“SORLOCK HOLMES” PREDICCIÓN DE ACTIVIDAD DELICTIVA EN EL MUNICIPIO DE TUMACO MEDIANTE TÉCNICAS DE MACHINE LEARNING

ANYELA SORANYI ALEGRIA CAMPAZ

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
SAN ANDRÉS DE TUMACO
2022

“SORLOCK HOLMES” PREDICCIÓN DE ACTIVIDAD DELICTIVA EN EL MUNICIPIO DE TUMACO MEDIANTE TÉCNICAS DE MACHINE LEARNING

ALEGRIA CAMPAZ ANYELA SORANYI

Trabajo de Investigación presentado como requisito para optar por el título de Ingeniero de Sistemas

Asesor
Ing. Héctor Andrés Mora
Docente Departamento de Sistemas

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
SAN ANDRÉS DE TUMACO
2022

NOTA DE RESPONSABILIDAD

“Las ideas y conclusiones aportadas en este Trabajo de Grado son responsabilidad exclusiva de los autores”.

Artículo 1° del Acuerdo No. 324 de octubre 11 de 1966, emanado del Honorable Consejo Directivo de la Universidad de Nariño.

Nota de aceptación

_____
_____
_____
_____
_____
_____
_____

_____
Firma del presidente del Jurado

_____
Firma del Jurado

_____
Firma del Jurado

San Andrés de Tumaco, febrero de 2022

Acuerdo No. 10 de 11 de febrero de 2022 de Consejo de Facultad de Ingeniería

ACUERDO No. 010
(11 de febrero de 2022)

EL CONSEJO DE FACULTAD DE INGENIERIA
DE LA UNIVERSIDAD DE NARIÑO

En uso de sus atribuciones reglamentarias, estatutarias y,

CONSIDERANDO

Que mediante Acuerdo No. 077 del 10 de diciembre del 2019, el Consejo Académico estableció y unificó la normatividad de los Trabajos de Grado de la Universidad de Nariño.

Que en el Artículo 9 del Acuerdo 077 del 10 de diciembre de 2019, el Consejo Académico de la Universidad de Nariño faculta a los comités curriculares de cada programa para definir y reglamentar los siguientes puntos relacionados con los Trabajos de Grado: a) Criterios de evaluación; b) Criterios para la selección de diplomados y créditos de postgrado; c) Número de estudiantes por Trabajo de Grado; d) Procedimiento de inscripción; e) Aspectos que debe contener cada modalidad de Trabajo de Grado; f) Jurados de evaluación; g) Presentación, sustentación y prerrequisitos; h) Criterios de inter y transdisciplinariedad; i) Consideraciones éticas.

Que el Comité Curricular del Departamento de Sistemas, junto con el estamento docente, realizaron sesiones conjuntas que permitieron definir, reglamentar y hacer operativas las modalidades de trabajos de grado acorde con su fundamento disciplinar, acogiéndose a la reglamentación más reciente que es el Acuerdo 077 de 2019 emitido por el Consejo Académico.

Que el cambio en la reglamentación del trabajo de grado no implica una modificación al Estatuto Estudiantil y la aprobación del presente Acuerdo no genera modificación de la normatividad existente, sino que por el contrario lo reglamenta, estableciendo lineamientos claros y definidos para las modalidades acogidas por el Departamento de Sistemas, en lo concerniente a la normatividad relacionada a la presentación, elaboración, seguimiento y calificación de los Trabajos de Grado.

Que el Comité Curricular del Programa de Ingeniería de Sistemas, mediante el Acuerdo No. 095 del 19 de agosto de 2021, aprueba el reglamento para la presentación, aprobación, desarrollo, sustentación, socialización y evaluación de trabajos de grado en las diferentes modalidades de graduación para los programas de pregrado adscritos al Departamento de Sistemas.

Que el Comité Curricular del Programa de Ingeniería de Sistemas, mediante el Acuerdo No. 095 del 19 de agosto de 2021, en el Capítulo 9 articulo 61 define el reconocimiento de las Distinciones para trabajos de grado cuando se realicen bajo la modalidad de Investigación o

ACUERDO No. 010
(11 de febrero de 2022)

Interacción Social, las cuales serán de Meritorio cuando obtengan una calificación entre 90 a 99 puntos y Laureado para 100 puntos; estos serán asignados si son dignos de presentarse a la comunidad académica en nombre de la Universidad de Nariño, por su aporte original en el campo de las ciencias, la tecnología, las humanidades, las artes o la pedagogía.

- El Consejo de Facultad será el responsable de decidir sobre dichas distinciones y podrá otorgarlas previa presentación de la proposición correspondiente por parte del Comité Curricular, en la cual se adjunte un informe elaborado por los jurados evaluadores que justifique dicho merecimiento.
- Para el otorgamiento de la distinción de Laureado, el Consejo de Facultad podrá solicitar nuevos conceptos de profesionales distintos a los jurados evaluadores.

Que el Comité Curricular del Programa de Ingeniería de Sistemas, mediante Proposición No. 003 de 27 de enero de 2022, propone al Consejo de Facultad: otorgar la mención de Laureado, al Trabajo de Grado: "SORLOCK HOLMES" predicción de actividad delictiva en el municipio de Tumaco mediante técnicas de Machine Learning, Modalidad Investigación, presentado por la estudiante Anyela Soranyi Alegría Campaz, bajo la dirección del Ingeniero Héctor Andrés Mora.

Que en Proposición No. 003 comunica que los ingenieros Camilo Lagos Mora y Álvaro Ricardo Cujar Rosero, recomiendan otorgar Mención de LAUREADO, al Trabajo de Grado presentado por la estudiante Anyela Soranyi Alegría Campaz, teniendo en cuenta lo siguiente:

- Solidez investigativa. El proceso investigativo realizado por la estudiante es riguroso, objetivo, válido, verificable e incluye publicaciones en eventos académicos como el 15 Congreso Colombiano en Computación - 15CCC y el Congreso Andino en Computación, Informática y Educación CACIED, para el 2021.
- Pertinencia social. Los datos utilizados en el proyecto son datos históricos de delitos en el municipio de Tumaco, los cuales fueron tratados en el desarrollo de esta investigación y que permitieron obtener mapas de calor que muestran cómo están distribuidos estos delitos en el municipio e inferir el comportamiento de los mismos y con una correcta utilización de por parte de los organismos gubernamentales respectivos el aporte sería de suma importancia toda vez que permite focalizar las zonas más vulnerables de este flagelo y concentrar las acciones gubernamentales en estas zonas y/o tomar las decisiones más acertadas en lo que a este flagelo respecta.
- Innovación. Mediante la aplicación de técnicas de minería de datos se obtiene conocimiento relevante que permitirá ser utilizado en aras del bienestar social.
- La temática trabajada tiene un alto grado de profundidad y se estudian y comparan diversos algoritmos de minería de datos de gran interés en la actualidad, adicionalmente

ACUERDO No. 010
(11 de febrero de 2022)

se seleccionan datos con un alto componente social para sus pruebas y cuyos resultados pueden impactar de forma positiva en la comunidad.

Que los Jurados otorgaron una calificación de (100) puntos al Trabajo de Grado.

Que el Trabajo de Grado cumple con las condiciones necesarias para otorgar la distinción solicitada de Tesis LAUREADO.

Que el Consejo de Facultad mediante sesión de 08 de febrero de 2022 conoció y avaló la Proposición 003 de 27 de enero de 2022 debido a la solidez investigativa, aporte social, innovación y el alto grado de profundidad de la temática debido a los diversos algoritmos de minería de datos que se utilizaron para el desarrollo del trabajo de grado.

En virtud de lo anterior, el Consejo de Facultad,

ACUERDA

ARTÍCULO 1°. Otorgar la mención de LAUREADO, al Trabajo de Grado: "SORLOCK HOLMES" predicción de actividad delictiva en el municipio de Tumaco mediante técnicas de Machine Learning, modalidad Investigación, presentado por la estudiante Anyela Soranyi Alegría Campaz, bajo la dirección del Ingeniero Héctor Andrés Mora.

ARTÍCULO 2°. Notificar del presente Acuerdo al estudiante Anyela Soranyi Alegría Campaz por medios electrónicos.

ARTICULO 3° Facultad de Ingeniería, OCARA, Departamento de Sistemas, anotarán lo de su cargo.

COMUNÍQUESE Y CÚMPLASE

Dado en San Juan de Pasto, a los once (11) días del mes de febrero de dos mil veintidós (2022)

ALEXANDER BARON SALAZAR
Presidente

SANDRA ORDOÑEZ CERÓN
Secretaria Académica

Aprobó: Consejo de Facultad

AGRADECIMIENTOS

A Dios ya que gracias a él he logrado concluir mi carrera.

A la Universidad de Nariño por haberme aceptado ser parte de ella y abierto sus puertas para formarme como profesional, así como también a los diferentes docentes que brindaron sus conocimientos y su apoyo para seguir adelante día a día.

A mi asesor Mg . Héctor Mora, por haberme brindado la oportunidad de recurrir a su capacidad y conocimiento científico, así como también haberme tenido toda la paciencia del mundo para guiarme durante todo el desarrollo de la tesis.

A mi madre en especial, mi mayor motivación y ejemplo de vida, quien ha estado en todo este proceso como el más fuerte pilar para el logro de mis objetivos.

A mi familia, en especial a mis hermanos Diego Alegría, John Eduar, Cristian Raúl y mi tía Marisol Campaz quienes han creído en mí siempre, dándome ejemplo de superación, humildad y sacrificio.

A mi esposo Edilson Cáceres por sus palabras y confianzas, por su amor y brindarme el tiempo necesario para realizarme profesionalmente.
Y para finalizar agradezco a mis dos amigas Leisy y Erika por su apoyo incondicional en este proceso.

DEDICATORIA

Dedico con todo mi corazón mi tesis a mi abuela Cecilia Calle, que, aunque ya no se encuentre en este mundo, fue una de las más orgullosas porque yo lograra ser una profesional, te amare por la eternidad mi vieja.

A mi madre Jenny María Campaz, por haber forjado en mí la mujer que ahora soy.

A mi hijo Edhian Josué, quien es mi mayor motivación por ser mejor cada día, y darle un mejor ejemplo de vida y poder ofrecerle un mejor futuro.
A mi esposo Edilson Cáceres, quien ha sido mi compañero en todo este proceso.

A mis hermanos Diego Alegría, John Eduar, Cristian Raúl, quienes están muy orgulloso de mi y son partes de la concepción de esta meta, la cual es el logro de todo el apoyo incondicional de una familia.

El éxito no es un accidente, es trabajo duro, perseverancia, aprendizaje, estudio, sacrificio y sobre todo amar lo que estás haciendo.
(Pelé)

RESUMEN

En la actualidad existen pocos estudios comparativos de algoritmos supervisados y no supervisados de machine learning, los cuales puedan ser aplicados para dar soluciones a problemas de carácter social, político, económico, etc; para de tal manera poder crear herramientas enfocadas en no solo para zonificar y agrupar delitos a través de los datos provenientes del Observatorio del Delito Colombiano, que ayuden en el proceso de investigación e identificación puntual de los delitos y de quienes los cometen, siendo este un factor muy desfavorable para contribuir en los planes de mitigación de delincuencia, si no también poder abarcar diversos problemas de carácter social, para lo cual es de gran importancia la aplicación de un estudio comparativo de machine learning para la obtención de un modelo de regresión y agrupación, aplicando descubrimiento de conocimiento en bases de datos (KDD) en históricos de actividad delictiva en el municipio de Tumaco, con el fin de brindar una herramienta que aporte a la investigación y permita ayudar en los planes de contingencia de los entes de control frente a los diferentes tipos de problemas de actividad delictiva.

Se relaciona la metodología KDD la cual consiste en una serie de pasos que permiten el desarrollo de dicha investigación. Primero, consiste en la abstracción del escenario, 2) selección de datos, 3) limpieza y preprocesamiento, 4) transformación de los datos, 5) elección de tareas de Minería de Datos, 6) elección del algoritmo, 7) aplicación del algoritmo, 8) evaluación e interpretación y 9) entendimiento del conocimiento.

Finalmente, cabe aclarar que las variables que se contemplan para el desarrollo, son datos que no son sensibles y que algunas de ellas tampoco son suministradas por las diferentes fuentes de información con el fin de mantener la seguridad e integridad de los datos frente a los diferentes casos de actividad delictiva.

Palabras claves: Algoritmos, delitos, datos, KDD, machine learning.

ABSTRACT

At present, there are few comparative studies of supervised and unsupervised machine learning algorithms, which can be applied to provide solutions to social, political, economic, etc. problems, in order to create tools focused not only on zoning and grouping crimes, using data from the Colombian Crime Observatory, which can help in the process of research and timely identification of crimes. crime and those who commit it, which is a very disadvantageous factor for contributing to crime-mitigation plans, if it is not also possible to cover various social problems, for which the use of a comparative machine-learning study to obtain a regression and clustering model is of great importance; the applying a comparative study of Machine Learning for the obtaining a regression and grouping model, applying Knowledge Disco-very in Databases (KDD) methodology over historical criminal activity in the Tumaco city, in this way it could be helped more successfully in the contingency plans of control entities against the different types of criminal activity.

The background of the study at the global and national level is presented, taking as sources books, publications, among others, which allow justifying many of the concepts covered during the research process. Next, the KDD methodology is related, which consists of a series of steps that allow the development of this research. First, it consists of the abstraction of the scenario, 2) data selection, 3) cleaning and preprocessing, 4) data transformation, 5) choice of Data Mining tasks, 6) choice of algorithm, 7) algorithm application, 8) evaluation and interpretation and 9) knowledge understanding.

Finally, it should be clarified that the variables contemplated for the development are non-sensitive data and that some of them are not provided by the different sources of information in order to maintain the security and integrity of the data in the face of the different cases of criminal activity.

Keywords: Algorithms, crimes, data, KDD, machine learning.

TABLA DE CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

Pag.

LISTA DE ECUACIONES

pág.

LISTA DE ANEXOS

Pag.

GLOSARIO

ÁRBOL DE DECISIÓN: es un mapa de los posibles resultados de una serie de decisiones relacionadas. Permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios.

ALGORITMO K-MEANS: K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamientOo se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática**.**

CLUSTERING: el Clustering es una tarea que consiste en agrupar un conjunto de objetos (no etiquetados) en subconjuntos de objetos llamados clúster. Cada clúster está formado por una colección de objetos que son similares (o se consideran similares) entre sí, pero que son distintos respecto a los objetos de otros clúster.[1]

EXTRAPOLACIÓN: el método de extrapolación es un método científico lógico que consiste en suponer que el curso de los acontecimientos continuará en el futuro, convirtiéndose en las reglas que se utilizarán para llegar a una nueva conclusión. Es decir, se afirma a ciencia cierta que existen unos axiomas y éstos son extrapolables a la nueva situación.[2]

ESTIMADOR INSESGADO: un estimador insesgado es aquel cuya esperanza matemática coincide con el valor del parámetro que sea desea estimar. En caso de no coincidir se dice que el estimador tiene sesgo.

GEORREFERENCIACIÓN: es la utilización de coordenadas de mapa para determinar una ubicación en el espacio a las diferentes entidades cartográficas. Todos los componentes de una capa de mapa poseen una ubicación geográfica y una extensión concretas que permiten emplazarlos en la superficie de la Tierra o próxima a ella.[3]

---

[1] MOYA, Ricardo: que es el Clustering. [en linea].(25 de Marzo de 2016). Obtenido de https://www.jarroba.com/que-es-el-clustering/

[2] PASCUZZO, Alda Extrapolación. [en linea].(11 de mayo de 2013). Obtenido de http://aldanalisis.blogspot.com/2013/05/extrapolacion.html

[3] LORENA. Georreferenciacion. [en linea].(12 de abril de 2018). Obtenido de https://www.certicalia.com/blog/georreferenciacion-que-es-y-para-que-se-utiliza

INTERPOLACIÓN: la interpolación es un proceso que utiliza mediciones realizadas sobre algún fenómeno (precipitación, temperatura o elevación) en determinados lugares, para hacer una predicción sobre un fenómeno en otros lugares donde no se han realizado mediciones.[4]

IMAGEN RÁSTER: un ráster consta de una matriz de celdas (o píxeles) organizadas en filas y columnas (o una cuadrícula) en la que cada celda contiene un valor que representa información, como la temperatura.

KDD (PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO): es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos.

KRIGING: es un método geoestadístico de interpolación que ha probado ser útil y popular en muchos campos.

K-PROTOTYPES: es un hermano menos conocido, pero ofrece la ventaja de trabajar con tipos de datos mixtos. Mide la distancia entre características numéricas usando la distancia euclidiana (como K-medias) pero también mide la distancia entre características categóricas usando el número de categorías coincidentes.

MACHINE LEARNING: es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos.

OBSERVATORIO DEL DELITO: es un grupo estratégico del área de Investigación Criminológica, encargado del monitoreo, diagnóstico, administración de la información, evaluación y análisis de la criminalidad.
PROYECCIÓN: se denomina mención al accionar y a los resultados de proyectar**,** es decir provocar el reflejo de una imagen ampliada en una superficie, lograr que la figura de un objeto se vuelva visible sobre otro, desarrollar una planificación para conseguir algo

REDES NEURONALES ARTIFICIALES: son un modelo inspirado en el funcionamiento del cerebro humano.

---

[4] FRANZPC. que es el Interpolación [en linea].(14 de Marzo de 2021). Obtenido de https://acolita.com/author/franzpc/

SOBREAJUSTE: se denomina sobreajuste al hecho de hacer un modelo tan ajustado a los datos de entrenamiento que haga que no generalice bien a los datos de test.

SUBREGISTRO (TÉRMINO ESTADÍSTICOS): diferencia entre el número de defunciones estimadas y el número de defunciones efectivamente registradas, expresada como porcentaje del total de defunciones estimadas, para un año dado, en un determinado país, territorio o área geográfica.

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 agro-industrial prometedor. Donde no suelen realizarse estudios comparativos de algoritmos ya sean supervisado o no supervisados con el fin de dar soluciones a problemáticas que presenta dicho municipio de carácter social, político u económico. Dado a ello siertos factores se ven opacados frente a 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"[5].

Es así como la amenaza y la extorsión, además de ser fuente de ingresos para los grupos armados ilegales, es 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 coronadas en muchos casos, con la desaparición y el homicidio. Toda esta violencia termina por desestructurar y disolver los procesos organizativos, dejando a las comunidades sumidas en la zozobra y la incertidumbre[6]; con el fin de contribuir en los planes de mitigación de delincuencia, se plantea obtener un modelo de predicción de actividad delictiva realizando un comparativo de algoritmos supervisados y de agrupación de machine learning, aplicando la metodología de descubrimiento de conocimiento en bases de datos (KDD) con históricos de actividad delictiva recopilados por el departamento del delito.

Creando de este modo una sinergia hombre máquina que asista y refuerce mediante inteligencia artificial a los mecanismos de inteligencia de las fuerzas militares.

[5] DEFENSORIA.COM "Informe especial: economías ilegales, actores armados y nuevos escenarios de riesgo en el posacuerdo,"; (Defensoría del Pueblo, 2017b, p. 183). Septiembre 2018. [Online]. Available: https://www.defensoria.gov.co/public/pdf/economiasilegales.pdf

[6] Ibid.

Siendo este un medio para contribuir en la disminución de problemas de carácter social, político, económico etc., logrando así la obtención de patrones que ayudaran a tomar decisiones con el fin de apoyar en la disminución de la actividad delictiva, siendo esta una herramienta clave para ayudar a la labor que realizan los diferentes entes de control frente a este tipo de problemática.

Es por esta razón 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, que sirva para los actores encargados de la seguridad en el municipio a realizar planes de mitigación y aprovechamiento eficiente del pie de fuerza.

El presente proyecto se desarrollará haciendo uso de la metodología KDD, para el desarrollo de este estudio se trabajará con algoritmos supervisados de regresión y agrupación, para obtención de patrones; Los datos a adquirir se extraerán de las bases de datos del observatorio del delito para los años 2010 a 2019. En la comparación objetiva se utilizará una metodología de sintonización de hiperparámetros mediante algoritmos evolutivos (algoritmos genéticos) y bayesianos. Para evaluar el compromiso en la predicción y coste computacional se utilizarán las métricas error cuadrático medio, error absoluto medio, coeficiente de determinación y coeficiente de correlación de Person.

En cuanto a la visualización geográfica de los datos se utilizará el interpolador de kriking utilizando el mejor Variograma dentro de los modelos lineales, gaussiano y esféricos.

Este documento describe el proyecto anteriormente enunciado mediante la definición de la descripción del problema, los objetivos a alcanzar, la justificación y los antecedentes, posteriormente se orienta en la definición de la metodología, los resultados esperados, los recursos y el tiempo necesarios para su realización.

1. PROBLEMA DE INVESTIGACIÓN

1.1 TEMA

Estudio comparativo de algoritmos supervisados y de agrupación de machine learning para la obtención de un modelo de regresión, aplicando descubrimiento de conocimiento en bases de datos (KDD) en históricos de actividad delictiva en el municipio de Tumaco.

1.2 ÁREA DE INVESTIGACIÓN

Este trabajo de grado está relacionado con el tratamiento de datos, programación de algoritmos, inteligencia artificial y específicamente aprendizaje automático.

1.3 LÍNEA DE INVESTIGACIÓN

Temáticas inscriptas en la "*línea software y manejo de información*"

1.4 FORMULACIÓN DEL PROBLEMA

Teniendo en cuenta los datos históricos del departamento del delito del municipio de Tumaco y un conjunto de algoritmos de machine learning pre seleccionados ¿Qué algoritmos de machine learning tienen el mejor compromiso en la exactitud para el apoyo al anális de actividad delictiva?

1.5 PLANTEAMIENTO DEL PROBLEMA

En la actualidad no existen herramientas eficaces que ayuden en el proceso de investigación e identificación puntual de los delitos y de quienes los cometen, siendo este un factor muy desfavorable para contribuir en los planes de mitigación de delincuencia que presenta el municipio de San Andrés De Tumaco; la falta de dichas herramientas es uno de los grandes problemas que hoy afronta el municipio, siendo esta una herramienta clave para ayudar a la labor que realizan los diferentes entes de control frente a este tipo de problemática de actividades delictivas.

Estos tipos de actividades delictivas se presentan entre los grupos armados en Tumaco, incluidos los grupos guerrilleros y paramilitares, quienes cometen violaciones sexuales, delitos asociados al narcotráfico, reclutamientos forzados y otros tipos de actividad delictiva ante la sociedad. Debido a la ola de criminalidad extendida por el territorio la Policía ha generado espacios para que el comandante del Distrito Especial de Tumaco, coronel José Luis Palomino López, sostuvo que la institucionalidad está generando espacios para que todas las entidades refuercen sus acciones para combatir la inseguridad que ronda al municipio. Según lo afirma en.[7]

Para aterrizar los elementos conceptuales de la problemática anterior, este estudio se enfocará en brindar una herramienta que ayude en el proceso de la investigación y la identificación de actividad delictiva la cual está relacionada con el cohecho, tráfico de influencias, malversación, fraudes y exacciones ilegales, negociación y actividad prohibida a los funcionarios públicos y por abusos en el ejercicio de su función, corrupción en las transacciones comerciales internacionales y delitos sobre la ordenación de territorios y el urbanismo, incluso todo acto de violencia hacia un individuo es conocido como una actividad delictiva. Así mismo, como un mecanismo posible para contribuir con las autoridades policiales dentro de los lineamientos expuestos anteriormente, se ve al aprendizaje automático como un elemento importante para caracterizar la zona de estudio. Ya que el aprendizaje automático es un tipo de inteligencia artificial (AI) que proporciona a las computadoras la capacidad de aprender, sin ser programadas explícitamente. El aprendizaje automático se centra en el desarrollo de programas informáticos que pueden cambiar cuando se exponen a nuevos datos;[8] el objetivo del machine learning es crear un modelo que permite resolver una tarea dada. Luego se entrena el modelo usando gran cantidad de datos.

El modelo aprende de estos datos y es capaz de hacer predicciones. Según la tarea que se quiera realizar, será más adecuado trabajar con un algoritmo u otro. Donde un algoritmo no es más que una serie de pasos ordenados que se dan para realizar una tarea. Hasta cierto punto, la mayoría de los científicos utiliza regularmente la predicción en la investigación como un elemento fundamental del método científico, cuando generan una hipótesis y predicen lo que sucederá; según, se afirma que una predicción es un anticipo de lo que ocurrirá de acuerdo al análisis de las condiciones existentes. Es frecuente que las predicciones surjan tras experimentos o investigaciones que permiten conocer las condiciones y estimar que, si se repiten,

[7] AUGUST, Extrema seguridad en 12 barrios de Tumaco. [en linea].(05 , 2021). Obtenido de https://diariodelsur.com.co/noticias/local/extreman-seguridad-en-12-barrios-de-tumaco-400202

[8] ARSYS, "Qué es Machine Learning y por qué es tan importante," . [en linea]. (25 febrero 2019). Obtenido en: https://www.arsys.es/blog/soluciones/infraestructura/machine-learning/

el resultado será el mismo. Siendo estos modelos una herramienta necesaria para identificar zonas donde es probable se presenten actividades delictivas.

Por otro lado, son pocas las herramientas disponibles que sirvan de apoyo a los entes de control inmediatos como se evidencia en el PETI 2019-2022 de la Policía Nacional donde se justifica la construcción de herramientas de análisis de datos frente a las carencias existentes[9]. Lo anterior refleja la importancia de las herramientas para el apoyo a la toma de decisiones para asistir a la mitigación de actividad delictiva ya que la delincuencia es un flageló que no solo azota al municipio de Tumaco, sino a Colombia y en consecuencia al mundo entero. Como evidencia, América es el continente con más asesinatos así se n de todo el planeta, señaló el estudio. Casi todos ellos ocurren en América Latina, que concentra apenas 8% de la población mundial. En cuanto a países en expansión con tasas de homicidio relativamente bajas; muchos de ellos, especialmente en Europa y Oceanía, han experimentado una disminución en los índices de homicidio desde 1990. En contraste, casi 750 millones de personas viven en países con niveles de homicidio elevados, lo que significa que casi la mitad de los homicidios suceden en países que representan alrededor del 11% de la población mundial y que la seguridad personal es aún una preocupación mayor para 1 de cada 10 personas en el mundo.

La tasa de homicidios de Haití se ha duplicado en seis años, de 5.1 en 2007 a 10.2 por cada 100 000 habitantes en 2012, en gran medida a causa de los elevados niveles de violencia y pandillerismo en la capital, para los países que van saliendo de un conflicto es decisivo prestar atención a la delincuencia y el homicidio en todas sus formas, ya que la violencia vinculada al crimen puede igualar, e incluso superar, aquélla generada por el conflicto mismo[10].

Colombia tiene una tasa de criminalidad excepcionalmente alta así lo afirman[11]. Esta información se sostiene al menos para los últimos 20 años y se puede comprobar con las cifras sobre asesinatos en Colombia y en otros países como Brasil, Nicaragua, Sri Lanka, Perú, Ecuador y estados Unidos. La operación estadística de “Conductas y Servicios de Policía” diferencia y cuenta los hurtos por la clase de bien; en el año 2015 se registraron 21.139 casos de hurto a residencias. La ciudad capital y cinco departamentos representaron el 53,73% del total de los

[9] PLAN ESTRATÉGICO DE TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES - PETI 2019 - 2022 . [en linea]. (2019). Obtenido en: https://www.policia.gov.co/contenido/plan-estrategico-tecnologias-informacion-y-comunicaciones

[10] O. D. D. PASTO, "La criminalidad en Nariño aumentó en un 60 por ciento según el Observatorio d(unodc,2013)el Delito," Diario del sur, 2017. [Online]. Available: https://diariodelsur.com.co/noticias/local/la-criminalidad-en-narino-aumento-en-un-60-por-ciento-segun-315758. [Accessed 13 marzo 2020].

[11] Ibíd., p. 2

casos: Bogotá (17,14%), Antioquia (12,01%), Valle (7,97%), Meta (6,01%), Tolima (5,96%) y Santander (4,64%). Asimismo, se registraron 22.455 casos de hurto a entidades comerciales; la ciudad de Bogotá (26,37%) y el departamento de Antioquia (14,99%) representaron el 41,36% del total de los registros. En el año 2015 la Policía Nacional registró 12.782 homicidios, con una reducción de al menos 4% con respecto al 2014. Los departamentos del Valle y Antioquia, y la ciudad de Bogotá, representaron el 46,48% (5.941) del total de los homicidios. Ocho departamentos registraron una reducción del homicidio mayor del 20% respecto del año anterior:

Caldas, Casanare, San Andrés, Guaviare, Guainía, Putumayo, Arauca y Amazonas. Cuatro meses del 2015 registraron menos de 1.000 homicidios: febrero, abril, septiembre y octubre. La tasa de homicidios fue de 28 personas muertas por cada cien mil habitantes. Doce departamentos registraron por encima del promedio de la tasa nacional: Valle (57), Quindío (49), Cauca (46), Putumayo (42), Caquetá (40), Chocó (35), Meta (35), Arauca (35), Risaralda (33), Guaviare (32), Antioquia (30) y Norte de Santander (29)[12].

Particularmente en el departamento de Nariño Las permanentes infracciones al Derecho Internacional Humanitario (DIH) y las graves violaciones a los Derechos Humanos se materializan a través (entre otras) de diferentes modalidades y las particularidades de cada zona, específicamente para el municipio de Ipiales se encontraron altos registros en los delitos de Lesiones personales, Hurto a personas y Hurto a comercio. Suministrados por la Policía Departamental (Nariño). Los registros sobre lesiones personales aumentaron, este delito según la información suministrada se ubica en la segunda posición con mayor número de casos. De acuerdo a las estadísticas el hurto común también registra un incremento. En el 2016 se presentaron 2.112, mientras que en los 5 primeros meses del año se reportan 1.541. La principal denuncia de robo es de equipos móviles o celulares. El hurto de vehículos es un delito que también se encuentra disparado. El año anterior se registraron 387 casos y en lo que va trascurrido de 2017 van 375. Si bien el índice de robo de ganado en las zonas rurales de Nariño es muy popular, casos de abigeato tienen una estadística muy baja. En 2016 se reportaron solo 12 casos, mientras que en este año se habla de 15, lo que igual significa que se presenta un incremento[13]. El hurto a personas se constituye frente a la información disponible, en el delito que mayor número de registros presenta. Dentro del contexto de los delitos de alto impacto que han afectado a gran parte de la población en el departamento de Nariño, en Túquerres se puede determinar que uno de los delitos de más preocupación es el hurto a personas que se está presentando en varias

[12] REVISTA CRIMINAL. vol.58 no.2 Bogotá May/Aug. 2016
[13] O. D. DELITO, ""Observatorio del Delito de la Policía Nacional"," Policia Nacional de Colombia, 2015. [Online]. Available: (https://www.policia.gov.co/observatoriodeldelito). [Accessed 03 04 2021].

modalidades (raponazo, cosquilleo, atraco a mano armada, entre otras) y que de alguna manera está afectando la seguridad y tranquilidad de los habitantes del municipio.

La situación es preocupante para la comunidad Tumaqueña el no contar con herramientas que brinden grandes aportes frente a esta problemática de actividades delictivas especialmente en la parte urbana se ha incrementado la presencia de grupos armados y principalmente los hechos de robos, raponazos y jaloneos. En promedio, los eventos delictivos de mayor impacto en el municipio son las lesiones personales (38,4%), seguido de los homicidios (19,5%) y el hurto a personas (18,7%), el restante 23% corresponde, en orden descendente, a delitos como: hurto a motocicletas, extorsión, hurto a residencias, hurto a comercio, hurto de automotores, hurto de cabezas de ganado, terrorismo y secuestro. Las víctimas de estos homicidios incluyen a líderes comunitarios. Desde 2015, en Colombia se ha registrado un aumento significativo de estos asesinatos. Tumaco, donde se han registrado al menos siete líderes comunitarios asesinados desde enero de 2017, es uno de los municipios más afectados. Si estos actos de violencias no se controlan, la comunidad Tumaqueña carece de herramientas capaces de aportar soluciones frente a este tipo de problemas de actividad delictivas como consecuencia vivirá con un ambiente de violencia incremental que va causando estragos en cada una de las familias, afectando la salud mental de sus habitantes, la economía del municipio, la falta de empleos; Según cifras del Dane, el 84% de la población de Tumaco vive en situación de pobreza y el desempleo predomina con un 74%, la falta de oportunidades legales hacen que algunos opten por la ilegalidad y se dediquen a delinquir con cultivos ilícitos, extorsiones, o atracos, donde estos delitos se ven generalmente en hombres, particularmente esta falta de oportunidades hace que además algunas mujeres opten por ser trabajadoras sexuales.

Estos factores afectan la inseguridad de los Tumaqueños y turistas que visitan con frecuencia este lugar, la deserción estudiantil en pregrado; debido a la violencia y el factor económico, generando así atraso en cuanto a la educación del municipio, incrementos de actividad delictiva a temprana edad, conflicto en los hogares, muertes de jóvenes y el ver a la delincuencia como un escape y parte del diario vivir mas no como muro que hay que derribar y sacar de sus vidas.
Desde el campo de la ingeniería se han hecho varios aportes para solventar distintas sintomatologías relacionadas con la violencia y delincuencia. Específicamente la inteligencia artificial ha aportado mediante los algoritmo de machine learning, como técnicas de minería de datos para la detección y prevención del lavado de activos y la financiación del terrorismo (la/ft), la cual fue aplicada en Bogotá D.C. EN EL 2014 registrado en otras como [14], las técnicas de aprendizaje

[14] YUMPU, "Tecnica y mineria de datos para la prevencion de lavado de activos y la financiacion de terroismo," Yumpu.com, 2017. [en linea]. Obtenito en:

automático para la detección de intrusos en redes de computadoras; El desarrollo de sistemas de detección de intrusos en redes de computadoras (del inglés NIDS). En la ciudad de Pasto se aplicó "Detección de Patrones de muerte por causa externa con técnicas de minería de datos en el observatorio del delito del municipio de Pasto" en Colombia [15] .

Es por ello que se pretende realizar un Estudio comparativo de algoritmos supervisados y no supervisados de machine learning para la obtención de un modelo de regresión, aplicando descubrimiento de conocimiento en bases de datos (KDD) en históricos de actividad delictiva en el municipio de Tumaco. Dado que se pretende realizar una contribución científica con significado social frente a la falta de herramientas capaces de dar solución a un problema y apoyo a la visualización de patrones delictivos alineados a las políticas nacionales debido a que esta problemática de actividad delictiva es un fenómeno social, de alta complejidad que no solo está afectando al municipio de Tumaco si no al mundo entero.

1.6 PREGUNTAS DE INVESTIGACIÓN

- ¿Cómo están estructurados los datos de actividad delictiva del municipio de Tumaco?
- ¿Qué variables ofrecen la mejor calidad en sus datos? Y ¿Cómo disminuir el ruido en los datos de las variables que presentan anomalías?
- ¿Qué variables son las más relevantes, para ser utilizadas en el modelo de entrenamiento?
- ¿Qué algoritmos de aprendizaje automático son los más adecuados para realizar predicciones sobre los datos de actividad delictiva del observatorio del delito?
- ¿De qué manera se puede comparar de manera justa y objetiva a los algoritmos de aprendizaje automático seleccionados?
- ¿Qué herramientas tecnológicas utilizar para implementar la metodología de comparación?
- ¿Cómo implementar la metodología de comparación?
- ¿Cómo validar los resultados obtenidos en la metodología de comparación implementada?
- ¿Cómo mostrar en mapas geográficos los mejores patrones de predicción hallados en la comparación?

---

https://www.yumpu.com/en/document/read/53701711/tecnicas-de-mineria-de-datos-para-la-prevencion-de. [Accessed 7 marzo 2020].

[15] VALENGA, F. "Aplicacion de mineria de datos para la exploracion y deteccion de patrones delictivos en argentina," sedici, [en linea]. Available: http://sedici.unlp.edu Timarán Pereira,2002.ar/bitstream/handle/10915/21783/Documento_completo.pdf?sequence=1. [accedido 4 abril 2021].

1.7 OBJETIVOS

1.7.1. Objetivo general. Implementar un marco experimental de comparación de 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**.**

1.7.2 Objetivos específicos.

- Preprocesar los datos de actividad delictiva del observatorio del delito que permita una selección de las variables más relevantes utilizando técnicas de extracción, transformación, limpieza, correlación y reducción de dimensión (RD).

- Desarrollar un marco experimental de comparación de algoritmos supervisados y no supervisados de machine learning que brinde un conjunto de modelos subóptimos de predicción de actividad delictiva.

- Analizar los resultados de los mejores modelos obtenidos mediante: gráficos de interpretabilidad, métricas de calidad y mapas de interpolación de la actividad delictiva en función de la ubicación geográfic

1.8 JUSTIFICACIÓN

Es de gran importancia la aplicación de un estudio comparativo de algoritmos supervisados de machine learning para la obtención de un modelo de regresión, aplicando descubrimiento de conocimiento en bases de datos (KDD) en históricos de actividad delictiva en el municipio de Tumaco, con el fin de brindar una herramienta que aporte a la investigación y permita ayudar a los planes de contingencia de los entes de control frente a los diferentes tipos de problemas de actividad delictiva; aportando en cuanto a una mayor seguridad a toda la comunidad Tumaqueña y a todos los turistas que visitan con frecuencia los diferentes barrios o lugares turísticos; de este modo poder contribuir al crecimiento de la economía y generación de empleo. Siendo este estudio de gran importancia para los entes encargado de velar por la seguridad ciudadana como lo es el Ministerio de Defensa Nacional siendo esta una herramienta de gran ayuda para predecir actividades delictivas en un rango de tiempo, y del mismo modo poder establecer estrategias, realizar planes de contingencia sobre los barrios de la comunidad donde más se presencien estos tipos de actividades delictiva, reforzar la seguridad en los lugares del municipio donde más se presencien actos delictivos con el fin de mitigar estos tipos de delincuencia que azotan al municipio de Tumaco y poder brindarles a los Tumaqueños y a sus visitantes la seguridad y poder disfrutar de un ambiente digno y tranquilo. La ola de violencia que vive el municipio de Tumaco es demasiado devastadora, la extorsión, robos, el tráfico de estupefacientes y otros aspectos que se desprenden de la actividad delictiva de las bandas criminales en el municipio han generado un ambiente de inseguridad y gran preocupación en la región Tumaqueña, “Aunque continúa siendo el municipio con mayor afectación y completa 16 años en la categoría de los 10 más afectados, el área sembrada de coca se redujo en un 16 %”, señala el monitoreo de cultivos ilícitos de la ONU contra la Droga y el Delito.

Desde finales de la década del noventa, Tumaco dejó de ser un lugar con una mínima presencia de actores armados y violencia para convertirse en uno de los casos emblemáticos de los nuevos escenarios del conflicto armado colombiano. En este municipio confluyen actualmente FARC y bandas criminales, acciones armadas de medio y bajo poder militar, una tasa de homicidios que supera más de tres veces la tasa nacional (130 hpch), un aumento en el número de víctimas por minas antipersonal y casos sistemáticos de micro extorsión, a lo que se suma que tiene el mayor número de hectáreas sembradas de coca a nivel nacional [16].

[16] INFANTE, D. "Dinámicas del conflicto armado en Tumaco y su impacto humanitario," FIP - FUNDACIÓN IDEAS PARA LA PAZ, 9 noviembre 2017. [Online]. Available: https://www.ideaspaz.org/publications/posts/926. [Accessed marzo 19 2020].

La aplicación de algoritmos supervisados y no supervisados de machine learning, es una rama de la inteligencia artificial que persigue el desarrollo de técnicas que permitan a las máquinas aprender de manera automática y mejorar a través de la experiencia, sin haber sido explícitamente programadas para ello lo establece en Arsys (2019), es por esta razón que se pretende por medio de esta técnica mejorar en cuanto a la búsqueda de patrones que permita saber más sobre la predicción de actividad delictiva que se presentan en el municipio de Tumaco con el fin de tomar decisiones en el futuro sobre nuevos conjunto de datos y como es el comportamiento de estos a lo largo de un periodo de tiempo y atravez de una comparación de estos algoritmos de los cuales se pueden extraer métricas y uso de recursos de la máquina, que pueden ser usadas para poder llevar a cabo una comparación y análisis[17].

Este tipo de estudio es muy novedoso y pretende hacer uso de la inteligencia artificial siendo su objetivo el crear máquinas con las mismas capacidades racionales que el ser humano a partir de la imitación de los procesos cognitivos, de esta manera se pretende aplicar machine learning.

Para predicción de actividades delictivas en Tumaco por primera vez; la cual Permita llevar el control de los tipos de problemas que presenta el municipio en el cual estos se manejan con frecuencias acudiendo a modelos de plan de mejoras, monitoreo, aumento de la fuerza armada, charlas de concientización, aumento de entes de control etc., las cuales estas se han convertido en pañitos de agua tibia para el manejo y control de esta gran problemática. Afirma Arsys (2019) que la combinación del aprendizaje automático con la IA y las tecnologías cognitivas puede hacer que sea aún más efectivo en el procesamiento de grandes volúmenes de información. Por ese motivo, el potencial de esta ciencia es enorme para la resolución de problemas complejos en todos los ámbitos. Además, se contará con la aplicación del (KDD) siendo para este su principal objetivo el proceso de minería de datos que consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.
Uno de los aspectos que generan viabilidad frente al problema que está presentando la comunidad Tumaqueña en cuanto a los diferentes tipos de actividades delictiva lo afirma Andalucía es Digital (2018) donde da a conocer que el Aprendizaje supervisado (supervised Machine Learning): En el que refiere a la interpretación de los datos del Big Data, previamente almacenados y clasificados (minería de datos o data mining); son de gran utilidad tanto desde el punto de vista de las empresas, en otras disciplinas como la Educación , Sanidad etc. .Estas

[17] ZAMORANO RUIZ, Juan. " Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea.,". [Online]. 2018 Available: https://eprints.ucm.es/id/eprint/48800/. [Accessed noviembre 24 2021].

referencias hacen que el estudio que se pretende realizar sea viable y arrojen buenos resultados en cuanto a la aplicación de este tipo de algoritmo que permitan dar soluciones al problema planteado. La aplicación de los sistemas de aprendizaje automático son uno de los grandes pilares y retos del futuro más inmediato, y con muy buenos resultados en soluciones de problemas de carácter social, científico, político y económico. Este estudio nos permite obtener una disminución de errores, acciones preventivas, ciberseguridad, automatización de procesos, siendo esta herramienta una de las más aplicadas en países de potencias desarrolladas con el fin de utiliza esta tecnología para facilitar el análisis de datos y obtener más y mejores insights (punto que nos lleva al camino de esa solución). Este tipo de estudio también fue aplicado por [18], aplicando el Descubrimiento de patrones, Minería de Datos, Observatorio del delito en el municipio de pasto a lo que se obtuvieron excelentes resultados.

La aplicación de este tipo de estudios aplicando algoritmos de aprendizaje automático en la comunidad Tumaqueña será una herramienta con alta calidad en cuanto a la IA que permitirá dar un aporte para resolver el problema como lo es la Actividad delictiva que se vive en el municipio de Tumaco, siendo este un lugar rico en agricultura, una de las mejores gastronomías, potencial sitio turístico, y uno de los municipios con alto índice de actividad delictivita e inseguridad. Estos son aplicados en la búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica, siendo este de gran beneficio para la ciencia, la educación; en cuanto a la creación de técnicas de aprendizaje. Beneficiando a si a las ramas de la investigación exploratoria; la cual es considerada como el primer acercamiento científico a un problema; siendo este estudio de gran aporte para las líneas Software y manejo de información, fundamentos de programación, auditoria de sistemas, sistemas distribuidos y lógicas matemáticas; de igual modo ayudar a los grupos de investigación como, GRIAS (Computación y Ciencias de la Información), Electrónica (GIIEE) (Ingeniería Eléctrica, Electrónica e Informática), Galeras.NET (Computación y Ciencias de la Información).

1.9 ALCANCE Y DELIMITACÓN

Para la realización de este estudio se emplearán técnicas de inteligencia artificial concernientes al aprendizaje automático o machine learning.

[18] VALENGA, Op. cit.

- Los algoritmos de aprendizaje automáticos a evaluar para la obtención de patrones serán algoritmos supervisados de regresión y algorimos no supervisados; arboles de decisión, bosques aleatorios, máquinas de soporte vectorial,k-means y k-Prototypes.

- Este estudio llevara a cabo las etapas del descubrimiento del conocimiento en bases de datos (KDD).

- Los datos adquirir se extraerán de las bases de datos del observatorio del delito para los años 2010 a 2019.

- Los algoritmos a escoger dentro de la comparación estarán relacionados a la relevancia expresada en artículos científicos similares.

- Para evaluar el compromiso en la predicción en los algoritmos de regresión se utilizará el coeficiente de determinación.

- Para la visualización geográfica de los datos se utilizará el interpolador de kriking utilizando el mejor variograma dentro de los modelos lineal, gaussiano y esféricos.

2. MARCO REFERENCIAL

2.1 MARCO TEÓRICO

2.1.1 Antecedentes. El estudio relacionado con aplicaciones de técnicas de minerías de datos, metodologías, limpieza de datos, transformación, caracterización de variables aplicación de la IA, encierra varios escenarios y trabaja desde distintos enfoques. Por ello, la investigación no solo requiere de un respaldo legal sino también teórico y conceptual que permitan a los diferentes estudios basados en la aplicación de machine learnig brindar grandes aportes en el desarrollo tecnológico de los diferentes municipios, permitiendo a si grandes estudios relacionados con algoritmos para dar solución a un problema. A continuación, se describen algunos proyectos u estudios con aportes relevantes incluyendo aspectos más precisos realizados a nivel global y nacional como soporte a lo anterior expuesto.

2.1.1.1 Antecedentes globales. Creación de un Sistema de predicción de hechos delictivos para la mejora del proceso de prevención del delito en el distrito de la molina utilizando de la Minería de Datos, realizado en Lima- Perú mediante algoritmos de aprendizaje automático, permiten crear patrones de predicción en xixbase a datos históricos, en este caso aplicando a la ocurrencia de hechos delictivos en el distrito de La Molina. El proyecto anterior citado aporta en cuanto a la aplicación de la minería de datos, lo que nos permite descubrir los patrones en grandes volúmenes de conjuntos de datos, utilizando los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos[19].

Modelo de aprendizaje automático para la predicción de la calidad del café, cuyo objetivo consiste en entrenar los algoritmos de aprendizaje automático escogidos con el conjunto de datos construido, bajo un enfoque de Clasificación y otro de regresión, realizado en la Universidad Distrital Francisco José De Caldas. Este estudio aporta en la aplicación de un enfoque de clasificación y regresión, el entrenamiento de los algoritmos y el poder validar la efectividad del modelo de aprendizaje automático implementándolo[20].

---

[19] JAULIS RUA y G. VILCARROMERO, Sistema de predicción de hechos delictivos,» 2015.[Enlínea].Available: https://repositorio.usmp.edu.pe/bitstream/handle/20.500.12727/2022/jaulis_vilcarromero.pdf?sequence=1&isAllowed=y. [Último acceso: 2021 08 06].

[20] SUAREZ PEÑA, Javier Andrés “Modelo de aprendizaje automático para la predicción de la calidad” [En línea] (diciembre, 2019). Obtenido en https://repository.udistrital.edu.co/bitstream/handle/11349/23560/SuarezPe%C3%B1aJavierAndres2019.pdf?sequence=1&isAllowed=y. [Último acceso: 13 marzo 2020].

