Datos y Criminalidad: Machine Learning aplicado en modelos predictivos en seguridad.

Elaborado por:

Geovanna López

Pedro Manosalvas

Contacto:

geovannalopez1994@gmail.com

manosalvas26@gmail.com

Diciembre, 2023

Contenido

Índice de Tablas

Índice de gráficas

Abstract

Este reporte se enmarca en una metodología rigurosa y sistemática destinada a analizar la delincuencia en Ecuador con el propósito de comprender sus patrones, tendencias y factores determinantes. Asimismo, busca explorar la viabilidad de construir modelos predictivos para enriquecer el entendimiento ciudadano, promover decisiones informadas, y evaluar la calidad de los datos abiertos que permitan enriquecer la investigación académica en Ecuador.

La etapa inicial de este reporte consiste en la recopilación de datos confiables y pertinentes del portal de Datos Abiertos de Ecuador. Los datos provienen de organismos como el Consejo de la Judicatura, el INEC y se someten a un proceso exhaustivo de limpieza y preprocesamiento, para asegurar la calidad necesaria para su respectivo análisis y modelado. Utilizando visualizaciones y estadísticas descriptivas, se identifican patrones temporales y geográficos relevantes. Posteriormente, la creación y validación de modelos de predicción determina la factibilidad de implementar modelos de aprendizaje automático, fundamentales para detectar patrones.

Los descubrimientos resultantes buscan aportar una comprensión más profunda de los determinantes de la delincuencia en Ecuador. Las conclusiones extraídas informan acerca de la viabilidad y utilidad de los modelos de predicción desarrollados en el informe, así como de la calidad de los datos extraídos de las plataformas de datos abiertos del país.

Se espera que este estudio genere resultados que incluyan la identificación de patrones y tendencias delictivas a lo largo del tiempo y en diversas regiones del Ecuador, a partir de las bases de datos proporcionadas por la plataforma de Datos Abiertos. Además, se anticipa la identificación de factores socioeconómicos y geográficos que ejercen influencia en los niveles de delincuencia. Esto resulta particularmente relevante para evaluar la eficacia de los datos proporcionados por las instancias gubernamentales. Además, se aspira a obtener una comprensión profunda de los desafíos subyacentes relacionados con la seguridad, y a informar tanto a la ciudadanía como a la sociedad civil sobre la utilidad de los datos abiertos en cuestiones de seguridad. Finalmente, se busca proporcionar información de relevancia a las instituciones interesadas en la seguridad ciudadana en Ecuador.

Palabras clave: Machine learning, Random Forest, Datos Abiertos, Seguridad.

1. Introducción

1.1. La crisis de seguridad

Ecuador enfrenta una compleja coyuntura caracterizada por la escalada de violencia, desafíos en seguridad, política y democracia. Históricamente considerado una "isla de paz", el país ha experimentado una transformación alarmante en su panorama. La evolución del proceso de paz en Colombia, la expansión de actividades ilícitas en regiones fronterizas y la desmovilización de insurgentes han contribuido a la fragilidad de las instituciones democráticas y a la creciente inseguridad. La convergencia de estos factores internos y externos ha transformado a Ecuador de "isla de paz" a uno de los países más inseguros de Latinoamérica, planteando desafíos significativos para la democracia y la estabilidad social.

1.2. Abordaje Analítico de la Crisis de Seguridad en Ecuador

La creciente crisis de seguridad en Ecuador, caracterizada por la escalada de los índices de violencia y desafíos en los ámbitos de seguridad, demanda un análisis crítico y sistemático para comprender y mejorar la seguridad ciudadana en el país. Históricamente considerado como un país sosegado, el drástico cambio en su panorama ha suscitado la necesidad de entender y abordar esta problemática desde las organizaciones civiles, la academia y las empresas privadas.

La adopción de medidas gubernamentales eficaces exige un enfoque basado en la evidencia. La utilización de datos abiertos y el análisis de la información disponible se revelan como elementos esenciales para una comprensión profunda de la situación actual. Este enfoque informado no solo permite a las instituciones tomar decisiones fundamentadas, sino también a la ciudadanía entender la magnitud de la problemática.

Investigaciones previas resaltan la importancia del análisis de datos en el contexto del crimen. La utilización de datos desglosados y la implementación de modelos predictivos, han demostrado su capacidad para enriquecer estrategias de prevención y mejorar las proyecciones de intervenciones. Si bien los estudios de predicción del crimen son limitados en América Latina, existen precedentes que muestran cómo estos enfoques pueden ser aplicados con éxito.

En el caso ecuatoriano, la delincuencia ha generado preocupación a nivel nacional e internacional. A pesar de que investigaciones anteriores han abordado factores socioeconómicos y políticas públicas como impulsores de comportamientos delictivos, existe

una necesidad imperante de implementar estudios basados en datos cuantitativos que brinden herramientas precisas para abordar la crisis de seguridad.

Según la Estrategia Nacional contra la Delincuencia Organizada 2023- 2030 (Ministerio del Interior, 2023), para el año 2021 la violencia criminal aumentó en un 93,2% en Ecuador, convirtiendo al país en un creciente foco de violencia en América del Sur. El aumento de los crímenes violentos y homicidios está directamente vinculado al crimen organizado, estableciendo una tasa de criminalidad al 6, 25. Dado la alarmante situación, en este trabajo se han seleccionado los crímenes catalogados cómo violentos, para delimitar el análisis.

Este reporte se enmarca en la necesidad de llenar el vacío de investigación en el campo de la seguridad ciudadana en Ecuador. El análisis de datos abiertos, la implementación de modelos predictivos y la comprensión profunda de las causas subyacentes de la inseguridad, permitirán formular estrategias efectivas de prevención del crimen, contribuyendo a restaurar la confianza en las instituciones, fortalecer la democracia y preservar la estabilidad social.

1.3. Aplicación de Modelos predictivos

La predicción y prevención del crimen son temas de suma importancia en la sociedad actual. La aplicación de técnicas de Machine Learning en este campo se ha convertido en una herramienta prometedora para abordar estos desafíos. A través de la recopilación y análisis de datos relacionados con incidentes delictivos, el Machine Learning puede ayudar a las agencias de aplicación de la ley y la sociedad civil a anticipar y prevenir la comisión de crímenes. En este trabajo se analizarán diversos estudios que abordan la predicción y prevención del crimen mediante el uso de técnicas de Machine Learning. Se explorarán las

contribuciones de cada estudio y se discutirá cómo el Machine Learning puede mejorar la eficacia de las estrategias de aplicación de la ley en diferentes contextos.

El estudio titulado "*Crime Prediction using Machine Learning with a Novel Crime Dataset*" se centra en la necesidad de una base de datos estructurada de crímenes en Bangladesh. Este estudio destaca cómo el aprendizaje automático puede aprovechar datos temporales, geográficos, climáticos y demográficos para predecir y prevenir el crimen. Asimismo, se menciona la evaluación de algoritmos de clasificación supervisada, lo que muestra resultados prometedores. Utilizando técnicas de Machine Learning, se diseñaron características que facilitan la predicción de crímenes y evaluaron varios algoritmos de clasificación supervisada. Los resultados obtenidos sugieren que esta base de datos podría ser valiosa para mejorar la predicción del crimen (Shohan et al., 2022).

Por otro lado, el estudio titulado "*Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention*" propone la integración de técnicas de Machine Learning y visión por computadora en la predicción y prevención del crimen. Los autores argumentan que estas tecnologías avanzadas pueden ofrecer valiosas herramientas para mejorar la detección y prevención del crimen. A través de estudios de casos del mundo real, ilustran cómo estas técnicas pueden transformar las estrategias de aplicación de la ley para abordar los desafíos de la delincuencia contemporánea (Shah et al., 2021).

En adición, el trabajo titulado *"Interpretable machine learning models for crime prediction*" aborda la necesidad de comprender las relaciones entre los patrones de crímenes y las variables asociadas. Los autores proponen el uso de modelos de Machine Learning avanzados y el método SHAP para mejorar la interpretabilidad de los modelos. Su enfoque se basa en

teorías criminológicas, y seleccionan cuidadosamente las variables para la predicción de crímenes. Esto permite una comprensión más profunda de los factores críticos para la predicción del crimen y su variación espacial (Zhang et al., 2022).

Otro estudio "*Machine Learning-based Soft Computing Regression Analysis Approach for Crime Data Prediction*" se centra en la creciente tasa de criminalidad en India. Los autores aplican diversos algoritmos de regresión para predecir el número de infracciones al Código Penal de India (IPC) por región y tipo de crimen. Sus resultados indican que el modelo de Random Forest Regression (RFR) funciona excepcionalmente bien, lo que ofrece información valiosa para asignar recursos de manera efectiva en la lucha contra el crimen en India (Aziz et al., 2022).

