Universidad Nacional de Loja

Facultad de la Energía, las Industrias y los Recursos Naturales No Renovables

Carrera de Computación

Optimización Bayesiana en modelos de clasificación: Árbol de Decisión y Support Vector Machine para determinar mediante Minería de Datos patrones en los asesinatos de la Zona 8 del Ecuador

Bayesian Optimization in classification models: Decision Tree and Support Vector Machine to determine through Data Mining patterns in the murders in Zone 8 of Ecuador

Trabajo de Integración Curricular, previa a la obtención del título de Ingeniera en Ciencias de la Computación

AUTORA:

Cecilia Fernanda Trueba Reyes

DIRECTORA:

Ing. Genoveva J. Suing-Albito, Mg.Sc

Loja – Ecuador

2025

Certificación de director

Ing, Genoveva Jackelinne Suing Albito, Mg.Sc.
DIRECTORA DEL TRABAJO DE INTEGRACIÓN CURRICULAR

CERTIFICO:

Que he revisado y orientado todo el proceso de elaboración del Trabajo de Integración Curricular denominado: **Optimización Bayesiana en modelos de clasificación: Árbol de Decisión y Support Vector Machine para determinar mediante Minería de Datos patrones en los asesinatos de la Zona 8 del Ecuador,** previo a la obtención del título de **Ingeniera en Ciencias de la Computación,** de la autoría de la estudiante **Cecilia Fernanda Trueba Reyes,** con cédula de identidad **1106087883**, una vez que se determina que el trabajo cumple con todos los requisitos exigidos por la Universidad Nacional de Loja, para el efecto, autorizo la presentación del mismo para su respectiva sustentación y defensa.

Firmado electrónicamente por:
GENOVEVA JACKELINNE SUING ALBITO

……

Ing. Genoveva Jackelinne Suing Albito, Mg.Sc.
DIRECTORA DEL TRABAJO DE INTEGRACIÓN CURRICULAR

Autoría

Yo, **Cecilia Fernanda Trueba Reyes**, declaro ser autora del presente Trabajo de Integración Curricular y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos, de posibles reclamos y acciones legales, por el contenido del mismo. Adicionalmente acepto y autorizo a la Universidad Nacional de Loja la publicación de mi Trabajo de Integración Curricular, en el Repositorio Digital Institucional – Biblioteca Virtual.

Firma: 

Cédula de identidad: 1106087883

Fecha: 06 de febrero de 2025

Correo electrónico: cecilia.trueba@unl.edu.ec

Teléfono: 0960934663

Carta de autorización por parte del autor, para consulta, reproducción parcial o total, y/o publicación electrónica del texto completo del Trabajo de Integración Curricular

Yo, **Cecilia Fernanda Trueba Reyes**, declaro ser el autor del Trabajo de Integración Curricular denominado: **Optimización Bayesiana en modelos de clasificación: Árbol de Decisión y Support Vector Machine para determinar mediante Minería de Datos patrones en los asesinatos de la Zona 8 del Ecuador**, autorizo al sistema Bibliotecario de la Universidad Nacional de Loja para que, con fines académicos, muestre la producción intelectual de la Universidad, a través de la visibilidad de su contenido en el Repositorio Institucional.

Los usuarios pueden consultar el contenido de este trabajo en el Repositorio Institucional, en las redes de información del país y del exterior con las cuales tenga convenio la Universidad.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia del Trabajo de Integración Curricular que realice un tercero.

Para constancia de esta autorización, suscribo, en la ciudad de Loja, a los seis días del mes de febrero de dos mil veinticinco.

Firma: Firmado electrónicamente por: CECILIA FERNANDA TRUEBA REYES

Autora: Cecilia Fernanda Trueba Reyes
Cédula de identidad: 1106087883
Dirección: Loja, Benjamín Pereira y Alfredo Mora
Correo electrónico: cecilia.trueba@unl.edu.ec
Teléfono: 0960934663

DATOS COMPLEMENTARIOS:

Director del Trabajo de Integración Curricular: Ing. Genoveva Suing Albito, Mg.Sc.

Dedicatoria

Dedico este trabajo de titulación a mis padres Gustavo Trueba Valdivieso y Cecilia Reyes Luna, quienes siempre me han brindado su amor y son mis razones para continuar. A mis abuelitos, Flavio Reyes e Isabel Luna, Gustavo Trueba y Fanny Valdivieso quienes desde niña me han apoyado y ayudado con su cariño todos los días de mi vida. A mis tíos que han estado conmigo siempre y me han dado sus consejos y enseñanzas a lo largo de mi camino. A mis docentes, quienes me han brindado conocimientos para poder culminar mis estudios y finalizar este trayecto de vida universitaria. Agradezco de todo corazón a todos los que han formado parte de esta etapa universitaria, por su ayuda y motivación para poder graduarme.

Cecilia Fernanda Trueba Reyes

Agradecimiento

Agradezco primeramente a Dios y a su hijo Jesucristo por todo lo que tengo, por la vida que me ha dado para poder cumplir mis metas. A la Universidad Nacional de Loja quien me ha permitido completar mis estudios universitarios en su prestigiosa institución, a la Facultad y a la Carrera de Ingeniería en Computación por permitirme obtener los conocimientos adecuados para poder desempeñarme en mi futuro profesional. A mi directora y docente Ing. Genoveva Suing, por su ayuda y conocimientos para poder culminar con éxito este presente Trabajo de Integración Curricular. Agradeciendo por el apoyo incondicional a mis padres y demás familiares, quienes me han ayudado a forjar mi camino y lograr cumplir el objetivo de graduarme y poder ser a futuro un buen profesional de la Ingeniería en Computación.

Cecilia Fernanda Trueba Reyes

Índice de Contenidos

Índice de Tablas

Índice de Figuras

Índice de Ecuaciones

Índice de Anexos

1. Título

Optimización Bayesiana en modelos de clasificación: Árbol de Decisión y Support Vector Machine para determinar mediante Minería de Datos patrones en los asesinatos de la Zona 8 del Ecuador.

Bayesian Optimization in classification models: Decision Tree and Support Vector Machine to determine through Data Mining patterns in the murders in Zone 8 of Ecuador.

2. Resumen

Este estudio se centró en mejorar los modelos de minería de datos mediante técnicas como la Optimización Bayesiana, con el objetivo de encontrar los hiperparámetros adecuados para configurar los clasificadores Árbol de Decisión (DT) y Support Vector Machine (SVM), además, se enfocó en determinar patrones en los asesinatos ocurridos en la Zona 8 del Ecuador, siguiendo la metodología CRISP-DM, que incluye las fases de comprensión del negocio, comprensión de los datos, preparación del dataset, modelado, evaluación y despliegue. El modelo SVM, configurado específicamente con la librería Optuna, obtuvo los mejores resultados, alcanzando un 83,78% con la validación cruzada en las tres métricas clave que fueron precisión, accuracy y recall, mientras que al evaluarlo con sklearn.metrics, el modelo alcanzó un 84,12% para cada una de estas métricas; este modelo permitió identificar patrones significativos, como que el arma de fuego es la más utilizada en los asesinatos, estos crímenes ocurren principalmente en áreas urbanas mayormente en la vía pública, los días sábados y domingos entre las 19:00 pm y las 00:59 am, especialmente en los Distritos de Nueva Prosperina, Sur y Pascuales, en cuanto a las víctimas de los asesinatos suelen ser hombres que sí presentan antecedentes penales, en edades que van entre 20 y 50 años. Con estos patrones encontrados, se proporcionó información valiosa a los organismos encargados, lo que contribuyó a una mejor comprensión del fenómeno estudiado. Además, este estudio evidenció que la Optimización Bayesiana mejora considerablemente los modelos de clasificación al ajustar sus hiperparámetros, incrementando los porcentajes de rendimiento, como la precisión, exactitud, sensibilidad, y fortaleciendo así su aplicabilidad práctica.

Palabras clave: CRISP-DM, Hiperparámetros, Configuración, Predicción, Algoritmos, Precisión.

Abstract

This study focused on improving data mining models through techniques such as Bayesian Optimization, with the objective of finding the appropriate hyperparameters to configure the Decision Tree (DT) and Support Vector Machine (SVM) classifiers, in addition, it focused on determining patterns in the murders occurred in Zone 8 of Ecuador, following the CRISP-DM methodology, which includes the phases of understanding the business, understanding the data, preparing the dataset, modeling, evaluation and deployment. The SVM model, specifically configured with the Optuna library, obtained the best results, reaching 83.78% with the cross validation in the three key metrics that were precision, accuracy and recall, while when evaluated with sklearn.metrics, the model reached 84.12% for each of these metrics; this model allowed identifying significant patterns, such as the firearm is the most used in the murders, these crimes occur mainly in urban areas mostly on public roads, on Saturdays and Sundays between 19:00 pm and 00:59 am, especially in the Districts of Nueva Prosperina, Sur and Pascuales, as for the victims of the murders are usually men who do have criminal records, in ages ranging between 20 and 50 years. With these patterns found, valuable information was provided to the agencies in charge, which contributed to a better understanding of the phenomenon studied. Furthermore, this study evidenced that Bayesian Optimization considerably improves classification models by adjusting their hyperparameters, increasing performance percentages, such as precision, accuracy, sensitivity, and thus strengthening their practical applicability.

Keywords: CRISP-DM, Hyperparameters, Configuration, Prediction, Algorithms, Precision.

3. Introducción

La Optimización Bayesiana (OB) es una técnica que actualmente se usa en el campo del aprendizaje supervisado para mejorar el rendimiento de los modelos de minería de datos, ya que permite encontrar los hiperparámetros adecuados que los configuran, superando a otros algoritmos de optimización [1]. La técnica de OB tiene una orientación probabilística para realizar las búsquedas, aprende continuamente y predice los valores favorables, de esta manera reduce el número de iteraciones que se necesitan para encontrar los mejores hiperparámetros [2], se la usa para examinar y optimizar las funciones objetivo (precisión, accuracy, recall, F1-Score, entre otras) [3], convirtiéndose en una herramienta adecuada para tratar los problemas de aprendizaje automático y optimización [4].

El uso de modelos de minería de datos como el Árbol de Decisión (DT) y Support Vector Machine (SVM), permiten evaluar las actividades delictivas de forma más efectiva al realizar clasificaciones y determinar patrones relacionados con los delitos [5].

Este estudio ofrece varios beneficios, como el análisis de grandes volúmenes de datos mediante técnicas de minería de datos. Al aplicar la Optimización Bayesiana en modelos como el Árbol de Decisión (DT) y el Support Vector Machine (SVM), se mejora la eficiencia en la selección de hiperparámetros que configuran a los clasificadores, lo que facilita investigaciones que presentan limitaciones de tiempo y recursos, así también, al analizar datos sobre crímenes, se identifican patrones que contribuyen a los organismos a una toma de decisiones más informada y planificada.

En trabajos previos como [30], [56], [57], [58], [59], se aplica la técnica de Optimización Bayesiana en modelos de minería de datos, estas investigaciones realizan comparaciones entre los resultados obtenidos antes y después de aplicar la optimización, demostrando un aumento en los porcentajes de desempeño al implementar esta técnica, sin embargo trabajan sobre un conjunto de datos diferente al utilizado en el presente TIC, por lo que este estudio se presenta como un referente sobre la aplicación de minería de datos junto con el uso de la técnica de Optimización Bayesiana, dirigida especialmente a la problemática de crímenes como los asesinatos que ocurren en la Zona 8 del Ecuador.

El actual Trabajo de Integración Curricular responde a la siguiente pregunta investigación: ¿Qué porcentaje de precisión se alcanza en los modelos Árbol de Decisión y SVM al usar la técnica de Optimización Bayesiana para configurarlos y encontrar, mediante minería de datos, los patrones de comportamiento en los asesinatos de la Zona 8 del Ecuador, correspondientes a los años de enero 2015 a febrero de 2024?, para responder esta interrogante, se plantea el siguiente objetivo general: Aplicar la técnica de Optimización Bayesiana para encontrar los mejores hiperparámetros de los modelos Árbol de Decisión y SVM, determinado mediante minería de datos los patrones de los asesinatos de la Zona 8 del

Ecuador, periodo enero 2015 - febrero 2024, siguiendo la metodología CRISP-DM, la cual incluye las fases de comprensión del negocio, comprensión de los datos, preparación del dataset, modelamiento, evaluación y despliegue. Este objetivo se cumple basado en los dos objetivos específicos propuestos: Mejorar la calidad del dataset de asesinatos del Ecuador mediante técnicas de preparación de datos, realizar el análisis exploratorio de datos (EDA) e implementación de los clasificadores optimizados Árbol de Decisión y SVM y Evaluar los clasificadores Árbol de Decisión y SVM con la métrica de precisión.

El alcance de este proyecto se centra en demostrar cómo el uso de la técnica de Optimización Bayesiana simplifica y automatiza el proceso de encontrar los mejores valores para los hiperparámetros, los cuales configuran los modelos clasificadores. Estas configuraciones son determinantes para los resultados en métricas como precisión, exactitud y recall al momento de evaluarlas. Además, este trabajo lleva a cabo un proceso de minería de datos, el cual permite descubrir información y patrones que no son fácilmente visibles a simple vista debido al extenso número de registros en la base de datos. Este análisis se enfoca en los asesinatos registrados en la Zona 8 del Ecuador, que incluye los cantones de Guayaquil, Durán y Samborondón, área que presenta el mayor número de registros de asesinatos en comparación con otras regiones del país, proporcionando información y conocimientos que permiten a organismos como la Policía Nacional del Ecuador aplicar mejoras en su planificación y toma de decisiones para enfrentar estos crímenes.

Las limitaciones presentes en este estudio incluyen la poca cantidad de recursos computacionales disponibles, como es el caso de la GPU, así como el tiempo prolongado requerido para obtener la base de datos de una fuente confiable, ya que es necesario esperar la respuesta de los organismos responsables para proporcionar la base de datos correspondiente a un período extenso, que incluye desde enero 2015 hasta febrero de 2024.

4. Marco Teórico

La presente sección describe los temas y conceptos que servirán para comprender la información concerniente al Trabajo de Integración Curricular (TIC). Los temas que se describieron son: Antecedentes, donde se detallaron conceptos relacionados a la problemática de los crímenes en Ecuador, especialmente los asesinatos y patrones de comportamiento de estos. En la fundamentación teórica, se trató las definiciones relacionadas a la minería de datos, modelos, metodología, técnica de Optimización Bayesiana y sobre las métricas que se usaron para la evaluación. En las herramientas de desarrollo se explicaron conceptos de las tecnologías y en la sección de trabajos relacionados, el conjunto de todas las investigaciones similares que se usaron para que sean un referente del actual TIC.

4.1. Antecedentes

El actual Trabajo de Integración Curricular está enfocado en determinar patrones de comportamiento en los asesinatos en la zona 8 del Ecuador, mediante modelos de clasificación de minería de datos, ya que los mismos se utilizan en proyectos donde se desee encontrar o descubrir información difícil de observar a simple vista. Además de estar enfocado en aplicar una técnica de Optimización Bayesiana que es útil para que modelos de clasificación de minería de datos puedan ser configurados correctamente de una forma automatizada, sin la necesidad de realizarlo de manera manual.

4.1.1. Homicidios Intencionales

Los homicidios intencionales son los actos que causan la muerte a una persona por parte del victimario, dentro de estos crímenes existen 4 tipos de muertes: homicidios, femicidios, sicariatos y asesinatos, siendo el último el crimen más frecuente en el país ecuatoriano [6]. En el año 2023 en Ecuador se tuvo un total de 8000 homicidios intencionales, duplicando el número en comparación con el año 2022 [7]. La provincia del Guayas tiene los cantones de Guayaquil, Samborondón y Durán con la mayor tasa de homicidios intencionales, especialmente de asesinatos como el tipo de muerte [8].

4.1.2. Asesinatos

Se considera asesinato al hecho de matar a otra persona a traición [9]. Otras características que están dentro de estos crímenes son el haber cometido el crimen por un monto o recompensa, con impiedad es decir generando un dolor extremo a la víctima o para impedir que un crimen diferente se llegue a saber [10]. En el país ecuatoriano el tipo de crimen que más se comete es el asesinato [11], por lo que este sigue siendo un fenómeno social preocupante en la población debido a la inseguridad que genera.

4.1.3. Patrones de Comportamiento

Los patrones de comportamiento en los asesinatos se definen como las características que se repiten en la realización de estos crímenes. Estos patrones pueden evidenciar

información relacionada con el perpetrador del hecho, así como sus motivaciones, métodos, localización e información relacionada con el perfil de la víctima [12]. Es viable descubrir patrones de comportamiento de crímenes mediante los registros que se han ido almacenando por varios años, para poder generar respuestas policiales efectivas que frenen estos actos frecuentes. Por tal motivo analizar estos datos es necesario para poder prevenir los crímenes tales como los asesinatos.

En Ecuador existen análisis e informes estadísticos con herramientas básicas pero gran parte de las veces no se refleja el correcto contraste de las variables, por eso hacer uso de algoritmos que clasifiquen y den predicciones sobre el comportamiento futuro de estos crímenes puede ser una herramienta valiosa al momento de combatir este problema [13]. Al obtener los patrones de comportamiento se puede analizar y comprender el modo como operan los asesinos, evidenciando sus formas de actuar, por ende, los patrones son actos fijos que muestran en la forma de actuar de las personas ante ciertas circunstancias [14].

4.2. Fundamentación Teórica

4.2.1. Minería de Datos (DM)

La minería de datos es un proceso en el cual se realiza el análisis y procesamiento de una gran cantidad de datos, mediante modelos específicos que permiten realizar tareas de clasificación y predicción [15]. La minería de datos permite trabajar con grandes volúmenes de registros, para poder obtener la información necesaria, es decir conocimiento que no es visible a simple vista y que se encuentra oculta en un dataset [16].

La minería de datos es una técnica básica porque nos brinda un conjunto de procedimientos para obtener patrones e información necesaria ya que está relacionada con la estadística, pero en este caso todo se realiza de manera automatizada mediante algoritmos o modelos. La minería de datos consta de varios procesos que van desde la obtención de datos, su limpieza y uso de modelos o algoritmos capaces de clasificar o predecir se ha vuelto muy utilizada para las empresas y organizaciones ya que permiten anticipar comportamientos sobre un determinado tema o situación repetitiva [17]. La minería de datos, en los proyectos que se desean descubrir patrones sobre un conjunto de datos, para poder determinar información relacionada a los asesinatos realizados en el Ecuador.

4.2.2. Inteligencia Artificial (IA)

La inteligencia artificial (IA) consiste en crear sistemas que puedan simular la inteligencia humana para poder resolver los problemas y tareas específicas, van aprendiendo continuamente conforme se los entrena [18]. La inteligencia artificial que está presente en el aprendizaje supervisado y en los modelos de minería de datos permiten extraer patrones avanzados [19]. La IA permite a las máquinas aprender continuamente para después tomar decisiones, esto evidencia la similitud con el humano [20]. La inteligencia artificial que se encuentra en los modelos y técnicas de optimización, facilitan el descubrimiento de

características ocultas, mejorando el análisis de datos y la automatización de procesos de minería de datos.

4.2.3. Aprendizaje Supervisado

El aprendizaje supervisado se encuentra dentro de las técnicas del aprendizaje automático, es considerada una de las más efectivas y una de las más usadas en la minería de datos para realizar el procesamiento de la base de datos [21]. El aprendizaje supervisado permite realizar el entrenamiento de los modelos de clasificadores, mediante datos que se encuentran etiquetados, para cada entrada está vinculada una salida que está etiquetada, permitiendo al modelo generar predicciones y clasificaciones sobre los datos que se hayan escogido, ayudando a obtener y descubrir información de manera confiable [22]. Permitiendo obtener clasificaciones más exactas en comparación con otros tipos de aprendizajes, ya que los modelos operan con datos que ya se encuentran dentro de una clase [23].

4.2.4. Árbol de Decisión

El árbol de decisión es un modelo de clasificación de minería de datos, tiene forma de árbol y sirve para clasificar y determinar patrones sobre un dataset [24]. El árbol de decisión permite realizar tareas de clasificación como de regresión en el mundo de la minería de datos [25], es un diagrama de flujo, su estructura parecida a un árbol está compuesta de nodos internos que representan a las variables, las ramas son las condiciones y los nodos terminales son las etiquetas o clases asignadas por el modelo como se visualiza en la **Figura 1** obtenida de [26] se observa la representación del Árbol de Decisión en donde se encuentra la estructura descrita anteriormente, se observan sus nodos iniciales, intermedios y finales.

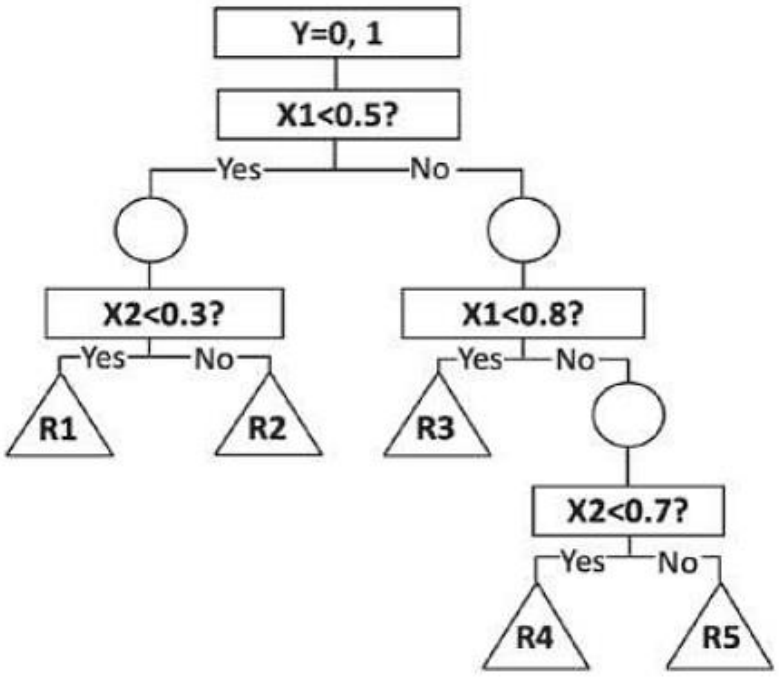


Figura 1. Árbol de Decisión simple.

4.2.5. Support Vector Machine

El Support Vector Machine (SVM) es un modelo de clasificación de minería de datos, usado para tareas de clasificación y regresión, cuenta con un conjunto de hiperparámetros, cuyo objetivo es encontrar un hiperplano que separe las clases lo más lejos posible, esto es el margen máximo que separa los punto lo más que se pueda [27]. El SVM usa diferentes kernels en su configuración para poder manejar datos y esto los hace menos vulnerables al sobreajuste, destacándose sobre otros modelos [28]. Este clasificador está incluido en los que procesan sus datos mediante el aprendizaje supervisado, funciona mediante el mapeo de puntos que son los datos originales del dataset [29]. En la **Figura 2** obtenida de [30] se muestra una representación del hiperplano del modelo SVM.

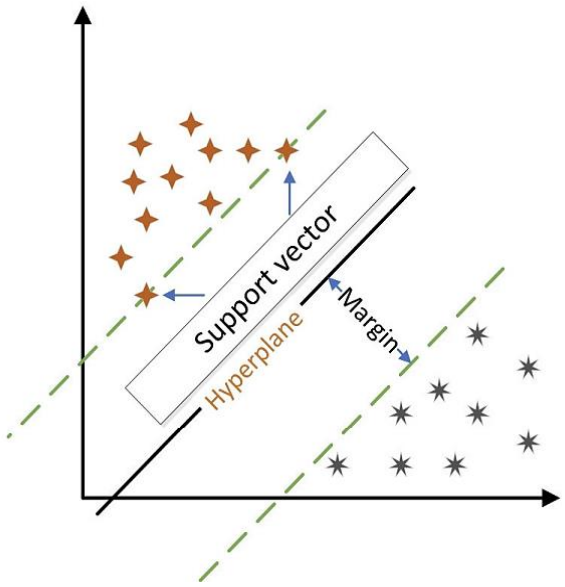


Figura 2. Hiperplano del modelo Support Vector Machine.

4.2.6. Optimización Bayesiana (OB)

La Optimización Bayesiana (OB) es una técnica en la cual mediante un conjunto de búsquedas e iteraciones se puede encontrar los mejores valores para los hiperparámetros de los modelos que posee varios de estos ajustes, de esta manera permite configurarlos adecuadamente [31]. Se basa en los valores que han tenido mejores resultados y fundamentándose en estos, busca otros que son probablemente adecuados y de esta manera encuentra el porcentaje más alto en sus métricas al momento de evaluarlos [32]. Esta técnica de optimización es adecuada cuando los modelos de minería de datos tienen bastantes hiperparámetros que configuran a estos clasificadores y al hacerlo manualmente sería complicado por lo que su función inteligente puede encontrar los valores óptimos de manera automatizada, siendo necesaria en los modelos Árbol de Decisión y SVM. Los hiperparámetros son un conjunto de valores que tienen los modelos para ajustarlos [33], estos pueden ser encontrados con la técnica de optimización Bayesiana para luego construir los modelos de clasificación de minería de datos adecuadamente.

❖ **Fórmula General de la Optimización Bayesiana**

Primeramente, la técnica de Optimización Bayesiana (OB) genera un Modelo de Probabilidad mediante un Proceso Gaussiano (GP), de esta manera encuentra las mejores configuraciones que permitan encontrar el porcentaje más alto de la función objetivo, en esta investigación se trabajará con la precisión, exactitud y sensibilidad.

A continuación, en la **Ecuación 1** se encuentra la fórmula general de la optimización Bayesiana obtenida de [32].

Ecuación 1. Fórmula general de la Optimización Bayesiana.

$$f(x) \sim N(\mu(x), \sigma 2(x))$$

Donde:

- **f (x):** función objetivo.
- **N:** distribución normal.
- **x:** valores o puntos para evaluar.
- **μ(x):** media del modelo en el punto x.
- **σ^2(x):** varianza en el valor x.

Luego se genera una Función de Adquisición (EI), la cual se puede observar en la **Ecuación 2,** que fue obtenida de [32], en donde explora nuevos puntos que todavía no se han evaluado, para obtener los valores óptimos que mejoren la función objetivo.

Ecuación 2. Fórmula para la Función de Adquisición (EI).

$$EI(x) = (\mu(x) - fbest)\ \Phi(z) + \sigma(x)\phi(z)$$

Donde:

- ***f*best:** el valor óptimo de la función objetivo que se ha encontrado.
- **Φ(z):** función que acumula.
- ***ϕ*(z):** función de probabilidad.

Finalmente se realiza lo siguiente:

- Elección del punto siguiente.
- Evaluar $f(X_{t+1})$.
- Actualización.
- Repetición.

❖ **Librerías de Optimización Bayesiana**

Existen varias librerías que se encuentran disponibles en el lenguaje de programación Python que sirven para implementar la técnica de Optimización Bayesiana. A continuación, en la **Tabla 1,** basada de [34] se encuentran librerías que se han usado en investigaciones similares, se describieron dos de ellas y las que se implementaron en el actual trabajo, además se encuentran los enlaces de estas para poder conocer más sobre cómo están desarrolladas.

Tabla 1. Librerías de Optimización Bayesiana.

Nro.	Librería	Descripción	URL
1	Optuna	Usa la optimización bayesiana, mediante búsquedas aleatorias de valores para los parámetros.	https://optuna.org/
2	Bayesian-Optimization	Realiza búsquedas con un número límite de iteraciones, basada en un proceso Gaussiano.	https://github.com/bayesian-optimization/BayesianOptimization

4.2.7. Matriz de Confusión

La matriz de confusión sirvió para evaluar el rendimiento de los modelos de clasificación de minería de datos ya que permitió visualizar el desempeño del modelo al contrastar las predicciones que fueron realizadas. En la **Figura 3** se visualiza la matriz de confusión con sus cuatro valores principales: TP (verdaderos positivos), FP (falsos positivos, TN (verdaderos negativos), FN (falsos negativos), en la diagonal principal se encuentran las predicciones correctas que fueron realizadas por el modelo [35].

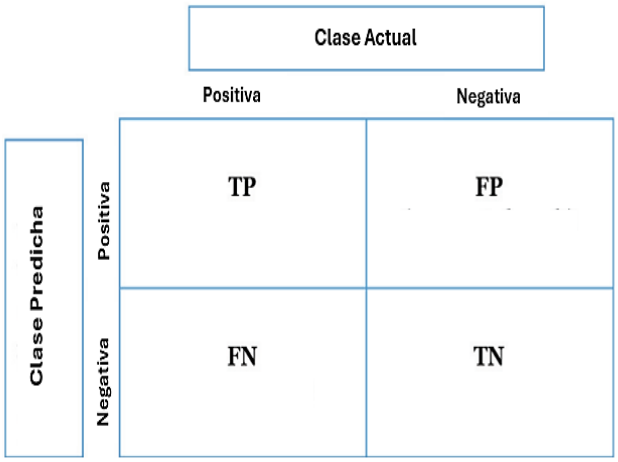


Figura 3. Composición de una Matriz de Confusión.

4.2.8. Precisión

La precisión es una métrica para evaluar los modelos de minería de datos, la cual sirve para medir la cantidad de verdaderos positivos sobre todos los positivos encontrados, de esta manera permite excluir los falsos positivos [36]. Esta métrica permite conocer el desempeño de un modelo al momento de realizar tareas de clasificación o regresión [37]. Se considera a la precisión como los porcentajes de las clasificaciones verdaderas que son realmente correctas [38]. Se pretende subir los porcentajes de precisión en los modelos de clasificación escogidos mediante la técnica de Optimización Bayesiana, para que estos puedan realizar clasificaciones más adecuadas y eficaces. A continuación, en la **Ecuación 3** se puede observar la fórmula matemática para la obtención de la precisión, la cual fue obtenida de [37].

Ecuación 3. Fórmula de cálculo de la precisión.

$$Precisión = \frac{TP}{TP + FP}$$

Donde:

- **TP:** Verdaderos Positivos.
- **FP:** Falsos Positivos.

4.2.9. Accuracy

La accuracy o exactitud es una métrica que evalúa la cantidad de predicciones correctas (verdaderos positivos y verdaderos negativos) con respecto al total de predicciones. A continuación, en la **Ecuación 4** se puede observar la fórmula matemática para el cálculo de la accuracy, la cual fue obtenida de [39].

Ecuación 4. Fórmula de cálculo del accuracy o exactitud.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Donde:

- **TP:** Verdaderos Positivos.
- **TN:** Verdaderos Negativos.
- **FN:** Falsos Negativos.
- **FP:** Falsos Positivos.

4.2.10. Recall

Él recall o sensibilidad es una métrica que evalúa la tasa de verdaderos positivos calculando el número de predicciones positivas correctas dividido por el número total de positivos. A continuación, en la **Ecuación 5** se observa la fórmula matemática para el cálculo del recall, la cual fue obtenida de [40].

Ecuación 5. Fórmula de cálculo del recall o sensibilidad.

$$Recall = \frac{TP}{TP + FN}$$

Donde:

- **TP:** Verdaderos Positivos.
- **FN:** Falsos Negativos.

4.2.11. Metodología para Minería de Datos

4.2.11.1. CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM es una metodología adecuada para proyectos de Minería de Datos, es considerada como un proceso continuo [41], ya que se la debe seguir durante todo el desarrollo de la investigación, pudiendo retroceder a cualquier paso anterior si se considera pertinente. En la **Figura 4**, obtenida de [42] se muestran las 6 fases que forman esta metodología, las cuales guían sobre cómo proceder para completar con éxito el proceso de Minería de Datos.

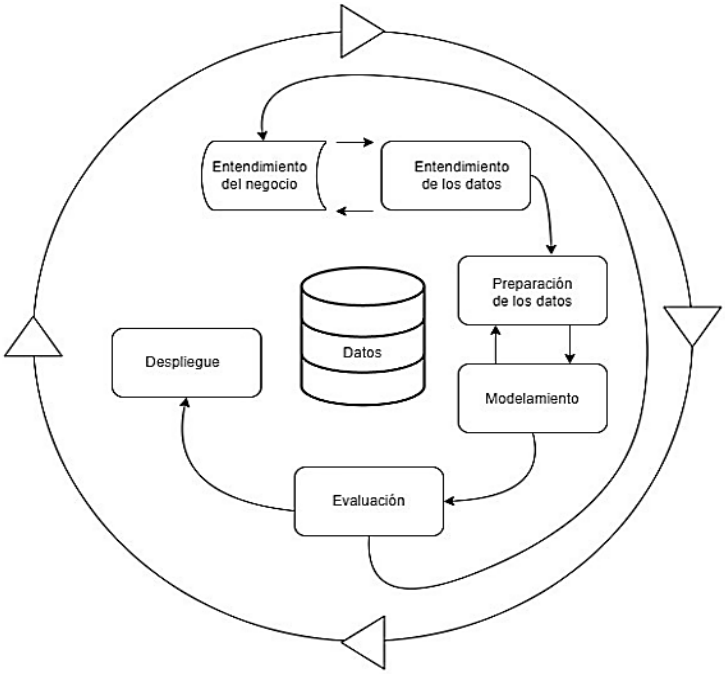


Figura 4. Fases de la Metodología CRISP-DM.

Las 6 fases se deben seguir durante todo el desarrollo de un proyecto de minería de datos, estas se describirán a continuación y fueron obtenidas de [42]:

- **Comprensión del negocio:** Determinación del problema inicial, los objetivos y requerimientos.
- **Comprensión de los datos:** Recolección de los datos, realización del Análisis Exploratorio de Datos (EDA).
- **Preparación de los datos:** Realización del filtrado, limpieza y estandarización de la base de datos.
- **Modelado:** Construir y entrenar los modelos de minería de datos.
- **Evaluación:** Evaluar con métricas de rendimiento los modelos de clasificación de minería de datos.
- **Despliegue:** Realización de informes para representar visualmente los conocimientos e información que fue obtenida mediante minería de datos, para ser usados para la toma de decisiones, planificaciones y modificaciones de una organización [43].

4.2.12. Herramientas de Desarrollo

4.2.12.1. Google Colab

Google Colab es una herramienta que se encuentra en la nube y se utiliza para programar en Python [44]. Esta herramienta sirve como entorno para realizar el proceso de Minería de Datos, facilita el procesamiento de una gran cantidad de datos sin necesidad de utilizar el almacenamiento del ordenador [45]. Google Colab permitió configurar el tipo de

entorno de la sesión ya que se tiene para elegir tanto GPU, CPU o TPU, siendo la primera de gran ayuda para poder realizar el procesamiento de una manera más rápida [46].

4.2.12.2. OpenRefine

OpenRefine es una herramienta que permite limpiar los datos de modo que estén libres de valores faltantes y nulos, ayudando a garantizar que los datos sean de calidad [47]. De la misma forma está herramienta es útil para realizar la correcta estandarización de la base de datos que se esté procesando [48]. OpenRefine brinda funcionalidades adicionales y mejoradas en comparación con otras herramientas que permiten de la misma manera realizar el gestionamiento de los datos [49].

4.2.12.3. Python

Python es un lenguaje de programación que nos brinda las herramientas para comenzar a codificar aplicaciones o modelos [50]. Es un lenguaje de programación fácil de entender y codificar [51], para poder llevar a cabo todas las fases de Minería de Datos que se realizarán en el presente trabajo. Python cuenta con librerías usadas para el aprendizaje supervisado, especialmente para tratar datos con algoritmos y obtener clasificaciones o predicciones mediante modelos implementados en pocas líneas de código [52].

4.2.12.4. Looker Studio

Looker Studio es una herramienta disponible en la nube desarrollada por Google, que permitió crear informes y representaciones gráficas mediante los datos que se le proporcionaron, permitió usar archivos provenientes de Microsoft Excel o archivos CSV para la creación de las visualizaciones [53].

4.3. Trabajos Relacionados

En la **Tabla 2** se describen cada uno de los trabajos relacionados que sirven para el desarrollo del presente Trabajo de Integración Curricular (TIC), las mismas que guían sobre la implementación de la técnica de Optimización Bayesiana aplicada en los modelos de minería de datos.

Tabla 2. Trabajos Relacionados.

Nro.	Título	Descripción
1	Optimización bayesiana de hiperparámetros [54].	En este trabajo se aplicó la técnica de optimización bayesiana al modelo CNN, logrando una precisión del 74%. Se puede apreciar un incremento en la precisión ya que anteriormente se tenía un 52% en esta métrica.
2	On Hyperparameter Optimization of Machine Learning Methods Using a Bayesian Optimization Algorithm to Predict Work Travel Mode Choice [30].	El estudio realizado en E.U para predicciones, utilizó modelos como K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Extreme Decision Tree (EDT) y Naive Bayes (NB), donde mediante el uso de la Optimización Bayesiana se logró una precisión para NB=97.38%, SVM=97.38% con la búsqueda de hiperparámetros de kernel: lineal, polinomial, sigmoid, rbf, C: rango de 1-100, multiclass technique: One-vs-One (ovo), One-vs-Rest ovr, KNN=97.23%, EDT=97.28%.

Nro.	Título	Descripción
3	Curvas de aprendizaje en la optimización bayesiana de hiperparámetros [55].	En este estudio realizado en España, se implementaron máquinas de vectores de soporte y redes neuronales. Antes el modelo SVM tenía una precisión inicial del 72%, luego llegó al 90%, y el modelo RN tenía una precisión inicial del 70%, luego de aplicar optimización bayesiana, la precisión aumentó al 85%, demostrando la efectividad de esta técnica.
4	Bayesian optimized hybrid kernel svm for rolling bearing fault diagnosis [56].	En este estudio realizado en China en la Universidad Agrícola de Nanjing, se aplicó la optimización bayesiana en el modelo SVM con un kernel híbrido para realizar clasificaciones y predicciones. El modelo inicialmente tenía una precisión de 88,67%, luego de aplicar la optimización bayesiana, la precisión alcanzó 92,50%. El F1-Score era previamente de 87% luego de aplicar esta técnica subió a 91%, la Sensibilidad era previamente de 87% luego aumentó a 90%.
5	Effective Coronary Artery Disease Prediction Using Bayesian Optimization Algorithm and Random Forest [57].	En este estudio se aplicó la optimización bayesiana para optimizar los hiperparámetros y mejorar los porcentajes de las métricas de desempeño del modelo Random Forest. En donde se usaron los siguientes rangos para los hiperparámetros que los configuran: n_estimators (10,100), max_depth (5,50), min_samples_split (2,11), min_samples_leaf (1,11), max_features (1,64), criterion (gini, entropy).
6	Efficient hyperparameter tuning for predicting student performance with Bayesian optimization [58].	En este estudio se aplicó la técnica de Optimización Bayesiana para mejorar la precisión de los modelos Random Forest y Decision Tree al momento de clasificar y predecir el rendimiento de los estudiantes. En donde se buscó los mejores hiperparámetros para el Árbol de Decisión, con los rangos siguientes para criterion: gini–entropy, max_depth:3-10, min_samples_split: 2-11, min_samples_leaf: 1-25, alcanzando una precisión del 89%.
7	Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification [59].	El estudio realizado en Taiwán para clasificar a personas que padecen Parkinson, con los modelos: SVM, RF, LR, NB, DT, en el cual antes de aplicar la técnica de optimización sus porcentajes en la métrica de precisión eran: SVM = 89.6%, RF = 87.2%, LR = 85.3%, DT = 85.7%, NB = 82.1%, RC = 80.9%. Después de aplicar la optimización Bayesiana los modelos tuvieron un incremento en su precisión como se evidencia a continuación: SVM = 92.3% encontrando los mejores valores para el kernel y el parámetro C, RF= 89.7%, LR = 87.2%, Decision Tree = 88.5%, NB=84.6%, RC = 83.3%.
8	Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier [60].	En este estudio, se utilizó la optimización bayesiana para ajustar los hiperparámetros del modelo de bosque aleatorio para realizar clasificaciones, se obtuvieron los siguientes porcentajes antes de aplicar la técnica de optimización: Precisión: 83.5%, Recall: 82.7%, F1-score: 83.1%, luego se aplicó la optimización bayesiana y se obtuvieron los siguientes porcentajes: Precisión: 87.3%, Recall: 91.8%, F1-score: 89.5%. Por lo tanto, la precisión aumentó en un 4%, el Recall (sensibilidad) en aproximadamente un 9% y el F1-score aumentó en un 6%, lo que demuestra que el modelo fue mejorado al aplicar la técnica de Optimización Bayesiana (OB), demostrado que es una técnica apta al momento de obtener los valores para los hiperparámetros.

Nro.	Título	Descripción
9	Enhanced machine learning tree classifiers for lithology identification using Bayesian optimization [61].	El estudio se realizó en China, para mejorar la identificación y clasificación de litologías en la exploración de petróleo y gas usando una técnica de optimización en este caso la Bayesiana. Se utilizó esta técnica en el modelo de Árbol de Decisión (DT), que tenía una precisión del 86,9% antes de la técnica de optimización, luego se aplicó la optimización bayesiana en los hiperparámetros min_samples_split y criterion, al encontrar los mejores valores aumentó la precisión al 89,8%.
10	Análisis de Correspondencias Múltiples para el Estudio de los Homicidios Intencionales en el Ecuador [12].	En este estudio se implementó el Análisis de Correspondencia Múltiple para determinar las variables más importantes y que más explican el comportamiento de los homicidios intencionales, al principio se tuvo un total de 13 variables, luego de usar esta técnica las variables más destacadas son: el arma, tipo de delito, edad, sexo de la víctima, provincia, hora, día, lugar, área, por ser irrelevantes se excluyeron las siguientes variables: estado civil, nacionalidad, etnia, año.
11	Minería de datos para determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en Ecuador en el año 2020 [62].	En este estudio se aplicó los modelos CHAID, CHAID Exhaustivo, CRT, Perceptrón Multicapa, Función de Base Radial, Naive Bayes y BayesNet, para realizar la minería de datos y poder determinar los factores que más influyen en la ocurrencia de accidentes de tránsito también se evaluó cada clasificación y predicción con las métricas de precisión, se concluyó que el algoritmo CHAID Exhaustivo obtuvo un mayor porcentaje en sus métricas teniendo 58,38% y 44,60% de precisión, superando a los otros algoritmos.
12	Minería de datos en la accidentabilidad vehicular en la Zona Urbana del Cantón Loja [63].	En este estudio se aplicó la minería de datos predecir la probabilidad de un accidente de tránsito en la zona urbana del cantón Loja, mediante los algoritmos J48 y CART, a través de las herramientas WEKA y Python, los hiperparámetros fueron: criterion y max_depth para la configuración de los Árboles de Decisión.

5. Metodología

En el presente Trabajo de Integración Curricular (TIC) se usó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), ya que cuenta con 6 fases que guiaron durante todo el desarrollo de la investigación y sirvió para obtener los resultados deseados al momento de realizar análisis de datos y descubrir patrones de comportamiento mediante modelos de minería de datos.

5.1. Área de Estudio

El presente Trabajo de Integración Curricular (TIC) está centrado en la Zona 8 del Ecuador que corresponde a los cantones de Guayaquil, Durán y Samborondón, ubicado en las coordenadas: latitud: 2.1817° S, longitud: 79.8975° W, siendo donde más números de crímenes se han registrado, se desarrolló con la base de datos de asesinatos obtenida de la Policía Nacional del Ecuador.

En la **Figura 5** obtenida de [64]**,** se visualiza el mapa que representa el lugar sobre el cuál se hizo el análisis de datos para determinar patrones en los asesinatos.

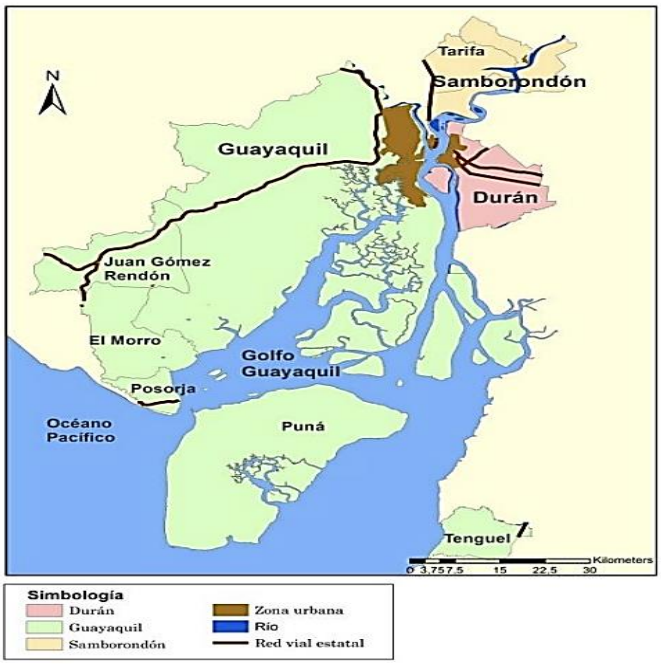


Figura 5. Zona 8 del Ecuador, cantones de Guayaquil, Durán y Samborondón.

5.2. Procedimiento

En este apartado se presentan las fases de la metodología para minería de datos que se siguió para poder cumplir los objetivos planteados para este Trabajo de Integración Curricular (TIC), además se evidencian los recursos científicos, técnicos de hardware y software que se usaron y se menciona los participantes que estuvieron involucrados en el actual trabajo.

5.2.1. Objetivo 1: Mejorar la calidad del dataset de asesinatos del Ecuador mediante técnicas de preparación de datos, realizar el análisis exploratorio de datos EDA e implementación de los clasificadores optimizados Árbol de Decisión y SVM.

El primer objetivo del presente TIC se cumplió siguiendo las 4 fases iniciales de la metodología CRISP-DM, las cuales sirvieron para desarrollar el trabajo de minería de datos. A continuación, se describen las tareas que se ejecutaron para realizar la investigación:

- **Fase 1: Comprensión del negocio**

Se realizó un oficio dirigido al Subteniente Sebastian Encalada quién pertenece a la Policía Nacional del Ecuador para solicitar la apertura de contestación de una encuesta de preguntas abiertas, esto se puede visualizar en **Anexo I**, con el fin de tener una comprensión de la problemática social, esta encuesta se llevó a cabo el día 17 de octubre del 2024, la misma que se encuentra formalizada en el **Anexo II**, con la finalidad de obtener información sobre las particularidades relacionadas a los asesinatos de la Zona 8 del Ecuador.

Para comprender la problemática de la parte técnica para el actual proyecto, se realizó una entrevista a la Ingeniera Genoveva Suing (ver en **Anexo III)**, la misma que sirvió para conocer sobre la Optimización Bayesiana (OB) en los modelos de minería de datos: Árbol de Decisión y Support Vector Machine (SVM), y como esta técnica de OB puede buscar los valores adecuados para los hiperparámetros que los configuran y así alcanzar un mayor porcentaje de precisión al momento de realizar tareas de clasificación y descubrir patrones de en los asesinatos de la Zona 8 del Ecuador.

- **Fase 2: Comprensión de los datos**

Se realizó un oficio el cual fue enviado mediante correo electrónico a: gestion.documental@ministeriodelinterior.gob.ec, del Ministerio del Interior para obtener la base de datos de homicidios intencionales periodo enero 2015 hasta febrero de 2024 (ver en **Anexo IV),** como contestación se recibió por correo en formato .xlsx[1] el dataset solicitado anteriormente, el mismo que fue otorgado por la Policía Nacional del Ecuador en conjunto con el Ministerio del Interior, teniendo la autorización para hacer uso de los datos en esta investigación (ver en **Anexo V)**. La base de datos contenía 34 variables con 22849 registros de los homicidios intencionales ocurridos en todo el Ecuador, luego la BD se la exportó al formato CSV[2], el contenido y descripción se puede visualizar en la **Tabla 3**.

Se realizó un análisis exploratorio de los datos (EDA) con la herramienta OpenRefine en el lenguaje GREL, mediante el cual se pudo analizar la dimensión del dataset, su número de variables y cantidad de registros, esto se puede visualizar con mayor detalle en la sección de **Resultados - Análisis Exploratorio de Datos**.

[1] https://bit.ly/4gKgdVc
[2] https://bit.ly/4gR8Qv8

- **Fase 3: Preparación de los datos**

Se filtró la base de datos de homicidios intencionales del Ecuador por que existían registros de homicidios, femicidios, sicariatos y asesinatos, este proceso se hizo para que aparezcan únicamente los registros de los asesinatos ocurridos en la Zona 8 del Ecuador ya que esta investigación está enfocada en este tipo de muerte (ver en **Resultados - Filtrado de Base de Datos**).

Se seleccionó las variables más relevantes, esto se pudo realizar mediante el EDA y los resultados de trabajos relacionados, en los que indicaron qué variables son las más adecuadas para explicar estos crímenes (ver en **Resultados - Selección de variables**).

Se estandarizó cada una de las variables a mayúsculas y se eliminaron caracteres especiales en los registros. Se procedió a limpiar los valores nulos que se encontraban en la base de datos, usando la herramienta OpenRefine (ver en **Resultados - Estandarización y limpieza)**.

- **Fase 4: Modelado**

Se implementó los modelos Árbol de Decisión y Support Vector Machine, usando el lenguaje de programación Python en la herramienta Google Colab, se llevó a cabo el entrenamiento de los mismos sin la técnica de Optimización Bayesiana (OB), luego fueron configurados con la técnica de OB para encontrar los mejores valores para los hiperparámetros que lograron alcanzar un mayor porcentaje de precisión y se determinó los patrones de comportamiento sobre la base de datos de asesinatos de la Zona 8 del Ecuador (Guayaquil, Durán y Samborondón), periodo enero 2015 - febrero 2024 (ver en **Resultados - Fase 4**)**.**

5.2.2. Objetivo 2: Evaluar los clasificadores Árbol de Decisión y SVM con la métrica de precisión.

El segundo objetivo del presente TIC se cumplió siguiendo las 2 últimas fases de la metodología CRISP-DM las cuales son la evaluación y el despliegue, estas sirvieron para desarrollar el trabajo de minería de datos. A continuación, se describen cada una de ellas y las tareas que se ejecutaron para realizar la investigación:

- **Fase 5: Evaluación**

Se evaluaron los porcentajes de precisión de cada uno de los modelos mediante la técnica de validación cruzada y a través de sklearn.metrics que es una librería de Python para calcular la precisión de los clasificadores de minería de datos, además se analizó los resultados y predicciones obtenidas de cada una de las variables en donde se realizó el proceso de minería de datos (ver en **Resultados - Fase 5**).

- **Fase 6: Despliegue**

Se realizó un informe exportado en formato PDF con los resultados de los patrones de los asesinatos de la zona 8 del Ecuador que fueron encontrados mediante minería de datos,

así también se realizó un dashboard en la herramienta Looker Studio en el cual se representó visualmente cada uno de los patrones encontrados, ambos fueron entregados a la Policía Nacional del Ecuador, (ver en **Resultados - Fase 6**).

5.3. Recursos

Para el desarrollo del Trabajo de Integración Curricular (TIC), con el fin de realizar el correcto cumplimiento de los objetivos se utilizó los siguientes recursos:

5.3.1. Recursos Científicos

- **Investigación de Bibliografía**

Se utilizó referencias y fuentes bibliográficas para validar la información descrita en el presente Trabajo de Integración Curricular, ya que esto permitió garantizar la ética de la investigación.

- **Método Experimental**

Se usó el método experimental al aplicar la técnica de optimización Bayesiana ya que se realizaron pruebas para configurar los modelos con la búsqueda de los mejores hiperparámetros, además este método ayudó en todas las fases que comprenden la metodología CRISP-DM, desde el ajuste hasta la evaluación de la precisión.

5.3.2. Recursos de Hardware

- **Laptop Asus ZenBook**

Posee un procesador i5-1135G7, con 8,00 GB de memoria, esta máquina de gama media se usó durante todo el desarrollo del Trabajo de Integración Curricular (TIC).

5.3.3. Recursos de Software

- **Microsoft Word**

Este software se utilizó para procesar textos ya que permitió la creación, edición, visualización de documentos para el desarrollo y culminación de la memoria del Trabajo de Integración Curricular.

- **OpenRefine**

Este software sirvió para realizar el procesamiento de datos ya que facilitó la limpieza, estandarización y filtrado de la base de datos de asesinatos correspondientes a la Zona 8 del Ecuador.

- **Mendeley Reference**

Se utilizó para gestionar artículos y poder obtener las referencias o bibliografías de las investigaciones relacionadas, mostrándose en formato IEEE que es la norma establecida para presentar el Trabajo de Integración Curricular.

- **Google Colab**

Sirvió para poder realizar el proceso de minería de datos, especialmente el entrenamiento, aplicación de Optimización Bayesian, modelado y evaluación de los clasificadores, empleando para esto el lenguaje de programación Python.

- **Looker Studio**

Se utilizó esta herramienta para la realización del dashboard que contiene gráficos y representaciones visuales interactivas, permitiendo subir archivos como Excel o CSV.

- **GitHub**

Se utilizó esta herramienta para guardar los códigos fuente y archivos como bases de datos, permitiendo realizar cambios y modificaciones realizadas durante todo el TIC.

5.3.4. Recursos Técnicos

- **Encuesta**

Se realizó una encuesta de preguntas abiertas al Subteniente de la Policía Nacional del Ecuador Sebastián Encalada, para conocer sobre la problemática de asesinatos en la Zona 8 del Ecuador, (ver en **Anexo II**).

- **Entrevista**

Se realizó la entrevista a la experta Ingeniera. Genoveva Suing, que posee conocimientos en las áreas de minería de datos y brindó sus conocimientos para poder desarrollar el actual TIC, (ver en **Anexo III**).

- **Metodología CRISP-DM**

Se realizó el presente Trabajo de Integración Curricular (TIC), siguiendo la metodología CRISP-DM, ya que cuenta con 6 fases que guiaron para el desarrollo del actual proyecto de minería de datos.

5.4. Participantes

En este apartado se describen el nombre de cada uno de los participantes que estuvieron involucrados en el desarrollo del presente trabajo, así como las tareas que desenvolvieron y que fueron de gran ayuda para realizar esta investigación:

- La estudiante Cecilia Fernanda Trueba Reyes como autora del Trabajo de Integración Curricular (TIC).
- Ingeniera Genoveva Jackelinne Suing Albito, Mg. Sc., como directora del Trabajo de Integración Curricular (TIC).
- El Subteniente Sebastián Encalada, perteneciente a la Policía Nacional del Ecuador, fue encuestado para conocer la problemática de asesinatos de la Zona 8 del Ecuador.
- Capt. David Estuardo Anrango Narváez, Subsecretario de Estudios y Estadísticas de la Seguridad, quién otorgó la base de datos de homicidios intencionales, periodo enero 2015 – febrero 2024.

