UNIVERSIDAD SAN PEDRO

ESCUELA DE POSGRADO

SECCION DE POSGRADO DE LA FACULTAD DE INGENIERIA

Modelo de predicción del estado delictivo de los distritos del Perú, 2020

Tesis para obtener el Grado Doctor en Estadística

Autor:

Alza Díaz José Alfredo

Código ORCID: 0009-0002-2072-3196

Asesor:

Sánchez Solorzano Roberto

Código ORCID: Código ORCID: 0000-0002-7689-961X

Chimbote – Perú

2024

Índice general

Índice de tablas

Índice de figuras

Palabras clave

Tema	:	Minería de datos, Estadística, Seguridad ciudadana.
Especialidad	:	Doctorado en Estadística.

Keywords

Thema	:	Data mining, Statistic, Public safety.
Specialy	:	Doctorate in Statistic.

Líneas de investigación

Línea de investigación	:	Gestión de organizaciones.
Área	:	Ingeniería, tecnología.
Subárea	:	Otros ingeniería y tecnologías
Disciplina	:	Ingeniería industrial.

USP UNIVERSIDAD SAN PEDRO | VICERRECTORADO DE INVESTIGACIÓN

CONSTANCIA DE ORIGINALIDAD

El que suscribe, Vicerrector de Investigación de la Universidad San Pedro:

HACE CONSTAR

Que, de la revisión del trabajo titulado **"Modelo de predicción del estado delictivo de los distritos del Perú, 2020"** del (a) estudiante: **ALZA DIAZ JOSE ALFREDO**, identificado(a) con Código Nº **3017100298**, se ha verificado un porcentaje de similitud del **20%**, el cual se encuentra dentro del parámetro establecido por la Universidad San Pedro mediante resolución de Consejo Universitario Nº 5037-2019-USP/CUpara la obtención de grados y títulos académicos de pre y posgrado, así como proyectos de investigación anual Docente.

Se expide la presente constancia para los fines pertinentes.

Chimbote, 01 de agosto de 2024

UNIVERSIDAD SAN PEDRO
VICERRECTORADO DE INVESTIGACIÓN

Dr. JAVIER MARTÍNEZ CARRIÓN
VICERRECTOR

Título

MODELO DE PREDICCIÓN DEL ESTADO DELICTIVO DE LOS DISTRITOS DEL PERÚ, 2020

Resumen

El estudio se llevó a cabo mediante un enfoque no experimental, transversal y exploratorio. La finalidad de la investigación consistió en identificar patrones delictuales y predecir la actividad delictiva de los distritos del Perú mediante técnicas de minería de datos.

En cuanto a la metodología utilizada, se realizó análisis exploratorio de datos, se transformó los atributos mediante el análisis de escalamiento multidimensional, se redujo la dimensionalidad del conjunto de datos mediante el análisis de componentes principales, y se agruparon los distritos según sus características mediante el análisis de clúster. Luego, se aplicaron algoritmos de árboles de decisión, naive bayes, k-vecinos más cercanos y máquinas de vectores de soporte para encontrar el mejor modelo de predicción de la actividad delictiva en los distritos.

Los resultados indican que el modelo de árboles de decisión es el más eficaz para predecir el estado delictivo de los distritos, con una precisión del 97%. Además, la investigación identifica 5 clúster de distritos organizados según el nivel de inseguridad y destaca 167 distritos con alta incidencia delictiva a nivel nacional. Se elabora un ranking de distritos según su actividad delictiva y, mediante la aplicación de sistemas de información geográfica, se logra visualizar la distribución espacial de la actividad delictiva en los distritos del país.

Abstract

The study was carried out through a non-experimental, cross-sectional, and exploratory approach. The purpose of the research was to identify criminal patterns and predict the criminal activity of Peruvian districts using data mining techniques.

Regarding the methodology employed, an exploratory data analysis was conducted, attributes were transformed through multidimensional scaling snalysis, the dimensionality of the dataset was reduced through principal component analysis, and districts were grouped based on their characteristics using cluster analysis. Subsequently, algorithms such as decision trees, naïve bayes, k-nearest neighbors, and support vector machines were applied to find the best predictive model for criminal activity in the districts.

The results indicate that the decision trees model is the most effective in predicting the criminal state of the districts, with an accuracy of 97%. Additionally, the research identifies 5 groups of districts organized according to the level of insecurity and highlights 167 districts with a high incidence of criminal activity at the national level. A ranking of districts is established based on their criminal activity, and through the application of GIS techniques, the spatial distribution of criminal activity in the districts of the country is visualized.

Introducción

Considerando que la seguridad ciudadana es un problema multidimensional que afecta la calidad de vida y bienestar del poblador peruano, nuestros esfuerzos para enfrentar este problema deben partir de la revisión de información estadística confiable, que nos permita conocer la realidad con precisión y abordar con mejor criterio dicho problema. En este sentido, la presente investigación, empleó herramientas de análisis de datos que permitan dimensionar el problema, analizar y evaluar indicadores de seguridad ciudadana, y diseñar intervenciones mejor focalizadas y con mayor nivel de efectividad. En este marco se consultó diversas investigaciones relacionadas a seguridad ciudadana que aplicaron técnicas de minería de datos con la finalidad de descubrir patrones subyacentes y realizar predicciones a partir de la data analizada, también se consultó bibliografía sobre metodologías para visualizar la distribución espacial de entidades respecto a indicadores de seguridad ciudadana, de este modo, se consultó las investigaciones siguientes:

Maggi (2023) a través de su investigación de tipo descriptivo explicó el control delictivo que realiza la Policía de Ecuador en la ciudad de Milagro los últimos 5 años en la cual se representa que ciertos patrones del crimen y comportamientos del delincuente variaron notablemente en términos de violencia (robo armado y homicidio) justificando el uso de modernas herramientas de software que permitan a los cuerpos policiales una mejor gestión en la seguridad ciudadana poniendo de ejemplo algunos sistemas informáticos usados como PredPol Skala demostrando la necesidad del uso de recursos técnicos informáticos especializados para desarrollar y anticiparse a hechos con probabilidades altas de ocurrir resaltando su aplicación de estrategias de prevención de hechos delictivos.

Colina (2022) propuso hacer un reconocimiento de la realidad de los datos abiertos en el Ecuador sobre delincuencia, y del proceso de minería de datos, utilizando herramientas de análisis de datos como Pentaho y Orange. Siguiendo el proceso de KDD (Descubrimiento de Conocimiento en Bases de Datos) para desarrollar el proceso de análisis de datos criminales y la correspondiente identificación de patrones relacionados con los delitos, permitiendo identificar la potencial utilidad de esta

herramienta para la exploración y detección de patrones delictivos y su consecuente beneficio en el poder de decisión de organismos competentes. Esto es especialmente relevante ante la poca disponibilidad de datos abiertos de interés social que puedan apoyar las estrategias de prevención de problemas sociales, permitiendo la identificación temprana y caracterización de amenazas presentes.

Oviedo (2022) analizó información de las denuncias de la Policía Judicial de Guayaquil utilizando algoritmos de clasificación como random forest, decision tree y netbayes, mediante el programa KNIME. La investigación se realizó en el marco de la clasificación del conocimiento y diseño de políticas de seguridad, produciendo un diagnóstico fiable de los delitos más comunes según la Policía de Guayaquil logrando así conocer los estratos de la población y analizar los factores que catalizan los índices delincuenciales dirigidos a mejorar y obtener reglas concretas para que sean analizadas por gestores en seguridad en la creación de estrategias para combatir la delincuencia organizada.

Ogbobe, y Okoronkw (2021) en su investigación titulada "Analysis of Crime Pattern using Data Mining Techniques", analizaron datos sobre delitos generados por el Sistema Integrado de Información sobre Delitos en Tiempo Real en apoyo a las Agencias de Aplicación de la Ley en las jurisdicciones de dos comisarías de policía en el sudeste de Nigeria. Esta investigación utilizó algoritmos de clasificación y reglas de asociación, con lo cual se logró identificar tendencias y patrones delictivos en dichas jurisdicciones.

Chaure (2021) en su estudio de criminología analizo las posibilidades de explotación que presentan las distintas tecnologías derivadas del desarrollo de la inteligencia artificial (IA) en su aplicación a las ciencias penales. En su revisión bibliográfica enfatizo la IA como un concepto englobante de las tecnologías se están utilizando para tratar de predecir ciertas características delictivas o la ocurrencia de delitos, así como el riesgo de reincidencia delictiva (Machine Learning, Big Data, Data Mining). Adicionalmente expone sobre la Policía predictiva en utilidad para las múltiples posibilidades de desarrollo, en especial el ámbito jurídico, otros como la industria y los servicios financieros, y, en concreto, las ciencias penales ofrecen una

oportunidad excelente para poner a prueba algoritmos y sistemas informáticos para el desarrollo de la sociedad.

Bhagat, y Shah (2021) llevaron a cabo un estudio que describen los principales resultados de investigaciones que predicen el crimen. En este contexto, los autores recomiendan el uso de algoritmos de machine learning y técnicas de visión por computadora en agencias legales para detectar, prevenir y resolver delitos a un ritmo mucho más preciso y rápido, con la finalidad de ayudar a los agentes de la policía a aliviar la carga de trabajo y a mejorar la prevención del delito. En la investigación, también se realizó un análisis del desempeño de los algoritmos de predicción utilizados en diferentes investigaciones para predecir el crimen, tales como: Árboles de decisión, k vecinos más cercanos, naive bayes, regresión, máquinas de vectores de soporte y bosque aleatorio. En dicho análisis, destaca el modelo de k vecinos más cercanos con un rendimiento de 0.87, naive bayes con 0.87 y máquinas de vectores de Soporte 0.84.

Muñoz (2021) llevo a cabo un análisis descriptivo y exploratorio de múltiples fuentes de información en la ciudad de Medellín y expuso la necesidad de desarrollar estrategias preventivas de vigilancia y control de los espacios públicos mediante el uso de tecnologías de aprendizaje automático (machine learning) para la predicción del crimen, integrando los resultados al desarrollo de la metodología CRISP-DM (Proceso estándar de la industria para la Minería de Datos) para la construcción de modelos predictivos aplicados en la ciudad de Medellín. El hurto a personas fue el delito seleccionado, específicamente en la modalidad de atraco, descuido y raponazo. Dentro de este proceso se definieron estrategias de priorización sobre diversos tipos de delitos, que se podrían replicar en otras ciudades; para que el desarrollo de algoritmos de predicción sea confiable, es de gran importancia un adecuado registro de la información y el procesamiento de los datos.

Mosquera (2021), con el objetivo de construir un modelo para predecir modalidades de crimen y el número de homicidios en la ciudad de Medellín, realizó una recopilación de información sobre el estudio del crimen y las principales técnicas de aprendizaje automático utilizadas para la predicción en una ciudad. A su vez, realizó un análisis descriptivo frecuentista y espacial para ofrecer una visión de la dinámica

del crimen en Medellín. Este análisis incluyó la clasificación de los tipos de delitos utilizando diversas técnicas de aprendizaje automático, como el logit multinomial, K-vecinos más cercanos, máquinas de vectores de soporte, bosques aleatorios y gradient boosting. El resultado indicó que el mejor modelo en términos de capacidad predictiva fue gradient goosting, con un 74.5% de precisión. Además, se destacó la importancia de disponer de información detallada para la clasificación de tipos de crímenes en una ciudad, ya que los delitos no son excluyentes; es decir, la ocurrencia de un delito no impide la ocurrencia de otro.

Según Norouzi y Ataei (2021), en su investigación "Application of data mining in identifying and discovering hidden patterns of theft" analizó una base de datos de 102 273 registros de victimarios que cometieron el delito de hurto durante los años 2010 y 2020 en Iran, el objetivo de la investigación fue encontrar relaciones ocultas en la base de datos, identificar y descubrir las reglas y patrones del delito de robo, con la finalidad de que la policía pueda predecir la ocurrencia de delitos y realizar tareas preventivas. En dicha investigación se utilizó técnicas de agrupamiento y reglas relacionales. Los datos se agruparon utilizando el método ji cuadrado medio y mediante el índice de Davis-Boldin se evaluó los diferentes modos de agrupación y se obtuvieron ocho agrupaciones como el mejor número de agrupaciones para los datos analizados.

Ordoñez (2020) nos presentó un modelo de machine learning basado en máquinas de soporte vectorial para regresión ajustado para predecir el aumento o disminución de actos delictivos (cantidad de hurtos) en Colombia y en sus 3 principales ciudades: Bogotá, Medellín y Cali; este modelo fue validado con un dataset tomado del sistema de información de la fiscalía nacional de Colombia. Los resultados de su ejecución con datos de las 3 ciudades tienen un comportamiento muy similar al comportamiento nacional en general, deduciendo que las estrategias y prácticas en seguridad pueden tener resultados similares, así se espera a futuro que el modelo pueda predecir el comportamiento de otros tipos de delitos.

Aravindan, Anusuya, y Ashok (2020) realizaron una predicción de la tasa de criminalidad en sectores policiales de Nigeria mediante algoritmos de Machine

Learning a partir de las denuncias de delitos de la policía. El procedimiento utilizó el 70 % de los datos para entrenar el modelo y el 30 % para evaluar su rendimiento. Los algoritmos aplicados fueron: Regresión logística, bosques aleatorios, vecinos más cercanos, clasificador de árboles de decisión, clasificador de bosques aleatorios y clasificador de vectores de soporte. El algoritmo que registró el mejor desempeño en la predicción de la tasa de criminalidad fue el modelo de regresión logística, el cual logró una precisión del 0.79.

Mangara (2020) analizó información de delitos del 2012 al 2015 en 47 condados de Kenia, el estudio aplicó técnicas de minería de datos como: agrupamiento k-means y algoritmo a priori con la finalidad de determinar grupos y reglas de asociación entre los diferentes delitos y condados de Kenia; en el estudio se determinó que los delitos decrecen a lo largo de los años analizados observándose concentración de delitos en los condados con mayor población.

Giraldo (2020) analizó 39 863 registros de víctimas de secuestros recopilados por la fiscalía nacional de Colombia para predecir la tendencia de secuestros anuales en los próximos 10 años. La investigación determinó que el algoritmo de redes neuronales presentó mayor precisión en la predicción que el algoritmo de máquinas de soporte vectorial.

Licona (2018) en su análisis cuantitativo, ejecuto la recolección de datos estadísticos para establecer patrones de comportamientos mediante la técnica de minería de datos con el fin de identificar zonas críticas y lugares con mayor probabilidad de ocurrencia de actos delictivos examinando 5 tipos de hurtos (automotores, motocicletas, residencias, establecimientos comerciales y entidades financieras), ello a su vez sirvió en la caracterización de hurtos que ocurren en el centro de la ciudad de Cartagena permitiendo a los ciudadanos estar informados a través de la divulgación del aplicativo web generado gracias al modelo matemático planteado para que pudieran tener una base sólida para tomar decisiones relacionadas con el lugar de vivienda, zonas a frecuentar, sectores peligrosos, entre otros aspectos que visualizan en el portal virtual del mismo aplicativo a través de un mapa geoespacial.

Deepika y Smitha (2018) realizan un análisis delictivo de estados indios

durante los años 2001 a 2012, en la investigación proponen un procedimiento para detectar la incidencia delictiva en la India utilizando metodologías de minería de datos, el planteamiento consta de los siguientes pasos: preprocesamiento, agrupación, clasificación y visualización de datos. Los delitos se identifican aplicando el procedimiento k-media, dicho procedimiento forma grupos en función de la similitud de los atributos delictivos. Posteriormente clasifica los ámbitos administrativos investigados mediante los algoritmos bosque aleatorio y redes neuronales. Finalmente logra visualizar los lugares de alta incidencia delictiva mediante la agrupación de marcadores de google en el mapa de la India.

Según el Latinobarómetro el estado delictivo del Perú ha sido reflejado como un producto de constantes vaivenes en los niveles de actos de violencia y de percepción de inseguridad, estos indicadores presentan una tendencia similar tanto a nivel regional como distrital. La OMS define la violencia como *"el uso intencional de la fuerza o el poder físico, de hecho o como amenaza, contra uno mismo, otra persona o un grupo o comunidad, que cause o tenga muchas probabilidades de causar lesiones, muerte, daños psicológicos, trastornos del desarrollo o privaciones"* (OMS, 2002, pág. 3). Complementariamente se concibe la violencia como toda acción u omisión que pueda interrumpir, impedir o retardar el desarrollo saludable de los seres humanos (Koller y De Antony) ya sea por exposición directa (víctima) o indirecta (testigo) mientras que autores como Galtung sostienen que "*la violencia es tan solo una arista de un conflicto que comprende subdivisiones en lo cultural, lo estructural y lo directo, las cuales son la afrenta evitable a las necesidades humanas y busca generar daño; en otras palabras, la violencia es entendida como el uso o amenaza de uso de la fuerza con la finalidad de afectar la integridad física, moral o psicológica del otro"* (Niño, 2020, pág. 209).

La violencia al ser un concepto de abordaje interdisciplinario puede moldearse desde distintos puntos de vista tanto para los agentes tomadores de decisiones, analistas, académicos, operadores de justicia, etc. Por su carácter general termina siendo de gran utilidad para elaborar un concepto-eje capaz de responder al abordaje de conductas delictivas como su objeto de estudio y según su criterio metodológico,

institucional, etc. siendo adecuado para un análisis claro, objetivo, exacto y operativo; ya que no existe una definición única ni universal para su aplicación sino desde múltiples enfoques.

En medio del debate sobre la categoría "Violencia delictiva" cuyo principal problema conceptual sería que contempla prácticas que son ilegales y violentas desde la perspectiva del sistema jurídico vigente, pero no desde el punto de vista de los actores (Isla, 2008) planteando de esta manera que para entender la violencia se debe ver a está interrelacionada con los procesos políticos, económicos y culturales (similar a la llamada violencia social) sobre los que están asentados los funcionamientos de otras instituciones como las cárceles, los centros juveniles, la policía y el sistema judicial.

Los países de América Latina y el Caribe (ALC) son considerados entre los de mayor violencia e inseguridad a nivel mundial, estos países presentan indicadores como la tasa de homicidios que triplican el promedio mundial (Muggah, 2017). A pesar de que la región de ALC ha logrado un buen desempeño económico y social en los últimos diez años, los indicadores de inseguridad como homicidios, percepción y victimización muestran cifras muy elevadas (Banco Interamericano de Desarrollo, 2021).

La Organización Mundial de la Salud (OMS) considera que un país tiene muy alta violencia cuando la tasa de homicidios es superior a 31 porcada 100,000 habitantes, en este sentido la región tiene en esta categoría a los países: El Salvador, Jamaica, Honduras y Venezuela. De los 50 países que registran las más elevadas tasas de homicidios a nivel mundial la región de ALC participa con 42, la misma preocupante situación se presenta con los indicadores que miden violencia familiar y contra la mujer (Banco Interamericano de Desarrollo, 2021).

Si bien los países ubicados en las zonas de Centroamérica y el Caribe presentan las tasas de homicidios más elevadas de la región, países de América del Sur, como Argentina, Chile, Paraguay y Perú, registran tasas de homicidios bajas, pero elevadas tasas de robos y hurtos, es decir, altos índices en la tasa de victimización, se estima en el último año que de cinco personas una fue víctima de robo y de 10 robos 5 se

realizaron con violencia (Banco Interamericano de Desarrollo, 2018).

El incremento de la violencia, la presencia del crimen organizado y la aparente incompetencia del estado para hacer frente a la incidencia criminal y la violencia se han convertido en uno de los vitales ejes de estudio del escenario que se vive hoy en América Latina (Dammert, 2017). El Perú no es ajeno a esta realidad, las noticias a diario están colmadas de robos, homicidios, feminicidios, estadísticas de seguridad ciudadana indican que la tasa de homicidios a nivel nacional se incrementó de 5.4 en el año 2011 a 5,8 víctimas por cada 100 mil habitantes en el año 2020, situación más complicada se presenta en ciudades como Barranca, Tumbes, Huaral, Huánuco y Pisco lugares donde los homicidios durante el año 2020 registraron tasas de 10.8, 22.9, 9.4, 6.8 y 14.0 homicidios por cada 100,000 habitante respectivamente, ciudades que por sus altos niveles de violencia requieren acciones focalizadas de lucha contra la inseguridad (Comité Estadístico Interinstitucional de la Criminalidad, 2020).

Otro indicador importante de inseguridad es la victimización cuyo objetivo es conocer si la población de 15 y más años del área urbana fue víctima de algún hecho delictivo en los últimos 12 meses antes de realizada la encuesta, este indicador en el año 2022 resultó 22.9%, es decir 23 de cada 100 peruanos fueron víctimas de algún hecho delictivo. Realizando una comparación frente al año 2019 (26.6%) observamos un decrecimiento de 3.7 puntos porcentuales (Instituto Nacional de Estadística e Informática, 2022).

Respecto a los hechos delictivos efectuados a nivel nacional, el más frecuente es el robo de dinero, cartera o celular que durante el año 2022 representó el 12.9%, menor al 17.3% de ciudadanos que fueron víctimas en al año 2019. En referencia al delito de estafa 3.1 % se vio afectado por este tipo de situación durante el año 2022, es decir, aumentó en 0.7% en comparación del mismo periodo del año 2019 (Instituto Nacional de Estadística e Informática, 2022).

Para operativizar los indicadores de criminalidad a través de las cifras de delitos registrados por los instrumentos mencionados se requiere cierta rigurosidad e innovación metodológica para sistematizar conforme a los lineamientos de la disciplina criminológica y en el proceso de investigación, clasificando según perfiles

de los distritos analizados para aproximarse a la elaboración de un modelo predictivo del delito, que sea capaz de realizar un proceso de simulación para medir y responder ante situaciones y condiciones específicas de lo que tipificamos como “distrito inseguro”

Partimos desde la concepción de que *"la seguridad es un bien público y el Estado, en especial sus fuerzas de policía, deben garantizar que el acceso a la seguridad sea justo y equitativamente distribuido"* (PNUD, 2013) revisamos y evaluamos (en torno al creciente estado delictivo) las metodologías así como las buenas prácticas que han sido empleadas por entidades que han incidido en materia de seguridad ciudadana para resaltar que el Estado es el único actor que tiene la responsabilidad de asegurar la provisión de seguridad ciudadana como un bien público.

El concepto de seguridad ciudadana se desprende de "seguridad nacional" tras superar contextos turbulentos de transición democrática en nuestra región, destacando la seguridad ciudadana en un marco de respeto a los derechos humanos, se la define como la capacidad de los Estados, en asocio con el sector privado, los particulares, la academia y asociaciones comunitarias, vecinales y ciudadanas, de proveer y coproducir un marco de protección de la vida y el patrimonio de los individuos, que les permita a los ciudadanos convivir pacíficamente, sin miedo, en aras de alcanzar una mejor calidad de vida. (Banco Interamericano de Desarrollo, 2018).

La inseguridad diferenciada del concepto del riesgo de Ulrich Beck, como un producto de la consciencia de peligros e inseguridades frente a procesos políticos, económicos, etc. en el contexto de modernización con el fin de sobrevivir (Beck, 2002), en dicho cálculo del riesgo para su anticipación se resalta el ámbito reflexivo del concepto ya que es generado por nosotros mismos al percibir como propia la amenaza desde un proceso cognitivo dentro del plano intersubjetivo, mientras que el término de “inseguridad” (surge desde la noción anglosajona de “miedo al crimen”) hacia determinados grupos sociales, al llegar a ser un problema público en especial por tratarse de una amenaza a la integridad física sin motivo específico, de esta forma se delimita la clara división entre “nosotros” (potencial víctima) y un “otros” (persona amenazante).

Sin embargo en el contexto de la puesta en marcha de políticas públicas para la lucha contra los crecientes índices de criminalidad e inseguridad, el Estado se ha visto superado por la dinámica actual del fenómeno delictivo producto de su deficiente respuesta para controlar el nivel de criminalidad hasta la gestión de información en materia de seguridad ciudadana dada la carencia o insuficiencia de estudios estadísticos que emplean recursos tecnológicos como algoritmos vinculados a la inteligencia artificial que permiten establecer estrategias de acción que van desde almacenar datos, operativizar gran cantidad de información, aplicar métodos predictivos con significativos niveles de precisión y visualización de resultados en mapas temáticos a nivel de distritos elaborados en base a SIG (Sistemas de Información Geográfica), resaltando la urgencia de promover la búsqueda de enfoques innovadores utilizando software estadísticos y otros programas relacionados al análisis de datos.

Entre uno de estos enfoques está el análisis predictivo, una parte de la Data Mining (minería de datos) cuyo fin es extraer información almacenada en los datos y utilizarla para hallar patrones de comportamiento utilizables en circunstancias con probabilidades de ocurrir a lo largo del tiempo. En este proceso se emplea el aprendizaje automático (Machine Learning), procesos estadísticos y bases de datos.

Machine Learning definida por Maggi (2023) como *el "campo de investigación que brinda a las computadoras la capacidad de aprender sin programación explícita. Primordialmente se usa para simplificar problemas que necesitan muchos ajustes de forma manual o extensas listas de requisitos, y para que la información sea obtenida, está sobre problemas complicados y con grandes cantidades de datos".* (Maggi, 2023, pág. 33)

Particularmente este mecanismo responde al problema de clasificación supervisada del análisis predictivo utilizando datos a través de algoritmos computacionales para realizar acciones como clasificar la información en diferentes categorías obteniendo resultados óptimos en el descubrimiento de patrones en el comportamiento de hechos delictivos. Ejecutado el algoritmo de Machine Learning, el modelo toma una entrada "A" de un conjunto de datos y produce una salida "B", el

modelo AB se fija en la fase de entrenamiento. De esta forma los modelos predictivos se utilizan para predecir qué harán las personas en una situación determinada, por ejemplo, si cambiarán sus hábitos o si comprarán un producto o servicio en particular. Los datos de una persona se introducen en el modelo y se obtiene una clasificación que indica la probabilidad de que el modelo investigue la situación (Maggi, 2023).

Dentro de esta línea de Análisis predictivo del crimen, también se encuentra el Big Data que imita el modus operandi de otros modelos predictivos de software basadas en la búsqueda de correlación y patrones en base a la probabilidad permitiendo optimizar la capacidad de los dispositivos digitales para poder procesar, almacenar grandes volúmenes de datos y ejecutar instrucciones dictadas de forma automática con el fin de obtener nueva información tras haber producido una salida (output) tomando previamente una entrada (input) de un conjunto de datos generando así una mayor necesidad de técnicas analíticas, algoritmos, enfoques sofisticados para las investigaciones estadísticas empleadas en la lucha contra el crimen.

En los informes de diagnóstico de criminalidad de un país se busca realizar análisis delictivos con información histórica fiable para el seguimiento de la evolución de la frecuencia de delitos y así dar cuenta de las variaciones de niveles delictivos en las diferentes regiones del país con la finalidad de justificar una intervención policial focalizada ajustándose a las necesidades que presenta un determinado distrito según su estado delictivo (Hot Spot Policing), sin embargo para una mejor intervención se necesita un modelo eficiente de predicción delictiva para estimar su situación futura y clasificación como inseguros o seguros, con la finalidad de realizar intervenciones diferenciadas en los distritos investigados.

Las cifras numéricas registradas como número de homicidios al año, tasa de victimización tradicionalmente han representado los indicadores privilegiados de la violencia y la inseguridad en términos generales así como fuentes de información para la generación de incidencia con la expectativa de contribuir en las políticas públicas intersectoriales, por ello, mediante la minería de datos se pueden desarrollar más funcionalidades para la generación de un modelo capaz de establecer una clasificación de distritos inseguros al crimen.

Revisando la aplicación de la noción de "seguridad" hacia el ámbito territorial incluimos analizar las encuestas de sentimientos de inseguridad en la cual se formula al encuestado la posibilidad de ser víctima de un robo en su zona de residencia, adicional a ello podemos rescatar de las teorías ecológicas de prevención del delito el concepto de "entornos seguros" planteado desde la corriente de Crime Prevention Through Environmental Design (Prevención del Crimen mediante diseño ambiental), según autores como Crowe, sostiene que desde el diseño del ambiente se puede generar una disminución de la oportunidad en la concreción de delitos a través de la modificación del espacio físico fortaleciendo 4 aspectos: vigilancia natural, acceso natural del espacio, refuerzo territorial y el mantenimiento-cuidado del espacio físico enfocado en la estrategia de producir efectos disuasivos en la conducta delictiva e impedir cualquier potencial actividad delictiva.

Exploramos un antecedente retratado con el concepto de "barrio inseguro", el cual no se pudo establecer un consenso en su definición ya que los autores consideran combinar diversas variables personales sociodemográficas (edad, género) incluyendo otras que van desde las relaciones sociales, presencia del estado, desorden físico, existencia de pandillas delictivas, etc. estableciendo modelos los cuales no se les podía circunscribir a determinados territorios una diferente frecuencia de delitos. En síntesis, se trataría de la percepción del problema del delito presente en la ciudad, que incluye también a la forma, textura y salud atribuible a las estructuras sociales y políticas de la urbe que termina constituyéndose en una amenaza a la integridad física, a la propiedad privada o a que otras personas cercanas puedan sufrir similar experiencia (Nuñez & Tocornal, 2012).

Tras ver el problema de consenso sobre la especificidad y formas de medición de la percepción de seguridad para territorios; en la presente investigación se optó por operativizar la definición de "Distrito inseguro" considerando las variables que expresan su medición operacional: Número de efectivos de la PNP, número de efectivos de serenazgo, muertes violentas asociadas a hechos delictivos dolosos, denuncias de delitos y faltas registradas por la PNP, último distrito de residencia del interno, y número de habitantes del distrito, data estadística oficial consolidada del

2020 proveniente de entidades como SIDPOL (Sistema de denuncias policiales de la Policía Nacional del Perú) y el INPE.

La presente investigación se realiza a través del diseño no experimental, transeccional y descriptivo. Emplea información estadística oficial, proveniente de censos encuestas y registros administrativos, desagregada a nivel distrital. El estudio tiene como objetivos identificar patrones delictivos y predecir el estado delictivo de los distritos del Perú mediante la aplicación de algoritmos de minería de datos; otro de los objetivos de la investigación es construir un ranking para identificar distritos inseguros y finalmente obtener la distribución espacial de la actividad delictiva a nivel distrital. Los resultados de la investigación son de mucha importancia a nivel de gobierno nacional y subnacional debido a que ayudará a tomar decisiones, planificar, orientar políticas y priorizar la ejecución de actividades en el sector interior.

A continuación, se presenta una breve descripción teórica de las técnicas empleadas para determinar el modelo de mejor precisión para predecir el estado delictivo de los distritos.

Árboles de Decisión

> Barrientos, Cruz, y Acosta (2009) refieren que árboles de decisión es una de las técnicas de aprendizaje supervisado no paramétrico, que se utiliza para la predicción, también se utiliza en el campo de la inteligencia artificial, donde se construyen diagramas de construcción lógica a partir de bases de datos, muy similares a los sistemas predictivos. Fineberg (1980) afirma que los árboles de decisión se basan en reglas que se utilizan para representar y clasificar un conjunto recurrente de condiciones para resolver un problema. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Al igual que con otros modelos de aprendizaje supervisado, las predicciones se realizan en función de un conjunto de variables de características y un umbral predeterminado, cuando se trata de árboles de decisión, se pueden usar múltiples umbrales; sin embargo, se debe tener en consideración el nivel de

impureza de la conformación de los nodos mediante el cálculo de la entropía y el índice de gini. Esta técnica tiene una representación intuitiva, los resultados se pueden expresar mediante reglas, son robustos a los outliers y son fáciles de ser estimados.

Un árbol de decisión particiona el espacio de variables predictoras en un conjunto de hiper-rectangulos y en cada uno de ellos ajusta un modelo sencillo, generalmente una constante. Es decir, $y = c$, donde y es la variable respuesta. La estimación de un árbol de decisión se basa en 4 elementos: un conjunto de preguntas binarias, el método usado para particionar los nodos, la estrategia requerida para particionar los nodos y la asignación de cada nodo terminal a una clase de la variable respuesta.

Coeficiente de gini: El índice de gini es la probabilidad de no clasificar correctamente cuando las variables se eligen al azar. El índice de gini tiende a favorecer particiones más grandes y, por lo tanto, es computacionalmente intensivo.

$$i_G(t) = \sum_{j=1}^{J} p(j|t)[1 - p(j|t)]$$

$$i_G(t) = 1 - \sum_{j=1}^{J} p(j|t)^2)$$

Coeficiente de entropía: La entropía es un método utilizado para dividir un árbol de decisión en subconjuntos más pequeños; al dividir el árbol, actúa como un umbral para los nodos del árbol. La entropía es una medida de la impureza de un conjunto de datos. La entropía es la suma de la probabilidad de ocurrencia de un valor multiplicada por el logaritmo en base dos de esa probabilidad. Dado que las probabilidades caen en el rango de 0 a 1, el valor agregado siempre es negativo, por lo que debe multiplicarse por -1.

$$i_E(t) = -\sum_{j=1}^{J} p(j|t)\log[p(j|t)]$$

Naive Bayes

En el aprendizaje automático, la clasificación Naive Bayes es un algoritmo sencillo y potente para la tarea de clasificación. El clasificador Naive Bayes se basa en la aplicación del teorema de Bayes con un supuesto de fuerte independencia entre las características. Los modelos de Naive Bayes también se conocen como Bayes simple o Bayes independiente.

El clasificador naive bayes utiliza el teorema de Bayes para predecir las probabilidades de pertenencia a cada clase, como la probabilidad de que un registro o punto de datos determinado pertenezca a una clase concreta. La clase con la probabilidad más alta se considera la más probable. Esto también se conoce como Máximo A Posteriori (MAP).

Características:

- Busca modelar la relación probabilística entre atributos y clases (Seguro/inseguro valores 0 y 1).
- Asume una distribución conjunta entre X y Y.
- Tiene atributos independientes dado la clase
- Dado un record con atributos $(D_1, D_2, \dots, D_n)$, el objetivo es predecir la clase H. (seguro / inseguro)

Estimación:

Bassett, y Deride (2018) menciona que el estimador MAP elige la clase con la probabilidad posterior más alta. Para la clasificación binaria, se trata simplemente de la probabilidad posterior positiva.

