Trabajo de grado sobre predicción de los tipos de delitos en Medellín

Jennifer Mosquera Cabra

Trabajo de grado presentado para optar al título de Economista

Asesor

Jaime Alberto Montoya Arbeláez

Universidad de Antioquia

Facultad de Ciencias Económicas

Economía

Medellín, Antioquia, Colombia

Resumen:

El crimen es un fenómeno que impacta de manera directa el bienestar de una sociedad por lo que comprender la racionalidad básica del accionar criminal, sus causas y razones da lugar a la posible disminución de éste y "aún más" si se trata de un análisis predictivo capaz de efectuar alertas tempranas para mitigar el nivel de criminalidad en un momento especifico. El objetivo de este informe es recopilar información sobre el estudio del crimen y las principales técnicas de aprendizaje automático utilizadas para la predicción del crimen en una ciudad. Inclusive, sobre el pronóstico del número de homicidios en función de covariables con el propósito de entender la estructura de cada método y la configuración de las variables que se utilizan. Todo lo anterior con la finalidad de construir un modelo que prediga algunos tipos de crimen o solo el número de homicidios en la ciudad de Medellín.

Palabras claves: Crimen, Meachine Learning, homicidios, patrones delictivos, utilidad, costos, penalidad.

1. Introducción

Según la convención de Palermo del 2000, el accionar contra el crimen debe tener en cuenta su definición. La delincuencia se ha denominado grupo delictivo organizado si cumple las siguientes condiciones: pluralidad determinada, temporalidad, voluntariedad y finalidad delictiva común. Este concepto es extenso y puede contener problemas de interpretación provenientes de los diferentes sistemas políticos que lo regulan, esto es porque cada sociedad sufre distintos flagelos. En Colombia se ha esbozado dos fuentes principales del crimen: el tráfico de drogas ilícitas y el terrorismo. Sin embargo, la ciencia de la criminalidad ha evidenciado varias conductas criminales derivadas de las dos anteriores, tales como extorsiones, trata de personas, hurtos o secuestros (D'angelo, 2019).

En Colombia, la Corte Constitucional estableció los elementos que definen la delincuencia organizada y cómo estos tienden a ser un tipo de concierto para delinquir. Para la Corte Suprema, este último presupone la existencia de una organización, conformada por un grupo de personas que se han puesto de acuerdo o han convenido llevar a cabo delitos que lesionan o ponen en peligro bienes jurídicos (D'angelo, 2019). Incluso, Zaffaroni (1991) precisa el delito como una acción humana de naturaleza contraria al orden jurídico, sancionable bajo la ley de las penas, que son claramente identificadas en los límites del comportamiento correcto de la sociedad, y reconocibles.

Según Winter, 2005, existen tres elementos principales que distinguen el razonamiento económico para tratar el crimen en comparación con otros razonamientos. El primero es que, dado que se requieren recursos costosos para desalentar el crimen, es muy probable que la cantidad óptima de crimen desde una perspectiva social sea positiva. En consecuencia, las cuestiones económicas claves para los costos de reducción de la criminalidad se enfocan en la cuantía de recursos que deben destinarse a la lucha contra esta. El segundo, es el supuesto que los criminales son racionales en términos de la conciencia de los costos o beneficios de sus actos. Y el tercero es que el reconocimiento de la probabilidad de ser aprehendido puede depender de la naturaleza del delito que también está unido a dos componentes básicos: la severidad y la certeza del castigo.

En la medida en la cual se comprenden dichos argumentos, las autoridades pueden afectar la tasa de delincuencia mediante el castigo, si conocen la racionalidad, los beneficios y los costos que enfrentan las personas que realizan estos actos (Winter, 2005). Existen varios estudios acerca de los beneficios y los costos del crimen, los cuales suponen que las personas dedican tiempo a la actividad criminal hasta que

la ganancia marginal sea igual a los costos marginales. Los tipos de utilidades obtenidas de un acto ilegal varían, dependiendo del tipo de delito y la persona que los comete: algunos son monetarios, por hurto, robo, fraude. Otros son psíquicos, es decir, que son por la sensación del peligro, la aceptación de los compañeros, la remuneración, el sentido de logro, o complacencias de deseos (Eide, 2006)

Para los costos, se pueden especificar de material, psíquicos, de penas y de oportunidad. Los costos de penas son todas las penalidades formales que son multas, encarcelamiento, entre otras. El costo de oportunidad del crimen es el beneficio neto de la actividad legal al que se renuncia por planear, hacer y encubrir el delito. Esto deja unas suposiciones, por ejemplo, que entre los delincuentes haya más jóvenes, minorías y trabajadores mal pagados. También es necesario mencionar la inmediatez de las ganancias en comparación al castigo del crimen, considerando que la pena será diferente de acuerdo con el delito (Eide, 2006).

La elección racional en la que se basa el siguiente modelo fue propuesta por Beccaria (1995) y Bentham (1843), los individuos que realizan actos criminales actúan como un agente que maximiza de manera racional su utilidad. Becker (1968), emplea el argumento de la utilidad esperada, asumiendo una función positiva de ingresos. La utilidad esperada del individuo $E[U]$ de cometer un crimen es:

.

$$E[U] = P * U(Y - f) + (1 - P) * U(Y),$$

donde $U(\cdot)$ es la función de utilidad von Neumann-Morgenstern del individuo, P es la probabilidad subjetiva de ser atrapado, Y es el ingreso monetario del delito y f es el valor monetario equivalente del castigo. El individuo cometerá el delito si la utilidad esperada es positiva y no lo hará si es negativa.

El análisis muestra que los aumentos en la probabilidad o rigor del castigo podrían cambiar la utilidad esperada de positiva a negativa. Becker también introduce una función de oferta de delito, donde los dos factores tienen un efecto sobre la cantidad total de delito. Por lo que la utilidad esperada se transforma de la siguiente manera.

.

$$E[U] = PU(W - f) + (1 - P)U(W + g)$$

donde W es el ingreso presente y g es la ganancia del crimen, el delito se cometerá si la utilidad esperada es mayor que la utilidad del ingreso inicial W. Entonces, la motivación del crimen está ligada al riesgo, a la utilidad y a las condiciones actuales de la persona que lleva a cabo los hechos delictivos. Por lo tanto, existe una relación entre el ingreso presente y la ganancia del crimen, la cual está afectada por el riesgo que se percibe de ser capturado Erling Eide, (2006). También se expone la idea que para los modelos que se han propuesto se obtienen el efecto sustitución, debido a que, un castigo más severo será menos frecuente, es decir, residirá en menos delitos y el efecto renta ya que el crimen dependerá de la actitud al riesgo Becker (1976).

En cuanto a la racionalidad del crimen en términos del beneficio, Blasco (2016) ha supuesto que el individuo escoge el proceso de su accionar que mejor se acomode a las preferencias, las cuales comprenden sus deseos, normas e incluso su entorno. Incluso, se puede resaltar el modelo tradicional de la Economía del delito, suponiendo la neutralidad al riesgo de los agentes que están motivados a cometer el delito y suponiendo racionalidad de estos. Se establece una variable dicotómica d donde toma el valor de 1 si se comete el delito y 0 en otro caso. Si el agente realiza el crimen, la variable d varía de acuerdo con el beneficio del delito. Y el beneficio neto estará dada por: $BN = X - W - C - P * F \geq 0$ si $d = 1$ y el $BN < 0$ si $d = 0$. Donde BN son los beneficios netos, X es el costo de oportunidad de cometer el delito, C es el costo material de llevar a cabo el crimen, P es la probabilidad de ser atrapado y F es la pena por la que puede ser acusado (Blasco, 2016). La situación de empleo también es un determinante para que

el crimen sea más atractivo, ya que la alternativa de no cometer el crimen es seguir en la pobreza más tiempo (Erling Eide, 2006). Inclusive en la encuesta de Chiricos (1987) se demuestra que, en la mayoría de los análisis, el desempleo parece aumentar la criminalidad.

El objetivo de este informe es recopilar información sobre el análisis de las dinámicas del crimen en un tiempo determinado, y de las técnicas que han sido utilizadas para la predicción e interpretación de su comportamiento. Con el propósito de presentar una propuesta de modelación predictiva del crimen o del número de homicidios en función de carácteristicas propias de los individuos en la ciudad de Medellín.

2. Revisión de literatura

2.1. Vizualización

La identificación de relaciones, patrones o tendencias en los datos es necesario para analizar conexiones, causalidades o cambios de un fenómeno en estudio. Dichos factores están determinados por inteligencia criminal, por la seguridad en la localidad o por el tiempo en que sucede estos actos. La técnica más común para la visualización es el *análisis espacial*, puede ser un análisis de frecuencia del crimen en una localidad, el tipo de crimen en diferentes áreas y visualización de puntos "calientes" según el tipo de crimen (Hitesh Kumar Reddy ToppiReddya, 2018).

También se han utilizado gráficos descriptivos para identificar dichas relaciones, tales como: tipos de crímenes durante un tiempo determinado, crímenes cometidos en diferentes localidades, detalles del tipo de crimen que tuvo mayor incidencia en la ciudad y número de crímenes por hora. Se puede evidenciar una visualización interesante en la investigación de Andersen y Malleson (2014)[1] sobre el desplazamiento del crimen enfocada a la identificación de cambios en los patrones o distribución espacial del delito; la metodología utilizada fue Spatial point analysis y otra aplicación en la cual identificaron patrones del crimen hecha por Phillips y Lee (2012)[2], utilizando Graph similarity [3]

2.2. Tratamiento de datos y modelación

En el campo del Meachine Learning, la predicción de este tipo de problemas se ha tratado en su mayoría como un problema supervisado específicamente de clasificación. Inclusive, algunos métodos de aprendizaje son utilizados para predecir la probabilidad de ocurrencia dada la estimación de una función de densidad. Otros también son empleados para reducir el tamaño del conjunto de datos, para caracterizar la muestra o para capturar la información más relevante. También se aplican modelos de regresión de conteo para predecir el número de homicidios en función de características relevantes. Algunos de los algoritmos más utilizados en la predicción del crimen en una ciudad, son el K-Nearest Neighbours, Redes Neuronales, Boosting, Arboles de Decisión, Bosques Aleatorios o Modelos Generalizados (Barnadas, 2016).