6. Resultados

A continuación, en esta sección se indica detalladamente los resultados obtenidos, se siguieron cada una de las fases que forman parte de la Metodología CRISP-DM para completar los dos objetivos específicos que fueron establecidos para desarrollar el presente Trabajo de Integración Curricular (TIC):

6.1. Objetivo 1: Mejorar la calidad del dataset de asesinatos del Ecuador mediante técnicas de preparación de datos, realizar el análisis exploratorio de datos EDA e implementación de los clasificadores optimizados Árbol de Decisión y SVM.

Para cumplir con el primer objetivo específico del presente Trabajo de Integración Curricular (TIC), se ejecutaron las cuatro primeras fases que componen la metodología CRISP-DM, las cuales son: comprensión del negocio, comprensión de los datos, preparación de los datos y modelado, las mismas que se muestran y detallan a continuación:

6.1.1. Fase 1: Comprensión del negocio.

Mediante el desarrollo de una encuesta de preguntas abiertas realizada a Sebastian Encalada Espinosa Subteniente de la Policía Nacional del Ecuador (ver en **Anexo II**), se estableció como problemática principal los asesinatos ocurridos en la Zona 8 (Guayaquil, Durán y Samborondón) del Ecuador, ya que en los últimos años se ha visto incrementado el número de asesinatos y violencia en el país, cambiando el estilo de vida de las personas debido a que se ha aumentado el miedo en los ciudadanos, produciendo el temor a salir de casa, el cierre de negocios y locales, cambios de vivienda porque las mafias se han apoderado de las mismas, todo esto en los lugares más afectados, que es donde estos grupos delictivos se han apoderado de estos sitios siendo donde mayor número de asesinatos existe. Los beneficiados con este estudio son la Policía Nacional del Ecuador específicamente el Departamento de Análisis de Información del Delito (DAID), ya que el Subteniente expresó: “Un informe con estadísticas e información relevante de los asesinatos ayuda a tomar decisiones y actuar de forma preventiva y disuasiva ya que se pueden conocer los puntos críticos, ubicaciones y distritos donde existe mayor número de asesinatos”. Por lo tanto, esta investigación y análisis estadístico con modelos de minería de datos contribuyó a comprender cómo operan los grupos delincuenciales al efectuar crímenes y poder realizar predicciones de comportamientos a futuro relacionados a los asesinatos.

De la misma manera para conocer el problema técnico de esta investigación se realizó una entrevista a la Ingeniera Genoveva Suing (ver en **Anexo III**), quien señaló que las técnicas de optimización pueden mejorar la precisión de los modelos de minería de datos al momento de evaluarlos.

6.1.2. Fase 2: Comprensión de los datos.

En esta fase se realizó la obtención de la base de datos .xlsx de Homicidios Intencionales del Ecuador periodo enero 2015 hasta febrero 2024, que luego se la exportó en formato CSV[3] para su respectivo análisis y procesamiento en la herramienta Google Colab en el lenguaje de programación Python.

En la **Tabla 3**, se visualiza el listado de variables o atributos con su descripción correspondiente, las mismas que pertenecen a la base de datos de Homicidios Intencionales de Ecuador, periodo enero 2015 - febrero 2024.

Tabla 3. Listado de variables de la Base de Datos.

Nro.	Variable	Tipo de variable	Descripción
1	Tipo Muert.	Categórica	Indica el tipo de muerte registrada en la base de datos, que incluye homicidios, femicidios, sicariatos y asesinatos.
2	Zona	Categórica	División del Ecuador en 9 zonas de planificación, que están formadas por varias provincias.
3	Subzona	Categórica	Subdivisiones dentro de cada zona de planificación.
4	Distrito	Categórica	Territorio de una provincia o ciudad que incluye varios circuitos.
5	Circuito	Categórica	Subdivisión dentro del distrito que incluye espacios reducidos.
6	Cod. Subcircu	Categórica	Combinación de números y letras que representan los subcircuitos.
7	Subcircuito	Categórica	Es la subdivisión del circuito, que son las áreas mucho más pequeñas que el propio circuito.
8	provincia	Categórica	División de las 24 provincias que conforman el Ecuador.
9	código de provincia	Numérica	Número que representa a cada provincia.
10	cantón	Categórica	Subdivisión de las provincias del Ecuador.
11	código de cantón	Numérica	Número que representa a cada cantón del país ecuatoriano.
12	Coord. Y	Categórica	Corresponde a la dimensión vertical (norte, sur) de la ubicación donde se realizó el crimen, ya sea homicidio, femicidio, asesinato o sicariato.
13	Coord. X	Categórica	Corresponde a la dimensión horizontal (este, oeste) que índica la ubicación donde se realizó el crimen, ya sea homicidio, femicidio, asesinato o sicariato.

[3] https://bit.ly/4gR8Qv8

Nro.	Variable	Tipo de variable	Descripción
14	Área del Hecho	Categórica	Índica como es la división territorial puede ser urbano o rural.
15	Lugar	Categórica	Sitio dónde se realizó el crimen, vía pública o en un lugar privado.
16	Tipo Lugar	Categórica	Espacio definido mediante coordenadas, puede ser público o privado.
17	Fecha Infracción	Categórica	Fecha (mes, día, año) en la que se produjo el crimen.
18	Hora Infracción	Categórica	Instante en el que se presume que se realizó el crimen.
19	Arma	Categórica	Medio usado para atacar a otras personas.
20	Tipo Arma	Categórica	Sirve para distinguir entre los diferentes tipos de armas que existen.
21	Presun. Motiva.	Categórica	Motivaciones generales que llevaron al victimario a realizar el crimen.
22	Presun. Motiva. Obser	Categórica	Motivaciones específicas que llevaron al victimario a realizar el crimen.
23	Probable Causa M.	Categórica	Causa que se sospecha que ha generado la muerte a una persona.
24	Edad	Numérica	Tiempo en el que un individuo está vivo contando desde que ha nacido.
25	Med. Edad	Categórica	Medida para describir la edad de la víctima desde su nacimiento hasta su muerte, puede ser (semanas, meses, días, años).
26	Sexo	Categórica	Características que definen a un individuo como hombre o mujer.
27	Género	Categórica	Identidad de una persona, es decir cómo se define.
28	Etnia	Categórica	Grupo con características culturales similares.
29	Estado Civil	Categórica	Entorno íntimo de la víctima.
30	Nacionalidad	Categórica	Vínculo de una persona con un país.
31	Discapacidad	Categórica	Limitaciones que tiene una persona.
32	Prof Reg Civ	Categórica	Profesión de la víctima que está registrada en el registro civil.
33	Instrucción	Categórica	Nivel de educación alcanzado por la víctima.
34	Antecedentes	Categórica	Historial de crímenes de un individuo.

6.1.2.1. Análisis Exploratorio de Datos

Se realizó un Análisis Exploratorio de Datos (EDA), a continuación, se describe cada parte del código para la realización de esta parte importante que sirvió para conocer información sobre los registros y atributos de la base de datos.

En la **Figura 6,** se visualiza como en Google Colab Notebook se realizó la carga del archivo CSV de homicidios intencionales en un DataFrame que es una estructura en la cual se guardan los datos, esto se logró accediendo al archivo que se guardó en Google Drive, ya que esta opción está disponible en el entorno de Colab.

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

```
Mounted at /content/drive
```

```
import pandas as pd
# Especifica la ruta al archivo en Google Drive
file_path = '/content/drive/My Drive/HOMICIDIOS_INTENCIONALES.csv'
# Cargar el archivo CSV en un DataFrame
df = pd.read_csv(file_path, delimiter=';', encoding='latin1', on_bad_lines='skip')
```

Figura 6. Carga de archivo CSV.

En la **Figura 7**, se procede a imprimir todo el dataset de homicidios intencionales desde enero 2015 hasta febrero 2024, con la finalidad de tener conocimiento sobre las variables y registros que posee este dataset, el mismo que ha sido cargado con anterioridad.

	Tipo Muert.	Zona	Subzona	Distrito	Circuito	Cod. Subcircu	Subcircuito	provincia	código de provincia
0	ASESINATO	ZONA 4	SANTO DOMINGO DE LOS TSACHILAS	SANTO DOMINGO OESTE	PLAN DE VIVIENDA	23D02C06S01	PLAN DE VIVIENDA 1	SANTO DOMINGO DE LOS TSACHILAS	23
1	ASESINATO	ZONA 8	D.M. GUAYAQUIL	PORTETE	SUBURBIO	09D04C06S03	SUBURBIO 3	GUAYAS	9
2	ASESINATO	ZONA 1	ESMERALDAS	ESMERALDAS	VALLE HERMOSO	08D01C08S01	VALLE HERMOSO 1	ESMERALDAS	8
3	ASESINATO	ZONA 9	D.M. QUITO	LA DELICIA	LA ROLDOS	17D03C05S04	LA ROLDOS 4	PICHINCHA	17
4	ASESINATO	ZONA 5	LOS RIOS	BABAHOYO	PIMOCHA	12D01C11S01	PIMOCHA 1	LOS RIOS	12
...	...	...	...	...	...	...	...	...	...
22844	ASESINATO	ZONA 7	EL ORO	MACHALA	LA FERROVIARIA	07D02C15S01	LA FERROVIARIA 1	EL ORO	7
22845	ASESINATO	ZONA 7	EL ORO	MACHALA	LA FERROVIARIA	07D02C15S01	LA FERROVIARIA 1	EL ORO	7
22846	ASESINATO	ZONA 5	GUAYAS	PLAYAS	PLAYAS	09D22C01S01	PLAYAS 1	GUAYAS	9
22847	ASESINATO	ZONA 5	LOS RIOS	BUENA FE	BUENA FE OESTE	12D06C03S02	BUENA FE OESTE 2	LOS RIOS	12
22848	ASESINATO	ZONA 8	D.M. GUAYAQUIL	PROGRESO	PROGRESO	09D10C03S01	PROGRESO 1	GUAYAS	9

22849 rows × 34 columns

Figura 7. Impresión del DataFrame cargado.

En la **Figura 8**, se muestra el código para ver las dimensiones del DataFrame, que tiene un total de 22,849 registros y 34 columnas.

```
print("Dimensiones del DataFrame (filas, columnas):")
print(df.shape)
```

```
Dimensiones del DataFrame (filas, columnas):
(22849, 34)
```

Figura 8. Dimensiones del DataFrame.

En la **Figura 9,** se indica el nombre de las variables o columnas del dataframe, estas son del tipo object es decir que son cadenas de caracteres y el tipo de dato int64 significa que son valores enteros, existiendo 3 variables con registros enteros y 31 variables con sus registros que son cadenas o textos.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22849 entries, 0 to 22848
Data columns (total 34 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   Tipo Muert.            22849 non-null  object
 1   Zona                   22849 non-null  object
 2   Subzona                22849 non-null  object
 3   Distrito               22849 non-null  object
 4   Circuito               22849 non-null  object
 5   Cod. Subcircu          22849 non-null  object
 6   Subcircuito            22849 non-null  object
 7   provincia              22849 non-null  object
 8   código de provincia    22849 non-null  int64 
 9   cantón                 22849 non-null  object
 10  código de cantón       22849 non-null  int64 
 11  Coord. Y               22849 non-null  object
 12  Coord. X               22849 non-null  object
 13  Area del Hecho         22849 non-null  object
 14  Lugar                  22849 non-null  object
 15  Tipo Lugar             22849 non-null  object
 16  Fecha Infracción       22849 non-null  object
 17  Hora Infracción        22849 non-null  object
 18  Arma                   22849 non-null  object
 19  Tipo Arma              22849 non-null  object
 20  Presun. Motiva.        22849 non-null  object
 21  Presun. Motiva. Obser  22849 non-null  object
 22  Probable Causa M.      22849 non-null  object
 23  Edad                   22849 non-null  int64 
 24  Med. Edad              22849 non-null  object
 25  Sexo                   22849 non-null  object
 26  Género                 22849 non-null  object
 27  Etnia                  22849 non-null  object
 28  Estado Civil           22849 non-null  object
 29  Nacionalidad           22849 non-null  object
 30  Discapacidad           22849 non-null  object
 31  Prof Reg Civ           22849 non-null  object
 32  Instrucción            22849 non-null  object
 33  Antecedentes           22849 non-null  object
dtypes: int64(3), object(31)
memory usage: 5.9+ MB
```

Figura 9. Información del Dataframe.

En la **Figura 10**, se presenta la descripción de las estadísticas de las variables numéricas, en este caso son “código de provincia”, “código de cantón” y “Edad”. En las cuales se imprimen los valores de: “count”, que indica el número total de registros para cada variable, “mean”, que indica la media aritmética correspondiente a la suma de todos los valores dividido

para la cantidad de valores de cada uno de los registros de las variables, “std” que es la media aritmética de cada registro de las variables, “min” es el valor mínimo de los registros de cada columna, “max” es el valor máximo de los registros de cada columna, “25%” es el cuartil primero que indica que el veinticinco por ciento de los datos es menor al valor mostrado para cada una de las variables, el “50%” es el cuartil segundo que indica que el cincuenta por ciento de los datos es menor al valor mostrado para cada una de las variables y el “75%” cuartil indica que el setenta y cinco por ciento de los datos es menor al valor mostrado para cada una de las variables.

```
print("\nEstadísticas descriptivas:")
display(df.describe())
```

Estadísticas descriptivas:

	código de provincia	código de cantón	Edad
count	22849.000000	22849.000000	22849.000000
mean	11.628343	1166.739201	32.600727
std	8.408528	840.564047	13.482937
min	1.000000	101.000000	0.000000
25%	9.000000	901.000000	24.000000
50%	9.000000	906.000000	30.000000
75%	13.000000	1303.000000	39.000000
max	90.000000	9004.000000	102.000000

Figura 10. Descripción del Dataframe.

En la **Figura 11**, se encuentra el código que permite obtener datos estadísticos de cada variable que se escoja y que se quiera obtener información, en este caso fue el “Tipo Muert,”, en donde el “ASESINATO” tiene un total de 19778 registros, siendo el tipo de muerte más frecuente y por lo cual se seleccionó este tipo de crimen para esta investigación ya que es el más reiterado, en comparación con el “HOMICIDIO”, “FEMICIDIO” y “SICARIATO” que cuentan con 2192, 713 y 166 registros respectivamente.

```
CountCategory = df['Tipo Muert.'].value_counts()
print(CountCategory)
```

```
Tipo Muert.
ASESINATO    19778
HOMICIDIO     2192
FEMICIDIO      713
SICARIATO      166
Name: count, dtype: int64
```

Figura 11. Datos estadísticos de cada categoría de la variable “Tipo Muert.”.

En la **Figura 12**, se encuentra el código que permite obtener datos estadísticos de cada categoría de la variable escogida, en este caso es la “Zona”, en donde la ZONA 8 es la que posee mayor número de registros, evidenciando que es donde más se producen los homicidios intencionales en el Ecuador.

```
CountCategory = df['Zona'].value_counts()
print(CountCategory)
```

```
Zona
ZONA 8               7183
ZONA 5               5316
ZONA 4               2997
ZONA 1               2328
ZONA 7               1633
ZONA 9               1388
ZONA 6                752
ZONA 3                645
ZONA 2                424
ZONA NO DELIMITADA    183
Name: count, dtype: int64
```

Figura 12. Datos estadísticos de cada categoría de la variable “Zona”.

En la **Figura 13**, se observa el histograma de la variable “Tipo Muert.” de los registros de todo el Ecuador, esta vez el código genera gráficos estadísticos de la variable que se desee visualizar. En la figura se observa que el asesinato es el que tiene mayor cantidad de registros, seguido del homicidio, femicidio y por último el sicariato.

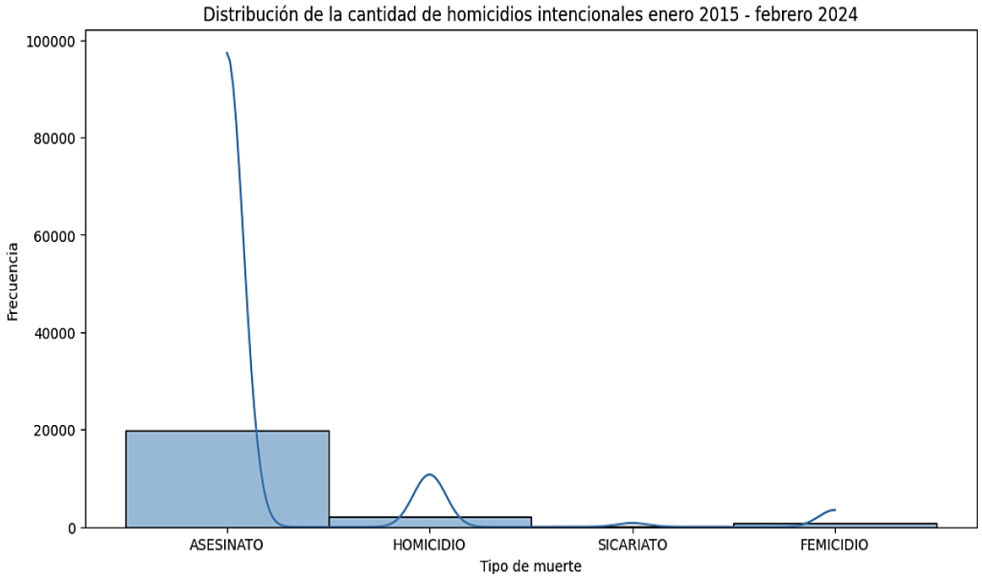


Figura 13. Histograma de variable categórica “Tipo Muert.”.

En la **Figura 14**, se observa el histograma de la variable “Zona” que corresponde a los registros de todo el Ecuador, está vez el código genera gráficos estadísticos de las variables que se quiera visualizar. Se puede ver que la Zona 8 es la que tiene mayor cantidad de homicidios intencionales, seguida por la Zona 5, Zona 4, Zona 1, Zona 7, Zona 9, Zona 6, Zona 3 y por último la Zona 2.

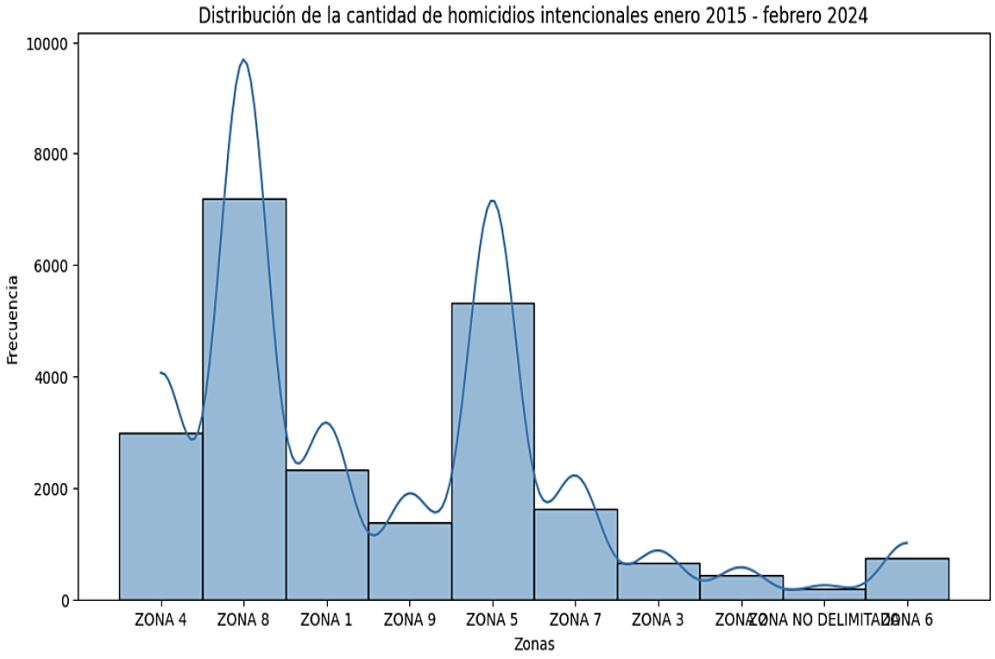


Figura 14. Histograma de variable categórica “Zona”.

En la **Figura 15**, se observa el histograma de las variables numéricas “código de provincia”, “código de cantón” y “Edad” que contiene los registros de todo el Ecuador, se puede ver que el “código de provincia” más usual es el que va en el rango de 0 a 10, el “código de cantón” más frecuente va de 0 a 1000, teniendo más de 10000 registros en cada uno, y la “Edad” más habitual va en el rango de 20 a 30 con más de 8000 registros.

```
df.hist()
```

```
array([[<Axes: title={'center': 'código de provincia'}>,
        <Axes: title={'center': 'código de cantón'}>],
       [<Axes: title={'center': 'Edad'}>, <Axes: >]], dtype=object)
```

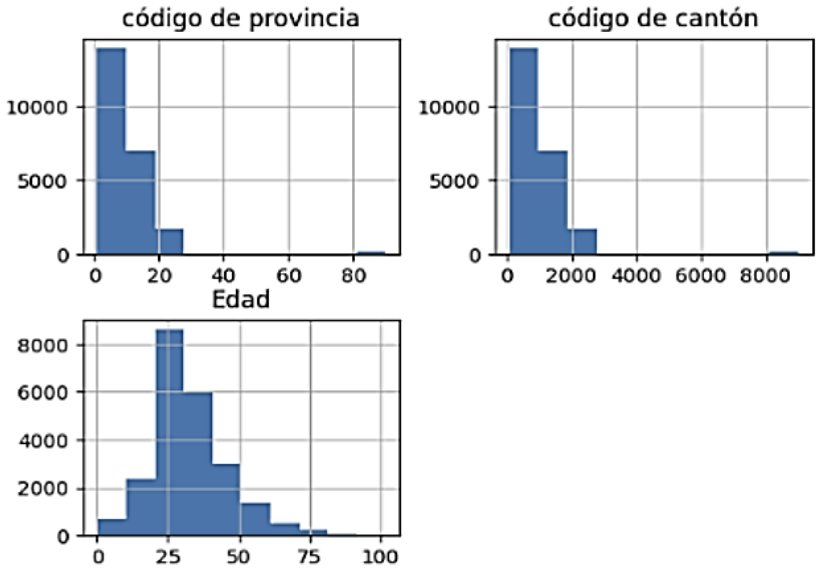


Figura 15. Histograma de variables numéricas.

6.1.3. Fase 3: Preparación de los datos

Para completar la fase 3 de la metodología CRISP-DM dedicada a la preparación de los datos, se completaron varios procesos como el filtrado de base de datos, la selección de variables relevantes para el estudio, la limpieza de valores faltantes y el balanceo de las clases desequilibradas de cada una de las variables, a continuación, se describen estos procesos.

6.1.3.1. Filtrado de Base de Datos

La base de datos de homicidios intencionales del Ecuador enero 2015 – febrero 2014, obtenida de la Policía Nacional del Ecuador contiene los 4 tipos de muertes: sicariato, homicidio, femicidio y asesinato que forman parte de la clasificación de los homicidios intencionales. El presente estudio está enfocado en comprender los patrones de los asesinatos de la Zona 8 del Ecuador, por lo que se realizó un filtrado de datos, para que solo se muestren los registros correspondientes que fueron delimitados para este estudio.

En la **Figura 16**, se visualiza como se creó un nuevo proyecto y se ingresó la base de datos correspondiente en la herramienta OpenRefine.

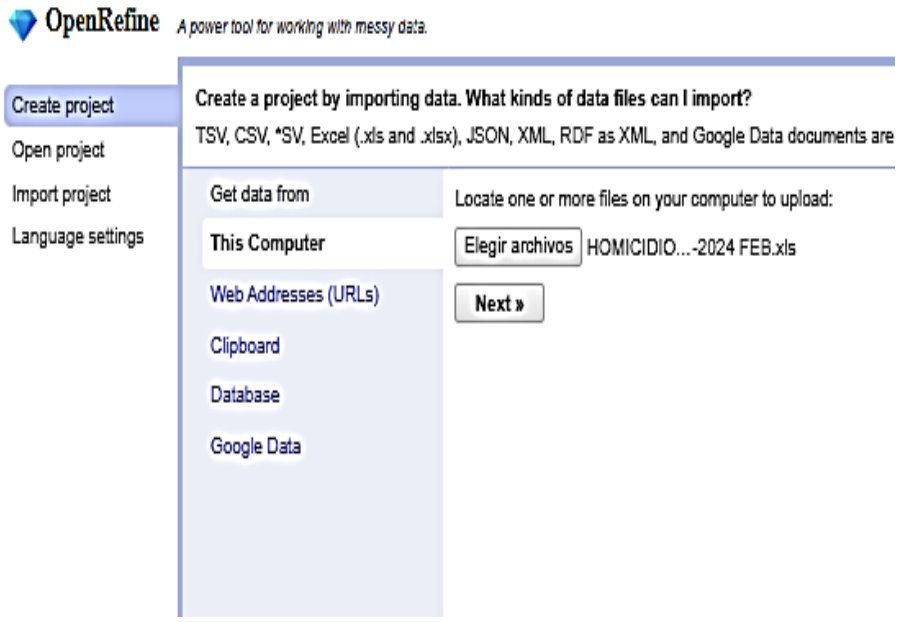


Figura 16. Creación de un nuevo proyecto en OpenRefine.

En la **Figura 17,** se visualiza como en la herramienta OpenRefine, que la base de datos que se ingresó posee un total de 22849 registros de los homicidios intencionales ocurridos en todo el Ecuador, por tal motivo se seleccionó la opción de filtrado de texto, esto se realizó en las variables “Tipo Muert.” y “Zona”, para que al momento que se realice el filtrado solamente se encuentra la Zona 8 y el asesinato como tipo de muerte, ya que existen otras categorías que no están comprendidas en el alcance de este estudio de minería de datos.

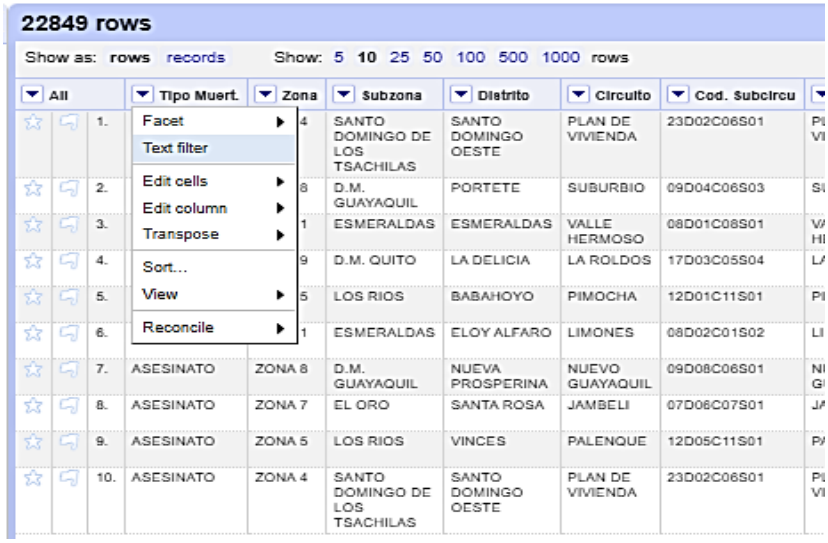


Figura 17. Filtrado de base de datos.

Para este caso fueron filtradas las variables “Tipo Muert.” y “Zona” como se muestra en la **Figura 18**, ya que para este trabajo se obtuvieron patrones en los asesinatos de la Zona 8 del Ecuador, en los cuales luego del filtrado se obtuvieron un total de 6920 registros de asesinatos con 34 variables iniciales correspondientes al periodo enero 2015 – febrero 2024.

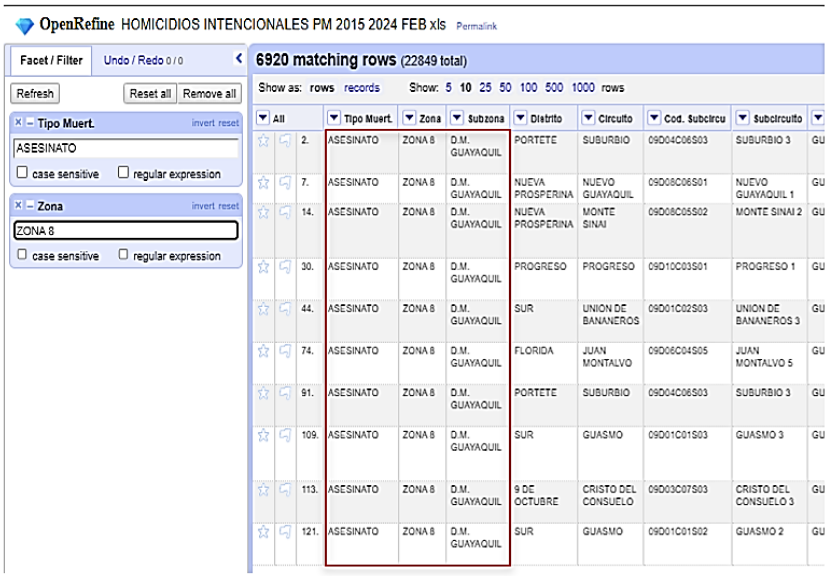


Figura 18. Base de datos filtrados por Zona 8 y asesinatos.

6.1.3.2. Selección de variables para el estudio

Mediante el análisis exploratorio de los datos se pudo evidenciar que existen variables redundantes como se observa en la **Figura 18,** presentada anteriormente, estos atributos son: “Tipo Muert.” es redundante, ya que contiene en todos sus registros los asesinatos, por lo que se la procedió a excluir. Se eliminó la variable “Zona” debido a que para todos los registros es la Zona 8, así mismo se evidenció que la variable “Subzona” es innecesaria porque contiene solo el D.M. Guayaquil en sus registros por lo que se procedió a eliminarla.

Se evidenció que la variable “provincia” en todos sus registros tiene la provincia del Guayas por lo que no es necesario analizarla y se procedió a eliminar. Lo mismo sucedió con la variable “codigo de provincia” en la cual poseía en todos sus registros el número 9 por lo que es necesario eliminarla por ser repetitiva. De la misma manera se descartó la variable “cantón” porque se consideró que es redundante para el estudio.

En la **Figura 19**, se realizó en Python el cálculo para la correlación de variables y se evidencio que la variable “Lugar” y “Tipo Lugar” presentan un nivel muy alto con respecto a su relación, por lo que se eliminó “Tipo Lugar”.

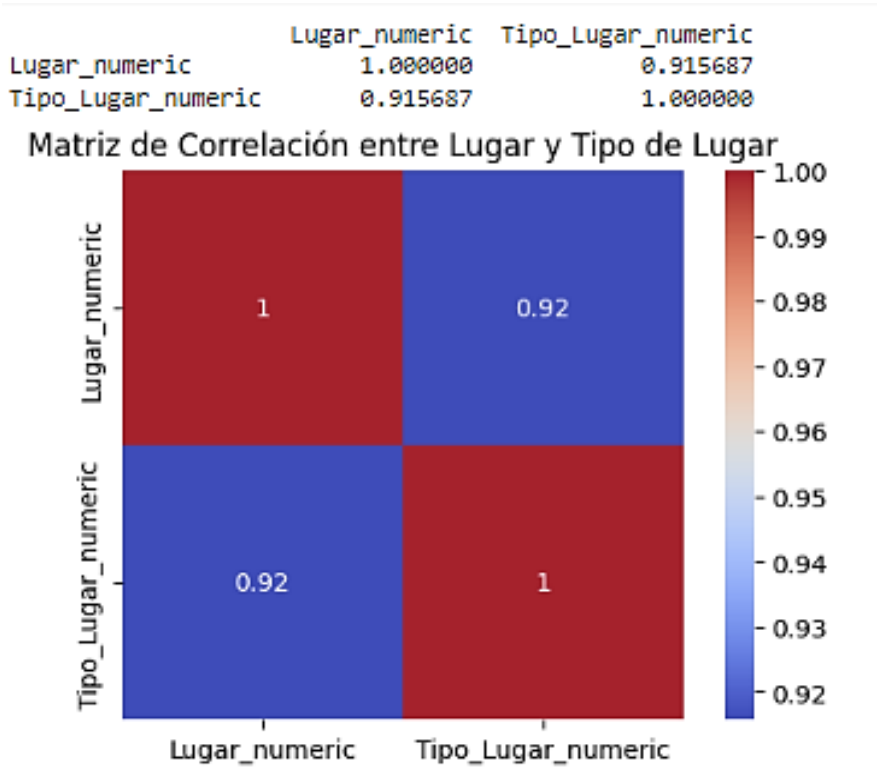


Figura 19. Correlación de variables “Lugar” y “Tipo Lugar”.

En base al estudio [50] se pudo analizar cuáles son las variables irrelevantes que se procedieron a eliminar, debido a que se consideran poco importantes para determinar patrones en los asesinatos ya que no aportan en gran medida a explicar el comportamiento de estos crímenes, estas variables son las siguientes:

- “Cod. Subcircu”
- “código de cantón”
- “Coord. Y”
- “Coord. X”
- “Genero”
- “Discapacidad”
- “Prof Reg Civ”
- “Instrucción”

- “Med. Edad”
- “Estado Civil”
- “Probable Causa M.”
- “Etnia”
- “Nacionalidad”
- “Circuito”
- “Tipo Arma” porque se consideró la variable “Arma” para el estudio.
- “Presun. Motiva. Obser” porque se consideró la variable “Presun. Motiva”.
- “Subcircuito”

En la **Tabla 4**, se puede visualizar todas las variables que se eliminaron ya que se consideran redundantes o irrelevantes para la presente investigación.

Tabla 4. Variables eliminadas.

Variables eliminadas	**Tipo de variable**
Tipo Muert.	Object
Zona	Object
Subzona	Object
provincia	Object
codigo de provincia	Object
cantón	Object
Tipo Lugar	Object
Cod. Subcircu	Object
código de cantón	Numérica
Coord. Y	Object
Coord. X	Object
Genero	Object
Discapacidad	Object
Prof Reg Civ	Object
Instruccion	Object
Med. Edad	Object
Estado Civil	Object
Probable Causa M.	Object
Etnia	Object
Nacionalidad	Object
Circuito	Object
Subcircuito	Object
Presun. Motiva. Obser	Object
Tipo Arma	Object

Luego del análisis correspondiente para excluir las variables innecesarias en el estudio, en la **Tabla 5** se tienen las siguientes que fueron seleccionadas y con las cuales se trabajó para determinar los patrones en los asesinatos.

Tabla 5. Variables seleccionadas para el estudio.

Variables seleccionadas	Tipo de variable
Distrito	Object
Area del Hecho	Object
Lugar	Object
Fecha infracción	Object
Hora infracción	Object
Arma	Object
Presun. Motiva.	Object
Edad	Numérica
Sexo	Object
Antecedentes	Object

6.1.3.3. Estandarización y limpieza de variables y registros de la base de datos

Se procedió a estandarizar los nombres de cada una de las variables, para que todas se encuentre en mayúsculas, esto se realizó en la herramienta OpenRefine, como se visualiza en la **Figura 20**, se seleccionó la variable que se quiso renombrar, luego se eligió la opción editar columna y finalmente se procede a renombrar la variable, realizando el cambio de cada uno de los nombres de los atributos.

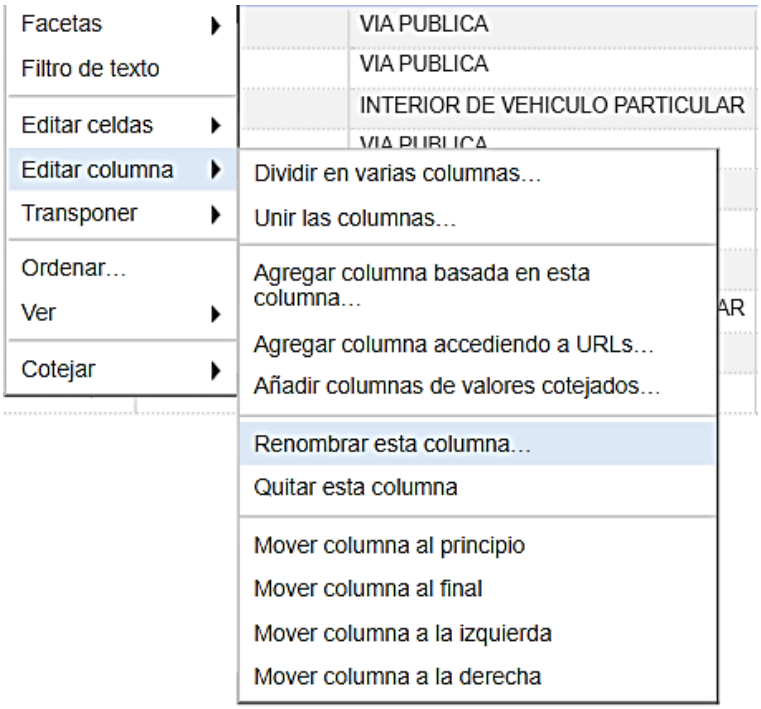


Figura 20. Renombre de variables de la base de datos.

En la **Tabla 6**, se visualizan las variables tanto antes y después de renombrarlas, esto permitió que no existan errores al momento de realizar el entrenamiento de los modelos de minería de datos, cada una de las variables que se encontraban en minúsculas se las cambió a mayúsculas.

Tabla 6. Variables Renombradas.

Nombre de variable original	Nombre de variable renombrada
Distrito	DISTRITO
Area del Hecho	AREA_DEL_HECHO
Lugar	LUGAR
Fecha infracción	FECHA_INFRACCION
Hora infracción	HORA_INFRACCION
Arma	ARMA
Presun. Motiva.	PRESUNTA_MOTIVACION
Edad	EDAD
Sexo	SEXO
Antecedentes	ANTECEDENTES

Luego de renombrar las variables de la base de datos, se realizó el reemplazo de tildes en los caracteres, en la **Tabla 7**, se visualiza el antes y después de los caracteres que se los reemplazó en el caso de existir tildes, lo mismo sucedió con la letra “Ñ/ñ”, que se la cambió por la N este proceso se realizó en todos los registros de la base de datos.

Tabla 7. Reemplazo de caracteres.

Carácter original	Caracter final
“Á/á”	“A”
“É/é”	“E”
“Í/í”	“I”
“Ó/ó”	“O”
“Ú/ú”	“U”
“Ñ/ñ”	“N”

En la **Figura 21**, se visualiza como en la herramienta OpenRefine se seleccionó la variable, eligiendo editar celdas y transformar, para proceder a realizar el reemplazo de caracteres.

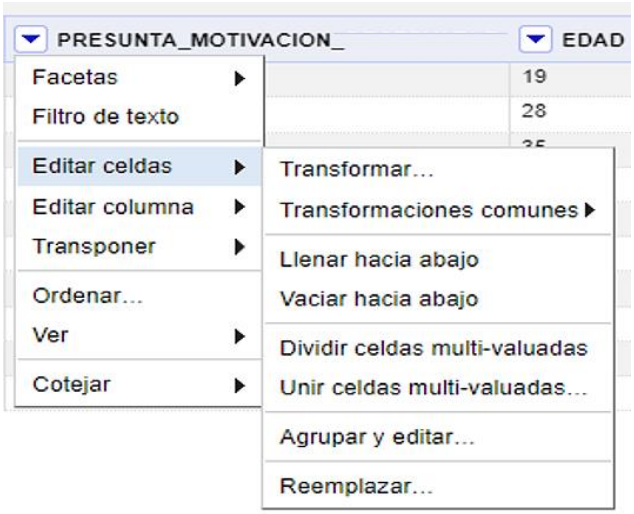


Figura 21. Pasos para cambiar caracteres con OpenRefine.

En la **Figura 22,** se visualiza como en OpenRefine se reemplazó cada una de las tildes en los registros de datos a través del comando value.replace(“valor original”, “valor reemplazado”), así: value.replace(“Ó”, “O”).

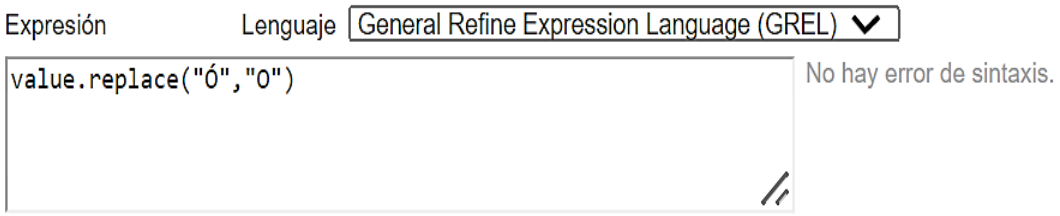


Figura 22. Limpieza de tildes en los registros de la base de datos.

En la **Figura 23,** se visualiza como en la herramienta OpenRefine se realizó el reemplazo del carácter “Ñ/ñ” por el carácter “N”, utilizando el comando value.replace(“Ñ”, “N”).

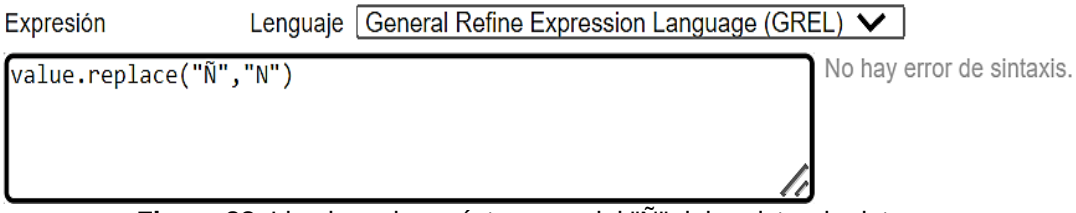


Figura 23. Limpieza de carácter especial "Ñ" del registro de datos.

En la **Figura 24**, se visualiza el resultado de eliminar la Ñ y tildes, este proceso se realizó en todos los registros y variables para que se encuentren estandarizados.

SEXO	ANTECEDENTES	DISTRITO	PRESUNTA_MOTIVACION
HOMBRE	SI	PORTETE	VIOLENCIA COMUNITARIA
HOMBRE	SI	NUEVA PROSPERINA	DELINCUENCIA COMUN
HOMBRE	SI	NUEVA PROSPERINA	DELINCUENCIA COMUN
HOMBRE	SI	PROGRESO	VIOLENCIA COMUNITARIA
HOMBRE	SI	SUR	DELINCUENCIA COMUN
HOMBRE	SI	FLORIDA	VIOLENCIA COMUNITARIA
HOMBRE	SI	PORTETE	VIOLENCIA COMUNITARIA
HOMBRE	SI	SUR	DELINCUENCIA COMUN
MUJER	SI	9 DE OCTUBRE	VIOLENCIA COMUNITARIA
HOMBRE	SI	SUR	DELINCUENCIA COMUN
HOMBRE	SI	SUR	DELINCUENCIA COMUN
HOMBRE	SI	FLORIDA	DELINCUENCIA COMUN
HOMBRE	SI	FLORIDA	DELINCUENCIA COMUN
HOMBRE	SI	FLORIDA	DELINCUENCIA COMUN
HOMBRE	SI	FLORIDA	DELINCUENCIA COMUN
HOMBRE	SI	PASCUALES	DELINCUENCIA COMUN

Figura 24. Resultados de eliminación de Ñ y tildes en los registros.

En la **Figura 25**, se visualiza como en la herramienta OpenRefine se procedió a realizar el reemplazo de caracteres en la variable “FECHA_INFRACCION” ya que contenía lo siguiente: “T10:00:00Z”, esto se realizó mediante el comando value.replace(“valor original”, “valor reemplazado”), así: value.replace(“T10:00:00Z”, “”).

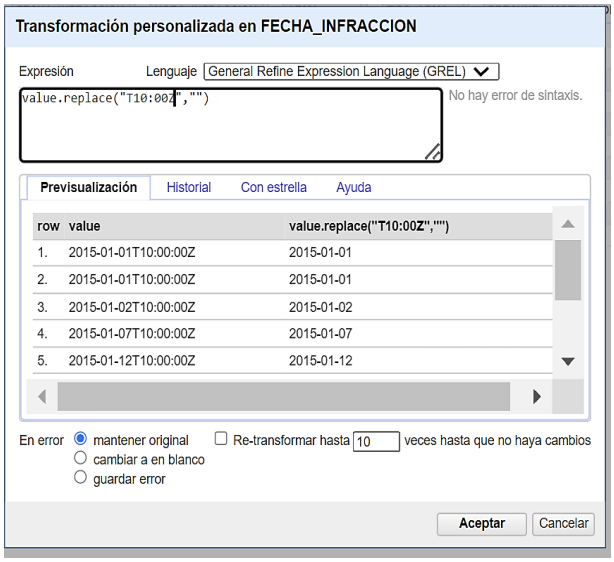


Figura 25. Transformación de variable “FECHA_INFRACCION”.

Se procedió a estandarizar los registros de la variable “HORA_INFRACCION” para poder simplificar la representación y visualización de las horas, en la **Tabla 8**, se puede observar cómo se estableció un identificador para cada rango horario.

Tabla 8. Reemplazo de variable “HORA_INFRACCION”.

Rango Horario Original	Rango Cambiado
00:00 - 00:59	**H00**
01:00 - 01:59	**H01**
02:00 - 02:59	**H02**
03:00 - 03:59	**H03**
04:00 - 04:59	**H04**
05:00 - 05:59	**H05**
06:00 - 06:59	**H06**
07:00 - 07:59	**H07**
08:00 - 08:59	**H08**
09:00 - 09:59	**H09**
10:00 - 10:59	**H10**
11:00 - 11:59	**H11**
12:00 - 12:59	**H12**
13:00 - 13:59	**H13**
14:00 - 14:59	**H14**
15:00 - 15:59	**H15**
16:00 - 16:59	**H16**
17:00 - 17:59	**H17**
18:00 - 18:59	**H18**
19:00 - 19:59	**H19**
20:00 - 20:59	**H20**
21:00 - 21:59	**H21**
22:00 - 22:59	**H22**
23:00 - 23:59	**H23**

En la **Figura 26**, se realizó la transformación de los registros de la variable "HORA_INFRACCION", en donde se siguió lo presentado en la **Tabla 8**, este proceso se realizó en la herramienta OpenRefine, mediante el comando: value.replace("valor original", "valor reemplazado"), así: value.replace("01:30", "H01").

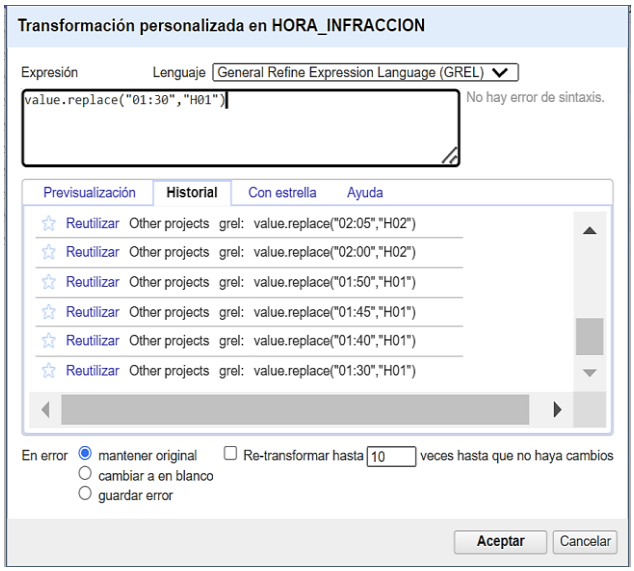


Figura 26. Transformación de registros de la variable "HORA_INFRACCION".

En la **Figura 27**, se visualiza como se creó la variable "DIA" a partir de "FECHA_INFRACCION", seleccionando la variable que se desee modificar, de las opciones desplegadas elegir editar columna y agregar una basada en la anterior.

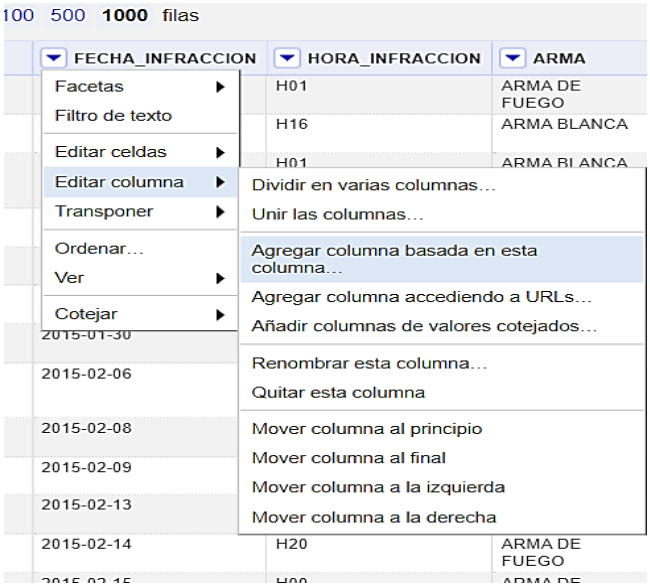


Figura 27. Pasos para la creación de la variable "DIA" en OpenRefine.

Se realizó la transformación de los registros de la variable "DIA", en donde se procesó desde el 01-01-2015 hasta el 28-02-2024, para cada fecha se asignó el día ya sea ("LUNES", "MARTES", "MIERCOLES", "JUEVES", "VIERNES", "SABADO", "DOMINGO"), para esto se usó el calendario para colocar el día correspondiente a la fecha que se encuentra en los registros de la Base de Datos, como se visualiza en la **Figura 28,** que para realizar este

proceso en la herramienta OpenRefine se usó el comando value.replace(“valor original”, “valor reemplazado”), así: value.replace(“2015-01-01”, “JUEVES”).

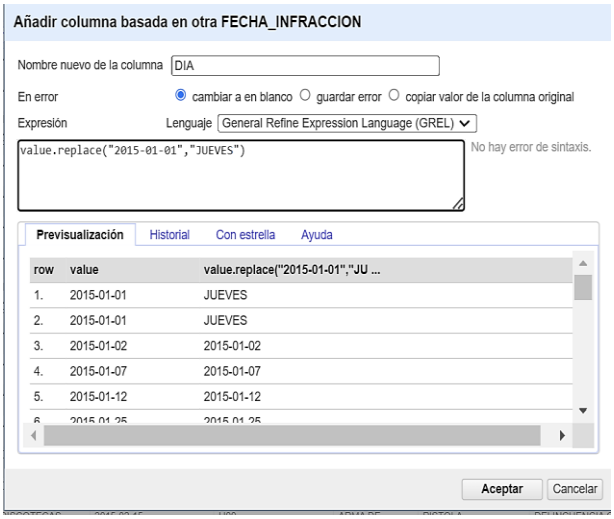


Figura 28. Transformación de registros de variable “DIA”.

En la **Figura 29**, se visualiza el resultado obtenido de las transformaciones realizadas a las variables “DIA” y “FECHA_INFRACCION”, en cada uno de los registros se realizó el proceso mencionado anteriormente.

DIA	HORA_INFRACCION
JUEVES	H01
JUEVES	H16
VIERNES	H01
MIERCOLES	H17
LUNES	H23
DOMINGO	H02
VIERNES	H22
VIERNES	H21
DOMINGO	H05
LUNES	H17
VIERNES	H17
SABADO	H20
DOMINGO	H00
DOMINGO	H00
DOMINGO	H00
MARTES	H16
MARTES	H16
SABADO	H14
DOMINGO	H23
MARTES	H10
JUEVES	H20
DOMINGO	H02

Figura 29. Resultados de transformaciones realizadas a las variables.

En la **Figura 30**, se observó que existen variables que contienen en sus registros valores “SIN_DATO”, y registros con “0” por lo que se procedió a exportar el dataset.

EDAD	SEXO
23	HOMBRE
0	HOMBRE
0	HOMBRE
34	HOMBRE
39	HOMBRE
48	HOMBRE
49	HOMBRE
48	HOMBRE
17	HOMBRE
36	HOMBRE
0	SIN_DATO
0	SIN_DATO
0	SIN_DATO

Figura 30. Valores “SIN_DATO” y “0” en registros de la base de datos.

Se procedió a cargar el dataset[4] exportado anteriormente en la herramienta Google Colab y se usó el lenguaje de programación Python para codificar, como se visualiza en la **Figura 31**, se procedió a reemplazar los registros de las variables que poseían valores “SIN_DATO” y “0” con la moda, estos cambios realizados se guardan en el dataset original.

```
import pandas as pd

# Especifica la ruta al archivo en Google Drive
file_path = '/content/drive/My Drive/ASESINATOS_ 10 VARIABLES.csv'

# Cargar el archivo CSV en un DataFrame
df = pd.read_csv(file_path)

# Verificar las primeras filas para entender la estructura del DataFrame
print("Primeras filas del DataFrame:")
print(df.head())
```

Figura 31. Carga de dataset exportado anteriormente.

En la **Figura 32**, se observa la codificación para realizar la limpieza de valores “SIN_DATO” que se encontraban en los registros de la variable “SEXO”, esto se realizó con el reemplazo de estos valores faltantes con la moda (valor que más se repite) de la columna, estos cambios realizados se guardan en el dataset original.

[4] https://bit.ly/40fCEdI

```python
import pandas as pd

# Asegurarse de que los valores de SEXO no tengan espacios ni diferencias en mayúsculas/minúsculas
df['SEXO'] = df['SEXO'].str.strip().str.upper()

# Verificar la moda de la columna SEXO
moda_sexo = df[df['SEXO'] != 'SIN_DATO']['SEXO'].mode()[0]
print(f"La moda de la columna SEXO es: {moda_sexo}")

# Reemplazar solo los registros que contienen "SIN_DATO" con la moda
df.loc[df['SEXO'] == 'SIN_DATO', 'SEXO'] = moda_sexo

# Contar el número de registros que contienen 'SIN_DATO' en la columna SEXO después del reemplazo
num_sin_dato_sexo = df[df['SEXO'] == 'SIN_DATO'].shape[0]

# Mostrar el número de registros que aún contienen 'SIN_DATO'
print(f"Número de registros que contienen 'SIN_DATO' en la columna SEXO después del reemplazo: {num_sin_dato_sexo}")

# Ver los valores únicos de la columna SEXO para asegurar que no haya variaciones de "SIN_DATO"
print("Valores únicos en la columna SEXO:")
print(df['SEXO'].unique())

# Guardar el DataFrame limpio en un nuevo archivo CSV
df.to_csv('ASESINATOS_CORRECTO.csv', index=False)  # Esto sobreescribe el archivo original
```

```
La moda de la columna SEXO es: HOMBRE
Número de registros que contienen 'SIN_DATO' en la columna SEXO después del reemplazo: 0
Valores únicos en la columna SEXO:
['HOMBRE' 'MUJER']
```

Figura 32. Limpieza de valores “SIN_DATO” en los registros de la variable “SEXO”.

En la **Figura 33**, se muestra el proceso para limpiar los valores “SIN_DATO” que se encontraban en los registros de la variable “ANTECEDENTES”, esto se realizó con el reemplazo de estos valores faltantes con la moda (valor que más se repite) de la columna, estos cambios realizados se guardan en el dataset original.

