# DETECCIÓN DE PATRONES DE ACTIVIDAD DELICTIVA MEDIANTE TECNICAS DE MACHINE LEARNING

# DETECTION OF CRIMINAL ACTIVITY PATTERNS THROUGH MACHINE LEARNING TECHNIQUES

Héctor Andrés Mora Paz[1]
Jorge Albeiro Rivera Rosero[2]

**Resumen**

San Andrés de Tumaco, presenta una situación de orden público que genera amenazas y extorsiones, siendo fuente de ingresos para los grupos armados ilegales(FIP - Ideas para la paz, 2022) e instrumento para el control social y económico. Es por ello que en este estudio se plantea realizar una aplicación que contribuya en la predicción de actividad delictiva en municipio de Tumaco, como herramienta para pronosticar la densidad de crecimiento delincuencial zonificada, como apoyo a los actores encargados de la seguridad en el municipio para realizar planes de mitigación y aprovechamiento eficiente del pie de fuerza. La situación de orden público que azota algunas zonas del departamento Nariño, entre ellas al municipio de Tumaco; Por su parte, las amenazas y las extorsiones: desafío a la paz territorial, muestra cómo "las redes extorsivas, configuran mecanismos a través de los cuales los grupos armados ilegales o delincuenciales se apropian de las actividades económicas de los territorios"(*Defensoría Del Pueblo*, 2022). El proyecto tiene como objetivo central el de Implementar un marco experimental de comparación del compromiso en métricas de calidad entre métricas de calidad y coste computacional en algoritmos supervisados y no supervisados de *Machine Learning* para la obtención de un modelo subóptimos de predicción de actividad delictiva para el municipio de Tumaco a través de la metodología KDD. La metodología se realizó mediante el Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Data-bases – KDD)(Data, 2022). Los resultados arrojan una base de datos de delitos para toda Colombia para los años 2010 a 2019; scripts de transformación y limpieza de la cual se obtienen las bases de datos de entrenamiento tanto para las técnicas supervisadas como para las no supervisadas. Se concluye que la aplicación de interpolación Kriging, para el desarrollo de esta investigación fue la más favorable ya que el procedimiento geoestadístico avanzado generó una superficie estimada a partir de un conjunto de puntos de actividad delictiva dispersados con valores *z*, obteniendo así una auto correlación, es decir, las relaciones estadísticas entre los puntos medidos.

**Palabras clave:** Actividad delictiva, machine learning, clustering, geo interpolación

Recepción: 10 de Febrero de 2024/ Evaluación: 05 de Marzo de 2024 / Aprobado: 01 de Abril de 2024

[1] Programa de Ingeniería de Sistemas, Facultad de Ingeniería, Universidad CESMAG, San Juan de Pasto Colombia. Miembro del grupo de investigación Tecnofilia. E-mail: hamora@unicesmag.edu.co ORCID: https://orcid.org/0000-0003-3097-4757

[2] Programa de Ingeniería de Sistemas, Facultad de Ingeniería, Universidad CESMAG, San Juan de Pasto Colombia Miembro del grupo de investigación Tecnofilia. E-mail: jarivera1@unicesmag.edu.co ORCID: https://orcid.org/0000-0002-0092-8226

**Abstract**

San Andrés de Tumaco presents a public order situation that generates threats and extortions, being a source of income for illegal armed groups (FIP - Ideas for Peace, 2022) and an instrument for social and economic control. That is why this study proposes to carry out an application that contributes to the prediction of criminal activity in the municipality of Tumaco, as a tool to forecast the density of zoned criminal growth, as support to the actors in charge of security in the municipality to carry out mitigation plans and efficient use of the force. The public order situation that plagues some areas of the Nariño department, including the municipality of Tumaco; For its part, threats and extortions: a challenge to territorial peace, shows how "extortion networks configure mechanisms through which illegal or criminal armed groups appropriate the economic activities of the territories" (Defensoría Del Pueblo, 2022). The main objective of the project is to implement an experimental framework for comparing the commitment in quality metrics between quality metrics and computational cost in supervised and unsupervised Machine Learning algorithms to obtain a suboptimal model for predicting criminal activity for the municipality of Tumaco through the KDD methodology. The methodology was carried out using Knowledge Discovery in Databases (KDD) (Data, 2022). The results provide a crime database for all of Colombia for the years 2010 to 2019; transformation and cleaning scripts from which the training databases are obtained for both supervised and unsupervised techniques. It is concluded that the application of Kriging interpolation for the development of this research was the most favorable since the advanced geostatistical procedure generated an estimated surface from a set of criminal activity points dispersed with z values, thus obtaining an autocorrelation, that is, the statistical relationships between the measured points.