En otro continente, el estudio "*Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics*" se enfoca en la aplicación de tecnologías geoespaciales, minería de datos y Machine Learning para la predicción y monitoreo de crímenes en Porto, Portugal. Los autores aprovechan datos policiales oficiales y análisis espaciales para identificar patrones y áreas críticas de crimen. Además, realizan un análisis de sentimientos en tweets relacionados con la inseguridad. Estas técnicas combinadas mejoran la capacidad de interpretación de patrones de crímenes, beneficiando tanto a las fuerzas del orden como a los planificadores urbanos (Saraiva et al., 2022).

Safat et al. (2021) publican su trabajo titulado "*Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques*" que se centra en la predicción y pronóstico de crímenes en áreas metropolitanas de Chicago. Los autores aplican una variedad de algoritmos de Machine Learning y Deep Learning para analizar datos de

crímenes. Destacan la eficacia de las redes neuronales LSTM en el análisis de series temporales, lo que proporciona predicciones precisas de patrones de crímenes en diferentes regiones.

Ansor et. Al (2022) investiga la aplicación de árboles de decisión y técnicas de conjunto para la clasificación y regresión de crímenes en su trabajo titulado *" Building Model for Crime Pattern Analysis Through Machine Learning Using Predictive Analytics*". Los autores abordan el desafío de datos desequilibrados utilizando una técnica de sobre muestreo sintético *(synthetic minority oversampling*). Destacan la eficacia de random forest en el análisis de patrones de crímenes. Este estudio enfatiza el potencial de los algoritmos de conjunto (ensemble algorithm) en la lucha contra el crimen.

Finalmente, el estudio titulado "*Crime Prediction and Analysis Using Machine Learning*" se centra en la aplicación de técnicas de Machine Learning para analizar datos de crímenes en los estados de la India. Los autores utilizan datos públicos disponibles en Kaggle para predecir la probabilidad de diferentes tipos de crímenes en momentos y lugares específicos. Reconocen el potencial de la IA y el Machine Learning en la predicción del crimen, aunque señalan la necesidad de abordar desafíos en la aplicación efectiva de estas tecnologías (Balu et al., 2022).

Cabe añadir que Wheeler, A.P., Steenbeek (2021), mencionan en su estudio "*Mapping the Risk Terrain for Crime Using Machine Learning",* que los modelos predictivos, basados en regresión, usan variables predictoras para medir la probabilidad de que un crimen ocurra en el futuro, siempre asumiendo relaciones lineales entre aquellos predictores. Por ello se recalca la particularidad de usar datos demográficos de censos poblacionales, lo cual no es habitual en

investigaciones basadas en micro lugares. Existen teorías del crimen que explican la distribución espacial del crimen basándose en tales factores demográficos. De igual manera, en este estudio, se ha añadido y explorado la correlación del delito con variable demográficas como pobreza multidimensional, población y desempleo, seleccionadas a partir de los factores de riesgo de delitos violentos según la *Estrategia Nacional Contra la Delincuencia Organizada 2023 – 2030* (Ministerio del Interior, 2023)

Los estudios analizados demuestran la creciente importancia del Machine Learning en la predicción y prevención del crimen en diferentes contextos geográficos como India, Bangladés, Portugal y Estados Unidos.

El uso de datos abiertos (o no) recopilados de manera exhaustiva y la aplicación de técnicas avanzadas de Machine Learning permiten a las agencias de aplicación de la ley, a la sociedad civil y a la academia a obtener valiosas perspectivas sobre patrones delictivos, factores predictivos y variaciones espaciales. Además, la integración de visión por computadora, análisis de texto y análisis espacial amplía las capacidades de predicción y prevención del crimen que deja una puerta abierta para el presente estudio.

Sin embargo, se deben abordar desafíos, como la interpretabilidad de los modelos de Machine Learning y la necesidad de enfrentar problemas de datos desequilibrados. A medida que estas técnicas continúan evolucionando, es esencial encontrar un equilibrio entre la precisión y la aplicabilidad de los modelos para garantizar su aplicabilidad práctica en el campo de la aplicación de la ley.

En última instancia, el Machine Learning tiene el potencial de mejorar significativamente la eficacia de las estrategias de aplicación de la ley en la predicción y prevención del crimen, lo que contribuye a la seguridad y la calidad de vida de las comunidades en todo el mundo.

2. Metodología

2.1. Datos

Los datos usados para este reporte se han recopilado desde el Portal de Datos Abiertos de Ecuador. El primer Dataset contiene información sobre los delitos que se han ingresado al Consejo de la Judicatura desde el año 2016, hasta la última fecha de corte (agosto 2023). El registro de datos del Consejo de la Judicatura contiene información sobre el número de artículo y nombre del delito infringido, la materia legal, el tipo de acción ingresado, provincia, cantón, estado del proceso e instancia. Los delitos seleccionados, para delimitar el estudio, debido a la amplitud del Dataset, fueron los delitos catalogados como violentos, siendo un total de 21 (Anexo 1).

El segundo Dataset contiene datos económicos, consiste en una recopilación de varios conjuntos de datos que se combinaron en uno solo, recopilados desde el portal del INEC. Este proceso fue bastante minucioso debido a que los datos fueron extraídos de más de 8 diferentes fuentes de datos no estandarizadas. Las metodologías trabajadas desde el INEC, muestran información descontinuada e incompleta, que varía de año a año, así como archivos que no cumplen los principios de datos abiertos para el uso de la sociedad civil[1]. Pese a las dificultades presentadas, se logró recopilar datos sobre la tasa de desempleo bruto en Ecuador

[1] Los datos, si bien se encuentran en una plataforma abierta, no cumplen con criterios de accesibilidad y tampoco son comparables.

desde el año 2016 al segundo periodo del 2023, la población nacional[2] y el indicador de pobreza multidimensional dentro de este mismo periodo.

2.2. Preprocesamiento de datos

Información del Dataset final

- Se incluyeron datos socioeconómicos tomados de la plataforma INEC: desempleo (tasa de desempleo), Población, y Pobreza multidimensional (tasa de P. M).
- Se tomó información existente desde el 2016 al 2023 (se tomó la información hasta del segundo trimestre del 2023, disponible a la fecha de este reporte)
- Debido al cambio de metodologías del INEC, los datos no están estandarizados, y se recopiló desde varias fuentes del INEC, donde la metodología cambia de un año a otro, esto puede crear sesgos, pero se trabajó con la información validada disponible.
- Cada variable se desagregó por Provincia y Cantón.
- Para cada Cantón (en las 3 variables socioeconómicas: desempleo, pobreza multidimensional, población) sólo existen datos de: Guayaquil, Quito, Cuenca, Machala, Ambato, por lo tanto, se tomó en cuenta sólo estos valores en la variable Cantón.

Tasa de desempleo

[2] A la fecha de recopilación de los datos, el nuevo censo nacional 2022, había sido publicado, en la plataforma de visualización de datos del INEC. Presentando inconsistencias entre las proyecciones poblacionales de años anteriores. Dado que la información era inconsistente e incompleta, para la construcción de la data set de este trabajo, se ingresó manualmente los datos actualizados del CENSO para el año 2022 y 2023, mientras que para años previos se mantuvieron los datos presentados por INEC previamente. Esto es un problema ya que crea sesgos en los resultados cuando se trabaja con grandes cantidades de información.

- Para la Tasa de Desempleo, se tomó 3 Datasets diferentes 2016- 2018, 2019-2021-2022, 2020 -2023 este último cuenta solo con datos de los Cantones más grandes mencionados anteriormente.
- Se tomaron datos de Desempleo del último trimestre de cada año, a excepción del 2023, donde se registró todo el trimestre II.
- Para los valores faltantes (Valores nulos) en tasa de desempleo se Imputo Datos con la Mediana, por cada columna, ya que esto no crea sesgos en los valores faltantes (provincia 2020 - 2023).

Población

- Para la población, se cuenta con datos completos para Quito, Guayaquil, Cuenca, Machala, Ambato del 2016 al 2023
- Hay datos faltantes del 2018 - 2019, 2020, 2021, para todas las Provincias.
- Se imputo los valores faltantes (Valores nulos) con la mediana de cada Provincia para no afectar los resultados posteriores.
- A la fecha de recopilación de datos no existía un archivo tabulado descargable, con la información obtenida del censo 2022, por lo que se insertó datos manualmente, obtenidos desde la plataforma de visualización del INEC.
- Los datos correspondientes al 2023- Provincia son los mismos que muestran para el 2022 - Provincia ya que son recientes.
- Existe una gran diferencia entre los resultados del Censo Poblacional 2010 y los nuevos resultados 2022, esto termina afectando y sesgando los resultados finales de estudios demográficos.

Pobreza multidimensional

- No existen datos desagregados de pobreza multidimensional por Provincia y Cantón antes del 2019, por lo tanto, no se incluyó.
- Se encontró datos para Cantón, y Provincia, solo para los años 2019, 2021, 2022, para los demás años la información está incompleta o no se encontró en el portal.
- Debido al número considerable de datos faltantes para esta variable, imputar los valores nulos con 0 o la mediana podría crear sesgos, por lo que en un principio se tratará de suprimir los valores nulos y trabajar con los valores existentes. De no ser posible se borrarán estas columnas.

