



UNIVERSIDAD TECNOLÓGICA ISRAEL

ESCUELA DE POSGRADOS “ESPOG”

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

Resolución: RPC-SO-32-No.536-2023.

PROYECTO DE TITULACIÓN EN OPCIÓN AL GRADO DE MAGISTER

Título del proyecto:
Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data
Línea de Investigación:
Ciencias de la ingeniería aplicadas a la producción, sociedad y desarrollo sustentable
Campo amplio de conocimiento:
Tecnologías de la Información y la Comunicación (TIC)
Autor:
Bryan Andrés Gualoto Barahona
Tutores:
Mg. Mario Pérez (Tutor Técnico) Phd. Renato Toasa (Tutor Metodológico)

Quito – Ecuador

2025

APROBACIÓN DEL TUTOR



Yo, Mario Rubén Pérez Cargua con C.I: 0603251984 en mi calidad de Tutor del proyecto de investigación titulado: **Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data**

Elaborado por: Bryan Andrés Gualoto Barahona, de C.I: 1723965289, estudiante de la Maestría en BIG DATA Y CIENCIA DE DATOS, de la UNIVERSIDAD TECNOLÓGICA ISRAEL (UISRAEL), como parte de los requisitos sustanciales con fines de obtener el Título de Magister, me permito declarar que luego de haber orientado, analizado y revisado el trabajo de titulación, lo apruebo en todas sus partes.

Quito D.M., septiembre de 2025

Firma

APROBACIÓN DEL TUTOR



Yo, Renato Mauricio Toasa Guachi, con C.I: 1804724167 en mi calidad de Tutor del proyecto de investigación titulado: **Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data**

Elaborado por: Bryan Andrés Gualoto Barahona, de C.I: 1723965289, estudiante de la Maestría en BIG DATA Y CIENCIA DE DATOS, de la UNIVERSIDAD TECNOLÓGICA ISRAEL (UISRAEL), como parte de los requisitos sustanciales con fines de obtener el Título de Magister, me permito declarar que luego de haber orientado, analizado y revisado el trabajo de titulación, lo apruebo en todas sus partes.

Quito D.M., septiembre de 2025

Firma

DECLARACIÓN DE AUTORIZACIÓN POR PARTE DEL ESTUDIANTE



Yo, *Bryan Andrés Gualoto Barahona* con C.I: 1723965289, autor/a del proyecto de titulación denominado: **Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data.** Previo a la obtención del título de Magister en Big Data y Ciencia de Datos.

1. Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar el respectivo trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.
2. Manifiesto mi voluntad de ceder a la Universidad Tecnológica Israel los derechos patrimoniales consagrados en la Ley de Propiedad Intelectual del Ecuador, artículos 4, 5 y 6, en calidad de autor@ del trabajo de titulación, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente. En concordancia suscribo este documento en el momento que hago entrega del trabajo final en formato impreso y digital como parte del acervo bibliográfico de la Universidad Tecnológica Israel.
3. Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de prosperidad intelectual vigentes.

Quito D.M., septiembre de 2025

Firma

Tabla de contenidos

APROBACIÓN DEL TUTOR	ii
APROBACIÓN DEL TUTOR	ii
DECLARACIÓN DE AUTORIZACIÓN POR PARTE DEL ESTUDIANTE	iii
INFORMACIÓN GENERAL	1
Contextualización del tema	1
Problema de investigación	1
Objetivo general	1
Objetivos específicos	1
Vinculación con la sociedad y beneficiarios directos:	1
CAPÍTULO I: DESCRIPCIÓN DEL PROYECTO	2
1.1. Contextualización general del estado del arte	2
1.2. Proceso investigativo metodológico	2
1.3. Análisis de resultados	2
CAPÍTULO II: PROPUESTA	3
2.1. Fundamentos teóricos aplicados	3
2.2. Descripción de la propuesta	3
2.3. Validación de la propuesta	3
2.4. Matriz de articulación de la propuesta	4
CONCLUSIONES	5
RECOMENDACIONES	6
BIBLIOGRAFÍA	7
ANEXOS	8

Índice de tablas

Tabla 1 Comparación de Metodología.....	12
Tabla 2 Comparación de herramientas	16
Tabla 3 Comparación de uso	17
Tabla 4 Comparación de herramientas de visualización.....	20
Tabla 5 Matriz de articulación.....	24

Índice de figuras

Figura 1 Estructura general de la propuesta. Metodología Crisp DM.	13
Figura 2 Identificación de columnas con similitud.....	14
Figura 3 Identificación de valores.....	19

INFORMACIÓN GENERAL

Contextualización del tema

El creciente aumento de los índices de violencia y criminalidad en América Latina ha hecho despertar la preocupación no solo en el ámbito académico, sino en el político y social. Uno de los países más golpeados ha sido Ecuador, que ha visto incrementar de forma significativa su inseguridad ciudadana, con los homicidios como principal indicador de violencia letal. De acuerdo con el Instituto Nacional de Estadística y Censos (INEC), las muertes violentas en el país han experimentado una tendencia al crecimiento en los últimos años lo que refleja una problemática social estructural que debe ser abordada si se quiere lograr soluciones a mediano plazo (INEC, 2024).

Al mismo tiempo, el análisis de las características sociodemográficas de los individuos aprehendidos conlleva una mejor comprensión del contexto social y territorial de la comisión del delito. Tener en cuenta estas variables, tales como edad, sexo, nivel educativo, lugar de detención y tipo de delito imputado es un eslabón clave para realizar diagnósticos integrales y así, implementar estrategias de prevención realmente efectivas.

Frente a esta situación, el uso de metodologías estructuradas de análisis, como CRISP-DM, puede brindar una guía a la hora de abordar procesos de minería de datos aplicada a problemas reales. Dicha metodología, muy empleada en proyectos de ciencia de datos, permite una comprensión total del problema, un procesamiento de grandes cantidades de información amigable y la generación de modelos analíticos que puedan luego ser visualizados e interpretados por usuarios no expertos en la materia.

Finalmente, el presente trabajo propone el análisis y visualización de dos datasets oficiales utilizando CRISP-DM: el primero de ellos, acerca de los homicidios registrados en Ecuador en el período 2020-2024, y el segundo, sobre los ciudadanos aprehendidos durante el mismo lapso. A través de Dockers, PySpark y Tableau como herramientas de tratamiento de datos masivos y visualización interactiva de los resultados respectivamente, la presente propuesta busca sentar una base sólida para la toma de decisiones en política de seguridad y prevención del delito en el país.

Problema de investigación

En la última década, el Ecuador ha enfrentado un aumento alarmante en los niveles de violencia, particularmente en los delitos relacionados con homicidios intencionales, sicariato y

feminicidios. Este fenómeno, que afecta tanto a zonas urbanas como rurales, se ha intensificado en determinadas provincias como Guayas, Esmeraldas, Manabí y Los Ríos, donde la tasa de homicidios ha superado la media nacional en varios periodos recientes (Ministerio del Interior, 2024). El crecimiento de la criminalidad ha generado lo antes dicho; una sensación de inseguridad generalizada y serios retos para las políticas públicas en seguridad ciudadana y control del delito

Aunque el país dispone de registros estadísticos históricos detallados sobre homicidios intencionales, que incluyen variables geográficas, temporales, sociodemográficas y circunstanciales, el uso efectivo de la información para el diseño de estrategias predictivas e intervención a nivel espacio-temporal es aún limitado. Herramientas de análisis tradicionales no son suficientes para identificación de patrones ocultos ni anticipación de eventos críticos espacio-temporales, razón por la cual las actuaciones son reactivas y complican la postura institucional de seguridad.

El análisis de gran volumen de datos Big Data y técnicas de ciencia de datos representan una oportunidad estratégica para transformar esos datos crudos en conocimiento útil. A través de herramientas como análisis geo-espacial, modelos de predicción en tiempo y visualizaciones interactivas, puede identificar zonas de riesgo, factores para biológica al homicidio y anticipar comportamientos futuros (PNUD, 2023; CEPAL, 2022). No obstante, en el contexto ecuatoriano, no existen aplicaciones sistemáticas de estas tecnologías para abordar el asunto de la violencia letal.

Uno de los problemas fundamentales es la poca disponibilidad, y variabilidad de calidad y volumen de datos que no permiten análisis profundos ni extracción de información relevante haciendo uso de herramientas de procesamiento. La necesidad de que las visualizaciones generadas tengan valor predictivo, sean comprensibles, interactivas por autoridades y audiencias promedio y que puedan ser utilizadas por investigadores y ciudadanos en general a la vez son otros retos a superar (Kitchin, 2014).

A pesar de que metodologías como CRISP-DM han demostrado ser efectivas para la estructuración de proyectos de análisis de datos complejos, su aplicación en el caso Ecuador, en marcos de seguridad ciudadana, es aún escasa. Además, tecnologías como PySpark o Tableau que permiten el manejo de grandes volúmenes de datos y la generación visualizaciones de alto nivel aún no está integradas en procesos institucionales de análisis delictivo.

Objetivo general

Aplicar la metodología CRISP-DM para el análisis predictivo y visualización de patrones de comportamiento delictivo en los homicidios y aprehensiones registrados en Ecuador entre los años 2020 y 2024, mediante técnicas de Big Data.

Objetivos específicos

- Contextualizar los fundamentos teóricos de la metodología CRISP-DM y de las herramientas de análisis de datos, así como su aplicabilidad en la predicción de patrones delictivos en el contexto ecuatoriano
- Diagnosticar los parámetros y variables clave necesaria para identificar y analizar las tendencias más relevantes sobre homicidios y aprehensiones en Ecuador entre 2020 y 2024.
- Desarrollar el proceso completo de tratamiento, limpieza y análisis predictivo de los datos delictivos mediante la metodología CRISP-DM, para su visualización.
- Valorar a través de criterio de especialistas el escenario propuesto.