**Keywords**: Criminal activity, machine learning, clustering, geo interpolation.

## Introducción

San Andrés de Tumaco, es una zona especial ambiental, donde confluyen variables naturales, gran biodiversidad, características físicas: terrestres, oceánicas y meteorológicas particulares y con una influencia humana y socioeconómica específicas, que la hacen un área de interés científico y con un futuro de desarrollo agroindustrial prometedor. Estos factores se ven opacados frente a la situación de orden público que azota algunas zonas del departamento Nariño, entre ellas al municipio de Tumaco(*Diario Del Sur*, 2022); Por su parte, las amenazas y las extorsiones: desafío a la paz territorial, muestra cómo “las redes extorsivas, configuran mecanismos a través de los cuales los grupos armados ilegales o delincuenciales se apropian de las actividades económicas de los territorios”(*Defensoría Del Pueblo*, 2022). Es así como la amenaza y la extorsión, además de ser fuente de ingresos para los grupos armados ilegales, son un instrumento para el control social y económico. En este marco, la labor de liderazgo se ve afectada mediante la vulneración sistemática de los derechos a la vida, la libertad, la integridad y la seguridad personal de quienes lo ejercen. De ahí que sean objeto de amenazas y hostigamientos directos contra ellos, sus familias o las organizaciones de las que hacen parte, a ello se suman estigmatizaciones, calumnias, vigilancia y seguimientos ilegales, hurto de información, violación y allanamiento ilegal de sus domicilios y oficinas, torturas, lesiones personales, detenciones arbitrarias y persecución judicial. Prácticas que terminan en muchos casos, con desapariciones y homicidios. Toda esta violencia termina por desestructurar y disolver los procesos organizativos, dejando a las comunidades sumidas en la zozobra y la incertidumbre; con el fin de contribuir en los planes de mitigación de delincuencia, se plantea obtener un modelo de predicción y otro de agrupación de actividad

delictiva(Galindo & Catalán, 2007) realizando un marco comparativo de algoritmos supervisados y no supervisados de machine learning(Arsys, 2019), aplicando la metodología de descubrimiento de conocimiento en bases de datos(Charte, n.d.) (KDD) con históricos de actividad delictiva recopilados por el observatorio del delito(Policía Nacional de Colombia, 2015). Creando de este modo una sinergia hombre máquina que asista y refuerce mediante inteligencia artificial a los mecanismos de inteligencia de las autoridades competentes.

Siendo un medio para contribuir en la disminución de problemas de carácter social, político y económico. Logrando así la obtención de patrones que puedan ayudar a tomar decisiones con el fin de apoyar en la disminución de la actividad delictiva(*En Tumaco No Cede Criminalidad*, n.d.), convirtiéndose en una herramienta clave para ayudar a la labor que realizan los diferentes entes de control frente a la problemática en mención. Es por ello, que en este estudio se plantea realizar una aplicación que contribuya en la predicción de actividad delictiva en municipio de Tumaco(*Los Datos Lo Demuestran: Tumaco No Sabe Qué Es El Posconflicto*, n.d.), como herramienta para pronosticar la densidad de crecimiento delincuencial zonificada, como apoyo a los actores encargados de la seguridad en el municipio a realizar planes de mitigación y aprovechamiento eficiente del pie de fuerza(BBC News Mundo, n.d.).

En el desarrollo de este estudio se trabajó con algoritmos de regresión y clustering, para obtención de patrones(Modeling reality generating software, 2020); los datos a adquirir se extrajeron de las bases de datos del observatorio del delito para los años 2010 a 2019(Policía Nacional de Colombia, 2015). Para evaluar el compromiso en la predicción se utilizaron las siguientes métricas: error cuadrático medio(*¿Qué Es El Error Cuadrático Medio RMSE? | El Blog de Franz*, n.d.) para comparar los algoritmos de regresión, e inercia y silueta para los algoritmos de agrupación (agrupación). En cuanto a la visualización geográfica de los datos se utilizó el interpolador de kriking utilizando el mejor variograma dentro de los modelos lineales, gaussiano y esféricos.

## Materiales y Métodos

Se tomó como fuente de extracción de datos, el Observatorio del Delito(Policía Nacional de Colombia, 2015), ya que es un grupo estratégico del área de Investigación Criminológica, encargado del monitoreo, diagnóstico, administración de la información, evaluación y análisis de la criminalidad en la cual se va a centrar la presente investigación.

Para una mejor ubicación de los puntos de actividad delictiva se necesitó complementar el conjunto de datos con la latitud y longitud de cada barrio, obtener el polígono (mapa shapefile) de la zona urbana del municipio y se ajustó hacienda uso de la herramienta ArgGis. A continuación, se detallan las herramientas utilizadas y la metodología empleada.

### Herramientas utilizadas

Los scripts con los que se han obtenido los resultados fueron escritos en el lenguaje de programación Python (López, 2020) y se utilizaron las bibliotecas detalladas en la Tabla 1.

| Bibliotecas | Descripción |
|---|---|
| **PyKrige** | Herramienta básica a la hora de interpolar datos con previa modelización estructural. |
| **Pandas** | Biblioteca de software escrita como extensión de Numpy para manipulación y análisis de datos. |

| | |
|---|---|
| **Numpy** | Es una biblioteca de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos. |
| **MatplotLib** | Biblioteca para la generación de gráficos a partir de datos contenidos en listas o arrays. |
| **Shapely** | Biblioteca para el tratamiento de shapefiles. |
| **Pyproj** | Biblioteca para reproyectar coordenadas geográficas. |

Tabla 1. Bibliotecas importadas en Python(López, 2020).
Fuente: Esta investigación

Se utilizaron 45 archivos correspondientes a los delitos comprendidos entre los años 2010 y 2019, los cuales fueron descargados desde el Observatorio Del Delito Colombiano. Por otro lado, se construyó un archivo con los barrios del municipio de Tumaco con su respectiva información geográficas (latitud, longitud). En la Tabla 2, se muestra las variables relacionadas en los anteriores archivos.