Aplicación de minería de datos como técnica para encontrar patrones que expliquen la tendencia de los datos, con el objetivo de extraer conocimiento de los trabajos de titulación de la Facultad de Ciencias de la Escuela Superior Politécnica de Chimborazo; se aplicaron cinco modelos de clasificación: Máquinas de Soporte Vectorial, Redes Neuronales, Árbol de Decisión, Bosque Aleatorio y Potenciación; considerando las líneas Diseño de Experimentos y Análisis Multivariable. Este estudio aporta en la aplicación de Árboles de Decisión donde se visualizan las variables sobre el problema planteado y sobre la aplicación de minería de datos[21].

2.1.1.2 Antecedentes nacionales. Estudio donde se aplicó el Descubrimiento de patrones, Minería de Datos, Observatorio del delito en el municipio de pasto a lo que se obtuvieron excelentes resultados, este fue un estudio realizado y aplicado en la ciudad de pasto de los cuales sus datos fueron tomados del observatorio de delitos de pasto, En este proyecto de investigación se plantea detectar patrones delictivos utilizando técnicas de descubrimiento de conocimiento a partir de los datos del Observatorio del Delito del municipio de Pasto, que facilite a los organismos gubernamentales y de seguridad tomar decisiones eficaces en lo relacionado a la seguridad ciudadana y prevención de delitos. Este es uno de los antecedentes más relevantes, brindando un gran aporte al estudio que se pretende realizar ya que está enfocado en el mismo problema de actividad delictiva y aplicación de las mismas técnicas de aprendizaje automático, minería de datos y contar con la información de observatorio de delito del municipio el cual se aplicará dicho estudio. Este estudio está relacionado con los mismos objetivos que se pretenden alcanzar para el desarrollo de este proyecto[22].

Estudio cuyo objetivo radica en la creación de un modelo de caracterización de delitos en Cartagena mediante la aplicación de la técnica de minería de datos con miras a analizar y describir los patrones de tendencias identificados en los hurtos cometidos en la ciudad durante los años 2015-2016, teniendo en cuenta los registros obtenidos del portal virtual del gobierno. Este trabajo de investigación presenta un método de caracterización de algunos delitos relacionados con hurtos en la ciudad de Cartagena, soportado por el desarrollo de técnicas de minería de datos (DM). Para el desarrollo de esta investigación se partió de información primaria suministrada por el gobierno. El desarrollo central de la investigación está basado en un modelo de aprendizaje no supervisado de datos, implementando las

[21] HARO, Silvia “Minería de datos para descubrir tendencias en la clasificación de los trabajos de titulación” [En línea] , (2018). Obtenido en https://journal.espe.edu.ec/ojs/index.php/cienciaytecnologia/article/view/739

[22] TIMÁRAN PEREIRA, Silvio Ricardo “GrupLAC - Plataforma SCienTI - Colombia” [En línea] , (2002). Obtenido en https://scienti.minciencias.gov.co/gruplac/jsp/visualiza/visualizagr.jsp?nro=00000000001538. [Último acceso: 13 marzo 2020].

técnicas de análisis de componentes principales en los datos. De este estudio se obtienen aportes relevantes en cuanto al descubrimiento de patrones, procesamiento de los datos de actividades delictivas y caracterización de variables[23].

Efectuaron una Metodología para el análisis de la violencia en el departamento de Bolívar mediante técnicas de machine learning, en la ciudad de Cartagena, del cual se enfocada en el uso de machine learning para crear dichos patrones, tendencias o hipótesis que ayuden a esclarecer que variables son las que influyen significativamente en la ocurrencia de estos hechos criminales. Este estudio realizado aporta en cuanto a la creación de patrones como tendencia que ayudan a esclarecer que variables son las que influyen significativamente en la ocurrencia de actividades delictivas[24].

2.1.2 Los orígenes del Machine Learning. Por moderno que pueda parecer este campo, nos debemos remontar al año 1950 cuando el gran Alan Turing creó el "Test de Turing". De forma que para pasar el test, una máquina debía engañar a un humano haciéndole creer que se encontraba delante de un humano en vez de un ordenador.

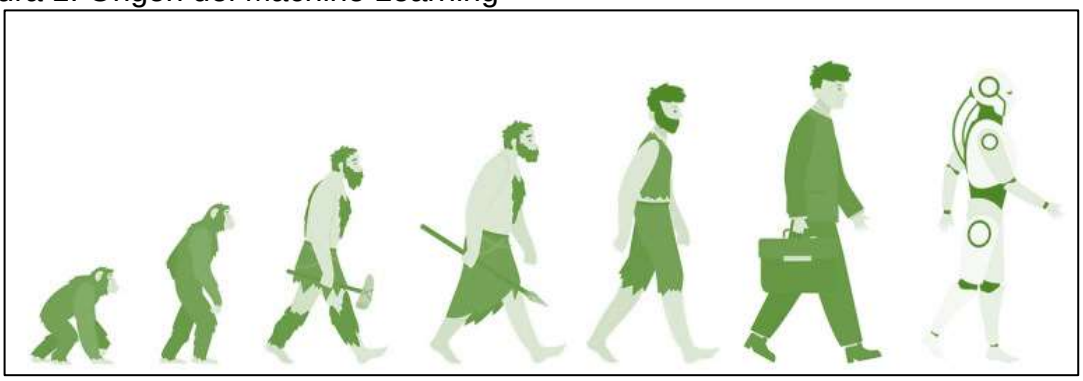
No debemos dejar de lado tampoco el año 1952, en el que Arthur Samuel escribe el primer algoritmo que es capaz de aprender; consistiendo este en un programa que jugaba a las damas y mejoraba tras cada partida su juego.
Posteriormente, en el seno de una conferencia nacerá el término 'Artificial Intelligence' (Inteligencia Artificial) para nombrar el nuevo campo que estudiaban en el verano de 1956[25].

[23] LICONA AGUILA, Aguilar y CONTRERAS, María Camila, (Septiembre – 2018)" Caracterización de los delitos en
Cartagena mediante la aplicación de minería de datos", (Septiembre – 2018),[En línea]. Available: https://biblioteca.utb.edu.co/notas/tesis/0074619.pdf

[24] FERNANDEZ CARABALLO, Franco. Metodología para el análisis de la violencia en el departamento de Bolívar mediante técnicas de machine learning",[En línea]. 2018. Available: https://repositorio.utb.edu.co/handle/20.500.12585/1118

[25] NALDA, Víctor "https://www.futurespace.es/machine-learning-los-origenes-y-la-evolucion/. [en linea].(29/ 09/ 2020). Obtenido dehttps://www.futurespace.es/machine-learning-los-origenes-y-la-evolucion/

Figura 1. Origen del Machine Learning



Fuente. future space (Anón 2020).

2.1.2.1 Evolución del Machine Learning. Los inicios del Machine Learning los encontramos en los años 50s, cuando Arthur Samuel, pionero en el campo de los juegos informáticos e IA, escribió el primer programa de aprendizaje informático matemáticos complejos al big data – una y otra vez, cada vez más rápido – es un logro reciente.

En los 60s, la creación del algoritmo conocido como “nearest neighbor” permitió a las computadoras utilizar un reconocimiento de patrones muy básico. Incluso tuvo fines comerciales, pues éste logró trazar un mapa de una ruta para vendedores ambulantes[26].

Los humanos pueden crear, por lo general, uno o dos buenos modelos por semana; el machine learning puede crear miles de modelos por semana. **Thomas H. Davenport,** Líder de pensamiento analítico Fragmento tomado de The Wall Street Journal

2.1.2.2 Importancia del Machine Learning. El resurgimiento del interés en el aprendizaje basado en máquina se debe a los mismos factores que han hecho la minería de datos y el análisis Bayesiano más populares que nunca. Cosas como los volúmenes y variedades crecientes de datos disponibles, procesamiento computacional más económico y poderoso, y almacenaje de datos asequible.

[26] RECLU IT “Historia y evolución del Machine Learning” (03/ 08/ 2020) [en linea]. Obtenido de https://recluit.com/historia-y-evolucion-del-machine-learning/#.YRNQDohKjIU

Todas estas cosas significan que es posible producir modelos de manera rápida y automática que puedan analizar datos más grandes y complejos y producir resultados más rápidos y precisos – incluso en una escala muy grande. Y con la construcción de modelos precisos, una organización tiene una mejor oportunidad de identificar oportunidades rentables – o de evitar riesgos desconocidos[27]

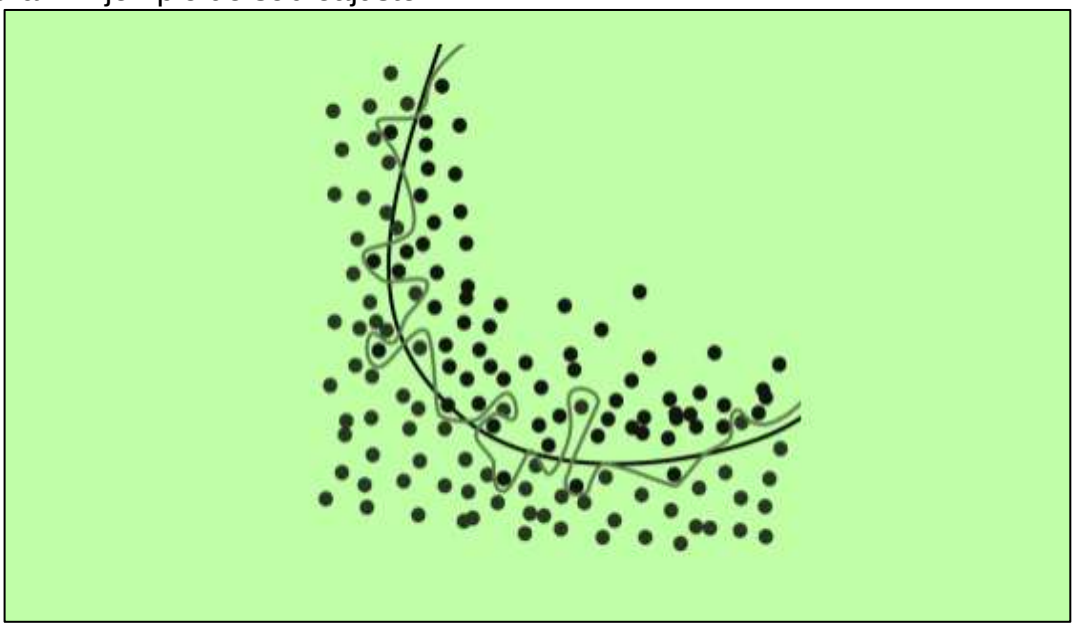
2.1.3 Data mining. Una de las técnicas de inteligencia de negocio que se ha venido utilizando dentro de las empresas es la minería de datos, la cual a partir de la exploración y el análisis se enfoca en descubrir conocimiento. Según Fayad, la minería de datos es "un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos" (Fayad, 1996, p. 5). La minería reúne ventajas de diversos campos como lo son: la estadística, la inteligencia artificial, la computación gráfica, las redes neuronales, entre otros" (Bramer, 2007, s. p.). La minería de datos brinda un acceso y navegación retrospectiva de los datos y de esta manera genera información precisa y oportuna a partir del apoyo de tres tecnologías (Bramer, 2007): (1) recolección de datos, (2) multiprocesador y (3) algoritmos de minería de datos[28].

2.1.4 Algoritmos de aprendizaje supervisado. Son entrenados utilizando ejemplos etiquetados, como una entrada donde se conoce el resultado deseado. Por ejemplo, una pieza de equipo podría tener puntos de datos etiquetados como “F” (fallidos) o “R” (corridas). El algoritmo de aprendizaje recibe un conjunto de entradas junto con los resultados correctos correspondientes, y el algoritmo aprende comparando su resultado real con resultados correctos para encontrar errores. Luego modifica el modelo en consecuencia. A través de métodos como la clasificación, regresión, predicción y aumento de gradiente, el aprendizaje supervisado utiliza patrones para predecir los valores de la etiqueta en datos no etiquetados adicionales. El aprendizaje supervisado se utiliza comúnmente en aplicaciones donde datos históricos predicen eventos futuros probables.[29]
“El aprendizaje automático es esa rama de la informática que otorga a la IA la capacidad de aprender tareas a través de los algoritmos del machine learning”[30]

[27] RUBY, Walker “Machine Learning” [en linea]. (07/ 05/ 2019) Obtenido de https://prezi.com/p/hrcxdfm91iaq/machine-learning/
[28] DUEÑAS, María Ximena “Minería de datos espaciales en búsqueda de la verdadera información” [en linea]. (Junio, 2009) Obtenido de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-21262009000100007
[29] AGENDA “Los algoritmos de aprendizaje supervisado“, (en línea)(13 agosto 2019) obtenido en https://www.revistaagenda.net/blog/machine-learning-el-motor-de-la-innovacion-en-las-empresas/
[30] REDACCIÓN APD “¿Cuáles son los tipos de algoritmos del machine learning? “, (en línea)(04 abril 2019) obtenido en https://www.apd.es/algoritmos-del-machine-learning/

Es de mencionar que se aplicaran algoritmos de tipo aprendizaje supervisados, empezaremos con los de clasificación, ya que son los que se ajustan a nuestra forma de representar las variables de entrada y salida, los patrones de actividad delictiva que queremos predecir. Antes de empezar con los algoritmos es necesario introducir el sobreajuste y la sobre generalización, o con sus términos en inglés, overfitting y underfitting. El sobreajuste es cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, pero generaliza mal cuando lleguen nuevos datos desconocidos. Por el contrario, la sobre generalización, que es cuando el modelo comete un error al no tener en cuenta ciertos patrones que son importantes para poder generalizar bien a nuevos datos y que funcione bien con los datos de entrenamiento como lo indica la figura 2.

Figura 2. Ejemplo de sobreajuste



Fuente.(Anón s. f., 2019)

2.1.5 Algoritmo-árbol de decisión. Un árbol de decisión es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente.

Es un algoritmo supervisado de aprendizaje automático porque para que aprenda el modelo necesitamos una variable dependiente en el conjunto de entrenamiento. Para dividir el espacio muestral en subregiones es preciso aplicar una serie de reglas o decisiones, para que cada sub-región contenga la mayor proporción posible de individuos de una de las poblaciones.

Si una sub-región contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en sub-regiones menores que integran datos de la misma clase.[31]

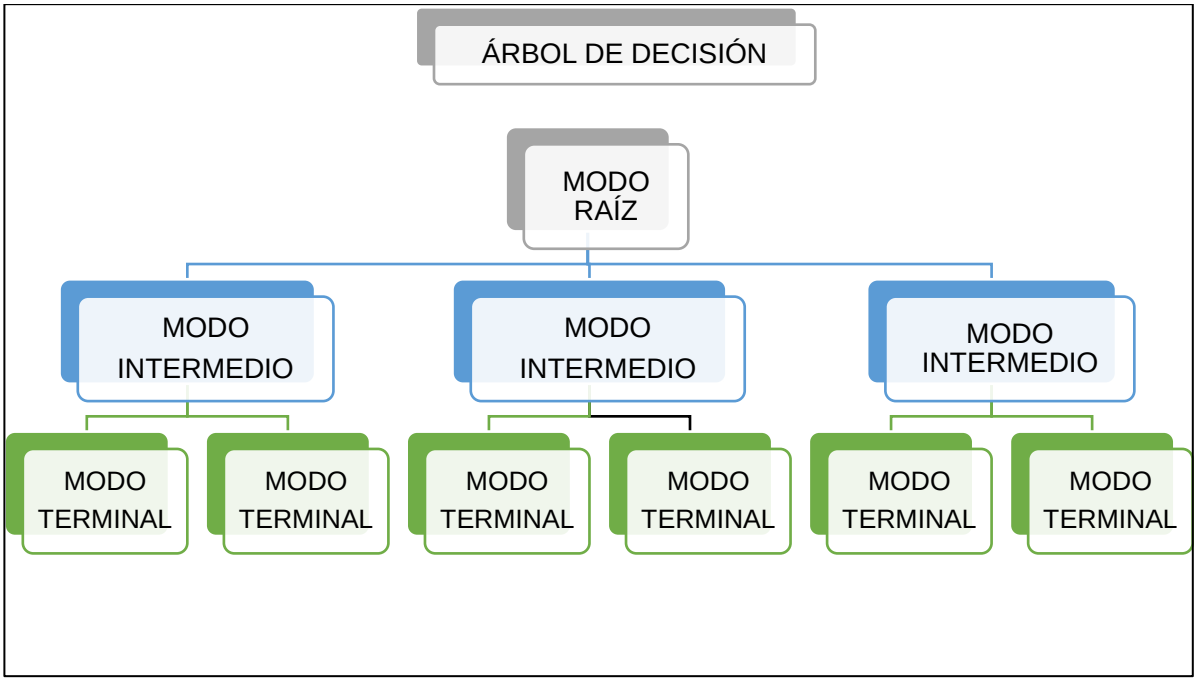
2.1.5.1 Estructura básica de un árbol de decisión. Los árboles de decisión están formados por nodos y su lectura se realiza de arriba hacia abajo.

Dentro de un árbol de decisión distinguimos diferentes tipos de nodos:

- Primer nodo o nodo raíz: en él se produce la primera división en función de la variable más importante.
- Nodos internos o intermedios: tras la primera división encontramos estos nodos, que vuelven a dividir el conjunto de datos en función de las variables.
- Nodos terminales u hojas: se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva.
- Otro concepto que debes tener claro es la profundidad de un árbol, que viene determinada por el número máximo de nodos de una rama.

[31] MERAYO, Patricia “Algoritmo-árbol de decisión”[en línea](20, mayo), obtenido en https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

Figura 3. Estructura de árbol de decisión, conformación de cada nodo, nodo intermedio y nodo terminal.



Fuente. (Data Science, 2020)

2.1.5.2 Ventajas del árbol de decisión.

- Son fáciles de construir, interpretar y visualizar.
- Selecciona las variables más importantes y en su creación no siempre se hace uso de todos los predictores.
- Si faltan datos no podremos recorrer el árbol hasta un nodo terminal, pero sí podemos hacer predicciones promediando las hojas del sub-árbol que alcancemos.
- No es preciso que se cumplan una serie de supuestos como en la regresión lineal (linealidad, normalidad de los residuos, homogeneidad de la varianza, etc.).
- Sirven tanto para variables dependientes cualitativas como cuantitativas, como para variables predictoras o independientes numéricas y categóricas. Además, no necesita variables dummys, aunque a veces mejoran el modelo.

- Permiten relaciones no lineales entre las variables explicativas y la variable dependiente.

- Nos podemos servir de ellos para categorizar variables numéricas.

En el caso de los árboles de decisión de un problema de regresión se utiliza el RSS (Residual Sum of Squares) que es una medida de la discrepancia entre los datos reales y los predichos por el modelo. Un RSS bajo indica un buen ajuste del modelo a los datos, es decir, se busca minimizar el RSS.
Se define el RSS como:

$$RSS = \sum_{i=1}^{n} \left(yi - \hat{y}i \right)^2$$

Ecuación 1. Residual Sum of Squeare, medida de la discrepancia entre el el dato real y los predichos por el modelo.

Donde yi es el valor real de la variable a predecir y ˆyi es el valor predicho.

2.1.6 Algoritmo k-Nearest Neighbor (KNN). Es un método que simplemente busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de datos que le rodean. Como dijimos antes, es un algoritmo:

Supervisado: esto -brevemente- quiere decir que se etiquetaron al conjunto de datos de entrenamiento, con la clase o resultado esperado dada “una fila” de datos.

Basado en Instancia: Esto quiere decir que el algoritmo no aprende explícitamente un modelo (como por ejemplo en Regresión Logística o árboles de decisión). En cambio, memoriza las instancias de entrenamiento que son usadas como “base de conocimiento” para la fase de predicción[32].

[32] JBAGNATO“¿Qué es el algoritmo k-Nearest Neighbor? ”[en línea], (2018), obtenido en https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/

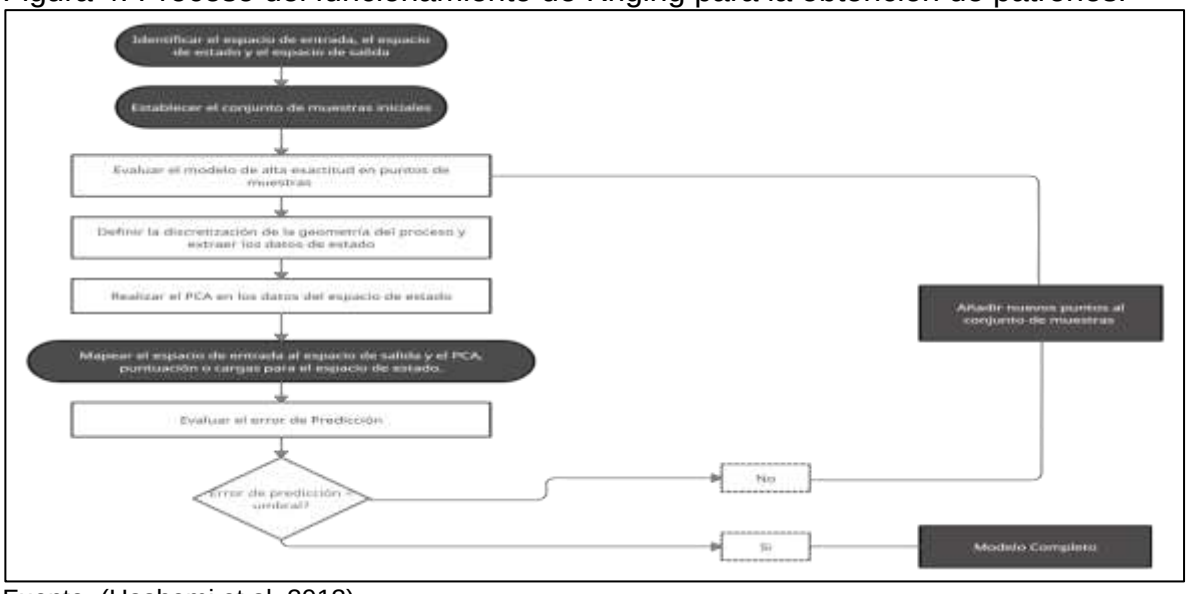
2.1.7 Cómo funciona KNN?.

1. Calcular la distancia entre el ítem a clasificar y el resto de ítems del dataset de entrenamiento.
2. Seleccionar los “k” elementos más cercanos (con menor distancia, según la función que se use)
3. Realizar una “votación de mayoría” entre los k puntos: los de una clase/etiqueta que <<dominen>> decidirán su clasificación final.

Teniendo en cuenta el punto 3, veremos que para decidir la clase de un punto es muy importante el valor de k, pues este terminará casi por definir a qué grupo pertenecerán los puntos, sobre todo en las “fronteras” entre grupos. Por ejemplo -y a priori- yo elegiría valores impares de k para desempatar (si las features que utilizamos son pares). No será lo mismo tomar para decidir 3 valores que 13. Esto no quiere decir que necesariamente tomar más puntos implique mejorar la precisión. Lo que es seguro es que cuantos más “puntos k”, más tardará nuestro algoritmo en procesar y darnos respuesta[33].

2.1.8 Interpolación Kriging.

Figura 4. Proceso del funcionamiento de Kriging para la obtención de patrones.



Fuente. (Hashemi,et al.,2013)

Las herramientas de interpolación IDW (Distancia inversa ponderada) y Spline son consideradas métodos de interpolación determinísticos porque están basados

[33] LIZARDO, B. A. ¿Cómo funciona kNN?,» Platzi, Enero 2021. [En línea]. Available: https://platzi.com/tutoriales/1841-probabilistica/9110-k-nearest-neighbor/.

directamente en los valores medidos circundantes o en fórmulas matemáticas especificadas que determinan la suavidad de la superficie resultante. Hay una segunda familia de métodos de interpolación que consta de métodos geoestadísticos, como **Kriging,** que está basado en modelos estadísticos que incluyen la autocorrelación, es decir, las relaciones estadísticas entre los puntos medidos. Gracias a esto, las técnicas de estadística geográfica no solo tienen la capacidad de producir una superficie de predicción, sino que también proporcionan alguna medida de certeza o precisión de las predicciones.

Kriging presupone que la distancia o la dirección entre los puntos de muestra reflejan una correlación espacial que puede utilizarse para explicar la variación en la superficie. La herramienta Kriging ajusta una función matemática a una cantidad especificada de puntos o a todos los puntos dentro de un radio específico para determinar el valor de salida para cada ubicación. Kriging es un proceso que tiene varios pasos, entre los que se incluyen, el análisis estadístico exploratorio de los datos, el modelado de Variograma, la creación de la superficie y (opcionalmente) la exploración de la superficie de varianza. Este método es más adecuado cuando se sabe que hay una influencia direccional o de la distancia correlacionada espacialmente en los datos. Se utiliza a menudo en la ciencia del suelo y la geología[34].

En los modelos de Kriging, la respuesta predicha asociado con un nuevo punto de entrada se determina como una suma ponderada de los valores de función conocidos asociados con entradas muestreadas previamente visualizado en la Ecuación (1). El peso atribuido a cada uno disminuye con el aumento de Euclidiana; distancia entre los puntos. Por lo tanto, Kriging se considera una ponderación de distancia inversa.

En general, una distancia máxima r se define de manera que solo los puntos dentro de cierta distancia r.

$$\boldsymbol{f(X_k) = \sum_{i=1}^{N} W_i f(X_i)}$$

Ecuación 1. Suma ponderada de los valores de la función

Los pesos de la Ecuación (1) son desconocidos y su determinación es un paso crítico en el desarrollo del modelo Kriging. El objetivo es seleccionar los pesos de manera que el cuadrado medio se minimiza el error de predicción. En la práctica,

[34] ArcMap, «Cómo funciona Kriging» esri, 2016. [En línea]. Available: https://desktop.arcgis.com/es/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm.

estos pesos se pueden obtener utilizando un variograma ajustado.modelo basado en N puntos que están suficientemente cerca de los pesos, se deben sumar la unidad, una restricción lo que surge en parte de la condición de que el predictor de Kriging sea insesgado,además, si los puntos de entrada están muy agrupados, se les da un peso menor para evitar una estimación sesgada. Para los propósitos de la discusión subsiguiente, la distancia euclidianaentre los puntos (h) y el variograma correspondiente a un conjunto de datos x que consta de N puntos demuestra γ (h) se definirá como en la ecuación (2).

$$\boldsymbol{h} = \|\boldsymbol{X_{i-}X_j}\|$$

$$\mathbf{r(h)=}\frac{\mathbf{1}}{\mathbf{2}}[\boldsymbol{var}(\boldsymbol{f}(\boldsymbol{X\_i}) - \boldsymbol{f}(\boldsymbol{X\_j}))]$$

Ecuación 2. Distancia euclidianaentre los puntos (h)

Como puede verse en la Ecuación (2), el variograma se calcula para pares de puntos individuales. Este tipo del variograma a veces se denomina semivarianza, mientras que el variograma se describe como 2γ (h), pero a los efectos de la discusión posterior, γ (h) se denominará variograma. Para conjuntos de datos que contiene N puntos de muestra hay un total de $\boldsymbol{N_T(N_{T-}}$ **1) / 2** distancias euclidianas y valores de γ correspondientes a determinar. Los datos **γ** contra **h** resultantes generalmente se suavizan y luego se instalan en uno de los cinco modelos básicos; esférico, gaussiano, exponencial, lineal o de potencia. El modelo se elige de manera que minimice el error de predicción, aunque la eficiencia computacional también puede ser considerado al hacer la selección. Si es necesario, se pueden utilizar combinaciones de los distintos tipos de modelos.

Para obtener un error apropiadamente bajo. Una vez que se ha obtenido una expresión para el variograma, se puede utilizar para obtener una función complementaria conocida como función de covarianza que se muestra en Ecuación (3).

$$\boldsymbol{cov}(\boldsymbol{h}) = \boldsymbol{\sigma^2}_{\boldsymbol{max}} - \boldsymbol{r}(\boldsymbol{h})$$

Ecuación 3. Funcion de covarianza

El término $\boldsymbol{\sigma^2}_{\boldsymbol{max}}$ corresponde a la varianza máxima de la función del variograma. Los pesos de kriging para un punto de prueba dado se pueden obtener a partir de la covarianza resolviendo el sistema de la Ecuación, como se muestra en la ecuación (4).

$$\begin{bmatrix} \mathbf{Cov(d1,1)} & \mathbf{Cov(d1,N)} & \mathbf{1} \\ \mathbf{Cov(dN,1)} & \mathbf{Cov(dN,N)} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} \end{bmatrix} \mathbf{x} \begin{bmatrix} \boldsymbol{W_1} \\ \boldsymbol{w_N} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{Cov(d1,k)} \\ \mathbf{Cov(dN,K)}) \\ \mathbf{1} \end{bmatrix}$$

Ecuación 4. Sistema de ecuación a partir de la covarianza

$\boldsymbol{d_{ij}}$ representa la distancia entre dos puntos $\boldsymbol{X_i}$ y $\boldsymbol{X_j}$ mientras que $\boldsymbol{d_{i,k}}$ representa la distancia entre un punto $\boldsymbol{X_i}$ y el punto de prueba $\boldsymbol{X_k}$. Al igual $\mathbf{Cov}\ (\boldsymbol{d_{ij}})$que representa la covarianza entre los datos correspondientes a vectores de entrada que están a una distancia $\boldsymbol{d_{ij}}$, obtenida a partir de la ecuación (3). λ son Multiplicadores de Lagrange asociados a la restricción de que las ponderaciones deben sumar la unidad. La varianza asociada al punto de prueba $\boldsymbol{X_k}$ se calcula a partir de la ecuación (5), en la que el término $\mathbf{Cov}\ (\boldsymbol{d_{ij}})$ es el lado derecho de la ecuación (4)[35].

$$\boldsymbol{\sigma^2{}_k = \sigma^2{}_{max} - \sum_j^N W_i cov(d_i, \mathbf{k}) - \lambda}$$

*Ecuación 5***.** Varianza asociada al punto de prueba

2.1.8.1 Variograma experimental. Es una herramienta que permite analizar el comportamiento espacial de una propiedad o variable sobre una zona dada. Ejemplo: detectar el comportamiento de actividad delictiva dada la variable "TIPO DE DELITO", teniendo en cuenta la latitud y longitud de cada uno de los barrios del municipio de Tumaco.

Obteniendo como resultado un Variograma experimental que refleja la distancia máxima y la forma en que un punto tiene influencia sobre otro punto a diferentes distancias.

El resultado de este análisis no puede ser aplicado directamente en los diferentes métodos de interpolación que lo ocupan como información base, es por esto que una vez calculado el Variograma experimental, debe ser realizado un modelo matemático que modele de la mejor forma posible al Variograma experimental, el cual es conocido como Variograma teórico.

[35] AMANDA J. Rogers; AMIR, Hashemi y MARIANTHI, G. Ierapetritou. Modelado de procesos de partículas para el continuo Fabricación de formas farmacéuticas de dosificación de base sólida. 2013 Available: file:///C:/Users/PcPersonal/Downloads/Modeling_of_Particulate_Processes_for_the_Continuo.pdf. (2021-11-27).

Propiedades del variograma experimental: El variograma experimental $\hat{\gamma}(\mathbf{h})$ es un estimador **insesgado** del variograma teórico:

$$E[\hat{\gamma}(\mathbf{h})]=\gamma(\mathbf{h}) \ .$$

Un indicador de la **robustez** de $\hat{\gamma}(\mathbf{h})$ es su varianza relativa

$$\mathrm{var}[\hat{\gamma}(\mathbf{h})]/[\gamma(\mathbf{h})]^2 \ .$$

Mientras más elevada dicha varianza, más susceptible es el variograma experimental de fluctuar en torno a su valor esperado (el variograma teórico $\gamma(\mathbf{h})$) y más difícil se vuelve la inferencia estadística. Aunque esta varianza relativa sólo puede ser expresada en algunos casos particulares, puesto que requiere conocer la función aleatoria hasta sus distribuciones quadrivariables, los principales factores que la influencian son:

- La distancia considerada (norma del vector **h**): la varianza relativa de $\hat{\gamma}(\mathbf{h})$ suele tomar valores considerables para las grandes distancias (para fijar las ideas, aquellas distancias mayores que la mitad del diámetro del campo).

- La irregularidad o el carácter preferencial de la malla de muestreo, que pueden provocar grandes fluctuaciones en el variograma experimental, incluso para pequeñas distancias.

- El número de pares de datos: mientras más bajo, mayores son las fluctuaciones.

- La presencia de datos extremos (*outliers*), los cuales tienen un impacto considerable en el cálculo del variograma experimental, pues este último eleva los valores al cuadrado.

2.1.8.2 Los métodos kriging. Existen dos métodos kriging: *ordinario y universal.*

El **kriging ordinario:** Es el más general y más utilizado de los métodos kriging y es el predeterminado. Presupone que el valor medio constante es desconocido. Esa es una presuposición razonable a menos que haya una razón científica para rechazarla.

El kriging universal: presupone que hay una tendencia de invalidación en los datos, por ejemplo, un viento prevaleciente, y puede modelarse a través de la función determinística polinómica. Esta función polinómica se resta de los puntos medidos originalmente y la autocorrelación se modela a partir de los errores aleatorios. Una vez que el modelo se ajusta a los errores aleatorios y antes de realizar una predicción, se vuelve a sumar la función polinómica a las predicciones para obtener resultados significativos. El kriging universal solo se debe utilizar si se conoce una tendencia en los datos y se puede dar una justificación científica para describirla[36].

2.2 MARCO CONCEPTUAL

Para el desarrollo de este proyecto ha sido necesario conceptualizar varios términos que ayudarán a entender mejor el propósito de esta investigación. Siendo estos explicados desde los aspectos más generales hasta llegar a los conceptos más específicos que se han de necesitar para la realización de esta tesis:

2.2.1 Actividad delictiva.

2.2.1.1. Actividad delictiva en Colombia. El fenómeno de los hechos de actividad delictiva y la violencia en las grandes zonas urbanas, principalmente de los países en desarrollo, se ha convertido en uno los grandes retos del siglo XXI. Por ejemplo, en América Latina, las tasas de delitos a la propiedad se encuentran entre las más altas del mundo y en general se observa que la población juzga al problema de la criminalidad como el problema social más importante de la región. Esta situación es particularmente alarmante si consideramos que la evidencia disponible muestra, para el caso de América Latina, que una vez que aumenta la actividad criminal, es

[36] ARCMAP, Los métodos kriging» esri, 2016. [En línea]. Available: https://desktop.arcgis.com/es/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm.

muy difícil hacerla decrecer, aun cuando se hayan eliminado los factores que causaron el incremento inicial.[37]

Los orígenes, causas y consecuencias de este fenómeno de la criminalidad son ciertamente múltiples y requieren analizar sus distintas dimensiones. Destaca, sin embargo, que el comportamiento de estos índices delictivos muestra ciertos patrones regulares que permiten identificar algunas relaciones básicas; en particular en aspectos referidos al desempeño de la policía y a factores económicos[38].

2.2.1.2. Evidencia empírica de actos delictivos en Colombia. Las actividades delictivas pueden considerarse como la consecuencia de un conjunto de factores que incluyen tanto condiciones económicas y sociales como factores demográficos, psicológicos y de respeto e imposición de la ley. En particular, los textos sobre economía (por ejemplo, Becker, 1968 y Ehrlich, 1977) plantean que el conjunto de los individuos responde a una estructura de incentivos y que, en este sentido, los incentivos, tanto positivos como negativos, determinan su participación o exclusión de las actividades criminales (Freeman, 1983 y Cameron, 1988). De este modo, la participación en actividades legales o ilegales puede plantearse como un problema económico asociado incluso a modelos de oferta y demanda del "mercado de trabajo" (Ehrlich, 1971. Estos modelos deben entonces incluir no sólo los incentivos económicos directos sino también considerar que las actividades criminales se realizan, desde luego, en un contexto de incertidumbre, donde existe una probabilidad (P) de ser atrapado y, por tanto, de ser castigado (sh) (s expresa la severidad del castigo y h el nivel de la actividad criminal). Así, el conjunto de las actividades criminales se considera una función de los pagos de la actividad ilegítima en referencia a los ingresos reales o potenciales de las actividades legales (por ejemplo, el salario de la actividad legítima), la probabilidad de aprehensión y la magnitud del castigo, los gustos o preferencias de riesgo y valores morales, culturales o incluso de capital social (Ehrlich, 2005). De esta manera, las personas participan en actividades delictivas dependiendo de los costos y ganancias potenciales, dada su estructura de preferencias y "valores morales". Ello implica entonces, desde el punto de vista económico, que existe un precio o incentivo de equilibro que determina la participación en actividades ilegales. En este sentido, las personas realizan actividades criminales no porque sus motivaciones básicas difieran sino porque sus beneficios y costos difieren[39] .

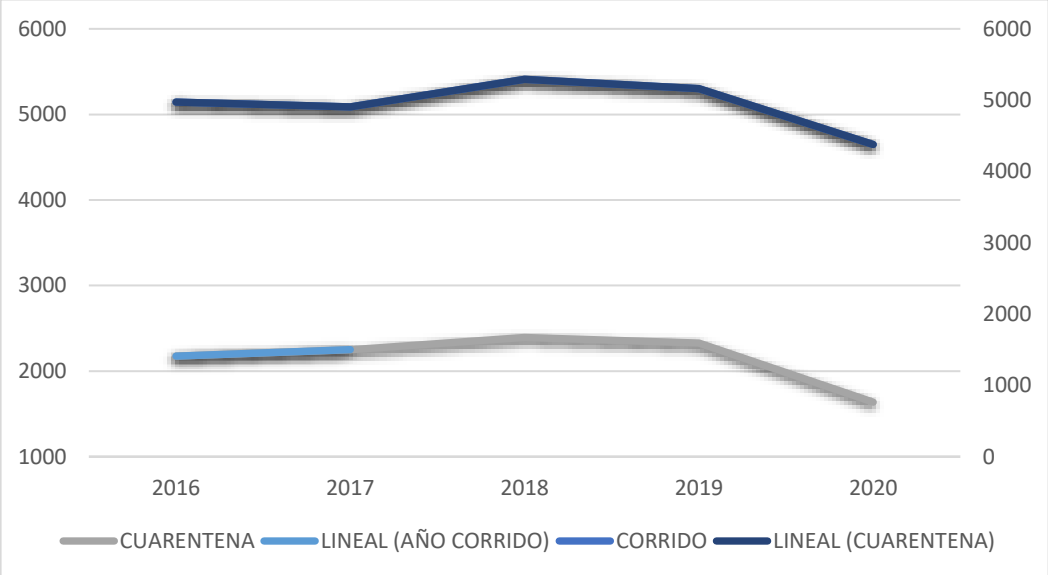
2.2.1.3 Caracterizaciones de actos delictivos en Colombia. El análisis de la violencia a partir de hechos como los homicidios, hurto a personas amenaza, donde

[37] ibid.
[38] GALINDO, L. y CATALÁN, H. Las actividades delictivas en el Distrito Federal,» julio-septiembre 2007. [En línea]. Available: https://www.redalyc.org/pdf/321/32112593003.pdf.
[39] Ibid.

el homicidio es uno de los delitos de mayor impacto social que posibilita la identificación de la mayoría de las víctimas, permite su interpretación en términos de comportamiento objetivamente observable a través de indicadores útiles para medir la probabilidad de riesgo y para comparar a nivel temporal y espacial los avances o retrocesos de la seguridad de una ciudad o un país. (Jerónimo Castillo, 2020)

Homicidio: aunque durante los meses de cuarentena se ha alertado sobre la reducción de los casos en diferentes comportamientos delictivos, sorprende que la disminución del homicidio no haya sido significativa u homogénea. Por el contrario, en algunas ciudades la reducción es modesta, y en otros municipios el número de casos aumentó. Revisando el periodo comprendido entre enero y mayo de los últimos cinco años, y el que va entre el 25 de marzo y el 31 de mayo (cuarentena), se observan múltiples hallazgos notables. En primer lugar, durante el periodo comprendido entre enero y mayo del último lustro, la tendencia del homicidio en Colombia venía en aumento, alcanzando su pico más alto en 2018, con 5.292 casos. Si bien en el periodo de 2019 hubo una ligera reducción, en lo corrido de 2020 se observa un mayor decrecimiento frente al año anterior (782 casos); es decir, una reducción del 15,2%. Estos datos generan el interrogante de por qué, frente a una circunstancia tan excepcional como una cuarentena nacional, el homicidio tan solo ha disminuido en un 15.2%.

Figura 5. Casos homicidios enero-mayo vs cuarentena 2016-2020



Fuente. (Jerónimo Castillo, 2020)

Hurto a personas: el hurto a personas ha venido aumentado en Colombia durante los últimos diez años, y su comportamiento contrasta con reducciones importantes que se dan en otros delitos.

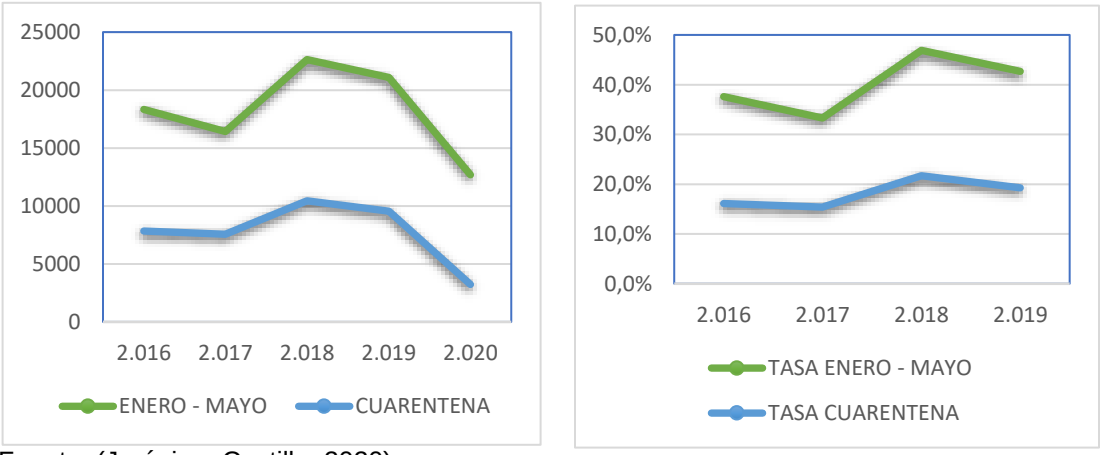
En todas sus modalidades (atraco, cosquilleo, arponazo, entre otros), el hurto es uno de los comportamientos delictivos que más preocupa a la ciudadanía y aumenta la percepción de inseguridad. Además, se ha convertido en uno de los mayores retos en materia de seguridad ciudadana, sobre todo en las grandes ciudades. El hurto —y en especial el hurto a personas—presenta un alto registro. A pesar de que se han impulsado mecanismos de denuncia en línea, los ciudadanos aún enfrentan diversas barreras para denunciar, algo que, a su vez, mantiene el subregistro. Llama la atención que las dificultades para presentar una denuncia son crónicas y no pareciera que las autoridades competentes (la Policía y Fiscalía) se tomaran en serio la necesidad de facilitar la captura y el acceso a información más amplia y precisa que permita entender este fenómeno y diseñar estrategias integrales para su control.

Amenaza: entre enero y mayo de 2020, las amenazas en Colombia disminuyeron significativamente respecto al mismo periodo del año anterior. Aunque ya se presentaba una tendencia decreciente en el comportamiento de este delito a nivel nacional desde 2019, la disminución registrada entre enero y mayo de 2020 fue relativamente mayor. Mientras que de enero a mayo de 2019 la reducción fue del 6,8% en el número de casos de este delito, durante los primeros cinco meses de 2020 su reducción alcanzó 39,8%, lo que sugiere un impacto significativo de la cuarentena, ya sea en la ocurrencia del delito o en su denuncia. Cabe destacar que, con estas reducciones, se alcanzó la menor cifra de amenazas reportadas en los últimos cinco años en el país: 12.692.

El comportamiento de la tasa de amenazas a nivel nacional sigue el mismo patrón decreciente que se registra desde 2018. En lo corrido de 2020 la tasa se redujo a 25 amenazas por cada cien mil habitantes, desde una tasa de 43 registrada en los primeros cinco meses del año anterior. De igual forma, en el lapso del 24 de marzo al 30 de mayo de 2019, se registró una tasa de amenazas de 19, mientras que en el periodo de cuarentena de 2020 disminuyó a 6 ,como lo indica la figura 6[40].

[40] J. C. e. al., « ¿Cómo se comporta el delito en Colombia en época de confinamiento?,» Septiembre 2020. [En línea]. Available: https://ideaspaz.org/media/website/FIP_DelitoyConfinamiento.pdf. [Último acceso: 16 Marzo 2020].

Figura 6. Comportamiento casos totales y tasa de amenazas, enero-mayo y periodo de cuarentena 2016-2020



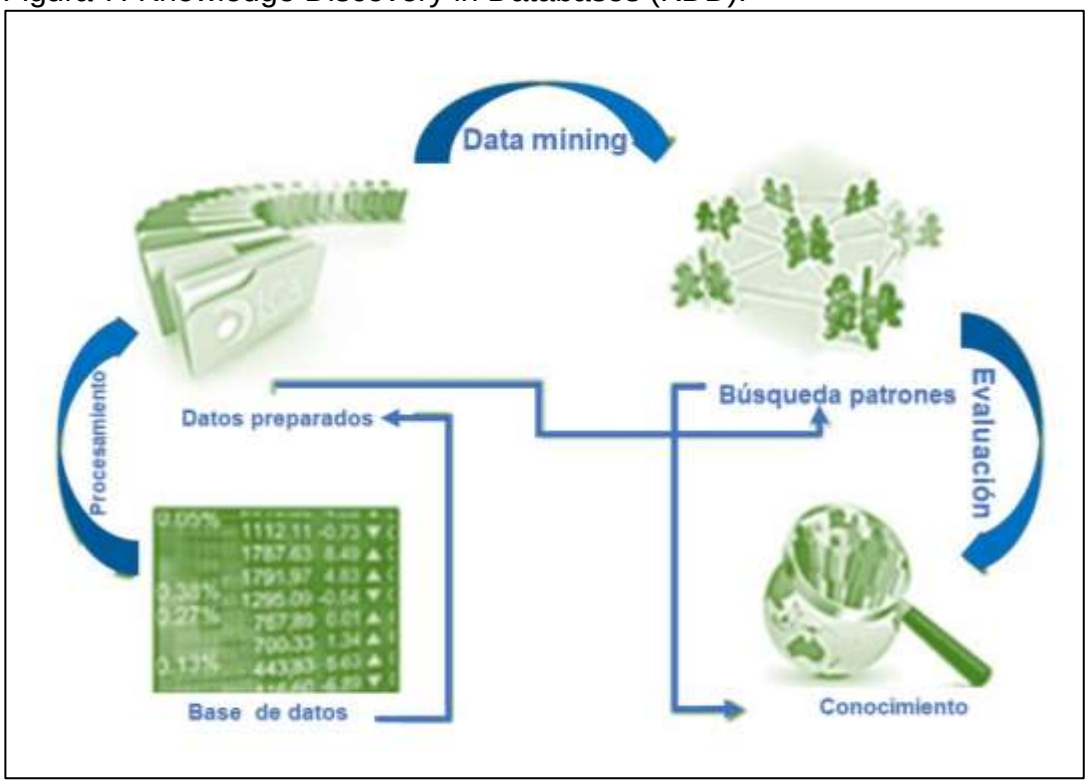
Fuente. (Jerónimo Castillo, 2020)

2.2.2 Descubrimiento de conocimiento de bases de datos:

2.2.2.1 (KDD, Knowledge Discovery in Databases). El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas compañías industriales como una [2]importante área y una oportunidad para obtener mayores ganancias (Timarán, 2009). Autores como Fayyad, Piatetsky-Shapiro y Smith (1996, p.89) lo definen como “El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos”. El Descubrimiento de conocimiento en bases de datos (kdd, del inglésKnowledge Discovery in Databases) es básicamente un proceso automático en el que se combinan descubrimiento y análisis, como lo indica la figura 7. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice.Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (data mining) y presentar resultados (Agrawal y Srikant, 1994) (Chen, Han y Yu, 1996) (Piatetsky Shapiro, Brachman y Khabaza, 1996) (Han y Kamber, 2001). Kdd se puede aplicar en diferentes dominios, por ejemplo, para determinar perfiles de clientes fraudulentos (evasión de impuestos), para descubrir relaciones implícitas existentes entre síntomas y enfermedades, entre características técnicas y diagnóstico del estado de equipos y máquinas, para determinar perfiles de estudiantes “académicamente exitosos” en términos de sus características

socioeconómicas y para determinar patrones de compra de los clientes en sus canastas de mercado[41].

