Automatización de rutas de patrullaje basados en modelos dinámicos y predictivos apoyados en el análisis de información delictual

Manjarres Moyano, Edison Orlando

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación previo a la obtención del título de Magíster en: Gestión de Sistemas de Información e Inteligencia de Negocios

Mgsbi. Jácome Paneluisa, Hernán

6 de noviembre del 2020

URKUND

Document Information

Analyzed document	tesisV1.8_rev4docx.docx (D85170985)
Submitted	11/13/2020 5:48:00 PM
Submitted by	
Submitter email	acbaldeon@espe.edu.ec
Similarity	2%
Analysis address	acbaldeon.espe@analysis.urkund.com

Sources included in the report

SA	**Universidad de las Fuerzas Armadas ESPE / TesisUrkundGuayasaminAdrian.docx** Document TesisUrkundGuayasaminAdrian.docx (D38681740) Submitted by: emcampania@espe.edu.ec Receiver: emcampania.espe@analysis.urkund.com	1
SA	**Universidad de las Fuerzas Armadas ESPE / Tesis_MSGIIN_MaricellaSinchiguano_2017.docx** Document Tesis_MSGIIN_MaricellaSinchiguano_2017.docx (D30682074) Submitted by: dmmarcillo@espe.edu.ec Receiver: dmmarcillo.espe@analysis.urkund.com	1
SA	**Universidad de las Fuerzas Armadas ESPE / Tesis_MSGBI_DianaPoma_2017_Final.docx** Document Tesis_MSGBI_DianaPoma_2017_Final.docx (D30726859) Submitted by: dmmarcillo@espe.edu.ec Receiver: dmmarcillo.espe@analysis.urkund.com	2
SA	**Proyecto de Grado MBI - Maestria v4.docx** Document Proyecto de Grado MBI - Maestria v4.docx (D44104363)	2
SA	**tesis FLORES PEREZ Y CESPEDES GARCIA).docx** Document tesis FLORES PEREZ Y CESPEDES GARCIA).docx (D33634658)	1
SA	**Universidad de las Fuerzas Armadas ESPE / Tesis.docx** Document Tesis.docx (D25211919) Submitted by: afhidalgo@espe.edu.ec Receiver: afhidalgo.espe@analysis.urkund.com	1
W	URL: https://docplayer.es/89789365-Tesis-para-optar-a-la-titulacion-de-postgrado-corres... Fetched: 11/5/2019 1:09:20 AM	3

Firmado electrónicamente por:
HERNAN JACOME

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Automatización de rutas de patrullaje basados en modelos dinámicos y predictivos apoyados en el análisis de información delictual**" fue realizado por el señor **Manjarres Moyano, Edison Orlando** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 6 de noviembre del 2020

Firmado electrónicamente por:
HERNAN JACOME

.....

Mgsbi. Jácome Paneluisa, Hernán

C.C.: 1707493459

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo **Manjarres Moyano, Edison Orlando**, con cédula de ciudadanía n° 0603091737, declaro que el contenido, ideas y criterios del trabajo de titulación*:* **Automatización de rutas de patrullaje basados en modelos dinámicos y predictivos apoyados en el análisis de información delictual,** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 6 de noviembre del 2020

Firmado electrónicamente por:
EDISON ORLANDO
MANJARRES MOYANO

Ing. Manjarres Moyano, Edison Orlando

C.C.: 0603091737

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo **Manjarres Moyano, Edison Orlando,** con cédula/ de ciudadanía n° 0603091737, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Automatización de rutas de patrullaje basados en modelos dinámicos y predictivos apoyados en el análisis de información delictual,** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi/nuestra responsabilidad.

Sangolquí, 6 de noviembre del 2020

Firmado electrónicamente por:
EDISON ORLANDO MANJARRES MOYANO

Ing. Manjarres Moyano, Edison Orlando

C.C.: 0603091737

Dedicatoria

Yo, dedico de todo corazón a mi madre, quien desde pequeño fue mi fuente de inspiración de pujanza y trabajo arduo para seguir con esa determinación en mi vida, personal y profesional y así conseguir las diferentes metas que me he propuesto.

A mi esposa quien de una y otra manera estuvo siempre con su apoyo, comprensión y tiempo, ayudándome de forma moral y personal para seguir a delante y poder conseguir uno más de nuestros objetivos.

Agradecimiento

Yo, que sería de la vida de una persona, si en el transcurso del paso por este mundo no existieran las personas correctas e incorrectas, las correctas para ayudarnos y apoyarnos a seguir nuestros sueños, guiarnos hasta alcanzarlos, las incorrectas porque pese su falta de apoyo e intentar desmoralizarnos, consiguieron sacar más fuerza para seguir adelante, por ello, MUCHAS GRACIAS Familia, Amigos, Dirigentes, Tutores.

Meta Cumplida.

Índice general

Índice de tablas

Índice de figuras

Resumen

La inseguridad durante los últimos años es uno de los principales problemas en el Ecuador, esto ha sido puesto en evidencia por prestigiosas instituciones y empresas dedicadas a llevar a cabo análisis y estudios de este tema, como en el caso de la encuesta de victimización y percepción de la inseguridad del 2011, realizada por la única institución gubernamental encargada de este tipo de estudios como el del Instituto Nacional de Estadísticas y Censos, publica en su sitio web (INEC, 2011) donde se da a conocer las incidencias de delitos cometidos por cada 100.000 habitantes, dentro del territorio Nacional. Es necesario contribuir de forma tecnológica con la entidad encargada de realizar este trabajo, la cual es la Policía Nacional del Ecuador, esta institución proporciona la información que ayuda a identificar delitos y concentración de los mismos, por esta causa el propósito de este proyecto se enfoca principalmente en analizar la información, pre procesar los datos y finalmente desarrollar modelos predictivos delictuales de las rutas de patrullaje. Se propone el desarrollo de modelos dinámicos predictivos mediante la realización de Estudio de Caso (Simons & Filella, n.d.) como metodología de investigación, y apoyados de KDD, SEMMA y CRISP-DM; siendo esta ultima la escogida por tener un 43% más de precisión que sus contrapartes; a su vez, el algoritmo utilizado fue SimpleKMeans con la metodología de codo. Como resultado final del proyecto se determinó que los clusters analizados después del proceso de filtro de la data set original se incorporó de modo satisfactorio a las rutas de patrullaje, promoviendo una disminución significativa del 31.20% de los casos delictuales.

- Palabras clave:
    - **MINERÍA DE DATOS**
    - **CASOS DELICTIVOS**
    - **ZONAS DE PATRULLAJE**
    - **SIMPLEKMEANS**
    - **BASE DE DATOS.**

Abstract

Insecurity in recent years is one of the main problems in Ecuador, this has been evidenced by prestigious institutions and companies dedicated to carrying out analyzes and studies on this issue, as in the case of the victimization and perception survey of insecurity in 2011, carried out by the only government institution in charge of this type of study, such as the National Institute of Statistics and Censuses, publishes on its website (INEC, 2011) where the incidences of crimes committed per 100,000 inhabitants, within the National territory. Taking into account this high percentage in the indices of criminal events in Ecuador, it is necessary to contribute in a technological way with the entity in charge of carrying out this work, which is the National Police of Ecuador, this provides the information that helps to identify crimes and concentration of the same, for this reason the purpose of this project focuses mainly on analyzing the information, before processing the data and finally developing predictive crime models of patrol routes. The development of dynamic predictive models is proposed by carrying out a Case Study (Simons & Filella, n.d.) as a research methodology, and supported by KDD, SEMMA and CRISP-DM; The latter being the one chosen for having 43% more precision than its counterparts. In turn, the algorithm used was SimpleKMeans with the elbow methodology. As a final result of the project, it was determined that the clusters analyzed after the filtering process of the original data set were successfully incorporated into the patrol routes within zone 9 of the metropolitan district of Quito, promoting a significant decrease of 17%. of criminal cases in the study area.

- **KEYWORDS:**
    - **DATA MINING**
    - **CRIMINAL CASES**
    - **PATROL ZONES**
    - **SIMPLEKMEANS**
    - **DATABASE.**

Capitulo i

Introducción

Los problemas más representativos con una afección en el país es la inseguridad ciudadana. Durante los últimos años se ha podido notar el aumento en el número de delitos y la impunidad con que son realizados, convirtiéndose en un problema a nivel nacional.

El distrito ecuatoriano no es ajeno a esta realidad, tal es así que, en base a información recogida en la comisaria nacional, las denuncias a nivel país son mayores en comparación al año pasado.

Esto aporta una primera indicación del proceso de prevención de delitos en el país se ve debilitado debido a diferentes factores, pero sobre todo a la falta de rondas, patrullajes o redadas realizadas de manera rápida por parte de la policía y la falta de recursos existentes en las comisarías.

La tecnología en grandes volúmenes da datos orientados a la predicción de sucesos de minería y datos, ya que ha sido muy utilizada últimamente en problemas de índole social, tales como predicción de desastres natrales, predicción del tiempo, predicción del crimen, etc.

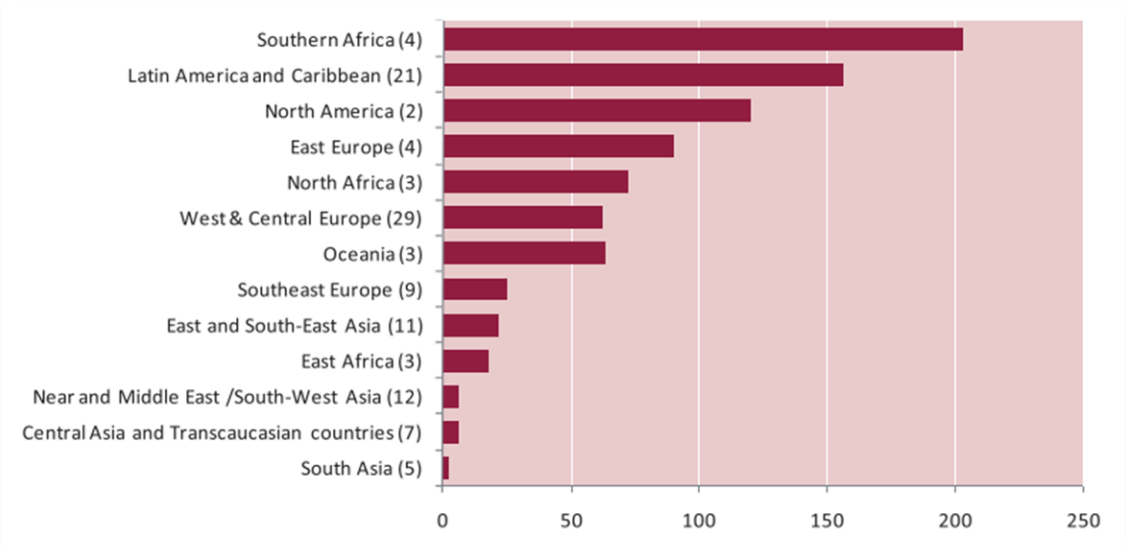
En esta investigación, se llevará a cabo una indagación en base a los registros históricos de las denuncias ocurridas dentro del país para poder crear un modelo de predicción de hechos delictivos que se pueda mostrar en un sistema de información precisa y manejable para los policías y así ellos puedan mejorar sus tomas de decisiones en cuanto a la prevención del delito.

Antecedentes

A nivel mundial se puede evidenciar los inconvenientes que ocasiona los diferentes tipos de delitos. Según la ONU ,Organización Mundial de las Naciones Unidas, por cada 100.000 habitantes se comete una media de 2006 actos delictivos a personas (Harrendorf, Heiskanen, & Malby, n.d.), como muestra la figura 1.

Figura 1.

Niveles de delincuencia a nivel mundial



Nota: El grafico muestra los datos deliectivos de las diferentes partes del mundo. (Harrendorf, Heiskanen, & Malby, 2016)

Se muestra un análisis de los incidentes delictivos, con un enfoque mundial, donde se evidencia de forma concisa los inconvenientes que ocasiona este fenómeno en la sociedad, así como las ponderaciones de las regiones en las cuales se concentra los índices delictivos más altos, así también es visible la situación de Latino América y el Caribe, en el cual ocupa el segundo lugar en incidencias delictivas, por lo que es indiscutible que Ecuador forme parte de

estas estadísticas, esto se pone en evidencia en el estudio de victimización y percepción de la inseguridad del 2011, realizado por parte del Instituto Nacional de Estadísticas y Censos INEC, en el que se realiza un análisis por provincia y la porción de hurtos a personas ocurridas, como se puede observar en la Tabla 1.

Tabla 1.

Robos a personas por provincia en ecuador

Provincia	Veces que fue víctima de robo a personas	Tasa de incidencia de delito x 100000 habitantes	Población total estudiada	Coeficientes de variación (%)
Tungurahua	42091	23496	179143	4,4%
Pichincha	365668	21287	1717835	4,2%
Azuay	63093	20619	305990	4,6%
Guayas	446542	19436	2297506	3,8%
Santo Domingo de los Tsáchilas	38112	19120	199334	5,9%
Nacional Urbano	1273444	17219	7395572	1,8%
Imbabura	32760	16314	200806	4,3%
El Oro	56718	16300	347966	3,8%
Chimborazo	21163	14495	145999	6,0%
Cotopaxi	10503	11847	88656	6,1%
Los Ríos	38537	11621	331631	5,4%
Loja	20567	11548	178094	5,6%
Manabí	70129	11509	609362	3,3%
Esmeraldas	25123	11296	222408	4,6%
Carchi	7051	10634	66302	6,1%
Cañar	6714	9315	72085	5,2%
Orellana	3190	8974	35552	11,2%
Sucumbíos	3863	8449	45726	12,5%
Santa Elena	14512	7553	192117	6,4%
Pastaza	2048	7361	27825	11,0%
Napo	1401	6235	22464	12,4%
Bolívar	1447	3716	38929	9,7%
Morona Santiago	1023	3351	30528	14,0%

Galápagos	346	2604	13276	17,8%
Zamora Chinchipe	394	1932	20390	17,3%
No responde	424	0	0	0,0%

Nota: Los datos presentados demuestran que, a mayor cantidad de habitantes en una provincia, el número de robos aumenta equitativamente. (INEC, 2019).

Planteamiento del problema

Los sucesos delictivos que se evidencia a nivel nacional en el Ecuador, han ocasionado un incremento en los índices delincuenciales, reflejando una baja prevención de estos acontecimientos por parte de la Policía Nacional.

Contexto del Problema

El aumento del índice delictivo a nivel nacional durante los últimos años ha sido un argumento que en la actualidad preocupa a los diferentes organismos de control. Este problema se genera por falta de oportunidades de trabajo, migración de personas de países vecinos o por las políticas gubernamentales. Esto ha ayudado al incremento en la taza de hechos delictivos, como, por ejemplo: robo a personas, robos domiciliarios, robo a unidades económicas, robo a vehículos, robo en ejes viales, robo a motos, robo a bienes accesorios y autopartes, entre los más relevantes (INEC, 2019), como se describe en la tabla 2.

Tabla 2.

Índice de inseguridad de las 24 provincias del Ecuador

PROVINCIA	VECES QUE FUE VÍCTIMA DE ROBO A PERSONAS	TASA DE INCIDENCIA DE DELITO X 100000 HABITANTES	POBLACION TOTAL ESTUDIADA	COEFICIENTES DE VARIACIÓN (%)

Tungurahua	42091	23496	179143	4,4%
Pichincha	365668	21287	1717835	4,2%
Azuay	63093	20619	305990	4,6%
Guayas	446542	19436	2297506	3,8%
Santo Domingo de los Tsáchilas	38112	19120	199334	5,9%
Nacional Urbano	1273444	17219	7395572	1,8%
Imbabura	32760	16314	200806	4,3%
El Oro	56718	16300	347966	3,8%
Chimborazo	21163	14495	145999	6,0%
Cotopaxi	10503	11847	88656	6,1%
Los Ríos	38537	11621	331631	5,4%
Loja	20567	11548	178094	5,6%
Manabí	70129	11509	609362	3,3%
Esmeraldas	25123	11296	222408	4,6%
Carchi	7051	10634	66302	6,1%
Cañar	6714	9315	72085	5,2%
Orellana	3190	8974	35552	11,2%
Sucumbíos	3863	8449	45726	12,5%
Santa Elena	14512	7553	192117	6,4%
Pastaza	2048	7361	27825	11,0%
Napo	1401	6235	22464	12,4%
Bolívar	1447	3716	38929	9,7%
Morona Santiago	1023	3351	30528	14,0%
Galápagos	346	2604	13276	17,8%
Zamora Chinchipe	394	1932	20390	17,3%
No responde	424	0	0	0,0%

Nota: En esta tabla se describe las veces que fue víctima de robo las personas en una determinada provincia por cada 100.000 habitantes Fuente: (INEC, 2019).

Problemática del problema

La Policía Nacional del Ecuador, encargada de precautelar la seguridad interna del País, por intermedio de uno de sus ejes principales, el eje preventivo, se encarga de realizar patrullajes frecuentes durante todas las horas del día en todos los puntos de la nación, con el

objetivo de combatir y prevenir el auge delictual, estos patrullajes por lo general se enfocan tan solo en los lugares, que por experiencia de permanencia en un respectivo lugar son conocidos, resultando ineficiente realizar prevención si no se dispone de la información adecuada para realizar esta actividad, como por ejemplo días de la semana que más se comenten robos, horas del día, forma de operar, concentraciones de hechos delictivos en un lugar etc.

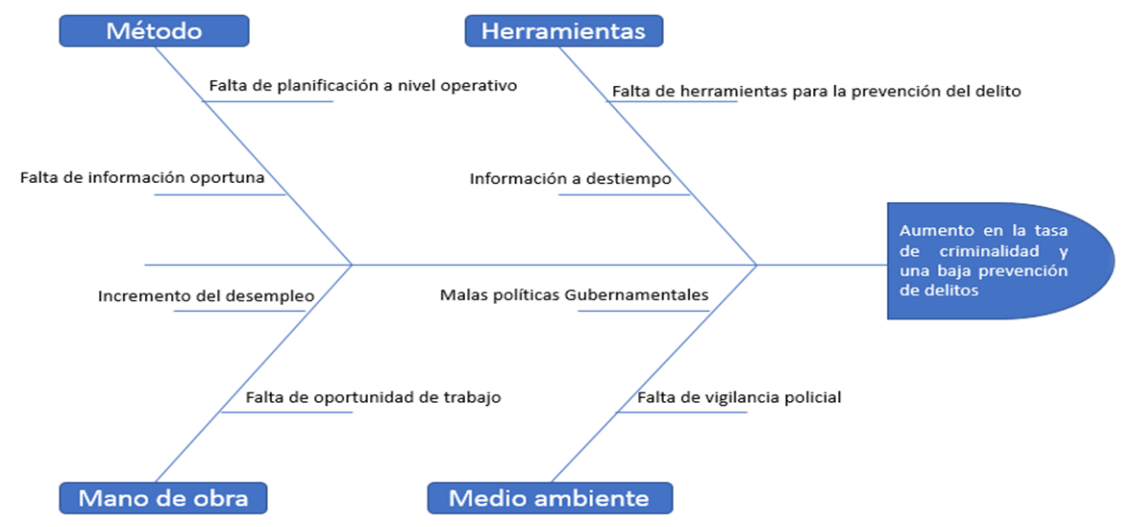
Por esta razón no es factible identificar de forma eficiente las áreas de mayor presencia delictual, resultando ineficaz definir rutas adecuadas para un correcto patrullaje preventivo, por parte de la Policía Nacional del Ecuador, lo que fomenta el incremento en la tasa de criminalidad y una baja prevención de delitos, provocando un incremento de los índices delictuales, y una alta percepción de inseguridad.

Diagrama Causa y Efecto

Hoy en día, la Policía Nacional centra sus esfuerzos en crear precaución hacia el delito, sin contar con una información adecuada y estructurada para llevar a efecto su trabajo, esto ocasiona que dichos esfuerzos sean infructuosos a la hora de prevenir los delitos, debido a que no tienen datos relevantes que proporcione una información adecuada que indique los lugares por donde debe dirigir sus esfuerzos para realizar sus patrullajes preventivos. Para ayudar a identifica él problema se ha optado hacer uso de diagrama de espina de pescado descrita en la figura 2.

Figura 2.

Diagrama de causa y efecto.



Nota: la figura muestra el diagrama de causa y efecto con modelo de pescado. (Elaboración propia, 2020)

Objetivos de la investigación

Objetivo Generales

Desarrollar modelos dinámicos para la predicción de delitos y generación de rutas de patrullaje preventivo de la Policía Nacional del Ecuador, basada en el histórico de la información delictual existente, y así conseguir una correcta prevención del delito.

Objetivo Específicos

- Análisis del estado del arte, mediante una revisión inicial de literatura de los últimos cinco años, con la finalidad de verificar si existe datos técnicos de minería que hayan participado en el campo de seguridad ciudadana.

- Analizar y pre procesar datos históricos existentes de numerosas bases de datos de la Policía Nacional, ECU 911 y otras instituciones.
- Analizar y seleccionar los modelos predictivos que más se ajuste a la automatización del patrullaje preventivo de la Policía Nacional.
- Evaluar los resultados mediante métodos estadísticos cualitativos, con el objetivo de verificar si los modelos utilizados se encuentran dando los resultados esperados.

Justificación, importancia y alcance del problema

Ante el incremento en los índices delictuales a nivel nacional, de lado de los medios de comunicación esto es transmitido a diario, esto ha generado un ambiente de temor en la sociedad ecuatoriana, por lo cual es de gran importancia el conocer cuáles son los tipos de delito que más se cometen, en qué lugar, los días a lo largo de la semana, los horarios, y de ahí poder optar por diferentes tipos de medidas que ayuden a prevenir el cometimiento de más delitos.

El presente proyecto pretende enfocarse en desarrollar modelos dinámicos para la predicción de delitos y generación de rutas de patrullaje preventivo de la Policía Nacional, basada en los datos históricos delictuales existentes, con la finalidad proporcionar información en línea, tratada y procesada que aporte inteligencia a las operaciones diarias de la Policía Nacional.

Además de identificar de forma clara los focos de criminalidad por medio del uso de la información tratada y geo localizada, así como por intermedio de la generación de modelos dinámicos disponer de la reproducción de rutas para el patrullaje preventivo.

Hipótesis

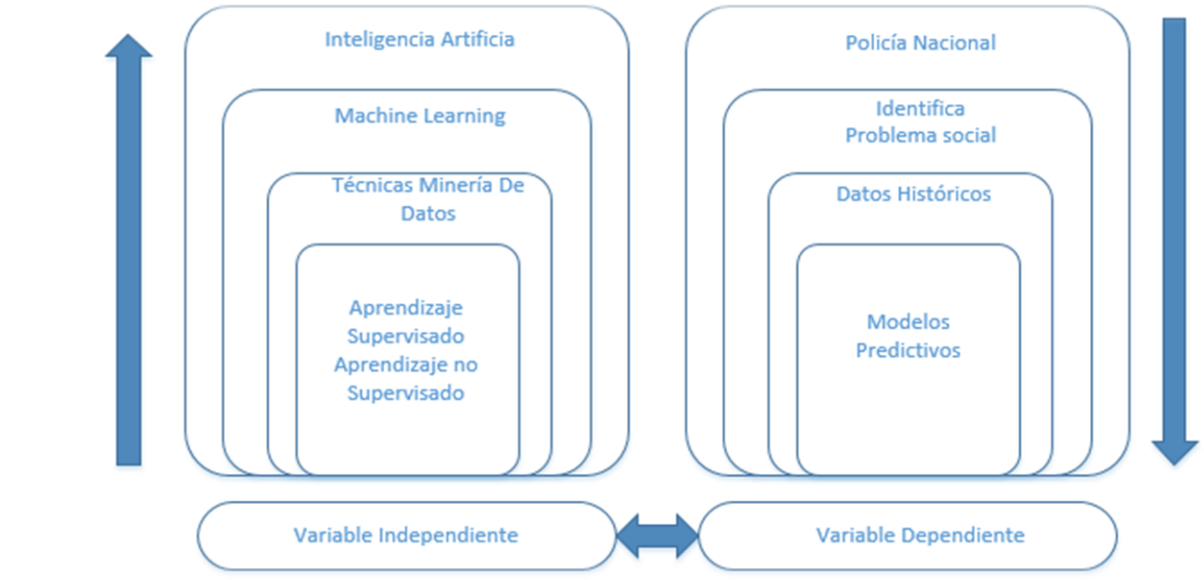
El desarrollo de modelos predictivos para la reproducción de rutas óptimas de patrullaje mejorará la tasa de detección de delitos, mediante de la utilización del método de inteligencia artificial, para eso se comparan las variables dependientes e independientes tal cual se detalla en la figura 3.