Vinculación con la sociedad y beneficiarios directos:

El apartado de vinculación con la sociedad es un componente esencial de todo trabajo investigativo, el cual permite que los resultados no permanezcan únicamente en el ámbito académico, sino que se compartan, discutan y validen con diferentes actores sociales. En este caso del presente proyecto, se buscó generar un primer espacio de socialización mediante un taller el cual fue dirigido a tres compañeros de trabajo de la empresa Security Data, los mismos que están involucrados en el área de TI y desarrollo, quienes representan un grupo de beneficiarios directos del estudio. La evidencia se adjunta en el anexo 1

La charla se llevó a cabo en el ámbito laboral y tuvo como propósito exponer los avances del proyecto de análisis de homicidios y aprehensiones en el Ecuador (2020–2024). La presentación se estructuró en tres fases para que el taller y el tema a tratar se entendible:

Exposición del problema: en esta fase se explicó la importancia de abordar e interpretar el problema social como lo es la inseguridad desde una perspectiva de datos, señalando cómo los homicidios y aprehensiones constituyen indicadores críticos en la seguridad pública.

Metodología aplicada: Se describió la utilización de la metodología CRISP-DM como marco analítico, así como el empleo de herramientas tecnológicas para el proceso técnico como Python, Apache Spark, Prophet y Tableau, resaltando la pertinencia de integrar modelos predictivos en tableros interactivos.

Impactos esperados: Se discutió el valor de contar con proyecciones y visualizaciones dinámicas que permitan anticipar tendencias delictivas y focalizar recursos de seguridad.

CAPÍTULO I: DESCRIPCIÓN DEL PROYECTO

1.1. Contextualización general del estado del arte

En los últimos diez años, los avances tecnológicos en la analítica de grandes volúmenes de datos han permitido que técnicas de predicción sean aplicadas a una gran variedad de aspectos de la vida humana, introduciendo herramientas como MOOC y laboratorios científicos. Desde ese momento, la utilización de Big Data ha demostrado ser un medio altamente efectivo para anticipar fenómenos sociales complejos tales como crimen y la violencia (Chen, Chiang y Storey, 2012)

Si se analiza a nivel internacional, un número apreciable de estudios centran su atención en la relación entre análisis de datos y predicción del comportamiento delictivo. Por ejemplo, en países como los Estados Unidos y el Reino Unido que utilizan modelos predictivos han podido reconocer zonas calientes delictivas, tendencias temporales de violencia, perfiles del comportamiento criminal, con lo cual se han establecido políticas públicas más efectivas (Mohler et al., 2015). Estas experiencias dan paso a sistemas como PredPol y CompStat, que utilizan datos históricos para respaldar decisiones gracias a un algoritmo de predicción que ayudan entes gubernamentales que mantienen el orden ciudadano.

Aunque en América Latina el uso de este tipo de tecnologías no han sido completamente desarrolladas, en países como Colombia, México y Brasil se han iniciado proyectos que integran técnicas de minería de datos y visualización interactiva con el fin de comprender mejor las dinámicas del crimen. Sin embargo, la falta de fuentes de información que generan datos reales, la escasa capacitación institucional y los vacíos legales han hecho que no sea posible llevar a cabo un análisis efectivo y profundo (González et al., 2020).

En el caso ecuatoriano, ha sido posible acceder a datos abiertos del Instituto Nacional de Estadística y Censos (INEC) y del Ministerio del Interior, que han servido para realizar análisis descriptivos sobre la violencia, especialmente desde el punto de vista de homicidios y aprehensiones. Sin embargo, la integración de estos datos en marcos metodológicos estructurados como CRISP-DM y la utilización de herramientas tecnológicas tales como Dockers, PySpark o Tableau para análisis predictivo y visualización de comportamientos y patrones de gran impacto aún no han sido suficientemente desarrolladas.

Esta investigación se sitúa en esa línea de innovación, proponiendo el uso de tecnologías avanzadas de análisis de datos para construir modelos predictivos sobre homicidios y aprehensiones en Ecuador, generando visualizaciones dinámicas que sirvan como insumo para la gestión pública y la toma de decisiones en materia de seguridad.

1.2. Proceso investigativo metodológico

La presente investigación adopta un enfoque cuantitativo, dado que se fundamenta en la recolección, procesamiento y análisis de datos estructurados sobre homicidios y aprehensiones en Ecuador, registrados entre los años 2020 y 2024. Este enfoque se caracteriza por el uso de técnicas estadísticas, algoritmos computacionales y modelos predictivos que permiten explicar fenómenos sociales mediante la medición objetiva y el análisis numérico (Hernández-Sampieri et al., 2014)

1.2.1. Diseño metodológico

Dado que el análisis se centra en datos históricos de homicidios y aprehensiones en Ecuador, así como en la evaluación de la aplicabilidad de técnicas de análisis predictivo para la gestión de la seguridad ciudadana, el diseño de la investigación se clasifica como descriptivo y no experimental. Esto implica que las variables de estudio no serán manipuladas, sino observadas y analizadas tal como se registraron en las bases de datos oficiales durante el periodo 2020–2024 (Hernández-Sampieri et al., 2014).

Para el componente de percepción experta, se aplicará la técnica de encuesta estructurada con escala de Likert, con el fin de obtener información primaria que permita medir su valoración sobre la utilidad de los modelos de análisis predictivo en la toma de decisiones estratégicas.

1.2.2 Población y muestra

La población objeto de estudio está conformada por el total de registros históricos de homicidios y aprehensiones ocurridos en Ecuador durante el período 2020–2024, recopilados a partir de fuentes oficiales como la Policía Nacional, el Instituto Nacional de Estadística y Censos (INEC) y el Ministerio del Interior. Esta información incluye variables como fecha, provincia, sexo, edad de la víctima o persona aprehendida, tipo de delito, motivación y ubicación geográfica, lo que permite realizar un análisis descriptivo y predictivo de las tendencias delictivas.

Para el componente de percepción experta, la población estará compuesta por profesionales entendidos en el tema público como lo es homicidios y aprehendidos los cuales serán un total de 15 agentes policiales del distrito Calderón los mismos que podrán aportar su criterio sobre el

tema a tratar y esta información será de gran utilidad para realizar el análisis y predicción sobre el tema planteado en este estudio.

1.2.3. Instrumento de Recolección de Datos

En esta investigación, el instrumento principal de recolección de datos fue la observación directa aplicada a las fuentes oficiales de información estadística y bases de datos que se encuentran para uso público relacionadas con homicidios y aprehensiones en el Ecuador.

La observación directa, según Sampieri, Fernández y Baptista (2014), es una técnica que hace posible la recolección de datos de forma ordenada, a través de la percepción controlada del investigador acerca de fenómenos o registros, sin tocar las variables de estudio.

La observación directa sobre registros documentales es ideal para investigaciones basadas en datos masivos (Big Data), ya que permite recopilar información completa, continua y estructurada. Esto posibilita el desarrollo de modelos de análisis descriptivo y predictivo sin necesidad de aplicar encuestas o entrevistas, reduciendo así tiempos y costos de recolección.

1.3. Análisis de resultados

El análisis de resultados se centra en el procesamiento, exploración y visualización de datos sobre homicidios intencionales y personas detenidas o aprehendidas en el Ecuador, a partir de los registros recopilados durante los periodos 2020-2024, respectivamente. Para ello, se aplicaron técnicas de limpieza, transformación y análisis descriptivo con apoyo de herramientas como Dockers, Python, Pyspark, herramientas de predicción de datos y Tableau, con el fin de identificar patrones de comportamiento y a su vez visualizar estos datos.

Luego, se aplicarán técnicas de análisis geoespacial para visualizar los homicidios sobre el territorio nacional mediante mapas de calor, gráficos por provincia, así como la detección de zonas con alta concentración delictiva (hot spots). Para ello se utilizarán datos de provincias y cantones incluidas en el dataset.

Después de los pasos previamente planteados, se implementará un modelo de predicción basado en series temporales (usando Prophet o ARIMA), con el propósito de estimar el número de homicidios futuros por mes o por provincia. Además, se evaluarán posibles correlaciones entre variables para identificar factores asociados a una mayor probabilidad de ocurrencia.

Finalmente, los resultados se organizarán en panel interactivos concebidos con Tableau, lo que posibilitará ver dinámicamente los indicadores esenciales, comparaciones por temporada, puntos geográficos cubiertos en este análisis gráfico. Estos informes visuales harán más fácil comprender los datos y por tanto son en realidad una herramienta necesaria tanto para personas a cargo de la toma de decisiones, académicos y personas que participan en el sistema de seguridad ciudadana.

A continuación, se presentan los indicadores que fueron evaluados en el análisis

Dashboard análisis descriptivo

Este primer tablero busca mostrar, de la manera más clara posible, qué ha pasado en Ecuador en los últimos años en materia de homicidios y aprehensiones. La idea es que, de un vistazo, cualquier persona pueda entender la magnitud del problema y sus características más importantes.

Lo primero que aparece son los indicadores principales (KPI):

- Cuántos homicidios se han registrado en total.
- Cuántas aprehensiones en el mismo periodo.
- Cómo ha sido la variación de un año a otro (si los casos aumentan o disminuyen).

Luego vienen los gráficos y mapas que cuentan la historia

- Una línea temporal que muestra la evolución de homicidios y aprehensiones mes a mes, para identificar los picos y caídas.
- Un mapa del Ecuador, coloreado por provincias, donde se puede ver de inmediato en qué zonas el problema es más fuerte.
- Un gráfico de barras horizontales con las 5 provincias más críticas.
- Un gráfico que muestra la distribución por sexo (hombres/mujeres), que nos permite analizar diferencias de género.
- Otro que presenta las principales motivaciones de los homicidios.
- Una vista de los tipos de armas más utilizadas, para dimensionar la presencia de armas de fuego frente a otras.
- Y, finalmente, un análisis por grupo de edad, que nos dice qué rangos están siendo más afectados

Dashboard 2: Análisis Predictivo

El segundo tablero es distinto: ya no se centra en lo que pasó, sino en lo que podría pasar en el futuro. Aquí entran en juego los modelos de predicción que permiten estimar cómo se comportarán los homicidios y aprehensiones en los próximos meses o años.