El algoritmo K-vécinos más cercanos

El método de clustering facilita la clasificación no paramétrica que particiona un conjunto de las observaciones o unidades en k grupos y su fin es catalogar las observaciones con base a una similitud entre ellas por medio de una función de distancia apropiada. Dentro de la estructura del cluster se puede utilizar el procedimiento de K-vecinos más cercanos para estimar las semejanzas. La idea general de este mecanismo es calcular el promedio de las distancias de cada punto a sus k vecinos más cercanos. Un aspecto importante para mencionar es que este método no hace ninguna suposición sobre las distribuciones de

[1] Véase bibliografía : 52
[2] Véase bibliografía : 53
[3] Véase: Alkesh Bharati1, 2018.

clases. La formulación matemática general es la siguiente:

$$\arg\max_{L} \sum_{i=0}^{N-1} p_i p_i = \sum_{j \in C_i} p_{ij}$$

$$\mathrm{p}_{ij} = \frac{\exp(-||Lx_i - Lx_j||^2)}{\sum\limits_{k \neq i} \exp -(||Lx_i - Lx_k||^2)}, \quad p_{ii} = 0$$

En este caso, se supone que el algoritmo está aprendiendo una métrica de distancia de Mahalanobis al cuadrado, $||L(x_i - x_j)||^2 = (x_i - x_j)^T M(x_i - x_j)$. Donde $M = L^T L$ es una matriz de tamaño semidefinida positiva simétrica. [4]. La finalidad de $K-Nearest$ es construir una matriz de transformación lineal óptima capaz de maximizar la suma de la probabilidad de correcta clasificación en todas las muestras, donde N es el número de muestras y pi es la probabilidad de que la muestra i esté correctamente clasificada. Ci es el conjunto de puntos en la misma clase y pij es la función exponencial generalizada sobre las distancias. En general, el fundamento detrás del método es encontrar un número de muestras de entrenamiento más cercanas en distancia al nuevo punto y predecir la categoría a partir de ellas. [5]

Una de las aplicaciones que ha usado el procedimiento de cluster es de (Kim, 2018) el cual utiliza el *K-nearest neighbour y boosted decision tree* con el fin de crear un modelo que pueda pronosticar con precisión el crimen en Vancouver. Otra aplicación fue una descripción general hecha por Bachner (2013) para la predicción del crimen con *Clustering y Social Network Analysis.* El estudio de patrones de crimen hecha por Nath (2006), usando *k-means clustering.* La agrupación de datos sobre delitos por tipo de delito y predicción de su frecuencia hecha por Malathi y Baboo utilizando *DBSCAN clustering (density based), k-means y árboles de decisión.* Y en Hossain (2020) que propusó *Decision Tree y K-nearest neighbor* para predecir el crimen, de forma tal que pudieran encontrar zonas calientes en un momento específico del día. Estos últimos, utilizaron el *bosques aleatorios* y *Adaboost* para mejorar la capacidad predictiva.

El algoritmo *K-Nearest Neighbours* es ventajoso en los casos en los cuales la estimación de la función de regresión no es lineal en x. Un modelo polinomial puede ser lineal en sus coeficientes, aunque sea una función no lineal de x. *K-Nearest* es no lineal y no paramétrico porque no acepta una estructura conocida de la función que modela el comportamiento de los datos (f) en cada iteración, donde $f \in F$. [6]

En general, este método es sencillo de implementar e interpretar. Sin embargo, tiene la desventaja de que al realizar la estimación se utiliza todo el conjunto de datos para entrenar cada punto y esto requiere mucho procesamiento y memoria, así mismo, el ajuste se torna más lenta en la medida en la cual el volumen de los datos aumenta. También es muy sensible a los atributos irrelevantes, su predicción depende mucho de que la elección de la métrica de distancia sea la correcta. Otra desventaja muy común en estos algoritmos es la maldición de la dimensionalidad, debido a que cuando aumenta la información, el volumen del espacio aumenta exponencialmente haciendo que los datos disponibles se dispersen más Kassambara (2017).

Los Modelos de redes neuronales

Es un modelo supervisado que puede aprender una función no lineal para clasificación y regresión. Es una metodología basada en la idea de interconexiones y transformación de información incluso con dirección. Por lo tanto, el fin de este método es la acumulación y la transición de datos.

En función de lo planteado, se obtiene un conjunto de inputs o dendritas y salidas que se causan por medio del axon, por lo cual, el output es $y = \in \{-1, +1\}$, en cada input se asigna un peso o ponderador w_j donde j son las características o las variables independientes; estos pesos en términos de una red, serán las conexiones, es decir, van a representar la importancia del input en la conexión con otras neuronas. El procedimiento de la acumulación de información es una suma ponderada de los pesos con las covariables: se suman los ponderadores hasta la p característica respectivamente.

[4] Véase: Nearest Neighbors, taken from: https://scikit-learn.org/stable/modules/neighbors.html
[5] Véase: Alboukadel Kassambara (2017).Practical Guide To Cluster Analysis in R
[6] Véase: J. Friedman, 2009

$$Z = X^t w = \sum_{j=1}^{p} w_j X_j$$

$$Y = \text{sign}(w_0 + X^t w) = \begin{cases} +1, w_0 + X^t w \geq 0 \\ -1, w_0 + X^t w < 0 \end{cases}$$

La función sign en la literatura es la función de activación que responde el tipo de clase a la que pertenece. Es decir, la función de activación en este caso es una combinación lineal que activa la información a una clase determinada. Resumiendo lo planteado, la idea de conexión de capas de entrada y salida continua cuando se construye un modelo de múltiples capas y también tenemos capas ocultas. Las capas se identifican por L, lo que significa que es un hiperparámetro. En una metodología de multiples capas,$L = 1$ representa la capa "input", $L = L$ es la respectiva capa de salida, por lo cual hay $L - 2$ capas ocultas. Es necesario señalar que cuando se construye una red neuronal tipo regresión y tipo clasificación de dos clases, solo se tendrá una capa, pero cuando se tiene múltiples clases, se debe tener en cuenta el mismo número de capas.[7]

Se denomina perceptrón multicapa a un mecanismo basado en redes neuronales en el cual, la capa de entrada consta de un conjunto de neuronas $X_i : (x1, x2, .., xm)$ que representa las características de entrada, este número de neuronas se representa con p_l. Y se tiene $L - 2$ capas ocultas. En síntesis, la estructura de una capa es; $W_j^l \hookleftarrow k$ la neurona k que está en la capa l y logra conectarse con la neurona j en la capa $l+1$. El algoritmo general de una red neuronal es: para $i = 1, ..., n, j = 1, ..., p_l$ se construye la función de activación.

$$Z_{i_j}^{l} + 1 = \sum_{k=1}^{p} {}_l((W_{jk})^l (O_{ij})^l + w_{0,j}^{l}), l = 1, 2, ..., L - 1,$$

$$O_{i_j}^{l} + 1 = f^{l+1}(z_{i_j}^{l+1})$$

Donde $O_{i_j}^1 = x_{ij}$ y $O_{i_j}^L = f^L(z_{i_j}^L)$, Z es la función que transforma la información y depende de w, pero Z se transforma por medio de la función f para generar el output de la siguiente capa, por lo tanto el O^{l+1} que es el output en la siguiente capa es función de w, es decir, si conozco w, entonces conozco Z y O. Por último, se necesita una función de perdida que evalúa el desajuste de lo observado y lo pronosticado. De manera iterativa se ajusta con una nueva ponderación los pesos de las características, se ajusta una "pérdida", se compara dicha pérdida con la anterior hasta que se llegue a un modelo con una buena capacidad predictiva, es decir, una función en la cual la pérdida no cambie, así: $w^l \to z^{l+1} \to O^{l+1} \to L(w)$. Cabe mencionar que en este método se asume una función de activación homogénea.[8]

Una aplicación en la cual utilizaron la idea de las redes neuronales fue en el descubrimiento de patrones de series criminales en un caso de estudio de robos hecha por Dahbur y Muscarello (2003) en la cual utilizaron Kohonen neural networks y heuristics. Y en un análisis y predicción de robo hecha por Oatley y Ewart (2003), utilizando Logistic regression, neural networks y Bayesian Network.

El algoritmo Artificial Neural Networks es ventajoso siempre que el objetivo del problema sea de clasificación. Este modelo tiene una alta capacidad de aprender de modelos no lineales y en tiempo real. Sin embargo, también existe unas desventajas, es sensible a la escala de las características, requiere un ajuste de una serie de hiperparámetros, como la cantidad de neuronas o las capas e iteraciones ocultas. Y hay una gran variabilidad en su validación, debido a que ante iniciaciones diferentes pueden llevar a la función de pérdida a diferentes mínimos [9]

Ramdom Forest

Ramdom Forest, se caracteriza por ser bastante aleatorio en las submuestras obtenidas y en las observaciones utilizadas. Debido a que el algoritmo selecciona una muestra aleatoria de las observaciones i de tamaño n con reemplazamiento. Para cada muestra se hace un entrenamiento de un árbol, de manera

[7] Véase: Rosenblatt, F. (1958).
[8] Véase: Rosenblatt, F. (1958)
[9] Véase: Bibliografía 61

repetitiva hasta la última muestra. Es decir para cada árbol, se implementa una inyección de aleatoriedad, lo que quiere decir es que, en cada nodo, se debe de seleccionar un subconjunto menor al número de características completa. En la construcción de un árbol, se tiene el subconjunto de variables seleccionadas, sobre estas variables se busca cuál es la variable x_j en la cual el decrecimiento de la impureza es mayor en ese nodo. Para en cada nodo se repite este proceso, es decir, se busca la variable de otra submuestra menor a la original con mayor decrecimiento de manera repetitiva hasta cumplir el criterio de número mínimo de observaciones en los nodos y este proceso se repite para cada árbol.