```python
import pandas as pd

# Asegurarse de que los valores de ANTECEDENTES no tengan espacios ni diferencias en mayúsculas/minúsculas
df['ANTECEDENTES'] = df['ANTECEDENTES'].str.strip().str.upper()

# Verificar la moda de la columna ANTECEDENTES
moda_antecedentes = df[df['ANTECEDENTES'] != 'SIN_DATO']['ANTECEDENTES'].mode()[0]
print(f"La moda de la columna ANTECEDENTES es: {moda_antecedentes}")

# Reemplazar solo los registros que contienen "SIN_DATO" con la moda
df.loc[df['ANTECEDENTES'] == 'SIN_DATO', 'ANTECEDENTES'] = moda_antecedentes

# Contar el número de registros que contienen 'SIN_DATO' en la columna ANTECEDENTES después del reemplazo
num_sin_dato_antecedentes = df[df['ANTECEDENTES'] == 'SIN_DATO'].shape[0]

# Mostrar el número de registros que aún contienen 'SIN_DATO'
print(f"Número de registros que contienen 'SIN_DATO' en la columna ANTECEDENTES después del reemplazo: {num_sin_dato_antecedentes}")


# Ver los valores únicos de la columna ANTECEDENTES para asegurar que no haya variaciones de "SIN_DATO"
print("Valores únicos en la columna ANTECEDENTES:")
print(df['ANTECEDENTES'].unique())

df.to_csv('ASESINATOS_CORRECTO.csv', index=False)  # Esto sobreescribe el archivo original
```

```
La moda de la columna ANTECEDENTES es: SI
Número de registros que contienen 'SIN_DATO' en la columna ANTECEDENTES después del reemplazo: 0
Valores únicos en la columna ANTECEDENTES:
['SI' 'NO']
```

Figura 33. Limpieza de valores “SIN_DATO” en los registros de la variable “ANTECEDENTES”.

En la **Figura 34**, se observa la codificación para realizar la limpieza de valores “0” que se encontraban en los registros de la variable “EDAD”, esto se realizó con el reemplazo de estos valores faltantes con la moda (valor que más se repite) de la columna, estos cambios realizados se guardan en el dataset original.

```
moda_edad = df[df['EDAD'] != 0]['EDAD'].mode()[0]
print(f"La moda de la columna EDAD es: {moda_edad}")

# Reemplazar los valores 0 en la columna EDAD con la moda
df.loc[df['EDAD'] == 0, 'EDAD'] = moda_edad

# Verificar la cantidad de registros que contenían 0 en EDAD
num_ceros_edad = df[df['EDAD'] == 0].shape[0]
print(f"Número de registros que contenían 0 en la columna EDAD después del reemplazo: {num_ceros_edad}")

# Ver los valores únicos de la columna TIPO_ARMA para asegurar que no haya variaciones de "SIN_DATO"
print("Valores únicos en la columna TIPO_ARMA:")
print(df['EDAD'].unique())

# Guardar el DataFrame limpio en un nuevo archivo CSV
df.to_csv('ASESINATOS_CORRECTO.csv', index=False)  # Esto sobreescribe el archivo original
```

```
La moda de la columna EDAD es: 30
Número de registros que contenían 0 en la columna EDAD después del reemplazo: 0
Valores únicos en la columna TIPO_ARMA:
[19 28 35 58 27 24 20 34 22 18 26 30 29 33 21 39 47 25 23 51 37 40 32 78
 49 41 43 46 69 36 45 42 14 17 16 60 55 38 59 44 54 50 65  7 31 52 67 48
 79 72 82  5 57 15  8 56 68  3  4 10 12 87 61 62 64 63 81 53 76 70  1 80
 66 73 86  6  2 74 71 13 11  9 95 94 89 77 75 85]
```

Figura 34. Limpieza de valores “0” en los registros de la variable “EDAD”.

En la **Figura 35**, se observa la codificación para realizar la exportación del dataset luego de haber realizado la limpieza de los registros que poseían valores faltantes, tanto en las que se tenía valores “SIN_DATO”, como también en las que se tenía “0”, esto se realizó guardando y luego generando el archivo[5] en formato CSV.

```
# Guardar el DataFrame limpio en un archivo CSV y XLSX
df.to_csv('ASESINATOS_BD_LIMPIA.csv', index=False)  # Guardar como CSV
```

Figura 35. Exportación de base de datos limpia.

6.1.3.4. Balanceo de clases de las variables seleccionadas para el estudio

Se procedió a balancear[6] las clases de las 10 variables seleccionadas para el presente estudio, se realizó este proceso porque existieron clases con pocos registros y otras con demasiados, como se visualiza en la **Figura 36**, en donde las clases de la variable “ARMA” se encuentran desbalanceadas y esto puede afectar las clasificaciones realizadas por los modelos de minería de datos.

[5] https://bit.ly/4fVtObd
[6] https://bit.ly/3E0WHFm

```
df['ARMA'].value_counts()
```

ARMA	count
ARMA DE FUEGO	6103
ARMA BLANCA	503
OTROS	118
ARMA CONTUNDENTE	104
CONSTRICTORA	89
SUSTANCIAS	3

dtype: int64

Figura 36. Desbalanceo de clases de variable “ARMA”.

En la **Figura 37,** se visualiza como se transformó cada una de las clases de tipo texto a numérico, este proceso se realizó para cada una de las variables del DataFrame, en este caso la transformación fue al atributo “ARMA”, se agruparon las clases “SUSTANCIAS Y OTROS” en una sola (clase 3).

```
df['ARMA'] = df['ARMA'].apply(lambda x:
                              1 if x == 'ARMA DE FUEGO' else
                              2 if x == 'ARMA BLANCA' else
                              3 if x == 'OTROS' else
                              4 if x == 'ARMA CONTUNDENTE' else
                              5 if x == 'CONSTRICTORA' else
                              3 #SUSTANCIAS
                              )
```

Figura 37. Transformación de texto a numérico de las clases de la variable “ARMA”.

Se agruparon las clases que poseían muy pocos registros y que eran multiclase, se unieron en rangos y según sus características similares, esto sirvió para simplificar y reducir la complejidad al momento de realizar el balanceo, en la **Tabla 9,** se visualizan las variables “DIA”, “EDAD”, “HORA_INFRACCION”, “DISTRITO”, “PRESUNTA_MOTIVACION” y “ARMA” con sus clases antes y después de transformarlas, como por ejemplo la variable “HORA_INFRACCION” inicialmente tenía 24 clases, por tal motivo para reducir su número y facilitar el balanceo con SMOTE, se las agrupó en rangos que van desde (H01 hasta H06 en clase 0), (H07 hasta H12 en clase 1), (H13 hasta H18 en clase 2) , (H19 hasta H00 en clase 3), así también se fue agrupando de acuerdo al número de registros que poseían las clases, las que tenían rangos mayores, menores y bajos se las agrupó respectivamente de acuerdo a este análisis, las variables con clases binarias como “LUGAR”, “AREA_DEL_HECHO”, “ATECEDENTES” y “SEXO”, no se las agrupó debido a que poseían 2 clases y el proceso de balanceo es más simple, para visualizar con más detalle visitar el enlace.[7]

[7] https://bit.ly/3E0WHFm

Tabla 9. Clases antes y después de agruparlas y transformarlas.

TRANSFORMACIÓN DE VARIABLES Y CLASES ANTES Y DESPUÉS DEL BALANCEO											
DIA		**EDAD**		**HORA_IN FRACCION**		**DISTRITO**		**PRESUNTA_MOTIVACION**		**ARMA**	
ANTES	**DES**	**ANTES**	**DES**	**ANTES**	**DES**	**ANTES**	**DES**	**ANTES**	**DES**	**ANTES**	**DES**
LUNES	3	1 - 19	1	H00	3	NUEVA PROSPERINA	1	VIOLENCIA COMUNITARIA	1	ARMA DE FUEGO	1
MARTES	2	20 - 50	2	H01	0	SUR	1	DELINCUENCIA COMUN	2	ARMA BLANCA	2
MIERCOLES	4	51 - 65	3	H02	0	PASCUALES	1	VIOLENCIA INTRAFAMILIAR	3	OTROS	3
JUEVES	3	66 - 95	4	H03	0	DURAN	2	VIOLENCIA SEXUAL	3	SUSTANCIAS	3
VIERNES	2	-	-	H04	0	ESTEROS	2	TRANSNACIONAL	4	ARMA CONTUNDENTE	4
SABADO	1	-	-	H05	0	PORTETE	2	SICOPATOLOGIAS	5	CONSTRICTORA	5
DOMINGO	1	-	-	H06	0	9 DE OCTUBRE	2	TERRORISMO	6	-	-
-	-	-	-	H07	1	FLORIDA	3	-	-	-	-
-	-	-	-	H08	1	PROGRESO	3	-	-	-	-
-	-	-	-	H09	1	MODELO	3	-	-	-	-
-	-	-	-	H10	1	SAMBORONDON	3	-	-	-	-
-	-	-	-	H11	1	CEIBOS	3	-	-	-	-
-	-	-	-	H12	1	-	-	-	-	-	-
-	-	-	-	H13	2	-	-	-	-	-	-
-	-	-	-	H14	2	-	-	-	-	-	-
-	-	-	-	H15	2	-	-	-	-	-	-
-	-	-	-	H16	2	-	-	-	-	-	-
-	-	-	-	H17	2	-	-	-	-	-	-
-	-	-	-	H18	2	-	-	-	-	-	-
-	-	-	-	H19	3	-	-	-	-	-	-
-	-	-	-	H20	3	-	-	-	-	-	-
-	-	-	-	H21	3	-	-	-	-	-	-
-	-	-	-	H22	3	-	-	-	-	-	-
-	-	-	-	H23	3	-	-	-	-	-	-

En la **Figura 38**, se visualiza el script del código que sirvió para realizar el balanceo[8] de clases con SMOTE, que genera registros simulados para cada clase, en este caso se aumentaron el número de muestras para las 5 clases de la variable ARMA, este procedimiento se realizó para todas las 10 variables seleccionadas para el presente estudio ya que todas poseían clases desequilibradas antes del balanceo, 4 eran variables con clases binarias y 6 poseían multiclases, se aplicó el mismo procedimiento a cada una de ellas.

```
import pandas as pd
from imblearn.over_sampling import SMOTE

# Supongamos que 'df' es tu DataFrame
# Variables binarias y multiclase
binary_vars = ['AREA_DEL_HECHO', 'ANTECEDENTES', 'SEXO', 'LUGAR']
multiclass_vars = [ 'DISTRITO', 'PRESUNTA_MOTIVACION',
                    'DIA', 'HORA_INFRACCION', 'EDAD']
target = 'ARMA'  # Variable target


# Paso 2: Codificar las variables categóricas usando OneHotEncoder
X = df.drop(columns=[target])  # Excluye la variable target
Y = df[target]  # Variable target

# Realizamos la codificación de las variables categóricas
X_encoded = pd.get_dummies(X, drop_first=False)


sampling_strategy = {
    1: 6103,
    2: 2003,
    3: 1305,
    4: 1205,
    5: 1152
}

# Paso 4: Aplicar SMOTE con la estrategia de muestreo personalizada
smote = SMOTE(sampling_strategy=sampling_strategy, random_state=42)
X_res, Y_res = smote.fit_resample(X_encoded, Y)

# Paso 5: Verificar el balance de clases después de aplicar SMOTE
print("Distribución de clases después de SMOTE:")
print(pd.Series(Y_res).value_counts())

# Verificar que el número total de registros es 11,768
print(f"Tamaño total del dataset después del balanceo: {X_res.shape[0]}")

# Paso 6: Exportar los datos balanceados de la variable target a archivos CSV y Excel
df_resampled_target_only = pd.DataFrame({target: Y_res})

output_file_path_csv = '/content/drive/My Drive/ARMA_BALANC_VAR_LISTO.csv'
output_file_path_xlsx = '/content/drive/My Drive/ARMA_BALANC_VAR_LISTO.xlsx'

df_resampled_target_only.to_csv(output_file_path_csv, index=False)
df_resampled_target_only.to_excel(output_file_path_xlsx, index=False)

print(f"La variable target balanceada ha sido exportada a '{output_file_path_csv}' y '{output_file_path_xlsx}'")
```

Figura 38. Aplicación de SMOTE para balancear las clases de la variable “ARMA”.

En la **Figura 39**, se visualiza el resultado de la variable “ARMA” luego de balancear sus clases con SMOTE, la distribución de registros quedó de la siguiente manera: la clase 1 con 6103 registros, la clase 2 con 2003, la clase 3 con 1305, la clase 4 con 1205 y la clase 5 con 1152 registros, sumando un total de 11,768 filas.

[8] https://bit.ly/3E0WHFm

```
Distribución de clases después de SMOTE:
ARMA
1    6103
2    2003
3    1305
4    1205
5    1152
Name: count, dtype: int64
Tamaño total del dataset después del balanceo: 11768
```

Figura 39. Distribución de clases después de balancear la variable “ARMA” con SMOTE.

En la **Figura 40,** se visualiza la combinación de todos los DataFrame generados en uno solo, este archivo[9] contiene las 10 variables con sus clases balanceadas, cuenta con un total de 11768 registros.

```python
import pandas as pd

# Lista de rutas a los archivos
input_file_paths = [
    '/content/drive/My Drive/DIA_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/HORA_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/ARMA_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/EDAD_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/LUGAR_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/SEXO_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/ANTECEDENTES_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/AREA_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/DISTRITO_BALANC_VAR_LISTO.csv',
    '/content/drive/My Drive/MOTIV_BALANC_VAR_LISTO.csv'
]

# Lista para almacenar los DataFrames
dataframes = []

# Leer cada archivo y almacenarlo en la lista
for file_path in input_file_paths:
    df = pd.read_csv(file_path)
    # Seleccionar la primera columna si el archivo tiene más de una
    if df.shape[1] > 1:
        df = df.iloc[:, 0]
    dataframes.append(df)

# Combinar todos los DataFrames como columnas en uno solo
combined_df = pd.concat(dataframes, axis=1)

# Renombrar las columnas para identificar de dónde provienen
column_names = [
    'DIA', 'HORA_INFRACCION', 'ARMA', 'EDAD', 'LUGAR',
    'SEXO', 'ANTECEDENTES', 'AREA_DEL_HECHO', 'DISTRITO', 'PRESUNTA_MOTIVACION'
]
combined_df.columns = column_names

# Guardar el DataFrame combinado en un archivo CSV
output_csv_path = '/content/drive/My Drive/DATASET_BALANCEADO_LIST.csv'
combined_df.to_csv(output_csv_path, index=False)

# Guardar el DataFrame combinado en un archivo Excel
output_excel_path = '/content/drive/My Drive/DATASET_BALANCEADO_LIST.xlsx'
combined_df.to_excel(output_excel_path, index=False, engine='openpyxl')

print(f"Dataset combinado guardado en {output_csv_path} y {output_excel_path}")
```

Figura 40. Creación de DataFrame con clases de las variables balanceadas.

[9] https://bit.ly/3C8eb25

6.1.4. Fase 4: Modelado

A continuación, se detalla cada uno de los resultados obtenidos en la fase de modelado, para esto se usó la herramienta Google Colab con el lenguaje Python para la implementación de las librerías y técnicas para el modelado y uso de la Optimización Bayesiana para construir los modelos Árbol de Decisión y Support Vector Machine con los hiperparámetros encontrados.

6.1.4.1. Uso de librerías y técnicas para modelado de datos

En la **Figura 41**, se evidencia como se importó las bibliotecas y librerías de Python que son adecuadas para llevar a cabo el proceso de minería de datos y para aplicar la técnica de Optimización Bayesiana, estas sirvieron para implementar el modelo Árbol de Decisión y Support Vector Machine, así también para generar gráficos y evaluar la precisión de los clasificadores.

```
!pip install bayesian-optimization
!pip install optuna
import numpy as np #Operaciones matemáticas rápidas sobre matrices
import pandas as pd #biblioteca de análisis y manipulación de datos para Pythor
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns #permite generar fácilmente gráficos
import statsmodels.api as sm

# Preprocesado y modelado
# ------------------------------------------------------------------------------
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
from sklearn.tree import export_graphviz
from sklearn.tree import export_text
from sklearn.model_selection import GridSearchCV
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score
```

Figura 41. Librerías de Optimización Bayesiana instaladas.

La base de datos que fue balanceada y exportada en formato CSV[10] se la guardó en el Drive, en la **Figura 42**, se visualiza el script del código que sirvió para conectar el entorno de Google Colab con el Drive personal para acceder a todos los archivos que se encuentran almacenados en el mismo.

```
from google.colab import drive
drive.mount('/content/drive')
import pandas as pd
```

```
Mounted at /content/drive
```

Figura 42. Conexión con Drive desde Google Colab.

[10] https://bit.ly/3C8eb25

En la **Figura 43**, se visualiza como se especificó la ruta de ubicación del archivo CSV, que se encontraba en Google Drive, y se lo cargó en un Dataframe que es una estructura de Python para almacenar y gestionar un dataset, luego se imprimió el contenido del DataFrame de asesinatos de la Zona 8 del Ecuador.

```
# Especifica la ruta al archivo en Google Drive
file_path = '/content/drive/My Drive/DATASET_BALANCEADO_LIST.csv'
df = pd.read_csv(file_path)
df
```

	DIA	HORA_INFRACCION	ARMA	EDAD	LUGAR	SEXO	ANTECEDENTES	AREA_DEL_HECHO	DISTRITO	PRESUNTA_MOTIVACION
0	3	0	1	1	1	1	1	1	2	1
1	3	2	2	2	1	1	1	1	1	2
2	2	0	2	2	2	1	1	1	1	2
3	4	2	1	3	1	1	1	1	3	1
4	3	3	1	2	1	1	1	1	1	2
...	...	...	...	...	...	...	...	...	...	...
11763	4	2	5	4	2	2	2	2	3	6
11764	4	2	5	4	2	2	2	2	3	6
11765	4	2	5	4	2	2	2	2	3	6
11766	4	2	5	4	2	2	2	2	3	6
11767	4	2	5	4	2	2	2	2	3	6

11768 rows × 10 columns

Figura 43. Registros y variables del DataFrame.

En la **Figura 44**, se mostró en pantalla el número de variables (10) y el número de registros (11,768) que se utilizaron en la investigación, los cuales forman parte del DataFrame ingresado al inicio.

```
ds=pd.DataFrame(df)
#Presenta el numero de filas
print("El número de filas(observaciones) es: ",ds.shape[0])

#Presenta el numero de columnas
print("El número de columnas(variables) es: ",len(ds.columns))
```

```
El número de filas(observaciones) es:  11768
El número de columnas(variables) es:  10
```

Figura 44. Registros y variables del DataFrame.

En la **Figura 45,** se imprimió cada una de las variables que se usaron en este estudio, las mismas que se cambiaron a datos numéricos, estas sirvieron para realizar el entrenamiento de los clasificadores, en total son 10 variables.

	DIA	HORA_INFRACCION	ARMA	EDAD	LUGAR	SEXO	ANTECEDENTES	AREA_DEL_HECHO	DISTRITO	PRESUNTA_MOTIVACION
0	3	0	1	1	1	1	1	1	2	1
1	3	2	2	2	1	1	1	1	1	2
2	2	0	2	2	2	1	1	1	1	2
3	4	2	1	3	1	1	1	1	3	1
4	3	3	1	2	1	1	1	1	1	2

Figura 45. Variables transformadas a datos numéricos.

En la **Figura 46**, se visualiza una matriz de correlación de las variables que a través de un mapa de calor que indica con el número 1 que existe una correlación positiva, es decir si una variable incrementa la otra también, -1 señala una correlación negativa es decir si una disminuye la otra aumenta, y un valor de 0 indica que no existe correlación entre las variables, por ejemplo el “ARMA” y el “LUGAR” tienen una correlación positiva de 0.62, en donde indica que los tipos de arma son usados dependiendo del lugar donde se realizan los asesinatos. El “AREA_DEL_HECHO” y “HORA_INFRACCION” tienen una correlación negativa de -0.48 lo que sugiere que los asesinatos pueden ocurrir en cualquier área ya sea urbana o rural y a cualquier hora del día.

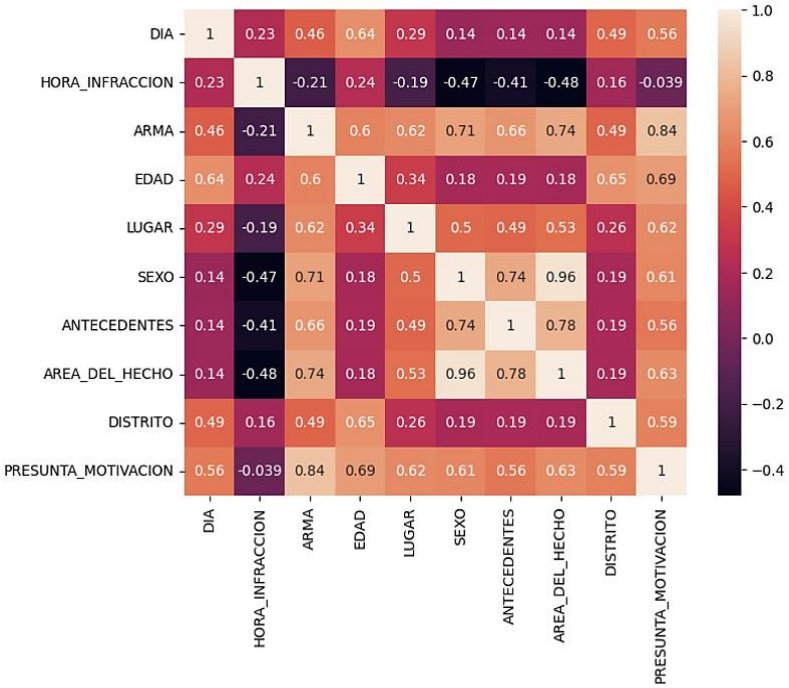


Figura 46. Matriz de correlación de variables seleccionadas para el estudio de asesinatos de la Zona 8 del Ecuador.

En la **Figura 47**, se visualiza la división de los datos del DataFrame, el 20% fue designado para el conjunto de prueba es decir 2354 registros, y para el conjunto de entrenamiento fue designado el 80% de los datos es decir 9414 registros, se realizó esto ya que es una tarea común que se realiza en el aprendizaje supervisado, para entrenar y evaluar los modelos de minería de datos.

```
from sklearn.model_selection import train_test_split


X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.20, random_state=7)

print('Son {} datos para entrenamiento y {} datos para prueba'.format(X_train.shape[0], X_test.shape[0]))
```

```
Son 9414 datos para entrenamiento y 2354 datos para prueba
```

Figura 47. División de los datos en conjunto de entrenamiento y prueba.

Se realizó la aplicación de la técnica de Optimización Bayesiana tanto con la librería Optuna y Bayesian-Optimization, para realizar la búsqueda de los hiperparámetros adecuados para construir los modelos Árbol de Decisión y Support Vector Machine (SVM), esto se encuentra detallado a continuación:

6.1.4.2. Optimización Bayesiana en Árbol de Decisión

Para aplicar la técnica de Optimización Bayesiana en el modelo Árbol de Decisión se usaron dos librerías que se encuentran disponibles en Python, la primera llamada Bayesian Optimization y la segunda Optuna, en ambos casos los rangos de búsqueda de los hiperparámetros fueron establecidos mediante los trabajos relacionados (ver en **Tabla 2**), estos rangos fueron: criterion: (gini, entropy) en donde “gini” es el criterio para medir la impureza que existe en un nodo del árbol y “entropy” mide el grado de desorden de un nodo, max depth: (1 hasta 10) es la profundidad máxima que tendrá el árbol o en cuántos niveles se dividió el mismo, min_samples spleat: (2 hasta 11) es el número mínimo de muestras en que un nodo se debe dividir en nodos hijos, min_samples_leaf: (1 hasta 25) siendo el número mínimo de muestras de un nodo hoja, también se encuentran los valores para los hiperparámetros sin usar la técnica de optimización, en la **Tabla 10,** se visualizan los rangos establecidos para la búsqueda con la técnica de OB para el Árbol de Decisión.

Tabla 10. Espacio de búsqueda de valores para los hiperparámetros de el Árbol de Decisión con la Optimización Bayesiana.

Árbol de Decisión - Configuración	
Sin Optimización Bayesiana	**Con OB - Rangos de Búsqueda**
criterion='entropy'	criterion = (gini, entropy)
max_depth=6	max_depth = (3 - 10)
min_samples_split=4	min_samples_split = (2 - 10)
min_samples_leaf=2	min_samples_leaf = (1 - 25)

En la **Figura 48**, se visualiza el script del código en Python para la aplicación de la librería Bayesian Optimization, que sirvió para realizar la búsqueda de los hiperparámetros antes descritos y seleccionar de manera automatizada el mejor, realizando para ello 100 iteraciones, evaluando cada valor de los rangos establecidos para la búsqueda, de esta manera los hiperparámetros que obtengan un mayor porcentaje de precisión se imprimieron

al final, con estos valores se pudo configurar y construir el modelo Árbol de Decisión, para esto se hizo uso de los datos de entrenamiento y prueba ya divididos con anterioridad.

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import precision_score
from bayes_opt import BayesianOptimization

def evaluate_model(criterion, max_depth, min_samples_split, min_samples_leaf):
    # Opciones para el criterio
    criterion_options = ['gini', 'entropy']
    criterion = criterion_options[int(round(criterion))]

    # Crear y entrenar el modelo
    model = DecisionTreeClassifier(
        criterion=criterion,
        max_depth=int(round(max_depth)),
        min_samples_split=int(round(min_samples_split)),
        min_samples_leaf=int(round(min_samples_leaf)),
    )
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    # Retornar la métrica objetivo (precisión)
    return precision_score(y_test, y_pred, average='micro')

# Definir el espacio de búsqueda
pbounds = {
    'criterion': (0, 1),              # 0 = 'gini', 1 = 'entropy'
    'max_depth': (3, 10),
    'min_samples_split': (2, 11),
    'min_samples_leaf': (1, 25)
}

# Configurar y ejecutar la optimización bayesiana
optimizer = BayesianOptimization(
    f=evaluate_model,
    pbounds=pbounds,
)

# Ejecutar la optimización
optimizer.maximize(init_points=10, n_iter=90)

# Obtener los mejores parámetros encontrados
best_params = optimizer.max['params']

# Convertir el índice del criterio a su nombre
best_criterion_index = int(round(best_params['criterion']))
best_criterion = ['gini', 'entropy'][best_criterion_index]

# Obtener la precisión encontrada
best_precision = optimizer.max['target']

# Imprimir los resultados
print("Mejores hiperparámetros encontrados:")
print(f"Criterion: {best_criterion}")
print(f"Max Depth: {int(round(best_params['max_depth']))}")
print(f"Min Samples Split: {int(round(best_params['min_samples_split']))}")
print(f"Min Samples Leaf: {int(round(best_params['min_samples_leaf']))}")
print(f"Precisión alcanzada: {best_precision:.4f}")
```

Figura 48. Aplicación de la librería Bayesian Optimization en el Árbol de Decisión.

En la **Figura 49,** se imprimieron los mejores valores para los hiperparámetros del modelo Árbol de Decisión, encontrados mediante la búsqueda realizada con la librería Bayesian Optimization, en este caso los más óptimos son: “criterion:entropy”, “max_depth:7”, “min_samples_split:5” y “min_samples_leaf:19”, con estos hiperparámetros se logró obtener un 98.64% para la variable “AREA_DEL_HECHO”, los hiperparámetros y porcentajes de precisión varían ya que esto depende de la variable que se esté evaluando.

```
Mejores hiperparámetros encontrados:
Criterion: entropy
Max Depth: 7
Min Samples Split: 5
Min Samples Leaf: 19
Precisión alcanzada: 0.9864
```

Figura 49. Mejores valores para los hiperparámetros encontrados con la librería Bayesian Optimization en el Árbol de Decisión.

En la **Figura 50**, se visualiza el script del código en Python para la aplicación de la librería Optuna, que sirvió para realizar la búsqueda de los hiperparámetros antes descritos y seleccionar de manera automatizada el mejor, es decir evalúa el valor que logre obtener el mayor porcentaje de precisión en las 100 iteraciones que realizó y los hiperparámetros encontrados se imprimieron al final junto al porcentaje de precisión alcanzado por los mismos, con estos valores se pudo configurar y construir el modelo Árbol de Decisión.

```python
import optuna
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import precision_score

# Función de objetivo para la optimización de hiperparámetros
def objective(trial):
    # Definición de los hiperparámetros
    criterion = trial.suggest_categorical('criterion', ['gini', 'entropy'])
    max_depth = trial.suggest_int('max_depth', 3, 10)
    min_samples_split = trial.suggest_int('min_samples_split', 2, 10)
    min_samples_leaf = trial.suggest_int('min_samples_leaf', 1, 25)

    # Crear y entrenar el modelo
    model = DecisionTreeClassifier(
        criterion=criterion,
        max_depth=max_depth,
        min_samples_split=min_samples_split,
        min_samples_leaf=min_samples_leaf
    )

    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    # Devuelve la métrica de precisión
    return precision_score(y_test, y_pred, average='micro')

# Crear un estudio
study = optuna.create_study(direction='maximize')  # Maximizar la precisión
study.optimize(objective, n_trials=100)  # Ejecutar la optimización

# Mostrar los mejores parámetros encontrados
print("Mejores hiperparámetros encontrados:")
print(study.best_params)
print("Mejor precisión:", study.best_value)
```

Figura 50. Aplicación de la librería Optuna en el Árbol de Decisión.

En la **Figura 51,** se puede visualizar el script de cómo se imprimieron los mejores valores para los hiperparámetros encontrados mediante la búsqueda realizada con la librería Optuna en el clasificador Árbol de Decisión, en este caso los más óptimos son “criterion:gini”, “max_depth:6”, “min_samples_split:8” y “min_samples_leaf:13”, con estos hiperparámetros se logró obtener un 98.64% en la métrica de precisión, esto se obtuvo para la variable “AREA_DEL_HECHO”, los hiperparámetros y porcentaje de precisión varían ya que esto depende de la variable que se esté evaluando.

```
Mejores hiperparámetros encontrados:
{'criterion': 'gini', 'max_depth': 6, 'min_samples_split': 8, 'min_samples_leaf': 13}
Mejor precisión: 0.9864061172472387
```

Figura 51. Mejores valores para los hiperparámetros encontrados con la librería Optuna en el Árbol de Decisión.

6.1.4.3. Construcción del Árbol de Decisión con los hiperparámetros encontrados mediante la técnica de Optimización Bayesiana (OB)

En la **Figura 52**, se visualiza el script del código que índica cómo se realizó la construcción del Árbol de Decisión (DT), haciendo uso de la librería sklearn.tree de Python[11] llamada DecisionTreeClassifier, configurando el árbol con los valores de los hiperparámetros antes encontrados con las librerías de Optimización Bayesiana llamadas Bayesian-Optimization y Optuna.

Los valores con los cuales se configuró el Árbol de Decisión fueron “max_depth: 7”, “min_samples_split: 5”, “min samples_leaf: 19” y “criterion: entropy”, ya que con ambas librerías se encontraron los mismos valores para los hiperparámetros, en este caso se realizó el proceso en la variable “AREA_DEL_HECHO”, pero este procedimiento se llevó a cabo en cada una de las 10 variables que fueron seleccionadas para este estudio de minería de datos para determinar patrones en el conjunto de datos de los asesinatos de la Zona 8 del Ecuador.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, classification_report
#Llamamos al constructor del arbol de decision
classifier = DecisionTreeClassifier(max_depth=7, min_samples_split=5,min_samples_leaf=19, criterion='entropy')
#Entrenamos el modelo
arbol_modelo = classifier.fit(X_train,y_train)
y_pred = classifier.predict(X_test)
```

Figura 52. Configuración del Árbol de Decisión con los hiperparámetros encontrados.

[11] https://bit.ly/4fX5tlb

6.1.4.4. Optimización Bayesiana en Support Vector Machine

Para aplicar la técnica de Optimización Bayesiana en el modelo Support Vector Machine (SVM), se usaron dos librerías que se encuentran disponibles en Python, la primera llamada Bayesian Optimization y la segunda Optuna. en ambos casos los rangos de búsqueda de los hiperparámetros fueron establecidos mediante los trabajos relacionados (ver en **Tabla 2**), estos rangos fueron: C: (1 hasta 100) es un parámetro de penalización de errores en las clasificaciones, kernel: (linear, poly, rbf, sigmoid) que indica la configuración del kernel ya sea una función lineal, polinómica, de base radial y sigmoidal, decision_function_shape: (ovo, ovr) indica la forma de la función de decisión ya sea ovo (one-vs-one, que realiza comparaciones de las clases una a una de manera más específica), u ovr (one-vs-rest realiza comparaciones de las clases más generales es decir una para todas), también se encuentran los valores para los hiperparámetros sin usar la técnica de optimización.

En la **Tabla 11**, se visualizan los rangos establecidos para realizar la búsqueda con la técnica de Optimización Bayesiana (OB) para el modelo Support Vector Machine (SVM).

Tabla 11. Espacio de búsqueda de valores para los hiperparámetros del Support Vector Machine con la Optimización Bayesiana.

Support Vector Machine - Configuración	
Sin Optimización Bayesiana	**Con OB - Rangos de Búsqueda**
C=10	C = (1 – 100)
kernel= linear	Kernel = (linear, poly, rbf, sigmoid)
decision_function_shape='ovr')	decision_function_shape = (ovo, ovr)

En la **Figura 53,** se visualiza el script del código en Python para la aplicación de la librería Bayesian Optimization, que sirvió para realizar la búsqueda de los hiperparámetros antes descritos y seleccionar de manera automatizada el mejor, es decir evalúa cada uno de los valores a través de las 25 iteraciones que realiza y los hiperparámetros que presenten el mayor porcentaje de precisión se imprimieron al final, se construyó el modelo Support Vector Machine (SVM) con los valores encontrados, para esto se hizo uso de los datos de entrenamiento y prueba ya divididos con anterioridad.

```python
import numpy as np
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score
from bayes_opt import BayesianOptimization
# Función para evaluar el modelo SVM con diferentes kernels, parámetros C y decision_function_shape
def svm_evaluate(C, kernel, decision_function_shape):
    kernel_map = {
        0: 'linear',    # Kernel lineal
        1: 'poly',      # Kernel polinómico
        2: 'rbf',       # Kernel de función de base radial
        3: 'sigmoid'    # Kernel sigmoidal
    }
    # Mapeo del valor decision_function_shape (0 = 'ovo', 1 = 'ovr')
    decision_function_shape_map = {
        0: 'ovo',       # Uno contra uno
        1: 'ovr'        # Uno contra todos
    }
    # Convertimos el kernel en un valor entero y lo mapeamos
    svc = SVC(C=C, kernel=kernel_map[int(kernel)], decision_function_shape=decision_function_shape_map[int(decision_function_shape)], random_state=7)
    # Ajustamos el modelo con los datos de entrenamiento
    svc.fit(X_train, y_train.values.ravel())
    # Predicción en el conjunto de prueba
    y_pred = svc.predict(X_test)
    # Devolvemos el precision_score ponderado
    return precision_score(y_test, y_pred, average='micro')
# Definir los límites para los hiperparámetros, incluyendo decision_function_shape
pbounds = {
    'C': (1, 100),                    # El parámetro C varía entre 1 y 100
    'kernel': (0, 3),                  # Valor entero entre 0 y 3 para mapear los kernels ('linear', 'poly', 'rbf', 'sigmoid')
    'decision_function_shape': (0, 1)    # 0 para 'ovo' y 1 para 'ovr'
}
# Inicializar el optimizador bayesiano
optimizer = BayesianOptimization(
    f=svm_evaluate,  # Función objetivo que queremos maximizar
    pbounds=pbounds, # Límites de búsqueda para los hiperparámetros
    random_state=7
)
# Ejecutar el proceso de optimización
optimizer.maximize(
    init_points=5,   # Número de exploraciones aleatorias iniciales
    n_iter=20        # Número de iteraciones de optimización
)
# Imprimir los mejores parámetros encontrados redondeados
best_params = optimizer.max['params']
best_params['C'] = round(best_params['C'], 2)
best_params['kernel'] = int(round(best_params['kernel']))
best_params['decision_function_shape'] = int(round(best_params['decision_function_shape']))
# Mapeo final de los valores encontrados
decision_function_shape_map = {0: 'ovo', 1: 'ovr'}
kernel_map = {0: 'linear', 1: 'poly', 2: 'rbf', 3: 'sigmoid'}
best_params['decision_function_shape'] = decision_function_shape_map[best_params['decision_function_shape']]
best_params['kernel'] = kernel_map[best_params['kernel']]
# Imprimir los mejores hiperparámetros encontrados
print("Mejores hiperparámetros encontrados:", best_params)
# Evaluar el modelo final con los mejores hiperparámetros
svc_best = SVC(C=best_params['C'], kernel=best_params['kernel'], decision_function_shape=best_params['decision_function_shape'], random_state=7)
svc_best.fit(X_train, y_train.values.ravel())
y_pred_best = svc_best.predict(X_test)
# Calcular precisión final
precision_final = precision_score(y_test, y_pred_best, average='micro') * 100
# Imprimir la precisión final como porcentaje
print(f"Precisión final del modelo: {precision_final:.2f}%")
```

Figura 53. Aplicación de la librería Bayesian Optimization en el SVM.

En la **Figura 54,** se imprimieron los mejores valores para los hiperparámetros del Support Vector Machine (SVM), encontrados mediante la búsqueda realizada con la librería Bayesian Optimization, en este caso los mejores hiperparámetros encontrados son: “C: 8.55”, “decision_function_shape: ovr”, “kernel: poly”, obteniendo un 98.64% en la precisión de la variable “AREA_DEL_HECHO”, estos porcentajes y valores encontrados varían dependiendo la variable que se evalúe.

```
Mejores hiperparámetros encontrados: {'C': 8.55, 'decision_function_shape': 'ovr', 'kernel': 'poly'}
Precisión final del modelo: 98.64%
```

Figura 54. Mejores valores para los hiperparámetros encontrados con la librería Bayesian Optimization en el SVM.

En la **Figura 55,** se visualiza el script del código en Python de la aplicación de la librería Optuna que sirvió para realizar la búsqueda de los hiperparámetros antes descritos y seleccionar de manera automatizada el mejor, es decir evalúa cada uno de los valores a través de las 25 iteraciones que realiza y los hiperparámetros que presenten el mayor porcentaje de precisión se imprimen al final, estos se usaron para construir el modelo SVM, haciendo uso de los datos de entrenamiento y prueba ya divididos con anterioridad.

```python
import optuna
import numpy as np
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score

# Función para evaluar el modelo SVM con diferentes kernels, parámetros C y decision_function_shape
def svm_evaluate(trial):
    # Sugerir un valor para el parámetro C entre 1 y 100
    C = trial.suggest_float('C', 1, 100)

    # Sugerir el tipo de kernel como un valor categórico
    kernel = trial.suggest_categorical('kernel', ['linear', 'poly', 'rbf', 'sigmoid'])

    # Sugerir la estrategia 'ovo' (uno contra uno) o 'ovr' (uno contra todos)
    decision_function_shape = trial.suggest_categorical('decision_function_shape', ['ovo', 'ovr'])

    # Crear el clasificador SVM con los parámetros sugeridos
    svc = SVC(C=C, kernel=kernel, decision_function_shape=decision_function_shape, random_state=7)

    # Ajustar el modelo con los datos de entrenamiento
    svc.fit(X_train, y_train.values.ravel())

    # Realizar predicciones en el conjunto de prueba
    y_pred = svc.predict(X_test)

    # Retornar el precision_score como métrica a optimizar
    return precision_score(y_test, y_pred, average='micro')

# Definir el estudio de Optuna para maximizar la precisión
study = optuna.create_study(direction='maximize')

# Ejecutar la optimización (n_trials = número de intentos)
study.optimize(svm_evaluate, n_trials=25)

# Obtener los mejores parámetros encontrados
best_params = study.best_params
print("Mejores hiperparámetros encontrados:", best_params)

# Obtener la mejor precisión alcanzada
best_precision = study.best_value
print(f"La mejor precisión alcanzada es: {best_precision * 100:.2f}%")
```

Figura 55. Aplicación de la librería Optuna en el SVM.

En la **Figura 56,** se imprimieron los mejores valores para los hiperparámetros encontrados mediante la búsqueda realizada con la librería Optuna en el clasificador SVM, en este caso los mejores hiperparámetros encontrados son para el parámetro “C: 11.98”, “decision_function_shape: ovo”, “kernel: poly”, obteniendo un 98.68% en la precisión de la variable “AREA_DEL_HECHO”, estos porcentajes y valores encontrados varían dependiendo la variable que se evalúe.

```
Mejores hiperparámetros encontrados: {'C': 11.9879863850263, 'kernel': 'poly', 'decision_function_shape': 'ovo'}
La mejor precisión alcanzada es: 98.68%
```

Figura 56. Mejores valores para los hiperparámetros encontrados con la librería Optuna en el SVM.

6.1.4.5. Construcción del Support Vector Machine con los hiperparámetros encontrados mediante la OB.

En la **Figura 57,** se visualiza como se realizó la construcción del Support Vector Machine (SVM) para la variable "AREA_DEL_HECHO", haciendo uso de la librería sklearn.svm de Python[12] llamada SVC, configurando el modelo con los valores de los hiperparámetros antes encontrados con las librerías de Optimización Bayesiana y Optuna, estos valores fueron "kernel: poly", "C: 11.98", "decision_function_shape: ovo".

```
# Cargamos la librería Support Vector Classifier
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix, classification_report

# Llamamos al constructor de Support Vector Machine
classifier = SVC(kernel='poly', C=11.98, decision_function_shape='ovo')

