PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Postgrado

Modelo de regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las Comisarías de la Policía Nacional de Perú en Lima Metropolitana

Tesis para optar el grado de Magíster en Informática con mención en Ciencias de la Computación que presenta:

Roger Chipa Sierra

Asesores:

Dr. César Armando Beltrán Castañón

Mg. Rodrigo Ricardo Maldonado Cadenillas

Lima, 2023

Informe de Similitud

Yo, Rodrigo Ricardo Maldonado Cadenillas, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulada Modelo de regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las Comisarías de la Policía Nacional de Perú en Lima Metropolitana, del autor Roger Chipa Sierra, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 21%. Así lo consigna el reporte de similitud emitido por el software Turnitin el 20/02/2023.

- He revisado con detalle dicho reporte y la Tesis, y no se advierte indicios de plagio.

- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

Lima, 20 de febrero del 2023.

<table>
<tr><td colspan="2">Apellidos y nombres del asesor / de la asesora:
Maldonado Cadenillas, Rodrigo Ricardo</td></tr>
<tr><td>DNI: 72535457</td><td rowspan="2">Firma</td></tr>
<tr><td>ORCID: 0000-0002-3845-4919</td></tr>
</table>

DEDICATORIA

A mis padres don Santos Justino Chipa y doña Catalina Sierra por enseñarme que todo sacrificio tiene recompensa.

AGRADECIMIENTO

A los señores Dr. César Armando Beltrán Castañón y Mg. Rodrigo Ricardo Maldonado Cadenillas, por haber aceptado este reto y haberme guiado en todo momento, demostrando la calidad de profesionales.

RESUMEN

En nuestro país, la cantidad promedio de denuncias por delitos de manera mensual en el año 2019 presenta 30,000 casos, lo cual se ha ido incrementando a lo largo de los años, por lo que es de crucial importancia generar estrategias de seguridad ciudadana que ayuden a mejorar el bienestar de la sociedad peruana, para ello es crucial tener información fiable y de calidad para la toma de decisiones; asimismo, tiene una Dirección de Tecnologías de la Información y Comunicación de la Policía (DIRTIC), la cual dentro de su estructura tiene a su cargo a la División de Estadística (DIVEST) y la División de Informática (DIVINFOR), las cuales tienen como función el control del Sistema de Denuncias Policiales en la cual el 80 % de las comisarías de la Policía Nacional del Perú se encuentran interconectadas y existe un 20 % que no se encuentran interconectadas, por lo cual el 20 % de comisarías no registra información en el Sistema de Denuncias Policiales, generando una incertidumbre para la toma de decisiones.

Es por tal motivo que la División de Estadística (DIVEST) propone un modelo para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana, en base a la técnica de regresión RANDOM FOREST REGRESSOR Y ÁRBOL DE DECISIÓN DE REGRESIÓN, con la cual se podrá conocer cuál de las dos técnicas genera un mejor pronóstico en base a la información que proviene del SIDPOL y del INEI.

Abstract

In our country, the average number of complaints for crimes on a monthly basis in 2019 presents 30,000 cases, which has been increasing over the years, so it is of crucial importance to generate citizen security strategies that help to improve the well-being of Peruvian society, for this it is crucial to have reliable and quality information for decision-making; likewise, it has a Police Information and Communication Technologies Directorate (DIRTIC), which within its structure has in charge of the Statistics Division (DIVEST) and the Information Technology Division (DIVINFOR), whose function is to control the Police Complaints System in which 80 % of the Peruvian National Police stations are interconnected and there are 20 % that are not interconnected, for which reason 20 % of police stations do not record information in the Police Complaints System, generating uncertainty for decision-making.

It is for this reason that the Statistics Division (DIVEST) proposes a model for forecasting the number of complaints for crimes that are registered in the Peruvian National Police stations in Metropolitan Lima, based on the RANDOM FOREST regression technique. REGRESOR AND REGRESSION DECISION TREE, with which it will be possible to know which of the two techniques generates a better prognosis based on the information that comes from SIDPOL and INEI.

Índice

Índice de figuras

Lista de abreviaturas

SIDPOL : Sistema Informático de Denuncias Policiales

DIRTIC : Dirección de Tecnologías de la Información y Comunicación de la Policía

DIVEST : División de Estadística

DIVINFOR : División de Informática

DIRTEL : Dirección de Telecomunicaciones de la Policía Nacional del Perú

PNP : Policía Nacional del Perú

MININTER : Ministerio del Interior del Perú

OPES : Oficina de Planeamiento Estratégico Sectorial

DGIS : Dirección de Gestión del Conocimiento y Seguridad del Estado

INEI : Instituto Nacional de Estadística e Informática

CEIC : Comité Estadístico Interinstitucional de la Criminalidad

PEA : Población económicamente activa

PDF : Formato de Documento Portátil

VB : Visual Basic

PYTHON : Lenguaje de programación de alto nivel que se utiliza para desarrollar aplicaciones de todo tipo

SCRAPING : Proceso de recopilar información de forma automática

SIRDIC : Sistema de registro de denuncias que usan las comisarias

IA : Inteligencia artificial

ANN : Red neuronal artificial

SVR : Regresión de vectores de soporte

GTB : Árbol de gradiente

EEUU : Estados Unidos

Lista de símbolos

R^2 : Coeficiente de Determinación

Alfa : Nivel de significancia

M : Regiones separadas

SRR: Suma de cuadrados residuales

p : Tamaño de espacio de características

T : t de student

n : Tamaño de muestra

$\bar{x}$: Media muestral

σ : Desviación estándar

μ : Media poblacional

H_o: Hipótesis Nula

H_1: Hipótesis alternativa

W_n: Media en una región especifica

V_m: Fracción de cada variable en un valor de umbral especifico

R_m: Región especifica

ns : Nivel de satisfacción de la información obtenida

te : Tolerancia al error del nivel de satisfacción de la información obtenida

d: Valor absoluto de la diferencia entre la cantidad de denuncias de la validación y la predicción

1. CAPÍTULO I: INTRODUCCIÓN

1.1. Aspectos sobre el sistema de denuncias Policiales

Es importante mencionar que, en la Policía Nacional del Perú se ha puesto en práctica el Sistema Informático de Denuncias Policiales "SIDPOL", con el cual se busca automatizar las funciones y procesos que son parte del registro de denuncias Policiales; de esta manera, proporciona información que servirá en la toma de decisiones.

Es por tales motivos que, se ha creado el sistema de denuncias que en la actualidad le permite a la ciudadanía ir a una comisaría a presentar la denuncia y recogerla en otra si es necesario, en menos de 24 horas debido a que se tiene este sistema llamado SIDPOL.

El SIDPOL se elaboró en base a las experiencias acumuladas por el personal de la Policía Nacional del Perú, que se ha desempeñado en este tipo de labores. El SIDPOL se va modernizarlo empleando las tecnologías existentes en informática, para la cual se requiere el empleo de un Modelo de *Machine Learning* para la estimación del nivel de los registros del Sistema de Denuncias Policiales.

Evidentemente, el acceso al sistema sólo está permitido a miembros de la Policía Nacional del Perú, los usuarios son proporcionados por responsables del Departamento de Mantenimiento y Desarrollo de Sistemas Informáticos de la División de Informática de la DIRTIC.

Es por tales motivos, que la Policía Nacional ha logrado un sistema integrado de servicios que está en permanente cambio tecnológico, a fin de elevar la calidad de este sistema de registro que tiende a brindar medidas de protección a los ciudadanos en todas las regiones y localidades en nuestro país. Así mismo, esta herramienta cada vez debe ser más efectiva ya que ayuda a la Policía Nacional en su lucha contra la delincuencia y el crimen organizado, entre otros males.

1.2. Importancia de los registros de datos de las denuncias Policiales

Es importante para las distintas áreas del Ministerio del Interior, contar con un sistema de registro de datos de las denuncias que se presentan en la Policía Nacional, por lo cual existe un mejoramiento en el nivel de calidad de información de los registros de datos del Sistema de Denuncias Policiales, en consecuencia existen iniciativas por parte del Ministerio del Interior para la implementación de diversos métodos y técnicas para la estimación de la cantidad de denuncias, que tiene como finalidad de reducir el error con lo cual se obtiene información y la cual sea mucho más confiable para los usuarios de la información, como son: La Dirección de Gestión del Conocimiento y Seguridad del Estado, La Oficina de Planeamiento Estratégico Sectorial, entre otras áreas.

Precisamente, la intención de conseguir información adecuada y confiable sobre la

violencia y criminalidad, en base a ello crear un sistema estadístico integrado de la criminalidad que se constituyó, mediante Decreto Supremo Nº 013-2013-MINJUS, el Comité Estadístico Interinstitucional de la Criminalidad (CEIC), en soporte al Consejo Nacional de Política Criminal, al Consejo Nacional de Seguridad Ciudadana y a varias instituciones del Estado. Las instituciones que conforman dicho son: INEI quien lo Preside, Poder Judicial, Ministerio Público, Ministerio del Interior, Policía Nacional del Perú, Instituto Nacional Penitenciario, y el Ministerio de Justicia y Derechos Humanos.

El estudio inicial sobre las comisarías se inició el 2010, en el cual se registró información del Programa de Accidentes de Tránsito. El año 2011 se ejecutó la Encuesta Nacional de Comisarías en la cual se acopió información de las primordiales características de infraestructura, equipamiento en áreas de comunicación e informática, labores de mantenimiento correctivo y preventivo. Desde al año 2012 al 2015 se ejecutaron el I al IV Censo Nacional de Comisarías. Además se elaboró un registro de delitos en las dependencias policiales 2014 con el fin de adquirir uno de los indicadores de criminalidad, como la tasa de homicidios, en los años 2011 y 2013 y entre los meses de marzo y mayo del 2016 se ejecutó el "REGISTRO NACIONAL DE DELITOS Y FALTAS, 2016".

1.3. Aspectos importantes sobre el INEI

Es importante mencionar sobre los Sistemas Nacionales de Estadística e Informática que poseen por propósito asegurar, en sus respectivos campos, sus actividades que se desarrollen en forma integrada, coordinada, racionalizada y bajo una normatividad técnica común, contando para ello con autonomía técnica y de gestión. Dichos sistemas poseen como ámbitos de competencia a los Sistemas Nacionales de Estadística e Informática. De este modo, están integrados principalmente por el Instituto Nacional de Estadística e Informática, el Consejo Consultivo Nacional de Estadística e Informática; el Comité de Coordinación Interinstitucional de Estadística e Informática y las Oficinas Sectoriales de Estadística e Informática y restantes Oficinas de Estadística o Informática de los Ministerios, de los Organismos Centrales, Instituciones Públicas Descentralizadas y Empresas del Estado.

Por tanto, el Instituto Nacional de Estadística e Informática (INEI), es una Entidad Público Descentralizado con personería jurídica de derecho público interno, con autonomía técnica y de gestión, anexa del presidente del Consejo de Ministros. También, es el organismo central y rector de los Sistemas Nacionales de Estadística e Informática, responsable de normar, planear, dirigir, coordinar y supervisar las actividades de estadística e Informática oficiales del país.

1.4. Planteamiento de la problemática

El 80 % de las comisarías de la Policía Nacional del Perú se encuentran interconectadas y existe un 20 % que no se encuentra interconectadas, por lo cual el 20 % de comisarias no registra información en el Sistema de Denuncias Policiales lo cual genera una incertidumbre para la toma de decisiones. Es por tales motivos, que las Unidades como: la Dirección de Gestión del Conocimiento y Seguridad del Estado (DGIS), Oficina de Planeamiento Estratégico Sectorial (OPES), Dirección de Seguridad Ciudadana, entre otras áreas que son parte del Ministerio del Interior del Perú (MININTER), presentan cierta incertidumbre generada por la falta de información sobre la cantidad de denuncias en determinadas comisarias que tienden a limitar efectivas decisiones adecuadas y oportunas en los planes de la Policía Nacional del Perú. La División de Estadística (DIVEST) propone un modelo de regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana, lo cual permitirá tener una herramienta para la estimación de la cantidad de denuncias por delitos en los lugares donde se requiera conocer. Para la cual comparan dos técnicas de modelo de regresión RANDOM FOREST REGRESSOR y ÁRBOL DE DECISIÓN DE REGRESIÓN, con lo cual se podrá conocer cuál de las dos técnicas genera un mejor pronóstico en base a la información que proviene del SIDPOL y del INEI.

1.5. Problema de investigación

1.5.1. Problema general

No se tiene una estimación precisa de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana, en el año 2019.

1.5.2. Problemas especificos

- No existe información estructurada sobre en número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, en los distritos de Lima Metropolitana
- La información captada se encuentra a nivel de distrito.
- No se ha aplicado un método para conocer buenos resultados para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana

1.6. Objetivo de la investigación

1.6.1. Objetivo general

Predecir la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional del Perú en Lima Metropolitana a través del modelo de regresión.

1.6.2. Objetivos especificos

- Obtener información sobre el número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, cantidad de denuncias de delitos en los distritos de Lima Metropolitana.
- Generar información estructurada adecuada para la aplicación del método de regresión.
- Comparar los métodos de regresión en el pronostico de la cantidad de denuncias por delitos que se registran en las comisarias de la Policía Nacional del Perú en Lima Metropolitana.

1.7. Justificación de la tesis

1.7.1. Implicaciones teóricas

Incorporar la teoría de regresión en las investigaciones para el pronóstico de la cantidad de denuncias por delitos.

1.7.2. Implicaciones prácticas

Las distintas unidades del Ministerio del Interior del Perú, como: Dirección de Gestión del Conocimiento y Seguridad del Estado, Oficina de Planeamiento Estratégico Sectorial, Dirección de Seguridad Ciudadana, entre otras áreas requieren conocer la cantidad de denuncias por delitos que ocurren en ciertos lugares que por ciertas razones no existe registro alguno en el SIDPOL, el modelo de regresión proporcionara una estimación de la cantidad de denuncias de los lugares en los cuales se requiera

1.7.3. Implicaciones metodológicas

Las áreas de investigación de las distintas instituciones dedicadas a la estimación del pronóstico de la cantidad de denuncias tengan un modelo a seguir, y de esta manera se pueda conocer además las aportaciones de los diferentes modelos.

1.8. Limitaciones

Se encontraron:

- El acceso a la información que se registran en el SIDPOL es muy restringido, debido a que se requiere un permiso especial que es otorgada por la División de Informática y la División de Estadística de la DIRTIC.
- La información acerca del número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, en los distritos de Lima Metropolitana se encuentra en documentos en formato PDF y se encuentran a nivel de distrito.

2. CAPÍTULO II: MARCO TEÓRICO

2.1. Antecedentes de trabajos nacionales y extranjeros

2.1.1. Antecedentes de trabajos nacionales

ESPINOZA, Sergio Ernesto (2020) en su tesis "PREDICCIÓN DE POSTULANTES QUE COMETERAN FRAUDE INTERNO CON ALGORITMO DE APRENDIZAJE SUPERVISADO" para optar el título de ingeniero de sistemas en la Universidad de Lima, Facultad de Ingeniería y arquitectura en abril del 2020. Esta tesis tiene como objetivo Pronosticar el fraude interno cometido por los postulantes para lo cual se utilizó algoritmos de aprendizaje supervisado en la etapa de selección. Debido a que uno de los procesos más transcendentales es la selección de personal por parte del área de recursos humanos, el cual incorporara personal apto a la organización. Lo cual tendrá como efecto las personas que edifican la institución. Los expertos de recursos humanos afrontan diversos desafíos en el proceso de selección debido a que requieren tomar varias medidas de tipo gerencial, para lo cual se propone que la existencia de variables en información sobre postulantes a una empresa los que permiten establecer con mayor seguridad si estos efectuarán fraude interno o no. Al aceptarse la hipótesis, se estaría demostrando que se puede utilizar la minería de datos con el fin de pronosticar qué posibles ingresantes a una empresa serían perjudiciales para esta, y como consecuencia se evitaría su admisión. Las variables "Estado marital" y los "Meses de promedio entre trabajos" deberían ser buenos predictores de un trastorno de personalidad antisocial, y por ende de fraude. Terceras variables que se escogen son ubigeo, pretensiones laborales, sexo, número de hijos y edad. Los algoritmos usados fueron Random Forest (Precision 79.4 %), Decision Tree (Precision 71.4 %), Red Neuronal con Regularizacion Bayesiana (96.8 %) y Red Neuronal con Back – Propagation (98.4 %). Mediante la cual el algoritmo de Red Neuronal con Back – Propagation presento mejores resultados.

DEL CARMEN, Walter Santiago (2021) en su trabajo de investigación "PROPUESTA DE UN SISTEMA PREVENTIVO PREDICTIVO DE DELITOS PATRIMONIALES TIPO X-LAW PARA EL DISTRITO DE PUEBLO LIBRE", para optar el título de Magister en Gestión Publica en la Universidad San Ignacio de Loyola en Lima 2021. El objetivo de esta investigación es diseñar una política pública en temas de prevención de delitos patrimoniales que, al ser acoplada con otras de carácter reactiva, creen una disminución en el índice de delitos patrimoniales en el distrito de Pueblo Libre, delitos como robo, hurto, estafa y entre otros similares. Para lo cual se tomó como modelo del caso de Italia con el sistema X-LAW que fue diseñado por la policía Napolitano Elia Lombardo, el cual utiliza el aprendizaje autónomo al investigar casos criminales que se sostienen en base a la información procedentes de las denuncias y de un algoritmo; los mismos elementos que detectan y predicen las mismas y similares escenarios en la que se comenten ilícitos. Debido a que es alarmante el aumento de la delincuencia en Lima Metropolitana y la ciudadanía del

distrito de Pueblo Libre no están ajeno a sufrir ilícitos. Existen planes de seguridad, sin embargo están orientados a la reacción, reducción y sanción; más no en la prevención de delitos como hurto, roba, estafa y demás. Se concluye que la ejecución de un sistema preventivo predictivo de delitos patrimoniales tipo X-LAW genera un impacto positivo en la disminución de delitos a partir diferentes ópticas, así como desde la prevención, como de la reacción, es importante la inmediatez del mismo.

GUTIERREZ, Mariella Vicky (2018) en su tesis “SISTEMA DE DISTRIBUCIÓN DE CARGA POLICIAL MEDIANTE DE PREDICCIÓN DE DELITOS” para optar el título de Ingeniero Informático en la Pontificia Universidad Católica del Perú, Facultad de Ciencias e Ingeniería.

Esta tesis tiene como objetivo desarrollar un sistema que permita la recepción de denuncias informales y que proponga una propuesta de distribución óptima de vehículos policiales. El proyecto sitúa su ámbito de aplicación en la capital del Perú, Lima, para que pueda mejorar su capacidad de reacción ante los delitos. Se ha elegido este ámbito por la importancia de agilizar la toma de decisiones en cuanto a la distribución de vehículos policiales, ya que el patrullaje puede prevenir los robos y actuar de una manera anticipada sin esperar que se concreten los delitos. Esta solución va dirigida a los ciudadanos de Lima, quienes podrán informar e informarse de los delitos cometidos en la capital, y a aquellas personas relacionadas con las 25 comisarías y serenazgos, como policías y serenos, para el uso de las propuestas de distribuciones óptimas de carga policial. La delincuencia es uno de los mayores problemas que hay en el Perú, especialmente, en Lima. En nuestra capital, el aumento de robos y delitos genera una gran preocupación ya que los policías no actúan rápido y no llegan a tiempo para impedir un robo o capturar ladrones. La situación actual es que las comisarías vigilan con sus patrullas los distritos sin considerar las zonas con mayor índice de criminalidad. Este criterio es de suma importancia pues puede ayudar a que los vehículos policiales velen por la seguridad de los ciudadanos de manera más estratégica. El presente proyecto contribuye ante esta problemática es desarrollar un sistema integral que permita la recepción de denuncias informales por medio de los ciudadanos y que genere, mediante un algoritmo que se retroalimente con los delitos registrados, una propuesta de distribución cercana a la óptima de vehículos policiales. La solución brindada por el sistema tiene como variableslos datos propios de la comisaría, cantidad de vehículos, horarios y las frecuencias de los delitos con el fin de convertir la labor policial de manera proactiva. Se concluye que se ha logrado construir una herramienta para la distribución de vehículos de la policía y de Serenazgo.

CHÁVEZ Enrique, RIVERA Yéssica, CHÁVEZ Elizabeth (2014) en la investigación “FACTORES DE RIESGO DE CONDUCTA DELICTIVA EN ALUMNOS DE NIVEL SECUNDARIO DE LAS ZONAS URBANO-MARGINALES DE LOS DISTRITOS DE HUÁNUCO, PILLCO MARCA Y AMARILIS”, la Universidad Nacional Hermilio Valdizán Pillco Marca, Perú. La finalidad de la investigación fue determinar los factores

de riesgo predominantes y el nivel de riesgo de conducta delictiva en alumnos de secundaria de los distritos de Huánuco, Amarilis y Pillco Marca. La metodología empleada fue de tipo descriptiva, a través de un diseño descriptivo simple, con una población de los alumnos del tercer año del nivel secundario y una muestra no probabilística intencional de 673 alumnos. El instrumento utilizado para recopilar la información fue el Cuestionario de Factores de Riesgo de la Conducta Delictiva (FRCD), elaborado para identificar factores biográficos, de déficit de desarrollo, familiares y socio-culturales, y que obtuvo una confiabilidad de 0.80 con el Alfa de Cronbach y una validez ítem test. Los resultados identifican factores de riesgo en sus diversos componentes, siendo los predominantes: vínculo afectivo inseguro en los primeros años de vida (63,64 %), falta de supervisión en el desarrollo académico de los padres hacia los hijos (55,72), dificultad en la expresión de sentimientos (51,26 %), presencia de pandillas juveniles en el entorno (47.85 %), dificultad en el razonamiento moral (45,77 %) y sector urbano con pocas alternativas de desarrollo (40,27 %). Asimismo los niveles de riesgo determinados corresponde a Muy alto (28.08 %), Alto (25.26 %), Bajo (24.96 %), Muy bajo (21.69 %). Destaca que el 53.34 % se encuentra en factor de riesgo Alto y Muy alto. Finalmente que los alumnos de las instituciones educativas que presentan más factores de riesgo en niveles alto y muy alto son: I.E. Héroes de Jactay (100 %), I.E. Potracancha (80 %), I.E. César Vallejo (79.31 %), I.E. Illatupa (72.73 %), I.E. Hermilio Valdizán (62.96 %), I.E. Marcos Durand Martel (61.76 %), I.E. Las Mercedes (59.37 %), I.E. Leoncio Prado (55.29 %); y las instituciones educativas que presentan menos fac-tores de riesgo en niveles bajo y muy bajo son: I.E. Aplicación UNHEVAL (77.78 %), I.E. Pedro Sánchez Gavidia (69.57 %), I.E. Javier Pulgar Vidal (68.57 %). I.E. Mariano Dámaso Beraún (66.67 %).