La regla de Bayes para este problema es de la forma:

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)}$$

Donde $P(H|D)$ es la "probabilidad posterior", es decir, la probabilidad de la hipótesis H para los datos dados D. $P(D|H)$ es la probabilidad de los datos D dada la hipótesis H es verdadera. $P(H)$ es la probabilidad previa de que H sea verdadera independientemente de los datos. $P(D)$ es la probabilidad de los datos independientemente de la hipótesis.

El MAP para una hipótesis con 2 eventos H y D es:

$$MAP(H) = \max P(H|D)$$

$$MAP(H) = \max \frac{P(D|H) * P(H)}{P(D)}$$

$$MAP(H) = \max P(D|H) * P(H)$$

$P(D)$ es la probabilidad de evidencia. Se utiliza para normalizar el resultado. Sigue siendo la misma, por lo que eliminarla no afectaría al resultado. El clasificador Naive Bayes asume que todas las características no están relacionadas entre sí. La presencia o ausencia de una característica no influye en la presencia o ausencia de cualquier otra característica. En los conjuntos de datos del mundo real, comprobamos una hipótesis a partir de múltiples pruebas sobre características. Por tanto, los cálculos se complican bastante. Para simplificar el trabajo, se utiliza el enfoque de la independencia de rasgos para desacoplar las pruebas múltiples y tratar cada una como independiente.

Máquinas de Vectores Soporte (SVM)

Boser (1992) afirma que las máquinas de vectores soporte son algoritmos de aprendizaje automático que se utilizan con fines de clasificación y regresión. Las SVM son uno de los algoritmos de aprendizaje automático más potentes para la clasificación, la regresión

y la detección de valores atípicos. Un clasificador SVM construye un modelo que asigna nuevos puntos de datos a una de las categorías dadas. Por lo tanto, puede considerarse como un clasificador lineal binario no probabilístico.

Un hiperplano es un límite de decisión que separa un conjunto dado de puntos de datos con diferentes etiquetas de clase. El clasificador SVM separa los puntos de datos utilizando un hiperplano con la máxima cantidad de margen. Este hiperplano se conoce como hiperplano de margen máximo y el clasificador lineal que define se conoce como clasificador de margen máximo.

Los vectores soporte son los puntos de datos de muestra más cercanos al hiperplano. Estos puntos de datos definirán mejor la línea de separación o el hiperplano calculando los márgenes.

Un margen es una distancia de separación entre las dos líneas sobre los puntos de datos más cercanos. Se calcula como la distancia perpendicular de la línea a los vectores de apoyo o puntos de datos más cercanos. En SVM, tratamos de maximizar esta brecha de separación para obtener el máximo margen.

Cada punto de entrenamiento $x_i \, \epsilon \, \mathcal{R}^2$ pertenece a alguna de dos clases y se le ha dado una etiqueta $y_i \, \epsilon \, \{-1,1\}$ para $i = 1, 2, \dots, I$ (Fletcher, 2009). En la mayoría de los casos, la búsqueda de un hiperplano adecuado en un espacio de entrada es demasiado restrictivo para ser de uso práctico. Una solución a esta situación es mapear el espacio de entrada en un espacio de características de una dimensión mayor y buscar el hiperplano óptimo allí. Deseamos encontrar el hiperplano.

$$wx + b = 0$$

Agregamos una restricción al problema

$$y_i(wx_i + b) \geq 1$$

El problema del hiperplano óptimo es entonces redefinido como la solución al problema

$\min_{w} \frac{1}{2} w^t w$; sujeto a: $y_i(wx_i + b) - 1 = 0$

Buscando el hiperplano óptimo, que puede ser resuelto construyendo un Lagrangiano y transformándolo en el dual

max $L_D = \sum_{i=1}^{I} \alpha_i + \sum_{i=1,j=1}^{I} \alpha_i \alpha_j y_i y_j x_i x_j$; sujeto a: $w = \sum_{i=1}^{I} \alpha_i y_i x_i$, $\sum_{i=1}^{I} \alpha_i y_i = 0$

Donde $\alpha = (\alpha_1, \ldots, \alpha_I)$ es un vector de multiplicadores de Lagrange positivos.

K Vecinos Más Cercanos

Es un algoritmo no paramétrico utilizado para tareas de clasificación y regresión. No paramétrico significa que no es necesario asumir la distribución de los datos. Por lo tanto, KNN (Fix & Hodges, 1951) requiere ninguna suposición subyacente. Tanto en tareas de clasificación como de regresión, la entrada consiste en los k ejemplos de entrenamiento más cercanos en el espacio de características. El resultado depende de si KNN se utiliza con fines de clasificación o de regresión

Beckmann, Ebecken, y Pires (2015) refiere que en la clasificación KNN, el resultado es la pertenencia a una clase. El punto de datos dado se clasifica en función del tipo mayoritario de sus vecinos. El punto de datos se asigna a la clase más frecuente entre sus k vecinos más cercanos. Normalmente, k es un número entero positivo pequeño. Si k=1, el punto de datos se asigna simplemente a la clase de ese único vecino más cercano.

En la regresión KNN, el resultado es simplemente un valor de propiedad del objeto. Este valor es la media de los valores de los k vecinos más cercanos.

KNN es un tipo de aprendizaje basado en instancias o aprendizaje perezoso. Aprendizaje perezoso significa que no requiere ningún punto de datos de entrenamiento para la generación del modelo. Todos los datos de entrenamiento se utilizarán en la fase de prueba. Esto hace que la formación sea más rápida y las pruebas más lentas y costosas. Por tanto, la fase de prueba requiere más tiempo y recursos de memoria.

En KNN, los vecinos se toman de un conjunto de objetos de los que se conoce la clase o el valor de la propiedad del objeto. Esto puede considerarse como el conjunto de entrenamiento para el algoritmo KNN, aunque no se requiere un paso de entrenamiento explícito. Tanto en el algoritmo KNN de clasificación como en el de regresión, podemos asignar un peso a las contribuciones de los vecinos. Así, los vecinos más cercanos contribuyen más a la media que los más lejanos.

Para clasificar cada muestra del conjunto de prueba, hay que realizar las siguientes operaciones en orden:

- Calcular la distancia a cada una de las muestras del conjunto de entrenamiento.
- Seleccionar k muestras del conjunto de entrenamiento con la distancia mínima a ellas.
- La clase de la muestra de prueba será la clase más frecuente entre esos k vecinos más cercanos.

El método se adapta con bastante facilidad al problema de la regresión: en el paso 3, no devuelve la clase, sino el número: una media (o mediana) de la variable objetivo entre los vecinos.

La calidad de la clasificación/regresión con KNN depende de varios parámetros:

- El número de vecinos k.
- La medida de distancia entre las muestras (las más comunes

son las distancias Hamming, Euclidiana, coseno y Minkowski). Tenga en cuenta que la mayoría de estas métricas requieren que los datos estén escalados.

- Pesos de los vecinos (cada vecino puede aportar distintos pesos).

Respecto a la justificación teórica de la investigación, de acuerdo con la revisión de bibliografía en el contexto nacional e internacional se ha podido evidenciar que mediante la aplicación de técnicas de minería de datos es posible generar un modelo eficiente de predicción delictiva para estimar la situación futura de los distritos, encontrar patrones, tendencias o reglas de clasificación; contando con dicha información las instituciones competentes podrán realizar intervenciones focalizadas y diferenciadas en distritos inseguros.

Así mismo, uno de los objetivos de la investigación es la construcción de un ranking que permita identificar los distritos más inseguros, desde el punto de vista de la justificación práctica, la presente investigación contribuirá a una adecuada focalización de la actividad delictivas en el Perú, dicha focalización permite establecer niveles de inseguridad en los distritos lo cual facilitará, a nivel de gobierno nacional y sub nacional la planificación, la orientación de políticas públicas de seguridad ciudadana, la toma de decisiones en el sector interior y la realización de una mejor asignación de recursos presupuestales para enfrentar la inseguridad en el país.

Considerando una justificación desde la percepción social, la presente investigación beneficia a la población por que a través de una adecuada clasificación e identificación de patrones delictivos a nivel de distritos las instituciones competentes podrá ejecutar con mejor criterio sus políticas, estrategias, programas y actividades para enfrentar la inseguridad, situación que impactará directamente en la población que

gozará de una mejor convivencia pacífica y disminución de la violencia en las vías y espacios públicos.

Adicionalmente la investigación brinda un importante aporte metodológico al análisis de datos de seguridad ciudadana considerando que aplica técnicas de minería de datos como algoritmos de aprendizaje supervisados de clasificación para predecir el estado delictivo de los distritos del Perú. Así mismo, es relevante el aporte metodológico en la aplicación de técnicas GIS en información de inseguridad ciudadana, lo cual permitirá elaborar mapas temáticos que muestren una clara visualización del comportamiento de las variables investigadas en una representación geográfica del Perú a nivel distrital. Estas metodologías podrán ser replicadas en los próximos años para realizar un seguimiento de las tendencias de los indicadores delictivos a nivel distrital.

Dentro de un marco analítico general que nos permite moldear múltiples relaciones existentes en torno a las diferentes variables de nuestro tema de investigación, y reforzando el establecimiento del modelo predictivo consideramos la implementación de modelos multivariantes con el fin de realizar análisis descriptivos y de frecuencias para las variables del estudio junto con procesos comparativos mediante softwares estadísticos para establecer la relación predictiva entre las variables, tras un respectivo tratamiento de data realizado a nivel multidimensional.

Según Meneses (2019, pág. 22) podemos definir el análisis multivariante como *el conjunto de técnicas estadísticas que tienen como objetivo analizar e interpretar las relaciones entre distintas variables de manera simultánea, mediante la construcción de modelos estadísticos complejos que permiten distinguir la contribución independiente de cada una de ellas en el sistema de relaciones y, de este modo, describir, explicar o predecir los fenómenos que son objeto de interés para la investigación*.

Considerando las diversas técnicas multivariantes y con el fin de analizar relaciones de interdependencia para describir la estructura de los datos elegimos utilizar las siguientes:

- El análisis de componentes principales, en medio de la identificación grupos de características similares de las diversas variables del estudio, reduciendo la complejidad de los datos a través de la obtención de un conjunto limitado de factores que permitiría representar eficientemente la variabilidad en las características de los individuos, de esta forma, conservando el máximo de la información recogida originalmente en las variables involucradas.
- El escalamiento multidimensional en medio de la identificación de grupos de objetos similares a partir de las valoraciones proporcionadas por los participantes de la investigación de acuerdo con sus percepciones de similitud aplicado tanto en variables cuantitativas como cualitativas.

A modo de complemento en el estudio de las relaciones de interdependencia también se emplea el análisis de conglomerados o análisis clúster para identificar grupos de individuos de características similares pudiendo ser aplicada a variables tanto cuantitativas como cualitativas.

Al implementar el análisis de escalamiento multidimensional (AEM) esta metodología nos permite determinar el número de dimensiones que mejor representa los datos ya que coadyuva en el desarrollo del modelo de referenciación de este estudio.

A su vez, con el fin de complementar el análisis con un nivel de visualización geográfica distrital que represente cartográficamente la interrelación de las variables del estudio para el modelo predictivo, optamos por el uso de un Sistema de Información Geográfica (SIG).

La configuración espacial del territorio resulta fundamental ya que la implementación de los SIG apoya en la toma de decisiones del ámbito de las ciencias políticas, en investigación en salud pública, incluyendo en criminología ya que los análisis de patrones que interrelacionan fenómenos sociales, económicos o ambientales con ocurrencias de delitos cuentan con una larga tradición en el uso del componente geográfico. Del Bosque (2012) señala que la perspectiva geográfica afecta de forma muy decisiva en los comportamientos de los individuos y de los grupos que lo habitan, observando que las técnicas geoespaciales de modelado, análisis y representación de datos georreferenciados favorecen la comprensión de las variables involucradas en las relaciones de los individuos configurados por características similares según pertenencia territorial, objeto e interdependencia describiendo así la variabilidad de la situación analizada en intensidad y magnitud para la construcción de dicho modelo.

Con las técnicas del SIG nos remitimos, según Víctor Olaya (2014, pág. 7), a los *«sistemas de información diseñado para trabajar con datos referenciados mediante coordenadas espaciales o geográficas. En otras palabras, un SIG es tanto un sistema de base de datos con capacidades específicas para datos georreferenciados, como un conjunto de operaciones para trabajar con esos datos. En cierto modo, un SIG es un mapa de orden superior.*

En torno a nuestro estudio podemos concebir al SIG como un elemento integrador de información geográfica, de tecnologías, de personas, de procesos y de conceptos de ciencias de geomática establecidos en una serie de subsistemas interrelacionados para las operaciones de entrada y salida de información, visualización y creación cartográfica y análisis de datos geográficos optimizando así su gestión mediante herramientas informáticas.

La problemática de la presente investigación es la inseguridad, Muggah (2017) señala que en las estadísticas mundiales de criminalidad los países de ALC presentan un grave problema de inseguridad, son considerados los países con mayor inseguridad y violencia del mundo, la tasa de homicidios de esta región es el triple que la tasa mundial, 8 de los 10 países con mayor violencia a nivel mundial, sin considerar zonas de guerra, se encuentran en ALC, más de 40 de un total de 50 ciudades más peligrosas del mundo se encuentran en esta región, los ciudadanos de ALC también registran una alta percepción de inseguridad.

En el Perú la tasa de homicidios aumentó de 5.4 en el año 2011 a 7.8 homicidios por cada 100,000 habitantes en el año 2017, situación más complicada se presenta en ciudades como Tambopata, Tumbes, Pisco, Provincia Constitucional del Callao y Barranca lugares donde los homicidios durante el año 2020 registraron tasas de 16.4, 15.8, 12.1, 12.0 y 11.9 homicidios por cada 100,000 habitante respectivamente, ciudades que por sus altos niveles de violencia requieren acciones focalizadas de lucha contra la inseguridad (Comité Estadístico Interinstitucional de la Criminalidad [CEIC], 2023).

Según el Instituto Nacional de Estadística e Informática (INEI) en los resultados obtenidos mediante la Encuesta Nacional de Hogares (ENAHO), en el Módulo de Gobernabilidad, Democracia y Transparencia de diciembre 2022 - mayo 2023, revela que el 55% de la población peruana mayores de 18 años de edad considera a la corrupción como el principal problema del país, seguido de la delincuencia (INEI, 2023). Otro estudio que señala que la seguridad ciudadana en el Perú se encuentra en serios problemas es la encuesta “Barómetro de las Américas” dicha publicación indica que, en la encuesta realizada en veinte países de la región, el Perú se ubica en el primer lugar, con mayor tasa de victimización, al registrar 35.8% de la

población afectada por algún hecho delictivo, seguido de México con 32.9%, Argentina 31.1%, Uruguay 29.3%, entre otros (Carrión, Zárate, Boidi, & Zechmeister, 2020). Dicho problema es de gran relevancia para el gobierno actual y frente a esta situación el Poder Ejecutivo ha considerado entre los lineamientos prioritarios de la Política General de Gobierno es "mejorar la seguridad ciudadana" (Decreto Supremos Nº 056, 2018).

En este sentido, considerando que la seguridad ciudadana es un problema multidimensional y generalizado en el ámbito nacional se requiere focalizar las intervenciones del Ministerio del Interior en distritos cuyo estado delictivo es crítico, en esa misma línea, el presente estudio plantea realizar un análisis de información estadística de seguridad ciudadana del Perú que permita priorizar los distritos que se encuentran en un estado delictivo crítico y de esta manera abordar con mejor criterio la inseguridad ciudadana en nuestro país.

La investigación plantea responder a la pregunta ¿Cuál es el mejor modelo de aprendizaje supervisado de clasificación para predecir el estado delictivo de los distritos del Perú? En este sentido, el estudio considerará toda la información oficial disponible a nivel distrital elaborada por instituciones vinculadas a la seguridad ciudadana como son: Instituto Nacional de Estadística e Informática, Instituto Nacional Penitenciario, Policía Nacional del Perú y Comité Estadístico Institucional de la Criminalidad; dichas instituciones generan información estadística de la cual se investigará las variables siguientes: homicidios, victimización, denuncias de delitos, personas privadas de la libertad según distrito de procedencia, policías que laboran en comisarías y efectivos que brindan servicio de serenazgo. Dichas variables serán analizadas en forma conjunta con el objetivo de identificar patrones delictivos y predecir el estado delictivo de los distritos del Perú utilizando algoritmos de minería de datos,

posteriormente, se construirá un ranking para identificar distritos inseguros y finalmente mediante el uso de un sistema de información geográfica se obtendrá la distribución espacial de la actividad delictiva de los distritos del Perú.

De esta manera se obtendrá información importante a nivel nacional y subnacional que será de utilidad para que el Ministerio del Interior ejecute con mejor criterio sus políticas, estrategias, programas y actividades para enfrentar la inseguridad, situación que impactará directamente en la población que gozará de una mejor convivencia pacífica y disminución de la violencia en las vías y espacios públicos del país.

En el marco de la conceptualización y operacionalización de variables la actividad delictiva es un conjunto de acciones relacionadas a hechos penados por la ley y asociados al crimen y a la violencia. Galindo (2007) sostiene que es la consecuencia de un conjunto de factores que incluyen tanto condiciones económicas y sociales como factores demográficos, psicológicos y de respeto e imposición de la ley.

La actividad delictiva utiliza 4 dimensiones: 1) Actores, 2) Delitos/Faltas, 3) Rehabilitación y 4) Población.

Sus indicadores son:

- Número de efectivos de la PNP
- Número de efectivos de serenazgo
- Muertes violentas asociadas a hechos delictivos dolosos
- Denuncias de delitos y faltas registradas por la PNP
- Último distrito de residencia del interno
- Número de habitantes del distrito

El estado delictivo, describe la situación de los territorios estudiados y nos ayudará a clasificarlos en función a la dinámica espacial del delito desarrollado según las variables seleccionadas de

inseguridad, crimen y violencia. Es decir, es la clasificación de los distritos de acuerdo con la inseguridad frente al crimen y a la violencia. El estado delictivo utiliza una dimensión, la inseguridad y cuyo indicador es distrito inseguro al crimen y a la violencia.

La hipótesis general que se debe comprobar en la presente investigación es: La predicción del estado delictivo de los distritos del Perú se realizará mediante la aplicación de algoritmos de aprendizaje supervisados de clasificación; así mismo se plantea las hipótesis especificas siguientes:

- La identificación de patrones delictivos se realizará empleando algoritmos de minería de datos.
- La identificación de grupos homogéneos de distritos se efectuará aplicando técnicas de agrupamiento.
- La visualización de la distribución espacial de la actividad delictiva en los distritos del Perú se efectuará empleando un sistema de información geográfica.

La investigación plantea como objetivo general predecir el estado delictivo de los distritos del Perú aplicando algoritmos de aprendizaje supervisados de clasificación y considera los objetivos específicos siguientes:

- Comparar los resultados de los modelos de predicción propuestos en base a indicadores de clasificación.
- Identificar patrones delictivos de los distritos del Perú mediante algoritmos de minería de datos.
- Visualizar la distribución espacial de la actividad delictiva en los distritos del Perú.
- Elaborar un ranking de los distritos del Perú según su actividad delictiva.

Metodología

Tipo de Diseño de Investigación

En relación con la metodología de este estudio, es de tipo descriptivo y posee un diseño no experimental transversal. La naturaleza **descriptiva** del estudio se manifestó al explorar y reconocer las características y distribución de las variables investigadas, identificar patrones delictivos, agrupar distritos con características similares y categorizar los distritos según las particularidades de sus actividades delictivas. **El diseño no experimental** se justifica porque se examinaron variables relacionadas con la seguridad ciudadana en su entorno real, sin hacer variaciones intencionadas para observar sus efectos en otras variables. Se denomina **transversal** ya que se recolectó información sobre actividades delictivas en distintos distritos del Perú durante un año específico.

1.1. Población y Muestra

La población está constituida por 1754 distritos (Anexo 01). Es importante precisar que el número total de distritos a nivel nacional en el año 2020 fue de 1874, se excluyó 120 distritos debido a que no cuentan con información en materia de seguridad ciudadana.

1.2. Técnicas e Instrumentos de Investigación

Para la construcción del estado del arte, se llevó a cabo un análisis documental, utilizando métodos inductivos y deductivos en textos, artículos científicos y trabajos de investigación relacionados con el tema. En cuanto a la recopilación y a la integración de datos, se utilizó información estadística relacionada a seguridad ciudadana del año 2020, información de libre disposición en las páginas web de las instituciones encargadas de generarla.

Las fuentes de información oficiales consultadas son las

siguientes: Sistema de Información de Unidades Policiales (SIUP) del Ministerio del Interior, Sistema de Denuncias Policiales (SIDPOL) de la Policía Nacional del Perú, Instituto Nacional Penitenciario (INPE), Comité Estadístico Institucional de la Criminalidad (CEIC) e Instituto Nacional de Estadística e informática (INEI).

Respecto al procesamiento de la información, primero, seleccionamos las variables de estudio y se realizó la limpieza de datos. Luego, se efectuó el análisis exploratorio de dichas variables, es decir, se aplicó técnicas estadísticas para describir la distribución, variabilidad, patrones, tendencias, valores atípicos y relación entre variables e identificar sus características. Posteriormente, se transformó los atributos utilizando el análisis de escalamiento multidimensional (AEM), generando coordenadas sintéticas a partir de los datos ya escalados. Luego se redujo la dimensionalidad de la base datos mediante el análisis de componentes principales (ACP).

En una segunda fase, agrupamos los distritos basándonos en sus similitudes en cuanto a características delictivas. Utilizamos el análisis clúster sobre las proyecciones resultantes del AEM y ACP. Gracias a esto, generamos dos nuevas variables que enriquecieron la base de datos. Luego, aplicamos algoritmos como árboles de decisión, naive bayes, k-vecino más cercano y máquinas de vectores soporte para identificar el modelo más preciso en la predicción del estado delictivo de los distritos.

Después, elaboramos un ranking con la finalidad de determinar un orden correlativo de los distritos en base a su actividad delictiva. Finalmente, empleamos un sistema de información geográfica para representar la distribución espacial del crimen en los distritos peruanos. El análisis estadístico se llevó a cabo utilizando los lenguajes de programación R y Python.

Resultados

Para evidenciar el objetivo específico 1, identificar patrones delictivos de los distritos del Perú mediante algoritmos de minería de datos, se procesó la información en lenguajes de programación R y Python, con la finalidad de descubrir el conocimiento de la base de datos. La base de datos está compuesta por 1,754 observaciones (distritos) y 7 variables, respecto a las variables 6 son cuantitativas (todas de escala de razón) y 01 cualitativa (de escala nominal), es importante precisar que la base de datos no registra datos faltantes.

Respecto a la variable "Inseguro", es la variable que se va a predecir, tiene como definición aquellos distritos que son inseguros al crimen y a la violencia; el total de registros es de 1,754 distritos, donde 120 son considerados distritos inseguros, los cuales representan el 7% y 1,634 considerados distritos no inseguros con un porcentaje del 93%.

Análisis Exploratorio por Variable

A. Variable "muertes violentas asociadas a hechos delictivos dolosos"

En la figura 1 se observa la distribución de la variable, con una mediana en cero muertes. El pico de la distribución nos representa los valores más comunes, es decir, la concentración de los distritos se presenta en el primer decil (de 0 a 3 muertes) ya que hay mayor frecuencia de distritos en los primeros valores de la variable, situación que origina que la distribución de la variable muertes sea sesgada a la derecha. Como se aprecia en la Figura 1, el promedio de la variable muertes es de 1.08, además, se aprecia en la gráfica que los datos de la variable son dispersos, presentan una desviación´ estándar de alrededor 4 muertes, la cantidad mínima de muertes es de 0 y la máxima de 80 muertes.

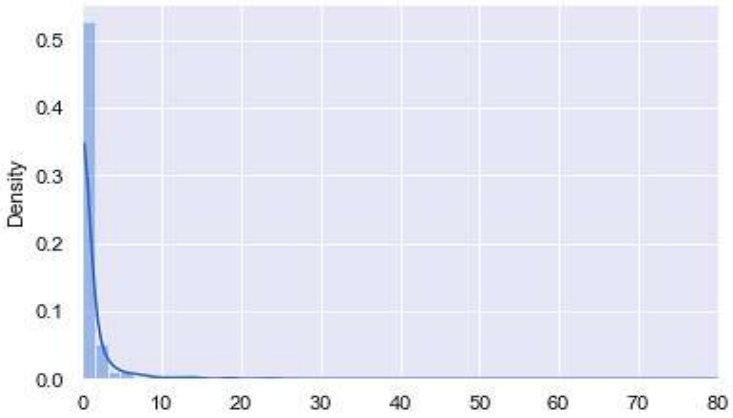


Figura 1. *Distribución de la variable muertes violentas asociadas a hechos delictivos dolosos 2020.*
Nota: Elaboración propia

Según la Figura 2, la variable "Muertes violentas asociadas ahechos delictivos dolosos", tiene una alta concentración de distritos en los primeros valores de la variable originando que dichos valores se acumulen en la parte inferior de la gráfica, en este sentido la gráfica muestra una asimetría hacia la derecha. También se aprecia una alta proporción de valores atípicos, gráficamente representados en los puntos negros del diagrama de cajas, la caja está conformada por el bigote inferior y superior; en este caso 0 y 2.5.

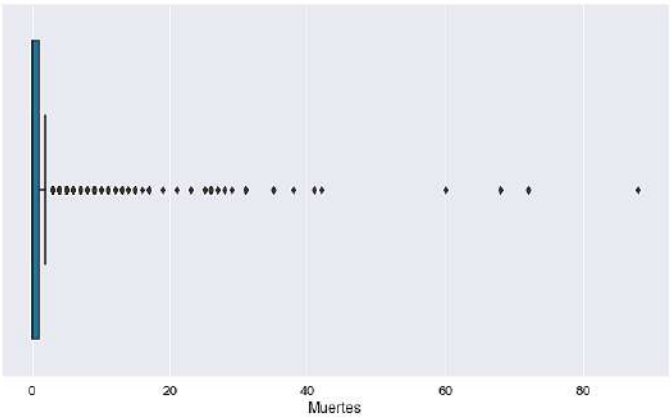


Figura 2. *Diagrama de cajas de la variable muertes violentas asociadas a hechos delictivos dolosos 2020.*
Nota: Elaboración propia

Luego se identificó aquellos distritos que son atípicos respecto a la variable muertes, es decir, se consideró valores atípicos

aquellos distritos que están fuera de los limites inferior y superior del diagrama de cajas (0 a 2.5 muertes), en este sentido, se realizó un análisis priorizando los distritos atípicos de la variable muertes con la finalidad de identificar patrones en los datos analizados.

El departamento de Lima registra el mayor número de muertes violentas asociadas a hechos delictivos dolosos (613), de los cuales 571 muertes (93%) ocurrieron en distritos atípicos. Le sigue El Callao que registró 132 muertes de los cuales 129 muertes (98%) ocurrieron en distritos atípicos. Seguido de Junín con 109 muertes violentas de los cuales 81 muertes (74%) ocurrieron en distritos atípicos. Los departamentos que registran el menor número de muertes violentas son Pasco y Moquegua con 13 y 12 muertes violentas respectivamente, como se puede apreciar en la gráfica siguiente:

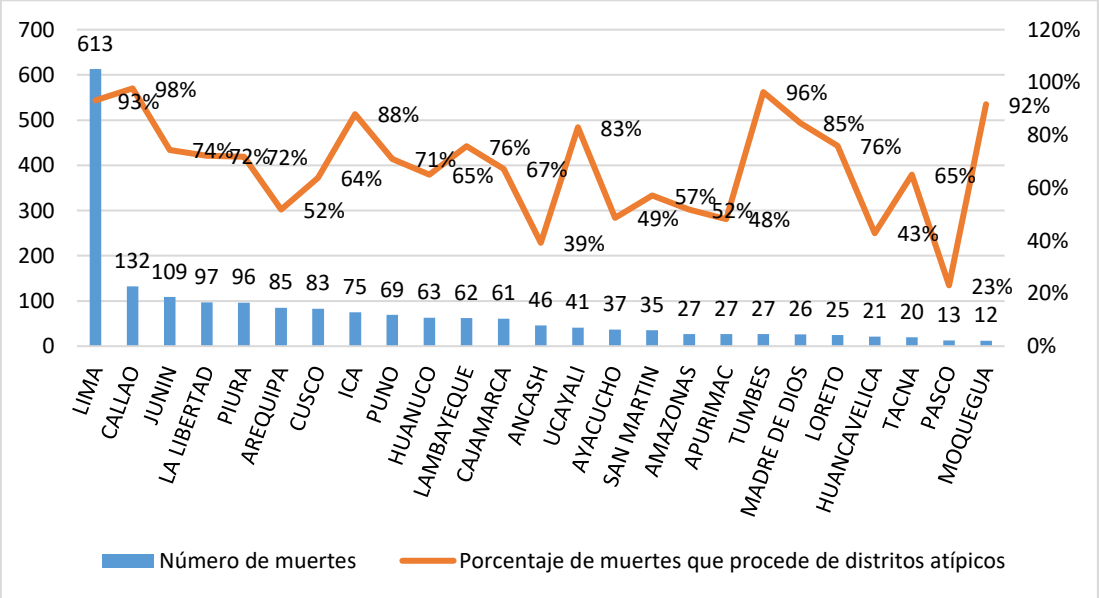


Figura 3. *Número de muertes violentas asociadas a hechos delictivos dolosos y porcentaje de muertes ocurridas en distritos atípicos según departamento.*
Nota: Elaboración propia

Respecto a las regiones con mayor porcentaje de distritos atípicos podemos mencionar al Callao que registró 4 (57%) distritos atípicos de los 7 que la conforman. Seguido de Ucayali que registró 6 (35%) distritos atípicos de los 17 que conforman la región. Luego, Tumbes que registró 4 (31%) distritos atípicos de los 13 que la conforman. Seguido de Lima que registró 44 (31%) distritos atípicos de los 144. Las regiones que registran menor porcentaje de distritos atípicos son Huancavelica y Ancash como se puede apreciar en la Figura Nº4.

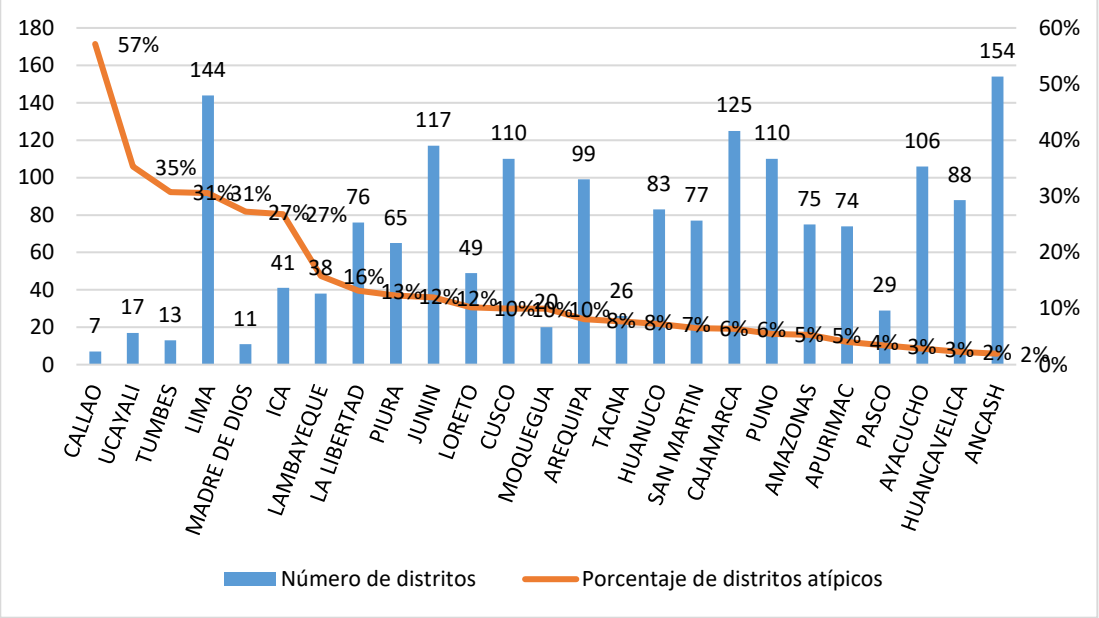


Figura 4. *Número de distritos por regiones respecto al porcentaje de distritos atípicos (respecto al número de muertes) por departamentos.*
Nota: Elaboración propia

Respecto a las 25 provincias que registraron el mayor número de muertes a nivel nacional, se encuentran en las primeras posiciones Lima y Callao. Lima con 526 muertes, de los cuales 512 (97%) proceden de distritos atípicos, seguido del Callao con 132 muertes de los cuales 129 (98%) proceden de distritos atípicos, seguido de Trujillo con 53 muertes de los cuales 52 (98%) proceden

de distritos atípicos, seguido de Piura con 43 muertes de los cuales 41 (95%) proceden de distritos atípicos. Es importante precisar que Lima, Callao y las principales provincias del norte del país lideran el ranking de provincias de mayor violencia en el país, como podemos observar en el gráfico siguiente:

En relación con las provincias que presentan el mayor número de muertes violentas, podemos mencionar Lima (526), seguido de Callao (132), Trujillo (53), Arequipa (51), Chiclayo (44); dichas provincias están conformadas por más de 65% de distritos atípicos como se puede apreciar en la gráfica siguiente:

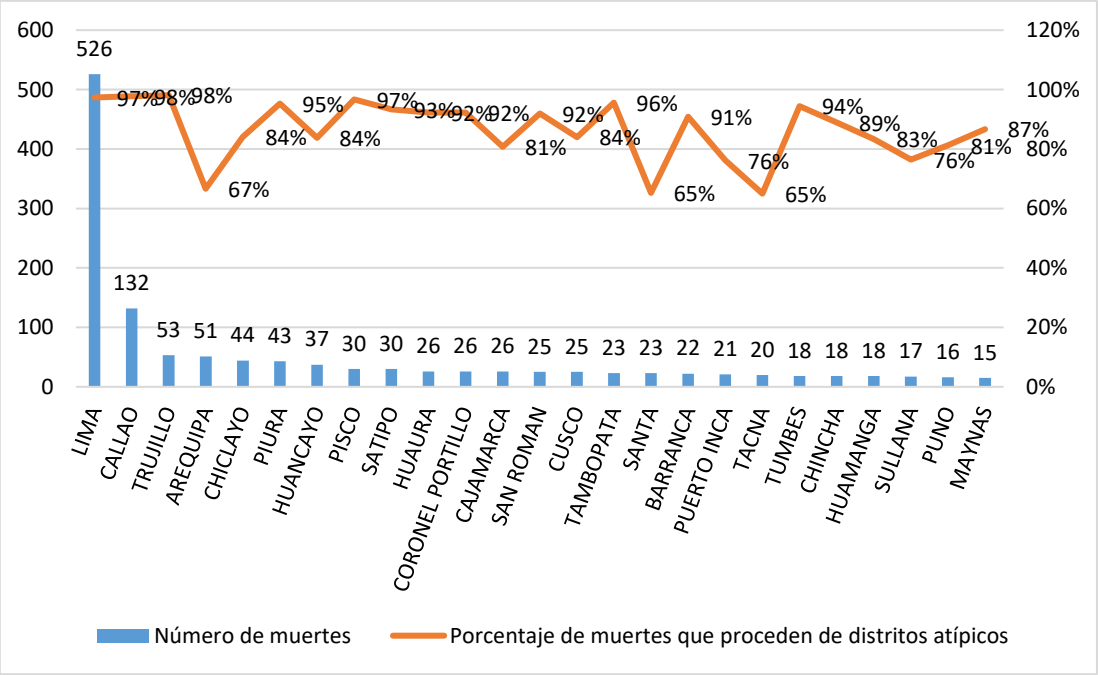


Figura 5. *Número de muertes y porcentaje de muertes ocurridas en distritos atípicos según 25 provincias con mayor número de muertes.*
Nota: Elaboración propia

En referencia a las provincias que presentan los más altos porcentajes de distritos atípicos podemos mencionar Barranca 80%, Tambopata 75%, Lima 74%, Trujillo 64%, Callao 57%. De las provincias señaladas Lima registra el mayor número de distritos atípicos es decir 32 de los 43 distritos según la figura n°6.