Arriba se muestran indicadores clave:

- La proyección de homicidios para el año siguiente.
- La proyección de aprehensiones.
- El mes que, según el modelo, tendrá más delitos (lo que funciona como una alerta temprana).
- Y un dato técnico pero importante: el error del modelo (RMSE o MAE), que nos dice qué tan confiable es la predicción.

Las visualizaciones ayudan a entender las proyecciones:

- Una línea histórica y proyectada: en azul se ven los datos reales y en rojo los valores predichos.
- Un mapa predictivo, que resalta las provincias donde se espera un aumento en los delitos.
- Gráficos de barras comparativas entre lo ocurrido y lo que se espera que ocurra.
- Y un gráfico con bandas de confianza (del modelo Prophet), que no solo da una predicción, sino también un rango de incertidumbre, mostrando los escenarios posibles.

CAPÍTULO II: PROPUESTA

1.1. Fundamentos teóricos aplicados

A continuación, se presentan conceptos teóricos que serán manejados dentro del presente proyecto.

Big Data

Big Data es el término utilizado para describir el procesamiento y análisis de enormes volúmenes de datos que no pueden ser administrados a través de los mismos métodos tradicionales debido a la cantidad, la variedad y la velocidad de generación de información. De acuerdo con Marr (2015), Big Data se distingue por las 5 “V” de: volumen, velocidad, variedad, veracidad y valor. Específicamente, en términos de seguridad, permite la visualización de datos y alertas tempranas, segmentación basada en perímetros de áreas temporales y geográficas que no pueden llevar a cabo inconsistencias de detección y series de tiempo.

Machine Learning y predicción

El Machine Learning o conocido como aprendizaje automático es una rama de la inteligencia artificial, que se centra en el desarrollo de algoritmos que tienen como objetivo aprender patrones a partir de datos y hacer predicciones sin ser programados previamente para cada tarea en específico (Bishop, 2006).

En esta propuesta, se utilizan modelos de series temporales, como ARIMA y Prophet, con la finalidad de predecir la ocurrencia futura de homicidios o aprehensiones. La utilidad principal de este enfoque es anticipar tendencias que permitan generar alertas tempranas y fortalecer políticas de seguridad

Dockers

Para el trabajo con el Análisis de datos y el Big Data, la reproducibilidad y la portabilidad son críticas. Docker es una solución tecnológica para crear contenedores, que son entornos que empaquetan todas las dependencias y la configuración necesarias para ejecutar una aplicación. A diferencia de las máquinas virtuales convencionales, los contenedores de Docker no requieren que se tenga un sistema operativo independiente para cada proceso, sino que comparten el kernel del sistema anfitrión, lo que los hace más ligeros, rápidos y eficientes en el uso de recursos (Merkel, 2014).

Apache PySpark

Apache Spark es un motor de análisis distribuido que permite el procesamiento masivo de datos en memoria RAM, optimizando de mejor manera tareas de procesamiento que serían inviables en un solo equipo (Zaharia et al., 2016). La implementación en Python, conocida como PySpark, facilita la integración con varias librerías una de ellas y la más conocida denominada pandas.

En la presente propuesta, PySpark es considerada una herramienta fundamental para la manipulación de datos que contienen registros históricos de homicidios y aprehensiones, ya que posibilita y facilita el ejecutar operaciones de limpieza, agregación y modelado de forma escalable y ágil.

Modelos de predicción: ARIMA y Prophet

Los modelos de series temporales son esenciales para anticipar la evolución de fenómenos delictivos.

- ARIMA: se basa en la combinación de tres componentes: autorregresivo (AR), integración (I) y media móvil (MA). Este modelo es el idóneo y el más adecuado para series estacionarias y permite capturar tendencias históricas con precisión (Box et al., 2016).
- Prophet: es un modelo aditivo desarrollado por Facebook que usa la tendencia, la estacionalidad y los factores externos. Prophet tiene la capacidad de indicar series con variabilidad no lineal y múltiples estacionalidades (Taylor y Letham, 2018).

Para el presente proyecto, se utilizará Prophet, para realizar las predicciones correspondientes que se presentarán en el tablero con referencia al año 2025, esto basado en los datos históricos tanto de homicidios como aprehensiones.

Tableau y visualización de datos

La siguiente etapa crítica en cualquier análisis es la visualización de datos, ya que traduce la información compleja en representaciones gráficas comprensibles para diferentes tipos de usuarios. Tableau es una de las principales herramientas de Business Intelligence en el mercado moderno que permite una creación sencilla de tableros interactivos, filtros dinámicos y visualizaciones avanzadas sin la necesidad de programación compleja (Murray, 2013). En este proyecto, Tableau se utiliza para construir tableros comparativos entre homicidios y aprehensiones, mapas de calor y gráficos predictivos, ofreciendo una plataforma accesible para analista de datos.

Visualización geoespacial

La visualización geoespacial es una herramienta clave en la criminología moderna, ya que permite mapear e identificar los delitos sobre el territorio, identificar zonas calientes (hot spots) y representar gráficamente concentraciones delictivas como provincias donde se tienen más actos delictivos. Según Chainey y Ratcliffe (2005), este tipo de modelo de visualización se ha demostrado ser crucial en el análisis geográfico de datos delictivos y, por lo tanto, explora más el espacio geográfico del crimen e influye en las soluciones de seguridad futuras.

Ciencia de datos aplicada a la seguridad

La ciencia de datos es un campo interdisciplinario que combina estadística, programación y conocimiento del dominio para extraer conocimiento útil a partir de grandes volúmenes de información. Su aplicación en el ámbito de la seguridad permite analizar tendencias delictivas, prever eventos, clasificar tipologías y generar reportes automatizados (Provost y Fawcett, 2013). En esta propuesta, la ciencia de datos se propone como un soporte metodológico fundamental para la integración y transformación de los datos de homicidios en insumos estratégicos.

Justificación del método utilizado

En el área de la minería de datos y el análisis previsional, hay varios marcos de trabajo que orientan el desarrollo del proyecto. Entre los más comunes se hallan CRISP-DM (Proceso Estándar Inter industrial para la Minería de Datos), SEMMA (Seleccionar, Explorar, Modificar, Modelar, Evaluar), Ralph Kimball y Bill Inmon. Aunque todos buscan organizar las etapas de un proyecto analítico, tienen diferencias notables en su enfoque, alcance y modo de aplicación. A continuación, se presenta en la figura 1, la tabla comparativa.

Tabla 1

Comparación de Metodología.

	CRISP-DM	SEMMA	Bill Inmon	Ralph Kimball
Tipo	Metodología minera de datos	Analítica Técnica	Top-down	Bottom-up
Etapas	Negocio → Datos → Preparación → Modelado →	Muestreo → Exploración → Modificación	Data Warehouse central	Data Marts → Integración → Data Warehouse

	Evaluación → Despliegue	→ Modelado → Evaluación	normalizado → Data Marts	
Enfoque	Flexible y estratégico	Estadístico	Corporativo	Orientado a negocio.
Ventajas	Alineación y técnicas adaptables	Rapidez en pruebas.	Visión unificada de datos.	Fácil de graficar.
Limitaciones	En ocasiones más lento en implementación.	No considera las necesidades del negocio	Alto costo	Riesgo de inconsistencia con datos

Por lo cual para este proyecto se tomó la metodología de Crisp-DM ya que es ideal para el desarrollo de proyectos de ciencia de datos, sin necesidad de que tan grande sea el tamaño del proyecto.

La elección de CRISP-DM (Cross Industry Standard Process for Data Mining) como metodología central en este proyecto se justifica en varios aspectos clave.

En primer lugar, CRISP-DM es la metodología mundial más utilizada en proyectos de minería de datos y analítica predictiva. Según Chapman et al. (2000), afirmaron que, debido a su estructura flexible y adaptable, se puede adaptar a casi todas las demandas en una amplia gama de industrias, incluyendo desde la seguridad hasta la salud y el comercio. Otra ventaja de CRISP-DM en el contexto de la minería de datos es su enfoque equilibrado entre los objetivos comerciales y el proceso analítico: en contraste con SEMMA, CRISP-DM se usa frecuentemente en proyectos en los que las prioridades de análisis son críticas para la empresa- toma de decisiones relacionadas.

Otro punto a favor es que sus seis fases (comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación y despliegue) constituyen un ciclo iterativo. Esto significa que el analista puede regresar a fases anteriores en caso de encontrar limitaciones en los datos o en el modelado. Esta característica resulta muy valiosa en proyectos como el presente, donde los registros de homicidios y aprehensiones pueden tener vacíos o inconsistencias que exigen retroalimentación constante.

1.2. Descripción de la propuesta

La propuesta presentada se centra en aprovechar el uso de Big Data, la minería de datos y el análisis predictivo para abordar el estudio de los homicidios intencionales y las aprehensiones cometidas en Ecuador durante el período 2020-2024. Por un lado, el objetivo principal del proyecto es aplicar enfoques y técnicas comprobadas en el análisis de datos para no solo describir, sino también predecir las relaciones delictivas utilizando modelos predictivos y visualizaciones interactivas.

En este sentido, el concepto de Big Data juega un papel crucial, ya que permite procesar grandes cantidades de datos de diferentes fuentes con altos valores de variedad, velocidad y veracidad (Marr, 2016). A su vez, los datos recopilados sobre homicidios y aprehensiones en Ecuador contienen datos como fuente de dato, fechas, edades, sexos, armas utilizadas, motivación y lugar, lo que implica un sistema de clasificación de la información y su uso en términos de análisis

La propuesta se enmarca dentro del modelo CRISP-DM, la cual proviene de la combinación de las palabras en inglés Cross Industry Standard Process for Data Mining, el cual fue creado por un consorcio europeo SPSS en 2000 el mismo que permite un marco flexible y estructurado para proyectos que están basados en minería de datos. La metodología consta de seis fases: comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación e implementación (Chapman et al., 2000). Esta metodología es ampliamente utilizada por su capacidad de adaptarse a diferentes dominios y por su enfoque iterativo, lo que permite ajustar el análisis según los hallazgos encontrados.