Variable dependiente: tiempo de identificación de riesgo, creación de modelos predictivos, generación de rutas de patrullaje.

Variable independiente: prevención de casos delincuenciales.

Figura 3.

Comparación de variable independiente y dependiente.



Nota: la figura muestra una comparación entre la variable independiente y su relación con la variable dependiente. (Elaboración propia, 2020)

Para la demostración de la hipótesis, se la realizará por comprobaciones estadísticas evaluando los resultados de los modelos creados. Esta comprobación se la realizará utilizando técnicas existentes en minería de datos como la matriz de confusión y la tasa de precisión en la detección, tasa de falsos positivos, falsos negativos que verificará la eficiencia de los algoritmos del modelo propuesto.

Justificación, importancia y alcance del problema

Ante el incremento en los índices delictuales a nivel nacional, lo que por parte de los medios de comunicación es transmitido a diario, esto ha generado un ambiente de temor en la sociedad ecuatoriana, por lo que resulta de especial interés conocer cuáles son los tipos de delito que más se cometen, en qué lugar, los días de la semana, los horarios, y de ahí poder optar por diferentes tipos de medidas que ayuden a prevenir el cometimiento de más delitos.

El presente proyecto pretende enfocarse en desarrollar modelos dinámicos para la predicción de delitos y generación de rutas de patrullaje preventivo de la Policía Nacional, basada en los datos históricos delictuales existentes, con la finalidad proporcionar información en línea, tratada y procesada que aporte inteligencia a las operaciones diarias de la Policía Nacional.

Además de identificar de forma clara los focos de criminalidad por medio del uso de la información tratada y geo localizada, así como por intermedio de la generación de modelos dinámicos disponer de la generación de rutas para el patrullaje preventivo.

Capitulo ii

Estado del arte

Marco Referencial

Es esencial revisar investigaciones y documentos relacionados al área temática, para así verificar el funcionamiento de sistemas ya consolidados, que hayan resuelto un determinado problema de la sociedad ya que esto va a permitir tener una vista panorámica sobre la temática a desarrollar y así alcanzar los objetivos que fueron señalados previamente.

Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina

Esta investigación realizada en Argentina, estuvo referida a la implementación de datos en minería en base a información histórica de hechos criminales que ocurrían en el mencionado país y a la vez comprobar su efectividad del mismo.

El autor Perversi (2007) en su investigación se identifican patrones de homicidio cometidos en Argentina durante el año 2005 en sustento esta información fue suministrada por la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación, institución encargada de realizar las oficiales estadísticas de criminalidad en Argentina. Se puso énfasis en los registros criminales como valor fundamental para la prevención del delito.

Otro objetivo destacado está relacionado a los planes de prevención y diseño de políticas a realizar, ya que en Argentina esta variedad de distinción se había ejecutado

transcendentalmente con herramientas estadísticas descriptivas básicas, considerando básicamente variables y relaciones primarias.

En general, la tesis abarcó el descubrimiento de patrones delictivos y la utilización de mesura preventiva para atacar los mismos (Perversi, 2007).

Caracterización y predicción espacio temporal de patrones delictivos mediante modelos lógicos combinatorios

Martínez (2009) afirma que su investigación:

Se basa en técnicas de aprendizaje inductivo y busca hacer un análisis y predicción delictiva en México D.F. El análisis delictivo se realiza mediante la fabricación de definiciones inductivas para cada familia delictiva que se plantea en la investigación; y, en base a estas definiciones se logra determinar la conducta de la actividad delictiva dentro de un espacio y tiempo seleccionado.

Luego del estudio, se llegó a determinar la cantidad de hechos delictivos esperados para cada familia de hechos delictivos. Los resultados son orientados como ayuda a las autoridades de seguridad pública, para que estos mejoren en temas de prevención y asignación de recursos humanos y financieros (Martínez, M., 2005, pág. 15).

Aplicando minería de datos al marketing educativo

Cadena (2011) muestra en su investigación realizada en Colombia que los datos de minería pueden abarcar distintos campos de estudio, en este caso, enfocándose al sector educativo; donde en base al análisis de los registros históricos de los estudiantes se pudo

establecer la tasa de abandono de los estudiantes y de los factores que influyeron en la misma, para ello abarcaron una muestra a un grupo de estudiantes de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda.

Al finalizar el estudio se resaltó aspectos importantes, algunos de ellos son; la mayoría de alumnos desertores eran de sexo femenino y que además pertenecían a los 3 primeros ciclos de la carrera, quienes alegaban que el motivo era el bajo rendimiento académico y problemas económicos. Otro factor importante la cual es que en su mayoría vivían cerca de la urbe de Bogotá, y sus edades oscilaban desde los 19 a los 22 años.

Con respecto a los varones, estos generalmente desertaban entre los primeros dos ciclos con un promedio de edad de 19 años, el motivo principal era el bajo rendimiento académico y la falta de atención en el Idioma Inglés.

Otros factores analizados fue la proveniencia de los alumnos, registros de notas de colegio, nivel cultural, deportes practicados con esta información la Universidad supo tomar mejores decisiones para posicionar sus carreras en el sector social adecuado (Cadena Pinzón, 2011).

Uso de redes neuronales para la Asistencia de voz en base a aplicaciones de Minería de Datos

Según Engels y Theusinger (2015) quienes describen un proyecto realizado en las inmediaciones del país de Cataluña España donde en base a datos históricos se ha podido crear un asistente de voz inteligente, este actualmente es en vías de culminación ya que uno de los principales problemas que se enfrentó fue en la realización de testeos, y la denominada ley de

Turing, el objetivo de este test es verificar y comprobar si un ente artificial es capaz de simular el comportamiento humano en este caso su raciocinio.

Es asistente de voz en base a distintos patrones de reconocimiento de Voz e Inteligencia Artificial tiene la capacidad de asistir a las diversas dudas de los usuarios, este proyecto se ha realizado para que el asistente sepa contestar a las distintas temáticas con las que interactúe, actualmente sigue en etapa de prueba, y en esta investigación, se enfocan en hacer referencia a las tecnologías que utilizaron para desarrollar este proyecto (Engels y Theusinger , 2015).

Desarrollando aplicaciones de minería de datos acopladas a sistemas gestores de bases de datos relacionales

Según Agrawall y Shim (2015) quienes mencionan que uno de los grandes avances en la tecnología fue la creación de repositorios integrados en nuestro caso popularmente denominados bases de Datos, en su principal oficio estas almacenan información y han servido de muchas formas para dirigirse a tomar decisiones y un gran ahorro en trabajos manuales muchas veces.

Realizar una aplicación de datos de minería no es algo sumamente difícil si es que se siguen los principales principios y distintos patrones de desarrollo, estos no se encuentran altamente relacionado a un determinado lenguaje de programación ya que trabajan de forma independiente y más se enfoca en el algoritmos de análisis de datos, este trabajo menciona las buenas prácticas de desarrollo como son CRISP-DM y de diversos datos de minería utilizados para su desarrollo de aplicaciones (Agrawal y Shim, 2015).

Mecanismos de aprendizaje y minería de datos

El autor Mitchell (2013) detalla en su investigación que:

Los distintos tipos de algoritmos de aprendizaje que existen para el estudio de datos y que se han surgido en las últimas décadas. Si bien muchas de las técnicas estadísticas ya empleadas en la actualidad ya existían, gracias al avance de la tecnología en especial del hardware que son capaces de realizar cálculos estadísticos sumamente complejos que al realizarlo humanos simplemente sería demasiado laborioso, y como esta gran cantidad de información se ha ido aumentado año a año teniendo esta nueva tecnología de Datos de Minería, en el cual se obtuvo una gran cantidad de información que tal vez no es tan relevante pero que dentro de todo ese universo existen pequeñas informaciones de gran valor significativo y que permitan tomar decisiones anticipadas (Mitchell Gary, 2013, pág. 14).

Preguntas de Investigación

Para la consecución de todos los objetivos específicos planteados, es indispensable responder las siguientes preguntas:

- OE1-RQ1.1: ¿Existen estudios previos sobre análisis históricos de información delictual?
- OE1-RQ1.2: ¿Existen técnicas utilizadas en otros estudios, que se pueda aplicar en el análisis de información aplicando inteligencia artificial con algoritmos, para predecir posibles delitos?
- OE2-RQ1.1: ¿Existen metodologías que ayude la automatización del pre procesamiento de la información histórica de delitos?

- OE2-RQ1.2: ¿Existen herramientas que ayuden a pre procesar datos históricos de diferentes fuentes?
- OE3-RQ1.1: ¿Qué herramientas se puede utilizar para verificar el modelo predictivo para un patrullaje inteligente?
- OE3-RQ1.2: ¿Qué metodología es la más aplicada para el desarrollo proyectos de inteligencia artificial?
- OE4-RQ1.1: ¿Cuáles son las mejores técnicas de validación a un modelo propuesto para la decisión de un problema?
- OE4-RQ1.2: ¿Cuáles son los márgenes de error aceptables para la validación de un modelo propuesto?

Metodológico de la Investigación y Técnica

Estudio del caso

En este apartado se describen tres de las metodologías más utilizadas en el mercado, esto con la finalidad de elegir el objetivo que más se adecue a las necesidades del presente proyecto. Resaltando criterios de selección y pesos a cada uno para poder llegar exitosamente a su respectivo análisis.

Crisp-DM

La metodología CRISP-DM distribución y durabilidad de un proyecto de aprovechamiento de datos en seis niveles de secuencia no rígida, interactúan de manera comunicativa a lo largo del progreso del proyecto. Las flechas muestran la frecuencia entre las fases y la dependencia más importante.

En la Figura 4 Chapman (2000) define una figura que aportara en el descubrimiento y la descripción los procesos del modelo CRISP – DM. El círculo exterior representa la calidad de tipo de proyectos. Enumerando las siguientes fases (Chapman, 2000):

Fase 1. Comprensión del comercio: Se caracteriza por la comprensión de las metas y requisitos del proyecto desde un punto de vista empresarial, para luego transformarlos en metas técnicas con un plan de proyecto. El conocimiento obtenido del negocio se convierte en un problema de Data Mining (DM). Las labores aplicables son: establecer las metas de un preciso comercio, medirlas circunstancias, señalar los objetivos del DM y crear un plan de proyecto.

Fase 2. Comprensión de los datos: se inicia con la recopilación de información para luego fijar una primera conexión que se estima con el problema, Lugo adaptarse a los datos, reconocer las cualidades de los datos e implantar relaciones visibles que permitan precisar la información oculta de la hipótesis. Las labores creadas son: recopilar datos iniciales, explicar los datos, buscar los datos y verificar la calidad.

Fase 3. Organización de datos: se concentra en los ejercicios de organización de los datos para la adaptarlos a los métodos de DM, así como: la observación, la indagación de lazos entre variables para la investigación. Las labores a ejecutar son: clasificación de datos, eliminación de datos innecesarios, organizar los datos, integración de datos y formatear los datos.

Fase 4. Moldeado de datos: Selección y aplicación de parámetros calibrados a valores óptimos de diversas técnicas de modelado, un ejemplo, el estudio de regresión, tejido neuronal,

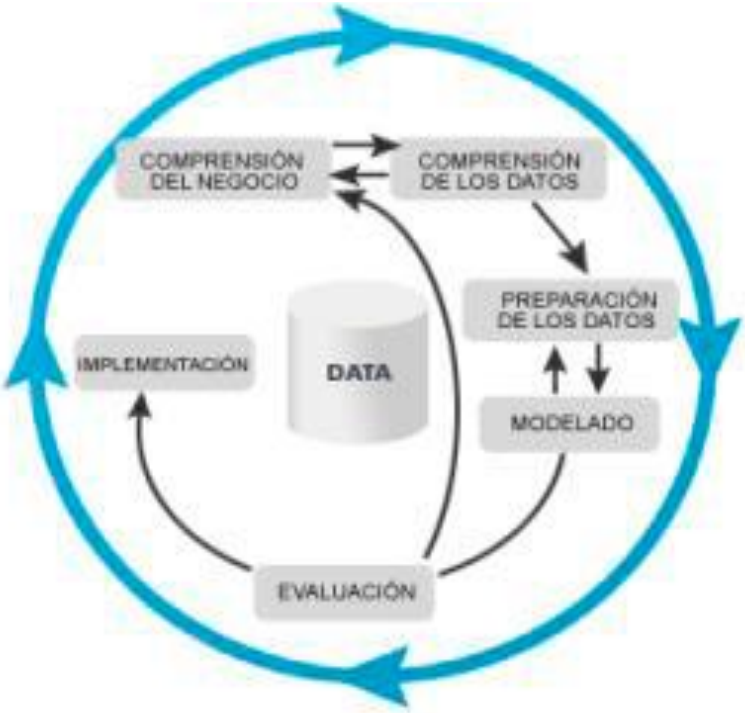
casos basados en razonamiento (RBC), etc. Algunos métodos requieren específicamente características de los datos y la tensión que se da. Las labores a ejecutar son: la recopilación metodológica del moldeado, generar un plan de pruebas, creación y evaluación del modelo.

Fase 5. Prueba del modelo: Una evaluación precisa del modelo y la inspección de los procesos realizados para crear un modelo y asegurar un alcance final de negocio. Es importante determinar algún objetivo no considerado. Las labores a ejecutarse son: la evaluación de resultados, verificar el proceso y precisar para poner en marcha.

Fase 6. Despliegue o implementación: La información adquirida debe ser informado de manera precisa y representado de manera comprensible para que el usuario pueda utilizarla. En cambio, la cantidad que necesitamos ira dependiendo, este proceso de la fase puede ser variada. Las labores a realizan son: crear un plan de implantación, realizar monitoreo y mantenimiento al plan, comunicar sobre el proyecto y la revisión del mismo.

Figura 4.

Ciclo de vida del modelo CRISP – DM



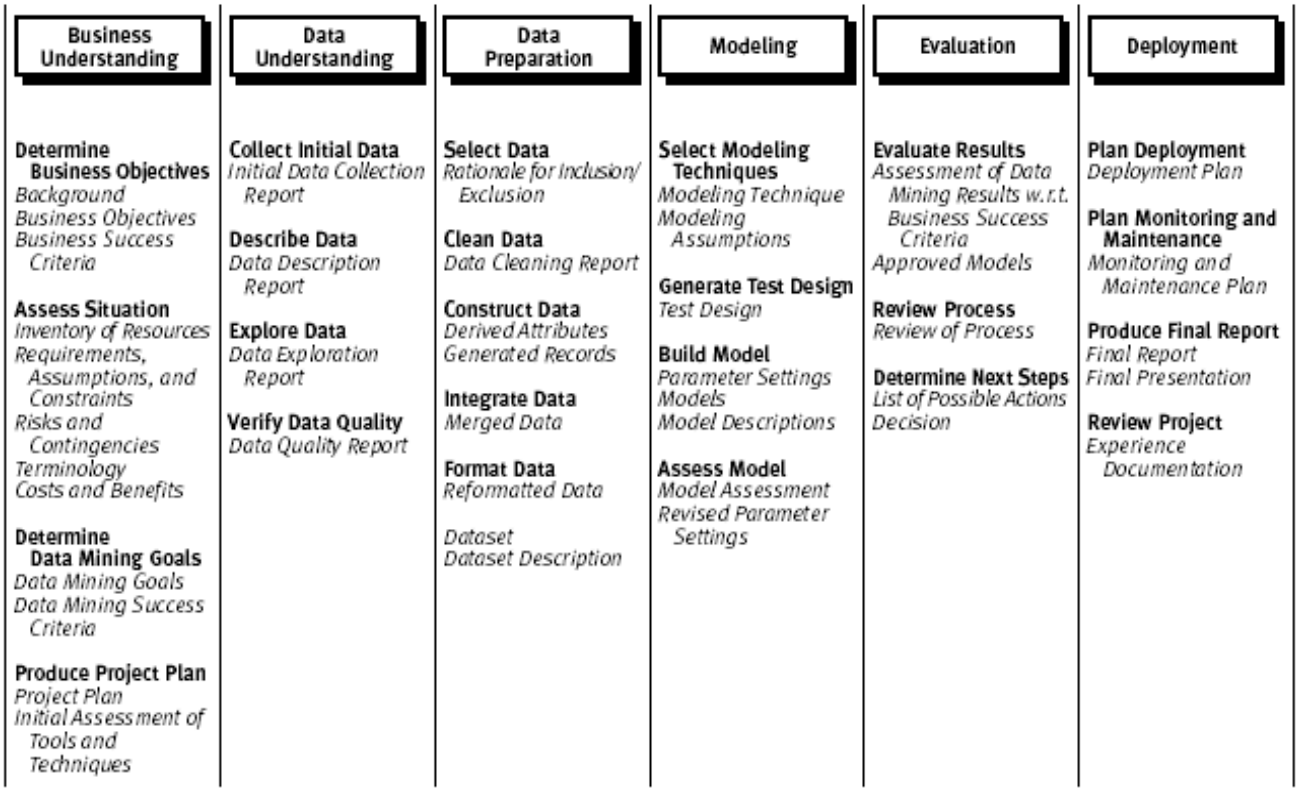
Nota: esta figura muestra el ciclo que cumple un modelo de minería de datos. (Chapman, 2000).

Estas fases que se identifican en el grafico anterior, a su vez se subdividen en actividades bien definidas (en algunos casos predecesoras unas de las otras, mientras que, en otros, son independientes), que hacen posible una planificación sobre la minería de dados eficiente. La figura 5 muestra las fases con sus respectivas actividades y posibles reportes o entregables finales según cada proyecto.

Algunos entregables podrían variar según el negocio y la solución de minería de datos que se esté desarrollando, pero las fases y actividades se deben respetar en cualquier caso si se sigue esta metodología.

Figura 5.

Las fases de la metodología CRISP-DM.



Nota: esta figura describe cada una de las fases con las que cuenta la metodología CRIPS-DMCRIPS-DM. (Chapman, 2000).

Semma

Para SAS – España (2006):

SEMMA es una sigla que significa Sample, Explore, Modify, Model and Assess que en español significa Muestra, Exploración, Modificación Modelo y Valoración. Es un registro de pasos secuencialmente desarrollados por SAS Institute Inc., es uno de los principales creadores

de estadísticas y la inteligencia en software de negocios. Orienta a poner en funcionamiento las aplicaciones de minería de datos. A pesar que SEMMA es a menudo considerado como una metodología general de minería de datos, es más bien una organización lógica del sistema de herramienta funcional de uno de sus productos, SAS Enterprise Miner, para la aplicación de las labores principales de los datos de minería.

SEMMA es un método corto en comparación a el CRISP-DM porque se acumula en mayoría al desarrollo del proceso de Minería de datos y no se dirige a las metas empresariales

La metodología presentada tiene 5 fases, cada una interpretando las siglas SEMMA (SAS - España, 2006):

Sample: Nacimiento de una representativa muestra

Como primer punto tenemos un método, que se aplica extrayendo un conjunto de datos que sean representativos de una población, las metas ayudaran a suministrar el transcurso de minado sobre los datos, eliminando poco a poco los tiempos que se necesita para concretar la información para el negocio.

Explore: investigación de los datos en la muestra.

Luego en este periodo, se hace un trayecto mediante datos obtenidos en la muestra para hallar, determinar y erradicar datos erróneos, favoreciendo a los procesos descubiertos en las fases siguientes. El proceso, la búsqueda se puede cambiar y aplicar a través de medios visuales, muchas veces no es muy eficiente este sistema, es por eso, la visualización puede

manejar diferentes recursos estadísticos como estudio de elementos, indagación de correspondencias, etc.

Modify: Transformación de los datos.

Esta transformación de los datos se puede crear eligiendo las variables en las que se tienen que enfocar a partir del modelo elegido, en varias ocasiones se tendrá que aplicar diversas modificaciones en los datos que se están investigando. Esto se debe al entorno en el que se labora dinámicamente los datos de minería.

Model: Modelación de los datos

El software realiza una búsqueda completa intercalando la información de los principales grupos, dando como resultado un pronóstico de información fiable.

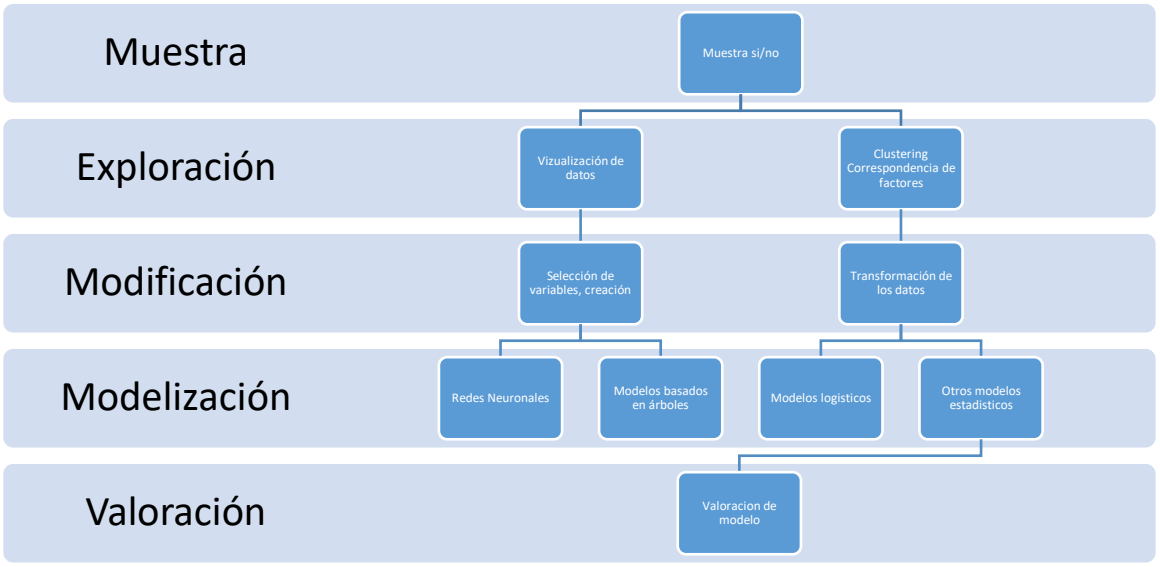
Los procesos y métodos de minería de datos iniciaran un papel sumamente importante para la resolución de problemas identificados en el proyecto de minería de datos.

Assess: Valoración de los datos obtenidos

En esta fase de modelación se presenta la aplicación de los procesos de minería de datos en los resultados obtenidos. Se debe realizar un estudio de resultados para ver si estos tuvieron éxito según la entrada que tuvieron en el análisis del problema. Se identifica los resultados esperados con el modelo creado, el aplicar este modelo a una diferente porción de datos. Si el modelo funciona de una buena manera esta muestra utilizada ayudará al proceso de creación del modelo y se obtendrá una posibilidad de tener un modelo válido (s/p).

Figura 6.

Fases de la metodología SEMMA.



Nota: esta figura describe las fases con las que cuenta la metodología SEMMA. (SAS - España, 2006).