TORRES Irma Luisa, CAMONES Libia Justina (2013) en su tesis “ANÁLISIS ESTADÍSTICO Y PRONÓSTICOS CON SERIES TEMPORALES DE LA INFORMACIÓN DE LAS DENUNCIAS DE ACCIDENTE DE TRÁNSITO, COMISARIA DISTRITAL PNPHUARAZ: 2007- 2012” para optar el título de Licenciado en estadística e informática en la Universidad Nacional Santiago Antúnez de Máyalo, Facultad de Ciencias Escuela Académica Profesional de Estadística e Informática.

La presente tesis se desarrolló en la comisaria distrital PNP de Huaraz, tiene por objetivo el estudio de las denuncias de los accidentes de tránsito registrados en la comisaria de Huaraz. La investigación se justifica en la siguientes razones: Es importante conocer las estadísticas de las denuncias de los accidentes de tránsito ocurridos en la ciudad de Huaraz, interés para la comisaria distrital PNP de Huaraz y otras Instituciones. La finalidad es determinar y analizar los accidentes de tránsito, comportamiento de las denuncias por accidente de tránsito; cuantificar el total por el tipo de denuncias por accidente de tránsito, conocer los lugares que se suscitaron los accidentes de Tránsito, las causas, las consecuencias, características del vehículo y del conductor, con el detalle análisis de las denuncias registradas entre año 2008 y 2012. Se Utiliza la metodología de Box-Jenkins como apropiado para predicciones a largo Plazo que para corto plazo, las variables que se toman se

cuenta son: total de los accidentes de tránsito, tipo de accidentes de tránsito, lugar de la ocurrencia de los accidentes de tránsito, causas de los accidentes de tránsito, consecuencias de los accidentes tránsito, tipo de vehículo, tipo de licencia de conducir, edad del conductor, estado civil del conductor, edad del afectado. Se Pronostica las denuncias de los accidentes tránsito para el año 2013, utilizando los datos registrados desde el año 2007 hasta el año 2012. En esta investigación se persevero en una temática de interés de las autoridades: La Policía Nacional del Perú, Ministerio de Transportes y Comunicaciones, las Municipalidades provinciales, las Municipalidades Distritales y El Instituto Nacional de Defensa de la Competencia y de la Protección de la propiedad Intelectual-INDECOPI. Esta investigación ayudara a tomar decisiones, a dar soluciones, este problema que afecta a la ciudadanía Huaracina. Se concluye que el comportamiento de las Denuncias por Accidente de Tránsito en la comisaria distrital PNP de la ciudad de Huaraz, 2007 - 2012 tiene una tendencia creciente, y las variaciones determinadas nos muestran el aumento de casos respecto a los años anteriores, en tanto afirmamos que durante los seis años los accidentes de tránsito registrados se han incrementado en promedio de 73 % (274); esto se debe al aumento excesivo de los vehículos que brindan el servicio de taxi. En cuanto a tipo de accidente de tránsito los que mayormente ocurren son los accidentes de delitos peligro común (28.2 %), atropello con lesiones personales (23.2 %) y el de menor ocurrencia es el de caída de personas desde el vehículo (2.2 %) lo cual conlleva a afirmar que los accidentes ocurridos se debe a las fallas ocasionadas por el conductor. Asimismo, los accidentes de tránsito son uno de los principales problemas de salud pública y de desarrollo en el mundo, y afectan de forma desproporcionada a determinados grupos vulnerables de usuarios de la vía pública se producen a consecuencia de una acción riesgosa, irresponsable o negligente de un conductor, pasajero o peatón, ya sea en las vías de una ciudad o en carretera. Se puede decir que gran parte de los accidentes de tránsito son predecibles y evitables.

Rafael Caparó, Elizabeth Pari (2017) en su publicación “MODELOS DE ECONOMETRÍA ESPACIAL PARA LA LUCHA CONTRA LA DELINCUENCIA EN EL PERÚ: UN ENFOQUE DE OPTIMIZACIÓN EN TIEMPO REAL”, en la Facultad Ingeniería Económica, Estadística y Ciencias Sociales de la Universidad Nacional de Ingeniería. Dicho estudio tiene como propósito generar modelos de econometría espacial para la lucha contra la delincuencia en el Perú. El estudio se realiza en el departamento de Lima y periféricos, dado que el grado de correlación entre la regiones, evidencia que existe un efecto contagio de la delincuencia en las regiones periféricas de Lima y con una tendencia positiva y con una clara dependencia con los niveles de ingreso y densidad poblacional, por lo cual el modelo puede contrastar la realidad y predecir escenarios futuros en cuanto a niveles de delincuencia. Los modelos que se usan son: retardo espacial y los modelos del error espacial; este análisis se centra en la estimación clásica donde las variables recogidas son el ingreso promedio y la tasa de desempleo observadas en cada uno de los años desde el 2012 hasta el 2015. Como resultado se obtuvo, un alto índice de correlación espacial entre distritos como San

Juan de Lurigancho y distritos aledaños muestra que la delincuencia se expande de manera territorial hacia ese distrito, esto prueba la idea de un efecto del territorio sobre los actos delictivos, de manera simple se puede concluir que distritos con altos índices delincuenciales tendrían influencia en el incremento de incides delincuenciales en distritos limítrofes.

2.1.2. Antecedentes de trabajos extranjeros

ORDÓÑEZ Hugo, COBOS Carlos, BUCHELI Víctor (2020) en el artículo "MODELO DE MACHINE LEARNING PARA LA PREDICCIÓN DE LAS TENDENCIAS DE HURTO EN COLOMBIA" Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca, Departamento de Sistemas y la Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería, Universidad del Valle.

En Colombia uno de los actos delictivos sustanciales en muchas áreas es el hurto, acto que ha generado alto impacto en la sociedad, ya que los altos índices de este delito en los últimos años hacen que la gente se sienta insegura. En este artículo, se presenta un modelo de machine learning basado en Maquinas de Soporte Vectorial para regresión (SRV) ajustado para la predicción de la tendencia de hurtos en Colombia y sus tres principales ciudades (Bogotá, Medellín y Cali). El modelo fue validado con un dataset tomado del sistema de información de la Fiscalía Nacional de Colombia, el cual contiene 2,662,402 registros de delitos realizados en Colombia desde el año 1960 hasta el 2019. Los resultados fueren confrontados frente a un modelo de regresión lineal estándar y un modelo de SRV sin ajuste, los resultados obtenidos son prometedores y pueden servir para la toma de decisiones a las autoridades competentes en la lucha contra el hurto. Los resultados de la ejecución del modelo con los datos de las tres principales ciudades de Colombia tienen casi el mismo comportamiento, y muy similar al comportamiento general del país, lo que permite evidenciar que la preocupación general del país y sus diferentes regiones en usar políticas, estrategias y tácticas similares, a manera de mejores prácticas, tiene resultados similares

VENERI, Federico (2019) en su tesis: "METODOS PARA LA PREDICCIÓN DE ROBOS VIOLENTOS" para optar el grado obtención del título de Magister en Ingeniería Matemática de la escuela de Posgrado en Ingeniería Matemática, Facultad de la Ingeniería de la Universidad de la Republica, Uruguay.

Los delitos son eventos que no se distribuyen de manera homogénea en las ciudades. Estos tienden a concentrarse en algunas unidades geográficas denominados puntos calientes. Por este motivo, las agencias policiales han adoptado estrategias focalizadas de patrullaje intentando priorizar ciertas zonas. La piedra angular de este tipo de estrategia es la correcta identificación de estas zonas. Concentrándose en los robos violentos (rapiñas), este trabajo tiene como objetivo contribuir en dos aspectos. En primer lugar, caracterizar la evolución del delito y su comportamiento espacio temporal, y en segundo lugar comparar distintos métodos para

la predicción y selección de zonas a patrullar. Los resultados muestran que los delitos de rapiña presentan un alto nivel de concentración y un comportamiento de aglomeración espacio temporal. Una vez que un delito de rapiña es cometido, es probable que se observe otro a una distancia pequeña y en un breve periodo de tiempo. Esto puede responder a que algunas zonas son más atractivas para los ofensores (mecanismo endémico) y a la existencia de un patrón de contagio a zonas cercanas (mecanismo epidémico). Basado en estos dos mecanismos, se presentan y compararan cinco métodos para la selección de zonas a patrullar: conteo, estimación de densidad, un método prospectivo y dos modelos endémico epidémico, uno que utiliza la presencia de factores de riesgo (eg: bancos, cajeros, bares) que vuelvan más atractivas ciertas zonas. La comparación de los métodos en un día fijo y sobre una muestra aleatoria de días permitió identificar sus fortalezas y desventajas. El método basado en conteo presentó mejores resultados en cuanto a su tasa de éxito, sin embargo, la dispersión de celdas seleccionadas para patrullar puede dificultar una estrategia de patrullaje necesitando una mayor cantidad de personal. El método basado en la estimación de densidad selección´ o áreas más compactas, pero estas tienden a repetirse en el tiempo y presentó un menor nivel de tasa de éxito. Los métodos prospectivos presentaron los mejores niveles de cobertura en cuanto a celdas patrulladas, aunque también presentaron bajos niveles de tasa de éxito. Los modelos endémicos epidémicos aplicados en este trabajo se encuentran en una situación intermedia. Proporcionaron áreas de patrullaje compactas y celdas alejadas, con un nivel de tasa de éxito intermedio, por lo que pueden considerarse como una alternativa de compromiso entre eficacia predictiva y eficiencia de patrullaje. Es importante destacar que la introducción de covariables no represento una mejoría respecto a un modelo simple de contagio. Este trabajo es el primero en realizar una comparación de distintos métodos para seleccionar zonas de patrullaje en Uruguay, abriendo una línea de investigación a futuro para desarrollar otros métodos que puedan contribuir a mejorar la seguridad ciudadana uruguaya.

CHAMELCO, San Juan (2016) en su tesis “PREDICCIÓN Y PREVENCIÓN DE LA DELINCUENCIA JUVENIL DENTRO DE LA SEGURIDAD CIUDADANA” sustentada en licenciatura en investigación criminal y forense (fds) facultad de ciencias jurídicas y sociales UNIVERSIDAD RAFAEL LANDÍVAR campus ”san pedro claver, s. j.”de la Verapaz en Guatemala en marzo del 2016. Dicha posee el propósito de servir como un insumo para que los diferentes actores de la sociedad civil y del Gobierno puedan tomar decisiones y formulen soluciones, y se contribuya al diseño de una predicción y prevención en materia de violencia juvenil en Guatemala. La violencia juvenil tiene varias dimensiones, se encuentran las personas que sufren la violencia y los victimarios. Asimismo, existe otro grupo, el cual aún no es victimario, no obstante tienen una elevada posibilidad de serlo, este grupo es conocido como la población en riesgo. A la relación víctima y victimario se suma otro tipo de actores: los que se encargan de prevenir y penar actos delictivos. Evidentemente, en relación con la propuesta de la investigación, se pretende entender de una manera más clara la

violencia, especialmente la violencia juvenil. La metodología que se utiliza para la investigación consistió en realizar una revisión bibliográfica de investigaciones de este tema, la recolección de información y entrevistas a personas cercanas que laboran en las instituciones que se relacionan con el progreso de la seguridad ciudadana en cuanto a la prevención de la violencia y de rehabilitación de jóvenes que hayan cometido algún acto criminal, asimismo, la delincuencia juvenil se encuentra dentro de la problemática, donde se hace referencia a delitos cometidos por menores de edad, y es muy probable que los criminales adultos tuvieron su inicio en ese periodo de la vida, lo que refleja la importancia de la predicción para reducir y poder planificar la convivencia humana, ahí la importancia de la presente investigación, por tanto, el tipo de predicción y prevención de la delincuencia juvenil es muy importante prevenir y ello será viable por la Policía Nacional de Perú respecto al sistema de denuncias mediante los delitos en las Comisarias de Lima Metropolitana.

ALIF RIDZUAN KHAIRUDDIN, RAZANA ALWEE y HABIBOLLAH (2020) en su Artículo “UN ANÁLISIS COMPARATIVO DE LAS TÉCNICAS DE INTELIGENCIA ARTIFICIAL PARA PRONOSTICAR LA TASA DE DELITOS VIOLENTOS” sustentada en, SCHOOL OF COMPUTING, FACULTY OF ENGINEERING, UNIVERSITI TEKNOLOGI MALAYSIA, 81310, JOHOR BAHRU, JOHOR, MALAYSIA este estudio tiene como objetivo aplicar, comparar y analizar el rendimiento de pronóstico de tres técnicas de IA seleccionadas, a saber, la red neuronal artificial (ANN), la regresión de vectores de soporte (SVR) y el aumento del árbol de gradiente (GTB) en el pronóstico de datos de tasas de delitos violentos en los EE. UU, para lo cual se toma los delitos robo, asalto agravado, violación forzada y asesinato y homicidio no negligente.

Debido a que el reciente aumento de las tasas de delincuencia, especialmente los delitos violentos, ha sido un factor preocupante y una gran preocupación para todos los países del mundo. Los crímenes causan inmensas pérdidas económicas e infligen daños a diversas partes e individuos. Los crímenes también son una gran amenaza para la estabilidad de las comunidades y sociedades.

En consecuencia, en el esfuerzo por reducir los delitos violentos se han implementado y propuesto las técnicas para analizar y observar patrones de delitos violentos. Las técnicas aplicadas para la previsión o predicción de delitos se han convertido en una orientación que no solo analiza el patrón de delitos violentos, sino que también puede extrapolar la posible ocurrencia de delitos violentos en el futuro. La ventaja de la previsión de delitos es que puede ayudar a muchas agencias de aplicación de la ley federales y estatales, como las fuerzas policiales y los servicios de seguridad, proporcionando información importante en la planificación y gestión de medidas eficaces de prevención del delito.

En este estudio, se aplicaron tres técnicas de inteligencia artificial, a saber, ANN, SVR y GTB, para pronosticar cuatro tipos de tasas de delitos violentos en EE. UU asesinato y homicidio no negligente, violación forzada, agresiones agravadas y robo, los resultados mostraron que GTB superó a las otras técnicas de IA, como lo demues-

tran sus valores de medición de error cuantitativos más pequeños. Esto muestra que GTB se considera más apropiado en el manejo de datos de tasas de criminalidad de series de tiempo limitadas en comparación con ANN y SVR.

TOLEDO, Rodrigo (2005) en su Tesis: "MÉTODOS ECONOMÉTRICOS PARA EL PRONÓSTICOS DE DELITOS EN EL GRAN SANTIAGO" para optar al título de ingeniero comercial con mención en economía en la Universidad de Chile Facultad de Ciencias Económicas y Administrativas Escuela de Economía y Administración en Santiago de Chile agosto del 2005. Esta Tesis tiene como objetivo comprobar cuan efectivos son los modelos multiecuacionales del tipo vectores auto regresivos (VAR) y los modelos de sistemas de ecuaciones, para poder desarrollar pronósticos de corto plazo de delitos, aplicado a la comuna de Santiago en el período de enero de 2001 y el 30 de junio de 2004. De esta manera, se encontró que los modelos apropiados para la formulación de pronósticos presentan diferencias, dependiendo del sector que se esté tratando. Los errores de pronósticos para las series diarias son cercanos al 27 %. Similares a los estudios obtenidos para otros países como Inglaterra y Estados Unidos que han investigado la utilidad de los modelos econométricos para de pronósticos de delitos a corto plazo, en los cuales se alcanzaron excelentes resultados.

Los modelos han permitido detectar con mayor claridad en donde se llevarán a cabo los crímenes, los sectores que poseen mayor probabilidad a sufrir ataques en las distintas fechas, las horas y días de la semana; en donde se concentran los asaltos según zonas geográficas. La idea que se esconde detrás de la utilización de modelos econométricos para realizar pronósticos de delitos es que los crímenes tienden a presentar patrones de comportamiento definidos a lo largo del tiempo, los cuales pueden ser estudiados y capturados a través de técnicas estadísticas y matemáticas.

Existen múltiples técnicas de series temporales, según su estructura así, se tiene los modelos uniecuacionales que describen el comportamiento de una serie a lo largo del tiempo sin tener en consideración lo que pasa en las otras, dentro de estos modelos se puede encontrar todos aquellos pertenecientes a la familia ARIMA. También es posible clasificarlos en modelos multiecuacionales los que consideran que existe una interrelación en las distintas series que son materia de estudio, afectándose cada una de ellas ante cambios en las otras series, pues, dichos modelos son efectivos como el de vectores auto regresivos (VAR) y los modelos de sistemas de ecuaciones para el desarrollo de pronósticos de corto plazo, aplicado a la predicción de crímenes en Santiago.

PÉREZ, Meritxell; REDONDO, Santiago; MARTÍNEZ, Marian; GARCÍA, Carlos; AN- DRÉS, Antonio (2008) en el artículo "PREDICCIÓN DE RIESGO DE REINCIDENCIA EN AGRESORES SEXUALES" Universidad de Oviedo, España. La evaluación del riesgo de conducta violenta es un campo emergente en la actual Psicología de la Delincuencia. A partir de la investigación sobre carreras criminales y predic-

tores de riesgo, durantelos últimos años se han desarrollado diferentes escalas de evaluación del riesgo de violencia. Uno de estos instrumentos es el Sexual Violence Risk Assessment-20 (SVR-20), traducido y adaptado al español en el Grupo de Estudios Avanzados en Violencia de la Universidad de Barcelona. El objetivo de este estudio es evaluar la capacidad del SVR-20 para predecir la reincidencia sexual en una muestra española de delincuentes sexuales. Para ello se ha aplicado el SVR-20 de forma retrospectiva a un grupo de 163 agresores sexuales ya excarcelados. La capacidad discriminativa del instrumento ha sido evaluada a través del modelo de regresión logística. Se obtuvo un porcentaje de clasificaciones correctas de los sujetos no reincidentes del 79,9 % y de los sujetos reincidentes del 70,8 %. La curva ROC obtenida muestra una buena capacidad discriminativa del SVR-20 con un valor de área bajo la curva (AUC) de 0.83. La principal conclusión de este estudio es que el SVR-20 es un instrumento de utilidad para mejorar los pronósticos de riesgo de violencia sexual

2.2. Actores en la generación de información estadística policial

Para el desarrollo de la Tesis ha de ser prioritario, plantear un marco teórico y conceptual que establezca principalmente la participación de la entidad de la Policía Nacional del Perú en su sistema de denuncias, las mismas que de manera interrelacionada, dinámica y sistémica han de estar comprendidas en las siguientes:

- Policía nacional del Perú
- Sistema de denuncias policiales
- Instituto Nacional de Estadistica e Informatica.
- Comisarias en Lima Metropolitana
- Modelos de regresion para el pronostico de las denuncias
- Metodo arbol de decisión regresion
- Metodo random forest regresor

2.2.1. Respecto a la Policía Nacional del Perú

La Policía es una entidad estatal fundada para avalar el orden interno, el libre ejercicio de los derechos fundamentales de las personas y el normal desarrollo de las acciones de la ciudadanía. Es profesional y jerarquizada. Sus miembros simbolizan la ley, el orden y la seguridad en todo el Perú y posee competencia para intervenir en asuntos que corresponden con el cumplimiento de sus funciones.

El objetivo primordial de la Policía es garantizar, mantener y restablecer el orden interno. Protege y ayuda a las personas y a la comunidad. Garantiza la obediencia de las leyes y la seguridad del patrimonio público y privado. Previene, investiga y combate la delincuencia. Vigila y controla las fronteras.

Por tanto, la Policía es una institución que ha sido creada para garantizar la seguridad ciudadana, principalmente, para batallar la delincuencia, así como, también, para disminuir la delincuencia. Es por ello por lo que la Policía es una entidad estatal creada para garantizar el orden interno, el libre ejercicio de derechos fundamentales de las personas y el normal progreso de las actividades de la ciudadanía, es decir, garantizar el cumplimiento de las leyes y la seguridad del patrimonio público y privado. (Poder Ejecutivo Decreto Legislativo N° 1267, 2016)

2.2.2. Respecto a las comisarias en Lima Metropolitana

Una comisaría es una institución policial ofrece diversos servicios a la comunidad. Si bien, las comisarías se distribuyen en el territorio con la intención de que su alcance cubra la totalidad de la superficie. Por lo usual los territorios se dividen en distritos, cada uno de los cuales cuenta con una comisaría, la misma que tiene la obligación de garantizar la seguridad de su distrito.

Una comisaría básica es aquella que se encuentra tipificada en A, B, C, D y E de acuerdo con el número de efectivos policiales, densidad poblacional, servicios requeridos y área mínima requerida de construcción. La comisaría especializada es aquella que desarrolla un servicio específico, comprende comisarías de mujeres, turismo, aeropuertos, terminales terrestres y protección de carreteras.

Las comisarías cuentan con varias oficinas, en las cuales se reciben denuncias hechas por los ciudadanos y se realizan diversos trámites, como la gestión de documentación. También puede disponer de una sala de interrogatorio y de calabozos para alojar, de manera temporal, a personas que se encuentran detenidas. Como institución, la comisaría es una unidad de gran importancia debido a que se trata de la edificación más visitada en la vida cotidiana, y eso le atribuye un cierto grado de importancia necesaria para los ciudadanos que van a denunciar.

Las comisarías están distribuidas para que ningún espacio se quede sin vigilancia de la policía. Pero, en Lima Metropolitana, hay una notoria desproporción entre estas jurisdicciones policiales. Esto origina que en una misma ciudad existan zonas menos protegidas que otras, es por ello por lo que, en Lima Metropolitana existe una comisaría para 3.500 manzanas, que, si bien, Lima tiene una extensión de 2.600 kilómetros cuadrados y 105.833 manzanas registradas una a una en el Censo Nacional de Población y Vivienda del 2017, elaborado por el INEI.