El Dataset final de las variables Socioeconómicas se unió con los datos abiertos obtenidos del Consejo de la Judicatura, con los valores Año, Delito, Cantón y Provincia, descartando las otras variables que no fueron necesarias para este estudio. Para la Variable Delito se tomaron en cuenta los delitos catalogados como Violentos (Anexo 1).

Tras la recopilación de la información, y la limpieza de los datos, donde se eliminaron los valores faltantes y duplicados, y en el caso de los datos socioeconómicos, se procedió a realizar el proceso de imputación de datos con la mediana para no afectar o crear sesgos en los resultados. Se procedió a concatenar los datos en un solo Dataset donde se encuentren solo las variables que se tomarán en este análisis. El Dataset filtrado final contiene diez columnas, distribuidas entre variables categóricas y numéricas: delito, año, provincia, cantón, tasa de desempleo por provincia, tasa de desempleo por cantón, población por provincia, población por cantón, e índice de pobreza por provincia e índice de pobreza por

cantón. El número final de columnas es 10, y el número filas es 55.770, siendo un set de datos limpio, con información robusta, delimitada y correctamente procesada.

Tabla 1. Dataset Final

	Delito	Provincia	Canton	Año	Tasa de Desempleo Provincia	Tasa de Desempleo Canton	Poblacion Provincia	Población Canton	Pobreza Provincia	Pobreza Canton
0	47	2382	1128	2016	2.90	5.7	837801.0	386346.0	0.0	0.0
1	47	2382	1128	2017	4.00	5.7	1004108.0	393502.0	0.0	0.0
2	47	2382	1128	2018	2.60	5.7	819705.0	400655.0	0.0	0.0
3	47	2382	1128	2019	3.40	5.7	819705.0	405270.0	24.7	7.2
4	47	2382	1128	2020	3.35	7.3	819705.0	410401.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
55765	4936	1613	986	2019	2.20	5.9	568452.5	193859.0	33.5	7.7
55766	4936	1613	986	2020	2.65	8.0	568452.5	196313.0	0.0	0.0
55767	4936	1613	986	2021	3.10	5.9	568452.5	199012.0	38.0	11.4
55768	4936	1613	986	2022	2.20	5.9	563532.0	201758.0	39.6	9.9
55769	4936	1613	986	2023	2.65	3.8	563532.0	203141.0	0.0	0.0

55770 rows × 10 columns

La metodología adoptada en este estudio consiste en el análisis de datos sobre delincuencia en Ecuador y el modelado de un modelo de aprendizaje automático (Random Forest) para la predicción del delito en Ecuador. El software para la traducción y modelado de los datos fue Python, usando librerías de análisis como Numpy, Pandas, Skylearn, Seaborn, Matplotlip, entre otras.

El siguiente paso fue la prueba de correlaciones entre variables y se realizó el análisis exploratorio de datos (EDA), donde se busca entender un panorama general entre la correlación de las variables que inciden en la delincuencia en Ecuador, mediante visualizaciones y estadísticas descriptivas, identificando patrones temporales, geográficos que permitan entender la distribución de la información.

Debido que se busca crear un modelo predictivo, las variables categóricas, debían transformarse a variables numéricas para poder ser correctamente interpretadas, por ello, se

aplicó la codificación de características, el equilibrio de los datos y el escalado de características. Este último paso es especialmente importante en el preprocesamiento de datos, ya que estandariza los valores y garantiza que todas las características tengan una influencia justa en el modelo, evitando el predominio de las características con valores más altos.

3. Resultados del Análisis Exploratorio de Datos

3.1. Correlaciones

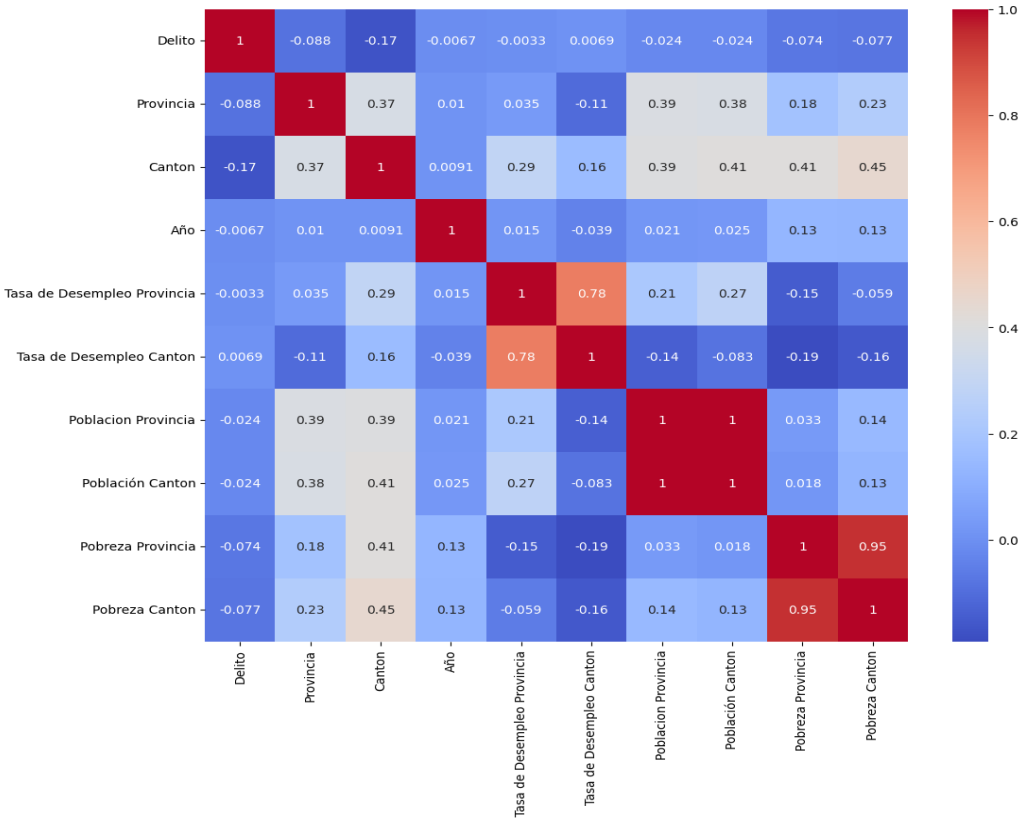


Gráfico 1. Matriz de correlación

La matriz de correlación[3] indica en general que la correlación entre Delito y otras variables es bastante baja, ya que los valores son cercanos a 0. La tasa de desempleo por Cantón es la variable más cercana a 1, esto puede deberse a que es la variable con más información, no solo a la correlación existente. Además, la correlación entre Delito y Pobreza representa la variable más cercana a 0 por lo que en un inicio se podría descartar la correlación, pero

[3] Para revisión del código, y bases de datos ir a https://github.com/GeovaLopez/Datos-y-Criminalidad---OESN

tomando en cuenta la cantidad de valores nulos para esta columna, es claro que este resultado se debe a falta de datos suficientes para analizar.

3.2. Visualización de datos

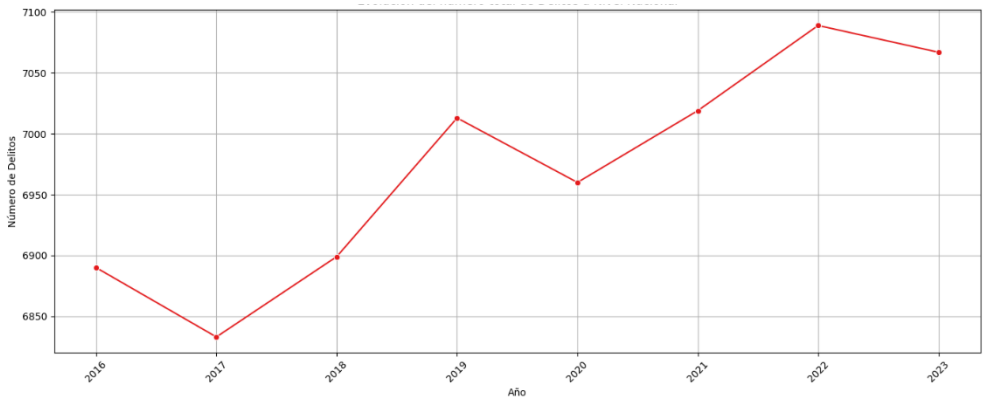


Gráfico 2. Evolución del Delito a Nivel Nacional

En el gráfico se puede apreciar un significativo el aumento de casos en el año 2022 y que se mantiene hasta la fecha de corte (agosto 2023). Claramente en el año 2020, debido a la pandemia la cifra disminuye, y se vuelve a incrementar y sigue en ascenso hasta la actualidad.