| Variable | Descripción |
|---|---|
| **Tipo de delito** | Esta variable llamada tipo de delito, contiene el tipo de actividad delictiva que se describe en los datos obtenidos, donde los tipos de delitos pueden ser “Amenaza”, “Homicidio”, “Hurto-persona “y “Terrorismo”. |
| **Fecha** | Esta variable contiene la fecha descrita en día, mes y año en el cual ocurre cada actividad delictiva. |
| **Día** | Esta variable contiene el día en el cual ocurrió dicha actividad delictiva. |
| **Hora** | En esta variable se almacena la hora en la que ocurrió dicha actividad delictiva |
| **Barrio** | Esta variable contiene el barrio o lugar en donde ocurrió dicha actividad delictiva. |
| **Clase-Sitio** | Esta variable contiene la referencia del sitio o lugar ya sea este un establecimiento o un punto de referencia, en donde ocurrieron dichas actividades delictivas. |
| **Arma empleada** | Esta variable contiene el tipo de arma que se empleó en dicha actividad delictiva. Ejemplo, arma de fuego, arma blanca / cortopunzante etc. |
| **Móvil Agresor** | Esta variable contiene el medio o tipo de transporte que utilizó el agresor quien cometió dicha actividad delictiva. |
| **Móvil Victima** | Esta variable contiene el medio o tipo de transporte que utilizaba la victima a quien se le cometió dicha actividad delictiva. |
| **Edad** | Esta variable contiene la edad de quien cometió dicha actividad delictiva. |

| | |
|---|---|
| **Sexo** | Esta variable contiene el tipo de sexo, ya sea “femenino” o “masculino” de la persona que cometió dicho delito |
| **Estado civil** | Esta variable contiene el estado civil de la persona que cometió dicho delito, ya sea “soltero”, “unión libre”, “casado”, etc. |
| **Clase empleado** | Esta variable contiene el tipo de empleo de la persona la cual cometió dicha actividad delictiva. |
| **Profesión** | Esta variable contiene el tipo de profesión que ejercía la persona que cometió dicho delito. |
| **Escolaridad** | Esta variable contiene el nivel de estudio de la persona que cometió dicha actividad delictiva. |
| **Latitud** | Esta variable contiene el dato geográfico “latitud” del barrio en donde ocurrió dicha actividad delictiva.<br>Esta variable nos va a permitir ordenar los puntos en el “mapa. Shp” que representan a los casos de actividad delictiva. |
| **Longitud** | Esta variable contiene el dato geográfico “longitud” del barrio en donde ocurrió dicha actividad delictiva. |

Tabla 2. Descripción de las variables
Fuente: Esta investigación

Igualmente, se adquirió un polígono del municipio de Tumaco inicialmente en XML, el cual fue digitalizado con la herramienta ArcGIS con la cual se organizó, administró, analizó, y transformo dicho polígono. Igualmente, por medio de esta herramienta, se ubicaron en el mapa shapefile los polígonos correspondientes al casco urbano de Tumaco unificados en un solo shapefile (shp) para facilitar la manipulación del polígono en Python(Minera, 2019). Una vez obtenido el shapefile final se reprojectaron tanto los datos del Observatorio del Delito(Policía Nacional de Colombia, 2015) como el polígono (alusivo al mapa interno en el shapefile) a coordenadas de Mercator EPSG:3857 para trabajar con distancias cartesianas en metros. En la Figura 1 se muestra el resultado de este procesamiento, con el porcentaje de cantidad de delitos. Este mapa es el conjunto de partida para realizar los experimentos de extrapolación e interpolación (*Interpolacion y Extrapolacion – Julianapinzon*, n.d.) de delitos.

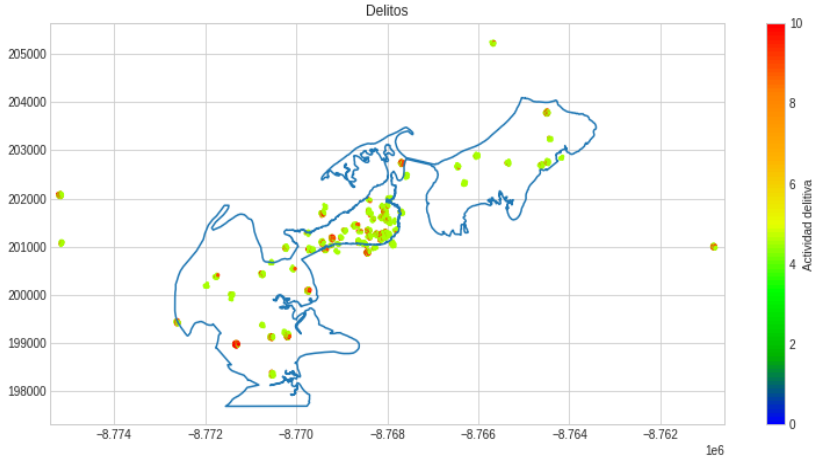


Figura 1. Mapa de la cantidad de actividad delictiva en el municipio de Tumaco.
Fuente: Esta investigación.

**Metodología**

Para el desarrollo de este estudio se utilizó la metodología, Descubrimiento de Conocimiento en Base de Datos(Alfredo, 2017) (Knowledge Discovery in Databases – KDD) (ver Figura 2), el cual es un proceso metodológico para encontrar un "modelo" válido, útil y entendible que describa patrones de acuerdo a la información. Esta metodología sigue una serie de fases ordenadas donde se aplican un conjunto de técnicas para adquirir los datos, preprocesarlos, experimentar con modelos de minería de datos, evaluar el desempeño e interpretar los resultados, en estas últimas etapas son fundamentales los modelos estadísticos, técnicas de muestreo y validación (*Tecnicas-de-Mineria-de-Datos-Para-La-Prevencion-De*, n.d.).