Por tales motivos, Lima tiene 114 comisarías básicas de las cuales 69 están en las áreas periféricas. Una de ellas es la de Huaycán, en Ate. Su responsabilidad como comisaría se extiende a lo largo de 3.503 manzanas, muchas de estas en cerros inaccesibles donde se levantan asentamientos humanos. Ahí viven más de 160 mil personas: el 13 % no accede a luz eléctrica y el 25 % habita casas precarias de madera. Sol de Oro es la comisaría de Los Olivos que más denuncias de robo registró en los últimos cuatro años en todo el país, pues cada año se denuncian unos 6 mil asaltos.

En general, las jurisdicciones policiales del norte, sur y este son más grandes que de la zona más céntrica y moderna de la ciudad, debido a que las 69 comisarías básicas ubicadas en los distritos de la periferia deben patrullar territorios que en promedio superan las 1.370 manzanas, es decir, que son cuatro veces más grandes que las Comisarías céntricas, pues estas en promedio patrullan territorios de 327 manzanas, con casos extremos como en la de la comisaría de Alfonso Ugarte que patrulla 76 manzanas del Cercado de Lima. Las comisarías de José Carlos Mariátegui (Villa María del Triunfo), Puente Piedra, El Progreso (Carabayllo), Huaycán y otras recorren territorios 10, 20 y hasta 40 veces más grandes. La desproporción es notoria. Asimismo, se ha identificado a otras 23 Comisarías de Lima y Callao custodiando territorios de más de un distrito. La mayoría están en El Agustino, San Martín de Porres, Villa María del Triunfo, Chorrillos, Villa El Salvador, Carabayllo y Comas.

Es por ello que en los distritos con las más altas concentraciones poblacionales se requerirían nuevas comisarías, pero, entendiendo que la capacidad del Estado es limitada, se debe hacer una redistribución de los recursos materiales y tecnológicos a las zonas más inseguras y pobladas y las que ya existen tengan capacidad operativa en sus jurisdicciones por más extensas que estas sean, debido a que la comisaría de Flor de Amancaes, construida en los cerros del Rímac, esta última que se creó en Lima en el año 2014, atendió 688 denuncias por robo en el período del 2015 al 2018 a pesar de los estar implementada por 50 policías por el Ministerio del Interior.

Lamentablemente, en los últimos cinco años no se han creado más comisarías en Lima pese a que el número actual de estos locales y cómo se distribuyen en el territorio está generando que el servicio policial prestado sea desproporcionado, otras comisarias no han sido priorizadas en su construcción debido a que el presupuesto es insuficiente respecto a la inversión, que incluye la construcción de los locales, equipos de comunicación, mobiliario, vehículos y su dotación de policías. (Paz Campusano, 2019)

2.2.3. Respecto al Sistema de Denuncias Policiales SIDPOL

En primer lugar, el SIDPOL busca automatizar funciones y procedimientos que se relacionan al registro de la denuncia; que ayudara en la toma de decisiones. Es decir, en la Policía Nacional del Perú se busca realizar una mejora en la calidad de información de los registros de datos del Sistema de Denuncias Policiales, mediante la comparación de los modelos, los algotirmos construyen modelos *Machine Learning* en los registros de datos del Sistema de Denuncias Policiales, con respecto al método tradicional. Pues, se trata de generar un modelo de regresión para el pronóstico de la cantidad de denuncias por delitos, para la cual comparan dos técnicas de modelo de regresión: Random Forest Regressor y Árbol de Decisión de Regresión, mediante las cuales se podrá conocer cuál de las dos técnicas genera un mejor pronóstico.

En segundo lugar, es importante mencionar que en la Policía Nacional se ha puesto en práctica el SIDPOL, que busca automatizar los procedimientos en el registro de la

denuncia; además, proporciona información para la toma de decisiones a nivel de las Comisarías, Distritos y Divisiones Territoriales. Es por tales motivos que, se ha creado el SIDPOL que hoy en día le permite a la ciudadanía ir a una comisaría a presentar la denuncia y recogerla en otra si es necesario. El sistema se diseñó en base a las experiencias del personal de la PNP, que se ha desempeñado en estas labores.

En tercer lugar, con el sistema de denuncias solo permite a los usuarios que se encuentren registrados con una cuenta de usuario y la contraseña que son proporcionados por el administrador de usuarios del Departamento de Mantenimiento y Desarrollo de Sistemas Informáticos de la DIVINFOR - DIRTEL PNP. Que es el Departamento que está dando oportunidad con la modernización del servicio y desde allí contribuir a crear el sistema de denuncias que usan las comisarías; del registro (SIRDIC), que utiliza la unidad de investigación criminal; y de la información policial, que sirve para consultar las requisitorias y antecedentes (SINPOL)

Por tanto, la Policía Nacional viene implementando un sistema integrado de servicios que está en permanente cambios tecnológicos a fin de elevar la calidad de este sistema de registro que tiende a brindar medidas de protección a los ciudadanos en todas las regiones y localidades en nuestro país, que cada vez debe ser más efectiva esta herramienta que ayuda a la Policía Nacional, en su lucha frente a la delincuencia y el crimen organizado, entre otros males. (Dirección de Telemática de la PNP, 2018), (Benites y Cervantes ,2017)

2.2.4. Respecto al Instituto Nacional de Estadística e Informática respecto a la información sobre los sistemas de criminalidad y seguridad ciudadana

En principio, las estadísticas nos revelan entre otros, sobre el problema nacional y como perjudica a los derechos fundamentales de la ciudadanía, como es el derecho a vivir en calma y en circunstancias apropiadas; por lo que es transcendental el análisis de la predilección del crimen y la ausencia de seguridad ciudadana en el país. En este contexto, el INEI, en coordinación con el Comité Estadístico Interinstitucional de la Criminalidad - CEIC, contribuyen con las autoridades, entre las cuales está la Policía Nacional del Perú e entidades públicas como privadas, en documento general el Anuario Estadístico de la Criminalidad y Seguridad Ciudadana, 2011-2017 con una visión a nivel de todo el país, la cual fue con base a los resultados de estudios que realiza el INEI y los registros administrativos sectoriales de las instituciones que miembros del CEIC.

Esta información contribuye en la creación de la políticas públicas en esta materia, debido a que afronta temas sobre la violencia y la criminalidad, comisión de los delitos, denuncias penales, delitos que ingresan en las Fiscalías Provinciales Penales y Mixtas; también el perfil de las personas que son ingresadas en el Registro Nacional de Detenidos y Sentenciados a pena privativa de libertad efectiva. De esta manera contiene los resultados del Censo Nacional de Comisarías 2017, estadísticas municipales. (Instituto Nacional de Estadística e Informática, 2018)

2.2.5. Respecto a la División de Estadística e Informática (DIVEST) en la Policía Nacional en el Perú

La DIRTIC, es un órgano de apoyo de la Policía Nacional del Perú, encargado brindar soporte técnico para optimizar los servicios de telecomunicaciones e informática en las unidades policiales y realizar el procesamiento de la información estadística para la toma de decisiones en las diferentes dependencias de la Policía Nacional y otros usuarios con fines de apoyo, concordante con el artículo 6° Estructura Orgánica de la Policía Nacional del Perú, según el Decreto Supremo N° 026-2017-IN Reglamento del Decreto Legislativo N° 1267 Ley de la Policía Nacional del Perú.

Asimismo, en el artículo 82° del reglamento en la DIRTIC cumple entre otra función la siguiente: “Dirigir y supervisar de manera articulada el desarrollo y mantenimiento de los servicios informáticos, de telecomunicaciones y estadística en los órganos y unidades orgánicas de la Policía Nacional del Perú”.

Igualmente, en el artículo 87° se norma sobre la División de Estadística, que es la unidad orgánica responsable de desarrollar las actividades vinculadas con los procesos de recopilación, producción, análisis, monitoreo y difusión de la información estadística de la Policía Nacional del Perú. Asimismo, es la encargada de dirigir el Sistema Estadístico Policial conforme a los lineamientos del Ministerio del Interior y las disposiciones del Director General de la Policía Nacional del Perú en el marco de la normativa sobre la materia. (Ministerio del Interior Decreto Supremo N° 026.2017-IN, 2017)

2.3. Bases teóricas de los modelos de regresión

El modelo de regresión para el pronóstico comprende los siguientes métodos fundamentales:

2.3.1. Método de Arbol de Decisión de Regresión

Un árbol de decisión presenta semejanzas en la vida, e influye en una vasta área del Machine Learning. Son técnicas de aprendizaje supervisado que pronostica las respuestas en base al aprendizaje de una serie de reglas de decisión procedentes de características. Los árboles trabajan mediante la división del espacio de características en diversas regiones rectangulares simples, divididas por divisiones paralelas de ejes. Para lograr un pronóstico para una observación específica, se utiliza el promedio o el modo de las respuestas de las observaciones de entrenamiento, dentro de la partición a la que corresponde la nueva observación (Espinoza, 2014). Se muestra la función del árbol de decisión:

$$f(x) = \sum_{m=1}^{M} W_m \, \varphi(x; V_m)$$

Donde:

W_m: *Es la respuesta media en una región especifica* (R_m)

V_m: *Representa cómo se fracciona cada variable en un valor de umbral especifico*

Las divisiones definen cómo el espacio de características en R-cuadrado en “M” regiones separadas, hiperbloques. Es decir, el concepto de la partición paralela de ejes se trasciende claramente a dimensiones mayores a dos. Para un espacio de “p” características, el espacio se divide en regiones “M”, cada una de las cuales es un hiperbloque p-dimensional. La heurística elemental para crear un árbol de decisión es la siguiente:

- Las características p, divide el área de características p-dimensional, en “M” regiones cambiantes que se mezclan totalmente el subconjunto del espacio de características y no se incorporan.
- Una observación nueva que caiga en un fraccionamiento específico tiene la respuesta apreciada dada por el promedio de todas las observaciones de entrenamiento con la partición.

Este proceso no relata efectivamente cómo crear el fraccionamiento de un modo algorítmico. Para ello se requiere utilizar la técnica de división binaria recursiva. El propósito de este algoritmo es disminuir algún error. Se desea minimizar la suma de cuadrados residual (RSS).

Desgraciadamente, es muy costoso computacionalmente que se considere a todos los fraccionamientos en el espacio de la característica “M”, consecuentemente, se tiene que usar un rumbo de exploración menos intensivo en materia computacional, pero con mucha mayor sofisticación, aquí interviene la división binaria recursiva la cual afronta el problema iniciando en la parte superior del árbol y fragmentando en dos ramas, lo cual establece una fraccionamiento de dos espacios. Lleva a cabo esta partición en particular en la parte superior varias veces y se elige la partición de las características que minimiza la suma de cuadrados residual.

El árbol crea una nueva rama en el fraccionamiento específico y se lleva a cabo el equivalente proceso, es decir, valúa el RSS en cada fraccionamiento de la partición y opta por el adecuado. Lo cual lo convierte en un algoritmo codicioso.

En la figura 1 ilustra el árbol en el lado derecho y la partición del espacio en el izquierdo. La partición del espacio se hace de manera repetitiva para encontrar las variables

y los valores de corte “c” de tal manera que se minimice la función de costos en base al error cuadrático medio (ECM).

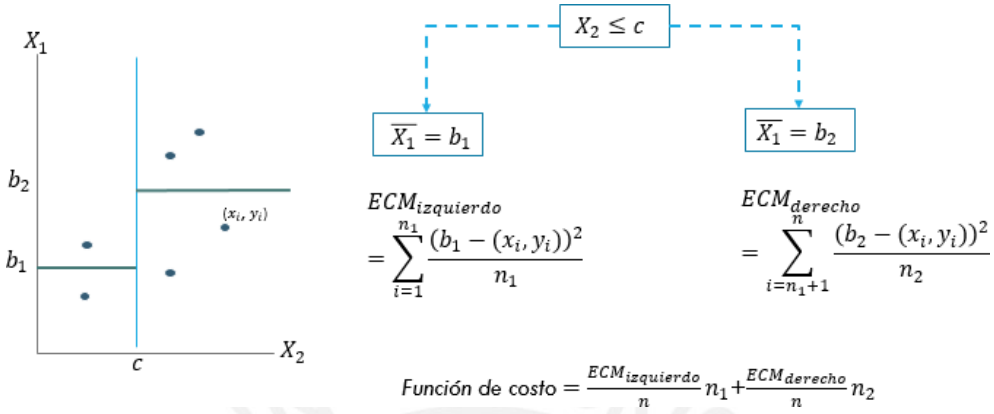


Figura 1: Árbol de Decisión de Regresión y función de costo

2.3.2. Método Random Forest Regressor

Es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación de BAGGING que construye a una colección de árboles no correlacionados y luego los promedia. Es un algoritmo de clasificación supervisado. Este algoritmo crea el bosque con varios árboles, que, cuanto mayor sea el número de árboles en el bosque, mayor será la precisión (Espinoza, 2014).

Las ventajas son las siguientes:

- Puede resolver problemas de clasificación y regresión, y realiza una estimación decente en ambos frentes.
- Maneja grandes cantidades de datos con mayor dimensionalidad. Puede manejar miles de variables de entrada e identificar las variables más significativas.
- Posee un método para estimar datos faltantes y mantiene la precisión cuando falta una gran proporción de los datos.
- Ser uno de los algoritmos más certeros que hay disponible. Para datos lo suficientemente grande produce un clasificador certero.
- Ser eficiente en bases de datos grandes.
- Manejar muchas variables de entrada sin exclusion.
- Mostrar las estimaciones de las variables son importantes en la clasificación.

- Tener un eficacia en la estimación de datos perdidos y mantiene la exactitud cuándo existe una elevada proporción de los datos perdidos.

Sin embargo, también este algoritmo posee desventajas:

- Realiza un adecuado trabajo en la clasificación, pero no es tan bueno como para los problemas de regresión, debido a que no proporciona predicciones precisas y continuas sobre la naturaleza. En el caso de la regresión, no predice más allá del rango en los datos de entrenamiento, y que pueden sobre ajustar los datos que son ruidosos.
- Se puede parecer este algoritmo a una caja negra, debido a que se tiene muy poco de control sobre lo que realiza el modelo.

A continuación la figura 2 ilustra como se unifican las predicciones.

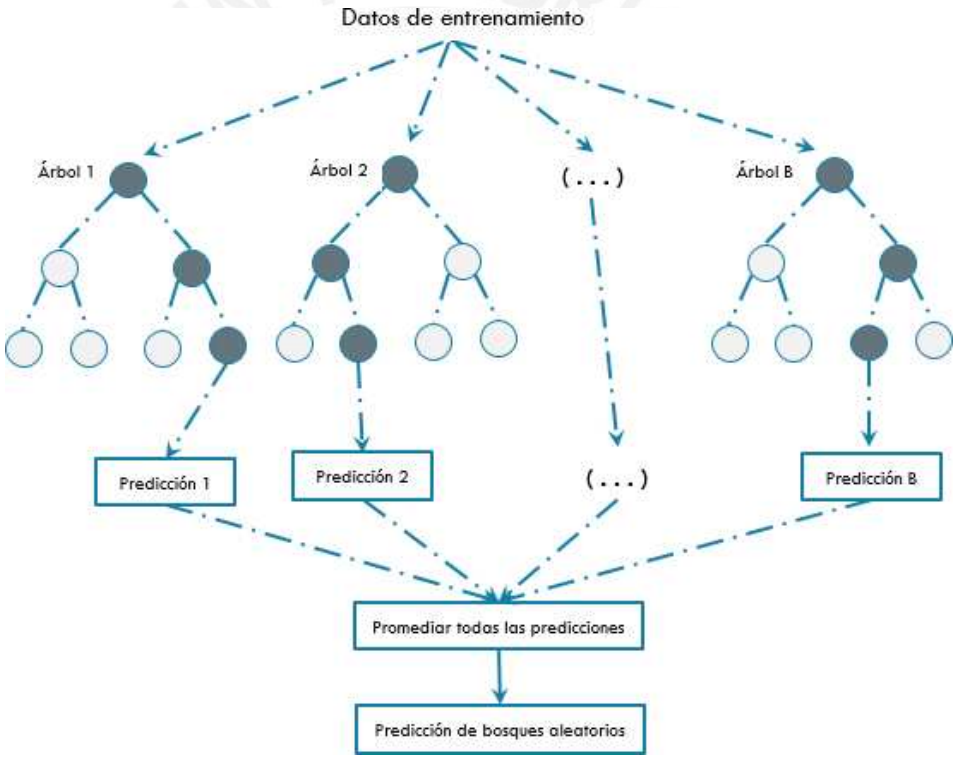


Figura 2: Método de Random Forest Regressor

2.3.3. Prueba de hipótesis

En el presente capitulo se desarrollan procesos de inferencia estadística centrados en la prueba de hipótesis, los cuales involucran uno o más parámetros desconocidos; se inicia con algunos conceptos básicos asociados al tema de hipótesis y luego se

plantea un algoritmo que posibilita llevar a cabo diversos perecimientos de inferencia estadística.

Clases de hipótesis

Los procesos de inferencia estadística suelen involucrar dos clases de hipótesis denominadas hipótesis nula e hipótesis alternativa; generalmente, la primera se simboliza con Ho, y la segunda, con H1 (Burbano y Valdivieso, 2016).

La hipótesis nula Ho se formula como una afirmación que indica que un determinado parámetro se mantiene en un valor; esta hipótesis es aquella que se acepta o se rechaza (Burbano y Valdivieso, 2016).

La hipótesis alternativa H1 se plantea en términos de que el parámetro ha cambiado (aumentado o disminuido); esta es la hipótesis de investigación, la cual se prueba utilizando los datos de una muestra aleatoria.

Tipos de errores

En general, se suelen cometer dos tipos de errores al aceptar o rechazar la hipótesis nula: el error tipo I y el error tipo II.

El error tipo I consiste en rechazar la hipótesis nula Ho, dado que es verdadera. La probabilidad de rechazar la hipótesis nula, siendo verdadera, se denomina nivel de significancia de la prueba y se denota con alpha; de aquí se desprende que 1-alpha sea la probabilidad de no rechazar Ho, dado que esta es verdadera (Burbano y Valdivieso, 2016).

El error tipo II consiste en aceptar la hipótesis nula Ho, dado que es falsa. La probabilidad de aceptar la hipótesis nula, siendo falsa, se denota Beta; de aquí se sigue que 1-Beta sea la probabilidad de rechazar Ho dado que esta es falsa (Burbano y Valdivieso, 2016).

Algoritmo para desarrollar un proceso de prueba de hipótesis

(1). Plantear las hipótesis Ho y H1.

(2). Establecer el nivel de significancia alpha menor o igual a 0.05

(3). Determinar la dirección de la prueba: unilateral izquierda, unilateral derecha o bilateral.

(4). Determinar la estadística de prueba “Z”, “t”, “X”, “F” y calcularla.

(5). Comparar el valor de la estadística de prueba con el valor en la distribución teórica.

(6). Tomar una decisión: aceptar Ho o, en caso contrario, rechazar Ho.

(7). Escribir una conclusión.

Prueba de hipótesis para la diferencia de medias poblacionales

En esta sección se indica el proceso de inferencia estadística para la prueba de hipótesis asociadas con la diferencia de medias poblacionales, el cual ha de desarrollarse a través de los siguientes pasos (Moya, 1998).

(1). Planteamiento de hipótesis.

$$i) \quad H_o: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 > 0$$

$$ii) \quad H_o: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 < 0$$

$$iii) \quad H_o: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

(2). Fijación del nivel de significancia alpha

(3). Dirección de prueba. Se determina en concordancia con cada una de las parejas de hipótesis.

(4). Estadística de prueba. Se presentan cuatro casos, a saber.

Caso 1. Si se conocen las desviaciones estándar poblacionales, se ha de utilizar

$$Z = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (1)$$

Sin embargo, bajo la hipótesis Ho: u1=u2, la ecuación (1) se simplifica y reduce a la siguiente estadística de prueba:

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (2)$$

Caso 2. Si las desviaciones estándar poblacionales son desconocidas, pero se suponen iguales, para muestras inferiores a 30 se utiliza una estadística de prueba asociada con la distribución t student.

$$t = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (3)$$

Nuevamente, bajo la hipótesis nula Ho: u1=u2, la ecuación (3) se simplifica y reduce a la siguiente estadística de prueba:

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (4)$$

Donde Sp se obtiene por medio de la expresión:

$$S_p = \sqrt{\frac{(n_1-1)\hat{S}_1^2 + (n_2-1)\hat{S}_2^2}{n_1+n_2-2}} \qquad (5)$$

Caso 3. Si las desviaciones estándar poblacionales son desconocidas, pero se suponen distintas, para muestras inferiores a 30 se utiliza una estadística de prueba asociada con la distribución t student con g grados de libertad.

$$t = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \qquad (6)$$

Otra vez, bajo la hipótesis nula Ho: u1=u2, la ecuación (6) se simplifica y reduce a la siguiente estadística de prueba:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \qquad (7)$$

Caso 4. Si las desviaciones estándar poblaciones son desconocidas, pero las muestras tienen tamaño superior a 30, se utiliza una estadística de prueba asociada con la distribución normal estándar dada en la expresión:

$$Z = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \qquad (8)$$

Bajo la hipótesis nula Ho: u1=u2, la ecuación (8) se simplifica; luego se usa la siguiente estadística de prueba:

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \qquad (9)$$

(5). Comparar el valor de la estadística de prueba con el valor en la distribución teórica.

(6). Toma de decisión: aceptar Ho, en caso contrario, rechazar Ho.

(7). Conclusión.

2.3.4. Comparación de modelos

Curva ROC

Una curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la

especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este gráfico es la representación de la razón o proporción de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o proporción de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo). El Análisis ROC es una metodología desarrollada para evaluar la capacidad de un modelo para clasificar de manera correcta. Un valor de 1 significa que el método es perfecto; un valor de 0.5 indica que el método no es útil, y valores intermedios miden la capacidad del método para discriminar. Una de las principales ventajas de usar las curvas ROC es que además de entregarnos un valor de decisión de manera automática, nos muestra una representación gráfica de fácil interpretación y rápido entendimiento visual (Fawcett, 2005).

Coeficiente de determinación

El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar. Es importante saber que el resultado del coeficiente de determinación oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será (Montgomery, 2005).