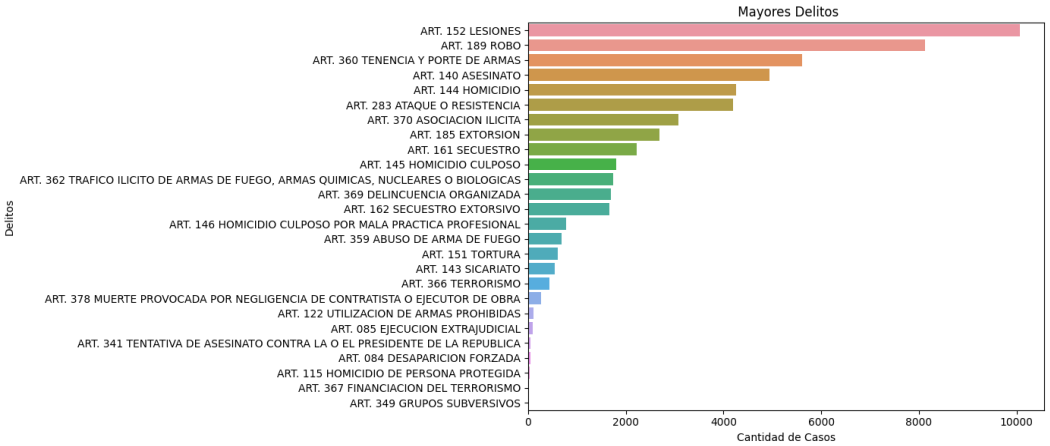


Gráfico 3. Delitos violentos recurrentes.

Los delitos violentos más recurrentes, son las lesiones, con una cantidad mayor a diez mil causas ingresadas, seguidas por el robo, con más de ocho mil causas y la tenencia y el porte de armas en tercer lugar. Otros delitos como la delincuencia organizada están cercanos a dos mil casos.

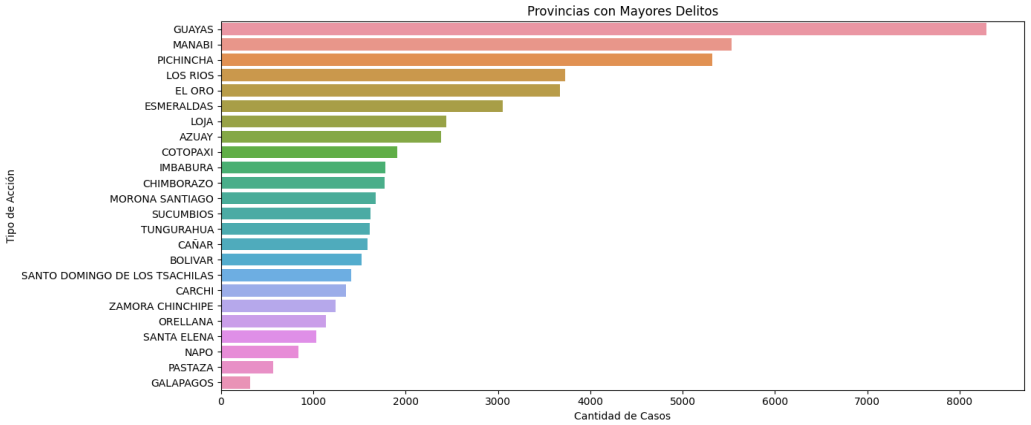


Gráfico 4. Provincias con Mayores Delitos

Los delitos violentos ocurren en Guayas con más de ocho mil casos, Manabí y Pichincha con más de cinco mil casos, Galápagos, Pastaza y Napo son las provincias con menos casos.

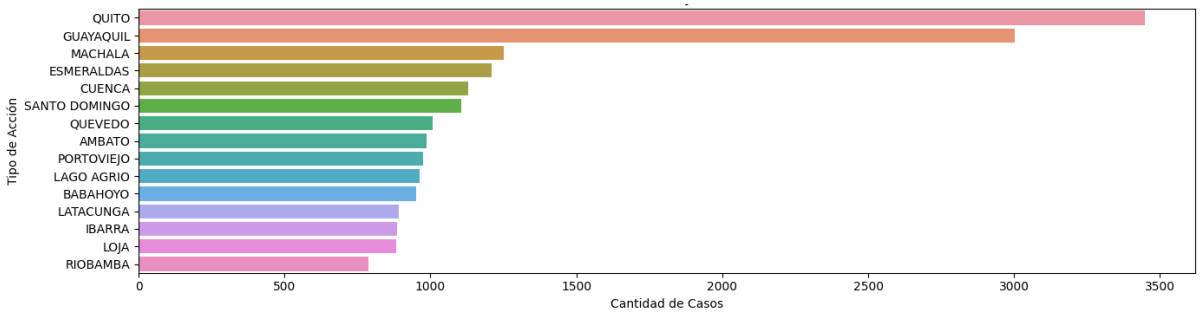
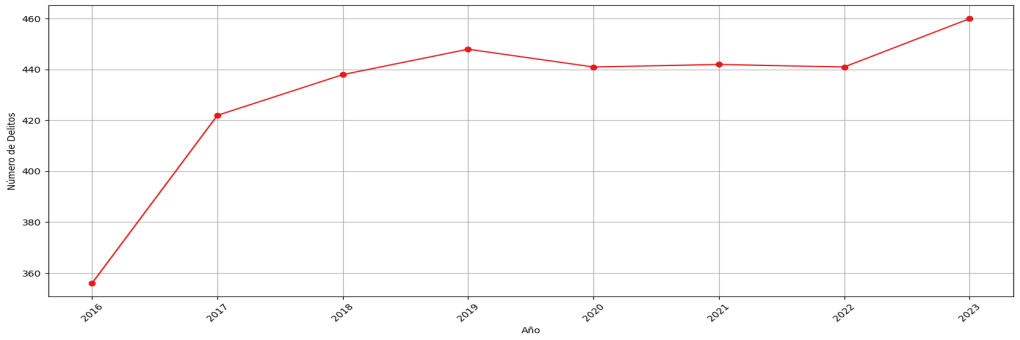


Gráfico 5. Cantones con mayores delitos

Las ciudades con más delitos son Quito, Guayaquil, seguida de Machala y Esmeraldas. Estas ciudades son ciudades con una alta densidad poblacional. La evolución y distribución de delitos violentos desde el año 2016, hasta mediados del 2023, en Quito, presenta un gran aumento de casos en el año 2023, siendo la ciudad con más delitos por lo tanto la más peligrosa.

Seguida por Guayaquil, que mantiene esta tendencia, que llega a su punto máximo en el 2022 y que, sin embargo, en el 2023 desciende, considerando que la fecha de corte fue a la mitad de este año, por lo que en el futuro podría aumentar.



Gráficos 6. Evolución del crimen en la ciudad de Quito

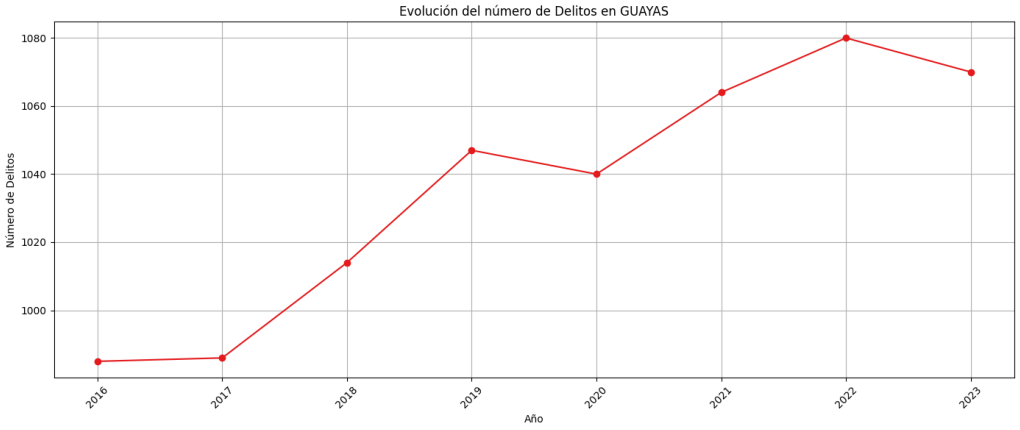


Gráfico 7. Evolución del crimen en la ciudad de Guayaquil

Los delitos extremadamente violentos como el asesinato presenta mayores números en las Provincias como Guayas y Manabí, seguidos por Pichincha. Por su parte el delito de sicariato, presenta igual o mayor número de casos en la provincia de Pichincha a comparación de Guayas, catalogada como la ciudad más violenta.