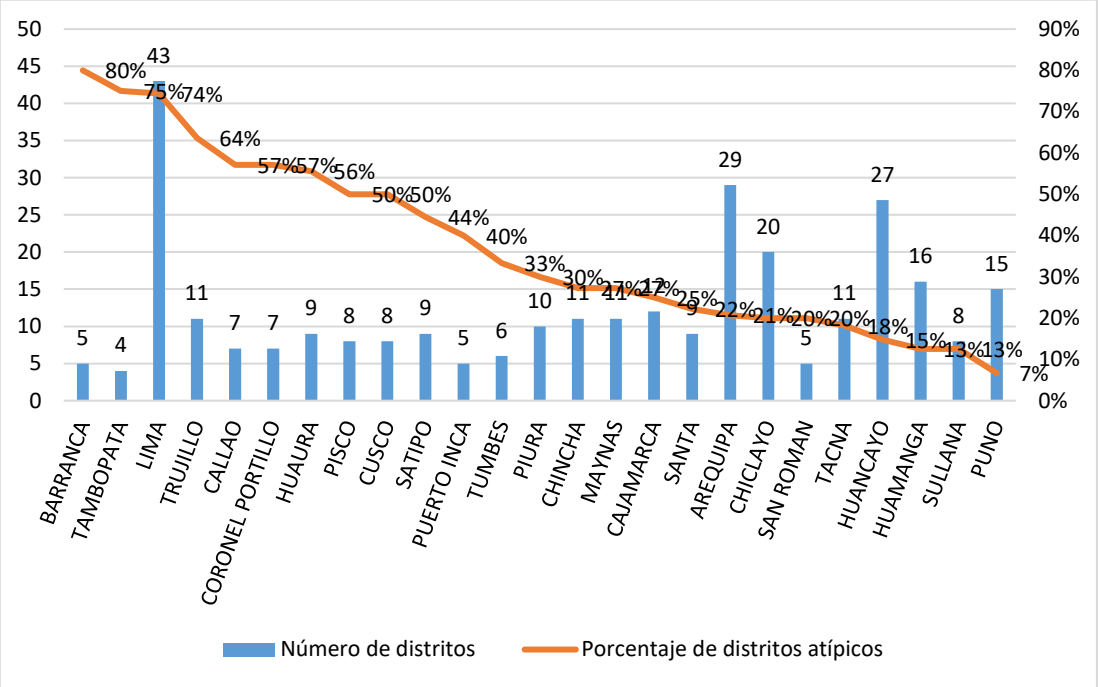


Figura 6. *Número de distritos respecto al porcentaje de distritos atípicos (respecto al número de muertes) según las 25 provincias con mayor número de muertes.*
Nota: Elaboración propia

Respecto a los 25 distritos con mayor número de muertes violentas, la figura n°7 señala que el Callao (80) lidera el ranking, seguido de Comas (59), Lima (52), San Juan de Lurigancho (41), San Martín de Porres (37), entre otros

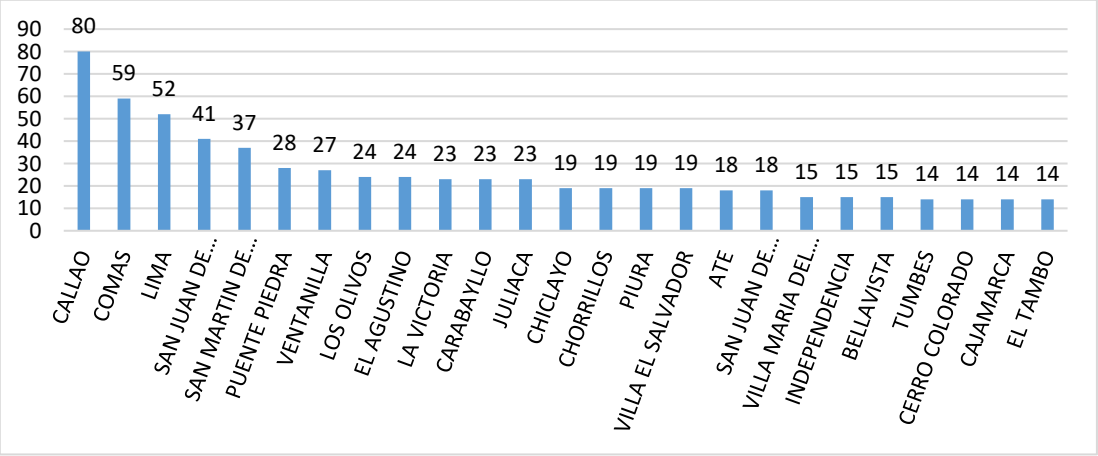


Figura 7. *25 distritos con el mayor número de muertes.*
Nota: Elaboración propia

B. Variable "denuncias de delitos y faltas registradas por la PNP"

En la figura 8 se observa la distribución de la variable "Denuncias", con una mediana en ocho delitos. La concentración de las observaciones está en el primer decil ya que hay más frecuencia de distritos, así mismo, la distribución de dicha variable es sesgada a la derecha. Respecto a las estadísticas descriptivas se observa que el promedio de denuncias es de 538.31, con una desviación estándar de 2,111.73 denuncias. Por otro lado, la cantidad mínima de delitos es de 0 y la máxima de 30,272 delitos.

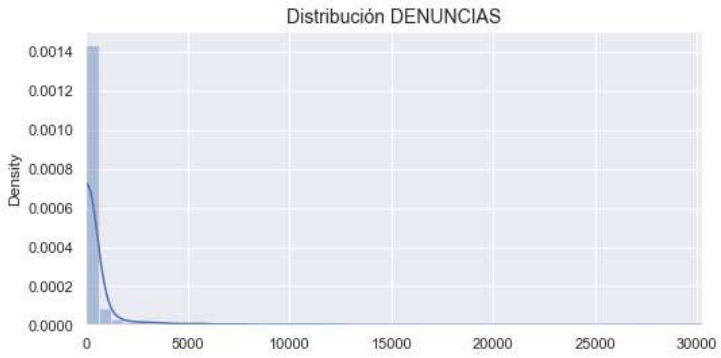


Figura 8. *Distribución de la variable denuncias de delitos y faltas registradas por la PNP.*
Nota: Elaboración propia

Según la Figura 9, la variable "denuncias" se aprecia gran proporción de datos atípicos, en el gráfico de cajas está representado por los puntos negros, la caja está conformada por el bigote inferior y superior; en este caso 0 y 425.

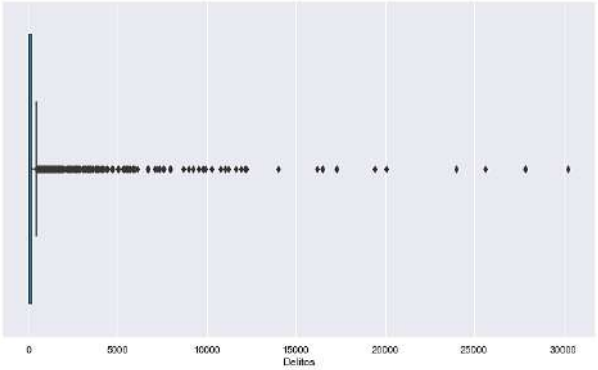


Figura 9. *Diagrama de cajas de la variable denuncias de delitos y faltas registradas por la PNP.*

Nota: Elaboración propia

Luego se identificó los distritos atípicos respecto a la variable "denuncias", es decir, se consideró valores atípicos aquellos distritos que están fuera de los limites inferior y superior del diagrama de cajas (0 a 425 denuncias).

En la figura 10 tratando de hallar los porcentajes de denuncias de delitos procedentes de distritos atípicos según departamentos, notamos que Lima con 311 774 denuncias, Arequipa con 116 813 denuncias y Lambayeque con 46 6620 denuncias encabezan el gráfico según el número de denuncias y en relación con el porcentaje de denuncias procedentes de distritos atípicos presentan respectivamente el 99%, 94% y 91% mientras que visualizamos que Pasco (69%) con 6289 , Moquegua con 5726 (92%) y Huancavelica con 5276 (62%) se ubican en la menor concentración según la cantidad de denuncias; de esta manera se refuerzan las consecuencias de la centralización de instituciones en las regiones, y se puede observar una ligera pero mayor efectividad y logística en lo referente a canales de recepción de denuncias en las regiones más pobladas del país en desmedro de las de menor densidad demográfica.

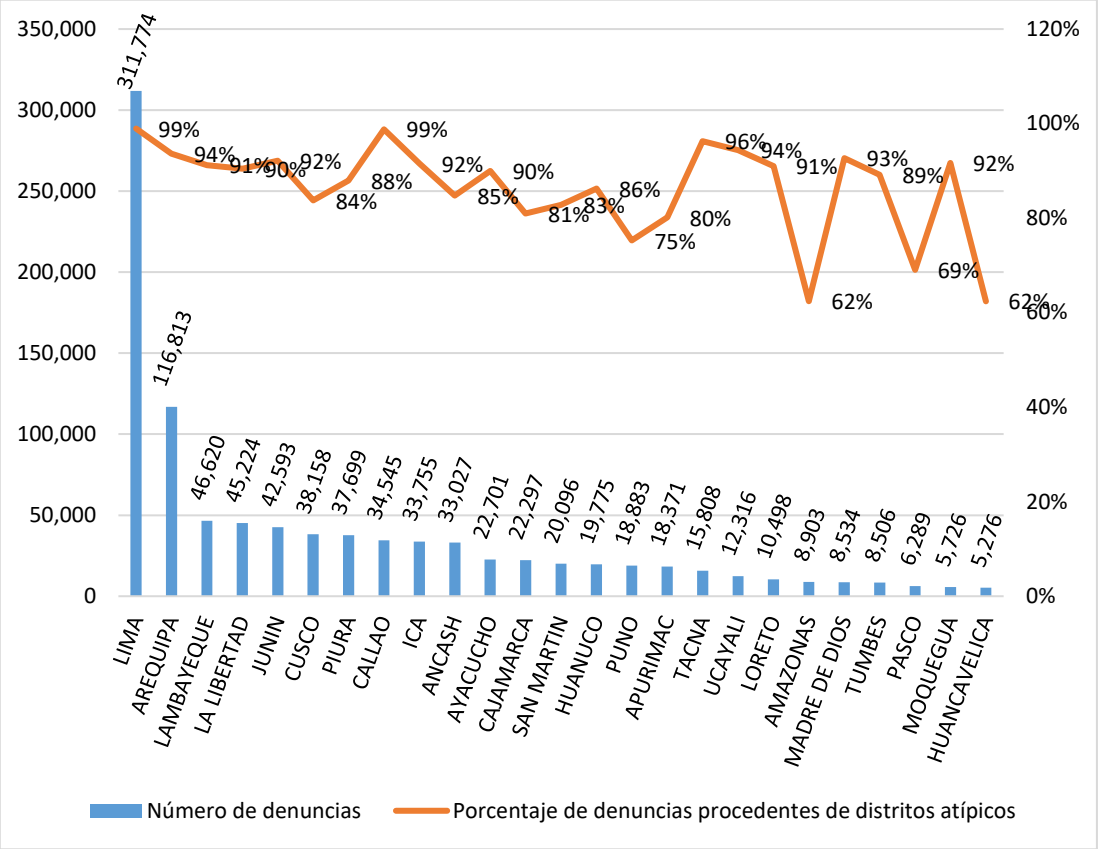


Figura 10. *Denuncias de delitos y faltas registradas por la PNP y porcentaje de denuncias ocurridas en distritos atípicos según departamento.*
Nota: Elaboración propia

Complementando el gráfico anterior, exploramos el porcentaje de distritos atípicos en relación al número de distritos por departamentos (considerando el número de denuncias y de delitos) la figura 11 nos señalan al Callao con 71% (7 distritos) seguido de Ica con 49% (41 distritos) y Lima con 44% (144 distritos) como los departamentos con mayor porcentaje mientras que cierran el gráfico de barras los departamentos de Amazonas con 4% (75 distritos), Puno con 3% (110 distritos) y Huancavelica con 2% (88 distritos) son los de menor porcentaje de distritos atípicos.

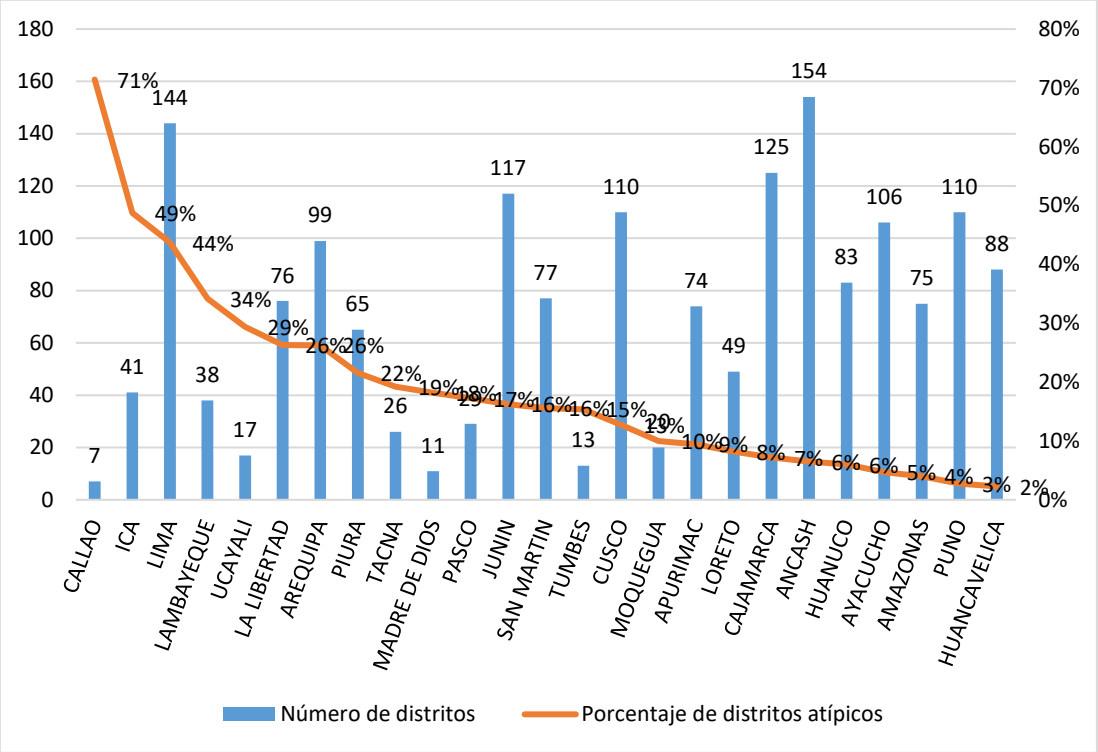


Figura 11. *Número de distritos por regiones respecto al porcentaje de distritos atípicos (considerando al número de denuncias de delitos y faltas) por departamentos.*
Nota: Elaboración propia

Al revisar en la figura 12 el número de denuncias de delitos y faltas registradas por la PNP y porcentaje de muerte ocurridas en distritos atípicos (25 provincias con mayor número de denuncias) encontramos la mayor concentración encabezada por Lima con el 100% seguido por Arequipa (98%), Callao (99%) Trujillo (99%) mientras que las que presentan menor porcentaje de distritos atípicos son Barranca (94%), Ascope (59%) y Huarochirí (84%).

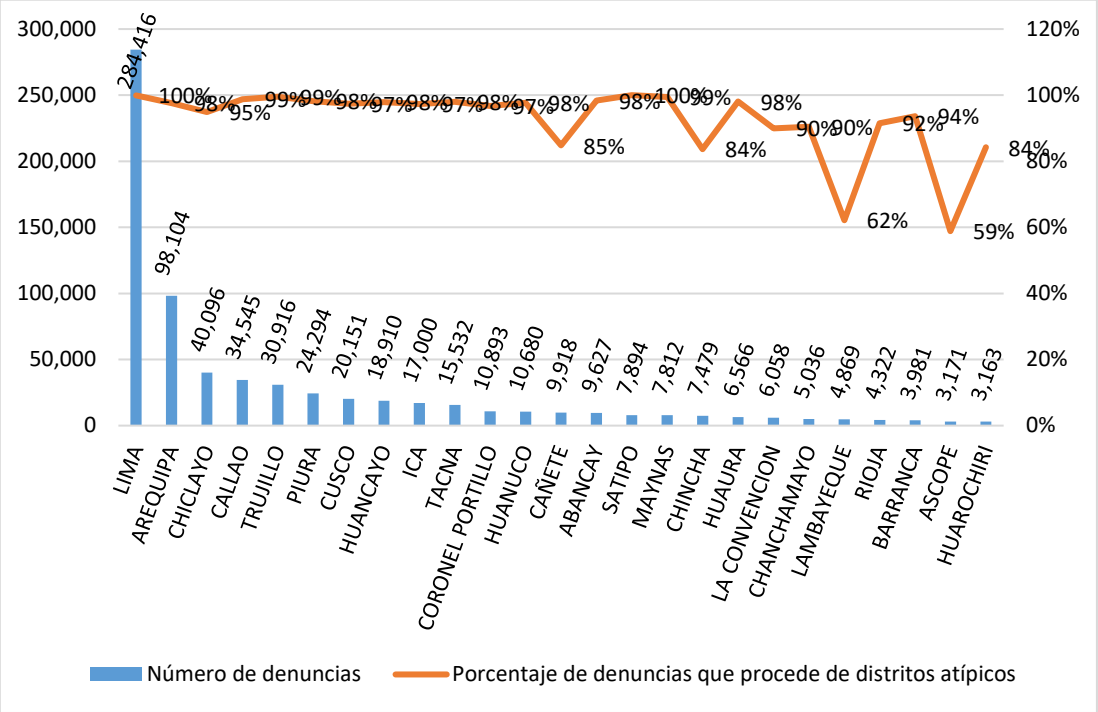


Figura 12. *Denuncias de delitos y faltas registradas por la PNP y porcentaje de muertes ocurridas en distritos atípicos según 25 provincias con mayor número de denuncias.*
Nota: Elaboración propia

Examinando a detalle la figura 13 y focalizando el porcentaje del número de distritos atípicos (respecto al número de denuncias) encontramos a Lima con 98% (de 48 distritos), Trujillo con 82% (de 11 distritos), Barranca con 80% (de 5 distritos) y Callao con 71% (7 distritos) encabezan la lista, mientras que Huancayo con 26% (de 27 distritos), Huánuco con 25% (de 12 distritos) y Huarochiri con 12% (de 26 distritos) se ubican al final de la misma, según la data de las 25 provincias con mayor número de denuncias de delitos y faltas registradas por la Policía.

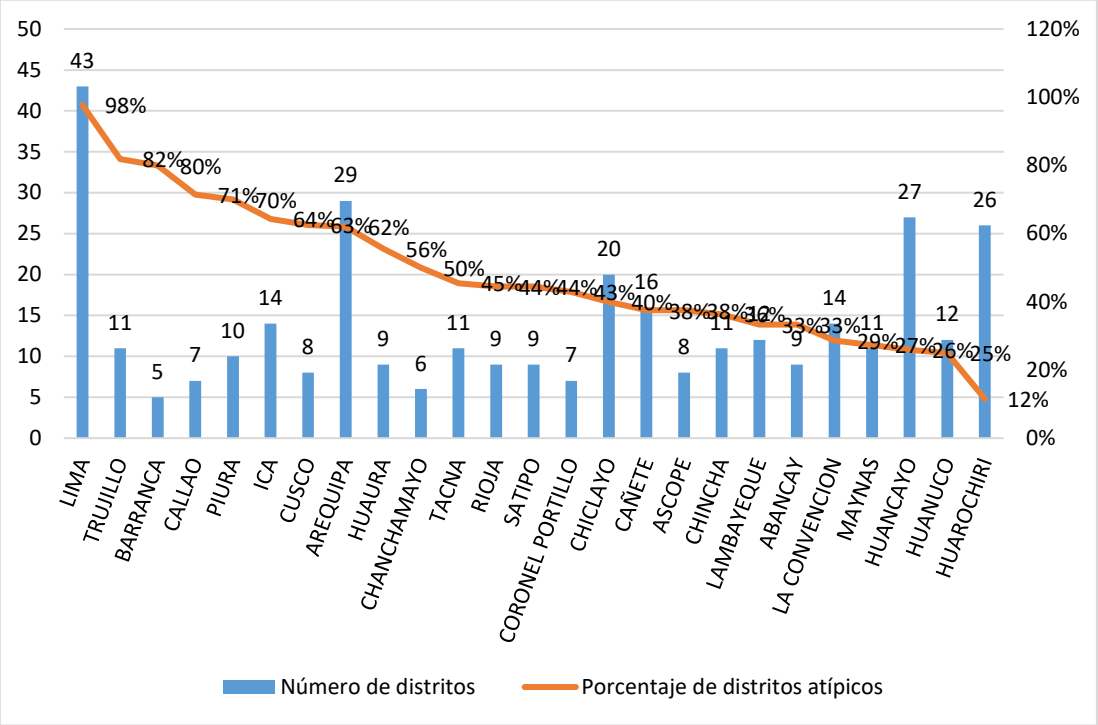


Figura 13. *Número de distritos respecto al porcentaje de distritos atípicos (respecto al número de denuncias) según las 25 provincias con mayor número de denuncias de delitos y faltas registradas por la PNP.*
Nota: Elaboración propia

En la figura 14 al final de la variable "Distritos con el mayor número de denuncias de delitos registradas por la PNP", comparando los 25 distritos con el mayor número de denuncias de delitos y faltas registradas por la PNP encontramos los distritos de mayor número de denuncias a Lima con 30 272, San Juan de Lurigancho con 27 888, Callao con 25 653 y Villa El Salvador con 23 972 mientras que Paucarpata con 9 768, Tacna con 9529, Ica con 9 224 y La Victoria con 8947 son los distritos de menor concentración de denuncias.

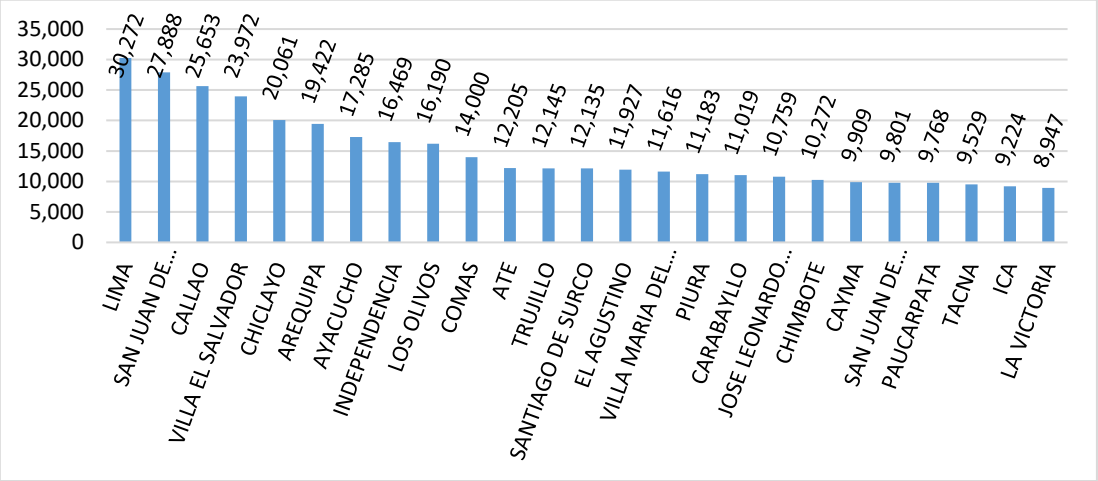


Figura 14. *25 distritos con el mayor número de denuncias de delitos y faltas registradas por la PNP.*
Nota: Elaboración propia

C. Variable "población"

Se observa en la figura 09 la distribución de la variable "Población", con una mediana en 4 507 habitantes. La concentración de las observaciones está en el primer decil ya que hay más frecuencia de distritos con población pequeña, así mismo, la distribución de dicha variable es sesgada a la derecha. Según las estadísticas descriptivas el promedio de la población a nivel distrital es de 18 519 habitantes, con una desviación´ estándar de 58 362 habitantes. Por otro lado, la cantidad mínima de habitantes es de 153 y la máxima de 1 177 626

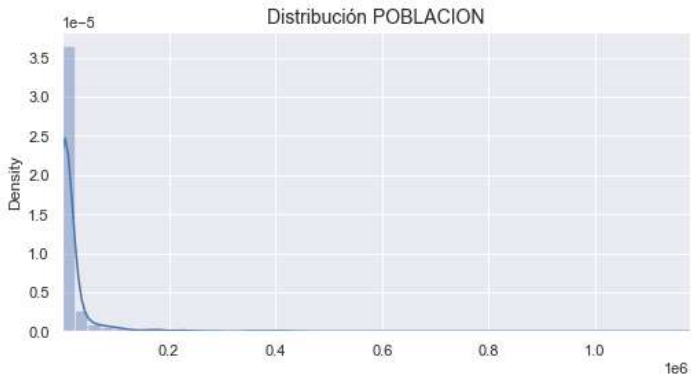


Ilustración 15. *Distribución de la variable población.*
Nota: Elaboración propia

Según el diagrama de cajas, la variable “Población” presenta distritos atípicos, representados por los puntos negros, la caja está conformada por el bigote inferior y superior; en este caso 0 y 26 293.

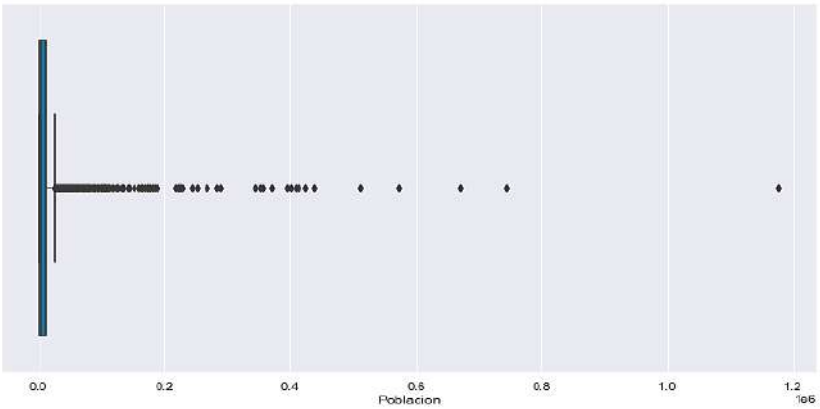


Figura 16. *Diagrama de cajas de la variable población.*
Nota: Elaboración propia

En la figura 17 revisando a detalle la variable población de las 25 provincias en el diagrama de cajas para el presente estudio, siendo los departamentos más poblados Lima, Piura, La Libertad.

Lima con 10 609 278 de habitantes nos muestra que el 96% de dicha cifra procede de distritos atípicos, mientras que Piura (2 047 954) y La Libertad (2 002 669) presentan 96% y 78% respectivamente como porcentaje de la población que procede de distritos atípicos con respecto del total. Del lado opuesto en los departamentos menos poblados: Tumbes, Moquegua y Madre de Dios que concentran respectivamente el 45% (de 251 521), 75% (de 192 740) y 58% (de 173 811) de la población que procede de distritos atípicos.

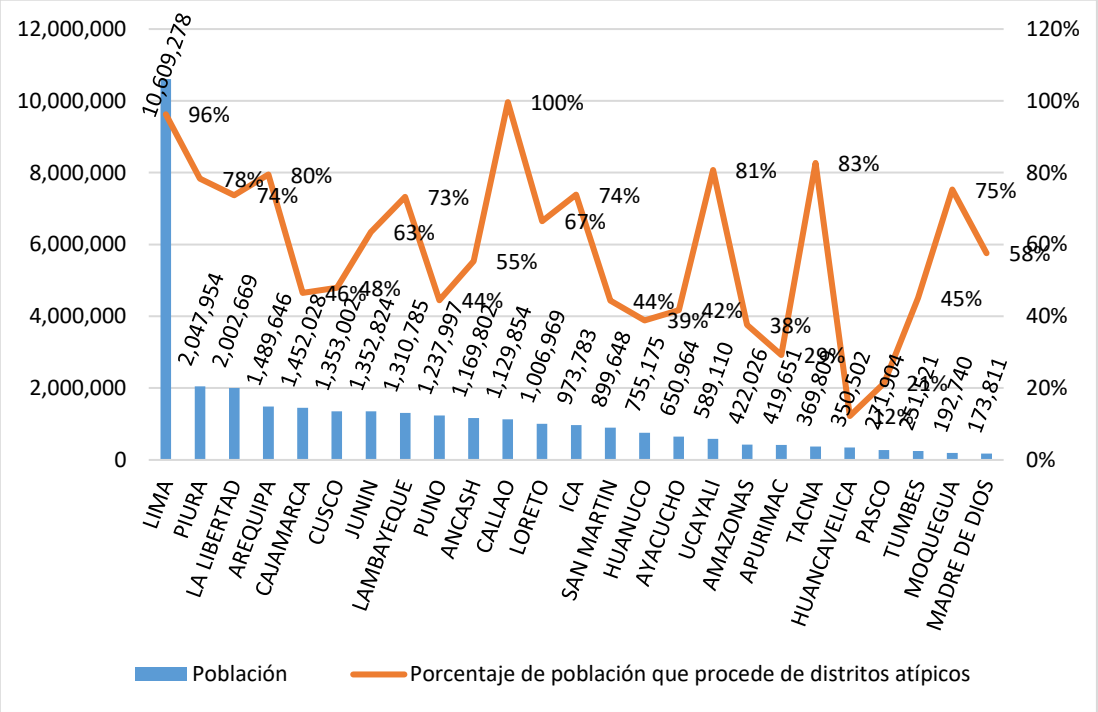


Figura 17. *Población y porcentaje de la población procedente de distritos atípicos según departamentos*
Nota: Elaboración propia

Dentro de la figura 18 evaluamos los 25 departamentos según el número de distritos, siendo las que concentran mayor cantidad de distritos Callao, Lima e Ica.

Del porcentaje de distritos atípicos respecto al tamaño de la población vemos que el Callao concentra el 86% (de 7 distritos) Lima, el 35% (de 144 distritos) e Ica, el 32% (de 41 distritos). Mientras que en las regiones con menor número de distritos como son Puno, Apurímac y Huancavelica hallamos que concentran el 5% (de 110 distritos), el 5% (de 24 distritos) y el 3% (de 88 distritos) respectivamente, con relación al porcentaje presente de distritos atípicos.

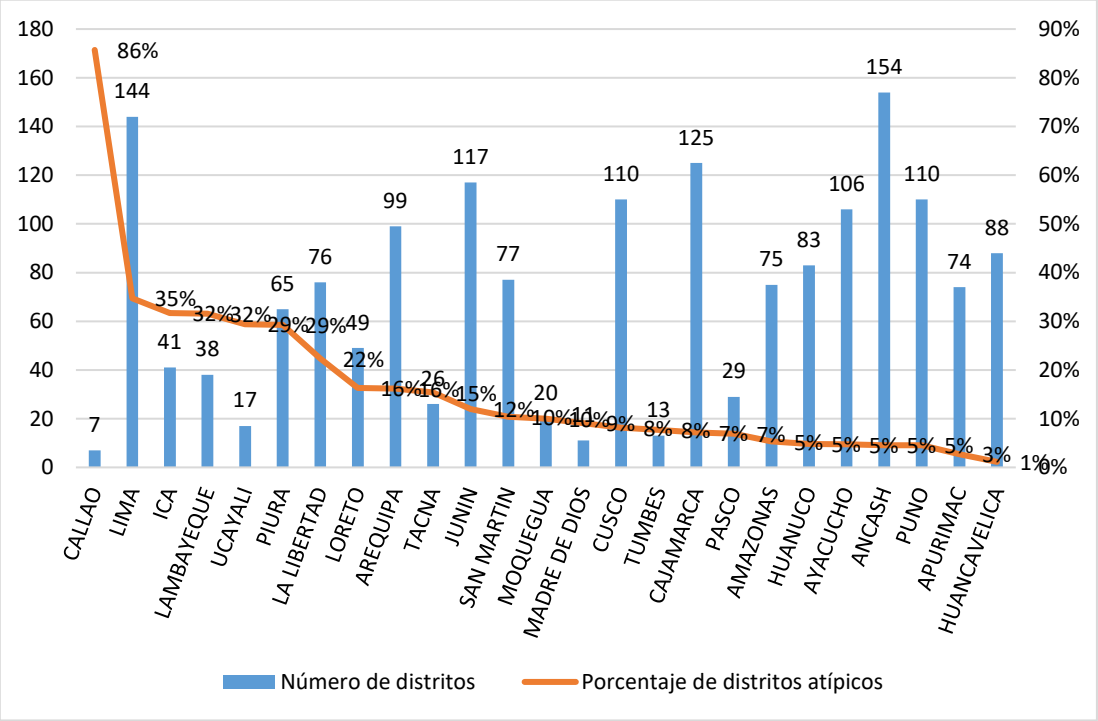


Figura 18. *Número de distritos por regiones respecto al porcentaje de distritos atípicos (respecto al tamaño de la población) por departamentos.*
Nota: Elaboración propia

Al seguir revisando la población consolidada de las provincias, en la figura 19, de dichas cifras podemos extraer el porcentaje de población procedente de distritos atípicos según las 25 provincias con mayor número de denuncias, siendo encabezadas por Lima que concentra un porcentaje del 99% de distritos atípicos de una población de 9 674 755 seguidas por Arequipa con 92% de distritos atípicos (de 1 175 765), Callao con el 100% de distritos atípicos (de 1 129 854) y Trujillo con 97% (de 1 118 724).