2.4. Hipótesis de investigación

2.4.1. Hipótesis General

La aplicación de un modelo de regresión en el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana genera un buen método para una buena estimación en el pronóstico

2.4.2. Hipótesis Especificas

- La aplicación del método de scraping para obtener información sobre el número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, en los distritos de Lima Metropolitana, genera la información deseada.
- La aplicación de las técnicas X, Y, Z en el preprocesamiento de la información extraída de los distritos de Lima Metropolitana, genera una información adecuada para la aplicación del método de regresión (Las técnicas X, Y, Z se describen en el capítulo 4, en la sección 4.3 Identificación de las técnicas).

- En la aplicación de los métodos de Random Forest Regressor y Árbol de decisión regresión se encuentra que una aplica mejor para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

2.5. Variables independientes y dependientes

2.5.1. Variable dependiente

Cantidad de denuncias de delitos por Comisaria (W)

2.5.2. Variables independientes

- Tipo de denuncia (X1)
- Sub tipo de denuncia (X2)
- Modalidad (X3)
- Número de habitantes (X4)
- Área en kilómetros (X5)
- Cantidad de mercados de abastos (X6)
- Pea ocupada (X7)
- Pea no ocupada (X8)
- Ingreso per cápita por hogar (X9)
- Nivel socio económico (X10)
- Número de habitantes por sereno (X11)

2.5.3. Identificación del tipo de variable

Variable	Tipo
Cantidad de denuncias de delitos por Comisaria (W)	Cuantitativa Discreta
Tipo de denuncia (X1)	Cualitativa Nominal
Sub tipo de denuncia (X2)	Cualitativa Nominal
Modalidad (X3)	Cualitativa Nominal
Número de habitantes (X4)	Cuantitativa Discreta
Área en kilómetros (X5)	Cuantitativa Continua
Cantidad de mercados de abastos (X6)	Cuantitativa Discreta
Pea ocupada (X7)	Cuantitativa Discreta
Pea no ocupada (X8)	Cuantitativa Discreta
Ingreso per cápita por hogar (X9)	Cuantitativa Continua
Nivel socio económico (X10)	Cualitativa Ordinal
Número de habitantes por sereno (X11)	Cuantitativa Discreta

Tabla 1: Identificación del tipo de variable

2.6. Matriz de consistencia

PROBLEMA GENERAL: No se tiene una estimación precisa de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional del Perú en Lima Metropolitana, en el año 2019.	**OBJETIVO GENERAL:** Predecir la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional del Perú en Lima Metropolitana a través del modelo de regresión.	**HIPOTESIS GENERAL:** La aplicación de un modelo de regresión en el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana, genera un buen método para una buena estimación en el pronóstico.
PROBLEMA ESPECIFICO 1: No existe información estructurada sobre en número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, cantidad de denuncias de delitos en los distritos de Lima Metropolitana.	**OBJETIVO ESPECIFICO 1:** Obtener información sobre el número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, cantidad de denuncias de delitos en los distritos de Lima Metropolitana.	**HIPOTESIS ESPECIFICO 1:** La aplicación del método de scraping para obtener información sobre el número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, en los distritos de Lima Metropolitana, genera la información deseada.
PROBLEMA ESPECIFICO 2: La información captada se encuentra a nivel de distrito	**OBJETIVO ESPECIFICO 2:** Generar información estructurada adecuada para la aplicación del método de regresión.	**HIPOTESIS ESPECIFICO 2:** La aplicación de las técnicas X, Y, Z en el preprocesamiento de la información extraída de los distritos de Lima Metropolitana, genera una información adecuada para la aplicación del método de regresión.
PROBLEMA ESPECIFICO 3: No se ha aplicado un método para conocer buenos resultados para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.	**OBJETIVO ESPECIFICO 3:** Comparar los métodos de regresión en el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional del Perú en Lima Metropolitana.	**HIPOTESIS ESPECIFICO 3:** En la aplicación de los métodos de **Random Forest Regressor y Árbol de decisión regresión** se encuentra que una aplica mejor para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la policía nacional de Perú en Lima Metropolitana.

Tabla 2: Matriz de consistencia de modelo de regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

3. CAPÍTULO III: METODOLOGÍA DE LA INVESTIGACÍON

3.1. Clasificación de la investigación

Es un estudio de tipo aplicativo ya que planea solucionar problemas prácticos. Con un nivel de investigación descriptivo dado que reside en la identificación del fenómeno con el objetivo de establecer su estructura y comportamiento.

3.2. Cobertura de estudio

3.2.1. Población

La población está compuesta por las comisarías en la Policía Nacional del Perú, en la cual se registran las denuncias de la población en la plataforma del SIDPOL. Se tomó como muestra la cantidad de denuncias por delitos que se registran en la comisaría de Lima Metropolitana en el año 2019, de los registros que se hallan en la División de Estadística de la DIRTIC, se tomó 5154 registros de denuncias de las distintas comisarías de Lima metropolitana.

3.2.2. Unidad muestral

Una comisaria

3.3. Fuentes de recolección de información

La primera fuente para la recolección de datos es el SIDPOL, de la cual se obtienen las siguientes variables: cantidad de denuncias de delitos, tipo de denuncia, subtipo de denuncia y modalidad de la denuncia en las comisarías de lima metropolitana.

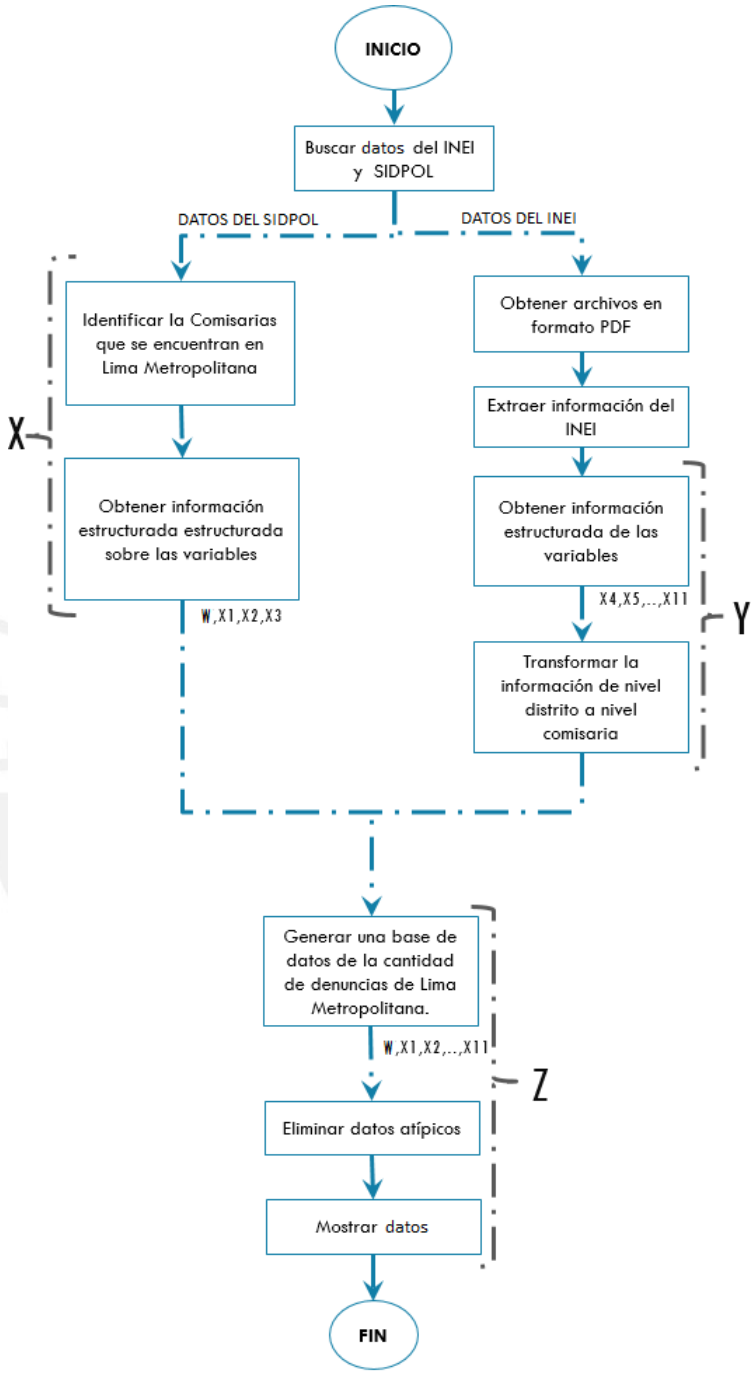
La segunda fuente es el INEI, de la cual se obtiene las variables: número de habitantes, área en kilómetros, cantidad de mercados de abastos, pea ocupada, pea no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno de los distritos de lima metropolita.

3.4. Instrumento de recolección de información

La herramienta para recolectar datos en cuanto al Sistema de Denuncias Policiales son los formatos de registro de la denuncia, como se muestra (Anexo Figura 32).

En cuanto a la información que se toma del INEI, vendría a ser una fuente secundaria, debido a que se obtendría de los distintos boletines estadísticos en los cuales se encuentra la información, como se muestra (Anexo Figura 14).

3.5. Técnicas de recolección y procesamiento de datos



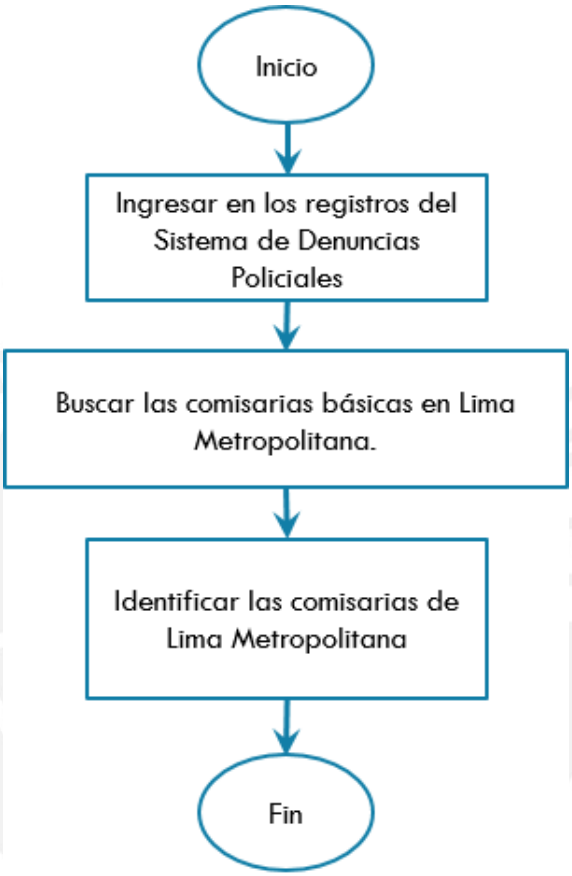
FUENTE: ELABORACIÓN PROPIA

Figura 3: Diagrama de recolección y procesamiento de información del SIDPOL y del INEI

3.5.1. Buscar datos del INEI y SIDPOL

Se realiza la búsqueda en los archivos en formato PDF del INEI y en el SIDPOL.

3.5.2. Identificar las comisarias que se encuentran en Lima Metropolitana

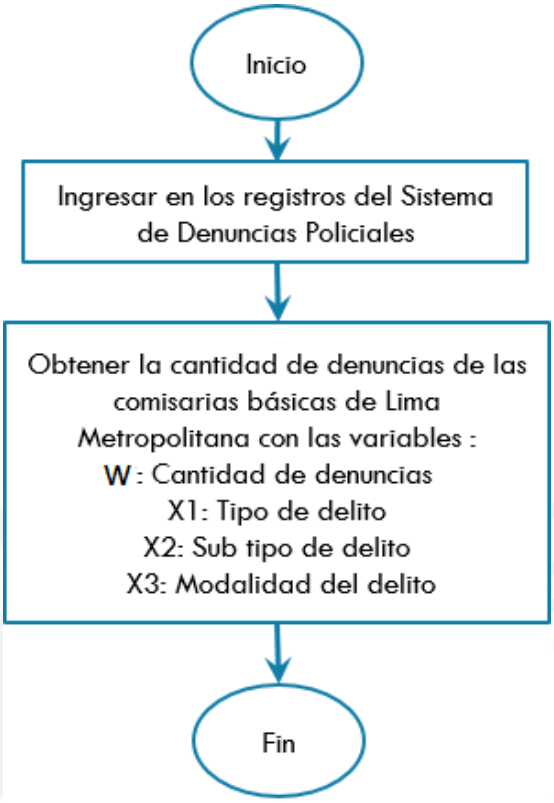


FUENTE: ELABORACIÓN PROPIA

Figura 4: Diagrama para la identificación de las comisarias básicas de Lima Metropolitana

El diagrama de flujo muestra la identificación de las 143 comisarías básicas de Lima Metropolitana para lo cual se ingresa en los registros del Sistema de Denuncias Policiales y se realiza la búsqueda de las comisarias correspondientes (Anexo Figura 12).

3.5.3. Obtener información estructurada sobre las variables w, x1, x2, x3

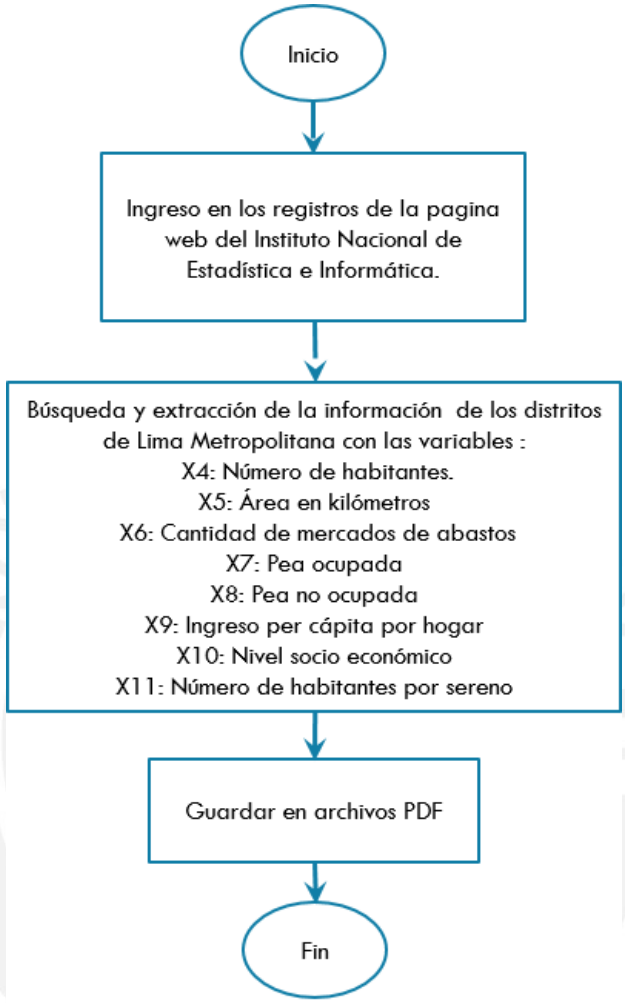


FUENTE: ELABORACIÓN PROPIA

Figura 5: Diagrama para la obtención de la cantidad de denuncias por delitos en las comisarías de Lima Metropolitana

El diagrama muestra la obtención de la cantidad de denuncias por delitos en las comisarías de Lima Metropolitana que fueron identificadas en la sección (3.5.2), para lo cual se ingresa al registro del Sistema de Denuncias Policiales y se realiza un filtrado de la información de Lima Metropolitana con las variables W, X1, X2, X3 (Anexo Figura 13).

3.5.4. Obtener archivos en formato PDF

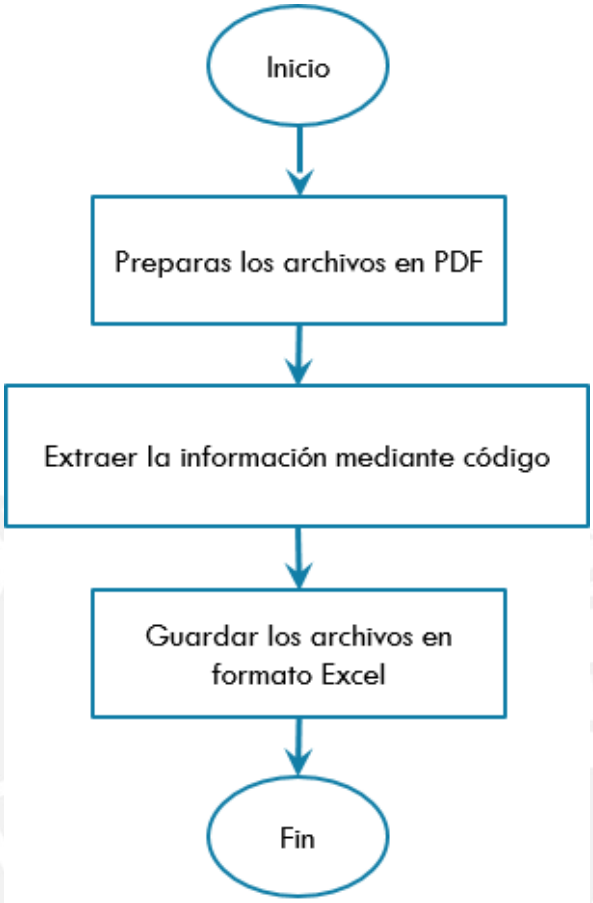


FUENTE: ELABORACIÓN PROPIA

Figura 6: Diagrama para la obtención de información del INEI

El diagrama muestra la forma de obtener datos cuya fuente es el INEI, los cuales se descargan y se guardan en formatos PDF, para lo cual se ingresa a la página web del INEI en la cual se realiza la búsqueda de información con las variables: Número de habitantes (X4), Área en kilómetros (X5), Cantidad de mercados de abastos (X6), Pea ocupada (X7), Pea no ocupada (X8), Ingreso per cápita por hogar(X9), Nivel socio económico (X10), Número de habitantes por sereno (X11); de los distritos de Lima Metropolitana, posteriormente se descargan y guardan los archivos en formato PDF (Anexo Figura 14).

3.5.5. Extraer la información del INEI

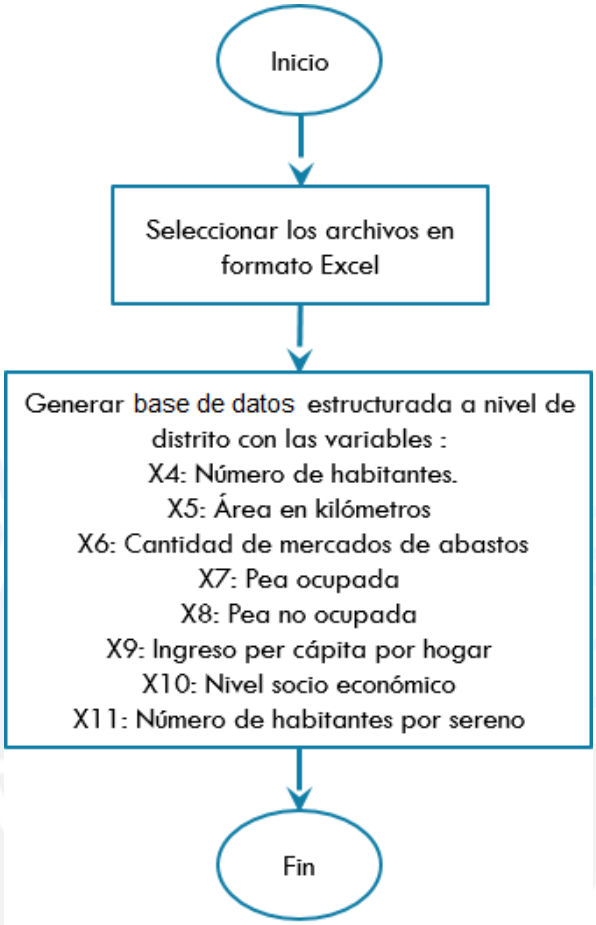


FUENTE: ELABORACIÓN PROPIA

Figura 7: Diagrama para la extracción de información del INEI mediante scraping

El diagrama muestra la extracción de información del INEI, para lo cual se utiliza los documentos en PDF que se obtuvieron en la sección (3.5.4) a la cual se le aplica un código en PYTHON con el procedimiento de scraping y se guarda en un archivo de formato Excel (Anexo Figura 15).

3.5.6. Obtener información estructurada sobre las variables x4, x5, x6, x7, x8, x9, x10, x11

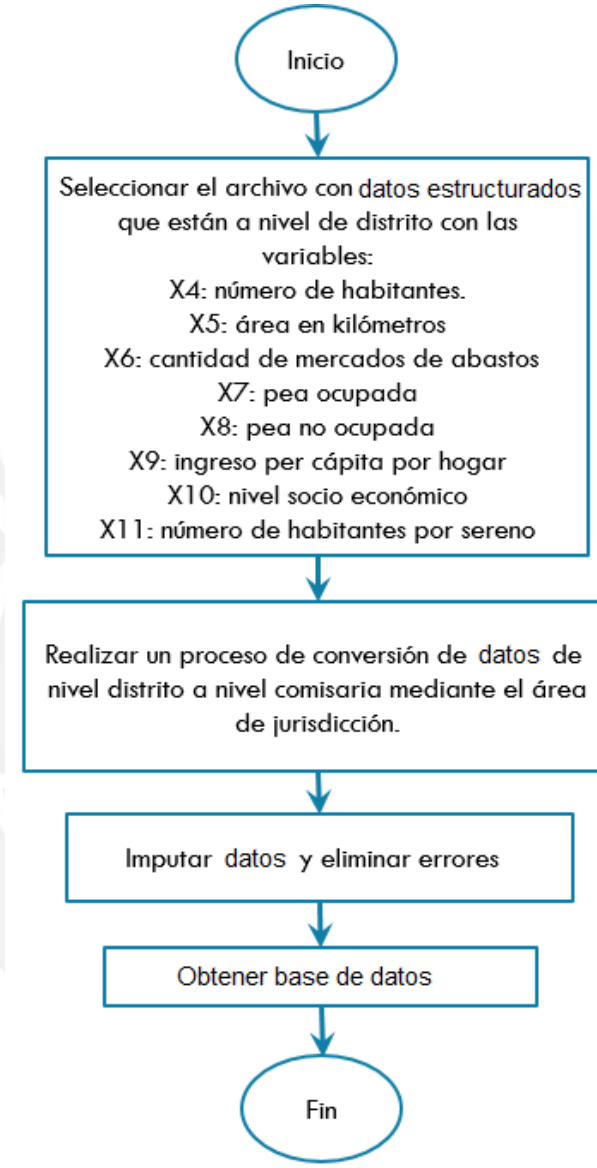


FUENTE: ELABORACIÓN PROPIA

Figura 8: Diagrama para la obtención de información estructurada

El diagrama muestra la obtención de información estructurada mediante la aplicación de un código, para lo cual se cuenta con el archivo en formato Excel que se obtuvo en la sección (3.5.5) al cual se aplica códigos con lo cual se genera una información estructurada de las variables: Número de habitantes (X4), Área en kilómetros (X5), Cantidad de mercados de abastos (X6), Pea ocupada (X7), Pea no ocupada (X8), Ingreso per cápita por hogar(X9), Nivel socio económico (X10), Número de habitantes por sereno (X11) (Anexo Figuras 16, 18, 20, 22 y 24).