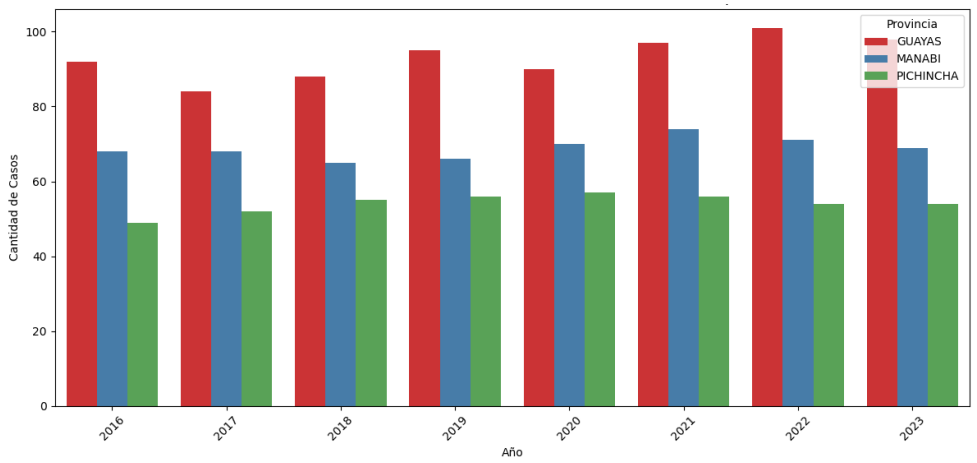


Gráfico 8. Distribución de Asesinatos en Guayas, Manabí y Pichincha por año

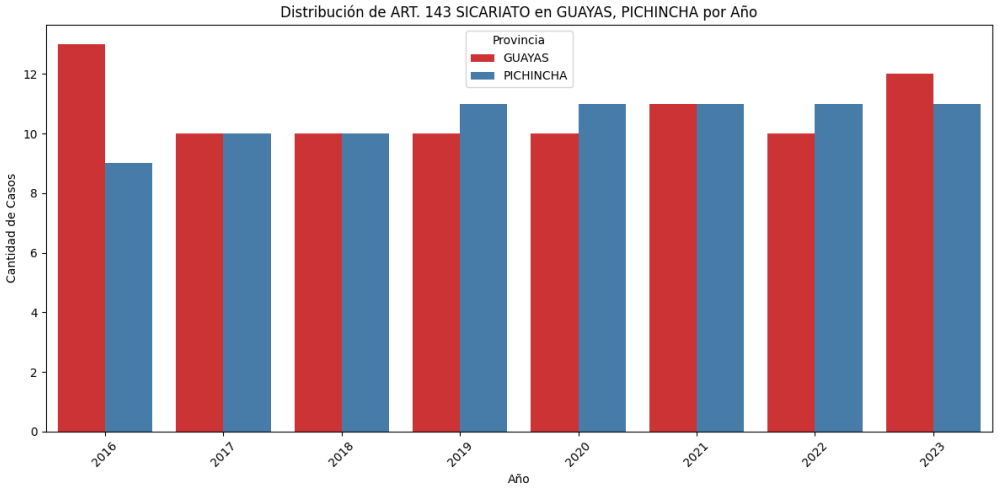


Gráfico 9, Distribución de delitos de sicariato en Guayas y Pichincha por año

Finalmente, para otros delitos violentos asociados con narcotráfico, como el tráfico ilícito de armas, el homicidio culposo, la tenencia y el porte de armas, y la asociación ilícita. Se encontró que en todos los casos la provincia de Guayas reincidió por número de registro de causas ingresadas. Siendo la provincia y población más afectada de todo el país, seguida por Manabí y Pichincha. Los gráficos de la recurrencia del delito en Guayas y en la ciudad de

Quito, así como el aumento del delito de sicariato a escala nacional pueden visualizarse en el Anexo 2.

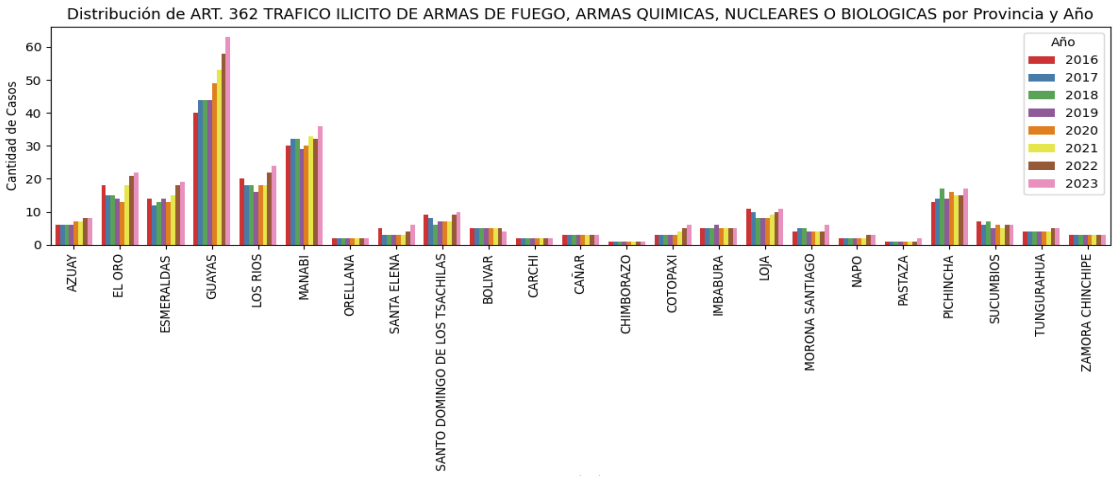


Gráfico 10. Distribución del delito de tráfico de ilícito a nivel nacional por año

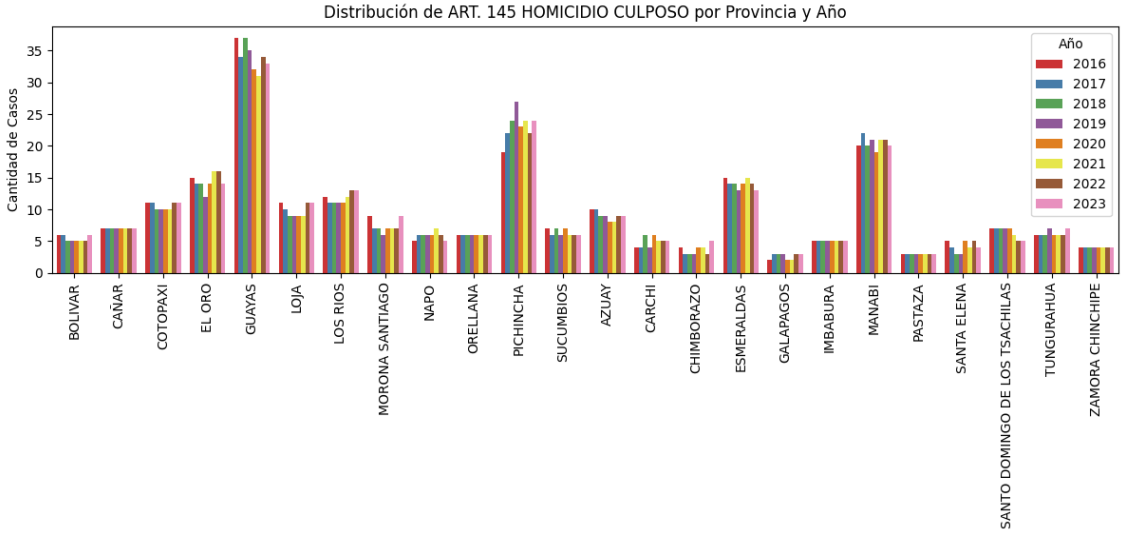


Gráfico 11. Distribución del delito de Homicidio Culposo a nivel nacional por año

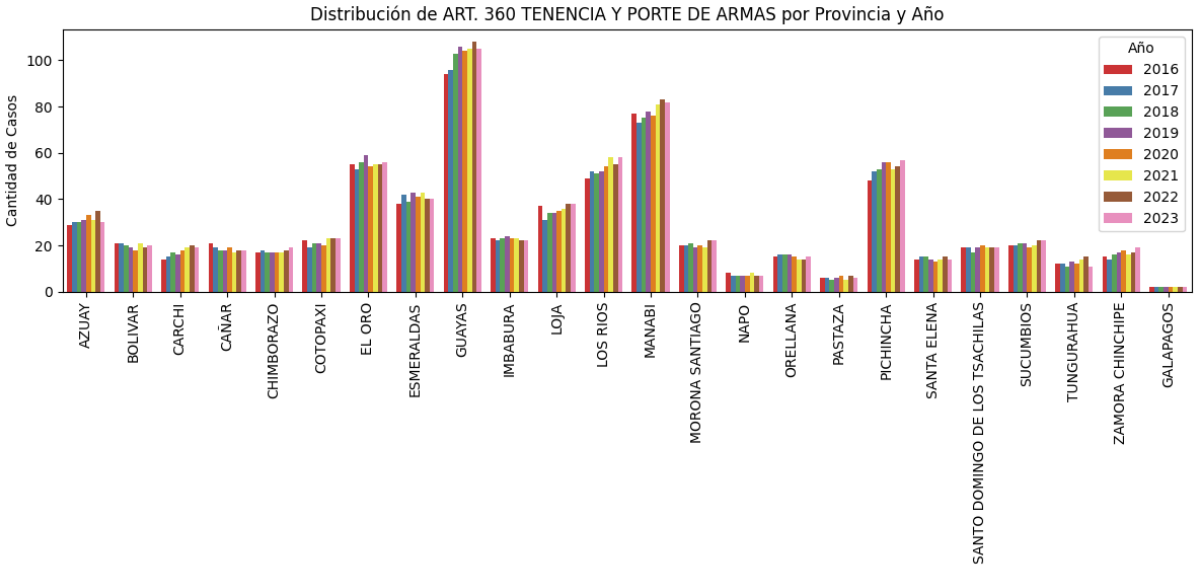


Gráfico 12. Distribución del delito de tenencia y porte de armas a nivel nacional por año