Por otro lado, el análisis predictivo se basa en aplicar Prophet, creador por Facebook/Meta el cual ofrece ventajas como la facilidad de implementación, lo que lo hace particularmente adecuado para datos delictivos con variaciones mensuales y anuales futuras (Shmueli y Koppius, 2011).

Además de estas técnicas, se proponen soluciones de procesamiento distribuido basadas en Apache Spark y PySpark, ya que permite procesar datos en paralelo y, por lo tanto, mejorar los tiempos de ejecución. En la etapa de implementación, se recomienda utilizar Docker para garantizar la productibilidad y portabilidad del entorno.

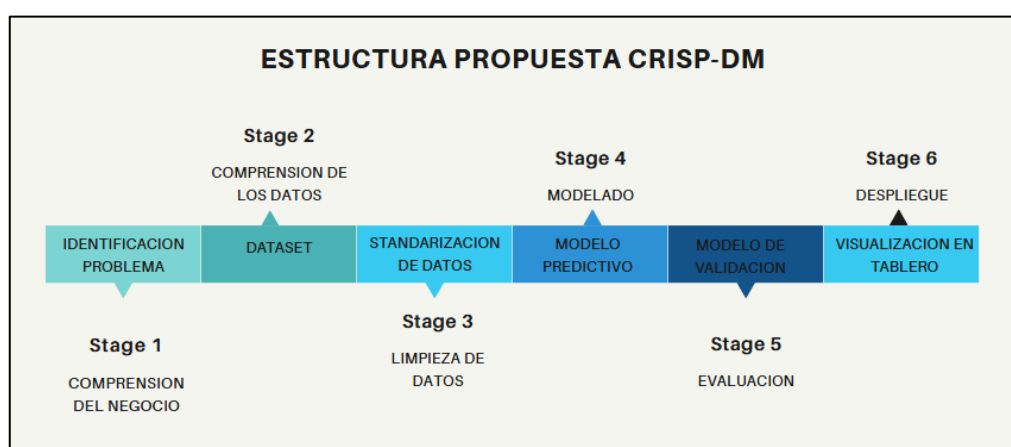
Con respecto a la visualización, se propone Tableau, una de las herramientas más populares para visualización de datos, ya que permite construir dashboards interactivos con mapas y filtros dinámicos.

a. Estructura general

A continuación, podemos visualizar las fases en las que se divide el presente proyecto, el mismo que está basado en la metodología Crisp DM.

Figura 1

Estructura de la propuesta basado en Crisp DM



Nota. La Figura 1 indica las fases de la metodología Crisp DM del presente proyecto.

b. Explicación del aporte

Comprensión del negocio:

Identificación del problema de investigación: altos niveles de homicidios intencionales y aprehensiones en Ecuador durante los últimos años, se defino los objetivos del presente proyecto, con la finalidad de definir patrones delictivos que mantengan relación entre homicidios y aprehensiones.

Entendimientos de los datos:

Revisión de fuentes oficiales de información, para nuestro estudio se tomó dos datasets inicialmente uno que comprende homicidios intencionales con registros históricos desde el 2014 hasta el 2024, con un total de 1.036.706 datos y otro dataset que abarca casos de aprehensiones desde el año 2020 hasta el año 2024 el cual consta con 17.144.644 de datos, los mismos que comparten columnas en común que serán identificadas para su análisis.

Identificación de variables en: fecha, año, mes, provincia, cantón, sexo, edad, grupo de edad, etnia, arma, fuente de dato, tipo de evento y motivación.

Adicional se identificaron datos faltantes, codificación de caracteres, los mismos que serán tratados posteriormente.

Figura 2

Identificación de columnas con similitud.

A	B	C	D	E	F	G	H	I	J	K	L	M
fuente_dato	tipo_evento	fecha	año	mes	provincia	cantón	sexo	edad	grupo_edad	etnia	arma	motivacion
HOMICIDIO	ASESINATO	2024-01-10	2024	1	1STO DGO DE LOS TSAO	SANTO DOMINGO	M	45-54		MESTIZO/A	BARRA	DELINCUENCIA COMUN
HOMICIDIO	ASESINATO	2024-07-28	2024	7	7STO DGO DE LOS TSAO	SANTO DOMINGO	M	20-25-34		AFRO	BARRA	VIOLENCIA COMUNITARIA
HOMICIDIO	HOMICIDIO	2024-11-23	2024	11	11STO DGO DE LOS TSAO	SANTO DOMINGO	M	34-25-34		MESTIZO/A	BATE	VIOLENCIA COMUNITARIA
HOMICIDIO	FEMICIDIO	2020-11-28	2020	11	11STO DGO DE LOS TSAO	SANTO DOMINGO	F	37-35-44		MESTIZO/A	CABLE	VIOLENCIA INTRAFAMILIAR
HOMICIDIO	HOMICIDIO	2023-05-11	2023	5	5STO DGO DE LOS TSAO	LA CONCORDIA	M	43-35-44		MESTIZO/A	CABLE	DELINCUENCIA COMUN
HOMICIDIO	HOMICIDIO	2021-12-14	2021	12	12STO DGO DE LOS TSAO	LA CONCORDIA	M	18-12-17		MULATO/A	CARTUCHE	DELINCUENCIA COMUN
HOMICIDIO	ASESINATO	2022-03-30	2022	3	3STO DGO DE LOS TSAO	SANTO DOMINGO	M	38-35-44		MULATO/A	CARTUCHE	VIOLENCIA COMUNITARIA
HOMICIDIO	ASESINATO	2023-02-26	2023	2	2STO DGO DE LOS TSAO	SANTO DOMINGO	M	34-25-34		MESTIZO/A	CARTUCHE	VIOLENCIA COMUNITARIA
HOMICIDIO	ASESINATO	2023-05-17	2023	5	5STO DGO DE LOS TSAO	SANTO DOMINGO	M	53-45-64		MESTIZO/A	CARTUCHE	DELINCUENCIA COMUN
HOMICIDIO	ASESINATO	2023-11-12	2023	11	11STO DGO DE LOS TSAO	SANTO DOMINGO	M	29-25-34		AFRO	CARTUCHE	DELINCUENCIA COMUN
HOMICIDIO	ASESINATO	2024-02-20	2024	2	2STO DGO DE LOS TSAO	SANTO DOMINGO	F	52-45-64		MESTIZO/A	CARTUCHE	DELINCUENCIA COMUN
HOMICIDIO	ASESINATO	2024-10-03	2024	10	10STO DGO DE LOS TSAO	SANTO DOMINGO	M	22-18-24		MESTIZO/A	CARTUCHE	DELINCUENCIA COMUN
HOMICIDIO	HOMICIDIO	2023-10-24	2023	10	10STO DGO DE LOS TSAO	SANTO DOMINGO	M	80-65+		MESTIZO/A	CORTO PUN	DELINCUENCIA COMUN
HOMICIDIO	ASESINATO	2022-04-02	2022	4	4STO DGO DE LOS TSAO	LA CONCORDIA	M	67-65+		MESTIZO/A	CUCHILLO	VIOLENCIA COMUNITARIA
HOMICIDIO	ASESINATO	2022-06-29	2022	6	6STO DGO DE LOS TSAO	SANTO DOMINGO	F	27-25-34		MESTIZO/A	CUCHILLO	DELINCUENCIA COMUN

Nota. La Figura 2 indica la selección de columnas a trabajar.

Preparación de los datos:

Carga y limpieza

Comparación de herramientas de procesamiento

Para el procesamiento de datos y gestión de los mismos es importante tomar diferencias entre los diferentes tipos de herramientas, las herramientas previamente analizadas fueron ApacheSpark(PySpark) y Hadoop, las mismas que se hacen una evaluación a continuación en la Tabla 2.

Tabla 2

Comparación de herramientas.

Criterio	Apache Spark (PySpark)	Hadoop
Uso de memoria	Si	No
Escalabilidad	Alta	Media
Velocidad	Alta	Baja
Compatibilidad	Alta	Alta
Facilidad de manejo	Alta	Baja

Para la carga y limpieza de datos se hace uso de Docker Desktop, para un mejor procesamiento de datos al momento de hacerlo desde la herramienta en línea que incluye el mismo software, el cual usará código Python con librerías PySpark haciendo uso de Jupyter

notebook, las mismas que son usadas para procesamiento de grandes volúmenes de datos por su eficacia y velocidad de procesamiento.

- Estandarización de texto.
- Transformación de caracteres (ej. reemplazo de tildes y “ñ”).
- Eliminación de columnas innecesarias.
- Eliminación de datos nulos.
- Eliminación de valores anómalos.

Modelado:

Justificación del modelo a usar

Tabla 3

Comparación de uso.

Criterio	Propeth	ARIMA
Tipo de modelo	Series temporales con tendencia, estacionalidad.	Enfocado a procesos lineales y estacionarios.
Requisitos	Ideal para manejo de valores faltantes.	Requiere que la serie sea estacionaria.
Parámetros	Fácil de configurar, sin parámetros complejos.	Requiere definir métodos más específicos.
Estacionalidad	Soporta multiples estacionalidades(diaria, mensual, anual).	Modela una única estacionalidad.
Robustez	Ideal ante valores atípicos	Sensible a cambios.
Interpretación	Entrega bandas de confianza(yhat_low, yhat_upper).	Ofrece una predicción en específico.
Escalabilidad	Se integra correctamente con herramientas como Pyspark, Python.	Cada serie requiere un ajuste.

Si bien ARIMA es un método adecuado para trabajar con predicciones de series temporales, de forma estable, Prophet es más flexible, interpretable y robusto. La capacidad de capturar

tendencias no lineales, trabajar con una cantidad considerable de datos faltantes y proporcionar intervalos de confianza lo hacen más adecuado para el propósito de esta tesis, predecir patrones delictivos y visibilizarlos en dashboards interactivos bajo la metodología CRISP-DM.