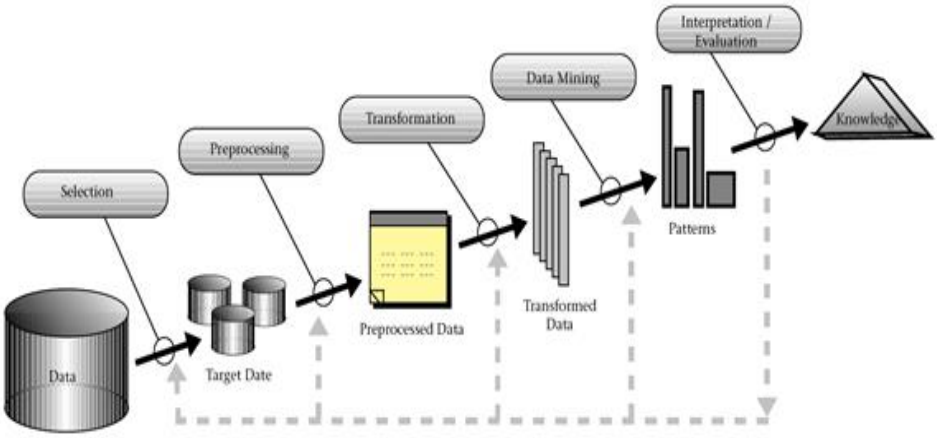


**Figura 2.** Esquema del Descubrimiento de Conocimiento en Base de Datos.
Fuente: Maribell 2010.

A continuación, se detalla groso modo las actividades desarrolladas en cada una de las fases de esta metodología.

**Abstracción del escenario**: Para entender la problemática y entender el contexto se analizaron las bases de datos obtenidas del Observatorio del Delito(Policía Nacional de Colombia, 2015) y se tomó como escenario de estudio la zona urbana del municipio de Tumaco.

**Selección de los datos**: Como se mencionó anteriormente la información se extrajo del Observatorio del Delito(Policía Nacional de Colombia, 2015), dado a que esta dependencia es la encargada de la investigación en criminología, consolidando, procesando y difundiendo los registros administrativos con fines estadísticos de delitos. Se descargaron 45 bases de datos, para los años comprendidos desde el 2010 al 2019, una vez descargada dicha información, se dejó las variables coincidentes por delito, se consolidaron las bases de datos(Toledo, 2016), y se filtró para dejar solo la información del municipio de Tumaco. Para realizar un mejor tratamiento de lo obtenido, se creó una base de datos con todos los barrios de la zona urbana del municipio de Tumaco con sus respectivos atributos geográficos, "latitud"," longitud" y por otro lado se consiguió el polígono del municipio de Tumaco, (Shapefile de la zona urbana de Tumaco).

**Limpieza y preprocesamiento**: En esta etapa se determinó la confiabilidad de la información, para esto se transformaron y eliminaron varias columnas, entre ellas la variable "fecha" la cual se transformó por, "día", "mes" y "hora", la variable "hora" por "día", "tarde" y "noche" y también se quitaron otras variables, dado a que no contaban con calidad en su distribución de datos, como lo son las variables "departamento", "municipio", "zona", "código DANE" y "delito", puesto que estas variables no brindaban información necesaria porque sus datos eran repetitivos, carecían de variabilidad y calidad.

**Transformación de los datos**: En esta etapa se mejoró la calidad de los datos con transformaciones que involucraron convertir toda la información de la base de datos, para ello se obtuvieron las coordenadas geográficas "latitud", "longitud" de cada uno de los barrios de la zona urbana del municipio de Tumaco y para una mejor comprensión de los datos se trabajó con coordenadas 3857, ya que es más favorable para obtener métricas de similitud y distancias entre los vectores de características trabajando en metros. Para la aplicación del algoritmo Kmeans se codificaron las variables categóricas y se normalizaron los datos.

**Minería de Datos**: En esta fase de minería de datos se eligieron algoritmos de regresión y agrupación (clustering) a aplicar. Una vez seleccionadas las tareas de minería de datos, se entrenaron los modelos sintonizando manualmente los hiperparametros, a saber se utilizó; KNN(F., n.d.), SVM, Arboles de decisión, Redes Neuronales(C. R. T. en el departamento de G. D. C. de T. G. where the puck is going to be, 2020), Random Forest, Procesos gaussianos y Kriging como algoritmos de regresión. Por otro lado, se experimentó con Kmeans y K-prototype (Kmeans más K Medias) como algoritmos de agrupación.

**Evaluación**: Para evaluar los algoritmos de regresión se compararon los coeficientes de determinación con datos de entrenamiento y de prueba, para luego contrastar si la métrica incidía en la calidad de la visualización. Luego para clustering, se examinó mediante la técnica del codo y la silueta que cantidad de clusters era la adecuada y posteriormente, se tomó de los mejores modelos los que brindaban el mejor balance entre grupos.

**Interpretación**: Finalmente se interpretaron los resultados obtenidos comparando sus métricas de calidad y respectivas visualizaciones.