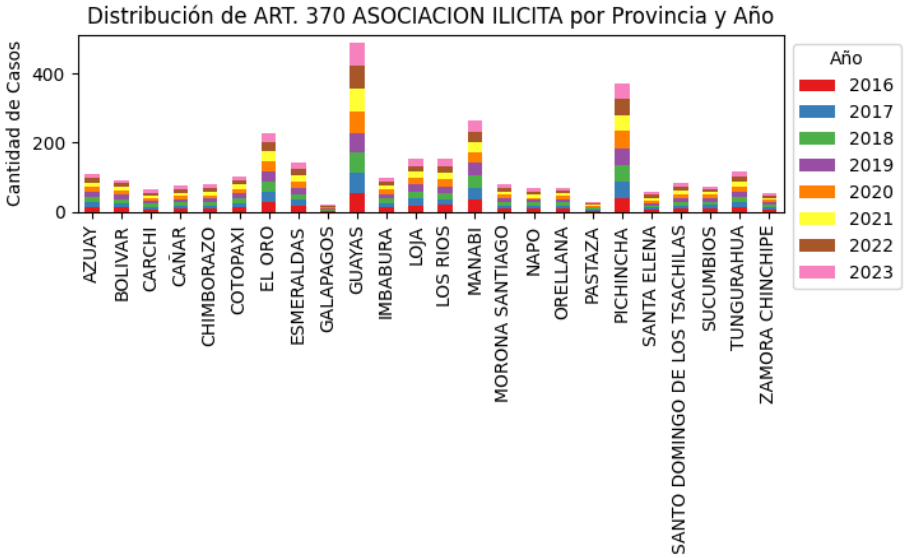


Gráfico 13. Distribución del delito de asociación ilícita a nivel nacional por año

3.3. Construcción del Modelo de Predicción "Random forest"

Random Forest es un algoritmo de aprendizaje automático basado en el ensamblaje de muchos árboles de decisión. Funciona creando una colección de árboles de decisión durante el entrenamiento y toma decisiones basadas en la mayoría de votos de estos árboles. Cada árbol crea un bosque utilizando subconjuntos aleatorios de datos y características. Esta diversidad y aleatorización de los datos ayudan a mejorar la robustez del modelo y a reducir el sobreajuste.

En este estudio, se aplica Random Forest por varias razones. Primero, porque es capaz de manejar conjuntos de datos complejos con múltiples características y relaciones no lineales entre variables. Además, Random Forest es resistente al sobre ajuste gracias a su capacidad para generalizar bien a datos no vistos. Comparado con otros modelos, Random Forest puede ofrecer un equilibrio entre precisión y capacidad de generalización. A menudo supera a los modelos individuales de árboles de decisión y es menos propenso a sobre ajustarse en comparación con métodos más complejos. Esto lo convierte en una elección ideal para este estudio, donde la capacidad para manejar datos complejos es de mucha importancia. Estudios como "*Mapping the Risk Terrain for Crime Using Machine Learning*" (Wheeler, A.P., Steenbeek, 2021), mencionan cómo entre varios modelos predictivos, el modelo algorítmico Random Forest puede proporcionar predicciones de largo término sobre la delincuencia.

En las etapas iniciales del proceso de modelado con Random Forest, es esencial preparar los datos y dividir el conjunto en conjuntos de entrenamiento y prueba. La preparación de datos implica asegurarse de que estén limpios y procesados adecuadamente, como la codificación de variables categóricas y la normalización de características. La división del conjunto de

datos se realiza utilizando la función `train_test_split` de la librería scikit-learn, donde se asignan conjuntos de características y etiquetas para entrenamiento y prueba. Este procedimiento es crucial para garantizar que el modelo se entrene adecuadamente y pueda generalizar bien a datos no vistos durante la evaluación.

Ecuación 1. Dimensiones de los datos de prueba y evaluación

```
Dimensiones de X_train: (44616, 7)
Dimensiones de X_test: (11154, 7)
```

Precisión del modelo

En la fase de entrenamiento del modelo, se empleó un clasificador de Random Forest con parámetros específicos, (n_estimators=50, max_depth=5, random_state=42). Este clasificador utiliza un ensamblaje de árboles de decisión, donde `n_estimators` define la cantidad de árboles en el bosque y `max_depth` controla la profundidad máxima de cada árbol. El valor `random_state=42` asegura reproducibilidad al fijar la semilla aleatoria.

La precisión del modelo, se obtiene mediante la comparación de las predicciones del modelo con las etiquetas reales del conjunto de prueba (test). La precisión es un indicador de qué tan bien el modelo identifica correctamente las clases. Un valor de 0.1805, como el que se obtuvo en este estudio, indica que aproximadamente el 18% de las predicciones del modelo son correctas en el conjunto de prueba. Este sin duda es un valor muy bajo, que tiene relación con la calidad de los datos con los que se trabajó.

La matriz de confusión visualiza la cantidad de predicciones correctas y erróneas, mientras que el informe de clasificación proporciona métricas detalladas para cada clase, como

precisión, recall y f1-score, ofreciendo una comprensión más profunda de cómo el modelo clasifica diferentes categorías del crimen. Como podemos observar los resultados de este estudio no fueron idóneos.

Tabla 2. Precisión del Modelo, Informe de clasificación y matriz de confusión

Precisión del modelo: 0.18056302671687288

Informe de clasificación:

	precision	recall	f1-score	support
8	0.00	0.00	0.00	3
16	0.00	0.00	0.00	1
32	0.00	0.00	0.00	9
47	0.00	0.00	0.00	7
48	0.00	0.00	0.00	12
96	0.00	0.00	0.00	19
114	0.00	0.00	0.00	21
263	0.00	0.00	0.00	54
438	0.00	0.00	0.00	100
553	0.00	0.00	0.00	105
609	0.00	0.00	0.00	116
689	0.00	0.00	0.00	148
773	0.00	0.00	0.00	147
1661	0.00	0.00	0.00	333
1699	0.00	0.00	0.00	330
1743	0.00	0.00	0.00	343
1801	0.00	0.00	0.00	361
2231	0.00	0.00	0.00	417
2684	0.00	0.00	0.00	526
3075	0.00	0.00	0.00	630
4198	0.00	0.00	0.00	815
4255	0.00	0.00	0.00	821
4936	0.00	0.00	0.00	997
5612	0.00	0.00	0.00	1112
8129	1.00	0.00	0.01	1719
10060	0.18	1.00	0.31	2008
accuracy			0.18	11154
macro avg	0.05	0.04	0.01	11154
weighted avg	0.19	0.18	0.06	11154

Con el fin de mejorar los resultados, se realizó la evaluación de los mejores parámetros, resultando los detalles descritos a continuación. Sin embargo, no se logró mejorar el rendimiento del modelo ya que la precisión se mantiene en 18%.

Ecuación 2, Mejores parámetros para el modelo

```
Mejores parámetros: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
```

Finalmente se construyó la importancia de las características, resultando entre las variables socioeconómicas, la población del cantón como una de las más relevantes, seguida por la población de la provincia y la tasa de desempleo en la provincia.

Tabla 3. Características más importantes

Importancia de las características:

	Feature	Importance
1	Canton	0.487548
0	Provincia	0.204448
2	Año	0.193677
6	Población Canton	0.045450
5	Poblacion Provincia	0.026559
3	Tasa de Desempleo Provincia	0.025585
4	Tasa de Desempleo Canton	0.016732

Por último, la visualización del árbol de decisión, indica los resultados proporcionados, donde el cada conjunto de valores representa las características de un nodo de términos de impureza Gini, el número de muestras de cada nodo y la distribución de las clases. Dado la amplitud de este árbol de decisión, debido al gran cantidad de datos, el archivo independiente puede visualizarse bajo un formato específico (Anexo 3)[4].