Figura 7. Knowledge Discovery in Databases (KDD).



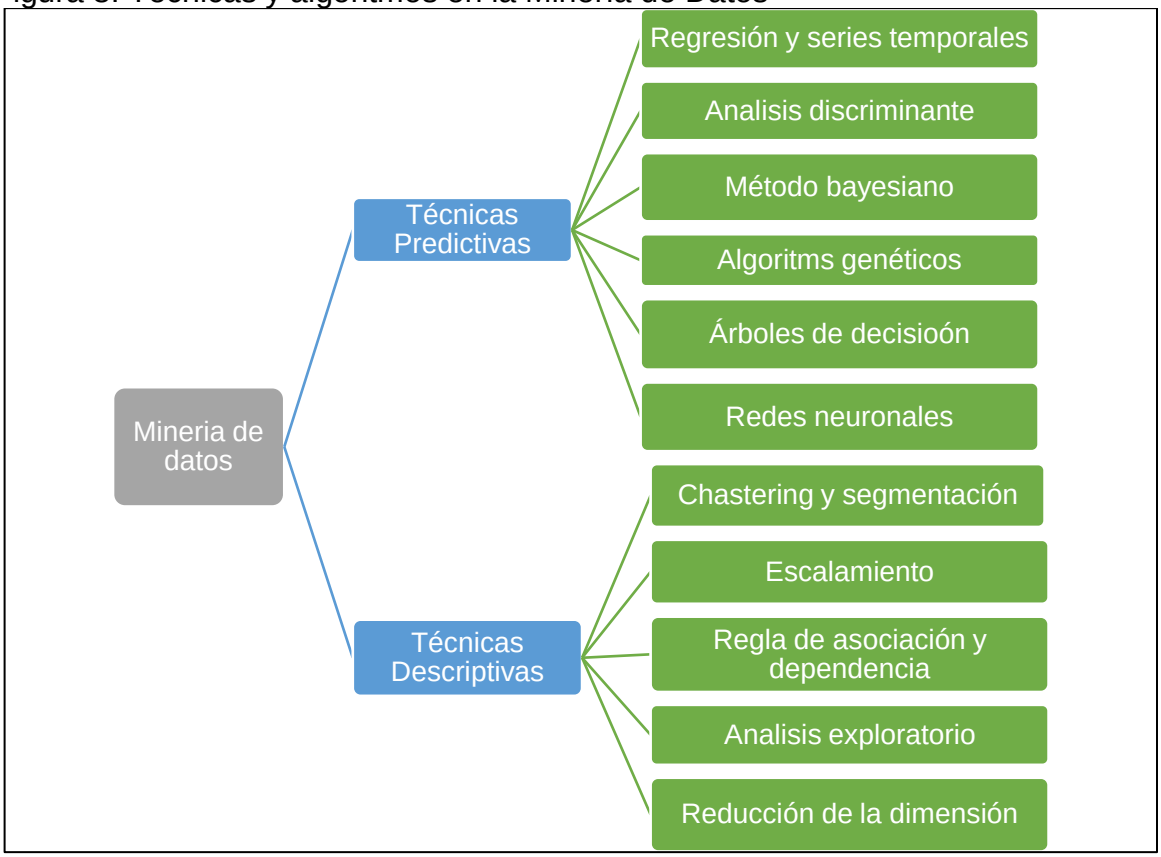
Fuente. (Han y Kamber, 2017)

Minería de datos: el objetivo de la etapa minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación (Quinlan, 1986) (Wang, Iyer y Scott, 1998), clustering (Ng y Han, 1994), (Zhang, Ramakrishnan, Livny, 1996), patrones secuenciales (Agrawal y Srikant, 1995) y asociaciones (Agrawal y Srikant, 1994), (Srikant y Agrawal, 1996), entre otras. Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas, como por ejemplo predecir para nuevos clientes si son buenos o malos basados en su estado civil, edad, género y profesión, o determinar para nuevos estudiantes si desertan o no en función de su zona de procedencia, facultad, estrato, género, edad y promedio de notas. Entre las tareas predictivas

[41] T.-P. e. al., «El proceso de descubrimiento de conocimiento en bases de datos,» 2016. [En línea]. Available: https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230-1?inline=1.

están la clasificación y la regresión. Los modelos descriptivos identifican patrones que explican o resumen los datos; sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos, como identificar grupos de personas con gustos similares o identificar patrones de compra de clientes en una determinada zona de la ciudad. Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los clustering y las correlaciones. Por lo tanto, la escogencia de un algoritmo de minería de datos incluye la selección de los métodos por aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) por utilizar[42].

Figura 8. Técnicas y algoritmos en la Minería de Datos



Fuente. (Hernández Orallo, 2007)

[42] Ibid.

Clasificación: la clasificación de datos permite obtener resultados a partir de un proceso de aprendizaje supervisado. Es, además, el proceso por medio del cual se encuentran propiedades comunes entre un conjunto de objetos de una base de datos y se los cataloga en diferentes clases, de acuerdo con el modelo de clasificación (Agrawal, Ghosh, Imielinsky, Iyer y Swami, 1992). Este proceso se realiza en dos pasos: en el primero se construye un modelo, en el cual cada tupla de un conjunto de tuplas de la base de datos tiene una clase conocida (etiqueta), determinada por uno de los atributos de la base de datos llamado atributo clase. El conjunto de tuplas que sirve para construir el modelo se denomina conjunto de entrenamiento y se escoge randómicamente del total de tuplas de la base de datos. A cada tupla de este conjunto se denomina ejemplo de entrenamiento (Han y Kamber, 2001). En el segundo paso se usa el modelo para clasificar. Inicialmente, se estima la exactitud del modelo utilizando otro conjunto de tuplas de la base de datos, cuya clase es conocida, denominado conjunto de prueba. Este conjunto es escogido randómicamente y es independiente del conjunto de entrenamiento. A cada tupla de este conjunto se denomina ejemplo de prueba (Han y Kamber, 2001). La exactitud del modelo, sobre el conjunto de prueba, es el porcentaje de ejemplos de prueba que son correctamente clasificadas por el modelo. Si la exactitud del modelo se considera aceptable, se puede usar para clasificar futuros datos o tuplas para los cuales no se conoce la clase a la que pertenecen. Se han propuesto varios métodos de clasificación: rough sets, árboles de decisión, redes neuronales, Bayes, algoritmos genéticos entre otros[43].

2.2.2.2 Interpolación KNN. Imagine que una computadora es un niño, somos su supervisor (por ejemplo, padre, tutor o maestro) y queremos que el niño (computadora) sepa cómo es un cerdo. Le mostraremos al niño varias imágenes diferentes, algunas de las cuales son cerdos y el resto podrían ser imágenes de cualquier cosa (gatos, perros, etc.)

Cuando vemos un cerdo, gritamos “¡cerdo!” Cuando no es un cerdo, gritamos “¡no, no cerdo!” Después de hacer esto varias veces con el niño, les mostramos una imagen y preguntamos “¿cerdo?” y ellos dicen (“la mayoría de las veces”) “¡cerdo!” o “no, no” cerdo! “dependiendo de lo que es la imagen. Eso es aprendizaje automático supervisado [44].

[43] Ibid.

[44] S. B. DATA, «Conceptos básicos de aprendizaje automático con el algoritmo de vecinos más cercanos a K,» 24 Diciembre 2019. [En línea]. Available: https://sitiobigdata.com/2019/12/24/algoritmo-de-aprendizaje-automatico-de-aprendizaje-automatico/.

2.2.3 El algoritmo KNN.

1. Carga los datos
2. Inicializa K a tu número elegido de vecinos
3. Para cada ejemplo en los datos
3.1 Calcule la distancia entre el ejemplo de consulta y el ejemplo actual a partir de los datos.
3.2 Agregue la distancia y el índice del ejemplo a una colección ordenada
4. Clasifique la colección ordenada de distancias e índices desde el más pequeño hasta el más grande (en orden ascendente) por las distancias.
5. Elija las primeras K entradas de la colección ordenada.
6. Obtenga las etiquetas de las entradas K seleccionadas.
7. Si es regresión, devuelva la media de las etiquetas K .
8. Si la clasificación es, devuelva el modo de las etiquetas K [45].

2.2.4 Arcgis. ArcGIS es un completo sistema que permite recopilar, organizar, administrar, analizar, compartir y distribuir información geográfica. Como la plataforma líder mundial para crear y utilizar sistemas de información geográfica (SIG), ArcGIS es utilizada por personas de todo el mundo para poner el conocimiento geográfico al servicio de los sectores del gobierno, la empresa, la ciencia, la educación y los medios. ArcGIS permite publicar la información geográfica para que esté accesible para cualquier usuario. El sistema está disponible en cualquier lugar a través de navegadores Web, dispositivos móviles como Smartphone y equipos de escritorio[46]

Aportes de la herramienta ArcGIS:

Crear, compartir y utilizar mapas inteligentes: los mapas constituyen una forma muy efectiva de organizar, comprender y proporcionar grandes cantidades de información de un modo comprensible universalmente. ArcGIS permite crear una amplia variedad de mapas, entre ellos, mapas Web accesibles en navegadores y dispositivos móviles, diseños de mapa impresos de gran formato, mapas incluidos en informes y presentaciones, libros de mapa, atlas, mapas integrados en aplicaciones, etc. Independientemente de cómo se publica, un mapa de ArcGIS es un mapa inteligente que muestra, integra y sintetiza completas capas de información geográfica y descriptiva de diversas fuentes[47].

[45] Ibid.
[46] «ARCGIS» ArcGis Resources Abril 2014. [En línea]. Available: https://resources.arcgis.com/es/help/getting-started/articles/026n00000014000000.htm.
[47] esri, « Aportes de la herramienta ArcGIS:» ArcGis Resources Abril 2014. [En línea]. Available: https://resources.arcgis.com/es/help/getting-started/articles/026n00000014000000.htm.

2.2.5 Geopandas. El objetivo de GeoPandas es facilitar el trabajo con datos geoespaciales en Python. Combina las capacidades de los pandas y la forma, proporcionando operaciones geoespaciales en pandas y una interfaz de alto nivel para múltiples geometrías para dar forma. GeoPandas le permite realizar fácilmente operaciones en Python que, de otro modo, requerirían una base de datos espacial como PostGIS [48].

2.2.6 Pykrige. PyKrige, es un Kit de herramientas de Kriging para Python, desarrollada por Benjamín S. Murphy (geólogo), la cual se presenta como herramienta básica a la hora de interpolar datos con previa modelización estructural.

La biblioteca si bien es escueta en cuanto a la cantidad de utilidades que posee, cumple con el propósito de generar mapas de variables a partir de muestras de una manera bastante sencilla.

Lo positivo de PyKrige, es que ha tenido constantes actualizaciones desde su lanzamiento en el año 2014, esto al punto de ser compatible con la actual versión de Python (v3.7) y bibliotecas masificadas en el análisis de datos[49].

2.2.7 Pyproj. La librería pyproj consiste en la interfaz de la librería PROJ4 de OSGeo traída a Python desde C y su uso está centrado básicamente en la proyección y conversión de geometrías entre sistemas de referencia de coordenadas.

Permite trabajar con cientos de sistemas de coordenadas distintos, realizando cálculos y transformaciones tanto cartográficas como geodésicas mediante las clases Proj y Geod respectivamente[50].

2.2.8 Rasterio. El propio nombre ofrece pistas acerca de su cometido: la librería Rasterio para Python permite leer, manipular y escribir archivos de tipo ráster.

---

[48] VILMOS, «Geopandas» ICHI.PRO, 2020. [En línea]. Available: https://ichi.pro/es/uso-de-geopandas-para-visualizacion-espacial-37278328347703.

[49] H. H. G. –. N. MINERA, "PRUEBA DEL KIT PYKRIGE EN PYTHO," Nube minera, 2019. [Online]. Available: https://nubeminera.cl/kit-pykrige-en-python/.

[50] ESTÉVES, R. «<<Librerías Python GIS para manipular y analizar datos espaciales>>,» Análisis Gis,Desarrollo Gis, 16 Septiembre 2019. [En línea]. Available: http://www.geomapik.com/desarrollo-programacion-gis/librerias-python-gis/.

Es una buena alternativa a GDAL como librería para trabajar con imágenes ráster pues permite ejecutar procesos similares escribiendo menos código, con una sintaxis algo más sencilla y elegante[51].

2.2.9 Patrones. Un patrón es una expresión, definida en un lenguaje, que describe una colección de objetos (Kumar 2013). Los patrones son usualmente expresados como combinaciones de valores de un atributo, tales como (Color= verde; Sexo = masculino; Edad = 23) o como propiedades lógicas, tales como: [Color = verde] ^ [Sexo = masculino] ^ [Edad > 23]. Se puede decir que el patrón P cubre el objeto x, o que el objeto x soporta al patrón P, si el objeto satisface la propiedad expresada por el patrón (García 2010). Una característica útil de un patrón P es la cantidad de objetos de una colección X que soportan a P, denominado soporte del patrón y se denota como soporte (P; X). En un problema de clasificación supervisada, se define un patrón discriminativo si este incluye propiedades que ayudan a la diferenciación de clases. Los patrones emergentes son un tipo de patrón discriminativo. Se dice que un patrón discriminativo es emergente si su soporte es significativamente mayor para una clase que para las demás. Adicionalmente, el soporte de un patrón discriminativo en esa clase debe ser mayor que un valor mínimo de soporte μ. La intuición detrás de utilizar un soporte mínimo es que un patrón emergente con soporte bajo puede ser ruidoso o casual, lo que puede perjudicar a la clasificación.

Por esta razón, los patrones emergentes son herramientas efectivas para resolver problemas reales en campos como Bioinformática (Pasquier 2008), análisis de flujo de datos (Alhammady 2007), detección de intrusos, reconocimiento de actividades humanas (Gu 2009), detección de anomalías en datos de conexiones de red (Ceci 2008), pronóstico de eventos raros (Gavrishchaka 2007) y extracción de relaciones de minado de espacio temporal (Celik 2006). Para medir la calidad de los patrones emergentes obtenidos por un algoritmo, es necesario analizar un amplio abanico de medidas de calidad utilizadas a lo largo de la literatura[52].

2.2.10 Máquinas de Vectores de Soporte (SVM). También son conocidas con el acrónimo SVM por sus siglas en inglés (Support Vector Machines), son herramientas fundamentales en sistemas de aprendizaje automático, permitiendo el tratamiento de problemas actuales en reconocimiento de patrones minería de datos tales como, reconocimiento y caracterización de texto manuscrito, detección ultrasónica de fallas en materiales, clasificación de imágenes médicas, sistemas

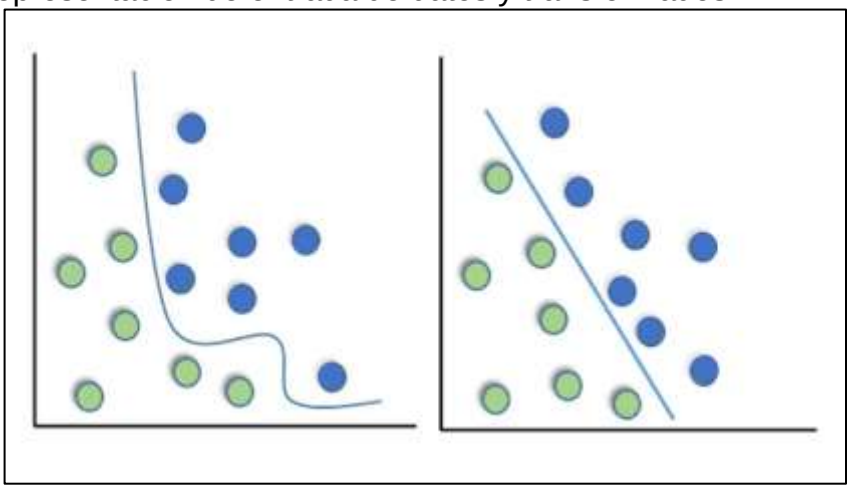
[51] Ibid.

[52] TOLEDO, A. "Métodos de selección de atributos para clasificación supervisada basados en teoría de información," researchgate, 2016. [Online]. Available: https://www.researchgate.net/publication/331155838_Metodos_de_seleccion_de_atributos_para_clasificacion_supervisada_basados_en_teoria_de_informacion.

biométricos, clasificación en bioinformática y en física de altas energías. Las SVM implementan reglas de decisión complejas, por medio de una función no lineal que permite mapear los puntos de entrenamiento a un espacio de mayor dimensión. Conceptualmente, los SVM son más fáciles de explicar para problemas de clasificación. (Chen 2006)

Vamos a usar estos datos de ejemplo. Suponemos que los puntos azules corresponden a la clase «azul» y los puntos rojos a la clase «rojo». Ahora intentaremos dibujar una línea que separe los puntos azules de los rojos. De esta forma, cuando haya un punto nuevo, podemos decir qué color va a tener, dependiendo del lado de la línea en el que se encuentre. Como se muestra en la figura 9[53].

Figura 9. Representación de entrada de datos y transformados

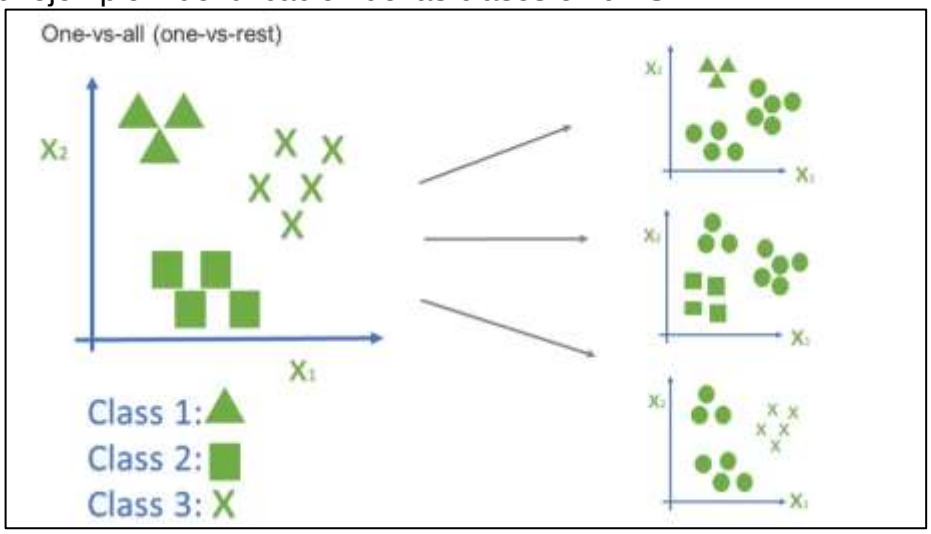


Fuente.(Yuh-jye Lee ets, 2018)

Las máquinas de vectores de soporte son una técnica de machine learning que encuentra la mejor separación posible entre clases, como se observa en la figura 9. Con dos dimensiones es fácil entender lo que está haciendo. Normalmente, los problemas de aprendizaje automático tienen muchísimas dimensiones. Así que en vez de encontrar la línea óptima, el SVM encuentra el hiperplano que maximiza el margen de separación entre clases. (Máquinas de Vectores de Soporte (SVM) - IArtificial.net, 2018).

[53] HERAS, J. M. "<<Máquinas de Vectores de Soporte (SVM)>>," IArtificial.net, 28 Mayo 2019. [Online]. Available: https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/.

Figura 10. ejemplo –identificación de las clases en un SVM.

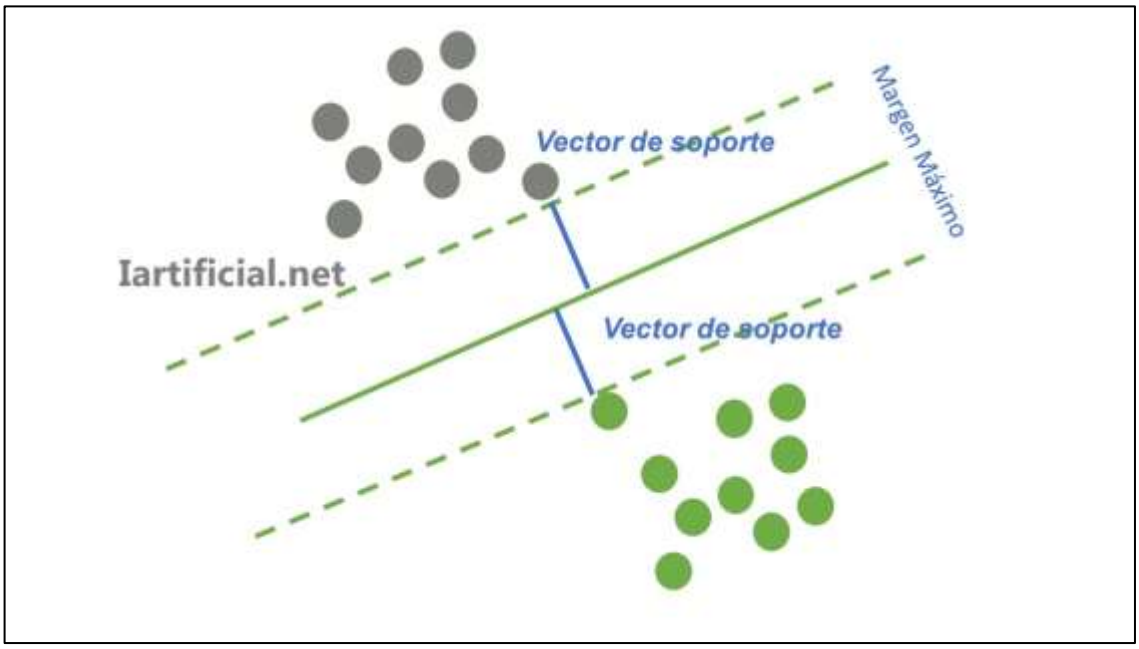


Fuente. (Máquinas de vectores de soporte (SVM) - Iartificial.Net, 2018)

2.2.10.1 ¿Por qué se llaman Máquinas de Vectores de Soporte?. Se llama «máquina» en español por la parte de «machine» learning. Los vectores de soporte son los puntos que definen el margen máximo de separación del hiperplano que separa las clases, como lo indica la figura11. Se llaman vectores, en lugar de puntos, porque estos «puntos» tienen tantos elementos como dimensiones tenga nuestro espacio de entrada. Es decir, estos puntos multi-dimensionales se representan con vector de n dimensiones[54].

[54] Ibid.

Figura 11. ejemplo, puntos multidimensionales se representan con vector de n dimensiones.



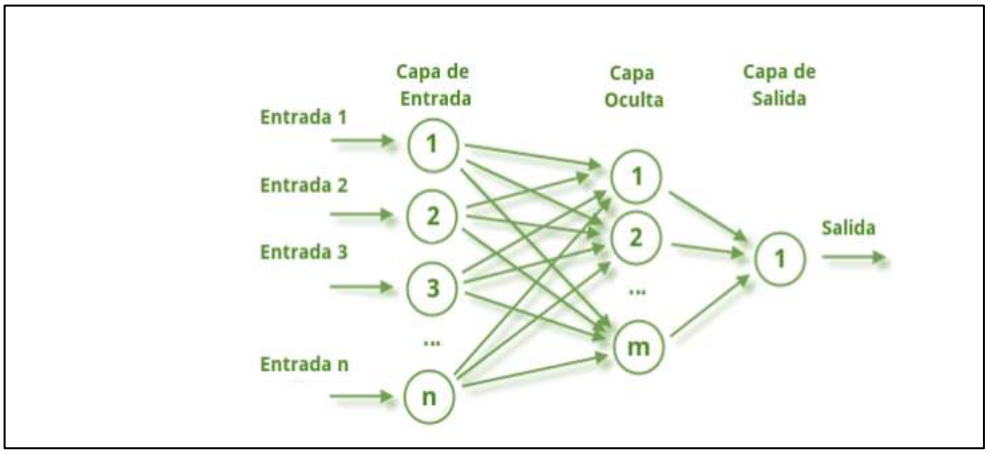
Fuente. (Máquinas de Vectores de Soporte (SVM) - IArtificial.net, 2018)

2.2.11 Redes Neuronales Artificial. Las redes neuronales artificiales son un modelo inspirado en el funcionamiento del cerebro humano. Está formado por un conjunto de nodos conocidos como neuronas artificiales que están conectadas y transmiten señales entre sí. Estas señales se transmiten desde la entrada hasta generar una salida[55].

2.2.11.1 ¿Cómo funcionan las redes neuronales? Como se ha mencionado el funcionamiento de las redes se asemeja al del cerebro humano. Las redes reciben una serie de valores de entrada y cada una de estas entradas llega a un nodo llamado neurona. Las neuronas de la red están a su vez agrupadas en capas que forman la red neuronal. Cada una de las neuronas de la red posee a su vez un peso, un valor numérico, con el que modifica la entrada recibida. Los nuevos valores obtenidos salen de las neuronas y continúan su camino por la red. Este funcionamiento puede observarse de forma esquemática en la figura 12.

[55] I. 4.0, « <<¿Qué son las redes neuronales y sus funciones?>>,» ATRIA INNOVATION, 22 Octubre 2019. [En línea]. Available: https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/.

Figura 12. Funcionamiento de las redes neuronales



Fuente: esta investigación

Las redes neuronales se han convertido en una pieza clave para el desarrollo de la Inteligencia Artificial, es uno de los principales campos de investigación y el que más está evolucionando con el tiempo, ofreciendo cada vez soluciones más complejas y eficientes[56].

Utilidad de las redes neuronales artificiales: Las redes neuronales se diferencian de otros modelos de IA en tener la capacidad de aprender en forma automática. Este proceso también es conocido como machine learning o aprendizaje de máquina.

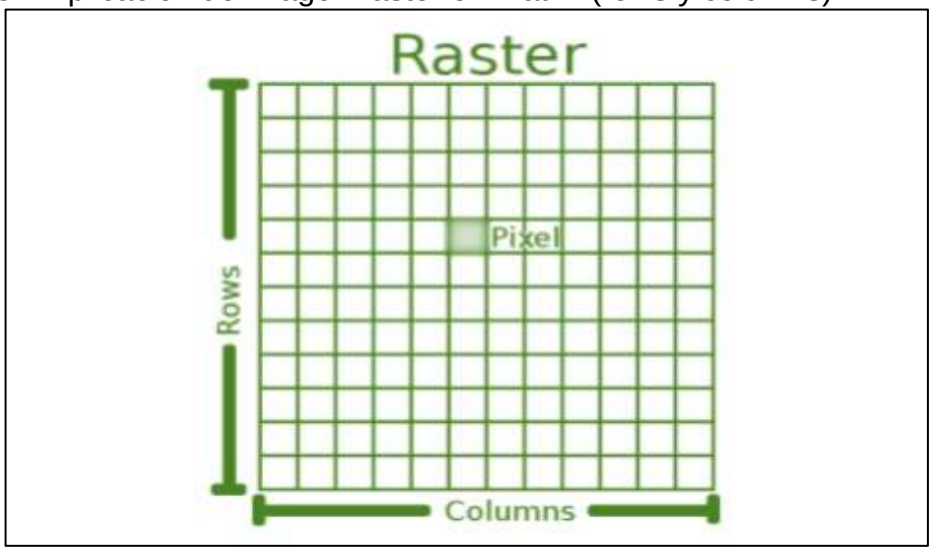
Algunas de las aplicaciones generales de las redes neuronales artificiales son:

- Sistemas inteligentes para la toma de decisiones en la gestión empresarial.
- Predicción.
- Reconocimiento de tendencias.
- Reconocimiento de patrones y gestión de riesgo, aplicados por ejemplo en la detección de fraude.
- Artefactos inteligentes con capacidad de aprendizaje, por ejemplo, los homepods o altavoces inteligentes.
- Hogar inteligente o domótica.
- Sistemas de visión computacional y detección.
- **Vehículos autónomos y energías renovables**[57]**.**

[56] I. 4.0, « <<¿Cómo funcionan?>>,» ATRIA INNOVATION, 22 Octubre 2019. [En línea]. Available: https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/.
[57] ROBERTO, C. << Utilidad de las redes neuronales artificiales>> - Thinkbig," Telefonica Tech , 12 febrero 2020. [Online]. Available: https://empresas.blogthinkbig.com/redes-neuronales-artificiales/.

2.2.12 Imágenes Raster. Los datos raster se componen de píxeles (también conocidos como celdas de la cuadrícula). Por lo general son cuadradadas y están regularmente espaciadas, pero no tiene por qué. La malla define el espacio geográfico como una matriz de puntos de cuadrícula cuadrados de igual tamaño dispuestos en filas y columnas. Cada punto de la cuadrícula almacena un valor numérico que representa un atributo geográfico (tales como elevación o superficie de la pendiente) para esa unidad de espacio. Cada celda de la malla se referencia por sus coordenadas x e y [58].

Figura 13. Explicación de imagen raster en matriz (rows y columns)



Fuente. (Datos Raster, 2016)

Un conjunto de datos ráster está compuesto de filas (corriendo de un lado a otro) y columnas (corriendo hacia abajo) de píxeles (también conocidos como celdas). Cada píxel representa una región geográfica, y el valor en ese píxel representa alguna característica de dicha [2]región.

2.2.12.1 Características de imágenes Raster:

- Los datos raster consisten en una cuadrícula de píxeles de tamaño regular.
- Los datos ráster son buenos para mostrar Información que varía continuamente.
- El tamaño del pixel en una imagen raster determina su resolución espacial.
- Las imágenes ráster pueden contener una o más bandas, cada una cubre la misma área espacial, pero contiene diferente información.

[58] AURELIO, M. « Los formatos GIS ráster más populares,» MmappingGIS, 17 Diciembre 2015. [En línea]. Available: https://mappinggis.com/2015/12/los-formatos-gis-raster-mas-populares/.

- Cuando los datos ráster contienen bandas de diferentes partes de espectro electromagnético, estas se llaman imágenes multiespectrales.
- Tres de las bandas de una imagen múlti-espectral se puede mostrar en los colores Rojo, Verde y Azul para que podamos verlas.
- Las imágenes compuestas de una sola banda se denominan imágenes en escala de gris.
- Las imágenes de banda única o de escala de grises, se pueden mostrar con pseudocolores por GIS.
- Las imágenes de raster pueden consumir grandes cantidades de espacio de almacenamiento[59].

2.2.13 Proyecciones. Proyección procede del latín proiectio y hace mención al accionar y a los resultados de proyectar (provocar el reflejo de una imagen ampliada en una superficie, lograr que la figura de un objeto se vuelva visible sobre otro, desarrollar una planificación para conseguir algo).

Proyección procede del latín proiectio y hace mención al accionar y a los resultados de proyectar (provocar el reflejo de una imagen ampliada en una superficie, lograr que la figura de un objeto se vuelva visible sobre otro, desarrollar una planificación para conseguir algo)[60].

2.2.14 Sobreajuste. Este concepto es uno de los conceptos clave en aprendizaje automático. Se denomina sobreajuste al hecho de hacer un modelo tan ajustado a los datos de entrenamiento que haga que no generalice bien a los datos de test.

Hay que recordar que el objetivo de los modelos de aprendizaje automático es el de obtener patrones de los datos de entrenamiento disponibles de cara a predecir o inferir correctamente datos nuevos. Es decir, el concepto clave es el de entrenar y obtener patrones generales que sean extrapolables a nuevos datos. Algo similar ocurre en el aprendizaje de los seres humanos, el sobreajuste se produciría cuando aprendemos las cosas de memoria, sin entender el concepto[61].

---

[59] Ibid.

[60] RENATA, A. «<<Respuesta -Proyección>>,» BRAINLY, 10 Febrero 2021. [En línea]. Available: https://brainly.lat/app/profile/19619938/answers.

[61] ÁLVARO, "<<¿Qué es el sobreajuste u overfitting y por qué debemos evitarlos?>>," MachineLearningParaTodos.com, 25 Mayo 2020. [Online]. Available: https://machinelearningparatodos.com/que-es-el-sobreajuste-u-overfitting-y-por-que-debemos-evitarlo/.

2.2.15 Error cuadrático medio (RMSE). El error cuadrático medio (RMSE) mide la cantidad de error que hay entre dos conjuntos de datos. En otras palabras, compara un valor predicho y un valor observado o conocido.

También se lo conoce como Raíz de la Desviación Cuadrática Media y es una de las estadísticas más utilizadas en SIG.

A diferencia del error absoluto medio (MAE), utilizamos RMSE en una variedad de aplicaciones cuando comparamos dos conjuntos de datos[62].

$$RMSE = \sqrt{\sum \frac{(Ypred - Yref)^2}{N}}$$

Ecuación 6. Error cuadrático medio (RMSE), También se lo conoce como Raíz de la Desviación Cuadrática Media y es una de las estadísticas más utilizadas en SIG.

2.2.16 Error Absoluto Medio (MAE). En MAE, el error se calcula como un promedio de diferencias absolutas entre los valores objetivo y las predicciones. El MAE es una puntuación lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio. Por ejemplo, la diferencia entre 10 y 0 será el doble de la diferencia entre 5 y 0. Sin embargo, lo mismo no es cierto para RMSE. Matemáticamente, se calcula utilizando esta fórmula:

$$\boldsymbol{MAE} = \frac{1}{N} \sum_{i=1}^{N} |yi - \hat{y}i|$$

Ecuación 7. Error absoluto medio (MAE), es una medida de errores entre observaciones emparejadas que expresan el mismo fenómeno**.**

Lo importante de esta métrica es que penaliza errores enormes que no tan mal como lo hace MSE. Por lo tanto, no es tan sensible a los valores atípicos como el error cuadrático medio[63] .

[62] GABRI [29], "¿Qué es el error cuadrático medio RMSE?," acolita, 22 mayo 2018. [Online]. Available: https://acolita.com/que-es-el-error-cuadratico-medio-rmse/.
[63] S. DATA, «Aprendizaje automático y las Métricas de regresión,» sitiobigdata.com, 27 Agosto 2018. [En línea]. Available: https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/

2.2.17 Coeficiente de determinación. El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. Es también denominado R cuadrado y sirve para reflejar la bondad del ajuste de un modelo a la variable que se pretende explicar.

El coeficiente de determinación puede adquirir resultados que oscilan entre 0 y 1. Así, cuando adquiere resultados más cercanos a 1, mayor resultará el ajuste del modelo a la variable que se pretende aplicar para el caso en concreto. Por el contrario, cuando adquiere resultados que se acercan al valor 0, menor será el Ajuste del modelo a la variable que se pretende aplicar y, justo por eso, resultará dicho modelo menos fiable[64].

$$\overline{R}^2 = 1 - \left(\frac{T-1}{T-K} \right)(1 - R^2)$$

Ecuación 8. Coeficiente de determinación, adquiere resultados que oscilan entre 0 y 1.

2.2.18 Coeficiente de Correlación de Person. El coeficiente de correlación de Pearson es una prueba que mide la relación estadística entre dos variables continuas. Si la asociación entre los elementos no es lineal, entonces el coeficiente no se encuentra representado adecuadamente.

El coeficiente de correlación puede tomar un rango de valores de +1 a -1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva. Es decir, a medida que aumenta el valor de una variable, también lo hace el valor de la otra. Un valor menor que 0 indica una asociación negativa; es decir, a medida que aumenta el valor de una variable, el valor de la otra disminuye[65].

[64] FRANCISCO, L. «Coeficiente de determinación (R cuadrado),» economipedia, 02 Octubre 2017. [En línea]. Available: https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html.

[65] QUESTIONPRO, "<<¿Qué es el coeficiente de correlación de Pearson?>>," QuestionPro, 28 mayo 2019. [Online]. Available: https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/

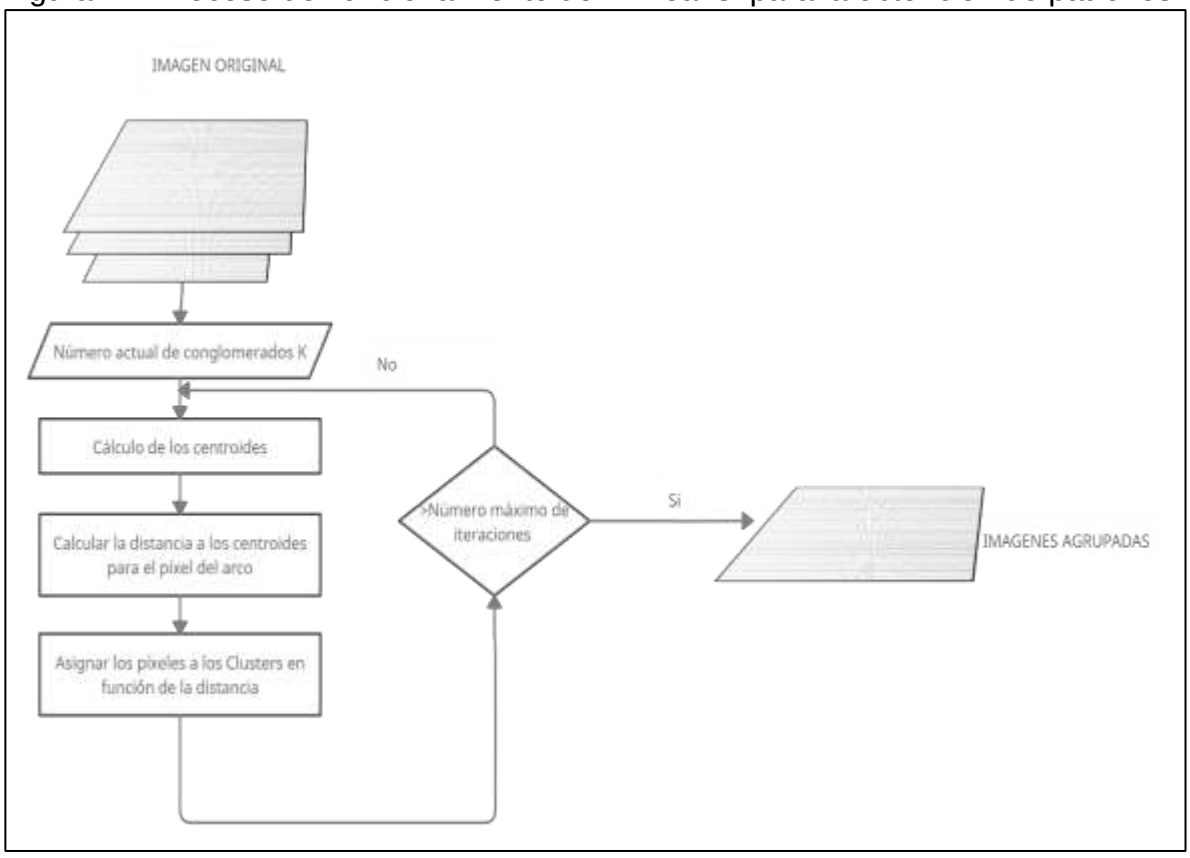
$$r = \frac{n..\sum xi - yi - \sum xi - \sum yi}{\sqrt{[n.\sum xi^2 - (\sum xi)^2].[n.\sum xi^2 - (\sum xi)^2]}}$$

$$-1 \leq r \leq 1$$

Ecuación 9. Coeficiente de correlación de Person, es una prueba que mide la relación estadística entre dos variables continuas.

2.2.19 Algoritmo k-means.

Figura 14. Proceso del funcionamiento de K-means para la obtención de patrones



Fuente**.** (Tripatía et.,2015)[66]

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se

[66] TRIPATÍA, B. Sudhir Kumar Sahu, Kamal Kumar Barik . (Enero, 2015). Una máquina de vectores de soporte de clasificación binaria y segmentación de imágenes de datos de teledetección de Chilika Lagloon. Obtenido de: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

El algoritmo consta de tres pasos:

Inicialización: una vez escogido el número de grupos, k, se establecen k centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.

Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano.

Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

El algoritmo k-means resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster.

Los objetos se representan con vectores reales de d dimensiones (x1,x2,…..,xn) y el algoritmo k-means construye k grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo S={S_1,S_2,……,S_K }, a su centroide. El problema se puede formular de la siguiente forma:

$$\min_{s} \boldsymbol{E}\,(\boldsymbol{\mu i}) = \min_{s} \sum_{i=1}^{k} \sum_{x_j \epsilon s_i} \left\| \boldsymbol{x_j} - \boldsymbol{\mu_i} \right\|^2 (1)$$

Ecuación 10. Suma de las distancias cuadráticas de cada objeto al centroide del cluster**.**

Donde S es el conjunto de datos cuyos elementos son los objetos X_j representados por vectores, donde cada uno de sus elementos representa una característica o atributo.

Tendremos k grupos o clusters con su correspondiente centroide μ_i. En cada actualización de los centroide, desde el punto de vista matemático, imponemos la condición necesaria de extremo a la función E(μ_i) que, para la función cuadrática (1) es:

$$\frac{\partial E}{\partial \mu_i} = 0 \Longrightarrow \mu_i^{(t+1)} = \frac{1}{|S_t^{(t)}|} \sum_{x_{j \epsilon S_i}^{(t)}} X_i$$

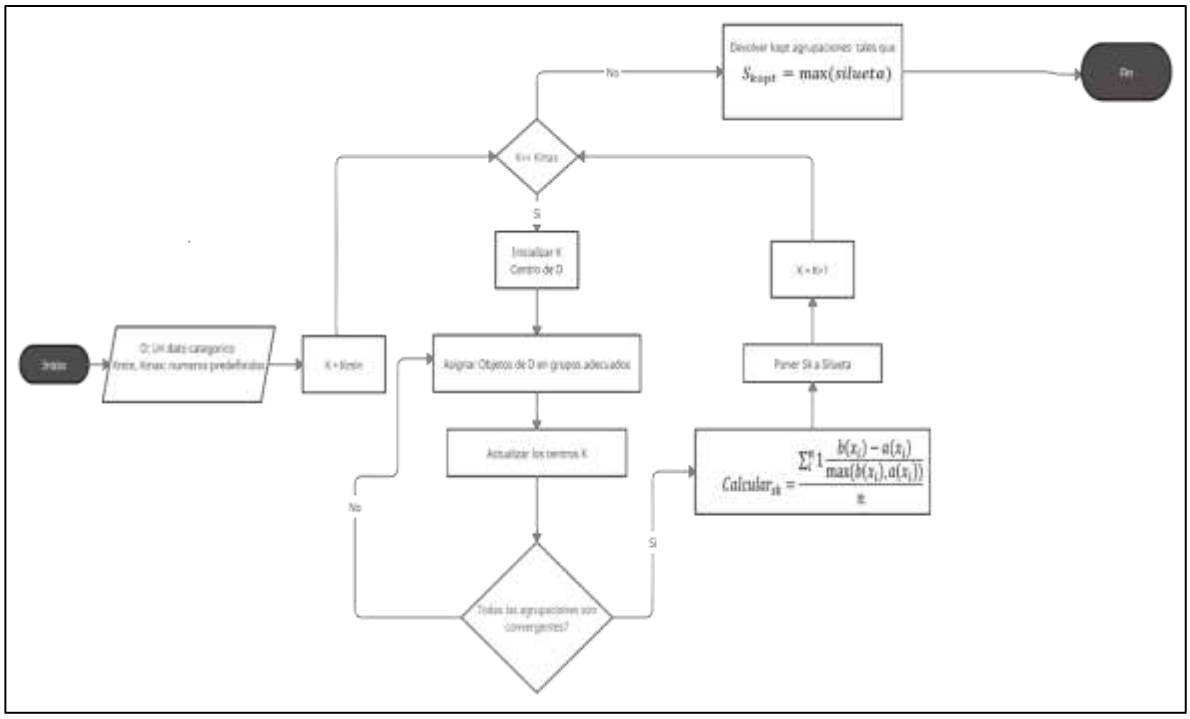
Ecuación 11. E(μi) Centroides

Y se toma el promedio de los elementos de cada grupo como nuevo centroide. Las principales ventajas del método k-means son que es un método sencillo y rápido. Pero es necesario decidir el valor de k y el resultado final depende de la inicialización de los centroide. En principio no converge al mínimo global sino a un mínimo local[67].

[67] UNIOVIEDO. (2020). Unioviedo. obtenido de unioviedo: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

2.2.20 Algoritmo K-modas.

Figura 15. Proceso del funcionamiento de K-modas para la obtención de patrones



Fuente. (Dinh,2019)[68]

El algoritmo k-modas es una versión del kmedias para datos categóricos. En k-modas se hacen 3 modificaciones a k-medias:

- Uso de diferentes medidas de disimilaridad.
- Sustitución de k medias por k modas para formar los centros.
- El método basado en las frecuencias de los datos para actualizar las modas.