# Entrenamos el modelo
svm_modelo = classifier.fit(X_train, y_train)

# Realizamos predicciones en el conjunto de prueba
y_pred = classifier.predict(X_test)
```

Figura 57. Configuración del SVM con los hiperparámetros encontrados.

6.2. Objetivo 2: Evaluar los clasificadores Árbol de Decisión y SVM con la métrica de precisión.

Para cumplir con el segundo objetivo específico del presente Trabajo de Integración Curricular (TIC), se ejecutó las dos últimas fases llamadas evaluación y despliegue de la metodología CRISP-DM, las mismas que se las mostrará a continuación:

6.2.1. Fase 5: Evaluación

En esta fase trabajada sobre el conjunto de prueba (20%) se evidenció y comparó los porcentajes alcanzados en las métricas de precisión, recall y accuracy en el modelo Árbol de Decisión y en el modelo Support Vector Machine, se obtuvo el promedio final de las 10 variables evaluadas tanto con la técnica de validación cruzada así también con la librería sklearn.metrics que ofrece Python en Google Colab para evaluar la métricas de rendimiento de los modelos de clasificación.

[12] https://bit.ly/40stXhv

6.2.1.1. Comparación de los porcentajes de precisión en los clasificadores Árbol de Decisión y Support Vector Machine

En la **Tabla 12,** se muestra la comparación de los porcentajes de precisión alcanzados tanto antes y después de aplicar la técnica de Optimización Bayesiana (OB) con 100 iteraciones en el Árbol de Decisión, se observa que si existe un aumento de la precisión al aplicar la técnica de OB.

Tabla 12. Comparación de los porcentajes de precisión obtenidos antes y después de aplicar la Optimización Bayesiana con 100 iteraciones en el Árbol de Decisión.

Árbol de Decisión						
Precisión	**Sin Optimización Bayesiana**		**Con Optimización Bayesiana – *100 Iteraciones***			
	Sin Librerías de Optimización		**Bayesian Optimization (Librería de OB)**		**Optuna (Librería de OB)**	
Variables - Evaluación	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**
AREA_DEL_HECHO	98,48	98,47	98,55	98,64	98,55	98,64
DIA	58,45	59,13	58,3	60,11	58,38	60,11
ARMA	89,95	89,71	89,97	90,01	90,03	90,05
EDAD	88,00	88,91	87,92	89,03	88,15	89,03
SEXO	97,94	98,21	97,93	98,21	97,94	98,21
HORA_INFRACCION	66,33	66,10	66,65	66,73	66,61	66,73
LUGAR	87,05	87,21	87,00	87,17	87,11	87,42
ANTECEDENTES	90,00	90,61	89,89	90,39	90,06	90,82
PRESUNTA_MOTIVACION	89,40	89,12	89,88	89,46	90,51	90,61
DISTRITO	68,12	68,18	68,97	68,86	68,97	68,86
TOTAL:	**83,37**	**83,56**	**83,50**	**83,86**	**83,63**	**84,04**

En la **Figura 58,** se muestra la gráfica estadística de la comparación de los porcentajes de precisión alcanzados tanto antes y después de aplicar la técnica de Optimización Bayesiana con 100 iteraciones en el Árbol de Decisión, se observa que si existió una mejora en los porcentajes de precisión al aplicar la OB, además se visualiza que con la librería Optuna se alcanzó un porcentaje en la precisión del 83.63% al evaluarlo con la validación cruzada y un 84.04% al evaluarlo con la librería sklearn.metrics de Python, con la librería Bayesian Optimization se alcanzó un porcentaje en la precisión del 83.50% al evaluarlo con la validación cruzada y un 83.86% al evaluarlo con la librería sklearn.metrics de Python. Sin aplicar la OB los porcentajes de precisión son menores alcanzando un 83.37% y 83.56% al evaluarlos con la validación cruzada y con la librería sklearn.metrics respectivamente.

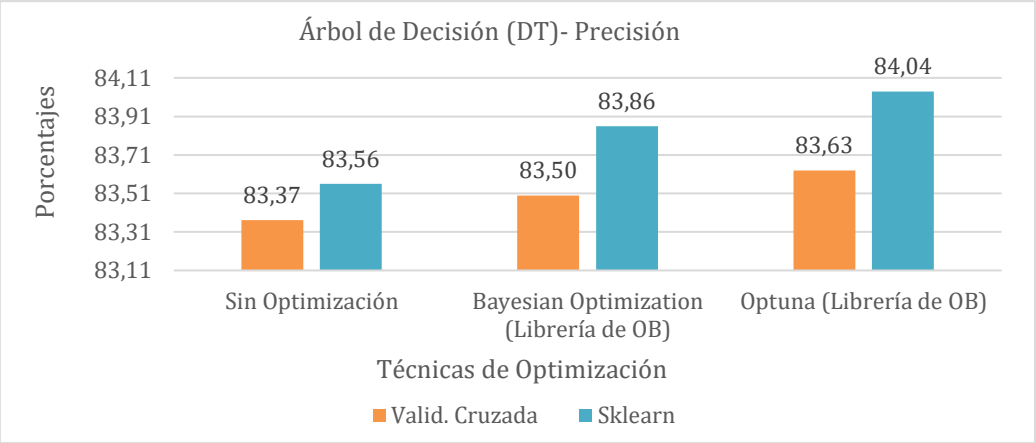


Figura 58. Gráfico de barras de precisión alcanzada antes y después de aplicar la Optimización Bayesiana con 100 iteraciones en el modelo Árbol de Decisión.

En la **Tabla 13,** se muestra la comparación de los porcentajes de precisión alcanzados tanto antes y después de aplicar la técnica de Optimización Bayesiana (OB) con 25 iteraciones en el modelo Support Vector Machine, se observa que si existe un aumento de la precisión al aplicar la técnica de OB.

Tabla 13. Comparación de los porcentajes de precisión obtenidos antes y después de aplicar la Optimización Bayesiana con 25 iteraciones en el Support Vector Machine.

Support Vector Machine (SVM)						
Precisión	**Sin Optimización Bayesiana**		**Con Optimización Bayesiana – *25 iteraciones***			
	Sin Librerías de Optimización		**Bayesian Optimization (Librería de OB)**		**Optuna (Librería de OB)**	
Variables - Evaluación	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**
AREA_DEL_HECHO	98,31	98,47	98,51	98,64	98,51	98,68
DIA	57,79	57,60	59,88	59,81	59,08	60,91
ARMA	89,32	88,82	90,20	90,27	90,20	90,14
EDAD	85,79	85,34	88,21	89,25	88,19	89,25
SEXO	97,94	98,21	97,94	98,21	97,94	98,21
HORA_INFRACCION	64,89	65,20	67,19	66,86	67,25	66,86
LUGAR	85,05	85,08	87,00	87,00	87,06	87,17
ANTECEDENTES	88,60	88,53	90,35	90,61	90,35	90,69
PRESUNTA_MOTIVACION	88,52	88,57	90,53	90,99	90,40	90,99
DISTRITO	67,19	66,27	68,86	68,22	68,85	68,30
TOTAL:	**82,34**	**82,20**	**83,86**	**83,98**	**83,78**	**84,12**

En la **Figura 59,** se muestra la gráfica estadística de la comparación de los porcentajes de precisión alcanzados tanto antes y después de aplicar la técnica de Optimización Bayesiana con 25 iteraciones en el modelo Support Vector Machine, se observa que si existió una mejora en los porcentajes de precisión al aplicar la OB en el modelo SVM, además se visualiza que con la librería Optuna se alcanzó un porcentaje en la precisión del 83.78% al evaluarlo con la validación cruzada y un 84.12% al evaluarlo con la librería sklearn.metrics de Python, con la librería Bayesian Optimization se alcanzó un porcentaje en la precisión del

83.86% al evaluarlo con la validación cruzada y un 83.98% al evaluarlo con la librería sklearn.metrics de Python. Sin aplicar la OB los porcentajes de precisión son menores alcanzando un 82.34% y 82.20% al evaluarlos con la validación cruzada y con la librería sklearn.metrics respectivamente.

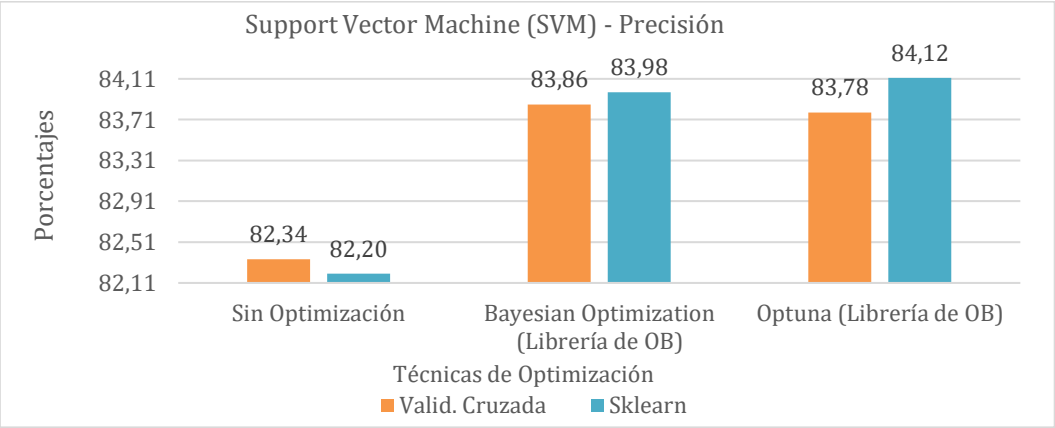


Figura 59. Gráfico de barras de precisión alcanzada antes y después de aplicar la Optimización Bayesiana con 25 iteraciones en el modelo Support Vector Machine.

6.2.1.2. Comparación de los porcentajes de accuracy "exactitud" en los clasificadores Árbol de Decisión y Support Vector Machine

En la **Tabla 14,** se presenta una comparación detallada de los porcentajes de *accuracy* obtenidos antes y después de aplicar la técnica de Optimización Bayesiana (OB) con un total de 100 iteraciones en el modelo de Árbol de Decisión. Los resultados permiten observar un incremento significativo en los niveles de accuracy tras la implementación de esta técnica, lo que evidencia la eficacia de la Optimización Bayesiana para ajustar de manera más precisa los hiperparámetros del modelo.

Tabla 14. Comparación de los porcentajes de accuracy obtenidos antes y después de aplicar la Optimización Bayesiana con 100 iteraciones en el Árbol de Decisión.

Árbol de Decisión						
Accuracy	**Sin Optimización Bayesiana**		**Con Optimización Bayesiana – *100 Iteraciones***			
	Sin Librerías de Optimización		**Bayesian Optimization (Librería de OB)**		**Optuna (Librería de OB)**	
Variables - Evaluación	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**
AREA_DEL_HECHO	98,48	98,47	98,55	98,64	98,55	98,64
DIA	58,46	59,13	58,31	60,11	58,37	60,11
ARMA	89,92	89,71	89,97	90,01	90,04	90,05
EDAD	88,00	88,91	87,93	89,03	88,15	89,03
SEXO	97,94	98,21	97,93	98,21	97,94	98,21
HORA_INFRACCION	66,33	66,10	66,62	66,73	66,59	66,73
LUGAR	87,04	87,21	87,00	87,17	87,07	87,42
ANTECEDENTES	90,00	90,61	89,92	90,39	90,08	90,82
PRESUNTA_MOTIVACION	89,38	89,12	89,86	89,46	90,51	90,61
DISTRITO	68,13	68,18	68,97	68,86	68,97	68,86
TOTAL:	**83,36**	**83,56**	**83,50**	**83,86**	**83,62**	**84,04**

En la **Figura 60,** se muestra la gráfica estadística de la comparación de los porcentajes de accuracy alcanzados, se visualiza que con la librería Optuna se alcanzó un porcentaje del 83.62% al evaluarlo con la validación cruzada y un 84.04% al evaluarlo con la librería sklearn.metrics de Python, con la librería Bayesian Optimization se alcanzó un porcentaje de accuracy del 83.50% al evaluarlo con la validación cruzada y un 83.86% al evaluarlo con la librería sklearn.metrics de Python. Sin aplicar la Optimización Bayesiana los porcentajes de accuracy son menores alcanzando un 83.36% y 83.56% al evaluarlos con la validación cruzada y con la librería sklearn.metrics respectivamente.

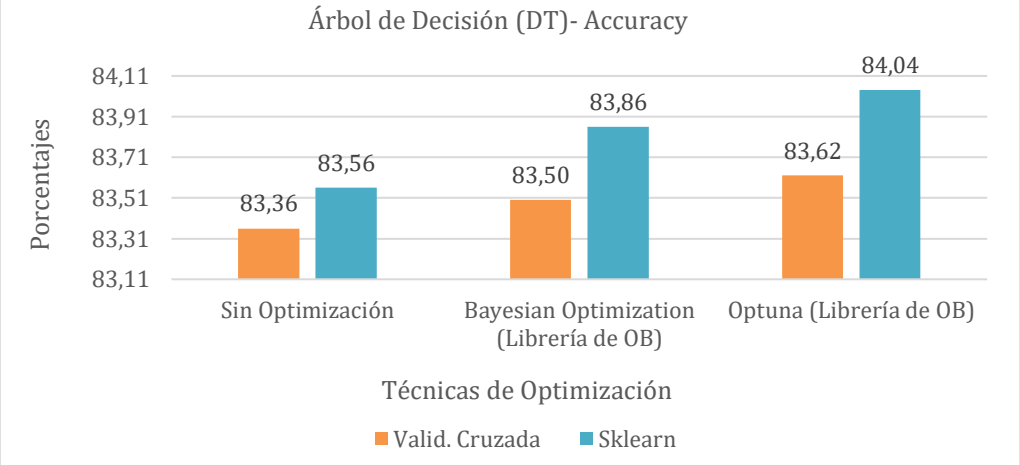


Figura 60. Gráfico de barras de accuracy alcanzada antes y después de aplicar la Optimización Bayesiana con 100 iteraciones en el modelo Árbol de Decisión.

En la **Tabla 15,** se muestra la comparación de los porcentajes de accuracy alcanzados tanto antes y después de aplicar la técnica de Optimización Bayesiana (OB) con 25 iteraciones en el modelo Support Vector Machine.

Tabla 15. Comparación de los porcentajes de accuracy obtenidos antes y después de aplicar la Optimización Bayesiana con 25 iteraciones en el Support Vector Machine.

Support Vector Machine (SVM)						
Accuracy	**Sin Optimización Bayesiana**		**Con Optimización Bayesiana – *25 iteraciones***			
	Sin Librerías de Optimización		**Bayesian Optimization (Librería de OB)**		**Optuna (Librería de OB)**	
Variables - Evaluación	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**
AREA_DEL_HECHO	98,31	98,47	98,51	98,64	98,51	98,68
DIA	57,79	57,6	59,88	59,81	59,08	60,91
ARMA	89,32	88,82	90,20	90,27	90,20	90,14
EDAD	85,79	85,34	88,21	89,25	88,19	89,25
SEXO	97,94	98,21	97,94	98,21	97,94	98,21
HORA_INFRACCION	64,89	65,20	67,19	66,86	67,25	66,86
LUGAR	85,05	85,08	87,00	87,00	87,06	87,17
ANTECEDENTES	88,60	88,53	90,35	90,61	90,35	90,69
PRESUNTA_MOTIVACION	88,52	88,57	90,53	90,99	90,40	90,99
DISTRITO	67,19	66,27	68,86	68,22	68,85	68,30
TOTAL:	**82,34**	**82,20**	**83,86**	**83,98**	**83,78**	**84,12**

En la **Figura 61,** se muestra la gráfica estadística de la comparación de los porcentajes de accuracy alcanzados tanto antes y después de aplicar la técnica de Optimización Bayesiana con 25 iteraciones en el modelo Support Vector Machine (SVM), se observa que sí existió una mejora en los porcentajes de las métricas evaluadas al aplicar la Optimización Bayesiana (OB) en el modelo SVM.

Se visualiza que con la librería de OB llamada Optuna, se alcanzó un porcentaje de accuracy del 83.78%, al evaluarlo con la validación cruzada y un 84.12%, al evaluarlo con la librería sklearn.metrics de Python, mientras que, con la librería Bayesian Optimization se alcanzó un porcentaje del 83.86%, al evaluarlo con la validación cruzada y un 83.98%, al evaluarlo con la librería sklearn.metrics de Python.

Sin aplicar la Optimización Bayesiana los porcentajes de accuracy son menores alcanzando un 82.34% y un 82.20%, al evaluarlos con la validación cruzada y con la librería sklearn.metrics respectivamente.

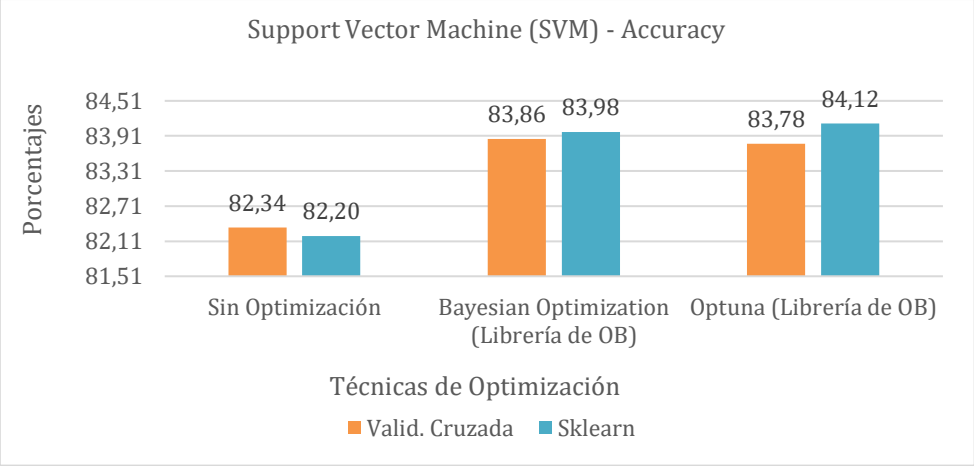


Figura 61. Gráfico de barras de accuracy alcanzada antes y después de aplicar la Optimización Bayesiana con 25 iteraciones en el modelo Support Vector Machine.

6.2.1.3. Comparación de los porcentajes de recall en los clasificadores Árbol de Decisión y Support Vector Machine

En la **Tabla 16**, se muestra la comparación de los porcentajes de recall alcanzados tanto antes y después de aplicar la Optimización Bayesiana (OB) con 100 iteraciones en el Árbol de Decisión, se observa que si existe un aumento del recall al aplicar la técnica de optimización. La OB ajusta de manera precisa los hiperparámetros del modelo, optimizando su desempeño y logrando un equilibrio más adecuado entre las métricas de evaluación.

Estos resultados confirman que la Optimización Bayesiana no solo mejora la precisión global, sino que también fortalece aspectos específicos del modelo, como su sensibilidad, haciendo que sea más efectivo en contextos prácticos.

Tabla 16. Comparación de los porcentajes de recall obtenidos antes y después de aplicar la Optimización Bayesiana con 100 iteraciones en el Árbol de Decisión.

Árbol de Decisión						
Recall	**Sin Optimización Bayesiana**		**Con Optimización Bayesiana – *100 Iteraciones***			
	Sin Librerías de Optimización		**Bayesian Optimization (Librería de OB)**		**Optuna (Librería de OB)**	
Variables - Evaluación	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**
AREA_DEL_HECHO	98,47	98,47	98,55	98,64	98,55	98,64
DIA	58,45	59,13	58,31	60,11	58,37	60,11
ARMA	89,95	89,71	89,96	90,01	90,04	90,05
EDAD	88,00	88,91	87,92	89,03	88,16	89,03
SEXO	97,94	98,21	97,93	98,21	97,94	98,21
HORA_INFRACCION	66,33	66,10	66,62	66,73	66,58	66,73
LUGAR	87,04	87,21	87,00	87,17	87,07	87,42
ANTECEDENTES	90,00	90,61	89,89	90,39	90,06	90,82
PRESUNTA_MOTIVACION	89,40	89,12	89,86	89,46	90,51	90,61
DISTRITO	68,12	68,18	68,97	68,86	68,97	68,86
TOTAL:	**83,37**	**83,56**	**83,50**	**83,86**	**83,62**	**84,04**

En la **Figura 62,** se muestra la gráfica estadística de la comparación de los porcentajes de recall alcanzados, se visualiza que con la librería Optuna se alcanzó un porcentaje del 83.62% de recall, esto fue al evaluarlo con la validación cruzada y se alcanzó un 84.04% al momento de evaluarlo con la librería sklearn.metrics de Python, mientras que con la librería Bayesian-Optimization se logró alcanzar un porcentaje del 83.50% en la métrica de recall al momento de evaluarlo con la validación cruzada y un 83.86% al evaluarlo con la librería sklearn.metrics de Python. Sin aplicar la técnica de OB los porcentajes en la métrica de recall son menores, alcanzando un 83.37% y 83.56% al evaluarlos con la validación cruzada y con la librería sklearn.metrics respectivamente.

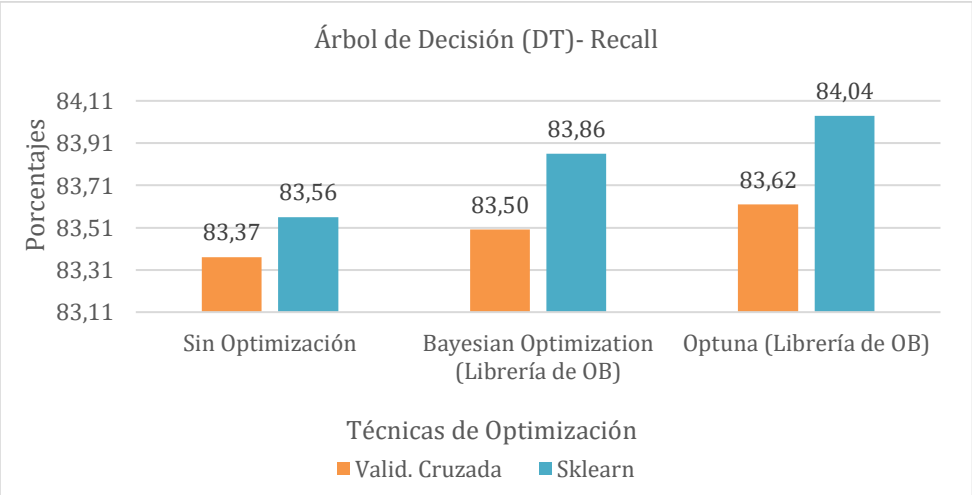


Figura 62. Gráfico de barras de recall alcanzada antes y después de aplicar la Optimización Bayesiana con 100 iteraciones en el modelo Árbol de Decisión.

En la **Tabla 17,** se muestra la comparación de los porcentajes de recall alcanzados tanto antes y después de aplicar la técnica de Optimización Bayesiana (OB) con 25 iteraciones en el modelo Support Vector Machine.

Tabla 17. Comparación de los porcentajes de recall obtenidos antes y después de aplicar la Optimización Bayesiana con 25 iteraciones en el Support Vector Machine.

Support Vector Machine (SVM)						
Recall	**Sin Optimización Bayesiana**		**Con Optimización Bayesiana – *25 iteraciones***			
	Sin Librerías de Optimización		**Bayesian Optimization (Librería de OB)**		**Optuna (Librería de OB)**	
Variables - Evaluación	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**	**Valid. Cruzada (%)**	**Sklearn (%)**
AREA_DEL_HECHO	98,31	98,47	98,51	98,64	98,51	98,68
DIA	57,79	57,6	59,88	59,81	59,08	60,91
ARMA	89,32	88,82	90,20	90,27	90,20	90,14
EDAD	85,79	85,34	88,21	89,25	88,19	89,25
SEXO	97,94	98,21	97,94	98,21	97,94	98,21
HORA_INFRACCION	64,89	65,20	67,19	66,86	67,25	66,86
LUGAR	85,05	85,08	87,00	87,00	87,06	87,17
ANTECEDENTES	88,60	88,53	90,35	90,61	90,35	90,69
PRESUNTA_MOTIVACION	88,52	88,57	90,53	90,99	90,40	90,99
DISTRITO	67,19	66,27	68,86	68,22	68,85	68,30
TOTAL:	**82,34**	**82,20**	**83,86**	**83,98**	**83,78**	**84,12**

En la **Figura 63,** se muestra la gráfica estadística de la comparación de los porcentajes de recall alcanzados, con la librería Optuna se obtuvo un porcentaje de recall del 83.78% al evaluarlo con la validación cruzada y un 84.12% al evaluarlo con sklearn.metrics, con la librería Bayesian Optimization se alcanzó un porcentaje del 83.86% (validación cruzada) y un 83.98% con sklearn.metrics. Sin aplicar la OB los porcentajes de recall son menores alcanzando un 82.34% y 82.20% al evaluarlos con validación cruzada y con sklearn.metrics respectivamente.

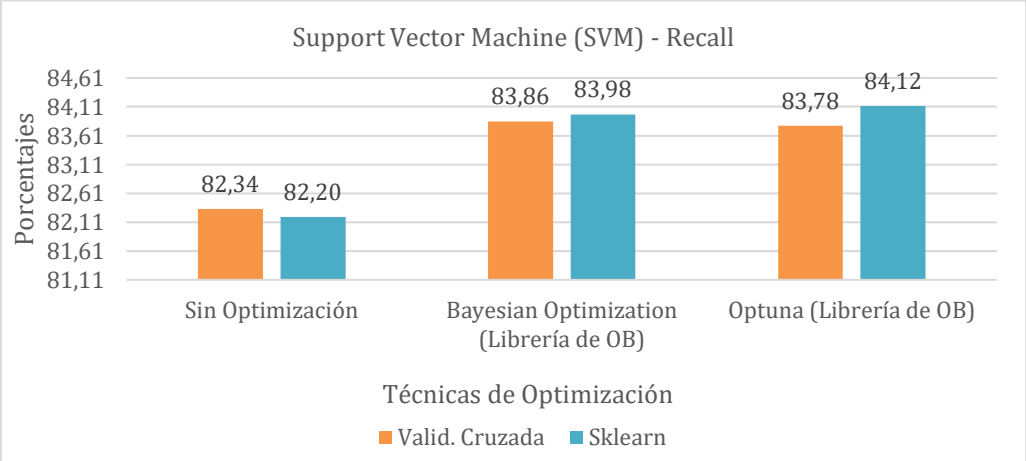


Figura 63. Gráfico de barras de recall alcanzada antes y después de aplicar la Optimización Bayesiana con 25 iteraciones en el modelo Support Vector Machine.

En la **Tabla 18**, se detallan los hiperparámetros identificados por la librería Optuna, seleccionados por su capacidad para lograr mayores porcentajes en las métricas del modelo de Árbol de Decisión (DT). Estos hiperparámetros se presentan para cada una de las 10 variables seleccionadas en este estudio, ya que cada variable posee un conjunto único de parámetros optimizados mediante la técnica de Optimización Bayesiana, la cual permite explorar eficientemente el espacio de búsqueda de hiperparámetros, mejorando la precisión del modelo y reduciendo el riesgo de sobreajuste. Además, la selección de estos parámetros permite una configuración más precisa.

Tabla 18. Hiperparámetros obtenidos con la librería Optuna en el modelo Árbol de Decisión para cada una de las variables.

Árbol de Decisión - Optuna				
Variables	**Hiperparámetros**			
	max depth	**min samples split**	**min samples leaf**	**criterion**
AREA_DEL_HECHO	6	7	17	entropy
DIA	7	8	6	entropy
ARMA	6	8	8	gini
EDAD	5	3	2	entropy
SEXO	9	5	11	gini
HORA_INFRACCION	7	3	1	entropy
LUGAR	9	5	12	entropy
ANTECEDENTES	7	7	4	entropy
PRESUNTA_MOTIVACION	10	10	25	gini
DISTRITO	10	10	12	entropy

En la **Tabla 19,** se detallan los hiperparámetros identificados por la librería Optuna, seleccionados por su capacidad para lograr mayores porcentajes en las métricas del modelo Support Vector Machine (SVM). Estos hiperparámetros se presentan para cada una de las 10 variables seleccionadas en este estudio, ya que cada variable posee un conjunto único de parámetros optimizados mediante la técnica de Optimización Bayesiana.

Tabla 19. Hiperparámetros obtenidos con la librería Optuna en el modelo SVM para cada una de las variables.

Support Vector Machine - Optuna			
Variables	**Hiperparámetros**		
	kernel	**C**	**decision function shape**
AREA_DEL_HECHO	poly	11,98	ovo
DIA	poly	69,70	ovo
ARMA	poly	77,49	ovr
EDAD	poly	1,59	ovr
SEXO	linear	55,00	ovo
HORA_INFRACCION	rbf	49,48	ovr
LUGAR	rbf	44,78	ovr
ANTECEDENTES	rbf	9,22	ovo
PRESUNTA_MOTIVACION	rbf	14,30	ovo
DISTRITO	rbf	66,13	ovo

6.2.1.4. Interpretación de resultados de cada variable obtenidos con el modelo Árbol de Decisión configurados mediante la librería Optuna

El modelo Árbol de Decisión (DT), que se configuró con los hiperparámetros obtenidos con la librería de Optimización Bayesiana Optuna, fue la que alcanzó los mayores porcentajes de precisión, accuracy y recall, siendo mejor que los resultados obtenidos por la librería Bayesian-Optimization.

A continuación, se realizó la interpretación de los resultados alcanzados para cada una de las 10 variables seleccionadas para este estudio con las clasificaciones obtenidas luego de configurar con Optuna el modelo Árbol de Decisión. Este análisis detallado permitió una mejor comprensión de los patrones resultantes en el conjunto de datos, mostrando la importancia de la selección adecuada de hiperparámetros para obtener modelos más precisos y exactos.

- **Variable "AREA_DEL_HECHO"**

En la **Figura 64**, se visualiza la estructura del árbol para la variable "AREA_DEL_HECHO", cuyo nodo raíz es la variable "SEXO <= 1.5", que divide las muestras en dos ramas: True y False, separando aquellas que cumplen o no con la condición principal, es decir, las que son menores o iguales a 1.5 y las que son mayores. A partir de esta división, el árbol continúa con nodos intermedios que evalúan variables como "HORA_INFRACCION", "LUGAR", "DISTRITO", "EDAD", "ANTECEDENTES", "ARMA" y "DIA", entre otras, mostrando métricas clave como "entropy" (desorden del nodo), "simples" (número de muestras), y "value" (distribución de las clases). A medida que el árbol avanza, las decisiones se vuelven más específicas, llevando a nodos finales que revelan las predicciones del modelo. En algunos casos, se observan nodos con entropía cercana a cero, lo que indica predicciones puras para una clase específica, mientras que, en otros, la diversidad de muestras genera nodos con entropía más alta, para ver la estructura completa del árbol visitar el enlace[13].

[13] https://bit.ly/3DRYUTL

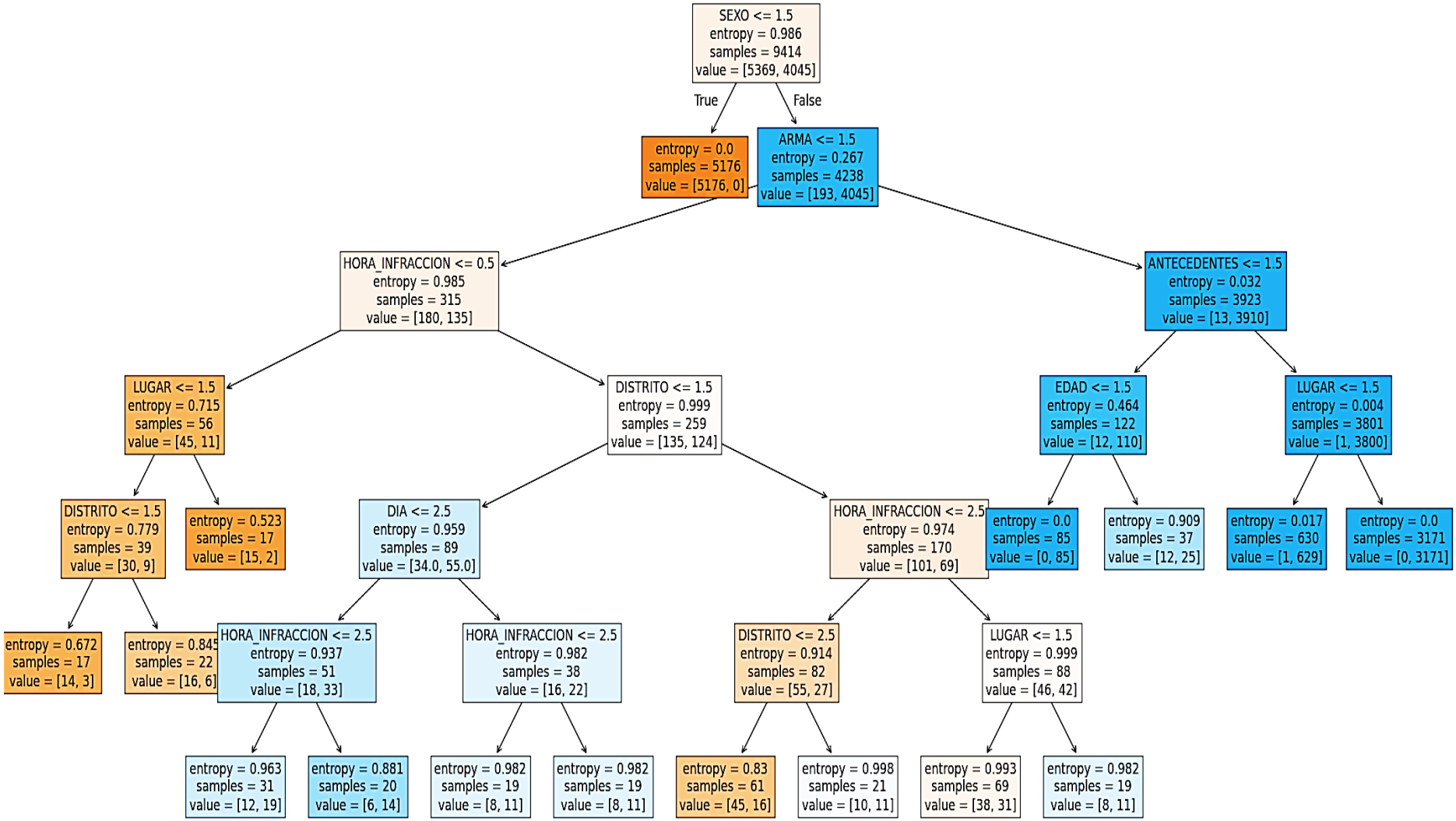


Figura 64. Árbol de decisión para la variable “AREA_DEL_HECHO”.

En la **Figura 65,** se muestra la matriz de confusión para la variable "AREA_DEL_HECHO", en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1330 verdaderos negativos y 10 falsos positivos, mientras que la clase 1 cuenta con 992 verdaderos positivos y 22 falsos negativos.

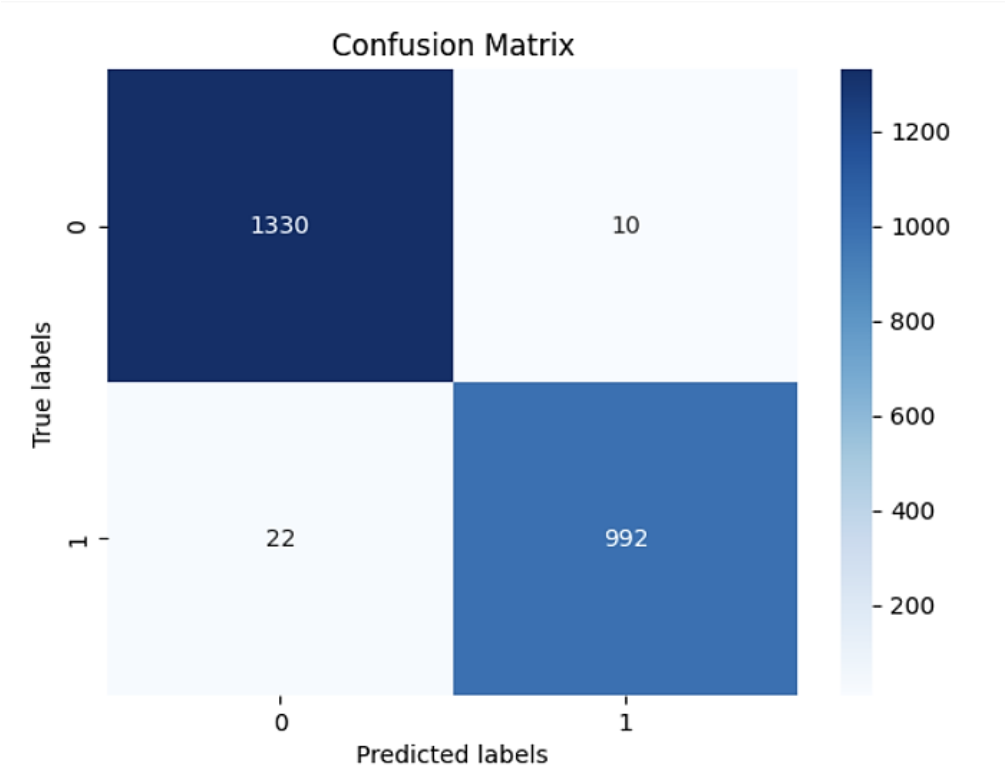


Figura 65. Matriz de confusión para la variable "AREA_DEL_HECHO" con el modelo Árbol de Decisión configurado con Optuna.

La **Tabla 20**, muestra el desempeño del modelo Árbol de Decisión para la variable "AREA_DEL_HECHO", teniendo dos clases: Urbano siendo la clase 0 y Rural la clase 1. La precisión para la clase Urbano es del 98,37%, teniendo un recall del 99,25%, mientras que la clase Rural tuvo una precisión del 99,00% y un recall del 97,83%, el modelo presentó una accuracy global de 98,64%. Para esta variable se evidenció un excelente desempeño del modelo en la clasificación de las clases.

Tabla 20. Desempeño del modelo Árbol de Decisión para la variable "AREA_DEL_HECHO" configurada con los hiperparámetros de la librería Optuna.

AREA_DEL_HECHO		
Clase	**Precisión**	**Recall**
0: Urbano	98,37 %	99,25 %
1: Rural	99,00 %	97,83 %
Accuracy	98,64 %	98,64 %

En la **Figura 66,** se visualiza el gráfico de barras para la variable "AREA_DEL_HECHO" que está representando la distribución de asesinatos en dos

categorías: Urbano y Rural, la primera tuvo un 57,40% del total, mientras que la segunda tuvo un 42,60% respectivamente, por lo tanto, se deduce que la mayoría de los asesinatos ocurren en áreas urbanas de la Zona 8 del Ecuador.

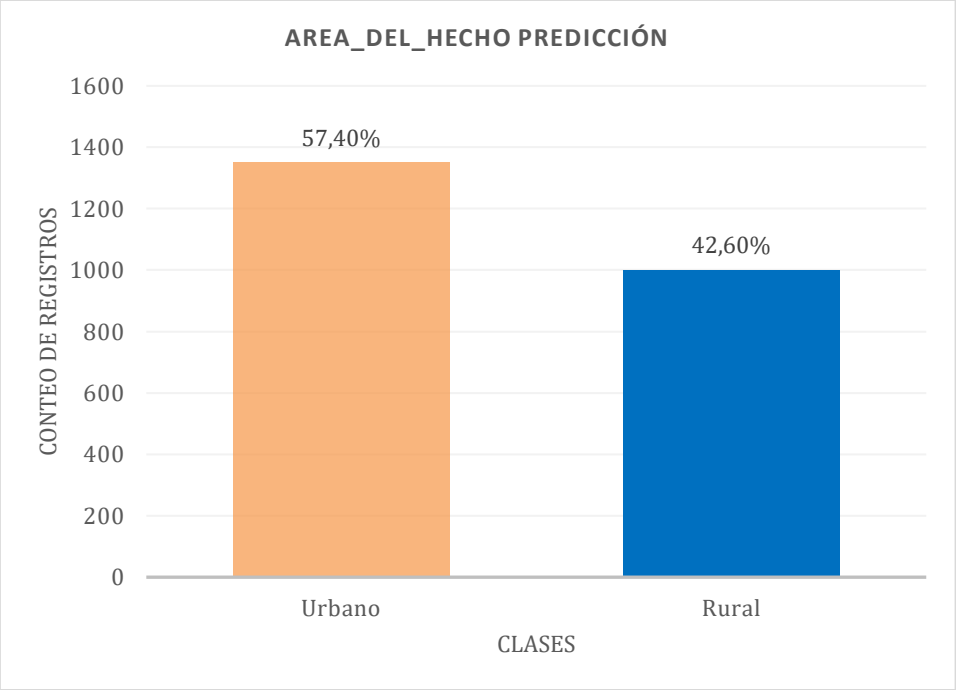


Figura 66. Distribución de asesinatos para la variable "AREA_DEL_HECHO" con el modelo Árbol de Decisión configurado con Optuna.

- **Variable "ANTECEDENTES"**

La estructura del árbol para la variable "ANTECEDENTES", tuvo en el nodo raíz "AREA_DEL_HECHO <= 1.5", las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como "entropy" (desorden del nodo), "samples" (número de muestras) y "value" (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[14].

En la **Figura 67,** se visualiza la matriz de confusión para la variable "ANTECEDENTES", en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1177 verdaderos negativos y 23 falsos positivos, mientras que la clase 1 cuenta con 961 verdaderos positivos y 193 falsos negativos.

[14] https://bit.ly/4h6QO7U

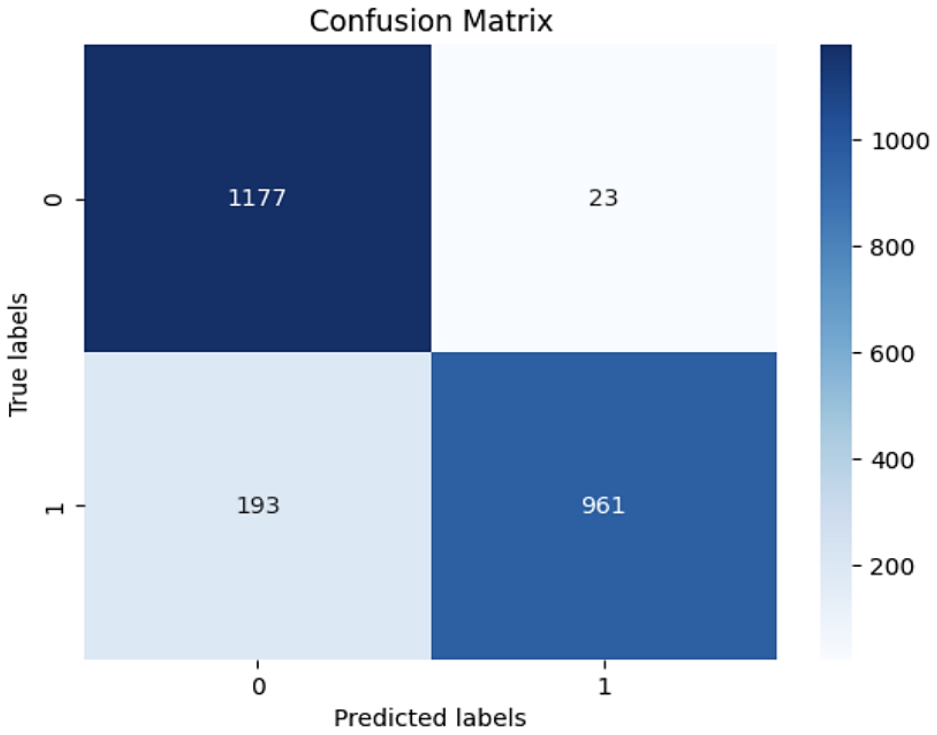


Figura 67. Matriz de confusión para la variable “ANTECEDENTES” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 21,** muestra el desempeño del Árbol de Decisión para la variable "ANTECEDENTES", donde se clasifican dos clases: Sí (Clase 0) y No (Clase 1). La precisión para la clase Sí es del 85,91%, con un recall del 98,08% mientras que la clase No tuvo una precisión del 97,66%, y un recall del 83,27%, el modelo Árbol de Decisión alcanzó una exactitud global (accuracy) del 90,82%, evidenciando un buen desempeño.

Tabla 21. Desempeño del modelo Árbol de Decisión para la variable “ANTECEDENTES” configurada con los hiperparámetros de la librería Optuna.

ANTECEDENTES		
Clase	**Precisión**	**Recall**
0: Si	85,91 %	98,08 %
1: No	97,66 %	83,27 %
Accuracy	90,82 %	90,82 %

En la **Figura 68,** se visualiza el gráfico de barras de la distribución de asesinatos según los antecedentes de la víctima, los cuales están divididos en dos categorías o clases que son: "Sí" y "No", la primera clase “Si” tuvo un 58,20% del total, la segunda clase “NO” alcanzó un 41,80% del total. Esto evidencia que los asesinatos se asocian mayormente a individuos que sí presentan antecedentes.

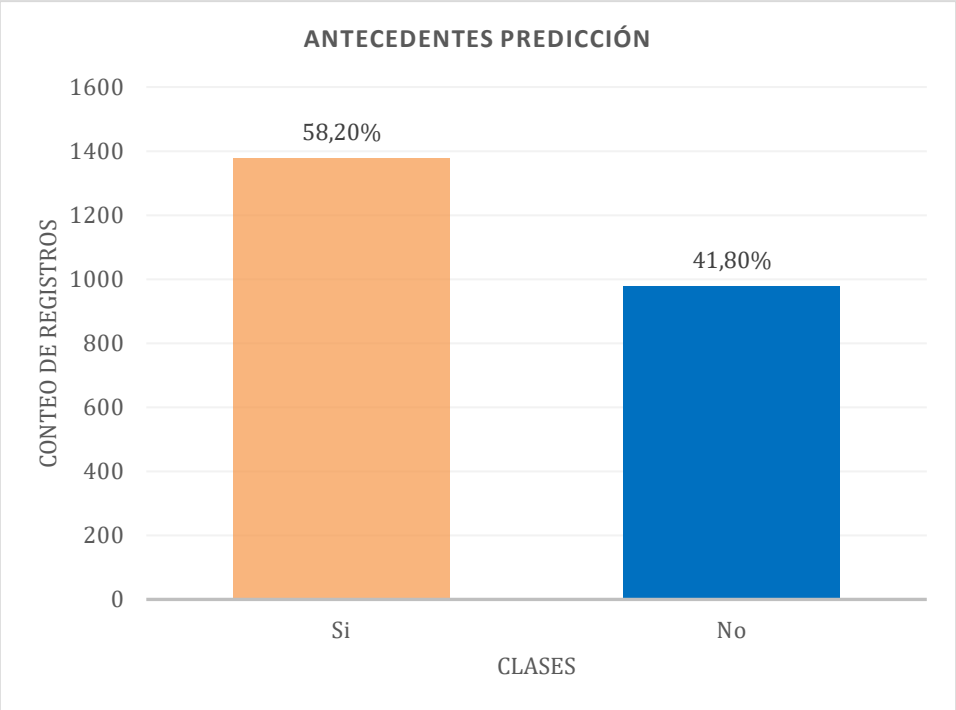


Figura 68. Distribución de asesinatos para la variable “ANTECEDENTES” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “LUGAR”**

La estructura del árbol para la variable “LUGAR”, tuvo en el nodo raíz “PRESUNTA_MOTIVACION <= 2.5”, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “entropy” (desorden del nodo), “samples” (número de muestras) y “value” (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[15].

En la **Figura 69,** se visualiza la a matriz de confusión para la variable “LUGAR”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1165 verdaderos negativos y 33 falsos positivos, mientras que la clase 1 cuenta con 893 verdaderos positivos y 263 falsos negativos.

[15] https://bit.ly/4hcgmkc

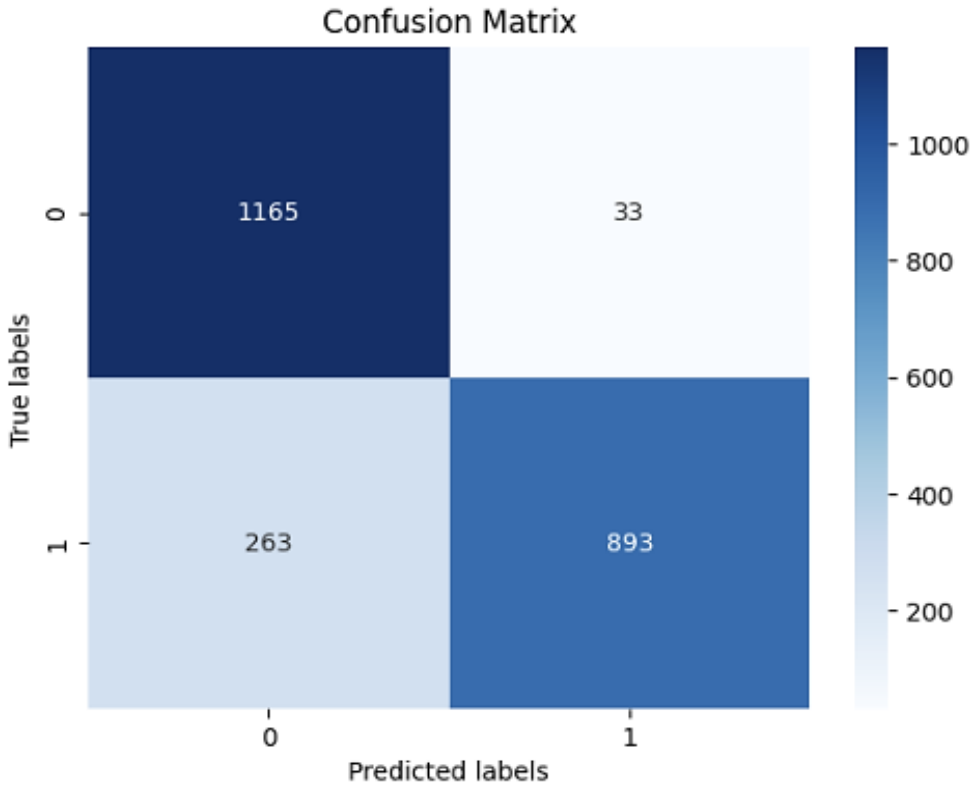


Figura 69. Matriz de confusión para la variable “LUGAR” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 22**, se muestra el desempeño del modelo para la variable "LUGAR", donde se clasifican dos clases: Vía Pública (Clase 0) y Lugares Privados (Clase 1). La precisión para la clase 0 es del 81,58% con un recall del 97,24%, mientras que la clase 1 presenta una precisión del 96,43% con un recall del 77,24%, el modelo Árbol de Decisión presentó una exactitud global (accuracy) del 87,42%, evidenciando un buen desempeño en las clasificaciones.

Tabla 22. Desempeño del modelo Árbol de Decisión para la variable “LUGAR” configurada con los hiperparámetros de la librería Optuna.

LUGAR		
Clase	**Precisión**	**Recall**
0: Vía Pública	81,58 %	97,24 %
1: Lugares Privados	96,43 %	77,24 %
Accuracy	87,42 %	87,42 %

En la **Figura 70,** se visualiza el gráfico de barras que muestra la distribución de asesinatos según los lugares donde ocurrieron los crímenes, divididos en dos categorías: "Vía Pública" y “Lugares Privados", la primera tuvo 60,70% del total, mientras que segunda un 39,30% respectivamente, evidenciando que la mayoría de los crímenes ocurren en la vía pública.

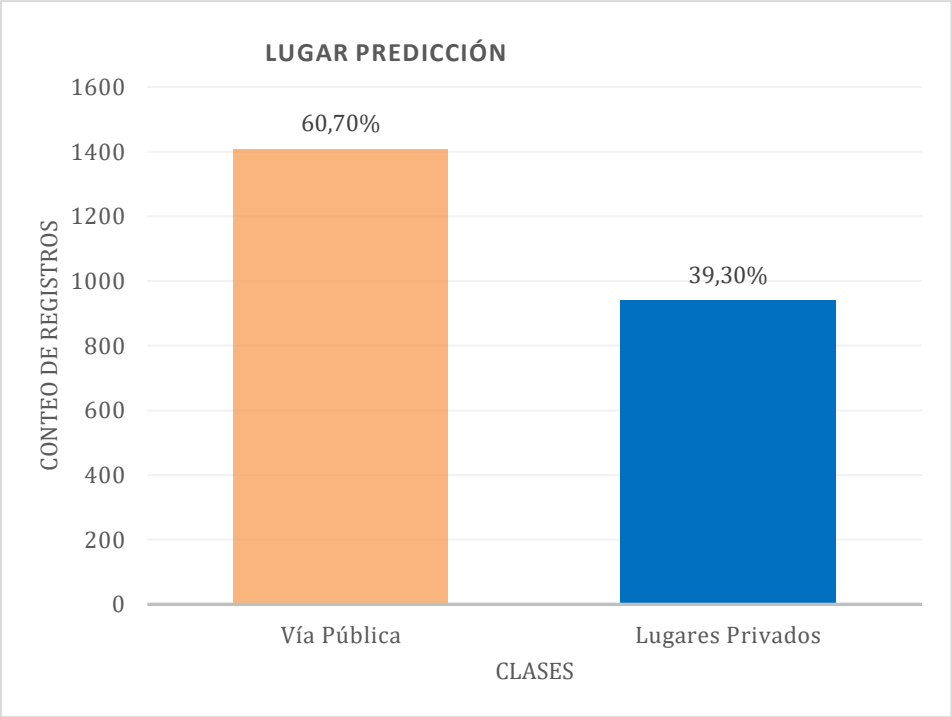


Figura 70. Distribución de asesinatos para la variable “LUGAR” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “SEXO”**

La estructura del árbol para la variable “SEXO” tuvo en el nodo raíz “AREA_DEL_HECHO <= 1.5”, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “gini” (pureza del nodo), “samples” (número de muestras) y “value” distribución de las clases, posteriormente se dividió en los nodos finales que exponen las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[16].

En la **Figura 71**, se visualiza la matriz de confusión para la variable “SEXO”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1298 verdaderos negativos y 0 falsos positivos, mientras que la clase 1 cuenta con 1014 verdaderos positivos y 42 falsos negativos.

[16] https://bit.ly/4gOoXtD

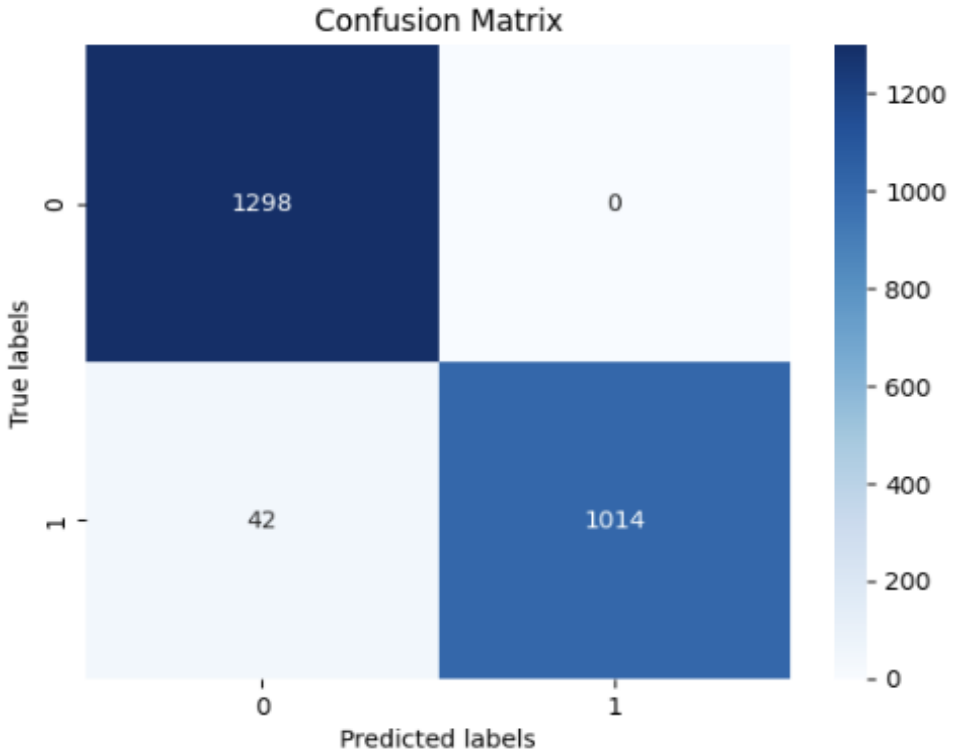


Figura 71. Matriz de confusión para la variable “SEXO” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 23**, muestra el desempeño del modelo para la variable "SEXO", donde se clasifican dos clases: Masculino (Clase 0) y Femenino (Clase 1). La precisión para la clase 0 es del 96,86%, con un recall del 100,00%, mientras que la clase 1 tuvo una precisión del 100,00%, con un recall del 96,02%, la exactitud global (accuracy) del modelo Árbol de Decisión fue del 98,21%, evidenciando un excelente desempeño al momento de realizar clasificaciones.

Tabla 23. Desempeño del modelo Árbol de Decisión para la variable “SEXO” configurada con los hiperparámetros de la librería Optuna.

SEXO		
Clase	**Precisión**	**Recall**
0: Masculino	96,86 %	100,00 %
1: Femenino	100,00 %	96,02 %
Accuracy	98,21 %	98,21 %

En la **Figura 72**, se visualiza un gráfico de barras que muestra la distribución de asesinatos según el sexo de las víctimas, divididos en dos categorías: "Masculino" y “Femenino", la primera tuvo 56,90% del total, mientras que la segunda un 43,10% respectivamente, evidenciando que la mayoría de las víctimas de asesinatos son de sexo masculino.

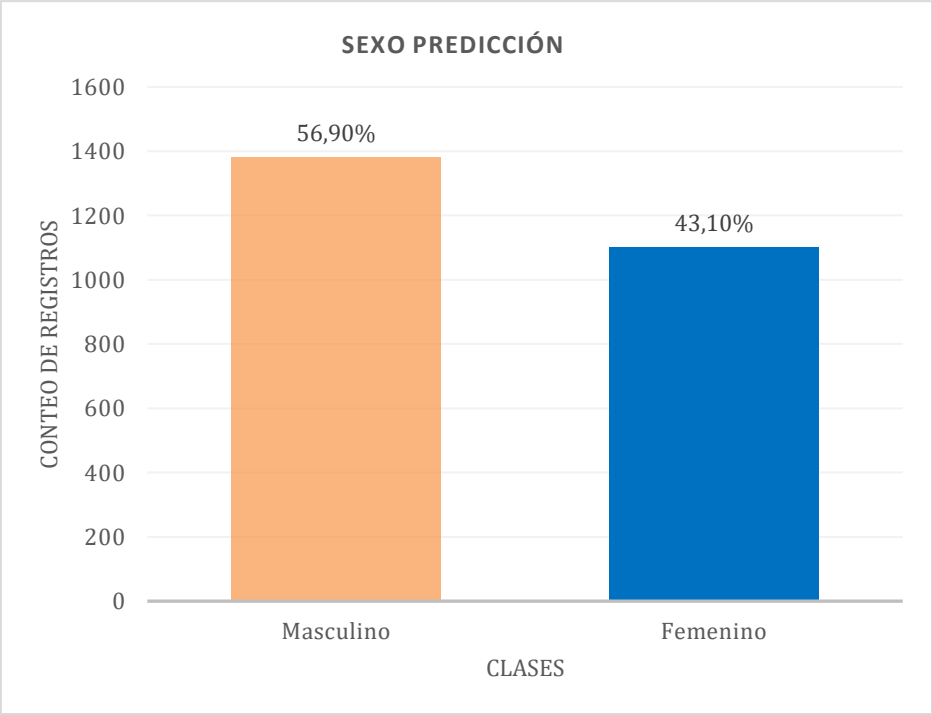


Figura 72. Distribución de asesinatos para la variable “SEXO” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “PRESUNTA_MOTIVACION”**

La estructura del árbol para la variable “PRESUNTA_MOTIVACION” tuvo en el nodo raíz ARMA <= 1.5, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “gini” (pureza del nodo), “samples” (número de muestras) y “value” (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[17].

En la **Figura 73**, se visualiza la matriz de confusión para la variable “PRESUNTA_MOTIVACION”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 225 instancias correctamente clasificadas, la clase 1 tuvo 1150, la clase 2 tuvo 171, la clase 3 tuvo 187, la clase 4 tuvo 189 y la clase 5 tuvo 221 clasificaciones correctas.

[17] https://bit.ly/3BRAfOA

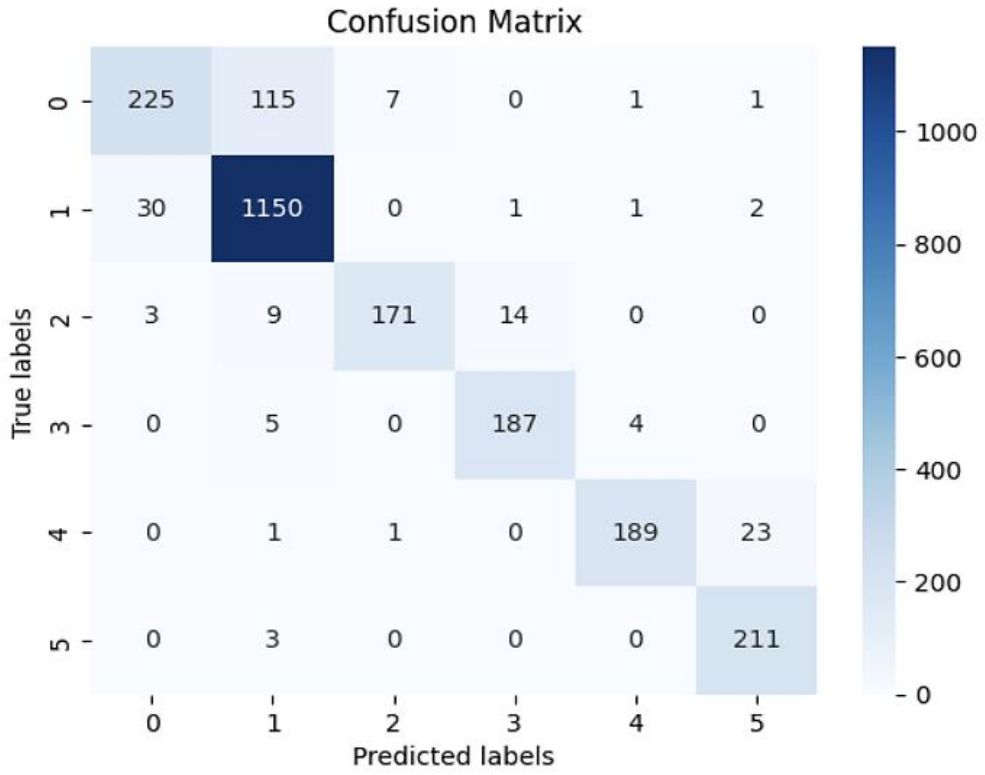


Figura 73. Matriz de confusión para la variable “PRESUNTA_MOTIVACION” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 24,** muestra el desempeño del modelo para la variable "PRESUNTA_MOTIVACION", en donde la clase "Delincuencia común" tuvo una precisión del 89,63% y un recall de 97,12%, la clase "Violencia intrafamiliar y sexual" muestra una precisión del 95,53 % y un recall del 86,80 %, la clase "Terrorismo" tuvo un recall del 98,59% y una precisión del 89,02 %, la clase de "Violencia comunitaria" presentó un recall del 64,46% y una precisión del 87,20%, la clase “Transnacional” tuvo una precisión del 92,57% y un recall del 95,40%, la clase “Psicopatologías” presentó una precisión del 96,92% y un recall del 88,31%. El modelo Árbol de Decisión tuvo una exactitud global (accuracy) del 90,61%, indicando un nivel aceptable en su clasificación.

Tabla 24. Desempeño del modelo Árbol de Decisión para la variable “PRESUNTA_MOTIVACION” configurada con los hiperparámetros de la librería Optuna.

PRESUNTA_MOTIVACION		
Clase	**Precisión**	**Recall**
0: Violencia comunitaria	87,20 %	64,46 %
1: Delincuencia común	89,63 %	97,12 %
2: Violencia intrafamiliar y sexual	95,53 %	86,80 %
3: Transnacional	92,57 %	95,40 %
4: Psicopatologías	96,92 %	88,31 %
5: Terrorismo	89,02 %	98,59 %
Accuracy	90,61 %	90,61 %

En la **Figura 74**, se visualiza un gráfico de barras que muestra la distribución de asesinatos según la presunta motivación, divididos en seis categorías: la primera de las

categorías es "Delincuencia común" que tuvo un 54,30% del total, mientras que la categoría "Transnacional" presentó un 8,60%, la categoría "Terrorismo" presentó un 10,10%, la categoría "Psicopatologías" tuvo un 8,30%, la categoría "Violencia intrafamiliar y sexual" tuvo un 7,60 % y la categoría "Violencia comunitaria" alcanzó un 11,30%. Esta representación gráfica permite observar que la mayoría de los asesinatos están motivados por la delincuencia común.

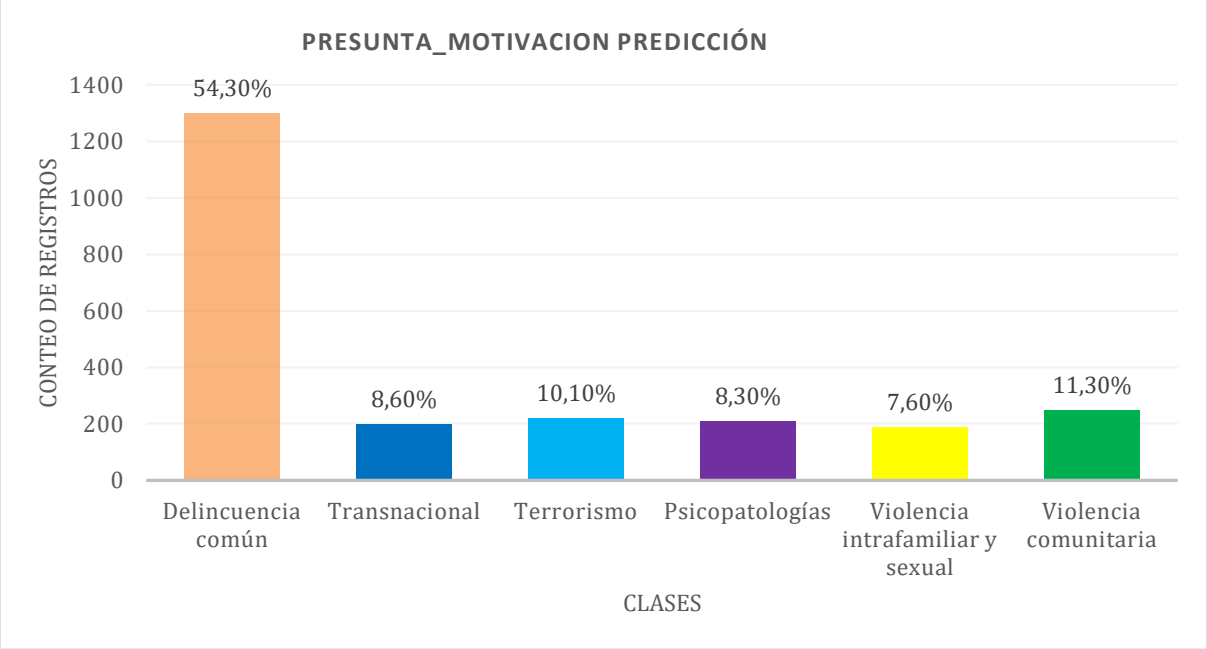


Figura 74. Distribución de asesinatos para la variable “PRESUNTA_MOTIVACION” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “ARMA”**

Para la estructura del árbol para la variable “ARMA”, en este caso tuvo en el nodo raíz “AREA_DEL_HECHO <= 1.5”, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “gini” (pureza del nodo), “samples” (número de muestras) y “value” (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[18].

En la **Figura 75**, se visualiza la matriz de confusión para la variable “ARMA”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 1216 instancias correctamente clasificadas, la clase 1 tuvo 285, la clase 2 tuvo 207, la clase 3 tuvo 178, la clase 4 tuvo 234 clasificaciones correctas.

[18] https://bit.ly/3DS159J

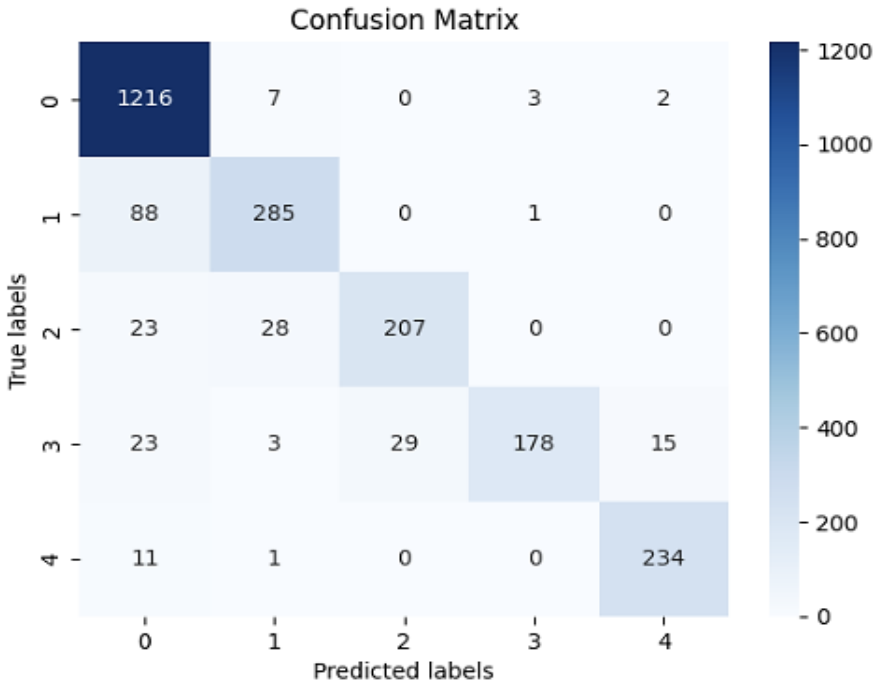


Figura 75. Matriz de confusión para la variable “ARMA” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 25**, muestra el desempeño del modelo para la variable “ARMA” en donde la clase "Arma de fuego" tuvo una precisión del 89,34% y un recall del 99,02%, la clase "Arma blanca" obtuvo una precisión del 87,96% y un recall del 76,20%, mientras que la clase "Sustancias y otros" alcanzó una precisión del 87,71% y un recall del 80,23%, la clase "Arma contundente" tuvo una precisión de 97,80%, pero un recall más bajo de 71,77%, finalmente la clase "Arma constrictora" alcanzó una precisión del 93,22% y un recall del 95,12%, el modelo obtuvo una exactitud global (accuracy) del 90,05%, evidenciando un nivel aceptable en su clasificación.

Tabla 25. Desempeño del modelo Árbol de Decisión para la variable “ARMA” configurada con los hiperparámetros de la librería Optuna.

ARMA		
Clase	**Precisión**	**Recall**
0: Arma de fuego	89,34 %	99,02 %
1: Arma blanca	87,96 %	76,20 %
2: Sustancias y otros	87,71 %	80,23 %
3: Arma contundente	97,80 %	71,77 %