KDD

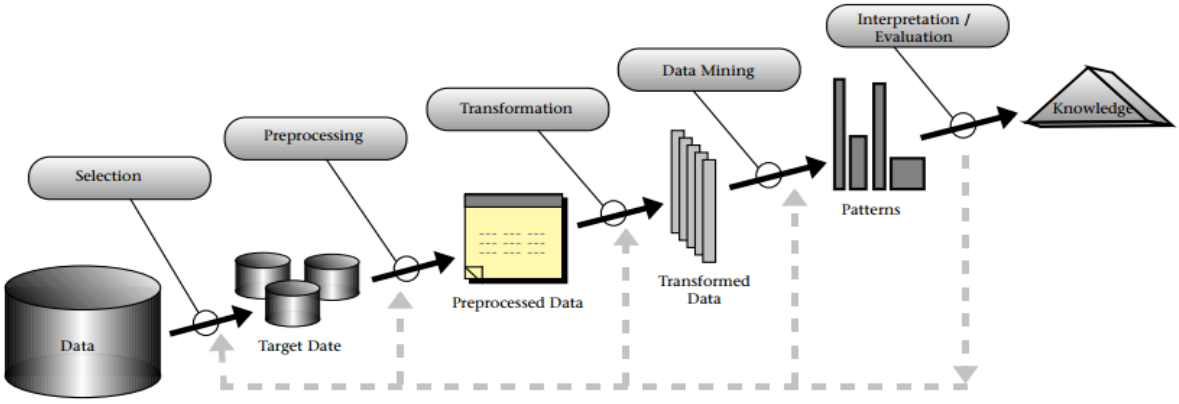
El proceso KDD, según lo presenta Fayyad (1996), es un proceso de aplicación de métodos de Data Mining (DM) para sacar conocimiento, usando la base de datos en conjunto a cualquier pre procesamiento, sub- muestreo y con cualquier variación requerida de la base de datos. Se toman en cuenta cinco etapas (Fayyad , 1996):

- Selección: en esta etapa se crea un conjunto de datos objetivos, que se enfocan en la agrupación de variables o muestras de datos, en la cual se debe realizar el descubrimiento.
- Procesamiento previo: consiste en la eliminación de datos innecesarios y pre - procesamiento de datos de trayectoria de datos consistentes.

- Transformación: es en la variación de datos usando procedimientos de disminución de dimensionalidad.
- Minería de Datos: es la exploración de patrones en una forma particular, según el objetivo de minería de datos (por lo general, la predicción).
- Interpretación / Evaluación - es la interpretación y evaluación de los patrones.

Figura 7.

Método KDD.



Nota: esta figura explica cada una de las fases con las que cuenta la metodología KDD (Fayyad , 1996)

Trabajos Relacionados

Para el sustento de una búsqueda adecuada de los trabajos relacionados, se realizó un análisis del estado del arte, con los parámetros requeridos, en los repositorios académicos IEEExplore, ACM Digital Library.

Para el análisis de este capítulo se usaron varias fases del Systematic Mapping Study, (SMS), que permite realizar un análisis de literatura, sobre el tema que se está investigando. Como el criterio de inclusión y exclusión, además estrategias de búsqueda.

Definición de los criterios exclusión e inclusión: Al realizar una búsqueda en las diferentes bases digitales se debe considerar que, al ejecutar exploraciones de uno u otro tema sin un criterio claramente definido, estas devolverán una larga lista de artículos, que superan de forma amplia nuestros requerimientos y necesidades, por lo que para nuestro análisis se toma en cuenta los siguientes criterios:

Criterios de inclusión:

- ✓ Para el efecto de la búsqueda se consideró todos los artículos con cinco años de antigüedad, a partir del 2013.
- ✓ Artículos que hablen del uso de la minería de datos en el análisis delictual.
- ✓ Artículos que indiquen técnicas de predicción delictual.
- ✓ Artículos desarrollados en idioma inglés.

Criterios de exclusión:

- ✓ Artículos que se encuentren fuera del rango históricos de fechas requeridas
- ✓ Artículos que no se encuentren en idioma inglés.
- ✓ Artículos que, enfocándose a la minería de datos, no hablen de análisis de crimen.

Revisión inicial:

Esta revisión inicial se la enfoca a las preguntas de investigación propuestas, y se las realiza en las librerías digitales indicadas al inicio de esta sección.

Validación cruzada de estudios:

En esta fase se verifica que los artículos encontrados, cumplan con los criterios de exclusión e inclusión planteados, obteniendo una lista de artículos con los que de aquí en adelante se trabajara en las siguientes fases.

Integración del grupo de control:

En esta fase se conforma los estudios que cumplan los criterios de inclusión y exclusión previamente definidos, realizando un análisis de las características principales del estudio, como son el título, la introducción, palabas claves y conclusiones como se presenta en la siguiente tabla:

Tabla 3.

Artículos seleccionados

Grupo de Control	Título	Palabras Claves
EC1	Forecasting Crimes Using Autoregressive Models	Predictive models, Urban areas, Forecasting, Law enforcement, Time series nalysis, Data mining, Data models

EC2	A review: Crime analysis using data mining techniques and algorithms	Data Mining, crime analysis, Naive Bayes Classifiers, Predictive approach
EC3	Crime pattern detection, analysis & prediction	Data mining, Correlation, Tools, Clustering algorithms, Data visualization, Aerospace electronics, Algorithm design and analysis
EC4	Crime prediction using patterns and context	Law enforcement, Prediction algorithms, Urban areas, Bandwidth, Software, Software algorithms, Sociology
EC5	CRIMETRACER: Activity space based crime location prediction	Roads, Computational modeling, Space exploration, Mathematical model, Vectors, Social network services, Predictive models
EC6	Cluster based zoning of crime info	Data analysis, data mining, pattern clustering, police data processing
EC7	A supervised learning approach for criminal identification using similarity measures and K-Medoids clustering	Data mining, Clustering algorithms, Tools, Law enforcement, Testing, Training, Databases
EC8	A fuzzy clustering algorithm to detect criminals without prior information	Data mining, Clustering algorithms, criminals
EC9	Text Mining and Recommender Systems for Text Mining and Recommender Systems for Predictive Policing Policing	Text Mining, Predictive algorithms, system

Nota: Esta tabla describe 9 artículos relacionados con el tema que se está estudiando, el estado del arte.

(Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo y Edison Manjarres, 2020)

Construcción de la cadena de búsqueda:

Para poder definir la cadena de búsqueda se verifica las palabras que más se repitan en los contextos ya definidos en los grupos de control.

Tabla 4.

Cuantificación de las palabras claves

Palabras Claves	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6	EC 7	EC 8	EC 9	EC1 0	Conteo
Crime	x	x	x	x	x	x					6
Criminal							x	x			2
Data Mining	x	x	x			x	x	x	x	x	8
Clustering						x	x	x	x		4

Nota: En esta tabla se cuantificar la aparición de las palabras claves para sustentar y encontrar la cadena de búsqueda más adecuada. (Elaboración propia, 2020)

Una vez verificado las palabras que más se repite se forman la cadena de búsqueda, usando los conectores AND, OR y para segmentar y agrupar “(paréntesis)”

(((((CRIME) OR criminal) AND Data Mining) AND Data Mining) AND Clustering)

Definido y encontrado la cadena de búsqueda que se adapta de mejor forma a nuestro estudio se lo aplica en los repositorios digitales con los siguientes resultados.

Tabla 5.

Resultados de la búsqueda en librerías digitales

IEEExplore	ACM Digital Library
90	36

Nota: en la tabla podemos verificar un total de 90 librerías encontradas y 36 librerías son las que se adaptan al proyecto. (Elaboración propia, 2020)

Obtenidos los resultados se procede a realizar la revisión de los documentos encontrados que más se acercan a la realidad del estudio que se realiza:

Forecasting Crimes Using Autoregressive Models:

Este artículo fue publicado el 13 de octubre del 2016, por los autores Eugenio Cesario, Charlie Catlett, Domenico Talia, en el cual tomando en cuenta la creciente urbanización se proyecta que para el 2030, una proyección de un 60% de la población va a vivir en las ciudades, lo que conllevará a un aumento en la delincuencia, así como el aumento en la recolección de datos que puede ser usada para comprender patrones de tendencia, lo que es útil para anticipar actividades delictivas, y para optimizar la asignación de recursos de seguridad pública. Por lo que proponen diseñar un modelo predictivo para pronosticar la cantidad de delitos que sucederán en los horizontes temporales, realizando un estudio de caso en un área de Chicago, tomando en cuenta variedad de fuentes de datos abiertos y disponibles, obteniendo como resultados de la evaluación experimental muestra que la metodología propuesta predice un 84% en un año y de un 80% en dos años. (Cesario, Catlett, & Talia, 2016)

A review: Crime analysis using data mining techniques and algorithms

Este artículo fue publicado el 21 de diciembre del 2017 por su autor Chhaya Chauhan, en el cual proponen por medio de un enfoque metódico, el análisis de patrones y tendencias del crimen, enfocándose en concepto de minería de datos, para ayudar a analizar grandes volúmenes de información, en este documento, el enfoque principal está en la revisión de algoritmos y técnicas utilizadas para identificar a los delincuentes, y que en sus resultados demuestran que el algoritmo teorema de bayes tuvo el porcentajes más alto 90% al momento de la evolución de los datos, lo que para este estudio fue el mejor (Chauhan & Sehgal, 2017)

Crime pattern detection, analysis & prediction

Este artículo fue publicado el 18 de diciembre del 2017 por sus autores, Sunil Yadav, Conoce Timbadia, Ajit Yadav, Rohit Vishwakarma, Nikhilesh Yadav, en este estudio los autores toman como objetivo principal el análisis históricos de acontecimientos delictivos en la India, de un histórico de14 años, enfocándose en la utilización de técnicas de aprendizaje supervisada, semi-supervisada y no supervisada, aplicadas a la información histórica de los delitos registrados, para el descubrimiento de conocimientos y para ayudar a aumentar la precisión predictiva del delito. (Yadav, Timbadia, Yadav, Vishwakarma, & Yadav, 2017)

Crime prediction using patterns and context

Este artículo fue publicado el 16 de octubre del 2017, constando varios autores, Nelson Baloian, Col. Enrique Bassaletti, Mario Fernández, Oscar Figueroa, Pablo fuentes, Raúl manasevich, Marcos Orchard, Sergio peñafiel, José pino, Mario Vergara, artículo en el cual se analiza datos de varias fuentes para predecir la ocurrencia de varios tipos de delitos, teniendo

como objetivo principal realizar una solución que ayude a la predicción del crimen, enfocada principalmente para las grandes ciudades de Chile. Con un novedoso enfoque que incluye tres módulos de software independientes los mismos que realizan predicciones basadas en diferentes algoritmos. El sistema desarrollado ha sido probado con datos históricos y su rendimiento se ha considerado aceptable para uso de la policía en el campo. Con una media de 45.29% de predicción. (Baloian et al., 2017)

Crimetracer: Activity space based crime location prediction

Este artículo fue publicado el 16 de octubre del 2014, por sus autores, Mohammad A. Tayebi, Martin Ester Uwe Glässer, Patricia L. Brantingham, documento en el cual hacen su estudio enfocado en la prevención del delito, ya sean estos de tipo oportunistas o violentos, por lo que proponen un modelo probabilístico de comportamiento espacial de delincuentes conocidos dentro de su espacio de actividad. Este estudio concluye que la delincuencia no crese de forma uniforme en las áreas urbanas, sino sé que se concentra solo en ciertas áreas. (Tayebi, Ester, Glasser, & Brantingham, 2014)

Cluster based zoning of crime info

Este artículo se publicó el 24 de abril del 2017, por sus autores, Lalitha Saroja Thota, Mohrah Alalyan, AL-Otaibi Awatif Khalid, Fabiha Fathima, Suresh Babu Changalasetty, Mohammad Shiblee, en este documento se realiza el análisis de información histórica criminal, y proponen realizar un análisis de conglomerados, para lo que utiliza el algoritmo de conglomerados k-means, en un conjunto de datos históricos de criminales de la India. Con la

finalidad de crear un mapa personalizado de criminales de la India, con las zonas más conflictivas en sus diferentes estados. (Thota et al., 2017)

A supervised learning approach for criminal identification using similarity measures and K-Medoids clustering

Este artículo fue publicado el 23 de abril del 2018, siendo una de los más actuales en nuestro listado, por sus autores S. T. Bharathi, B. Indrani, M. Amutha Prabakar, documento donde hablan de la importancia de las técnicas de minería de datos para descubrir patrones e información que no se conoce. El esquema propuesto se compara con el algoritmo de agrupamiento K-Means relacionado con el mismo conjunto de entrenamiento y prueba. Obteniendo como resultados un 97% de precisión con datos de entrenamiento y un 92% con los datos de test, usando el algoritmo K-Means. (Bharathi, Indrani, & Prabakar, 2017)

A fuzzy clustering algorithm to detect criminals without prior information

Publicado el 20 de agosto del 2014 por sus autores Changjun Fan, Kaiming Xiao, Baoxin Xiu, Guodong Lv, documento en el cual se enfocan en una forma diferente de agrupación de crímenes y realizan pruebas por medio de experimentos que se describe. Proponen un algoritmo de agrupamiento difuso para detectar criminales ocultos de la red temática, que no utilizó la información de identidad previa de los individuos, obteniendo resultados altos en los sospechosos conocidos (Fan, Xiao, Xiu, & Lv, 2014)

Text Mining and Recommender Systems for Text Mining and Recommender Systems for Predictive Policing Policing

Publicado el 28 de agosto del 2018, por sus autores, Isabelle Percy, Alexander Balinsky, Helen Balinsky, Steve Simske, en este documento se realiza un análisis de texto para un sistema de recomendaciones, considerando los delitos registrados, Evalúan el rendimiento de varias medidas de similitud para los algoritmos de agrupación de textos y documentos. (Percy, Balinsky, Balinsky, & Simske, 2018)

Model of Dynamic Routes for Intelligent Police Patrolling

Publicado el 24 de octubre del 2018, autores Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo, Edison Manjarres. Este artículo propone el desarrollo de un modelo que genere rutas dinámicas de patrullaje, para la policía nacional basados en inteligencia artificial con el algoritmo K-means, con el objetivo de identificar puntos críticos dentro del territorio ecuatoriano. (Guevara et al., 2018)

Conclusión

Los artículos citados en la revisión de la literatura, dan una visión clara de las soluciones exploradas al momento de combatir la delincuencia, por parte de investigadores en diversas regiones, consiguiendo mejorar la efectividad al momento de identificar posibles delitos, en los lugares que fueron aplicados. Es por esto que se propone el desarrollo de modelos predictivos en el Ecuador, con la finalidad de identificar los focos donde se concentra la mayor cantidad de delitos, y a sus ves proporcione los lugares por donde se debería realizar los patrullajes que se encarga la Policía Nacional del Ecuador y así prevenir la posibilidad del cometimiento de delitos

Marco Teórico

Seguridad ciudadana

Las discusiones sobre la seguridad ciudadana en el país comienzan en la década de los noventa, por la crisis social y económica que se experimentó y la decadencia social institucional por la manifestación de un auge delictivo nunca antes visto.

La obligación inquietante de los derechos de protección y los bienes privados afectados por el hampa. La poca dependencia de la seguridad nacional se da por la pérdida política socialista y se encuentra un riesgo menor para el Estado burgués y oligárquico. También nacen a partir de la desaparición de las dictaduras en Latino América, irrumpiendo a la política derrocando al sistema capitalista por los movimientos populares. Luego de las discusiones se interroga a la doctrina de seguridad nacional, pues no se debería perder de vista simplemente se mantiene un perfil bajo, sin ser eliminada de los enfoques internos de las fuerzas del orden. En caso de riesgo capitalista en cualquier variable el Estado recurre a la violencia institucional.

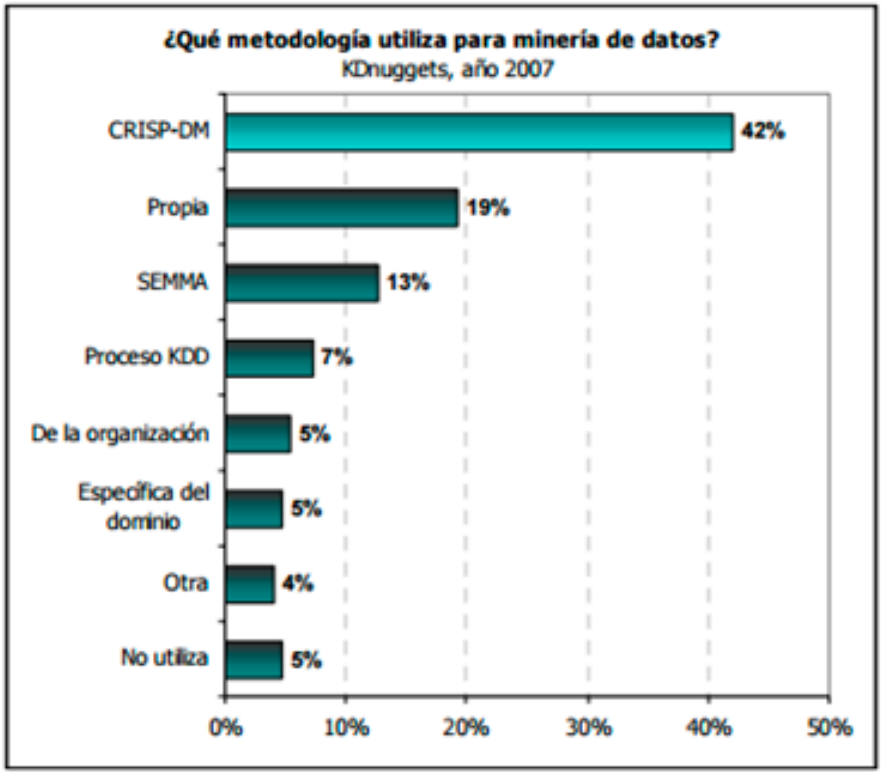
En extracto, la seguridad ciudadana se concentra en términos generales de un indispensable desdoblamiento del Estado en el área de la seguridad, para lograr evolucionar una dimensión que lo reconozca frente a la sociedad, en una circunstancia evidente de crisis social. Se confiere a la defensa de la propiedad privada, elemento formado por el capitalismo y más aún en un momento neoliberal, en donde el egoísmo habla en su máxima expresión.

Minería de datos

Según Mainmon y Rokach (2016) la teoría de minar datos se refiere a un campo dentro de la computación científica centrada en el proceso que descubre patrones a grandes volúmenes de un grupo de datos. Utilizando un método sobre la inteligencia artificial, el aprendizaje automático, los sistemas de datos compartidos y la estadística referencial; con el objetivo del proceso de minería de datos centrado en sacar información de un grupo de datos y evolucionarla en una estructura comprensible para posteriormente usarla. También en esta etapa de análisis en bruto, que incluyen conocimientos sobre gestión, tratamiento u procesamiento de bases de datos y sus diferentes métodos de modelamiento. También se basa en la métrica de datos sobre campos de interés y en consideraciones con interferencias entre conjunto de datos, así como teorías complejas computacionales de post-procesamiento de estructuras encontradas en observación y actualización en línea (Maimón y Rokach, 2016). Los modelos actuales de minería de datos se sustentan realizando un ciclo de vida, en el cual se moldea el comportamiento y el conocimiento de los datos y así lograr obtener un resultado significativo en el mundo real (Valcárcel Ascencios, 2011). Según la figura 8 CRISP-DM (Cross – Industry Standard Process for Data Mining) una metodología más usada.

Figura 8.

Comparación entre las metodologías de minería de datos



Nota: La figura muestra una comparación estadística entre las metodologías mas utilizadas. (By Gregory Piatetsky, 2017)

Comparación metodología CRISP –DM Y SEMMA.

SEMMA fue desarrollado para un paquete específico de software: SAS Enterprise Miner y, pone menos énfasis en las fases de planificación inicial cubiertas en las fases de CRISP-DM (Business Understanding y Data Understanding) y omite totalmente la fase de implementación.

Tabla 6.

Comparación CRISP –DM Y SEMMA

SEMA	CRIP-DM
Muestra y Exploración	Fase de Entendimiento de Datos
Modificar	Fase de preparación de datos
Modelo	Fase de modelado,
Evaluar paralelos	Fase de evaluación
Ambos modelos pretenden ser algo cíclicos en lugar de lineales.	

Nota: La metodología CRIP-DM muestra algunas ventajas ante la metodología SEMA, siendo la mejor opción. (Elaboración propia, 2020)

Comparación metodología KDD y SEMMA.

Se puede decir que las cinco etapas del proceso SEMMA pueden ser vistas como una implementación práctica de las cinco etapas del proceso KDD, ya que está directamente vinculada al software SAS Enterprise Miner.

Tabla 7.

Comparación KDD y SEMMA

KDD	SEMA
Muestra	Selección.
Explorar	Pre-procesamiento
Modificar	Transformación
Modelo	Data Mining
Evaluación	Interpretación / Evaluación

Nota: La metodología SEMMA muestra una ligera mejoría ante la metodología KDD, siendo la mejor entre estos dos métodos analizados. (Elaboración propia, 2020)

Comparación metodología KDD y CRISP-DM.

Se puede observar que la metodología CRISP-DM incorpora los pasos que deben preceder y seguir el proceso KDD:

Tabla 8.

Comparación metodología KDD y CRISP-DM

KDD	CRISP-DM
La fase de Entendimiento de Negocios	Puede identificarse con el desarrollo de una comprensión del dominio de la aplicación, el conocimiento previo relevante y los objetivos del usuario final.
La fase de implementación	Puede identificarse con la consolidación incorporando este conocimiento en el sistema.
La fase de Entendimiento de Datos	Puede ser identificada como la combinación de Selección y Pre procesamiento.
La fase de preparación	Se puede identificar con Transformación.
La fase de modelado	Se puede identificar con Data Mining
La fase de Evaluación	Puede ser identificada con Interpretación/ Evaluación.

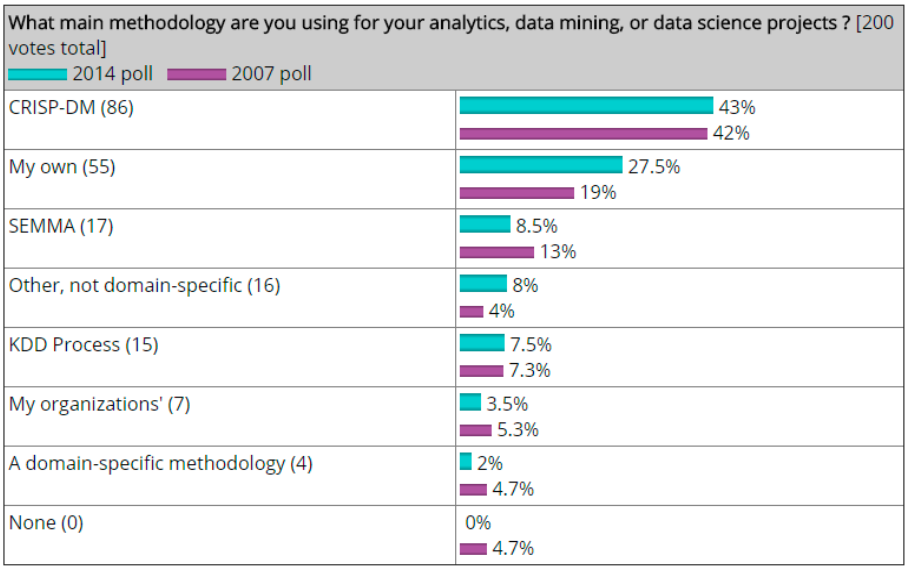
Nota: metodología KDD es notoriamente inferior a las características que ofrece la metodología CRISP-DM, siendo esta ultima la mejor entre estos 2 métodos a analizados. (Elaboración propia, 2020)

Para elegir la metodología a implementar se realizó un análisis sobre las características necesitadas, después se elige la metodología que se quiere implementar para satisfacer de forma completa los criterios de selección. Se necesita un amplio uso de profesionales en la metodología por lo cual se debe realizar un análisis de kdnuggets de preferencia a 200 usuarios que utilizaron una metodología de minería de datos.

Se puede observar como primer lugar de preferencia la metodología CRISP –DM con un 43%, en segundo lugar, SEMMA con un 8.5% y en tercer lugar KDD con 7.5%

Figura 9.

Metodologías más utilizadas.



Nota: esta figura muenstra valores estadisticos de todas las metodologias analizadas. (By Gregory Piatetsky, 2017)

Tabla 9.

Análisis de metodología de minería de datos

CRITERIOS/METODOLOGÍAS	CRISP-DM	SEMMA	KDD
Metodología Estructurada	✓	✓	✓
Metodología Independiente	✓	X	X
Ampliamente Usada	✓	X	X
Mejora la calidad de resultados en proyectos de Data Mining.	✓	✓	✓
Herramientas y técnicas independientes	✓	✓	✓

Finalidad diversa (Ej. Ampliamente estable en la resolución de problemas variados).	✓	✓	✓
Fácil de implementar	✓	✓	✓

Nota: Entre las 3 metodologías más utilizadas por los usuarios, se presenta una comparativa para poder analizar cual tiene mejores resultados y se adapta al presente proyecto. (Elaboración propia, 2020)

Luego del análisis realizado, se selecciona y se implementa la metodología CRISP-DM, ya que es una metodología imparcial y de fácil de implementación, mejorando el valor de los resultados en los proyectos, es un método estructurado y ha sido ampliamente aceptado en análisis de minería de datos.