## Resultados

En este estudio se tiene como resultado la base de datos de delitos para toda Colombia para los años 2010 a 2019; scripts de transformación y limpieza de la cual se obtienen las bases de datos de entrenamiento tanto para las técnicas supervisadas como para las no supervisadas(Toledo, 2016); igualmente se tiene los mapas de delitos, métricas de calidad, los hiperparámetros con los que se obtuvieron los mejores resultados por algoritmo, como también las agrupaciones y mapas de agrupaciones logrados con los algoritmos de agrupación.

**Mapas de delitos**: Para la obtención del mapa que logro una buena aproximación de la actividad delictiva interpolando los delitos conocidos sobre puntos cercanos, se normalizaron los datos geográficos y se obtuvo el r2 para cada algoritmo en estudio como lo indica la Tabla 3, donde KNN es K Vecinos Cercanos, SVR Maquinas de soporte Vectorial para regresión, MLPR Perceptron Multi Capa para regresión, GP Procesos Gaussianos, TD Árbol de decisión, RF Bosques Aleatorios y KO Kriging Ordinario.

| **Algoritmos** | **Hyperparámetros** | **$r^2$ prueba** | **$r^2$ entrenamiento** |
|---|---|---|---|
| **KNN** | Vecinos=8 | 0.41 | 0.40 |
| **SVR** | Kernel=RBF<br>Gamma=20<br>Cte Regularización=100 | 0.40 | 0.25 |
| **MLPR** | Iteraciones=100K<br>Capas ocultas=4<br>Topología=64,32,8,16 | 0.42 | 0.31 |
| **GP** | Kernel=RBF | 0.36 | 1 |
| **TD** | Profundidad = 4 | 0.39 | 0.37 |
| **RF** | Profundidad=4<br>Arboles=50 | 0.41 | 0.39 |
| **KO** | Variograma=hole-effect | **0.85** | **0.87** |

Tabla 3. Métricas $r^2$
Fuente: Esta investigación

Una vez obtenidas las métricas de $r^2$, las que más se ajustaron al modelo de datos de entrenamiento fue la del algoritmo Kriging Ordinario, con un coeficiente de determinación de 0.85 con datos de prueba y 0.87 con datos de entrenamiento.

En la Figura 3 se puede apreciar la visualización de la interpolación de los primeros 6 algoritmos de la Tabla 3, esto se obtuvo desde una malla de puntos con resolución de 50 metros, estructurados desde los vértices dados por los puntos máximos y mínimos de las latitudes y longitudes del polígono del municipio de Tumaco, de los cuales solo se tomaron aquellos puntos que se encuentran dentro del polígono de datos. El algoritmo de Kriging ordinario obtuvo el mejor compromiso en sus métricas de calidad y su visualización puede apreciarse en la Figura 4.

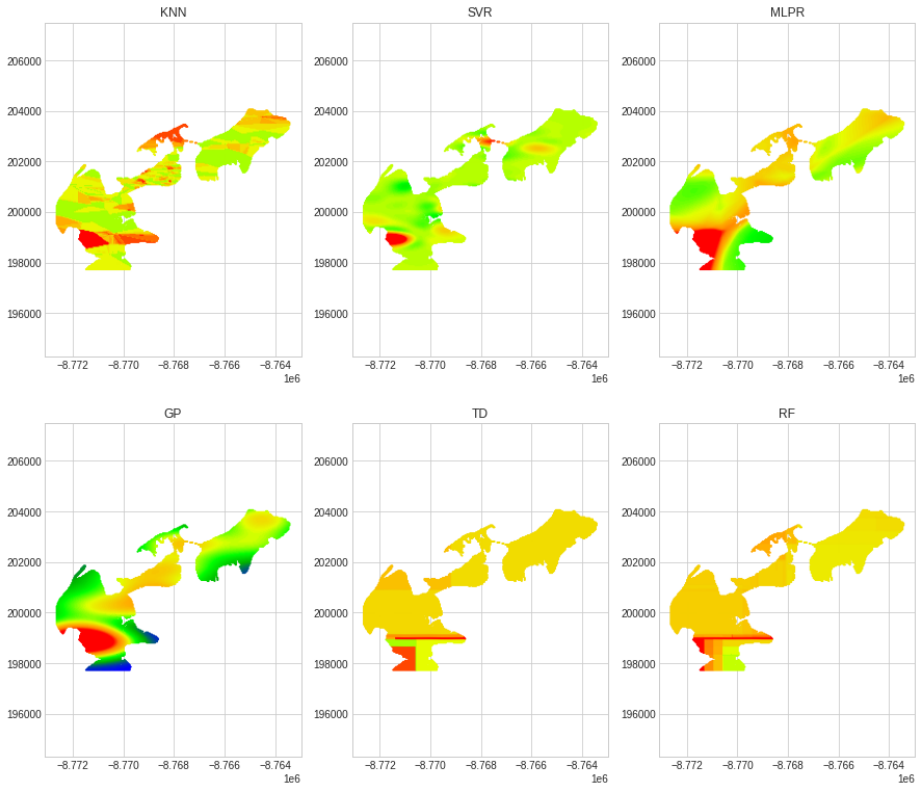


Figura 3. Interpolación de malla de puntos en los algoritmos aplicados.
Fuente: esta investigación.

En cuanto a la implementación del algoritmo Kriging, el cual es un método de interpolación geoestadístico de estimación de puntos que utiliza un modelo de Variograma para obtener ponderadores que se dan a cada punto de referencia usado en la estimación, cabe resaltar que este algoritmo es el que recomienda la literatura para poder realizar interpolaciones geográficas, así entonces, se inició a construir el Variograma (herramienta que permite analizar el comportamiento).

