos y de los nombres de las columnas, y la evaluación de la cantidad de nulos por feature o muestra a fin de determinar cuáles de los datos que contuvieran esos nulos servirían para la posterior experimentación. Como resultado de la unión de ambos obtuvimos el DataFrame “delitos_total”.

Por otro lado, como tercer dataset que nos serviría luego para la evaluación de ciertos parámetros respecto de la ocurrencia de delitos en los distintos barrios, utilizamos el dataset “pob_barrios” y creamos el DataFrame “barrios”. Dicho dataset contendría la cantidad de personas por barrio de la Ciudad Autónoma de Buenos Aires, que utilizaríamos para calcular la cantidad de delitos por cantidad de población de cada barrio.

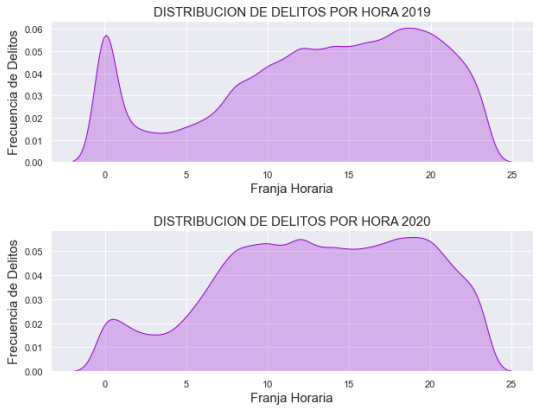
ANÁLISIS EXPLORATORIO DE DATOS

Luego de realizar el acondicionamiento de los distintos datasets para formar nuestros dos principales DataFrames – “delitos_total” y “barrios” - comenzamos con el Análisis Exploratorio de Datos, al que quisimos darle tres enfoques distintos con los que llegamos a conclusiones parciales diferentes.

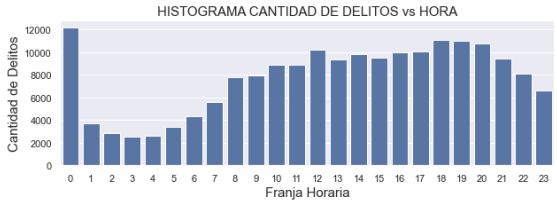
Delitos vs. Tiempo

Analizando los delitos y su distribución según la franja horaria de ocurrencia y en los distintos años 2019 y 2020, tal como se puede ver en los siguientes gráficos, la distribución es similar para los horarios diurnos, con la gran diferencia de que en el año 2020 se da una significativa disminución de

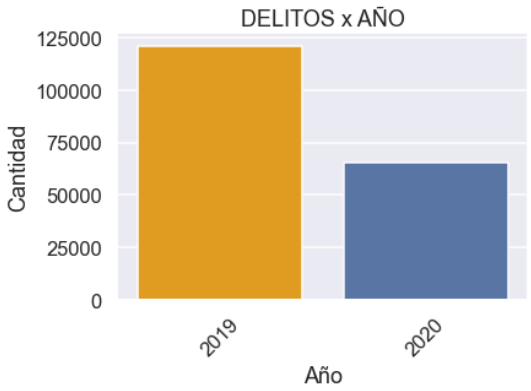
ocurrencia en el horario de medianoche. Consideramos que esto puede deberse a las restricciones aplicadas durante todo ese año por el Gobierno Nacional a raíz de la pandemia COVID-19, que impedían la circulación de la población en horarios nocturnos.



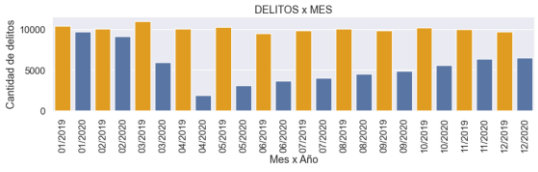
Si realizamos el análisis de la cantidad de delitos ocurridos por franja horaria sumando ambos años, tal como se puede ver en el siguiente gráfico de barras, en el horario de la media noche es cuando más ocurren los delitos, siendo los mismos alrededor de 12.000 para ambos años en conjunto. Durante las primeras horas del día hasta las 8 de la mañana, se da un incremento paulatino para luego generarse una meseta de la tasa de delitos durante el resto del día, con distintos picos por ejemplo a las 12 del mediodía y a las 18/19 horas. Consideramos que esto puede deberse a que ambos momentos del día mencionados, son aquellos en los que más circulación hay en la calle debido a los horarios de almuerzo y salida de las oficinas/vuelta a los hogares.