3.5.7. Transformar la información de las variables x4, x5, x6, x7, x8, x9, x10, x11 de nivel distrito a nivel comisaria



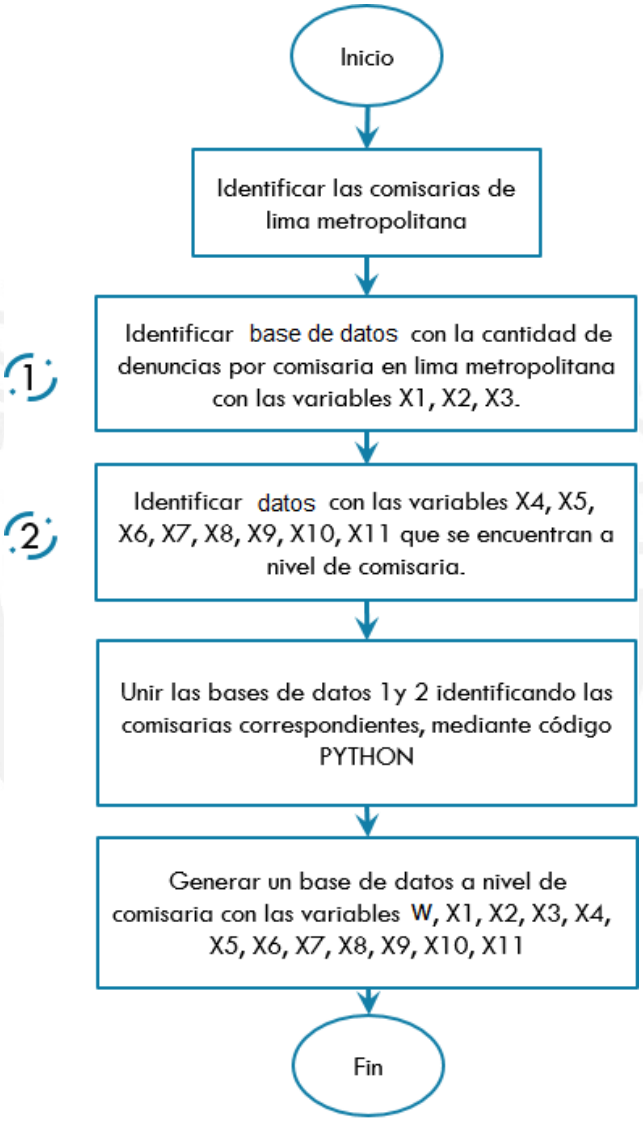
FUENTE: ELABORACIÓN PROPIA

Figura 9: Diagrama para la transformación de información de variables de nivel distrito a nivel comisaria

El diagrama muestra la transformación de la información de las variables de nivel distrito a nivel comisaria, para lo cual se cuenta con el archivo en formato Excel que se obtuvo en la sección (3.5.6), al cual se le aplica un código el cual realiza el proceso de conversión de los datos de nivel distrito a nivel comisaria mediante el

área de jurisdicción, para posteriormente realizar la imputación de datos y obtener una base de datos (Anexo Figura 26).

3.5.8. Generar una base de datos de la cantidad de denuncias en las comisarías de Lima Metropolitana con las variables x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11

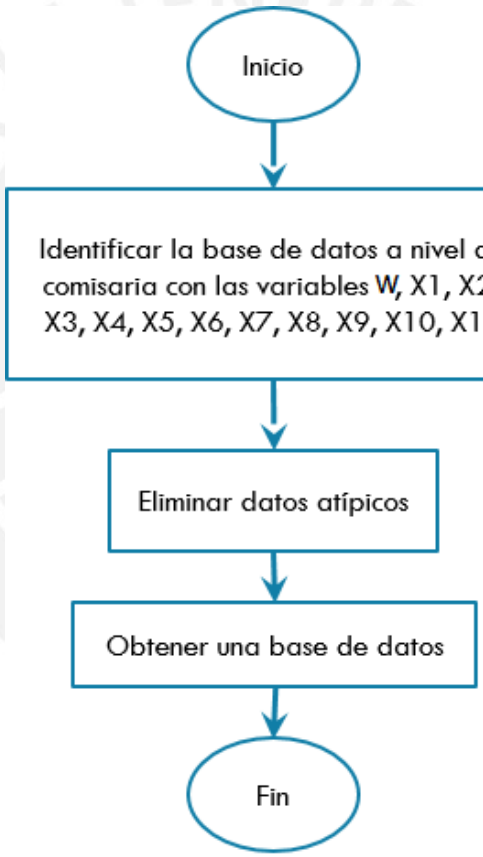


FUENTE: ELABORACIÓN PROPIA

Figura 10: Diagrama para la unión de bases de datos de la cantidad de denuncias de Lima Metropolitana

El diagrama muestra la unión de las bases de datos, para lo cual se identifica las comisarías de Lima Metropolitana, se identificar las bases de datos de la cantidad de denuncias por comisaria en Lima Metropolitana con las variables W, X1, X2, X3 que se obtuvo en la sección (3.5.4), también identifica la base de datos con las variables X4, X5, X6, X7, X8, X9, X10, X11 que se obtuvo en la sección (3.5.8) con lo cual se une las bases de datos anteriormente identificadas a través de un código en PYTHON, con lo cual se obtiene una base de datos a nivel comisaria con las variables W, X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11 (Anexo Figura 27).

3.5.9. Eliminar los datos atípicos a la base de datos con las variables w, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11



FUENTE: ELABORACIÓN PROPIA

Figura 11: Diagrama para la limpieza de bases de datos

El diagrama muestra el proceso de limpieza de la base de datos, para lo cual se identifica la base de datos que se obtuvo en la sección (3.5.8) a la cual se aplica

un código para la eliminación de datos atípicos con lo cual se obtiene una base de datos final (Anexo Figura 28).

3.5.10. Mostrar base de datos final

Se muestra la base de datos final (Anexo Figura 29).

4. CAPÍTULO IV: ANÁLISIS Y EXPLICACIÓN DE LAS VARIABLES EN ESTUDIO E IDENTIFICACIÓN DE LAS TÉCNICAS X,Y,Z

4.1. Análisis y explicación de la variable dependiente

Cantidad de denuncias de delitos por comisaria (W): Representa la cantidad de denuncias que se registra por comisaria en el SIDPOL.

4.2. Análisis y explicación de la variables independientes

- Tipo de denuncia (X1): Es el tipo de denuncia según el código procesal civil y penal.
- Sub tipo de denuncia (X2): Es el subtipo de denuncia que se encuentra dentro de una denuncia
- Modalidad (X3): Es la modalidad por la cual se generó el subtipo de denuncia
- Número de habitantes (X4): Es la cantidad de habitantes, en la jurisdicción de una comisaría
- Área en kilómetros (X5): Representa el área en kilómetros de una jurisdicción de una comisaría.
- Cantidad de mercados de abastos (X6): Representa la cantidad de mercados de abastos dentro de una jurisdicción de una comisaría.
- Pea ocupada (X7): Cantidad de personas que estén dentro de la población económicamente activa ocupada dentro de una jurisdicción de una comisaría
- Pea no ocupada (X8): Cantidad de personas que estén dentro de la población económicamente activa no ocupada dentro de una jurisdicción de una comisaría
- Ingreso per cápita por hogar (X9): Representa el ingreso per cápita por hogar dentro de una jurisdicción de una comisaría
- Nivel socio económico (X10): Representa el estrato que tiene mayor prevalencia dentro de una jurisdicción de una comisaría
- Número de habitantes por sereno (X11): Representa la cantidad de habitantes por sereno dentro de una jurisdicción de una comisaría

4.3. Identificación de las Técnicas

- TECNICA X : La técnica está compuesto por los procedimientos (1): Identificación de las comisarias que se encuentran de Lima Metropolitana y (2): Obtener información sobre la cantidad de denuncias por comisaria en lima metropolitana.

- TECNICA Y : La técnica está compuesto por los procedimientos (3): Obtener información de las distintas variables independientes (Establecer una base de datos estructurada) y (4): Transformar los datos de las variables independientes de nivel distrito a nivel comisaria mediante el área la jurisdicción e imputación de errores.
- TECNICA Z : La técnica está compuesto por los procedimientos (5): Identificar cada comisaria con las variables independientes correspondientes (Generar una base de datos de la cantidad de denuncias de Lima Metropolitana), (6): Eliminar datos atípicos y (7): Obtención de la base de datos lista.

5. CAPÍTULO V: CONTRASTE E INTERPRETACIÓN DE RESULTADOS

5.1. Contraste de las hipótesis especificas

5.1.1. Contraste de las hipótesis especifica 1

Planteo de hipotesis:

- Ho: La aplicación del método de SCRAPING para obtener información sobre el número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, en los distritos de Lima Metropolitana, genera la información deseada
- H1: La aplicación del método de SCRAPING para obtener información sobre el número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, en los distritos de Lima Metropolitana, NO genera la información deseada.

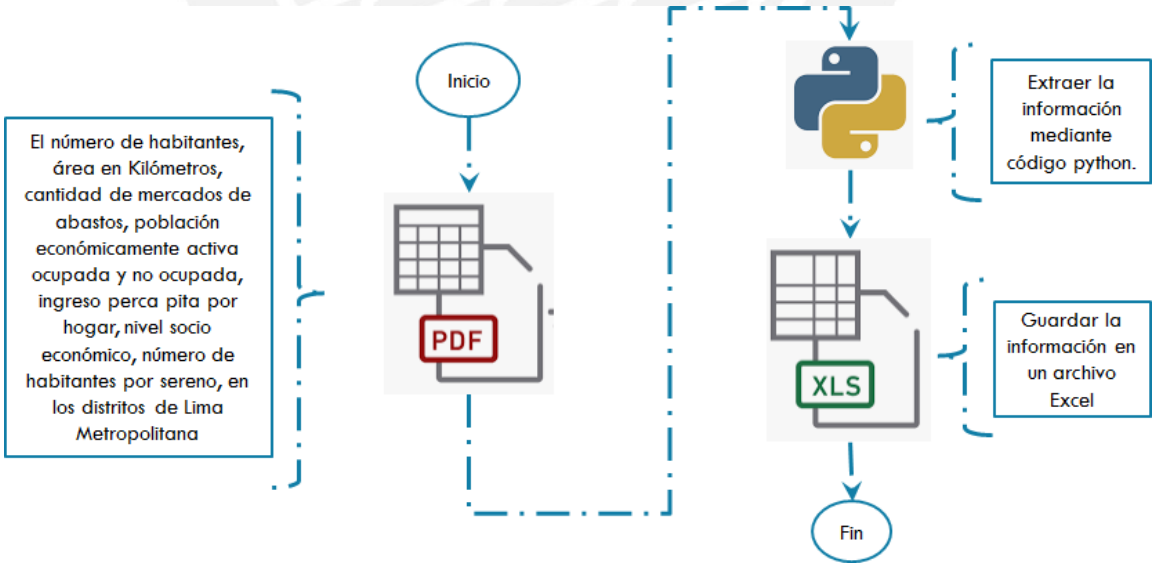
Análisis:

Para que se cumpla la hipótesis se debe cumplir las condiciones:

(1). Obtener un archivo en formato PDF con información requerida.

(2). Extraer la información mediante código en PYTHON.

Según como se muestra en el diagrama:



Los puntos (1) y (2) se obtienen a partir de las secciones 3.5.4 y 3.5.5.

Para lo cual se analiza el nivel de satisfacción de la información obtenida, con el fin de medir si la información es adecuada o no, los resultados son:

SCRAPING			
ns X4	: 95%	te X4	: 5%
ns X5	: 90%	te X5	: 10%
ns X6	: 90%	te X6	: 10%
ns X7	: 85%	te X7	: 15%
ns X8	: 85%	te X8	: 15%
ns X9	: 95%	te X9	: 5%
ns X10	: 90%	te X10	: 10%
ns X11	: 85%	te X11	: 15%

Tabla 3: Nivel de satisfacción y tolerancia al error de las variables X4, X5, X6, X7, X8, X9, X10 y X11 en la aplicación del SCRAPING

Según lo planteado en la hipótesis nula conllevaría a que el nivel de satisfacción de la información obtenida es mayor al 90 % (La tolerancia al error del nivel de satisfacción es menor a 10 %), por lo que las hipótesis en términos de tolerancia al error estarían planteadas como:

- Ho: El promedio de la tolerancia del error del nivel de satisfacción es menor o igual al 10 %.
- H1: El promedio de la tolerancia del error del nivel de satisfacción es mayor al 10 %.

Alfa=0.05

Estadístico de prueba t stutent.

$T = \frac{(\bar{x}-u)\cdot\sqrt{n}}{\sigma} \sim t - Student(n-1)$

Los resultados de la Tabla 3 son:

$\bar{x}$= 10.6 %; σ= 4.17 %; n=8; T tabla= 1.8946; T calculado= 0.424

Dado que T calculado es menor al T tablas entonces la hipótesis nula es aceptada. Por lo que el promedio de la tolerancia del error del nivel de satisfacción es menor o igual al 10 %, con lo que a un nivel de confianza del 95 % existe evidencia estadística suficiente para afirmar que obtener un archivo en formato PDF con información requerida y extraer la información mediante código en PYTHON generan resultados favorables.

Conclusión:

Dado que los condiciones (1) y (2) se cumplen con lo cual la hipótesis nula es aceptada con lo cual se puede afirmar que la aplicación del método de SCRAPING genera buenos resultados en la extracción de la información, por lo tanto, la hipótesis planteada es aceptada.

5.1.2. Contraste de las hipótesis especifica 2

Planteo de hipotesis:

- Ho: La aplicación de las técnicas X, Y, Z en el preprocesamiento de la información extraída de los distritos de Lima Metropolitana, genera una información adecuada para la aplicación del método de regresión.
- H1: La aplicación de las técnicas X, Y, Z en el preprocesamiento de la información extraída de los distritos de Lima Metropolitana, NO genera una información adecuada para la aplicación del método de regresión.

Análisis:

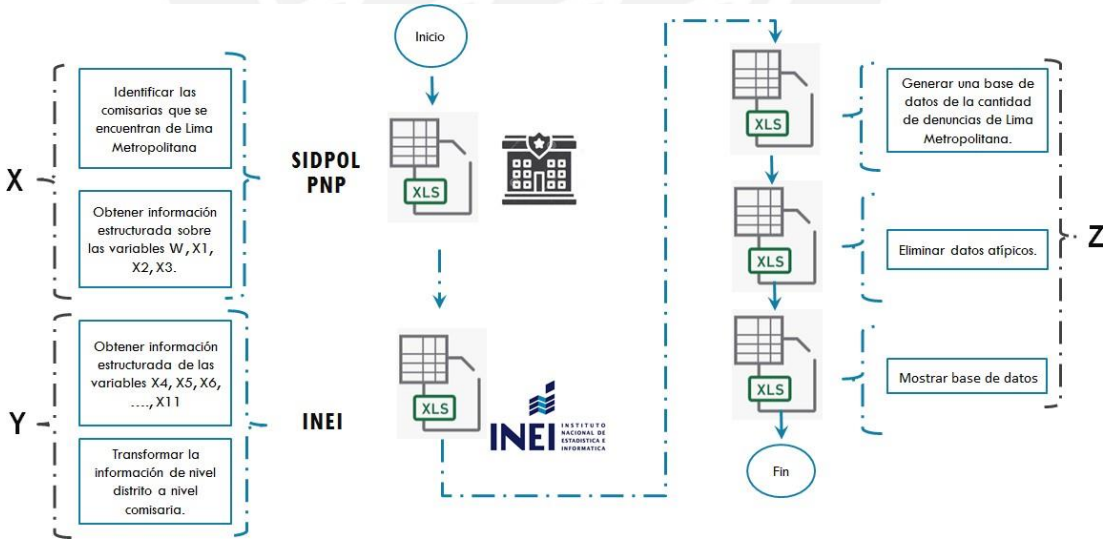
Para que se cumpla la hipótesis se debe cumplir las condiciones:

(1). La TECNICA X genera resultados favorables.

(2). La TECNICA Y genera resultados favorables.

(3). La TECNICA Z genera resultados favorables.

Según como se muestra en el diagrama:



Los puntos (1), (2) y (3) se obtienen a partir de las secciones 3.5.2, 3.5.3, 3.5.6, 3.5.7, 3.5.8, 3.5.9 y 3.5.10.

Para lo cual se analiza el nivel de satisfacción de la información obtenida, con el fin de medir si la información es adecuada o no, los resultados son:

Técnica X				Técnica Y				Técnica Z			
ns X1 : 100%	te X1 : 0%			ns X4 : 90%	te X4 : 10%			ns : 95%	te : 5%		
ns X2 : 100%	te X2 : 0%			ns X5 : 80%	te X5 : 20%						
ns X3 : 100%	te X3 : 0%			ns X6 : 95%	te X6 : 5%						
ns W : 100%	te W : 0%			ns X7 : 95%	te X7 : 5%						
				ns X8 : 80%	te X8 : 20%						
				ns X9 : 95%	te X9 : 5%						
				ns X10 : 85%	te X10 : 15%						
				ns X11 : 90%	te X11 : 10%						

Tabla 4: Nivel de satisfacción y tolerancia al error de las técnicas X, Y, Z

Según lo planteado en la hipótesis nula conllevaría a que el nivel de satisfacción de la información obtenida es mayor al 90 % (La tolerancia al error del nivel de satisfacción es menor a 10 %), por lo que las hipótesis en términos de tolerancia al error estarían planteadas como:

- Ho: El promedio de la tolerancia del error del nivel de satisfacción es menor o igual al 10 %.
- H1: El promedio de la tolerancia del error del nivel de satisfacción es mayor al 10 %.

Alfa=0.05

Estadístico de prueba t stutent.

$$T = \frac{(\bar{x}-u)\cdot\sqrt{n}}{\sigma} \sim t - Student(n-1)$$

Los resultados de la Tabla 4 son:

$\bar{x}$= 7.3 %; σ= 7.25 %; n=13; T tabla= 1.7171; T calculado= -1.339

Dado que T calculado es menor al T tablas entonces la hipótesis nula es aceptada. Por lo que el promedio de la tolerancia del error del nivel de satisfacción es menor o igual al 10 %, con lo que a un nivel de confianza del 95 % existe evidencia estadística suficiente para afirmar que las Técnica X, Y, Z. generan resultados favorables.

Conclusión:

Dado que las condiciones (1), (2), (3) se cumplen con lo cual la hipótesis es aceptada, con lo cual se puede afirmar que la aplicación de las técnicas X, Y, Z en el preprocesamiento de la información extraída genera una información adecuada para la aplicación del método de regresión, por lo tanto, la hipótesis planteada es aceptada.

5.1.3. Contraste de las hipótesis especifica 3

Planteo de hipótesis:

- Ho: En la aplicación de los métodos de Random Forest Regressor y Árbol de decisión de regresión se encuentra que una aplica mejor para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.
- H1: En la aplicación de los métodos de Random Forest Regressor y Árbol de decisión de regresión NO se encuentra que una aplica mejor para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

Alfa=0.05

Análisis:

Para que se cumpla la hipótesis se debe cumplir las condiciones:

(1). Aplicar el método de Random Forest Regressor para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

(2). Aplicar el método de Árbol de decisión regresión para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

(3). Encontrar el mejor método para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

La condición (1) se realiza a partir de la aplicación mediante el código (Anexo Figura 30), se realizó una división de datos 67 % en el entrenamiento y un 33 % para su validación, en el modelo de Random Forest Regressor el R^2 en el entrenamiento es de 0,8841 y en la validación de 0.8205 y los factores más importantes para el modelo son: Modalidad, Ingreso per cápita por hogar, área en kilómetros, Pea ocupada y cantidad de mercados de abastos, asi como se muestra en los gráficos 1 y 2.

```
30 Árboles
C-star: 0.8777107549629888

50 Árboles
C-star: 0.8818099588453366

100 Árboles
C-star: 0.8841541187406778

200 Árboles
C-star: 0.8871146615643278

500 Árboles
C-star: 0.8869185199455374

1000 Árboles
C-star: 0.886997713921823

2000 Árboles
C-star: 0.887053808863177
```

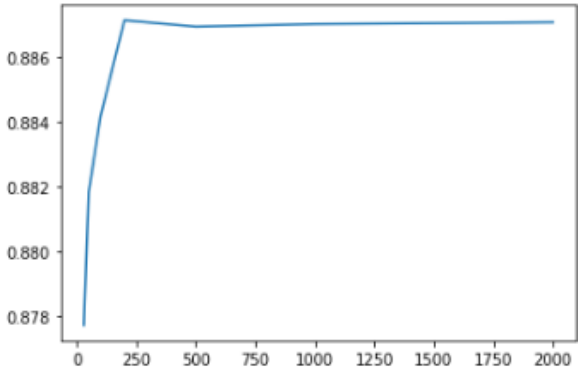


Gráfico 1 :Precisión para el entrenamiento de los datos según el número de estimadores mediante el modelo de Random Forest Regressor.