Según Duncan Ross, un especialista en análisis avanzado de CRIPM-DM, copatrocinador de Teradata, la tecnología de minería de datos se ha desarrollado mucho desde que se lanzó en 1999.

CRISP-DM fue financiado inicialmente por una subvención de la Unión Europea a mediados de los años noventa, a diferencia de los enfoques elaborados por grupos de servicios profesionales de proveedores, fue diseñado para ser neutral en cuanto a herramientas y proveedores.

Weka

Weka se desarrolló bajo el lenguaje JAVA por la universidad Waikato en 1993 y es una herramienta de modelo software con relación a la minera de datos y al aprendizaje automático. Este instrumento tiene su nombre por las siglas de: Waikato Environment for Knowledge Analysis y es un software con licencia de distribución libre. Este programa cuenta con una serie

de algoritmos para ejecutar diferentes tipos de análisis y modelamientos de datos predictivos, también posee herramientas de visualización de registros.

Hall (2015) afirma que Weka está en constante evolución, la cual está enfocada principalmente en las nuevas funcionalidades del software, actualmente se tiene más de 4 años de desarrollo continuo, donde se tiene varios contribuidores ya que es un software Open Source y posee de un respaldo de la Universidad de Waikato, en este documento aparte de las nuevas funcionalidades también se habla de las futuras funcionalidades que se desean implementar y de los casos de uso a los que se ha empleado, figurando sus casos de éxito en distintos áreas temáticas y orientándolos a una variedad de instituciones que han aprobado la utilización del software (Mark Hall, 2015).

Esta es una herramienta muy variable que sobrelleva varias labores normales de filtros con información o minería en labores del procesamiento de datos, retroceso, clasificación, clusterin entre otras, del mismo modo permite visualizar y seleccionar los datos.

Todos los procesos en Weka están sustentados en la fusión de datos disponibles con la relación a un fichero plano, cada lista de datos esta descrito y fijado numéricamente con atributos nominales o numéricos. Permitiendo la entrada a otras peticiones de datos mediante sentencias SQL y gracias a la estructura JDBC, también puede procesar resultados generados en la base de consultas realizada en una base de datos.

Hoy en día además de weka existen a nivel comercial un sin número de herramientas prácticas para la salida de resultados de datos minados, varias poseen buenas e incluso

mejoradas características, pero en su mayoría todas son diseñadas con el mismo objetivo, se muestran tablas comparativas a continuación:

Tabla 10.

Comparación de herramientas de minería de datos

Característica	**Clementine**	**SAS Enterprise Miner**	**Tariykdd**	**Weka**
Licencia libre	No	No	Si	Si
Requiere conocimientos avanzados	No	No	No	No
Acceso a SQL	Si	No	Si	Si
Multiplataforma	No	Si	Si	Si
Requiere bases de datos especializadas	No	---	No	No
Métodos de máquinas de soporte vectorial	Si	Si	No	Si
Métodos bayesianos	Si	---	No	Si
Puede combinar modelos	Si	Si	No	Si (no resulta muy eficiente)
Modelos de clasificación	Si	Si	Si	Si
Implementa arboles de decisión	Si	Si	Si	Si
Modelos de regresión	Si	Si	No	Si
Clusterin y agrupamiento	Si	Si	No	Si
Interfaz amigable	Si	Si	Si	Si

Nota: Entre las herramientas más buscadas del mercado, se presenta a WEKA como la opción que cumple con los requerimientos necesarios que se adaptan a la metodología CRIPM-DM. (Elaboración propia, 2020)

Metodología de gestión del proyecto: PMBOK

Según el PMI (2015), PMBOK es una guía que contiene estándares globales para la gestión de Proyectos, estas son dictadas por el PMI (Project Management Institute).

Según PMBOK Fifth Edition (2014) un proyecto es un servicio relacionado con un producto único y tiene bien definidos un principio y un final. El final se logra cuando se realizan objetivos del proyecto, al terminar el proyecto debemos identificar por qué su finalidad no se cumplirá o no pueden ser cumplida, o el por qué ya no hace falta la precisión que dio origen al proyecto. De igual forma, el por qué se puede poner fin a un proyecto o si el cliente desea terminar el proyecto. También (PMBOK Fifth Edition, 2014) dice que existen 47 procesos dedicados a la trayectoria del proyecto que al mismo tiempo agrupa en diez Áreas de Conocimiento claramente diferenciadas.

El Área de Conocimiento represente en un grupo total de términos, conceptos y ejercicios que conforman un ámbito profesional, con una dirección a los proyectos áreas de especialización. Estas diez Áreas de Conocimiento se aplican en la mayoría de proyectos, durante la mayoría de la fracción de tiempo. Los grupos del proyecto deben aplicar estas diez Áreas de Conocimiento, también otras áreas de conocimiento, de la manera más conveniente y especifica según su proyecto.

Estas áreas son: Gestión integracional del proyecto, Gestión del alcance del proyecto, Gestión del tiempo del proyecto, Gestión de la calidad del proyecto, Gestión de recursos humanos del proyecto, Gestión de las comunicaciones del proyecto, Gestión de los riesgos del proyecto, Gestión de las adquisiciones del proyecto y Gestión de los interesados del proyecto.

Cada una de las Áreas de Conocimiento se trata en una sección específica de la Guía del PMBOK.

RapidMiner Studio

Es un software de minería dado por la compañía que tiene el mismo nombre. El proyecto se creó en 2001 dentro de la universidad de Dortmund en el país de Alemania, con el crecimiento de un programa poderoso y flexible de minería de datos, con el nombre de YALE (Yet Another Learning Enviroment). Programa muy conocido a un nivel tan alto que el equipo decidió lanzar su compañía.

Una de las principales ventajas de este instrumento es la plataforma unificada que tiene un entorno de programación visual fácil de utilizar, permitiendo soltar y arrastrar, de manera acelerada en el enfoque de análisis predictivo para aumentar la productividad. De igual manera, la plataforma es de código abierto, en la cual más de 250,000 expertos en datos crean y mantienen nuevas tendencias y necesidades del mercado. Puede acceder a trabajar desde cualquier lugar, ya que proporciona un repositorio principal basado en la nube.

Base de datos MySql

El sistema de MySQL es una herramienta que ayuda a la administración de bases de datos relacionales, el cual está en código abierto y está respaldado por la empresa Oracle y se basa en lenguaje de consulta estructurado, más conocido por sus siglas “SQL”. Al ser uno de los más populares gestores de bases de datos, MySQL se puede ejecutar en todos sistemas operativos, incluidos: Linux, UNIX y Windows. La fortaleza de esta base de datos es amplia,

aunque su uso se ha visto enfocado con mayor frecuencia en aplicaciones web y publicaciones en línea (Gallego, J., 2008).

Definición de términos básicos

- Bloquer: Conocido por ser una falla que, al ocurrir, no es posible continuar con el proceso de la función que se había seleccionado.
- Data Set: Se denomina Data Set al conjunto de datos a analizar con el software de minería de datos.
- ETL: Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes para, limpiarlos, unificarlos y cargarlos en un repositorio, este generalmente es una base de datos, data Warehouse o Data Mart.
- Matriz de confusión: Es una herramienta muy usada en inteligencia artificial, que sirve para ver las predicciones correctas de cada clase y haca donde fueron las predicciones erradas. Cada columna representa el número de predicciones de cada clase y cada fila representa las instancias reales de la clase.
- MD: Siglas pertenecientes a la tecnología de Minería de datos.
- Minería de Datos: Es el medio para hallar la información en grandes agrupaciones de datos, utilizando un análisis matemático para entender tendencias y los patrones que existen en los datos. Naturalmente, los patrones no pueden hallar mediante la búsqueda tradicional de los datos debido a las relaciones excesivamente complejas y también porque existen demasiados datos (Microsoft corporation, 2015).
- Redes Neuronales Artificiales: es un paradigma de aprendizaje y procesamiento automático derivado de la forma en que se procesa el sistema nervioso de los animales.

Es un sistema de neuronas interconectadas que se ayudan entre sí para crear un estímulo de salida (Norvig y Russell, 2014).

- Servicio Policial: El servicio policial se basa en mantener, garantizar y restaurar el orden interno, de igual manera presta protección y ayuda a la comunidad y a las personas. Garantizando el cumplimiento de las leyes y la seguridad del patrimonio público y privado. investiga, previene y batalla con la delincuencia. Controla y vigila las fronteras.
- UAT Son las siglas de User Aceptance Testing, que se refiere a todas las personas (usuarios) que van a probar la aplicación manualmente.
- Weka: Software de aprendizaje automático y Minería de Datos desarrollado por la Universidad de Waikato en Nueva Zelanda.

Capítulo iii

Automatización de rutas de patrullaje basados en modelos dinámicos y predictivos apoyados en el análisis de información delictual

Fase 1. Selección del caso

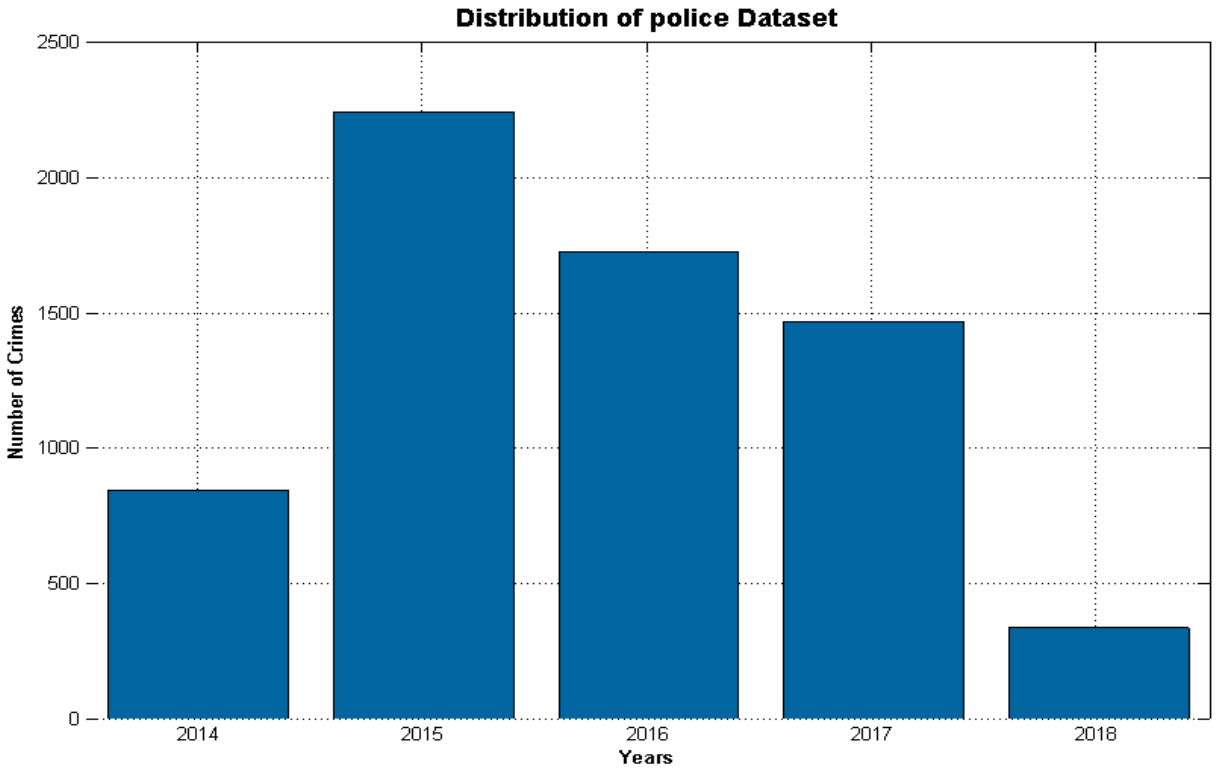
Se tomó una muestra de los registros de la base de datos original que es proporcional a la totalidad de casos criminalísticas ocurridos durante el año 2018 en el país del Ecuador.

Cabe aclarar que, si bien se seleccionó un dataset de pocos registros para que pueda ser visualizado e interpretado por el lector, no tiene sentido práctico aplicar técnicas de minería de datos sobre tan poca cantidad de información. Por lo tanto, el análisis y las conclusiones de este caso serán la medida que se tenga como muestra para implementar en la totalidad de los datos.

Para este artículo se utilizará una recopilación de información dada por la Policía Nacional de Ecuador llamada "PoliciaEc". Estos datos recopilan la información sobre delitos ejecutados entre los años 2014 y 2018 dividido por circuitos. Los circuitos desde donde se recopiló la información son los más conflictivos dentro del país y con la tasa más alta de delitos denunciados. Esta información es detallada y exacto al denunciar un delito cometido, ubicación geográfica (latitud y longitud), fecha y hora, tipo de delito y características genéricas de la víctima. Este conjunto de datos contó con 6605 registros en un período de 4 años, donde los delitos fueron registrados durante las 24 h del día y los 365 días del año. En la Figura 10, se presenta la distribución de los datos de los años 2014 a 2018.

Figura 10.

Datos de crímenes en Ecuador.



Nota: esta figura muestra el número de crímenes realizados en el ecuador desde el año 2014 hasta el 2018. (Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo y Edison Manjarres, 2020)

Usando este conjunto de datos, se propuso desarrollar un algoritmo que permitiera la creación de una ruta de la patrulla dinámica. Esta ruta debe cubrir de manera eficiente la mayor parte del área del circuito (distribución territorial Figura 11) donde se han denunciado delitos. Por otro lado, estas rutas deben variar según la hora del día.

Figura 11.

Zonas de patrullaje

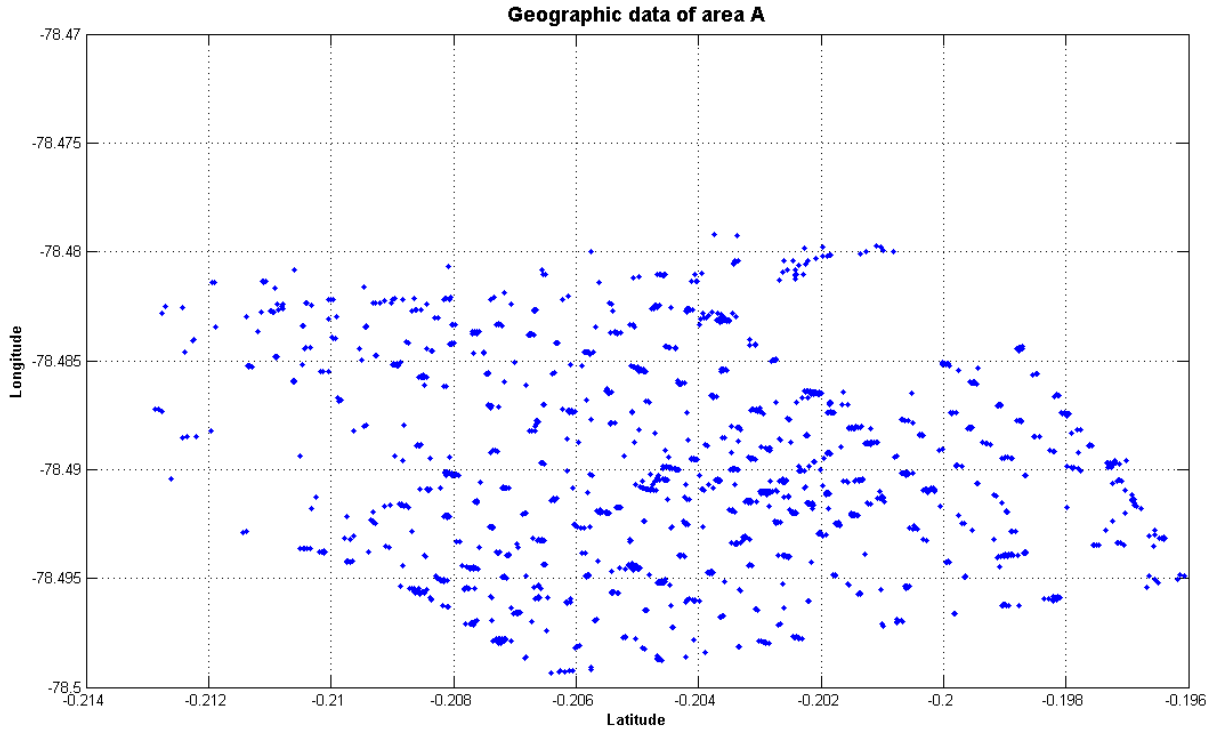
Nota: esta figura muestra la distribución de zonas de patrullaje. (Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo y Edison Manjarres, 2020)

El estudio se centró en reunir las parroquias que forman parte de un circuito, con una extensión de 5 km^2 y va hasta 5000 personas. Ecuador cuenta con 1134 circuitos, conformados para 2028 subcircuitos con una extensión aproximada de 1 km^2 de 5000 a 10000 personas.

En la figura 12 se muestran la distribución de datos geográficos por cada circuito. Esto permite identificar en una mejor forma de ubicar los lugares más conflictivos y, al final, poder diseñar un sistema eficiente modelo de patrullaje.

Figura 12.

Distribución de datos geográficos



Nota: esta figura muestra los datos por el área de patrullaje. (Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo y Edison Manjarres, 2020)

Fase 2. Elaboraciones de preguntas

Determinar los objetivos de prevención

La primera actividad de esta fase es determinar los objetivos de los subcircuitos de los diferentes distritos dentro del Ecuador, esto debido a que nuestra solución de predicción debe apoyar directamente a estos objetivos generando valor a los mismos.

Los objetivos de los circuitos son los que cumple la Policía Nacional del Ecuador y son los siguientes:

- Garantizar, mantener y restablecer el orden interno de la jurisdicción.
- Prestar protección y ayuda a las personas de la comunidad.
- Garantizar el cumplimiento de las leyes y la seguridad del patrimonio público y privado.
- Prevenir, investigar y combatir la delincuencia.
- Vigilar y controlar las fronteras con el propósito de defender a la sociedad y las personas en el marco de una cultura de paz y de respecto a los derechos humanos

Determinar los objetivos de minería de datos

Determinar un modelo eficaz de minería de datos para predecir los hechos delictivos dentro del Ecuador.

- Conseguir como un porcentaje mínimo de aciertos según los objetivos de prevención, en este caso los circuitos (detallado como criterio de éxito).
- Determinar las zonas donde podría ocurrir hechos delictivos según una selección de variables de entrada.
- Mostrar los resultados de las predicciones en un mapa geográfico.
- Criterios claves de éxito:
- Según los objetivos del negocio, y las denuncias que se registran mensualmente podemos tener como criterio de éxito que:
- Se obtenga como mínimo un 90% de probabilidad de aciertos en las predicciones.

- Se tenga un detalle de las zonas con los delitos más comunes con su probabilidad de ocurrencia.
- Se pueda realimentar el modelo con más data histórica.
- Se puedan predecir más hechos delictivos según vaya pasando el tiempo (no solo para uno o dos meses siguientes, sino que se pueda volver continuo en el futuro).

Fase 3. Localización de fuentes

En el presente capítulo se describe la estructura de la información recibida y la conformación del dataset. Se denomina data set al conjunto de datos a analizar con el software de minería de datos. El proceso de conformación del dataset a partir de una base de datos involucra diversas etapas:

- Consolidación de la información de interés en una única tabla;
- Selección de los campos de interés;
- Depuración de registros en busca de completitud y consistencia;
- Modificación de las variables de los campos en función del software y los algoritmos a utilizar y/o de la visión del especialista.

En los siguientes puntos se desarrollan estos pasos en base a la información suministrada por la base de datos de la policía nacional del Ecuador.

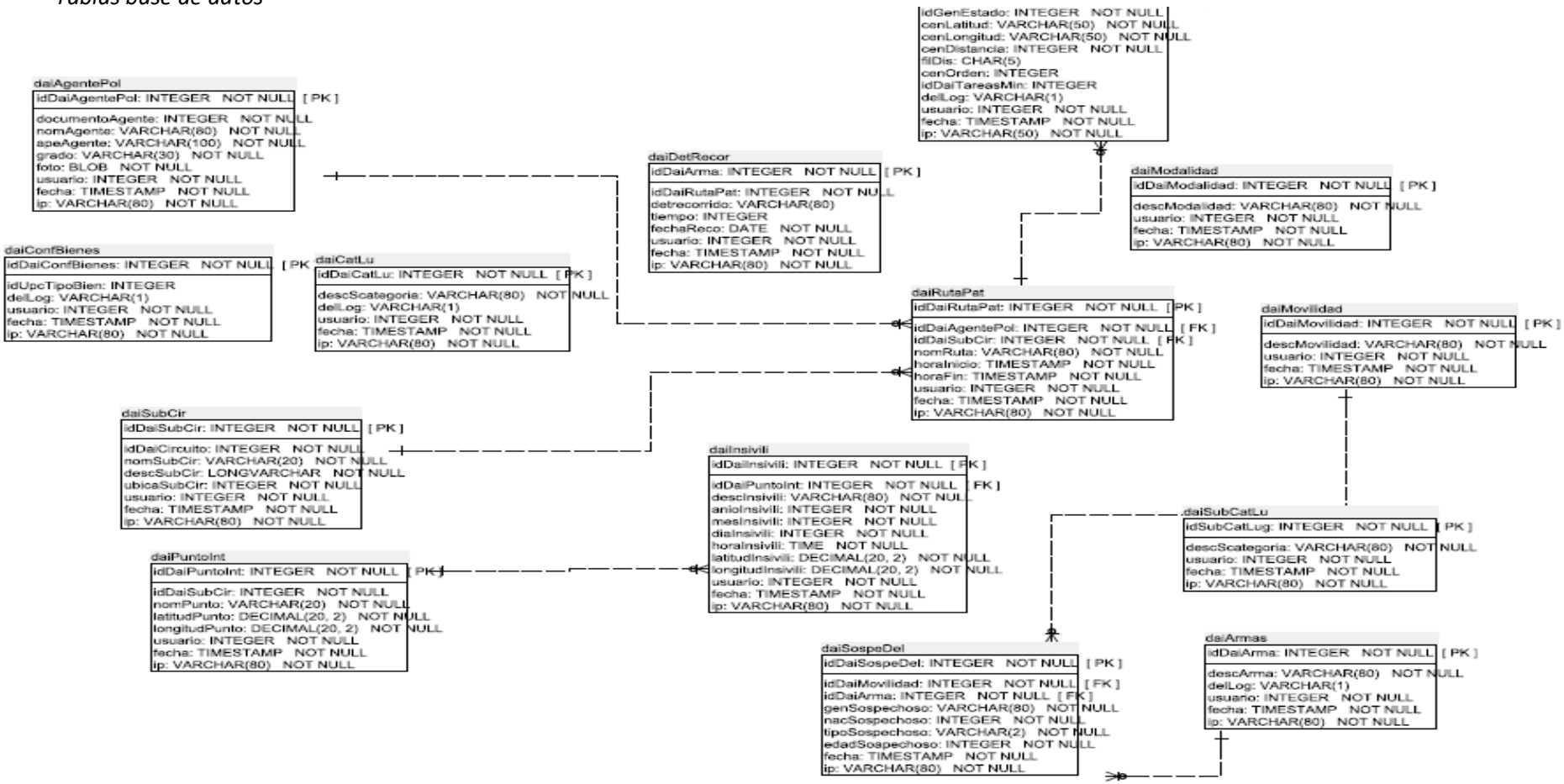
Consolidación de la información

Se tiene un registro que contiene los 65968 hechos delictivos ocurridos en 2018 registrados en la base de datos interna mostrada en la Figura 13. Está compuesta por 4 secciones:

- Sección 1: Como tabla central tenemos esta “daiRutaPat”, la cual está encargada de guardar un registro de las rutas que se hacen en cada uno de las secciones. Esta tabla tiene como herederas a otras 7 tablas, quienes captaran la información del agente que está asignado a cada ruta, las diferentes rutas prestablecidas, datos relevantes sobre recorridos y vehículos que se utilizan.
- Sección 2: En este grupo de 2 tablas se detalla toda la información de un delito y quien fue responsable del mismo.
- Sección 3: Se guarda en un grupo de 4 tablas, los incidentes que no llegaron a ser considerados como delitos pero que quedan con un estado de sospechas para un futuro rastreo. De igual manera se almacenan los datos personales, así como el sector y ruta captada.
- Sección 4: Aquí se encuentran 2 tablas flotantes que sirven como referencia interna del sistema. Una de ellas es el estado, el cual permitirá cambiar a un usuario como activo o inactivo. Y la segunda, sirve para saber si existe un tipo de movilidad acorde a un vehículo.