Adicional a ello vemos en las provincias con menor población que comprende a Chanchamayo, Ascope y Bagua concentran respectivamente porcentajes de 97% (de 167 385 habitantes), el 48% (de 123 480 habit.) y el 73% (de 84 672 habit.) en relación a la población procedente de distritos atípicos

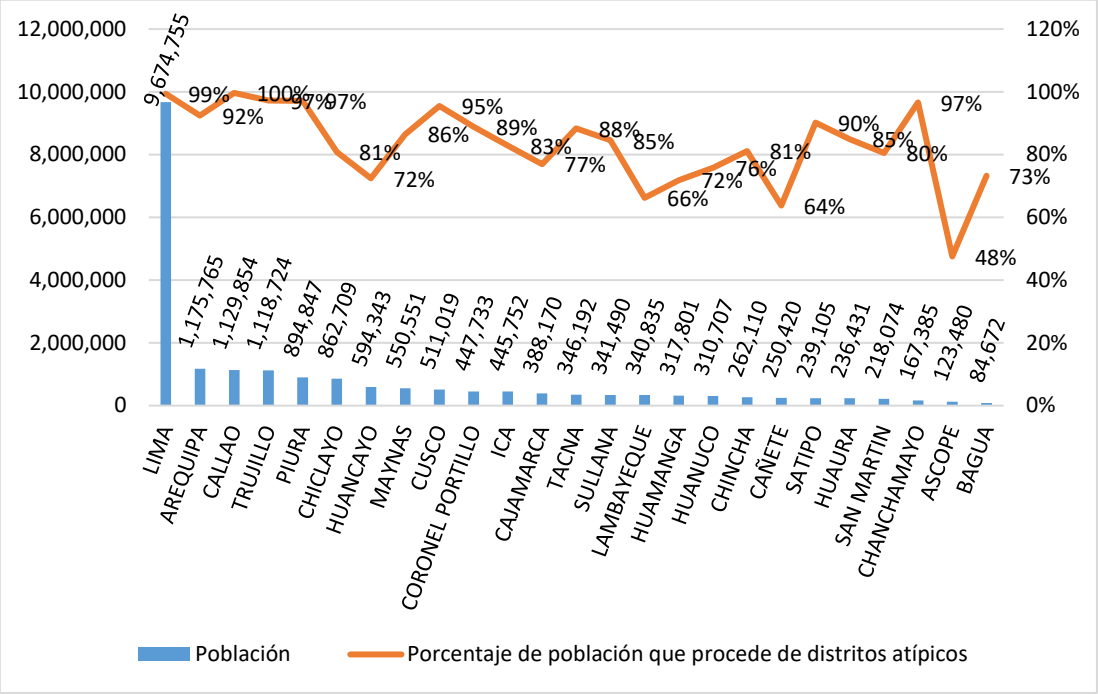


Figura 19. *Población y porcentaje de población procedente de distritos atípicos según 25 provincias con mayor número de denuncias.*
Nota: Elaboración propia

Examinando la figura 20, de las 25 provincias con mayor número de denuncias registradas por la PNP, vemos según el número de distritos, como registrado por mayor número a Lima, Callao, Piura y Trujillo a su vez en ellos cotejamos en relación de distritos atípicos porcentajes de 88% (con 43 distritos), el 86% (con 7 distritos), el 80% (con 10 distritos) y 73% (con 11 distritos)

Mientras como menor número de distritos están Ascope, San Martín, Cajamarca y Huancayo, en estas provincias hallamos respectivamente los siguientes porcentajes en relación con distritos atípicos: 25% (con 8 distritos), 21% (con 14 distritos), 17% (con 12 distritos) y 15% (con 27 distritos).

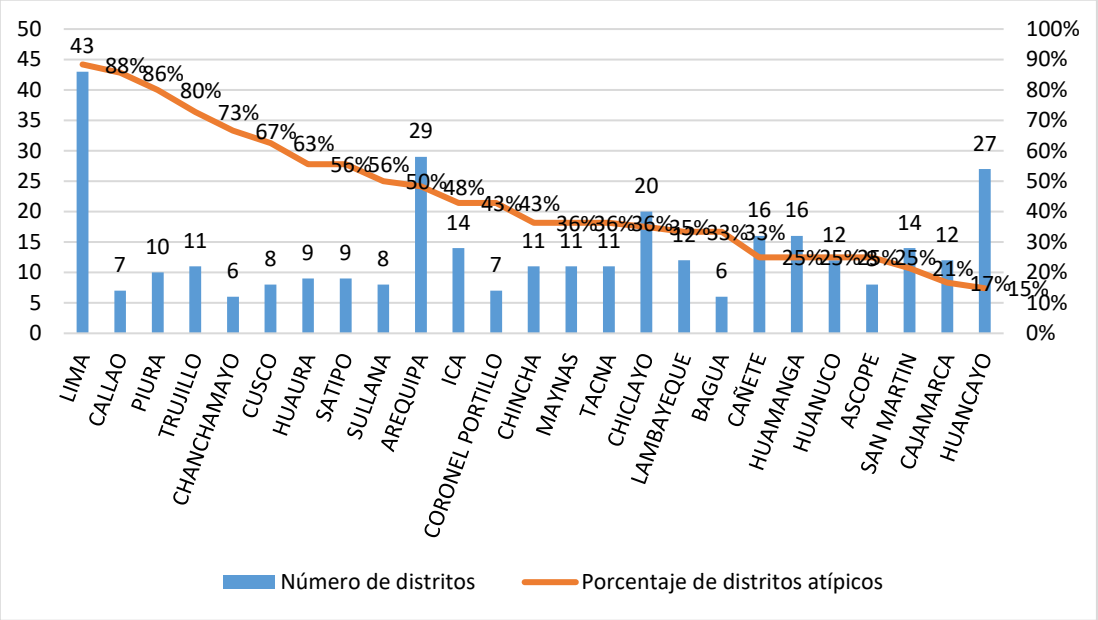


Figura 20. *Número de distritos respecto al porcentaje de distritos atípicos (respecto al tamaño de la población) según las 25 provincias con mayor número de denuncias de delitos y faltas registradas por la PNP.*
Nota: Elaboración propia

Adentrándonos a un ámbito más interno, en la figura 21 podemos observar a los 25 distritos con el mayor número de población, encontrando en los puestos más altos a San Juan de Lurigancho con 1 177 629 habitantes, San Martin de Porres con 744 055 habit. y Ate con 670 818 habit. mientras que en los puestos más bajos de esta lista se ubica La Esperanza con 224 427 habitantes, Independencia con 222 850 habitantes y El Agustino con 221 974 habitantes.

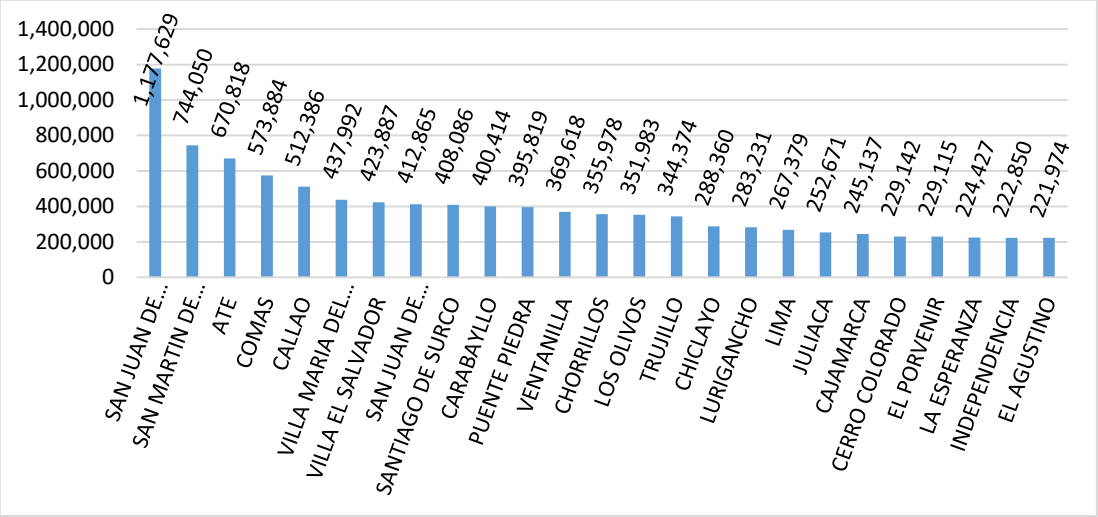


Figura 21. *25 distritos con el mayor tamaño de población.*
Nota: Elaboración propia

D. Variable "número de efectivos de la Policía Nacional del Perú"

En la figura 22 se observa la distribución de la variable, con una mediana en 11 policías. El pico de la distribución nos representa los valores más comunes, es decir, la concentración de los distritos se presenta en el primer decil (de 0 a 60 policías) ya que hay mayor frecuencia de distritos en los primeros valores de la variable, situación que origina que la distribución de la variable policías sea sesgada a la derecha, el promedio de la variable policías es de 28.67, además, se aprecia en la gráfica que los datos de la variable son dispersos, presentan una desviación estándar de alrededor 70.80 policías, la cantidad mínima de muertes es de 0 y la máxima de 1096 muertes.

Nota: Elaboración propia

Nota: Elaboración propia

Nota: Elaboración propia

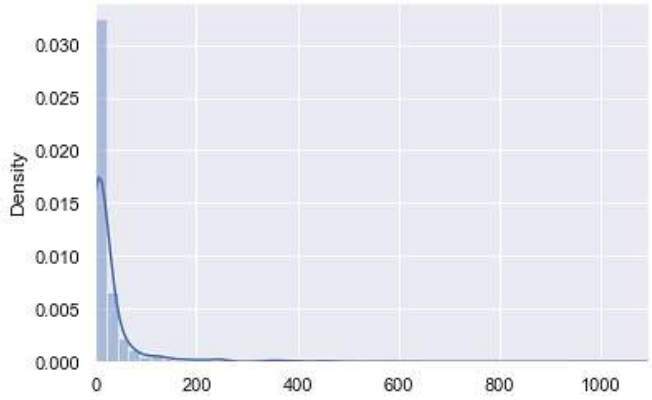


Figura 22. *Distribución de la variable número de efectivos policiales en el distrito.*

Nota: Elaboración propia

Según la Figura 23, la variable "policías", tiene una alta concentración de distritos en los primeros valores de la variable originando que dichos valores se acumulen en la parte inferior de la gráfica, en este sentido la gráfica muestra una asimetría hacia la derecha. También se aprecia una alta proporción de valores atípicos, gráficamente representados en los puntos negros del diagrama de cajas, la caja está conformada por el bigote inferior y superior; en este caso 0 y 60.

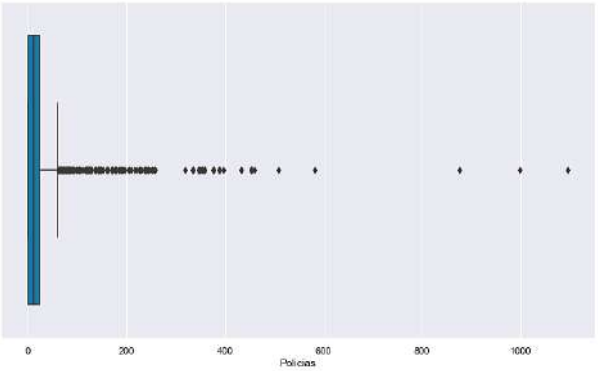


Figura 23. *Diagrama de cajas de la variable número de efectivos policiales en el distrito.*

Nota: Elaboración propia

Dentro de la figura 24 se presenta el número de policías de los departamentos encabezando la lista: Lima con 12 915, Arequipa con 3256 y Piura 2976, los cuales presentan cifras de 91%, 65% y

63% respectivamente como porcentajes de efectivos policiales procedentes de distritos atípicos.

Mientras que los departamentos más bajos (según número de policías) que cierran la lista: Ucayali con 613, Moquegua con 489 y Madre de Dios con 262 presentan sus respectivos porcentajes de policías procedentes de distritos atípicos siendo 64%, 53% y 45% respectivamente.

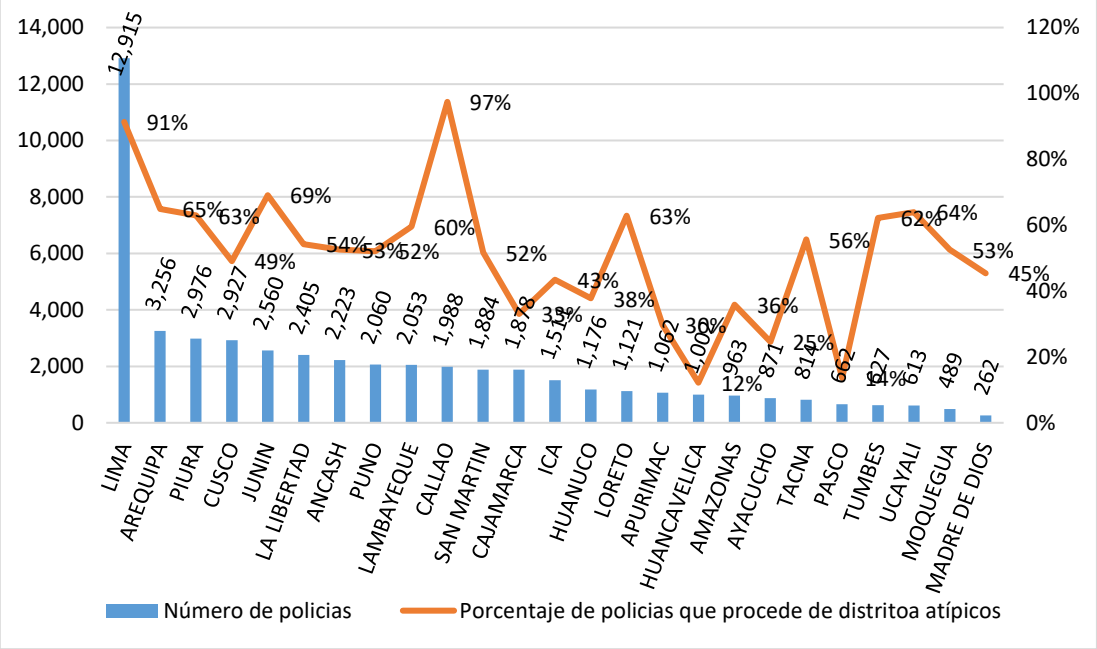


Figura 24. *Número de efectivos policiales y porcentaje de efectivos policiales procedentes de distritos atípicos según departamentos*
Nota: Elaboración propia

En la figura 25 se nos muestra una lista de los departamentos según el número de distritos con la cual pasamos a cotejar el porcentaje de policías procedentes de distritos atípicos. Guiándonos por porcentajes se presenta el número de policías procedentes de distritos atípicos por departamentos, siendo Callao con 86% de 7 distritos, Lima con 34% de 144 distritos y Tumbes con 23% de 13 distritos los que encabezan la enumeración de la gráfica mientras que Pasco con 3% de 29 distritos Ayacucho con 2% de 106 distritos y

Huancavelica con 1% de 88 distritos se ubican en los últimos puestos.

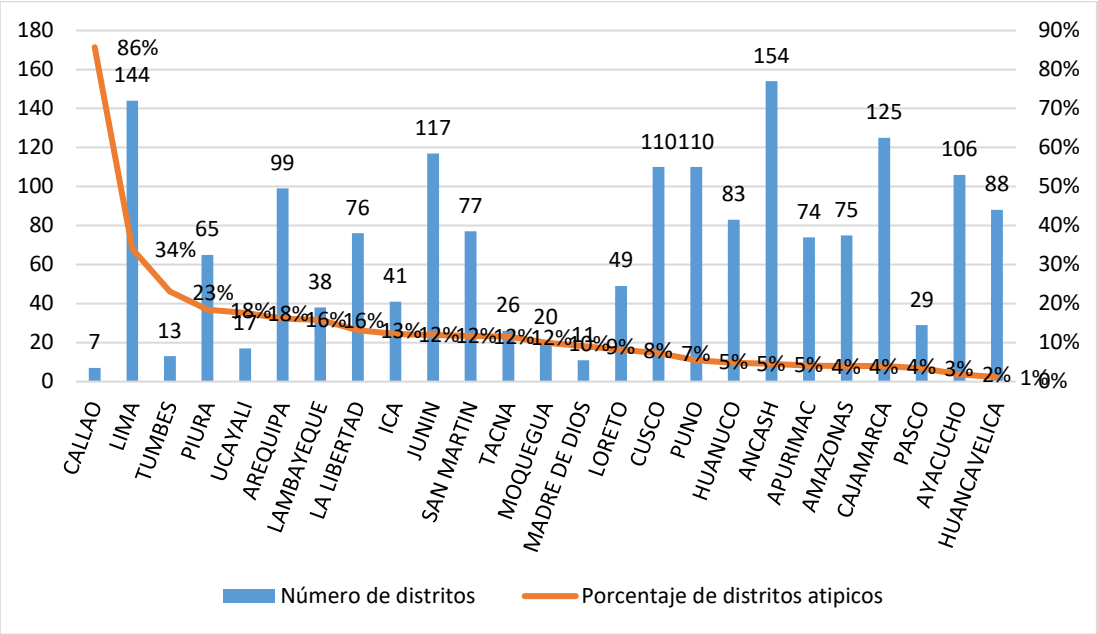


Figura 25. *Número de efectivos policiales por regiones respecto al porcentaje de distritos atípicos (respecto al número de efectivos policiales) por departamentos.*
Nota: Elaboración propia

En la figura 26 se procede a señalar el número de policías según las 25 provincias, dicha lista está encabezada por Lima con 12 915, Arequipa 2216 y Callao 1980 cuyo porcentaje de policías que proceden de distritos atípicos son respectivamente 91%, 65% y 63%; y en la parte final se ubican según Huamanga con 287 Huaura con 265 y Huaral con 191 y respectivamente con sus porcentajes de 74.6%, 50.2% y 83.6%.

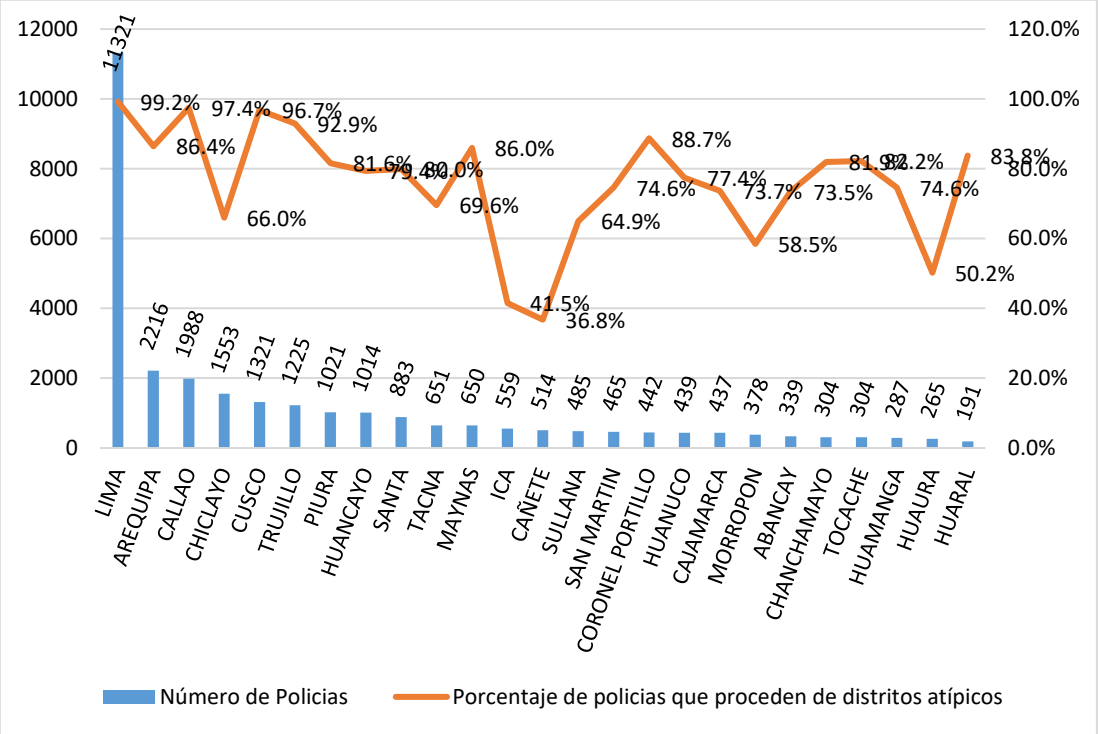


Figura 26. *Número de efectivos policiales procedente de distritos atípicos según 25 provincias con mayor número de efectivos.*
Nota: Elaboración propia

Examinando la figura 27 podemos retratar la gráfica cruzando el número de distritos con el porcentaje de distritos atípicos. Guiándonos según dichos porcentajes respecto al número de efectivos policiales vemos que encabezan la lista: Lima con 95% de 43 distritos, Callao con 86% de 7 distritos y Cusco con 75% de 8 distritos y cerrando las últimas posiciones se sitúan Cajamarca con 17% de 12 distritos, Ica con 14% de 14 distritos y Huamanga con 13% de 16 distritos, habiendo contemplado las 25 provincias.

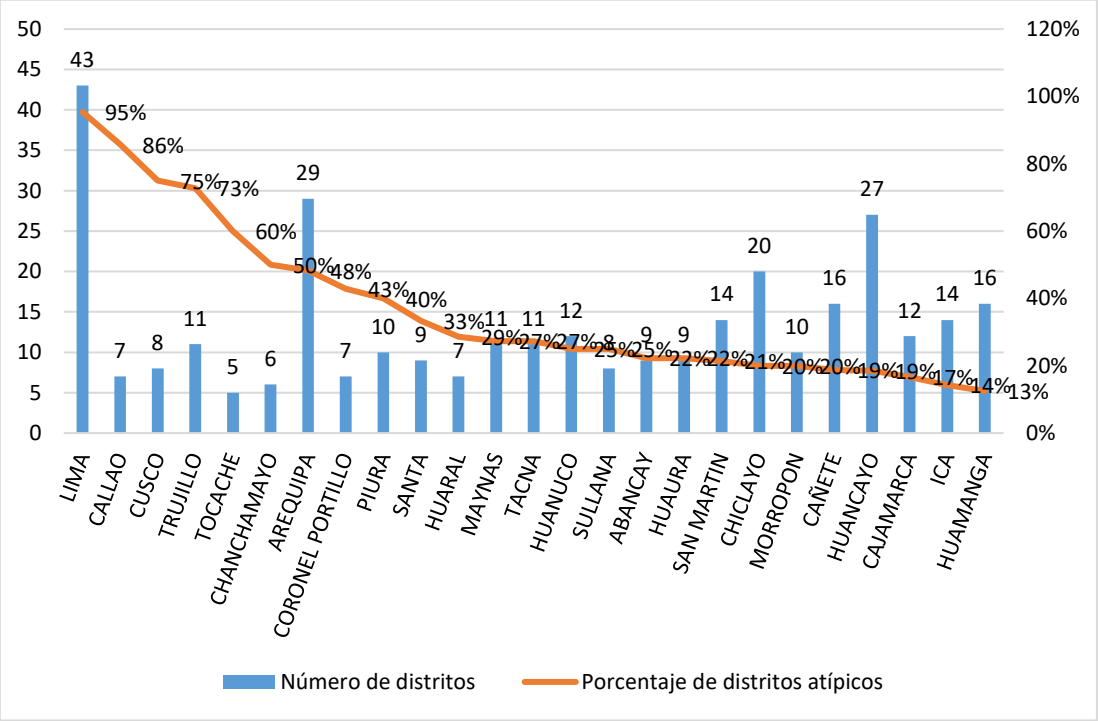


Figura 27. *Número de distritos respecto al porcentaje de distritos atípicos (respecto al número de efectivos policiales) según las 25 provincias con mayor número de efectivos policiales.*
Nota: Elaboración propia

Por otro lado, en la figura 28 se realiza una subdivisión para indicarnos el ordenamiento de los 25 distritos con el mayor número de efectivos policiales, siendo San Juan de Lurigancho con 1096, seguido de Callao 999 y de Lima con 876, mientras que Independencia con 347, Juliaca con 335 y Trujillo con 334 ubican las últimas posiciones.

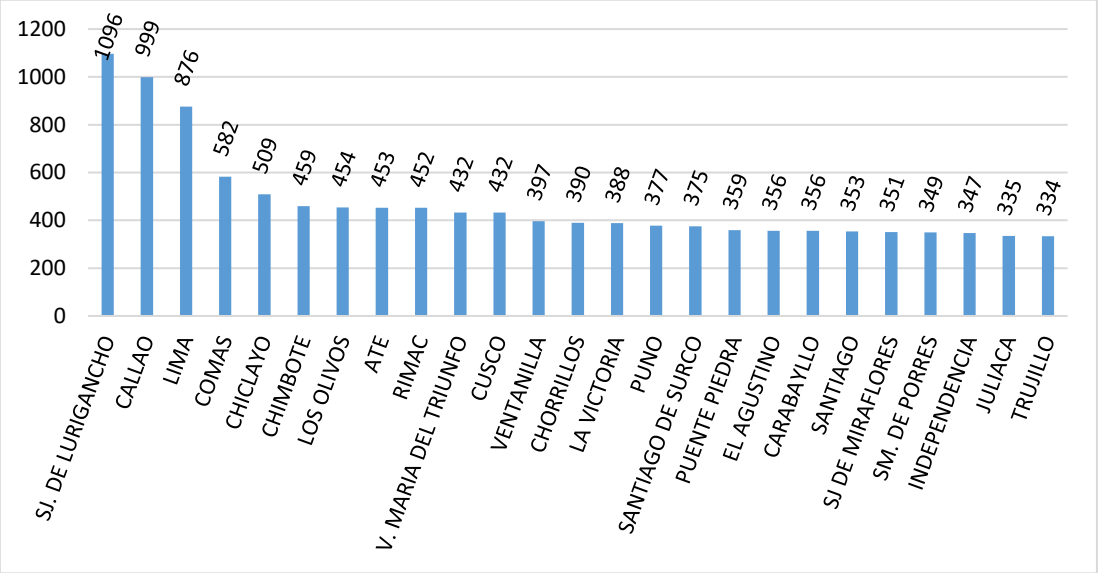


Figura 28. *25 distritos con el mayor número de efectivos policiales*
Nota: Elaboración propia.

E. Variable "número de efectivos de serenazgo"

Se observa en la figura 29 la distribución de la variable "Serenos", con una mediana en 3 serenos. La concentración de las observaciones está en el primer decil ya que hay más frecuencia de serenos, así mismo, la distribución de dicha variable es sesgada a la derecha. El promedio de la cantidad de serenos es de 18.08, además, con una desviación´ estándar de alrededor 77 serenos. Por otro lado, la cantidad mínima de serenos es de 0 y la máxima de 1 594 serenos.

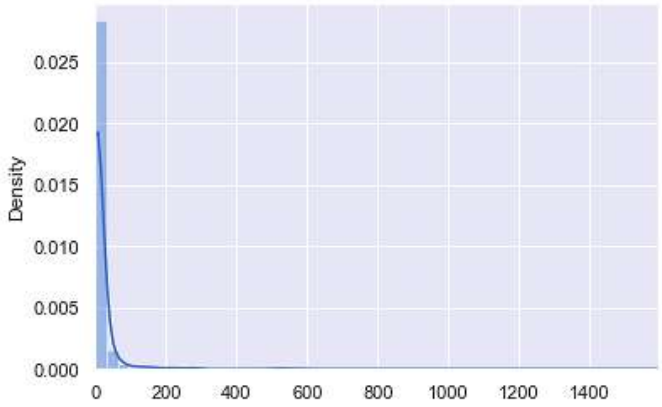


Figura 29. *Distribución de la variable número de efectivos de serenazgo en el distrito.*
Nota: Elaboración propia

Según la figura 29, el diagrama de cajas, la variable "Serenos" presenta un elevado número de valores atípicos, gráficamente representa los puntos negros, la caja está conformada por el bigote superior e inferior; en este caso 0 y 26.87.

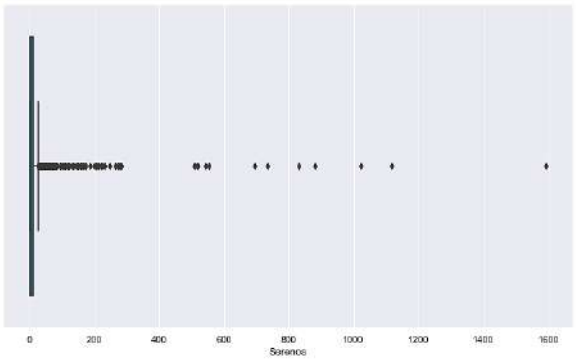


Figura 30. *Diagrama de cajas de la variable número de efectivos de serenazgo en el distrito.*
Nota: Elaboración propia

En la figura 31 podemos visualizar el número de serenos según las 25 regiones, cuya lista está encabezada por Lima con 14 182, La Libertad con 1 924 y Cusco con 1 656, mientras que en los últimos lugares se encuentra Madre de Dios, Amazonas y Moquegua respectivamente con 98, 95 y 91 serenos.

Adicionalmente la gráfica muestra los porcentajes de serenos que proceden de distritos atípicos, siendo los de mayor porcentaje Callao, Lima y Piura con 100%, 97% y 85% respectivamente; mientras que Tumbes (24%) Huánuco (23%) y Huancavelica (0%) serían las regiones de menor porcentaje.

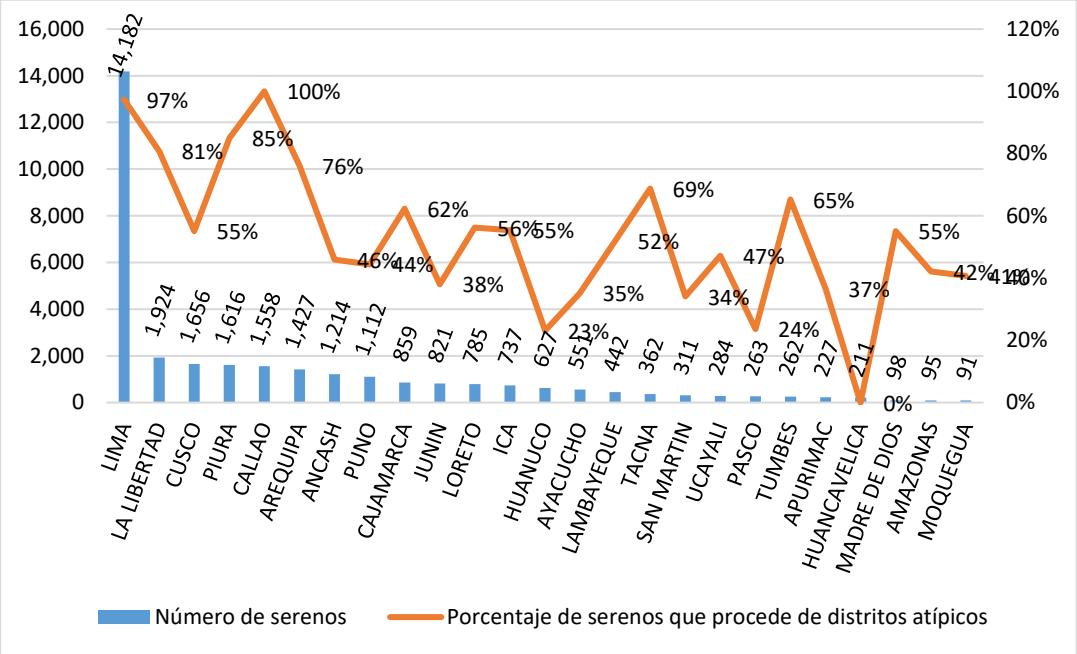


Figura *31* . *Número de efectivos de serenazgo y porcentaje de efectivos de serenazgo procedentes de distritos atípicos según departamentos.*
Nota: Elaboración propia

Viendo la figura 32 considerando tanto el número de distritos respecto al número de serenos por departamentos, pasamos a cotejar el porcentaje de los distritos atípicos. Según esto último hallamos entre las regiones con mayor porcentaje de distritos atípicos a Lima con el 100 % (de 144 distritos), Piura con 42% (de 65 distritos) e Ica con 31% (de 41 distritos). Por otra parte, entre las regiones con menor porcentaje de distritos atípicos figuran Huánuco con 3% (de 83 distritos), Amazonas con 2% (de 75 distritos) y Huancavelica con 1% (de 88 distritos).