[4] Para entender mejor la interpretación de los resultados se sugiere revisar https://towardsdatascience.com/interpreting-random-forests-638bca8b49ea

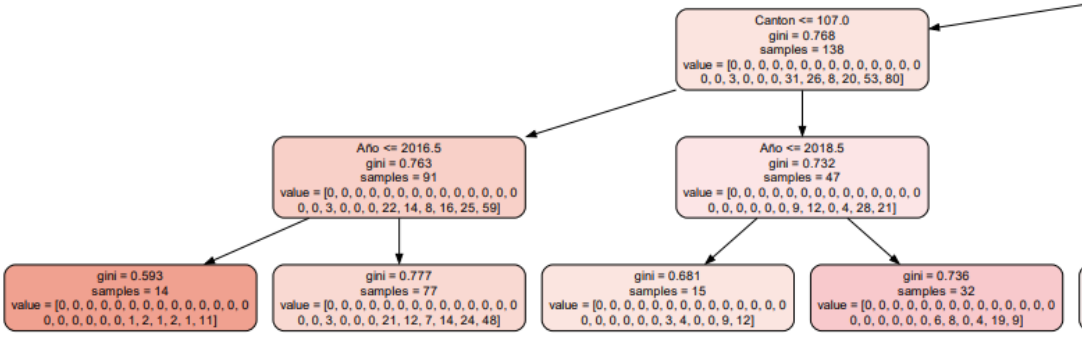


Gráfico 14. Fragmento del último nodo del Árbol de decisión

En la ilustración se puede ver el casillero superior "Canton <= 107.0" , el cual pertenece al último nodo izquierdo en un árbol de decisión. La métrica Gini de 0.768 indica la impureza del nodo, es decir un valor cercano a 1 sugiere una mayor impureza. También se puede ver que el nodo se construyó con 138 muestras y se distribuye en varias clases "value". Por ejemplo, hay 31 muestras de una clase, 26 de otra, y así sucesivamente.

4. Análisis de resultados y conclusiones

La situación de seguridad en Ecuador es, sin duda, preocupante y requiere una atención inmediata y sostenida. La conclusión que se puede extraer de la información proporcionada es que Ecuador ha pasado de ser una mal llamada “isla de paz” a enfrentar una crisis de seguridad significativa, con un aumento del 93.2% en la violencia criminal para el año 2021. La tasa de criminalidad se ha establecido en 6.25, lo que indica una situación alarmante. Es esencial que se adopten medidas basadas en un análisis crítico y sistemático de los datos disponibles. La implementación de modelos predictivos y el análisis de datos abiertos son fundamentales para comprender las causas subyacentes de la inseguridad y formular estrategias efectivas de prevención del crimen. Esto no solo ayudará a mejorar la seguridad

ciudadana, sino que también contribuirá a restaurar la confianza en las instituciones democráticas, fortalecer la democracia y preservar la estabilidad social en Ecuador.

La prueba de correlación que se realizó no arrojó resultados significativos, la baja correlación entre delitos y otras variables, reflejó valores cercanos a 0, esto significa que los factores que contribuyen al delito no pueden ser explicados únicamente a través de las variables seleccionadas. Se destaca que el modelo construido presenta limitaciones significativas en su capacidad predictiva, logrando predecir únicamente el 18% de los resultados probables. Esta baja eficiencia se atribuye principalmente a la falta de calidad de los datos recopilados. Se observa que las columnas con menos valores nulos, es decir, aquellas que contenían información disponible, exhibieron una mayor relevancia en el modelo. La decisión de retirar completamente las columnas de Pobreza por cantón y provincia del dataframe se tomó debido a su impacto significativo en los resultados y su contribución desfavorable al desempeño del modelo, especialmente considerando la presencia de varios valores nulos en esta característica.

Se evidencia la insuficiencia de datos demográficos proporcionados, ya que factores cruciales como el lugar del crimen, la hora, la descripción detallada del crimen, el móvil y el distrito no estaban disponibles, limitando considerablemente la información con la que se trabajó. Esto afectó la capacidad del modelo para capturar la complejidad del fenómeno delictivo. En estudios referentes, estas características también son tomadas en cuenta y además permiten crear mapas de concentración del crimen en una ciudad, lo cual es un gran aporte.

La selección de variables desde el inicio presentó desafíos, la integridad de las variables contribuyó a la dificultad en la construcción de un modelo robusto y preciso. Estos hallazgos

subrayan la importancia de la calidad y la diversidad de los datos abiertos en la construcción de modelos predictivos, así como la necesidad de considerar cuidadosamente la selección de variables para obtener resultados más confiables y generalizables.

La profunda evaluación de la evolución de los delitos violentos destaca un notorio incremento desde el año 2022, manteniéndose constante hasta la fecha límite en agosto de 2023. Entre las infracciones más recurrentes, las lesiones lideran con más de diez mil casos, seguidas por el robo con más de ocho mil registros. La tenencia y porte de armas ocupan el tercer lugar, mientras que la delincuencia organizada se aproxima a los dos mil casos. En cuanto a la distribución geográfica, Guayas encabeza la lista de casos de delitos violentos, seguido por Manabí y Pichincha. Quito y Guayaquil, reflejo de su elevada densidad poblacional, son las ciudades con mayor incidencia delictiva. La evolución de los delitos violentos revela un notable aumento en Quito en 2023, consolidándose como la ciudad más afectada y potencialmente la más peligrosa. Guayas lidera en delitos vinculados al narcotráfico, sugiriendo que es la provincia más afectada a nivel nacional. Estos hallazgos subrayan la necesidad de medidas preventivas y estrategias específicas para abordar la creciente problemática delictiva.

En conclusión, el modelo de Random Forest utilizado para predecir crímenes muestra una baja efectividad de precisión. La eliminación de las columnas, destaca la importancia de gestionar variables con valores nulos. Si bien se logró realizar series históricas sobre la distribución del delito geográficamente a escala nacional, los resultados finales también subrayan la importancia de la diversidad de los datos abiertos desde las instituciones estatales

para el uso de la ciudadanía y sectores académicos que busquen realizar futuras investigaciones de este tipo.

5. Bibliografía

Asor, Jonard, et al. "Building Model for Crime Pattern Analysis Through Machine Learning Using Predictive Analytics." International Journal of Science, Technology, Engineering and Mathematics, vol. 2, no. 1, Mar. 2022, pp. 61-73.

Aziz, Rabia Musheer; Hussain, Aftab; Sharma, Prajwal; and Kumar, Pavan (2022) "Machine Learning-based Soft Computing Regression Analysis Approach for Crime Data Prediction," Karbala International Journal of Modern Science: Vol. 8: Iss. 1, Article 1. Available at: https://doi.org/10.33640/2405-609X.3197

Balu, V., Navya Sri, T., & Bupathi, M. A. (2022). Crime Prediction and Analysis Using Machine Learning. *International Journal of Computer Science and Mobile Computing*, *11*(3), 95–101. https://doi.org/10.47760/ijcsmc.2022.v11i03.011

Miguel Marinho Saraiva, Matijošaitienė, I., Mishra, S., & Amante, A. (2022). Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics. *ISPRS International Journal of Geo-Information*, *11*(7), 400–400. https://doi.org/10.3390/ijgi11070400

Ministerio del Interior. (2023). *Estrategia nacional contra la delincuencia organizada 2023 - 2030.* Presidencia de la República , UNODC, UE, Quito Ecuador.

Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1). https://doi.org/10.1186/s42492-021-00075-z

Shohan, Faisal Tareque, Akash, Abu Ubaida, Ibrahim, M., & Alam, M. S. (2022). *Crime Prediction using Machine Learning with a Novel Crime Dataset*. ArXiv.org. https://arxiv.org/abs/2211.01551

Wheeler, A.P., Steenbeek, W. Mapping the Risk Terrain for Crime Using Machine Learning. *J Quant Criminol* **37**, 445–480 (2021). https://doi.org/10.1007/s10940-020-09457-7

Wajiha Safat, Asghar, S., & Gillani, S. (2021). Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *IEEE Access*, *9*, 70080–70094. https://doi.org/10.1109/access.2021.3078117

Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., & Chen, J. (2022). Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, *94*, 101789–101789. https://doi.org/10.1016/j.compen