El modelado se llevó a cabo en un entorno de Jupyter Notebook, integrando PySpark para la preparación y limpieza de datos, y pandas para el manejo de series temporales a nivel agregado. La construcción del modelo fue llevada a cabo utilizando los registros correspondientes a los años 2020 a 2024, normalizados y agrupados previamente, en función de la fuente del dato, provincia, grupo de edad, sexo, etnia, arma y motivación. Para cada combinación de estas variables, se generaron series temporales mensuales, lo que permitió la construcción de modelos y la generación de predicciones diferenciales con predicción al año 2025. Con este enfoque, el proceso generó una predicción general a nivel nacional al mismo tiempo que una más diferencial por territorio y características sociodemográficas.

Evaluación:

La etapa de evaluación en la metodología CRISP-DM tiene como propósito determinar la calidad, validez y utilidad del modelo construido en la fase de modelado. Según Chapman et al. (2000), En esta etapa consiste en verificar que el modelo funcione de manera correcta y se alinee a la investigación planteada.

En el análisis de series temporales aplicado a fenómenos sociales y delictivos, la evaluación cobra gran importancia debido a la variabilidad y la presencia de valores atípicos. En este sentido se pueden utilizar los siguientes métricos:

RMSE (Root Mean Square Error): mide la desviación promedio entre los valores reales y los valores pronosticados.

MaPE (Mean Absolute Percentage Error): hace referencia al error en términos porcentuales.

El procedimiento fue:

- Entrenar el modelo Prophet usando datos hasta un año de corte (2024) y predecir el siguiente año (2025).
- Calcular RMSE y MAPE métricas sobre el rango de validación, para medir el error de pronóstico.
- Revisar en el .csv que se hayan exportado de manera correcta las bandas de confianza, (yhat_lower, yhat_upper), las cuales nos sirven para visualizarlo gráficamente, con el valor pronosticado (yhat).

Figura 3

Identificación de valores yhat_lower, yhat_upper.

motivacion	anio	real	yhat	yhat_lower	yhat_upper
BOLETAS	2020	11167	13409.7015405067	9603.79440251619	17280.575121801
BOLETAS	2021	15566	12937.4103611141	9124.14518960766	16659.8982967715
BOLETAS	2022	12824	12416.3731002245	8605.4122122571	16141.056050075
BOLETAS	2023	11625	11894.1393367213	8049.65424058644	15696.9013959751
BOLETAS	2024	10779	11291.9306401891	7552.66105166679	15157.3236721926
BOLETAS	2025	0	10810.2769136793	6970.12352277911	14553.3453197189
CONTRAVENCION CONTRA EL DERECHO AL TRABAJO	2020	10	10.0356944974272	9.34536783509113	10.7176259737213
CONTRAVENCION CONTRA EL DERECHO AL TRABAJO	2021	11	11.0606199238159	10.3781177072468	11.7641181438369
CONTRAVENCION CONTRA EL DERECHO AL TRABAJO	2022	6	5.99981230895088	5.55534178492183	6.45939284723751
CONTRAVENCION CONTRA EL DERECHO AL TRABAJO	2023	4	4.01533525189149	3.56374656797059	4.45856693267203
CONTRAVENCION CONTRA EL DERECHO AL TRABAJO	2024	3	7.94473892492015	6.64798016173932	9.2082402694828
CONTRAVENCION CONTRA EL DERECHO AL TRABAJO	2025	0	2.60685963671436	2.15848075736296	3.07557986805467

Nota. La Figura 3 presenta un ejemplo de los datos exportados en el análisis de predicción donde se puede validar la evaluación del modelado.

Despliegue:

Comparación de herramientas de visualización

Las visualizaciones de los datos generados constituyen una parte fundamental en el proceso del análisis predictivo y visualización de los patrones, ya que nos permite visualizar de manera gráfica grandes volúmenes de datos que nos facilitan la interpretación y la toma de decisiones, en este estudio se consideró dos herramientas Power bi y Tableau. A continuación, se presenta una comparación de estas herramientas en la Tabla 4.

Tabla 4

Comparación de herramientas de visualización.

Criterio	Tableau	Power Bi
Facilidad de uso	Interfaz intuitiva	Aprendizaje más suave
Integración de datos	Amplio soporte	Integración nativa
Capacidad de análisis	Potente en análisis visual	Adecuado para análisis descriptivo
Personalización	Alta flexibilidad	Personalización restringida
Costo	Licencias más costosas	Más económico
Escalabilidad	Grande manejo de datos	Buen rendimiento en entornos corporativos

Dado a la gran adaptabilidad, escalabilidad y capacidad de visualización en el presente proyecto se usará la herramienta Tableau la misma que nos permitirá visualizar el análisis de predicción y la visualización de patronas previamente planteados.

Desarrollo de dashboards interactivos en Tableau con filtros que se ajustan automáticamente al análisis que se desea realizar.

Difusión de resultados como herramienta de apoyo al entendimiento de patrones delictivos en el Ecuador.

c. Estrategias y/o técnicas

En esta sección se define un conjunto de estrategias y técnicas que componen la propuesta metodológica para abordar la sistematización y eficiencia del proceso de análisis de datos.

Procesamiento y preparación de datos

Limpieza y transformación: depuración de registros incompletos, estandarización de variables categóricas (ej. provincias, motivaciones), y manejo de valores atípicos.

Integración de fuentes: en este paso, integramos diversos datasets de homicidios y aprehensiones, ambos en un único repositorio. Este paso es crítico debido a la necesidad de garantizar la coherencia, lo que nos permite comparar y analizar mejor las tendencias y regulares en la información reunida.

Uso de herramientas: aprovechamos las herramientas avanzadas, como Dockers, Python, especialmente las bibliotecas pandas, que son esenciales para procesar y manipular datos efectivamente. Cuando se trata de grandes volúmenes de información, recurrimos a Apache PySpark, que es una plataforma que nos permite procesar la gran cantidad de datos optimizando recursos.

El análisis descriptivo y exploratorio

La estadística descriptiva tiene como objetivo la recopilación, presentación y análisis de datos. A través de distintas métricas como la frecuencia, el porcentaje y medidas de tendencia central es posible describir y caracterizar las propiedades más relevantes del conjunto de datos.

La estructura de los patrones espacio-temporales se logra mediante el análisis y la estructura de área geográfica en las provincias, cantones. Este enfoque permitirá visualizar y comprender cómo se comportan las variables en diferentes situaciones y momentos y encontrar patrones y comportamientos significativos que puedan influir en la toma de decisiones futuras.

Además, la segmentación sociodemográfica implica el estudio minucioso y exhaustivo de los datos en base a ciertas categorías. Por ejemplo, se consideran: sexo, grupo de edad y etnia.

Modelado predictivo

Series de Tiempo: Uso de Prophet como método de aprendizaje para obtener la proyección de las tendencias de homicidios y aprehensiones a corto y mediano plazo.

Validación: Separación de los elementos de la muestra completa a prueba y entrenamiento, medida por RSME, MAE.

Visualización de resultados

Construcción de paneles de control interactivos utilizando Tableau que contienen mapas, gráficos de barras, series de tiempo, diagramas circulares entre otras herramientas útiles para el análisis respectivo.

Filtros dinámicos: Basándonos en criterios filtrables dinámicos como el año, la provincia y el tipo de evento y la motivación, facilitamos a cualquier usuario la visualización de escenarios específicos.

1.3. Validación de la propuesta

La validación de la propuesta del presente trabajo se la realizó mediante el criterio de especialistas en el área de datos, esto con la finalidad de obtener criterios de personas con experiencia en el área de datos, esto se puede evidenciar en el Anexo 3.

Especialista 1

Validado por: STEVEN GABRIEL SIMBAÑA DELGADO

Título obtenido: INGENIERO EN SISTEMAS

C.I.: 1720437639

E-mail: simbanagabriel14@gmail.com

Institución de Trabajo: SECURITY DATA S.A

Cargo: INGENIERO DE DESARROLLO

Años de experiencia en el área: 3 AÑOS

Especialista 2

Validado por: JORGE LUIS CHINCHERO IZA

Título obtenido: MAGISTER EN TELECOMUNICACIONES MENCIÓN EN GESTION DE LAS
TELECOMUNICACIONES

C.I.: 1719310326

E-mail: jorge_fxf@hotmail.com

Institución de Trabajo: SECURITY DATA S.A

Cargo: INGENIERO HELPDESK SR

Años de experiencia en el área: 3 AÑOS

Especialista 3

Validado por: STEVEN JHOLAUS ALAVA GUANOPATIN

Título obtenido: INGENIERO ELECTRONICO

C.I.: 1726198110

E-mail: stnalava9396@gmail.com

Institución de Trabajo: SECURITY DATA S.A

Cargo: INGENIERO DE DATOS JR

Años de experiencia en el área: 3 AÑOS

1.4. Matriz de articulación de la propuesta

En la presente matriz se sintetiza la articulación del producto realizado con los sustentos teóricos, metodológicos, estratégicos-técnicos y tecnológicos empleados.