Siguiendo con lo mencionado anteriormente, en la situación coyuntural de la pandemia del COVID-19 cuyo impacto en Argentina para el año 2020 fue desde fines marzo a diciembre –continúa en 2021 pero escapa del análisis de la investigación- se pudo observar en el Análisis Exploratorio de los Datos la gran diferencia cuantitativa de ocurrencia de delitos en ese año respecto del anterior. Analizando mes a mes, la cantidad de delitos guarda estrecha relación con las distintas medidas y restricciones aplicadas por el Gobierno Nacional sobre la circulación de las personas en la vía pública. De forma anual, se puede ver claramente lo mencionado en el siguiente gráfico:



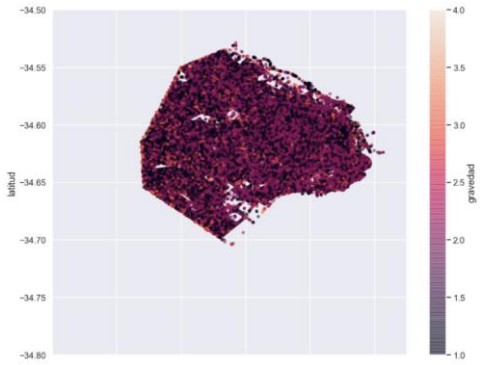
Comparando mensualmente, se puede ver la gran disminución de la ocurrencia de delitos en el mes de abril de 2020 respecto el mismo período del año anterior, dado que fue el momento de mayor restricción en la circulación. Con el pasar de los meses, las medidas fueron siendo menos estrictas y eso llevó a un incremento paulatino de la cantidad de delitos.



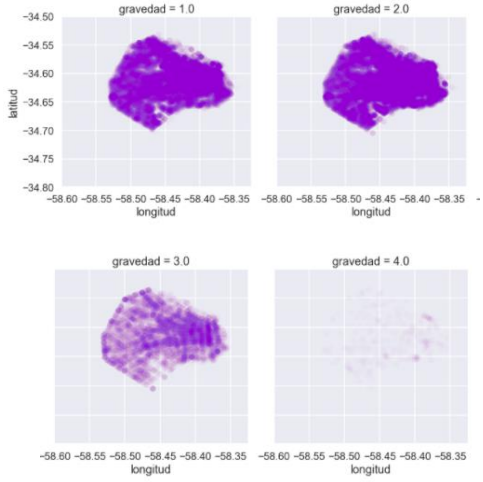
<u>Delitos vs. Tipo</u>

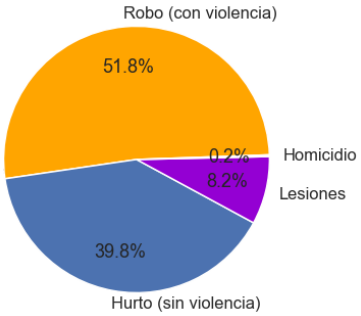
Para comenzar con la exposición del segundo enfoque del Análisis, debemos conocer cuáles son los cuatro tipos distintos según los que las comisarías de la Ciudad de Buenos Aires catalogan a los delitos: Hurto (sin violencia), Robo (con violencia), Lesiones y Homicidios. Para la realización del EDA, se asignó a cada tipo una gravedad distinta a fin de numerar lo mencionado anteriormente del 1 al 4 respectivamente.