Figura 13.

Tablas base de datos



La primera sección 1 y la 2 son las bases para poder realizar un análisis delictivo y minar los datos necesarios en rutas de patrullaje, ya que contiene la mayor cantidad de información relevante, serán las secciones bases para formar el data set. Según la opinión del departamento interno, las tablas secundarias son de baja calidad (hay muchos registros incompletos) y aportan poca información sobre la víctima y el imputado. Por esta razón, sumado a la dificultad para consolidarlas junto a la sección principal de forma que cada registro represente un hecho, estas tablas serán excluidas del análisis.

Selección de los campos de interés

La información registrada en la base de datos relacionada con el perímetro de estudio, el cual está en la zona 9 de la subzona distrito metropolitano de quito, del distrito Eugenio espejo, del circuito Universitario, división que se tiene en los registros, para lo cual contamos con 47 atributos los cuales se indica a continuación:

1. id_distrito
2. distrito
3. codigo_circuito
4. circuito
5. cod_subcir
6. subcircuito
7. dpa_zona
8. provincia
9. canton
10. parroquia
11. sector
12. fecha_infraccion
13. hora_infraccion
14. tipo_delito
15. delito
16. modalidad
17. movilidad_victimario
18. latitud
19. longitud

20. numfiscalia
21. direccion_infraccion
22. categoria
23. subcategoria
24. origen_noticia
25. victima_denunciante
26. sexo
27. edad
28. estado_civil
29. nacionalidad_victima
30. pertenencia_etnica
31. condicion_victima
32. profesion
33. instruccion
34. consumo_alcohol
35. detenido_sospechoso
36. sospechoso_sexo
37. nacionalidad_sospechoso
38. edad_sospechoso
39. tipo_automotor_robado
40. vehiculo_robado
41. marca
42. modelo_robado
43. anio_fabricacion
44. color
45. no_motor
46. no_chasis
47. placas_robado

Campos seleccionados

Es necesario realizar un filtro de los campos que se deben omitir para reducir el tamaño del dataset final, para lo cual se procede a eliminar los 11 primeros atributos y direccion_infraccion, por ser un análisis focalizado a un determinado circuito no tendrá mayor relevancia ni peso y al contrario solo ocasionará ruido quedando los siguientes:

1. fecha_infraccion: determina la fecha en la que se cometió una infracción.
2. hora_infraccion: determina la hora de la infracción.
3. tipo_delito: es el dato que clasifica el tipo de delito como robo, asalto, etc.

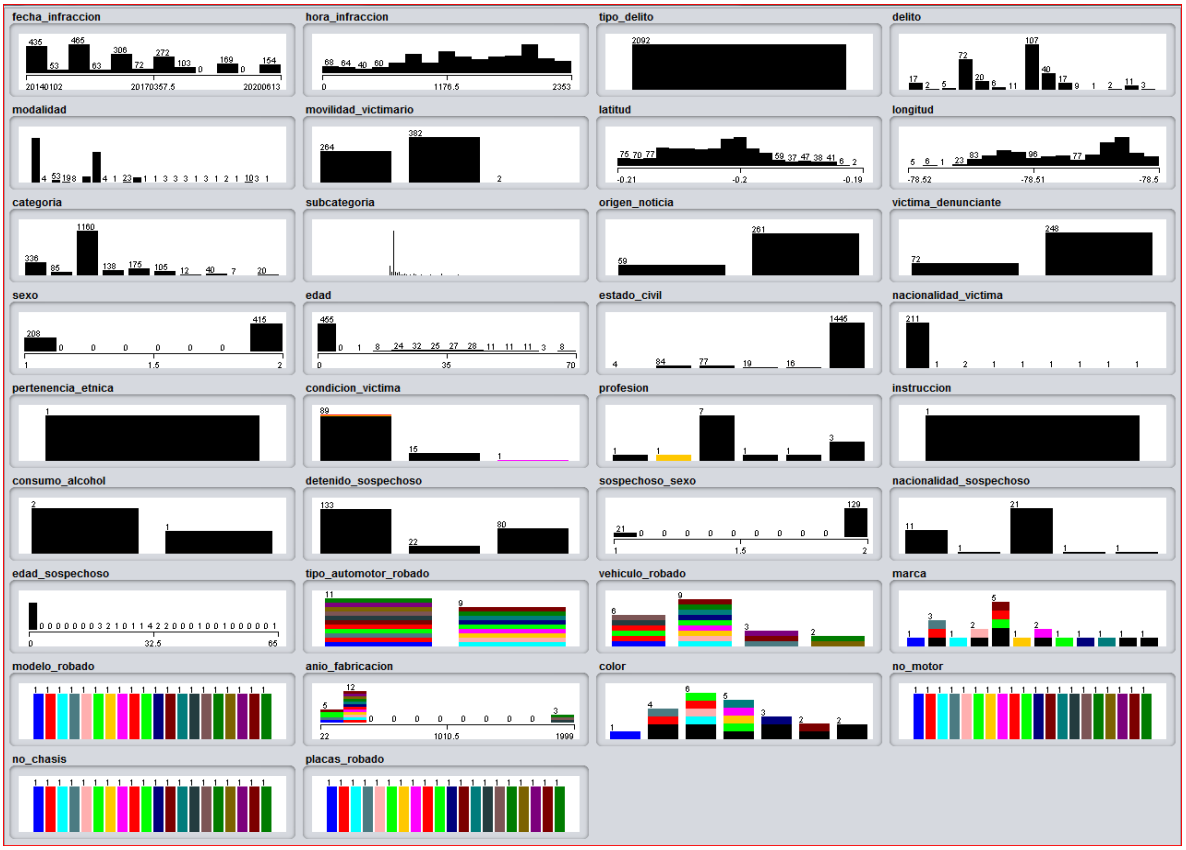
4. delito: es la descripción del delito cometido.
5. modalidad: es el modo en el que se realizó el delito.
6. movilidad_victimario: describe cual fue el tipo de movilidad en la que se encontraba la víctima como vehículo, a pie, moto, etc.
7. latitud: coordenada acorde a la latitud en donde se realizó el delito.
8. longitud: coordenada acorde a la longitud en donde se realizó el delito.
9. numfiscalia: número de la fiscalía en la que se realizó la denuncia del delito.
10. categoría: categoría en la que entra el delito como robo, asesinato, etc.
11. subcategoría: categoría en específico del delito como robo a mano armada, robo en moto, etc.
12. origen_noticia: lugar donde se origina la noticia del delito.
13. victima_denunciante: nombre del denunciante o la víctima.
14. sexo: género de la persona que denunció el delito.
15. edad: edad de la persona que denunció el delito.
16. estado_civil: estado de la persona que denunció el delito.
17. nacionalidad_victima: dato que expone de que nacionalidad es la persona que denunció el delito.
18. pertenencia_etnica: tipo de etnia a la que forma el denunciante.
19. condicion_victima: descripción si el denunciante o victima tiene algún tipo de condición como lesiones, dolencias, traumas, etc.
20. profesión: descripción del tipo de trabajo que realiza el denunciante.
21. Instrucción: grado académico del denunciante o víctima.
22. consumo_alcohol: valor que permite determinar si la victima ha ingerido alcohol.
23. detenido_sospechoso: nombre del sospecho del delito.
24. sospechoso_sexo: sexo del sospechoso del delito.
25. nacionalidad_sospechoso: nacionalidad perteneciente del sospechoso.
26. edad_sospechoso: edad del sospechoso del delito.
27. tipo_automotor_robado: determina cual es el tipo de automor que se denuncia como carro, moto, etc.
28. vehiculo_robado: descripción del automotor robado.
29. marca: determina la marca del automotor.
30. modelo_robado: determina cual es el modelo del automotor.
31. anio_fabricacion: determina el año en el que se fabricó el automotor.
32. color: describe el color o combinación de colores del automotor.
33. no_motor: dato que describe el número del motor.
34. no_chasis: dato que describe el número del chasis.
35. placas_robado: placa del automotor que fue robado.

Con los atributos filtrados en una primera parte, se procede con la discriminación de una forma más técnica hasta tener las características más optimas de los datos y que aporten información relevante al modelo planteado, dejando como resultado en la dataset un total de

2902 instancias. Cada uno de los campos seleccionados serán trasladados al dataset para poder tener una estructura de tabla donde se pueda observar los datos específicos de cada delito registrado en la base de datos.

Figura 14.

Flujo de datos en de los campos origen



Nota: la figura muestra la frecuencia de cada uno de los campos de la data set. (Elaboración propia, 2020)

Mediante la figura de distribución de datos se pueden determinar cuáles campos tienen mayor frecuencia, lo que será de utilidad para hacer una selección previa al pre procesamiento de datos, se puede ir identificando los tipos de campos con los que se puede trabajar.

Figura 15.

Dataset de la base de datos de crímenes

distrito	codigo_circui	circuito	cod_subcir	subcircuito	dpa_zona	provincia	canton	parroquia	sector	fecha_infrac	hora_infracci	tipo_delito	delito	modalidad	movilidad_vi	latitud	longitud
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/07 0	19:00	DELITOS CON	ROBO DE BIE	ESTRUCHE		-0.203062235	-7.850.024.541.859
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/02 0	17:50	DELITOS CON	ROBO A PERS	ASALTO	A PIE	-0.202610150	-78.500.011.239.738
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/11 0	5:40	DELITOS CON	ROBO A UNI(	ESTRUCHE		-0.205414949	-78.508.514.839.529
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/04 0	19:35	DELITOS CON	ROBO A UNI(	ASALTO	A PIE	-0.196154793	-78.499.431.336.373
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/07 0	6:40	DELITOS CON	ROBO A PERS	ASALTO	A PIE	-0.197641167	-78.500.389.934.154
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/07 0	16:10	DELITOS CON	ROBO A PERS	ASALTO	VEHICULO	-0.203264605	-78.498.337.541.312
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/07 0	13:30	DELITOS CON	ROBO A PERS	CARTERISTAS		-0.198775640	-78.496.283.279.634
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/05 0	19:15	DELITOS CON	ROBO A PERS	ASALTO		-0.195313023	-78.507.009.626.004
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/06 0	16:28	DELITOS CON	ESTAFA	ENGAÃ'O		-0.197675219	-7.850.025.943.874
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/09 0	9:00	DELITOS CON	ROBO A PERS	ASALTO	VEHICULO	-0.196621738	-78.501.938.573.928
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/09 0	22:20	DELITOS CON	ROBO A PERS	ASALTO	A PIE	-0.196194954	-78.507.947.939.388
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/14 0	12:00	DELITOS CON	ESTAFA	ENGAÃ'O		-0.199845817	-78.500.908.479.798
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/09 0	23:30	DELITOS CON	ROBO A PERS	ASALTO	A PIE	-0.195314031	-78.507.725.430.936
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/02/08 0	9:04	DELITOS CON	ROBO DE BIE	ESTRUCHE		-0.191219715	-78.509.174.016.985
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/02/03 0	21:45	DELITOS CON	ROBO A PERS	ASALTO	A PIE	-0.200135300	-785.007.154.941.559
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/02/10 0	14:00	DELITOS CON	ROBO A PERS	ASALTO	A PIE	-0.196446016	-78.502.946.733.516
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/02/05 0	20:30	DELITOS CON	ROBO A PERS	ATURDIMIEN	VEHICULO	-0.203264605	-78.498.353.634.568
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/16 0	17:15	DELITOS CON	ABUSO DE C(	DISTRACCIÃ"N		-0.198221689	-78.496.047.767.264
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/21 0	21:40	DELITOS CON	ROBO A CAR	ESTRUCHE		-0.192801812	-78.511.155.415.591
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/02/11 0	14:00	DELITOS CON	ROBO A PERS	ASALTO	A PIE	-0.196316151	-78.499.433.525.878
EUGENIO ESF	17D05C04	UNIVERSITAR	17D05C04S0	UNIVERSITAR	ZONA 9	PICHINCHA	QUITO	QUITO	URBANO	2020/01/18 0	17:27	DELITOS CON	ROBO DE BIE	ESTRUCHE	VEHICULO	-0.200594060	-78.510.919.457.938

Nota: la figura muestra el data set completo con cada uno de los campos y sui información. (Elaboración propia, 2020).

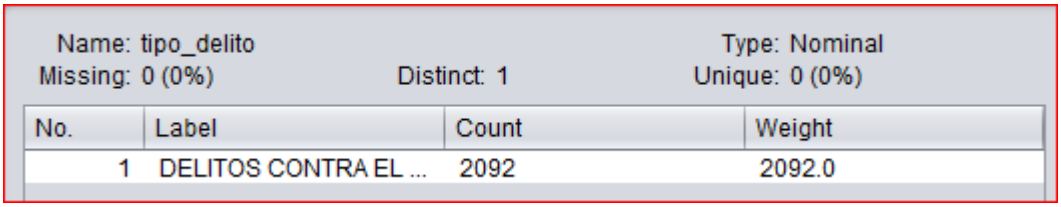
Campos omitidos

Se describe a continuación todos los campos que se consideran innecesarios para la construcción del dataset final:

tipo_delito: se elimina ya que todos los datos se tratan de DELITOS CONTRA EL DERECHO A LA PROPIEDAD los 2092 son de un solo tipo.

Figura 16.

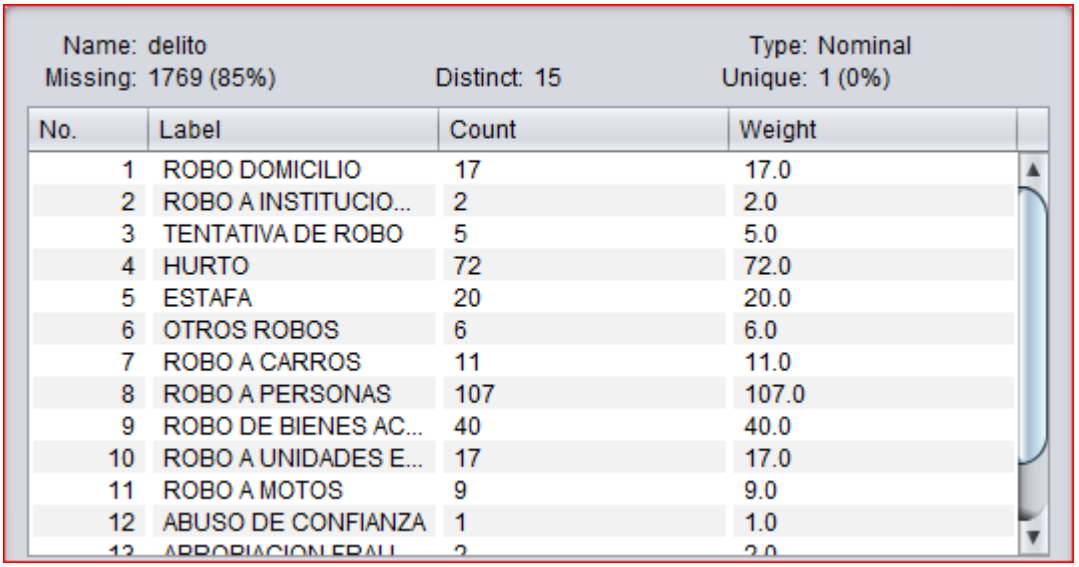
Campo tipo de delito



delito: se elimina ya que el 85% de los datos son desconocidos por lo que hace complicado aplicarle algún filtro.

Figura 17.

Campo delito



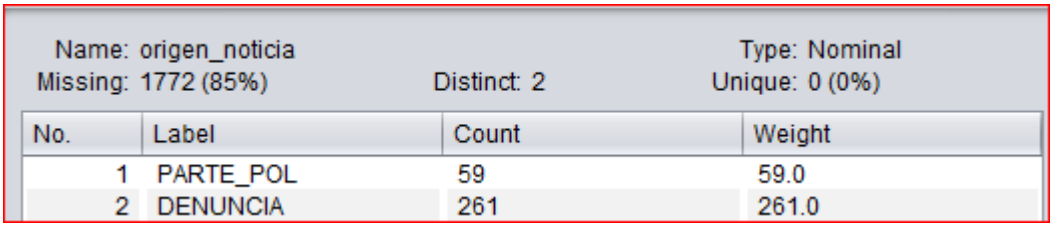
Name: delito
Missing: 1769 (85%)
Distinct: 15
Type: Nominal
Unique: 1 (0%)

No.	Label	Count	Weight
1	ROBO DOMICILIO	17	17.0
2	ROBO A INSTITUCIO...	2	2.0
3	TENTATIVA DE ROBO	5	5.0
4	HURTO	72	72.0
5	ESTAFA	20	20.0
6	OTROS ROBOS	6	6.0
7	ROBO A CARROS	11	11.0
8	ROBO A PERSONAS	107	107.0
9	ROBO DE BIENES AC...	40	40.0
10	ROBO A UNIDADES E...	17	17.0
11	ROBO A MOTOS	9	9.0
12	ABUSO DE CONFIANZA	1	1.0
13	APROPIACION FRAU...	2	2.0

origen_noticia: se elimina porque existe un 85% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 18.

Campo origen de noticia



Name: origen_noticia
Missing: 1772 (85%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	PARTE_POL	59	59.0
2	DENUNCIA	261	261.0

victima_denunciante: se elimina porque existe un 85% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 19.

Campo denunciante

Name: victima_denunciante Type: Nominal
Missing: 1772 (85%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	DENUNCIANTE	72	72.0
2	VICTIMA	248	248.0

estado_civil: se elimina porque existe más del 21% de datos desconocidos, pero también existe 1445 con un estado que no existe y 16 sin dato lo que conlleva más de un 90% por lo que hace complicado aplicarle algún filtro.

Figura 20.

Campo estado civil

Name: estado_civil Type: Nominal
Missing: 447 (21%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count	Weight
1	VIUDO	4	4.0
2	CASADO	84	84.0
3	SOLTERO	77	77.0
4	DIVORCIADO	19	19.0
5	SIN DATO	16	16.0
6	\N	1445	1445.0

nacionalidad_victima: se elimina porque existe un 89% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 21.

Campo nacionalidad de la victima

Name: nacionalidad_victima | Type: Nominal
Missing: 1872 (89%) | Distinct: 9 | Unique: 7 (0%)

No.	Label	Count	Weight
1	ecuatoriana	211	211.0
2	colombiana	1	1.0
3	estadounidense	2	2.0
4	venezolana	1	1.0
5	cubana	1	1.0
6	alemana	1	1.0
7	checa	1	1.0
8	canadiense	1	1.0
9	italiana	1	1.0

pertenencia_etnica: se elimina porque existe un 100% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 22.

Campo pertenencia étnica

Name: pertenencia_etnica | Type: Nominal
Missing: 2091 (100%) | Distinct: 1 | Unique: 1 (0%)

No.	Label	Count	Weight
1	MESTIZA	1	1.0

condicion_victima: se elimina porque existe un 95% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 23.

Campo condición de la victima

Name: condicion_victima
Missing: 1987 (95%)
Distinct: 3
Type: Nominal
Unique: 1 (0%)

No.	Label	Count	Weight
1	TIENE TRABAJO O NE...	89	89.0
2	ESTUDIANTE	15	15.0
3	OTRO	1	1.0

profesión: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 24.

Campo profesión de la victima

Name: profesion
Missing: 2078 (99%)
Distinct: 6
Type: Nominal
Unique: 4 (0%)

No.	Label	Count	Weight
1	NO MENCIONA	1	1.0
2	DOCENTE	1	1.0
3	POLICIA NACIONAL	7	7.0
4	ABOGADO	1	1.0
5	SERVIDOR POLICIAL	1	1.0
6	COMERCIANTE	3	3.0

Instrucción: se elimina porque existe un 100% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 25.

Campo instrucción de la victima

Name: instruccion Type: Nominal
Missing: 2091 (100%) Distinct: 1 Unique: 1 (0%)

No.	Label	Count	Weight
1	SECUNDARIA	1	1.0

consumo_alcohol: se elimina porque existe un 100% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 26.

Campo consumo de alcohol

Name: consumo_alcohol Type: Nominal
Missing: 2089 (100%) Distinct: 2 Unique: 1 (0%)

No.	Label	Count	Weight
1	SI	2	2.0
2	NO	1	1.0

detenido_sospechoso: se elimina porque existe un 89% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 27.

Campo detenido o sospechoso

Name: detenido_sospechoso Type: Nominal
Missing: 1857 (89%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	sospechoso	133	133.0
2	detenido	22	22.0
3	S/D	80	80.0

sospechoso_sexo: se elimina porque existe un 93% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 28.

Campo sexo del sospechoso

Name: nacionalidad_sospechoso Type: Nominal
Missing: 2057 (98%) Distinct: 5 Unique: 3 (0%)

No.	Label	Count	Weight
1	venezolana	11	11.0
2	cubana	1	1.0
3	ecuatoriana	21	21.0
4	guatemalteca	1	1.0
5	colombiana	1	1.0

nacionalidad_sospechoso: se elimina porque existe un 98% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 29.

Campo nacionalidad del sospechoso

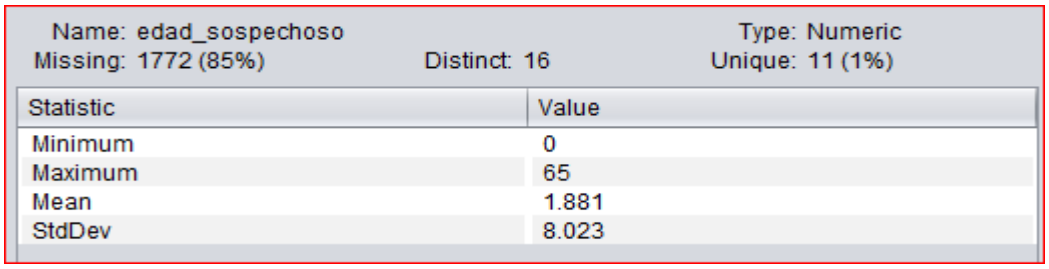
Name: sospechoso_sexo Type: Numeric
Missing: 1942 (93%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	2
Mean	1.86
StdDev	0.348

edad_sospechoso: se elimina porque existe un 85% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 30.

Campo edad del sospechoso



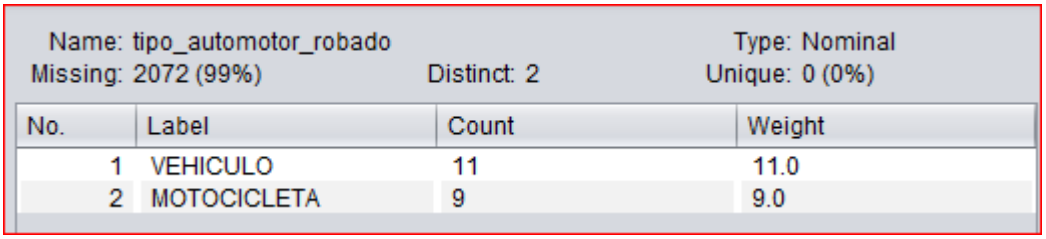
Name: edad_sospechoso — Type: Numeric
Missing: 1772 (85%) — Distinct: 16 — Unique: 11 (1%)

Statistic	Value
Minimum	0
Maximum	65
Mean	1.881
StdDev	8.023

tipo_automotor_robado: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 31.

Campo auto robado



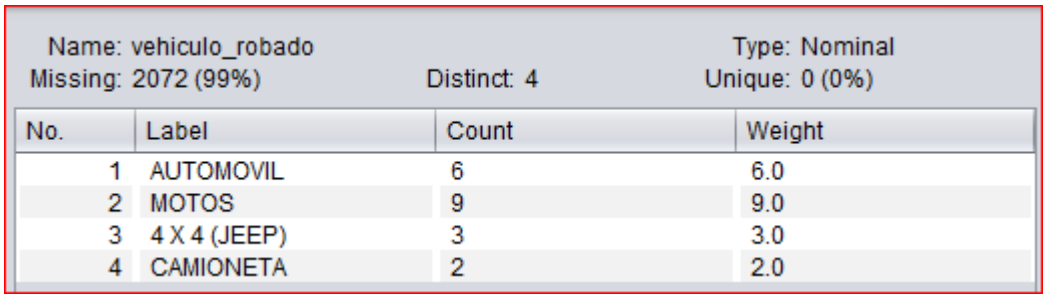
Name: tipo_automotor_robado — Type: Nominal
Missing: 2072 (99%) — Distinct: 2 — Unique: 0 (0%)

No.	Label	Count	Weight
1	VEHICULO	11	11.0
2	MOTOCICLETA	9	9.0

vehiculo_robado: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 32.