Tabla 5.
Matriz de articulación

EJES O PARTES PRINCIPALES	SUSTENTO TEÓRICO	SUSTENTO METODOLÓGICO	ESTRATEGIAS / TÉCNICAS	DESCRIPCIÓN DE RESULTADOS	INSTRUMENTOS APLICADOS
Comprensión del negocio	Enfoque de minería de datos	Definición de objetivos y requerimientos en CRISP-DM.	Análisis de la problemática de homicidios y aprehensiones en Ecuador.	Marco conceptual claro y objetivos del análisis predictivo	Revisión documental, normativa, reportes oficiales.
Comprensión de los datos	La importancia de la calidad y cantidad de datos (Han, Pei y Kamber, 2022).	Identificación de fuentes y exploración inicial de datos.	Revisión de registros oficiales de homicidios (2014–2024) y aprehensiones (2019–2024).	Identificación de variables clave: sexo, etnia, provincia, fecha, tipo de delito, grupo de edad, fuente de dato, motivaciones.	Python(pandas), Docker, Apache Pyspark.
Preparación de datos	Procesos basados en ETL en analítica (Kimball, 1996).	Fase de Data Preparation.	Limpieza, transformación, integración de datasets.	Base de datos consolidada y lista para análisis	Python, apache pyspark herramientas de normalización de dato
Modelado	Modelos de series temporales (Hyndman	Construcción del modelo Prophet con	Entrenamiento de modelos predictivos segmentados	Obtención de proyecciones para el	Jupyter Notebook, PySpark, Prophet,

	y Athanasopoulos, 2021)	estacionalidad anual y variables históricas (2020–2024).	por fuente de dato (homicidios y aprehensiones), provincia, grupo etario y motivaciones.	año 2025, incluyendo valores centrales (yhat) y bandas de confianza (yhat_lower)	exportación a CSV, integración con Tableau.
Evaluación	Métodos de validación de modelos predictivos	Se debe realizar la evaluación del modelo al contrastar las predicciones que se han realizado bajo el uso de métricas. (Jaramillo, 2021).	Medición con RMSE, MAE,	Jupyter Notebook, PySpark, Prophet, exportación a CSV, integración con Tableau.	Python, herramientas estadísticas.
Despliegue	Visualización de datos para toma de decisiones (Inmon, 2005).	Realización de los dashboards interactivos en Tableau para mostrar las predicciones y visualización de patrones	Construcción de dashboards interactivos.	Tableros con KPIs, mapas geográficos y filtros dinámicos.	Utilización de Tableau.

fuentes: Elaboración propia

CONCLUSIONES

Desde el punto de vista teórico, se pudo identificar que CRISP-DM es un método estructurado y adaptable para extraer datos y, como se utilizó ampliamente en proyectos de pronóstico, es adecuado. De manera similar, se comprobó que Apache Spark(Pyspark), Docker Desktop, Facebook Prophet y Tableau también son recursos aceptables para trabajar con grandes volúmenes de datos, pronósticos y visualización de conjuntos de datos, respectivamente.

Se pudo verificar que la metodología CRISP-DM es un marco estructurado y aplicable para la predicción de patrones delictivos, ya que organiza las fases de proceso en lógicas adecuadas desde la comprensión del negocio hasta la evaluación de los resultados. Además, el proceso realizado con PySpark, Prophet y Tableau son adecuados para el contexto ecuatoriano, ya que permiten tanto el procesamiento masivo de datos en un breve tiempo como la representación interactiva y comprensible para la toma de decisiones.

El diagnóstico de los datos como la fuente de los datos, homicidios y aprehensiones, provincia, grupo de edad, sexo, etnia, tipo de arma y motivación. El análisis histórico puso en perspectiva un futuro decreciente de las aprehensiones y un crecimiento acelerado de los homicidios, concentrados especialmente en personas entre 18 y 34 años de edad y en las provincias costeras de Guayas, Manabí y Los Ríos.

La etapa de preparación y modelado de datos, bajo los fundamentos de CRISP-DM, ha hecho posible transformar los datos históricos en fuentes para análisis predictivos. A través del modelo de Prophet se lograron construir proyecciones confiables que originan a un incremento de 10,000 casos más en homicidio para el año 2025 y una disminución de 69.000 registros en aprehensión. Esto ratifica que un buen uso de datos puede ser de gran uso para la analítica predictiva en cualquier tema de interés social.

RECOMENDACIONES

Dadas las variables críticas ya diagnosticadas como provincias, la edad, el sexo, la motivación, otros estudios futuros podrían incluir nuevas variables externas que influyan en el análisis delictivo, tales como variables socioeconómicas, educativas y laborales y factores de movilidad. Este enfoque permitiría crear modelos más sólidos que no solo estarían basados en la tasa de criminalidad, sino que también incluirían variables explicativas que ayudarían a comprender las razones subyacentes de la violencia.

Para futuras investigaciones, se recomienda realizar análisis de escenarios a más largo plazo y estudiar más a fondo la evaluación de políticas públicas alternativas en diferentes supuestos. También se sugiere desarrollar sistemas de alerta temprana mediante el uso de modelos predictivos como Prophet para apoyar la toma de decisiones basadas en el tiempo.

Se recomienda en futuras investigación profundizar la comparación de Prophet con otros modelos de predicción de series temporales, como ARIMA, LSTM y otros enfoques de machine learning, para evaluar su rendimiento en contextos complejos delictivos; y expandir la aplicación de CRISP-DM a otros fenómenos sociales, estableciendo el modelo CRISP-DM como una práctica estándar de análisis en el contexto ecuatoriano.

Se recomienda continuar la validación participativa de los resultados, extendiendo la consulta a otros actores sociales, entre ellos autoridades locales, organizaciones comunitarias y entes académicos.

BIBLIOGRAFÍA

Instituto Nacional de Estadística y Censos (INEC). (2024). *Anuario de estadísticas de criminalidad 2023*. <https://www.ecuadorencifras.gob.ec>

Ministerio del Interior del Ecuador. (2024). *Operación estadística de homicidios intencionales 2014–2024 [Base de datos]*. <https://www.datosabiertos.gob.ec/>

PNUD. (2023). *Informe sobre seguridad ciudadana y desarrollo humano en América Latina y el Caribe. Programa de las Naciones Unidas para el Desarrollo*. <https://www.undp.org/es/latin-america>

CEPAL. (2022). *La seguridad ciudadana en América Latina: desafíos para una agenda pública regional. Comisión Económica para América Latina y el Caribe*. <https://www.cepal.org/es/publicaciones>

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.

Chen, H., Chiang, R. H. L., Storey, V. C. (2012). *Business Intelligence and Analytics: From Big Data to Big Impact*. *MIS Quarterly*, 36(4), 1165–1188. <https://doi.org/10.2307/41703503>

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., y Brantingham, P. J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110(512), 1399–1411.

González, D., Medina, H., y Rosero, A. (2020). *Análisis geoespacial del crimen en ciudades latinoamericanas: oportunidades y desafíos*. *Revista Latinoamericana de Seguridad*, 12(1), 45–61.

Hernández-Sampieri, R., Fernández-Collado, C., y Baptista-Lucio, P. (2014). *Metodología de la investigación* (6.ª ed.). McGraw-Hill.**

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., y Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM consortium.

Hernández-Sampieri, R., Fernández-Collado, C., y Baptista-Lucio, P. (2014). *Metodología de la investigación* (6.ª ed.). McGraw-Hill.

Otzen, T., y Manterola, C. (2017). *Técnicas de muestreo sobre una población a estudio*. *International Journal of Morphology*, 35(1), 227–232. <https://doi.org/10.4067/S0717-95022017000100037>

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., y Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM Consortium.

Shmueli, G., y Koppius, O. R. (2011). *Predictive analytics in information systems research*. *MIS Quarterly*, 35(3), 553–572.

Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239), 2.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Murray, D. (2013). *Tableau your data!: Fast and easy visual analysis with Tableau software*. Wiley.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., y Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11), 56–65. <https://doi.org/10.1145/2934664>.

Taylor, S. J., y Letham, B. (2018). *Forecasting at scale*. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., y Ljung, G. M. (2016). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., y Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Wirth, R., y Hipp, J. (2000). CRISP-DM: *Towards a standard process model for data mining*. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39.

ANEXOS

ANEXO 1



ANEXO 2

RESULTADOS EN TABLEU SOBRE EL TEMA APLICACIÓN DE LA METODOLOGÍA CRISP-DM PARA EL ANÁLISIS PREDICTIVO Y VISUALIZACIÓN DE PATRONES DE HOMICIDIOS Y APREHENSIONES EN ECUADOR ENTRE 2020 Y 2024 MEDIANTE TÉCNICAS DE BIG DATA

Realizado: Bryan Gualoto

**UNIVERSIDAD TECNOLÓGICA ISRAEL
2025**

Descripción de la propuesta

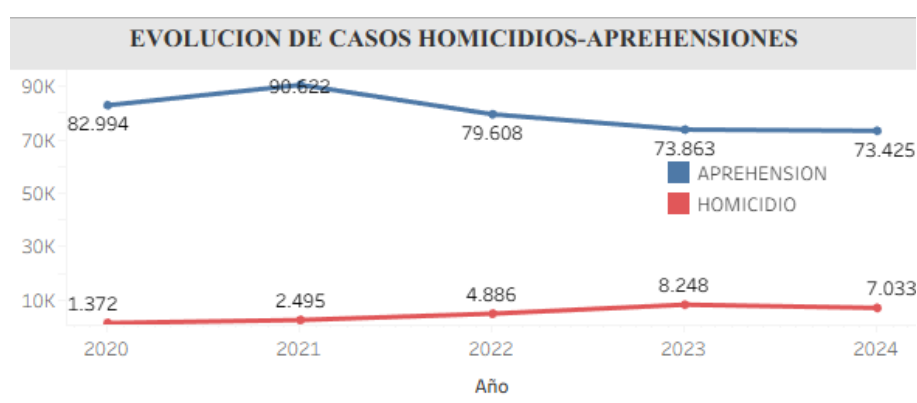
En el siguiente tablero se muestra dos dashboards, el uno abarca en la página principal la visualización la cual es clave para identificar y analizar las tendencias más relevantes sobre homicidios y aprehensiones en Ecuador entre 2020 y 2024 y en el segundo dashboard el análisis predictivo con respecto al incremento de homicidios y aprehensiones para el año 2025

PAGINA PRINCIPAL

Para la presentación del tablero se ha planteado varias preguntas para poder identificar patrones de tendencias, mediante la recolección de datos históricos obtenidos que abarcan desde el 2020. hasta el 2024 tanto para aprehensiones como para homicidios para posteriormente visualizarlos.