Especificamente lo anterior es; predice Y aprendiendo de reglas de decisión con base a las características $X = (X1, ..., Xp) \in X$, la idea es particionar las características de forma binaria tal que: $X = \bigcup_{t=1}^{T} R_t$ donde $R_t \cap R'_t = \emptyset, \forall t \neq t'$, cada región tiene dos segmentaciones, dos clases en un problema de clasificación y dos constantes asignadas si el problema es de regresión.

La idea global de árboles de decisión es la siguiente: sea $D = (x_i, y_i), x_i \in X \subset \mathbb{R}^p, y_i \in Y$, donde X se particionará en T regiones. La forma funcional de un árbol de decisión es $f(x_i) := f(x_i, (w_t, R_t)) = \sum_{t=1}^{T} w_t I(x_i \in R_t)$, donde w_t es una constante que representa la ponderación en cada nodo y $I(.)$es la condición en la cual se basa la segmentación de las características. Si x_i no pertenece a la región Rt, entonces tomará el valor de 0, y 1 en otro caso. [10]

$$\mathrm{I}(\mathrm{x}_i \in R_t) = \begin{cases} 0 & x_i \notin R_t \\ 1 & x_i \in R_t \end{cases}$$

Si la orientación es para un problema de clasificación del tipo de crimen, entonces la función objetivo es minimizar la función de pérdida representada por w_i, maximizando el decrecimiento de una impureza en cada partición. Para problemas de clasificación esta impureza puede ser Hinge, Squared hinge, Squared hinge y Binary cross entropy. Para problemas de regresión esta función puede ser Mean Squared Error, Poisson, Logcosh, entre otras.

$$\mathrm{w}_i = argminE_x, y|_t I(Y, k)$$

Donde w_i es la ponderación en cada nodo, teniendo en cuenta las observaciones en el nodo k. Si la orientación es de regresión, se recomienda que la regresión sea con una respuesta Poisson, entonces la especificación es minimizar el error para la configuración Poisson, el error viene dado por $L(D, \lambda)$.

$$L(D,\lambda) = \frac{1}{n} \sum_{i=1}^{n} 2N_i[\frac{\lambda(x_i)\nu_i}{N_i} - 1 - log(\frac{\lambda(x_i)\nu_i}{N_i})] \geq 0$$

La función de regresión óptima, estima $x \to \lambda\hat{(}x)$, minimizando la función de pérdida $L(D, \lambda)$, especificamente la función de desviación del promedio Poisson en λ, lo que es equivalente a determinar el estimador maximum likelihood de λ. Donde D se refiere a los datos de entrenamiento, N_i son los conteos independientes de un evento, $\lambda(.)$ es la función de regresión a estimar y ν_i es la frecuencia esperada.[11]

En la aplicación de Luiz G.A. Alves, 2017, se estimó una regresión Random Forest para predecir el crimen y cuantificar la influencia de indicadores urbanos sobre los homicidios. Y en el cual concluyen que el crimen depende mucho de los indicadores urbanos, también muestran que este modelo es una buena solución para predecir delitos e identificar la importancia de los indicadores que los afectan, incluso bajo pequeñas perturbaciones en los datos. En la aplicación de Ginger Saltos, 2017 tuvo como objetivo predecir la frecuencia de algunos tipos de delitos con aprendizaje basado en *instancias, regresión y árboles de decisión.* En las aplicaciones en las cuales el objetivo es predecir el tipo de crimen en una ciudad. Los procedimientos más utilizados son: *Decision Trees, Ramdom Forest, Naive Bayes, regresión lineal,*

[10] Véase: Louppe (2014)

[11] Véase: Alexander Noll, (2020).Case Study: French Motor Third-Party Liability Claims

regresión logística, Support Vector Machine (SVM), Bayesian methods y Multinomial model.

El algoritmo de Decision Trees y Random Forest son los algoritmos que se ha utilizado más en la aplicación de este tipo de problemas. Por su parte, los Decision Trees no necesitan supuestos formales de distribución, es simple de entender y de interpretar, puede mostrarse gráficamente, requiere poca preparación de datos, puede manejar grandes conjuntos de información, puede manejar características cuantitativas y cualitativas, ignora fácilmente las variables redundantes.

Sin embargo, tiene dos desventajas: los árboles pueden cambiar mucho si cambiamos un poco los datos, es decir, es muy inestable y generalmente no tienen un buen desempeño predictivo. Pero estas últimas dos desventajas dan pie al algoritmo de ensamblaje de Ramdom Forest, debido a que en conjunto, muchos Decision Trees mejoran la precisión de un modelo de aprendizaje estadístico dado y disminuye su varianza, incluso el Ramdom Forest, mejora esta varianza a través de la reducción de correlación entre los árboles que es causa de la inyección de aleatoriedad que obtiene del muestreo Boostraping. Inclusive se ha utilizado el *Random Forest* para cuantificar el papel que juegan los indicadores urbanos en la predicción del crimen en Brasil. [12]

Modelo Aditivo Generalizado

El modelo aditivo generalizado permite una mayor flexibilidad entre las relaciones de las variables de control y la variable de respuesta. Tiene la capacidad de ser una metodología estadística que permite inferencia y predicción.

$$Y_t = \alpha + \sum_{j=1}^{p} f_j(X_{jt}) + \xi_t$$

donde ξ es un término de perturbación aleatorio independiente de cada X_j, con $\mathbb{E}(\xi) = 0, Var(\xi) = \sigma^2$. Las f_j's representan funciones desconocidas univariantes las cuales pueden reflejar de los predictores X_j. Note que si $f_j(X_{jt}) = \beta_j X_{jt}$ el modelo se reduce al caso del modelo de regresión lineal.[13]

De acuerdo con Wood (2017), la función $f_j(X_{jt})$ puede ser representada entre otras funciones por un spline cúbico, que es una curva formada por secciones de polinomios cúbicos conectados de forma tal que exista la primera derivada para que se garantice la continuidad de la función en el punto y que exista la segunda derivada para que no se den cambios de concavidad de un lado a otro del punto, garantizando de esta manera la suavidad de curva alrededor del punto. A los puntos de conexión se les conoce como nudos del spline, que pueden ser igualmente espaciados en el rango de X_{jt} o estar posicionados en sus cuantiles.

Meghanathan (2015) usa una técnica llamada Waikato Environment for Knowledge Analysis (WEKA) para comparar estudios de patrones de violencia y delitos no normalizados de un repositorio de California. Esta aplicación implementó Linear Regression, Additive Regression y Decision Stump. Esto con el fin de demostrar la efectividad y precisión de los algoritmos utilizados para predecir los patrones de delitos violentos. Su aplicación consistió en asociar e identificar correlaciones en los datos de manera que se pueda establecer una relación, en clasificar de tal manera que se pudiera descubrir similitudes, en predicción basadas en el reconocimiento de patrones y la en la visualización.

En síntesis, se han propuesto varios tipos de modelos para predecir varios delitos que en su mayoría han sido de ensamblaje y de identificación de segmentos. Por consiguiente, se han abordado para problemas de clasificación, de tal manera que el objetivo es catalogar los crímenes en función de ciertas características, como el lugar, la hora, el perfil de la víctima del crimen, entre otros. Es decir, que la variable a predecir será el tipo de crimen, lo que se traduce en que la variable es multinomial, puede tomar un valor dentro

[12] Véase: Supervised learning, scikit-learn(2021) & Louppe (2014)
[13] Véase Wood (2017)

de un conjunto de clases.

En la identificación de las dinámicas del crimen cabe la idea de una predicción tipo frecuentista, especificamente una estimación de solamente el número de homicidios que se producen en un intervalo de tiempo o de espacio, de modo idéntico se usaría los modelos antes expuestos pero enfocados a regresión con respuesta de conteo o con una función de pérdida tipo poisson. Se podría extrapolar el problema en función de una regresión adaptada a una distribución de probabilidad, dependiendo del objetivo a seguir, en este caso sería la distribución Poisson. No obstante, si el objetivo del problema es determinar el número de pruebas necesarias para conseguir K éxitos u homicidios en este caso, la distribución de probabilidad a utilizar es la distribución Binomial Negativa. Si el objetivo del problema es establecer el tiempo que transcurre hasta que se produce el homicidio, la distribución a modelar sería exponencial [14].