Al aplicar Kriging se obtuvieron resultados favorables en cuanto a la manipulación de los datos delictivos de entrenamientos manejados en este estudio, por lo tanto los resultados obtenidos están basados en la realidad, la ubicación de los puntos o lugares identificados de color rojo el cual indica que hay mayor actividad delictiva en dicho lugar, coincidiendo con los datos reportados en el Observatorio del Delito(Policía Nacional de Colombia, 2015), de igual manera el color naranja que hace referencia al lugar donde se cometen actividades delictivas de una manera moderada pero no es un lugar altamente riesgoso para transitar.

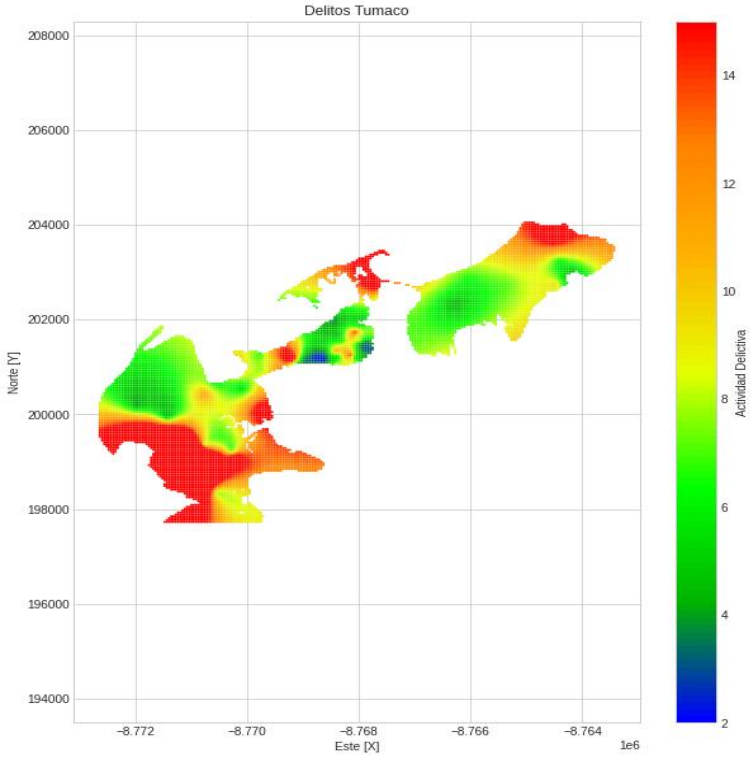


Figura 4. Interpolación de Kriging.
Fuente: esta investigación.

Una vez obtenidos los mapas de interpolaciones mencionados anteriormente, se procedió a la creación de 3 grupos de Clustering cuya cantidad de grupos se obtuvo utilizando la técnica del codo y silueta para K-Means (Figura 3 A y B) y codo para K-Prototype (Figura 3 C).

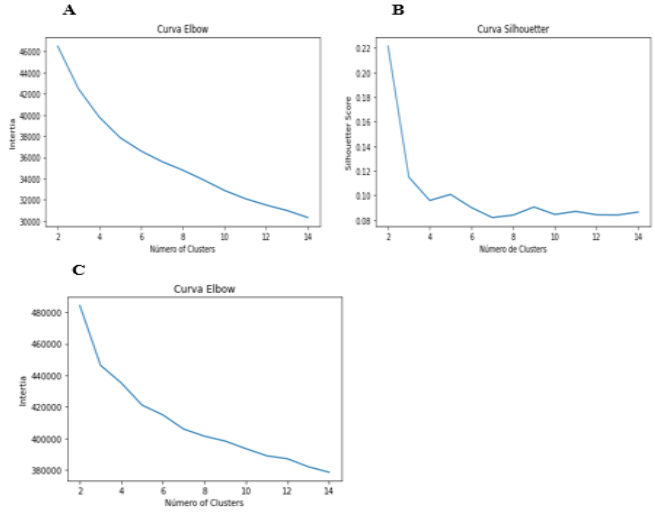


Figura 5. Curvas Codo y Silueta.
Fuente: esta investigación.

De la Figura 5 se optó por construir grupos de 2 y 3 para K-Means y 3 para K-Prototype; el que mejor balance obtuvo en sus agrupaciones fue K-prototype, como se muestra en la figura 6, indicada por la gráfica “C”, donde se observa que las agrupaciones están mejor balanceadas que el modelo K-Means (ver Figura 6 A y B).

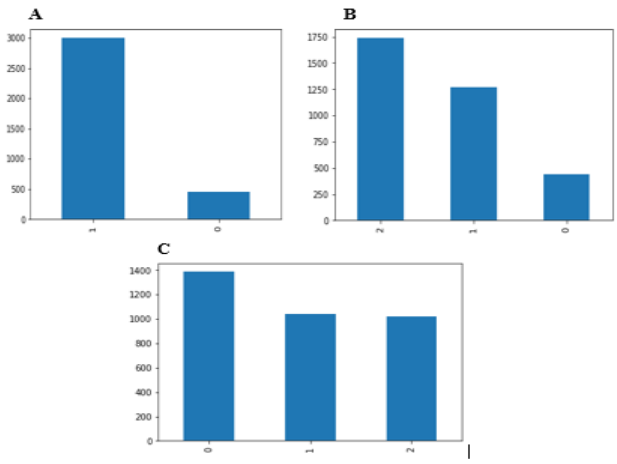


Figura 6. Grupos de clustering-kmean y k-prototype.
Fuente: esta investigación.