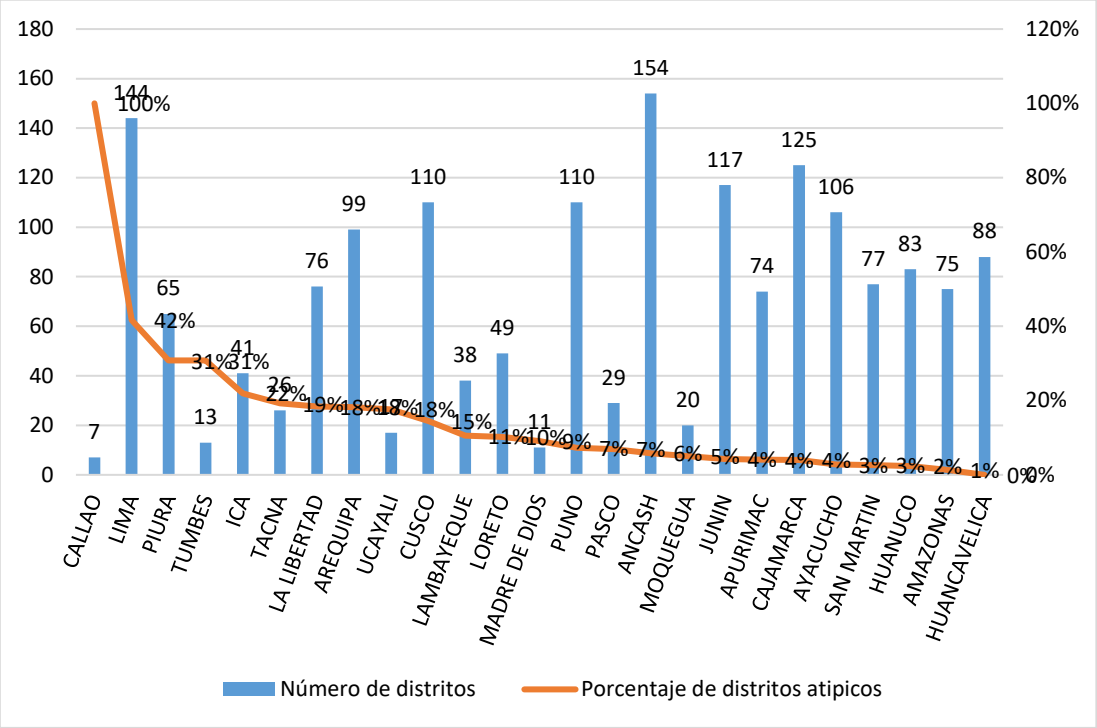


Figura 32. *Número de distritos y porcentaje de distritos atípicos (respecto al número de efectivos de serenazgo) por departamentos.*
Nota: Elaboración propia

Examinando las cifras descritas en la figura 33 podemos extraer el porcentaje de serenos que procede de distritos atípicos según las 25 provincias con mayor número de efectivos de serenazgo. Entre las provincias con mayor número de serenos están Lima con 13 046, Callao con 1 558 y Trujillo 1 266 mientras que en las últimas plazas se ubican Islay, Huaral y Alto Amazonas respectivamente con 143, 140 y 83 serenos.

Y principalmente, acerca del porcentaje de serenos que proceden de distritos atípicos vemos que la lista la encabezan Lima, Callao y Talara con el mismo porcentaje 100%, y finalizan la lista La Convención (62%), Ica (61%) y Huarochirí (52%) con el menor porcentaje.

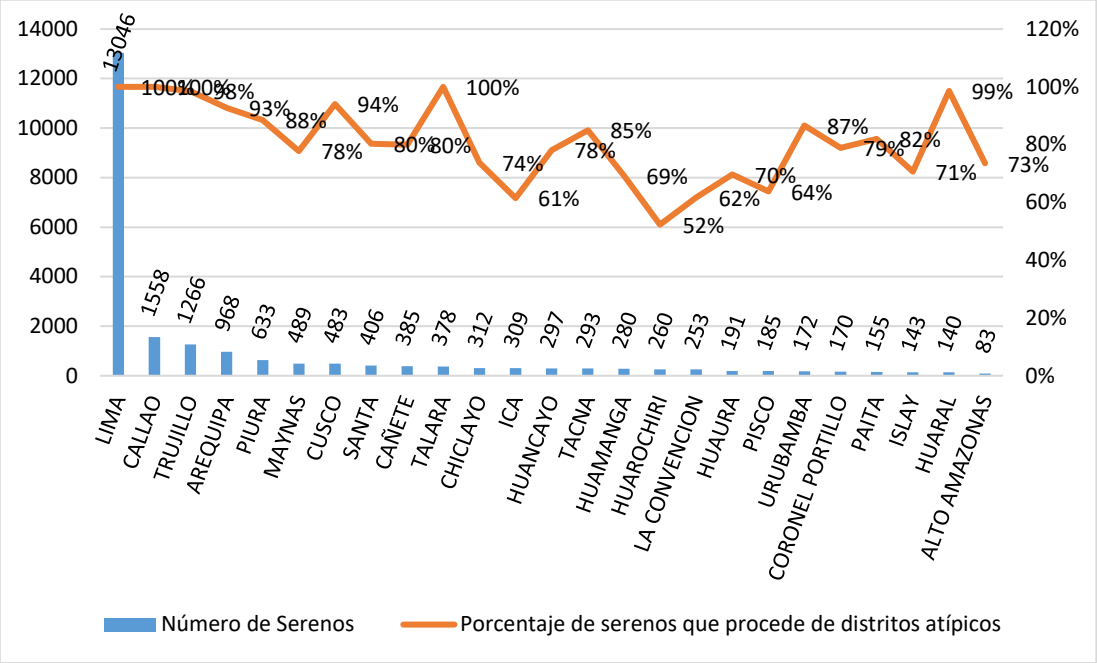


Figura 33. *Número de efectivos de serenazgo respecto al porcentaje de efectivos de serenazgo procedente de distritos atípicos según 25 provincias con mayor número de efectivos.*
Nota: Elaboración propia

En la figura 34 se nos muestra el número de distritos respecto al porcentaje de distritos atípicos (con relación al número de efectivos de serenazgo) según las 25 regiones con mayor número de serenos. Cotejando el porcentaje de distritos atípicos encontramos encabezando la lista con el mayor porcentaje por igual (100%) a Lima, Callao y Talara con 43 distritos, 7 distritos y 6 distritos respectivamente. Por otro lado, con menor porcentaje de distritos atípicos figuran Huamanga con 19% (de 16 distritos) Huarochirí con 12% (de 26 distritos) y Huancayo con 11% (de 27 distritos).

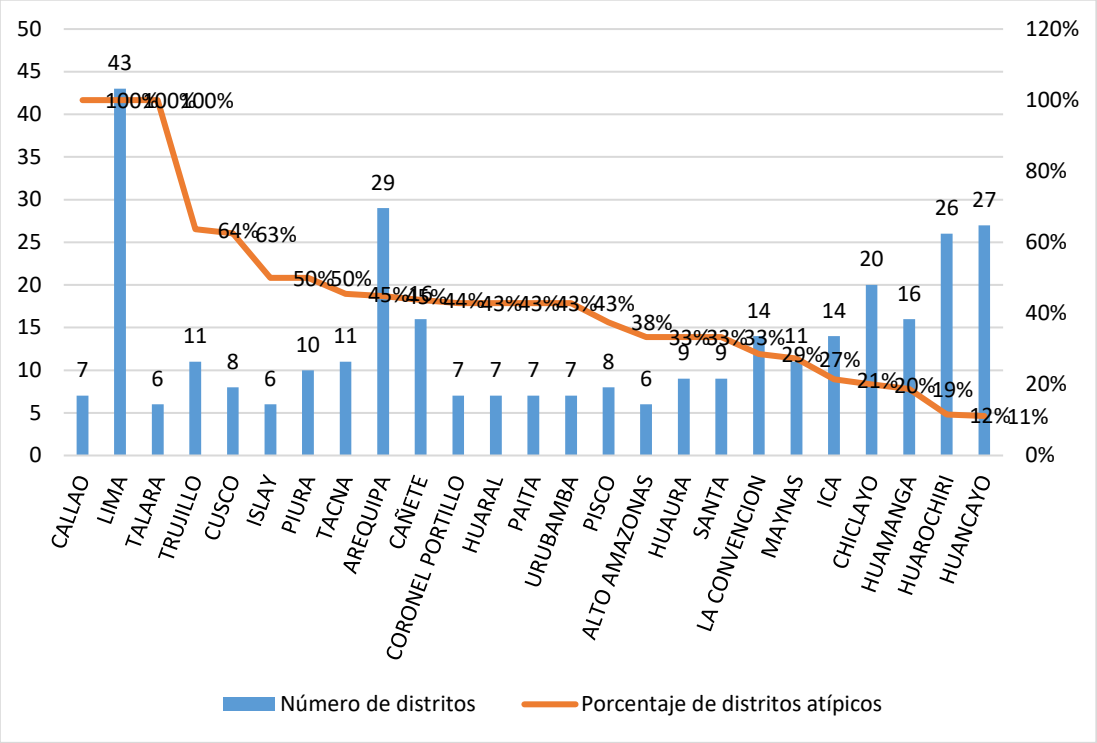


Figura 34. *Número de distritos respecto al porcentaje de distritos atípicos (respecto al número de efectivos de serenazgo) según las 25 provincias con mayor número de efectivos.*
Nota: Elaboración propia

Más internamente, en la figura 35 podemos ver un ordenamiento de los 25 distritos según el mayor número de efectivos de serenazgo, encabezando la lista Lima con 1 594, seguido de San Isidro con 1 116 y Santiago de Surco con 1 9999999022; mientras que Chimbote con 231, Callao con 227 y Cusco con 220 serenos se ubican en las últimas posiciones.

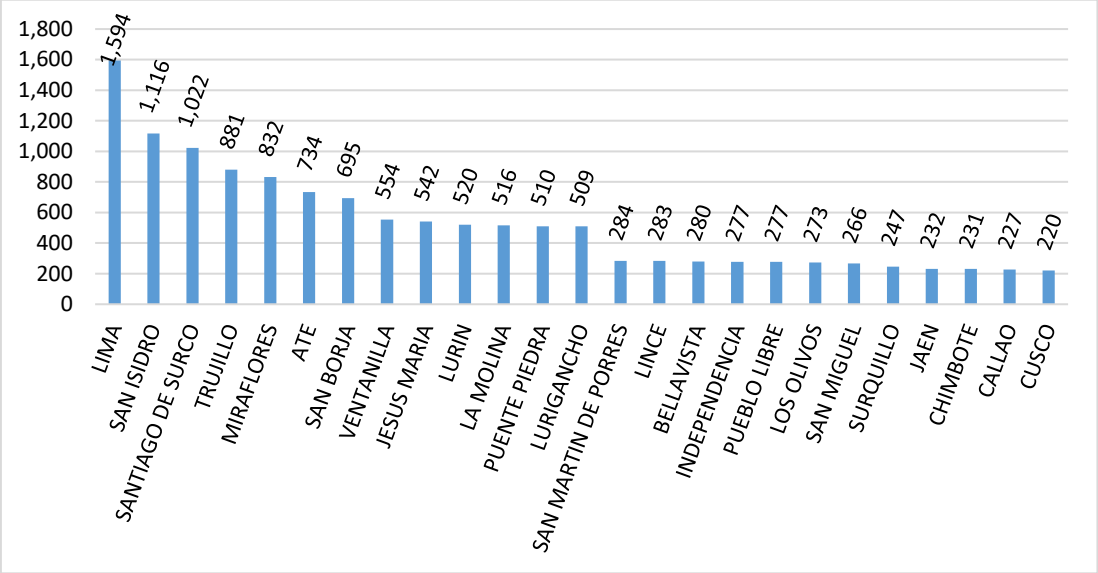


Figura 35. *25 distritos con el mayor número de efectivos de serenazgo.*
Nota: Elaboración propia

F. Variable "último distrito de residencia del interno"

Se observa en la figura 36 la distribución de la variable "Internos", con una mediana en 3 internos. La concentración de las observaciones está en el primer decil ya que hay más frecuencia de internos, así mismo, la distribución de dicha variable es sesgada a la derecha. El promedio de la cantidad de internos es de 47.98, además, con una desviación´ estándar de alrededor 186.78 internos. Por otro lado, la cantidad mínima de internos es de 0 y la máxima de 3,388 internos.

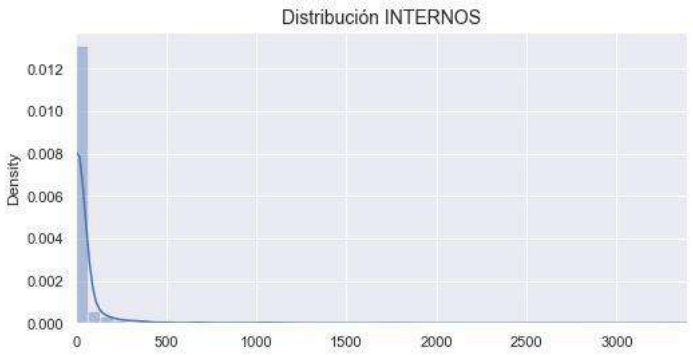


Ilustración 36. *Distribución de la variable último distrito de residencia del interno.*

Nota: Elaboración propia

Según la figura 37, el diagrama de cajas, la variable "Internos" presenta un elevado número de valores atípicos, gráficamente representa los puntos negros, la caja está conformada por el bigote superior e inferior; en este caso 0 y 39.5.

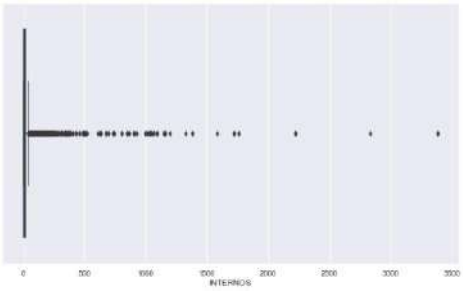


Figura 37. *Diagrama de cajas de la variable último distrito de residencia del interno.*
Nota: Elaboración propia

Analizando la figura 38 según el número de internos procedentes de alguna de las 25 regiones, con el mayor número de internos hallamos a Lima con 28 426, seguida de La Libertad con 5672 y Callao con 4560; mientras que Pasco, Huancavelica y Moquegua ocupan los últimos lugares con 654, 588 y 340 internos, respectivamente.

Adicionalmente, con el mayor porcentaje de internos que proceden de distritos atípicos se encuentran el Callao con 100 % seguido de Lima con 98% y Tacna con 96%. Y con el menor porcentaje de internos procedentes de distritos atípicos figuran Apurímac, Amazonas y Huancavelica con 55%, 44% y 16% respectivamente.

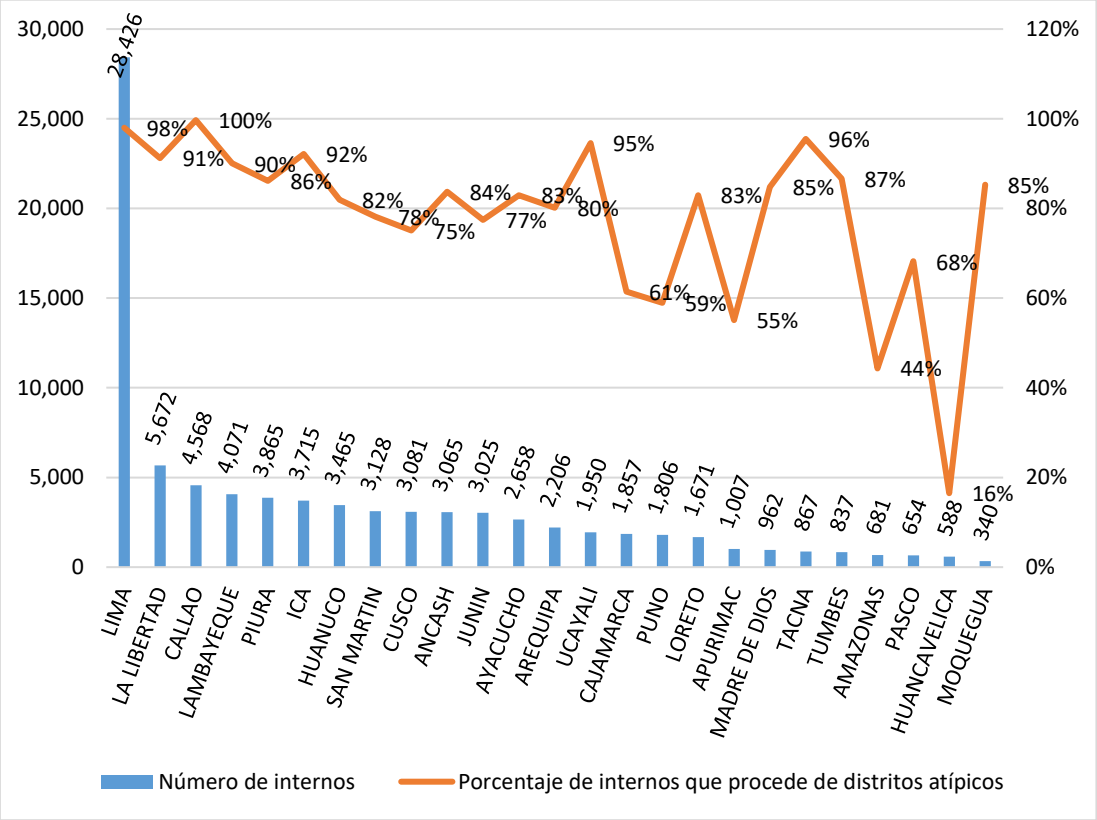


Figura 38. *Número de internos que indicaron que residieron en la región y porcentaje de internos procedentes de distritos atípicos según departamentos*
Nota: Elaboración propia

En la figura 39 se nos expone el número de distritos y porcentaje de distritos atípicos (respecto al número de internos que residieron en el distrito) según departamentos. En función a los distritos atípicos hallamos, encabezando la lista de mayor porcentaje, a Callao con 71% (de 7 distritos) Ica con 41% (de 41 distritos) y Ucayali con 41% (de 17 distritos). Mientras que con menor porcentaje de distritos atípicos figuran Puno con 5% (de 110 distritos), Amazonas con 4% (de 75 distritos) y Huancavelica con 1% (de 88 distritos) cerrando los últimos puestos.

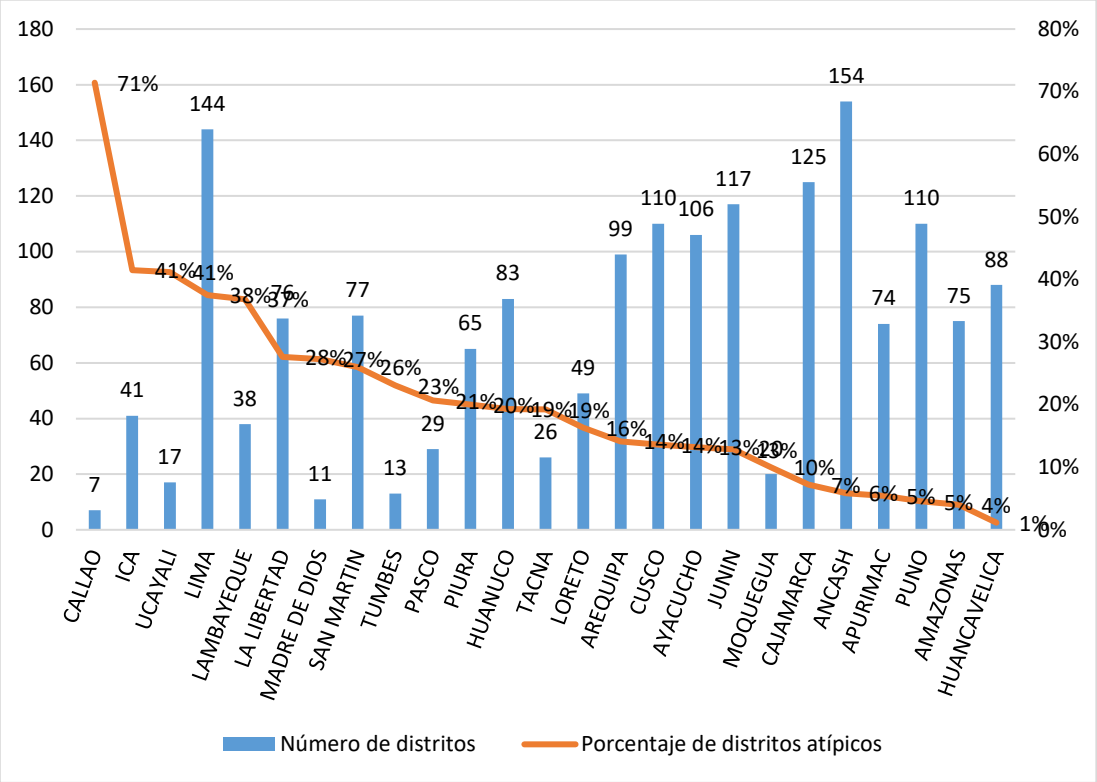


Figura 39. *Número de distritos y porcentaje de distritos atípicos (respecto al número de internos que residieron en el distrito) según departamentos*
Nota: Elaboración propia

En la figura 40 pasamos a cotejar tanto el número de internos con el porcentaje de internos que proceden de distritos atípicos en las 25 provincias con mayor número de internos. Según esto último entre las provincias con mayor porcentaje de distritos atípicos hallamos al Callao (4560 internos) y Trujillo (3927 internos) con el 100% seguido de Lima (25 339 internos) con el 99%. Mientras que con el menor porcentaje que proceden de distritos atípicos hallamos a Huaura (691 internos), Ascope (444 internos) y Oxapampa (401 internos) con 86%, 80% y 78% respectivamente.

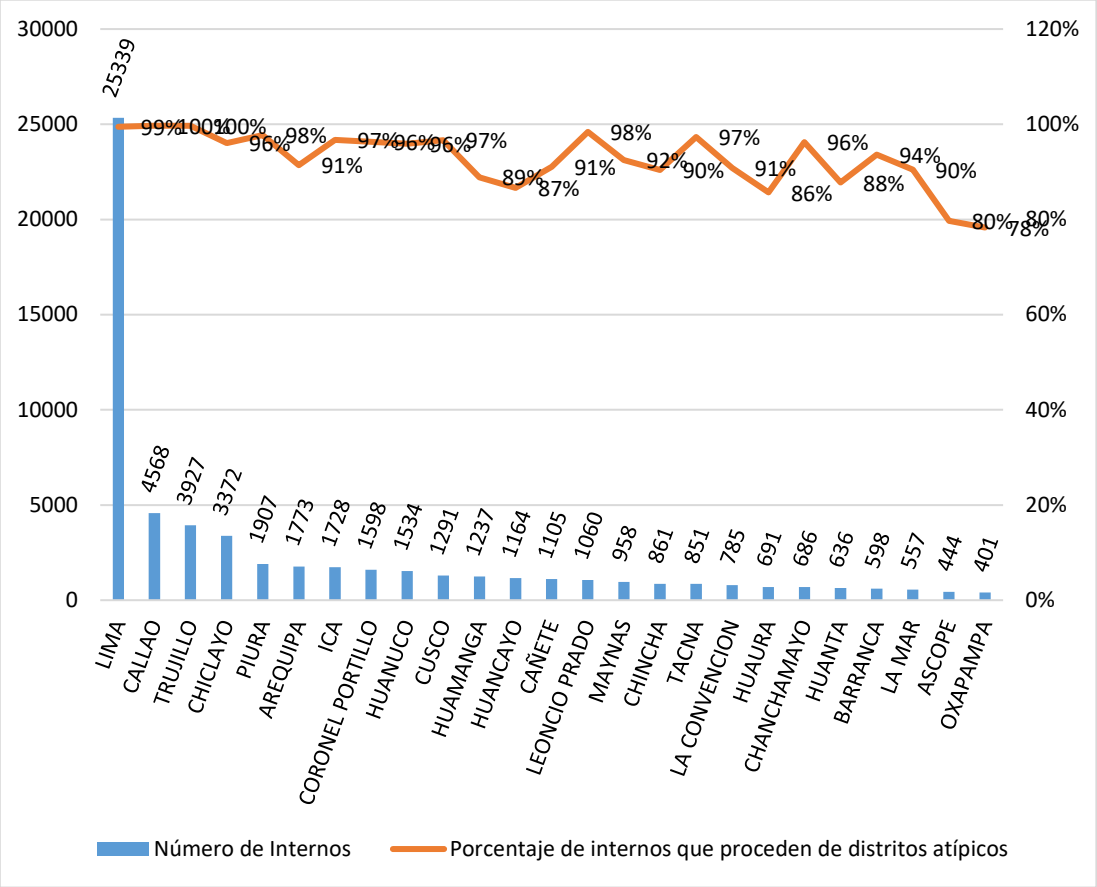


Figura 40. *Número de internos que residieron en la provincia y porcentaje de internos procedentes de distritos atípicos en las 25 provincias con mayor número de internos*
Nota: Elaboración propia

En la figura 41 observamos por un lado el número de distritos y por otro el porcentaje de distritos atípicos con relación al número de internos que residieron en el distrito según las 25 provincias con mayor número de efectivos. En función a los distritos atípicos, encabezando la lista de mayor porcentaje, hallamos a Lima con el 84% (de 43 distritos) seguido de Trujillo con 82% (de 11 distritos) y Barranca con 80% (de 5 distritos). Mientras que con el menor porcentaje de distritos atípicos figuran La Convención con 36% (de 14 distritos), Huamanga con 25% (de 17 distritos) y Huancayo con 22% (de 27 distritos).

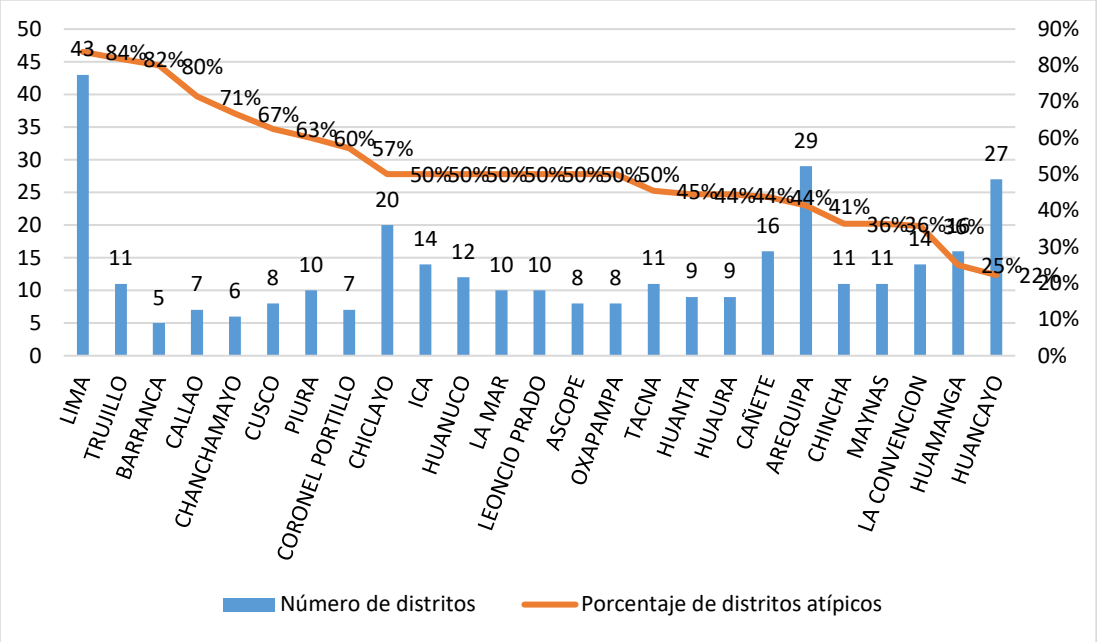


Figura 41. *Número de distritos según las 25 provincias con mayor número de efectivos.*
Nota: Elaboración propia

En cifras consolidadas, la figura 42 nos muestra el listado de los 25 distritos con el mayor número de internos que residieron en dichos distritos, encabezada por San Juan de Lurigancho con 3388, seguido por el Callao con 2837 y Lima con 2223 internos; mientras que Callería con 804, Huánuco con 744 y Los Olivos con 733 internos son los distritos con menor cantidad.

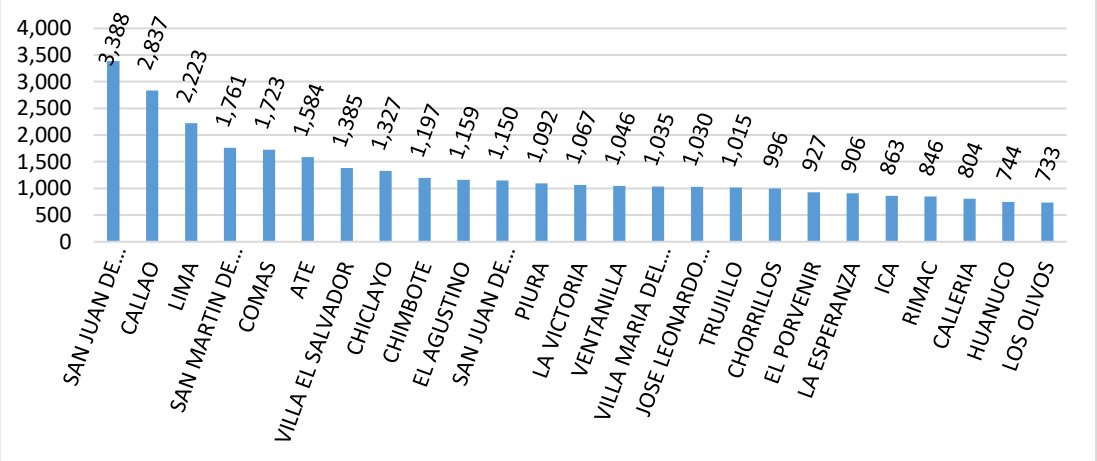


Figura 42. *25 distritos con el mayor número de internos que residieron en dicho distrito.*

Nota: Elaboración propia

La figura 43 se aprecia la correlación lineal entre las variables del estudio, en la cual se observa que la variable internos es la que muestra mayor nivel de asociación lineal con el resto de las variables, presentando su más alto nivel de asociación con población (0.90), seguido de policías (0.89) y muertes (0.88); mientras que su menor nivel de asociación es con el número de serenos en los distritos (0.52).

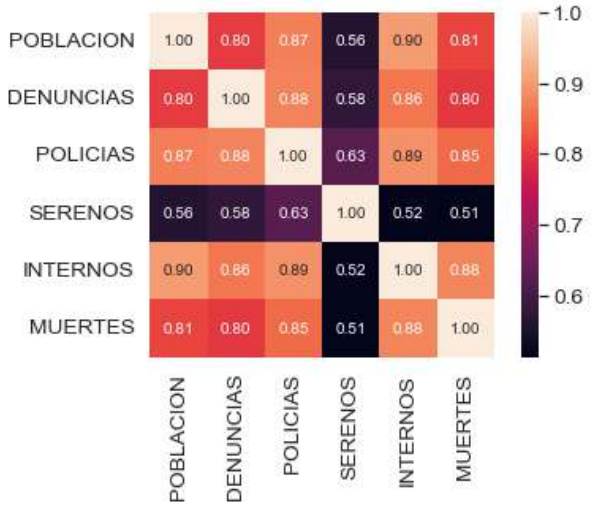


Figura 43. *Correlaciones de las variables investigadas*
Nota: Elaboración propia. Reporte generado del R

1.1. Escalamiento Multidimensional (EMD)

El objetivo principal del EMD es crear coordenadas (sintéticas) en los datos previamente escalados, a partir una matriz de distancias. “D” representa la matriz de distancia para el análisis de EMD, k representa la dimensión y el valor predeterminado es 2 dimensiones.

EMD también se puede utilizar para revelar un patrón oculto en una matriz de correlación. Según la figura 44, podemos agrupar los comportamientos en las dimensiones de la siguiente forma:

- Dimensión 1: Internos, Población INEI, Muertes y Policías.

- Dimensión 2: Denuncias.

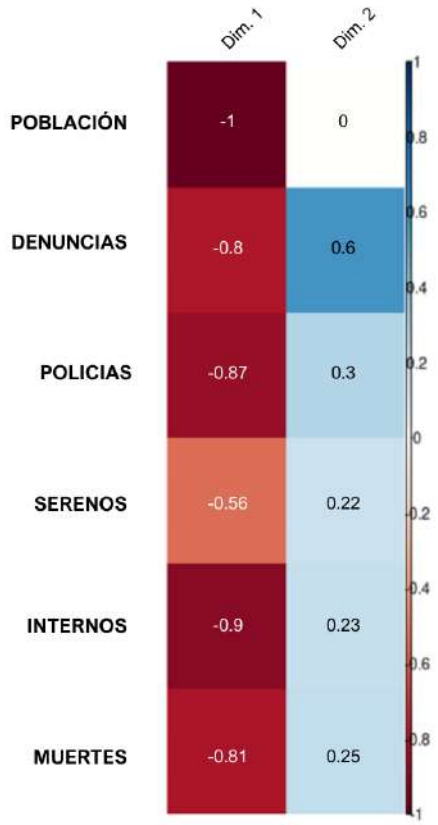


Figura 44. *Correlaciones de las variables investigadas según dimensiones del EMD.*
Nota: Elaboración propia. Reporte generado del R

En términos generales y con alguna excepción podemos interpretar que la dimensión uno recoge comportamientos demográficos como población e internos. La dimensión dos recoge comportamientos asociados al riesgo delictivo como en este caso de la variable denuncias.

1.3. Clúster por Escalamiento Multidimencional

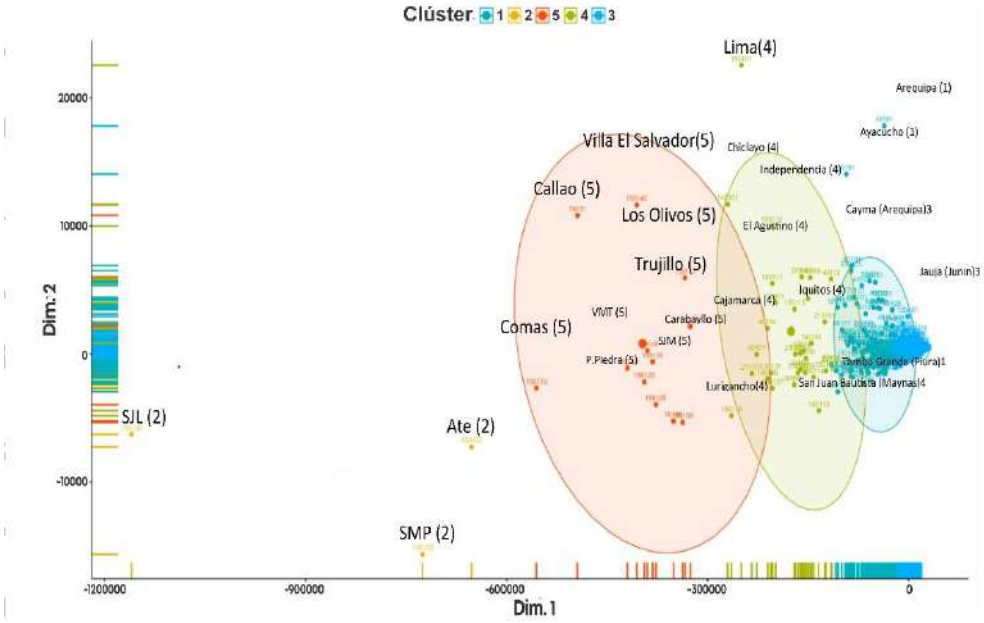
Desde el punto de vista matemático y conceptual, existen estrechas correspondencias entre el análisis de EMD y otros métodos utilizados para reducir la dimensionalidad de datos complejos, como el análisis de componentes principales (ACP).