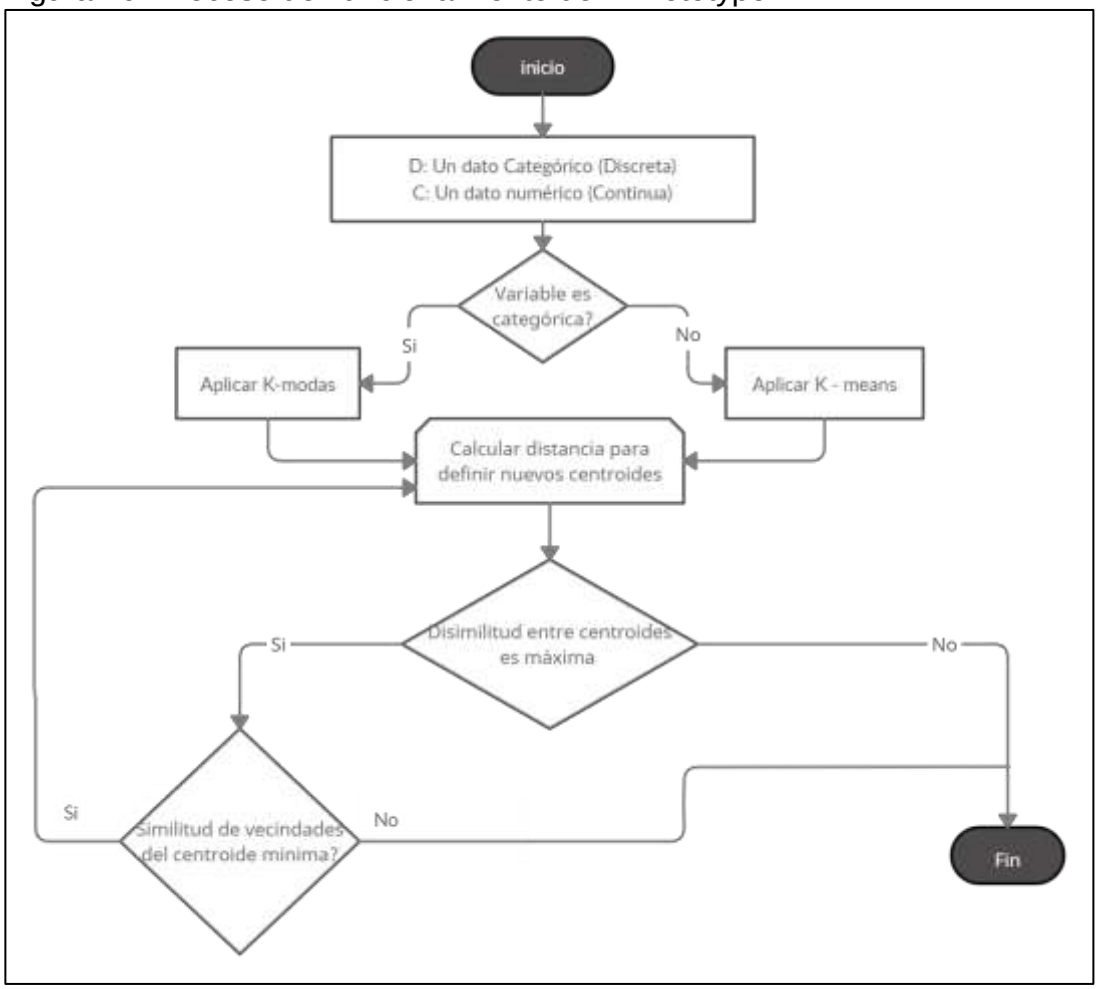
La actualización de las modas se realiza en cada asignación de un objeto a su grupo, mientras que en k-medias es al final de cada iteración del algoritmo.

[68]TAI DINH, Tsutomu Fujinami, Van-Nam Huynh (Noviembre 2019), “Estimación del número óptimo de clústeres en agrupamiento de datos categóricos por coeficiente de silueta” Obtenido de: https://www.researchgate.net/publication/336980455_Estimating_the_Optimal_Number_of_Clusters_in_Categorical_Data_Clustering_by_Silhouette_Coefficient

El algoritmo k-modas al igual que el algoritmo k-medias produce soluciones óptimas locales, que dependen del conjunto de modas iniciales y el orden de los objetos en el conjunto de datos[69]

2.2.21 Algoritmo k-Prototype.

Figura 16. Proceso del funcionamiento de k-Prototype



Fuente. esta investigación

El algoritmo **k − Prototype** es un algoritmo de agrupamiento restringido que permite agrupar grandes conjuntos de datos mezclados. Este algoritmo básicamente

[69] RENDÓN, Eréndira†; ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez. (05 julio 2015). El algoritmo de agrupamiento K-Modas. Obtenido de: https://www.ecorfan.org/bolivia/researchjournals/Tecnologia_e_innovacion/vol2num5/Tecnologia_e_Innovacion_Vol2_Num5_2.pdf

constituye una integración de los algoritmos k-Modes y k-Means. El algoritmo k-Modes fue la primera extensión del algoritmo k-Means orientada al agrupamiento de datos categóricos. Sigue la misma idea que el algoritmo k-Means y la estructura del algoritmo no cambia, siendo la principal diferencia la medida de similitud usada para comparar objetos.

Las principales características de este algoritmo son:

- Usa una medida de disimilaridad para comparar objetos.
- Reemplaza el uso de promedios por el de modas.
- Usa un método basado en frecuencias para actualizar las modas.

El algoritmo k-Modes fue diseñado para agrupar grandes conjuntos de datos categóricos exclusivamente.

El algoritmo k-Prototypes integra al algoritmo k-Means y k-Modes para remover la limitación de poder trabajar ´únicamente con un solo tipo de datos.

Esto lo hace de la siguiente forma: se asume que s r es la medida de disimilaridad entre atributos numéricos definida por el cuadrado de la distancia Euclidiana y s c es la medida de disimilaridad entre atributos categóricos definida por el número de coincidencias (mismatches) de categorías entre objetos. La disimilaridad entre dos objetos se define como:

$$\boldsymbol{S^r + \gamma S^c}$$

Ecuación 12. Disimilaridad entre dos objetos

En donde γ es un peso usado para equilibrar las dos partes, lo que evita favoritismo entre los dos tipos de atributos. Un pequeño valor de γ indica que el agrupamiento está dominado por los atributos numéricos, mientras que un valor grande implica que los atributos categóricos dominan el agrupamiento.

Los algoritmos k-Prototypes y k-Modes son demasiado inestables debido a la no unicidad de las modas, es decir, el resultado depende fuertemente de la selección de las modas durante el proceso de agrupamiento. Por lo que una mala elección de la moda puede llevar a errores en el agrupamiento y considerar todas las modas implica un alto costo computacional. Esto se debe a que un solo valor de un atributo,

con la frecuencia más alta, no es suficiente para representar efectivamente la distribución del atributo en el agrupamiento. Asimismo, estos algoritmos heredan uno de los principales problemas del algoritmo k-Means el cual consiste en una fuerte dependencia de las condiciones iniciales[70].

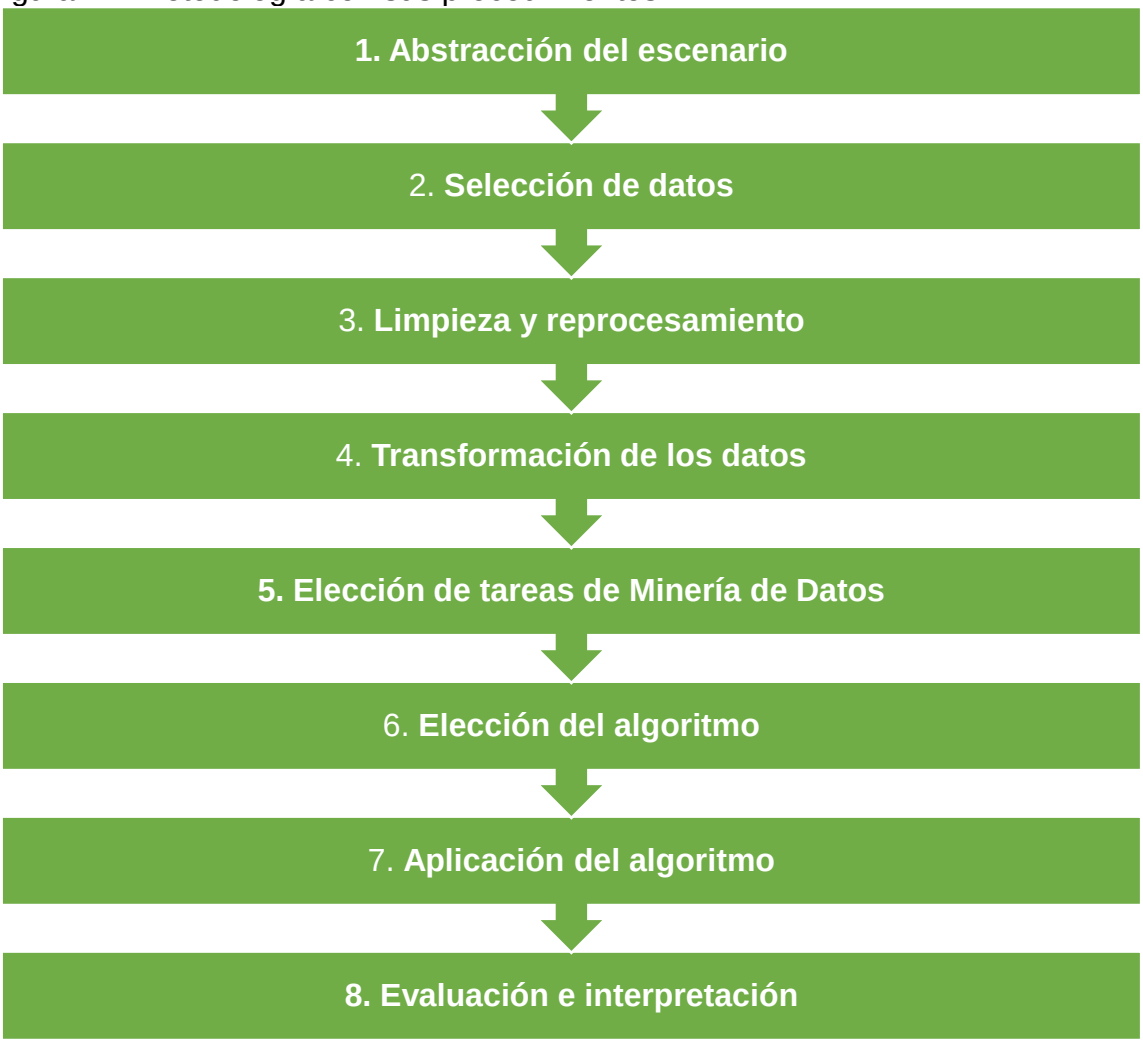
[70] ESCOBAR, S. L. Algoritmos de Agrupamiento. Recuperado el 06 de agosto de 2021, de https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/628/1/LopezES.pdf

3. METODOLOGÍA

Para el desarrollo de este estudio se va a trabajo con la metodología KDD, en realidad es el núcleo de todo un proceso llamado Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases – KDD), el cual es un proceso metodológico para encontrar un “modelo” válido, útil y entendible que describa patrones de acuerdo a la información, y como modelo entendemos que es la representación que intenta explicar ese patrón en los datos.

Es importante mencionar que hablar de “modelo” como fórmula mágica no significa que existe un modelo para cualquier problemática, sino todo lo contrario, pues existen muchos métodos o algoritmos que podrían satisfacer las necesidades dependiendo de los objetivos del estudio y de los datos que se quieran analizar. Es por esta razón que un requisito para poder adentrarse en esta área es tener conocimiento de conceptos de Estadística (LANDA, 2016).

Figura 17. Metodología con sus procedimientos



Fuente. esta investigación

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

3.1 ABSTRACCIÓN DEL ESCENARIO

Para entender la problemática y entender el contexto, se analizaron las bases de datos obtenidas por el Observatorio del Delito y se tomó como escenario de estudio la zona urbana de dicho municipio, excepto la zona rural de la Guayacana ya que es un lugar de Tumaco en el que se presenta gran parte de los delitos cometidos y además se encuentra relativamente cerca al casco urbano del municipio.

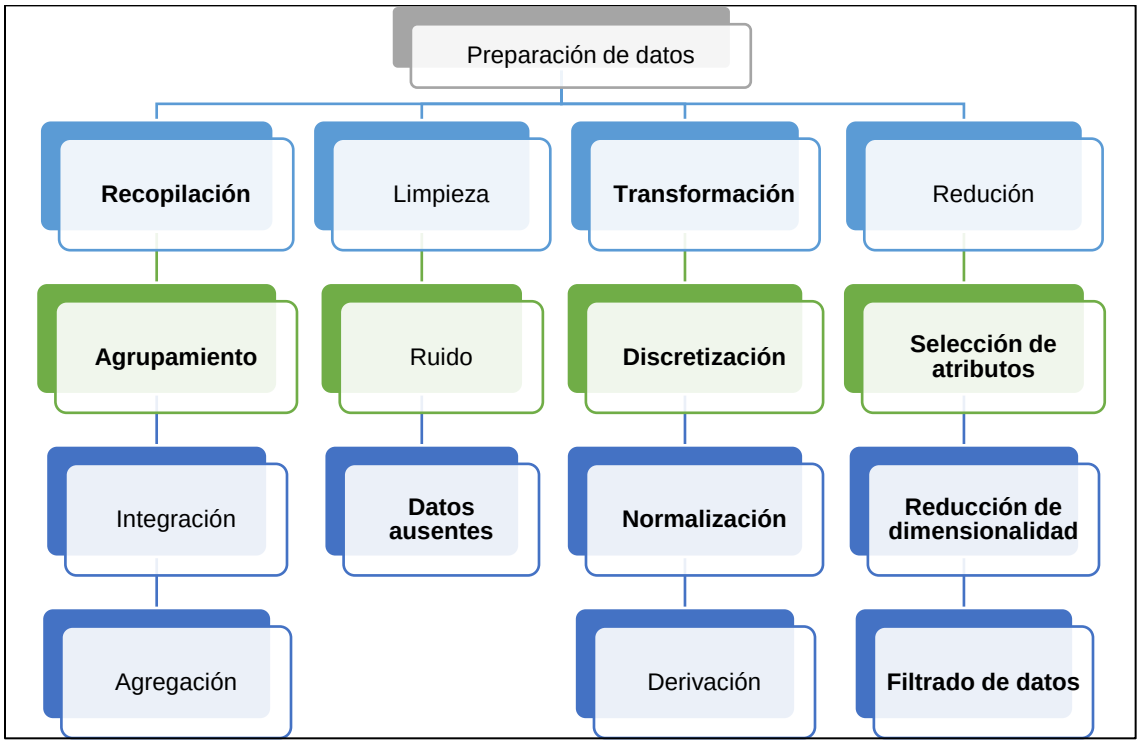
3.2 SELECCIÓN DE LOS DATOS

Como se mencionó anteriormente la información se extrajo del Observatorio del Delito, dado a que es la dependencia es el área de investigación criminología encargada de consolidar, procesar y difundir los registros administrativos con fines estadísticos de delitos de los cuales se descargaron 45 bases de datos, por un rango de año desde el 2010 al 2019, una vez descargada dicha información, se dejó las variables coincidentes por delito, se consolidaron las bases de datos, y se filtró para dejar solo la información del municipio de Tumaco. Para realizar un mayor tratamiento de los datos, se creó una base de datos con todos los barrios de la zona urbana del municipio de Tumaco con sus respectivos datos geográficos, “latitud”, “longitud” y por otro lado se consiguió el polígono del municipio de Tumaco, (Shapefile de la zona urbana de Tumaco).

3.3 LIMPIEZA Y PREPROCESAMIENTO

En esta etapa se determinó la confiabilidad de la información, para ello se transformaron y eliminaron varias columnas, entre ellas la variable “fecha” la cual se transformó por, “día”, ”mes” y “hora”, la variable “hora” por “dia”, ”tarde” y “noche” y también se quitaron otras variables por no ofrecer buena calidad en su distribución de datos, como lo son las variables “departamento”, ”municipio”, ”zona”, ”codigo dane” y ”delito”, ya que estas variable no brindaban información necesaria porque sus datos eran repetitivos, carencia de variabilidad y calidad.

Figura 18. Proceso de preparación de datos



Fuente**.** esta investigación

3.4 TRANSFORMACIÓN DE LOS DATOS

En esta etapa se mejorarón la calidad de los datos con transformaciones que involucrarón 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.

3.5 SELECCIÓN DE LA APROPIADA TAREA DE MINERÍA DE DATOS

En esta fase de minería de datos se eligieron algoritmos de regresión y agrupación (clustering) a aplicar.

3.6 ELECCIÓN DEL ALGORITMO DE MINERÍA DE DATOS

Una vez seleccionadas las tareas de minería de datos, se entrenaron los modelos sintonizando manualmente los hiperparámetros y se utilizó; KNN, SVM, Arboles de decisión, Redes Neuronales, 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.

3.7 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 luego se tomó de los mejores modelos los que brindaban el mejor balance entre grupos.

3.8 APLICACIÓN E INTERPRETACIÓN

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

4. PLAN DE ACCIÓN

Tabla 1. Plan de acción

Fases	Actividad	Entregables	Recursos
FASE 1: Preprocesamiento de datos.	Seleccionar las herramientas para el almacenamiento y manipulación de datos.	Datos depurados y obtención de las variables más relevantes, recopilados en un sistema gestor de base de datos. Conjuntos de entrenamiento.	**Internet** **Computador** **Instalar datacleaning.** **Visitar página web-observatorio de delitos.** **Lenguaje de programación Python.** **Motor de base de datos (PostgreSQL).** **Software pare la gestión de hojas cálculo.**
	Adquirir los datos sobre actividad delictiva del observatorio del delito**.**		
	Estudiar los metadatos sobre actividad delictivas brindadas por el observatorio del delito del municipio de Tumaco.		
	Seleccionar aquellos datos que se encuentre dentro del rango a manejar.		

	Aplicar técnicas de extracción, transformación, limpieza, correlación y reducción de dimensión en los datos.		
FASE 2: Desarrollo de marco experimental	Estudiar materiales y métodos más relevantes relacionados con el objeto de estudio**.**	Diagramas Arquitectónicos del marco experimental	**Artículos, pdf. Lenguaje de programación (Python). Motor de base de datos (PostgreSQL). Internet, Computador**.
	Diseñar un marco experimental para comparación de algoritmos de aprendizaje automático**.**	Scripts para la obtención de un marco Experimental.	
	Codificar un marco experimental para la sintonización de hiperparámetros.	Configuración de modelos subóptimos**.**	
	Probar el marco experimental con un conjunto de bases de datos conocidas.	Resultados del compromiso entre la eficiencia computacional y gráficos de interpretabilidad de bases de datos conocidas.	

FASE 3: Análisis de resultados	Almacenar los agentes inteligentes con mejor compromiso en sus métricas de calidad.	Resultados del compromiso entre la eficiencia computacional y gráficos de interpretabilidad de actividad delictiva en Tumaco.	**Internet** **Computador** **Artículos, pdf.** **Lenguaje de programación Python.** **Servidor web.** **Editor de datos**.
	Obtener matrices de métricas de calidad, gráficos de interpretabilidad y mapas de interpolación.	Modelo para predicción de actividad delictiva**.**	
	Publicar el agente con mejor evaluación en sus métricas de calidad e interpretabilidad.	Mapas de interpolación de actividad delictiva. Artículo científico	

Fuente**.** esta investigación

En la figura 19: se muestra el esfuerzo que requiere cada fase del proceso de KDD.

Figura 16. Esfuerzo requerido por cada fase del proceso de KDD

Fuente**:**(Han y Kambert,2001).

4.1 PREPROCESAMIENTO DE LOS DATOS DE ACTIVIDAD DELICTIVA DEL OBSERVATORIO DEL DELITO

El proyecto se compone de tres partes fundamentales con el fin de dar cumplimiento a los objetivos planteados.

La primera consta de preprocesar los datos de actividad delictiva del observatorio del delito que permita una selección de las variables más relevantes utilizando técnicas de extracción, transformación, limpieza, correlación y reducción de dimensión (RD).

La segunda parte es desarrollar un marco experimental de comparación de algoritmos supervisados de machine learning que brinde un conjunto de modelos subóptimos de predicción de actividad delictiva mediante una sintonización de hiperparámetros.

La tercera parte es Analizar los resultados de los mejores modelos obtenidos mediante: gráficos de interpretabilidad, métricas de calidad y mapas de interpolación de la actividad delictiva en función de la ubicación geográfica.

4.1.1 Fuentes de extracción y sus variables. Preprocesar los datos de actividad delictiva del observatorio del delito se contaron con diferentes fuentes de extracción las cuales comprenden en total 45 bases de datos, formato Excel con 4 variables globales referenciadas al "tipo de delito" más 18 variables especificas descriptivas (fecha, departamento, etc.) a utilizar, que van a permitir ser materia prima para la búsqueda de datos. Por tal motivo, se procede analizar las fuentes de extracción y sus variables.

4.1.2 Fuentes de extracción. Se tomó como fuente de extracción de datos del Observatorio del Delito, ya que es un grupo estratégico del área de Investigación Criminológica, encargado del monito-reo, diagnóstico, administración de la información, evaluación y análisis de la criminalidad de 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.

Fuentes de extracción:

1. Observatorio de delitos.
2. Alcaldía Municipal de Tumaco.

4.1.3.Criterios de selección. Por consiguiente, de las anteriores fuentes de Información y con el objetivo de definir cuales se van a trabajar, se tienen en cuenta los siguientes criterios de selección que se encuentran enfocados a la calidad de datos:

- **Utilidad:** Los datos pueden utilizarse como una entrada para la obtención del mejor algoritmo. (R. Y. Wang et al., 1995). Por ejemplo: Los datos asociados con el tipo de delitos para saber el número de actividad delictiva ocurrida en los diferentes barrios del municipio de Tumaco, permiten identificar el barrio específico y la cantidad de actividades delictivas ocurridas en dicho espacio o lugar.

- **Credibilidad:** Medida en que el que se pueden utilizar los datos como una entrada de decisión que tenga validez. (R. Y. Wang et al., 1995). Por ejemplo: En la ejecución de los algoritmos supervisados, se Identifiquen los patrones con los datos que se están asignando como entrada.

- **Cantidad:** Capacidad de tener un número considerable de datos dentro de una fuente de información. Por ejemplo: El rango de los datos Consolidados a evaluar es del periodo 2010 a 2019.

- **Precisión:** Implica que los datos tienen detalles suficientes y apropiados. (Observatorio del Delito de la Policía Nacional, 2015) Por ejemplo: La obtención de los diferentes actos delictivos tienen la fecha detallada indicando el día y el año.

Criterios de Selección en cuanto a los patrones obtenidos:

- **Simplicidad,** que se mide mediante el número de patrones y el número de variables que conforman el patrón.
- **Poder discriminativo,** que se puede medir mediante el índice de crecimiento y la confianza de los patrones.
- **Generalidad,** que se mide mediante la sensibilidad y el soporte.
- **Ganancia de información,** que se puede medir mediante la atipicidad y la ganancia del patrón.

4.1.4.Selección fuentes de Extracción. Se tomó como fuente de extracción de datos al Observatorio de los delitos, dado a que es la dependencia del área de investigación criminología encargada de consolidar, procesar y difundir los registros administrativos con fines estadísticos de delitos y actividad operativa institucional de la cual se va a centrar dicha investigación.

Una vez se accede a la página principal del observatorio de delitos en Colombia, nos ubicamos en el área de ESTADISTICAS –DELICTIVAS para realizar la descarga de cada una de las bases de datos de los diferentes años que contengan relación con los tipos de actividad delictiva para el desarrollo principal de este

estudio, para lo cual se realizo una organización en carpetas de cada una de las bases de datos por “nombre de actividad delictiva” y “año”.

4.1.5 Pasos para la obtención de variables. Las variables que se contemplaron para el desarrollo, son datos que una vez descargados, se procedieron a la realización de un estudio de dicha información. Luego se Iniciaron los respectivos filtros de cada una de las bases de datos por “municipio”, de esta forma nos centramos en nuestro primer paso para la aplicación de nuestra metodología basada en KDD,**” abstracción del escenario”;** por consiguiente, es necesario delimitar nuestro campo de estudio para obtener mejores resultados.

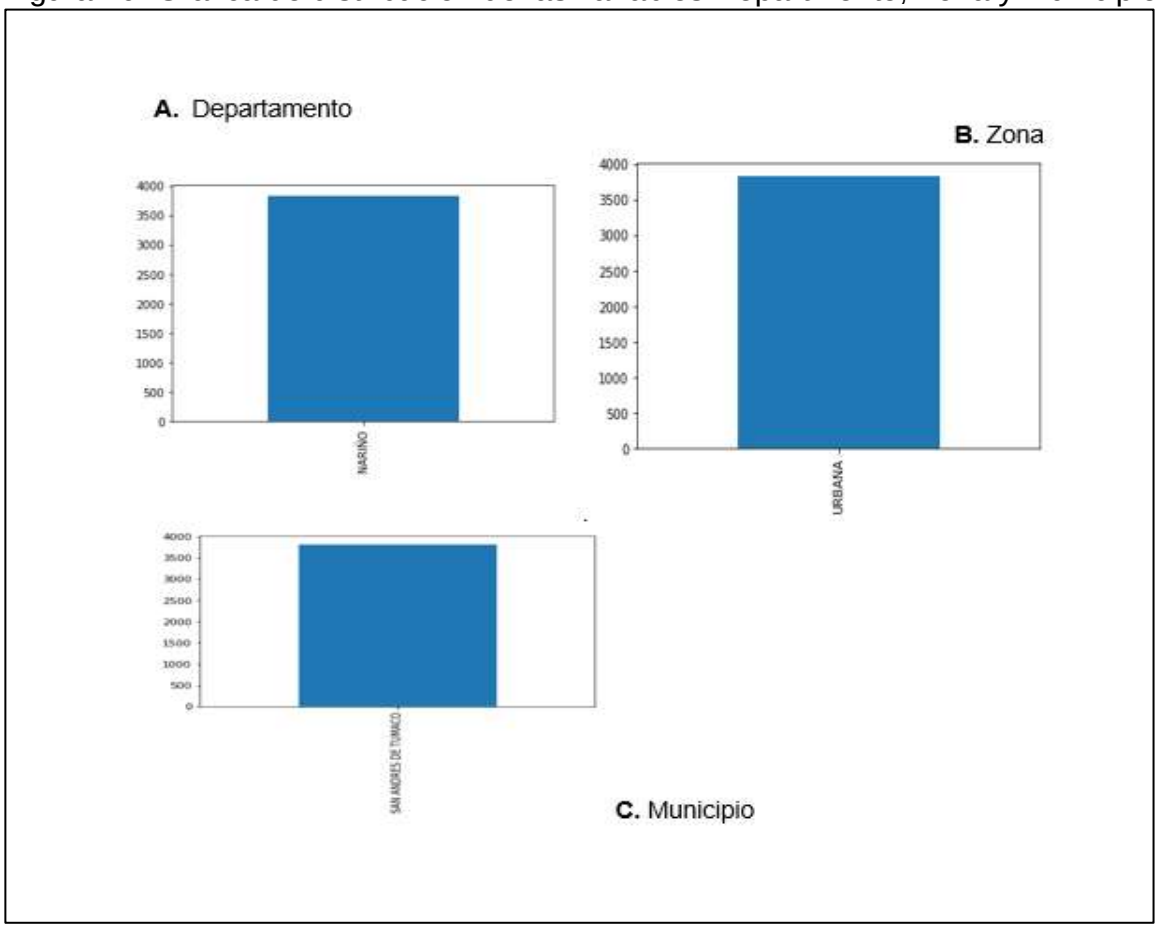
- En el estudio de toda la información suministrada en las diferentes bases de datos se observó que la zona rural del municipio de Tumaco no contaba con suficientes reportes frente a los casos de actividad delictiva ocurridas en dicha zona y no era identificada su ubicación, como tal en las diferentes bases de datos descargadas del Observatorio del delito Colombiano , por tanto solo se pudiese contar con un lugar referente o cercano al sitio donde ocurren los hechos ,dada esta situación se realizaron filtros del campo
  “zona”, centrándonos en el campo de estudio a la “Zona Urbana” del municipio de Tumaco, del cual se obtuvieron la gran cantidad de información para el desarrollo de esta investigación.

- Una vez obtenido dichos procesos, se seleccionó las bases de datos con las variables globales frente a los actos delictivos de dicho municipio, con el fin de organizarlas por año y ubicarlas en carpetas desde drive.

4.1.6 Procesamiento de limpieza de las variables. Se realizó un procesamiento de limpieza de cada una de estas variables con el fin de obtener información relevante, que aporten al desarrollo de dicho estudio, para ello se aplicó graficas de distribución de variables con el fin de analizar que variables serian mas relevantes y cuales serian innecesarias como se puede observar en las figuras 20 **A,B,C**.

En cuanto a las variable departamento, zona y municipio como se observa en la gráfica de distribución indicada en la figura 18, estas no varian, por tanto no arroja mucha información de la cual se pueda inferir algo para el desarrollo del estudio, por ello se procedio a quitarlas.

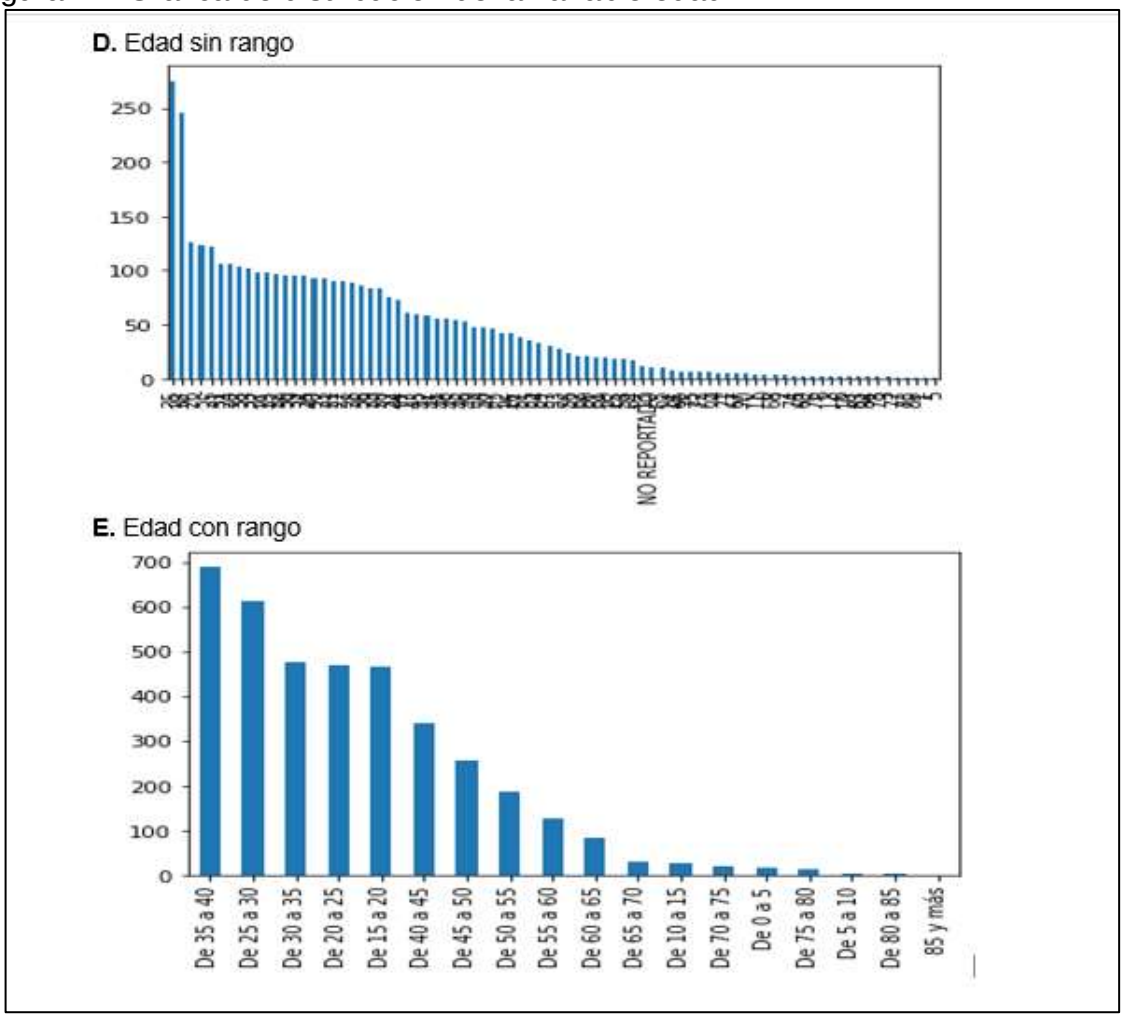
Figura 20. Gráfica de distribución de las variables Departmento, Zona y Municipio



Fuente. esta investigación

De este modo como se observa en la gráfica 19 D de distribución de la variable edad, de la cual no se arroja información relevante, dado a ello no se puede deducir algo muy claro o verídico, por tanto fue transformada aplicandole rango a la variable edad como se indica en la figura 21 E; de esta forma se obtuvo información verídica de la cual se pueda deducir algo y que sirva como aporte para el desarrollo de este estudio.

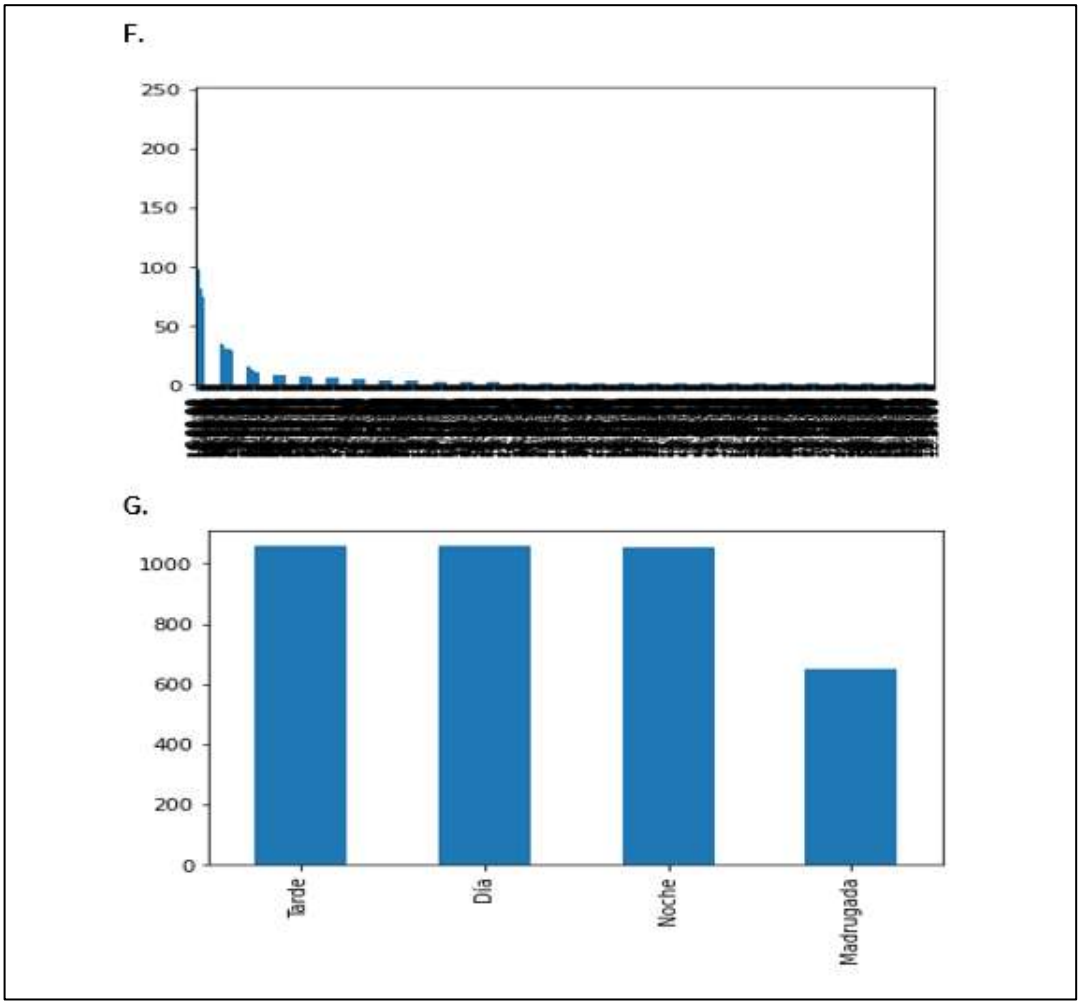
Figura 21. Gráfica de distribución de la variable edad.



Fuente. esta investigación

De esta manera la variable “hora”, como lo indica la gráfica de distribución indicada en la figura 22 F, no era posible obtener información clara, dado a esto fue necesario que la variable “hora” se transformará en “tarde”,”dia”,”noche” y “madrugada”, como se observa en la figura 20G, de esta manera se obtuvo información mas clara y útil para la solución de dicho estudio.

Figura 22. Gráfica de distribución de la variable hora.

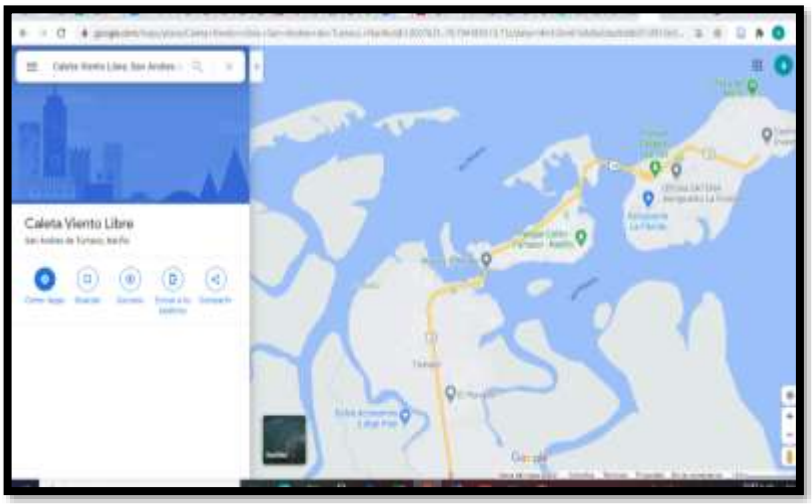


Fuente. esta investigación

4.1.7.Obtención de los datos geográficos de cada uno de los Barrios. Una vez organizadas las bases de datos, se procedió a realizar una investigación sobre cada uno de los barrios del municipio de Tumaco, para de esta forma suministrarle a cada uno de ellos los datos geográficos; como la “latitud”, “longitud”.

En la obtención de dicha información en cuanto a las latitudes y longitudes de los diferentes barrios del casco urbano del municipio de Tumaco, se procedió a la búsqueda de dicha información en Google Maps. Como lo indica en la figura 23.

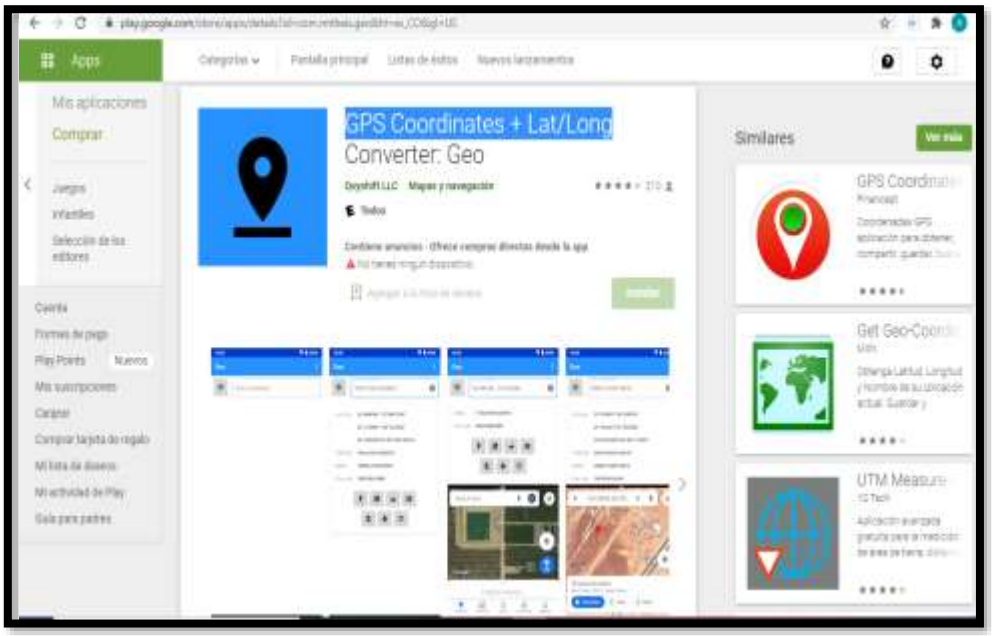
Figura 23. Google maps-obtención de datos geográficos (latitud-longitud)



Fuente. esta investigación

Dado que en dicha herramienta de Google Mapas no se encontraron suministrados todos los datos geográficos en cuanto a la “latitud” y “longitud” de los diferentes barrios del municipio de Tumaco, debido a que no se encuentra actualizada en dicha herramienta geográfica los datos de algunos barrios; se procedió a la descarga desde play store una herramienta u programa, llamada
GPS Coordinates+Lat/Long y dirigirse a cada uno de los barrios de los cuales no había registros de la “latitud” y “longitud” de forma presencial y capturar dicha información, de esta forma obtener la información geográfica para el respectivo avance de la investigación expuesta con anterioridad.

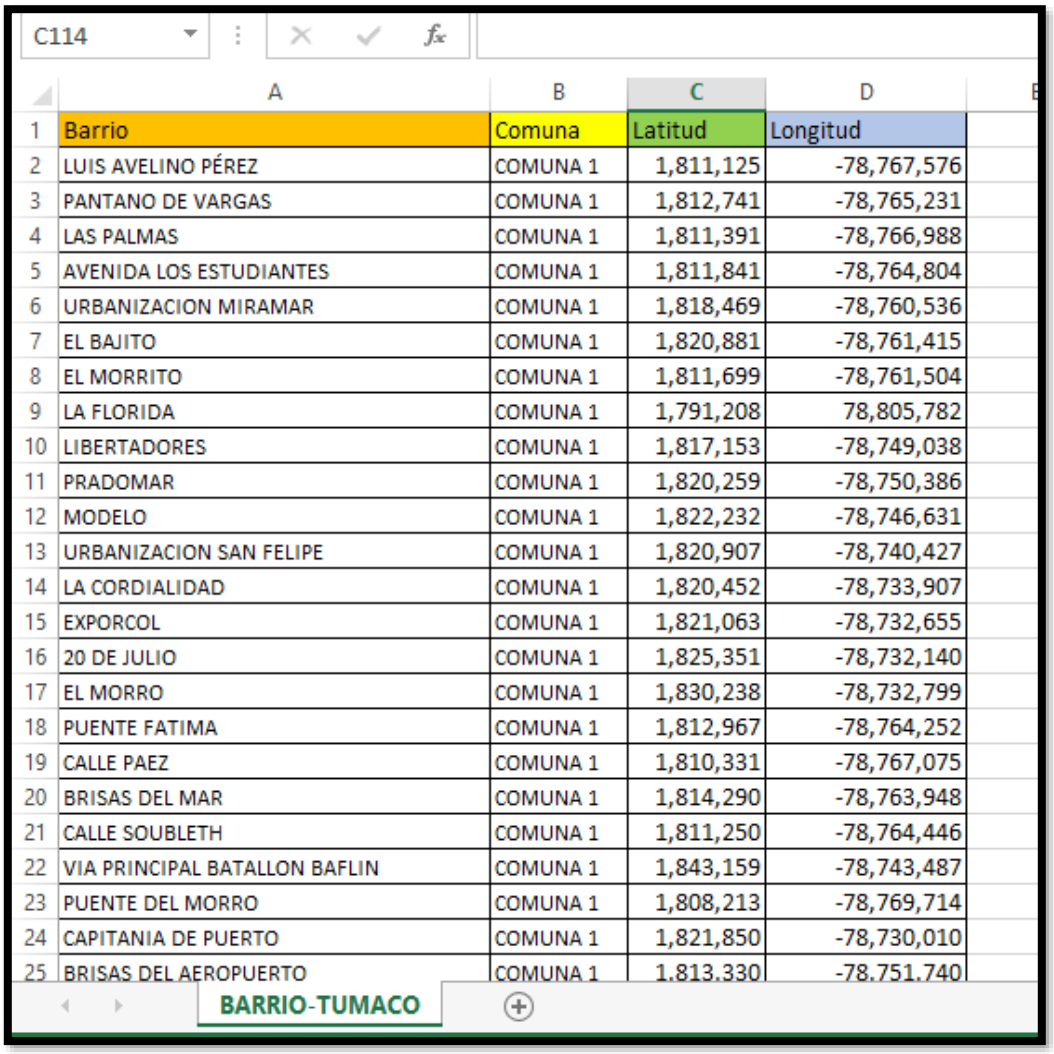
Figura 24. GPS Coordinates + Lat/Long, para la obtención de la información geográfica de los diferentes barrios de Tumaco.



Fuente. esta investigación

Una vez obtenida dicha información geográfica de cada uno de los barrios del casco urbano del municipio de Tumaco, se creó una base de datos con el fin de organizar dicha información en donde se suministran por “nombres de los barrios”, “comuna”, “latitud” y “longitud”. Como se ilustra en la figura 25.

Figura 25. Base de datos de los barrios de Tumaco y sus respectivos datos geográficos "latitud" y "longitud"



C114

	A	B	C	D
1	Barrio	Comuna	Latitud	Longitud
2	LUIS AVELINO PÉREZ	COMUNA 1	1,811,125	-78,767,576
3	PANTANO DE VARGAS	COMUNA 1	1,812,741	-78,765,231
4	LAS PALMAS	COMUNA 1	1,811,391	-78,766,988
5	AVENIDA LOS ESTUDIANTES	COMUNA 1	1,811,841	-78,764,804
6	URBANIZACION MIRAMAR	COMUNA 1	1,818,469	-78,760,536
7	EL BAJITO	COMUNA 1	1,820,881	-78,761,415
8	EL MORRITO	COMUNA 1	1,811,699	-78,761,504
9	LA FLORIDA	COMUNA 1	1,791,208	78,805,782
10	LIBERTADORES	COMUNA 1	1,817,153	-78,749,038
11	PRADOMAR	COMUNA 1	1,820,259	-78,750,386
12	MODELO	COMUNA 1	1,822,232	-78,746,631
13	URBANIZACION SAN FELIPE	COMUNA 1	1,820,907	-78,740,427
14	LA CORDIALIDAD	COMUNA 1	1,820,452	-78,733,907
15	EXPORCOL	COMUNA 1	1,821,063	-78,732,655
16	20 DE JULIO	COMUNA 1	1,825,351	-78,732,140
17	EL MORRO	COMUNA 1	1,830,238	-78,732,799
18	PUENTE FATIMA	COMUNA 1	1,812,967	-78,764,252
19	CALLE PAEZ	COMUNA 1	1,810,331	-78,767,075
20	BRISAS DEL MAR	COMUNA 1	1,814,290	-78,763,948
21	CALLE SOUBLETH	COMUNA 1	1,811,250	-78,764,446
22	VIA PRINCIPAL BATALLON BAFLIN	COMUNA 1	1,843,159	-78,743,487
23	PUENTE DEL MORRO	COMUNA 1	1,808,213	-78,769,714
24	CAPITANIA DE PUERTO	COMUNA 1	1,821,850	-78,730,010
25	BRISAS DEL AEROPUERTO	COMUNA 1	1,813,330	-78,751,740

BARRIO-TUMACO

Fuente. esta investigación

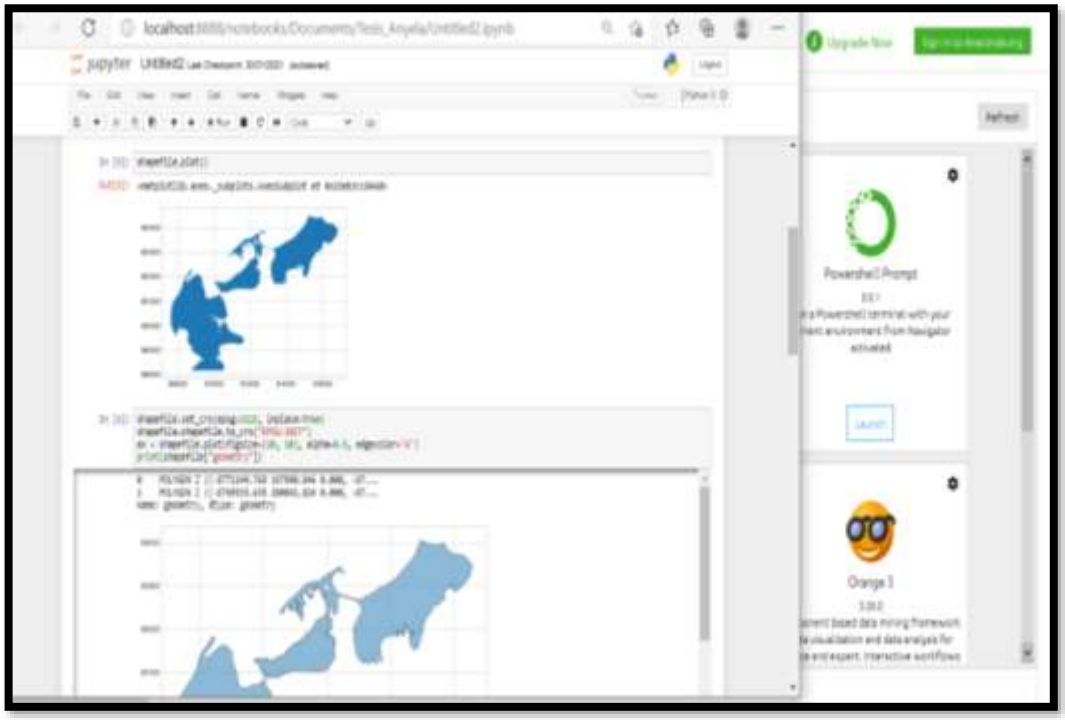
- Una vez organizadas las bases de datos correspondientes a los registros de actividades delictivas de dicho municipio y aquella en la cual esta suministrada la respectiva información geográfica; “latitud” y “longitud” de los diferentes barrios del casco urbano de Tumaco.