En el gráfico 1 se muestra los valores de la precisión del entrenamiento de los datos según el número de árboles, a medida que la cantidad de árboles se incrementa la precisión también lo hace, cuando el número de árboles se encuentra entre 200 y 500 la precisión encuentra su pico más alto con un valor superior a 0.887

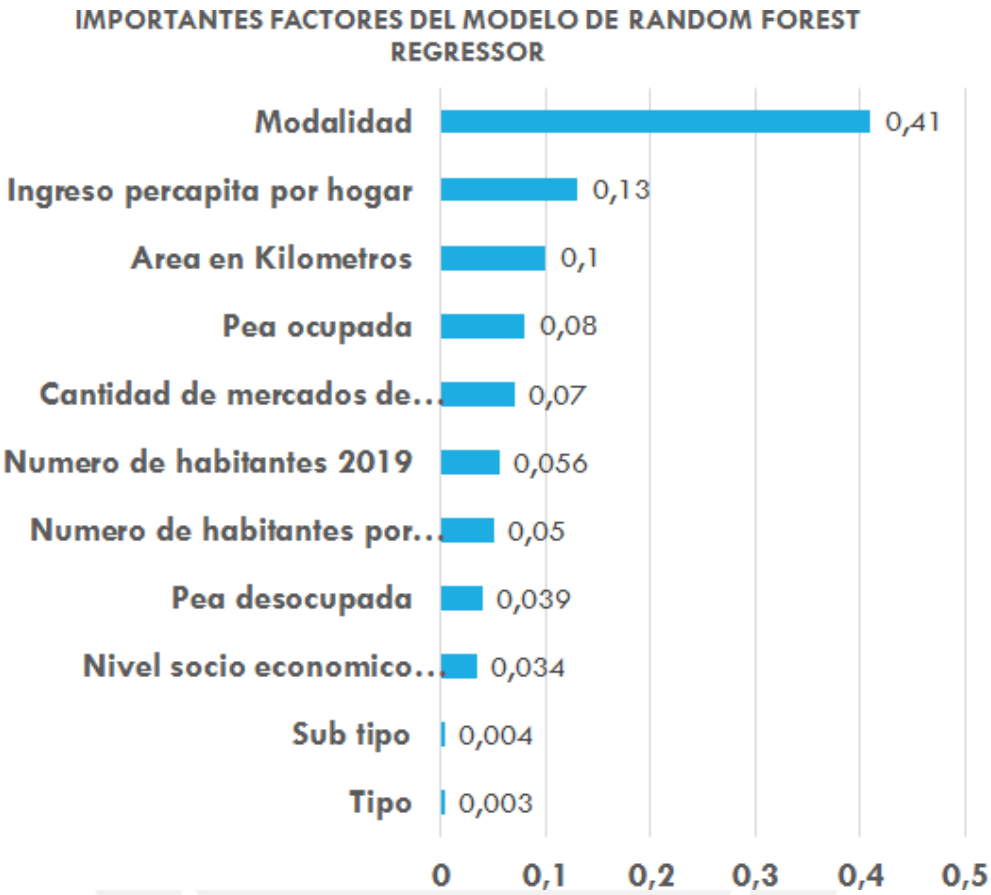


Gráfico 2: Factores del modelo de Random Forest Regressor.

Los resultados de la predicción para los 1701 valores de la validación de los datos son:

w_validación	w_predicción
78	71.89
2	2.16
3	1.7684772727272726
87	105.7
2	4.5
13	11.32
2	2.6746269841269843
9	9.545
2	2.97
65	65.59
4	4.97
2	3.2145115440115433
3	5.2996689976689995
10	25.44214285714286
13	8.267142857142858
1	1.0
96	314.6532142857143
1	1.29
5	4.76
65	58.4

Cantidad de denuncias de la validación y predicción con el modelo de Random Forest Regressor.

Cuyos parámetros del valor absolutos de la diferencia entre la cantidad de denuncias de la validación y la predicción son: $\bar{d}$ 1= 9.2471; σ1= 28.3921; n=1701, Con lo cual se cumple la condición (1).

La condición (2) se realiza a partir de la aplicación mediante el código (Anexo Figura 31), Se realizó una división de datos 67 % en el entrenamiento y un 33 % para su validación. En el modelo de Árbol de decisión regresión el R^2 en el entrenamiento es de 0,5172 y en la validación de 0.3519 y los factores más importantes para el modelo son: Modalidad, cantidad de mercados de abastos, número de habitantes, pea desocupada y pea ocupada, como se muestra en el gráfico 3.

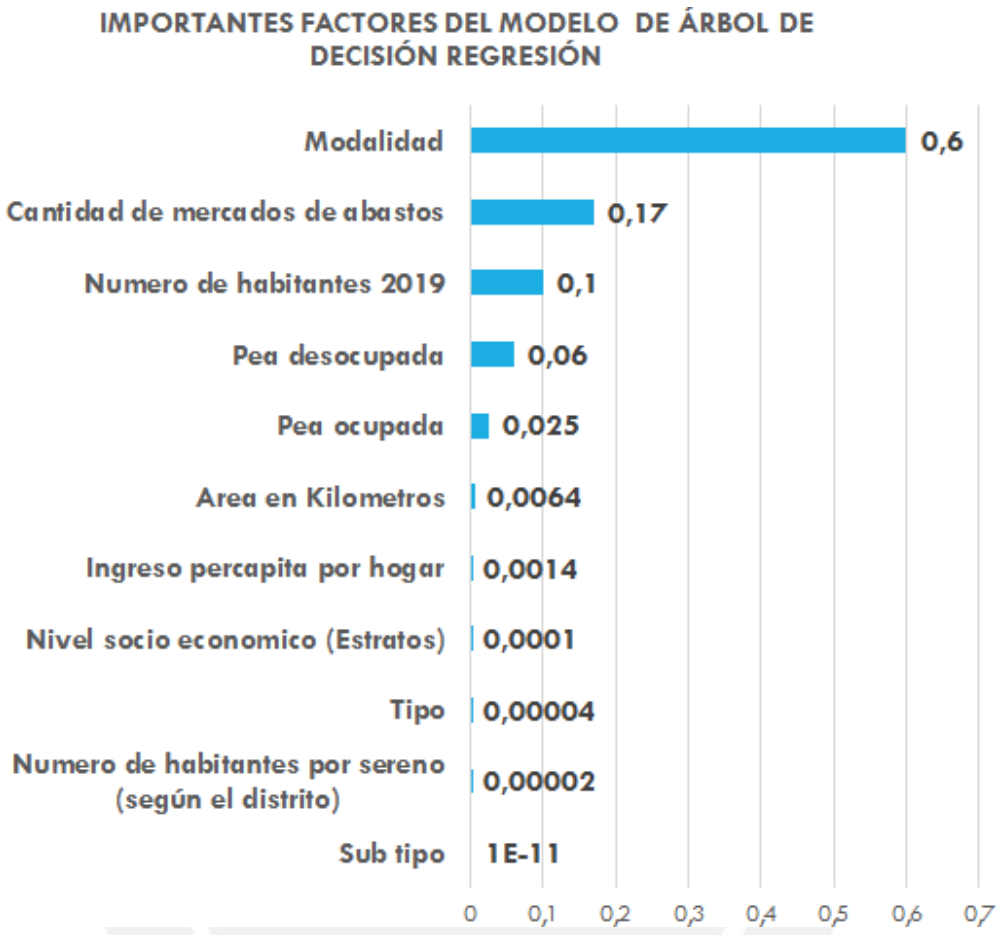


Gráfico 3: Factores del modelo de Árbol de Decisión Regresión.

Los resultados de la predicción para los 1701 valores de la validación de los datos son:

w_validación	w_predicción
96	62.90553745928339
14	13.683498647430117
37	79.29032258064517
15	13.683498647430117
70	324.22727272727275
14	62.90553745928339
1	1.8408408408408408
2	1.8408408408408408
218	103.71428571428571
2	10.571428571428571
8	13.683498647430117
7	5.147058823529412
165	324.22727272727275
21	5.147058823529412
78	22.338709677419356
553	324.22727272727275
15	103.71428571428571
1	2.3118811881188117
11	62.90553745928339
22	6.788461538461538

Cantidad de denuncias de la validación y predicción con el modelo de Árbol de Decisión Regresión.

Cuyos parámetros del valor absoluto de la diferencia entre la cantidad de denuncias de la validación y la predicción (d) son: $\bar{d}$ 2= 20.972; σ2= 43.013; n=1701, Con lo cual se cumple la condición (2).

La condición (3), se cumple a partir de la comparación de los resultados de los pronósticos de los métodos, así como se plantea:

- Ho: El promedio del valor absolutos de la diferencia entre la cantidad de denuncias de la validación y la predicción con el método de Random Forest Regresor es menor o igual que con el método de Árbol de Decisión Regresión. ($\bar{d}$ 1 menor o igual a $\bar{d}$ 2)
- H1: El promedio del valor absolutos de la diferencia entre la cantidad de denuncias de la validación y la predicción con el método de Random Forest Regresor es mayor que con el método de Árbol de Decisión Regresión. ($\bar{d}$ 1 mayor a $\bar{d}$ 2)

Alfa=0.05

Estadístico de prueba Z.

$$Z = \frac{(\bar{d}1 - \bar{d}2)}{\sqrt{\frac{\sigma 1^2 + \sigma 2^2}{n}}} \sim Z - N(0, 1)$$

Donde:

$\bar{d}1$= 9.2471; σ1= 28.3921; n=1701, $\bar{d}2$= 20.972; σ2= 43.013

Z tabla= 1.65; Z calculado= -9.40

Dado que Z calculado es menor a Z de tablas entonces la hipótesis nula es aceptada. Por lo que el promedio del valor absolutos de la diferencia entre la cantidad de denuncias de la validación y la predicción con el método de Random Forest Regresor es menor o igual que con el método de Árbol de Decisión Regresión, con lo que a un nivel de confianza del 95 % existe evidencia estadística suficiente para afirmar que el mejor método para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana es el Random Forest Regresor, con lo cual se cumple la condición (3).

Conclusión:

Dado que las condiciones (1), (2) y (3) se cumplen favorablemente, lo que conlleva que la hipótesis nula sea aceptada, por lo tanto se puede afirmar que en la aplicación de los métodos de Random Forest Regressor y Árbol de Decisión Regresión se encuentra que una aplica mejor para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional del Perú en Lima Metropolitana.

5.2. Contraste de la hipótesis general

Planteo de hipótesis:

- Ho: La aplicación de un modelo de regresión en el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana, genera un buen método para una buena estimación en el pronóstico.
- H1: La aplicación de un modelo de regresión en el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana, NO genera un buen método para una buena estimación en el pronóstico.

Análisis:

Para que se cumpla la hipótesis se debe cumplir las condiciones:

(1). La hipótesis especifica 1 sea aceptada.

(2). La hipótesis especifica 2 sea aceptada.

(3). La hipótesis especifica 3 sea aceptada.

Dado que las hipótesis especificas 1, 2 y 3 son aceptadas, por lo cual la hipótesis general es aceptada.

Conclusión:

Dado que los condiciones (1), (2) y (3) se cumplen con lo cual la hipótesis nula es aceptada, con lo cual se puede afirmar que el modelo de Random Forest Regressor, presentó mejor capacidad de pronóstico que el modelo de Árbol de Decisión Regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

6. CAPÍTULO VI: CONCLUSÌON, DISCUSÍON Y RECOMENDACIONES

6.1. Conclusiones

6.1.1. Conclusión 1

Esta conclusión resulta del análisis de la hipótesis especifica 1. La aplicación del método de SCRAPING genera buenos resultados para obtener información sobre el número de habitantes, área en kilómetros, cantidad de mercados de abastos, población económicamente activa ocupada y no ocupada, ingreso per cápita por hogar, nivel socio económico, número de habitantes por sereno, en los distritos de Lima Metropolitana, genera la información deseada.

6.1.2. Conclusión 2

Esta conclusión resulta de análisis de la hipótesis especifica 2. La aplicación de las técnicas X, Y, Z en el preprocesamiento de la información extraída de los distritos de Lima Metropolitana, genera una información adecuada para la aplicación del método de regresión.

6.1.3. Conclusión 3

Esta conclusión resulta de análisis de la hipótesis especifica 3. En la aplicación de los métodos de Random Forest Regressor y Árbol de Decisión Regresión se encuentra que una aplica mejor para conocer el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

El modelo de Random Forest Regressor presenta un Coeficiente de determinación (R^2) en el entrenamiento de un 0,8841 y en la validación de 0.8205 y los factores más importantes para el modelo son: Modalidad, Ingreso per cápita por hogar, área en kilómetros, población económicamente activa ocupada y cantidad de mercados de abastos, como se muestra en el gráfico 4.

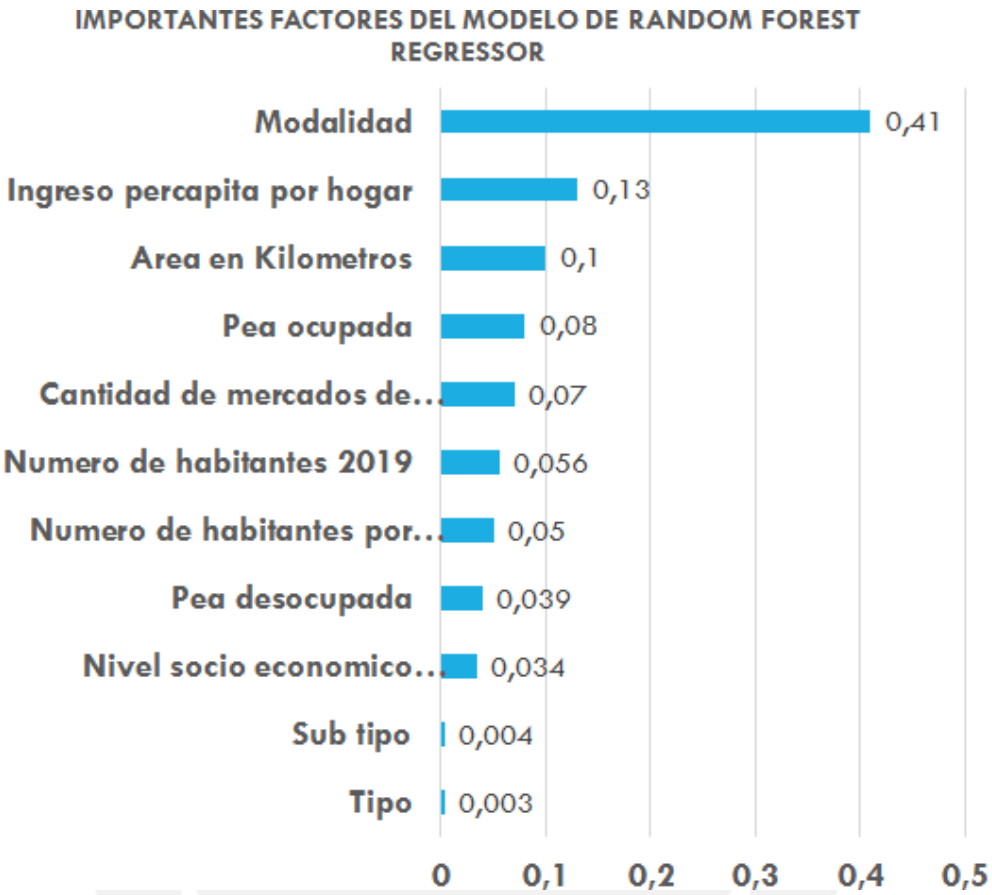


Gráfico 4: Factores del modelo de Random Forest Regressor.

El modelo de Árbol de Decisión Regresión presenta un Coeficiente de determinación (R^2) en el entrenamiento de un 0,5172 y en la validación de 0.3519 y los factores más importantes para el modelo son: Modalidad, cantidad de mercados de abastos, número de habitantes, población económicamente activa desocupada y ocupada, como se muestra el gráfico 5.

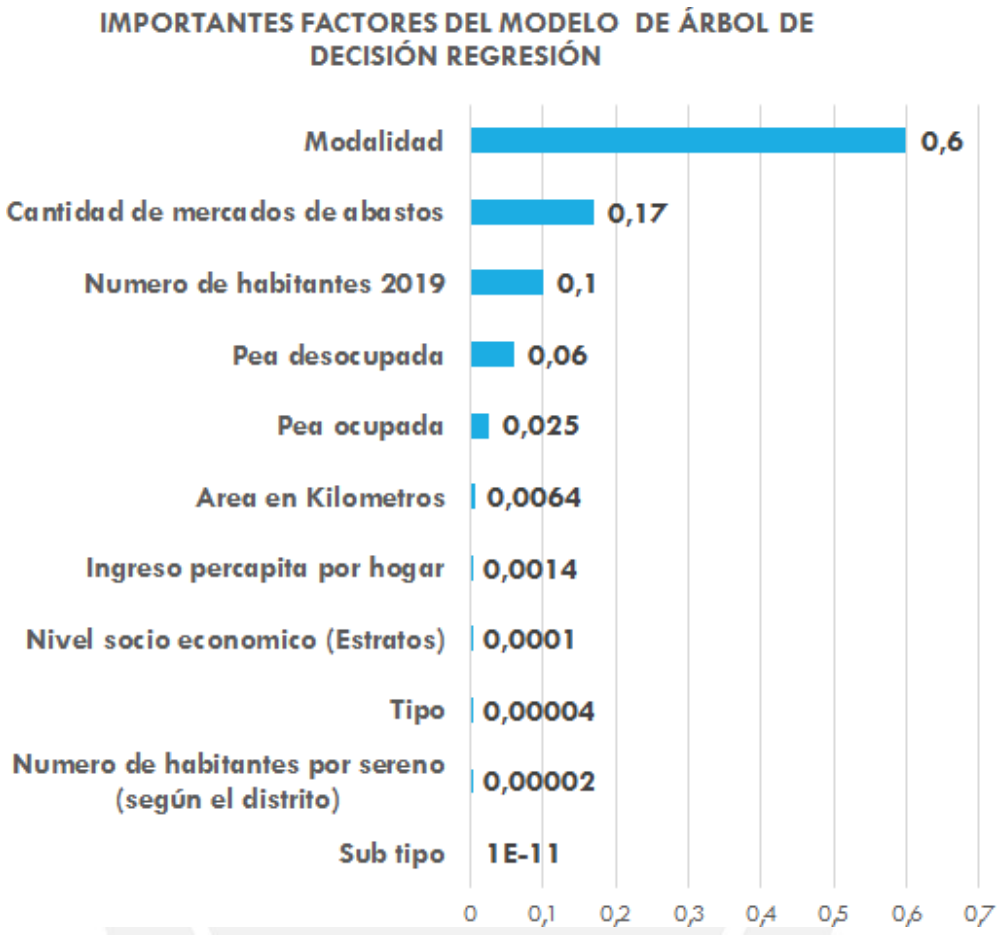


Gráfico 5: Factores del modelo de Árbol de Decisión Regresión.

Por lo que modelo de Random Forest Regressor posee mejor capacidad de pronóstico que el modelo de Árbol de Decisión Regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías en Lima Metropolitana.

Los factores que hacen posible la construcción de los modelos se describen a continuación:

El tipo de denuncia (X1) presenta un comportamiento homogéneo con un datos atípico, lo cual no genera un efecto en la generación de los modelos de regresión, como se muestra en el gráfico 6.

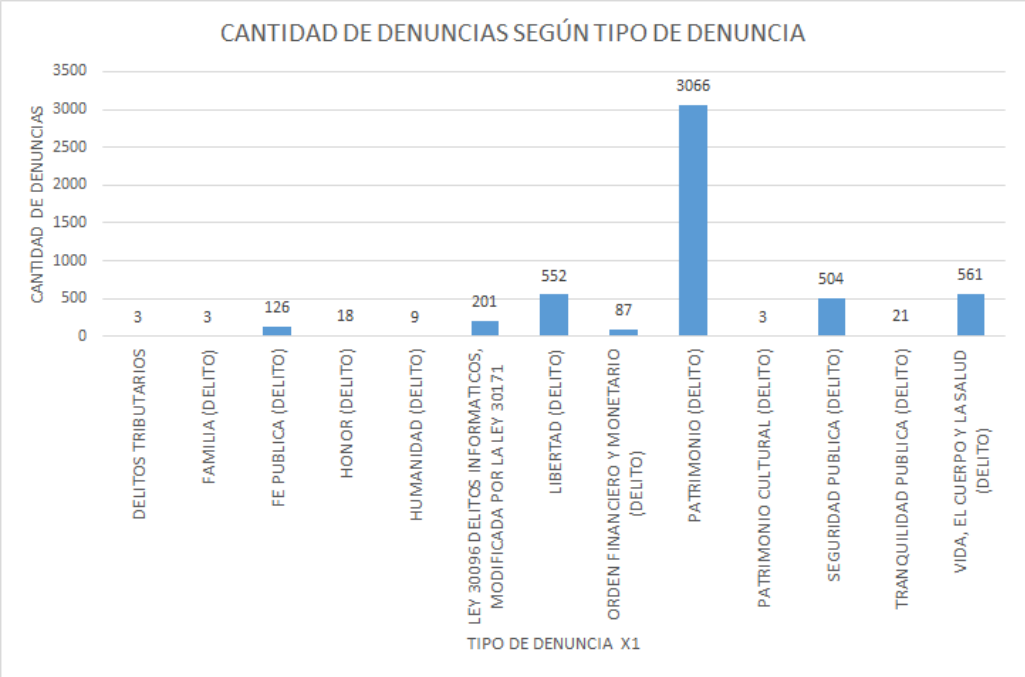


Gráfico 6: Cantidad de denuncias según tipo de denuncia.

El sub tipo de denuncia (X2) presenta un comportamiento homogéneo con dos datos atípicos, lo cual no genera un efecto en la generación de los modelos de regresión, como se muestra en el gráfico 7.