4: Arma constrictora	93,22 %	95,12 %
Accuracy	90,05 %	90,05 %

En la **Figura 76,** se visualiza el gráfico de barras para las cinco categorías que son: la primera es la categoría “Arma de fuego” que alcanzó un 57,80%, mientras que la categoría “Sustancias y otros” presentó un 10,00%, la categoría “Arma constrictora” tuvo un 10,70%, la categoría “Arma contundente” llegó a un 7,70% y la categoría “Arma blanca” presentó un 13,80%. Por lo tanto, la mayoría de los asesinatos se realizaron con armas de fuego.

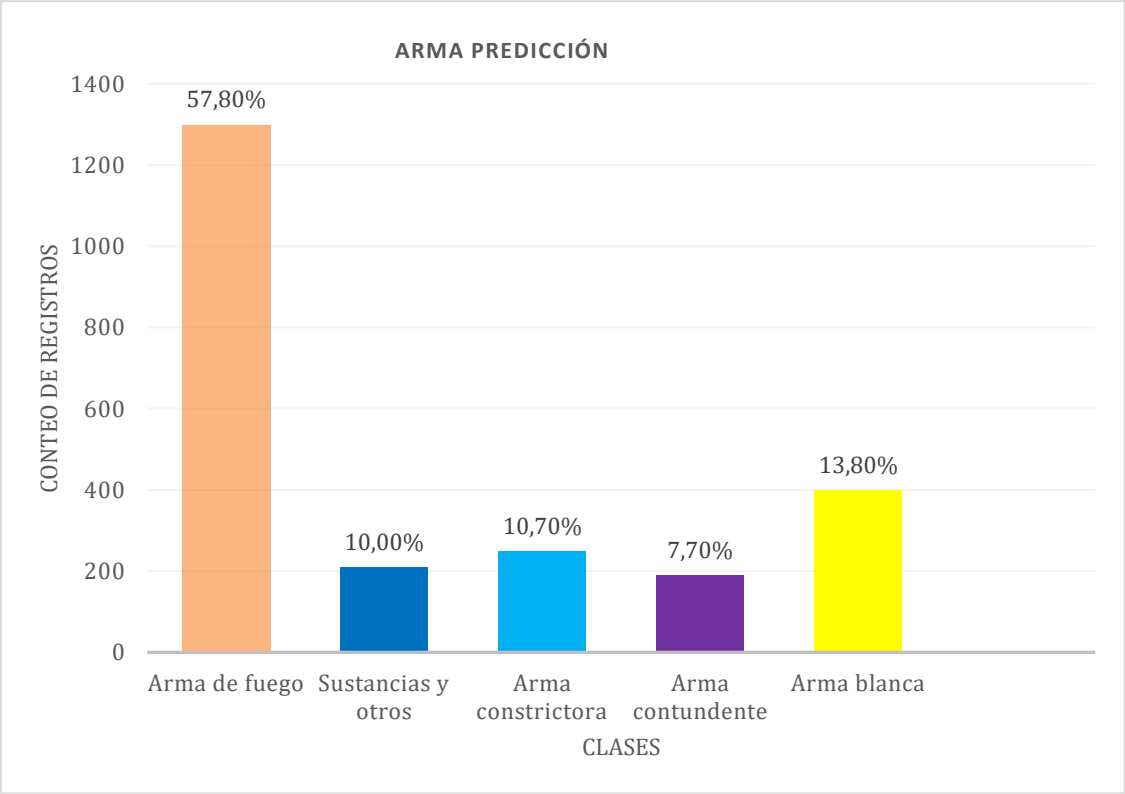


Figura 76. Distribución de asesinatos para la variable “ARMA” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “DIA”**

La estructura del árbol para la variable “DIA”, en este caso tuvo en el nodo raíz “PRESUNTA_MOTIVACION <= 4.5”, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “entropy” (desorden del nodo), “samples” (número de muestras) y “value” (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[19].

En la **Figura 77**, se visualiza la matriz de confusión para la variable “DIA”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 662 instancias correctamente clasificadas, la clase 1 tuvo 261, la clase 2 tuvo 202, la clase 3 tuvo 290 clasificaciones correctas.

[19] https://bit.ly/3Ce4UFF

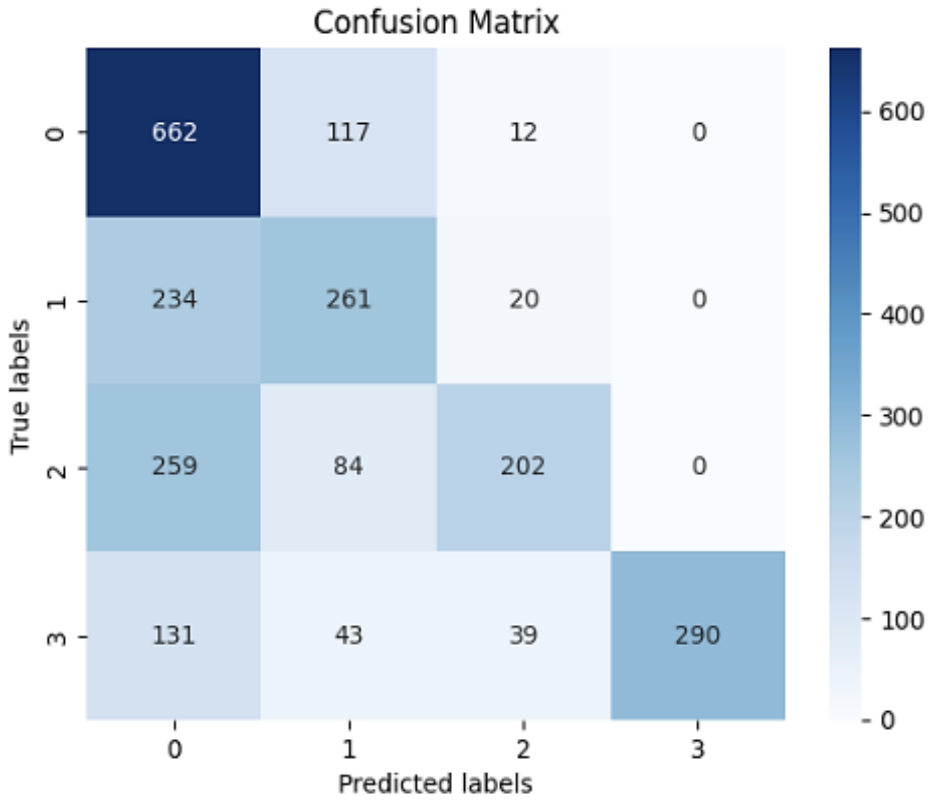


Figura 77. Matriz de confusión para la variable “DIA” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 26**, se muestra el desempeño del modelo para la variable “DIA”, en donde la clase 0 tuvo una precisión del 51,47% y un recall del 83,69%, la clase 1 obtuvo una precisión del 51,68% y un recall del 50,67%, mientras que la clase 2 alcanzó una precisión del 73,99% y un recall del 37,06%, la clase 3 tuvo una precisión de 100,00%, pero un recall más bajo de 57,65%, el modelo obtuvo una exactitud global (accuracy) del 60,11%, evidenciando un nivel regular en su clasificación.

Tabla 26. Desempeño del modelo Árbol de Decisión para la variable “DIA” configurada con los hiperparámetros de la librería Optuna.

DIA		
Clase	**Precisión**	**Recall**
0: Sábado, Domingo	51,47 %	83,69 %
1: Martes, Viernes	51,68 %	50,67 %
2: Miércoles	73,99 %	37,06 %
3: Lunes, Jueves	100,00 %	57,65 %
Accuracy	60,11 %	60,11 %

En la **Figura 78**, se visualiza un gráfico de barras que muestra la distribución de asesinatos según el día en dónde más se cometen, divididos en cuatro categorías: la primera es la categoría “Sábado, Domingo” que alcanzó un 54,60%, mientras que la categoría “Martes, Viernes” presentó un 21,50%, la categoría “Miércoles” tuvo un 11,60%, la categoría “Lunes,

Jueves” llegó a un 12,30%. Esto evidenció que la mayoría de los asesinatos se cometen los días sábado y domingo.

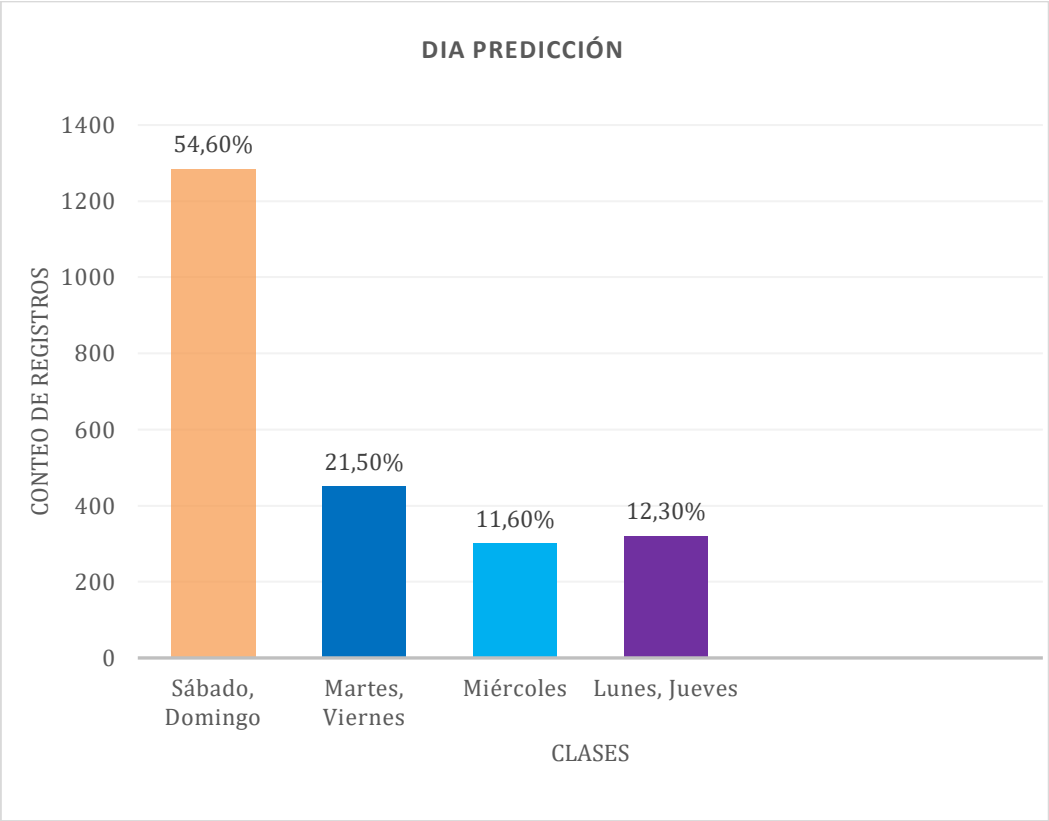


Figura 78. Distribución de asesinatos para la variable “DIA” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “EDAD”**

La estructura del árbol para la variable “EDAD”, en este caso tuvo en el nodo raíz “AREA_DEL_HECHO <= 1.5”, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “entropy” (desorden del nodo), “samples” (número de muestras) y “value” (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[20].

En la **Figura 79**, se visualiza la matriz de confusión para la variable “EDAD”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 436 instancias correctamente clasificadas, la clase 1 obtuvo 1181, la clase 2 alcanzó 230 y la clase 3 tuvo 249 clasificaciones correctas.

[20] https://bit.ly/40gfeVR

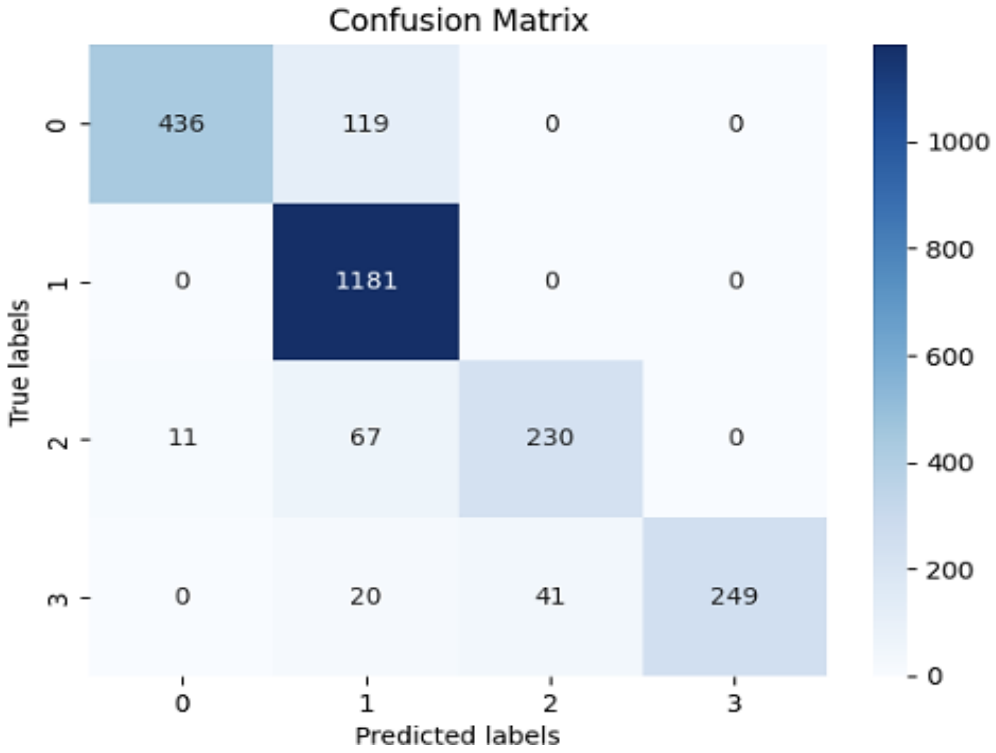


Figura 79. Matriz de confusión para la variable “EDAD” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 27**, se muestra el desempeño del modelo para la variable “EDAD”, en donde la clase 0 tuvo una precisión del 97,53% y un recall del 78,55%, la clase 1 obtuvo una precisión del 85,14% y un recall del 100,00%, mientras que la clase 2 alcanzó una precisión del 84,87% y un recall del 74,67%, la clase 3 tuvo una precisión de 100,00%, pero un recall más bajo de 80,32%, el modelo obtuvo una exactitud global (accuracy) del 89,03%, evidenciando un nivel bueno en su clasificación.

Tabla 27. Desempeño del modelo Árbol de Decisión para la variable “EDAD” configurada con los hiperparámetros de la librería Optuna.

EDAD		
Clase	**Precisión**	**Recall**
0: 1-19	97,53 %	78,55 %
1: 20-50	85,14 %	100,00 %
2: 51-65	84,87 %	74,67 %
3: 66-95	100,00 %	80,32 %
Accuracy	89,03 %	89,03 %

En la **Figura 80**, se visualiza un gráfico de barras que muestra la distribución de asesinatos basándose en la edad de la víctima, divididos en cuatro categorías: la primera es la categoría que va en un rango de “20 – 50” alcanzó un 58,90%, la categoría “1 -19” presentó un 19,00%, la categoría “66 – 95” tuvo un 10,60%, la categoría “51 – 65” llegó a un 11,50%. Esto evidenció que la mayoría de las víctimas de asesinatos tienen edades que van desde los 20 hasta los 50 años.

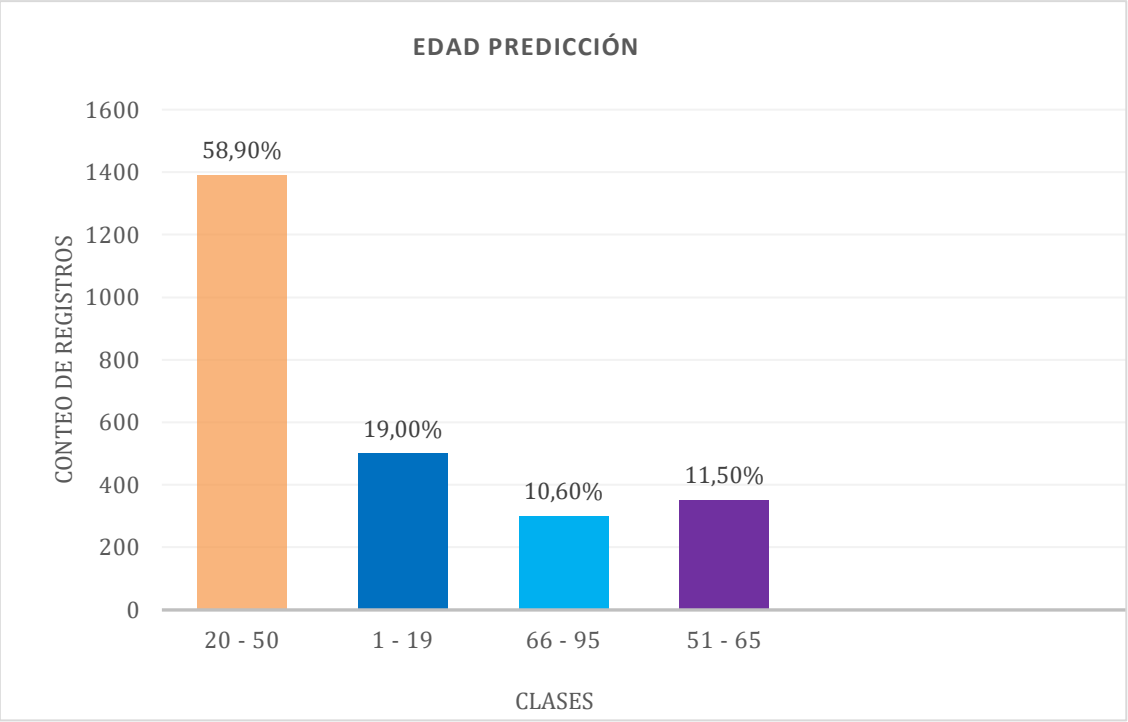


Figura 80. Distribución de asesinatos para la variable “EDAD” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “HORA_INFRACCION”**

La estructura del árbol para la variable “HORA_INFRACCION”, en este caso tuvo en el nodo raíz “PRESUNTA_MOTIVACION <= 3.5”, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “entropy” (desorden del nodo), “samples” (número de muestras) y “value” (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[21].

En la **Figura 81,** se visualiza la matriz de confusión para la variable “HORA_INFRACCION”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 361 instancias correctamente clasificadas, la clase 1 tuvo 378, la clase 2 tuvo 257, la clase 3 tuvo 575 clasificaciones correctas.

[21] https://bit.ly/3C4yQEe

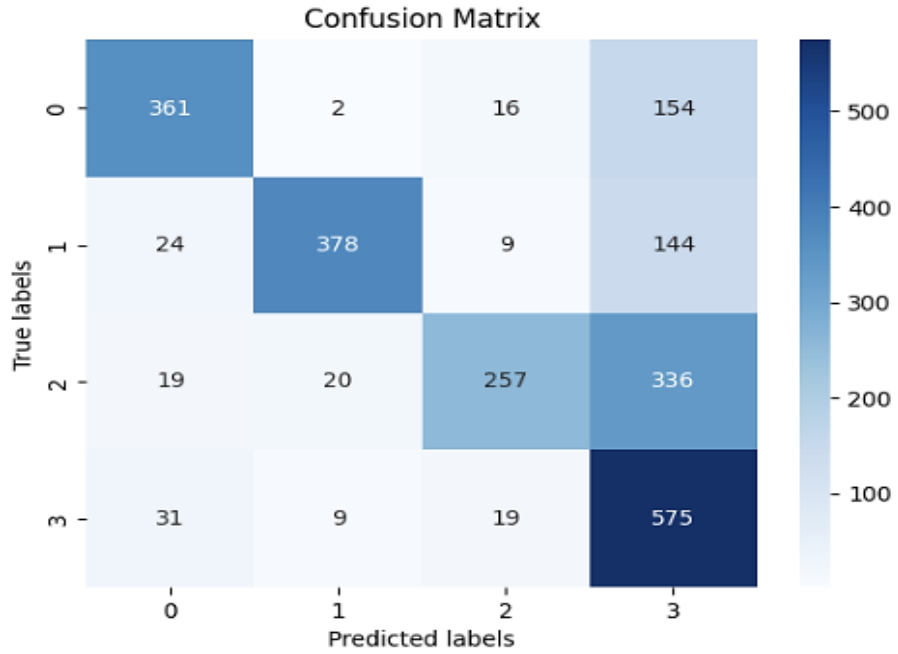


Figura 81. Matriz de confusión para la variable “HORA_INFRACCION” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 28**, se muestra el desempeño del modelo para la variable “HORA_INFRACCION”, en donde la clase 0 tuvo una precisión del 82,98% y un recall del 67,72%, la clase 1 obtuvo una precisión del 92,42% y un recall del 68,10%, mientras que la clase 2 alcanzó una precisión del 85,38% y un recall del 40,66%, la clase 3 tuvo una precisión de 47,56%, pero un recall más bajo de 90,69%, el modelo obtuvo una exactitud global (accuracy) del 66,73%, evidenciando un nivel regular en su clasificación.

Tabla 28. Desempeño del modelo Árbol de Decisión para la variable “HORA_INFRACCION” configurada con los hiperparámetros de la librería Optuna.

HORA_INFRACCION		
Clase	**Precisión**	**Recall**
0: H01 – H06	82,98 %	67,72 %
1: H07 – H12	92,42 %	68,10 %
2: H13 – H18	85,38 %	40,66 %
3: H19 – H00	47,56 %	90,69 %
Accuracy	66,73 %	66,73 %

En la **Figura 82**, se visualiza un gráfico de barras que muestra la distribución de asesinatos basándose en la hora que se cometen los crímenes, divididos en cuatro categorías: la primera es la categoría que va en un rango horario de “H19 - H00” esta alcanzó un 51,40%, mientras que la categoría “H07 - H12” presentó un 17,40%, mientras que la categoría “H01 – H06” tuvo un 18,50%, la categoría “H13 – H18” llegó a un 12,80%. Esto evidenció que la mayoría de los asesinatos se realizan en horas que van desde las 19:00 pm hasta las 00:00 am (medianoche).

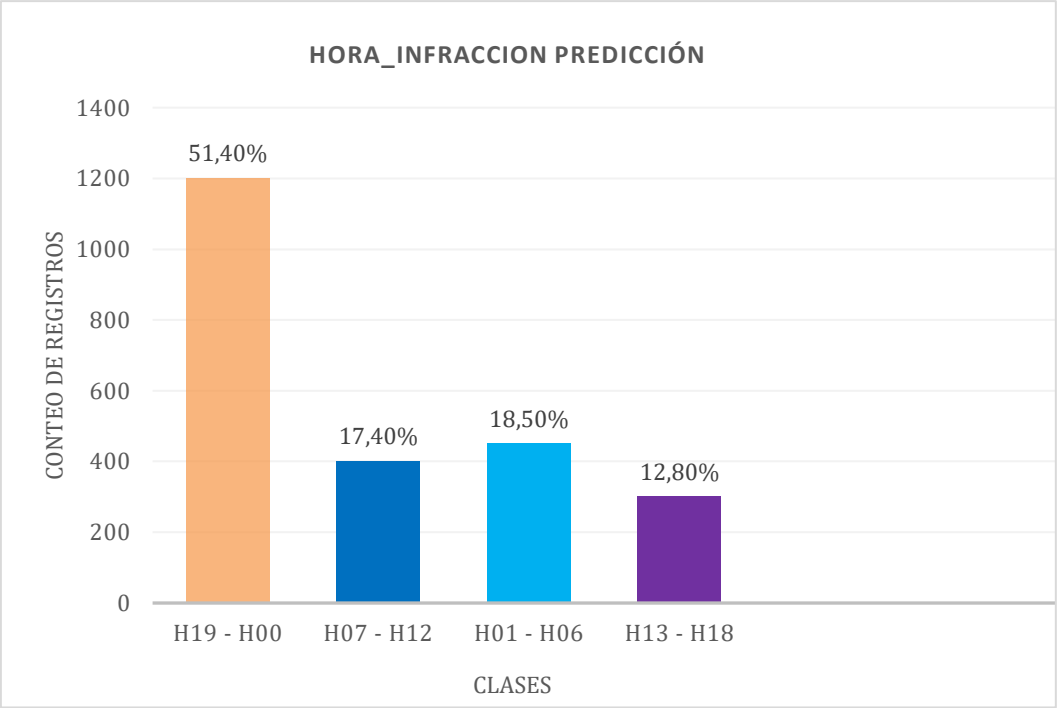


Figura 82. Distribución de asesinatos para la variable “HORA_INFRACCION” con el modelo Árbol de Decisión configurado con Optuna.

- **Variable “DISTRITO”**

La estructura del árbol para la variable “HORA_INFRACCION”, en este caso tuvo en el nodo raíz “PRESUNTA_MOTIVACION <= 4.5”, las ramas se dividieron en True y False, es decir separa las que si cumplen o no con la condición principal, luego se dividió en nodos intermedios en donde va evaluando otras variables, indicando métricas como “entropy” (desorden del nodo), “samples” (número de muestras) y “value” (distribución de las clases), posteriormente se dividió en los nodos finales que revelan las predicciones realizadas por el modelo, para ver la estructura completa visitar el enlace[22].

En la **Figura 83**, se visualiza la matriz de confusión para la variable “DISTRITO”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 obtuvo 750 instancias correctamente clasificadas, la clase 1 tuvo 439, la clase 2 obtuvo 432 clasificaciones correctas.

[22] https://bit.ly/4acCJDO

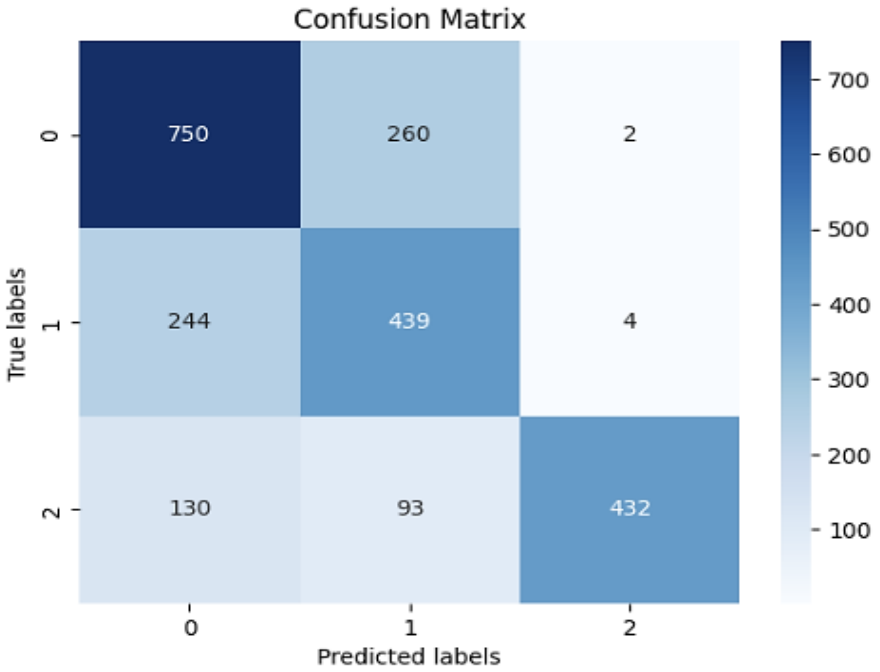


Figura 83. Matriz de confusión para la variable “DISTRITO” con el modelo Árbol de Decisión configurado con Optuna.

En la **Tabla 29**, se muestra el desempeño del modelo para la variable “DISTRITO”, en donde la clase 0 tuvo una precisión del 66,72% y un recall del 74,11%, la clase 1 obtuvo una precisión del 55,42% y un recall del 63,90%, mientras que la clase 2 alcanzó una precisión del 98,63% y un recall del 65,95%, el modelo obtuvo una exactitud global (accuracy) del 68,86%, evidenciando un nivel regular en su clasificación.

Tabla 29. Desempeño del modelo Árbol de Decisión para la variable “DISTRITO” configurada con los hiperparámetros de la librería Optuna.

DISTRITO		
Clase	**Precisión**	**Recall**
0: Nueva Prosperina, Distrito Sur, Pascuales	66,72 %	74,11 %
1: Portete, 9 de Octubre, Durán, Estero	55,42 %	63,90 %
2: Progreso, Florida, Modelo, Ceibos, Samborondón	98,63 %	65,95 %
Accuracy	68,86 %	68,86 %

En la **Figura 84,** se visualiza un gráfico de barras que muestra la distribución de asesinatos basándose en el Distrito, divididos en tres categorías: la primera es la categoría que consta de "Nueva Prosperina, Distrito Sur, Pascuales" esta alcanzó un 47,70% del total, mientras que la categoría "Portete, 9 de Octubre, Durán, Estero" presentó un 33,60%, la categoría "Progreso, Florida, Modelo, Ceibos, Samborondón" tuvo un 18,60%. Esto evidenció que la mayoría de los asesinatos se realizan en los Distritos de Nueva Prosperina, Distrito Sur y Pascuales, de la Zona 8 del Ecuador.

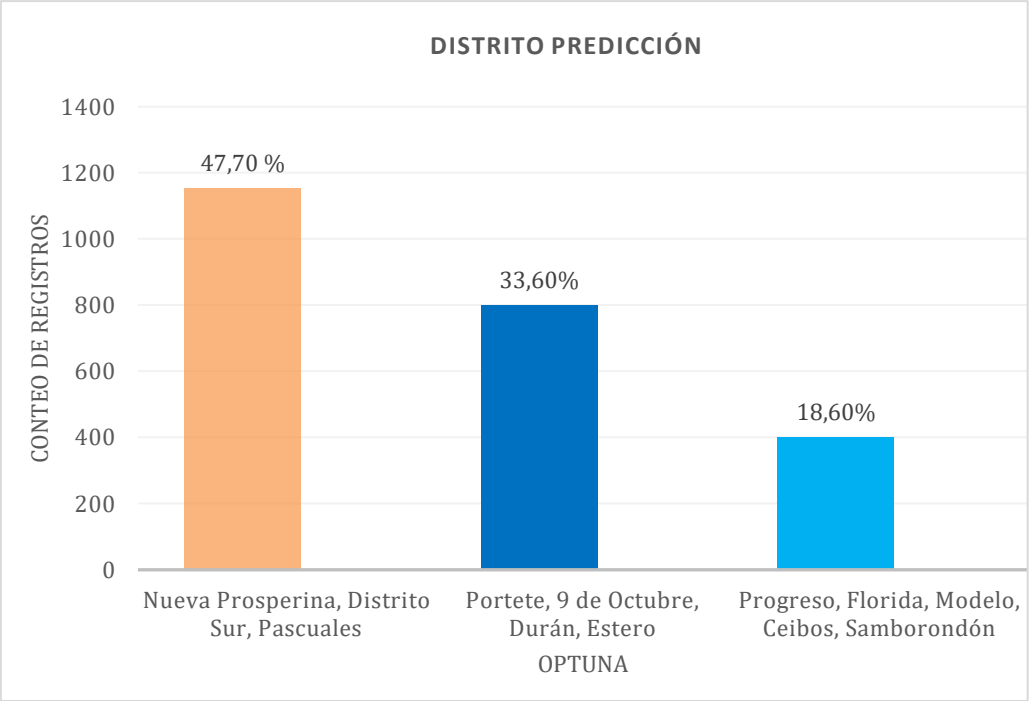


Figura 84. Distribución de asesinatos para la variable “DISTRITO” con el modelo Árbol de Decisión configurado con Optuna.

6.2.1.5. Interpretación de resultados de cada variable obtenidos con el modelo Support Vector Machine configurados mediante la librería Optuna

El modelo Support Vector Machine que estuvo configurado con los hiperparámetros obtenidos mediante la librería Optuna, fue el que presentó mayores porcentajes en las métricas de precisión, accuracy y recall, en comparación con los obtenidos con la librería Bayesian-Optimization, por tal motivo se realizó una interpretación de los resultados conseguidos para cada una de las 10 variables seleccionadas para este estudio. A continuación, se describe cada uno de los resultados de las clasificaciones para cada atributo.

- **Variable “AREA_DEL_HECHO”**

En la **Figura 85**, se muestra la matriz de confusión para la variable “AREA_DEL_HECHO”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1336 verdaderos negativos y 4 falsos positivos, mientras que la clase 1 cuenta con 987 verdaderos positivos y 27 falsos negativos.

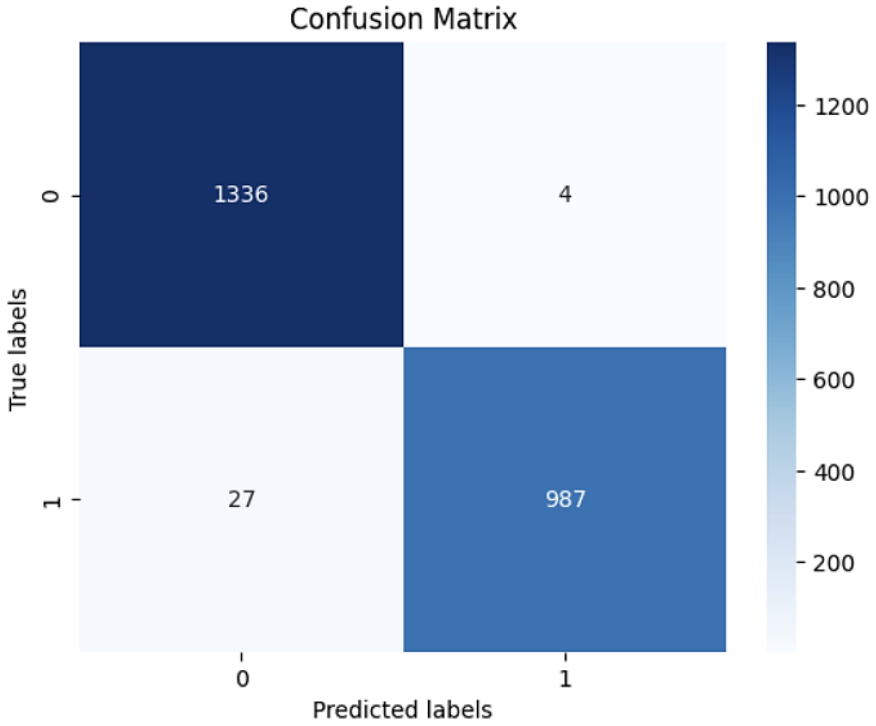


Figura 85. Matriz de confusión para la variable “AREA_DEL_HECHO” con el modelo SVM configurado con Optuna.

En la **Tabla 30**, se muestra el desempeño del modelo SVM para la variable “AREA_DEL_HECHO”, donde se clasifican dos clases: Urbano que fue la clase 0 y Rural la clase 1. La precisión para la clase Urbano es del 98,01%, teniendo un recall del 99,70%, mientras que la clase Rural tuvo una precisión del 99,59% y un recall del 97,33%, el modelo presentó una accuracy global de 98,68%. Para esta variable se evidenció un excelente desempeño del modelo en la clasificación de las clases.

Tabla 30. Desempeño del modelo SVM para la variable “AREA_DEL_HECHO” configurada con los hiperparámetros de la librería Optuna.

AREA_DEL_HECHO		
Clase	**Precisión**	**Recall**
0: Urbano	98,01 %	99,70 %
1: Rural	99,59 %	97,33 %
Accuracy	98,68 %	98,68 %

En la **Figura 86,** se visualiza el gráfico de barras para la variable “AREA_DEL_HECHO” que está representando la distribución de asesinatos en dos categorías: Urbano y Rural, la primera tiene un 57,90% mientras que la segunda tiene un 42,10% respectivamente, por lo tanto, se deduce que la mayoría de los asesinatos ocurren en áreas urbanas de la Zona 8 del Ecuador.

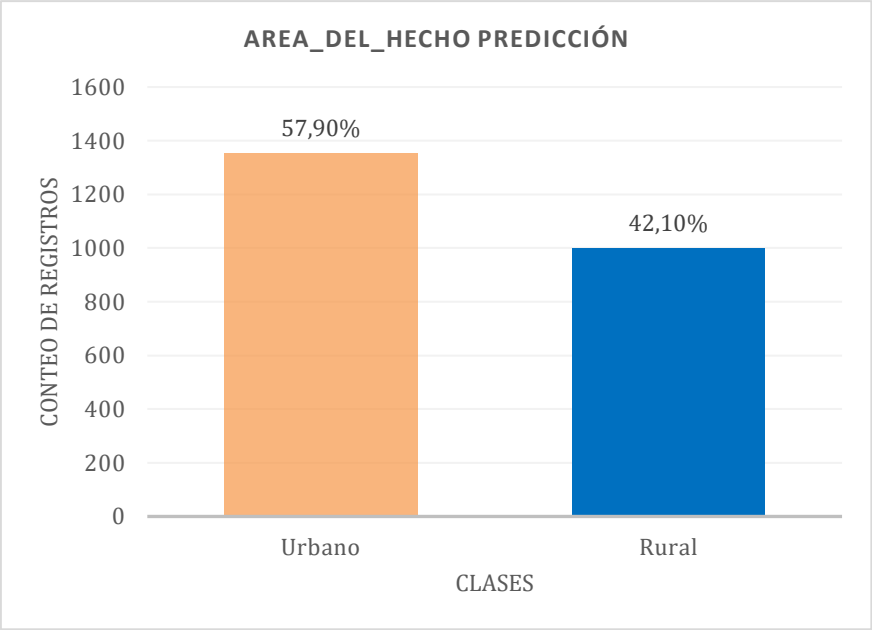


Figura 86. Distribución de asesinatos para la variable “AREA_DEL_HECHO” con el modelo SVM configurado con Optuna.

- **Variable “ANTECEDENTES”**

En la **Figura 87**, se visualiza la a matriz de confusión para la variable “ANTECEDENTES”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1181 verdaderos negativos y 19 falsos positivos, mientras que la clase 1 cuenta con 954 verdaderos positivos y 200 falsos negativos.

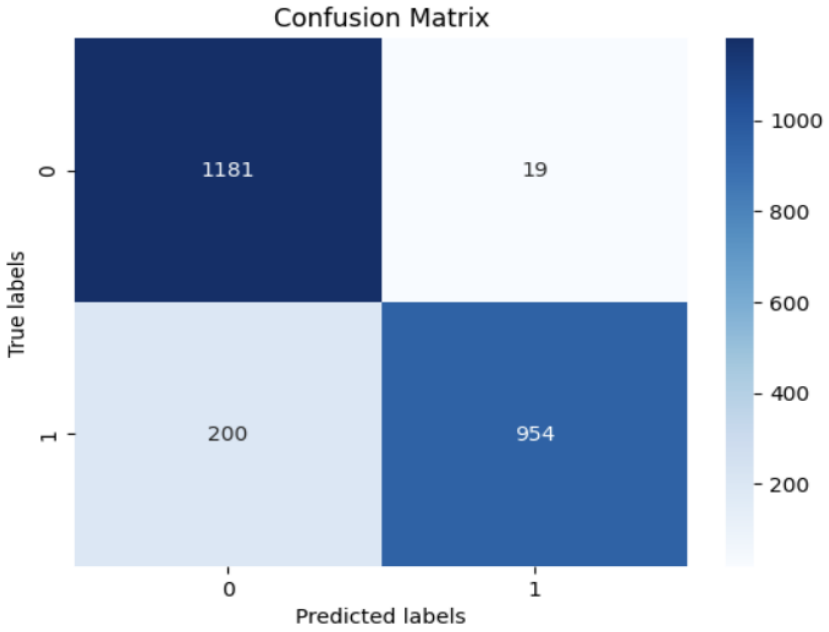


Figura 87. Matriz de confusión para la variable “ANTECEDENTES” con el modelo SVM configurado con Optuna.

En la **Tabla 31**, se muestra el desempeño del modelo SVM para la variable "ANTECEDENTES", donde se clasifican dos clases: Sí siendo la clase 0 y No la clase 1. La precisión para la clase Sí es del 85,51%, con un recall del 98,41% mientras que la clase No tuvo una precisión del 98,04%, y un recall del 82,66%, el modelo SVM alcanzó una exactitud global (accuracy) del 90,69%, evidenciando un buen desempeño.

Tabla 31. Desempeño del modelo SVM para la variable "ANTECEDENTES" configurada con los hiperparámetros de la librería Optuna.

ANTECEDENTES		
Clase	**Precisión**	**Recall**
0: Si	85,51 %	98,41 %
1: No	98,04 %	82,66 %
Accuracy	90,69 %	90,69 %

En la **Figura 88**, se visualiza el gráfico de barras de la distribución de asesinatos según los antecedentes de la víctima, divididos en dos categorías: "Sí" y "No", la primera tuvo un 58,70% del total, la segunda un 41,30% respectivamente. Esto evidencia que los asesinatos se asocian mayormente a individuos que sí presentan antecedentes.

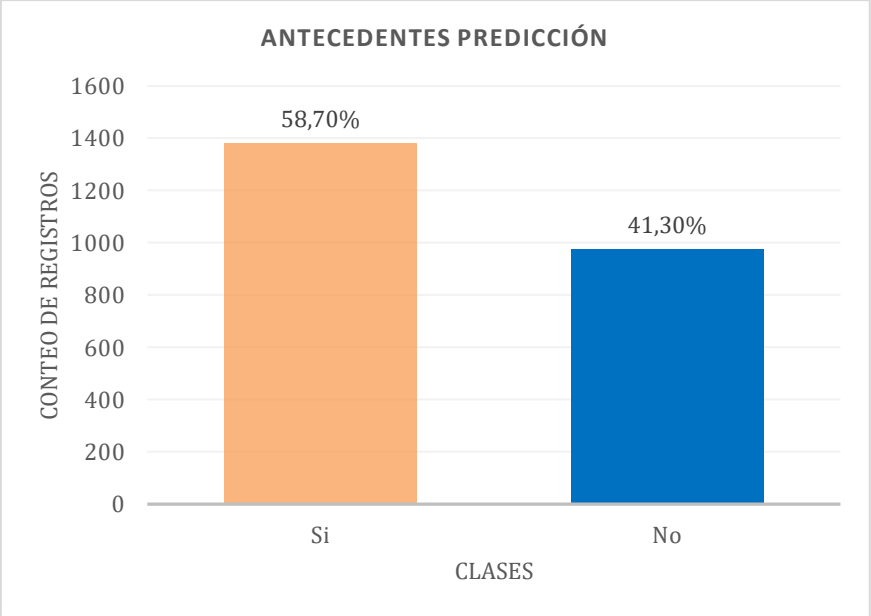


Figura 88. Distribución de asesinatos para la variable "ANTECEDENTES" con el modelo SVM configurado con Optuna.

- **Variable "LUGAR"**

En la **Figura 89**, se visualiza la a matriz de confusión para la variable "LUGAR", en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1156 verdaderos negativos y 42 falsos positivos, mientras que la clase 1 cuenta con 896 verdaderos positivos y 260 falsos negativos.

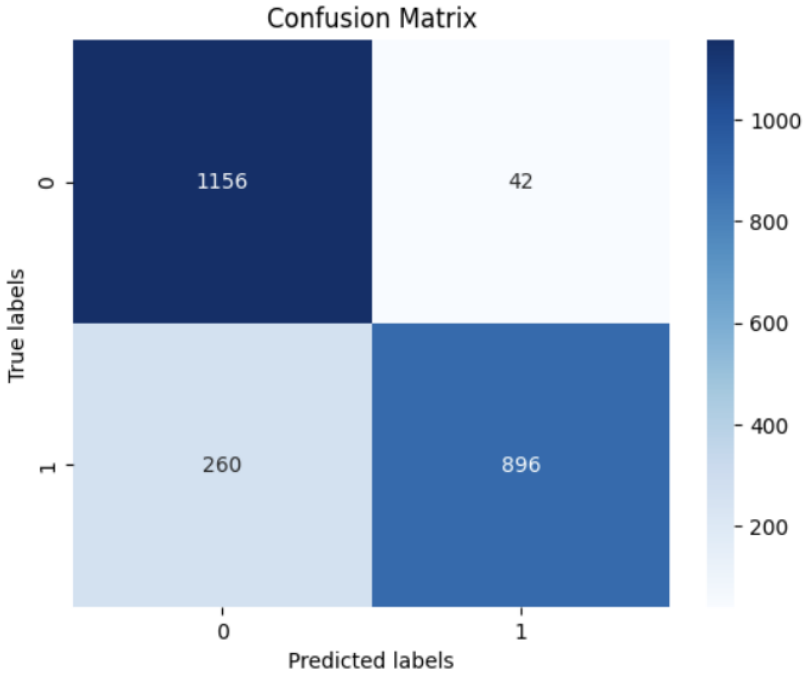


Figura 89. Matriz de confusión para la variable “LUGAR” con el modelo SVM configurado con Optuna.

En la **Tabla 32**, se muestra el desempeño del modelo SVM para la variable "LUGAR", donde se clasifican dos clases: Vía Pública siendo la clase 0 y Lugares Privados la clase 1. La precisión para la clase 0 es del 81,63% con un recall del 96,49%, mientras que la clase 1 presenta una precisión del 95,52% con un recall del 77,50%, el modelo SVM presentó una exactitud global del 87,17%, evidenciando un buen desempeño en las clasificaciones.

Tabla 32. Desempeño del modelo SVM para la variable “LUGAR” configurada con los hiperparámetros de la librería Optuna.

LUGAR		
Clase	**Precisión**	**Recall**
0: Vía Pública	81,63 %	96,49 %
1: Lugares Privados	95,52 %	77,50 %
Accuracy	87,17 %	87,17 %

En la **Figura 90**, se visualiza un gráfico de barras que muestra la distribución de asesinatos según los lugares donde ocurrieron los crímenes, divididos en dos categorías: “Vía Pública” y “Lugares Privados”, la primera clase tuvo 60,20% y la segunda un 39,80%, por lo tanto, los asesinatos ocurren con mayor frecuencia en la vía pública.

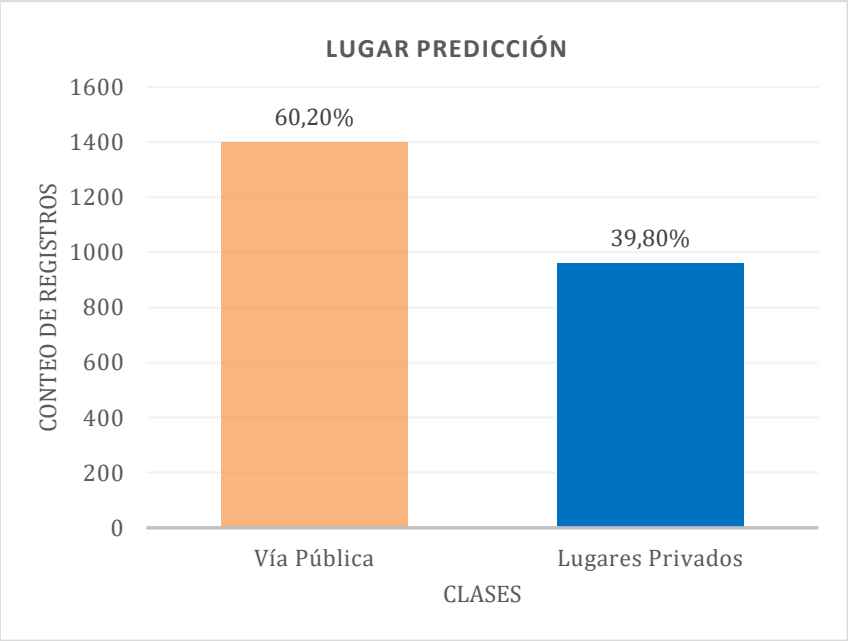


Figura 90. Distribución de asesinatos para la variable “LUGAR” con el modelo SVM configurado con Optuna.

- **Variable “SEXO”**

En la **Figura 91**, se visualiza la a matriz de confusión para la variable “SEXO”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que la diagonal secundaria son las predicciones incorrectas, la clase 0 tiene 1298 verdaderos negativos y 0 falsos positivos, mientras que la clase 1 cuenta con 1014 verdaderos positivos y 42 falsos negativos.

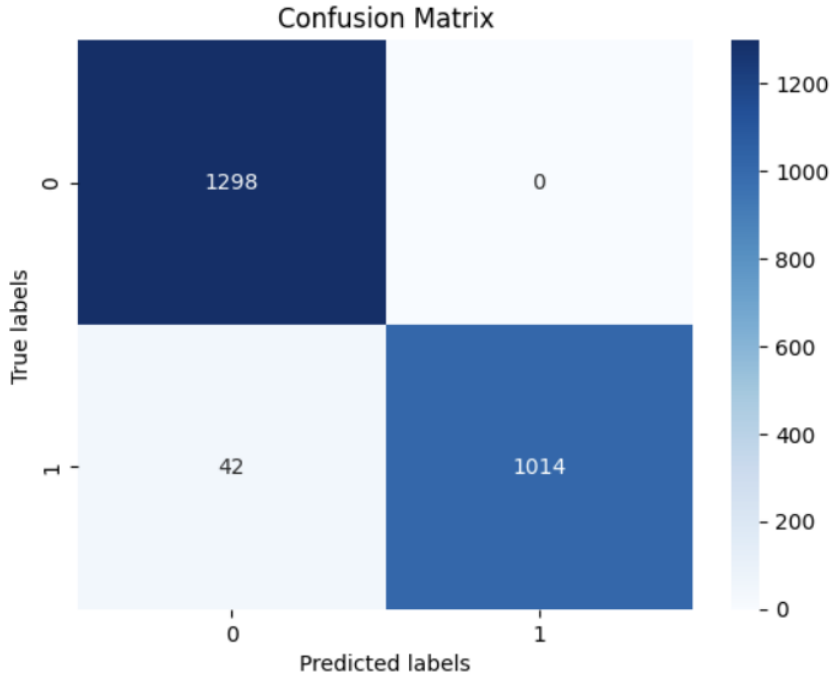


Figura 91. Matriz de confusión para la variable “SEXO” con el modelo SVM configurado con Optuna.

En la **Tabla 33**, se muestra el desempeño del modelo SVM para la variable "SEXO", donde se clasifican dos clases: Masculino siendo la clase 0 y Femenino la clase 1. La precisión para la clase 0 es del 96,86%, con un recall del 100,00%, mientras que la clase 1 tuvo una precisión del 100,00%, con un recall del 96,02%, la exactitud global del modelo SVM fue del 98,21%, evidenciando un excelente desempeño al momento de realizar clasificaciones.

Tabla 33. Desempeño del modelo SVM para la variable “SEXO” configurada con los hiperparámetros de la librería Optuna.

SEXO		
Clase	**Precisión**	**Recall**
0: Masculino	96,86 %	100,00 %
1: Femenino	100,00 %	96,02 %
Accuracy	98,21 %	98,21 %

En la **Figura 92**, se visualiza un gráfico de barras que muestra la distribución de asesinatos según el sexo de las víctimas, divididos en dos categorías: “Masculino” que llegó a tener un 56,90% y la clase “Femenino” obtuvo un 43,10%, por lo tanto, la mayoría de las víctimas pertenecen al sexo masculino.

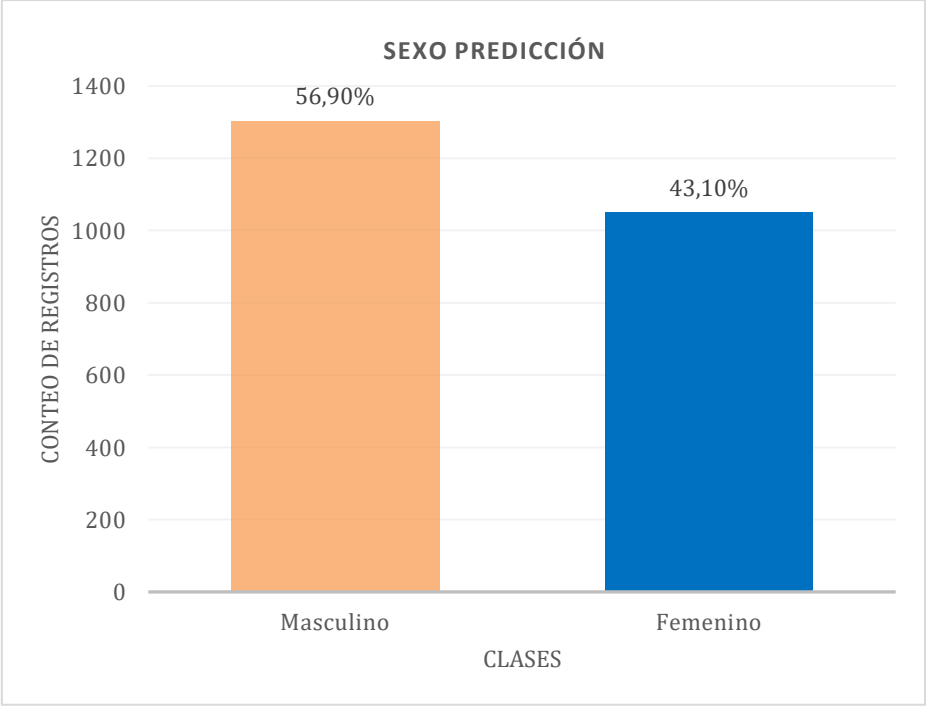


Figura 92. Distribución de asesinatos para la variable “SEXO” con el modelo SVM configurado con Optuna.

- **Variable “PRESUNTA_MOTIVACION”**

En la **Figura 93**, se visualiza la a matriz de confusión para la variable “PRESUNTA_MOTIVACION”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 217 instancias

correctamente clasificadas, la clase 1 tuvo 1166, la clase 2 tuvo 171, la clase 3 tuvo 187, la clase 4 tuvo 190 y la clase 5 tuvo 211 clasificaciones correctas.

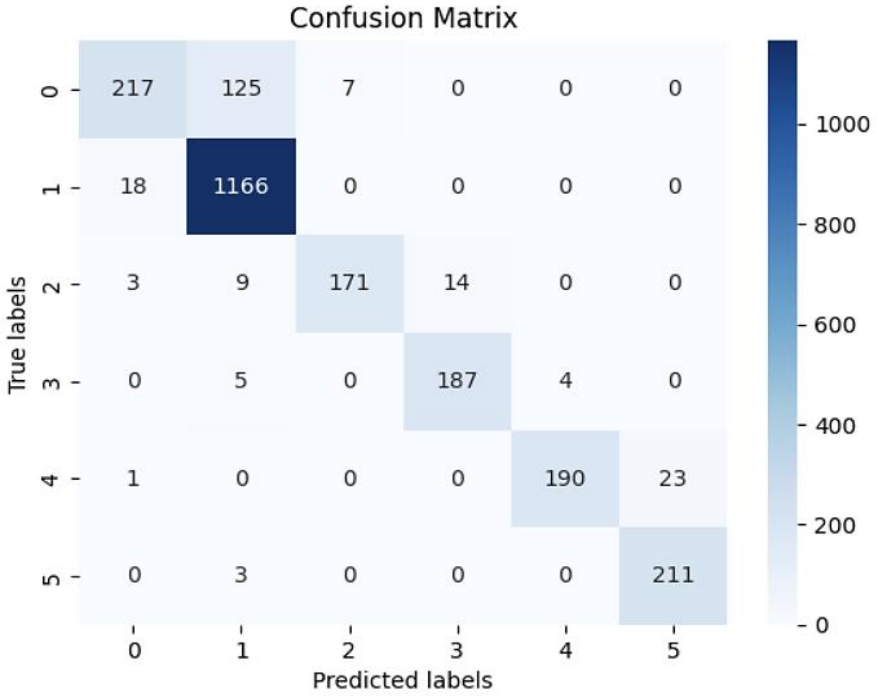


Figura 93. Matriz de confusión para la variable “PRESUNTA_MOTIVACION” con el modelo SVM configurado con Optuna.

En la **Tabla 34**, muestra el desempeño del modelo SVM para la variable "PRESUNTA_MOTIVACION", en donde la clase "Delincuencia común" tuvo una precisión del 89,14% y un recall de 98,47%, la clase "Violencia intrafamiliar y sexual" alcanzó una precisión del 96,06 % y un recall del 86,80 %, la clase "Terrorismo" tuvo un recall del 98,59% y una precisión del 90,17 %, la clase "Violencia comunitaria" presentó un recall del 62,17% y una precisión del 90,79%, la clase “Transnacional” tuvo una precisión del 93,03% y un recall del 95,40%, la clase “Psicopatologías” presentó una precisión del 97,93% y un recall del 98,59%. El modelo SVM tuvo una exactitud global (accuracy) del 90,99%, indicando un nivel aceptable en su clasificación.

Tabla 34. Desempeño del modelo SVM para la variable “PRESUNTA_MOTIVACION” configurada con los hiperparámetros de la librería Optuna.

PRESUNTA_MOTIVACION		
Clase	**Precisión**	**Recall**
0: Violencia comunitaria	90,79 %	62,17 %
1: Delincuencia común	89,14 %	98,47 %
2: Violencia intrafamiliar y sexual	96,06 %	86,80 %
3: Transnacional	93,03 %	95,40 %
4: Psicopatologías	97,93 %	88,78 %
5: Terrorismo	90,17 %	98,59 %
Accuracy	90,99 %	90,99 %

En la **Figura 94**, se visualiza un gráfico de barras que muestra la distribución de asesinatos según la presunta motivación, divididos en seis categorías: la primera de las categorías es "Delincuencia común" que tuvo un 55,60% del total, mientras que la categoría "Transnacional" presentó un 8,50%, la categoría "Terrorismo" presentó un 9,90%, la categoría "Psicopatologías" tuvo un 8,20%, la categoría "Violencia intrafamiliar y sexual" tuvo un 7,60 % y la categoría "Violencia comunitaria" alcanzó un 10,20%. Esta representación gráfica permite observar que la mayoría de los asesinatos están motivados por la delincuencia común.

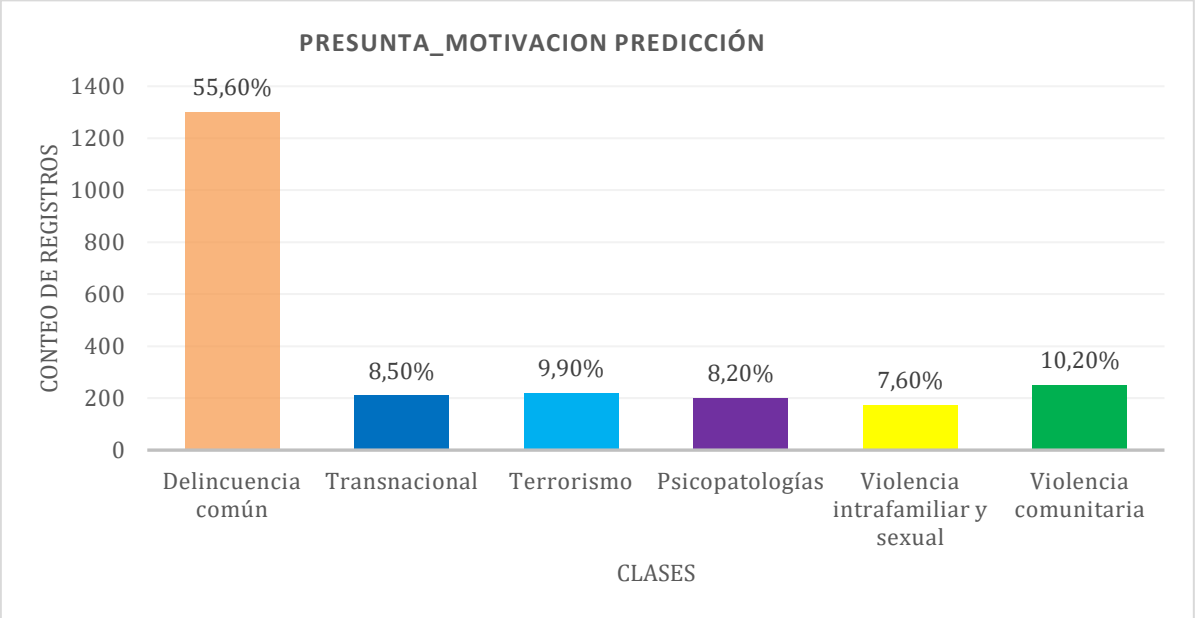


Figura 94. Distribución de asesinatos para la variable “PRESUNTA_MOTIVACION” con el modelo SVM configurado con Optuna.

- **Variable “ARMA”**

En la **Figura 95**, se visualiza la matriz de confusión para la variable "ARMA" que mostró el desempeño del modelo Support Vector Machine (SVM), evidenciando las predicciones correctas e incorrectas realizadas. En ella, la diagonal principal, que va desde la esquina superior izquierda hasta la inferior derecha, refleja las clasificaciones correctas del modelo, indicando su habilidad para identificar cada clase. Las celdas fuera de la diagonal muestran las predicciones incorrectas, en cuanto a las clases específicas, la clase 0 presentó 1223 instancias correctamente clasificadas, lo que demuestra un alto nivel de precisión. La clase 1 tuvo 279 aciertos. La clase 2, con 208 aciertos, mientras que la clase 3, con 178 instancias correctas, reveló mayor dificultad en la clasificación. Finalmente, la clase 4 presentó 234 clasificaciones correctas.

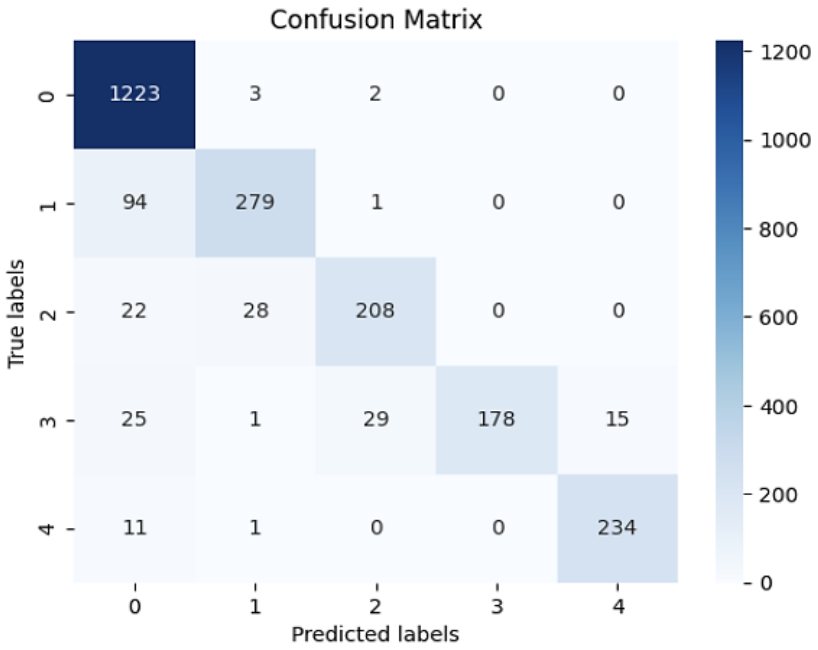


Figura 95. Matriz de confusión para la variable “ARMA” con el modelo SVM configurado con Optuna.

En la **Tabla 35**, se muestra el desempeño del modelo para la variable “ARMA” en donde la clase "Arma de fuego" tuvo una precisión del 88,94% y un recall del 99,59%, la clase "Arma blanca" obtuvo una precisión del 89,42% y un recall del 74,59%, mientras que la clase "Sustancias y otros" alcanzó una precisión del 86,66% y un recall del 80,62%, la clase "Arma contundente" tuvo una precisión de 100%, pero un recall más bajo de 71,77%, finalmente la clase "Arma constrictora" alcanzó una precisión del 93,97% y un recall del 95,12%, el modelo SVM obtuvo una exactitud global (accuracy) del 90,14%.

Tabla 35. Desempeño del modelo SVM para la variable “ARMA” configurada con los hiperparámetros de la librería Optuna.

ARMA		
Clase	**Precisión**	**Recall**
0: Arma de fuego	88,94 %	99,59 %
1: Arma blanca	89,42 %	74,59 %
2: Sustancias y otros	86,66 %	80,62 %
3: Arma contundente	100,00 %	71,77 %
4: Arma constrictora	93,97 %	95,12 %
Accuracy	90,14 %	90,14 %

En la **Figura 96,** se visualiza un gráfico de barras que muestra la distribución de asesinatos según el tipo de arma utilizada, divididos en cinco categorías: la primera es la categoría "Arma de fuego" que alcanzó un 58,40% del total, mientras que la categoría "Sustancias y otros" presentó un 10,20%, la categoría "Arma constrictora" tuvo un 10,60%, la categoría "Arma contundente" llegó a un 7,60% y la categoría "Arma blanca" presentó un 13,30% del total. Esto evidenció que la mayoría de los asesinatos se cometen con armas de fuego.

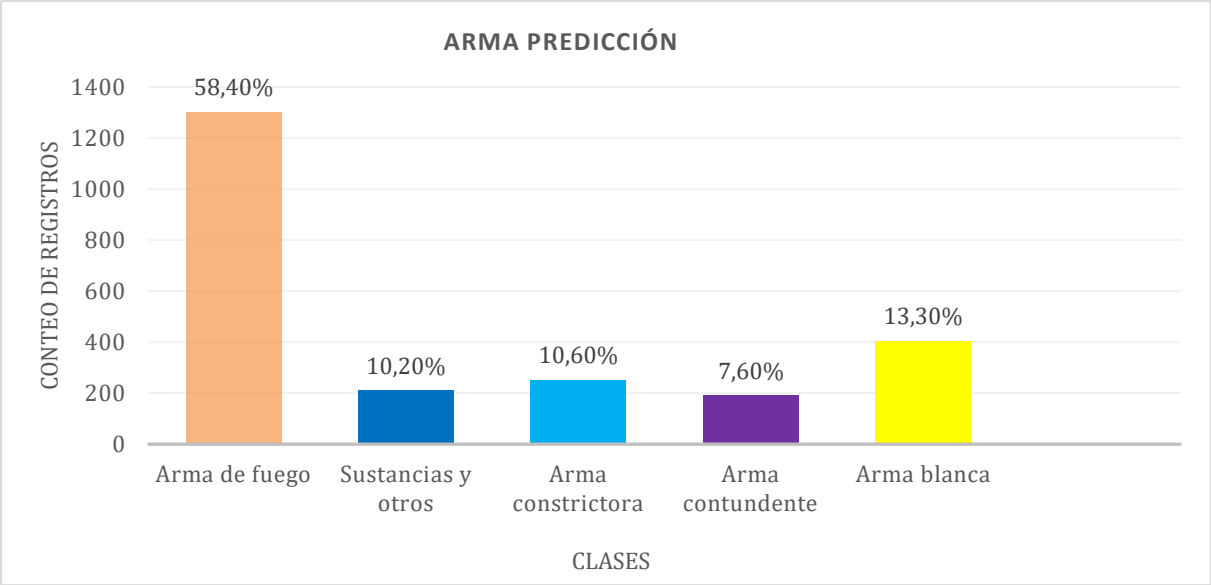


Figura 96. Distribución de asesinatos para la variable “ARMA” con el modelo SVM configurado con Optuna.

- **Variable “DIA”**

En la **Figura 97**, se visualiza la matriz de confusión para la variable “DIA”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 649 instancias correctamente clasificadas, la clase 1 tuvo 254, la clase 2 tuvo 240, la clase 3 tuvo 291 clasificaciones correctas.

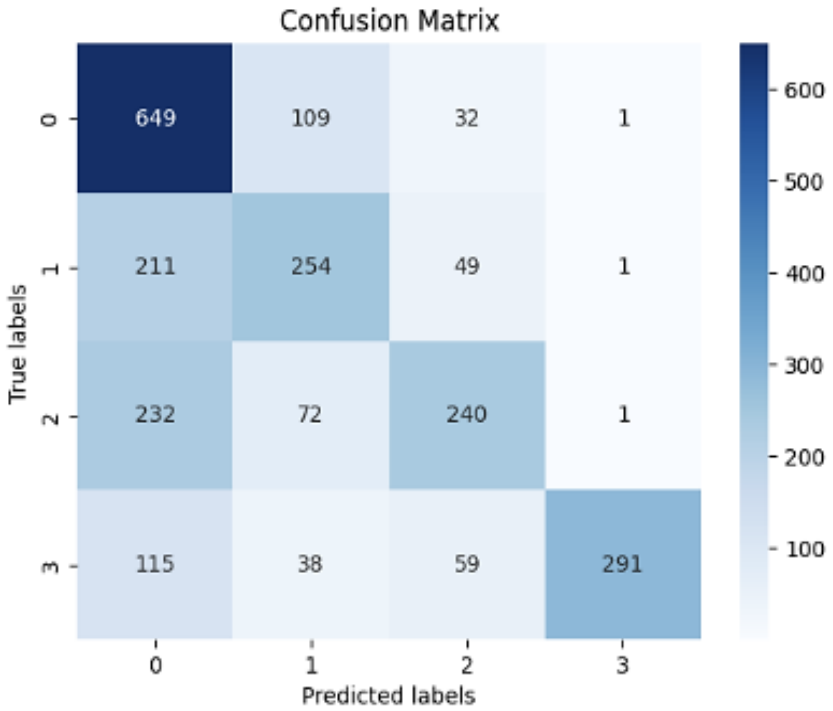


Figura 97. Matriz de confusión para la variable “DIA” con el modelo SVM configurado con Optuna.

En la **Tabla 36**, se muestra el desempeño del modelo SVM para la variable “DIA”, en donde la clase 0 tuvo una precisión del 53,76% y un recall del 82,04%, la clase 1 obtuvo una precisión del 53,69% y un recall del 49,32%, mientras que la clase 2 alcanzó una precisión del 63,15% y un recall del 44,03%, la clase 3 tuvo una precisión de 98,97%, pero un recall más bajo de 57,85%, el modelo SVM obtuvo una exactitud global (accuracy) del 60,91%, evidenciando un nivel regular en su clasificación.

Tabla 36. Desempeño del modelo SVM para la variable “DIA” configurada con los hiperparámetros de la librería Optuna.

DIA		
Clase	**Precisión**	**Recall**
0: Sábado, Domingo	53,76 %	82,04 %
1: Martes, Viernes	53,69 %	49,32 %
2: Miércoles	63,15 %	44,03 %
3: Lunes, Jueves	98,97 %	57,85 %
Accuracy	60,91 %	60,91 %

En la **Figura 98**, se visualiza un gráfico de barras que muestra la distribución de asesinatos según el día en dónde más se cometen, divididos en cuatro categorías: la primera es la categoría "Sábado, Domingo" que alcanzó un 51,30% del total, mientras que la categoría "Martes, Viernes" presentó un 20,10%, la categoría "Miércoles" tuvo un 16,10%, la categoría "Lunes, Jueves" llegó a un 12,50%. Esto evidenció que la mayoría de los asesinatos se cometen los días sábado y domingo.

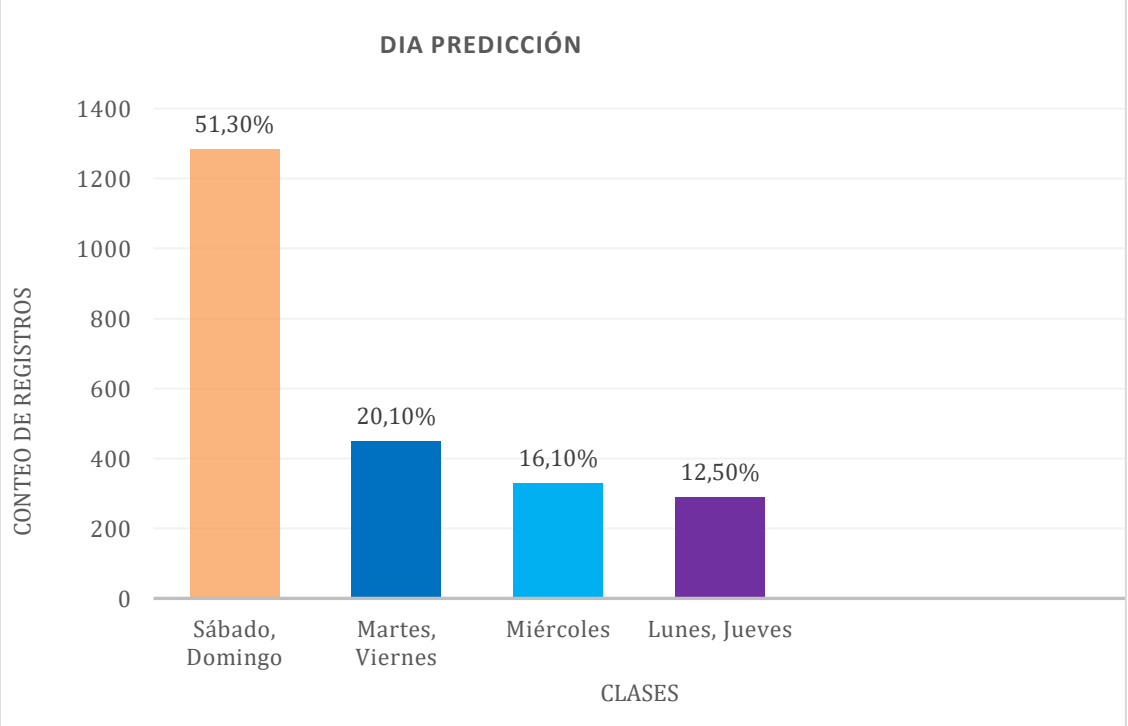


Figura 98. Distribución de asesinatos para la variable “DIA” con el modelo SVM configurado con Optuna.

- **Variable “EDAD”**

En la **Figura 99**, se visualiza la matriz de confusión para la variable “EDAD”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 435 instancias correctamente clasificadas, la clase 1 tuvo 1181, la clase 2 tuvo 195, la clase 3 tuvo 290 clasificaciones correctas.

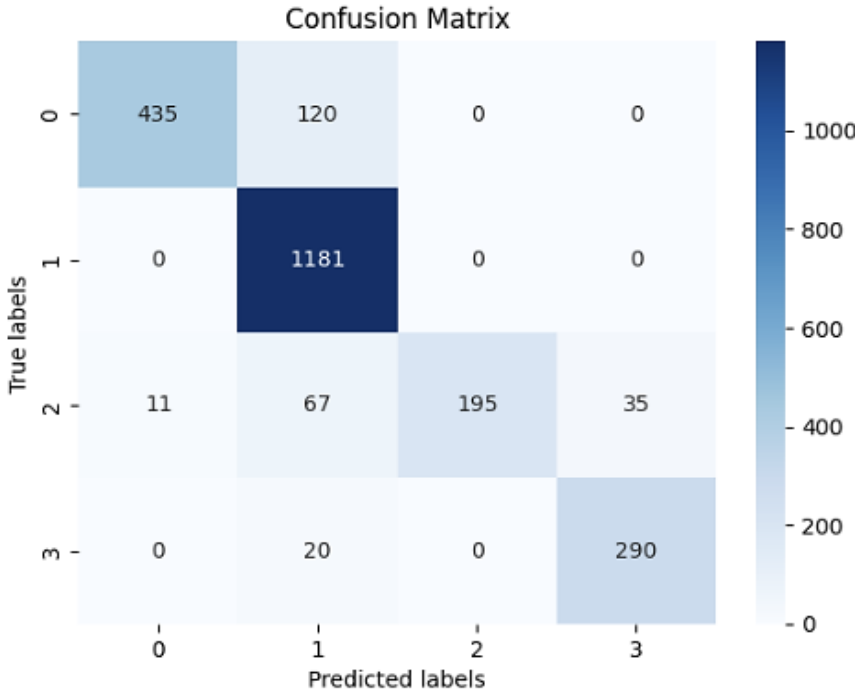


Figura 99. Matriz de confusión para la variable “EDAD” con el modelo SVM configurado con Optuna.

En la **Tabla 37**, se muestra el desempeño del modelo SVM para la variable “EDAD”, en donde la clase 0 tuvo una precisión del 97,53% y un recall del 78,37%, la clase 1 obtuvo una precisión del 85,08% y un recall del 100,00%, mientras que la clase 2 alcanzó una precisión del 100,00% y un recall del 63,31%, la clase 3 tuvo una precisión de 89,23%, pero un recall de 93,54%, el modelo SVM obtuvo una exactitud global (accuracy) del 89,25%, evidenciando un nivel bueno en su clasificación.

Tabla 37. Desempeño del modelo SVM para la variable “EDAD” configurada con los hiperparámetros de la librería Optuna.

EDAD		
Clase	**Precisión**	**Recall**
0: 1-19	97,53 %	78,37 %
1: 20-50	85,08 %	100,00 %
2: 51-65	100,00 %	63,31 %
3: 66-95	89,23 %	93,54 %
Accuracy	89,25 %	89,25 %

En la **Figura 100**, se visualiza un gráfico de barras que muestra la distribución de asesinatos basándose en la edad de la víctima, divididos en cuatro categorías: la primera es la categoría que va en un rango de “20 – 50”, alcanzó un 59,00%, mientras que la categoría que va desde “1 -19” presentó un 18,90%, la categoría “66 – 95” tuvo un 13,80% y la categoría “51 – 65” llegó a un 8,30%. Por lo tanto, la mayoría de las víctimas tienen edades que van desde los 20 hasta los 50 años.

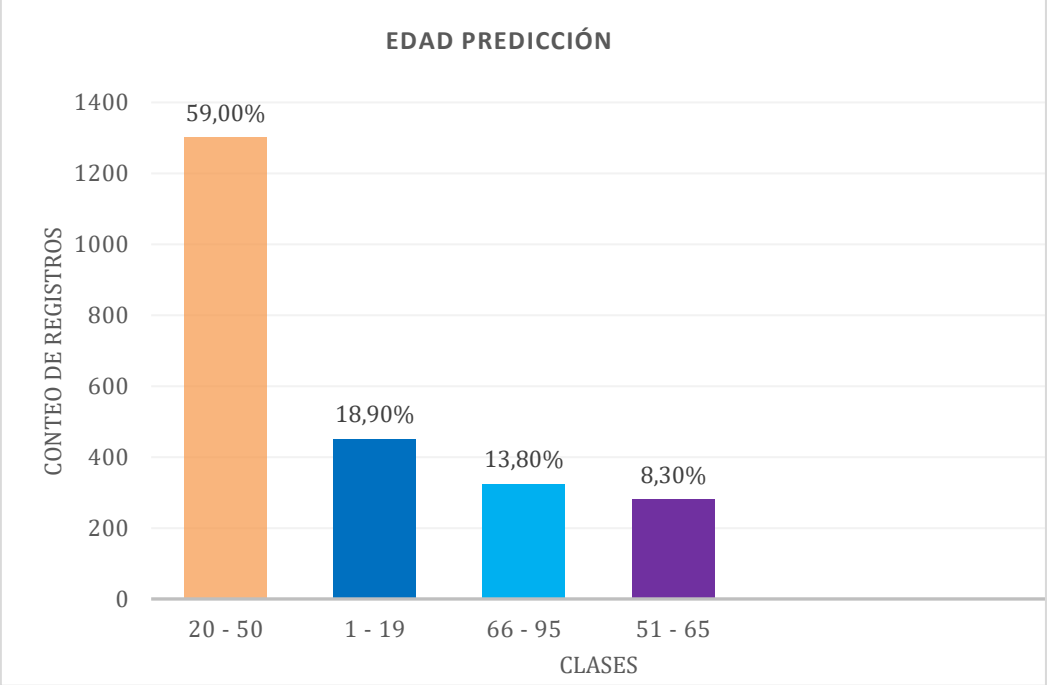


Figura 100. Distribución de asesinatos para la variable “EDAD” con el modelo SVM configurado con Optuna.

- **Variable “HORA_INFRACCION”**

En la **Figura 101**, se visualiza la matriz de confusión para la variable “HORA_INFRACCION”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo Árbol de Decisión, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 366 instancias correctamente clasificadas, la clase 1 tuvo 372, la clase 2 tuvo 265, la clase 3 tuvo 571 clasificaciones correctas.

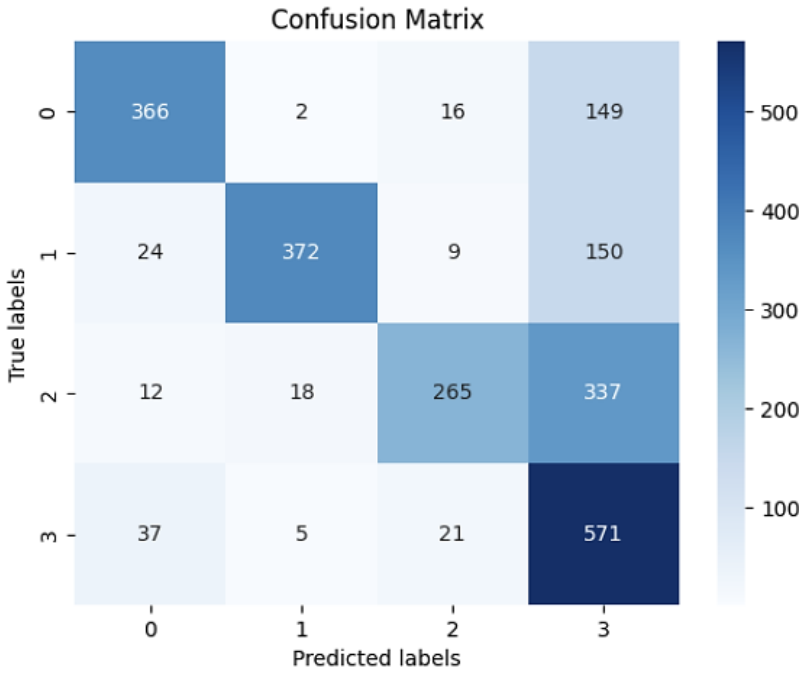


Figura 101. Matriz de confusión para la variable “HORA_INFRACCION” con el modelo SVM configurado con Optuna.

En la **Tabla 38**, se muestra el desempeño del modelo SVM para la variable “HORA_INFRACCION”, en donde la clase 0 tuvo una precisión del 83,37% y un recall del 68,66%, la clase 1 obtuvo una precisión del 93,70% y un recall del 67,02%, mientras que la clase 2 alcanzó una precisión del 85,20% y un recall del 41,93%, la clase 3 tuvo una precisión de 47,30%, pero un recall más bajo de 90,06%, el modelo obtuvo una exactitud global (accuracy) del 66,86%, evidenciando un nivel regular en su clasificación.

Tabla 38. Desempeño del modelo SVM para la variable “HORA_INFRACCION” configurada con los hiperparámetros de la librería Optuna.

HORA_INFRACCION		
Clase	**Precisión**	**Recall**
0: H01 – H06	83,37 %	68,66 %
1: H07 – H12	93,70 %	67,02 %
2: H13 – H18	85,20 %	41,93 %
3: H19 – H00	47,30 %	90,06 %
Accuracy	66,86 %	66,86 %

En la **Figura 102**, se visualiza un gráfico de barras que muestra la distribución de asesinatos basándose en la hora que se cometen los crímenes, divididos en cuatro categorías: la primera es la categoría que va en un rango horario de “H19 - H00” tuvo un 51,30%, mientras que la categoría “H07 - H12” obtuvo un 16,90%, la categoría “H01 – H06” alcanzó un 18,60%, la categoría “H13 – H18” consiguió un 13,20%. Por lo tanto, la mayoría de los asesinatos se realizan en horas que van desde las 19:00 pm hasta las 00:00 am (medianoche).

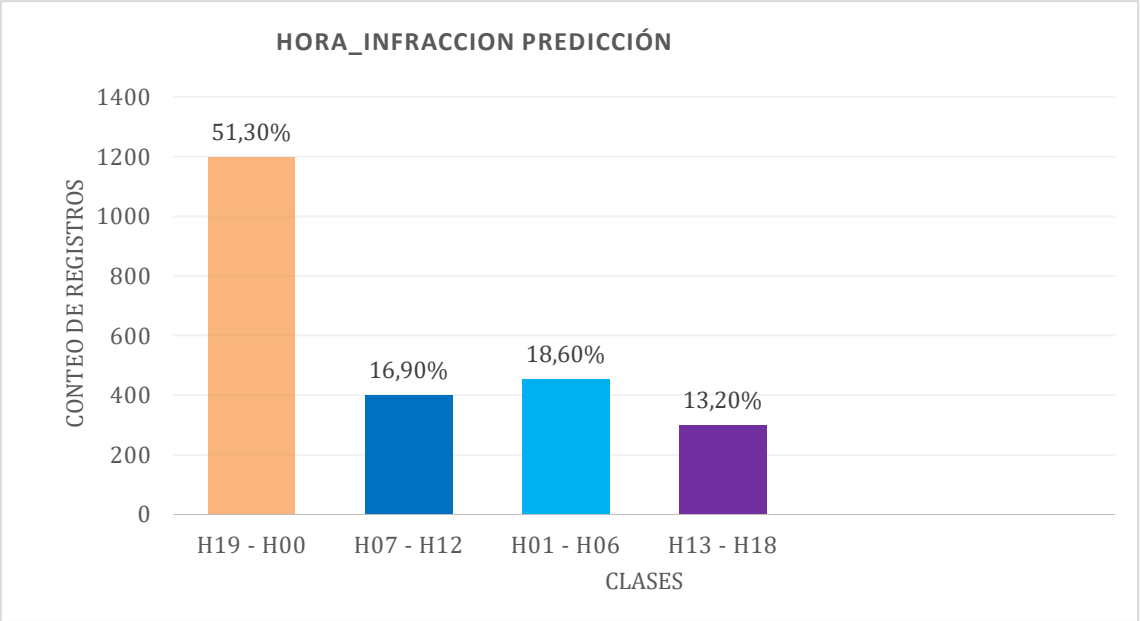


Figura 102. Distribución de asesinatos para la variable “HORA_INFRACCION” con el modelo SVM configurado con Optuna.

- **Variable “DISTRITO”**

En la **Figura 103**, se visualiza la matriz de confusión para la variable “DISTRITO”, en donde la diagonal principal (de izquierda arriba a derecha abajo) evidencia las predicciones correctas realizadas por el modelo SVM, mientras que las restantes son las predicciones incorrectas, por lo tanto, la clase 0 tuvo 854 instancias correctamente clasificadas, la clase 1 tuvo 319, la clase 2 tuvo 435 clasificaciones correctas.

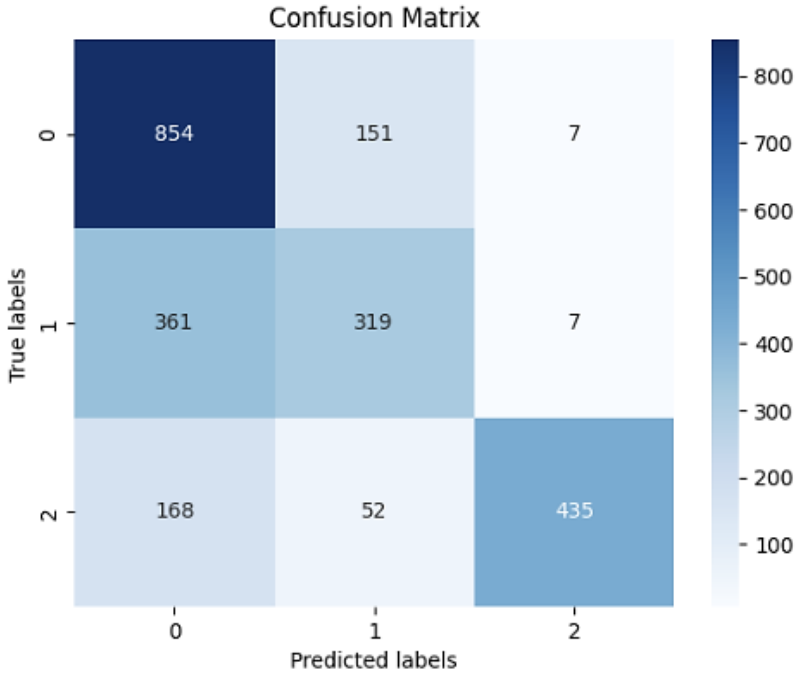


Figura 103. Matriz de confusión para la variable “DISTRITO” con el modelo SVM configurado con Optuna.

En la **Tabla 39**, se muestra el desempeño del modelo SVM para la variable “DISTRITO”, en donde la clase 0 tuvo una precisión del 61,74% y un recall del 84,38%, la clase 1 obtuvo una precisión del 61,11% y un recall del 46,43%, mientras que la clase 2 alcanzó una precisión del 96,88% y un recall del 66,41%, el modelo obtuvo una exactitud global (accuracy) del 68,30%, evidenciando un nivel regular en su clasificación.

Tabla 39. Desempeño del modelo SVM para la variable “DISTRITO” configurada con los hiperparámetros de la librería Optuna.

DISTRITO		
Clase	**Precisión**	**Recall**
0: Nueva Prosperina, Distrito Sur, Pascuales	61,74 %	84,38 %
1: Portete, 9 de Octubre, Durán, Estero	61,11 %	46,43 %
2: Progreso, Florida, Modelo, Ceibos, Samborondón	96,88 %	66,41 %
Accuracy	68,30 %	68,30 %

En la **Figura 104**, se visualiza un gráfico de barras que muestra la distribución de asesinatos basándose en el distrito, divididos en tres categorías: la primera es la categoría que consta de "Nueva Prosperina, Distrito Sur, Pascuales" esta alcanzó un 58,80% del total, mientras que la categoría "Portete, 9 de Octubre, Durán, Estero" presentó un 22,20%, la categoría " Progreso, Florida, Modelo, Ceibos, Samborondón " tuvo un 19,10%. Esto evidenció que la mayoría de los asesinatos se realizan en los distritos de Nueva Prosperina, Distrito Sur y Pascuales, de la Zona 8 del Ecuador.

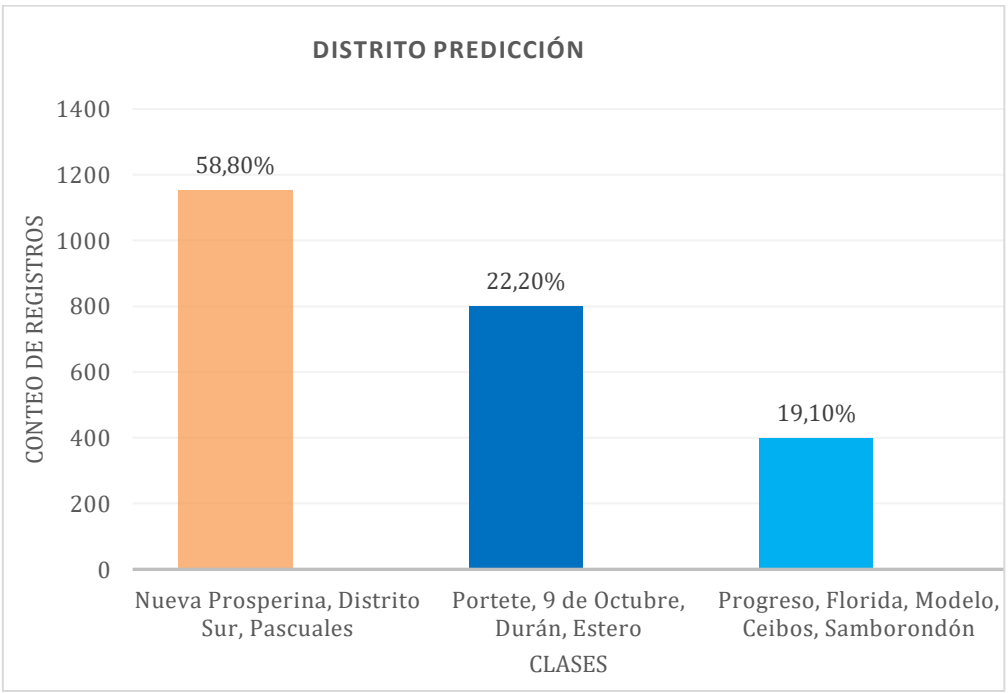


Figura 104. Distribución de asesinatos para la variable “DISTRITO” con el modelo SVM configurado con Optuna.

6.2.1.6. Elección del modelo de minería de datos para la fase de Despliegue

- **Comparación del desempeño en las métricas de rendimiento entre los modelos Árbol de Decisión y SVM**

Los modelos de minería de datos: Árbol de Decisión y SVM presentaron mayores porcentajes al configurarlos con los hiperparámetros obtenidos con la librería Optuna de OB, se observó que el modelo SVM alcanzó un porcentaje mayor en sus métricas de precisión, accuracy y recall, en la **Tabla 40**, se puede visualizar los valores alcanzados por ambos modelos de minería de datos.

Tabla 40. Desempeño de los modelos de minería de datos configurados con los hiperparámetros de la librería Optuna.

Métricas de evaluación	Árbol de Decisión - Optuna		Support Vector Machine - Optuna	
	Validación Cruzada (%)	Sklearn.metrics (%)	Validación Cruzada (%)	Sklearn.metrics (%)
Precisión	83,63	84,04	83,78	84,12
Accuracy	83,62	84,04	83,78	84,12
Recall	83,62	84,04	83,78	84,12

- **Comparación del tiempo de ejecución entre el Árbol de Decisión y SVM**

En el proceso de Optimización Bayesiana de hiperparámetros, se observó una mínima diferencia en el tiempo de ejecución entre los modelos de minería de datos Árbol de Decisión (DT) y Support Vector Machine (SVM).

En la **Figura 105**, se muestra el tiempo de ejecución que tomó realizar la búsqueda de hiperparámetros con la librería de OB llamada Optuna en el modelo Árbol de Decisión, tomando un tiempo de entre 3 a 60 segundos.

Figura 105. Tiempo de ejecución empleado para realizar la búsqueda de hiperparámetros con Optuna en el modelo Árbol de Decisión.

En la **Figura 106**, se muestra el script del tiempo usado para construir el modelo Árbol de Decisión (DT), el cual tuvo un tiempo de ejecución de entre 0 y 5 segundos para cada una de las 10 variables que fueron seleccionadas para este estudio.