Pregunta 1

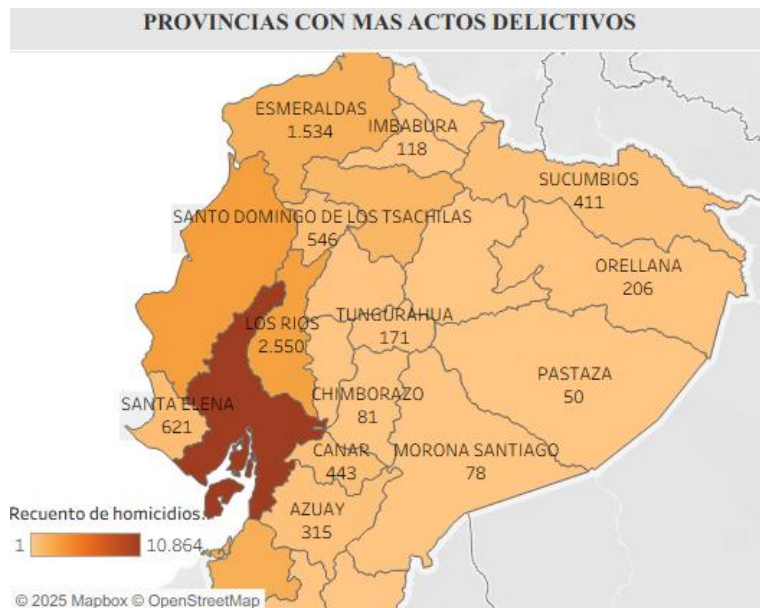
¿Cómo han evolucionado los homicidios y aprehensiones en Ecuador entre 2020 y 2024?



En cuanto a aprehensiones, se identificaron 82.994 en 2020 y 90.622 en 2021, cifras que se proyectaron a la baja los años subsiguientes: 79.608, 73.863 y 73.425. Se estima un descenso acumulado del 19% entre 2021 y 2024, los homicidios exhiben el comportamiento opuesto. Al finalizar el año base, 2020, se contabilizaron 1.372 casos, incrementando de manera constante para 2023 con 8.248. Esta tasa de homicidios fue más de cinco veces superior a la del año base, pese a que en 2024 manifestó una ligera tendencia a disminuir, registrándose 7.033, se encuentra muy por encima de la tasa en 2020. Según estas cifras, las aprehensiones en los últimos tres años muestran un descenso progresivo y los homicidios uno acelerado; en 2023, el incremento en los homicidios fue crítico. En este caso, verdaderas variables diferenciadas que favorecen la necesidad de medidas anticipadas.

Pregunta 2

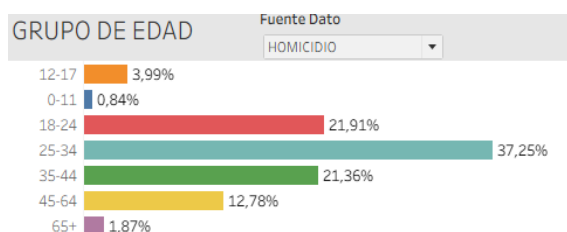
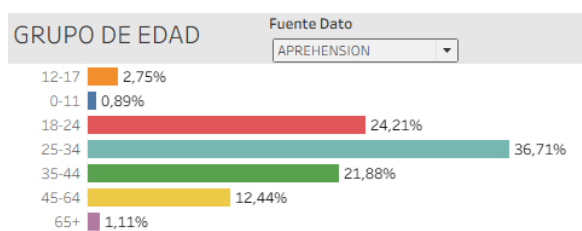
¿Qué provincias concentran la mayor cantidad de homicidios?



El análisis sobre los homicidios en Ecuador evidencia una concentración territorial marcada en el período analizado, presentando los siguientes resultados: Guayas lidera con la mayor cantidad de casos homicidios, superando los 10,000 homicidios y por lo tanto, marcando con claridad el epicentro de la violencia letal en el país, Manabí 2.704 casos y Los Ríos 2.550 casos son los segundos “focos críticos”, consolidando a la región litoral como el área con mayor incidencia de actos violentos. Otras provincias con cifras relevantes son Esmeraldas 1.534 casos y Santo Domingo de los Tsáchilas 1.388 casos, quienes poseen similitudes geográficas con las provincias anteriormente mencionadas y presentan niveles de violencia superiores al promedio nacional. La tendencia de las aprehensiones recoge un descenso progresivo en los últimos tres años, en cambio, los homicidios evidencian un incremento acelerado, llegando a un punto crítico en 2023.

Pregunta 3

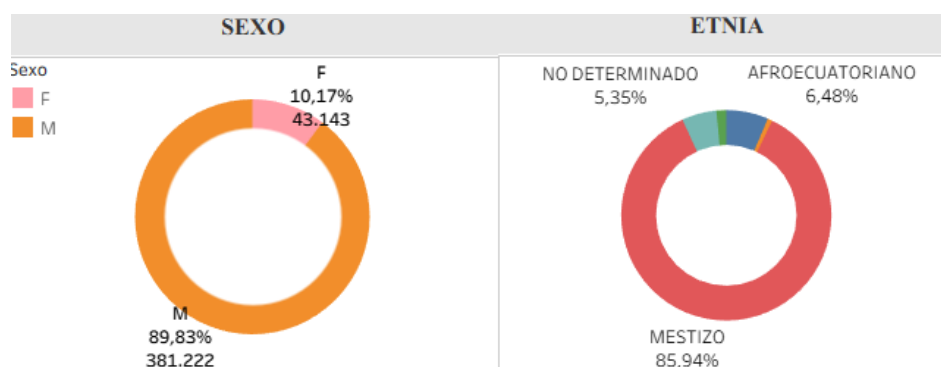
¿Qué grupos de edad han estado más involucrados en los actos delictivos entre 2020 y 2024?



El análisis realizado demuestran que la población joven-adulta 18 – 44 no es solo la más involucrada a ser víctima y agresora de homicidio, sino además ese rango de población es el más aprehendido por distintas motivaciones. Esto refleja que la violencia y la delincuencia están proporcionalmente dirigidas a un segmento de la población activa, lo que puede tener una amplia repercusión social y económica, así como en materia de seguridad.

Pregunta 4

¿Cuáles son las características de población que comete homicidios e infracciones?



El análisis revela una marcada diferencia entre hombres y mujeres en relación con los delitos registrados durante el período 2020–2024:

- Hombres (M): representan el 89,83% de los casos (381.222 registros).
- Mujeres (F): representan únicamente el 10,17% de los casos (43.143 registros).
- Esto indica que 9 de cada 10 casos están asociados al sexo masculino, lo que confirma que los hombres son los principales involucrados en actos delictivos, ya sea como víctimas o como procesados en aprehensiones.

En cuanto a la variable etnia, la información muestra que:

- El grupo mestizo concentra la gran mayoría de los registros, con un 85,94% de los casos.

- La población afro ecuatoriana representa el 6,48%, mientras que los casos con etnia no determinada suman un 5,35%.
- Otras etnias aparecen con porcentajes residuales, que no superan el 2% en conjunto.

Pregunta 5

¿Cuál son las principales motivaciones para casos de aprehensión en el periodo 2020-2025?

TOP 5 APREHENSIONES COMUNES					
Motivación	Fecha				
	2020	2021	2022	2023	2024
DELITOS CONTRA EL DERECHO DE PROPIEDAD PRIVADA (ROBOS)	14.826	14.792	14.350	14.188	12.391
BOLETAS	11.167	15.566	12.824	11.625	10.779
DELITOS POR LA PRODUCCIÓN Y TRÁFICO DE DROGAS	11.873	12.800	12.150	10.674	11.231
CONTRAVENCIONES DE TRÁNSITO	7.392	7.495	6.764	6.127	7.046
CONTRAVENCION DE VIOLENCIA INTRAFAMILIAR	6.424	6.980	5.709	4.761	4.502

En base a los delitos que se repiten en las motivaciones de aprehensiones entre 2020 y 2024, es posible identificar ciertas tendencias en los delitos que sostienen mayor regularidad a lo largo de los años. En primer lugar, los delitos contra el derecho de propiedad privada (robos), que en todos los años analizados se encuentran como mayor causa de aprehensión con montos entre 14.826 en 2020 y 12.391 en 2024. A pesar de su tendencia decreciente mantienen el primer lugar. Las boletas están en segundo lugar, con un máximo en 15.566 casos en 2021 y un decrecimiento mínimo a 10.779 en 2024.