A continuación, se pueden observar los distintos tipos de delitos que en conjunto grafican la Ciudad de Buenos Aires:



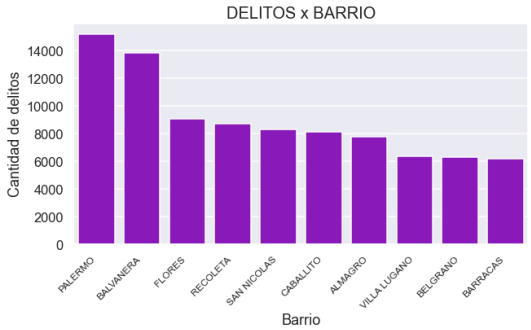
Para identificar con claridad las distintas densidades de cantidad de delitos por tipo, podemos observar los siguientes gráficos: el primero donde se ve claramente que los delitos de tipo 1 y 2 (Robo con y sin violencia) son los de mayor ocurrencia, y el segundo donde se dimensionan porcentualmente los datos vistos gráficamente.



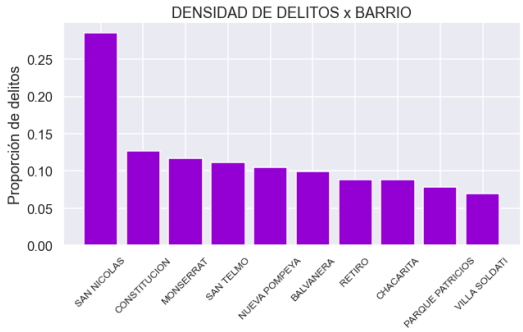


Delitos vs. Barrio

El tercer y último enfoque de análisis apunta a identificar aquellos barrios "más peligrosos", considerando que la peligrosidad se asocia a la cantidad de delitos, de la Ciudad Autónoma de Buenos Aires. Con ese objetivo, se pueden identificar dos perspectivas en las que los resultados van a ser diferentes: en primer lugar, si se analiza la cantidad de delitos por barrio sin tener en cuenta ninguna otra variable intrínseca a los mismos, el resultado de los 10 barrios con mayor ocurrencia será:



En cambio, si se considera la densidad poblacional de los distintos barrios y el número de ocurrencia de delitos, los "más peligrosos" serán los siguientes barrios, cuyos resultados aportan mayor veracidad respecto de la hipótesis inicial del enfoque:



MATERIALES Y MÉTODOS (ALGORITMOS UTILIZADOS)

Aplicamos aprendizaje no supervisado haciendo Clustering con K-Means. Para ello, disponemos de la matriz "delitos_total" caracterizada por 186.927 filas y 7 columnas a la que modificaremos para tomar las features que creamos necesarias. El aprendizaje no supervizado consiste en un aprendizaje sin variable dependiente "Y", es decir que se infieran propiedades y estructuras de la distribución de los datos.

En nuestro caso, utilizaremos 2 métodos y estrategias de Clustering: en primer lugar, aplicaremos K-Means y evaluaremos el Silhouette Index y Rand index en relación a la hipótesis

planteada. En segundo lugar, realizaremos reducción de la dimensionalidad con PCA y volveremos a aplicar K-Means para evaluar si el método genera mayor optimización de clusterizado.

K-Means es un método de clusterizado en donde cada cluster está caracterizado por un centroide y cada muestra se asigna al centroide que más cercano se encuentre. Cada centroide se va recalculando durante el proceso de aprendizaje y por lo tanto también la pertenencia de las muestras a los clusters. La distancia euclidiana cuadrática es la medida de similaridad entre muestras:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

Para medir qué tan buenos son los resultados del clustering, utilizaremos Silhouette Index, que mide cuán similar es una muestra respecto de su propio Cluster. Este indicador puede variar entre -1 y 1. Cuando es cercano a 1, se dice que los clusters han asignado a grupos de muestras bien separadas y definidas, cuando es 0 los clusters están superpuestos y es dificultoso encontrar grupos bien definidos, y cuando se acerca a -1, el algoritmo de clustering asignó erróneamente las etiquetas con respecto a la estructura/distribución de los datos. El Silhouette Index está dado por:

$$S_X = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{n_k} \sum_{x \in C_k} S(x_i) \right]$$