```
#Llamamos al constructor del arbol de decision
classifier = DecisionTreeClassifier(max_depth=7, min_samples_split=8,min_samples_leaf=6, criterion='entropy')

#Entrenamos el modelo
arbol_modelo = classifier.fit(X_train,y_train)

y_pred = classifier.predict(X_test)
```

✓ 0 s completado a las 16:36

Figura 106. Tiempo de ejecución empleado para construir el modelo Árbol de Decisión con los hiperparámetros encontrados.

En la **Figura 107**, se muestra el tiempo de ejecución que tomó realizar la búsqueda de hiperparámetros con la librería Optuna de Optimización Bayesiana en el modelo Support Vector Machine (SVM), en donde varió el tiempo para terminar la ejecución entre 2 a 3 horas.

```
01-04 20:33:25,408] Trial 24 finished with value: 0.6079014443500425 and parameters: {'C': 77.1805889893365
hiperparámetros encontrados: {'C': 99.16715122292081, 'kernel': 'poly', 'decision_function_shape': 'ovr'}
 precisión alcanzada es: 60.92%
```

✓ 3 h 13 min 55 s completado a las 15:32

Figura 107. Tiempo de ejecución empleado para realizar la búsqueda de hiperparámetros con Optuna en el modelo Support Vector Machine.

En la **Figura 108**, se muestra el tiempo usado para construir el modelo Support Vector Machine (SVM), el cual tuvo un tiempo de ejecución de entre 5 y 16 minutos para cada una de las 10 variables que se seleccionó para este estudio.

+ Código + Texto

```
# Cargamos la librería Support Vector Classifier
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix, classification_report

# Llamamos al constructor de Support Vector Machine
classifier = SVC(kernel='poly', C=99.16, decision_function_shape='ovr')

# Entrenamos el modelo
svm_modelo = classifier.fit(X_train, y_train)

# Realizamos predicciones en el conjunto de prueba
y_pred = classifier.predict(X_test)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1408: DataConversionWarning: A column
  y = column_or_1d(y, warn=True)
```

```
# Resumen de las predicciones hechas por el clasificador
from sklearn import metrics
reporte = metrics.classification_report(y_test, y_pred,output_dict=True)
pre = pd.DataFrame(reporte).transpose()
print(pre)
pre.to_excel("resumen_día.xlsx")
```