Como se observa en la Figura 6 las agrupaciones realizadas con 3 grupos tanto para K-Means (Figura 6 B), como para K-Prototype (Figura 6 C) ofrecen el mejor balance en sus grupos. A continuación, se describen los patrones encontrados en estos dos modelos.

**Interpretación con Kmeans**

Para realizar la interpretación sobre este algoritmo, primero se hayo los puntos más cercanos a los centroides encontrados y se decodificó su información. De ahí se obtuvo que la actividad delictiva se concentra en los homicidios, al rededor del segundo cuatrimestre del año en los días martes en horas de la madrugada y noche (6pm a 6am).

El cluster 0 tiene el 13% de los datos y agrupa a aquellos delitos cometidos al rededor del barrio la Carbonera en las coordenadas (1.80709,-78.7671), a empleados públicos hombres separados, con estudios de secundaria, de 35 a 40 años de edad, generalmente permutados con granada de mano, con desplazamiento no definido tanto para la víctima como agresor.

El cluster 1 tiene el 37% de los datos y agrupa a aquellos delitos cometidos al rededor del barrio el Triunfo en las coordenadas (1.80911,-78.7616), a empleadas particulares mujeres separadas, con estudios superiores, de 40 a 45 años de edad, generalmente sin reportar arma, con desplazamiento a pie tanto para la víctima como agresor.

El cluster 2 tiene el 50% de los datos y agrupa a aquellos delitos cometidos al rededor del barrio Libertad en las coordenadas (1.80069,-78.7832), a empleados públicos hombres solteros, con estudios de secundaria, de 30 a 35 años de edad, generalmente sin reportar arma, con desplazamiento a pie tanto para la víctima como agresor.

**Interpretación K-Prototype**

En este algoritmo se encontró un mejor balance en sus agrupaciones y arrojaron que la actividad delictiva se concentra en victimas solteras, que transitan a pie al igual que el agresor.

El cluster 0 tiene el 37% de los datos y agrupa a aquellas amenazas concentradas en el barrio Obrero en las coordenadas (1.799773500389714, -78.78134212860543), en el mes de julio los días lunes en la tarde, a empleados particulares hombres, con estudios de secundaria, de 35 a 40 años de edad, generalmente perpetuados con arma de fuego.

El cluster 1 tiene el 32% de los datos y agrupa a aquellos homicidios concentrados en el barrio el Bajito en las coordenadas (1.8066203929236473,-78.77252556424568), en el mes de diciembre los días domingo en la noche, a empleados independientes hombres, con estudios de primaria, de 25 a 30 años de edad, generalmente perpetuados con arma de fuego.

El cluster 2 tiene el 31% de los datos y agrupa a aquellos hurtos concentrados en el barrio Calle del Comercio en las coordenadas (1.8083773912248629, -78.76403454753219), en el mes de agosto los días martes en el día, a empleadas particulares mujeres, con estudios de secundaria, de 35 a 40 años de edad, generalmente sin emplear armas.

Como esta agrupación arrojó el mejor balance en sus grupos se procedió a la generación del mapa de agrupaciones como se visualiza en la Figura 7.

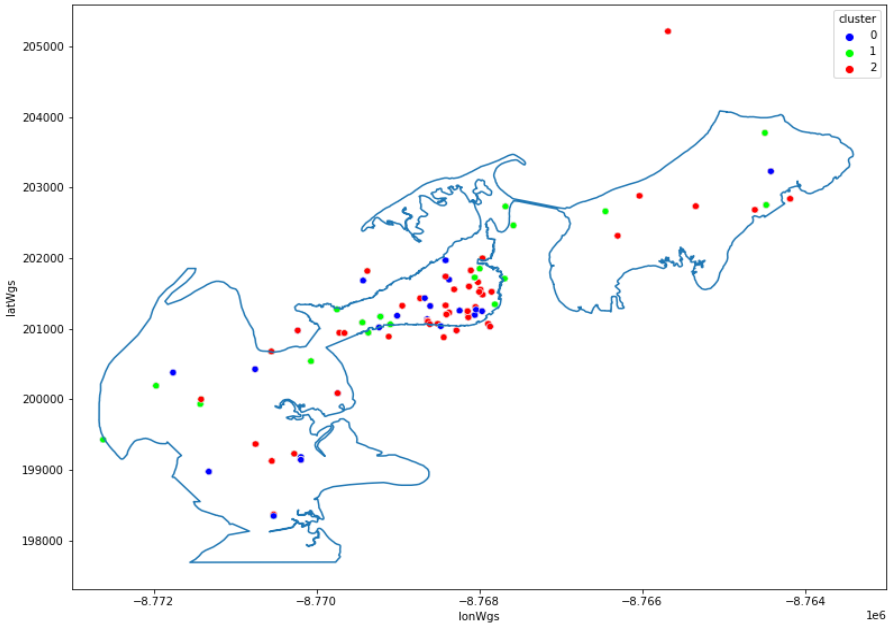


Figura 7. Mapa de agrupaciones.
Fuente: esta investigación.