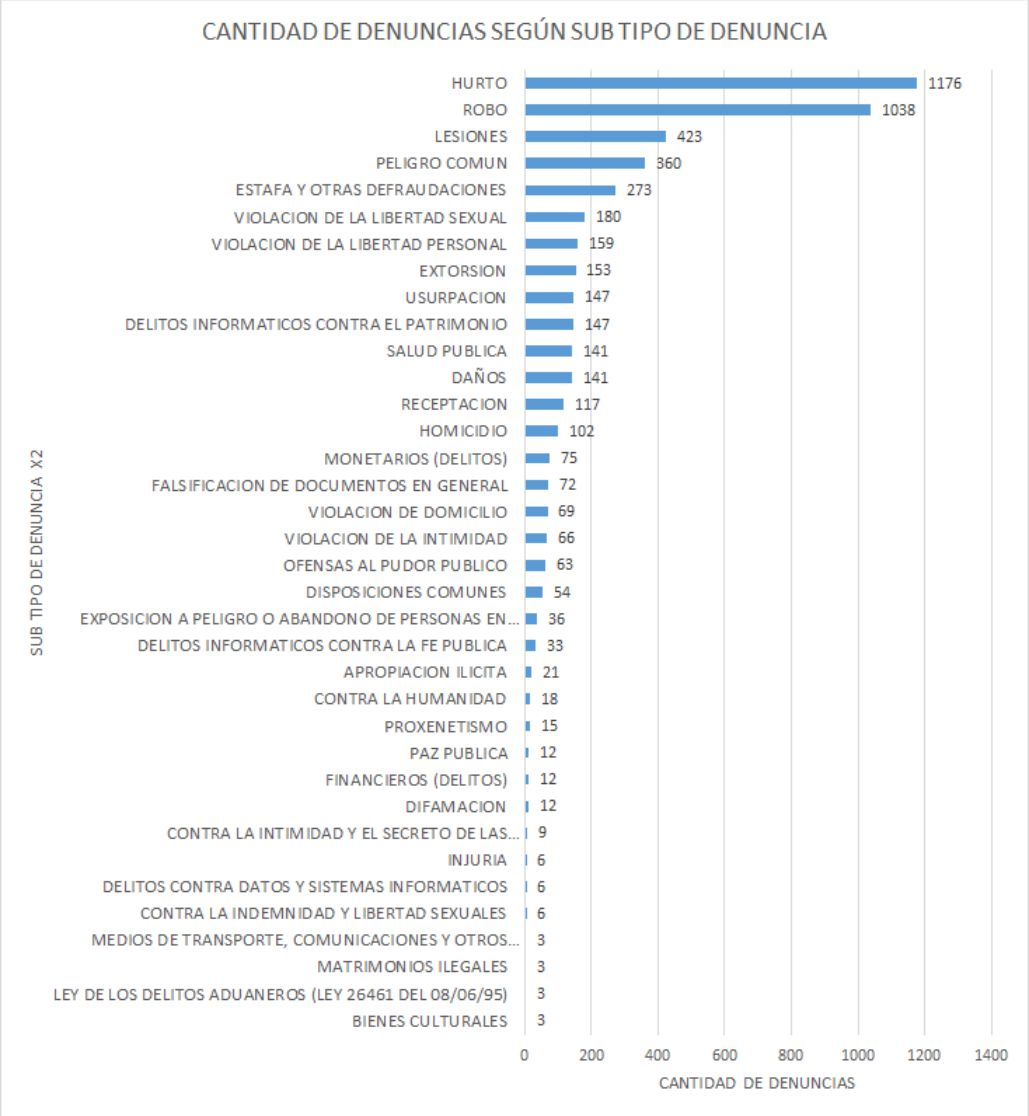


Gráfico 7: Cantidad de denuncias según sub tipo de denuncia.

La modalidad (X3) presenta un comportamiento heterogéneo con tres grupos de datos, lo cual genera un efecto en la generación de los modelos de regresión, como se muestra en el gráfico 8.

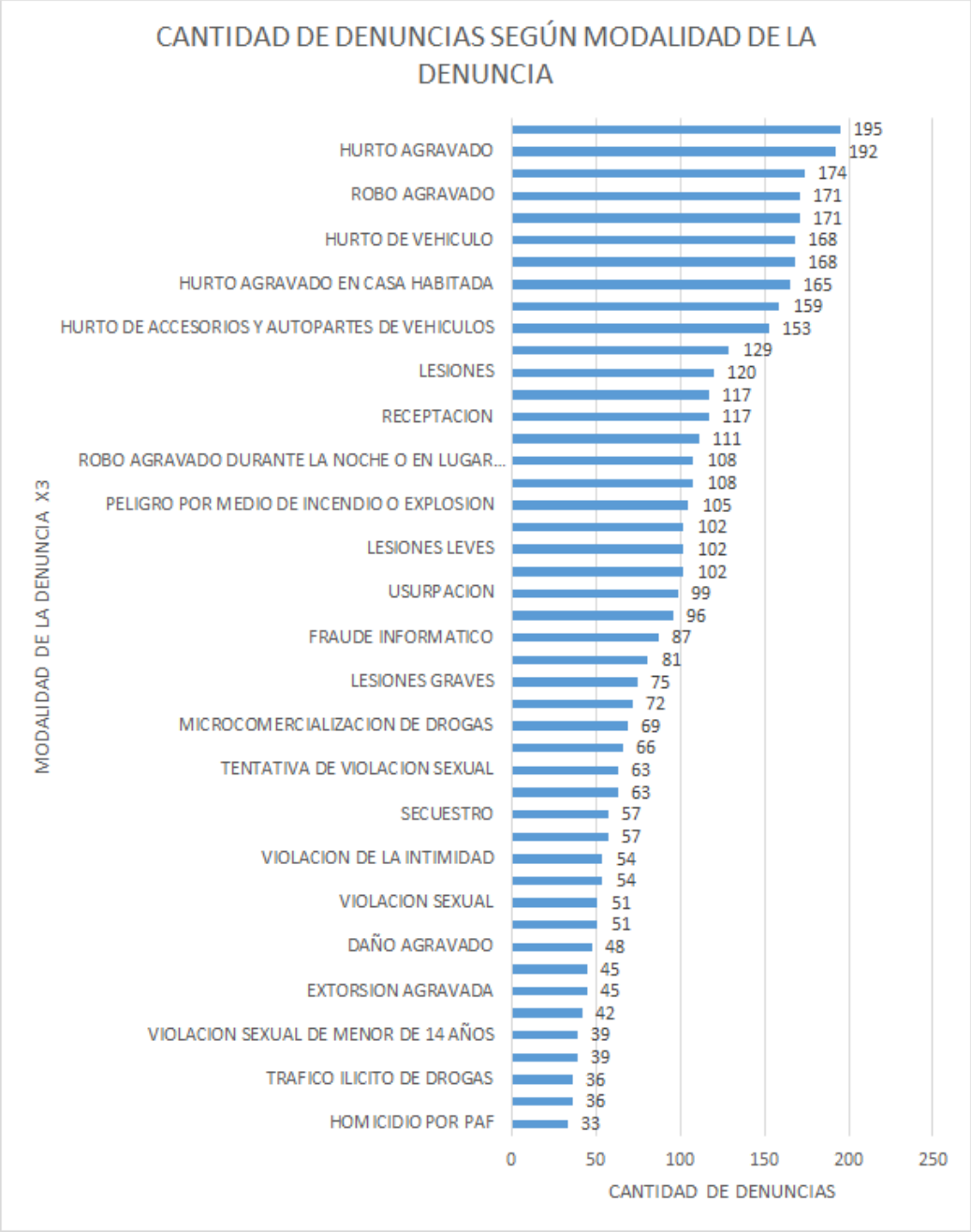


Gráfico 8: Cantidad de denuncias según modalidad de la denuncia.

El número de habitantes (X4) presenta un comportamiento heterogéneo con un grupo de datos concentrados, lo cual genera un efecto en la generación de los modelos de regresión, como se muestra en el gráfico 9.

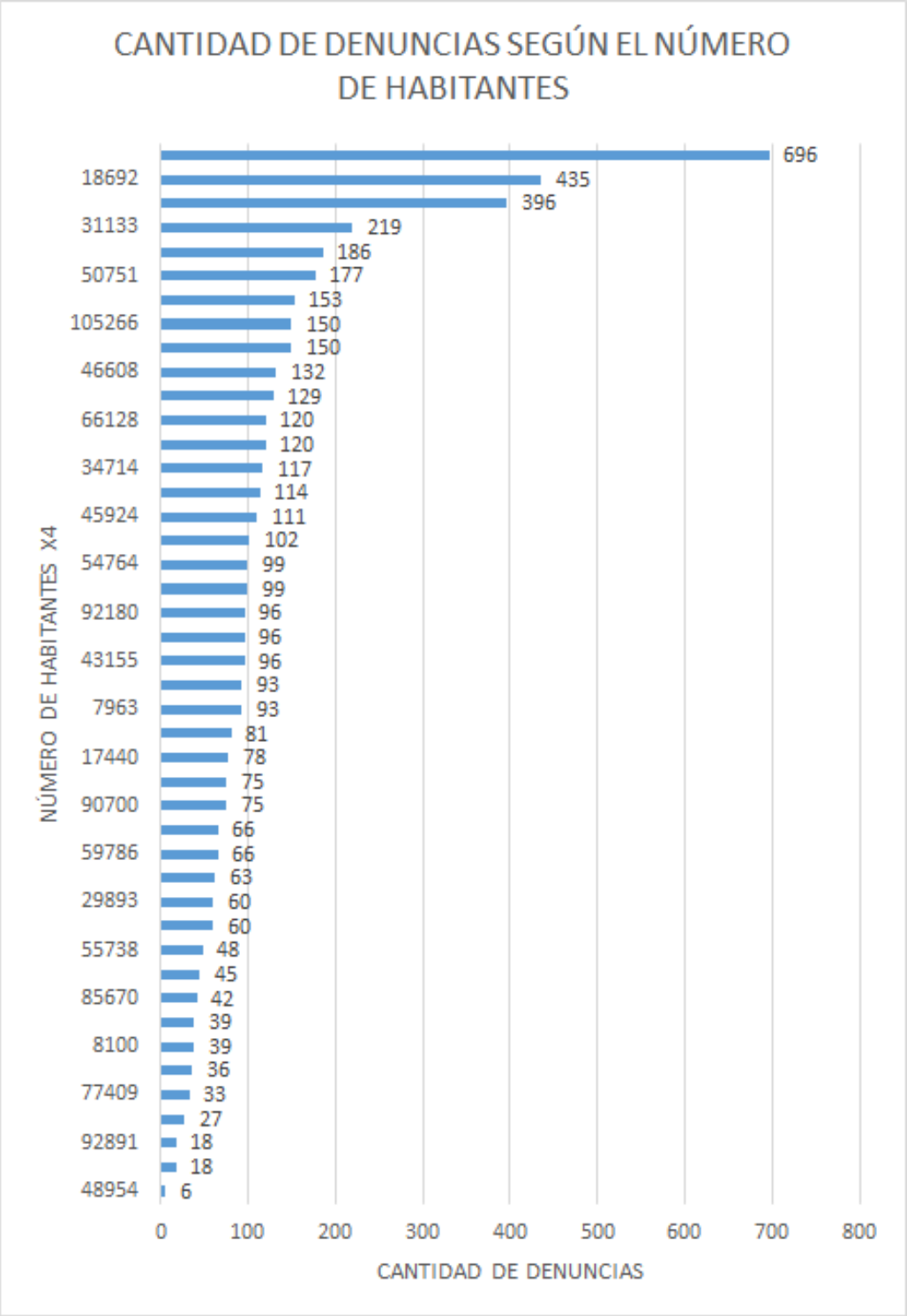


Gráfico 9: Cantidad de denuncias según número de habitantes.

Área en kilómetros (X5) presenta un comportamiento homogéneo con tres datos atípicos, lo cual genera poco efecto en la generación de los modelos de regresión, como se muestra en el gráfico 10.

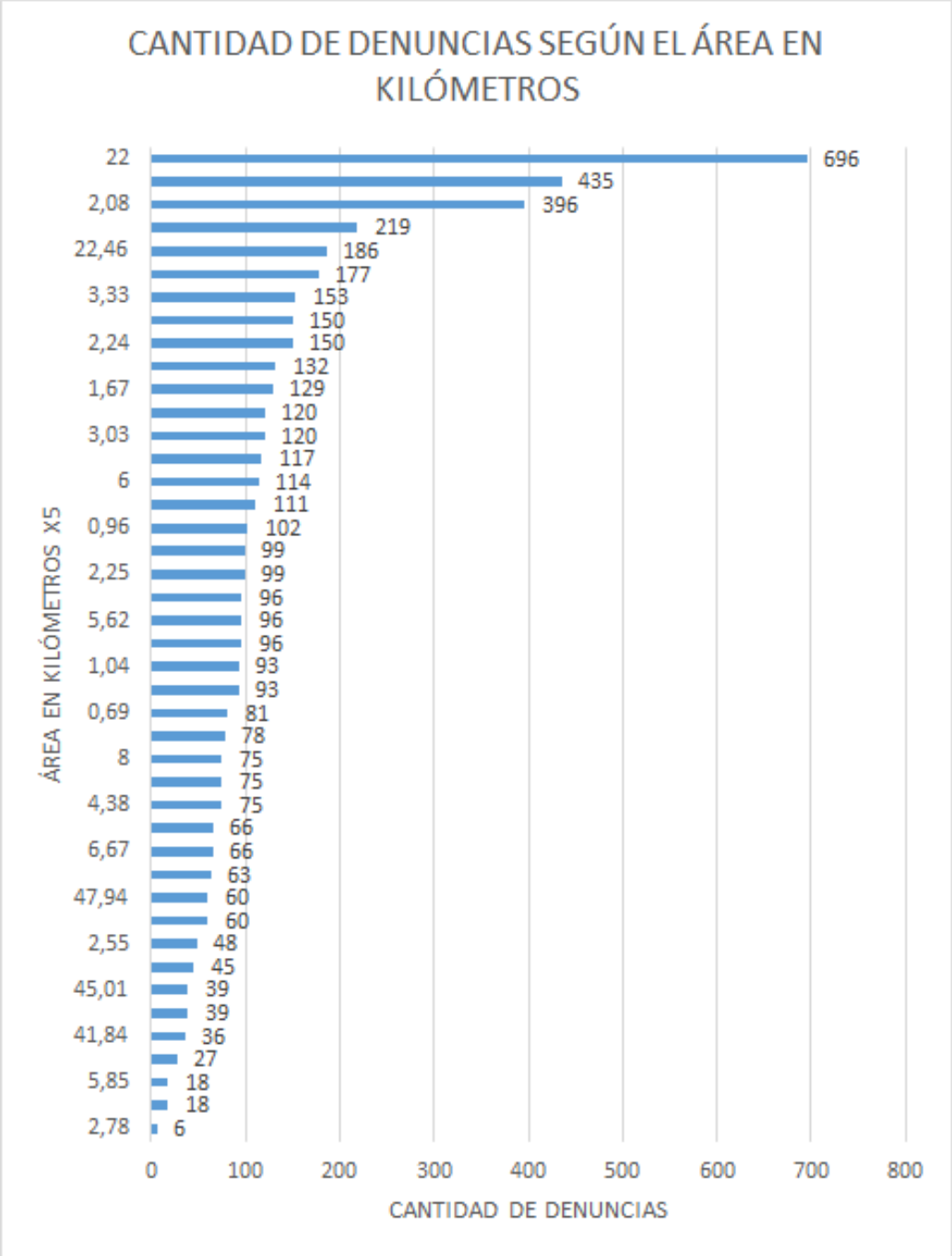


Gráfico 10: Cantidad de denuncias según el área en kilómetros.

Cantidad de mercados de abastos (X6) presenta un comportamiento heterogéneo, lo cual genera efecto en la generación de los modelos de regresión, como se muestra en el gráfico 11.

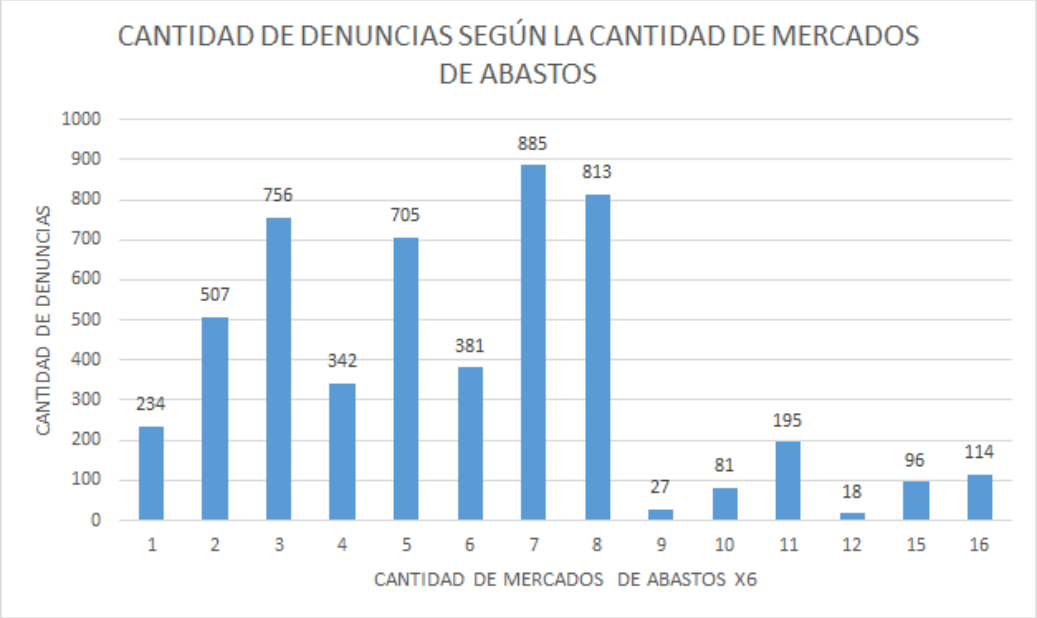


Gráfico 11: Cantidad de denuncias según la cantidad de mercados de abastos.

Pea ocupada (X7) presenta un comportamiento homogéneo con tres datos atípicos, lo cual genera poco efecto en la generación de los modelos de regresión, como se muestra en el gráfico 12.

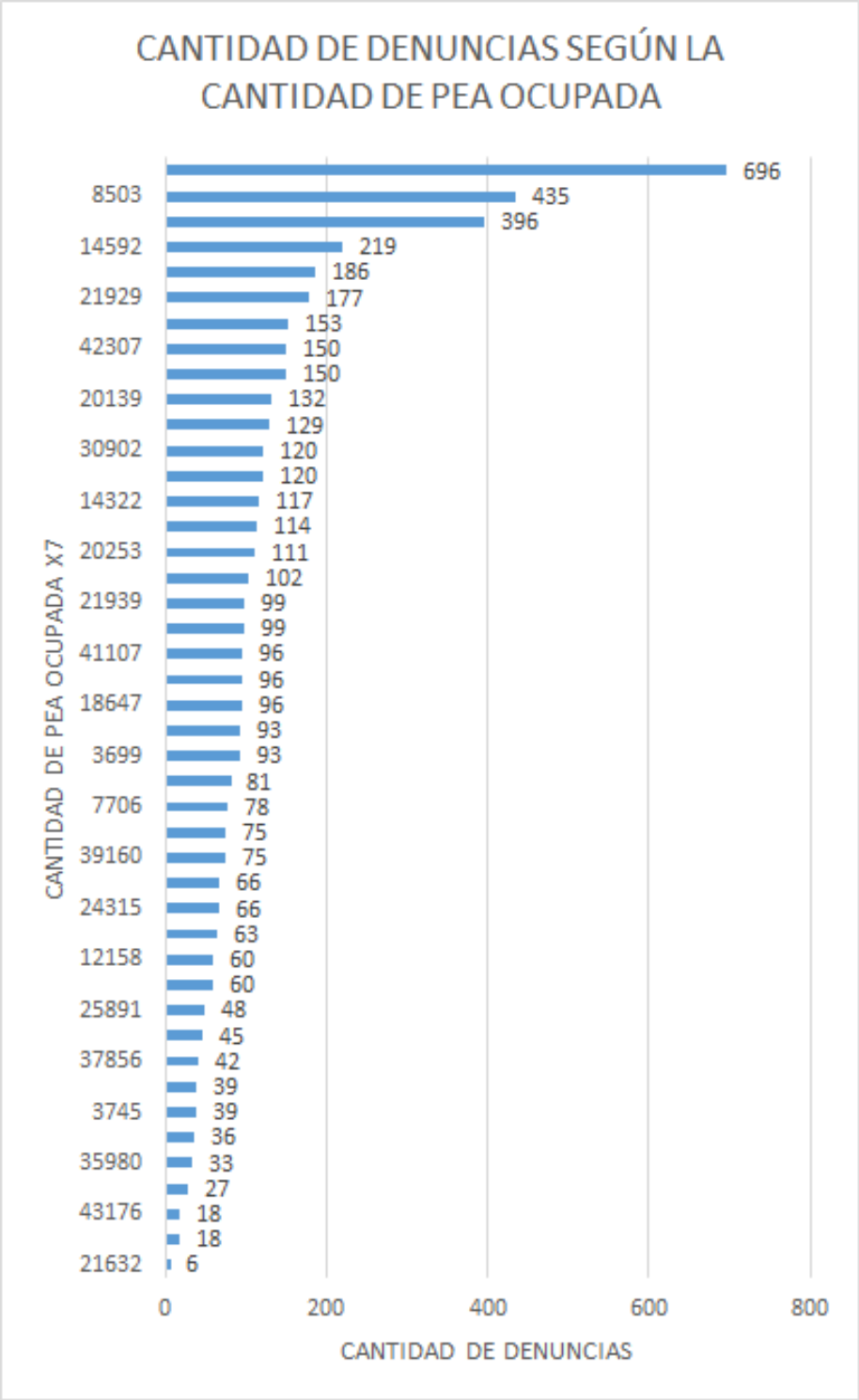


Gráfico 12: Cantidad de denuncias según la cantidad de PEA ocupada.

Pea no ocupada (X8) presenta un comportamiento homogéneo con tres datos atípicos, lo cual genera poco efecto en la generación de los modelos de regresión, como se muestra en el gráfico 13.