Campo vehículo robado



Name: vehiculo_robado — Type: Nominal
Missing: 2072 (99%) — Distinct: 4 — Unique: 0 (0%)

No.	Label	Count	Weight
1	AUTOMOVIL	6	6.0
2	MOTOS	9	9.0
3	4 X 4 (JEEP)	3	3.0
4	CAMIONETA	2	2.0

marca: se elimina existe porque un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 33.

Name: modelo_robado
Missing: 2072 (99%)
Distinct: 20
Type: Nominal
Unique: 20 (1%)

No.	Label	Count	Weight
1	SENTRA 1.6 A/T	1	1.0
2	I1 5DR 1.1 AC	1	1.0
3	DY25GY	1	1.0
4	TERRACAN GL CRDI...	1	1.0
5	FX 15 FORTE	1	1.0
6	AVEO ACTIVO 1.6L 4P...	1	1.0
7	TRACKER 25	1	1.0
8	BROTHER 25R	1	1.0
9	AVEO FAMILY STD 1.5...	1	1.0
10	2GY	1	1.0
11	MAX	1	1.0
12	VITARA 3P STD T/M IN...	1	1.0
13	XY25GY	1	1.0

Campo marca de auto

modelo_robado: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 34.

Name: marca
Missing: 2072 (99%)
Distinct: 12
Type: Nominal
Unique: 8 (0%)

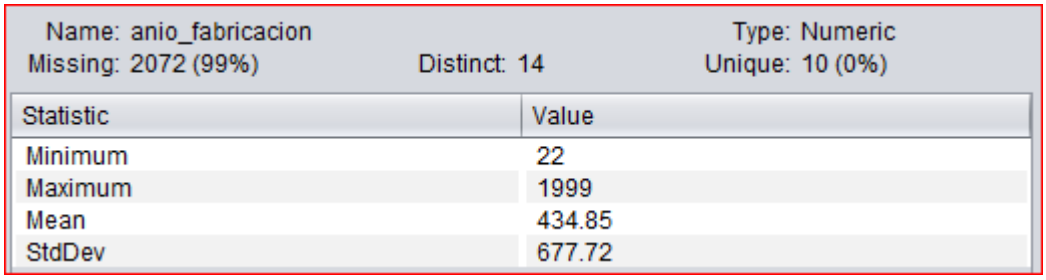
No.	Label	Count	Weight
1	Nissan	1	1.0
2	Hyundai	3	3.0
3	Daytona	1	1.0
4	Motor Uno	2	2.0
5	Chevrolet	5	5.0
6	Axxo	1	1.0
7	Otros	2	2.0
8	Ranger	1	1.0
9	UM	1	1.0
10	Shineray	1	1.0
11	Suzuki	1	1.0
12	Lada	1	1.0

Campo modelo de auto robado

anio_fabricacion: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 35.

Campo año de fabricación de auto



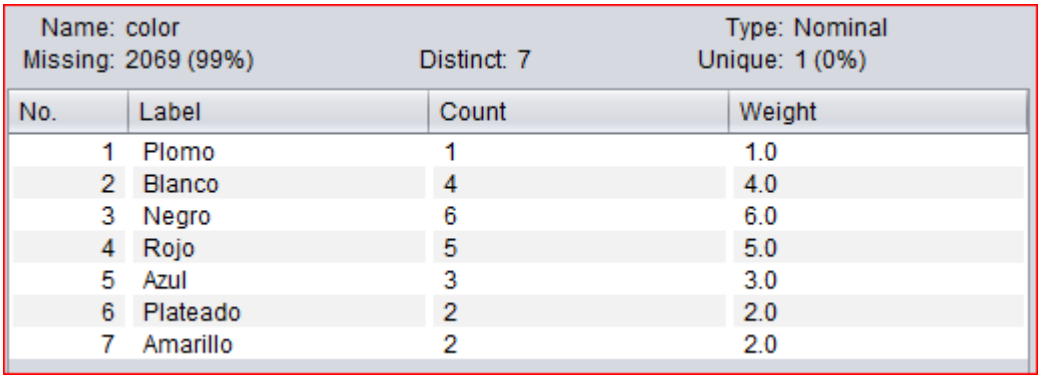
Name: anio_fabricacion Type: Numeric
Missing: 2072 (99%) Distinct: 14 Unique: 10 (0%)

Statistic	Value
Minimum	22
Maximum	1999
Mean	434.85
StdDev	677.72

color: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 36.

Campo color del auto



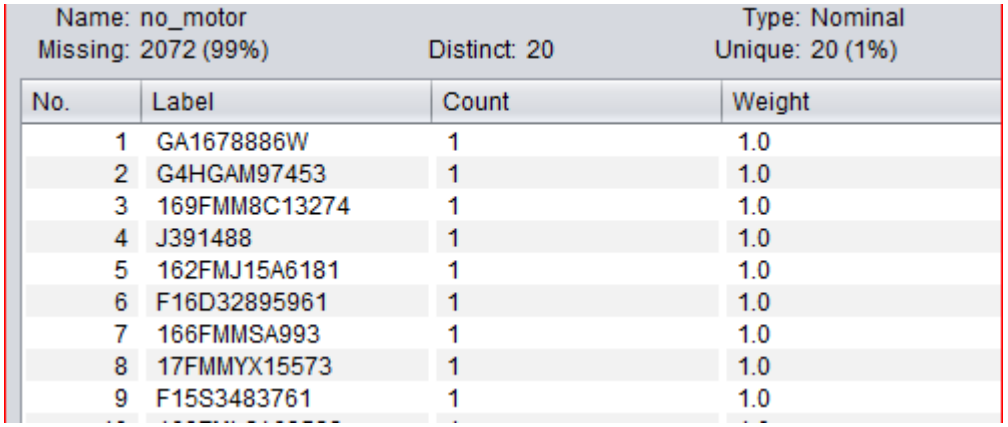
Name: color Type: Nominal
Missing: 2069 (99%) Distinct: 7 Unique: 1 (0%)

No.	Label	Count	Weight
1	Plomo	1	1.0
2	Blanco	4	4.0
3	Negro	6	6.0
4	Rojo	5	5.0
5	Azul	3	3.0
6	Plateado	2	2.0
7	Amarillo	2	2.0

no_motor: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro.

Figura 37.

Campo motor de auto



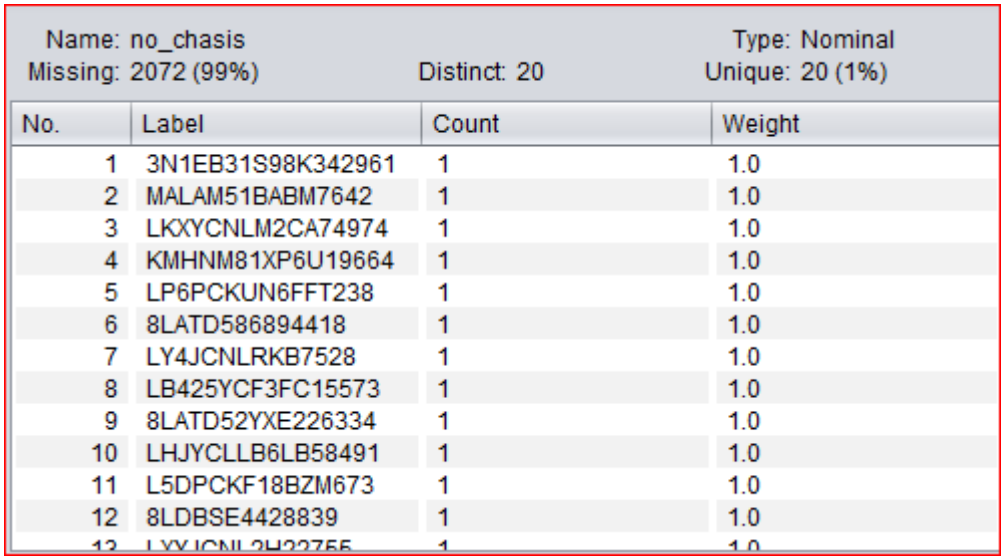
Name: no_motor
Missing: 2072 (99%)
Distinct: 20
Type: Nominal
Unique: 20 (1%)

No.	Label	Count	Weight
1	GA1678886W	1	1.0
2	G4HGAM97453	1	1.0
3	169FMM8C13274	1	1.0
4	J391488	1	1.0
5	162FMJ15A6181	1	1.0
6	F16D32895961	1	1.0
7	166FMMSA993	1	1.0
8	17FMMYX15573	1	1.0
9	F15S3483761	1	1.0

no_chasis: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro

Figura 38.

Campo casis del auto



Name: no_chasis
Missing: 2072 (99%)
Distinct: 20
Type: Nominal
Unique: 20 (1%)

No.	Label	Count	Weight
1	3N1EB31S98K342961	1	1.0
2	MALAM51BABM7642	1	1.0
3	LKXYCNLM2CA74974	1	1.0
4	KMHNM81XP6U19664	1	1.0
5	LP6PCKUN6FFT238	1	1.0
6	8LATD586894418	1	1.0
7	LY4JCNLRKB7528	1	1.0
8	LB425YCF3FC15573	1	1.0
9	8LATD52YXE226334	1	1.0
10	LHJYCLLB6LB58491	1	1.0
11	L5DPCKF18BZM673	1	1.0
12	8LDBSE4428839	1	1.0

placas_robado: se elimina porque existe un 99% de datos desconocidos lo que hace complicado aplicarle algún filtro

Figura 39.

Campo placas del auto

Name: placas_robado Type: Nominal
Missing: 2072 (99%) Distinct: 20 Unique: 20 (1%)

No.	Label	Count	Weight
1	PBB5811	1	1.0
2	LBA6837	1	1.0
3	HG68Q	1	1.0
4	HBB4958	1	1.0
5	IF545K	1	1.0
6	TBA337	1	1.0
7	IV321X	1	1.0
8	IM436C	1	1.0
9	PCI2169	1	1.0
10	JA515F	1	1.0
11	HB687W	1	1.0
12	PBD255	1	1.0
13	IO2I	1	1.0

Una vez removido estos campos por las razones indicadas se reducen a 10 atributos con los que se procede a trabajar. Al no poseer una clase definida por la que se quiera predecir, se enfocara en clusterizar la data, lo que hace necesario que todos los atributos sean de un tipo específico (numeric) por lo que con la utilización de filtros de weka se procede a tratar los atributos con la finalidad de conseguir lo que se requiere. Las variables resultantes son las siguientes:

- fecha_infraccion: determina la fecha exacta en la que se realizó el delito.
- hora_infraccion: determina la hora exacta en la que se realizó el delito.
- modalidad: determina cual es el tipo de modalidad con la cual fue asaltado la víctima.
- movilidad_victimario: determina cual es el tipo de movilidad que estaba usando la victima cuando se produjo el delito.

- latitud: coordenada exacta de la latitud donde se produjo el delito.
- longitud: coordenada exacta de la longitud donde se produjo el delito.
- categoría: tipo de delito.
- subcategoría: subtipo del delito.
- sexo: género de la persona que denuncio o victima el delito.
- edad: edad de la persona que denuncio o victima el delito.

Fase 4. Análisis e interpretación de la información y resultados

Se analizó el data set obtenido en el capítulo anterior y con el software Weka 3.5.5. En primer lugar, se aplicó el algoritmo K-means para agrupar los 2092 registros en varias pruebas utilizando de entre 5 a 10 clusters, hasta encontrar el número necesario que se ajusten a la realidad, razón por la cual para identificar de una forma más técnica estas pruebas se hizo uso del método de codo (Elbow Method), el que proporciona el número más exacto los clusters con los que debería trabajar, obteniendo un número de 5 clusters. Luego, con las herramientas descriptas, se obtuvo una primera caracterización de los clusters y finalmente se utilizó el algoritmo J48 para una interpretación formal.

Algoritmo Kmeans

Este algoritmo es un procedimiento de clustering o un conjunto de datos. Se define al clustering como un método dentro de un conjunto un determinado de datos y los clasifica. Es de esta manera como los datos que comparten semejanzas serán agrupados y separados de otros grupos que no cuenten con datos parecidos. Para identificar cual es el rango de semejanza o diferencia entre datos, el algoritmo de K-means desarrolla la comparación estructural. En resumen, la distancia media entre los datos será el valor a tomar en un modelo euclideana que cuentan con algoritmos definidos de clustering con aprendizaje sin supervisión, como en efecto

para el proyecto se utiliza la distancia manhattan. La cual busca una secuencia lógica de patrones en los campos sin contar con una variable dependiente y realizando una predicción con un objetivo en específico.

El algoritmo de K-means precisa de un tipo de dato como entrada que especifica el número de clusters que tendrá la población. Al tener este ajuste de K clusters, el algoritmo ubica primero k puntos aleatorios que se denominan como: centroides; despues asigna a estos puntos todas las muestras con las distancias que tienen entre ellas.

Una vez obtenidos los puntos centrales, estos se desplazarán a las muestras más cercanas. Este proceso generará una nueva asignación entre las muestras originales, puesto que algunos datos están ahora más cerca entre los centroides. Esta secuencia se repite de modo reiterado y los grupos se van alineando hasta que la asignación no cambie entre los puntos céntricos. Este es el resultado el cual se considera final por el tipo de ajuste que maximiza la distancia que se encuentra entre los distintos grupos y minimiza las diferencias de datos.

El algoritmo consta de los siguientes pasos:

1. Apertura: después de elegir un conjunto de números, k, se determina k centroides en un determinado espacio de los datos, escogido aleatoriamente.
2. Fijar objetos a los centroides: cada dato tiene como finalidad ser asignado a su centroide más próximo.
3. Actualización centroides: la posición de los centroides ubicados en cada grupo se actualiza teniendo en cuenta como nuevo centroide como figura del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que no se mueven los centroides, o tengan un movimiento por debajo de una trayectoria umbral en cada paso. El algoritmo k-means soluciona una dificultad de optimización, teniendo una función de optimizar (minimizar) el aumento de las distancias cuadráticas de cada uno de los objetos del centroide de su cluster.

Los objetos se presentan como vectores existentes en d dimensiones (x1,x2,…,xn) mientras que el algoritmo k-means conforma k conjunto que se minimiza la acumulación de distancias de los objetos, dentro de los diferentes conjuntos S={S1,S2,…,Sk} , a su centroide. La cuestión se puede plantear de la siguiente forma:

$$\min_S E(\mu_i)=\min_S \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j-\mu_i\|_2 \quad (1)$$

Donde S es la agrupación de datos en la cual sus elementos son los objetos xj representados por vectores, cada una de sus partes se figura como un atributo. El comprender los conjuntos k o clusters con un semejante centroide μi .

Durante las actualizaciones de los centroides, desde una perspectiva matemática, se exige como condición indispensable a la función E(μi) que, para la función cuadrática es:

$$\partial E \partial \mu_i = 0 \Longrightarrow \mu_{(t+1)i} = 1 \| S_{(t)i} \| \sum_{x_j \in S_{(t)i}} x_j$$

Una de las principales ventajas del procedimiento k-means son que es un procedimiento sencillo. Y es indispensable argumentar que la valoración de k y el resultado final dependerá del inicio de los centroides. Al principio no converge al mínimo global sino a un mínimo local.

Método de codo

En este método se utilizan las valorizaciones de inercia adquiridos tras aplicar el K-means a números diferentes de Clusters (desde 1 a N Clusters), siendo la inercia la adición de las distancias al cuadrado de los diferentes objeto del Cluster a su centroide:

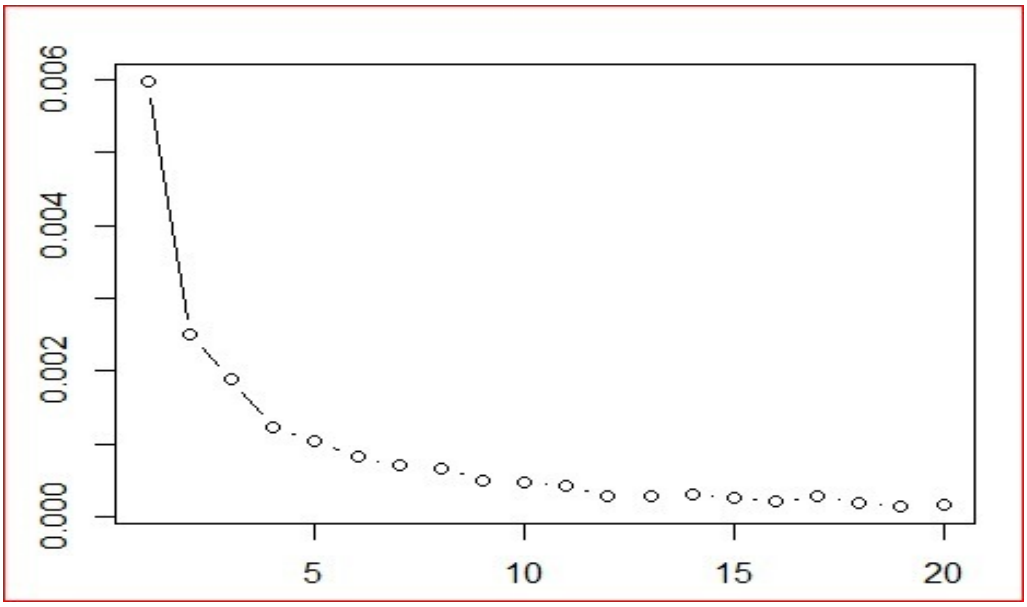
$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$

Al obtener las valoraciones de la inercia tras aplicar el K-means de 1 a N Clusters, se representa gráficamente y lineal la inercia en relación de los números de Clusters. En la gráfica se debe apreciar una gran transformación en la evolución de la inercia, obteniendo una línea representada de una similar forma a la de un codo. En este punto se observa una muy grande transformación en la inercia se expresa el número óptimo de Clusters a escoger para el data set; en otras palabras: el punto que representa al codo del brazo es el número óptimo de Clusters para esa data set.

En la siguiente grafica se mostrarán los resultados obtenidos para las tres diferentes datas sets. El script solo devuelve la gráfica lineal. los Clusters que se muestran a un lado de la gráfica lineal ha adquirido implementa el EM en el script y se mostrara para lograr visualizar el número de Clusters que se señala una coherencia en el proceso del codo:

Figura 40.

Metodología de codo



Nota: la figura muestra como la gráfica se desvía para formar una curva que se asemeja a un codo. (Chapman, 2000)

En la figura 40 se puede identificar el decrecimiento del punto k, el cual que inicia desde el punto 4 y continua de forma pronunciada hasta el punto 12; y, de ahí en adelante no existe mayor incidencia. Por tal razón, mediante este método se ha identificado un rango en la cual se podrá verificar el número de k más óptimo para ser utilizado.

Algoritmo implementado

Se probaron algunos algoritmos de cauterización configurando diferentes hierparámetros y se pudo definir que el de mejor rendimiento para el caso particular de esta investigación se utilizó a SimpleKMeans con una distancia Manhattan y con 5 clúster los que se

implementaron en conjunto con la técnica del codo y con una inicialización del método random obteniendo los siguientes resultados:

Cluster 0: 2203,15,1,-0.197585,-78.500295,1,33,1.666132
Cluster 1: 630,20,2,-0.19843,-78.512439,1,33,2
Cluster 2: 1030,15,1,-0.20262,-78.499973,2,38,1.666132
Cluster 3: 1400,20,2,-0.202925,-78.504489,8,61,2
Cluster 4: 1730,14,1,-0.196825,-78.499754,10,30,1.666132

Estos son resultados que una vez configurado los hiperparámetros (número de k, algoritmo de distancia, algoritmo iterativo de inicialización), arroja el algoritmo KMeans, que para la investigacion se trata de los 5 centroides de los cluster generados, que serán utilizados para las rutas de patrullaje y predicciones.

Se presenta el algoritmo de K-Means con K=5 que ha sido agrupado con el filtro de rutas y sus registros, teniendo las dimensiones y sus valores en la siguiente tabla:

Tabla 11.

Atributos del dataset

Attribute	**Full Data**	**0**	**1**	**2**	**3**	**4**
	(2092.0)	(444.0)	(469.0)	(479.0)	(188.0)	(512.0)
hora_infraccion	1420	1933	100	1000	1405	1347.5
modalidad	16	15	20	156	20	20

movilidad_victimario	2	2	2	2	2	2
Latitud	-0.198	-0.1999	-0.1958	-0.1999	-0.1999	-0.1964
Longitud	-78.502	-78.4995	-78.5088	-78.4995	-78.5019	-78.5063
categoria	1	1	1	1	8	11
subcategoria	33	33	33	33	61	34
sexo	1.6661	1.6661	1.6661	1.6661	1.6661	1.6661

Nota: *se muestra en la tabla la selección de los campos con cada uno de sus valores para los clústers. elaboración propia.*

Una de las formas más claras de analizar los datos que me arroja el algoritmo KMeans, es la presentación de la tabla 11, donde se puede identificar las instancias generadas por el algoritmo, también se puede ver los nombres de los atributos, lo que ayuda a identificar de forma inequívoca las nuevas instancias de los 5 centroides que forma cada clúster.

Cuando se utiliza un algoritmo para clusterizar datos es recomendable normalizar los campos. En este contexto, normalizar se define a poner la información con valores semejantes a una escala igual. Este primero paso ayuda al clúster ya que implementa grupos de campos a partir de las instancias originales y si encuentra atributos con escalas de diferentes valores, el de mayor rango tendrá preferencia ante los demás.

Dentro del análisis de los datos del proyecto, se va a agrupar los datos para que sea más útil en aplicaciones tales como segmentación distancias, esto reducirá la perdida de datos por el agrupamiento de entidades con comportamientos similares.

Tabla 12.

Resultados del clúster

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
0,991	**0,004**	**0,987**	**0,991**	**0,989**	**0,986**	**0,999**	**0,996**	**cluster0**
0,981	**0,003**	**0,989**	**0,981**	**0,985**	**0,981**	**0,997**	**0,992**	**cluster1**
0,992	**0,006**	**0,981**	**0,992**	**0,987**	**0,982**	**0,998**	**0,989**	**cluster2**
0,979	**0,000**	**1,000**	**0,979**	**0,989**	**0,988**	**1,000**	**0,998**	**cluster3**
0,994	**0,003**	**0,992**	**0,994**	**0,993**	**0,991**	**1,000**	**0,998**	**cluster4**
0,989	**0,003**	**0,989**	**0,989**	**0,989**	**0,985**	**0,999**	**0,994**	**Weighted Avg.**

*Nota: en la tabla se muestra el resultado de cluster con el campo extra llamado "class". (*elaboración propia, 2020)

En la tabla 12 se determina un análisis general de los resultados, vasados en ciertas métricas que arroja la herramienta y se toma como ejemplo al cluster0 como referencia.

El TP Rate: indica la tasa de verdaderos positivos, o las instancias correctamente clasificados, que para este ejemplo es de 0,991

El FP Rate: indica la tasa los falsos positivos o instancias falsamente clasificadas que para el ejemplo es 0,004

Precisión: da la proporción que verdaderamente pertenecen a un determinado clouster que para el ejemplo es de 0,987

Recall: pasa a ser un equivalemte a la tasa TP, que para el ejemplo es 0,991 por ende la misma medida

F-Measure: es una medida combinada de precisión y recuperación, que para el ejemplo es 0,989

MCC: se utiliza como medida de calidad de una clasificación tomando en cuenta los falsos positivos y los verdaderos positivos, para el ejemplo es 0,986.

ROC Área: una medida que sirve para verificar como está funcionando la clasificación, que para el ejemplo es 0,999

PRC Área: pasa más de información pero que igual sirve para verificar como está el algoritmo, que para el ejemplo es 0,996

Como se ha podido identificar en la descripción de cada una de las medidas, se deja claro que las conclusiones del algoritmo se encuentran correctamente ejecutadas, por, lo sé que obtiene datos considerados relevantes para la predicción de las rutas.