El ACP se centra más en las propias dimensiones y trata de maximizar la varianza explicada, mientras que el análisis EMD se centra más en las relaciones entre los objetos escalados.

El EMD proyecta puntos de datos n-dimensionales a un espacio (normalmente) bidimensional, de forma que los objetos similares en el espacio n-dimensional estarán próximos en el gráfico bidimensional, mientras que el ACP proyecta un espacio multidimensional a las direcciones de máxima variabilidad utilizando la matriz de covarianza/correlación para analizar la correlación entre los puntos de datos y las variables.

Los resultados del análisis clúster por EMD se aprecia la conformación de 05 clústeres como se evidencia en la Figura 45, se observa que el **clúster 2** se encuentra ubicado en el extremo negativo de la dimención 1, esta conformado por solo 3 distritos SJL, SMP y Ate, distritos que registran las más altas frecuencias de las variables analizadas, es decir estan asociados a la mayor intensidad de la actividad delictiva, Luego sigue el **clúster 5** conformado por los 12 distritos siguientes: Callao, Comas, Los Olivos, Ventanilla, Puente Piedra, Trujillo, San Juan de Miraflores, Villa María del Triunfo, Carabayllo, Santiago de Surco, Villa e Salvador y Chorrillos; estos distritos siguen presentando alta insidencia delictiva, es decir, altas frecuencia de variables investigadas pero en menor medida que el grupo anterior. Luego, alejandonos del extremo del eje negativo de la dimensión 1 esta el **clúster 4**, conformado por 32 distritos con menor frecuencia de la actividad delictiva. Posteriormente, más proximo al centro del eje de coordenadas esta representado el **clúster 1,** conformado por 120 distritos con menor actividad delictiva. Finalmente encontramos el **clúster 3,** es el mas numeroso con 1587 distritos y se caracterisa por tener poca frecuencia en las variables investigadas, es decir poca frecuencia de actividades delictivas.

Es importante resaltar que el **clúster** 2, **clúster** 5, **clúster** 4 y **clúster** 1 en total suman 167 distritos que representan los ambitos territoriales de mayor insidencia delictiva a nivel nacional. De los 1875 distritos investigados a nivel nacional, 167 representan solo el 9% de distritos, sin embargo en este grupo de distritos concentra la mayor actividad delictiva del Perú, tal es asi que los ambitos territoriales de los 167 distritos identificados concentra el 81.1% de las denuncias registradas por la PNP a nivel nacional, 75.9% de internos de los penales y el 67.4% de muertes asociadas a hechos delictivos dolosos.



Nota: Elaboración propia. Reporte generado del R.

Figura 45. *Clústeres de distritos generados a partir del análisis de EMD.*

1.4. Análisis de Componentes Principales (ACP)

El método de componentes permite "condensar" la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como el análisis clúster.

En la tabla 01 y en la figura 46 se muestra la varianza explicada por los componentes, en la cual se puede precisar que con solo dos componentes se puede capturar el 91% de la varianza, prácticamente la totalidad de la información contenida en la expresión de las 6 variables. Viendo en detalle el valor de las dos primeras componentes se observa que en la primera componente tiene un 80% de varianza y la segunda un 10%.

Tabla 1. *Resultados del análisis de componentes principales*

	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.1959	0.7812	0.45149	0.43705	0.31797
Proportion of variance	0.8037	0.1017	0.03397	0.03184	0.01685
Cumulative Proportion	0.8037	0.9054	0.93936	0.97119	0.98805

Fuente: Elaboración propia. Reporte generado del R.

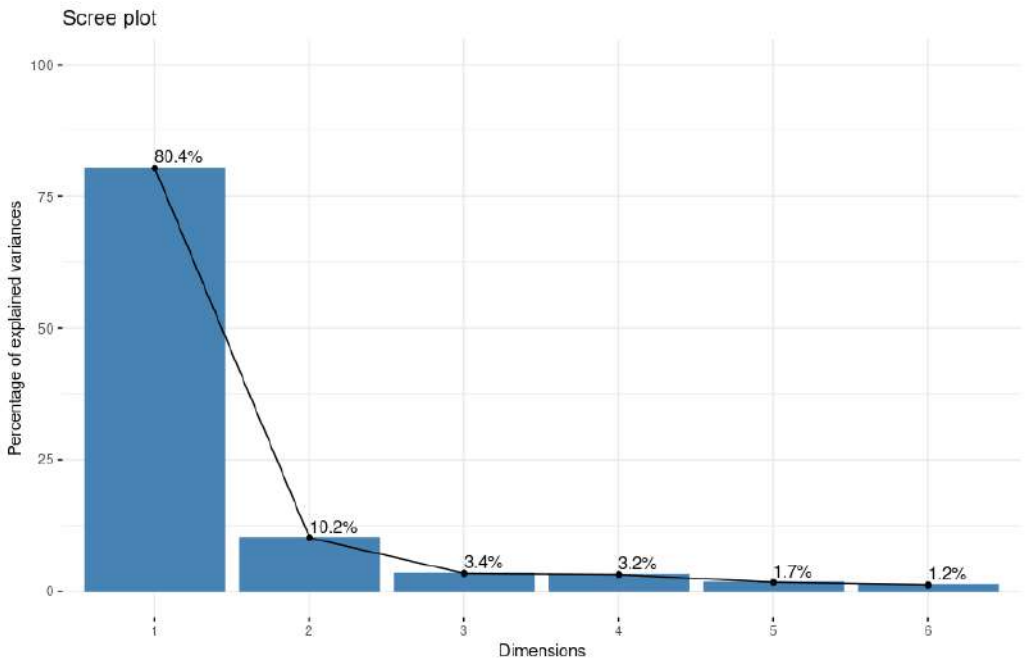


Figura 46. *Varianza explicada por el ACP.*
Nota: Elaboración propia. Reporte generado del R.

Respecto a la figura 47, se puede apreciar muy buena representación de las variables policías, internos, población, denuncias y muertes con la componente 1, respecto a la componente 2 se observa buena representación de la variable serenos.

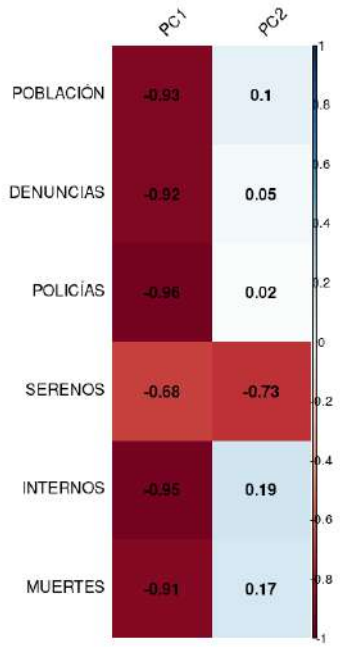


Figura 47. *Correlaciones de las variables investigadas según componentes del ACP.*

Nota: Elaboración propia. Reporte generado del R

1.5. Clúster por componentes principales

Este algoritmo de clasificación no supervisado agrupa objetos en k grupos basándose en la mínima suma de distancias entre cada objeto y el centroide de su clúster. A pesar de que la agrupación de variables puede ayudar en el análisis de la totalidad de datos, el proceso se hace cada vez más demandante entre mayor sea el número de variables. Por esta razón el ACP se hace una herramienta importante para identificar las variables que mayor aportan a la variabilidad de datos y trabajar con estas únicamente. Respecto al número óptimo de clúster, se puede realizar una clasificación máxima de 10 clúster, pero se evidencia que dos clústeres pueden ser suficientes, como se observa en la figura 48.

Los distritos que integran el grupo 2 y se encuentran más relacionados con el CP1 se caracterizan por tener mayor número de policías, internos, denuncias, mayor tamaño de población y a medida que se relacionan con el CP2 se aproxima a un mayor número de serenos. Respecto a la figura 48 se puede indicar que el clúster 2 tiene mayor representación frente al clúster 1, se puede precisar que el 70% de distritos inseguros, 69% de policías y 61% de la población investigada se encuentran en el grupo 2.

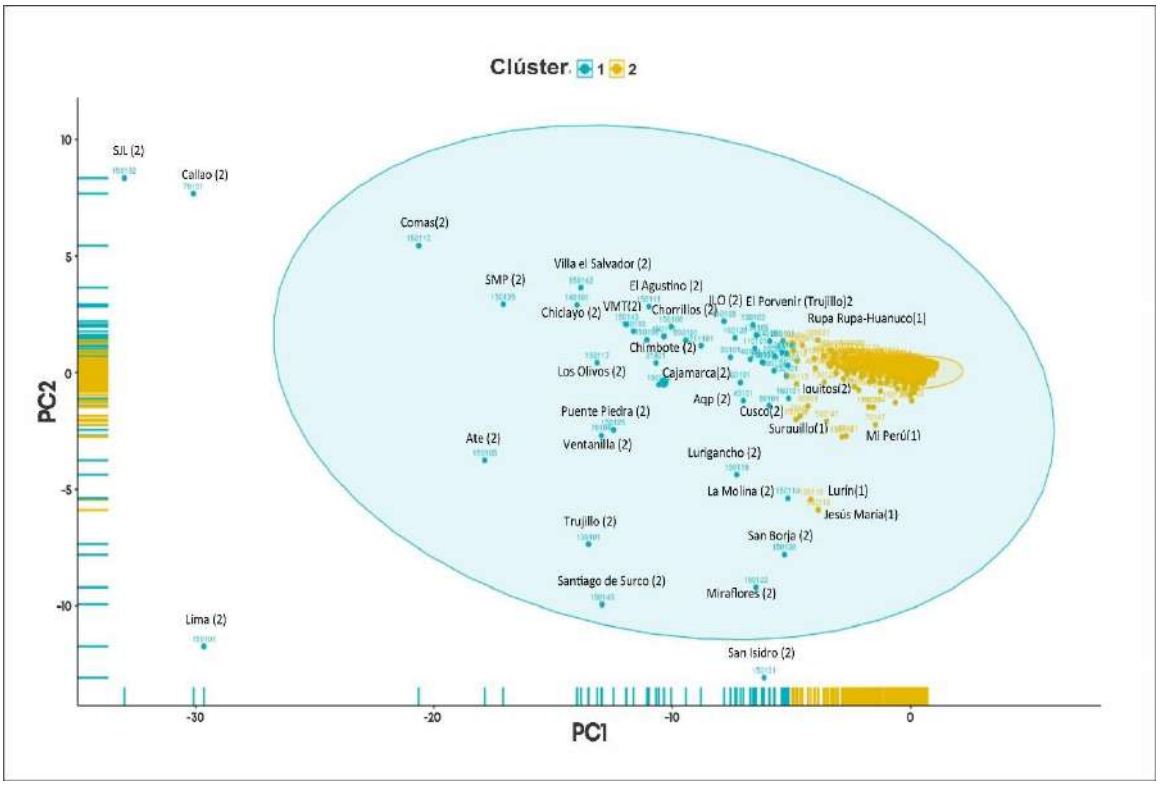


Figura 48. *Clústeres de distritos generados a partir del análisis de componentes principales.*
Nota: Elaboración propia. Reporte generado del R.

1.6. Relación entre variables y el EMD

En relación con las Figuras 49 y 50, se puede observar que la correlación más fuerte se encuentra entre los pares de variables "internos - población", con una correlación de 0.9, "internos - policías" (0.89) e "internos - muertes" (0.88). Además, se destaca la correlación de "policías - denuncias" (0.88) y "policías - población" (0.87). Las

variables "internos" y "policías" muestran una correlación más intensa que con el resto de las variables, mientras que la variable "serenos" muestra una relación menos intensa. En todas las variables investigadas se observa una fuerza de asociación que va entre magnitud moderada y fuerte; en todos los casos se presenta una relación positiva, es decir, a medida que aumenta la variable en el eje de coordenadas, también aumenta la variable en el eje de las abscisas.

En la Figura 49, se puede observar que las variables con mayor correlación presentan una clara formación de clúster mediante EMD. Estos puntos se distribuyen de acuerdo con el nivel de inseguridad de los distritos. Es decir, en la base de los diagramas de dispersión, se encuentran los puntos de color lila (1587 distritos), caracterizados por un nivel muy bajo de inseguridad, seguidos por los puntos rojos (120 distritos) con un bajo nivel de inseguridad. Los puntos anaranjados (32) representan un nivel intermedio de inseguridad, los puntos verdes (12) indican un nivel alto, y los puntos azules (3) reflejan un nivel muy alto de inseguridad.

En la figura 50, se aprecia un comportamiento similar a la figura anterior; en este caso, ubica en la parte superior de los diagramas de dispersión a los puntos de color azul que representan los distritos inseguros de la variable predictora.

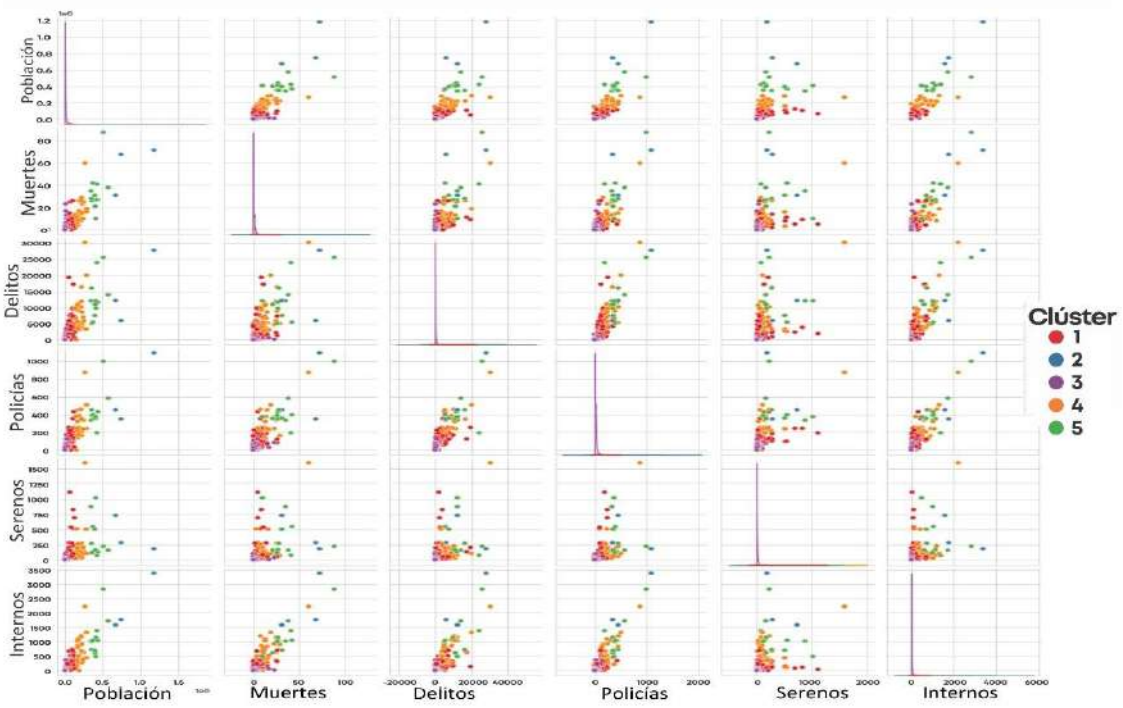


Figura 49. *Diagramas de dispersión distinguiendo la conformación de clústeres del análisis de EMD.*
Nota: Elaboración propia. Reporte generado del Python.

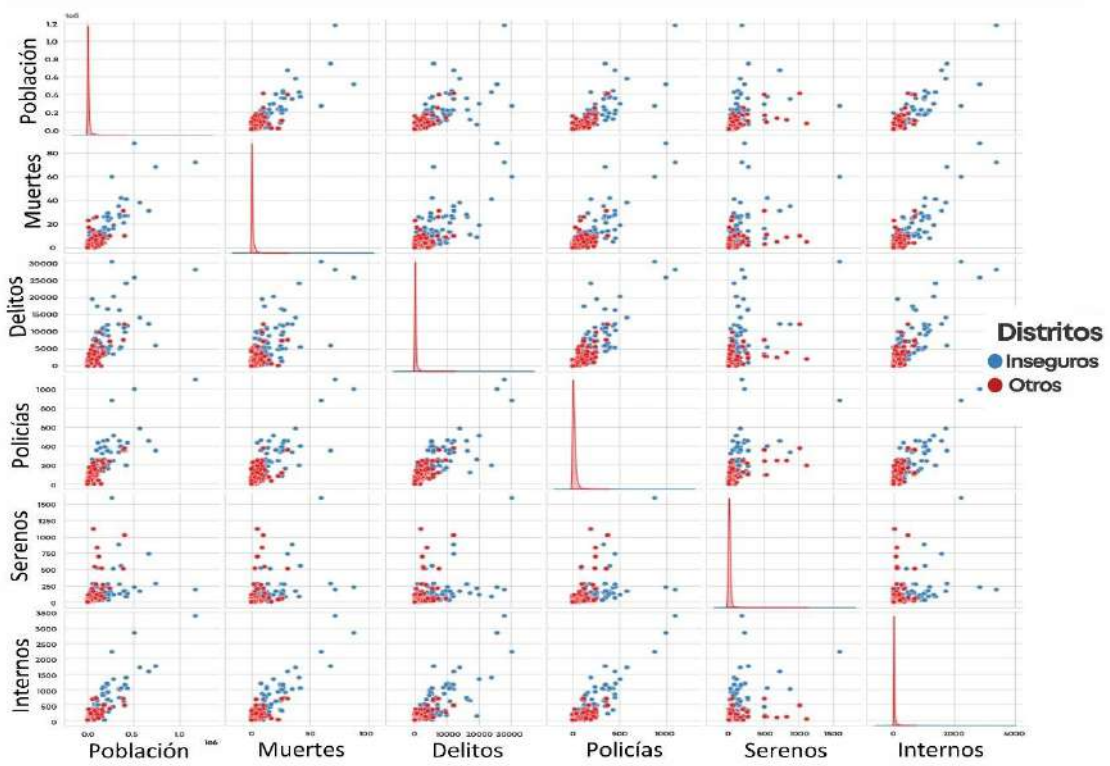


Figura 50. *Diagramas de dispersión de la variable predictora "inseguro".*
Nota: Elaboración propia. Reporte generado del Python.

1.7. Distribución espacial de la actividad delictiva en los distritos del Perú

La investigación examina la distribución geográfica de la actividad delictiva en el Perú a través de la elaboración de mapas temáticos, los cuales muestran una representación visual nítida y fácil de entender de las variables investigadas a nivel del ámbito territorial de distritos. A continuación, se describen los mapas temáticos elaborados.

En la Figura 51, se observa que la costa norte y centro son las áreas con la mayor concentración de población. A nivel de departamentos, es notable que, en Piura y Lambayeque, prácticamente todos los distritos que los conforman presentan niveles elevados de

población representados por el color rojo. En contraste, Tacna y Moquegua registran distritos con bajos niveles de población. Después de la región costa, sigue la Selva, con una presencia significativa de distritos con altos niveles de población. Finalmente, la región Sierra muestra distritos con una menor concentración de población. Se pueden observar distritos en las regiones de Ayacucho y Apurímac con una baja frecuencia de población, es decir, una mayor presencia de distritos representados en color plomo.

En la Figura 52 se representan distintas realidades para las regiones en relación con la variable "denuncias registradas por la PNP". En la Costa, se evidencia una concentración significativa de denuncias a lo largo de la región, con mayor incidencia en los distritos de departamentos como Lambayeque, Piura y La Libertad, que exhiben una alta frecuencia de denuncias identificada con el color rojo. En la Sierra, los departamentos muestran una menor presencia del color rojo, indicando variabilidad en la pertenencia a diferentes escalas de los distritos, incluyendo aquellos sin denuncias, como es el caso de departamentos como Ayacucho y Huancavelica, que presentan una baja frecuencia de denuncias. Finalmente, la región Selva muestra una distribución más equitativa de las denuncias, con una presencia moderada de distritos en color rojo, siendo el departamento de Madre de Dios un ejemplo que ilustra esta situación.

En el mapa representado en la figura 53, en relación con la variable número de policías, observamos una distribución desigual a lo largo de la región de la Costa. Piura y Lambayeque son los departamentos con mayor presencia de distritos de alta frecuencia, identificados en color rojo. En cuanto a la región Sierra, notamos que la distribución de policías es más diversa, siendo Pasco y Junín los departamentos con mayor concentración de distritos de alta frecuencia. Finalmente, la región selva persiste la problemática de no

contar con una cantidad adecuada de efectivos policiales en sus departamentos, siendo Madre de Dios la única excepción. En los demás departamentos de la Selva, se observa una mayor cantidad de distritos con poca frecuencia de policías en comparación con el resto de los departamentos.

En el mapa de la figura 54, en relación con el número de efectivos de serenazgo, se observa una alta frecuencia en los departamentos de la región Costa, especialmente en los distritos cercanos al litoral. Por otro lado, en la sierra se aprecia una menor concentración de serenos, salvo en los casos de los departamentos de Cuzco y Pasco, donde se encuentran distritos con mayor presencia del servicio de serenazgo. Finalmente, en la región selva se evidencia una menor concentración de efectivos de serenazgo, atribuible a la menor densidad demográfica de dicha región en comparación con la extensión de su territorio, lo cual podría explicar esta situación.

La figura 55 presenta la distribución geográfica de la variable correspondiente a la última residencia de los internos. Las zonas en color rojo indican la concentración más alta de personas que actualmente se encuentran internadas en algún penal y que previamente residían en dicho distrito. Los departamentos ubicados en la costa norte (Tumbes, Piura, Lambayeque y La Libertad) muestran una concentración significativa de internos, mientras que en la costa centro la concentración es menor; en la costa sur, los distritos en color rojo están más dispersos. En la sierra, la concentración de internos disminuye, pero se observa que departamentos como Cusco, Junín, Pasco y Huánuco registran una concentración alta de internos. En la selva, el color rojo en los distritos se presenta de manera más dispersa, con la excepción de San Martín, que muestra una alta concentración de internos en gran parte de su territorio.

En la Figura 56 se presenta la distribución geográfica de las muertes violentas a nivel distrital. Se observa que los distritos con mayor concentración están dispersos en las regiones costa, sierra y selva. En la costa, Lima, Lambayeque y Piura son los departamentos que registran mayor número de casos, destacando los distritos de Callao, Comas y San Juan de Lurigancho con alta frecuencia de muertes. En la sierra, Huánuco, Junín y Cuzco son los departamentos con mayores registros, y Juliaca, El Tambo y Cerro Colorado son los distritos con alta frecuencia delictiva. Por último, en la selva, Madre de Dios y Ucayali son los departamentos con más registros, mientras que Callería, Tambopata e Inambari son los distritos con mayor incidencia de muertes violentas.

En la Figura número 57 se presentan los resultados del análisis de clúster por EMD. Se puede observar la conformación de 5 clústeres, siendo el tercero el más extenso y está representado en el mapa con los distritos en color gris, los cuales abarcan el 91% de los distritos a nivel nacional. Este clúster está compuesto por 1,587 distritos que se caracterizan por tener una baja frecuencia de actividad delictiva. El otro 9% de los distritos restantes, es decir, 167, se distribuyen entre los clústeres 1, 2, 4 y 5, los cuales se distinguen por tener una alta frecuencia de actividad delictiva. La mayoría de estos distritos se encuentran en la región costa, especialmente en la costa central (Lima y Callao) y la costa norte (Piura y Lambayeque), aunque en menor medida se ubican en la región sierra, en los departamentos de Cajamarca, Junín, Cusco y Puno, y tienen una presencia aún menor en la región selva, en los departamentos de Loreto, Ucayali y San Martín.

En el marco de los 167 distritos de alta frecuencia delictiva podemos diferenciarlos en 4 clúster, cada uno identificado en el mapa temático por un color específico. El clúster 2, representado en rojo

oscuro, está compuesto por los distritos SJL, SMP y Ate, y se caracteriza por tener la mayor actividad delictiva. Por otro lado, el clúster 5, marcado en rojo claro, está integrado por 12 distritos (Callao, Comas, Los Olivos, Ventanilla, Puente Piedra, Trujillo, San Juan de Miraflores, Villa María del Triunfo, Carabayllo, Santiago de Surco, Villa el Salvador y Chorrillos), y muestra una actividad delictiva menor en comparación con el grupo anterior. El clúster 4, mostrado en color anaranjado, consta de 32 distritos y también presenta una actividad delictiva menor que el clúster 2. Finalmente, el clúster 1, identificado en amarillo, abarca 120 distritos y exhibe una actividad delictiva aún menor que los grupos anteriores.

En la figura 58 el mapa temático distingue en color rojo los 167 distritos de alta frecuencia delictiva, es decir, como se mencionó en el párrafo anterior esta categoría delictiva lo conforman los distritos que integran los clústeres 2, 5, 4, 2 y 1. También se aprecia en el mapa en color gris los 1,587 distritos que se caracterizan por tener una baja frecuencia de actividad delictiva

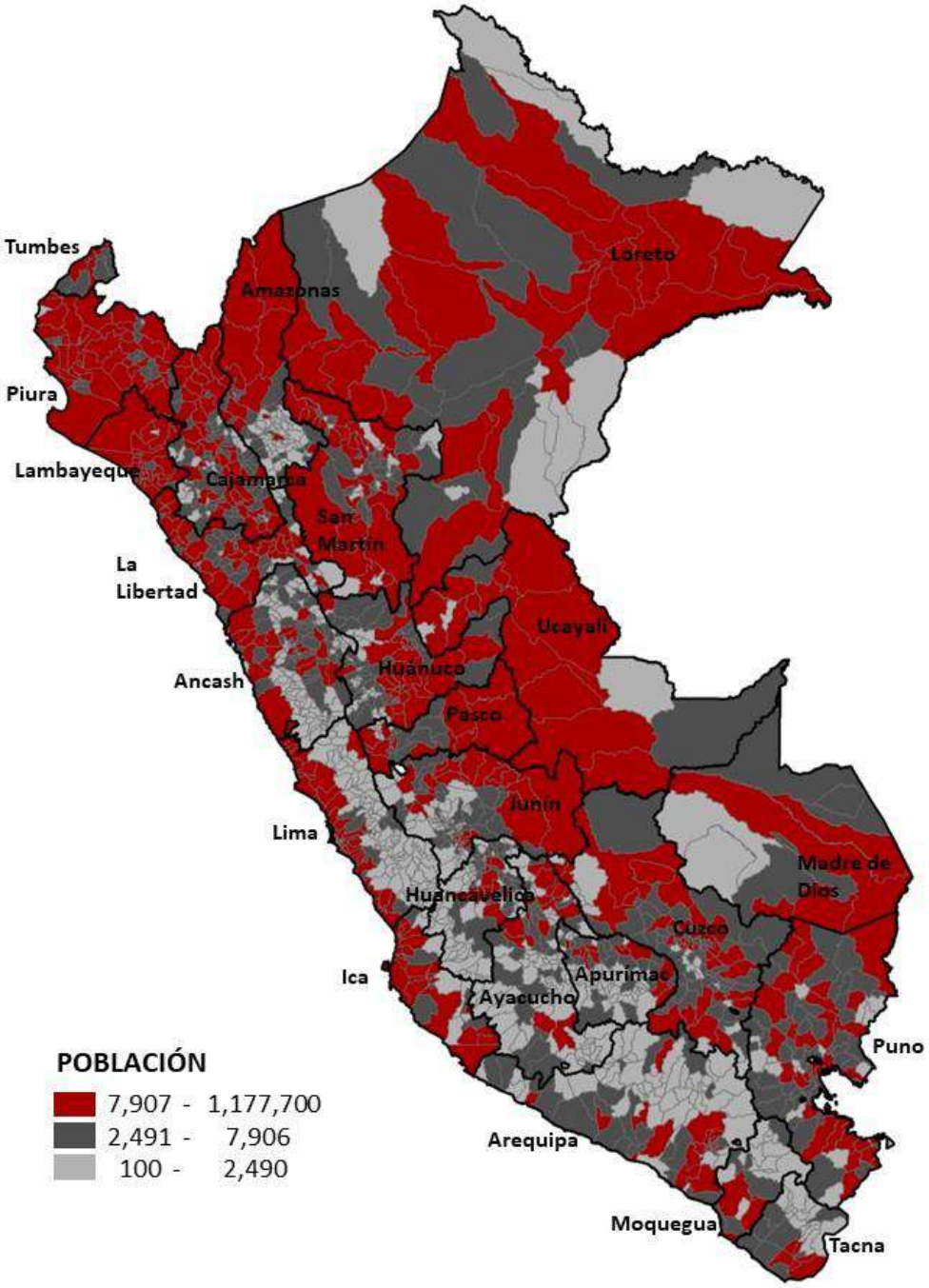


Figura 51. *Población INEI 2020*
Nota: Elaboración propia.

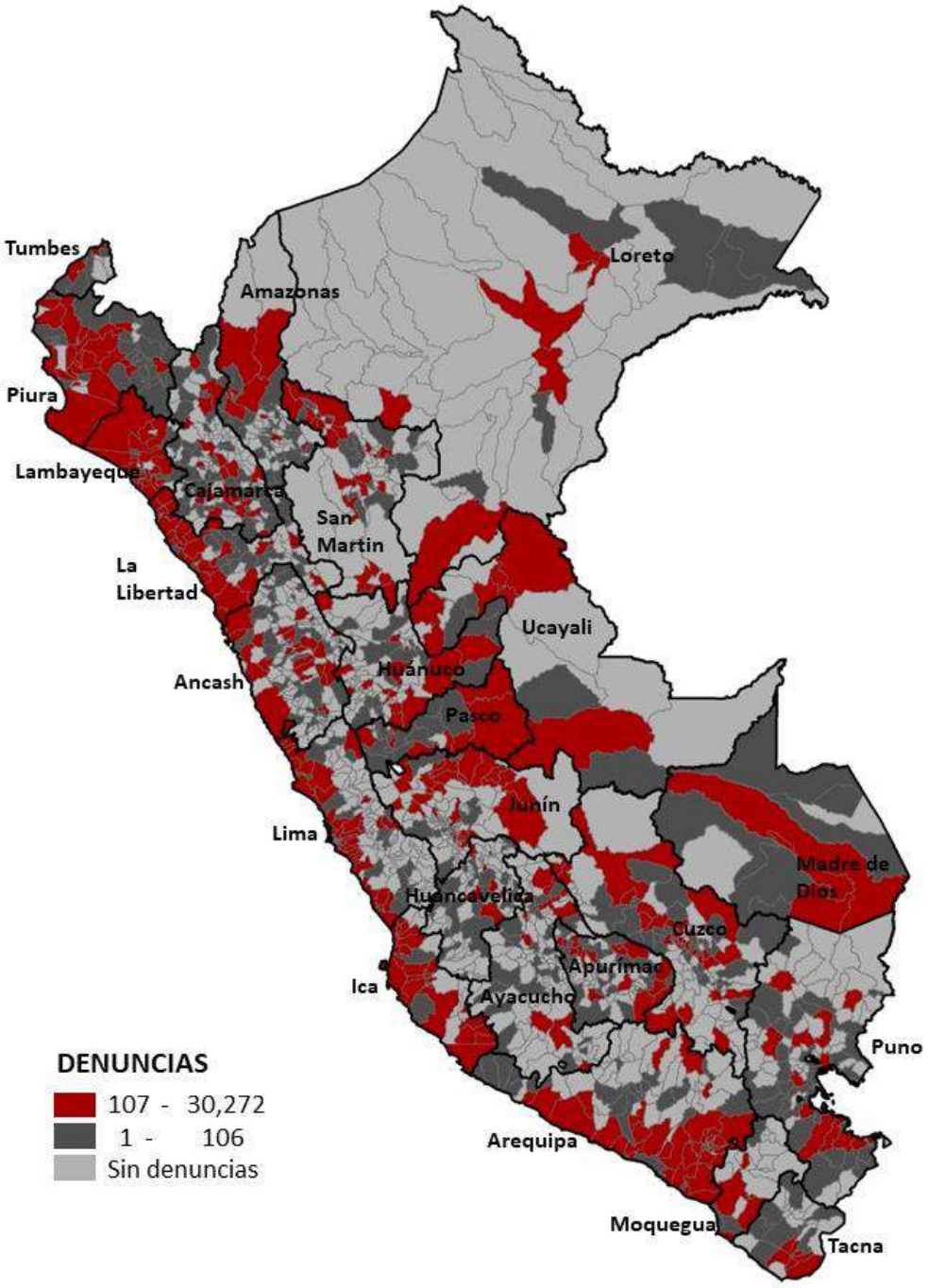


Figura 52. *Denuncias de delitos y faltas registradas por la PNP. 2020.*
Nota: Elaboración propia.