3. Metodología

3.1. Datos

En general los datos más usados para una metodología de clasificación han sido; el tipo de crimen, la temporalidad de ocurrencia del crimen, es decir, la hora, el mes, el día, la localidad, el nombre del distrito policial, la latitud y longitud del lugar del crimen, es decir, las coordenadas del área. En la aplicación de Hitesh Kumar Reddy ToppiReddya. B. (2018) se utilizó la ubicación, el distrito, el área comunitaria, coordenada X, coordenada Y, latitud, longitud, hora y mes. En la aplicación de Ginger Saltos, E.(2017), utilizaron el Id del crimen, la fecha del crimen, el nombre de quien reportó el crimen, la longitud, la latitud, el código del área de superproducción de capa inferior (LSOA) donde se cometió el crimen, 16 tipos de delitos, una referencia a cualquiera de los resultados asociados con el crimen ocurrido más recientemente e información adicional como el contexto. [15]

Los datos implementados para la construcción del modelo de predicción son provistos por el Sistema de Información para la Seguridad y Convivencia (SISC). La base de datos contiene cuatro tipos de delitos principales; Secuestro, Extorsión, Hurto y Homicidio, lesiones personales e incautación de armas y mercancias. De éstos se desprenden otros delitos, los tipos de delitos contenidos son: Hurto a persona, Hurto de moto, Lesión no fatal dolosa según Policía, Hurto a establecimiento comercial, Hurto de carro, Hurto a residencia, Hurto por piratería terrestre, Hurto a entidad financiera, Hurto de semoviente, Homicidio, Lesión no fatal dolosa, Extorsión y Secuestro. La temporalidad es desde el 2003 hasta el 2021, de manera diaria. Las variables obtenidas son: la fecha; de la cual se extrajo, el día, la hora, el minuto, la jornada y el año, también se tiene el sexo de la persona víctima del delito, la edad, el arma que se utilizó y las coordenadas en las cuales se realizó el hecho. Se cuenta con 309.883 observaciones y 15 variables.

3.2. Modelo tentativo de variable dependiente de conteo

Para estimar el efecto de algunas variables en el número de homicidios en Medellín, se propone un modelo de series de tiempo de conteo siguiendo modelos lineales generalizados. Esta es una clase flexible de modelos que pueden describir la correlación serial de una manera parsimoniosa. La media condicional está vinculada a sus valores pasados y a posibles efectos covariables. Se propone este método porque hay temporalidad en dos datos de homicidios y también se podría obtener el número de homicidios por día y modelarlo en función de información carácteristica del homicidio. La distribución condicional puede ser Poisson en las situaciones en las cuales la varianza y el valor esperado de la variable aleatoria son similares o Binomial negativo los cuales son más flexibles al no asumir que la media es igual a la varianza. Un caso especial importante de esta clase es el llamado modelo INGARCH y su extensión de log-lineal

Una regresión con respuesta Poisson también podría ser mediante el modelo Generalized Additive Model

[14] Véase Mendenhall Sheaffer,Estadística Matemática con Aplicaciones
[15] Véase: Anexo para saber de los tipos de datos que se utilizarían en un modelo de conteo.

(GAM). Este modelo podría funcionar muy bien, debido a su mayor flexibilidad entre las relaciones de las covariables y la variable explicada. Incluso, el método de regresión de árbol de decisión también puede ser adaptado para tener una función de pérdida asociada a una distribución Poisson como; la pérdida media de desviación de Poisson.[16]Sin embargo se propone un modelo que tiene en cuenta la temporalidad. Se supone una serie de tiempo de conteo dada por $y_t : \ t \in N$. Se denotará $X_t : \ t \in N$ como un vector covariable r-dimensional variable en el tiempo, $X_t : (X_t 1, ..., X_t, r)^T$. La media condicional de la serie temporal de conteo es; $E(Y_t|F_{t-1})) = \lambda_t$, donde F_t es el conjunto de información $Y_t, \lambda_t, X_{t+1} : \ t \in N$. El supuesto de distribución para Y_t está dada por F_{t-1}. La forma general del modelo es:

$$\mathbb{g}(\ \lambda_t) = B_0 + \sum_{k=1}^{p} B_k \tilde{g}(Y_{t-ik}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-jl}) + n^T X_t$$

Donde $g : \mathbb{R}^+ \to \mathbb{R}$ es la función link y $\tilde{g} : \mathbb{R}^+ \to \mathbb{R}$ es la función de transformación. n es un vector que corresponde a los efectos de las covariables. Por lo tanto el predictor lineal es $v_t = g(\lambda_t)$, Para esta regresión se tiene un conjunto P de observaciones pasadas de la variable respuesta. Analogamente, se define un conjunto Q para las medias latentes rezagadas.

3.3. Análisis Exploratorio

En la figura 1 se evidencia los delitos con mayor frecuencia en el periodo descrito antes. Sin duda, el delito más cometido en Medellín según la base de datos, es el hurto a persona, seguido de hurto de motos, más de la mitad de los delitos que se han cometido desde 2003 son de este tipo. La razón de esto es porque no requiere la misma capacidad instalada que un homicidio, una extorsión o secuestro. Lo anterior se sustenta en que los hurtos han sido el 77 % de los casos. Las situaciones de homicidios son el 21.5 % con respecto al total y por último las situaciones de extorsión y secuestro son el 1.3 % del total.[17]

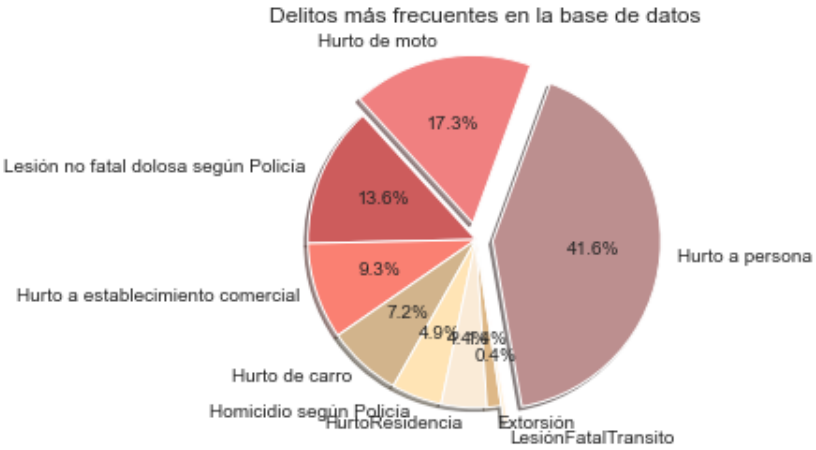

Figura 1: Proporción de los tipos de delitos desde 2003 a 2021. *Elaboración Propia*

En la figura 2 se refleja la proporción en la cual los hombres y las mujeres sufren de delitos por día para los delitos de hurto a personas y de motos respectivamente, en general, se puede afirmar que los hombres en su mayoría son las víctimas de delitos en Medellín, y los días en los que más se presentan estas situaciones son los viernes y sábados, en delitos a personas y miércoles y jueves en hurtos de motos.

En cuanto al homicidio, se refleja en la figura 3 que existe una mayor distancia en términos de homicidio entre hombres y mujeres, continuando con una mayor inclinación en los hombres. La proporción de víctimas por cuenta de Lesión no fatal dolosa, los hombres siguieron siendo el grupo mayoritario especialmente el domingo, en las lesiones hay un incremento considerable de mujeres victimas. En general, se puede evidenciar un flagelo de delincuencia muy pronunciado en los hombres y en los fin de semanas y

[16]Véase: Alexander Noll, (2020).Case Study: French Motor Third-Party Liability Claims
[17]Véase: Anexo para encontrar la tabla donde se demuestra lo dicho

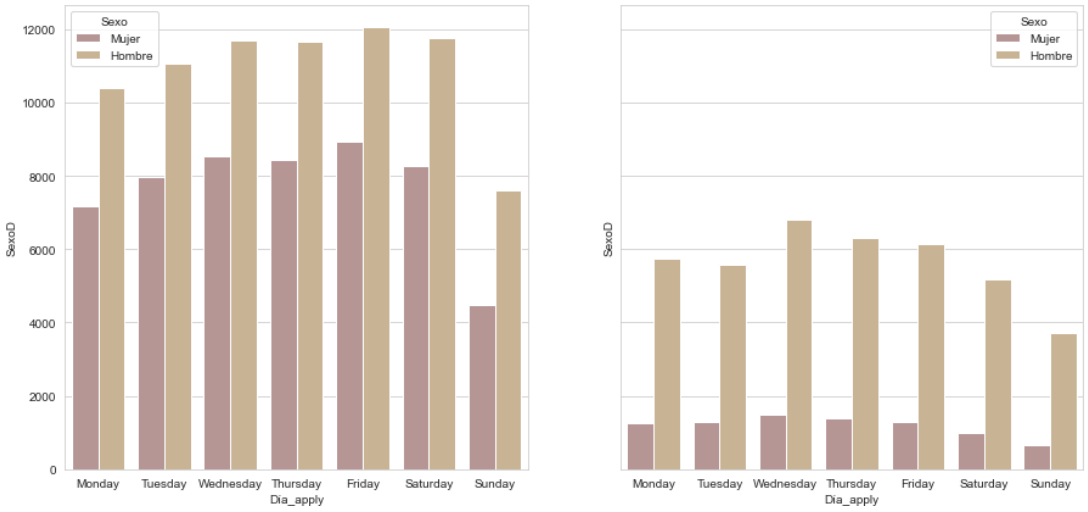

Figura 2: Frecuencia de los hurtos a personas y de motos por día y por sexo desde 2003 hasta 2021. *Elaboración Propia*

en algunas veces en mitad de semana o finalizando la misma.