Clustered Instances

0 444 (21%)
1 469 (22%)
2 512 (24%)
3 188 (9%)
4 479 (23%)

En los porcentajes que se puede identificar anteriormente, se verificar de la cantidad de instancias que integra cada uno de los clusters en el 100%.

Con la finalización de este primer paso, se ha identificado los clusters con los que se trabajará, por lo que se procede a generar un primer modelo, el mismo que con la utilización de los centroides y los atributos latitud y longitud proporcionados, se procede a identificar y

predecir los puntos por los que debe existir el patrullaje preventivo, con la finalidad de disminuir el índice delictual existente en ese punto.

Una vez clusterizado se procede con un segundo paso, que sería identificar un nuevo modelo que prediga o clasifique los posibles delitos basados en la información de cada clusters, por lo que se recuperan los datos en un nuevo set de datos, pero esta vez ya se posee un atributo esencial denominado class, que servirá para poder realizar un experimento con los mejores algoritmos y así inicializar las predicciones. Se muestra el análisis resultante del experimento mediante la siguiente tabla:

Tabla 13.

Análisis y experimentación para la selección del mejor algoritmo basado en un mismo set de datos

Algoritmo	Porcentaje	Comparación A/B	Selección
lazy.IBk	88.33%	Algoritmo base	NO
rules.ZeroR	24.47%	- relevante	NO
functions.SMO	93.80%	Significativamente relevante	NO
functions.Logistic	94.25%	Significativamente relevante	NO
bayes.NaiveBayes	89.47	No relevante	NO
trees.J48	96.64	Significativamente relevante	SI

*Nota: en la tabla se realiza la comparación de diferentes algoritmos para seleccionar el que se acople al dataset. (*Elaboración propia, 2020)

Para la búsqueda del mejor algoritmo se lo realizó experimentando con seis diferentes tipos de algoritmos en cualidades y configuraciones, como consta en la tabla 13, con la finalidad de ubicar el algoritmo que mejor resultados arroje, y así utilizarlo en nuestras predicciones.

Por lo que para este experimento se utilizó seis algoritmos diferentes tomando como base el algoritmo lazy. IBk lo que implica que el porcentaje de acierto no puede ser menor a este, al realizar la experimentación en un mismo set de datos se identificó que solo uno de estos algoritmos es sumamente inferior, y por lo tanto el menos relevante, así mismo uno no relevante y tres significativamente relevantes, entre los cuales se encuentra el algoritmo con el mejor resultado, el mismo que se seleccionó para utilizarlo. Una vez que se identificó el mejor algoritmo a utilizar, que, para el caso del presente proyecto fue J48, un algoritmo para clasificación por medio de árboles de decisión, y que al utilizarlo y entrenarlo se obtuvo los siguientes resultados:

Tabla 14.

Análisis deJ48

Correctly Classified Instances:	98.8528 %
Incorrectly Classified Instances	1.1472 %
Kappa statistic	0.9854%
Mean absolute error	0.0084%
Root mean squared error	0.065%
Relative absolute error	2.6908 %
Root relative squared error	16.4041 %
Total Number of Instances	2092

*Nota: en la tabla se realiza la comparación de diferentes algoritmos para seleccionar el que se acople al dataset. (E*laboración propia, 2020)

Los resultados obtenidos en la tabla 14, una vez que entreno el modelo con el algoritmo J48, denotan una mejora en comparación al experimento de búsqueda del mejor algoritmo detallado en la tabla 13, de 2,21%, en donde se obtuvo 96,64% de correcta clasificación, mientras tanto, entrenado el modelo se obtiene un 98,85%, como se puede identificar en la

tabla en mención, además de otras variables que demuestran la eficiencia de los resultados, como el error medio absoluto que se encuentra en un 0,0084%.

Una vez verificado los resultados del entrenamiento del algoritmo seleccionado, se procede a obtener el modelo que servirá para las predicciones. Para ello se separa un set de datos con resultados que no intervinieron en el entrenamiento del modelo, por lo que estos datos serán nuevos para el modelo al momento de las pruebas. Se describe el set de datos de prueba:

@attribute hora_infraccion numeric

@attribute modalidad numeric

@attribute movilidad_victimario numeric

@attribute latitud numeric

@attribute longitud numeric

@attribute categoria numeric

@attribute subcategoria numeric

@attribute sexo numeric

@attribute class {cluster0,cluster1,cluster2,cluster3,cluster4}

@data

800,16,2,-0.197758,-78.499986,10,30,1.666132,cluster4

945,20,2,-0.199221,-78.499657,2,19,1.666132,cluster2

1949,15,1,-0.199946,-78.500714,1,33,1,cluster0

2203,20,2,-0.196498,-78.502899,1,33,1,cluster1

1600,20,2,-0.196238,-78.503891,1,33,1,cluster1

1715,20,2,-0.197883,-78.498198,1,33,1.666132,cluster0

1320,15,2,-0.1997,-78.49678,10,29,1.666132,cluster4

240,15,4,-0.197149,-78.497837,1,33,1.666132,cluster2

Para un correcto entendimiento de este data set, que se encuentra tratado y clusterizado, es recomendable dar unos cuantos pasos atrás en donde la información no se encuentra tratada por ejemplo.

hora_infraccion: 15:00

modalidad: ESTRUCHE

movilidad_victimario: A PIE

latitud: -0.191371

longitud: -78.510253

categoría: VIVIENDAS/ALOJAMIENTO

subcategoría: CASA/VILLA

sexo:MASCULINO

*class:*cluster3

Este es el tipo de información que se encuentra representada en el data set anterior pero tratada de forma numérica, con una excepción de el atributo clase que ayuda a identificar la predicción en el siguiente paso.

Estos son los datos englobados en el data set original con el cual se realizará las pruebas de predicción del modelo ya entrenado. Previo a las pruebas se deben sustituir los resultados

del atributo clase(*cluster0, cluster1, cluster2, cluster3, cluster4*) por el signo de pregunta teniendo el siguiente resultado:

@attribute hora_infraccion numeric

@attribute modalidad numeric

@attribute movilidad_victimario numeric

@attribute latitud numeric

@attribute longitud numeric

@attribute categoria numeric

@attribute subcategoria numeric

@attribute sexo numeric

@attribute class {cluster0,cluster1,cluster2,cluster3,cluster4}

@data

800,16,2,-0.197758,-78.499986,10,30,1.666132,?

945,20,2,-0.199221,-78.499657,2,19,1.666132,?

1949,15,1,-0.199946,-78.500714,1,33,1,?

2203,20,2,-0.196498,-78.502899,1,33,1,?

1600,20,2,-0.196238,-78.503891,1,33,1,?

1715,20,2,-0.197883,-78.498198,1,33,1.666132,?

1320,15,2,-0.1997,-78.49678,10,29,1.666132,?

240,15,4,-0.197149,-78.497837,1,33,1.666132,?

Como se ha explicado anteriormente, estos datos se encuentran tratados, con la diferencia que el atributo clase, no existe para este set, puesto que es lo que se busca predecir.

El data set resultante es con el cual se trabajaran las pruebas, el mismo que consta de los siguientes atributos: hora de la infracción, modalidad del delito utilizada al momento del hecho, movilidad del victimario, latitud, longitud, categoría del lugar donde se comete el delito, subcategoría del lugar donde se comete el delito, y el sexo de la persona víctima del delito; además se sustituyó el atributo clase por un signo de pregunta porque ese es el tipo de formato que reconoce weka y que lo tomará para dar la predicción correspondiente. Se procede a cargar el modelo entrenado y se realizan las pruebas, para lo cual se carga un set de datos con una muestra de 8 instancias sin resultados en la clase, procediendo a ejecutar las pruebas como se detalla a continuación:

Tabla 15.

Pruebas del modelo con el dataset y sus resultados

inst#	actual	predicted	error	prediction
1	1:?	5:cluster4		0.977
2	1:?	3:cluster2		0.99
3	1:?	1:cluster0		0.997
4	1:?	2:cluster1		0.974
5	1:?	2:cluster1		0.974
6	1:?	1:cluster0		0.997
7	1:?	5:cluster4		0.977
8	1:?	3:cluster2		0.99

Nota: la tabla representa a los valores de error que contiene cada uno de los clusters. (Elaboración propia, 2020)

Como se puede observar en la tabla 15, las pruebas realizadas al modelo que se entrenó y creo, se ingresan ocho instancias a predecir. Los mismos que se encuentran detallados en el set de datos de prueba anteriormente indicado, arrojando una salida con sus respectivas predicciones en un rango de aceptación bastante alto, las mismas que varía entre 0.97% a 0.997%, comprobando así que nuestro modelo se encuentra prediciendo de forma correcta nuevos posibles delitos.

Elaboración de informe

La minería de datos es el método más extraordinario en la actualidad, ya que se encarga de quitar la trivialidad de los patrones ocultos y útiles que habitan los datos y su forma más rápido de investigar grandes niveles de información.

Las dos razones sustentaron esta justificación ya que estas técnicas de análisis fueran utilizadas dentro del proyecto de maestría: *Automatización de rutas de patrullaje basados en modelos dinámicos y predictivos apoyados en el análisis de información delictual,* con el empleo de la Minería de Datos y de igual modo se conforma el marco teórico, se decidió indagar en varios resultados obtenidos con las herramientas seleccionadas.

A pesar que el proceso de análisis de la minería de datos tiene estudios en ramas económicas hace algunos años en ámbitos internacionales, para el territorio policial del Ecuador, es una novedosa forma de análisis, por primera vez se está siendo aplicada en este sector y con sus resultados se ha adquirido la utilización de nuevos métodos en relación al patrullaje obteniendo un resultado con una gran disminución en posibles casos delictivos,

también la nueva calidad de la información de la base de datos ha permitido un mejoramiento funcional del Sistema Gestor de Información.

Para obtener estos resultados se realizaron en varios procesos de minería de datos a la información bibliográfica y se encontraron varios patrones, que permitirá realizar mejoras en las rutas originales de patrullaje e inclusive dejaron abierta la posibilidad de hacer otras investigaciones futuras. (Rueda-Clausen, Villa-Roel, & Rueda-Clausen, 2005)

Materiales y métodos

Dentro de los materiales empleados se puede hallar la base de datos con información bibliográfica, este respaldo de datos recoge campos como fecha, descinsivili, latitudinsivili – longitudinsivili, usuarioinsivili, anioinsivili – mesinsivili –diainsivili, nomruta, desmovilidad.

El instrumento digital seleccionado para realizar las técnicas de minería de datos fue el software Weka por la calidad de resultados, gráficas y además que ofrece las posibilidades de adjuntarle extensiones, que aumentaran las prestaciones digitales que ofrece el software en su forma original.

El método utilizado para la aplicación del proceso de minería de datos es CRISP-DM y se enumera sus seis procesos fundamentales, las cuales se ordenan de la siguiente manera según su autor (Cabena, 1998):

1) Determinación de los Objetivos o comprensión del comercio.
2) Preparación o comprensión de datos.

   a. Selección: Reconocer las diferentes fuentes de información internas y externas seleccionando el subconjunto de datos necesarios.
   b. Pre procesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.
   c. Transformación de datos: conversión de datos en un modelo analítico.

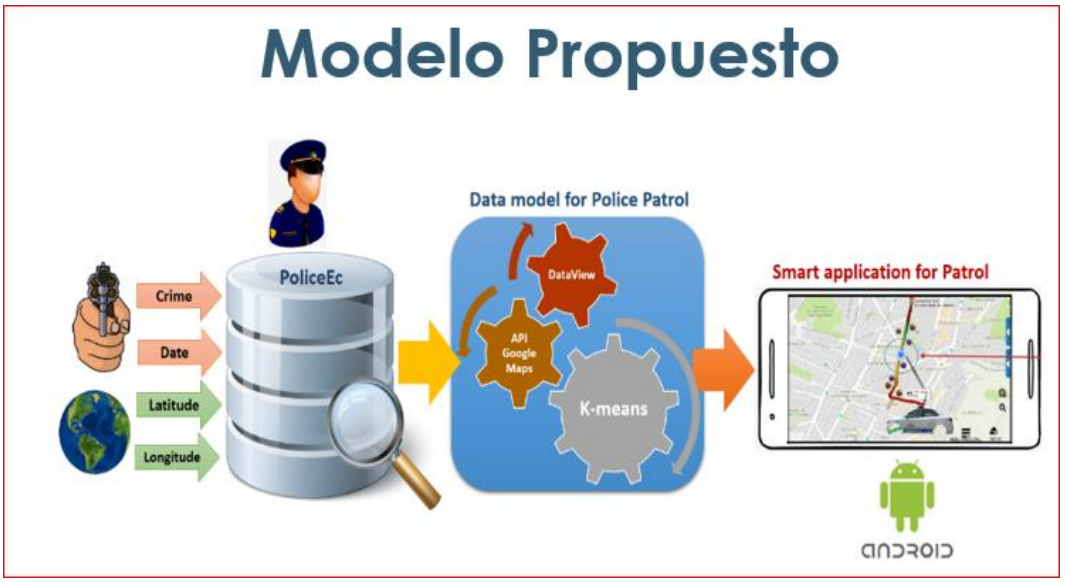
3) Organización o análisis de datos.
   a. Selección de campos o registros importantes y eliminación de información irrelevante.
4) Moldeado o minería de datos
   a. Tratamiento automatizado de los datos identificados con la combinación apropiada de algoritmos.
5) Pruebas de modelos y análisis de resultados.
   a. Interpretación de los resultados obtenidos en la anterior fase, generalmente con la ayuda de un procedimiento visual.
6) Despliegue, implementación. discusión, conclusiones y recomendaciones.
   a. Aplicación del conocimiento descubierto.

Resultados obtenidos

El proceso que se realiza para obtener los datos requeridos, son determinados por el recorrido lineal del método propuesto para adaptar los datos de la base original a un dataset que contenga campos específicos que sirvan como punto de partida para obtener mejores rutas de patrullaje.

Figura 41.

Modelo de minería de datos

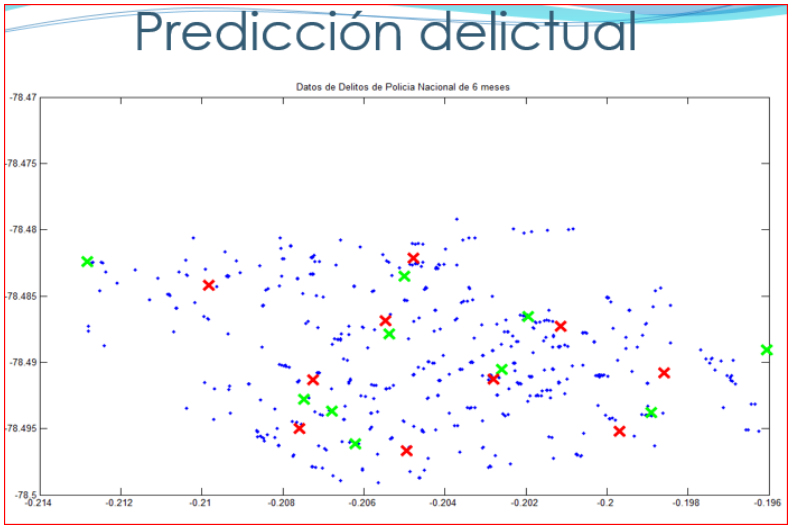


Nota: la figura representa al modelo que se propone para el traslado de minería de datos. ; (María Pérez Marqués, 2016)

Después de aplicar las diferentes técnicas de minería de datos, uno de los resultados más interesante logrados, son las salidas de predicción delincuencial, porque utilizan las técnicas de Clasificación, además de ser procesos de autoaprendizaje (Madrid, 2009), razón del porque cada resultado es diferente al anterior, también crea una salida en forma de gráfico, con todos los metadatos que entraron al análisis.

Figura 42.

Predicción delictual

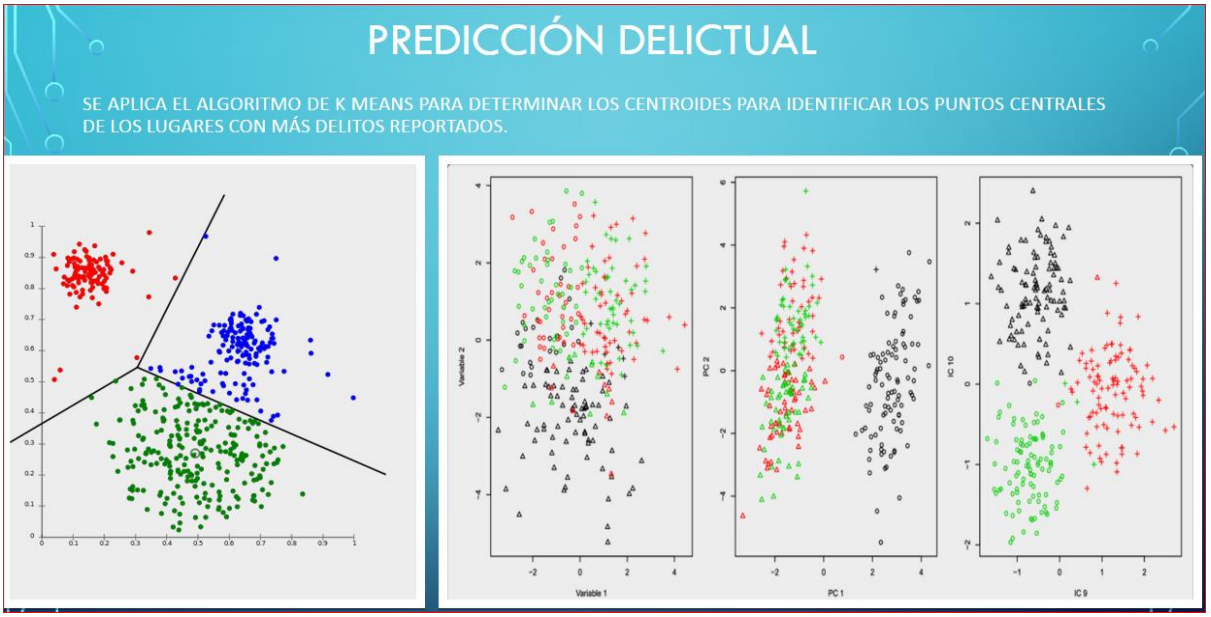


Nota: la figura representa la predicción de las rutas de patrullaje. (Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo y Edison Manjarres, 2020)

Este tipo de predicción se consideran dentro de los procesos supervisados (Gutiérrez Rodríguez, 2012), pero el algoritmo requiere conocer a la variable independiente, en este caso la variable fue casos potenciales de delitos dentro del sector, para a partir de ahí comenzar con la clusterización de las palabras claves y el aprendizaje, por eso se aprecia en la figura los cuatros segmentos de la ciencia, cada uno como una rama variable. Para este proceso se tomó una muestra de 396 registros.

Figura 43.

Variables de predicción delictual



Nota: La figura muestra cada dimensión de la predicción que tienen las rutas. (Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo y Edison Manjarres, 2020)

Una vez realizadas las pruebas se pueden identificar los resultados de la predicción mediante la siguiente tabla comparativa:

Tabla 16.

Comparativa de resultados

	Ingreso de datos sin resultado									Salida de predicción		
	Set de datos sin resultados									Orden	Predicción	Error Predicción
1	800	16	2	-0.197758	-78.499.986	10	30	1.666.132	?	1	cluster4	0.977
2	945	20	2	-0.199221	-78.499.657	2	19	1.666.132	?	2	cluster2	0.99
3	1949	15	1	-0.199946	-78.500.714	1	33	1	?	3	cluster0	0.997
4	2203	20	2	-0.196498	-78.502.899	1	33	1	?	4	cluster1	0.974
5	1600	20	2	-0.196238	-78.503.891	1	33	1	?	5	cluster1	0.974
6	1715	20	2	-0.197883	-78.498.198	1	33	1.666.132	?	6	cluster0	0.997
7	1320	15	2	-0.1997	-7.849.678	10	29	1.666.132	?	7	cluster4	0.977
8	240	15	4	-0.197149	-78.497.837	1	33	1.666.132	?	8	cluster2	0.99

Nota: en la tabla se observa la comparativa entre los datos de ingreso con la predicción de cada cluster. (Elaboración propia, 2020)

En la tabla 16 se puede identificar una comparativa entre, la entrada y salida, donde se ingresan ocho instancias, las cuales se son el rango óptimo para un análisis; cada uno de estos instancias consta de con nueve atributos, haciendo referencia a los campos de la base de datos filtrados y seleccionados como relevante, también se crea el atributo clase que es remplazado con el signo de pregunta, ya que es lo que se busca predecir, y como salida se obtiene el cluster al que fue clasificado con su respectivo porcentaje de predicción.

Tabla 17.

Datos finales

Datos Originales										Salida de predicción		
Set de datos sin resultados										Orden	Predicción	Error Predicción
1	800	16	2	-0.197758	-78.499.986	10	30	1.666.132	**cluster4**	1	**cluster4**	0.977
2	945	20	2	-0.199221	-78.499.657	2	19	1.666.132	**cluster2**	2	**cluster2**	0.99
3	1949	15	1	-0.199946	-78.500.714	1	33	1	**cluster0**	3	**cluster0**	0.997
4	2203	20	2	-0.196498	-78.502.899	1	33	1	**cluster1**	4	**cluster1**	0.974
5	1600	20	2	-0.196238	-78.503.891	1	33	1	**cluster1**	5	**cluster1**	0.974
6	1715	20	2	-0.197883	-78.498.198	1	33	1.666.132	**cluster0**	6	**cluster0**	0.997
7	1320	15	2	-0.1997	-7.849.678	10	29	1.666.132	**cluster4**	7	**cluster4**	0.977
8	240	15	4	-0.197149	-78.497.837	1	33	1.666.132	cluster2	8	cluster2	0.99

Nota: En la tabla se puede identificar la comparativa de la predicción obteniendo con un 100% de aciertos. (Elaboración propia, 2020)

En la tabla 17 se puede realizar la misma comparativa que en la tabla 16, pero con la diferencia que las ocho instancias de comparación se lo hace con la data y los resultados originales de la clusterización sin sustituir el atributo clase, y se compara con la salida de las predicciones arrojadas, lo que ayuda a identificar de forma clara que el porcentaje de perdición concuerda y afirma una correcta clasificación del modelo. Se observa que el atributo que contenía el símbolo de interrogación es remplazado por el cluster analizados, dejando la tabla lista para un depurado de datos en análisis posteriores.

En conclusión, el algoritmo seleccionado j48 predice con un alto grado de exactitud y los 5 clúster que son proporcionados por el algoritmo KMeans realizan dos significativos resultados, los cuales son:

1.- Los centroides los cuales se utilizan los campos latitud y longitud para proporcionar los lugares de patrullaje que se deben tomar en cuenta según lo proporcionado en el algoritmo no supervisado

2.- Prevé un nuevo atributo clase para poder utilizar unos algoritmos de predicción por medio de clasificaciones para lo que después de un fuerte análisis se optó por J48.

Comparación de resultados

Una vez determinado el algoritmo en el data set, se implementa las predicciones en las rutas reales para analizar los resultados en un periodo de tiempo desde enero del 2020 hasta septiembre del mismo año, con el fin de tener un porcentaje exacto de disminución delincuencial en la zona de patrullaje.

Tabla 18.

Comparación de delitos

Delitos	2019	2020	Diferencia	Porcentaje
Robo a personas	113	85	-28	-24.78%
Robo de bienes y accesorios	55	22	-33	-60.00%
Robo a domicilio	15	12	-3	-20.00%
Robo a unidades económicas	14	11	-3	-21.47%
Robo a motos	4	11	7	+175.00%
Robo a carros	17	9	-8	-47.06%
TOTAL	**218**	**150**	**-68**	**-31.20%**

Nota: en la tabla se puede observar los datos antes y después de la puesta en marcha del algoritmo en las rutas de patrullaje (Dirección General De Operaciones, 2020).

Los datos presentados como resultado demuestran la efectividad de la implementación de un algoritmo para el cambio de rutas de patrullaje, siendo una mejora de un 31.20% en disminución de los casos delictivos a comparación con las rutas anteriores. Se puede concluir el éxito en el cambio de rutas por la notoria disminución de los mismos.