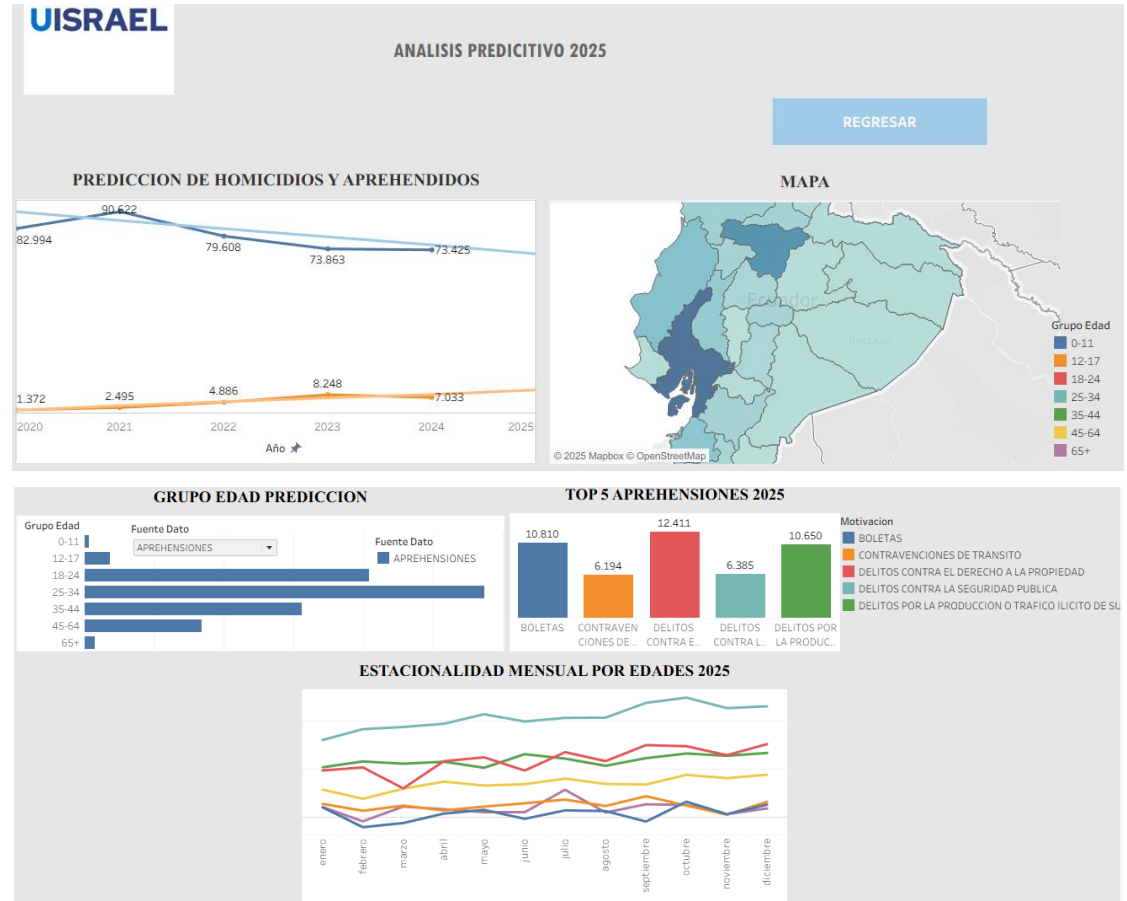
Los delitos por la producción y tráfico de drogas se encontrarán en tercer lugar con valores oscilantes entre 11.000-12.800 casos por año, lo cual da cuenta de la importancia del narcotráfico en la dinámica delictiva. La cantidad de contravenciones de tránsito es más baja, con valores entre 7.392 en 2020 y 7.046 en 2024, presentando cierta estabilidad en el período. Por último, las contravenciones por violencia intrafamiliar presentan, en base al TOP 5, la menor cantidad de casos, con un pico en 6.980 en 2021 y una baja a 4.502 en 2024.

Los resultados indican que, a pesar de pequeñas variaciones interanuales, el patrón de motivaciones de aprehensión se mantiene estable en el período 2020–2024. Esto revela que los delitos contra la propiedad y el narcotráfico son problemáticas estructurales del país, mientras que factores sociales como la violencia intrafamiliar también tienen un peso considerable dentro de las aprehensiones comunes.

DASHBOARD PREDICCIONES

El dashboard de predicción 2025 brinda una proyección de la evolución probable de homicidios y aprehensiones en Ecuador, cuyos resultados se pueden explorar a nivel nacional y a nivel provincial. A través del mapa interactivo, los usuarios pueden visualizar cómo se

distribuyen los casos en relación con los grupos etarios en cada provincia y se pueden identificar las TOP 5 motivaciones de aprehensiones proyectadas para el futuro. Asimismo, el análisis de la estacionalidad por edades establece en qué meses del año hay una mayor concentración de delitos. Estos modelos visualizan un enfoque corto y ágil los conceptos claves para apoyar decisiones en seguridad ciudadana.



ANEXO 3

VALIDACION DE LOS ESPECIALISTAS

UNIVERSIDAD TECNOLÓGICA ISRAEL

ESCUELA DE POSGRADOS "ESPOG"

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

INSTRUMENTO PARA VALIDACIÓN DE LA PROPUESTA

Estimado colega:

Se solicita su valiosa cooperación para evaluar la calidad del siguiente contenido digital "Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data ". Sus criterios son de suma importancia para la realización de este trabajo, por lo que se le pide que brinde su cooperación contestando las preguntas que se realizan a continuación.

Datos informativos

Validado por: STEVEN GABRIEL SIMBAÑA DELGADO
Título obtenido: INGENIERO EN SISTEMAS
C.I.: 1720437639
E-mail: simbanagabriel14@gmail.com
Institución de Trabajo: SECURITY DATA S.A
Cargo: INGENIERO DE DESARROLLO
Años de experiencia en el área: 3 AÑOS

Instructivo:

- Responda cada criterio con la máxima sinceridad del caso.
- Revisar, observar y analizar la propuesta de la plataforma virtual, blog o sitio web.
- Coloque una X en cada indicador, tomando en cuenta que Muy adecuado equivale a 5, Bastante Adecuado equivale a 4, Adecuado equivale a 3, Poco Adecuado equivale a 2 e Inadecuado equivale a 1.

Tema: "Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data".

Indicadores	Muy adecuado	Bastante Adecuado	Adecuado	Poco adecuado	Inadecuado
Pertinencia	X				
Aplicabilidad		X			
Factibilidad		X			
Novedad	X				
Fundamentación pedagógica		X			
Fundamentación tecnológica	X				
Indicaciones para su uso	x				
TOTAL	20	12			

Observaciones: El trabajo es de gran aporte al público en general ya que busca que la ciudadanía y los interesados tengan referencias con respecto a delitos públicos.

Recomendaciones: Para futuros trabajos se puede tomar como referencia distintas fuentes de datos que tengan el mismo enfoque y poder procesarlas.

Lugar, fecha de validación: Quito 24 de ago. de 25



Firma del especialista
Steven Simbaña

UNIVERSIDAD TECNOLÓGICA ISRAEL

ESCUELA DE POSGRADOS "ESPOG"

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

INSTRUMENTO PARA VALIDACIÓN DE LA PROPUESTA

Estimado colega:

Se solicita su valiosa cooperación para evaluar la calidad del siguiente contenido digital
"Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones
de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data
". Sus criterios son de suma importancia para la realización de este trabajo, por lo que se le pide
que brinde su cooperación contestando las preguntas que se realizan a continuación.

Datos informativos

Validado por: JORGE LUIS CHINCHERO IZA
Título obtenido: MAGISTER EN TELECOMUNICACIONES MENCIÓN EN GESTION DE LAS TELECOMUNICACIONE
C.I.: 1719310326
E-mail: jorge_fxf@hotmail.com
Institución de Trabajo: SECURITY DATA S.A
Cargo: INGENIERO HELPDESK SR
Años de experiencia en el área: 3 AÑOS

Instructivo:

- Responda cada criterio con la máxima sinceridad del caso.
- Revisar, observar y analizar la propuesta de la plataforma virtual, blog o sitio web.
- Coloque una X en cada indicador, tomando en cuenta que Muy adecuado equivale a 5, Bastante Adecuado equivale a 4, Adecuado equivale a 3, Poco Adecuado equivale a 2 e Inadecuado equivale a 1.

Tema: "Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data".

Indicadores	Muy adecuado	Bastante Adecuado	Adecuado	Poco adecuado	Inadecuado
Pertinencia	X				
Aplicabilidad		X			
Factibilidad		X			
Novedad	X				
Fundamentación pedagógica	X				
Fundamentación tecnológica	X				
Indicaciones para su uso	X				
TOTAL	25	8			

Observaciones: El presente trabajo se encuentra correctamente identificado por sus etapas las cuales ayudan a un entendimiento bastante amplio sobre el concepto que se desea manejar.

Recomendaciones: Sería interesante tomar en futuros trabajos otro tipo de metodología de análisis y hacer una comparación con la metodología de estudio del presente trabajo.

Lugar, fecha de validación: Quito 24 de ago. de 25



Jorge Luis Chinchero

123



Firma del especialista
Jorge Chinchero

UNIVERSIDAD TECNOLÓGICA ISRAEL
ESCUELA DE POSGRADOS "ESPOG"

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

INSTRUMENTO PARA VALIDACIÓN DE LA PROPUESTA

Estimado colega:

Se solicita su valiosa cooperación para evaluar la calidad del siguiente contenido digital
"Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones
de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data
". Sus criterios son de suma importancia para la realización de este trabajo, por lo que se le pide
que brinde su cooperación contestando las preguntas que se realizan a continuación.

Datos informativos

Validado por: STEVEN JHOLAUS ALAVA GUANOPATIN
Título obtenido: INGENIERO ELECTRONICO
C.I.: 1726198110
E-mail: stnalava9396@gmail.com
Institución de Trabajo: SECURITY DATA S.A
Cargo: INGENIERO DE DATOS JR
Años de experiencia en el área: 3 AÑOS

Instructivos:

- Responda cada criterio con la máxima sinceridad del caso.
- Revisar, observar y analizar la propuesta de la plataforma virtual, blog o sitio web.
- Coloque una X en cada indicador, tomando en cuenta que Muy adecuado equivale a 5, Bastante Adecuado equivale a 4, Adecuado equivale a 3, Poco Adecuado equivale a 2 e Inadecuado equivale a 1.

Tema: "Aplicación de la metodología CRISP-DM para el análisis predictivo y visualización de patrones de homicidios y aprehensiones en Ecuador entre 2020 y 2024 mediante técnicas de Big Data".

Indicadores	Muy adecuado	Bastante Adecuado	Adecuado	Poco adecuado	Inadecuado
Pertinencia	X				
Aplicabilidad		X			
Factibilidad		X			
Novedad	X				
Fundamentación pedagógica		X			
Fundamentación tecnológica	X				
Indicaciones para su uso	X				
TOTAL	20	12			

Observaciones: la integración de tecnologías nuevas como Docker, PySpark y Prophet también evidencia un enfoque moderno y escalable. Las presentes herramientas con las que se trabajó aseguran no solo una limpieza y procesamiento limpio de los datos masivos mismos, sino también como prevención en caso de escalamiento.

Recomendaciones: Para futuros trabajos se puede tomar como referencia fuentes de datos internacionales para un proyecto de igual magnitud.

Lugar, fecha de validación: Quito 24 de ago. de 25



Steven Alava Alava
Quito, Ecuador
24 de agosto de 2025

Firma del especialista
Steven Alava