- Se procede a realizar los diferentes cruces de bases de datos haciendo uso de la función **“BUSCARV”,** con el fin de obtener una sola base de datos de forma ordenada y poder visualizar toda la información que será utilizada en dicho estudio.

- Dada las respectivas bases de datos, tanto la que ya se encuentra cruzada que contiene toda la información y la que esta suministrada los diferentes barrios de Tumaco ,con sus respectivos datos geográficos ,"latitud" y "longitud", se necesita pasarlos en formato CSV, dado que los archivos CSV (del inglés comma-separated valúes) son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma en donde la coma es el separador decimal como en Chile, Perú, Argentina, España, Brasil, entre otros) y las filas por saltos de línea. Para de esta forma poderlas visualizar desde python por medio del editor jupyter notebook.

Anaconda (distribución de Python). Al iniciar dicho estudio se trabajó con Anaconda, siendo esta una distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos, y aprendizaje automático (machine learning). Esto incluye procesamiento de grandes volúmenes de información, análisis predictivo y cómputos científicos. Está orientado a simplificar el despliegue y administración de los paquetes de software. (Anaconda (distribución de Python) ,2018), se empezó trabajando con dicha distribución del lenguaje de python con editor jupyter notebook, como se observa en la figura 26.

Figura 26. Avances desde el distribuidor ANACONDA -EDITOR jupyter notebook



Fuente. esta investigación

Con el fin de garantizar la seguridad en los datos, se realizó todo el trabajo desde Google Colab. Donde los cuadernos de Colab son cuadernos de Jupyter alojados en Colab.

Con Colab, se puede aprovechar toda la potencia de las bibliotecas más populares de Python para analizar y visualizar datos. La celda de código de abajo utiliza NumPy para generar datos aleatorios y Matplotlib para visualizarlos. Uno de los aspectos más importantes es que Los cuadernos que se crean en Colab se almacenan en la cuenta de Google Drive (guardar nuestros datos en la Nube).

- **Aprendizaje automático**

Con Colab, puedes importar un conjunto de datos de imágenes, entrenar un clasificador de imágenes con dicho conjunto de datos y evaluar el modelo con tan solo usar unas pocas líneas de código. Los cuadernos de Colab ejecutan código en los servidores en la nube de Google, lo que te permite aprovechar la potencia del hardware de Google, incluidas las GPU y TPU, independientemente de la potencia de tu equipo. Lo único que se requiere es un navegador. (Google Colaboratory, 2018)

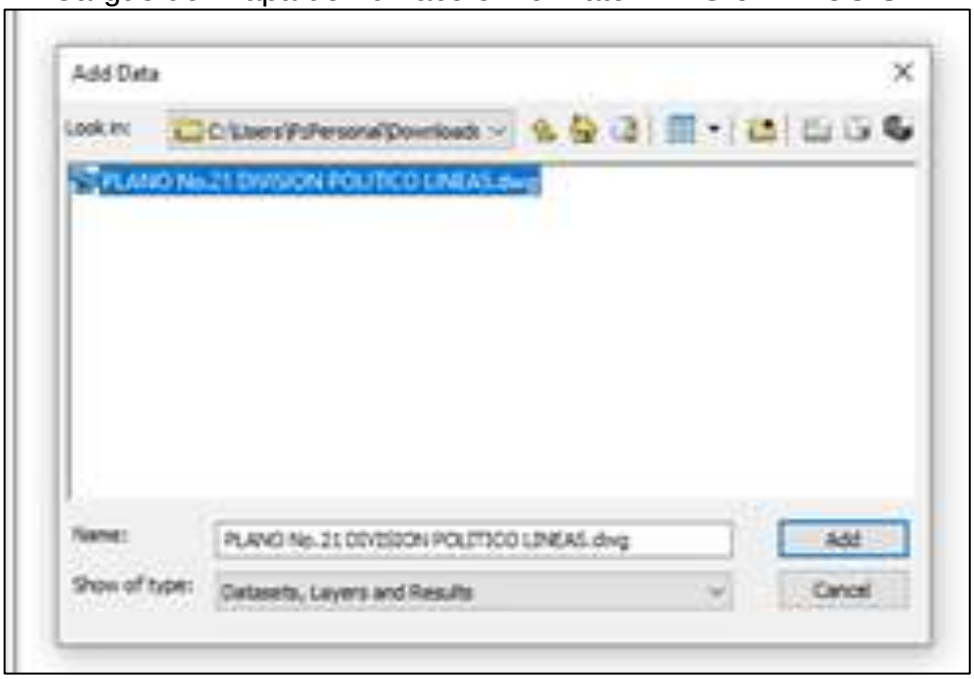
4.1.8.Obtención del mapa shapefile (*shp). Una vez solicitado el mapa del municipio en formato.shp a la Alcaldía municipal de Tumaco por medio de una solicitud escrita como se muestra en la figura, la encargada de dicha solicitud no da respuesta alguna a la petición reportada, debido a esta situación presentada, es necesario acudir a otro medio para la obtención de dicho mapa, una vez se logra obtener el mapa en formato.DWG, es necesario analizar el archivo para de este modo buscar la forma de obtener un archivo.SHP.

El archivo en formato .DWG, el cual contiene el mapa de Tumaco; se puede observar según la figura 25.

Se necesitaba:

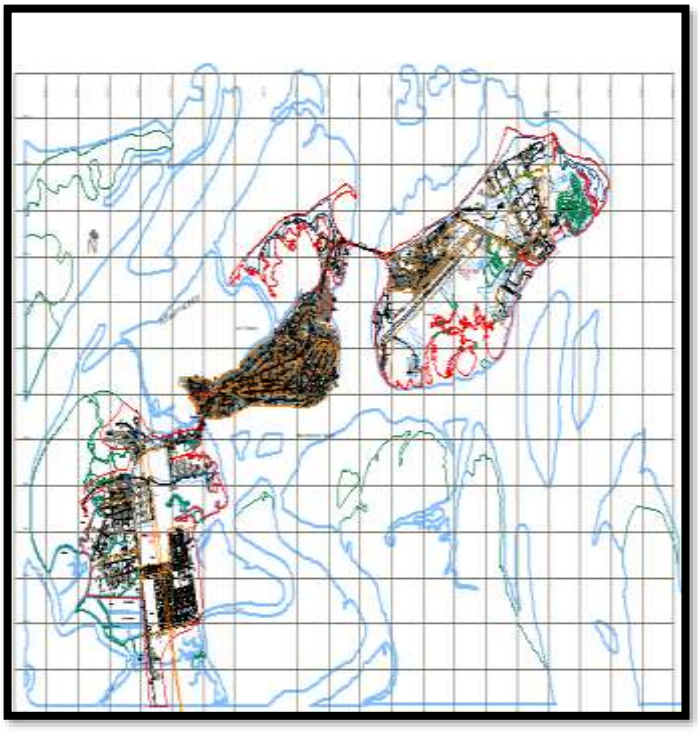
1. Pasar el archivo de .DWG a SHP.
2. Realizar un corte para la obtención de solo la parte urbana de dicho municipio.
Para obtener la realización de los puntos con anterioridad se utilizó la herramienta ARCGIS

Figura 27. Cargue del mapa de Tumaco en formato. DWG en ARCGIS



Fuente. esta investigación

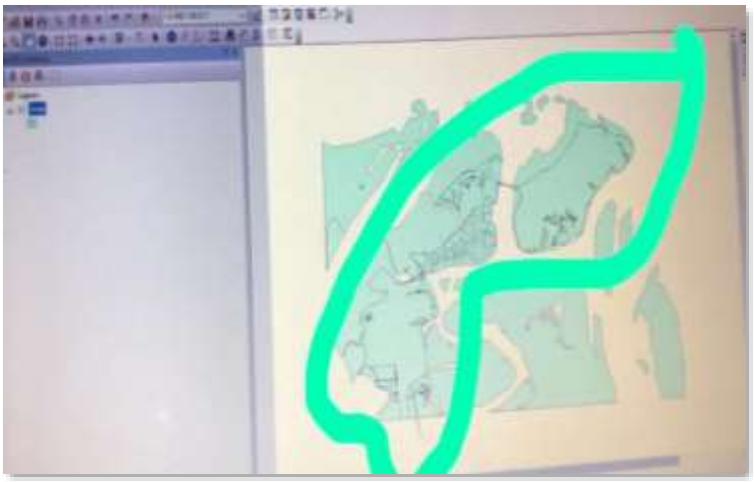
Figura 28. Visualización del mapa del casco urbano de Tumaco en formato.DWG



Fuente. esta investigación

Una vez se convierte el archivo a.”SHP” y se realiza el corte de solo la parte urbana, como se muestra en la figura 29.

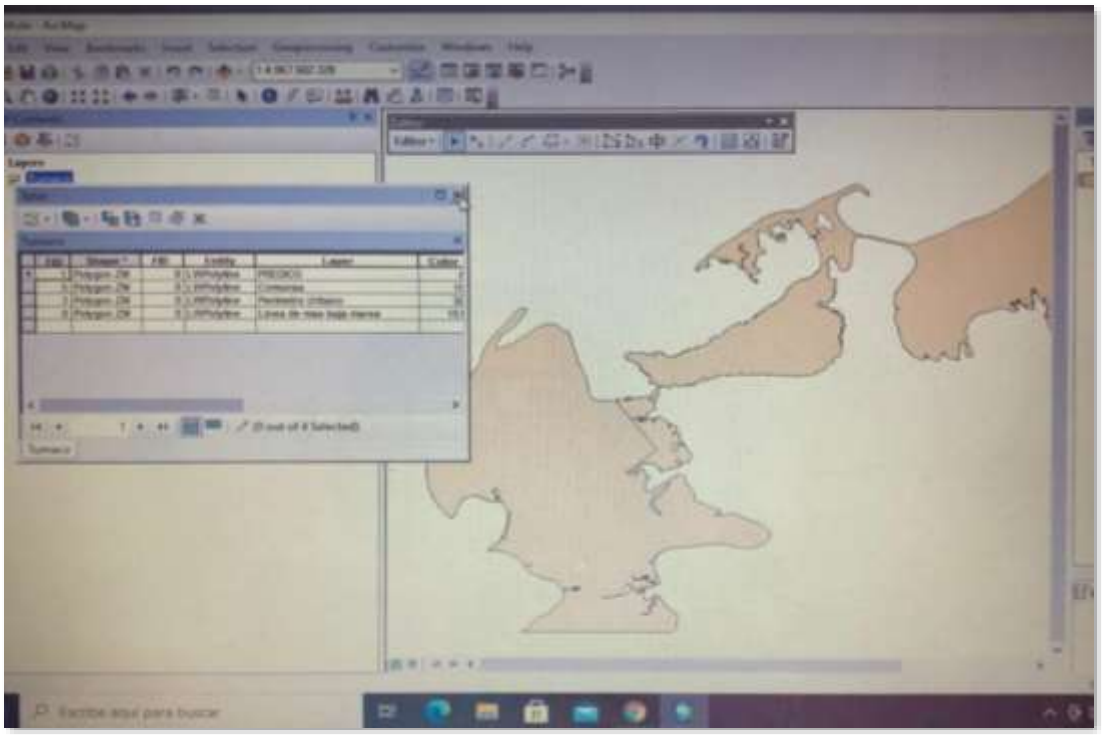
Figura 29. Corte de solo la parte urbana del mapa de Tumaco



Fuente. esta investigación

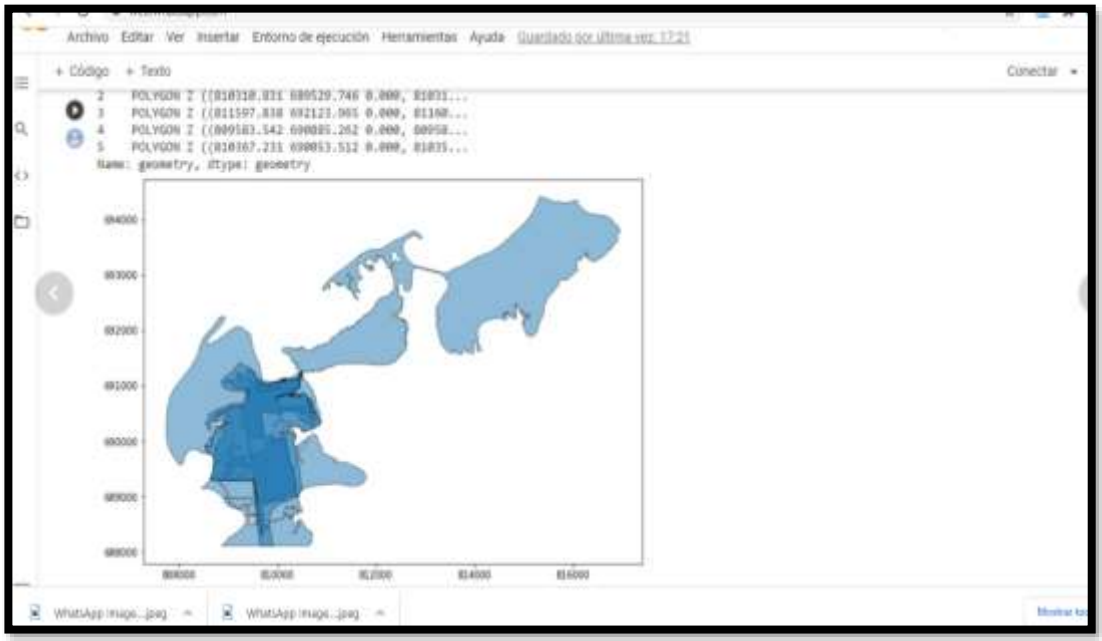
A la hora de cargar el archivo ya convertido en .SHP en Python , se contaron con varias dificultades , una de ellas y la mas importante fue la cantidada de poligonos que conformaban el mapa del casco urbano de tumaco, se contaban con 5 poligonos, como lo indica la figura 30 .

Figura 30. Cantidad de polígono del mapa del casco urbano de Tumaco



Fuente. esta investigación

Figura 31. Visualización de los 5 polígonos desde Google Colab



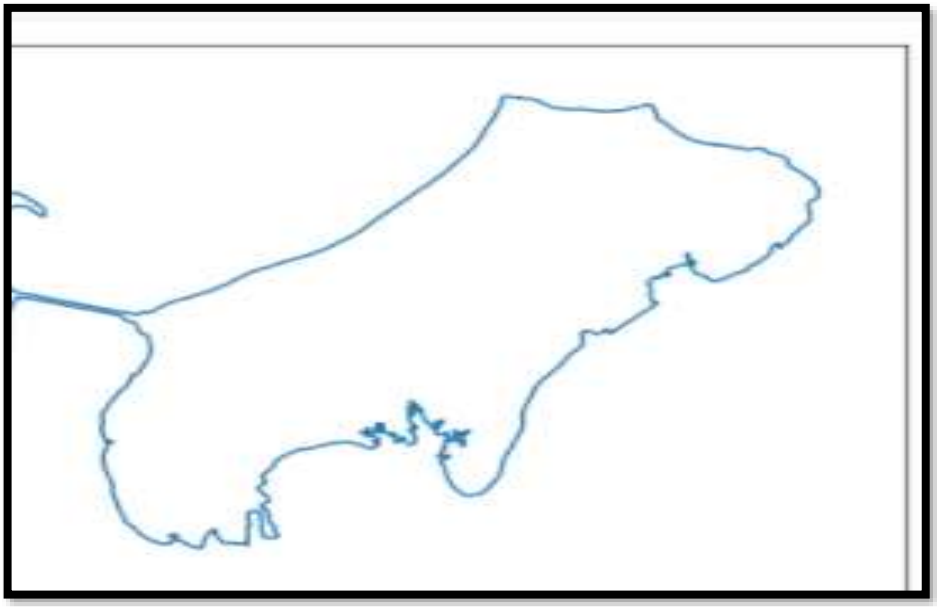
Fuente. esta investigación

luego se procede a dejar solo 2 poligonos tomando como referencia la parte superior e inferior del mapa , a la hora de cargar el archivo.shp a Python con sus 2 poligonos, no se lograba graficar por completo el mapa , como se observa en la figura 32 y 33.

Figura 32. captura de la parte inferior del mapa de Tumaco.

Fuente. esta investigación

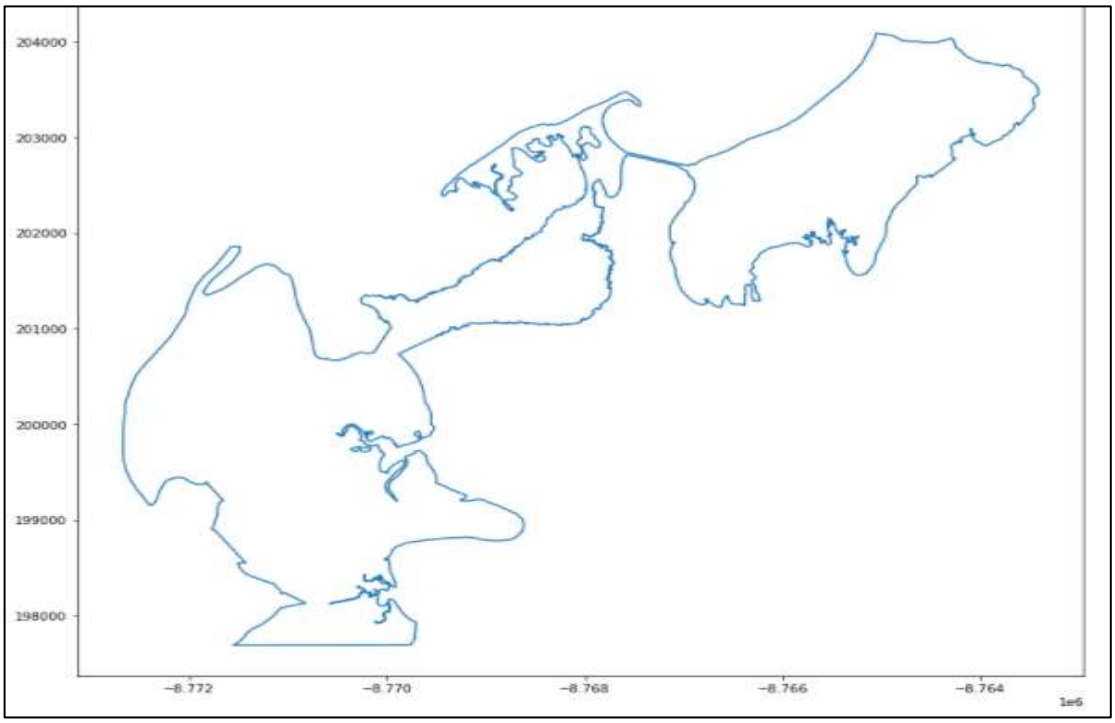
Figura 33. captura de la parte superior del mapa de Tumaco



Fuente. esta investigación

Se procedio a dejar tan solo un poligono para el mapa de Tumaco en formato.SHP ,de esta manera se ovtuvo un cargue exitoso y la correcta visualizacion del mapa del casco urbano de dicho municipio. Como se puede observar en la figura 34 .

Figura 34. Visualización del mapa del casco urbano de Tumaco con 1 solo polígono



Fuente. esta investigación

4.1.9 Definición de variables. Las variables que se tomaron para el desarrollo, algunas de ellas no son sensibles y por tanto no son suministradas por la fuente del observatorio del delito, ya que en algunos casos no son reportadas.

Los datos recolectados fueron obtenidos mediante las fuentes de extracción mencionadas anteriormente, las cuales permiten identificar comportamientos posteriormente realizado el análisis de los datos.

También se cuentan con variables que son útiles para la validación de los agentes inteligentes con mejor compromiso en sus métricas de calidad, como lo indica la tabla 2.

Tabla 2. Variables Globales

VARIABLES GLOBALES	
Variables	**Descripción**
Amenaza	Esta variable hace referencia al “tipo de delito”, dentro de la base de datos a analizar.
Homicidio	Esta variable hace referencia al “tipo de delito”, dentro de la base de datos a analizar.
Hurto-persona	Esta variable hace referencia al “tipo de delito”, dentro de la base de datos a analizar.
Terrorismo	Esta variable hace referencia al “tipo de delito”, dentro de la base de datos a analizar.

Fuente. esta investigación

Tabla 3. Tabla de descripción de cada una de las variables que serán utilizadas para el desarrollo de la investigación-suministradas en la base de datos del observatorio de delito colombiano

Numero	Variable	Descripción
1	**Tipo de delito**	Esta variable llamada tipo de delito, contiene el tipo de actividad delictiva que se describe en los datos obtenidos, la cual hace referencia a cada una de las filas que contiene la base de datos principal (cruzada) con su respectiva información suministrada, en donde los tipos de delitos pueden ser “Amenaza”, “Homicidio”, “Hurto-persona “y “Terrorismo”.
2	**Fecha**	Esta variable contiene la fecha, descrita en día, mes y año en el cual ocurre cada actividad delictiva.
3	**Departamento-(Se elimino)**	Esta variable contiene el nombre del departamento al cual pertenece el lugar de dicho estudio, y lugar de donde ocurren cada actividad delictiva; en este caso TUMACO, departamento NARIÑO.
4	**Municipio-(Se elimino)**	Esta variable contiene el municipio en el cual ocurren los diferentes tipos de actividad delictiva.
5	**Día**	Esta variable contiene el día en el cual ocurrió dicha actividad delictiva.
6	**Delito- (Se elimino)**	Esta variable contiene un soporte del tipo de delito que ocurrió, ejemplo en el caso de

		las filas del tipo de variable “Terrorismo”, se registran en el campo “Delito” el nombre del artículo en el cual contiene toda la información de como ocurre dicha actividad delictiva.
7	**Hora**	En esta variable se almacena la hora en la que ocurrió dicha actividad delictiva.
8	**Barrio**	Esta variable contiene el punto o lugar en donde ocurrió dicha actividad delictiva.
9	**Zona**	Este campo contiene el tipo de zona al cual se va aplicar dicho estudio, en este caso a la “ZONA URBANA”.
10	**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.
11	**Arma empleada**	Esta variable contiene el tipo de arma que se empleó en dicha actividad delictiva. Ejemplo, arma de fuego, arma blanca / cortopunzante etc.
12	**Móvil Agresor**	Esta variable contiene el tipo de arma que se empleó en dicha actividad delictiva. Ejemplo, arma de fuego, arma blanca / cortopunzante etc.
13	**Móvil Victima**	Esta variable contiene el medio o tipo de transporte que utilizaba la victima a quien se le cometió dicha actividad delictiva.
14	**Edad**	Esta variable contiene la edad de quien cometió dicha actividad delictiva.
15	**Sexo**	Esta variable contiene el tipo de sexo, ya sea “**femenino**” o “**masculino**” de la persona que cometió dicho delito.
16	**Estado civil**	Esta variable contiene el estado civil de la persona que cometió dicho delito, ya sea “soltero”, “unión libre”, “casado”, etc.
17	**País de nacimiento**	Esta variable contiene el país de nacimiento de la persona que cometió dicho delito.
18	**Clase empleado**	Esta variable contiene el tipo de empleo de la persona la cual cometió dicha actividad delictiva.

19	**Profesión**	Esta variable contiene el tipo de profesión que ejercía la persona que cometió dicho delito.
20	**Escolaridad**	Esta variable contiene el nivel de estudio de la persona que cometió dicha actividad delictiva.
21	**Código Dane -(Se elimino)**	Esta variable contiene el código Dane del municipio de Tumaco en el cual se va a centrar dicho estudio
22	**Cantidad-(Se elimino)**	Esta variable contiene la cantidad de actividad delictiva, según la fecha y el año en que esta ocurrió.
23	**Latitud**	Esta variable contiene el dato geográfico “latitud” del barrio en donde ocurrió dicha actividad delictiva. Esta variable nos va a permitir ordenar los puntos en el “**mapa. Shp”** que representan a los casos de actividad delictiva.
24	**Longitud**	Esta variable contiene el dato geográfico “longitud” del barrio en donde ocurrió dicha actividad delictiva. Esta variable nos va a permitir ordenar los puntos en el “**mapa. Shp”** que representan a los casos de actividad delictiva.

Fuente. esta investigación

4.2.DISEÑO

En esta sección se dará a conocer el diseño de cómo se manejan los datos a través del punto de vista de información con sus correspondientes diagramas, para lograr una mejor explicación de la fuente de extracción como aporte para el desarrollo de esta investigación.

4.2.1.Diseño de arquitectura de datos. Es necesario definir una arquitectura de datos, para identificar aspectos a tener en cuenta en el empleo adecuado de los datos: su recolección, agrupación y proceso. Para cumplir con el objetivo se propone utilizar array ya que estos permiten manipular datos de manera muy flexible. Combinándolas y anidándolas, es posible organizar información de manera estructurada para representar sistemas del mundo real.

En muchas aplicaciones de Ingeniería, por otra parte, más importante que la organización de los datos es la capacidad de hacer muchas operaciones a la vez sobre grandes conjuntos de datos numéricos de manera eficiente. Algunos ejemplos de problemas que requieren manipular grandes secuencias de números son: la predicción del clima, la construcción de edificios, y el análisis de indicadores financieros entre muchos otros.

La estructura de datos que sirve para almacenar estas grandes secuencias de números (generalmente de tipo float) es el arreglo.

Los arreglos tienen algunas similitudes con las listas: Los elementos tienen un orden y se pueden acceder mediante su posición, los elementos se pueden recorrer usando un ciclo for. Sin embargo, también tienen algunas restricciones: Todos los elementos del arreglo deben tener el mismo tipo,en general, el tamaño del arreglo es fijo (no van creciendo dinámicamente como las listas),Se ocupan principalmente para almacenar datos numéricos.

Seleccionar modelos de referencia, puntos de vista y herramientas: los puntos de vista dentro de la Arquitectura de datos, comprenden las Partes interesadas, es decir, la Facultad de Ingeniería de la Universidad De Nariño, a la estudiante correspondiente a la Investigación y a los entes de control frente a casos de actividad delictivas, Alcaldía Municipal De Tumaco.

1. **Diccionario de datos:**

Tabla 4. Diccionario de datos, en el cual se describe el tipo de dato de cada una de las variables a trabajar, ubicado con una x el tipo de variable respectivo.

Variables (tipo de variables)	Cadena	Fecha	Flotante	Entero
Fecha		x		
Departamento QUITAR	X			
Municipio QUITAR	x			
Día	x			
Delito QUITAR			x	
Hora	X			
Barrio	x			
Zona	X			
Clase-Sito	X			
Arma empleada	X			
Móvil agresor	X			
Móvil victima	X			
Edad				x
Sexo	X			
Estado civil	X			
País de nacimiento	X			
Clase de empleado	X			
Profesión	X			
Escolaridad			x	
Código Dane QUITAR				x
Cantidad QUITAR	X		x	
Latitud	X		x	
Longitud	X			

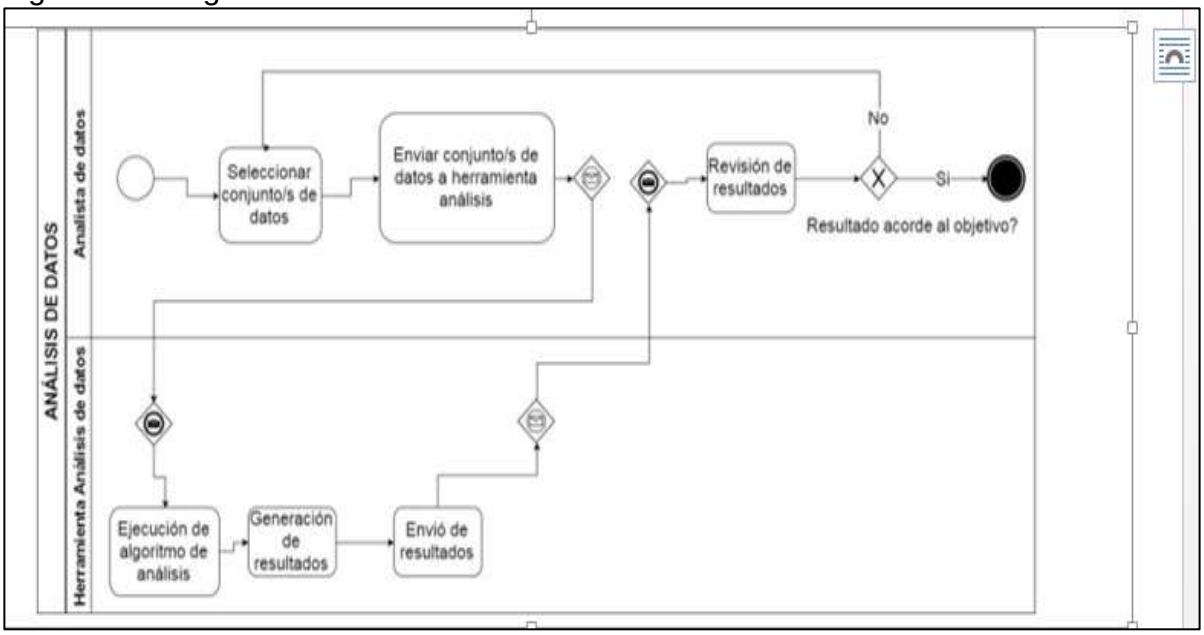
Fuente. esta investigación

En la tabla 4, se contaron con 18 variables en total. La especificación de cada una de las variables esta descrita en la **tabla 2.** Definición de variables de este documento.

2. **Diagrama de ciclo de vida:** En este diagrama se muestra el proceso del análisis de datos consta de los actores: Analista de datos y la herramienta de análisis de datos, los cuales interactúan para obtener el objetivo de la investigación. Para ello el analista de datos envía los datos a la herramienta de análisis de datos, la herramienta aplica el algoritmo de regresión y el analista revisa los resultados obtenidos en patrones.

El analista, después de revisar los resultados verifica si cumplen con el objetivo propuesto, si no cumplen con el objetivo propuesto, se repite el proceso hasta encontrar un resultado adecuado con el objetivo; si el analista obtiene un resultado acorde al objetivo, el proceso finaliza.

Figura 35. Diagrama de ciclo de vida



Fuente. esta investigación

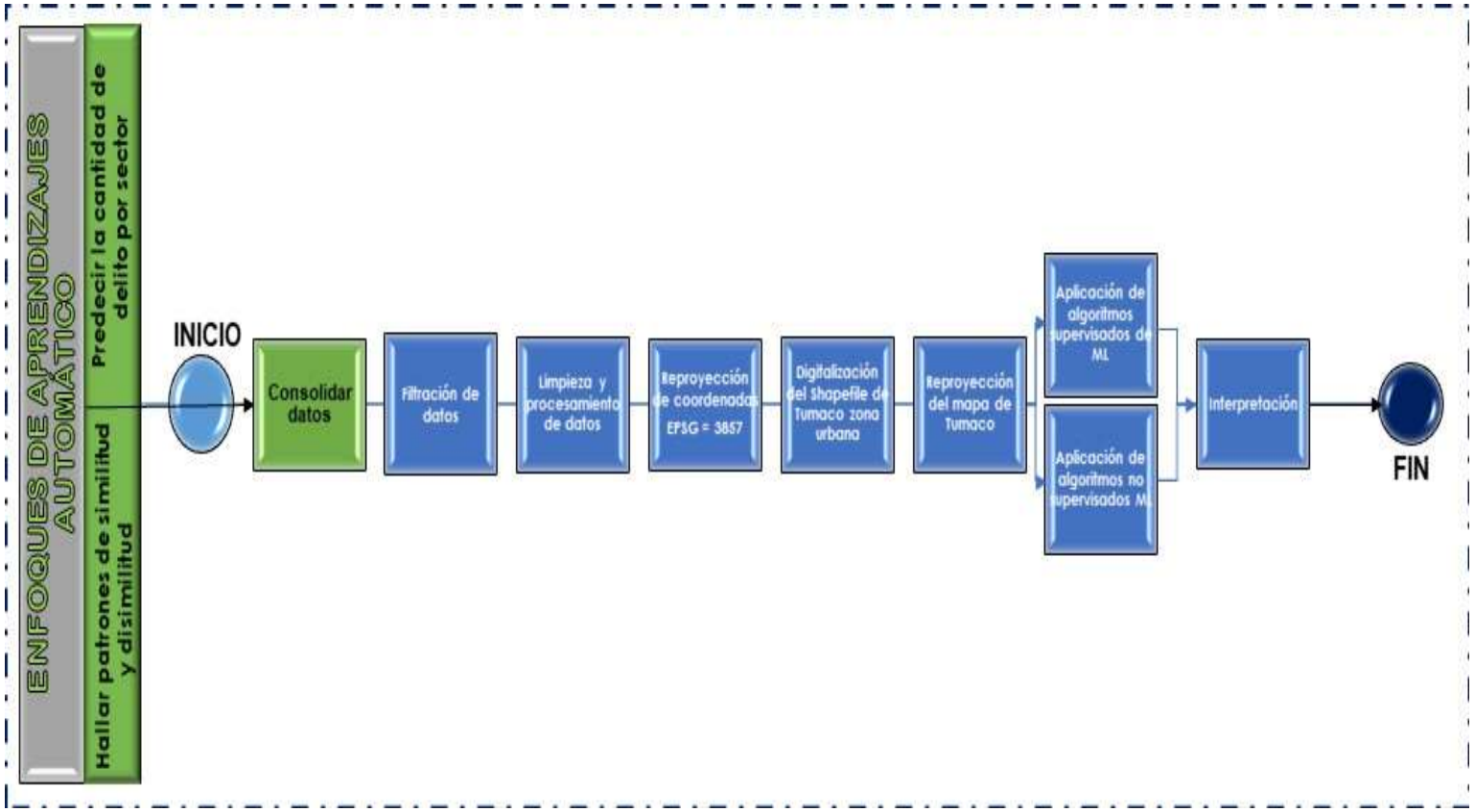
- **Analista de datos:** Se encargará de ejecutar el algoritmo de regresión en la herramienta de análisis de datos, los conjuntos de datos para la muestra de resultados.

- **Arquitecto de datos:** Se encargará de diseñar y mostrar a través de un conjunto de modelos subóptimos de predicción de actividad delictiva mediante una sintonización de hiperparámetros.

- **Investigador:** Se encargará de analizar los resultados de los mejores modelos obtenidos mediante: gráficos de interpretabilidad, métricas de calidad y mapas de interpolación de la actividad delictiva en función de la ubicación geográfica.
- Estos roles los tiene asignado la estudiante que realiza la investigación.

4.2.2. Desarrollo del marco de comparacion de algoritmos supervizados de Machine Learning:

Figura 36. Diagrama Arquitectónico del marco experimental



Fuente. esta investigación

4.2.3.Selección de algoritmos de regresión y Agrupación. Es necesario aplicar algoritmos de regresión y agrupación, cuyo objetivo es establecer un método para la relación entre un cierto número de características y una variable objetivo continua; para esta selección se utilizó la técnica de aprendizaje automático basados en:

- **Tecnicas Supervisadas de Machine Learning.**

    - Arboles de decisión
    - Bosques aleatorios
    - Máquinas de soporte vectorial
    - Redes neuronales artificiales.

- **Tecnicas no supervisadas de Machine Learning.**

    - Kmeans
    - K-Prototype.

4.2.4.Criterios de selección. A continuación, se definen los criterios de selección sobre los algoritmos de regresión y agrupación.

- Permite analizar volúmenes de datos. Por ejemplo: El número de casos de actividad delictiva a validar es del 2010 hasta 2019.

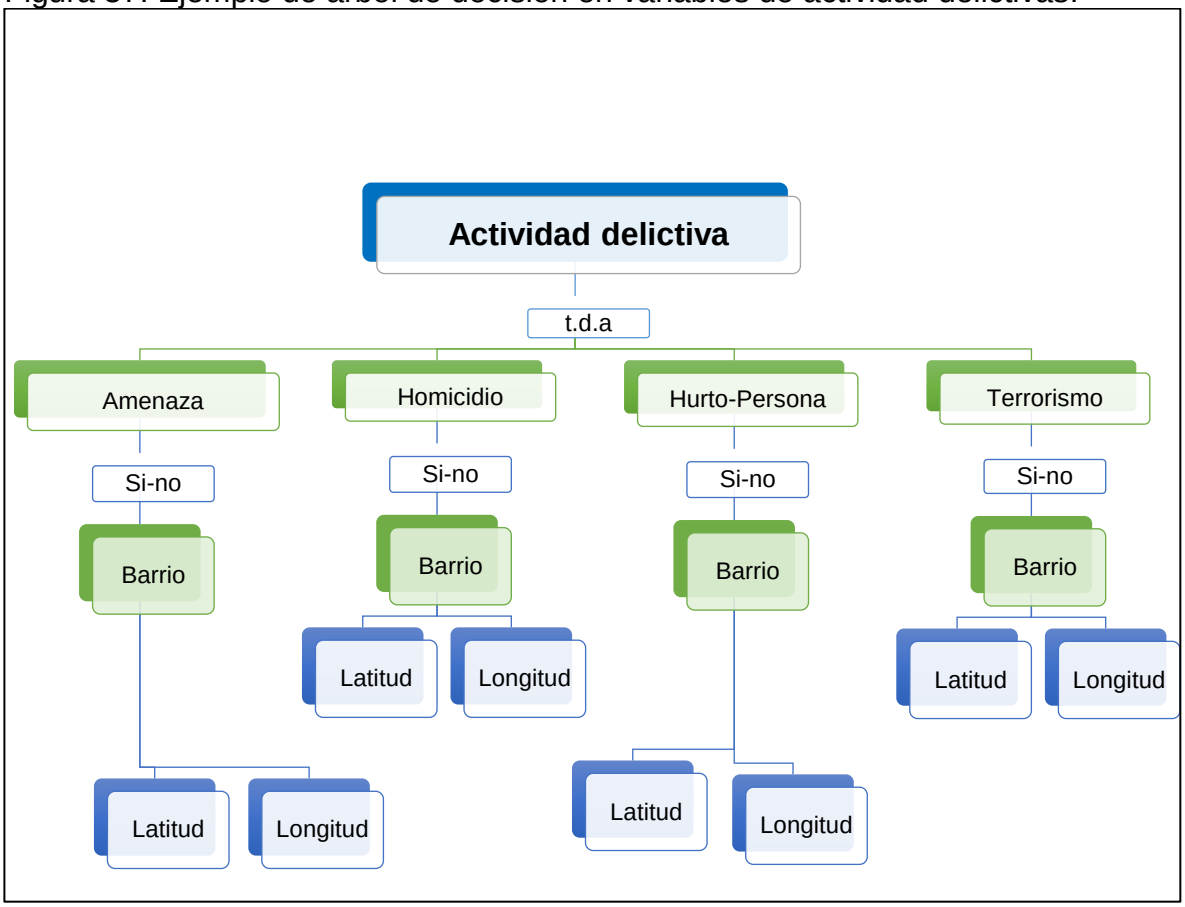
- Permite analizar de manera completa todas las posibles soluciones.

- Ayuda a realizar las mejores decisiones con base a la información existente y a las mejores suposiciones.

- Su estructura permite analizar las alternativas, los eventos, las probabilidades y los resultados.

- itera en los casos de un conjunto de datos para agruparlos en clústeres que contengan características similares.

Figura 37. Ejemplo de árbol de decisión en variables de actividad delictivas.



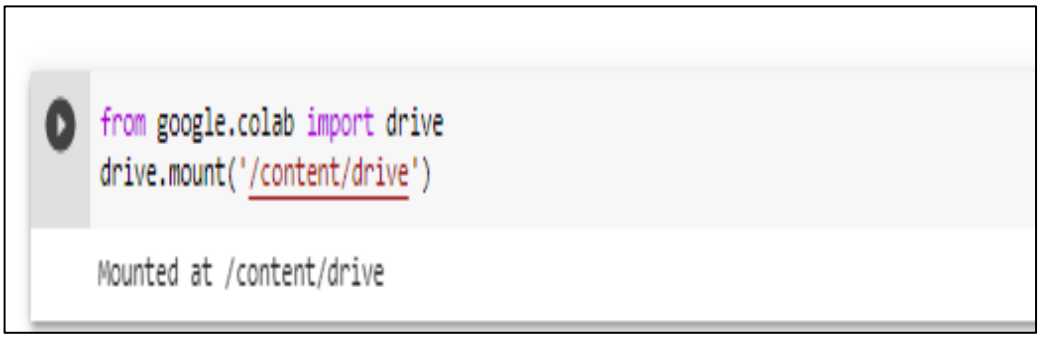
Fuente. esta investigación

4.2.5. Pasos para la obtención del marco experimental. En primer lugar, se montó la unidad de drive para poder acceder a los archivos que se van a utilizar desde la unidad de colab con las siguientes líneas de código.

```
*From google.colab import drive
*drive.mount ('/content/drive')
```

Una vez montada la unidad drive ya se pueden observar las diferentes carpetas a utilizar con sus respectivos archivos. Como indica la figura 35.

Figura 38. Conten/drive -desde google colab



Fuente. esta investigación

Instalación bibliotecas "Geopandas" y "pykrige":

- Se instaló la biblioteca geopandas, para el tratamiento de los datos, que facilita la labor de calcular datos masivos de sistemas de información geográfica, con la siguiente línea de código.

- La siguiente línea de código "**!pip install pykrige",** que es un conjunto de herramientas que interpolan datos para realizar geo-estadísticas, también permite exportar e importar archivos .asc, Numpy es una biblioteca de funciones matemáticas que admite el procesamiento de matrices, Lapack y Blas son bibliotecas para resolver sistemas de ecuaciones lineales, problemas de valores propios y descomposición de valores singulares. Ejemplo; en el caso que se tiene varios tipos de actividad delictiva y se sabe que en algunos puntos geográficos (del municipio de Tumaco) se hallaron unas cantidades importantes de actividad delictiva, realizando el PyKrige se puede saber en qué barrios específicos hay actos delictivos.

- Luego se procedió a importar las bibliotecas, cada script con los que se han obtenido los resultados fueron escritos en el lenguaje de programación Python y se utilizaron las bibliotecas detalladas en la Tabla 5.

Tabla 5. Tabla bibliotecas a importar en colab

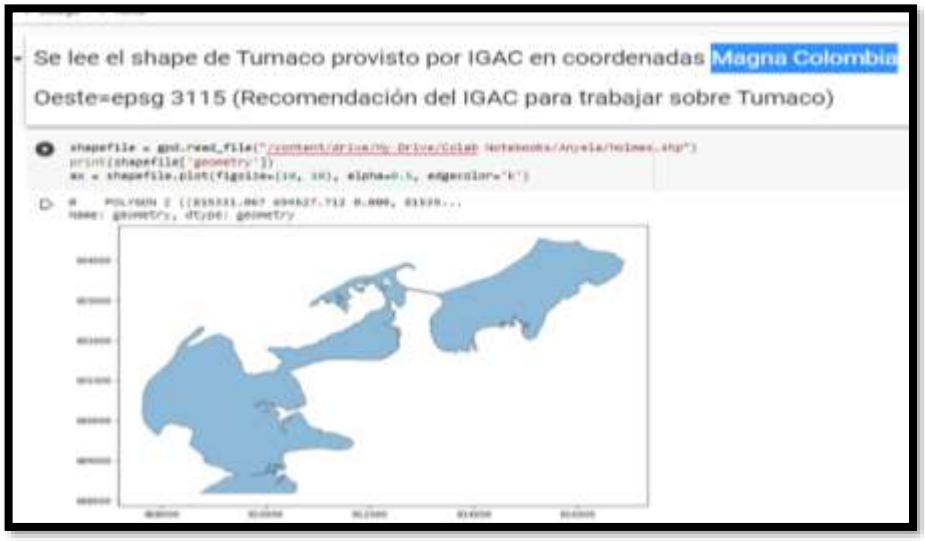
Bibliotecas	Descripción
PyKrige	Herramienta básica a la hora de interpolar datos con previa modelización estructural.
Pandas	Biblioteca de software escrita como extensión de Numpy para manipulación y análisis de datos.
Numpy	Es una librería de Python especializada en el cálculo numérico y el análisis 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 geograficas.

Fuente. esta investigación

4.2.6. Obtener un Modelo para la Prediccion de actividad delictiva utilizando la Metodologia KDD:

4.2.6.1 Lectura del archivo SHP en coordenadas Magna Colombia (oeste=epsg 3115). Las siguientes líneas de códigos son muy importantes para el desarrollo de dicha investigación ya que se realiza la lectura del archivo SHEPE DEL CASCO URBANO DEL MUNICIPIO DE TUMACO ,provisto por el IGAC en coordenadas Magna Colombia ; ya que generalmente las coordenadas que se trabajan en Colombia son las coordenadas magna Colombia 3115 oeste=epsg 3115, en este caso se necesitó trabajar con plano , es decir realizando conversiones(proyecciones) del globo terráqueo con el fin de **llevar dicha esfera a un plano** epsg 3115 . Como indica en la figura 39.