Capítulo iv

Discusión de resultados

Delimitación de los casos de evaluación

Este proyecto es de carácter experimental porque tiene un mayor énfasis y atención a la formulación y tratamiento de la hipótesis mediante procedimientos básicamente educativos, orientados en el procesamiento de datos informativos y los parámetros bajo los cuales esta establecidos el desarrollo del proyecto. Además, se toma énfasis en la obtención de datos vía encuestas al personal de seguridad que hace el patrullaje por la zona, delimitando los casos a una muestra para determinar si la implementación del proyecto tuvo los resultados deseados.

Población

La población se refiere al grupo de incidencias dentro del objeto del cual se va a hacer referencia para poder sacar los datos necesarios a ser analizados, para esto se muestra en la Tabla 19 las características de este grupo seleccionado.

El universo al que se hace referencia sobre la población de estudio, estará dirigido a una sesión de 100 pruebas realizadas en un entorno real con el prototipo desarrollado, donde se hará un sondeo de tiempo de respuesta sobre las rutas de patrullaje de la zona, con un registro de casos positivos y posibles negativos.

Se analizarán en diferentes horas del día para poder tener un margen de datos más amplio y poder ofrecer un análisis que abarque todas las posibles incidencias delincuenciales. Teniendo la variación de posibles sospechas según los parámetros presentados en el estudio.

Muestra

Tabla 19.

Muestra de estudio

POBLACIÓN Y MUESTRA	
Grupo Objetivo	Incidencias sobre rutas de patrullaje
Zona Geográfica	Distrito metropolitano de Quito – Zona central
Método de Investigación	Inductivo
Metodología	Observación y documentación
Tipo de recolección de datos	Personal
Zona objetiva a implementar en el análisis de resultados	100 incidencias en varias horas del día.

Nota: datos informativos para tomar la muestra que será utilizada en las encuestas. (Elaboración propia, 2020)

Técnica para análisis de datos

Es necesario dar un análisis completo a los resultados que se pretenden conseguir dentro del proyecto, realizando un seguimiento del prototipo en diversas ramas, lo cual dará datos que se pueden procesar y sacar un análisis completo que ayudará en concluir con los objetivos presentados. Por ello, en la Tabla 20 se presentan las siguientes preguntas a realizar cuando se esté en el capítulo de análisis de resultados:

Tabla 20.

Preguntas de investigación

Pregunta 1: ¿Los resultados presentados después del filtro de la minería de datos fueron satisfactorios?
Pregunta 2: ¿Se implementó de manera óptima el cambio de rutas de patrullaje?
Pregunta 3: ¿ Se ha encontrado el número de clusters adecuado para aplicar el algoritmo de minería de datos?
Pregunta 4: ¿Se ha tenido inconvenientes en encontrar posibles delitos dentro de la nueva ruta de patrullaje?
Pregunta 5: ¿La zona central se ha visto afectada a la hora de modificar las rutas de patrullaje?
Pregunta 6: ¿Se ha reducido la incidencia de delitos en las zonas de patrullaje?

Nota: se describen cada una de las preguntas que serán realizadas en las encuestas. (Elaboración propia, 2020)

Cada pregunta dentro de la encuesta tiene la finalidad de análisis y conocer si la investigación está correctamente encaminada al objetivo principal, estos objetivos se los presenta en la Tabla 21.

Tabla 21.

Objetivos de preguntas

Pregunta 1: En primera instancia se debe medir si los resultados recibidos después de hacer el filtro con los métodos de minería de datos son óptimos, para que se pueda realizar un cambio satisfactorio en las rutas de patrullaje.
Pregunta 2: En los diferentes distritos que utilizan rutas de patrullaje por un prolongado tiempo, el cambio de las mismas puede afectar en los resultados finales. Por ello es importante medir si las nuevas rutas se han implementado sin contratiempos.

Pregunta 3: Identificar el número preciso de clusters para la centralización de las nuevas rutas, es importante para poder aplicar el algoritmo de minería de datos y poder tener los resultados esperados. **Pregunta 4:** Al presentar una nueva ruta, es posible que en inicio no se presenten resultados favorables inmediatos, por ello se requiere estudiar los posibles inconvenientes. **Pregunta 5:** Las personas que transitan y residen dentro de una zona en específica, ya tienen dentro de su rutina un tipo constante de ruta de patrulleros. Al verse modificada la ruta, es posible que se presenten inconvenientes dentro de la zona. **Pregunta 6:** El resultado final es el poder disminuir en algún grado los índices delictivos dentro de la zona donde se aplica el estudio.

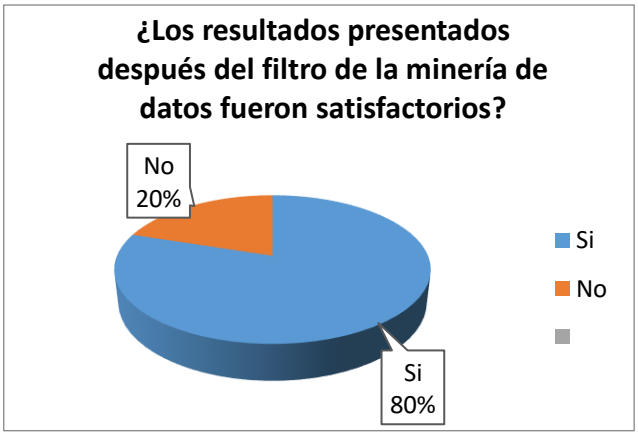
Nota: cada una de las preguntas de la encuesta está relacionada con un objetivo que se asigna para poder tener la información necesaria y realizar un análisis completo de los resultados del proyecto. (Elaboración propia, 2020)

Análisis de Resultados

Pregunta 1: ¿Los resultados presentados después del filtro de la minería de datos fueron satisfactorios?

Gráfico 1.

Pregunta 1



Si	80	80%
No	20	20%
TOTAL	100	100%

Fuente: (Elaboración propia, 2020)

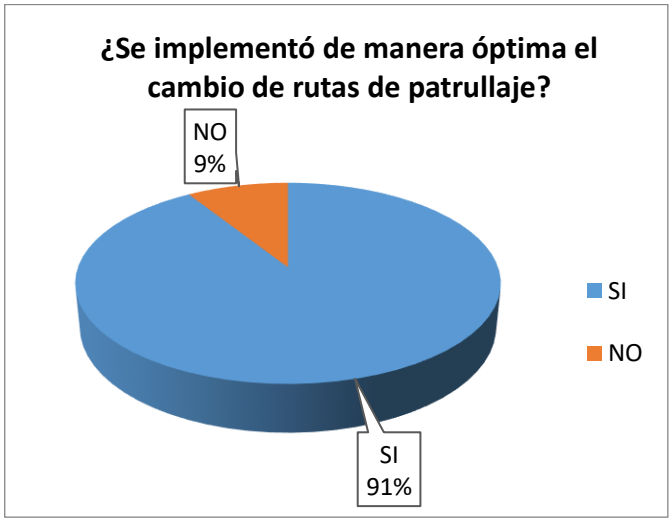
Análisis:

Para el desarrollo del proyecto existió la necesidad de medir la adaptabilidad de los datos filtrados, analizando si los campos seleccionados cumplen con los requerimientos para poder establecer una nueva ruta o si, por el contrario; los campos contienen información innecesaria que ralentice el proceso. Los datos mostrados en el Gráfico 1, demuestran que en un 80% de los casos, el procesamiento de minar los datos, pasando por los diferentes filtros y aplicando diversos algoritmos y métodos, se realizó de una manera óptima. Por otro lado, se puede ver que un 20% de las muestras han presentado algún tipo de fallo debido a campos vacíos, duplicidad de datos o información errónea dentro de los registros.

Pregunta 2: ¿Se implementó de manera óptima el cambio de rutas de patrullaje?

Gráfico 2.

Pregunta 2



SI	91	91%
NO	9	9%
TOTAL	100	100%

Fuente: (Elaboración propia, 2020)

Análisis:

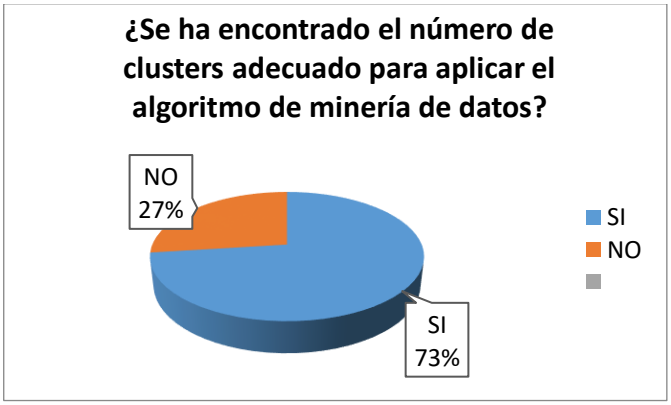
El objetivo principal del proyecto presentado, es el de implementar un cambio en las rutas de patrullaje para poder disminuir posibles casos delincuenciales, por lo tanto, al saber si las nuevas rutas se han implementado de manera estable y con bajos niveles de riesgo, se puede proceder al estudio de posibles casos.

Por ello se determinó si el cambio de las rutas ha conllevado algún tipo de inconveniente. En el Gráfico 2 se pueden ver que un 91% de los casos presentados, da como respuesta que las rutas se han implementado sin ningún tipo de novedad; en contra parte, tan solo un 9% de casos, han resultado con complejidades por la falta de conocimiento en nuevas rutas, las cuales serán solventadas con un constante recorrido de las mismas.

Pregunta 3: ¿Se ha encontrado el número de clusters adecuado para aplicar el algoritmo de minería de datos**?**

Gráfico 3.

Pregunta 3



SI	73	73%
NO	27	27%
TOTAL	100	100%

Fuente: (Elaboración propia, 2020)

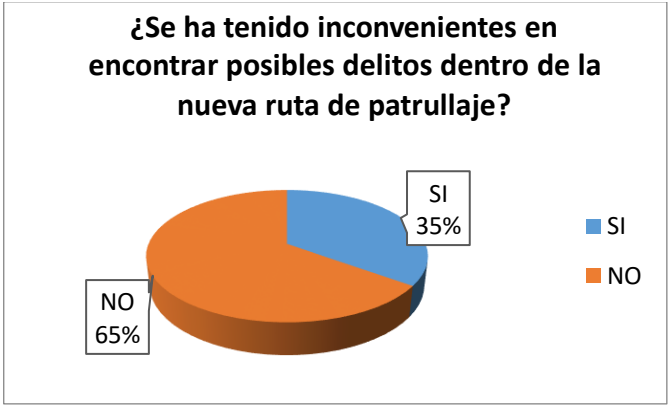
Análisis:

La pregunta se direcciona hacia el cambio en las rutas de patrullaje mediante métodos de minería de datos, acopla diferentes patrones para poder seleccionar el punto centralizado de cada ruta. Estos patrones dependen de los cluster seleccionados. Aún con una correcta instalación de módulos y algoritmos que filtren los datos, el análisis del número de los centroides que manejaran los clusters se debe adaptar para potenciar el manejo de resultados óptimos. Se puede observar en el Gráfico 3 que un 27% de los casos en los que se ha adaptado un número de cluster, han resultado con errores o insolvencias en el proceso de minar los datos. En contra parte, un gran porcentaje, el cual es del 73% de casos, un número entre 5 a 10 clusters da un buen resultado.

Pregunta 4: ¿Se ha tenido inconvenientes en encontrar posibles delitos dentro de la nueva ruta de patrullaje?

Gráfico 4.

Pregunta 4



SI	35	35%
NO	65	65%
TOTAL	100	100%

Fuente: (Elaboración propia, 2020)

Análisis:

Para un adecuado nivel de detección de posibles casos delictivos, es necesario tener una respuesta inmediata en los datos históricos recibidos por los patrulleros. En las diversas pruebas que se realizaron, uno de los inconvenientes fue la adaptación del personal de patrulla a una nueva ruta y la falta de conocimiento de calles, avenidas y horarios picos dentro de la zona. Después de algunas pruebas y con el conocimiento adquirido, se pudo determinar en el Gráfico 4 que hay varios lugares óptimos para enfocar las rutas de patrullaje, por ello es que el 35% de los resultados fueron pruebas que resultaron no recomendables. Y, por otro lado, el restante 65% de los casos, ya tuvieron una respuesta mucho más precisa y satisfactoria que resulta en una adaptabilidad de la nueva ruta.

Pregunta 5: ¿La zona central se ha visto afectada a la hora de modificar las rutas de patrullaje?

Gráfico 5.

Pregunta 5

SI	10	10%
NO	90	90%
TOTAL	100	100%

Fuente: (Elaboración propia, 2020)

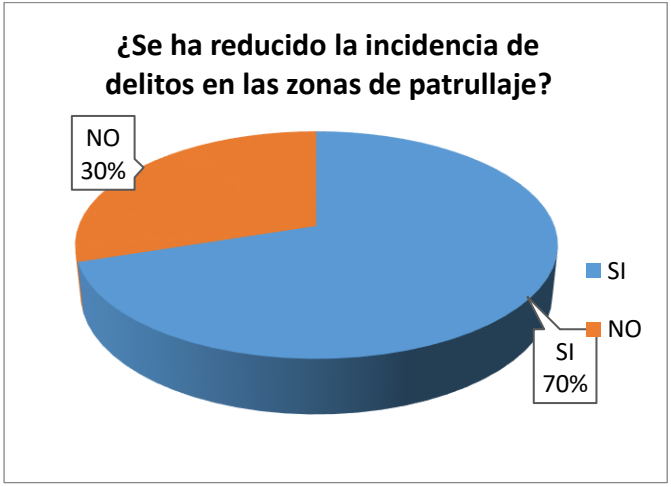
Análisis:

Después de haber encontrado los lugares más óptimos dentro de la nueva ruta y sin que interfiera con la movilidad de las personas. Se realizó un seguimiento para analizar si es adaptable en uso dentro de la zona y que tanto afecta a la población. Se pudo tener como resultados en el Gráfico 5 qué, una vez acostumbrado a la nueva ruta, la incomodidad tanto de la población como del oficial que realiza el patrullaje, se vio disminuida considerablemente, llegando casos en los que ni siquiera notaba que la ruta fue modificada; siendo el 90% de los casos en los que la zona no se vio afectada y un 10% en los que si sufrió algún tipo de incomodidad.

Pregunta 6: ¿Se ha reducido la incidencia de delitos en las zonas de patrullaje?

Gráfico 6.

Pregunta 6



SI	70	70%
NO	30	30%
TOTAL	100	100%

Fuente: (Elaboración propia, 2020)

Análisis:

El resultado final que se espera para la resolución del problema propuesto, es un decremento en los posibles casos o incidentes delincuenciales, aumentado así las posibilidades de evitar algún tipo de accidente o delito dentro de la zona central.

Para poder llegar a este último análisis del Gráfico 6, se tomó en cuenta el índice de casos reportados con las nuevas rutas en contra parte con los datos recibidos de las rutas antiguas. Se realizó un escaneo a patrulleros y su opinión con respecto a la disminución de casos. Dando como resultado que las nuevas rutas ponía en alerta a la población y disminuían así en un 70% los casos.

Discusión

En esta investigación se plantea el interés de realizar las técnicas de minería de datos a la información bibliográfica, para llegar a la finalidad de:

- Mejoramiento de la calidad de la información de la base de datos, para obtener rutas de patrullaje para la minimización de posibles casos criminales.
- Hallar patrones ocultos que ayuden al mejoramiento de las rutas originales.
- Mejorar la gestión del conocimiento y de la información.

Los resultados alcanzados luego de la realización de las técnicas de minería de datos, demuestran que la totalidad de los objetivos presentados se alcanzaron. Se puede decir que los diferentes patrones hallados mediante los Árboles de Decisión, la Matriz de Correlación, los algoritmos de Kmeans y la metodología de codo, pueden indicar que:

El proceso de desarrollo de minería de datos que alcanzan a los aspectos relacionados a la Preparación de datos, demostrado la selección de la sub agrupación de datos y la elección de los campos identificados son correctos para la aplicación de filtros de minería de datos.

Mientras ocurre el Pre procesamiento de la información, se obtendrá una mejora de la calidad de la información obtenida en la base de datos, mediante la estandarización de la información, que permite caracterizar la información que puede ser empleada, por un diferente sistema de gestión de información.

Los Árboles de Decisión demostraron que las temáticas más investigadas a nivel nacional, se ordena por segmentos campales. Se encuentra información propuesta por la instalación de un repositorio que ayuda a obtener nuevas rutas de servicio.

Capítulo v

Conclusiones

Conclusiones

- En el algoritmo de asociación se puede ver que se conforman ciertas reglas con ocurrencia e importancia similares, estas normas ayudan hallar patrones sólidos dentro de la base de datos con relación a las rutas de patrullaje y los delitos realizados. El algoritmo KMeans indica que tan determinante es una variable de latitud y longitud para resolver el clúster que se agrupa de acuerdo con los parámetros y asigna a cada una un nivel de ocurrencia.
- La base de datos se analiza y refina antes de realizar una técnica de minado o un proyecto de Data Mining. Por lo tanto, es indispensable clasificar los datos de la inicial data set para que pueda adaptarse a las variables de asociación y exploración que se presentan al aplicar un agrupamiento y análisis de minería de datos óptimo. Por ello es importante realizar el número de clusters que sea significativo para el grupo de registros, ya que los algoritmos que se emplean en la minería de datos deben de identificar las tendencias para indicar resultados.
- Al utilizar una metodología de minería de datos es muy importante determinar cuál algoritmo es el que se adapta a los requerimientos presentados por la propuesta. Esta selección se la toma realizando un análisis de la información y el procesamiento de los datos ingresados. Se debe escoger la metodología correcta para ampliar el sesgo de éxitos y disminuir lo más posible la frecuencia de error. Para el caso en particular de los

datos presentados en el data set final del proyecto, se determinó que la metodología con mayor rango de resultados es *SimpleKMeans,* por su estructura adaptable al distanciamiento *manhatan,* de igual manera se adaptó a 5 clusters los cuales fueron el resultado de la técnica de codo.

- Mediante los datos recogidos en un primer vistazo de la base de datos original de Policía Nacional, se pudo analizar que las rutas de patrullaje eran constantes y sin cambios significativos en periodos cortos de tiempo. Al ejecutar el análisis mediante la implementación de los métodos y algoritmos de minería de datos, se pudo observar que la ejecución de rutas de patrullaje mejoró los índices delictivos, tal cual se muestra en las encuestan y sus resultados.

Recomendaciones

- Se recomienda instaurar los resultados de manera inmediata en las rutas de patrullaje del sector, estableciendo los cambios de manera paulatina y con análisis periódicos para determinar su eficiencia a mediano y largo plazo, de la misma manera es recomendable realizar una clasificación por medio de correo electrónico para que el departamento administrativo que pueda establecer nuevas reglas en el cambio de rutas de patrullaje.
- Es recomendable instaurar un sistema de minería de datos en todo el entorno del sub centro, tomando como base este estudio y los parámetros de análisis, permitiendo de esa manera adaptar más rutas en un corto periodo de tiempo. Todos estos cambios se los debe guardar en un servidor de información externo al que se usa para las rutas, de esta manera ayudará a tener un registro histórico independiente de la base de datos central.

- Del mismo modo es recomendable realizar una investigación de acuerdo a reglas identificadas, tanto en mailing por correo electrónico, como mailing físico; esto para tener alerta a los directivos de patrullaje en cada una de las zonas y puedan responder de manera inmediata a cualquier tipo de emergencia dentro de su distrito. Para alcanzar este objetivo a futuro es indispensable tener datos almacenados sin procesamiento, para que los registros minados puedan dar un resultado más limpio y preciso dotando de información relevante y con la cual se podrá tomar decisiones acertadas e informadas.
- La presenta investigación toma como parámetros los registros de una base de datos en específico, que son las rutas de patrullaje, es recomendable expandir la investigación adaptando los conocimientos y análisis en otros departamentos que involucren casos delictivos, esto se puede realizar mediante la alimentación y procesamiento de dataset con los algoritmos trabajados y realizando pruebas para determinar el número de clusters necesarios para cada departamento en específico.

Trabajos futuros

Esta investigación en conjunto con la propuesta, son una herramienta que beneficiará a los distritos que cuentan con un sistema de patrullaje motorizado, teniendo la capacidad de modificar dichas rutas para poder disminuir los casos de delitos que se procesan dentro del perímetro de revisión. Por ello se plantea que en investigaciones futuras se dé prioridad a la realización de pruebas de este sistema en otros distritos y que se cambien paulatinamente la alimentación de datos para que el proceso de minera pueda actualizarse y adaptarse a los cambios de entorno que sufren las zonas en relación con la tasa de delincuencia.

Los datos filtrados por el proceso de minera se convierten en información relevante para tener en cuenta tanto en las rutas de patrullaje como en los diferentes parámetros en cuestión de delitos realizados en diversas zonas. Es por ello que esta investigación puede ser la base para implementar nuevos filtros de búsqueda y tener datos reales que puedan prevenir delitos en otras áreas, se puede expandir con una secuencia de pruebas en diversas zonas y distritos que manejen más personal, que cuenten con una infraestructura de mayor tamaño y puedan automatizar un algoritmo que sea cambiante en circunstancias de emergencia.

Referencias bibliográficas

Agrawal y Shim. (2015). *Mathematical and Computational Approaches in Advancing.*

By Gregory Piatetsky. (2017). *KDnuggets*. Obtenido de https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

Cadena Pinzón. (2011). *Analítica del Aprendizaje y la Minería de datos.* Medellin.

Cesar Guevara, Janio Jadán, César Zapata, Luis Martínez, Jairo Pozo y Edison Manjarres. (2020). Model of Dynamic Routes for Intelligent Police Patrolling. *MDPI*.

Chapman. (2000). *CRISP-DM methodology.*

Dirección General De Operaciones. (28 de Sep de 2020). *Policia del Ecuador*. Obtenido de https://www.policia.gob.ec/dgo/

Diseño de investigación no experimental extraído. (25 de Enero de 2017). Obtenido de http://es.slideshare.net/conejo920/diseo-de-investigacion-no-experimental

Engels y Theusinger . (2015). *Point of care devices for assessings bleending algorithms.*

Fayyad . (1996). *From Data Mining to Knowledge Discovery in Databases.*

Gallego, J. (2008). *Desarrollo Web con PHP y MySQL.* España: Anaya.

Harrendorf, Heiskanen, & Malby. (25 de Agosto de 2019). *ONU*. Obtenido de https://www.un.org/es/

Hormazi, A. M.,. (2014). *DATA MINING: A COMPETITIVE WEAPON FOR BANKING AND RETAIL INDUSTRIES.*

INEC. (18 de Diciembre de 2019). *ecuadorencifras*. Obtenido de https://www.ecuadorencifras.gob.ec/estadisticas/

Maimón y Rokach. (2016). *Minería de datos aplicada a la detección de la deserción en adolescentes infractores.*

María Pérez Marqués. (2016). *Minería de datos a través de ejemplos.*

Mark Hall. (2015). *The WEKA Data Mining Software: An Update.*

Martínez, M. (2005). *TRANSICIÓN DESDE LA DELINCUENCIA JUVENIL A LA DELINCUENCIA ADULTA.* Mexico.

Meng, J. L., Yang, Y. Y., & Niu, W. H. (10 de Noviembre de 2016). *The research of commonly used technologies and application fields on data mining*. Obtenido de http://dx.doi.org/10.4028/www.scientific.net/AMM.130-134.282

Mitchell Gary. (2013). *Selecting the best theory to implement planned change.*

P. Porto. (25 de Enero de 2017). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. Obtenido de http://recipp.ipp.pt/bitstream/10400.22/136/1/KDD-CRISP-SEMMA.pdf

Pakdd. (2010). *Advances in Knowledge Discovery and Data Mining.* India.

Perversi. (2007). *IDENTIFICACIÓN Y DETECCION DE PATRONES DELICTIVOS BASADA EN MINERIA DE DATOS.* Buenos Aires.

Sandip Pa. (25 de 01 de 2016). *Happiest Minds Technologies Pvt.* Obtenido de https://www.happiestminds.com/whitepapers/a-predictive-approach-to-understand-online-buyers-behavior.pdf

SAS - España. (2006). *Minería de Datos en el Sector Seguros.* Madrid.

Vallejos, Sofía. (2016). *Minería de Datos.*

Víctor López. (2016). *Aplicación de técnicas de minería de datos para la fidelización y retención de clientes.*

Anexos