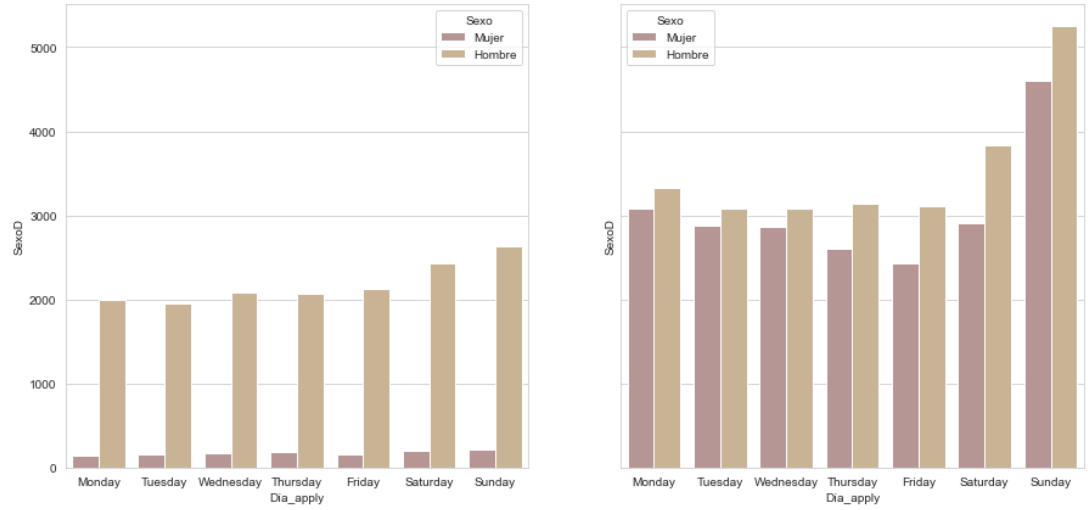

Figura 3: Frecuencia de los homicidios y lesiones personales no mortales por día y por sexo desde 2003 hasta 2021. *Elaboración Propia*

El 68 % de la población que ha sufrido de delitos son hombres. Los delitos con mayor número de casos son; el hurto a personas, el hurto a moto, la lesión no fatal dolosa, el hurto a establecimiento comercial, el hurto de carro, el homicidio y el hurto a residencia. Cuando ocurre un hurto a entidad financiera, hurto a establecimiento comercial, hurto por piratería terrestre, homicidio, hurto a moto o hurto a carro, entre el 80 % y 99.4 % son a hombres. Pero en la extorsión su participación disminuye con el 68 % y en lesión no fatal dolosa con el 53 %.

En la figura 4, se puede evidenciar que la mayoría de los delitos se cometen en la tarde y en la noche. La jornada de la tarde es la más frecuente en situaciones de hurto a persona. En situación de lesión no fatal dolosa se presenta en la jornada de la noche, en cuanto al homicidio y el hurto de moto la jornada más frecuente es en la noche. En la figura 5, refleja que la frecuencia en delitos aumenta en las horas de la tarde y noche, especialmente a las siete y ocho de la noche. Es de resaltar el hecho de que, aunque sigue siendo de noche y que posiblemente hay más oportunidad de criminalidad a las horas tardes, se refleja una disminución de éstos desde las nueve de la noche. Siendo las horas de madrugada las más “seguras”.

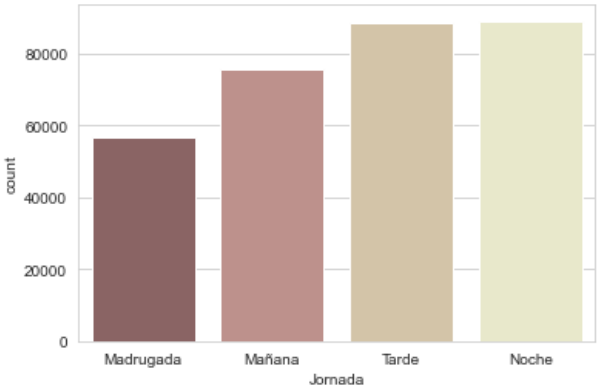

Figura 4: Jornada más frecuente en toda la base de datos *Elaboración Propia*

Tal es el caso, desde la una de la mañana a las cinco de la mañana.

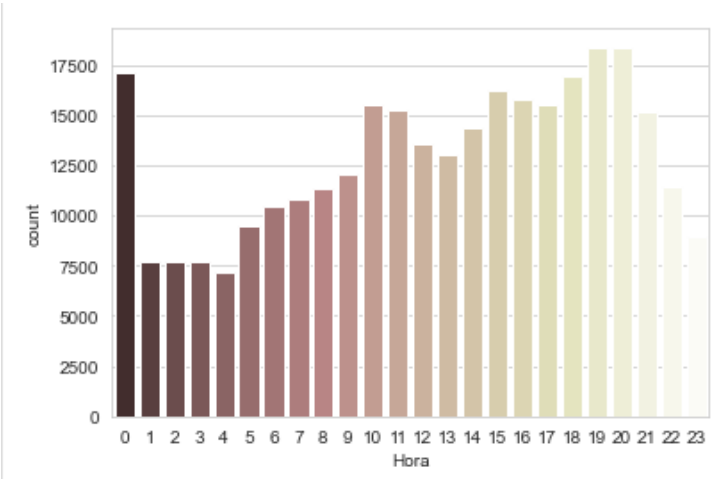

Figura 5: Las horas del día más frecuentes en toda la base de datos *Elaboración Propia*

En términos del uso de armas en los delitos, se puede evidenciar que en hurto de moto es más común usar como medio de arma una llave maestra y arma de fuego, en hurtos a personas es más frecuente no utilizar arma y si se utiliza, son las armas de fuego o cortopunzante. En cuanto a la lesión no fatal es más frecuente utilizar un arma contundente, cortopunzante o de fuego. Sin embargo, vale la pena mencionar que este delito es uno en los cuales se hace un mayor uso de varios medios como explosivos, químico, combustibles, pólvora, entre otras. Para situaciones de homicidio el arma más frecuente es de fuego, cortopunzante y objeto contundente. Las armas más utilizadas en jóvenes de menos de 20 años en delitos son; el ácido en delitos como Lesión no fatal dolosa y arma cortopunzante en situaciones de hurto y homicidio. Y para jóvenes de menos de 28 años, el arma más usada es el ácido y alimento vencido en situaciones de lesión no fata. Y arma cortopunzante en situaciones de homicidio y de hurto.

3.3.1. Análisis exploratorio espacial

Con el fin de dar un avistamiento del comportamiento de los delitos en las comunas de Medellín, se realiza una agrupación de las localidades de acuerdo a delitos; hurtos a personas, de carro, de moto, a establecimiento, y a residencia, violencia intrafamiliar, lesiones personales, extorsión, convivencia, y homicidios. Se decidió tener en cuenta todo este conjunto de crimenes de manera sincrónico, debido a que entre delitos hay relaciones que no se deberían de suponer como nulas.

El método k-means, es uno de los algoritmos no supervizados más utilizados para dividir un conjunto de datos en varios grupos teniendo en cuenta diferente información. La idea básica detrás de la agrupa-

ción de k-means consiste en definir agrupaciones de modo que se minimice la variación total dentro de la agrupación.[18], este método utiliza información intrínseca en los datos para evaluar la calidad de la agrupación, se utilizan algunas métricas para determinar el número optimo de agrupamientos; la conectividad, Silhouette y el índice de Dunn. El método sugerido por la evaluación fue hierarchical y k-means respectivamente, por lo cual se escogio k-means por tener una lógica más parsimoniosa y dado que la dimensionalidad de los datos no requería mucho proceso[19]. De igual manera mediante uso de los métodos Elbow y Silhouette, se determina que el número optimo de agrupaciones son cuatro.

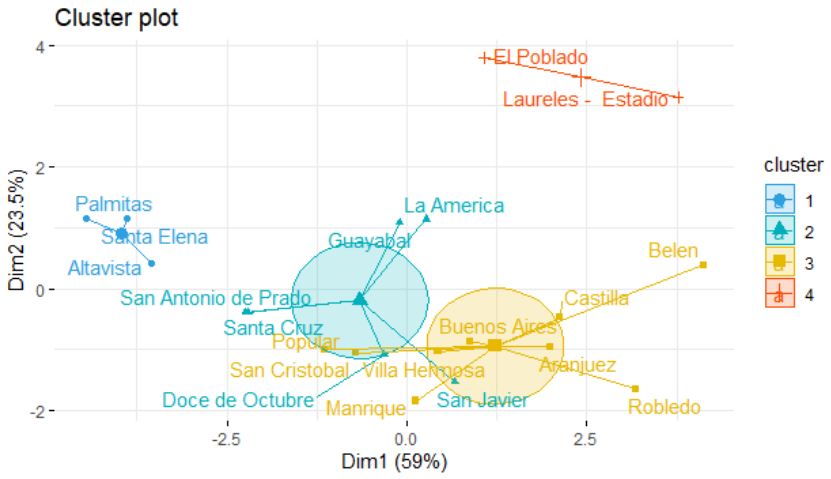

Figura 6: Cluster de comunas con kmeans *Elaboración Propia*

En la figura 6 se expone las agrupaciones de comunas en Medellín que se generan en términos de delitos. Lo que se puede resaltar es que esta agrupación tienen en cuenta más del 60 % de variabilidad de la información y que las comunas más atípicas en delincuencia son el poblado y Laureles. En cuanto a las comunas cercanas al centro de la ciudad presentan un comportamineto similar, pese a esto hay una disimilitud que se expone mediante dos agrupamientos.

En las figuras 7 y 8 se encuentran de manera espacial la densidad kernel de los homicidios(delito poco frecuente) y hurtos (delito muy frecuente) respectivamente. Esta técnica representa la densidad de puntos en el espacio geográfico. La densidad está basada en el número de puntos que caen dentro de cierto ancho de banda. En general, el homicidio se concentra en la comuna, La Candelaria con 300 a 400 homicidios desde 2016 hasta 2021, con comunas cercanas como Aranjuez, Manrique, Villa hermosa con homicidios entre 100 y 200 en el mismo periodo. En cuanto a los hurtos también se evidencia una concentración en la comuna La Candelaria con 1200 a 1400 hurtos en el mismo periodo descrito anteriormente. Aunque la desidad no es muy extensa se aprecia que la mayoría de localidades por fuera de esta concentración sí han tenido entre 0 a 400 hurtos.