## Discusión y conclusiones

Como resultado de las fuentes de extracción se obtuvieron los datos correspondientes a cada una de las variables definidas, contando con el Observatorio del Delito Colombiano como la fuente principal para el desarrollo de la investigación, sin embargo los datos de la fuente inicialmente, no contaban con suficientes registros claros y específicos para la aplicación de minería de datos, la cual es una de las etapas importantes de la metodología aplicada (KDD), para ello es de gran importancia cruzar las bases de datos con el fin consolidar una sola y proceder al análisis de los datos que serán de utilidad para dicho estudio, por ello fue de gran importancia aplicar limpieza y preprocesamiento de los datos, logrando eliminar y dividir las columnas innecesarias para realizar un mejor tratamiento de los datos. En cuanto la variable tipo “fecha” la cual se dividió por “día”, “mes” y “año” , se obtuvieron muy buenos patrones frente actividad delictiva, siendo de gran utilidad para el desarrollo de este estudio(Valenga et al., 2016).

En cuanto a los mapas de delitos, se observa que corresponden a la realidad actual que presenta el municipio de Tumaco frente a los casos de actividad delictiva, ya que contrastan con los datos que no se habían visto antes de la experimentación.

Se logra hacer una mejor agrupación de la actividad delictiva con el modelo K-prototype, ya que este modelo ofrece una agrupación que tiene un mejor balance y se ajustan mucho más a la realidad que el algoritmo K-means.

La aplicación de interpolación Kriging(gabri, 2018), para el desarrollo de esta investigación fue la más favorable ya que el procedimiento geoestadístico(BSG Institute, n.d.) avanzado, generó una superficie estimada a partir de un conjunto de puntos de actividad delictiva dispersados con valores z, obteniendo así una auto correlación, es decir, las relaciones estadísticas entre los puntos medidos. Gracias a esto, las técnicas de estadística geográfica no sólo tienen la capacidad de producir una superficie de predicción, sino que también proporcionan alguna medida de certeza o precisión de las predicciones(*Definición de Predicción — Definicion.De*, n.d.).

Como trabajos futuros se podrían adicionar una comparación de algoritmos de agrupación utilizando técnicas de agrupación jerárquicas; cruzar la información del departamento del delito con las de otros repositorios como la fiscalía general de la nación, defensoría del pueblo(*Defensoría Del Pueblo*, 2022); crear una herramienta para él apoyó a la actividad delictiva basada en recomendaciones sobre el perfil de usuario y movilidad; realizar un análisis de series de tiempo que permita predecir el porcentaje de delitos a partir de una serie de imágenes de mapas temporales(Pérez & Luis, 2014).

Otro de los trabajos futuros que pueden ser considerados para dar a conocer tanto la problemática como los modelos utilizados es aplicar las tecnologías asociadas con los Ambientes Virtuales de Aprendizaje como se manifiesta en (Jiménez et al., 2020).

## Referencias bibliográficas


Alfredo. (2017). ¡Descubriendo información en bases de datos!

Arsys. (2019). ¿Qué es Machine Learning y por qué es tan importante?

BBC News Mundo. (n.d.). Por qué América Latina es la región más violenta del mundo (y qué lecciones puede tomar de la historia de Europa).

BSG Institute. (n.d.). Geoestadística: Variograma Experimental.

C. R. T. en el departamento de G. D. C. de T. G. where the puck is going to be. (2020). ¿Qué son las redes neuronales artificiales y cómo funcionan? - Thinkbig.

Charte, F. (n.d.). Cómo es el proceso de extraer conocimiento a partir de bases de datos.

Data, M. (2022). KDD: Knowledge Discovery in Databases.

Defensoría del Pueblo. (2022).

Definición de predicción — Definicion.de. (n.d.).

Diario Del Sur. (2022).

En Tumaco no cede criminalidad. (n.d.).

F., S. (n.d.). NxKNN - K Barrios más cercanos (K-NN ) Regresión.FIP - Ideas para la paz. (2022). Dinámicas del conflicto armado en Tumaco y su impacto humanitario.

Gabri. (2018). Geoestadística, interpolación con Kriging.

Galindo, L. M., & Catalán, Y. H. (2007). Las actividades delictivas en el Distrito Federal. Revista Mexicana de Sociología, 29.

Interpolacion y Extrapolacion – julianapinzon. (n.d.).

Jiménez, J., Muñoz, A., Ramos, D., Muñoz, J., Gonzales, C., Guerrero, H., Santacruz, F., Leyton, G., & Eraso, A. (2020). Las tecnologías de las redes y los AVA. Editorial Universidad CESMAG. https://doi.org/10.15658/CESMAG20.12010129

López, B. R. (2020). Librerías de Python para trabajar con datos espaciales.

Los datos lo demuestran: Tumaco no sabe qué es el posconflicto. (n.d.).

Minera, N. (2019). PRUEBA DEL KIT PYKRIGE EN PYTHON.
Modeling reality generating software. (2020). Machine Learning Algorithms and Techniques.
Pérez, R., & Luis, J. (2014). Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. Revista Cubana de Ciencias Informáticas, 8(4), 52–73.
Policía Nacional de Colombia. (2015). Observatorio del Delito de la Policía Nacional.
¿Qué es el error cuadrático medio RMSE? | El blog de franz. (n.d.).
Tecnicas-de-mineria-de-datos-para-la-prevencion-de. (n.d.).
Toledo, A. (2016). Métodos de selección de atributos para clasificación supervisada basados en teoría de información.
Valenga, F., Perversi, I., Fernández, E., Merlino, H., Rodríguez, D., & Britos, P. (2016). APLICACION DE MINERIA DE DATOS PARA LA EXPLORACION Y DETECCION DE PATRONES DELICTIVOS EN ARGENTINA. 13.