```
              precision    recall  f1-score      support
1              0.538782  0.816688  0.649246   791.000000
2              0.536170  0.489320  0.511675   515.000000
3              0.627249  0.447706  0.522484   545.000000
```

✓ 16 min 31 s completado a las 16:30

Figura 108. Tiempo de ejecución empleado para construir el modelo Support Vector Machine con los hiperparámetros encontrados.

Se eligió el modelo Support Vector Machine (SVM) y los resultados obtenidos en sus clasificaciones y predicciones para realizar la fase de despliegue, ya que se evidenció que es mínima la diferencia en el tiempo de ejecución que presenta en comparación con el modelo Árbol de Decisión (DT), además el SVM presentó un 83,70% (validación cruzada) y un 84,12% (sklearn.metrics) tanto en las métricas de precisión, accuracy y recall, siendo mayores en comparación con las métricas obtenidas por el modelo DT que alcanzó un 83,62% (validación cruzada) y un 84,04% (sklearn.metrics) en las mismas métricas evaluadas.

6.2.2. Fase 6: Despliegue

En esta última fase de la metodología CRISP-DM, se realizó un informe[23] de resultados y un dashboard[24] que se puede visualizar en todos los navegadores, ambos contienen representaciones gráficas para entender los patrones encontrados, estos resultados fueron validados por el Subteniente de la Policía Nacional (ver en **Anexo VI**).

En la **Figura 109**, se muestra el dashboard realizado en la herramienta Looker Studio, el cual se encuentra en la nube, en el servidor que pertenece a la empresa de Google Cloud, en la parte izquierda se observa cada una de las variables que fueron seleccionadas, y se despliegan los gráficos que muestran las clasificaciones realizadas por el modelo SVM, también se observa el mapa geográfico de la Zona 8 del Ecuador, el cual fue obtenido mediante Google Maps.

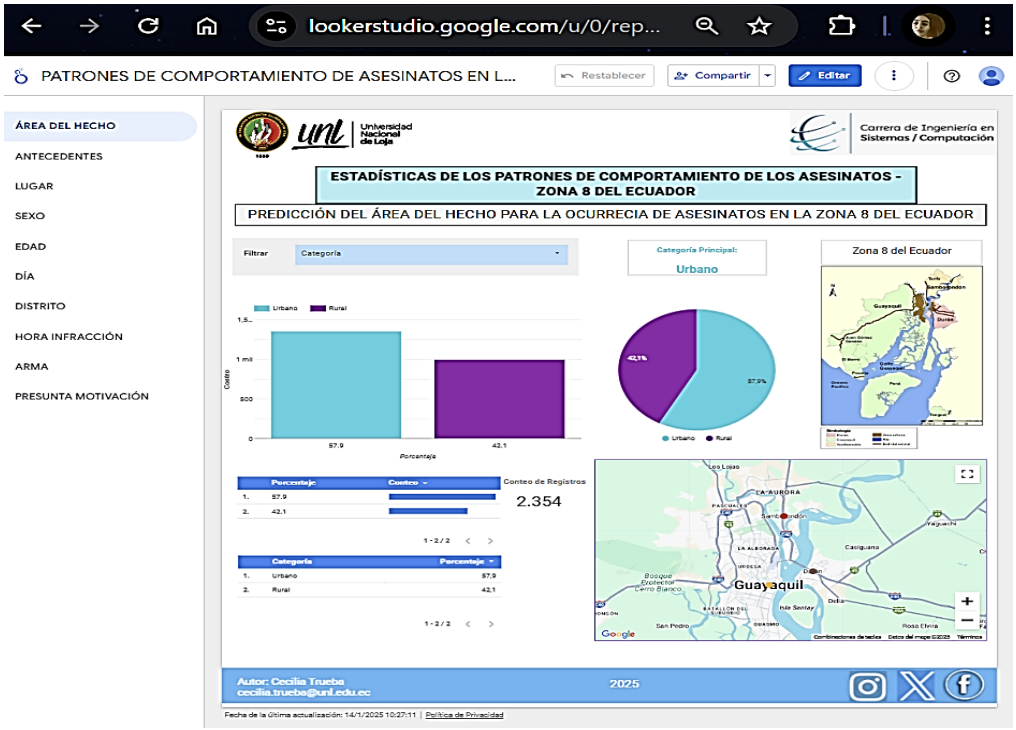


Figura 109. Dashboard generado mediante la herramienta Looker Studio.

[23] https://drive.google.com/file/d/1y-urC1JAfidRq2Sh1v0REG18JwCPu6-4/view?usp=sharing
[24] https://lookerstudio.google.com/reporting/83297a6d-05f2-4d29-ac6e-e3e3f99f35c7

7. Discusión

7.1 Primer objetivo: Mejorar la calidad del dataset de asesinatos del Ecuador mediante técnicas de preparación de datos, realizar el análisis exploratorio de datos EDA e implementación de los clasificadores optimizados Árbol de Decisión y SVM.

El Análisis Exploratorio de Datos (EDA) permite obtener una comprensión detallada de la base de datos inicial, identificando el número de registros y variables presentes, mediante este análisis se identifican 10 variables clave como las más relevantes para el estudio, pues en [7] se eligen atributos similares a los de este trabajo para realizar el proceso de minería de datos, igualmente en [55] mediante el EDA se explora en los datos la existencia de valores nulos, registros inusuales o errores, siendo un paso fundamental para guiar la etapa de preparación de los datos. La mejora de la calidad del dataset a través de técnicas de preparación de datos, permite tratar inconsistencias en los registros, por tal motivo en los estudios [62] y [63], al igual que con el presente trabajo se usa la herramienta OpenRefine para realizar diversas tareas que forman parte del procesamiento de datos, como son la limpieza de registros nulos o faltantes y la estandarización de los atributos y registros del dataset, pero a diferencia de estos estudios, en el actual trabajo se realiza el balanceo de clases desequilibradas mediante la técnica de SMOTE, ya que esta tarea evita que los clasificadores tengan sesgo hacia la clase mayoritaria, realizando predicciones incorrectas.

Se implementan los clasificadores Árbol de Decisión y Support Vector Machine (SVM), que son configurados con la técnica de Optimización Bayesiana (OB), mediante el uso de las librerías Optuna y Bayesian-Optimization (véase Tabla 18 y Tabla 19), al igual que en los estudios [30], [34], [57], [58], se aplicaron estas librerías de OB para ajustar los modelos DT y SVM, manteniendo los mismos espacios de búsqueda de valores para los hiperparámetros.

En los trabajos mencionados anteriormente, la Optimización Bayesiana (OB) ha sido aplicada en modelos de minería de datos, aunque sobre conjuntos de datos diferentes al utilizado en el presente estudio. En este caso, se trabaja con un dataset con 10 variables y 6920 registros que contiene información sobre los asesinatos ocurridos en la Zona 8 del Ecuador, en el período comprendido entre los años 2015 y 2024.

7.2 Segundo objetivo: Evaluar los clasificadores Árbol de Decisión y SVM con la métrica de precisión.

En los trabajos [58], [59] y [60] se evalúa la precisión antes y después de aplicar la OB en los modelos de minería de datos evidenciando una mejora en sus porcentajes, de igual manera en el presente trabajo al realizar la evaluación de los clasificadores mediante la validación cruzada y con la librería sklearn.metrics de Scikit-learn, se observa que la precisión, el accuracy y el recall aumentan al aplicar la técnica de Optimización Bayesiana (OB). Sin el uso de esta técnica, se alcanzan menores porcentajes en estas métricas, ya que se obtiene

un 83,37% (validación cruzada) y un 83,56% (sklearn.metrics) en precisión, accuracy y recall para el modelo Árbol de Decisión (DT), mientras que para el modelo Support Vector Machine (SVM), se obtiene un 82,34% (validación cruzada) y un 82,20% (sklearn.metrics) en los porcentajes de las métricas de precisión, accuracy y recall. En el caso del modelo Árbol de Decisión configurado con la librería Optuna, se alcanza un porcentaje similar en la precisión, accuracy y recall, del 83,63% al evaluarlo con validación cruzada y 84,04% con sklearn.metrics, frente al 83,50% (validación cruzada) y 83,86% con la librería Bayesian-Optimization. Para el modelo SVM, Optuna logra un porcentaje de precisión, accuracy y recall superior, alcanzando un 83,78% (validación cruzada) y 84,12% (sklearn.metrics), mientras que con la librería Bayesian-Optimization se obtiene un 83,86% (validación cruzada) y un 83,98% (sklearn.metrics).

En base a lo anterior descrito se procede a responder a la pregunta de investigación que versa: ¿Qué porcentaje de precisión se puede alcanzar en los modelos Árbol de Decisión y SVM al usar la técnica de Optimización Bayesiana para configurarlos y poder determinar mediante minería de datos los patrones en los asesinatos de la Zona 8 del Ecuador correspondientes a los años de enero de 2015 a febrero de 2024?, considerando que el modelo Árbol de Decisión alcanza un porcentaje de precisión del 83,63% (validación cruzada) y 84,04% (sklearn.metrics), mientras que el modelo Support Vector Machine alcanza un 83,78% (validación cruzada) y un 84,12% (sklearn.metrics), mediante la aplicación de la técnica de Optimización Bayesiana específicamente con la librería Optuna.

Además, se tiene que el tiempo de ejecución para construir el modelo DT demora entre 0 y 5 segundos, en cambio el modelo SVM muestra un tiempo de ejecución para la construcción del modelo de entre 15 y 30 minutos por cada variable. Aunque el tiempo de ejecución y el costo computacional son mayores en el modelo Support Vector Machine (SVM) en comparación con el Árbol de Decisión (DT), los porcentajes obtenidos en sus métricas evaluadas son superiores (véase Tabla 40). Al ser el SVM el mejor modelo, se determina con sus clasificaciones que los patrones de los asesinatos ocurridos en la Zona 8, indican que mayormente se realizan estos crímenes con armas de fuego, en los distritos de Nueva Prosperina, Sur y Pascuales, en áreas urbanas, en la vía pública, los fines de semana, desde las 19:00 pm hasta las 00:59 am, afectando al sexo masculino, en edades de entre los 20 y 50, que presentan antecedentes penales, por presuntos motivos relacionados a la delincuencia común.

Para futuras investigaciones, se puede aplicar la técnica de Optimización Bayesiana en diversos modelos de minería de datos y contrastar con los resultados y porcentajes obtenidos por los modelos SVM y DT que se implementan en el presente estudio.

8. Conclusiones

- La colaboración recibida por parte del Ministerio del Interior en conjunto con la Policía Nacional del Ecuador es necesaria para completar este estudio, especialmente en las fases de comprensión del negocio y evaluación ya que, al ser otorgada la base de datos de Homicidios Intencionales desde enero de 2015 hasta febrero de 2024 por parte de los organismos encargados, es decir a partir de una fuente confiable, se asegura la veracidad de los datos.
- Al aplicar las fases de la metodología CRISP-DM, como la ingeniería de datos, se mejora la calidad del dataset mediante técnicas como la estandarización de variables y registros, la limpieza de valores faltantes y el balanceo de clases desequilibradas, así también, se incluye la etapa de ingeniería de modelos, que facilita la implementación y aplicación de la Optimización Bayesiana (OB) en los clasificadores Árbol de Decisión y Support Vector Machine, que son fundamentales para llevar a cabo procesos de minería de datos y determinar patrones en los asesinatos ocurridos en la Zona 8 del Ecuador.
- Al ajustar los modelos Árbol de Decisión (DT) y Support Vector Machine (SVM) con los hiperparámetros obtenidos mediante la técnica de Optimización Bayesiana, utilizando la librería Optuna, se alcanzan los siguientes porcentajes en cada una de las tres métricas (precisión, accuracy y recall), para el modelo Árbol de Decisión (DT) se obtiene un 83,63% (validación cruzada) y 84,04% (sklearn.metrics), y en el caso del Support Vector Machine (SVM) se alcanza un 83,78% (validación cruzada) y 84,12% (sklearn.metrics). Estos resultados indican que, aunque ambos modelos presentan un desempeño similar, el modelo Support Vector Machine obtiene un rendimiento ligeramente superior, lo que lo hace el mejor para realizar las clasificaciones en el presente estudio.
- La herramienta Looker Studio resulta de gran ayuda para plasmar las representaciones visuales obtenidas mediante la minería de datos en un entorno en la nube interactivo, al cual se accede con facilidad, simplificando la visualización de los patrones encontrados, mejorando el entendimiento por parte de los usuarios.

9. Recomendaciones

- En la fase de preparación de los datos, realizar el balanceo de las variables del conjunto de datos que presentan clases desequilibradas mediante técnicas como SMOTE permite igualar la distribución de las clases, obteniendo una representación adecuada, de esta manera se evita que el modelo favorezca a las clases mayoritarias.
- Aumentar el número de registros del dataset, ya que esto favorece a los modelos de minería de datos al proporcionar una mayor cantidad de información y generalizar de manera adecuada al momento de trabajar con nuevos datos.
- Considerar la ampliación del análisis a otras zonas del país, incluyendo variables adicionales que permitan conocer y obtener información detallada sobre los crímenes en distintas ubicaciones.
- Ampliar la aplicación de la Optimización Bayesiana a otros modelos de minería de datos para comparar su desempeño en las métricas evaluadas y contrastar los resultados con los modelos analizados en este estudio.
- Usar la herramienta gratuita OpenRefine para la limpieza y estandarización de la base de datos, ya que es una opción fácil de usar al momento de realizar la fase de preprocesamiento de los datos que se incluye en la metodología CRISP-DM, ya que permite identificar y transformar las inconsistencias que están presentes en los datos.

10. Bibliografía

[1] M. Li, Z. Li, and C. Li, "Machine learning model with Bayesian optimization for ultrasonic flowmeter in-use measurement," Measurement: Sensors, vol. 101570, 2024. [Online]. Available: https://doi.org/10.1016/j.measen.2024.101570. [Accessed: 26-Jan-2025].

[2] S. Echabarri, P. Do, H.-C. Vu, and B. Bornand, "Machine learning and Bayesian optimization for performance prediction of proton-exchange membrane fuel cells," Energy and AI, vol. 17, p. 100380, 2024. [Online]. Available: https://doi.org/10.1016/j.egyai.2024.100380. [Accessed: 26-Jan-2025].

[3] C. Wellmann, A. R. Khaleel, T. Brinkmann, A. Wahl, C. Monissen, M. Eisenbarth, and J. Andert, "Electric machine co-optimization for EV drive technology development: Integrating Bayesian optimization and nonlinear model predictive control," *eTransportation*, p. 100392, 2025.

[4] N. Mahboubi, J. Xie, and B. Huang, "Point-by-point transfer learning for Bayesian optimization: An accelerated search strategy," Computers & Chemical Engineering, vol. 194, p. 108952, 2025. [Online]. Available: https://doi.org/10.1016/j.compchemeng.2024.108952. [Accessed: 26-Jan-2025].

[5] R. M. Saeed and H. A. Abdulmohsin, "A study on predicting crime rates through machine learning and data mining using text," Journal of Intelligent Systems, vol. 32, no. 1, pp. 20220223, 2023.

[6] D. Petsain, “Aplicación de agentes virtuales para consultas estadísticas sobre casos de homicidios intencionales en Ecuador reportados por el ministerio de gobierno,” 2021.

[7] G. A. L. Villagómez, “Las armas de fuego y su impacto en la inseguridad ciudadana en Ecuador,” Innovación & Saber, vol. 2, no. 1, pp. 16–28, 2020.

[8] G. P. Cabezas Uriarte, M. S. Rodríguez Barrero, J. I. Sierra Durán, and M. H. Flórez Guzmán, “Incidencia de factores sociales y económicos en la criminalidad en Guayaquil, Ecuador,” Revista Logos Ciencia & Tecnología, vol. 16, no. 2, pp. 10–23, 2024.

[9] R. L. Martínez, “Dualidad conceptual entre el homicidio y el asesinato,” REVISTA IUS, vol. 16, no. 50, 2022.

[10] A. Marín Redondo et al., “El impacto social del asesinato y la revisión bibliográfica del caso Marta del Castillo,” 2023.

[11] A. M. C. VARGAS and M. A. E. MINA, “Aplicación de minería de datos en datos abiertos de Ecuador: Delitos,” UCV Hacer, vol. 11, no. 1, pp. 79–93, 2022.

[12] M. E. A. Donoso, N. E. C. Maurisaca, and J. E. A. Reyes, “Análisis de correspondencias múltiples para el estudio de los homicidios intencionales en el Ecuador,” Revista Politécnica, vol. 50, no. 3, pp. 43–52, 2022.

[13] J. G. Balcázar Gonzales, “Minería de datos en la detección de patrones delictivos: una revisión sistemática,” 2021.

[14] P. T. Cardona Jiménez, M. P. Garzón Bustos, and A. V. López Muñoz, “Aproximación a las características psicológicas del feminicida desde una revisión documental de 45 publicaciones de Iberoamérica,” 2020.

[15] J. I. Pincay-Ponce, N. G. Angulo-Murillo, J. S. Herrera-Tapia, and W. R. Delgado-Muentes, “Técnicas de minería de datos como soporte para la gestión de un sistema de comercialización de energía eléctrica,” Mikarimin. Revista Científica Multidisciplinaria, vol. 6, no. 2, pp. 19–34, 2020.

[16] A.-L. Pérez-Suasnavas, B. F. Salgado-Proaño, W. Hasperué, K. L. Cela, and J. L. Santamaría, “Evolución de las técnicas de minería de datos para extraer datos provenientes de twitter aplicadas a la educación superior: una revisión sistemática,” South Florida Journal of Development, vol. 4, no. 1, pp. 33–55, 2023.

[17] M. J. B. Meza, A. A. P. Moreno, and C. D. B. Borja, “Revisión bibliográfica sobre la importancia de la minería de datos en las actividades empresariales,” Dominio de las Ciencias, vol. 10, no. 2, pp. 1815–1833, 2024.

[18] R. Selman-Alvarez, U. Figueroa-Fernández, E. Cruz-Mackenna, C. Jarry, G. Escalona, M. Corvetto, and J. Varas-Cohen, “Artificial intelligence in medical simulation: current state and future outlook,” Revista Latinoamericana de Simulación Clínica, vol. 5, no. 3, pp. 117–122, 2024.

[19] J. Lennox, 2084: Inteligencia artificial y el futuro de la humanidad. Publicaciones Andamio, 2022.

[20] P. D. Terrón, A. J. M. Guerrero, J. L. Belmonte, and J. A. M. Marín, "Inteligencia Artificial y Machine Learning como recurso educativo desde la perspectiva de docentes en distintas etapas educativas no universitarias," *RiiTE Revista interuniversitaria de investigación en Tecnología Educativa*, pp. 58–78, 2023.

[21] A. Araujo-Ahon, B. Cardenas-Mayta, O. Iparraguirre-Villanueva, J. Zapata-Paulini, and M. Cabanillas-Carbonell, “Técnicas y algoritmos para predecir el resultado de los partidos de fútbol utilizando la minería de datos, una revisión de la literatura,” 2023.

[22] J. R. C. Romero, J. J. E. Romero, and C. H. Miranda, “Aplicación de algoritmos de aprendizaje automático en geociencia: revisión integral y desafío futuro,” REVISTA AMBIENTAL AGUA, AIRE Y SUELO, vol. 14, no. 2, pp. 9–18, 2023.

[23] B. P. Cordero-Torres, “Algoritmos de aprendizaje supervisado para proyección de ventas de camarón ecuatoriano con lenguaje de programación python,” Economía y Negocios, vol. 13, no. 2, pp. 30–51, 2022.

[24] R. Tobar-Díaz, Y. Gao, J. F. Mas, and V. H. Cambrón-Sandoval, “Clasificación de uso y cobertura del suelo a través de algoritmos de aprendizaje automático: revisión bibliográfica,” Revista de Teledetección, no. 62, pp. 1–19, 2023.

[25] E. Cruz, M. González, and J. C. Rangel, “Técnicas de machine learning aplicadas a la evaluación del rendimiento ya la predicción de la deserción de estudiantes universitarios, una revisión.” Prisma Tecnológico, vol. 13, no. 1, pp. 77–87, 2022.

[26] D. N. Castillo Warnken, “Caracterización y predicción de conducta de usuarios de aplicación móvil enfocado a proceso on boarding utilizando herramientas de machine learning,” 2022.

[27] M. Tanveer, T. Rajani, R. Rastogi, Y.-H. Shao, and M. Ganaie, “Comprehensive review on twin support vector machines,” Annals of Operations Research, pp. 1–46, 2022.

[28] K. L. Yandar, O. R. Sánchez, and M. E. Bolaños-González, “Machine learning para la predicción de energía eléctrica: una revisión sistemática de literatura,” Ingeniería y Competitividad, vol. 26, no. 2, 2024.

[29] L. D. S. Riveros, W. P. Ríos, and I. M. M. Ramírez, “Técnicas estadísticas y logro de aprendizaje: revisión bibliográfica,” Eco Matemático, vol. 12, no. 2, pp. 112–125, 2021.

[30] M. Aghaabbasi, M. Ali, M. Jasinski, Z. Leonowicz, and T. Novák, “On hyperparameter optimization of machine learning methods using a bayesian optimization algorithm to predict work travel mode choice,” IEEE Access, vol. 11, pp. 19 762–19 774, 2023.

[31] J. Yin and N. Li, “Ensemble learning models with a bayesian optimization algorithm for mineral prospectivity mapping,” Ore geology reviews, vol. 145, p. 104916, 2022.

[32] X. Wang, Y. Jin, S. Schmitt, and M. Olhofer, “Recent advances in bayesian optimization,” ACM Computing Surveys, vol. 55, no. 13s, pp. 1–36, 2023.

[33] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon, “Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black box optimization challenge 2020,” in NeurIPS 2020 Competition and Demonstration Track. PMLR, 2021, pp. 3–26.

[34] H. Shao, X. Liu, D. Zong, et al., "Optimization of diabetes prediction methods based on combinatorial balancing algorithm," Nutr. Diabetes, vol. 14, no. 63, 2024.

[35] H. E. Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes prediction using machine learning algorithms and ontology," *Journal of ICT Standardization*, vol. 10, no. 2, pp. 319-337, 2022, doi: 10.13052/jicts2245-800X.10212.

[36] G. Varoquaux and O. Colliot, “Evaluating machine learning models and their diagnostic value,” Machine learning for brain disorders, pp. 601– 630, 2023.

[37] M. S. Asto-Lazaro and H. P. Bermejo-Terrones, “Revisión sistemática: Machine learning en la predicción de deserción académica,” Revista Ibérica de Sistemas e Tecnologías de Información, no. E64, pp. 463– 476, 2023.

[38] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, “Machine learning applications for precision agriculture: A comprehensive review,” IEEE Access, vol. 9, pp. 4843–4873, 2020.

[39] S. Senthivel and M. Chidambaranathan, "Machine learning approaches used for air quality forecast: A review," *Revue d'Intelligence Artificielle*, vol. 36, no. 1, pp. 73, 2022.

[40] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," *IEEE Access*, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[41] N. C. P. Moreno y E. d. C. N. Romero, “La gestión del riesgo integrada con la minería de procesos y la administración de procesos de negocios bpm: Una revisión de la literatura,” en la Nueva Era, p. 104, 2022.

[42] R. D. C. Mendoza, "Predicción del rendimiento académico utilizando modelos de aprendizaje automático: Una revisión sistemática de la literatura," 593 Digital Publisher CEIT, vol. 9, no. 6, pp. 1038-1054, 2024.

[43] C. IBM, “Guía de crisp-dm de ibm spss modeler: Despliegue,” August 2021, Última actualización: 2021-08-17. Accedido el 09-12-2024. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/ saas?topic=deployment-overview.

[44] M. Kuroki, “Using python and google colab to teach undergraduate microeconomic theory,” International Review of Economics Education, vol. 38, p. 100225, 2021.

[45] T. A. Meyer, C. Ramirez, M. J. Tamasi, and A. J. Gormley, “A user’s guide to machine learning for polymeric biomaterials,” ACS Polymers Au, vol. 3, no. 2, pp. 141–157, 2022.

[46] V. Jain, A. K. Saxena, A. Senthil, A. Jain, and A. Jain, “Cyber-bullying detection in social media platform using machine learning,” in 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART). IEEE, 2021, pp. 401–405.

[47] A. Makarov and D. Namiot, “Overview of data cleaning methods for machine learning,” International Journal of Open Information Technologies, vol. 11, no. 10, pp. 70–78, 2023.

[48] G. G. Diago, “Perspectivas para abordar la inteligencia artificial en la enseñanza de periodismo. Una revisión de experiencias investigadoras y docentes,” Revista Latina de Comunicación Social, no. 80, p. 26, 2022.

[49] J. Texier and J. Zambrano, “La relación entre el curriculum dl y las ciencias de la computación: Una revisión bibliográfica,” Interciencia, vol. 45, no. 2, pp. 68–75, 2020.

[50] F. S. Molina, D. P. Santamaría, A. G. Bernabéu, and N. U. González, “Una revisión de experiencias y recursos educativos para aprender economía y finanzas con python,” in In-Red 2023. IX Congreso Nacional de Innovación Educativa y Docencia en Red: Modelos transformadores docentes para un aprendizaje a lo largo de la vida. Universidad Politécnica de Valencia= Universitat Politècnica de València, 2023, pp. 308–322.

[51] F. Labora Gomez, “Desarrollo de un modelo de unit-commitment en python-pyomo para su ejecución mediante interfaz-web.” 2023.

[52] C. Hill, L. Du, M. Johnson, and B. McCullough, “Comparing programming languages for data analytics: Accuracy of estimation in python and r,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p. e1531, 2024.

[53] V. Gómez-Saavedra, G. Morales-Canelo, M. Ves-González, y J. Hernández-Espinoza, "Análisis y visualización de datos de infección por SARS-CoV-2 de la población atendida en los establecimientos de salud de atención primaria de la Corporación Municipal de Viña del Mar, durante los años 2021-2022," ARS médica (Santiago), vol. 48, no. 3, pp. 12–22, 2023.

[54] K. Ortiz Falcon, “Optimización bayesiana de hiperparámetros,” B.S. thesis, 2023.

[55] O. Ruiz Sarrias, “Curvas de aprendizaje en la optimización bayesiana de hiperparametros,” 2021.

[56] X. Song, W. Wei, J. Zhou, G. Ji, G. Hussain, M. Xiao, and G. Geng, “Bayesian-optimized hybrid kernel svm for rolling bearing fault diagnosis,” Sensors, vol. 23, no. 11, p. 5137, 2023.

[57] M. Amrullah and A. Yuniarti, “Effective coronary artery disease prediction using bayesian optimization algorithm and random forest,” Building of Informatics, Technology and Science (BITS), vol. 6, no. 2, pp. 785–796, Sep. 2024. [Online]. Available: https://ejurnal.seminar-id. com/index.php/bits/article/view/5554.

[58] S. Albahli, “Efficient hyperparameter tuning for predicting student performance with bayesian optimization,” Multimedia tools and applications, vol. 83, no. 17, pp. 52 711–52 735, 2024.

[59] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, “Bayesian optimization with support vector machine model for parkinson disease classification,” Sensors, vol. 23, no. 4, p. 2085, 2023.

[60] T. Zhang, J. Su, Z. Xu, Y. Luo, and J. Li, “Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier,” Applied Sciences, vol. 11, no. 2, p. 543, 2021.

[61] S. Asante-Okyere, C. Shen, and H. Osei, “Enhanced machine learning tree classifiers for lithology identification using bayesian optimization,” Applied Computing and Geosciences, vol. 16, p. 100100, 2022.

[62] Y. Torres-Quezada, “Minería de datos para determinar los factores más influyentes en la ocurrencia de siniestros de tránsito en ecuador en el año 2020,” CEDAMAZ, vol. 11, no. 2, pp. 124–132, 2021.

[63] P. Benítez and E. Coronel, “Minería de datos en la accidentabilidad vehicular en la zona urbana del cantón Loja,” Loja, Ecuador, 2023.

[64] Comité Nacional de Límites Internos, “Elaboración: Secretaría nacional de planificación y desarrollo zona 8,” En línea, 2018, accedido: 18-10-2024. [Online]. Available: https://www.planificacion.gob.ec/wp-content/uploads/downloads/2019/06/Agenda-Coordinaci%C3%B3n-Zonal-Z8-2017-2021.pdf.

11. Anexos

Anexo I. Solicitud emitida al Subteniente de la Policía Nacional del Ecuador.

UNIVERSIDAD NACIONAL DE LOJA
ÁREA DE LAS ENERGÍAS, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES
INGENIERÍA EN COMPUTACIÓN

Estimado:

Sebastian Alejandro Encalada

Subteniente de la Policia Nacional del Ecuador.

Mi nombre es Cecilia Fernanda Trueba Reyes soy estudiante de la carrera de Ingeniería en Computación. Actualmente, estoy desarrollando mi proyecto de integración curricular como requisito para mi graduación. El objetivo de mi trabajo es aplicar **técnicas de minería de datos** para analizar la base de datos de **asesinatos en la zona 8 del Ecuador**, la misma que ya me otorgo la Policía Nacional del Ecuador y poder descubrir **patrones de comportamiento** que puedan ser útiles en la planificación de estrategias de prevención de estos crímenes. A través de este análisis, buscó identificar características clave asociadas a los asesinatos en la zona y descubrir información que es difícil encontrar a simple vista debido a que existen miles de registros de estos crímenes en las bases de datos. Esta información resultante podría servir como un complemento para la labor de la Policía Nacional del Ecuador, ayudando a focalizar esfuerzos y optimizar los recursos destinados a reducir la violencia. Para llevar a cabo este estudio, es importante contar con su colaboración mediante la participación en una breve encuesta de preguntas abiertas que nos permitirá comprender mejor el contexto de la problemática, así como las necesidades de información de la Policía Nacional del Ecuador para abordar de manera más efectiva los asesinatos en la zona 8. Le agradezco de antemano su tiempo y apoyo en la recolección de datos, ya que su aporte es fundamental para el éxito de este proyecto. Adjunto, encontrará un espacio para registrar su firma como **encuestado**.

Nombre del Encuestador: Cecilia Fernanda Trueba Reyes
Firma del Encuestador:
Firmado electrónicamente por:
CECILIA FERNANDA TRUEBA REYES

Nombre del Subteniente (Encuestado):
Sebastian Alejandro Encalada
Firma del Subteniente (Encuestado):

Nombre: SEBASTIAN ALEJANDRO ENCALADA ESPINOZA
Emitido por: UANATACA CA2 2016

Fecha: 07/10/2024
Atentamente, Cecilia
Trueba Reyes
cecilia.trueba@unl.edu.ec

Anexo II. Encuesta realizada al Subteniente de la Policía Nacional del Ecuador para conocer la problemática de los asesinatos en la Zona 8 del Ecuador.

UNIVERSIDAD NACIONAL DE LOJA
FACULTAD DE LA ENERGÍAS, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES
INGENIERÍA EN COMPUTACIÓN

A. PLANIFICACIÓN DE LA ENCUESTA

- Fecha: 17/10/2024
- Encuestador: Tesista Cecilia Fernanda Trueba Reyes
- Encuestado: Sub. Sebastian Alejandro Encalada Espinoza

B. ENCUESTA DE PREGUNTAS ABIERTAS

1. **¿Qué cambios se han observado en la tendencia de los asesinatos en los últimos años?**

   El Ecuador en los últimos ha estado viviendo diversos cambios en torno a la violencia generada por los diferentes grupos de delincuencia organizada, siendo uno de los mayores problemas que vive actualmente el país que es el narcotráfico, donde ciertas bandas como los choneros y lobos están aliados con narcos colombianos, mexicanos, albaneses, entre otros, quienes en su lucha por tener el control de ciertos territorios dentro de nuestro país su grado de violencia ha ido aumentando existiendo hoy en día desmembramientos de personas, atentados con coches bombas, secuestros, ataques con explosivos a locales comerciales, asesinatos y más.

2. **¿Cuál es la percepción de seguridad en la zona 8 según la comunidad?**

   Actualmente en Guayaquil las personas su nivel de percepción en torno a lo que es seguridad, podemos decir que ha ido incrementando gracias a los diversas acciones y operaciones que realiza la Policía Nacional en coordinación con las autoridades locales y la comunidad, generando en las personas seguridad y confianza para salir a las calles, sin embargo no podemos ignorar que es un trabajo complejo y que necesita tiempo para poder seguir creando en la comunidad un ambiente seguro y confiable para ellos, debido a todos los acontecimientos que vive el país.

3. **¿Cree que los patrones de comportamiento encontrados a través de técnicas estadísticas de análisis de grandes volúmenes de datos, como la minería de datos, serían útiles para tomar decisiones estratégicas en la prevención de asesinatos?**

   Es muy importante siempre tener una estadística de datos ya que nos permita medir el nivel de productividad que se tiene y poder comparar con los datos anteriores, permitiéndonos tomar decisiones en pro de la seguridad de cada territorio y de esta manera enfocarnos en el principal problema de acuerdo a los datos que nos arroje el sistema.

4. **¿De qué manera la identificación de patrones de comportamiento mediante análisis de datos podría ayudar a planificar intervenciones más efectivas en seguridad?**
   Estos datos nos permiten medir aquellos delitos que se dan con mayor frecuencia en torno a lo que sucede cada semana, es así que mediante el departamento de análisis

de información del delito (DAID) con el departamento de operaciones de cada distrito fija estrategias, operaciones, acciones preventivas y disuasivas en los puntos mas críticos donde se dan estos hechos delictivos con el fin de ir disminuyendo estos delitos y obtener mejores resultados.

5. **¿Qué estrategias y programas están actualmente en marcha para reducir los asesinatos en la zona 8?**
   Existen diversas operaciones que se están dando para obtener resultados positivos e ir disminuyendo el índice delictivo en esta zona, como es el operativo Relámpago que se da en Durán cuyo fin es el de realizar operativos en los lugares más críticos de este cantón, cuyos datos nos brinda el personal del departamento de análisis de información del delito (DAID), también existe otros operativos que se realiza como son el operativo APC, operativo focalizados, entre otros, encargados de realizar registros a personas y vehículos.

6. **¿Qué se podría mejorar en las intervenciones actuales utilizando un enfoque basado en análisis de datos?**

   Es importante mencionar que la realidad que se vive en el territorio muchas veces no coincide con la base de datos, es por eso que se podría mejorar notoriamente si se trabajara conjuntamente con el personal preventivo ya que ellos son los primeros en llegar al lugar de los hechos y conocen más de cerca la problemática que existe en el territorio.

7. **¿Qué tipo de información o hallazgos específicos del análisis facilitarían la toma de decisiones operativas en el día a día y de manera particular en los asesinatos de la Zona 8 del Ecuador?**

   La información que se necesita para tomar decisiones y obtener resultados positivos podría ser las ubicaciones donde operan estás mafias, esto nos ayudaría a optimizar recursos, de igual manera los puntos críticos donde se dan la mayor parte de los asesinatos, puntos donde existe la venta y distribución de sustancias sujetas a fiscalización, esto nos permitiría actuar de manera preventiva y disuasiva.

8. **¿Cómo se coordina la Policía Nacional con otras instituciones para reducir los asesinatos?**

   Actualmente la Policía Nacional está trabajando conjuntamente con fuerzas armadas para poder tomar el control de los territorios donde las mafias han tomado posesión, de igual manera la Policía Nacional trabaja con las autoridades locales que en este caso puede ser el Municipio, GAD`s, Intendente, Comisaria, Personal de Salud, Agentes Municipales y Agentes de Tránsito cuya función es la de clausurar locales clandestinos que no cuentan con permisos y se prestan a la venta y distribución de sustancias sujetas a fiscalización donde además se dan los hechos delictivos, de igual manera asegurarse que las personas cuenten con los permisos de conducir y que no deban nada a la justicia y por último en la de verificar autos cuya numeración alfanumérica se encuentre alterada. Trabajar en conjunto nos permite obtener mejores resultados y poder cada día dar la seguridad y paz que necesita el territorio ecuatoriano.

9. **¿Cómo ha afectado el aumento de asesinatos la calidad de vida de los ecuatorianos?**

   Claramente las personas han tenido que abandonar sus hogares por qué las mafias se han apoderado de sus hogares, personas que han tenido que cerrar sus locales de trabajo debido a las vacunas que actualmente existe en el país, personas que por temor se quedan en casa para no ser víctimas de robo, entre otras, sin duda los acontecimientos que ha vivido el país afectado el estilo de vida de todos los ecuatorianos.

Firmado electrónicamente por: CECILIA FERNANDA TRUEBA REYES

Nombre: SEBASTIAN ALEJANDRO ENCALADA ESPINOZA
Emitido por: UANATACA CA2 2016

.....

Firma de Encuestador

Cecilia Trueba Reyes

ESTUDIANTE DE LA CARRERA DE COMPUTACIÓN

.....

Firma del Encuestado

Sebastian Encalada Espinoza

SUBTENIENTE DE LA POLICÍA NACIONAL DEL ECUADOR

Anexo III. Transcripción de la entrevista realizada a la Ingeniera Genoveva Suing.

UNIVERSIDAD NACIONAL DE LOJA
Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables
Carrera de Computación

Carrera de Ingeniería en Sistemas / Computación

Entrevista

Fecha: 17/06/2024

Problemática: Optimización Bayesiana en modelos de clasificación de Minería de Datos para determinar patrones en los asesinatos de la Zona 8 del Ecuador

Presentación (5 minutos):

La presente entrevista tiene como objetivo recolectar información y profundizar sobre las técnicas de optimización, en este caso la optimización Bayesiana, que se usará para mejorar el rendimiento de los modelos, al momento de evaluar con la métrica de precisión los modelos de Minería de Datos: Árbol de Decisión y SVM, para determinar los patrones en los asesinatos de la Zona 8 del Ecuador. Las respuestas de la Ingeniera en Computación, ayudarán a solventar cualquier duda, ya que estarán profundizando sobre el tema en cuestión, mejorando el entendimiento y análisis por parte del estudiante. Agradeciendo de antemano su colaboración y predisposición para brindar sus conocimientos en este campo de la Minería de Datos.

Actores:

- **Datos Personales del Entrevistador:**

Nombres: Cecilia Fernanda Trueba Reyes

Cargo: Estudiante de la Carrera de Ingeniería en Ciencias de la Computación

Correo electrónico: cecilia.trueba@unl.edu.ec

- **Datos Personales del Entrevistado:**

Nombres: Ing. Genoveva Suing Albito

Cargo: Docente de la materia de Machine Learning

Objetivos:

- Obtener conocimientos sobre la optimización Bayesiana que ayudaran a mejorar el rendimiento de los modelos de minería de datos.
- Obtener información valiosa que profundicen sobre la importancia de las métricas de rendimiento al momento de evaluar los modelos de Minería de Datos.
- Conocer las herramientas, métodos y técnicas de preferencia a usar, para realizar el análisis, entrenamiento, implementación y evaluación de los modelos de minería de datos, aplicadas a los datos de los homicidios de la zona 8 del Ecuador para determinar patrones.

Preguntas y Respuestas de la Entrevista:

1. **¿Por qué es importante la métrica de rendimiento específicamente la precisión para evaluar los modelos de minería de datos?**

   Al poder considerar algoritmos relacionados dentro de la rama a nivel general Machine Learning, existen algunas métricas, una de ellas parte importante y decisiva para poder saber en cuanto al rendimiento, es la métrica de precisión, pero siempre es aconsejable no únicamente considerar una métrica, sino hacer una comparativa, poder ver algunas métricas adicionales, para que en base a ello poder considerar la evaluación del rendimiento del modelo, que usted quiere involucrar en este caso relacionados con la minería de datos. Sí, considero que la precisión es muy importante, pero adicional podríamos complementarlas con evaluación de otras métricas también.

2. **¿Por qué es importante aplicar la optimización Bayesiana en el ajuste de parámetros de los modelos de minería de datos: Árbol de Decisión y SVM?**

   La optimización en todos los modelos minería de datos es un factor muy importante, ya que conlleva a la búsqueda de parámetros que mejor se ajusten a la exploración y explotación de datos, con la finalidad de llegar a obtener un modelo eficiente y efectivo, que logre el mejor rendimiento posible minimizando el tiempo y recursos computacionales necesarios.

3. **¿Qué ventajas presenta el uso de optimización Bayesiana en comparación con otras técnicas de optimización?"**

   Generalmente cuando se involucra la optimización Bayesiana, presenta ventajas que permiten tener ese criterio de optimización, ya que nos permite tener técnicas de optimización más ajustables al proyecto que usted está involucrado, es un algoritmo que por lo general sí está vinculado para hacer esta minería de datos.

4. **¿Qué beneficios trae la técnica de optimización en los modelos clasificadores de minería de datos?**

   Bueno, una de las consideraciones que como ya lo hemos visto, igualmente relacionados con modelos de Machine Learning, involucra el poder tener argumentos para poder hacer esa clasificación, hoy en día, los porcentajes a los cuales se puede llegar con algoritmos de Machine Learning, pueden estar superando un 90%, en el mismo escenario, se puede considerar la optimización Bayesiana, no únicamente quedarse con un solo algoritmo de configuración normal, siempre tratando de evidenciar o trabajar con esos hiperparámetros que hacen referencia para poder mejorar cada vez el rendimiento, por ende poder mejorar lo que involucra tener un algoritmo relacionado con optimización, existen algunas técnicas que hemos visto, relacionadas con descenso de gradiente, que nos permite de igual manera poder disminuir el error, siempre tratar de buscar esa representación del error mínimo, lo más aproximada posible a una solución analítica pero con porcentajes de error que sean lo menos posibles.

5. **¿Recomienda usar Google Colab en la nube o Python de manera local para realizar el análisis, procesamiento e implementación de los modelos?**

   El uso de herramientas para la implementación de modelos de minería de datos para su trabajo de titulación, se puede establecerse en base a un análisis comparativo de las ventajas y desventajas de varias herramientas existentes actualmente. Por otro lado, también se debe considerar el conocimiento que tenga en el uso de estas aplicaciones. Si es para estudio académico, tanto Google Colab en la nube o Python si son adecuadas, ya que proporcionan las librerías necesarias para la implementación de modelos de minería de datos.

6. **¿Qué aplicación o herramienta según su experiencia se podría usar para la limpieza de los datos?**

   Actualmente existen gran cantidad de herramientas para la limpieza de datos como OpenRefine, Trifacta, DataCleaner, pandas, R y Matlab, esta selección debe ser considera de acuerdo a un análisis comparativo entre ellas, tomando en cuenta la información disponible.

7. **¿Es más factible entrenar y validar los modelos programándolos en Python o simplemente usando la herramienta Weka?**

   La implementación mediante lenguaje de programación le facilita el manejo y control de arquitectura del modelo, para poder hacer los ajusten de hiperparámetros con la finalidad de obtener un modelo optimizado, mientras que las herramientas como Weka nos limitan en algunos cambios en la arquitectura.

8. **¿Por qué el porcentaje de las métricas de rendimiento de los modelos pueden ser muy bajos en algunos casos?**