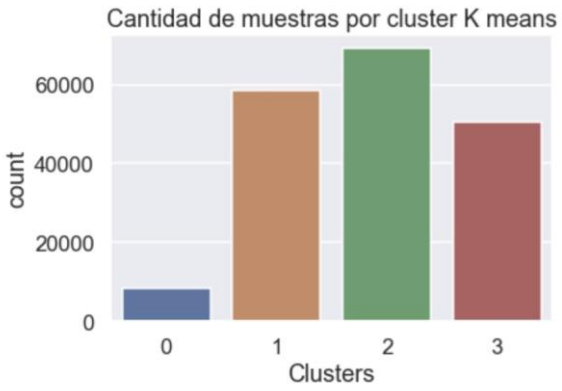
El Rand Index, que utilizaremos para evaluar qué tan cerca nos encontramos al clusterizar respecto de la hipótesis original, es una métrica utilizada cuando las etiquetas de las muestras están disponibles para validar luego de clusterizar la calidad de ello, pero no se usan durante el aprendizaje. En nuestro caso, verificaremos si mediante el método de K-Means aplicado, el algoritmo encuentra una relación entre las features aportadas para establecer la gravedad (tipo de delito) de cada muestra. Para ello, mediante el empleo de Rand Index evaluaremos el "accuracy" del clustering. Cuanto más cercano a 1 más "puro" serán los clusters.

Con la reducción de la dimensionalidad mediante el empleo de PCA – Principal Component Analysis - transformaremos la matriz empleada para el primer clusterizado con K-Means, en una de menor dimensión. Las nuevas dimensiones no representan nada del entorno real de los datos, sino que son combinaciones lineales de las features originales creadas con el fin de representar la mayor cantidad de variación de los datos.

EXPERIMENTOS Y RESULTADOS

Para la experimentación con K-Means en la primera instancia utilizamos los datos de barrio, franja horaria, mes, día, población, delitos por barrio y delitos por densidad poblacional. No utilizamos latitud y longitud como features dado que existen infinitos puntos en el mapa que asumimos complicaría la clusterización. Podría considerarse que los datos asociados al barrio como delitos por barrio, población y delitos por densidad poblacional son redundantes, pero dado que realizamos un label encoder para los barrios y buscamos evitar que el algoritmo tome el número de barrio jerarquizado por valor, pensamos que dichos datos aportarían peso a la hora de asociar los clusters. Para aplicar el método de K-Means definimos un n clusters igual a 4 para evaluar si existe la relación entre features tal que se pueda predecir el tipo de delito (gravedad) asociada a cada caso.

Como resultados en esta primera instancia, obtuvimos la cantidad de muestras por cluster:



Al evaluar el Rand Index, obtuvimos un valor igual a 0,93 y un Silhouette igual a 0,46.

Luego, aplicamos reducción de la dimensionalidad con PCA y el valor del Rand Index disminuyó considerablemente a 0,49 pero el Sillhouette se mantuvo igual.

DISCUSIÓN Y CONCLUSIONES

A partir de los resultados obtenidos con la implementación de los 2 métodos mencionados, podemos concluir que el Rand Index en la primera instancia es muy cercano a 1 por lo que se podría asumir que existe una fuerte relación entre la gravedad o tipo de delito de cada muestra y la forma en la que los clusters asocian cada una. Sin embargo, al ver que luego aplicando PCA el valor disminuye a menos de la mitad (0,49), consideramos que el índice no es representativo para asegurar que el cluster esté asociando por gravedad.

El Silhouette, por su parte, tiene un valor intermedio entre la unidad y 0, por lo que los clusters no están tan separados y definidos. Esto se mantiene al aplicar reducción de la dimensionalidad por lo que no implica una mejora.

REFERENCIAS

*Ghahramani, Z. (2003). Unsupervised learning. In *Summer School on Machine Learning* .Springer, Berlin, Heidelberg.

*Libro: Elements of Statistical Learning, Tibshirani et Al.

*Paper “Model-based evaluation of clustering validation measures”. *Pattern recognition*, *40*(3), 807-824
(2007).