Figura 39. Polígono oeste=epsg 3115

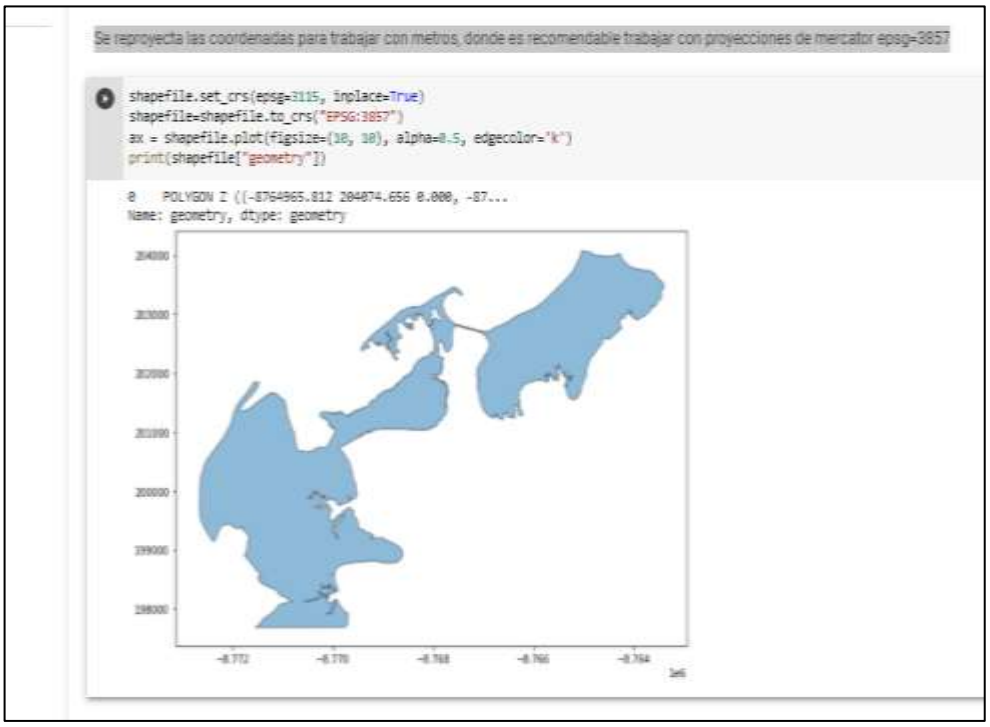


Fuente. esta investigación

Para una mejor comprensión de los datos no se trabajó con coordenadas 3115, ya que es más favorable trabajar en metros, para ello.

Se reproyecta las coordenadas para trabajar con metros, donde es recomendable trabajar con proyecciones de mercator epsg=3857. Como lo indica la figura 40.

Figura 40. Polígono Mercator epsg=3857



Fuente. esta investigación

4.2.7.Creación del método - convertir archivo “SHP” a una lista de puntos. Una vez convertidas la coordenada a metros, se procede a crear un método para dado un conjunto de polígonos del mapa (shape) para que retorne una lista de puntos. Como se muestra en la figura 41.

Figura 41. Método para dado un conjunto de polígonos de un mapa (shape) se retorne una lista de puntos.

```
from shapely.geometry import MultiPolygon

def points_from_polygons(polygons):
    points = []
    for mpoly in polygons:
        if isinstance(mpoly, MultiPolygon):
            polys = list(mpoly)
            print("YES")
        else:
            polys = [mpoly]
        for polygon in polys:
            for point in polygon.exterior.coords:
                points.append(point)

            for interior in polygon.interiors:
                for point in interior.coords:
                    points.append(point)

    return points
```

Type Markdown and LaTeX: $\alpha 2$

Fuente. esta investigación

- Luego se convierte el SHAPE en puntos y se aterrizan en las variables X y Y correspondientes a las coordenadas planas de (latitud y longitud).como se muestra en la figura 42.

Figura 42. Se convierte el shape en puntos y se aterriza en las variables x y y correspondientes a las coordenadas plana de (latitud y longitud).

```
[ ]  points = points_from_polygons(shapefile['geometry'])
     #points = path_from_polygons(shapefile['geometry']) [(1,5),(2,3),(4,5)] ->x [1,2,4] y [5,3,5]

     x = [point[0] for point in points]
     y = [point[1] for point in points]
```

```
[ ]  import pandas as pd
     #from evolutionary_search import EvolutionaryAlgorithmSearchCV
     from sklearn.model_selection import  GridSearchCV, RandomizedSearchCV
     #from KSVM import KSVC, KSVR

     from sklearn.model_selection import train_test_split,KFold, cross_val_score,ShuffleSplit
     from sklearn.preprocessing import MinMaxScaler
     from sklearn.preprocessing import Normalizer
     from sklearn.preprocessing import StandardScaler
     import numpy as np
     from time import time
     from multiprocessing.pool import Pool
     from sklearn.metrics import mean_absolute_error,mean_squared_error,median_absolute_error
     import matplotlib.pyplot as plt
     import matplotlib.tri as tri
     import numpy as np
     from sklearn.gaussian_process import GaussianProcessRegressor
     from matplotlib.colors import LinearSegmentedColormap, Normalize
```

Fuente. esta investigación

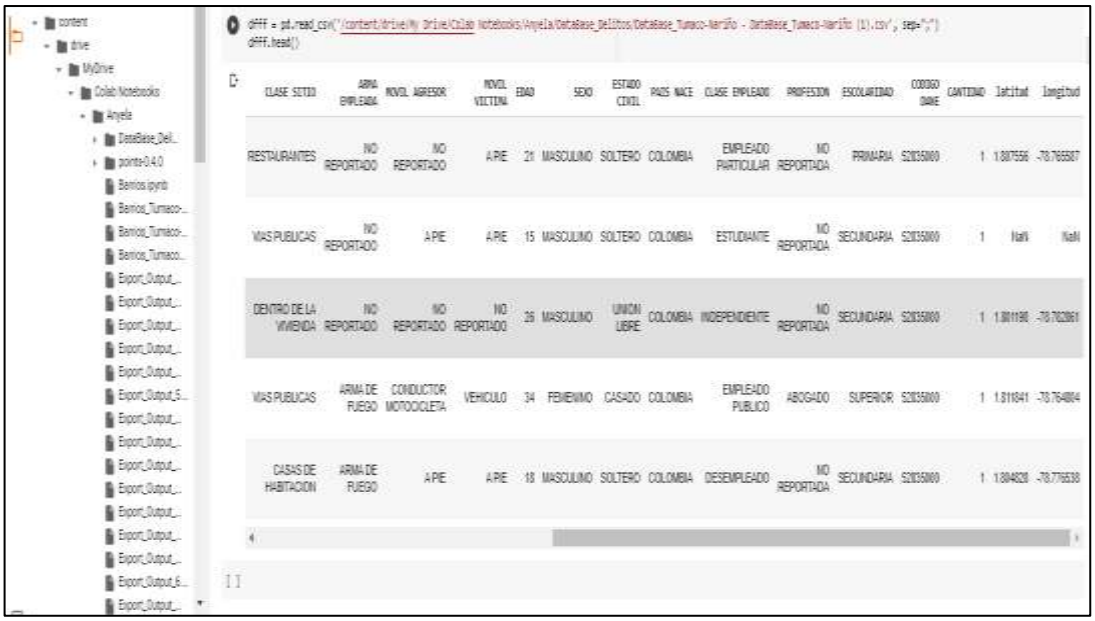
- Una vez ya convertido el archivo shape en puntos, se procede a cargar las 2 bases de datos que serán de gran importancia para el desarrollo de este estudio, la base de datos que contiene los barrios del casco urbano del municipio de Tumaco llamada “dff” y la base de datos ya cruzada que contiene todo el registro de los casos de actividad delictiva más los datos geográficos de cada barrio “latitud “ y “longitud”, llamada “dfff. En este caso se debe tener en cuenta cual es el “separador “del archivo, es “;”para realizar un cargue correcto de los datos.

Para lo cual es necesario aplicar muy bien la ruta desde drive.

Figura 43. Cargue de la base de datos de los barrios de tumaco con su información geográfica

Fuente. esta investigación

Figura 44. Cargue de la Base de datos cruzada con los diferentes casos de actividad delictiva más información de los barrios del casco urbano de Tumaco y su respectiva información geográfica



Fuente. esta investigación

- una vez ya cargada las 2 bases de datos anteriormente mencionadas, se analizan las coordenadas aplicadas a los diferentes barrios del casco urbano de Tumaco, teniendo en cuenta que no todas las coordenadas de los barrios se encontraban en Google Maps, la parte faltante para la obtención de esta información esta explicada en la pag 81. y en la figura 22 indica la herramienta que se utilizó para la obtención de dichas informaciones geográficas.

- Con la obtención de dicha información, se necesita realizar la reproyección ya que esos puntos obtenidos están en coordenadas de Google Maps 4326, siendo estos ángulos de “latitud” y “longitud”. Como se trabajó el mapa en 3857, se debe realizar la reproyección de las coordenadas.

- Para ello se importa la biblioteca de **“from pyproj import Transformer”**
  Luego se aplica la línea de código que indica que se va a transformar los datos geográficos de **4326 a 3857.**

- **Transformer = Transformer.from_crs ('epsg: 4326’,'epsg:3857’, always_xy=True).**

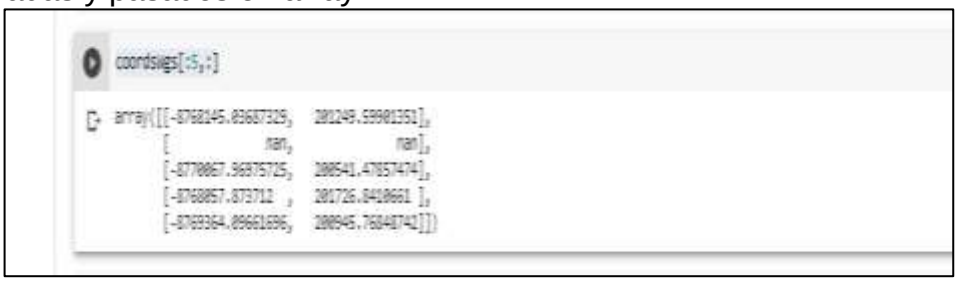
  Y los puntos que se tenían en el dataframe mencionado con anterioridad llamado “dfff”, se transforman con la “longitud” y “latitud”, como se indica en la línea de código.

- **points = list(zip(dfff.longitud,dfff.latitud))**

  Al obtener dichos puntos, le enviamos el transforme y los ubicamos en una coordenada, luego se transforma en array para que sea fácil su tratamiento, como lo indica la siguiente línea de código.

- **coordsWgs = np.array(list(transformer.itransform (points)))**
  Luego con la siguiente línea de comando verificamos si realizo la transformación como lo indica la figura 45.
  coordsWgs[:5,:]

Figura 45. Transformación de los puntos del dataframe “dfff “, convertidos a coordenadas y pasados en array.



Fuente. esta investigación

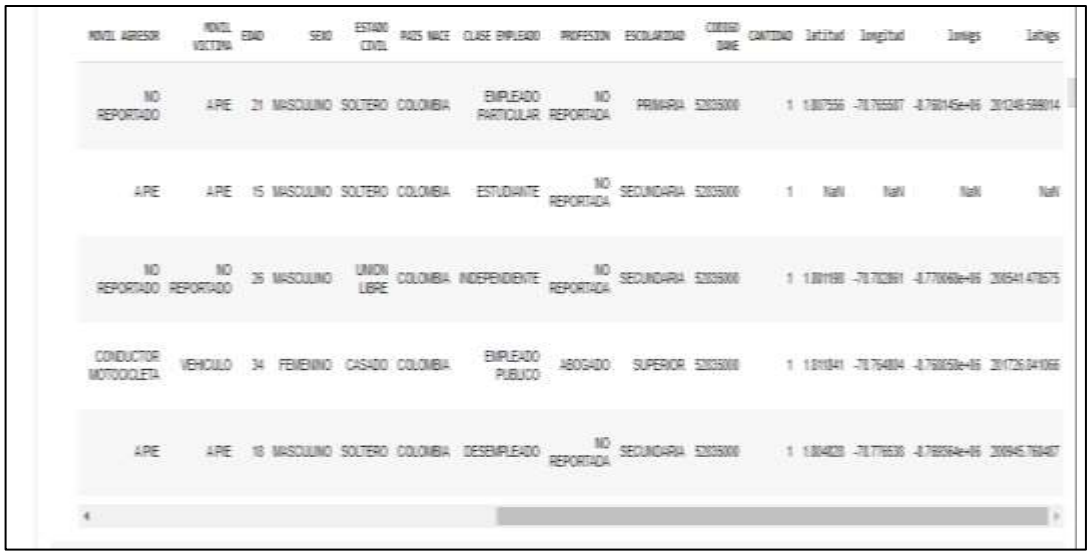
Al dataframe llamado “dfff”, se le asignan otras dos propiedades llamadas “latitud” y “longitud” , luego se asignan las coordenadas “0” para “longitud” y las coordenadas “1” para “latitud “, como se indica en la figura 46.

Figura 46. Asignación de 2 propiedades al dataframe "dfff”, "latitud “y "longitud" – reproyectadas

Fuente. esta investigación

Luego se imprime para verificar si se realizó bien la conversión, como se indica en la figura 474.

Figura 47. Creación de 2 nuevas columnas "latitud" y "longitud –reproyectada de 4326 a 3857.



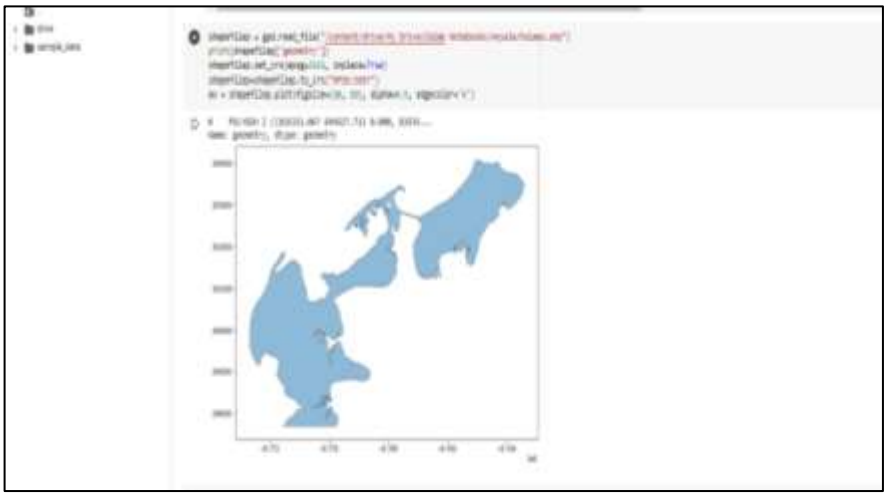
NIVEL AGRESOR	NIVEL VICTIMA	EDAD	SEXO	ESTADO CIVIL	PAIS NACE	CLASE EMPLEADO	PROFESION	ESCOLARIDAD	CODIGO DANE	CANTIDAD	latitud	longitud	lonWgs	latWgs
NO REPORTADO	APE	21	MASCULINO	SOLTERO	COLOMBIA	EMPLEADO PARTICULAR	NO REPORTADA	PRIMARIA	5203000	1	1.807556	-76.765587	-8.769145e+06	201249.598814
APE	APE	15	MASCULINO	SOLTERO	COLOMBIA	ESTUDIANTE	NO REPORTADA	SECUNDARIA	5203000	1	NaN	NaN	NaN	NaN
NO REPORTADO	NO REPORTADO	35	MASCULINO	UNION LIBRE	COLOMBIA	INDEPENDIENTE	NO REPORTADA	SECUNDARIA	5203000	1	1.801198	-76.782361	-8.770668e+06	200541.478575
CONDUCTOR MOTOCICLETA	VEHICULO	34	FEMENINO	CASADO	COLOMBIA	EMPLEADO PUBLICO	ABOGADO	SUPERIOR	5203000	1	1.811041	-76.764804	-8.768659e+06	201726.841066
APE	APE	18	MASCULINO	SOLTERO	COLOMBIA	DESEMPLEADO	NO REPORTADA	SECUNDARIA	5203000	1	1.804828	-76.779536	-8.769364e+06	200945.769407

Fuente. esta investigación

4.2.8.Graficación del polígono de la zona urbana de Tumaco a puntos. Una vez realizados los pasos anteriormente mencionados, se grafica el polígono de la zona urbana del municipio de Tumaco, durante la creación del grafico de dicho polígono surgieron algunos inconvenientes como se mencionó anteriormente.

Antes de graficar el archivo “SHP”, cargamos nuevamente el polígono en coordenada 3857, para verificar que contenga 1 solo polígono y poder graficarlo correctamente, como indica la figura 48.

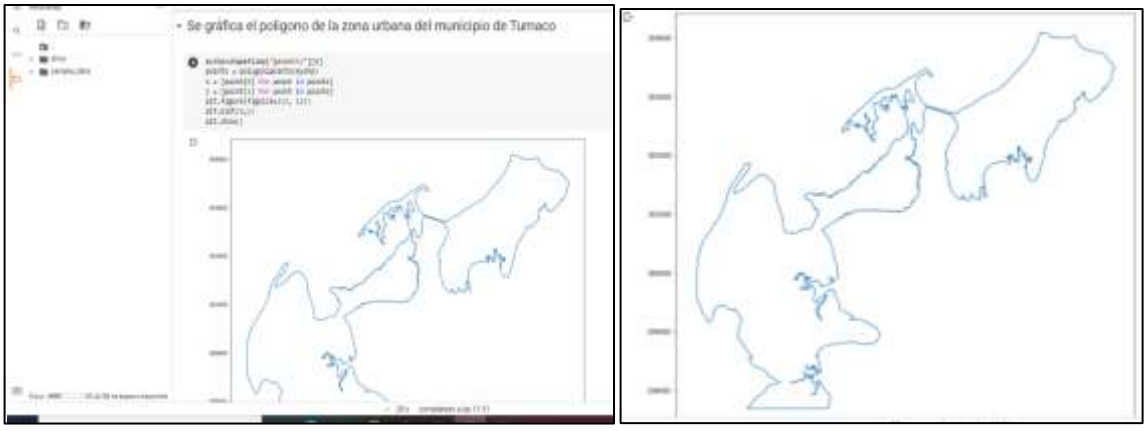
Figura 48. cargue del polígono shapefile en coordenadas 3857 para graficar



Fuente. esta investigación

Con la siguiente línea de comando se procede a realizar la gráfica de dicho polígono como se indica en las siguientes líneas de comando, pero no solo es graficar el polígono si no se sacar en “x” y “y” los puntos del archivo shapefile. Como se observa en la figura 49.

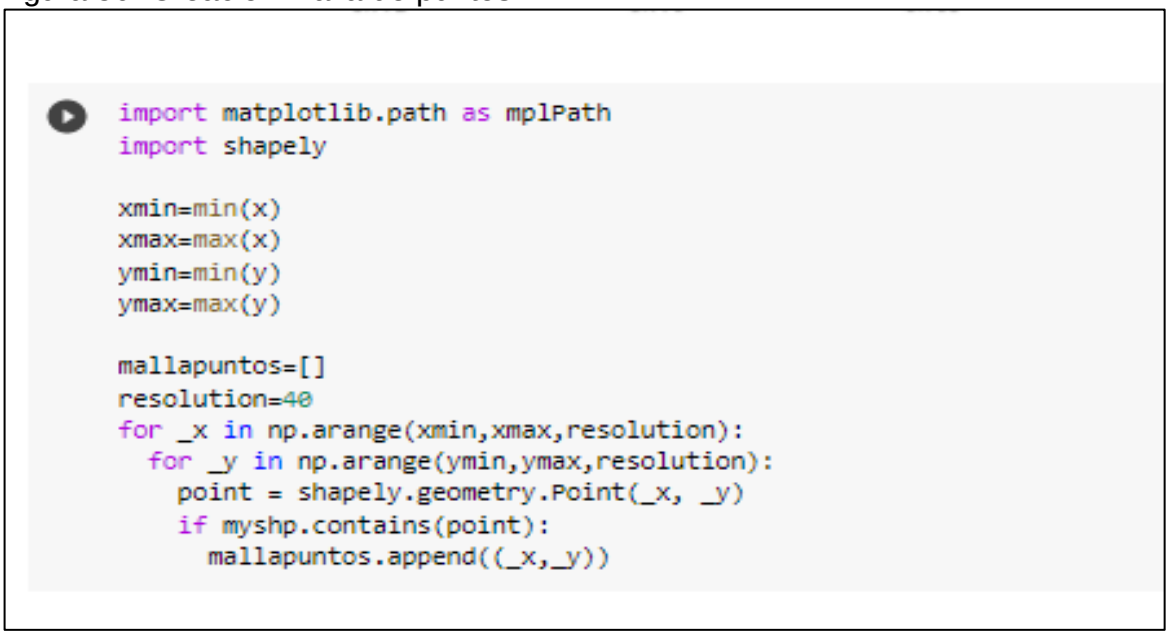
Figura 49. Graficando el polígono de la zona urbana del municipio de Tumaco.



Fuente. esta investigación

4.2.9.Creación de la malla de puntos. Una vez graficado el polígono a utilizar, se procede a crear una malla de puntos que se va a utilizar más adelante, como se observa en la figura 50.

Figura 50. Creación malla de puntos.



```
import matplotlib.path as mplPath
import shapely

xmin=min(x)
xmax=max(x)
ymin=min(y)
ymax=max(y)

mallapuntos=[]
resolution=40
for _x in np.arange(xmin,xmax,resolution):
  for _y in np.arange(ymin,ymax,resolution):
    point = shapely.geometry.Point(_x, _y)
    if myshp.contains(point):
      mallapuntos.append((_x,_y))
```

Fuente. esta investigación

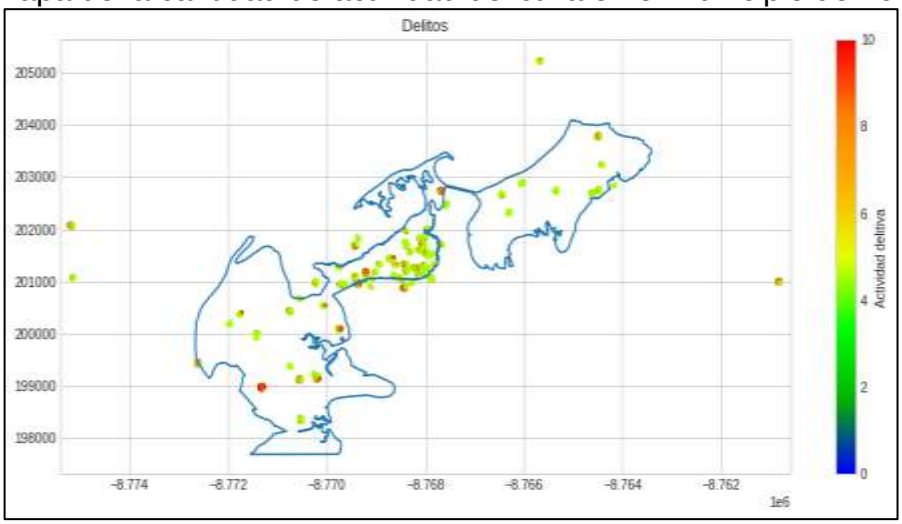
En la creación de la malla de puntos se saca la mínima y la máxima de cada una de las coordenadas “x” y “y”, de esta forma: la “xmin=min(x)”, “xmax=max(x)”, “ymin=min (y)”, “ymax=max(y)”, para poder construir una malla de puntos. Se gráfica solo el polígono de Tumaco, como se visualiza en la figura 46 de esta misma forma se grafica el polígono, pero estructurando la malla de puntos que está dentro del anterior polígono, para poder recorrer los puntos internos pixel por pixel con los de actividad delictivas de cada barrio, para comprobar si coinciden las conversiones del polígono y los datos.

- **Graficar cantidad de actividad delictiva en el polígono.**

Una vez obtenido el poligono del municipio de Tumaco incialmente en XML, el cual fue digitalizado con la herramienta ArcGIS con la se organizó, administró, se analizó, y transformo dicho poligono. Igualmente por medio de esta herramienta, se ubicarón en el mapa shapefile los poligonos correspondientes al casco urbano de Tumaco se unificaron a uno solo shapefile (shp) para facilitar la manipulación del polígono en Python. Una ves obtenido el shapefile final se reprojectaron tanto los datos del observatorio del delito como el poligono (alusivo al mapa interno en elsha-pefile) a

coordenadas de Mercator EPSG:3857 para trabajar con distancias cartecia-nas 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 de delitos como lo indica la figura 51**.**

Figura 51. Mapa de la cantidad de actividad delictiva en el municipio de Tumaco.



Fuente. esta investigación

4.2.10.Realización de una expansión cercana a radial 0 a 10 metros. Una vez verificado de que los puntos coincidieron con el polígono, se procede a realizar una expansión cercana con ruido radial de 0 a 10 metros para que los casos de actividad delictivas tengan una apariencia más real, debido a que las bases de datos suministradas por el observatorio del delito no brindan con exactitud la ubicación en donde ocurrieron dichos casos de actividad delictiva, contando como punto de referencia con el "barrio", pero puede haber sucedido que algunos casos de actividad delictiva ocurrieron 1 o 2 metros arriba o debajo de la ubicación de dicho barrio; es por tal razón que se realizó una expansión cercana a radial de 0 10 metros ,para que dichos casos de actividad delictiva tengan una apariencia más real , para realizar lo anteriormente mencionado aplicamos matemáticas básicas.

3. Se calcula un ángulo aleatorio para saber en qué parte del circulo van a quedar las coordenadas.

4. Se calcula un radio aleatorio de "50", para saber en qué parte del círculo van a quedar las coordenadas.

- Se le suman el radio por el coseno del ángulo para la "longitudes como indica la siguiente línea de código.

dfff["lonWgsr"]=dfff["lonWgs"]+radio*np.cos(angulo)

- Se le suman el radio por el seno del ángulo para la “latitudes” como indica la siguiente línea de código

dfff["latWgsr"]=dfff["latWgs"]+radio*np.sin(angulo)

En este proceso se explica un poco las coordenadas polares de un círculo.

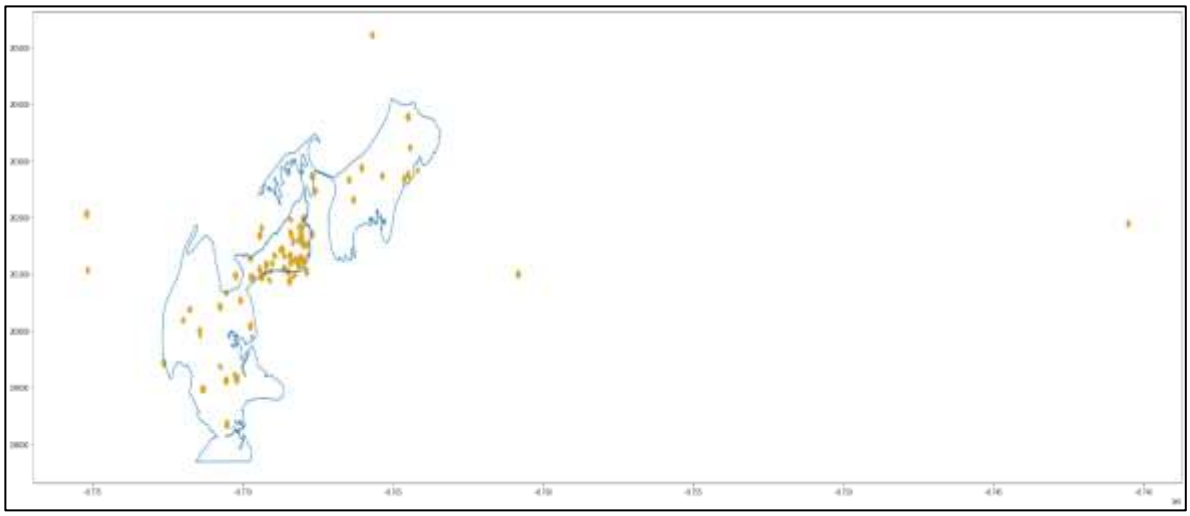
La aplicación de dicha expansión se puede observar en la figura 52.

Figura 52. aplicación de una expansión cercana con ruido radial de 0 a 10 metros para que los casos de actividad delictiva cometidos tengan una apariencia más real.

Fuente. esta investigación

Una vez realizada dicha expansión como se indicó anteriormente, se procede a graficar el polígono con sus respectivos puntos para observar que se efectuaron dichos ajustes de manera correcta, como lo indica la figura 53.

Figura 53. Polígono con sus respectivos puntos con expansión cercana con ruido radial de 0 a 10 metros



Fuente. esta investigación

5. ANALIZAR LOS RESULTADOS DE LOS MEJORES MODELOS OBTENIDOS

5.1 RECONOCER PATRONES A PARTIR DE LA INFORMACIÓN RECOPILADA

Para el análisis se tuvieron en cuenta la aplicación de algoritmos supervisados llamados, KNN es K Vecinos Cercanos, SVR Maquinas de soporte Vectorial para regresión, MLPR Perceptron Multi Capa para regresión, GP Procesos Gaussianos, TD Ar-bol de decisión, RF Bosques Aleatorios y KO Kriging Ordinario.

5.1.1. Análisis de resultados:

5.1.1.1 .Interpolación de los delitos. Para la obtención del mapa que logré 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 6

Tabla 6. Métricas r^2

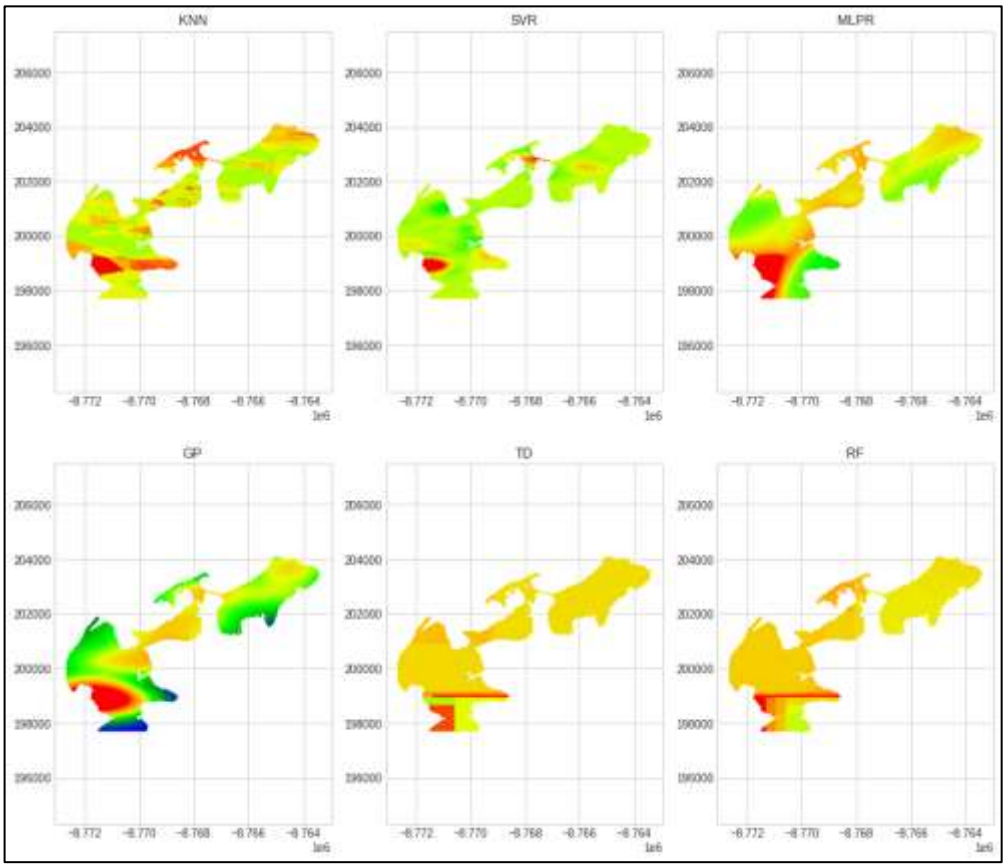
Algoritmos	Hiperparámetros	r^2 prueba	r^2 entrenamiento
KNN	Vecinos=8	0.32	0.45
SVR	Kernel=RBF Gamma=20 Cte Regularización=100	0.35	0.32
MLPR	Iteraciones=100K Capas ocultas=4 Topologia=64,32,8,16	-8.06	-0.00
GP	Kernel=RBF	0.36	1
TD	Profundidad = 4	0.38	0.37
RF	Profundidad=4 Arboles=50	0.32	0.88
KO	Variograma=hole-effect	**0.85**	**0.87**

Fuente. esta investigación

Una vez obtenidas las métricas de r2, 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 57 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 minimos de las latitudes y longitudes del polígono del municipio de Tumaco, de los cuales solo se tomarón 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 54.

Figura 54. Interpolación de los delitos en el polígono de la zona urbana del municipio de Tumaco

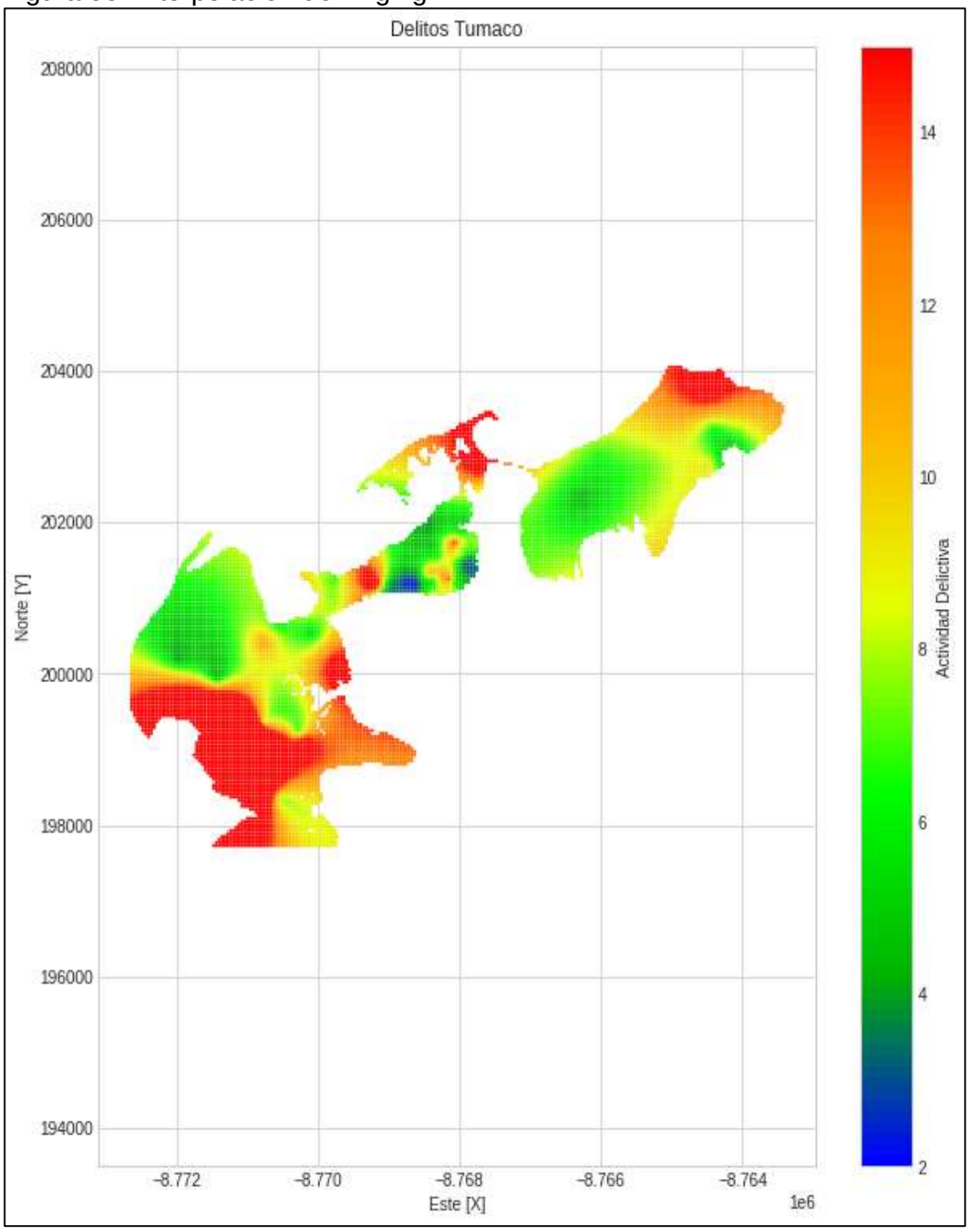


Fuente. esta investigación

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 actos de actividad delictiva en dicho lugar coinciden con los datos reportados en el Observatorio del Delito, de igual manera el color naranja que hace referencia al lugar donde se cometen actividad delictiva de una manera moderada pero no es un lugar altamente riesgoso para transitar, como lo indica la figura 53.

Figura 55. Interpolación de Kriging.

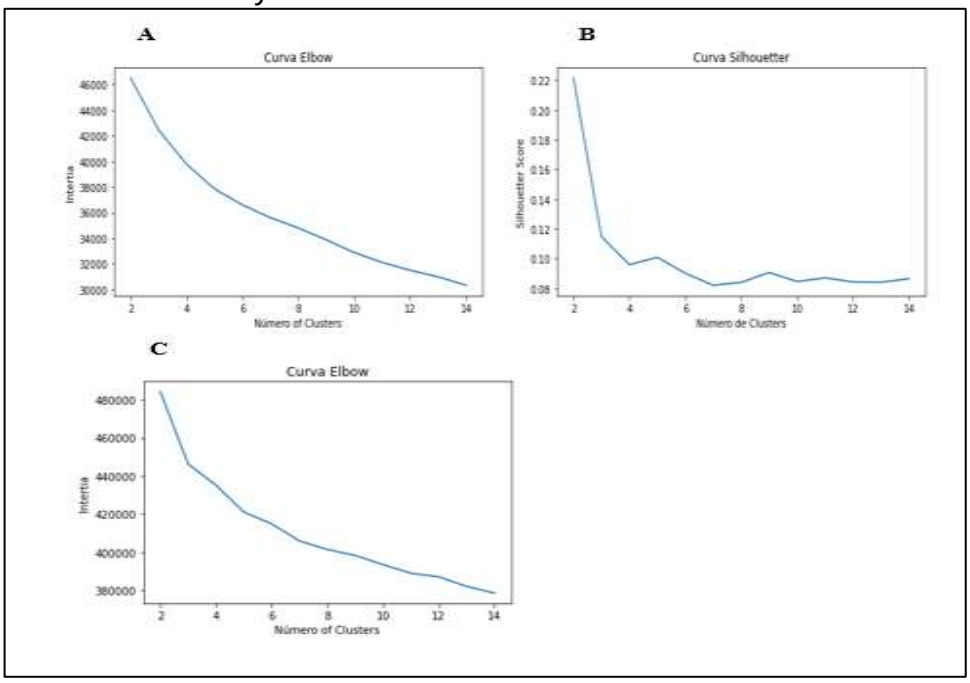


Fuente. esta investigación

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 y k-modas (Figura 54 A y B) y codo para K-Prototype (Figura 56 C).

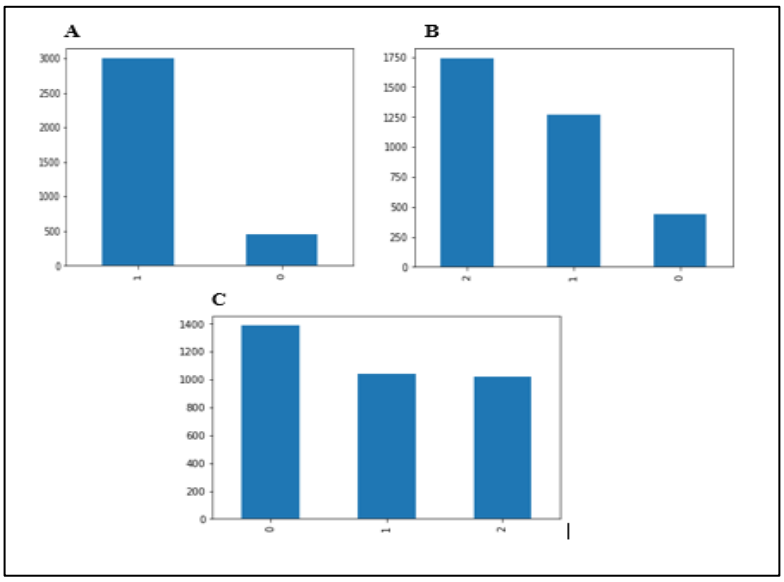
Figura 56. Curvas Codo y Silueeta.



Fuente. esta investigación

De la Figura 54 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 55, indicada por la gráfica “C”, donde se observa que las agrupaciones están mejor balanceados que el modelo K-protype (ver Figura 57).

Figura 57. Grupos de clustering-kmean y k-prototype..



Fuente. esta investigación

Como se observa en la Figura 55 las agrupaciones realizadas con 3 grupos tanto para K-Means (Figura 55A y B), como para K-Prototype (Figura 55C) ofrecen el mejor balance en sus grupos. A continuación, se describen los patrones encontrados en estos dos modelos.

5.1.2. Interpretación con Kmeans. Para realizar esta interpretación sobre este algoritmo primero se hallo 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 permetuados 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.

5.1.3. Interpretación K-Prototype. En este algoritmo se encontró un mejor balance en sus agrupaciones y arrojó que la actividad delictiva se concentra en victimas solteras, cuya movilidad es a pie al igual que la movilidad del agresor.

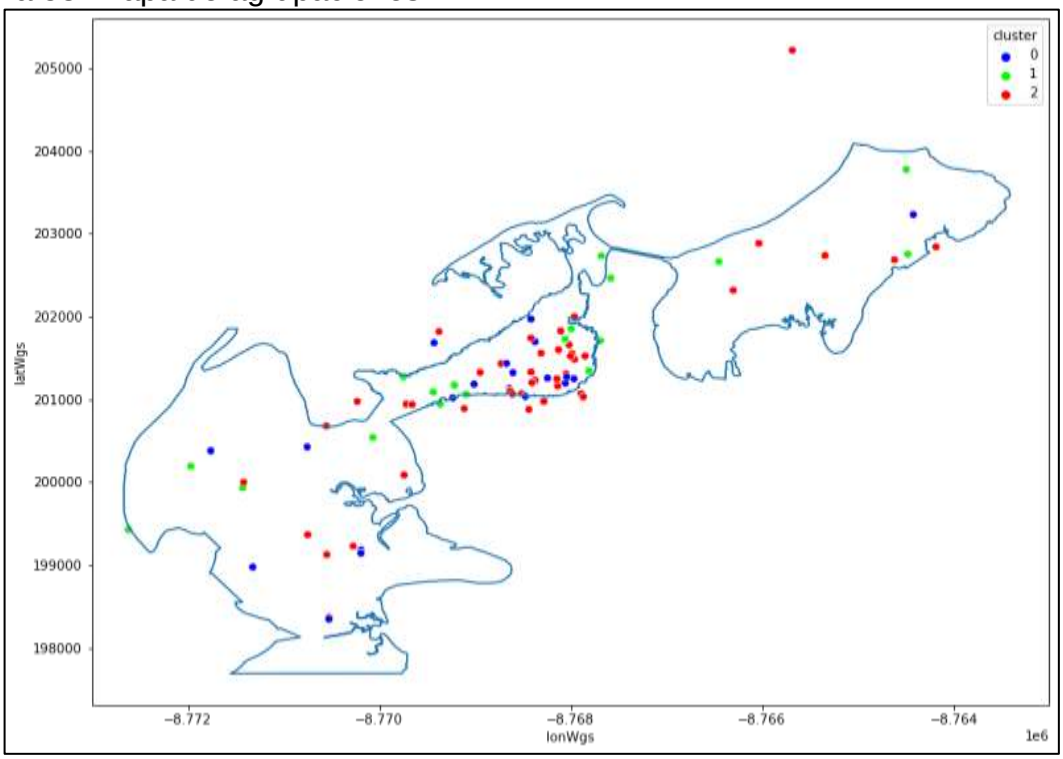
El cluster 0 tiene el 37% de los datos y agrupa a aquellas amenazas concentradas en el barrio Obrero en las coordenadas (1.799773500389714, -78.78134212860543), en el mes de julio los días lunes en la tarde, a empleados particulares hombres, con estudios de secundaria, de 35 a 40 años de edad, generalmente 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 idependientes hombres, con estudios de primaria, de 25 a 30 años de edad, generalmente perpetua-dos con arma de fuego.

El cluster 0 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, general-mente 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 58.