3.4. Modelación

3.4.1. Modelo de clasificación

Para predecir el tipo de crimen en Medellín, el modelo propuesto se enfocará en una estimación de clasificación como Stochastic Gradient Boosting, el cual tendría como algoritmo base árboles de decisión. Especificamente, un modelo de multiples clases. Su característica principal es la capacidad de aprender de los errores de tal manera que maneja muy bien el "trade off" entre sesgo y varianza.

El algoritmo de ensamblaje Boosting combina múltiples clasificadores débiles. Este algoritmo tiene dos ponderadores: un ponderador de las observaciones mal clasificadas y un ponderador de los métodos bases de clasificación. En esencia, pondera con mayor peso a las observaciones malas, pero también pondera con

[18] Véase: El paquete de R, clValid (G. Brock et al., 2008)
[19] Véase: Anexo, en el cual se muestra que en este caso k-means obtiene los mismos resultados de un método jerárquico

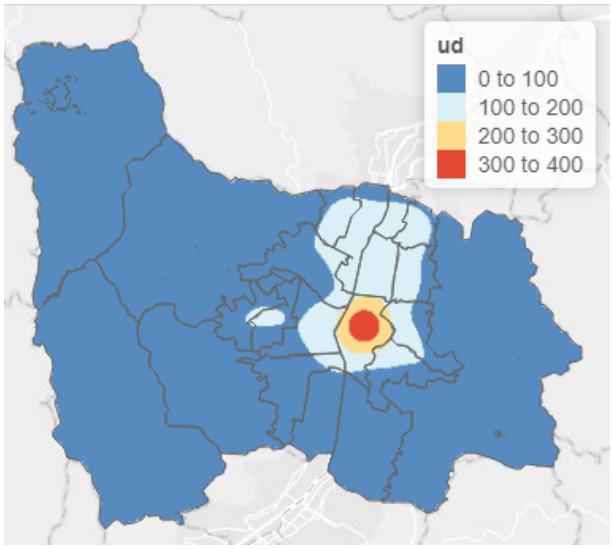

Figura 7: Análisis Kernel de homicidio desde 2016 hasta 2021 *Elaboración Propia*

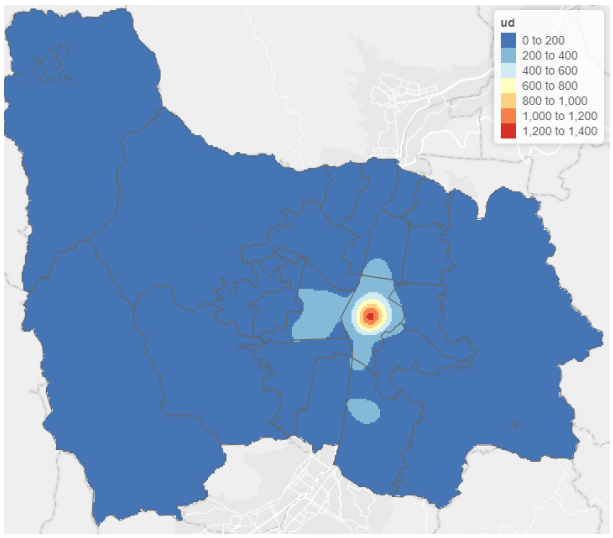

Figura 8: Análisis Kernel de hurtos desde 2016 hasta 2021 *Elaboración Propia*

mayor peso a los clasificadores buenos. Por lo general, el algoritmo base son árboles de decisiones, debido a que es el método más común en metodologías de ensamblaje y porque muchos árboles de decisión tiene un alto poder predictivo debido a la gran aleatoriedad que posee el algoritmo, inyecta un componente aleatorio al modelo para descorrelacionar la muestra, debido a que en cada iteración se extrae una submuestra aleatoria del conjunto de características sin reemplazamiento para el entrenamiento. Cabe mencionar que el método Gradient Boosting introduce métodos numéricos para que éste converja. La idea escencial es la siguiente; se empieza con un peso inicial de las observaciones $w_i = \frac{1}{n}$, donde $i = 1, ..., n$ observaciones. La aleatoriedad se refleja en las m submuestras adquiridas y para cada m submuestras se entrena m clasificadores usando los datos con sus respectivos pesos w_i.[20]

Visto de esta forma, se calcula la predicción del error, se obtiene un ponderador del ensamblaje, se actualiza los pesos de acuerdo con este ponderador y el proceso se repite hasta obtener un ensamblaje con un ponderador que refleje una buena calificación de dicho ensamblaje.

[20] Véase: Friedman(2001)

$$\text{err}_m = E[I(y_i \neq f_m(x))] = \frac{\sum_{i=1}^{n} w_i I(y_i \neq f_m(x_i))}{\sum_{i=1}^{n} w_i}$$

$$\alpha_m = log((1 - err_m)/err_m)$$

$$\text{w}_i \hookleftarrow w_i \cdot exp[\alpha_m I(y_i \neq f_m(x_i))]$$

$$\text{F(x)} = \text{sign}\ [\sum_{m=1}^{M} \alpha_m f_m(x)]$$

Donde err_m es el error para calcular, que está dado por la comparación entre las observaciones y la función clasificadora base evaluada teniendo en cuenta los pesos de dichas observaciones. α_m es el ponderador del ensamblaje, w_i es la actualización de los pesos de las observaciones que debe estar normalizado, es decir que $\sum_i w_i = 1$ Y $F(x)$ es el ensamblaje que es actualizado en cada iteración.

El objetivo es minimizar la pérdida para cada clasificación. Se utiliza el algoritmo forward sagewise, que tiene como objetivo minimizar la función de pérdida por medio de métodos numéricos, la forma en la cual encuentra ese mínimo es mediante el gradiente descendente, es decir, con la derivada que minimiza el error junto con una tasa de aprendizaje.

$$\widehat{f^{(0)}}(x) = arg\ min\ \sum_{i=1}^{n} L(y_i, \rho)$$

$$\widetilde{y}_{\pi(i),m}\ = -\hat{g}_m(x_{\pi(i)})$$

$$\widehat{\Upsilon m}(x) = arg\ min\ \sum_{i=1}^{n} (y(i), m - Bh(x(i); \Upsilon)^2$$

$$\rho_m = argmin \sum_{i=1}^{n} (y_{\pi i}, \hat{f}^{m-1}(x_{\pi i}) + \rho h(x_{\pi i}; \hat{\gamma_m})$$

$$f\ (x) = \widehat{f^{(0)}}\ (x) + \widehat{\sum_{m=1}^{M} \hat{\rho}_m h\ (x; \hat{\gamma}_m)}$$

El algoritmo empieza con valores iniciales minimizando la función de pérdida y actualizando las "pseudo-respuestas". Donde y_i son las observaciones a predecir, x_i son las P características, en cada construcción del árbol se minimiza $L(.)$, que es la función de pérdida dada por la comparación de las observaciones y la función evaluada la cual tiene una ponderación de la clasificación de ese método, $g(.)$ es la tasa de aprendizaje en la cual el algoritmo va convergiendo, como es un mínimo lo que se desea es que vaya cambiando de manera descendente, $\gamma(.)$ representa los hiperparámetros de la función base, que se actualiza minimizando el error de manera tal que pondera las clasificaciones de las observaciones mal asignadas, $\rho(.)$ es el ponderador de las clasificaciones bases, siguiendo la misma idea de comparar las observaciones con la función evaluada anterior sumando una moderación en términos de $\gamma(.)$ así mismo, $h(.)$ es el ponderador del ensamblaje. [21]

En general, para minimizar la pérdida exponencial por medio del algoritmo de forward stagewise additive, se debe de estimar la ponderación de cada actualización del método base de clasificación, se debe estimar los hiperparámetros de la función base junto con una tasa de aprendizaje hasta llegar a un ensamblaje en el cual haya convergido.

El algoritmo Boosting, Adaboost y el Extra Gradient Boosting, introducen mejoras para lidiar con la compensación de sesgo-varianza, es un sistema de aumento de gradiente de árbol eficiente, flexible y escalable, maneja fácilmente los valores perdidos, admite muchas funciones de pérdida para regresión y clasificación, permite personalizar las funciones de pérdida y las métricas de evaluación, permite el ajuste de parámetros y permite procesar cientos de millones de instancias en un escritorio, Chen and Guestrin (2016). Cabe mencionar que la técnica de *Gradient Boosting Meachine* se ha utilizado para encontrar relaciones ocultas de redes criminales[22].

[21] Véase: Chen and Guestrin (2016)

[22] Véase Friedman (2002) para profundizar en Boosting

3.4.2. Preprocesamiento

Se obtuvo cuatro bases de datos, de homicidios, de clases de hurtos, de secuestro y de extorsión, con estas cuatro se construyó una sola base, la cual contiene el día, que es una fecha con año, mes, día, hora, minutos y segundos. La variable sexo, si es hombre o mujer, la edad, la conducta que es el tipo de crimen, la ocupación, las coordenadas y el arma que se utilizó. Desde la variable día, se extrajo la variable nombre del día, hora, minuto, jornada y año para que fueran variables individuales. Se eliminaron conductas que no se considera un crimen externo, es decir, con voluntad como el suicidio, Lesión no fatal accidente de tránsito, Lesión no fatal accidental y Lesión fatal accidental. Se realizó una identificación de valores faltantes, encontrando que en la variable ocupación, el 80 % no tenía valor. Por lo cual, se eliminó, también se realizó una imputación con la moda para las variables sexo y arma. También se hizo un tratamiento a los nombres escritos de manera diferente pero que pertenece a una sola categoría, esto último con el fin de no contar más categorías que las verdaderas en cada variable.