   Generalmente podemos tener esos porcentajes bajos por algunos escenarios, los más comunes podrían ser, por falta de datos, es muy poca la información que se puede tener para poder generar una proyección, otra de las características pueden ser las configuraciones internas que podemos estar dando ya que no son adecuadas, si yo estoy modelando una regresión polinomial, a lo mejor yo visualizo que el escenario lo esté relacionando con alguna regresión cuadrática y realmente mi escenario está vinculado con una regresión polinomial de un mayor grado, por configuraciones internas, a lo mejor una particularidad gaussiana, entonces todas estas configuraciones permiten que nosotros mejoremos esos porcentajes, y uno de los factores es la configuración y lo que involucra los datos, las dos condiciones, que se pueden dar.

9. **¿Por qué es importante evaluar y comparar los modelos de Minería de Datos antes y después de aplicar la optimización Bayesiana?**

   Bueno a nivel general , no únicamente dentro de minería de datos, sino a nivel de todo algoritmo relacionado con el aprendizaje de máquina, podríamos considerar el poder crear un modelo, poderlo generar y tenerlo establecido, está correcto, pero siempre tiene que pasar por la validación, si está relacionado con el ámbito de la salud, tendría

que ir con el experto, poder probar, poder tener esos resultados, y ver si efectivamente tengo alguien quien lo valide, porque yo puedo como técnico en el área de Ingeniería puedo configurar a mi criterio pero generalmente trabajamos con requerimientos de usuario, así sea un modelo pequeño, para yo poderlo interpretar, para poder sacar y ver información, qué hiperparámetros puedo cambiar, cual es la predicción adecuada, entonces yo siempre tengo que hacer esa etapa de validación, el decir sí, esta correcto o no, de esta manera no únicamente quedarme con mi criterio técnico, porque recordemos nosotros creamos y diseñamos para clientes, finalmente validarlo con el experto o con el más cercano, para tener una validez de lo que he podido configurar.

10. **¿Por qué es importante determinar patrones en los asesinatos de la Zona 8 del Ecuador con modelos clasificadores de minería de datos?**

Realmente poder identificar este tipo de información, es algo que les permite a quienes estén vinculados, como en este caso la policía, el poder tener esa comparativa, poder visualizar cuales serían los factores, ya que esto es parte esencial de un análisis, el poder ver cuáles son esas características, esa información relevante, que les permita a ellos aplicar algunas correctivas y tener algunas alertas relacionadas con este tipo de investigaciones que usted está desarrollando y la optimización Bayesiana es uno de los algoritmos que se los ha podido vincular, ya que tiene sus características importantes para poder hacer este tipo de análisis.

GENOVEVA JACKELINNE SUING ALBITO

.....

Ing. Genoveva Suing Albito

Entrevistado

.....

Cecilia Trueba Reyes

Entrevistador

Anexo IV. Solicitud enviada por correo electrónico al Ministerio del Interior para obtener la base de datos de Homicidios Intencionales enero 2015- febrero 2024.

REPÚBLICA DEL ECUADOR

Ministerio del Interior

SOLICITUD DE ACCESO A LA INFORMACION PÚBLICA

Fecha:	19/03/2024
Ciudad:	Loja
Institución	MINISTERIO DEL INTERIOR
Autoridad:	Ministra del Interior (E), Dra. Mónica Palencia Núñez

IDENTIFICACIÓN DEL SOLICITANTE

Nombre: Cecilia Fernanda Apellido: Trueba Reyes

Cédula No. 1106087883

Dirección domiciliaria: Benjamín Pereira y Alfredo Mora

Teléfono (fijo o celular): 0960934663

PETICIÓN CONCRETA:

Pido muy comedidamente me envíen la Base de Datos de los Homicidios Intencionales de los siguientes años : 2015 Enero-Diciembre, 2016 Enero-Diciembre,2017 Enero-Diciembre,2018 Enero-Diciembre,2019 Enero-Diciembre,2020 Enero- Diciembre, 2021 Enero-Diciembre,2023 Enero-Diciembre,2024 Enero-Febrero, de antemano muchas gracias.

FORMA DE RECEPCIÓN DE LA INFORMACIÓN SOLICITADA:

Retiro de la información en la institución: ☐

Email: ceciliatrueba289928@outlook.com

FORMATO DE ENTREGA:

Copia en papel: ☐

Cd. ☐

Formato electrónico digital: ☒ PDF ☐

Word ☐

Excel ☒

Otros ☐

Anexo V. Certificación de obtención y acceso a base de datos de homicidios intencionales enero 2015 hasta febrero 2024.

REPÚBLICA DEL ECUADOR

Oficio Nro. MDI-VSC-SEES-2024-0045-OF

Quito, D.M., 22 de abril de 2024

Asunto: Respuesta a Cecilia Trueba Reyes mediante oficio No. MDI-CGAF-DA-2024-2331-EXT, remite solicitud de acceso a la información (Digital)

Señorita
Cecilia Fernanda Trueba Reyes
En su Despacho

De mi consideración:

Reciba un cordial saludo, en respuesta al su oficio No. MDI-CGAF-DA-2024-2331-EXT de 19 de marzo de 2024, mediante el cual se solicita información de Homicidios Intencionales, en cumplimiento a sumilla inserta por parte de autoridad me permito adjuntar en formato xls (Excel) la data de enero 2015 a febrero de 2024.

La información de Homicidios Intencionales de marzo 2024 se actualizará en el portal de Datos Abiertos Ecuador (https://www.datosabiertos.gob.ec/dataset/?organization=ministerio-del-interior) hasta el 26 de abril del presente año.

Con sentimientos de distinguida consideración.

Atentamente,

Documento firmado electrónicamente

Capt. David Estuardo Anrango Narváez
SUBSECRETARIO DE ESTUDIOS Y ESTADÍSTICA DE LA SEGURIDAD

Referencias:
- MDI-CGAF-DA-2024-2331-EXT

Anexos:
- homicidios_intencionales_pm_2015-2024_feb.rar

Copia:
Señorita Ingeniera
Yira Nataly Monge Viteri
Directora de Estadística y Economía de la Seguridad

Paulina Elizabeth Armendariz Pacheco
Asistente de Estadística y Economía de la Seguridad

Señor Economista
Alexander Javier Paredes Romero
Analista de Estadística y Economía de la Seguridad 1

ap/ym

DAVID ESTUARDO ANRANGO NARVAEZ

Anexo VI. Certificación de informe y dashboard entregados al Subteniente de la Policía Nacional del Ecuador.

Universidad Nacional de Loja

Carrera de Ingeniería en Sistemas / Computación

Loja, 17 de enero de 2025

Sebastian Encalada Espinoza

Subteniente de la Policía Nacional del Ecuador

De mis consideraciones:

Reciba un cordial saludo y a la vez deseándole toda clase de éxitos en las funciones a su cargo.
La presente tiene la finalidad de poner a su conocimiento y a la institución a la cual usted representa, los resultados del estudio que realicé en la Carrera de Ingeniería en Computación de la Universidad Nacional de Loja, titulado "Optimización Bayesiana en modelos de clasificación: Árbol de Decisión y Support Vector Machine para determinar mediante Minería de Datos patrones en los asesinatos de la Zona 8 del Ecuador", en el que se analizó la base de datos de Homicidios Intencionales del Ecuador periodo enero 2015 – febrero 2024. Los patrones de comportamiento determinados indican que el área urbana es el lugar con mayor cantidad de asesinatos, siendo la vía pública el sitio donde ocurren con mayor frecuencia, la principal motivación asociada a estos crímenes es la delincuencia común, los días con mayor incidencia de asesinatos son los sábados y domingos, en un rango horario comprendido entre las 19:00:00 pm y la 00:59:00 am, las armas de fuego son las más utilizadas en estos hechos, en cuanto a la ubicación geográfica, los distritos de Nueva Prosperina, Sur, Pascuales son los más afectados, en dónde la mayoría de las víctimas tiene antecedentes penales, siendo el sexo masculino el más afectado en un rango de edad comprendido entre los 20 y los 50 años.
Por esta razón pongo a su disposición estos resultados como una fuente de información, los mismos que se encuentran reflejados en el informe[1] estadístico y en el recurso web interactivo (dashboard[2]). Esperando que el presente informe sea de utilidad para la institución, le expreso mis sentimientos de gratitud.

Atentamente,

Cecilia Trueba Reyes.

FIRMAS DE RESPONSABILIDAD DE LO ACTUADO	
Realizado por: Cecilia Trueba Reyes **Estudiante de la Carrera de Computación**	Firmado electrónicamente por: CECILIA FERNANDA TRUEBA REYES ..... **Firma**
Aceptado por: Sebastian Encalada Espinoza **Subteniente de la Policía Nacional del Ecuador**	Nombre: SEBASTIAN ALEJANDRO ENCALADA ESPINOZA Emitido por: UANATACA CA2 2016 ..... **Firma**

[1] https://drive.google.com/file/d/1y-urC1JAfidRq2Sh1v0REG18JwCPu6-4/view?usp=sharing
[2] https://lookerstudio.google.com/reporting/83297a6d-05f2-4d29-ac6e-e3e3f99f35c7

Anexo VII. Informe final entregado al Subteniente de la Policía Nacional del Ecuador.

Facultad de Energía, las Industrias y los Recursos Naturales no Renovables

CARRERA DE INGENIERÍA EN COMPUTACIÓN

Informe de estadísticas sobre los patrones de comportamiento de los asesinatos ocurridos en la Zona 8 del Ecuador, periodo enero 2015 – febrero 2024.

Elaborado por:

- Cecilia Fernanda Trueba Reyes

Revisado y aprobado por:

- Ing. Genoveva Suing Albito, Mg.Sc.

LOJA - ECUADOR
2025

FIRMAS DE RESPONSABILIDAD DE LO ACTUADO	
Realizado por: Cecilia Trueba Reyes **Estudiante de la Carrera de Computación de la UNL**	 Firmado electrónicamente por: CECILIA FERNANDA TRUEBA REYES ..... **Firma**
Revisado por: Ing. Genoveva Suing **Docente de la Carrera de Computación de la UNL**	 Firmado electrónicamente por: GENOVEVA JACKELINNE SUING ALBITO ..... **Firma**

1. INTRODUCCIÓN

En la actualidad, el análisis y la interpretación de grandes volúmenes de datos han adquirido una relevancia sin precedentes en diversos campos, incluyendo la seguridad ciudadana. La minería de datos, como disciplina clave dentro del análisis de datos, permite descubrir patrones, tendencias y relaciones significativas en conjuntos de datos complejos y extensos. Estas capacidades son especialmente útiles en el ámbito de la criminología, ya que posibilitan identificar comportamientos recurrentes y generar información valiosa para la prevención y mitigación de delitos. En este contexto, el presente informe tiene como objetivo principal analizar los patrones relacionados con los asesinatos ocurridos en la Zona 8 del Ecuador, que abarca los cantones de Guayaquil, Durán y Samborondón, durante el periodo comprendido entre enero de 2015 y febrero de 2024.

El análisis de datos en el ámbito del crimen no solo proporciona una perspectiva clara sobre las dinámicas delictivas, sino que también permite la elaboración de estrategias efectivas que impacten positivamente en la seguridad de la ciudadanía. En este estudio, la minería de datos se utilizó como una herramienta fundamental para identificar patrones relacionados con variables críticas que inciden en los asesinatos. Estas variables fueron seleccionadas con base en su relevancia para entender los contextos y circunstancias delictivas, proporcionando un marco analítico que combina aspectos demográficos, geográficos, temporales y contextuales.

Como parte de este estudio, se desarrolló además un dashboard[1] que sirvió como herramienta de visualización para resumir y representar gráficamente los hallazgos obtenidos. Este dashboard permitió explorar los datos de manera más dinámica, proporcionando representaciones visuales claras y comprensibles que complementan los análisis estadísticos realizados.

2. DESARROLLO

Para el desarrollo del presente estudio se seleccionaron diez variables principales, cada una cuidadosamente escogida por su pertinencia para describir las dinámicas de los asesinatos en la Zona 8. Las variables consideradas son: Área del hecho, Antecedentes, Lugar, Sexo, Edad, Día, Distrito, Hora de la infracción, Arma, Presunta motivación. Estas variables, consideradas en conjunto, han permitido realizar un análisis profundo y estructurado que no solo identifica los patrones, sino que también arroja luz sobre los posibles factores de riesgo y las condiciones subyacentes que contribuyen a la ocurrencia de los asesinatos.

[1] https://lookerstudio.google.com/reporting/83297a6d-05f2-4d29-ac6e-e3e3f99f35c7

A continuación, se describirá cada una de las variables analizadas, destacando los hallazgos más relevantes y cómo estos aportan a la comprensión de los patrones de asesinatos en la Zona 8 del Ecuador.

2.1. VARIABLE ÁREA DEL HECHO

A continuación, en la **Tabla I** se presenta la clasificación del área del hecho, según los datos recolectados en Ecuador en el año periodo 2015 – febrero 2024.

Tabla I

CLASIFICACIÓN DEL ÁREA DEL HECHO DE LOS ASESINATOS

Área del Hecho	
Representa la limitación geográfica de un territorio de acuerdo con su ubicación urbana o rural [1].	
Categorías	**Descripción**
Urbano	Se entiende por paisaje o espacio urbano al paisaje interior de las ciudades, o sea, al espacio habitado, estructurado y organizado que compone los centros urbanos.
Rural	Se entiende por paisaje o espacios rurales al paisaje externo de las ciudades, o sea, al espacio que se ha mantenido en un estado rústico y es de índole productivo.

En la **Tabla II** se presentan los resultados con respecto al área del hecho más frecuente para la ocurrencia de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla II

ÁREA DEL HECHO DE ASESINATOS EN LA ZONA 8.

N°	**Área del Hecho**	**Probabilidad de Ocurrencia**
1	Urbano	57,9%
2	Rural	42,1%

La **Fig. 1** muestra la distribución de asesinatos según el área del hecho, la cual está representada por dos categorías: Urbano y Rural, la primera tiene un 57,9% del total, mientras que la segunda tiene un 42,1%, por lo tanto, se deduce que la mayoría de los asesinatos ocurren en áreas urbanas de la Zona 8 del Ecuador.

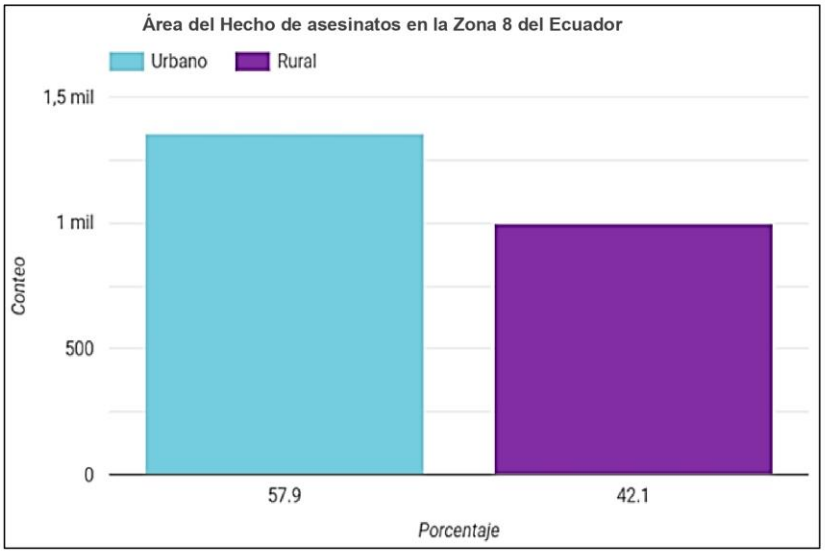


Fig. 1 Distribución del área del hecho de los asesinatos en la Zona 8 del Ecuador.

2.2. VARIABLE LUGAR

A continuación, en la **Tabla III** se presenta la clasificación de los lugares donde ocurren los asesinatos, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla III

CLASIFICACIÓN DEL LUGAR DE LOS ASESINATOS

Lugar	
Área o espacio donde se ha desarrolla una conducta antijurídica y culpable que ha derivado en la muerte de una persona por causas estructurales de violencia [1].	
Categorías	**Descripción**
Vía Pública	Una vía pública es cualquier espacio de dominio común por donde transitan los peatones o circulan los vehículos [2].
Lugares Privados	Aquel espacio en el cual una persona o grupo de personas puede establecer una regulación consciente y efectiva de su interacción social con los demás [3].

En la **Tabla IV** se presentan los resultados con respecto al lugar más frecuente para la ocurrencia de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla IV

LUGAR DE OCURRENCIA DE ASESINATOS EN LA ZONA 8.

N°	Lugar	Probabilidad de Ocurrencia
1	Vía Pública	60,2 %
2	Lugares Privados	39,8 %

La **Fig. 2** muestra la distribución de asesinatos según el lugar donde ocurren los asesinatos, la cual está representada por dos categorías: Vía Pública y Lugares Privados, la primera tuvo 60,2% del total, mientras que segunda un 39,8%, evidenciando que la mayoría de los crímenes ocurren en la vía pública.

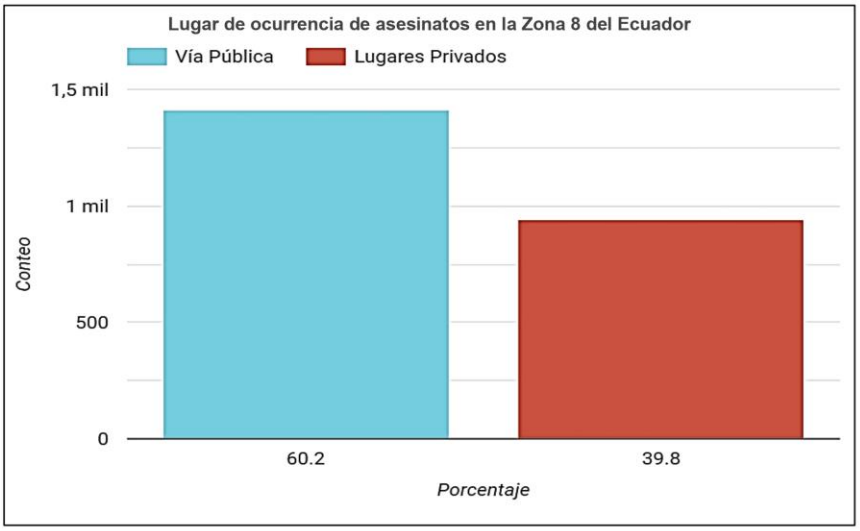


Fig. 2 Distribución del lugar de la ocurrencia de asesinatos en la Zona 8 del Ecuador.

2.3. VARIABLE ANTECEDENTES

A continuación, en la **Tabla V** se presenta la clasificación de los antecedentes de las víctimas de asesinatos, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla V

CLASIFICACIÓN DEL ANTECEDENTES DE LAS VÍCTIMAS DE ASESINATO

Antecedentes	
Historial de crímenes de un individuo [1].	
Categorías	**Descripción**
Si	La víctima de asesinato si presenta un historial de crímenes cometidos con anterioridad.
No	La víctima de asesinato no presenta un historial de crímenes cometidos con anterioridad.

En la **Tabla VI** se presentan los resultados con respecto a los antecedentes de las víctimas de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla VI

ANTECEDENTES DE VÍCTIMAS DE ASESINATOS EN LA ZONA 8.

N°	Antecedentes	Probabilidad de Ocurrencia
1	Si	58,7 %
2	No	41,3 %

La **Fig. 3** muestra la distribución de asesinatos según los antecedentes que presentan las víctimas, la cual está representada por dos categorías: Si y No, la primera tuvo 58,7% del total, mientras que segunda un 41,3%, evidenciando que la mayoría que los asesinatos se asocian mayormente a individuos que si presentan antecedentes.

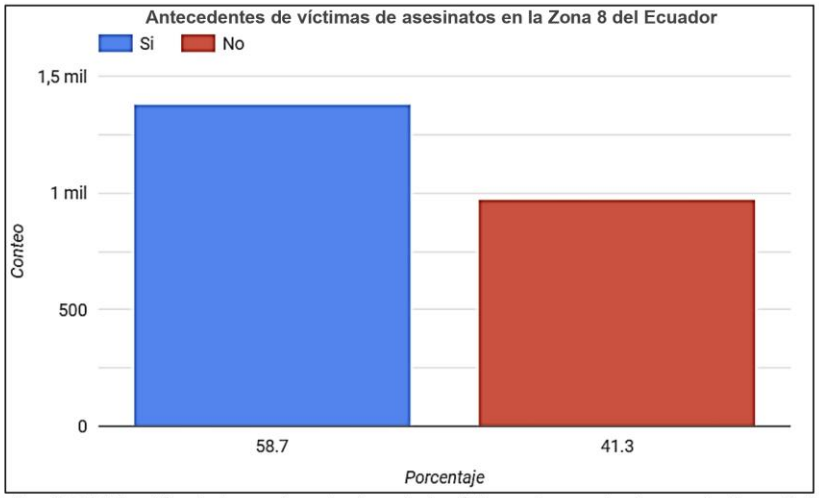


Fig. 3 Distribución de los antecedentes de la víctima de asesinatos en la Zona 8 del Ecuador.

2.4. VARIABLE SEXO

A continuación, en la **Tabla VII** se presenta la clasificación del sexo de las víctimas de asesinatos, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla VII

CLASIFICACIÓN DEL SEXO DE LAS VÍCTIMAS DE ASESINATO

Sexo	
Es el conjunto de las peculiaridades que caracterizan los individuos de una especie dividiéndolos en hombres y mujeres [1].	
Categorías	**Descripción**
Masculino	Es una palabra que puede referirse, de manera general, al ser animado racional, sea varón, que forma parte de la especie humana.
Femenino	Es una palabra que puede referirse, de manera general, al ser animado racional, sea mujer que forma parte de la especie humana.

En la **Tabla VIII** se presentan los resultados con respecto al sexo de las víctimas de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla VIII

SEXO DE VÍCTIMAS DE ASESINATOS EN LA ZONA 8.

N°	**Sexo**	**Probabilidad de Ocurrencia**
1	Masculino	56,9%
2	Femenino	43,1%

La **Fig. 4** muestra la distribución de asesinatos según el sexo que presentan las víctimas, la cual está representada por dos categorías: "Masculino" y “Femenino", la primera tuvo 56,9% del total, mientras que la segunda un 43,1%, evidenciando que la mayoría de las víctimas de asesinatos son de sexo masculino.

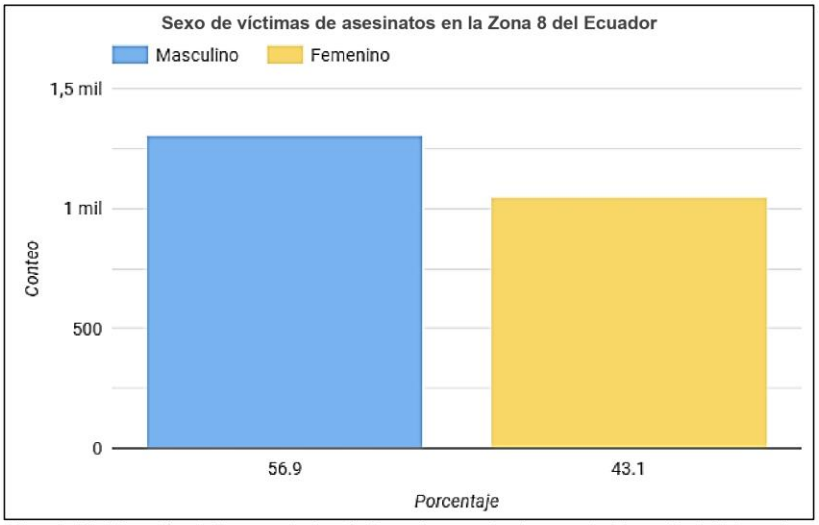


Fig. 4 Distribución del sexo de la víctima de asesinatos en la Zona 8 del Ecuador.

2.5. VARIABLE PRESUNTA MOTIVACIÓN

La **Fig. 5** obtenida de [4], clasifica los delitos de asesinato, homicidio, feminicidio y sicariato según su motivación y el tipo de violencia involucrado. Las motivaciones se agrupan en categorías como violencia comunitaria, intrafamiliar y sexual, así como sicopatologías, delincuencia común, transnacional y terrorismo. Cada categoría aborda causas específicas, desde conflictos emocionales, linchamientos y disputas familiares hasta delitos transnacionales como tráfico de personas y drogas. Asimismo, se distingue el tipo de violencia asociado: interpersonal, criminal o sociopolítica, proporcionando un marco claro para analizar las dinámicas detrás de estos actos.

Motivación			
TIPO DELITO/MUERTE	Motivacion (Porque?)	Observacion Motivacion	Tipo Violencia
ASESINATO/HOMICIDIO/FEMICIDIO/SICARITO	VIOLENCIA COMUNITARIA	ACTOS DE ODIO	VIOLENCIA INTERPERSONAL
		DEUDAS	
		EMOCIONAL	
		LINCHAMIENTO	
		LITIGIO DE TIERRAS	
		RIÑAS	
	VIOLENCIA INTRAFAMILIAR	SENTIMENTAL	
		MALTRATO	
		LITIGIO DE BIENES	
	VIOLENCIA SEXUAL	PERSONA CONOCIDA	
		VIOLACION FAMILIAR DIRECTO	
	SICOPATOLOGIAS	TRASTORNOS MENTALES	
	TRANSNACIONAL	CONTRABANDO	VIOLENCIA CRIMINAL
		SECUESTRO	
		TRAFICO DE ARMAS,MUNICIÓN Y EXPLOSIVOS	
		TRAFICOS DE MIGRANTES	
		TRAFICO DE ORGANOS	
		TRAFICO INTERNACIONAL DE DROGA	
		TRATA DE PERSONAS	
	DELICUENCIA COMUN	ABIGEATO	
		AMENAZA	
		DEFENSA PROPIA	
		EVASIÓN DE LA JUSTICIA	
		PROYECTIL SIN ORIGEN	
		RECEPTACIÓN ILEGAL (CACHINERIA)	
		ROBO A DOMICILIOS	
		ROBO A ENTIDADES FINANCEIRAS	
		ROBO A PERSONAS	
		ROBO A UNIDADES ECONÓMICAS	
		ROBO DE BIENES PATRIMONIALES	
		ROBO DE CARROS	
		ROBO DE MOTOS	
		ROBO EN EJES VIALES O CARRETERAS	
		SECUESTRO EXPRESS	
		TRAFICO INTERNOS DE DROGAS (MICROTRAFICO)	
		VIOLACION SEXUAL (DESCONOCIDO)	
	TERRORISMO	TERRORISMO	VIOLENCIA SOCIOPOLITICA

Fig. 5 Motivación y observación del Presunto Delito/Muerte.

A continuación, en la **Tabla IX** se presenta la clasificación de la presunta motivación para la ocurrencia de asesinatos, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla IX

CLASIFICACIÓN DE LA PRESUNTA MOTIVACIÓN PARA LA OCURRENCIA DE ASESINATOS.

Presunta Motivación	
Aquella sugestión temporal que guía la conducta humana para el cometimiento de una acción criminal, derivando al deceso de una persona [4].	
Categorías	**Descripción**
Delincuencia común	Aquellos actos de la cual se deriven decesos cuya motivación sea originada por el antecedente o el cometimiento de un hecho punible, o por acción de personas que buscando escapar de la justicia provoquen decesos de personas (civiles, agentes de seguridad, funcionarios públicos encargados de hacer cumplir la ley, entre otros).
Violencia comunitaria	Cuando la muerte de la víctima es originada por una persona ajena al núcleo familiar; es decir, por amigos o extraños.
Terrorismo	Cuando por la realización de actos de terrorismo se produzca la muerte de una o más personas.
Transnacional	Aquella conducta atípica, antijurídica de la cual se produzca la muerte tanto de aquellas personas víctimas de esta tipología del delito.
Psicopatologías	Cuando el victimario previo una valoración médica padezca un trastorno mental permanente o transitorio y cause la muerte de una persona (víctima), únicamente como producto de aquella condición.
Violencia intrafamiliar y violencia sexual	Cuando producto de la violencia intrafamiliar se produzca acciones que consistan en maltrato físico, psicológico o sexual, ejecutado por un miembro de la familia en contra de la mujer o demás integrantes del núcleo familiar y de tales acciones se deriven muertes. Es violencia sexual cuando el impulso de cometer un acto criminal que atente contra la sexualidad de la víctima, se produzca la muerte.

En la **Tabla X** se presentan los resultados con respecto a la presunta motivación para la ocurrencia de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla X

PRESUNTA MOTIVACIÓN PARA LA OCURRENCIA DE ASESINATOS EN LA ZONA 8.

N°	Presunta Motivación	Probabilidad de Ocurrencia
1	Delincuencia común	55,6%
2	Violencia comunitaria	10,2%
3	Terrorismo	9,9%
4	Transnacional	8,5%
5	Psicopatologías	8,2%
6	Violencia intrafamiliar y sexual	7,6%

La **Fig. 6** muestra la distribución de asesinatos según la presunta motivación, divididos en seis categorías: la primera de las categorías es "Delincuencia común" que tuvo un 55,6% del total, mientras que la categoría "Violencia comunitaria" alcanzó un 10,2%, la categoría "Terrorismo" presentó un 9,9%, la categoría "Transnacional" presentó un 8,5%, la categoría "Psicopatologías" tuvo un 8,2% y la categoría "Violencia intrafamiliar y sexual" tuvo un 7,6%. Esta representación gráfica permite observar que la mayoría de los asesinatos están motivados por la delincuencia común.

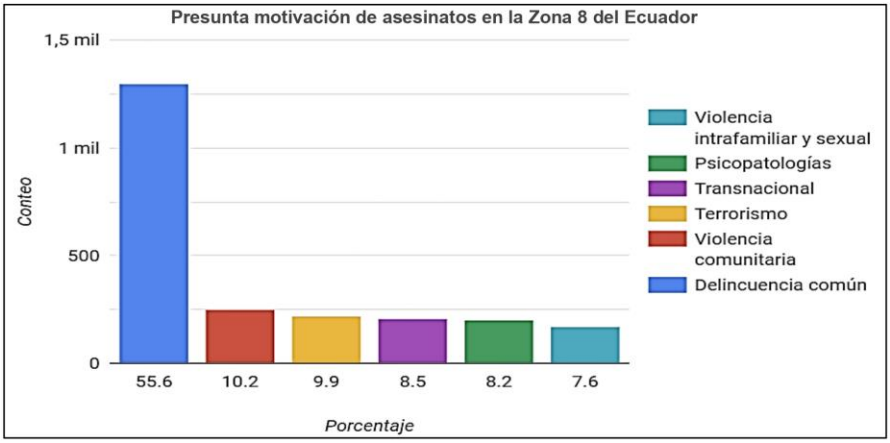


Fig. 6 Distribución de la presunta motivación de los asesinatos ocurridos en la Zona 8 del Ecuador.

2.6. VARIABLE ARMA

A continuación, en la **Tabla XI** se presenta la clasificación del arma usada en los asesinatos, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla XI

CLASIFICACIÓN DEL ARMA USADA EN LOS ASESINATOS.

Arma	
Medio que utiliza una persona para conseguir un fin determinado, especialmente el que utiliza para atacar a otras personas o defenderse de ellas [4].	
Categorías	**Descripción**
Arma de fuego	Es un dispositivo destinado a propulsar uno o múltiples proyectiles mediante la presión generada por la combustión.
Arma blanca	Es aquella arma o herramienta que se caracteriza por su capacidad de cortar, herir o punzar mediante bordes afilados o puntiagudos.
Arma constrictora	Medio para causar la inducción de dolor, sumisión, o muerte a través de la fuerza ejercida por un objeto que facilita la acción mecánica de apretar, siendo las más usuales: soga, cable, prenda textil, correa, cuerda.
Sustancias y otros	Una sustancia es una especie de materia homogénea de composición química definida, en este caso han sido empleadas con mayor frecuencia, como medios para neutralizar la voluntad de las víctimas para el cometimiento de delitos y en otros casos siendo aquella sustancia que de forma voluntaria o accidental (sobredosis) ha causado el deceso de un individuo. Dentro de la categoría de otros se encuentran los siguientes: explosivos, funda, cinta de embalaje, electricidad, hidrocarburos (gasolina, aceite), ácidos, armas de aire comprimido.
Arma contundente	Instrumento y acto que producen contusión. Así suele llamarse el arma o instrumento destinado a obrar por contusión o golpe: tales son, por ejemplo, entre las primeras el palo, la clava, la maza, el azote, entre otros.

En la **Tabla XII** se presentan los resultados con respecto al arma usada en los asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla XII

ARMA USADA EN LOS ASESINATOS DE LA ZONA 8.

N°	Arma	Probabilidad de Ocurrencia
1	Arma de fuego	58,4%
2	Arma blanca	13,3%
3	Arma constrictora	10,6%
4	Sustancias y otros	10,2%
5	Arma contundente	7,6%

La **Fig. 7** muestra la distribución de asesinatos según el tipo de arma utilizada, divididos en cinco categorías: la primera es la categoría "Arma de fuego" que alcanzó un 58,4% del total, la categoría "Arma blanca" presentó un 13,3% del total, la categoría "Arma constrictora" tuvo un 10,6%, mientras que la categoría "Sustancias y otros" presentó un 10,2% y la categoría "Arma contundente" llegó a un 7,6%. Esto evidenció que la mayoría de los asesinatos se cometen con armas de fuego, seguido por el arma blanca.

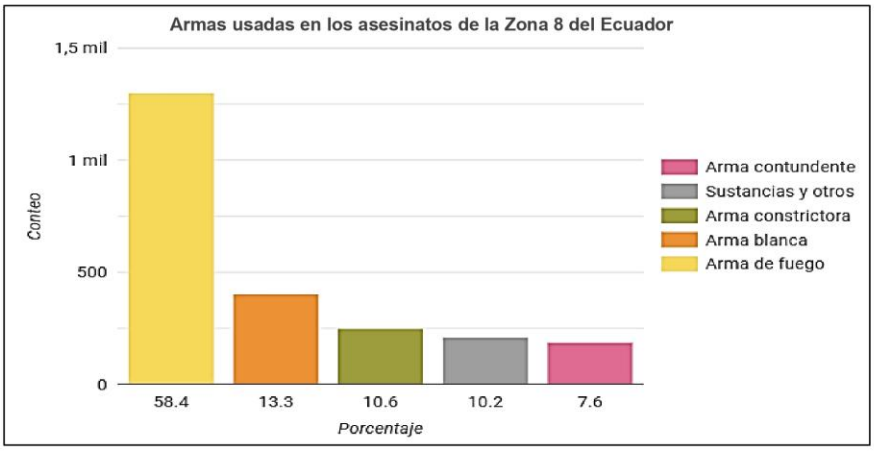


Fig. 7 Distribución del arma usada en los asesinatos ocurridos en la Zona 8 del Ecuador.

2.7. VARIABLE DÍA

A continuación, en la **Tabla XIII** se presenta la clasificación del día que ocurren los asesinatos, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla XIII

CLASIFICACIÓN DEL DÍA EN LA OCURRENCIA DE LOS ASESINATOS.

Día	
El día en el cual se suscita un hecho delictivo, en este caso se lo considerará con el día de la lesión fatal producido a la víctima [1].	
Categorías	**Descripción**
Lunes, Martes, Miércoles, Jueves, Viernes, Sábado, Domingo.	Representan los siete días de la semana, que se repiten continuamente.

En la **Tabla XIV** se presentan los resultados con respecto a los días de ocurrencia de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla XIV

DÍA EN LA OCURRENCIA DE ASESINATOS DE LA ZONA 8.

N°	Día	Probabilidad de Ocurrencia
1	Sábado, Domingo	51,3%
2	Martes, Viernes	20,1%
3	Miércoles	16,1%
4	Lunes, Jueves	12,5%

La **Fig. 8** muestra la distribución de asesinatos según el día en dónde más se cometen, divididos en cuatro categorías: la primera es la categoría "Sábado, Domingo" que alcanzó un 51,3% del total, mientras que la categoría "Martes, Viernes" presentó un 20,1%, %, la categoría "Miércoles" llegó a un 16,1% y la categoría "Lunes, Jueves" tuvo un 12,5%. Esto evidenció que la mayoría de los asesinatos se cometen los días sábado y domingo.

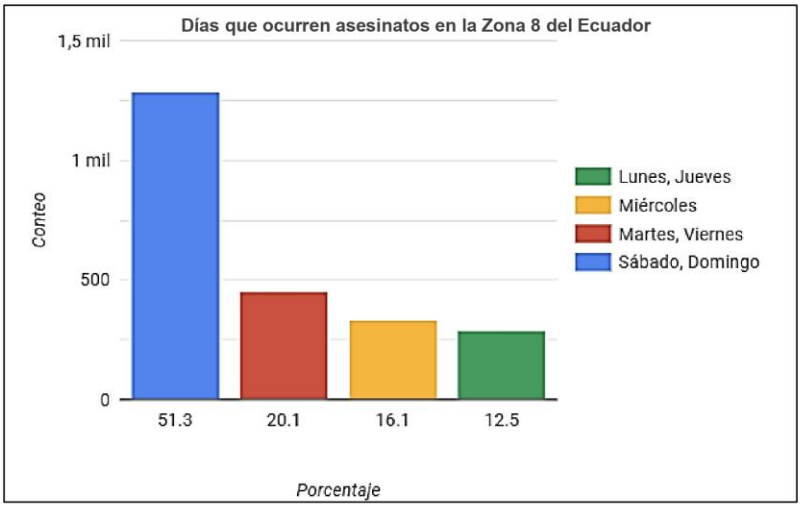


Fig. 8 Distribución de los días de la semana donde ocurren asesinatos en la Zona 8 del Ecuador.

2.8. VARIABLE EDAD

A continuación, en la **Tabla XV** se presenta la clasificación de la edad de las víctimas de asesinatos, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla XV

CLASIFICACIÓN DE EDAD DE LAS VÍCTIMAS DE ASESINATOS.

Edad	
Es el tiempo que ha vivido una persona u otro ser vivo contando desde su nacimiento, total de años que tiene un individuo [1].	
Categorías	**Descripción**
1 - 95	Representan la edad que posee una persona, edades que van desde 1 año hasta 95 años.

En la **Tabla XVI** se presentan los resultados con respecto a la edad de las víctimas de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla XVI

EDAD DE LAS VÍCTIMAS DE ASESINATOS DE LA ZONA 8.

N°	Edad (Rango)	Probabilidad de Ocurrencia
1	20 - 50	59,0 %
2	1 - 19	18,9 %
3	66 - 95	13,8 %
4	51 - 65	8,3 %

La **Fig. 9** muestra la distribución de asesinatos según la edad de la víctima, divididos en cuatro categorías: la primera es la categoría que va en un rango de "20 - 50 " esta alcanzó un 59,0% del total, mientras que la categoría "1 - 19 " presentó un 18,9%, la categoría "66 - 95" llegó a un 13,8% y la categoría "51 - 65" tuvo un 8,3%. Esto evidenció que la mayoría de las víctimas de asesinatos tienen edades que van desde los 20 hasta los 50 años.

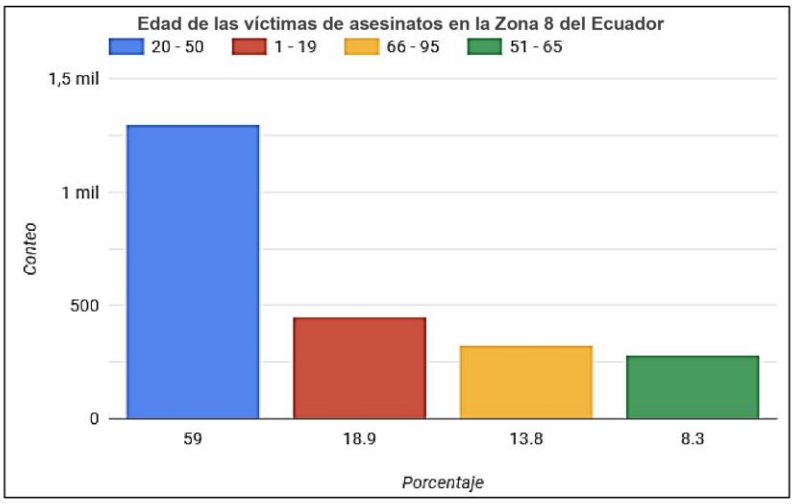


Fig. 9 Distribución del rango de edad de las víctimas de asesinatos en la Zona 8 del Ecuador.

2.9. VARIABLE HORA DE INFRACCIÓN

A continuación, en la **Tabla XVII** se presenta la clasificación de las horas de ocurrencia de asesinatos, según los datos recolectados en Ecuador.

Tabla XVII

CLASIFICACIÓN DE LAS HORAS DE OCURRENCIA DE ASESINATOS.

Horas	
Es una indicación del tiempo en que ocurre o se hace algo en hora y segundos [1].	
Categorías	**Descripción**
H00 - H23	Representan las horas en un rango, señalando el momento en donde se suscitó la lesión fatal.

A continuación, en la **Tabla XVIII** se muestra las horas de ocurrencia de asesinatos en la Zona 8, según los datos recolectados en Ecuador periodo enero 2015 – febrero 2024.

Tabla XVIII

HORAS DE OCURRENCIA DE ASESINATOS EN LA ZONA 8 DEL ECUADOR.

Hora	Identificador
00:00:00 A 00:59:00 AM	H00
01:00:00 A 01:59:00 AM	H01
02:00:00 A 02:59:00 AM	H02
03:00:00 A 03:59:00 AM	H03
04:00:00 A 04:59:00 AM	H04
05:00:00 A 05:59:00 AM	H05
06:00:00 A 06:59:00 AM	H06
07:00:00 A 07:59:00 AM	H07
08:00:00 A 08:59:00 AM	H08
09:00:00 A 09:59:00 AM	H09
10:00:00 A 10:59:00 AM	H10
11:00:00 A 11:59:00 AM	H11
12:00:00 A 12:59:00 PM	H12
13:00:00 A 13:59:00 PM	H13
14:00:00 A 14:59:00 PM	H14
15:00:00 A 15:59:00 PM	H15
16:00:00 A 16:59:00 PM	H16
17:00:00 A 17:59:00 PM	H17
18:00:00 A 18:59:00 PM	H18
19:00:00 A 19:59:00 PM	H19
20:00:00 A 20:59:00 PM	H20
21:00:00 A 21:59:00 PM	H21
22:00:00 A 22:59:00 PM	H22
23:00:00 A 23:59:00 PM	H23

En la **Tabla XIX** se presentan los resultados con respecto a las horas de ocurrencia de asesinatos, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla XIX

HORAS DE OCURRENCIA DE ASESINATOS DE LA ZONA 8.

N°	Horas (Rango)	Probabilidad de Ocurrencia
1	H19 - H00	51,3 %
2	H01 – H16	18,6 %
3	H07 – H12	16,9 %
4	H13 – H18	13,2 %

La **Fig. 10** muestra la distribución de asesinatos basándose en la hora que se cometen los asesinatos, divididos en cuatro categorías: la primera es la categoría que va en un rango horario de "H19 - H00" esta alcanzó un 51,3% del total, la categoría "H01 – H06" tuvo un 18,6%, mientras que la categoría "H07 - H12 " presentó un 16,9%, la categoría "H13 – H18" llegó a un 13,2%. Esto evidenció que la mayoría de los asesinatos se realizan en horas que van desde las 19:00:00 pm hasta las 00:59:00 am (medianoche).

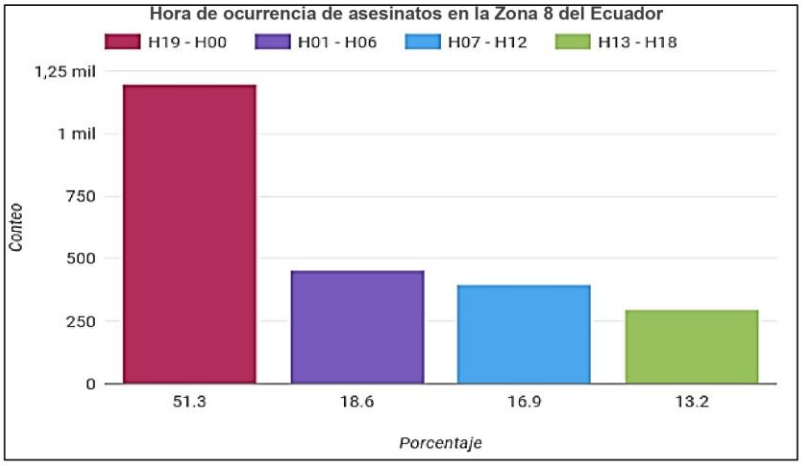


Fig. 10 Distribución del rango de horas más frecuentes para la ocurrencia de asesinatos en la Zona 8 del Ecuador.

2.10. VARIABLE DISTRITO

A continuación, en la **Tabla XX** se presenta la clasificación de los distritos donde ocurren asesinatos, según los datos recolectados en Ecuador.

Tabla XX

CLASIFICACIÓN DE LOS DISTRITOS CON MAYOR FRECUENCIA DE OCURRENCIA DE ASESINATOS.

Distritos	
Territorio de una provincia o ciudad que contiene diversos circuitos [1].	
Categorías	**Descripción**
Nueva Prosperina, Distrito Sur, Pascuales, Portete, 9 de Octubre, Durán, Estero, Progreso, Florida, Modelo, Ceibos, Samborondón.	12 distritos que corresponden a la Zona 8 (Guayaquil, Durán y Samborondón) del Ecuador.

En la **Tabla XXI** se presentan los resultados con respecto a los distritos donde ocurren con mayor frecuencia los asesinatos en la Zona 8 del Ecuador, organizadas de forma descendente según su porcentaje de probabilidad de incidencia.

Tabla XXI

DISTRITOS MÁS FRECUENTES EN LA OCURRENCIA DE ASESINATOS EN LA ZONA 8.

N°	**Horas (Rango)**	**Probabilidad de Ocurrencia**
1	Nueva Prosperina, Distrito Sur, Pascuales	58,8%
2	Portete, 9 de Octubre, Durán, Estero	22,2%
3	Progreso, Florida, Modelo, Ceibos, Samborondón	19,1%

La **Fig. 11** muestra la distribución de asesinatos basándose en el distrito, divididos en tres categorías: la primera es la categoría que consta de "Nueva Prosperina, Distrito Sur, Pascuales" esta alcanzó un 58,8% del total, mientras que la categoría "Portete, 9 de Octubre, Durán, Estero" presentó un 22,2%, la categoría " Progreso, Florida, Modelo, Ceibos, Samborondón " tuvo un 19,1%. Esto evidenció que la mayoría de los asesinatos se realizan en los distritos de Nueva Prosperina, Distrito Sur y Pascuales, de la Zona 8 del Ecuador.

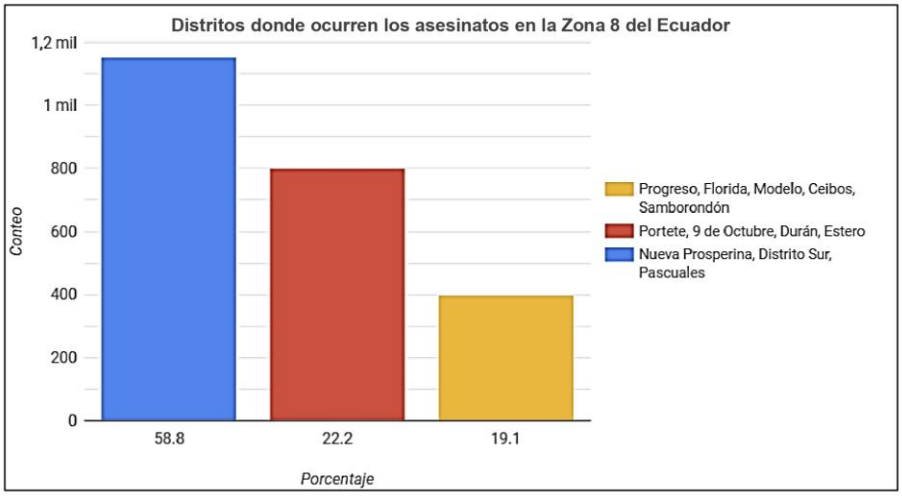


Fig. 11 Distribución de distritos con mayor frecuencia de ocurrencia de asesinatos en la Zona 8 del Ecuador.

3. CONCLUSIONES

De acuerdo con el presente estudio de minería de datos, se puede concluir lo siguiente:

- Los patrones de comportamiento identificados a través de la minería de datos indican que el área urbana es el lugar con mayor cantidad de asesinatos, siendo la vía pública el sitio donde ocurren con mayor frecuencia. La principal motivación asociada a estos crímenes es la delincuencia común.
- Los días con mayor incidencia de asesinatos son los fines de semana (sábado y domingo), en un rango horario comprendido entre las 19:00:00 pm y la 00:59:00 am (medianoche).
- Las armas de fuego son las más utilizadas en estos hechos. En cuanto a la ubicación geográfica, los distritos de Nueva Prosperina, Sur, Pascuales, todos parte de la zona 8, son los más afectados por estos crímenes.
- La mayoría de las víctimas tiene antecedentes penales, siendo el sexo masculino el más afectado por este tipo de delitos, especialmente en un rango de edad comprendido entre los 20 y los 50 años.

4. BIBLIOGRAFÍA

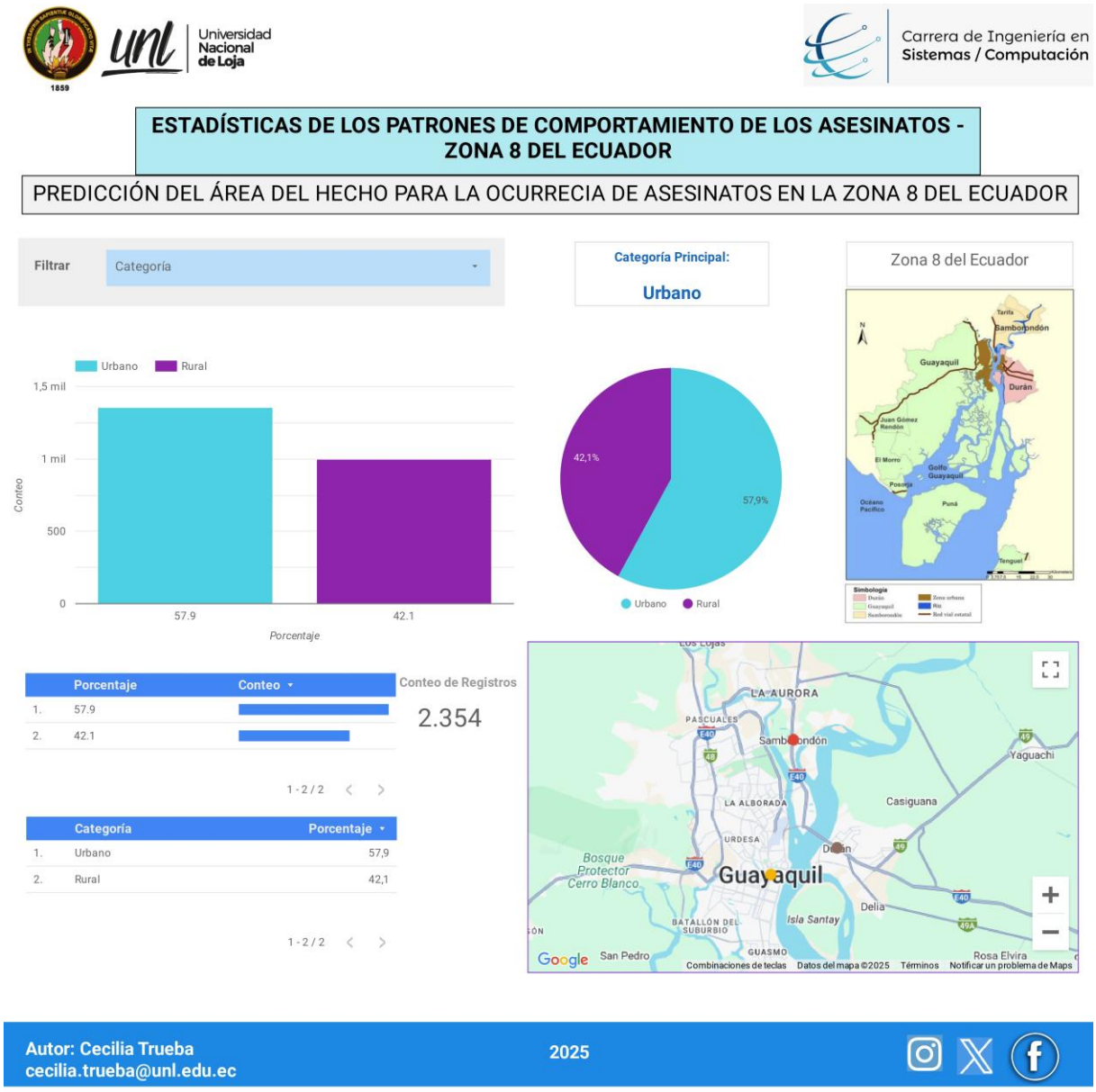
[1] Ministerio del Interior, "Estadísticas de seguridad", Subsecretaría de Estudios y Estadística de la Seguridad, Datos abiertos, [En línea]. Disponible en: http://181.113.21.13:8080/registroinicial-war/estadisticas.html. [Accedido: 29-dic-2024]. Correo: subsestudiosyestadisticas@ministeriodelinterior.gob.ec. Última actualización: 19-dic-2024, Fecha de creación: 17-ago-2021.

[2] Ministerio de Telecomunicaciones y de la Sociedad de la Información, "Definición de vía pública", República del Ecuador. [En línea]. Disponible en: https://www.gob.ec/tramites/buscar?search_api_fulltext=v%C3%ADa%20p%C3%BAblica#:~:text=Una%20v%C3%ADa%20p%C3%BAblica%20es%20cualquier,peatones%20o%20circulan%20los%20veh%C3%ADculos. [Accedido: 29-dic-2024].

[3] Universitat de Barcelona, "Introducción al concepto de privacidad - Espacios privados," http://www.ub.edu/psicologia_ambiental/unidad-3-tema-5-5-a#:~:text=En%20t%C3%A9rminos%20de%20privacidad%20se,interacci%C3%B3n%20social%20con%20los%20dem%C3%A1s. , accedido el 29 de diciembre de 2024.

[4] Ministerio del Interior, Oficina de Análisis de Información del Delito, Manual de Conceptualización de Muertes por Causas Externas, versión 1.3, enero 2015. [En línea]. Disponible en: https://anda.inec.gob.ec/anda/index.php/catalog/681/download/12480. [Accedido: 2-ene-2025]

Anexo VIII. Dashboard entregado al Subteniente de la Policía Nacional del Ecuador.



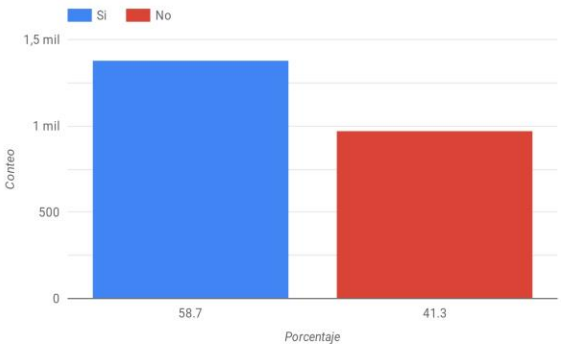
Universidad Nacional de Loja

Carrera de Ingeniería en Sistemas / Computación

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

PREDICCIÓN DE ANTECEDENTES DE LAS VÍCTIMAS DE ASESINATOS EN LA ZONA 8 DEL ECUADOR

Filtrar: Categoría



	Porcentaje	Conteo
1.	58.7	
2.	41.3	

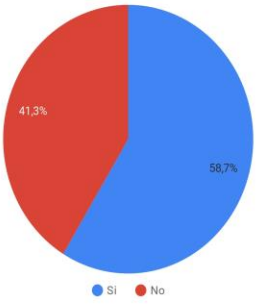
1 - 2 / 2

	Categoría	Porcentaje
1.	Si	58,7
2.	No	41,3

1 - 2 / 2

Categoría Principal:

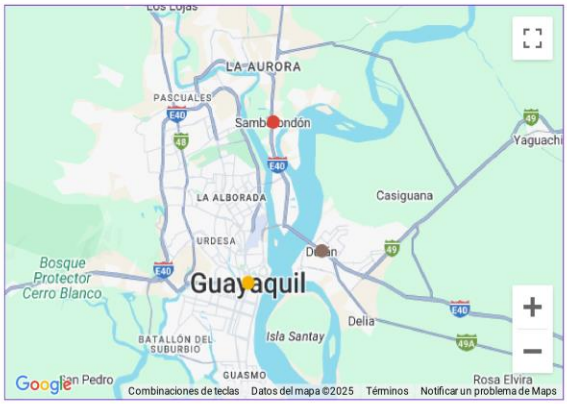
Si



Conteo de Registros

2.354

Zona 8 del Ecuador

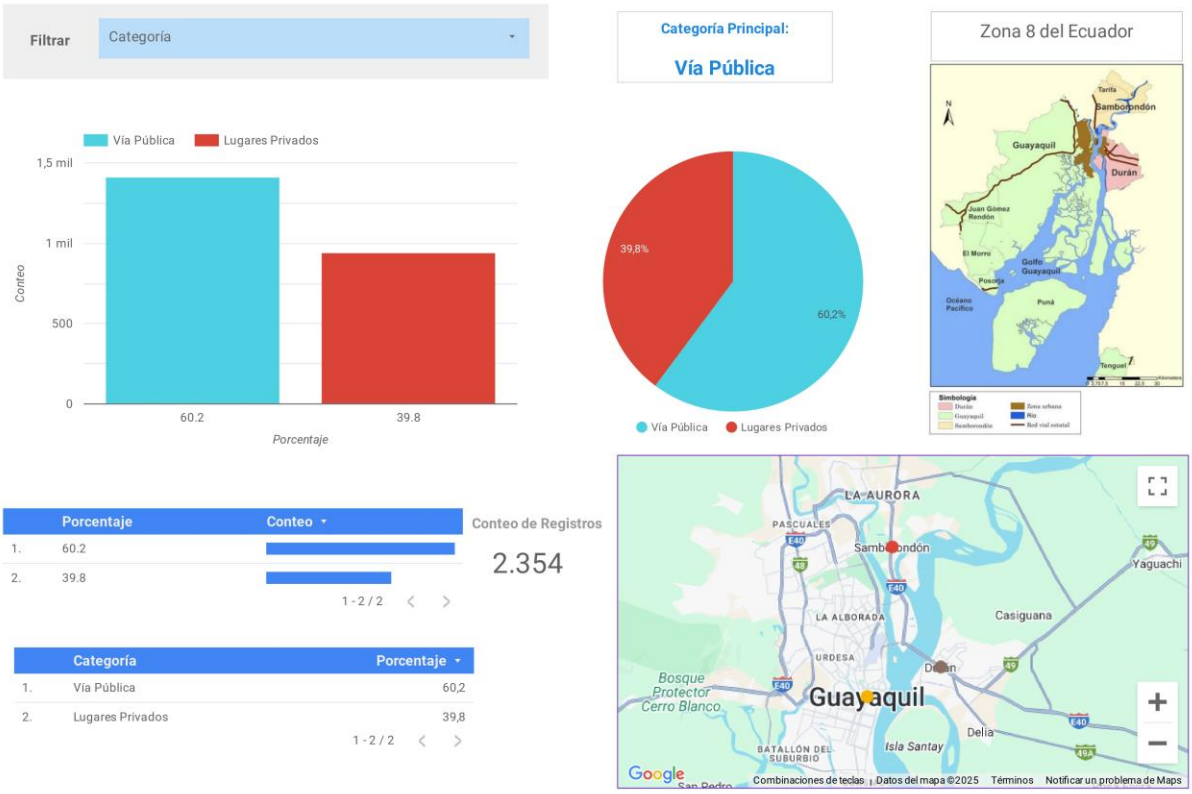


Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

PREDICCIÓN DEL LUGAR PARA LA OCURRENCIA DE ASESINATOS EN LA ZONA 8 DEL ECUADOR

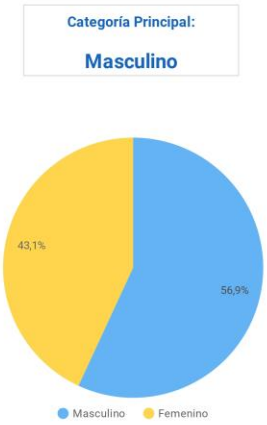


Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

PREDICCIÓN DE SEXO DE VÍCTIMAS DE ASESINATOS OCURRIDOS EN LA ZONA 8 DEL ECUADOR



	Porcentaje	Conteo
1.	56.9	
2.	43.1	

1 - 2 / 2

Conteo de Registros

2.354

	Categoría	Porcentaje
1.	Masculino	56,9
2.	Femenino	43,1

1 - 2 / 2

Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

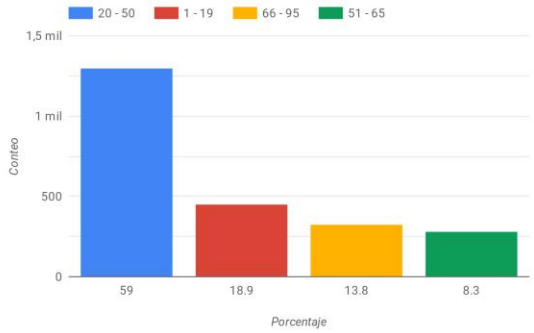
Universidad Nacional de Loja

Carrera de Ingeniería en Sistemas / Computación

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

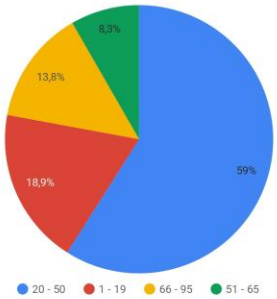
PREDICCIÓN DE EDAD DE VÍCTIMAS DE ASESINATOS OCURRIDOS EN LA ZONA 8 DEL ECUADOR

Filtrar: Categoría



Categoría Principal:

20 - 50 años



Zona 8 del Ecuador

	Porcentaje	Conteo
1.	59	
2.	18.9	
3.	13.8	
4.	8.3	

1 - 4 / 4

Conteo de Registros

2.354

	Categoría	Porcentaje
1.	20 - 50	59
2.	1 - 19	18,9
3.	66 - 95	13,8
4.	51 - 65	8,3

1 - 4 / 4

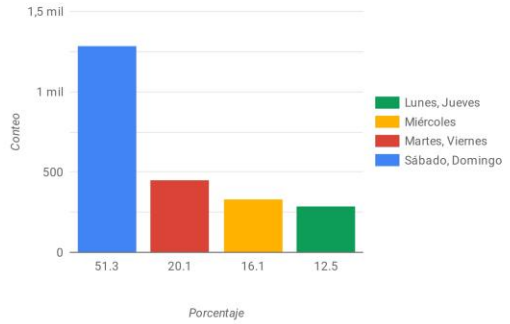
Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

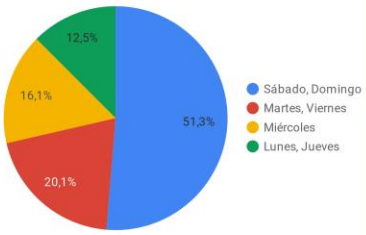
PREDICCIÓN DE DÍA DE LA SEMANA PARA LA OCURRENCIA DE ASESINATOS EN LA ZONA 8 DEL ECUADOR

Filtrar: Categoría



Categoría Principal:

Sábado, Domingo



Zona 8 del Ecuador

	Porcentaje	Conteo
1.	51.3	
2.	20.1	
3.	16.1	
4.	12.5	

1 - 4 / 4

Conteo de Registros

2.354

	Categoría	Porcentaje
1.	Sábado, Domingo	51,3
2.	Martes, Viernes	20,1
3.	Miércoles	16,1
4.	Lunes, Jueves	12,5

1 - 4 / 4

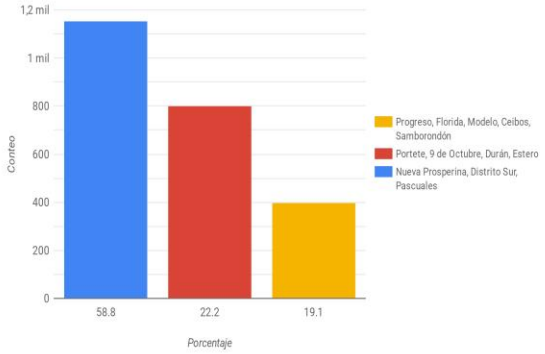
Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

PREDICCIÓN DE DISTRITOS PARA LA OCURRENCIA DE ASESINATOS EN LA ZONA 8 DEL ECUADOR

Filtrar Categoría



	Porcentaje	Conteo
1.	58.8	
2.	22.2	
3.	19.1	

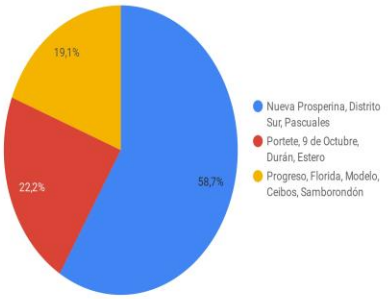
1 - 3 / 3

	Categoría	Porcentaje
1.	Nueva Prosperina, Distrito Sur, Pascuales	58,8
2.	Portete, 9 de Octubre, Durán, Estero	22,2
3.	Progreso, Florida, Modelo, Ceibos, Sambo...	19,1

1 - 3 / 3

Categoría Principal:

Nueva Prosperina, Distrito Sur, Pascuales



Conteo de Registros

2.354

Zona 8 del Ecuador

Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

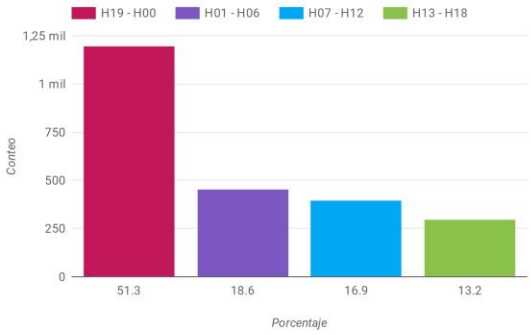
PREDICCIÓN DE HORAS PARA LA OCURRENCIA DE ASESINATOS EN LA ZONA 8 DEL ECUADOR

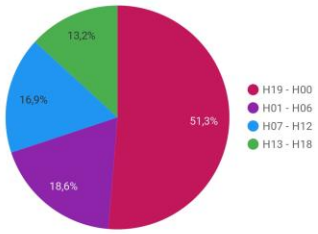
Filtrar: Categoría

Categoría Principal:

H19 - H00
19:00:00 pm - 00:59:00 am

Zona 8 del Ecuador





	Porcentaje	Conteo
1.	51.3	
2.	18.6	
3.	16.9	
4.	13.2	

1 - 4 / 4

Conteo de Registros

2.354

	Categoría	Porcentaje
1.	H19 - H00	51,3
2.	H01 - H06	18,6
3.	H07 - H12	16,9
4.	H13 - H18	13,2

1 - 4 / 4

Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

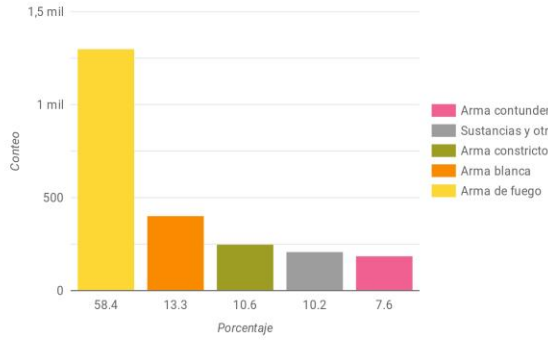
Universidad Nacional de Loja

Carrera de Ingeniería en Sistemas / Computación

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

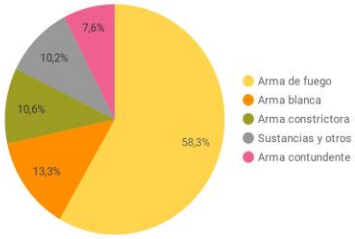
PREDICCIÓN DE ARMAS EN LOS ASESINATOS DE LA ZONA 8 DEL ECUADOR

Filtrar | Categoría



Categoría Principal:

Arma de Fuego



Zona 8 del Ecuador

	Porcentaje	Conteo
1.	58.4	
2.	13.3	
3.	10.6	
4.	10.2	
5.	7.6	

1 - 5 / 5

Conteo de Registros

2.354

	Categoría	Porcentaje
1.	Arma de fuego	58,4
2.	Arma blanca	13,3
3.	Arma constrictora	10,6
4.	Sustancias y otros	10,2
5.	Arma contundente	7,6

1 - 5 / 5

Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

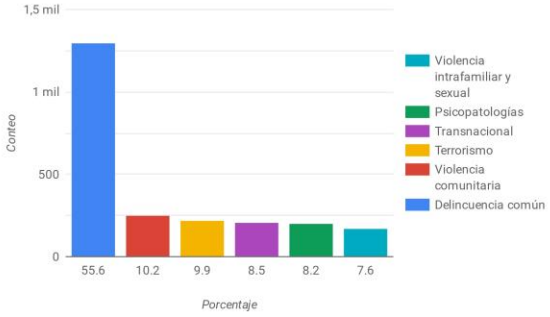
Universidad Nacional de Loja

Carrera de Ingeniería en Sistemas / Computación

ESTADÍSTICAS DE LOS PATRONES DE COMPORTAMIENTO DE LOS ASESINATOS - ZONA 8 DEL ECUADOR

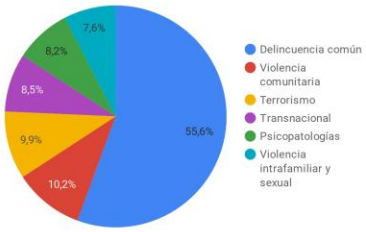
PREDICCIÓN DE MOTIVACIÓN PARA LA OCURRENCIA DE ASESINATOS DE LA ZONA 8 DEL ECUADOR

Filtrar: Categoría



Categoría Principal:

Delincuencia común



Zona 8 del Ecuador

	Porcentaje	Conteo
1.	55.6	
2.	10.2	
3.	9.9	
4.	8.5	
5.	8.2	
6.	7.6	

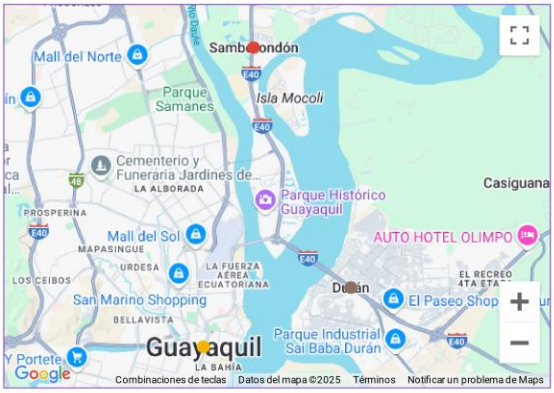
1 - 6 / 6

Conteo de Registros

2.354

	Categoría	Porcentaje
1.	Delincuencia común	55,6
2.	Violencia comunitaria	10,2
3.	Terrorismo	9,9
4.	Transnacional	8,5
5.	Psicopatologías	8,2
6.	Violencia intrafamiliar y sexual	7,6

1 - 6 / 6



Autor: Cecilia Trueba
cecilia.trueba@unl.edu.ec

2025

Anexo IX. Certificado de traducción del resumen al idioma inglés por parte del Profesional.

NEW YORK
ENGLISH LEARNING CENTER

CERTIFICACIÓN DE TRADUCCIÓN DEL RESUMEN (ABSTRACT)

Loja, 24 de enero de 2025

Yo, Josselyn Nicole Gualán Sarango, Licenciada en Pedagogía del Idioma Inglés, (Registro Senescyt 1008-2024-2841618) a petición de la parte interesada.

CERTIFICO:

Que el documento aquí compuesto es fiel traducción del idioma español al inglés, del resumen de la tesis bajo el nombre: **Optimización Bayesiana en modelos de clasificación: Árbol de Decisión y Support Vector Machine para determinar mediante Minería de Datos patrones en los asesinatos de la Zona 8 del Ecuador.**, de autoría de **Cecilia Fernanda Trueba Reyes**, con cédula de identidad: **1106087883**; egresada de la carrera de Ingeniería en Computación de la Universidad Nacional de Loja, previa a la obtención del título de Ingeniera en Ciencias de la Computación.

Particular que certifico en honor a la verdad, autorizando a la interesada hacer uso del presente para los fines académicos pertinentes.

Lic. Josselyn Nicole Gualán Sarango

C.I: 1150214037

Dirección: ATAHUALPA ENTRE 1 DE MAYO Y 10 DE AGOSTO