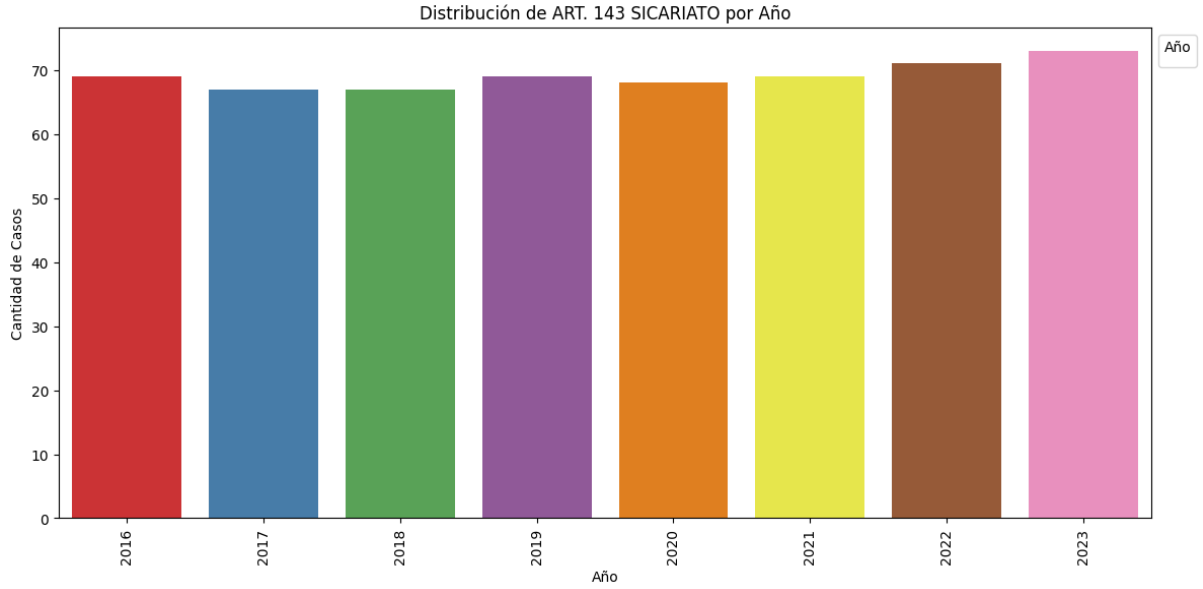
Anexos

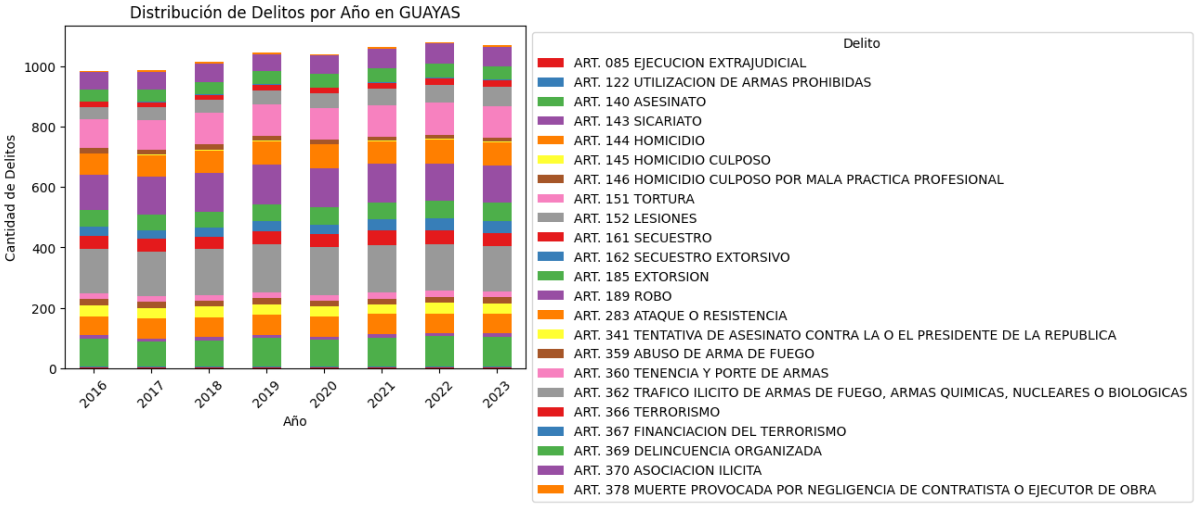
Anexo 1

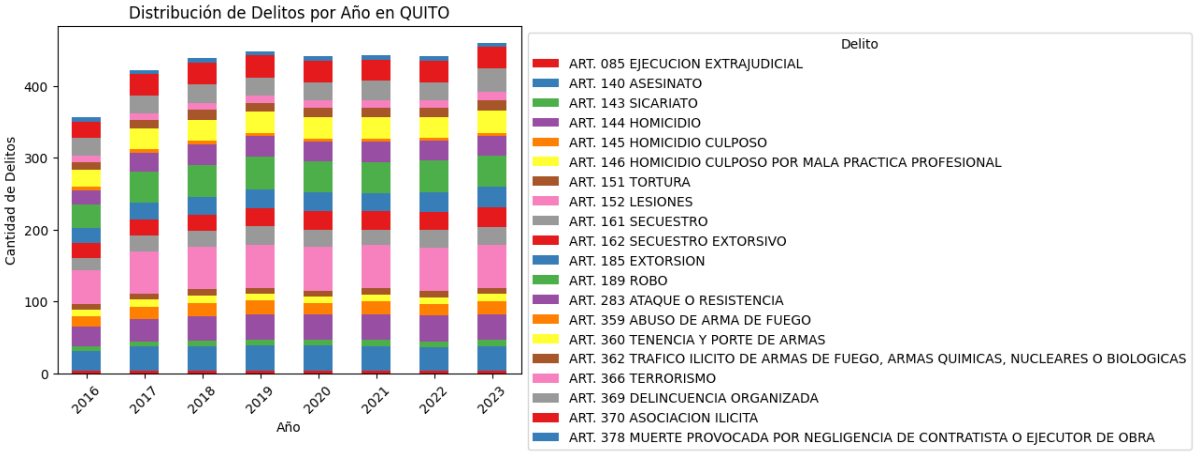
Delitos Violentos	
1. Art. 084 desaparición forzada	16. Art. 140 asesinato
2. Art. 161 secuestro	17. Art. 143 sicariato
3. Art. 162 secuestro extorsivo	18. Art. 341 tentativa de asesinato contra la o el presidente de la republica
4. Art. 185 extorsion	19. Art. 085 ejecución extrajudicial
5. Art. 151 tortura	20. Art. 146 homicidio culposo por mala práctica profesional
6. Art. 152 lesiones	
7. Art. 366 terrorismo	

8. Art. 369 delincuencia organizada 9. Art. 370 asociación ilícita 10. Art. 349 grupos subversivos 11. Art. 367 financiación del terrorismo 12. Art. 378 muerte provocada por negligencia de contratista o ejecutor de obra 13. Art. 115 homicidio de persona protegida 14. Art. 144 homicidio 15. Art. 145 homicidio culposo	21. Art. 122 utilización de armas prohibidas 22. Art. 359 abuso de arma de fuego 23. Art. 360 tenencia y porte de armas 24. Art. 189 robo 25. Art. 283 ataque o resistencia 26. Art. 361 armas de fuego, y explosivos no autorizados 27. Art. 362 tráfico ilícito de armas de fuego, armas químicas, nucleares o biológicas

Anexo 2




Distribución de Delitos por Año en GUAYAS
Cantidad de Delitos
Año
0
200
400
600
800
1000
2016
2017
2018
2019
2020
2021
2022
2023
Delito
ART. 085 EJECUCION EXTRAJUDICIAL
ART. 122 UTILIZACION DE ARMAS PROHIBIDAS
ART. 140 ASESINATO
ART. 143 SICARIATO
ART. 144 HOMICIDIO
ART. 145 HOMICIDIO CULPOSO
ART. 146 HOMICIDIO CULPOSO POR MALA PRACTICA PROFESIONAL
ART. 151 TORTURA
ART. 152 LESIONES
ART. 161 SECUESTRO
ART. 162 SECUESTRO EXTORSIVO
ART. 185 EXTORSION
ART. 189 ROBO
ART. 283 ATAQUE O RESISTENCIA
ART. 341 TENTATIVA DE ASESINATO CONTRA LA O EL PRESIDENTE DE LA REPUBLICA
ART. 359 ABUSO DE ARMA DE FUEGO
ART. 360 TENENCIA Y PORTE DE ARMAS
ART. 362 TRAFICO ILICITO DE ARMAS DE FUEGO, ARMAS QUIMICAS, NUCLEARES O BIOLOGICAS
ART. 366 TERRORISMO
ART. 367 FINANCIACION DEL TERRORISMO
ART. 369 DELINCUENCIA ORGANIZADA
ART. 370 ASOCIACION ILICITA
ART. 378 MUERTE PROVOCADA POR NEGLIGENCIA DE CONTRATISTA O EJECUTOR DE OBRA


Distribución de Delitos por Año en QUITO
Cantidad de Delitos
Año
0
100
200
300
400
2016
2017
2018
2019
2020
2021
2022
2023
Delito
ART. 085 EJECUCION EXTRAJUDICIAL
ART. 140 ASESINATO
ART. 143 SICARIATO
ART. 144 HOMICIDIO
ART. 145 HOMICIDIO CULPOSO
ART. 146 HOMICIDIO CULPOSO POR MALA PRACTICA PROFESIONAL
ART. 151 TORTURA
ART. 152 LESIONES
ART. 161 SECUESTRO
ART. 162 SECUESTRO EXTORSIVO
ART. 185 EXTORSION
ART. 189 ROBO
ART. 283 ATAQUE O RESISTENCIA
ART. 359 ABUSO DE ARMA DE FUEGO
ART. 360 TENENCIA Y PORTE DE ARMAS
ART. 362 TRAFICO ILICITO DE ARMAS DE FUEGO, ARMAS QUIMICAS, NUCLEARES O BIOLOGICAS
ART. 366 TERRORISMO
ART. 369 DELINCUENCIA ORGANIZADA
ART. 370 ASOCIACION ILICITA
ART. 378 MUERTE PROVOCADA POR NEGLIGENCIA DE CONTRATISTA O EJECUTOR DE OBRA