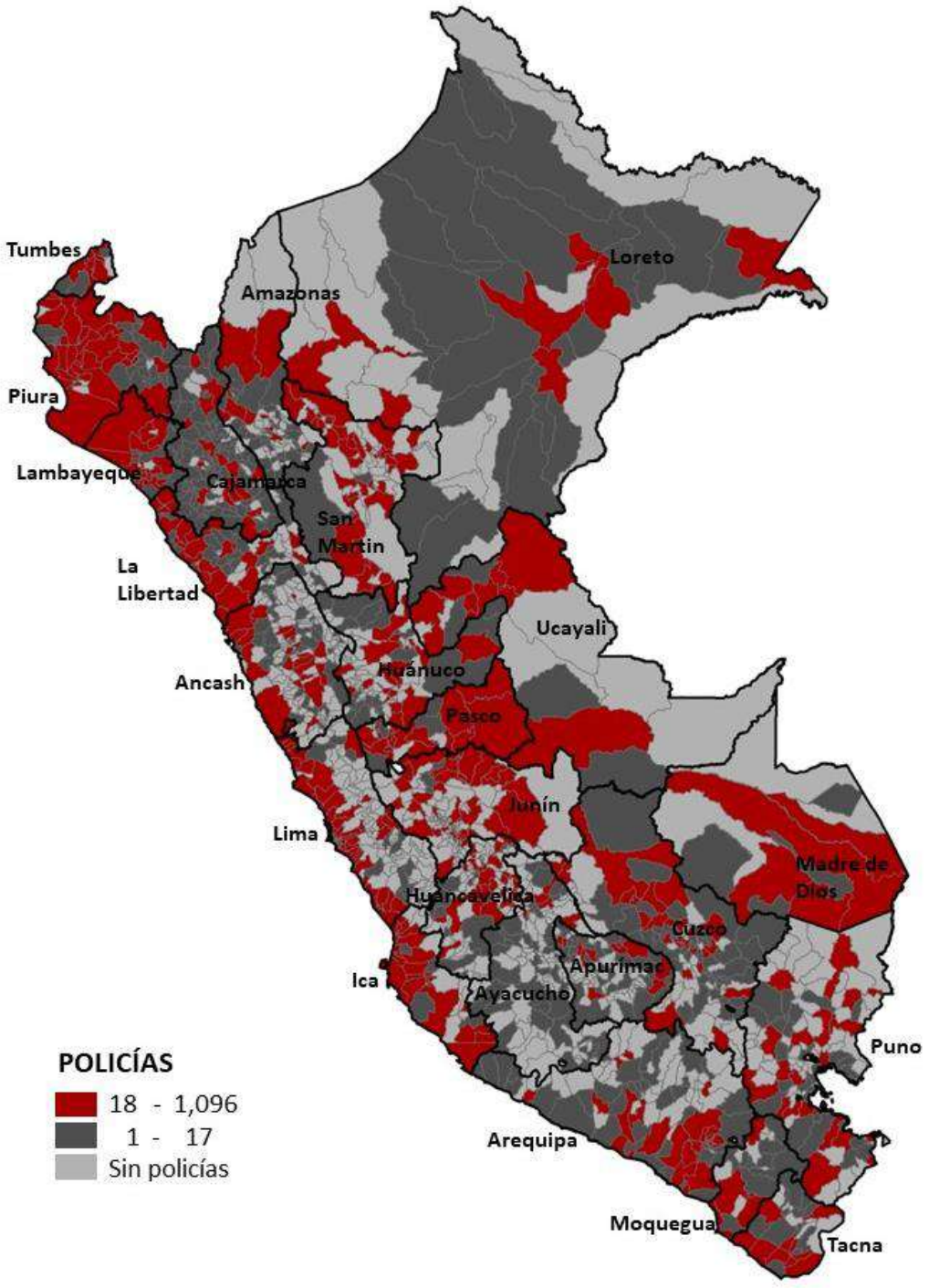


Figura 53. *Número de efectivos de la PNP. 2020.*
Nota: Elaboración propia.

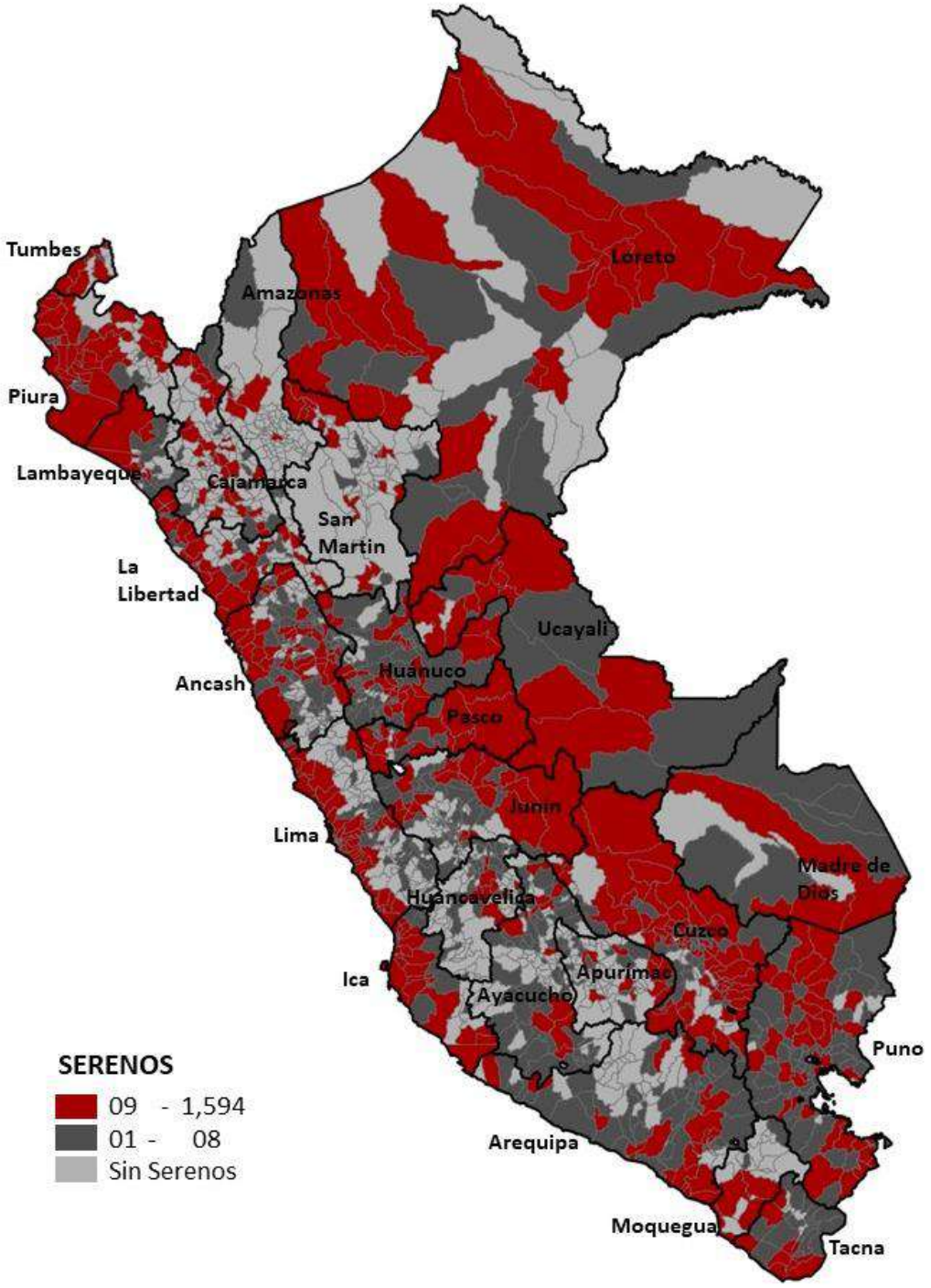


Figura 54. *Número de efectivos de serenazgo. 2020*
Nota: Elaboración propia.

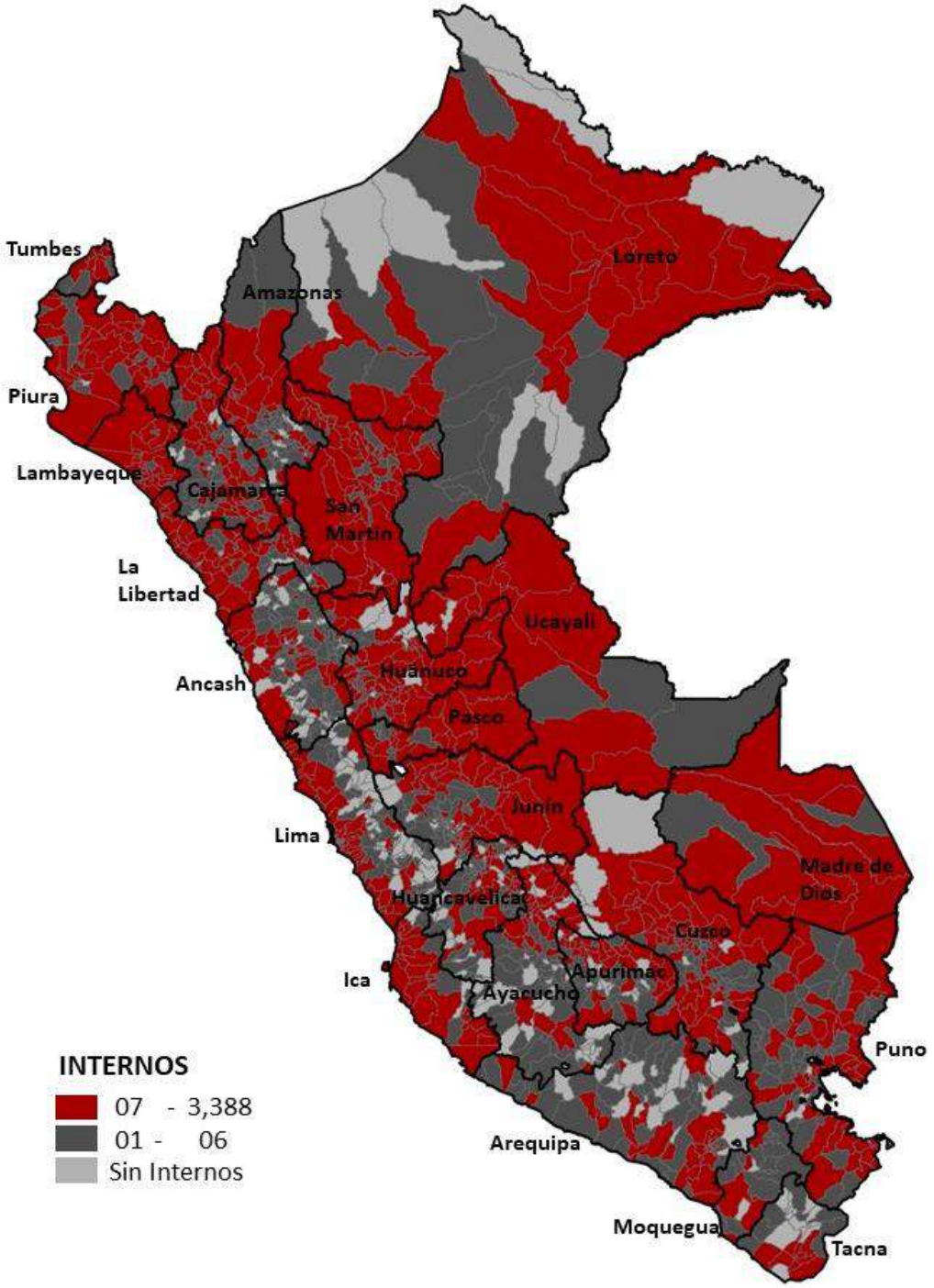


Figura 55. *Último distrito de residencia del interno. 2020.*
Nota: Elaboración propia.

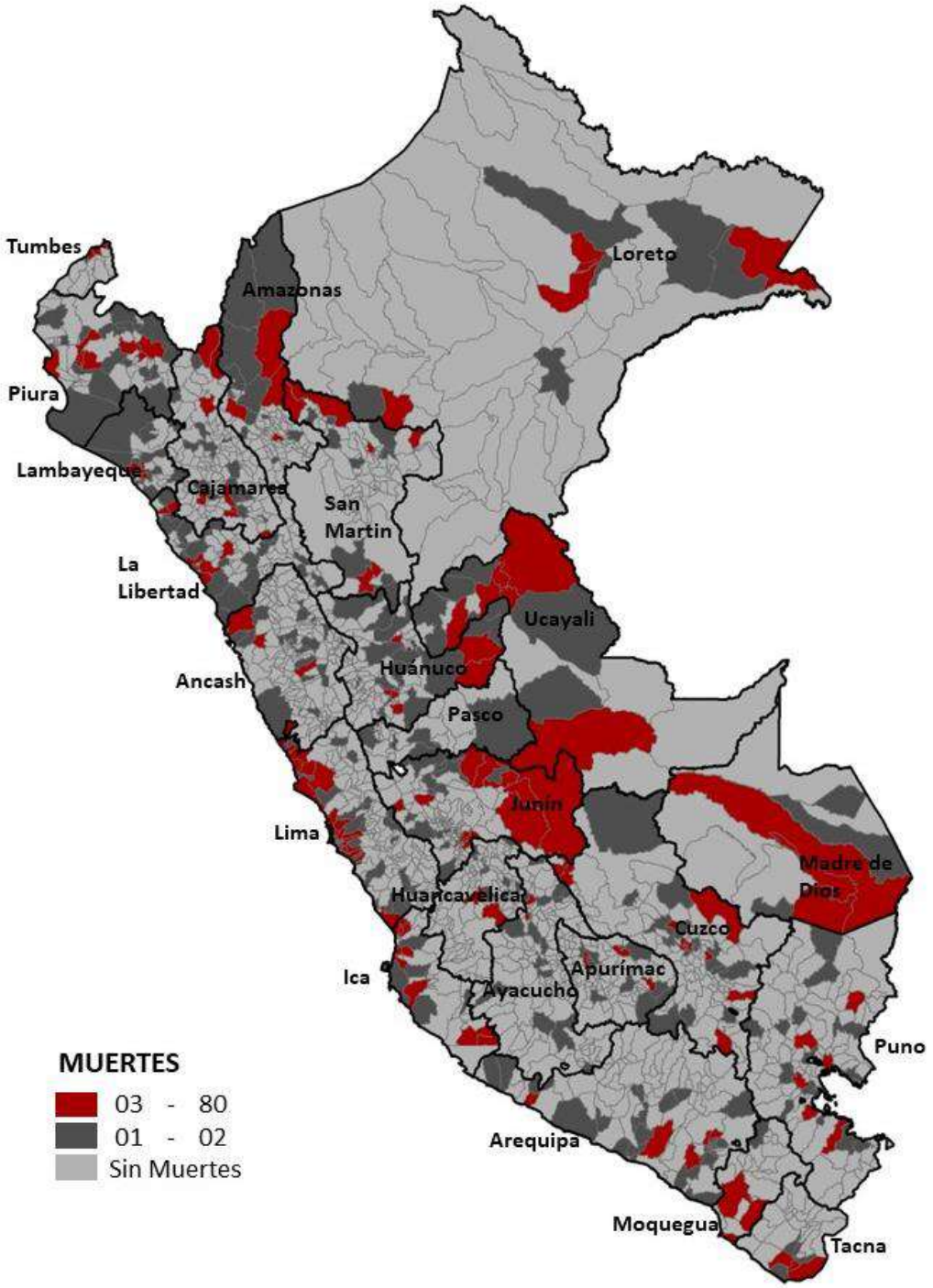


Figura 56. *Muertes violentas asociadas a hechos delictivos dolosos. 2020.*
Nota: Elaboración propia.

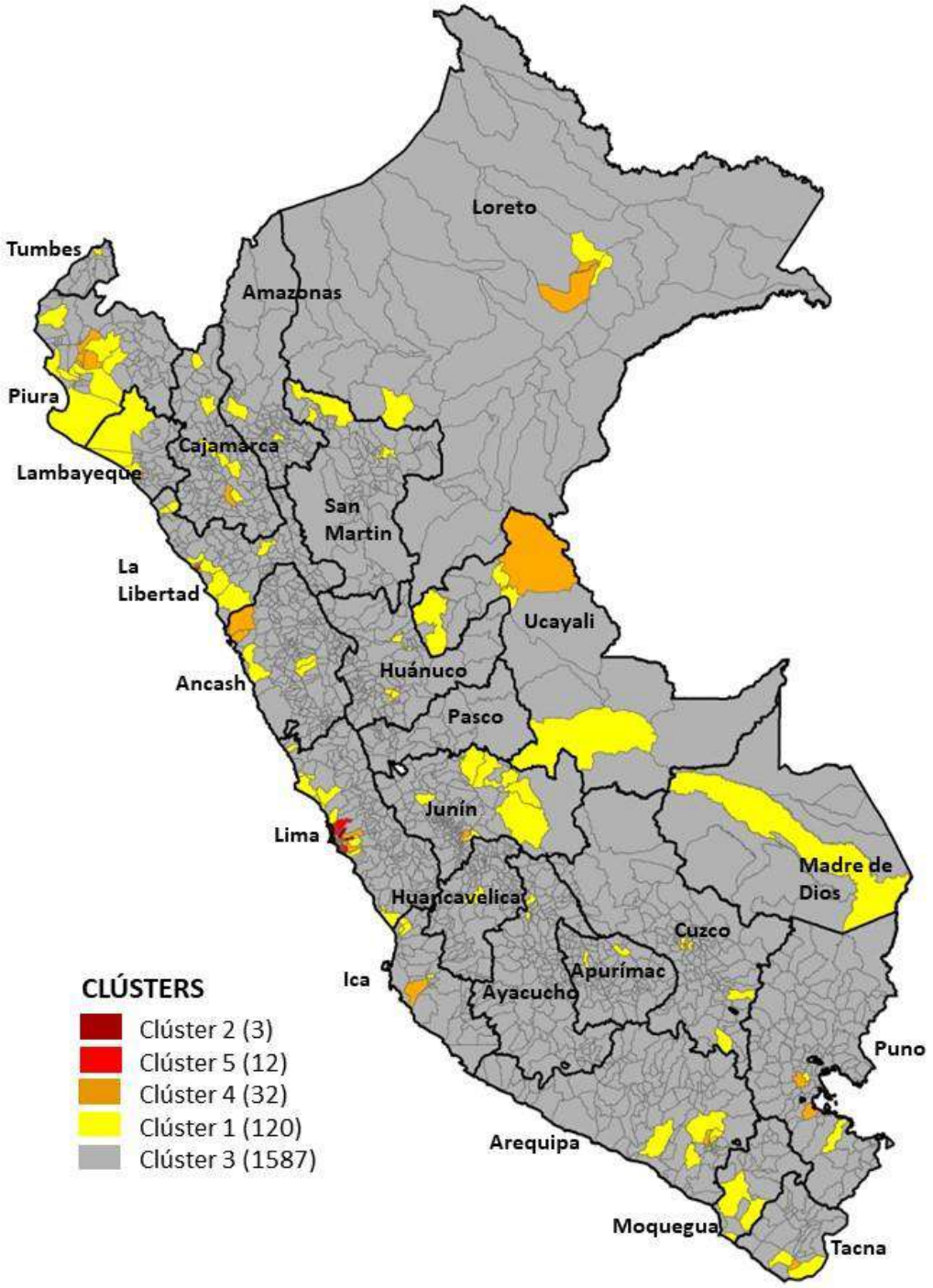


Figura 57. *Clúster de Distritos elaborado mediante el Análisis del Escalamiento Multidimensional.*
Nota: Elaboración propia.

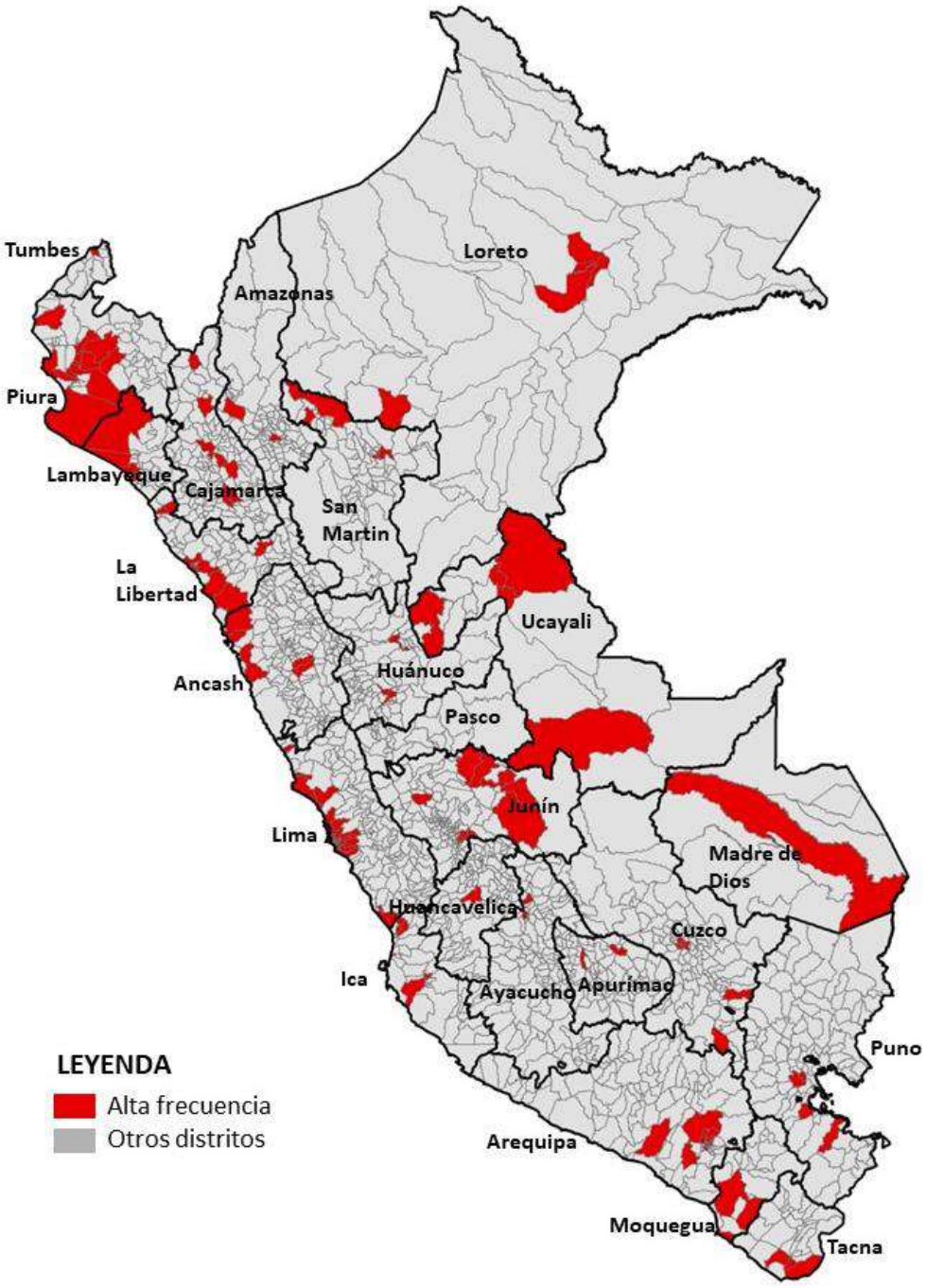


Figura 58. *Distritos de alta frecuencia delictiva elaborado mediante el análisis clúster a partir del análisis de escalamiento multidimensional.*
Nota: Elaboración propia.

1.8. Ranking de distritos del Perú según su actividad delictiva

Para la elaboración del ranking se utilizó las variables siguientes: Número de efectivos de la PNP (policías), número de efectivos de serenazgo (serenos), muertes violentas asociadas a hechos delictivos dolosos (muertes), denuncias de delitos y faltas registradas por la PNP (denuncias), último distrito de residencia del interno (internos) y población. El ranking fue elaborado mediante el cálculo de un índice a partir del resultado del ACP el cual se presenta en el Anexo 02. Es importante precisar que para el cálculo del índice se utilizó la formula siguiente:

$$\overline{\perp}_i = \frac{[\lambda_1 x_i + \lambda_2 \gamma_i]}{\lambda_1 + \lambda_2}$$

$$i = 1, 2, 3 \ldots 1754$$

Donde: λ_1: Autovalor asociado al primer factor

λ_2: Autovalor asociado al segundo factor

x_i: Valor de la coordenada x en el distrito i

y_i: Valor de la coordenada y en el distrito i

1.9. Resultados de los modelos de predicción

A. Modelo de Máquina de Vectores Soporte (MVS)

Los modelos de MVS es un tipo de algoritmo de machine learning supervisado aplicable a problemas de regresión y clasificación, aunque se usa comúnmente como modelo de clasificación. Para proceder con el modelo se consideró la variable "Inseguro" como variable de respuesta y el resto de variables como predictores. En el marco del algoritmo se ha descompuesto la matriz X en 2 columnas llamadas x_1 y x_2, el algoritmo consideró 94 vectores de apoyo y utilizando la validación cruzada con la muestra dividida

en dos partes, el modelo estimó una probabilidad de clasificación correcta del 93%.

La importancia de la característica se refiere a las técnicas que calculan una puntuación para todas las características de entrada para un modelo dado; las puntuaciones simplemente representan la "importancia" de cada característica. Una puntuación más alta significa que la característica específica tendrá un mayor efecto en el modelo que se utiliza para predecir una determinada variable. En la tabla 02 se muestra el ranking de las variables del modelo según su efecto en la contribución al modelo.

ROW_NAMES	RANK
CLÚSTER EMD	1
CLÚSTER ACP	2
MUERTES	3
INTERNOS	4
DELITOS	5
SERENOS	6
POLICIAS	7

Tabla 2. *Resultados de MVS, efecto de la variable en el modelo*
Fuente: Elaboración propia. Reporte generado del R.

La curva ROC también es conocida como la representación de sensibilidad frente a (1-especificidad). Cada resultado de predicción representa un punto en el espacio ROC. El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). Para el caso de la Curva ROC

de la predicción del modelo de MVS el área bajo la curva es de 0.98 como se puede apreciar en la figura 59.

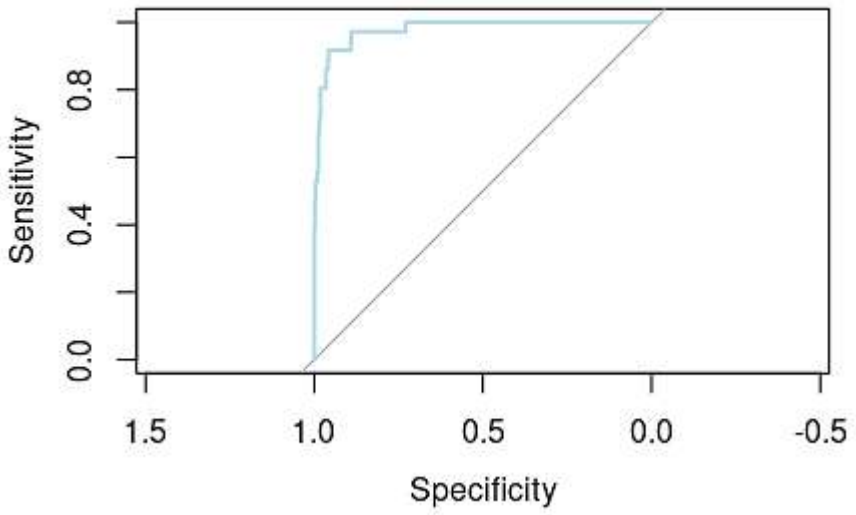


Figura 59. *Curva ROC de la predicción del modelo de MVS.*
Nota: Elaboración propia. Reporte generado del R.

B. Modelo de Arboles de Decisión

Los árboles de decisión son un método usado en distintas disciplinas como modelo de predicción. Estos son similares a diagramas de flujo, en los que llegamos a puntos en los que se toman decisiones de acuerdo con una regla. De manera general, lo que hace este algoritmo es encontrar la variable independiente que mejor separa nuestros datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla. A cada regla corresponde un nodo.

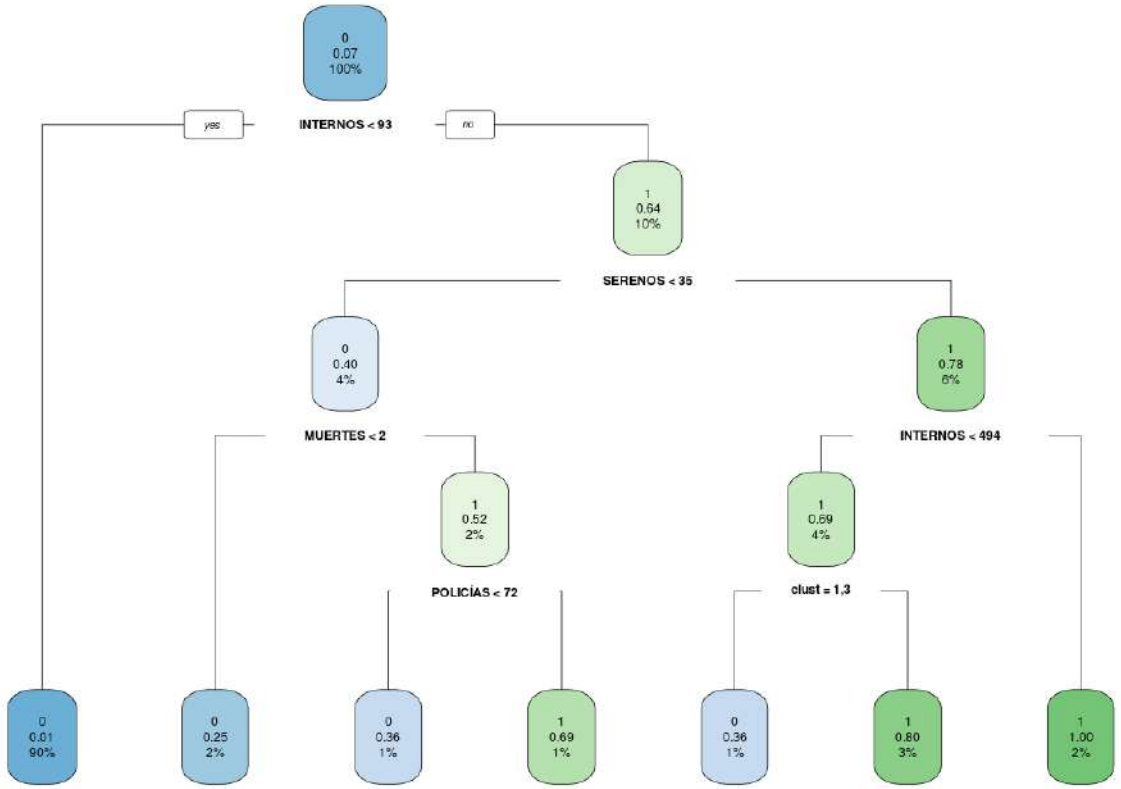


Figura 60. *Resultado del árbol de decisión para la predicción del modelo.*
Nota: Elaboración propia. Reporte generado del R.

Cada nodo está coloreado de acuerdo con la categoría mayoritaria entre los datos que agrupa. Esta es la categoría que ha predicho el modelo para ese grupo.

Dentro del rectángulo de cada nodo se muestra qué proporción de casos pertenecen a cada categoría y la proporción del total de datos que han sido agrupados allí. Por ejemplo, según el algoritmo de árboles de decisión, la variable que mejor discrimina es "internos", identificando que el mejor punto de corte para dicha variable es 93. La probabilidad media de que un distrito registre internos es de 0.07. Aquellos distritos con menos de 93 internos representan el 90% del total, con una probabilidad que disminuye de 0.07 a 0.01. Por otro lado, los distritos con más de 93 internos

constituyen el 10% restante, con una probabilidad que aumenta de 0.07 a 0.64.

Respecto a la contribución de las características específicas en el modelo de clasificación se puede observar que el mayor efecto en el modelo lo realiza la variable muertes, seguido de las variables denuncias, policías, tal como se describe en la tabla 03

ROW_NAMES	RANK
MUERTES	1
DENUNCIAS	2
POLICIAS	3
SERENOS	4
INTERNOS	5
CLÚSTER EMD	6
CLÚSTER ACP	7

Tabla 3. Efecto de la variable en el modelo de Árboles de Decisión. *Fuente*: Elaboración propia. Reporte generado del R.

De acuerdo con los resultados de la predicción, en la figura 61, se observa la curva de predicción ROC del modelo arboles de decisión el cual indica un área bajo la curva de 0.9395, es decir, el modelo estima una probabilidad de clasificación correcta de aproximadamente el 94%.

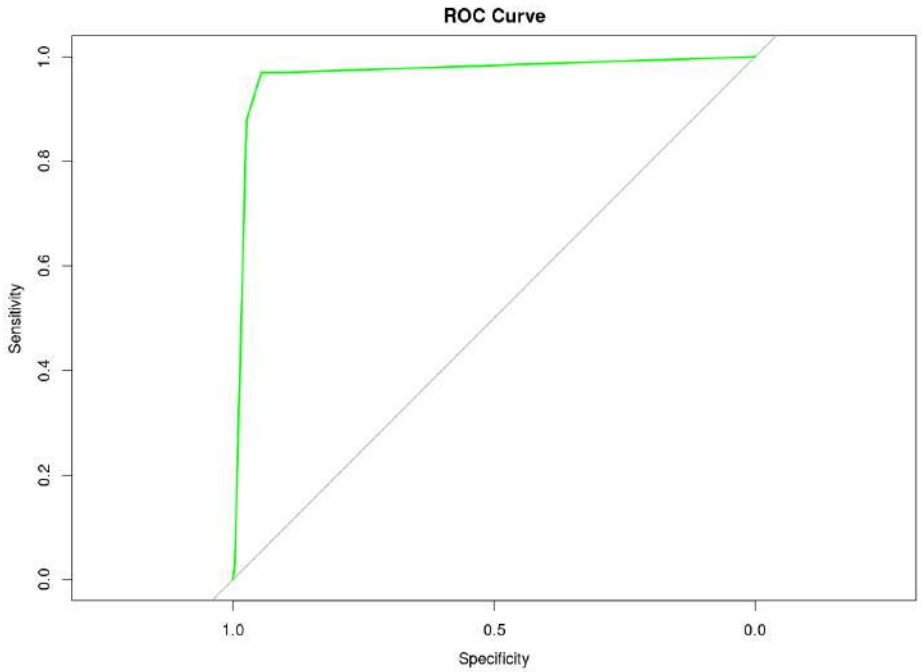


Figura 61. *Curva ROC de la predicción del modelo de Árboles de Decisión.*
Nota: Elaboración propia. Reporte generado del R.

C. Modelo de Naive Bayes

La clasificación de Naive Bayes es un algoritmo sencillo pero eficaz; es más rápido en comparación con muchos otros algoritmos iterativos; no necesita escalado de características; y su fundamento es el Teorema de Bayes.

Sin embargo, Naive Bayes se basa en el supuesto de que la probabilidad condicional de cada característica dada la clase es independiente de todas las demás características. La suposición de probabilidades condicionales independientes significa que las características son completamente independientes entre sí, sin embargo, las características categóricas no lo son. Suponiendo la independencia de todas las características, vamos a ajustar un modelo bayesiano ingenuo a nuestros datos de entrenamiento.