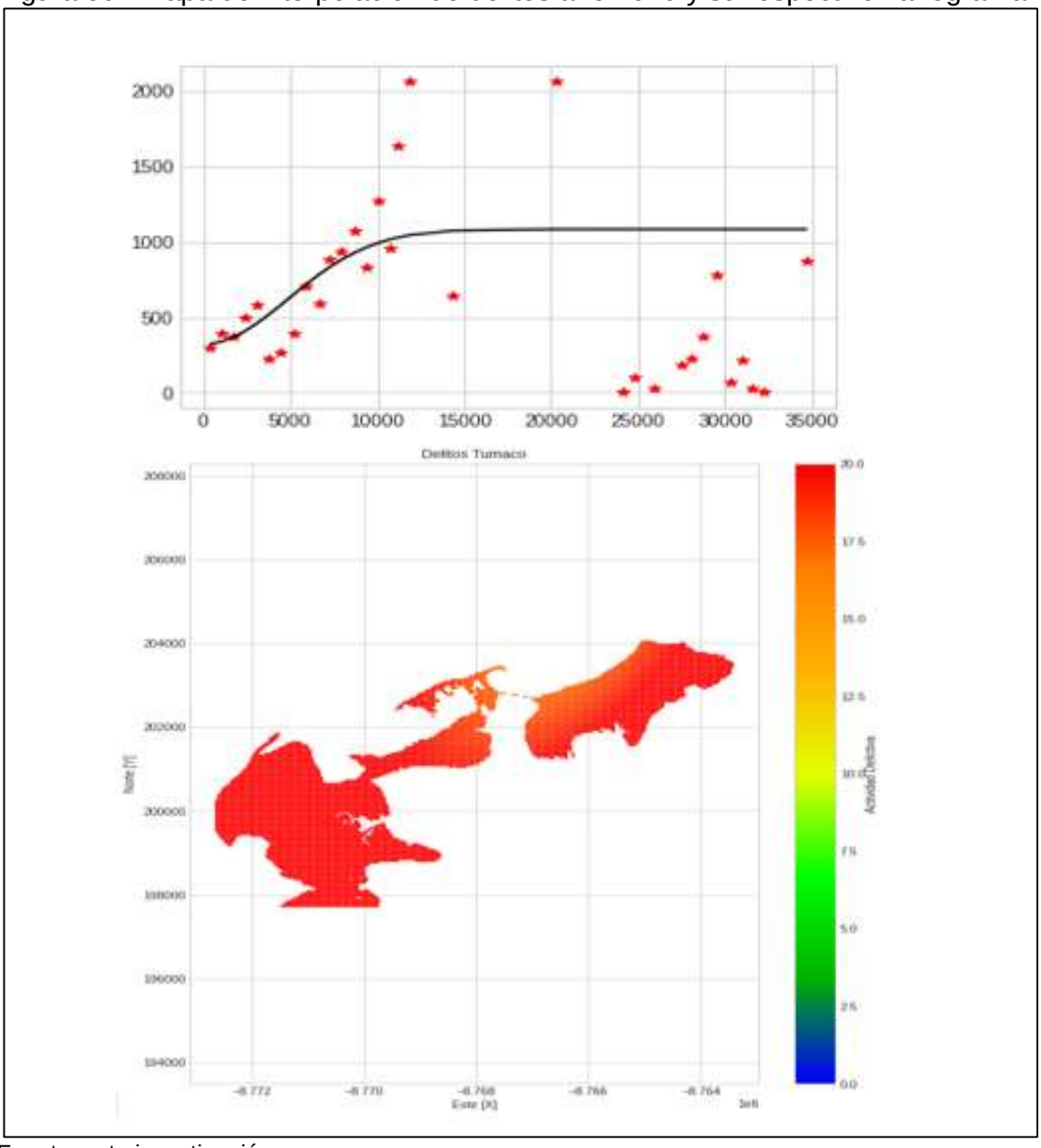
Figura 58. Mapa de agrupaciones



Fuente. esta investigación

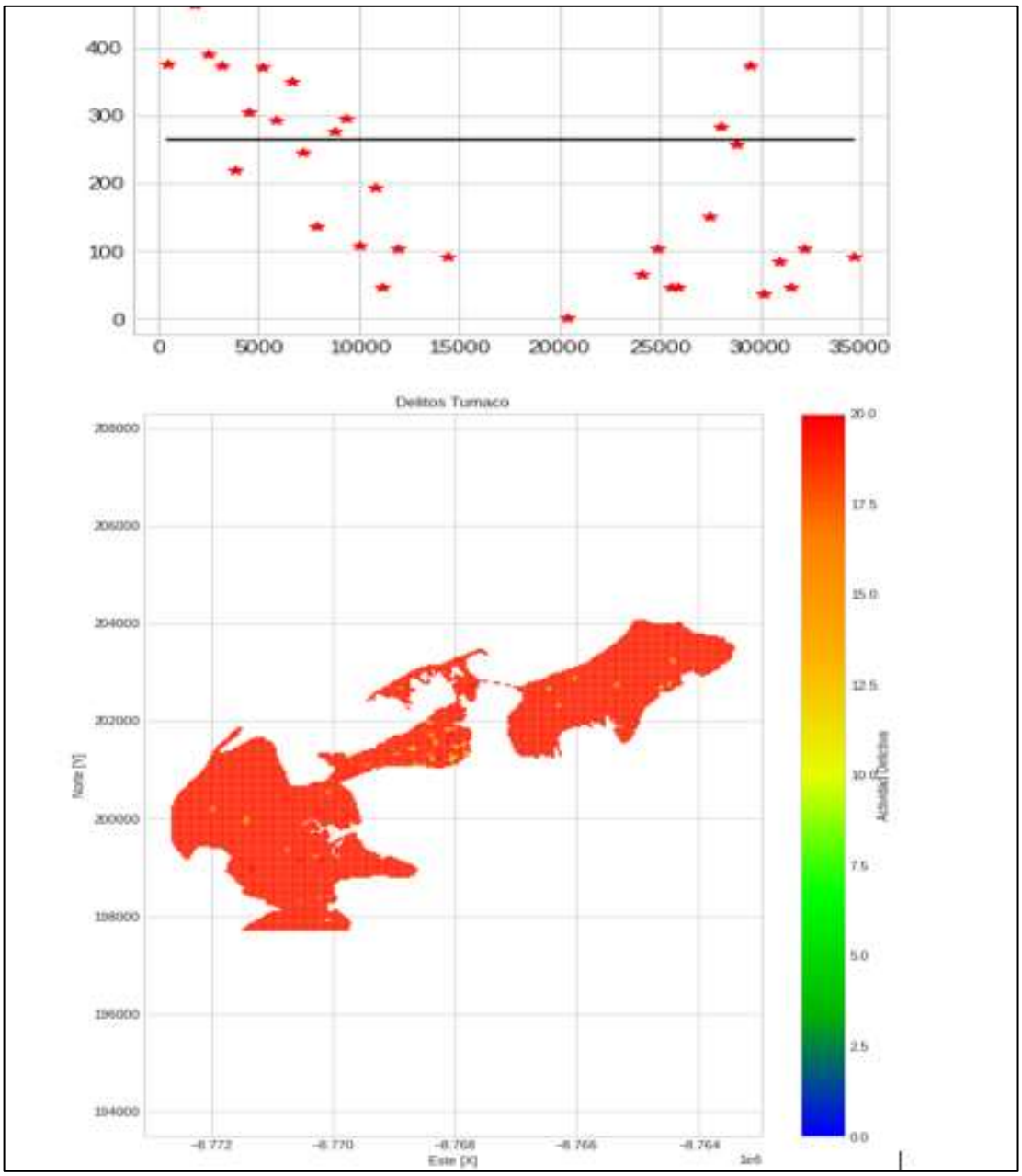
5.1.4. Mapas de interpolación de los delitos por años. Como ya se obtuvo el mapa general, este procedimiento se convierte a una función, permitiendo generar el raster y los Variograma gaussianos de cada año, donde en promedio la rasterización de cada mapa es alrededor de 6.30 minutos, y como resultado se muestran en las figuras desde la 59 hasta la 66, de las cuales según las escalas de los mapas mencionados a continuación van de 0 a 20 que se representan de menor cantidad de delitos al máximo de ellos, cada uno de estos se representaron por los colores (azul,verde,amarillo,naranja y rojo), donde el azul indica menores casos de actos delictivos hasta llegar a la zona de color rojo, el cual indica el máximo caso de actos delictivos, de esta manera se observa el comportamiento de las actividades delictivas en un rango de tiempo.

Figura 59. Mapa de interpolación de delitos año 2010 y su respectivo variograma



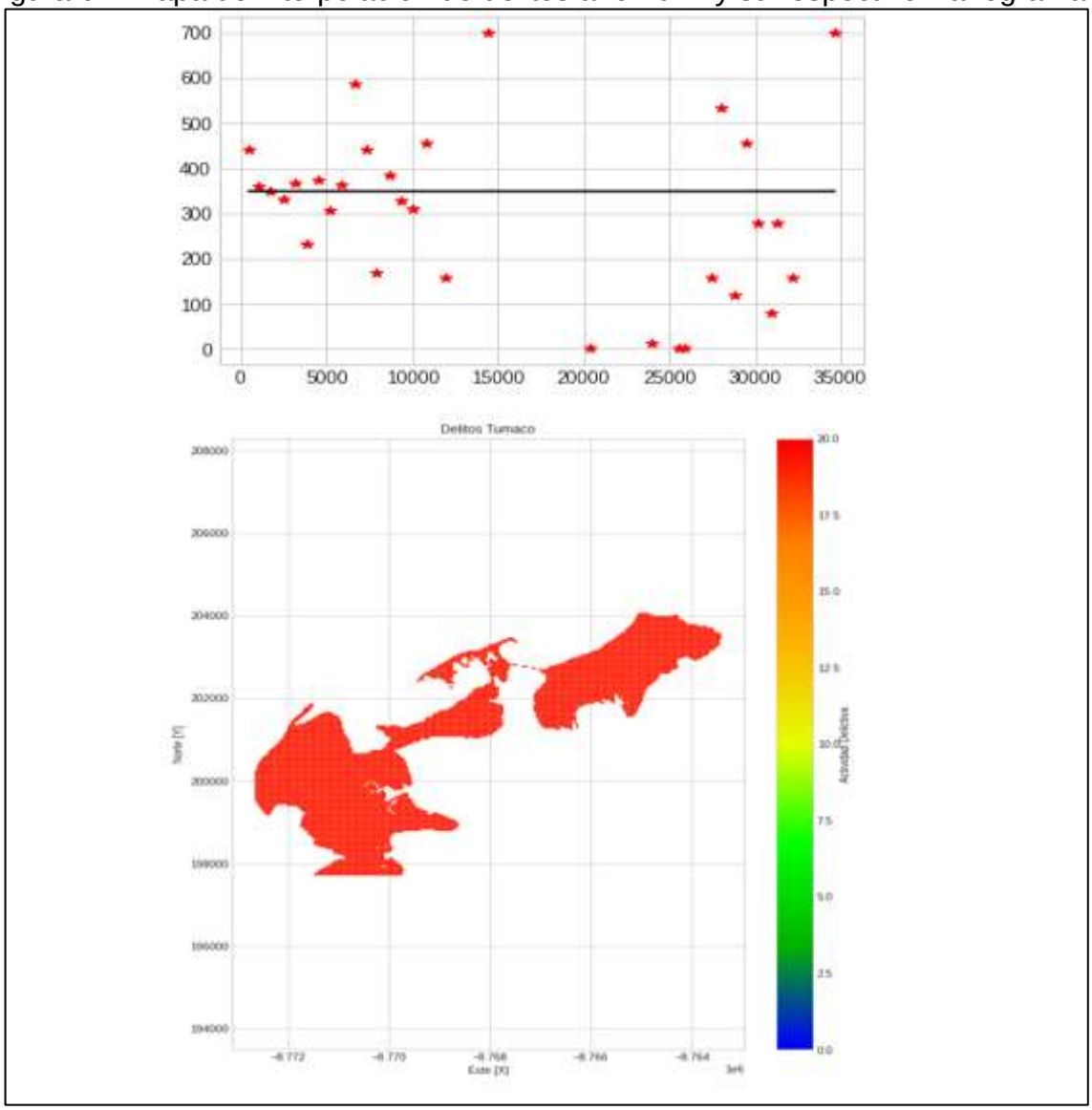
Fuente. esta investigación

Figura 60. Mapa de interpolación de delitos año 2011 y su respectivo variograma



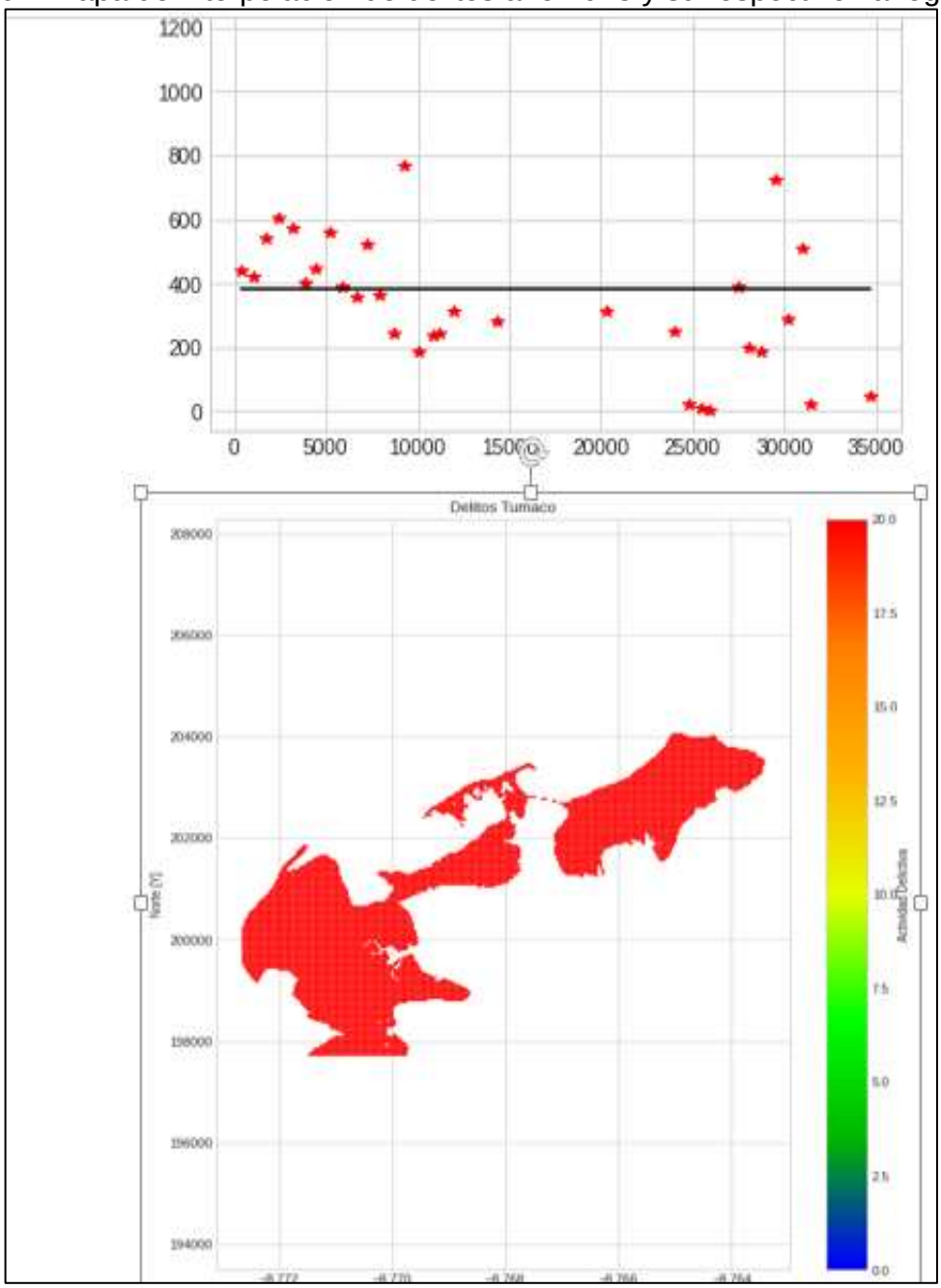
Fuente. esta investigación

Figura 61. Mapa de interpolación de delitos año 2012 y su respectivo Variograma



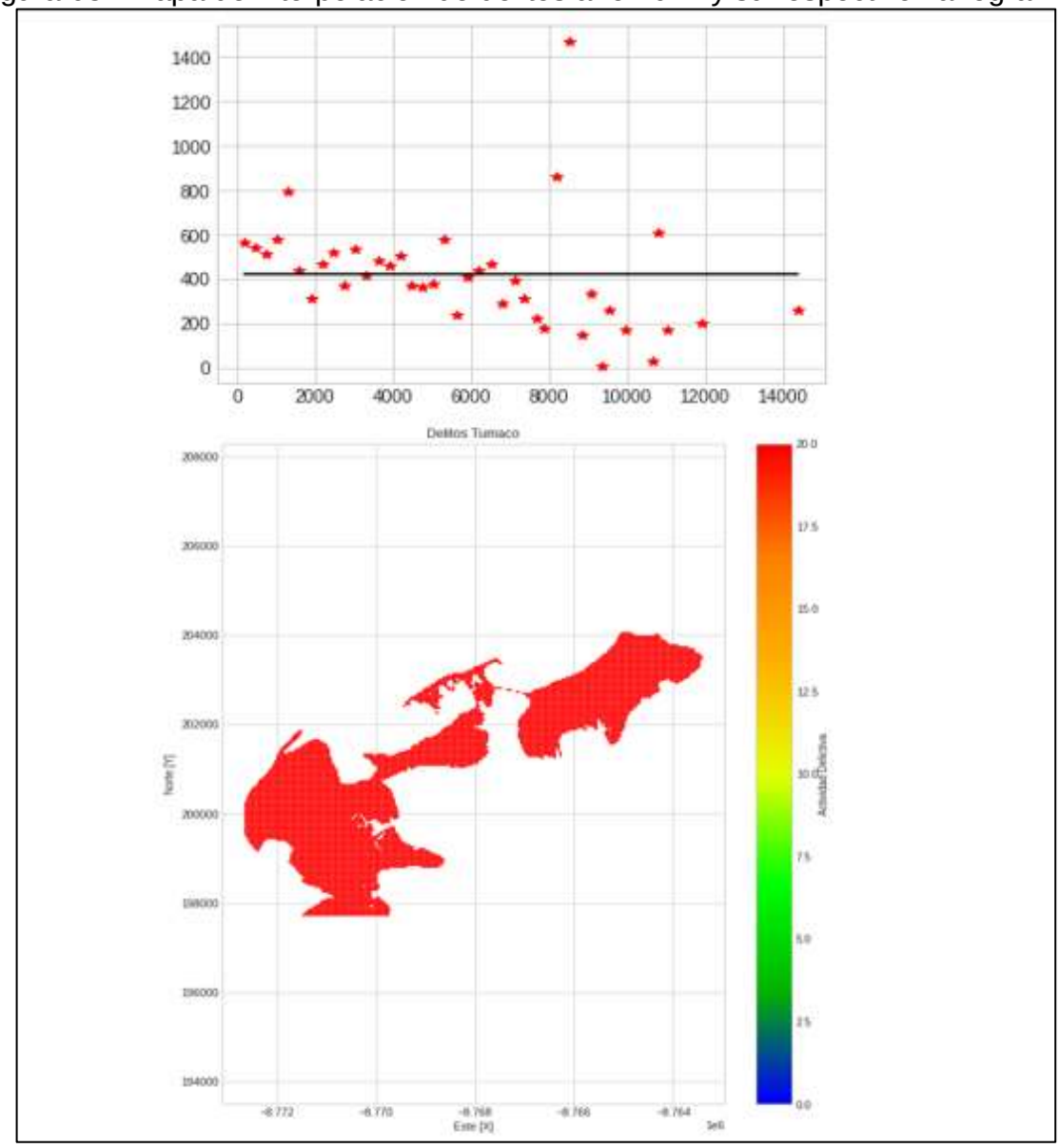
Fuente. esta investigación

Figura 62. Mapa de interpolación de delitos año 2013 y su respectivo variograma



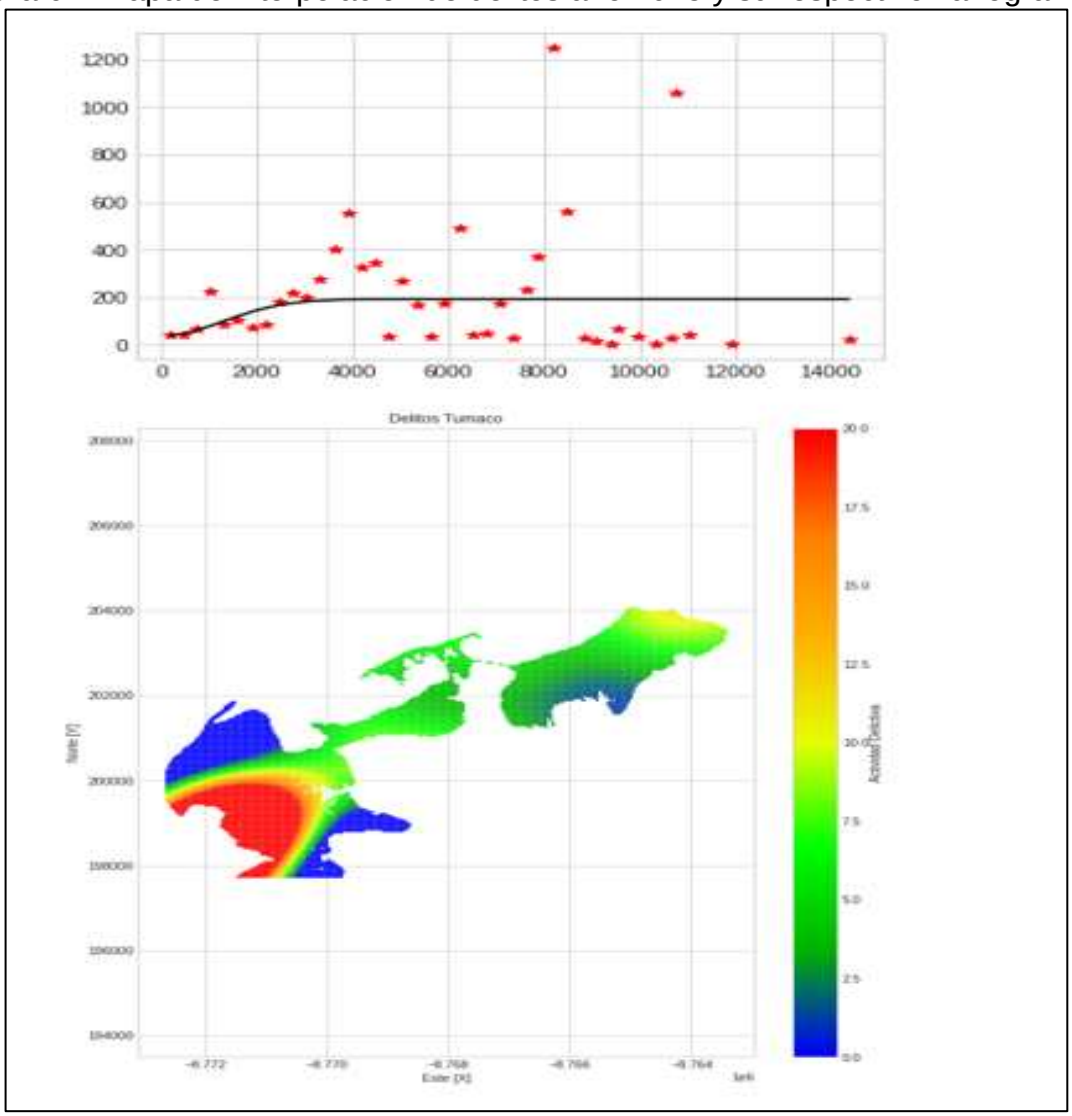
Fuente. esta investigación

Figura 63. Mapa de interpolación de delitos año 2014 y su respectivo variograma



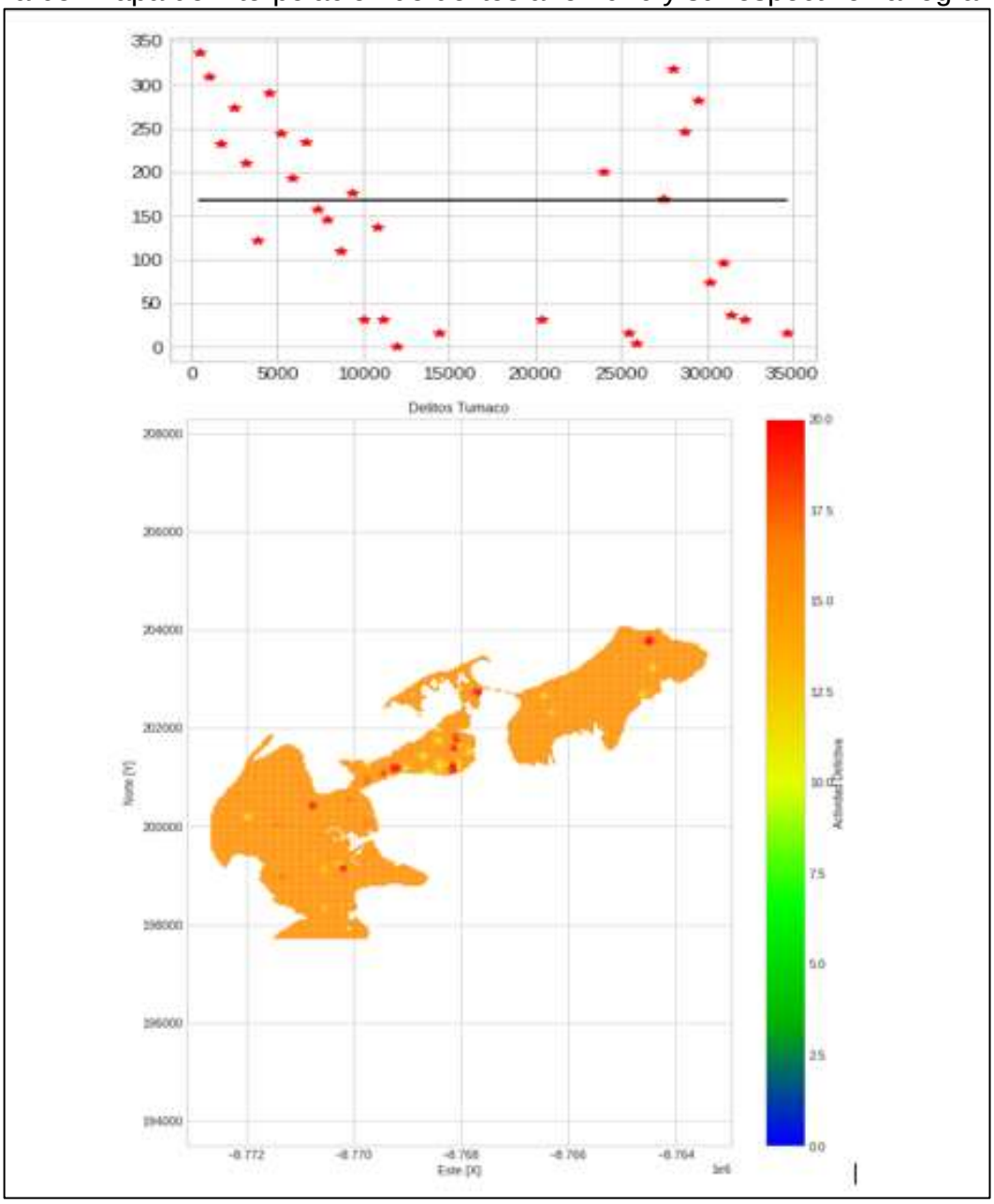
Fuente. esta investigación

Figura 64. Mapa de interpolación de delitos año 2015 y su respectivo variograma



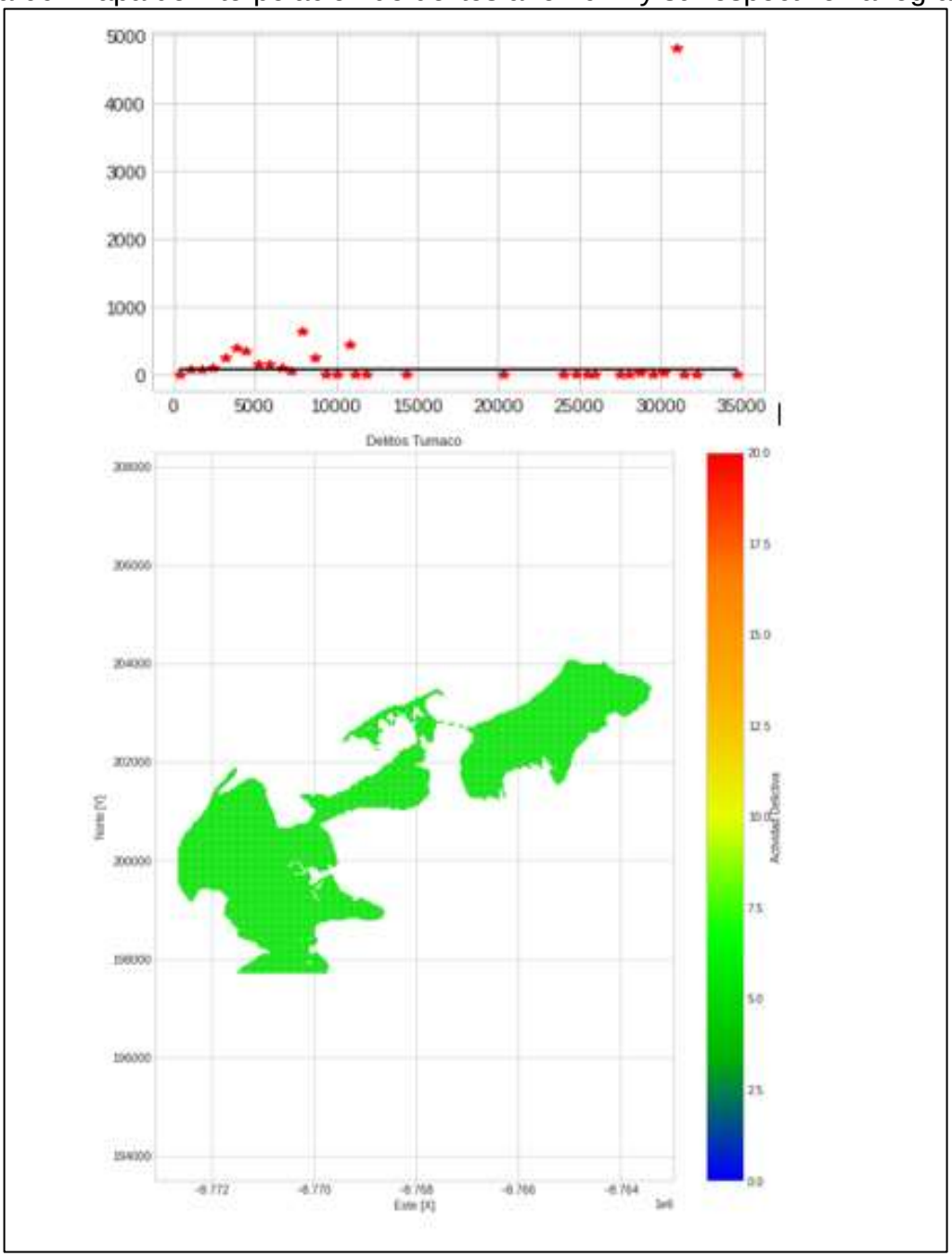
Fuente. esta investigación

Figura 65. Mapa de interpolación de delitos año 2016 y su respectivo variograma



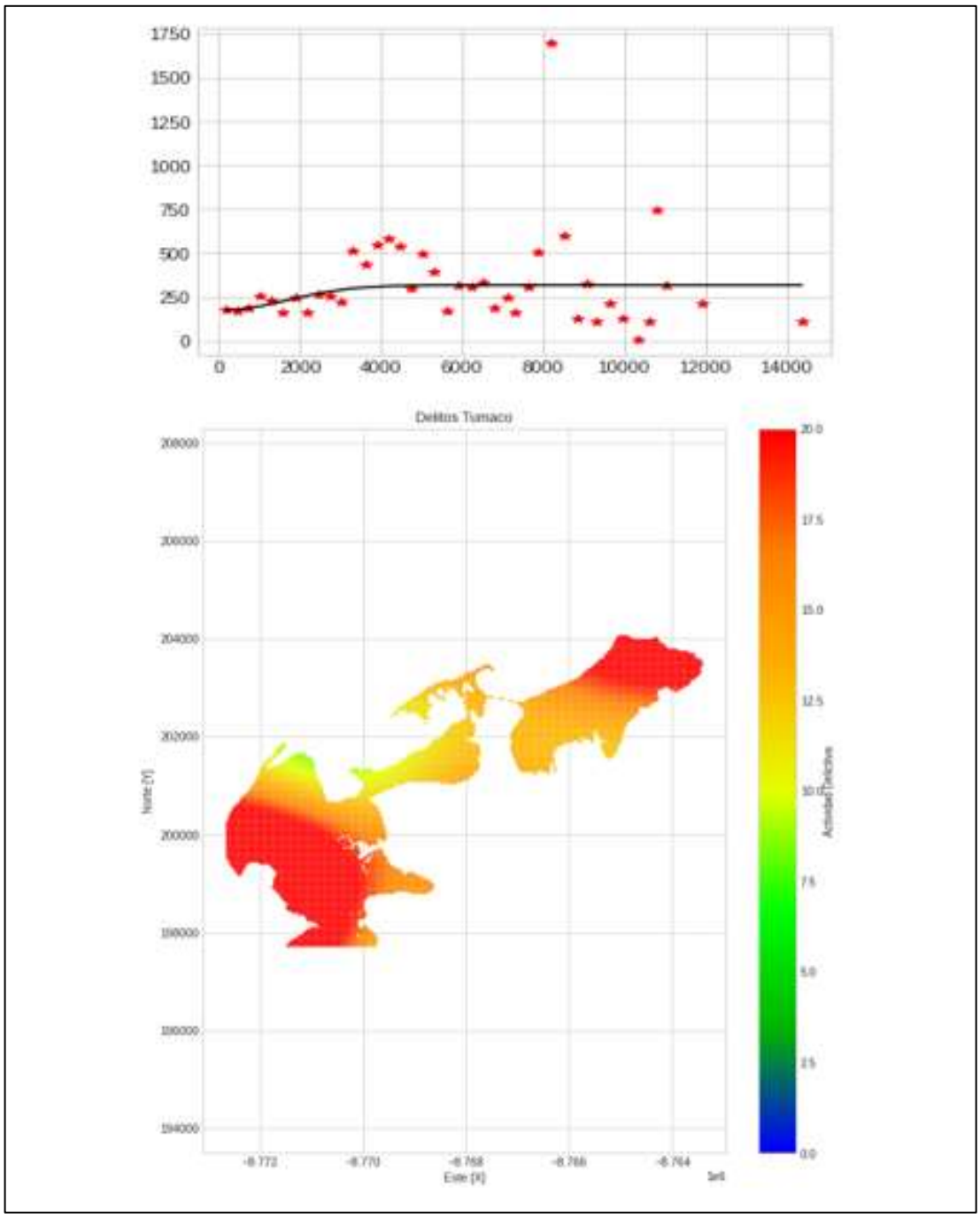
Fuente. esta investigación

Figura 66. Mapa de interpolación de delitos año 2017 y su respectivo variograma



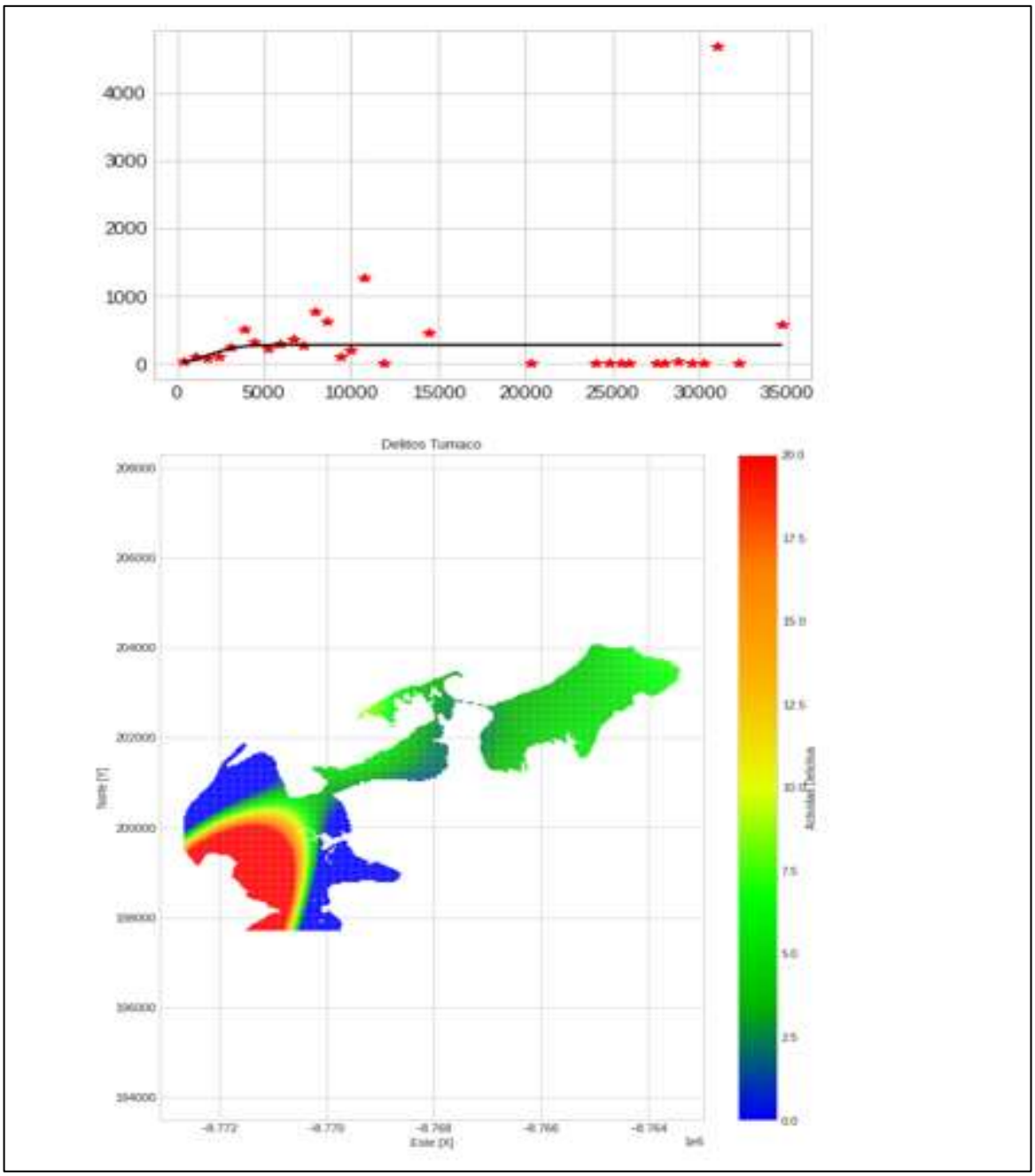
Fuente. esta investigación

Figura 67. Mapa de interpolación de delitos año 2018 y su respectivo variograma



Fuente. esta investigación

Figura 68. Mapa de interpolación de delitos año 2019 y su respectivo variograma

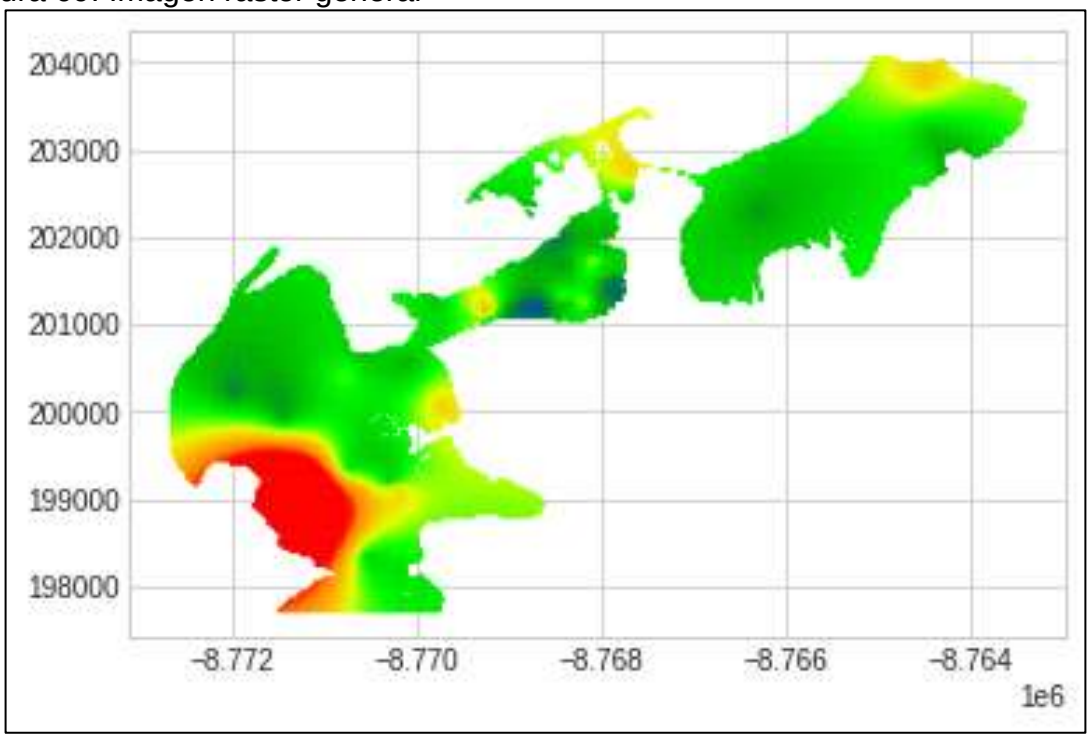


Fuente. esta investigación

5.1.5. Aplicación de imágenes raster. Una vez obtenidos los resultados brindados por los respectivos algoritmos de regresión y agrupación se aplicará imagen raster lo cual conlleva a definir el tamaño de la imagen que estará dado por valores máximos, mínimos y resolución, es decir se crea un vector que conforma todos los puntos geográficos que se tiene en X, Y con el fin de saber cuántas columnas y filas debe tener la imagen raster (malla de puntos), enseguida se convierte la malla de puntos original en una matriz array para agilizar y calcular el total de puntos en la matriz.

Para generar la imagen raster (Pixeles) se crea una lista de filas, donde se recorren matricialmente entre los límites dados anteriormente X (Filas), Y (Columnas) así se empieza a cargar los puntos de geometría para saber si tiene contenido, si es así se le asigna un color dentro de la malla de colores, de lo contrario se establece un punto en blanco (transparente) representado en la siguiente figura 69.

Figura 69. Imagen raster general

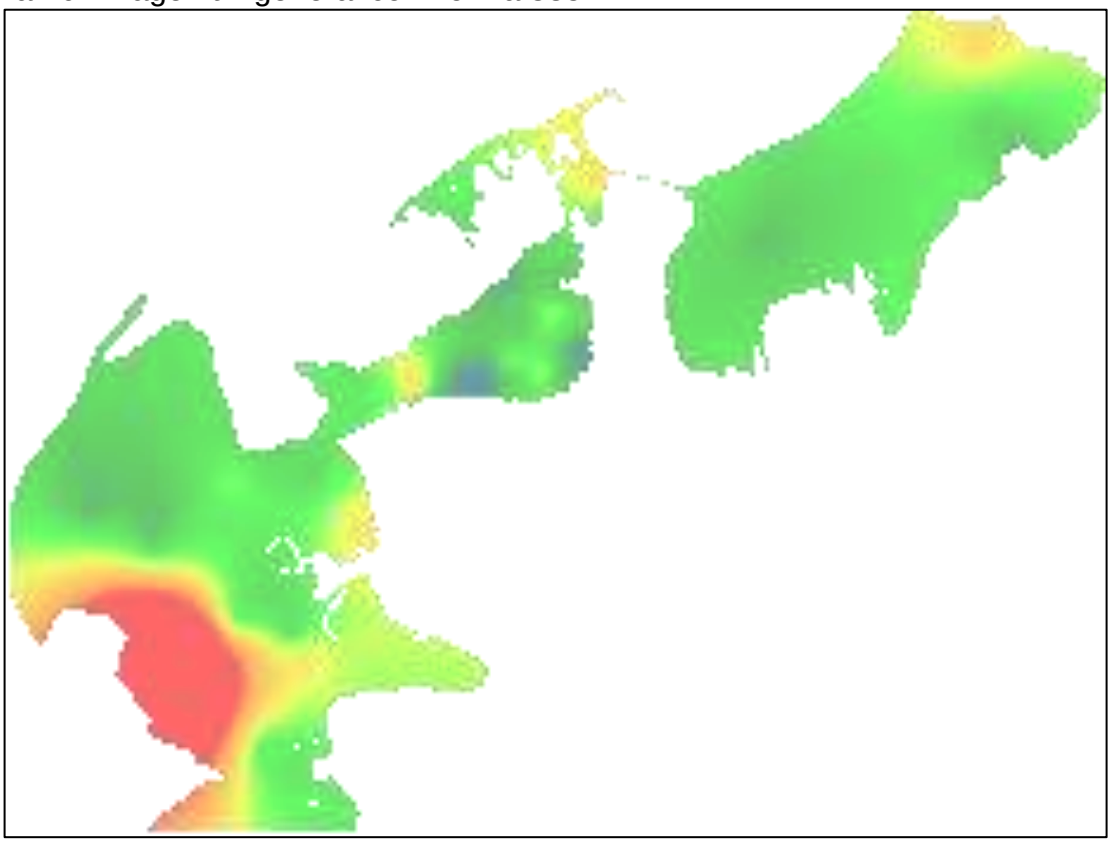


Fuente. esta investigación

5.1.6. Generación de imágenes.Tiff. Una vez obtenida la imagen raster, como se indica en la figura 68, se pasa a transformarlas a imágenes.Tiff, en el cual se determina el tamaño de la imagen, un número de valores y resolución en X, Y, se genera un geotransformador y se importan los datos epgs 3857 y 4326 como se observan en la figura 70.

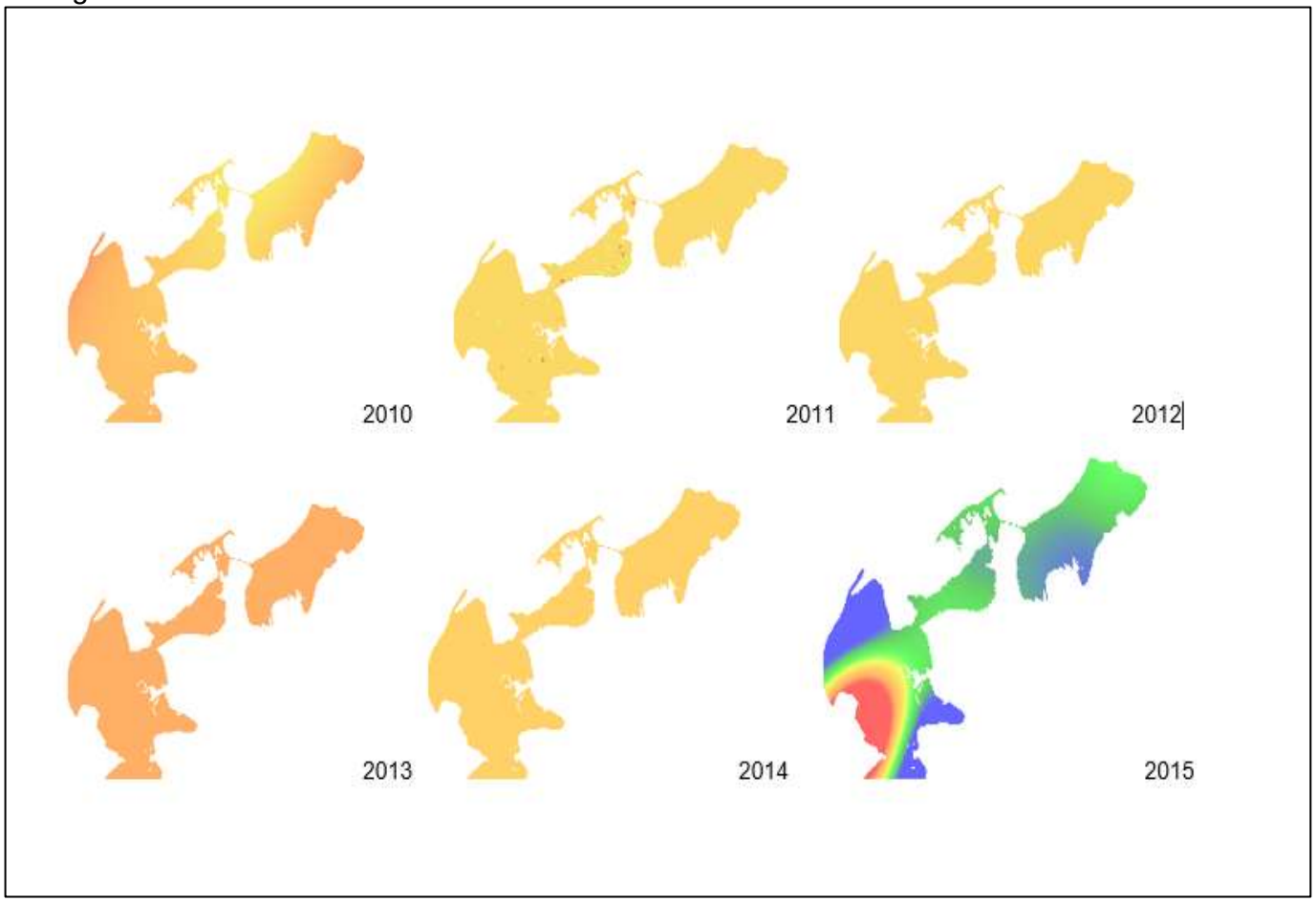
- mágenes**.Tiff** con norma 3857.