Por último, se realizó una codificación numérica de las variables categóricas, específicamente a las variables conducta, jornada, sexo, arma y día. Para evitar un ajuste excesivo y obtener una precisión más realista, el conjunto de datos se divide en dos partes: conjunto de datos de prueba y conjunto de datos de entrenamiento. El conjunto de datos de entrenamiento contiene todas las características junto con la etiqueta de destino. El conjunto de datos de prueba solo contiene las características a partir de las cuales un modelo de aprendizaje automático predice la etiqueta de destino. No se realizó una selección de características para no sesgar el modelo y también porque no se cuenta con muchas variables[23].

3.5. Resultados

Se utilizo las observaciones desde 2003 hasta 2021 y se realizó varios métodos de aprendizaje para clasificar los tipos de crimen. Entre ellos están: Logit Multinomial, con un accuracy de 67.5 %, K-vecinos más cercanos con un accuracy de 50 %, un máquina de vector de soporte con un accuracy de 43.6 %, Bosques aleatorios con un accuracy de 75.4 % y por último se implementó el Gradient Boosting con un accuracy del 74.3 %. Todos los algoritmos anteriores se realizaron sin búsqueda de hiperparámetros. Es decir que sus argumentos fueron los por defecto. Por consiguiente, se escogió el método de Gradient Boosting ya que éste utiliza árboles de decisión como estimador base y dado que el otro algoritmo que también tuvo una buena capacidad predictiva fue los árboles de decisión.

Este método tiene varios parámetros, como; la función de pérdida (loss), se utilizó la pérdida deviance, para el número de árboles (n_estimators), se utilizó 150 árboles, en cuanto a la tasa de aprendizaje que se utiliza para controlar la influencia que tiene cada weak learner en el conjunto del ensemble (learning_rate), fue de 0.1, respecto a la submuestra (subsample) fue de 0.7, en cuanto a la profundidad del árbol o tamaño del árbol (max_depth), número mínimo de muestras necesarias para dividir un nodo interno (min_samples_split), número mínimo de muestras necesarias para estar en un nodo hoja (min_samples_leaf), se especificaron los que el algoritmo tiene por defecto. Lo anterior como forma de mejorar su capacidad predictiva.

Como la ocurrencia de un tipo de delito no implica que los demás delitos no ocurran, también se tiene en cuenta la probabilidad de ocurrencia de cada tipo Por esta razón se estimó el modelo Multinomial el cual es la generalización de un modelo Logit pero con clasificación múltiple y en el Gradient Boosting se realiza una predicción de probabilidades obteniendo la probabilidad con la que el modelo considera que cada observación puede pertenecer a cada una de las clases. El resultado de esta estimación es un array con una fila por observación y tantas columnas como tipos de crimen hayan.

[23] Para el preprocesamiento se utilizó, Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

4. Conclusiones

Para la clasificación de tipos de crímenes en una ciudad es importante tener en cuenta la disposición de información detallada debido a que los delitos no son excluyentes, es decir, la ocurrencia de un delito no significa que no exista la ocurrencia de otro delito. Por lo cual, en este trabajo se implementó un análisis descriptivo frecuentista y espacial, para dar un avistamiento de la dinámica del crimen en la ciudad de Medellín. También se realizó la clasificación de los tipos de delitos con diferentes alternativas de aprendizaje entre ellas el Logit Multinomial, K-vecinos más cercanos, Maquinas de vector de soporte, Bosques aleatorios y Gradient Boosting, lo anterior se realizó para determinar cuál algoritmo aprendizaje era más preciso dado los datos disponibles, por lo cual, se concluyó que el mejor modelo en términos de capacidad predictiva fue Gradient Boosting, con un 74.5 % de accuracy.

Sería interesante agregar otras variables que puedan mejorar la calidad de estas como; Estrato económico, proporción de reportes de inseguridad por barrio, estaciones de policía, payment rate of security fee (vacuna), la percepción de los vecinos en el sector, índice de pobreza en el lugar del hecho, nivel de escolaridad y calidad de servicios públicos. Y también es pertinente complementar la clasificación anterior con un método de efectos espaciales en el cual se pueda predecir el lugar junto con la predicción del tipo de crimen, teniendo en cuenta que algunos delitos son más fáciles de predecir dada su naturaleza, por ejemplo, los hurtos.

5. Bibliografía

[1] M. A. Andresen and N. Malleson, Police foot patrol and crime displacement a local analysis, Journal of Contemporary Criminal Justice, 30 (2) (2014) 186–199.

[2] J. Bachner, Predictive policing: Preventing crime with data and analytics (IBM Center for The Business of Government, 2013).

[3] T. H. Grubesic, On the application of fuzzy clustering for crime hot spot detection, Journal of Quantitative Criminology, 22 (1) (2006) 77–105.

[4] M. Helbich, J. Hagenauer, M. Leitner, and R. Edwards, Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach, Cartography and Geographic Information Science, 40 (4) (2013) 326–336, http://dx.doi.org/10.1080/15230406.2013.779780.

[5] S.-T. Li, S.-C. Kuo, and F.-C. Tsai, An intelligent decision-support model using FSOM and rule extraction for crime prevention, Expert Systems with Applications, 37 (10) (2010) 7108–7119, http://www.sciencedirect.com/scie
S0957417410001855.

[6] K. Dahbur and T. Muscarello, Classification system for serial criminal patterns, Ar- tificial Intelligence and Law, 11 (4) (2003) 251–269.

[7] S. Lin and D. E. Brown, An outlier-based data association method for linking criminal incidents, Decision Support Systems, 41 (3) (2006) 604–615, http://www. sciencedirect.com/science/article/pii/S0167923604001344.

[8] D. E. Brown and S. Hagen, Data association methods with applications to law enforcement, Decision Support Systems, 34 (4) (2003) 369–378, http://www.sciencedirect. com/science/article/pii/S0167923602000647.

[9] A. Malathi and S. S. Baboo, An enhanced algorithm to predict a future crime using data mining, International Journal of Computer Applications, 21 (1) (2011) 1–6 .

[10] A. T. Murray and T. H. Grubesic, Exploring spatial patterns of crime using nonhierarchical cluster analysis, in Crime modeling and mapping using geospatial tech- nologies, ed. M. Leitner (Springer, 2013), pp. 105–124.

[11] S. V. Nath, Crime pattern detection using data mining, in Web Intelligence and Intel- ligent Agent Technology Workshops, 2006 IEEE/WIC/ACM International Conference on (IEEE, 2006), pp. 41–44.

[12] G. C. Oatley and B. W. Ewart, Crimes analysis software:ins in maps clustering and bayes net prediction, Expert Systems with Applications, 25 (4) (2003) 569–588.

[13] P. Phillips and I. Lee, Mining co-distribution patterns for large crime datasets, Expert Systems with Applications, 39 (14) (2012) 11 556–11 563, http://www. sciencedirect.com/science/article/pii/S0957417412005945.

[14] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, Crime data mining: a general framework and some examples, IEE Computer, 37 (4) (2004) 50–56.

[15] lkesh Bharati1, D. S. (2018). Crime Prediction and Analysis Using Machine Learning. International Research Journal of Engineering and Technology (IRJET), 6.

[16] Barnadas, M. V. (2016). MACHINE LEARNING APPLIED TO CRIME PREDICTION. Universitat Politècnica de Catalunya, 50.

[17] Ginger Saltos, E. H. (2017). AN EXPLORATION OF CRIME PREDICTION USING DATA. International Journal of Information Technology Decision Making, 29.

[18] Hitesh Kumar Reddy ToppiReddya, *. B. (2018). Crime Prediction Monitoring Framework Based on Spatial Analysis. International Conference on Computational Intelligence and Data Science, 10.

[19] Luiz G.A. Alves, H. V. (2017). Crime prediction through urban metrics and statistical. Physica A, 9.

[20] Meghanathan, L. M. (2015). USING MACHINE LEARNING ALGORITHMS TO. Machine Learning and Applications: An International Journal (MLAIJ), Vol.2, No.1, .

[21] Sohrab Hossain, A. A. (2020). Crime Prediction Using Spatio-Temporal Data. Department of Computer Science and Engineering, East Delta University, 13.

[22] Suhong Kim, P. J. (2018). Crime Analysis Through Machine Learning. Fraser International College, Simon Fraser University, 6.

[23] Alkesh Bharati1, D. S. (2018). Crime Prediction and Analysis Using Machine Learning. International Research Journal of Engineering and Technology (IRJET), 6.

[24] Barnadas, M. V. (2016). MACHINE LEARNING APPLIED TO CRIME PREDICTION. Universitat Politècnica de Catalunya, 50.

[25] Ginger Saltos, E. H. (2017). AN EXPLORATION OF CRIME PREDICTION USING DATA. International Journal of Information Technology Decision Making, 29.

[27] Luiz G.A. Alves, H. V. (2017). Crime prediction through urban metrics and statistical. Physica A, 9.

[28] Meghanathan, L. M. (2015). USING MACHINE LEARNING ALGORITHMS TO. Machine Learning and Applications: An International Journal (MLAIJ), Vol.2, No.1, .

[29] Sohrab Hossain, A. A. (2020). Crime Prediction Using Spatio-Temporal Data. Department of Computer Science and Engineering, East Delta University,, 13.

[30] Suhong Kim, P. J. (2018). Crime Analysis Through Machine Learning. Fraser International College, Simon Fraser University, 6.

[31] Wood, S. N. (2017) *Generalized Addictive Models An introduction with R.*

[32] Supervised learning¶ (2021). Scikit Learn, scikit-learn.org. Recuperado de:

$https://scikit-learn.org/stable/supervised_learning.html supervised-learning.$

[33] Becker, G. S. (1968), 'Crime and punishment: An economic approach'. Journal of Political Economy 76, 169–217.