Las posibilidades de que el distrito sea INSEGURO en los clústeres conformados por EMD, es alta para el clúster 3 (0.476190476) y en menor medida con el clúster 5 (0.214285714) y 4 (0.202380952), como se aprecia en la taba 4. En la tabla 5, para los clústeres conformados por el ACP se observa que la posibilidad de clasificación como inseguro es alta para el clúster 2 (0.698476190)

	1	2	3	4	5
0	0.000000000	0.001748252	0.040209790	0.005244755	0.952797203
1	0.011904762	0.095238095	0.476190476	0.202380952	0.214285714

Tabla 4. Posibilidades de éxito en la clasificación del modelo Naive Bayes para los clústeres conformados por EMD.
Fuente: Elaboración propia. Reporte generado del R.

	1	2
0	0.004370629	0.995629371
1	0.309523810	0.698476190

Tabla 5. Posibilidades de éxito en la clasificación del modelo Naive Bayes para los clústeres conformados por ACP.
Fuente: Elaboración propia. Reporte generado del R.

De acuerdo con los resultados del modelo Naive Bayes indica un área bajo la curva ROC = 0.9775, es decir, el modelo estima una probabilidad de clasificación correcta de aproximadamente el 98% como se observa en la figura siguiente:

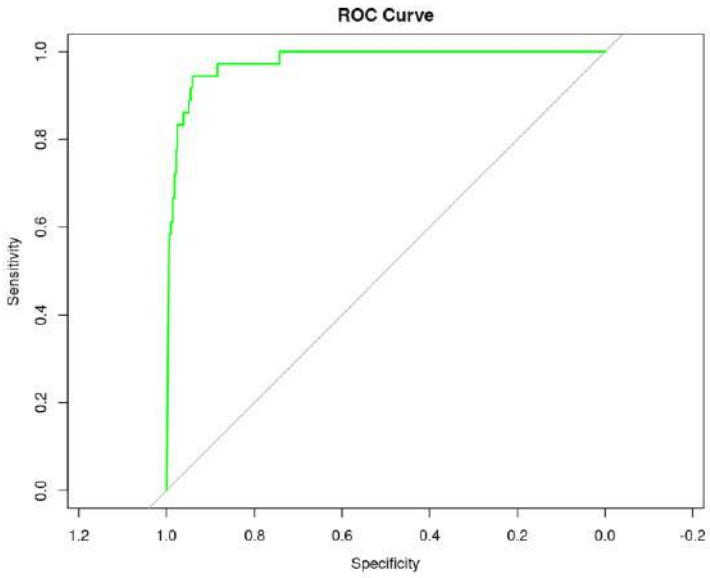


Figura 62. *Curva ROC de la predicción del modelo Naive Bayes.*
Nota: Elaboración propia. Reporte generado del R.

D. Modelo k Vecinos Más Cercanos (K-NN)

Es un algoritmo de clasificación no lineal supervisado. K-NN es un algoritmo no paramétrico, es decir, no hace ninguna suposición sobre los datos subyacentes o su distribución. Es uno de los algoritmos más simples que pueden ser utilizados para resolver problemas de clasificación y de regresión, la estimación que realiza depende de su valor k (vecinos). Los resultados mostrados en la tabla 04 se observa que para el valor 'K' de 11 obtenemos la máxima precisión con valores de accuracy y kappa de 0. 9539903 y 0.6239784

K	ACCURACY	KAPPA
1	0.9340349	0.4960431
3	0.9437977	0.5577185
5	0.9519544	0.6085446
7	0.9519561	0.6070465

9	0.9511480	0.6032878
11	0.9539903	0.6238632
13	0.9511381	0.5958916
15	0.9535788	0.6239784
17	0.9531673	0.6143007
19	0.9511397	0.5925789
21	0.9519561	0.5973852
23	0.9523592	0.6005910
25	0.9519577	0.5931566
27	0.9523626	0.5933762
29	0.9523592	0.5904994

Tabla 6. *Valores del accuray y kapa según K vecinos del modelo*
Fuente: Elaboración propia. Reporte generado del R.

En la figura 55, de acuerdo con los resultados del modelo k vecinos más cercanos indica un área bajo la curva = 0.957, es decir, el modelo estima una probabilidad de clasificación correcta de aproximadamente el 96%.

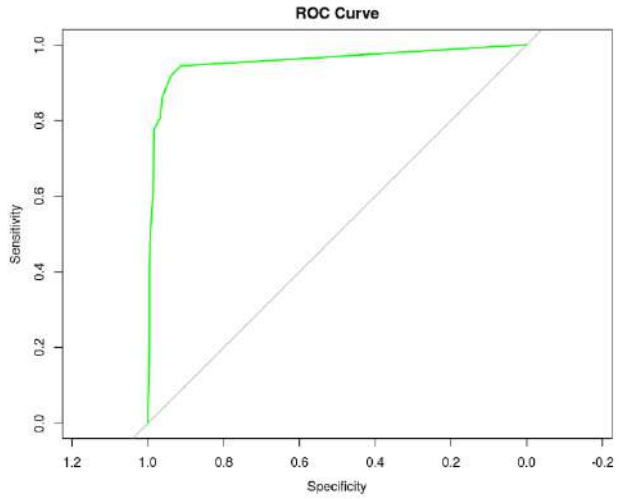


Figura 63. *Curva ROC de la predicción del modelo k Vecinos Más Cercanos*

Nota: Elaboración propia. Reporte generado del R.

Para elegir el mejor modelo de predicción se tuvo en consideración 5 criterios de los cuales el modelo de Arboles de Decisión presentó el mejor rendimiento debido a que evidenció los mejores resultados en 4 de los 5 criterios analizados, como se observa en la Tabla 07. Es importante precisar que para identificar el mejor modelo se considera el mayor valor de los criterios AUC, accuracy y sensibilidad; y el menor valor para los criterios gini y error.

MODELOS	AUC	GINI	ACCURACY	ERROR	SENSIBILIDAD
ÁRBOLES DE DECISIÓN	0.94	0.88	0.97	0.03	0.83
NAIVE BAYES	0.98	0.95	0.96	0.04	0.75
K VECINOS MÁS CERCANOS	0.96	0.92	0.96	0.04	0.61
MÁQUINA DE VECTORES DE SOPORTE	0.98	0.95	0.96	0.04	0.83

Tabla 7. *Comparación de los resultados de los modelos de predicción propuestos según AUC, GINI, accuracy, error y sensibilidad*
Fuente: Elaboración propia. Reporte generado del R.

Análisis y Discusión

El objetivo principal de la investigación fue: Predecir el estado delictivo de los distritos del Perú aplicando algoritmos de aprendizaje supervisados de clasificación, para lograr dicho objetivo se aplicó los modelos de clasificación supervisados que se detallan a continuación: Arboles de decisión, naibe bayes, K vecinos más cercanos, y máquinas de vectores de soporte. Para elegir el mejor modelo se evaluó los 5 criterios siguientes: AUC, gini, accuracy, error y sensibilidad; realizada la evaluación el modelo arboles de decisión presento el mejor rendimiento debido a que presentó mejores resultados en 4 de los 5 criterios como se puede apreciar en la Tabla 07. Es importante mencionar que Oviedo (2022) también encontró buenos

rendimientos al aplicar el modelo de árboles de decisión para la predicción de delitos que afectan la seguridad ciudadana en la ciudad de Guayaquil (Ecuador).

Respecto al primer objetivo específico: Identificar patrones delictivos de los distritos del Perú mediante algoritmos de minería de datos podemos precisar lo siguiente:

- Existe una alta correlación positiva entre los pares de variable, población-internos (0.90), población-muertes (0.81) y población-denuncias (0.80). Es decir, a mayor población en los distritos existe mayor incidencia delictiva. Concuerda con lo señalado por Mangara (2020), quien analiza datos de delitos en Kenia, en el cual indica que los condados altamente poblados reportan un mayor número de delitos en comparación con aquellos de baja población.
- Existe una baja correlación entre la variable serenos y el resto de las variables analizadas en la investigación. Al respecto, se puede precisar que el número de serenos en un distrito es la respuesta operativa de los gobiernos locales para enfrentar la inseguridad, y el número de serenos esta influenciado por la capacidad presupuestal con la que cuenta la municipalidad.
- El análisis clúster por EMD genera 5 clústeres de distritos organizados de acuerdo con el nivel de inseguridad, en la Figura 57 los podemos identificar de la siguiente manera:
    - ∴ **Clúster 3.** Puntos de color celeste, conformados por 1587 distritos, caracterizados por un nivel muy bajo de inseguridad.
    - ∴ **Clúster 1.** Puntos celestes oscuro, formado por 120 distritos, diferenciado con un nivel bajo de inseguridad.
    - ∴ **Clúster 4**. Puntos verdes, compuesto por 32 distritos, identificado con un nivel medio de inseguridad.
    - ∴ **Clúster 5.** Puntos rojos, conformados por 12 distritos, distinguido por un nivel alto de inseguridad.
    - ∴ **Clúster 2.** Puntos de color anaranjados, formado por 3 distritos, caracterizados por un nivel muy alto de inseguridad.
- Así mismo, la consolidación del clúster 2, clúster 5, clúster 4 y clúster 1

suman en total 167 distritos que representan los ámbitos distritales de mayor incidencia delictiva a nivel nacional. Estos 167 distritos constituyen tan solo el 9% del total de los 1875 distritos investigados. No obstante, este conjunto de distritos concentra la mayor actividad delictiva en el país. De hecho, los territorios de los 167 distritos identificados representan el 81.1% de las denuncias a nivel nacional, el 75.9% de la población carcelaria y el 67.4% de las muertes asociadas a hechos delictivos dolosos.

- La intensa relación positiva que existe entre las variables investigadas también se ven reflejadas en la estructura de los clústeres conformados por el análisis clúster por EMD, como se aprecia en la Figura 49 en la base de los diagramas de dispersión se encuentran los puntos de color lila (1587 distritos que conforman el clúster 3) caracterizados por un nivel muy bajo de inseguridad, sobre la base encontramos los puntos rojos (120 distritos del clúster 1) con un nivel bajo de inseguridad, a continuación los puntos anaranjados (32 distritos del clúster 4) que representan un nivel intermedio de inseguridad, en seguida los puntos verdes (12 distritos del clúster 5) que indican un nivel alto, y en la parte superior los puntos azules (3) que reflejan un nivel muy alto de inseguridad.

Respecto al segundo objetivo específico: Comparar los resultados de los modelos de predicción propuestos en base a indicadores de clasificación detallo lo siguiente:

- Para elegir el mejor modelo de predicción se consideró los 5 criterios siguientes: AUC, gini, accuracy, error y sensibilidad de los cuales el modelo de árboles de decisión presentó el mejor rendimiento debido a que evidenció los mejores resultados en 4 de los 5 criterios analizados, como se observa en la Tabla 07.
- Respecto al criterio **AUC**, considerado como el área bajo la curva ROC cuyo valor va de 0 a 1, el rendimiento logrado por el modelo de árboles de decisión es de 0.94, muestra un alto rendimiento a pesar de que el resto de los modelos registran mejores rendimientos. El índice **gini** cuantifica la probabilidad de

que un elemento elegido al azar sea clasificado incorrectamente, en este criterio el modelo de árboles de decisión obtiene el mejor valor (0.88) frente al resto de modelos. En relación con el **accuracy**, mide la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones, en este criterio el modelo de árboles de decisión obtiene el más alto valor (0.97) en comparación a los demás modelos. El **error**, medida que cuantifica cuan incorrecta son las predicciones del modelo, respecto a esta medida el modelo de árboles de decisión presenta el menor valor (0.03). La **sensibilidad** mide la capacidad del modelo para identificar correctamente todos los casos positivos en comparación con el total de casos positivos reales en el conjunto de datos criterio el modelo de árboles de decisión obtiene el más alto valor (0.83) en comparación del resto de modelos

En alusión al tercer objetivo específico: Visualizar la distribución espacial de la actividad delictiva en los distritos del Perú, al respecto preciso lo siguiente**:**

- Los mapas temáticos elaborados permiten identificar la distribución espacial de la actividad delictiva en los distritos del Perú.
- El sistema de información geográfica facilita el análisis simultáneo de la estructura cuantitativa de las variables investigadas en relación con el espacio geográfico de estudio.
- La fase de preparación y elaboración de la base de datos constituye el aspecto principal y más laborioso al aplicar un Sistema de Información Geográfica. Es crucial tener especial cuidado al unir la base de datos de las variables investigadas con la estructura cartográfica del mapa base para evitar problemas en el procesamiento de la información.
- En la Figura 58, se observa que los distritos identificados con un **nivel muy alto de inseguridad (03 distritos del clúster 2)** están ubicados principalmente en la provincia de Lima. Los distritos con un **nivel alto de inseguridad (12 distritos del clúster 5)** se encuentran en Lima y Callao, a excepción de Trujillo, que también forma parte de este grupo. Posteriormente, los distritos con un **nivel medio de inseguridad (32 distritos del clúster 4)**

están situados en la costa norte del Perú (Piura, Lambayeque, La Libertad y Ancash), así como en Lima, la zona sur (Arequipa, Ica, Tacna), la sierra (Cajamarca, Cusco, Junín, Puno) y la selva (Loreto y Ucayali). A continuación, se encuentran los **niveles bajo (120 distritos del clúster 1) y muy bajo de inseguridad (1587 distritos del clúster 3)**, distribuidos a lo largo de todo el país.

En referencia al cuarto objetivo específico: Elaborar un ranking de los distritos del Perú según su actividad delictiva, al respecto especifico lo siguiente:

- Para la elaboración del ranking usamos los resultados del ACP mediante el cual se proyecta las observaciones en 2 dimensiones las cuales explican el 91% de la información contenida en la base de datos, los resultados se muestran en el Anexo 02, es importante precisar que para el cálculo del índice se utilizó los autovalores asociados a cada dimensión con sus correspondientes coordenadas proyectadas en el espacio de las componentes principales de cada distrito investigado.
- Adicionalmente, se evidencia que el orden correlativo de los distritos en función del nivel de inseguridad elaborado mediante el ACP guarda una estrecha relación con los niveles de inseguridad de los clústeres de distritos conformados mediante el EMD.
- El índice elaborado está compuesto por las siguientes dimensiones:
    - ∴ Dimensión operativa:
        - Número de efectivos de la PNP (policías)
        - Número de efectivos de serenazgo (serenos)
    - ∴ Dimensión Delictiva
        - Denuncias de delitos y faltas registradas por la PNP (denuncias)
        - Muertes violentas asociadas a hechos delictivos dolosos (muertes)
    - ∴ Dimensión social
        - Población
        - Último distrito de residencia del interno (internos)

Conclusiones

Se obtuvo el mejor modelo de aprendizaje supervisado para predecir el estado delictivo de los distritos del Perú, utilizando árboles de decisión. Este modelo fue evaluado considerando cinco criterios: AUC (0.94), gini (0.88), accuracy (0.97), error (0.03) y sensibilidad (0.83). Dicho modelo presenta el mejor rendimiento, ya que evidencia los mejores resultados en 4 de los 5 criterios analizados, como se observa en la Tabla 07.

La investigación identifica 05 clústeres de distritos organizados de acuerdo con el nivel de inseguridad, en la Figura 45 y 57 podemos distinguir los clústeres siguientes: **Clúster 3**, conformados por 1587 distritos caracterizados por un nivel muy bajo de inseguridad; **clúster 1**, formado por 120 distritos diferenciados con un nivel bajo de inseguridad; **clúster 4**, compuesto por 32 distritos identificado con un nivel medio de inseguridad; **clúster 5**, conformados por 12 distritos distinguido por un nivel alto de inseguridad; y el **clúster 2**, formado por 3 distritos caracterizados por un nivel muy alto de inseguridad.

Los resultados destacan 167 distritos (Figura 58) que representan los ámbitos territoriales de mayor incidencia delictiva a nivel nacional (consolidación de los clúster 2, clúster 5, clúster 4 y clúster 1). Los 167 distritos constituyen tan solo el 9% del total de los 1875 distritos en Perú. No obstante, este conjunto de distritos concentra la mayor actividad delictiva en el país. De hecho, los territorios de los 167 distritos identificados representan el 81.1% de las denuncias a nivel nacional, el 75.9% de la población carcelaria y el 67.4% de las muertes asociadas a hechos delictivos dolosos.

Existe una alta correlación positiva entre los pares de variable, población-internos (0.90), población-muertes (0.81) y población-denuncias (0.80), situación que evidencia que en los distritos de mayor población se registra mayor incidencia delictiva (Figura 43).

Los mapas temáticos creados mediante un sistema de información geográfica permiten identificar la distribución espacial de la actividad delictiva en los distritos del Perú (Figuras 51 al 58).

Respecto al ranking, se evidencia que el orden correlativo de los distritos según el nivel de inseguridad elaborado mediante el ACP guarda estrecha relación con los niveles de inseguridad de los clústeres de distritos elaborado mediante el análisis de EMD (Anexo 02).

El índice desarrollado para construir el ranking de distritos del Perú es una herramienta metodológica coherente que facilita la identificación de distritos prioritarios para la implementación de políticas públicas en materia de seguridad ciudadana (Anexo 02).

Recomendaciones

Considerando el enfoque metodológico, científico, social y teórico de la presente investigación se formula la siguiente propuesta:

- El índice creado para la elaboración del ranking de distritos en el Perú representa una herramienta metodológica coherente. En este contexto, se recomienda su empleo para identificar distritos prioritarios en la implementación de políticas públicas en el ámbito de la seguridad ciudadana. Dicho índice facilita la clasificación de los distritos según niveles de inseguridad, lo cual simplifica la planificación de programas, estrategias o actividades en el sector interior, tanto a nivel nacional como subnacional. Además, el empleo del índice permite una asignación más eficiente de recursos presupuestales para abordar el problema de la inseguridad en el país.
- Considerando que en la presente investigación se ha evidenciado que la aplicación de técnicas SIG facilita el análisis simultáneo de la estructura cuantitativa de las variables investigadas en relación con el espacio geográfico de estudio, se recomienda emplear este tipo de técnicas cuando se analice información de seguridad ciudadana. Esta técnica proporciona el componente territorial en la variable investigada, aspecto fundamental para el desarrollo de planes, estrategias y acciones en materia de seguridad ciudadana.
- En relación con la implementación de un SIG para el análisis de datos, se recomienda prestar especial atención a la fase de preparación y creación de la

base de datos. Esta etapa representa el aspecto más crucial y laborioso, especialmente al integrar la base de datos con la estructura cartográfica del mapa base.

Referencias bibliográficas

Aravindan, S., Anusuya, E., & Ashok, M. (2020). Gui based prediction of crime rate using machine learning approach. *Int J Comput Sci Mob Comput, 9(3)*, 221 - 229. Obtenido de https://paper.researchbib.com/view/paper/241337#google_vignette

Banco Interamericano de Desarrollo. (2018). *Seguridad ciudadana en América Latina y el Caribe.* Obtenido de https://publications.iadb.org/es/seguridad-ciudadana-en-america-latina-y-el-caribe-desafios-e-innovacion-en-gestion-y-politicas

Banco Interamericano de Desarrollo. (2021). *Evaluación de Apoyo del BID en el Área de Seguridad Ciudadana y Justicia en la Región.* Obtenido de https://publications.iadb.org/es/documento-de-enfoque-evaluacion-del-apoyo-del-bid-en-el-area-de-seguridad-ciudadana-y-justicia-en

Barrientos, M., Cruz, R., Acosta, M., Rabatte, S., Gogeascoechea, T., Pavon, L., & Blazquez, M. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica de la Universidad Veracruzana, 9 (2)*, 19-24. Obtenido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf

Bassett, R., & Deride, J. (2018). *Maximum a posteriori estimators as a limit of Bayes estimators.* Obtenido de https://www.researchgate.net/publication/310594908_Maximum_a_Posteriori_Estimators_as_a_Limit_of_Bayes_Estimators

Beck, U. (2002). *La sociedad del riesgo. Hacia una nueva modernidad.* México: Paidos.

Beckmann, M., Ebecken, N., & Pires, B. (2015). A KNN Undersampling Approach for Data Balancing. *Journal of Intelligent Learning Systems and Applications, 7*, 104-116. Obtenido de https://www.scirp.org/journal/paperinformation?paperid=60996

Boser, B., Guyon, I., & Vapnik, V. (1992). *A Training Algorithm for Optimal Margin Classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92).* Pittsburgh.

Carrión, J., Zárate, P., Boidi, F., & Zechmeister, E. (2020). Cultura política de la democracia en Perú y en las Américas. *Tomándole el pulso a la democracia*, 104-107. Obtenido de https://www.vanderbilt.edu/lapop/peru.php

Chaure, P. (2021). *Aplicaciones y oportunidades de la inteligencia artificial para la justicia penal. Predicción del riesgo de reincidencia de reos y policía predictiva.* España: Universidad Pontificia Comillas. Obtenido de https://repositorio.comillas.edu/xmlui/handle/11531/49501

Chikodili, H., Ogbobe, P., & Okoronkwo, M. (2021). *Analysis of Crime Pattern using Data Mining Techniques.* International Journal of Advanced Computer Science and Applications. 12 (12). Obtenido de https://www.researchgate.net/publication/357454221_Analysis_of_Crime_Pattern_using_Data_Mining_Techniques

Colina, A., & Et al. (2022). Aplicación de minería de datos en datos abiertos de Ecuador: Delitos. UCV HACER. *Revista de Inv. Y Cult., 11*(1). Obtenido de https://doi.org/10.18050/RevUCVHACER.v11n1a8

Comité Estadístico Interinstitucional de la Criminalidad. (2020). *Homicidios en el Perú contándolos uno a uno 2011 - 2018.* Obtenido de https://www.inei.gob.pe/biblioteca-virtual/boletines/

Dammert, L. (2017). *Violencias y criminalidad en las principales ciudades andinas: Caracterización y políticas públicas.* Wilson Center. Obtenido de https://www.wilsoncenter.org

Deepika, K., & Smitha, V. (2018). Crime analysis in India using data mining techniques. *International Journal of Engineering & Technology. 7 (2.6)*, 253-258. Obtenido de https://www.sciencepubco.com/index.php/ijet/article/view/10779

Del Bosque, T., & Et al. (2012). *Los sistemas de información geográfica y la investigación en ciencias humanas y sociales.* Madrid.

Fineberg, H. (1980). Decision trees: contruction, uses and limits. *Bull Cancer (67)*, 395-404. Obtenido de https://pubmed.ncbi.nlm.nih.gov/7225598/

Fix, E., & Hodges, J. (1951). *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties.* Obtenido de https://www.jstor.org/stable/1403797

Galindo, L., & Et al. (2007). Las actividades delictivas en el Distrito Federal. *Revista Mexicana de Sociología 69, núm. 3 (julio-septiembre, 2007).*

Giraldo, S., Ordoñez, L., Guerrero, V., & Ordoñez, H. (2020). Modelo de redes neuronales para predecir la tendencia de víctimas de secuestro en Colombia.

Investigación e innovación en ingenierías, 8(3), 38-49. Obtenido de https://revistas.unisimon.edu.co/index.php/innovacioning/article/view/4702

Instituto Nacional de Estadística e Informática. (2022). *Informe técnico estadísticas de seguridad ciudadana junio 2021 - noviembre 2021.* Obtenido de https://www.inei.gob.pe/biblioteca-virtual/boletines/

Licona, A. (2018). *Caracterización de los delitos en Cartagena mediante la aplicación de minería de datos. [Tesis para obtener el grado de Ingeniería Industrial]*. Colombia: Universidad Tecnológica de Bolívar. Obtenido de https://primo.utb.edu.co/discovery/fulldisplay?vid=57UTB_INST:57UTB_INST&tab=Everything&docid=alma990000504570205731&lang=es&context=L&adaptor=Local%20Search%20Engine&offset=0

Maggi, R. (2023). *Análisis de factibilidad para la implementación de un sistema informático de gestión como modelo de prevención de delitos basado en algoritmos de simulación predictivos en el comando rural de Policía de la ciudad de Milagro [Tesis de Maestría]*. Obtenido de https://repositorio.unemi.edu.ec/bitstream/123456789/6906/1/MAGGI%20ORTIZ%20RODRIGO%20BERNABE.pdf

Mangara, S., Njuguna, J., Kyalo, R., & Mutai, N. (2020). Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya. *International Journal of Data Science and Analysis, 6(1),* 20-31. Obtenido de https://www.sciencepublishinggroup.com/journal/paperinfo?journalid=367&doi=10.11648/j.ijdsa.20200601.13

Meneses, J. (2019). *Introducción al análisis multivariante.* Barcelona: Universidad abierta de Cataluña. Obtenido de https://femrecerca.cat/meneses/publication/introduccion-analisis-multivariante/

Mosquera, J. (2021). *rabajo de grado sobre predicción de los tipos de delitos en Medellín. [Trabajo para obtener título de Economista]*. Medellín. Obtenido de https://hdl.handle.net/10495/20644

Muggah, R. (2017). The Rise of Citizen Security in Latin America and the Caribbean in Alternative Pathways to Sustainable Development: Lessons from Latin America. *International Development Policy series. (9),* 291-322. Obtenido de https://journals.openedition.org/poldev/2512

Muñoz, V. (2021). *Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín. [Tesis de Magíster en Ingeniería Analítica]*. Colombia. Obtenido de https://repositorio.unal.edu.co/handle/unal/80976

Niño, C. (2020). *Manual de Ciencia Política y Relaciones Internacionales.* (S. y. Fabio, Ed.) Bogotá: Universidad Sergio Arboleda. Obtenido de https://www.researchgate.net/publication/343658851_Manual_de_Ciencia_Politica_y_Relaciones_Internacionales

Norouzi, N., & Ataei, E. (2021). Application of data mining in identifying and discovering hidden patterns of theft. *International Journal of Innovative Research in the Humanities. 1 (1)*, 29 - 42. Obtenido de https://www.researchgate.net/publication/351986275

Nuñez, J., & Tocornal, J. (2012). Determinantes individuales y del entorno residencial en la percepción de seguridad en barrios del Gran Santiago,Chile. *Revista INVI N° 74 / May 2012*. Obtenido de https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-83582012000100003

Olaya, V. (2014). *Sistemas de información geográfica. OSGE versión web.* Obtenido de icog.es/TyT/files/Libro_SIG.pdf

OMS. (2002). *Informe Mundial sobre la violencia y la salud.* Obtenido de iris.who.int/bitstream/handle/10665/67411/a77102_spa.pdf;jsessionid=340FFF106CC8BE8B7D4C061F52E91538?sequence=1

Ordoñez, H. (2020). *Modelo de machine learning para la predicción de las tendencias de hurto en Colombia.* RISTI Revista Ibérica de Sistemas e Tecnologías de Informação N.° 29. Mayo 2020. Obtenido de https://www.proquest.com/openview/fb8bfe36673b48be2d035ee8a035c307/1.pdf?pq-origsite=gscholar&cbl=1006393

Oviedo, B., & Et al. (2022). *Utilización de algoritmos de clasificación para la predicción de los delitos que afectan la seguridad ciudadana. Centro Sur Social Science Journal.* Obtenido de https://www.centrosureditorial.com/index.php/revista/article/view/273

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), 9*. Obtenido de https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z

USP
UNIVERSIDAD SAN PEDRO

REPOSITORIO INSTITUCIONAL DIGITAL

FORMULARIO DE AUTORIZACIÓN PARA LA PUBLICACIÓN DE DOCUMENTOS DE INVESTIGACIÓN

1. Información del Autor		
Alza Díaz José Alfredo	16630965	josealza.d@gmail.com
Apellidos y Nombres	DNI	Correo Electrónico

2. Tipo de Documento de Investigación

X	Tesis		Trabajo de Suficiencia Profesional		Trabajo Académico		Trabajo de Investigación

3. Grado Académico o Título Profesional [1]

	Bachiller		Título Profesional		Título Segunda Especialidad		Maestría	X	Doctorado

4. Título del Documento de Investigación

Modelo de predicción del estado delictivo de los distritos del Perú, 2020

5. Programa Académico

Doctorado en Estadística

6. Tipo de Acceso al Documento

X	Abierto o Público [2] (info:eu-repo/semantics/openAccess)		Acceso restringido [3] (info:eu-repo/semantics/restrictedAccess) (*)

(*) En caso de restringido sustentar motivo	

A. Originalidad del Archivo Digital

Por el presente dejo constancia que el archivo digital que entrego a la Universidad, es la versión final del trabajo de investigación sustentado y aprobado por el Jurado Evaluador y forma parte del proceso que conduce a obtener el grado académico o título profesional.

B. Otorgamiento de una licencia CREATIVE COMMONS [4]

El autor, por medio de este documento, autoriza a la Universidad, publicar su trabajo de investigación en formato digital en el Repositorio Institucional Digital, al cual se podrá acceder, preservar y difundir de forma libre y gratuita, de manera íntegra a todo el documento. [5]

Lugar	Día	Mes	Año
Chimbote	22	08	2024

Firma

Original Firmado

1 Según Resolución de Consejo Directivo N° 033-2016-SUNEDU-CD Reglamento del Registro Nacional de Trabajos de Investigación para optar Grados Académicos y Títulos Profesionales, Art. 8, inciso 8.2.

2 Ley N° 30035, Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y D.S. 006-2017-PCM

3 Si el autor eligió el tipo de acceso abierto o público, otorga a la Universidad San Pedro una licencia no exclusiva, para que la pueda hacer uso según se indica y difunda en el Repositorio Institucional Digital Respetando siempre los Derechos de Autor y Propiedad Intelectual de acuerdo con el Marco de la Ley 822.

4 En caso de que el autor elija la segunda opción, únicamente se publicará los datos del autor y resumen de tesis, de acuerdo a la directiva N° 006-2018-CONCYTEC-DIGC (Numerales 7.2 y 8.1) que norma el funcionamiento del Repositorio Nacional Digital

5 Las licencias Creative Commons (CC) son una organización internacional sin fines de lucro que pone a disposición de los autores un conjunto de licencias flexibles y de herramientas tecnológicas que facilitan la difusión de información, recursos educativos, obras artísticas y científicas, entre otros. Estas licencias también garantizan que el autor obtenga el crédito por su obra.

6 Según el inciso 12.2. del artículo 12° del Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales RENATI "Las universidades, instituciones y escuelas de educación superior tienen como obligación registrar todos los trabajos de investigación y proyectos, incluyendo los metadatos en sus repositorios institucionales previamente si son de acceso abierto o restringido, los cuales serán posteriormente recolectados por el Repositorio Digital RENATI, a través del Repositorio ALICIA".

Nota: En caso de falsedad en los datos, se procederá de acuerdo a lo establecido en el Art. 32, num. 32.3).

Modelo de predicción del estado delictivo de los distritos del Perú, 2020

INFORME DE ORIGINALIDAD

20%	19%	%	8%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	hdl.handle.net Fuente de Internet	2%
2	rstudio-pubs-static.s3.amazonaws.com Fuente de Internet	1%
3	repositorio.unemi.edu.ec Fuente de Internet	1%
4	Submitted to Universitat Politècnica de València Trabajo del estudiante	1%
5	ichi.pro Fuente de Internet	1%
6	www.researchgate.net Fuente de Internet	1%
7	bibliotecadigital.udea.edu.co Fuente de Internet	1%
8	www.congreso.gob.pe Fuente de Internet	1%

9	revistas.ucv.edu.pe Fuente de Internet	1 %
10	Submitted to Universidad Europea de Madrid Trabajo del estudiante	1 %
11	tesis.pucp.edu.pe Fuente de Internet	< 1 %
12	es.slideshare.net Fuente de Internet	< 1 %
13	vdocuments.mx Fuente de Internet	< 1 %
14	repositorio.comillas.edu Fuente de Internet	< 1 %
15	dspace.unl.edu.ec Fuente de Internet	< 1 %
16	Submitted to Universidad de Santiago de Chile Trabajo del estudiante	< 1 %
17	repository.usergioarboleda.edu.co Fuente de Internet	< 1 %
18	repositorio.una.ac.cr Fuente de Internet	< 1 %
19	aprendeia.com Fuente de Internet	< 1 %

VII. ANEXO Y APÉNDICE

7.1. Base de datos

UBIGEO	DEPARTAMENTO	PROVINCIA	DISTRITO	POBLACIÓN	DENUNCIAS	POLICIAS	SERENOS	INTERNOS	MUERTES	INSEGURO
130103	LA LIBERTAD	TRUJILLO	FLORENCIA DE MORA	38334	2540	84	10	373	3	1
70102	CALLAO	CALLAO	BELLAVISTA	80704	926	185	280	322	15	1
240301	TUMBES	ZARUMILLA	ZARUMILLA	25390	505	69	32	125	6	1
150507	LIMA	CAÑETE	IMPERIAL	38559	1851	54	49	340	3	1
150101	LIMA	LIMA	LIMA	267379	30272	876	1594	2223	52	1
150115	LIMA	LIMA	LA VICTORIA	188619	8947	388	162	1067	23	1
100101	HUANUCO	HUANUCO	HUANUCO	95540	6687	117	80	744	8	1
130102	LA LIBERTAD	TRUJILLO	EL PORVENIR	229115	4700	243	36	927	9	1
150201	LIMA	BARRANCA	BARRANCA	72221	2213	83	51	343	10	1
150111	LIMA	LIMA	EL AGUSTINO	221974	11927	356	84	1159	24	1
200101	PIURA	PIURA	PIURA	177748	11183	240	159	1092	19	1
110501	ICA	PISCO	PISCO	78636	2858	206	52	508	14	1
150202	LIMA	BARRANCA	PARAMONGA	21938	509	53	22	73	4	1
80106	CUSCO	CUSCO	SANTIAGO	103817	4214	353	61	459	8	1
50101	AYACUCHO	HUAMANGA	AYACUCHO	111370	17285	122	137	636	12	1
70101	CALLAO	CALLAO	CALLAO	512386	25653	999	227	2837	80	1
120601	JUNIN	SATIPO	SATIPO	42647	3439	121	44	231	8	1
240101	TUMBES	TUMBES	TUMBES	113458	7082	251	54	506	14	1
110507	ICA	PISCO	SAN CLEMENTE	28904	883	51	10	108	3	1
80101	CUSCO	CUSCO	CUSCO	118127	5296	432	220	335	4	1
110207	ICA	CHINCHA	PUEBLO NUEVO	73510	1921	31	21	178	3	1