Figura 70. Imagen.tiff general con norma 3857



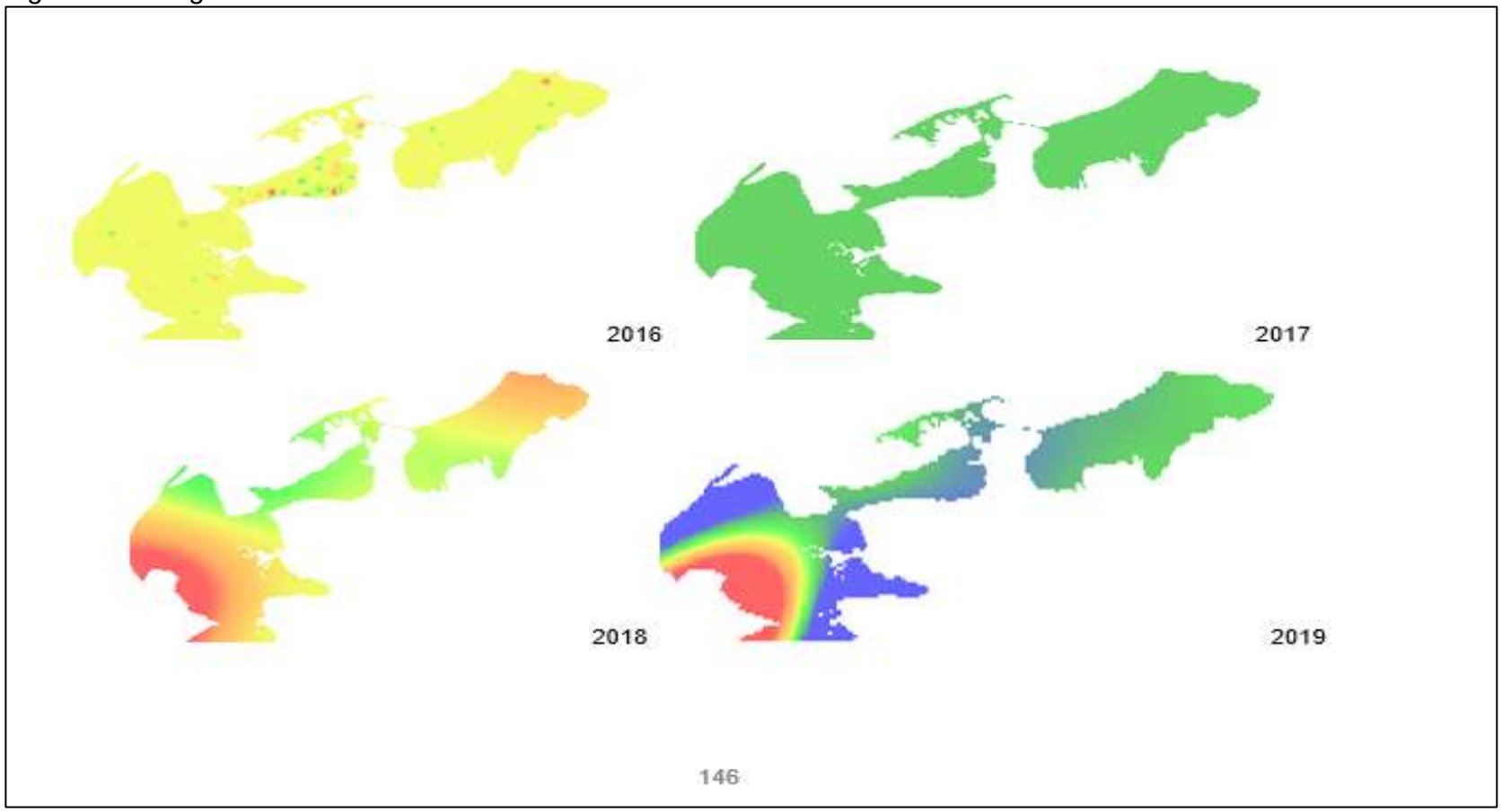
Fuente. esta investigación

Figura 71. Imágenes.tiff año 2010 hasta 2015 con normas 3857.



Fuente. esta investigación

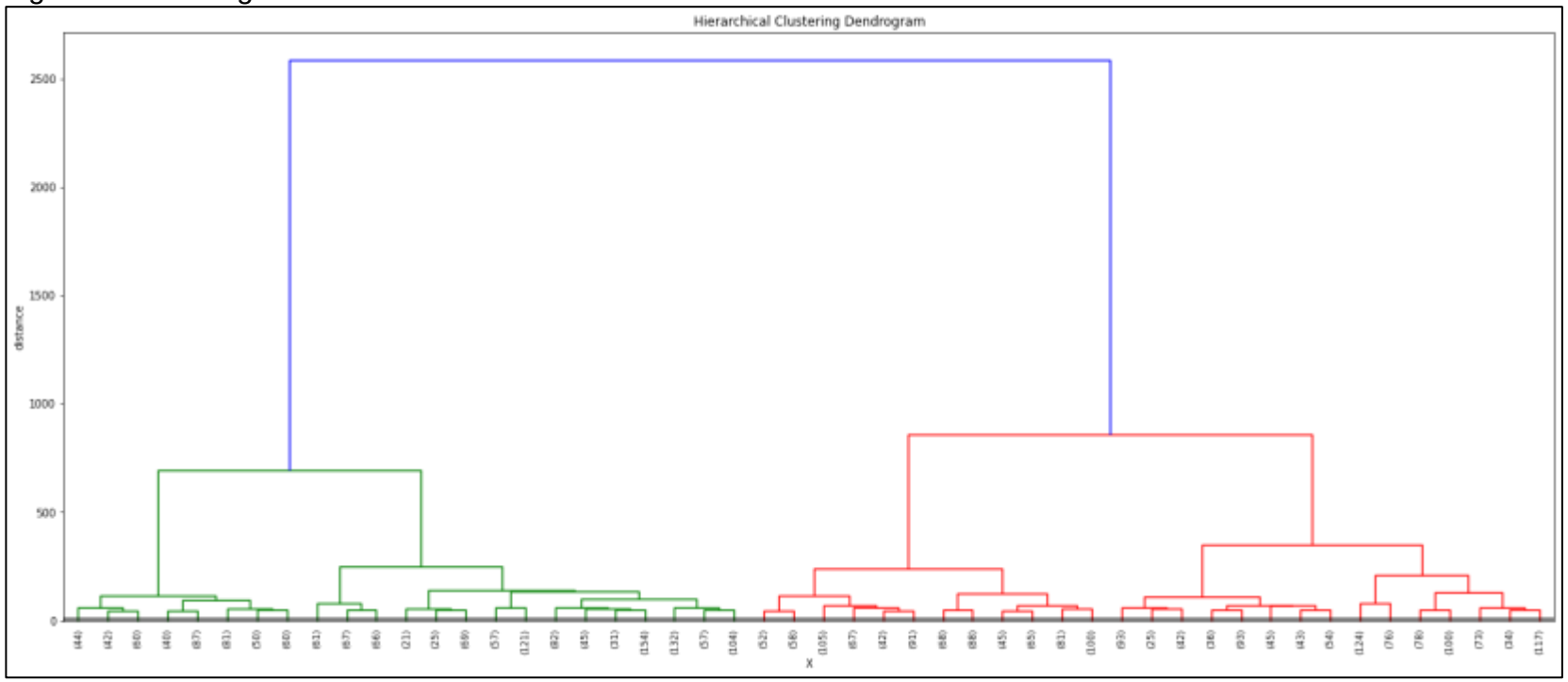
Figura 72. Imágenes.tiff año 2016 hasta 2019 con normas 3857



Fuente. esta investigación

5.1.7. Aplicación del Dendograma:

Figura 73. Dendograma hierarchical

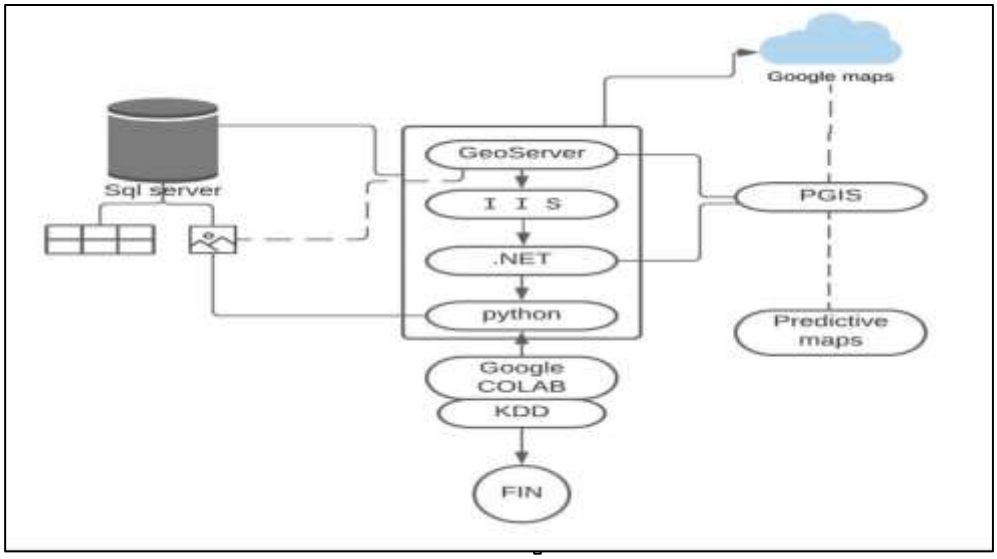


Fuente. esta investigación

5.2 CONSTRUCCIÓN DE LA PLATAFORMA TECNOLÓGICA PARA LA GESTIÓN DE DATOS

5.2.1.Diagrama de despliegue. En la figura 74 se observa la arquitectura del diagrama de despliegue que permite visualizar que se trabajó con un motor de base de datos sql server en conjunto con el gestor de servicios geográficos GeoServer, la capa de aplicaciones que está dentro del servidor consiste del ya mencionado GeoServer también de I I S (Internet Information Services) este módulo sirve para desplegar la aplicación web, framework.net para desarrollar la lógica de negocio del lado del pack, python se encarga de las predicciones para generar los mapas esos mapas que son presentados gracias a los servicios externos dados por google maps y google colab la mayoría del proceso KDD esta implementado en google colab donde se hizo la extracción, filtración, limpieza, cruce de datos y minería de datos en el cual se probaron varios modelos y finalmente la visualización de los mapas.

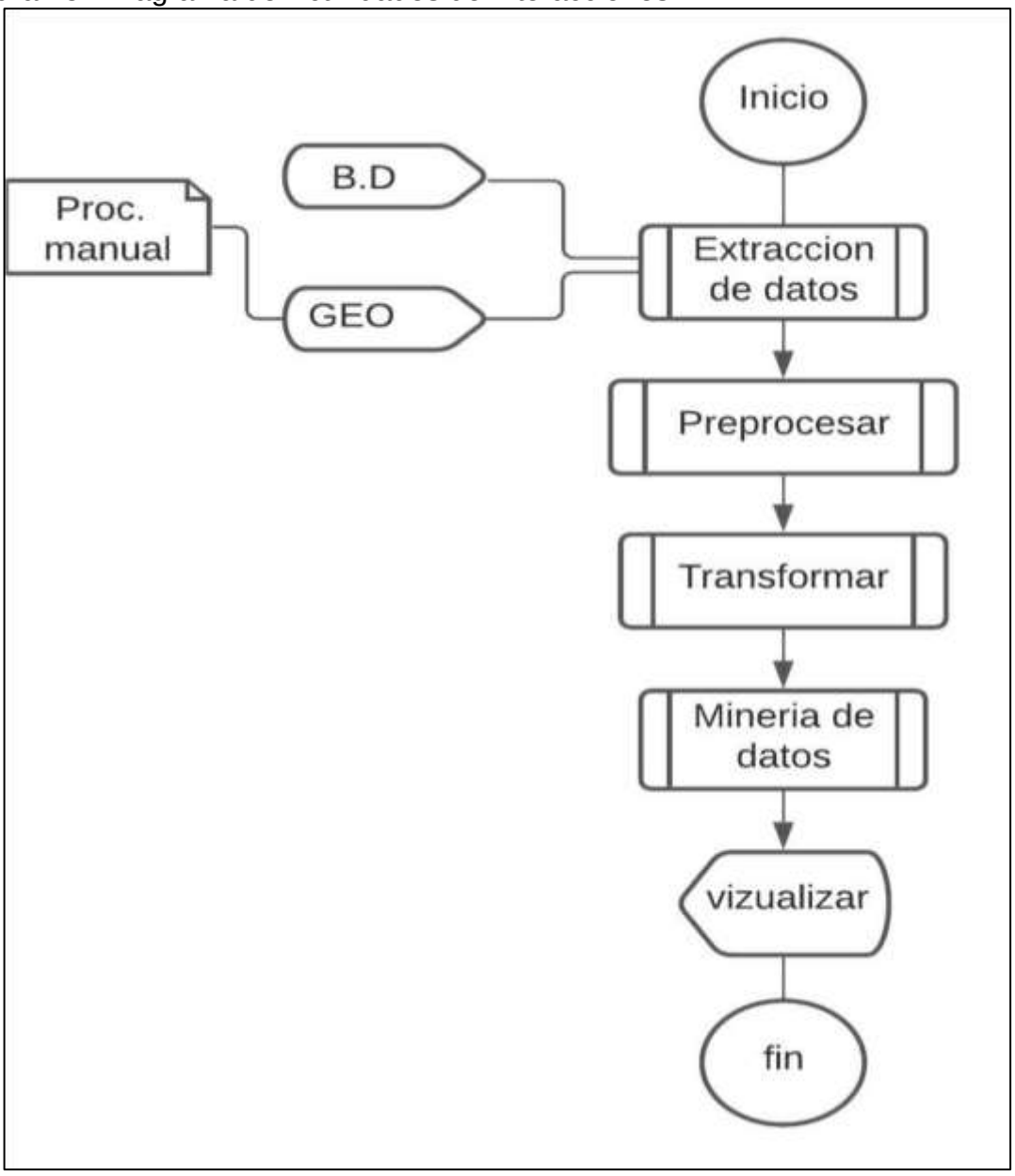
Figura 74. Diagrama de Despliegue



Fuente. Esta investigación

5.2.2. Diagrama de Comportamientos. En la figura 75 se observan los pasos de la metodología KDD utilizados, para la extracción de datos se utilizó dos fuentes de datos que provienen del observatorio del delito y datos georreferenciados manualmente luego se pre-proceso los datos, se transformó, se hizo minería y finalmente se los visualizó en un prototipo funcional de visualización de mapas.

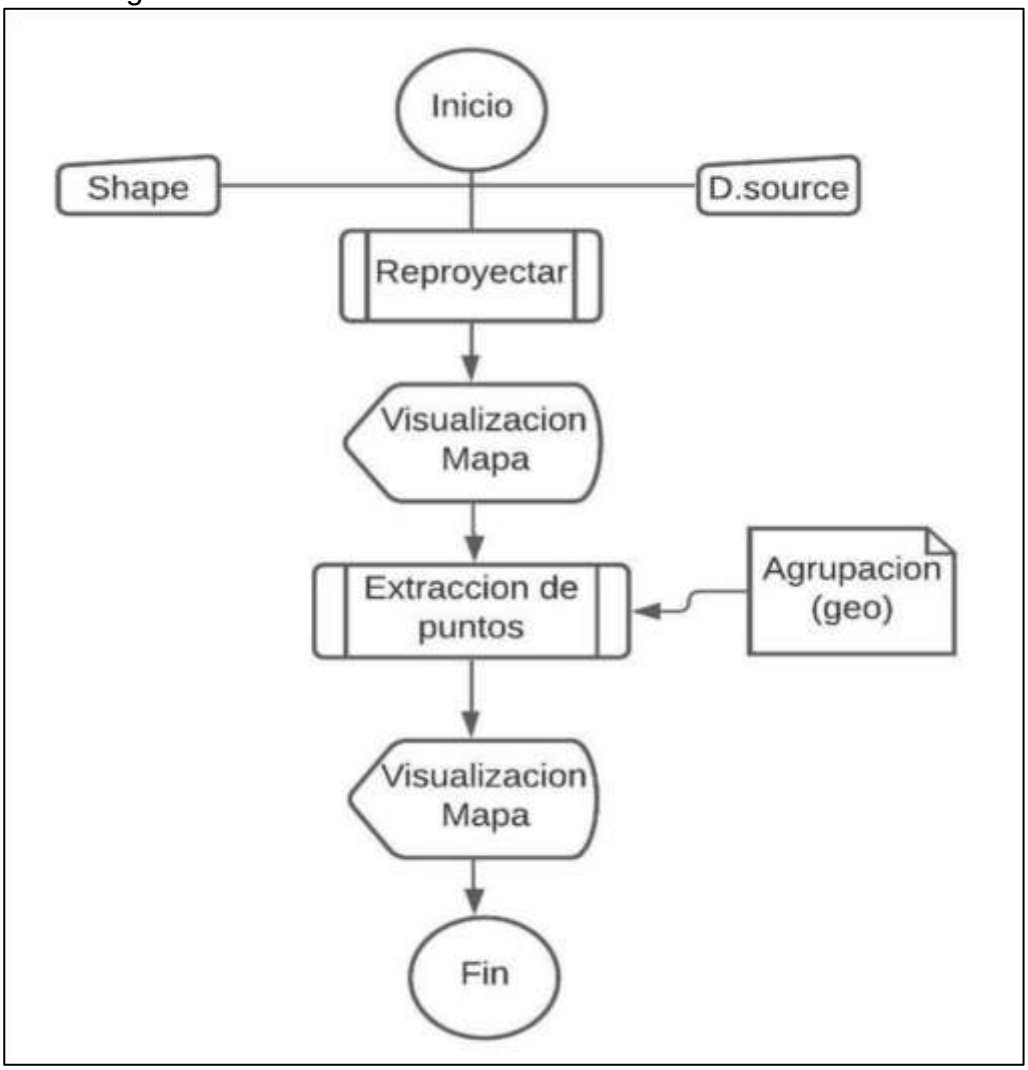
Figura 75. Diagrama de Actividades de Interacciones KDD



Fuente. esta investigación

5.2.3 Diagrama de actividades Transformación de Datos. Para la transformación se trajo el conjunto de datos y el Data Source y con la biblioteca de geopandas, se reproyecta el shape de magna Colombia Oeste 3115 y lo pasamos a Mercator 3857 para trabajar en metros, después realizamos el mismo proceso con el Data Source donde reproyecta de 3115 a 3857, esto se realizó en la misma transformación para que los puntos puedan visualizarse correctamente y coincidan, luego se hizo una expansión de puntos con ruido radial aleatorio para que los delitos se parezcan más al comportamiento real, ya que en la realidad no se presentan delitos en un mismo punto geográfico sino en dispersión, luego se agruparon los datos para mirar cuantos delitos se cometieron en las zonas donde realizamos una visualización por colores (Rojo, Amarillo, Verde y Azul) donde rojo es donde se cometen más delitos y Azul es donde menos delitos se cometen.

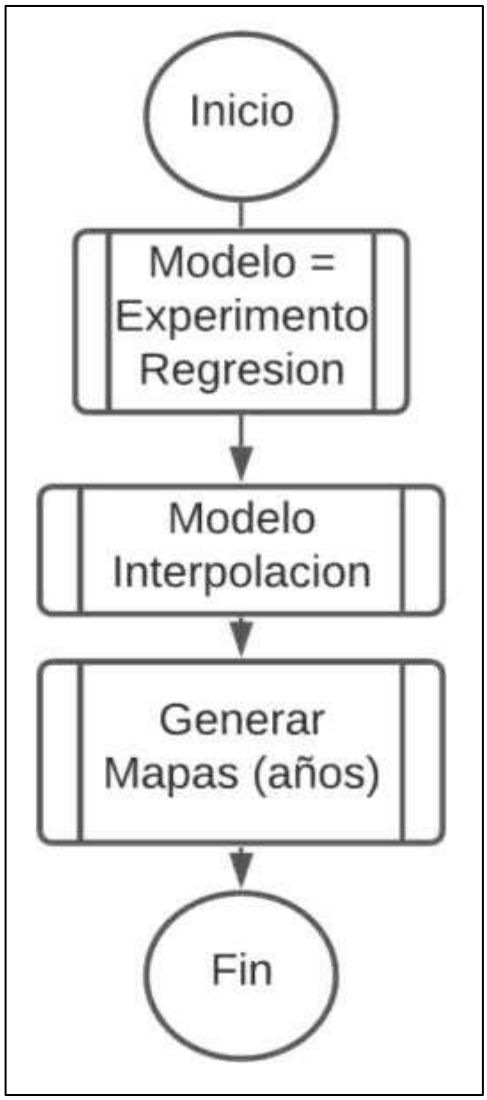
Figura 76. Diagrama Transformación de Datos



Fuente. esta investigación

5.2.4. Diagrama de actividades Minería de Datos. Se escogen todos los datos agrupados para probar con varios modelos de regresión (KNN, SVR, Arboles de decisión, Redes Neuronales, Random Forest) al que mejor le va es a SVR (Support Vector Regression). En la visualización se notó que los modelos por si solos no fueron capaces de generar una interpolación correcta de los datos, por lo tanto se usó una interpolación donde recibe este modelo de regresión (SVR) ya que obtuvo la mejor métrica, al usar kriging en conjunto con SVR como máquina de soporte vectorial mejora las predicciones y la interpolación de los datos, de esta manera se logró encontrar el variograma para interpolar con Kriging el cual fue el Gaussiano, con esto se conciben los demás modelos de interpolación para los diferentes años y como resultado de este modelo es generar mapas.

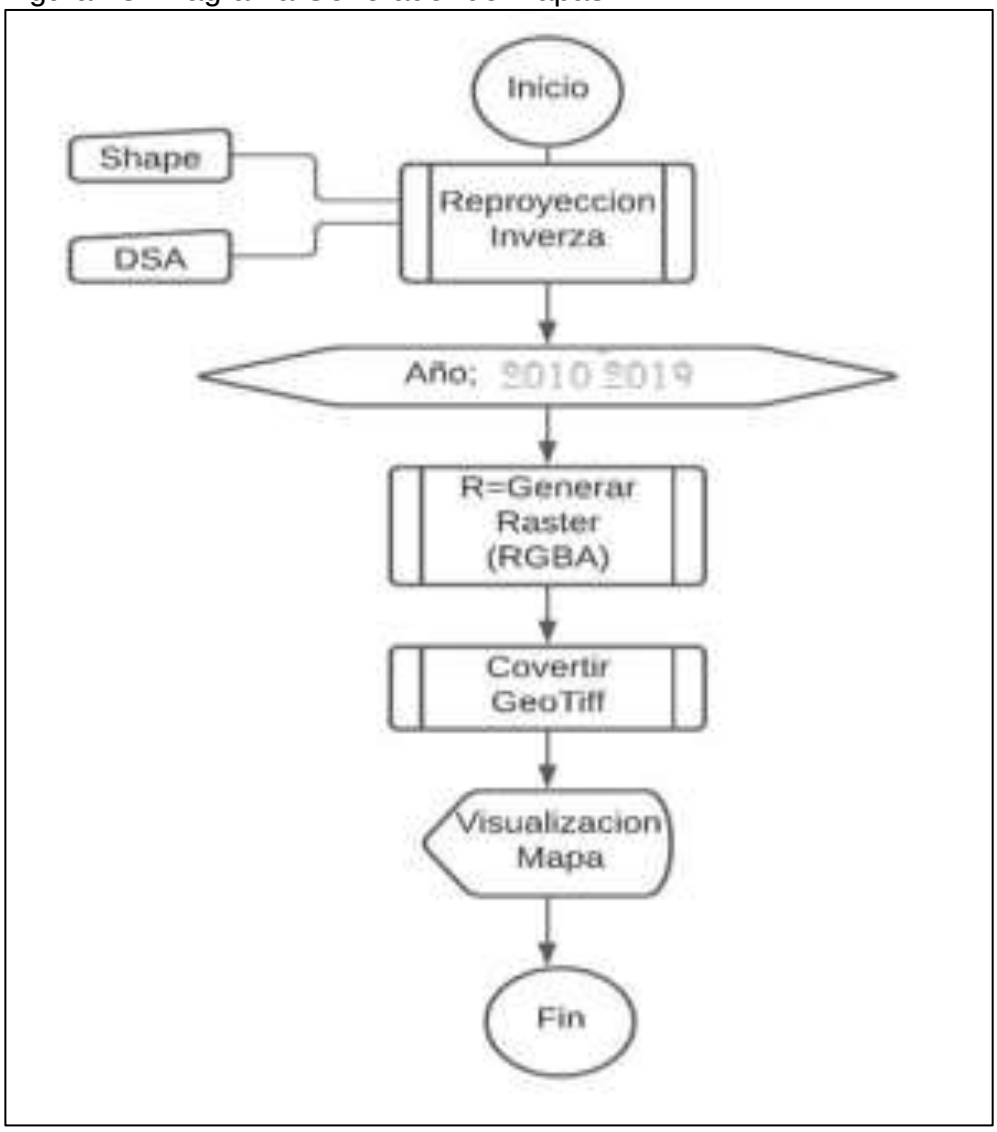
Figura 77. Diagrama Minería de Datos



Fuente. esta investigación

5.2.5. Diagrama de actividades Generador de Mapas. Luego de hacer la transformación de los datos se vuelve a realizar una retransformación inversa para pasar de Mercator 3857 a Oeste 3115 para poder trabajar con ángulos, luego de reproyectar se generan las imágenes desde el 2011 hasta el 2019, para ello se genera un raster para los canales rojo, verde y un alfa para colocar la transparencia (RGBA), y se convierte la imagen raster en un Geotiff.

Figura 78. Diagrama Generador de Mapas



Fuente. esta investigación

6. CONCLUSIONES

- Como resultado de las fuentes de extracción se obtuvieron los datos correspondientes a cada una de las variables definida,contando con el departamento del delito Colombiano como la fuente principal para el desarrollo de la investigación, sin embargo los datos de la fuente del departamento del delito 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 de tener una sola y proceder al análisis de los datos que van hacer útil para dicho estudio, por ello fue de gran importancia aplicar limpieza y preprocesamiento de los datos y así poder 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 que fueron de gran utilidad para el desarrollo de este estudio.

- 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 en sus agrupaciones y se ajustan mucho más a la realidad que el algoritmo K-means.

- 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 autocorrelación, 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.

- La aplicación de los algoritmos KNN(K Vecinos Cercanos), SVR (Maquinas de soporte Vectorial para regresión), MLPR (Perceptron Multi Capa para regresión), GP (Procesos Gaussianos), TD (Ar-bol de decisión), RF (Bosques Aleatorios), para el desarrollo de este estudio no arrojarón resultados favorables, ya que al realizar las comparaciones con las métricas de calidad desde un rango de 0 a 1%, donde 0 representan un menor ajuste con los datos de entrenamiento y cerca a 1, un mayor ajuste a los datos de entrenamiento, cada uno de ellos

obtuvieron un promedio de 0,3% según el ajuste de la métrica, es por ello que no son tan favorables para la aplicación del grupo de datos de entrenamiento aplicados en este estudio.

- El desarrollo e implementación de “Sorlock Holmes” predicción de actividad delictiva en el municipio de Tumaco mediante técnicas de machine learning, busca contribuir en el análisis de delitos para la realización de planes estratégicos cuyo objetivo sea mitigar actos de actividad delictiva que está azotando a la comunidad Tumaqueña con el fin de contribuir a la seguridad del municipio, además se obtienen los siguientes beneficios teniendo en cuenta los objetivos planteados en este proyecto:

    a) Servirá como punto de partidas a proyectos relacionados con actividad delictiva en el municipio de Tumaco ya que se obtendrá bases de datos sin procesar, preprocesados y modelos de entrenamiento que sirvan a diferentes proyectos acorde al estudio a realizar.

    b) “Sorlock Holmes” permitirá la Obtención de zonas o barrios donde se presencien actividades delictivas, para lo cual poder identificar los lugares más peligrosos y no muy actos para transitar.

    c) Facilitar el proceso u labor que realizan las fuerzas armadas u entes de control en cuanto a la seguridad de los Tumaqueños, ya que por medio de esta se pueden realizar planes de contingencia con el fin de mitigar actos delictivos en la comunidad, brindando así un entorno seguro.

    d) Permitirá contar por primera vez con “Sorlock Holmes” aplicación de técnicas de machine learning siendo esta un aporte al avance tecnológico como una herramienta de gran aporte a la investigación y a problemas de actividad delictivas para la comunidad Tumaqueña.

7. RECOMENDACIONES

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; 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.

BIBLIOGRAFÍA

AGENDA “Los algoritmos de aprendizaje supervisado“, (en línea)(13 agosto 2019) obtenido en

ÁLVARO, "<<¿Qué es el sobreajuste u overfitting y por qué debemos evitarlos?>>," MachineLearningParaTodos.com, 25 Mayo 2020. [Online]. Available: https://machinelearningparatodos.com/que-es-el-sobreajuste-u-overfitting-y-por-que-debemos-evitarlo/.

AMANDA J. Rogers; AMIR, Hashemi y MARIANTHI, G. Ierapetritou. Modelado de procesos de partículas para el continuo Fabricación de formas farmacéuticas de dosificación de base sólida. 2013 Available:

ARCGIS ArcGis Resources Abril 2014. [En línea]. Available: https://resources.arcgis.com/es/help/getting-started/articles/026n00000014000000.htm.

ARCMAP, «Cómo funciona Kriging» esri, 2016. [En línea]. Available: https://desktop.arcgis.com/es/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm.

ARCMAP, Los métodos kriging» esri, 2016. [En línea]. Available: https://desktop.arcgis.com/es/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm.

ARSYS, "Qué es Machine Learning y por qué es tan importante," . [en linea]. (25 febrero 2019). Obtenido en: https://www.arsys.es/blog/soluciones/infraestructura/machine-learning/

AUGUST, Extrema seguridad en 12 barrios de Tumaco. [en linea].(05 , 2021). Obtenido de https://diariodelsur.com.co/noticias/local/extreman-seguridad-en-12-barrios-de-tumaco-400202

AURELIO, M. « Los formatos GIS ráster más populares,» MmappingGIS, 17 Diciembre 2015. [En línea]. Available: https://mappinggis.com/2015/12/los-formatos-gis-raster-mas-populares/.

DEFENSORIA.COM "Informe especial: economías ilegales, actores armados y nuevos escenarios de riesgo en el posacuerdo,"; (Defensoría del Pueblo, 2017b, p. 183). Septiembre 2018. [Online]. Available: https://www.defensoria.gov.co/public/pdf/economiasilegales.pdf

DUEÑAS, María Ximena “Minería de datos espaciales en búsqueda de la verdadera información” [en linea]. (Junio, 2009) Obtenido de

ESCOBAR, S. L. Algoritmos de Agrupamiento. Recuperado el 06 de agosto de 2021, de https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/628/1/LopezES.pdf

ESTÉVES, R. «<<Librerías Python GIS para manipular y analizar datos espaciales>>,» Análisis Gis,Desarrollo Gis, 16 Septiembre 2019. [En línea]. Available: http://www.geomapik.com/desarrollo-programacion-gis/librerias-python-gis/.

FERNANDEZ CARABALLO, Franco. Metodología para el análisis de la violencia en el departamento de Bolívar mediante técnicas de machine learning”,[En línea]. 2018. Available: https://repositorio.utb.edu.co/handle/20.500.12585/1118 file:///C:/Users/PcPersonal/Downloads/Modeling_of_Particulate_Processes_for_the_Continuo.pdf. (2021-11-27).

FRANCISCO, L. «Coeficiente de determinación (R cuadrado),» economipedia, 02 Octubre 2017. [En línea]. Available: https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html.

FRANZPC, que es el Interpolación [en linea].(14 de Marzo de 2021). Obtenido de

GABRI [29], "¿Qué es el error cuadrático medio RMSE?," acolita, 22 mayo 2018. [Online]. Available: https://acolita.com/que-es-el-error-cuadratico-medio-rmse/.

GALINDO, L. y CATALÁN, H. Las actividades delictivas en el Distrito Federal,» julio-septiembre 2007. [En línea]. Available: https://www.redalyc.org/pdf/321/32112593003.pdf.

H. H. G. –. N. MINERA, "PRUEBA DEL KIT PYKRIGE EN PYTHO," Nube minera, 2019. [Online]. Available: https://nubeminera.cl/kit-pykrige-en-python/.

HARO, Silvia “Minería de datos para descubrir tendencias en la clasificación de los trabajos de titulación” [En línea] , (2018). Obtenido en

HERAS, J. M. "<<Máquinas de Vectores de Soporte (SVM)>>," IArtificial.net, 28 Mayo 2019. [Online]. Available: https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-21262009000100007

INFANTE, D. "Dinámicas del conflicto armado en Tumaco y su impacto humanitario," FIP -FUNDACIÓN IDEAS PARA LA PAZ, 9 noviembre 2017. [Online].

Available: https://www.ideaspaz.org/publications/posts/926. [Accessed marzo 19 2020].

JAULIS RUA y G. VILCARROMERO, Sistema de predicción de hechos delictivos,» 2015.[Enlínea].Available: https://repositorio.usmp.edu.pe/bitstream/handle/20.500.12727/2022/jaulis_vilcarromero.pdf?sequence=1&isAllowed=y. [Último acceso: 2021 08 06].

JBAGNATO“¿Qué es el algoritmo k-Nearest Neighbor? ”[en línea], (2018), obtenido en

LICONA AGUILA, Aguilar y CONTRERAS, María Camila, (Septiembre – 2018)” Caracterización de los delitos en

LIZARDO, B. A. ¿Cómo funciona kNN?,» Platzi, Enero 2021. [En línea]. Available: https://platzi.com/tutoriales/1841-probabilistica/9110-k-nearest-neighbor/.

LORENA, Georreferenciacion. [en linea].(12 de abril de 2018). Obtenido de https://www.certicalia.com/blog/georreferenciacion-que-es-y-para-que-se-utiliza

MERAYO, Patricia “Algoritmo-árbol de decisión”[en línea](20, mayo), obtenido en https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

MOYA, Ricardo: que es el Clustering. [en linea].(25 de Marzo de 2016). Obtenido de https://www.jarroba.com/que-es-el-clustering/

NALDA, Víctor “https://www.futurespace.es/machine-learning-los-origenes-y-la-evolucion/. [en linea].(29/ 09/ 2020). Obtenido dehttps://www.futurespace.es/machine-learning-los-origenes-y-la-evolucion/

O. D. D. PASTO, "La criminalidad en Nariño aumentó en un 60 por ciento según el Observatorio d(unodc,2013)el Delito," Diario del sur, 2017. [Online]. Available: https://diariodelsur.com.co/noticias/local/la-criminalidad-en-narino-aumento-en-un-60-por-ciento-segun-315758. [Accessed 13 marzo 2020].

O. D. DELITO, ""Observatorio del Delito de la Policía Nacional"," Policia Nacional de Colombia, 2015. [Online]. Available: (https://www.policia.gov.co/observatoriodeldelito). [Accessed 03 04 2021].

PASCUZZO, Alda Extrapolación. [en linea].(11 de mayo de 2013). Obtenido de http://aldanalisis.blogspot.com/2013/05/extrapolacion.html

PLAN ESTRATÉGICO DE TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES - PETI 2019 - 2022 . [en linea]. (2019). Obtenido en:

QUESTIONPRO, "<<¿Qué es el coeficiente de correlación de Pearson?>>," QuestionPro, 28 mayo 2019. [Online]. Available: https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/

RECLU IT “Historia y evolución del Machine Learning” (03/ 08/ 2020) [en linea]. Obtenido de

REDACCIÓN APD “¿Cuáles son los tipos de algoritmos del machine learning? “, (en línea)(04 abril 2019) obtenido en https://www.apd.es/algoritmos-del-machine-learning/

RENATA, A. «<<Respuesta -Proyección>>,» BRAINLY, 10 Febrero 2021. [En línea]. Available: https://brainly.lat/app/profile/19619938/answers.

RENDÓN, Eréndira†; ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez. (05 julio 2015). El algoritmo de agrupamiento K-Modas. Obtenido de: https://www.ecorfan.org/bolivia/researchjournals/Tecnologia_e_innovacion/vol2num5/Tecnologia_e_Innovacion_Vol2_Num5_2.pdf

REVISTA CRIMINAL. vol.58 no.2 Bogotá May/Aug. 2016

ROBERTO, C. << Utilidad de las redes neuronales artificiales>> - Thinkbig," Telefonica Tech , 12 febrero 2020. [Online]. Available: https://empresas.blogthinkbig.com/redes-neuronales-artificiales/.

RUBY, Walker “Machine Learning” [en linea]. (07/ 05/ 2019) Obtenido de

S. B. DATA, «Conceptos básicos de aprendizaje automático con el algoritmo de vecinos más cercanos a K,» 24 Diciembre 2019. [En línea]. Available: https://sitiobigdata.com/2019/12/24/algoritmo-de-aprendizaje-automatico-de-aprendizaje-automatico/.

S. DATA, «Aprendizaje automático y las Métricas de regresión,» sitiobigdata.com, 27 Agosto 2018. [En línea]. Available: https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/

SUAREZ PEÑA, Javier Andrés “Modelo de aprendizaje automático para la predicción de la calidad” [En línea] (diciembre, 2019). Obtenido en https://repository.udistrital.edu.co/bitstream/handle/11349/23560/SuarezPe%C3%B1aJavierAndres2019.pdf?sequence=1&isAllowed=y. [Último acceso: 13 marzo 2020].

TAI DINH, Tsutomu Fujinami, Van-Nam Huynh (Noviembre 2019), “Estimación del número óptimo de clústeres en agrupamiento de datos categóricos por coeficiente de silueta” Obtenido de: https://www.researchgate.net/publication/336980455_Estimating_the_Optimal_Number_of_Clusters_in_Categorical_Data_Clustering_by_Silhouette_Coefficient

TIMÁRAN PEREIRA, Silvio Ricardo “GrupLAC - Plataforma SCienTI - Colombia” [En línea] , (2002). Obtenido en

TOLEDO, A. "Métodos de selección de atributos para clasificación supervisada basados en teoría de información," researchgate, 2016. [Online]. Available: https://www.researchgate.net/publication/331155838_Metodos_de_seleccion_de_atributos_para_clasificacion_supervisada_basados_en_teoria_de_informacion.

TRIPATÍA, B. Sudhir Kumar Sahu, Kamal Kumar Barik . (Enero, 2015). Una máquina de vectores de soporte de clasificación binaria y segmentación de imágenes de datos de teledetección de Chilika Lagloon. Obtenido de: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

UNIOVIEDO. (2020). Unioviedo. obtenido de unioviedo: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

VALENGA, F. "Aplicacion de mineria de datos para la exploracion y deteccion de patrones delictivos en argentina," sedici, [en linea]. Available: http://sedici.unlp.edu Timarán Pereira,2002.ar/bitstream/handle/10915/21783/Documento_completo.pdf?sequence=1. [accedido 4 abril 2021].

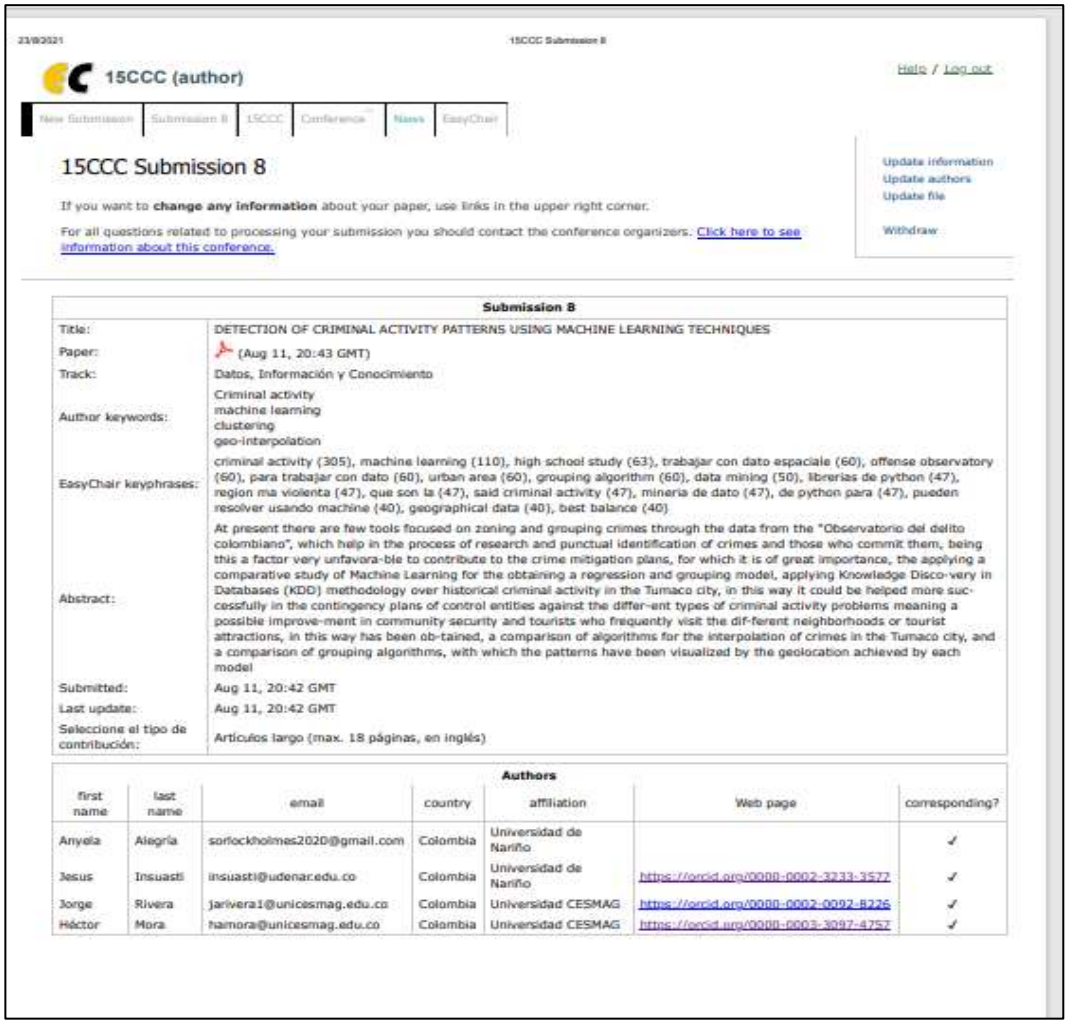
VILMOS, «Geopandas» ICHI.PRO, 2020. [En línea]. Available: https://ichi.pro/es/uso-de-geopandas-para-visualizacion-espacial-37278328347703.

YUMPU, "Tecnica y mineria de datos para la prevencion de lavado de activos y la financiacion de terroismo," Yumpu.com, 2017. [en linea]. Obtenito en: https://www.yumpu.com/en/document/read/53701711/tecnicas-de-mineria-de-datos-para-la-prevencion-de. [Accessed 7 marzo 2020].

ZAMORANO RUIZ, Juan. " Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea.,". [Online]. 2018 Available: https://eprints.ucm.es/id/eprint/48800/. [Accessed noviembre 24 2021].

ANEXOS

Anexo A. Inscripción de participación en el congreso Colombiano de Computación, modalidad inscripta en articulos largos en inglés.



23/8/2021 15CCC Submission 8

15CCC (author)

Help / Log out

New Submission | Submission 8 | 15CCC | Conference | News | EasyChair

15CCC Submission 8

If you want to **change any information** about your paper, use links in the upper right corner.

For all questions related to processing your submission you should contact the conference organizers. Click here to see information about this conference.

Update information
Update authors
Update file

Withdraw

Submission 8

Title:	DETECTION OF CRIMINAL ACTIVITY PATTERNS USING MACHINE LEARNING TECHNIQUES
Paper:	(Aug 11, 20:43 GMT)
Track:	Datos, Información y Conocimiento
Author keywords:	Criminal activity machine learning clustering geo-interpolation
EasyChair keyphrases:	criminal activity (305), machine learning (110), high school study (63), trabajar con dato espaciale (60), offense observatory (60), para trabajar con dato (60), urban area (60), grouping algorithm (60), data mining (50), librerias de python (47), region ma violenta (47), que son la (47), said criminal activity (47), mineria de dato (47), de python para (47), pueden resolver usando machine (40), geographical data (40), best balance (40)
Abstract:	At present there are few tools focused on zoning and grouping crimes through the data from the "Observatorio del delito colombiano", which help in the process of research and punctual identification of crimes and those who commit them, being this a factor very unfavora-ble to contribute to the crime mitigation plans, for which it is of great importance, the applying a comparative study of Machine Learning for the obtaining a regression and grouping model, applying Knowledge Disco-very in Databases (KDD) methodology over historical criminal activity in the Tumaco city, in this way it could be helped more suc-cessfully in the contingency plans of control entities against the differ-ent types of criminal activity problems meaning a possible improve-ment in community security and tourists who frequently visit the dif-ferent neighborhoods or tourist attractions, in this way has been ob-tained, a comparison of algorithms for the interpolation of crimes in the Tumaco city, and a comparison of grouping algorithms, with which the patterns have been visualized by the geolocation achieved by each model
Submitted:	Aug 11, 20:42 GMT
Last update:	Aug 11, 20:42 GMT
Seleccione el tipo de contribución:	Artículos largo (max. 18 páginas, en inglés)

Authors

first name	last name	email	country	affiliation	Web page	corresponding?
Anyela	Alegría	sorlockholmes2020@gmail.com	Colombia	Universidad de Nariño		✓
Jesus	Insuasti	insuasti@udenar.edu.co	Colombia	Universidad de Nariño	https://orcid.org/0000-0002-3233-3577	✓
Jorge	Rivera	jarivera1@unicesmag.edu.co	Colombia	Universidad CESMAG	https://orcid.org/0000-0002-0092-8226	✓
Héctor	Mora	hamora@unicesmag.edu.co	Colombia	Universidad CESMAG	https://orcid.org/0000-0003-3097-4757	✓

Fuente. esta investigación

Anexo B. Carta de solicitud para la obtencion del mapa.shp del casco urbano de Tumaco a la alcaldia municipal.

SAN ANDRES DE TUMACO

28 DE OCTUBRE DEL 2020

Dirección Territorial Nariño

Estimados

Soy la estudiante Anyela Soranyi Alegria Campaz, identificada con cc: 1087200115 de Tumaco, del programa de Ingeniería de Sistemas de la Universidad de Nariño.Por medio de la presente carta, quiero solicitarles muy amablemente el mapa del municipio de Tumaco en formato (extensión). "**shp**", con código de verificación de catastro 52835 ,el cual es de suma importancia para el desarrollo de un trabajo de grado llamado **"SORLOCK HOLMES" PREDICCIÓN DE ACTIVIDAD DELICTIVA EN EL MUNICIPIO DE TUMACO MEDIANTE TÉCNICAS DE MACHINE LEARNING**".

Le agradezco de antemano su rápida respuesta y me despido atentamente.

Anyela Soranyi Alegria Campaz

Correo: anyelaSoranyi06@hotmail.com

Cel: 3168631550

Fuente. esta investigación

Anexo C. Certificado sobre Introduction to machine learning

Fuente. esta investigación

Anexo D. Artículo en ingles de “DETECTION OF CRIMINAL ACTIVITY PATTERNS USING MACHINE LEARNING TECHNIQUES”

DETECTION OF CRIMINAL ACTIVITY PATTERNS USING MACHINE LEARNING TECHNIQUES

First Author[1][0000-1111-2222-3333] and Second Author[2][1111-2222-3333-4444]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com

Abstract. At present there are few tools focused on zoning and grouping crimes through the data from the “*Observatorio del delito colombiano*”, which help in the process of research and punctual identification of crimes and those who commit them, being this a factor very unfavorable to contribute to the crime mitigation plans, for which it is of great importance, the applying a comparative study of Machine Learning for the obtaining a regression and grouping model, applying Knowledge Discovery in Databases (KDD) methodology over historical criminal activity in the Tumaco city, in this way it could be helped more successfully in the contingency plans of control entities against the different types of criminal activity problems meaning a possible improvement in community security and tourists who frequently visit the different neighborhoods or Tourist attractions, in this way has been obtained, a comparison of algorithms for the interpolation of crimes in the Tumaco city, and a comparison of grouping algorithms, with which the patterns have been visualized by the geolocation achieved by each model.

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

Fuente. esta investigación

Anexo E. Link de los repositorios

Link de aplicación de algoritmos supervisados.

- https://colab.research.google.com/drive/1aSqNWao1ADTjO8lOOor1fTco9nz9ZTYE?authuser=1

Link de aplicación de Clustering

- https://colab.research.google.com/drive/1sypKlnXZHDxULpqrM-uHEZ4dUc4S21wB?authuser=1