[34] Becker, G. S. (1976), The Economic Approach to Human Behavior. Chicago: University of Chicago Press.

[35] Becker, G. S. and G. J. Stigler (1974), 'Law enforcement, malfeasance, and compensation of enforcers'. Journal of Legal Studies 3, 1–18.

[36] Chiricos, T. G. (1987), 'Rates of crime and unemployment: An analysis of aggregate research evidence'. Social Problems 34, 187–212.

[37] Winter, H. (2005). The Economics of Crime, An introduction to rational crime analysis.

[38] Erling Eide, P. H. (2006). Economics of Crime.

[39] D'angelo, A. S. (2019). Conceptualización del Crimen Organizado y su regulación en la legislación Penal Colombiana. 34.

[40] Tsybakov. Introduction to Nonparametric Estimation. Springer Series in Statistics. Springer, New York, 2008.

[41] Louppe. Understanding Random Forests: From Theory to Practice. PhD thesis, University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering Computer Science, Liège, Belgium, 2014.

[42] Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

[43] Morales (2018) GLM para respuesta Poisson, rstudio-pubs-static, sacado de: https://bookdown.org/j$_m$*orales/weblinm*

[44] Friedman. Stochastic gradient boosting. Computational Statistics Data Analysis, 38(4):367–378, 2002.

[45] Copyright © 2018 Javier Morales. Universidad Miguel Hernández de Elche.GLM para respuesta Poisson

[46] Louppe. Understanding Random Forests: From Theory to Practice. PhD thesis, University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering Computer Science, Liège, Belgium, 2014.

[47] Alexander Noll,Robert Salzmann, (2020). Case Study: French Motor Third-Party Liability Claims.

[48] Gareth James, Daniela Witten (2013). An Introduction to Statistical Learning with Applications in R.

[49] Ross.S.A first course in probability.Prentince Hall para profundizar en la distribuci on de Poisson

[50] Blasco (2016). Economía del delito y el efecto institucional: una aproximación teórica y empírica.

[51] Zaffaroni, E. R., (1991) Manual de derecho penal. México, Editorial Cárdenas

[52] M. A. Andresen and N. Malleson, Police foot patrol and crime displacement a local analysis, Journal of Contemporary Criminal Justice, 30 (2) (2014) 186–199

[53] P. Phillips and I. Lee, Mining co-distribution patterns for large crime datasets, Expert Systems with Applications, 39 (14) (2012) 11 556–11 563, http://www. sciencedirect.com/science/article/pii/S0957417412005945.

[54] K. Dahbur and T. Muscarello, Classification system for serial criminal patterns, Artificial Intelligence and Law, 11 (4) (2003) 251–269

[55] G. C. Oatley and B. W. Ewart, Crimes analysis software:ins in maps clustering and bayes net prediction, Expert Systems with Applications, 25 (4) (2003) 569–588.

[56] Tsybakov (2008). Introduction to Nonparametric Estimation.

[57] Santiago Gallón (2019). Introduction to Nonparametric Regression.

[58] Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Annals of Statistics, 28(2):337–407, 2000.

[59] Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Series in Statistics. Springer, New York, 2nd. edition, 2009

[60] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519

[61] Scikit-learn: Machine Learning in Python , Pedregosa et al. , JMLR 12, págs. 2825-2830, 2011.

[62] Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

6. Apéndice

La distribución Poisson

En la aplicación de Luiz G.A. Alves, H. V. (2017) para la predicción del número de homicidios a nivel de ciudades, se utilizaron algunos indicadores urbanos; trabajo infantil (fracción de la población de 10 a 15 años que se encuentra trabajando), población anciana (ciudadanos de 60 años o más), población femenina, producto interno, analfabetismo (ciudadanos de 15 años o más que no saben leer y escribir, ingreso familiar (ingreso promedio de los residentes de la familia), población masculina, saneamiento (número de casas que tienen agua corriente y alcantarillado), desempleo (ciudadanos de los 16 años o más que están sin trabajo o buscando trabajo), número de accidentes de tráfico, número de suicidios. Otra aplicación del modelo Poisson de Londoño (2009) utilizaron una base de datos de todos los homicidios ocurridos en el periodo 2007-2008 en el Municipio de Santiago de Cali, con 3003 observaciones. Y como covariables se usaron la edad (en años), el año, el género y el tipo de arma. La variable edad se categorizó utilizando los grupos etarios.

La distribución Poisson se caracteriza por la independencia entre los sucesos y cada suceso tiene una probabilidad p de suceder durante un tiempo determinado. [24]. La función de probilidad de una variable de conteo es $P(y)$, resulta lógico que en cualquier metodología de regresión se pueda adaptar a una función de enlace de tipo Poisson.

$$\mathrm{P(y)} = \lambda^{y} exp(-\lambda) \tfrac{1}{y!}$$

donde $Y \sim Po(\lambda)$ representaría el número de veces de ocurrencia de homicidios en un período determinado y λ es el parámetro de ajuste de la distribución, la cual, representa el parámetro de intensidad. Una variable aleatoria Poisson se aproxima a una variable binomial con parámetros n que es el número de observaciones y p, cuando n es grande y p es pequeña, por consiguiente $\lambda = np$. Una variable aleatoria binomial tiene valor esperado de np y varianza $np(1-p)$ por lo cual el valor esperado y la varianza de una variable aleatoria Poisson es λ, $Var(y) = E[y^2] - (E[y])^2 = \lambda$[25]. En este método se tiene un parámetro n_i que es la exposición al riesgo para la combinación lineal, dada por una variable base o una constante. También se tiene la tasa de riesgo λ_i con la información de las variables predictoras que está dada por $\frac{y}{n_i}$. La estimación es de tipo inferencial, la cual está representada por el riesgo relativo (RR) sobre la tasa de riesgo asociado a un cambio en una unidad de la covariable x_j.[26]

En la aplicación de Guerrero Escamilla, J. B., Franco Sánchez, L. M. y Bass Zavala, S. (2018), los cálculos de las variables son;

$$\mathrm{x_1} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} X_{gi}$$

$$\mathrm{X_2} = \tfrac{NoSentencias}{PT} * 100000$$

$$\mathrm{X_3} = 1 - \tfrac{X_2}{Y} * 100$$

$$\mathrm{X_4} = (\tfrac{X_R + X_u + X_c + X_t}{n}) * 100$$

$$\mathrm{y} = \tfrac{NoDelitos}{PT} * 100000$$

Donde PT se refiere a población total,X_R es la tasa de pobreza rural, X_u es la tasa de pobreza urbana, X_c es la tasa de pobreza crónica y X_t es la pobreza temporal, X_5 es la tasa de desempleo por municipio, es el total de desempleados sobre la población económicamente activa, y por último, X_6 que es el grado de gobernabilidad que es la calidad del municipio a tratar sobre la calidad de un municipio base.

[24] Véase GLM para respuesta Poisson (2018)
[25] Véase Ross.S.A first course in probability.Prentince Hall para profundizar en la distribución de Poisson
[26] Véase: Morales (2018)

La estructura de este modelo de regresión Poisson se expresa como una regresión múltiple $Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$. Donde Y es la tasa de delincuencia por cada cien mil habitantes, X_1 es el grado promedio de escolaridad de una población, X_2 es la tasa de sentencia por municipio (es el número de sentencias por cada cien mil habitantes), X_3 es el grado de impunidad por municipio (es el nivel de impunidad por cada cien mil habitantes), X_4 es la tasa de pobreza promedio por municipio, X_5 es la tasa de desempleo por municipio, X_6 es el grado de gobernabilidad.

$$\lambda_i = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}),$$
$$\log(\mathrm{E}(\mathrm{Y}_i)) = log(\mu_i) = log(n_i \lambda_i) = log(n_i) + log(\lambda_i) = log(n_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

De esta manera, la regresión tiene un offset el cual es fijo, tiene unos los coeficientes $exp(\beta_j)$, los cuales representan el riesgo relativo (RR) sobre la tasa de incidencia de los sucesos asociado a un incremento de una unidad en la covariable x_j. El riesgo relativo para determinar el cambio entre dos niveles de una variable categórica de valores A y B es; si $RR > 1$ el efecto aumenta y disminuye cuando $RR < 1$. Tenemos el mismo riesgo cuando $RR = 1$[27].

$$\mathrm{RR}(\mathrm{X}_{AB}) = \frac{\lambda(X=B)}{\lambda(X=A)} = exp(\text{coeficiente asociado a B} - \text{coeficiente asociado a A})$$

En la figura 9 se encuentra una tabla de frecuencias, en la cual se resalta el delito más frecuente por año desde el 2003 hasta el 2020.

	Year	ConductaMásFrecuente	NúmeroCasos
0	2003	Hurto de carro	2823
1	2004	Hurto a persona	3193
2	2005	Hurto a persona	1564
3	2006	Hurto a persona	1030
4	2007	Hurto a persona	750
5	2008	Lesión no fatal dolosa según Policía	1435
6	2009	Homicidio según Policía	1428
7	2010	Homicidio según Policía	1405
8	2011	Hurto de moto	2709
9	2012	Hurto de moto	3201
10	2013	Hurto de moto	3353
11	2014	Hurto a persona	3240
12	2015	Hurto a persona	7318
13	2016	Hurto a persona	13336
14	2017	Hurto a persona	17719
15	2018	Hurto a persona	21810
16	2019	Hurto a persona	27099
17	2020	Hurto a persona	17704

Figura 9: Tabla del delito más frecuente por año desde el 2003 hasta 2021. *SISC*

[27] Véase GLM para respuesta Poisson (2018)