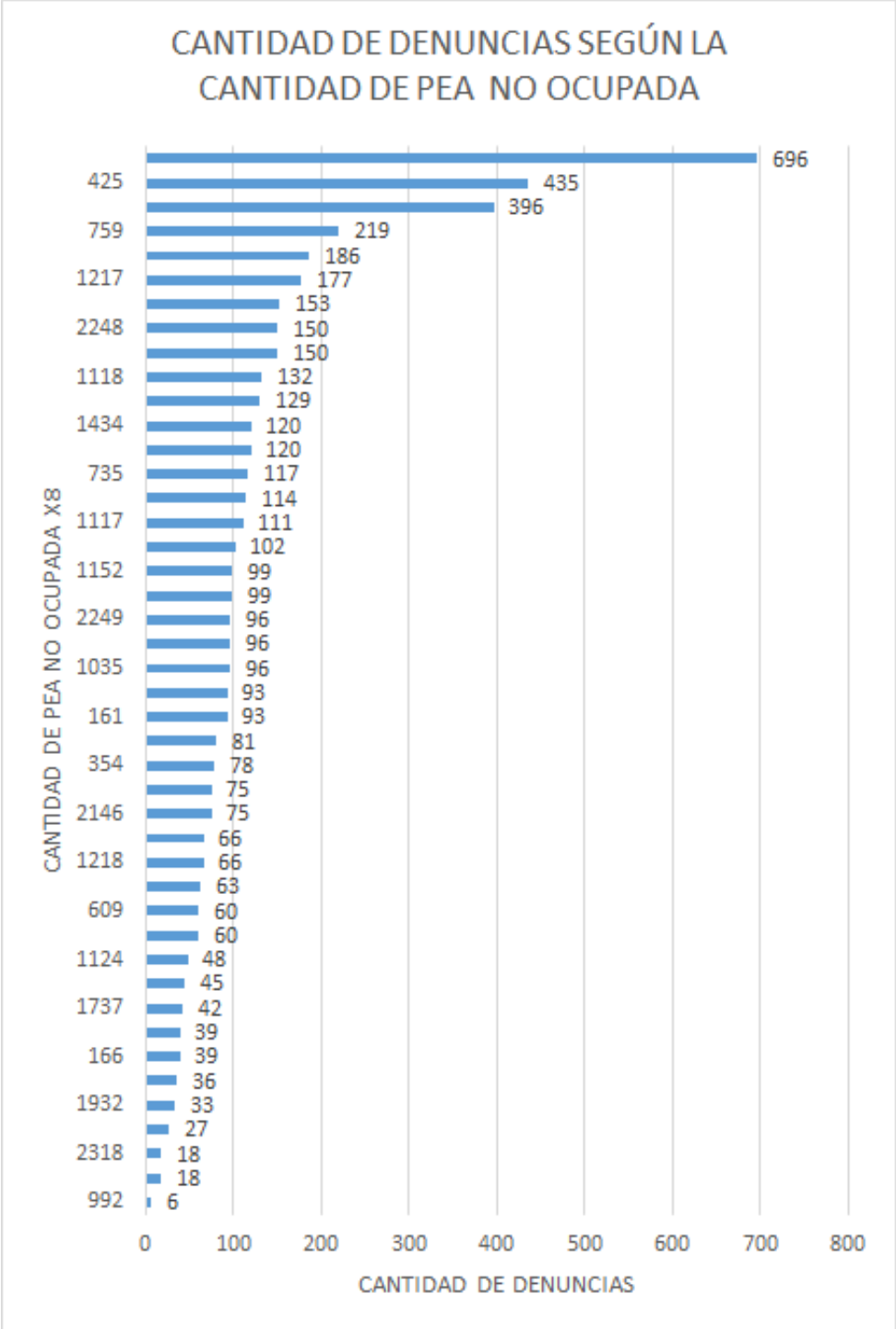


Gráfico 13: Cantidad de denuncias según la cantidad de PEA no ocupada.

Ingreso per cápita por hogar (X9) presenta un comportamiento heterogéneo, lo cual genera efecto en la generación de los modelos de regresión, como se muestra en el gráfico 14.

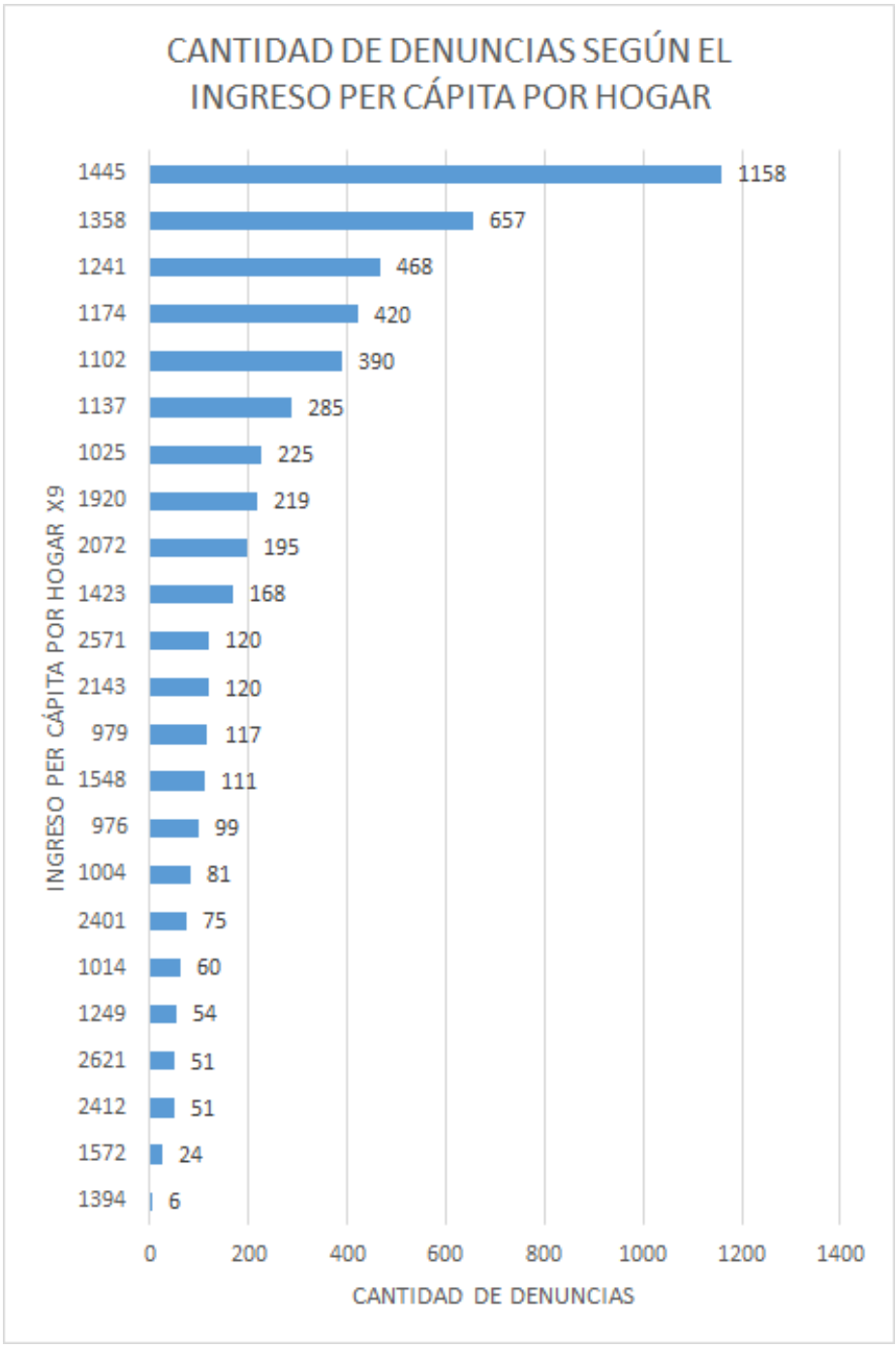


Gráfico 14: Cantidad de denuncias según ingreso per cápita por hogar.

Nivel socio económico (X10) presenta un comportamiento heterogéneo, lo cual genera efecto en la generación de los modelos de regresión, como se muestra en el gráfico 15.

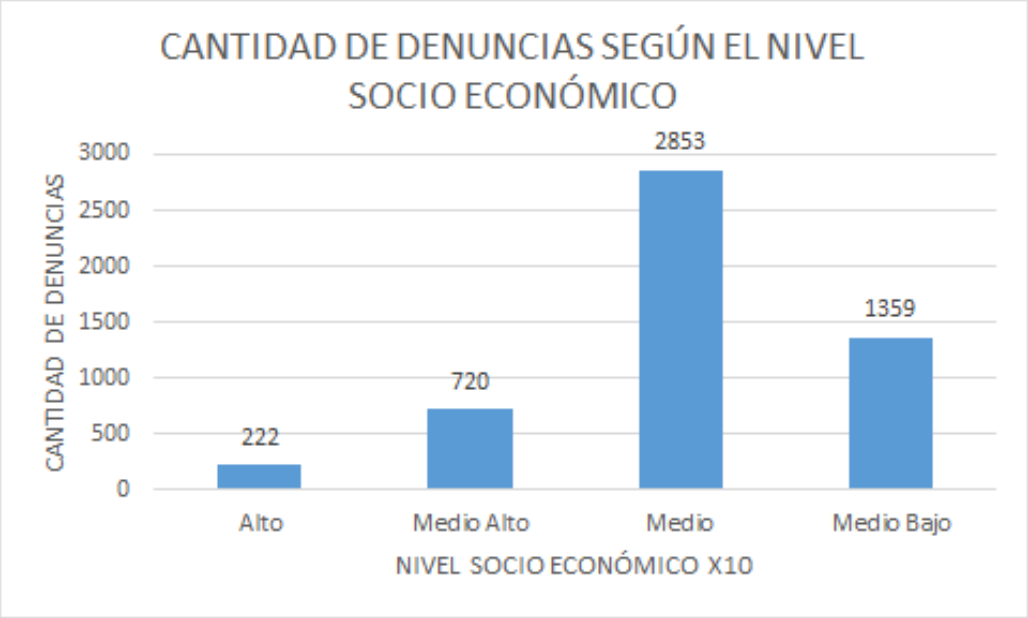


Gráfico 15: Cantidad de denuncias según nivel socio económico.

Número de habitantes por sereno (X11) presenta un comportamiento homogéneo en dos grupos y un dato atípico, lo cual genera poco efecto en la generación de los modelos de regresión, como se muestra en el gráfico 16.

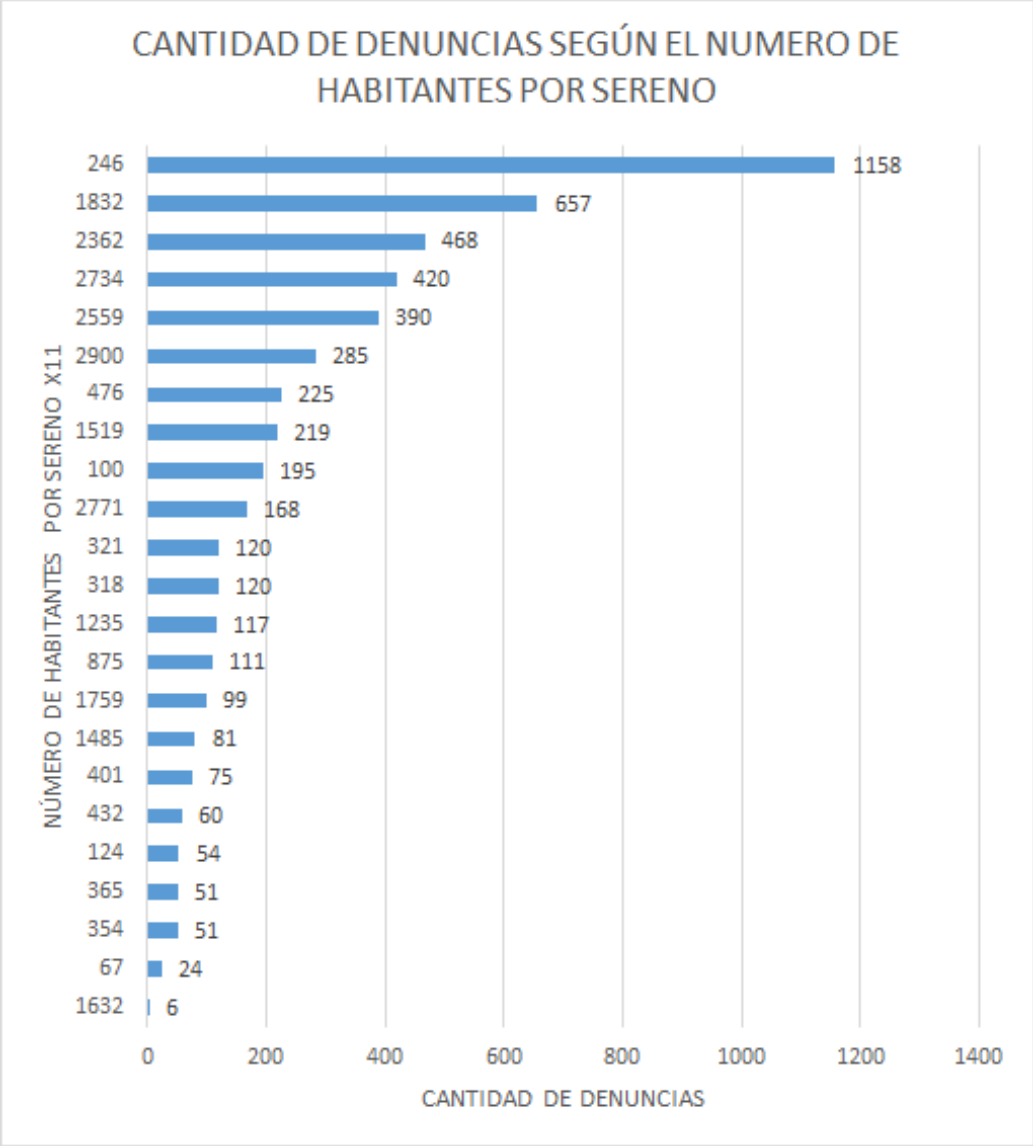


Gráfico 16: Cantidad de denuncias según el número de habitantes por sereno.

Cantidad de denuncias (W) presenta un comportamiento heterogéneo con una mayor concentración de casos en la cantidad de denuncias menores, como se muestra en el gráfico 17.

Gráfico 17: Cantidad de denuncias según la frecuencia de la denuncia.

6.1.4. Conclusión 4

Esta conclusión resulta de análisis de la hipótesis general.

La aplicación de un modelo de regresión en el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana genera un buen método para una buena estimación en el pronóstico.

El modelo de Random Forest Regressor, presentó mejor capacidad de pronóstico que el modelo de Árbol de Decisión Regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana.

6.2. Discusión

El método de Random Forest Regressor presento un R^2 en la validación de 0.822, lo que genera un error cuantitativo de 21.90 lo cual presento un mejor resultado en comparación al estudio de ALIF RIDZUAN KHAIRUDDIN, RAZANA ALWEE y HABIBOLLAH (2020) en su Artículo “UN ANÁLISIS COMPARATIVO DE LAS TÉCNICAS DE INTELIGENCIA ARTIFICIAL PARA PRONOSTICAR LA TASA DE DELITOS VIOLENTOS” sustentada en, SCHOOL OF COMPUTING, FACULTY OF ENGINEERING, UNIVERSITI TEKNOLOGI MALAYSIA, 81310, JOHOR BAHRU, JOHOR, MALAYSIA en la cual muestra que el Árbol de Gradiente GTB se considera más apropiado en el manejo de datos de tasas de criminalidad de series de tiempo limitadas en comparación con ANN y SVR, Así como se muestran en los resultados de la del error cuantitativo para el Robo con GTB 25.28 para Asaltos agravados con GTB 37.03.

En la presente investigación el método de Random Forest Regressor presento mejor capacidad para predecir con un 0.822 frente al SVR con un porcentaje de clasificaciones correctas de los sujetos no reincidentes del 0.799 y de los sujetos reincidentes del 0.708 del estudio de PÉREZ, Meritxell; REDONDO, Santiago; MARTÍNEZ, Marian; GARCÍA, Carlos; ANDRÉS, Antonio (2008) en el artículo “PREDICCIÓN DE RIESGO DE REINCIDENCIA EN AGRESORES SEXUALES” Universidad de Oviedo, España.

6.3. Recomendaciones

La aplicación del modelo de Random Forest Regressor para el pronóstico de la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú, genera muy buenos resultados para conocer la cantidad de denuncias por delitos que se registran en las comisarías de la Policía Nacional de Perú en Lima Metropolitana, por lo que se recomienda su aplicación en las distintas Comisarías de la Policía a nivel Nacional en las cuales se requiera su uso.

Referencias

Barajas, F. H. (2021). *Modelos predictivos* (Vol. 1).

Benites J, P., Cervantes D.J. (2017). *Mejora del sistema sidpol para la policía nacional del perú*. Universidad Peruana de Ciencias Aplicadas.

Burbano, V., Margoth, V. (2016). Inferencia estadística básica, apoyo al estudio independiente. , *1*.

Dirección de Telemática de la Policia Nacional del Perú, D. (2018). *Sistema informático de denuncias policiales version 1.1*. Manual de usuario DIRTIC.

Fawcett, T. (2005). An introducction to roc analysis. , *2*.

Géron, A. (2020). Aprende machine learning con scikit-learn, keras y tensorflow. *Anaya*.

Gutierrez Delgado, M. V. (s.f.). Sistema de distribución de carga policial mediante de predicción de delitos.

Instituto Nacional de Estadística e Informática, P. (2018). *El sistema integrado de estadísticas de la criminalidad y seguridad ciudadana*. Oficina Técnica de Difusión INEI.

Irma Luisa Torres, L. J. C. (2013). *Analisis estadistico y pronosticos con series temporales de la informacion de las denuncias de accidentes de transito, comisaria distrital de huaraz: 2007-2021*. Universidad Nacional Santiago Antunes de Mayolo.

MATOS, E. E. C., MANSILLA, Y. M. R., Huaman, E. C. (2014). Factores de riesgo de conducta delictiva en alumnos de nivel secundario de las zonas urbano marginales de los distritos de huánuco, pillco marca y amarilis. *Investigación Valdizana*, *8*(1), 57–60.

Merino, R. F. M., Chacón, C. I. Ñ. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas r y python. *Interfases*(10), 165–189.

Ministerio del Interior Decreto Supremo N° 026.2017-IN, P. (2017). *Reglamento del decreto legislativo n° 1267 ley de la policía nacional del perú*. Diario el Peruano.

Montgomery, D. C., Peck, E. A., vining, G. (2005). Introdución al análisis de regresión lineal. , *3*.

Moya, R., Saravia, G. (1988). Probabilidad e inferencia estadística. , *2*.

Oberti, A., Bacci, C. (2016). Metodología de la investigación.

Ordóñez, H., Cobos, C., Bucheli, V. (2020). Modelo de machine learning para la predicción de las tendencias de hurto en colombia. *Revista Ibérica de Sistemas e Tecnologias de Informação*(E29), 494–506.

Paz Campusano, O. (2019). *Estoy alerta: Una comisaria para 3500 manzanas de lima*. Artículo Periodístico.

Pértegas Díaz, S., Pita Fernández, S. (2001). La distribución normal. *Cad Aten Primaria*, *8*, 268–274.

Poder Ejecutivo Decreto Lejislativo Nª 1267, L. d. l. P. N. d. P. (2016). *Ley de la policia nacional del peru*. Diario El Peruano.

Rafael Caparo, E. P. (2017). *Modelos de econometría espacial para la lucha contra la delincuencia en el perú: Un enfoque de optimización en tiempo real* (Vol. 1). Universidad Nacional de Ingeniería.

Ramírez, M. P., Illescas, S. R., García, M. M., Forero, C. G., Pueyo, A. A. (2008). Predicción de riesgo de reincidencia en agresores sexuales. *Psicothema*, *20*(2), 205–210.

Rich, E., Knight, K., Calero, P. A. G., Bodega, F. T. (1994). *Inteligencia artificial* (Vol. 1). McGraw-Hill.

Sánchez Turcios, R. A. (2015). t-student: Usos y abusos. *Revista mexicana de cardiología*, *26*(1), 59–61.

Sepúlveda, J. F. D., Morales, J. C. C. (2013). Comparación entre árboles de regresión cart y regresión lineal. *Comunicaciones en Estadística*, *6*(2), 175–195.

Toledo, R. (2005). *Métodos econométricos para el pronóosticos de delitos en el gran santiago*. Universidad de Chile.

Veneri Guarch, F. A. (2019). Métodos para la predicción de robos violentos: ejercicio comparado para montevideo, uruguay.

7. ANEXOS

7.1. Técnicas de recolección y procesamiento de datos

Identificar las 143 comisarias básicas que se encuentran en Lima Metropolitana.

Nº	DISTRITO	COMISARIA
1	ANCON	ANCÓN
2	ATE	HUAYCAN
3	ATE	SALAMANCA
4	ATE	SANTA CLARA
5	ATE	VITARTE
6	BARRANCO	BARRANCO
7	BARRANCO	COMTUR LIMA SUR
8	BREÑA	BREÑA
9	BREÑA	CHACRA COLORADA
10	CARABAYLLO	CARABAYLLO
11	CARABAYLLO	PROGRESO
12	CARABAYLLO	SANTA ISABEL
13	CHACLACAYO	CHACLACAYO
14	CHORRILLOS	CHORRILLOS
15	CHORRILLOS	SAN GENARO
16	CHORRILLOS	VILLA
17	CIENEGUILLA	CIENEGUILLA
18	COMAS	CIE -COMAS
19	COMAS	COLLIQUE
20	COMAS	COMISARIA DE MUJERES DE COLLIQUE
21	COMAS	LA PASCANA
22	COMAS	SANTA LUZMILA
23	COMAS	TUPAC AMARU
24	COMAS	UNIVERSITARIA
25	EL AGUSTINO	COMISARIA MUJERES EL AGUSTINO
26	EL AGUSTINO	EL AGUSTINO
27	EL AGUSTINO	SAN CAYETANO
28	EL AGUSTINO	SAN PEDRO
29	EL AGUSTINO	SANTOYO
30	EL AGUSTINO	VILLA HERMOSA
31	INDEPENDENCIA	COMISARIA DE MUJERES DE INDEPENDENCIA
32	INDEPENDENCIA	INDEPENDENCIA
33	INDEPENDENCIA	LA UNIFICADA
34	INDEPENDENCIA	PAYET
...	...	...
141	VILLA MARIA DEL TRIUNFO	NUEVA ESPERANZA
142	VILLA MARIA DEL TRIUNFO	S.F. TABLADA L.
143	VILLA MARIA DEL TRIUNFO	VMT

Figura 12: Cuadro de la Comisarias básicas de Lima Metropolitana.

Cuadro con información estructurada sobre las variables w, x1, x2, x3.

N	Cantidad de denuncias (w)	Tipo (X1)	Sub tipo (X2)	Modalidad (X3)
1	1	ADMINISTRACION PUBLICA (DELITO)	COMETIDOS POR FUNCIONARIOS PUBLICOS	COHECHO ACTIVO ESPECIFICO
2	1	ADMINISTRACION PUBLICA (DELITO)	COMETIDOS POR FUNCIONARIOS PUBLICOS	PECULADO
3	14	ADMINISTRACION PUBLICA (DELITO)	COMETIDOS POR PARTICULARES	DESOBEDIENCIA O RESISTENCIA A LA AUTORIDAD
4	23	ADMINISTRACION PUBLICA (DELITO)	COMETIDOS POR PARTICULARES	INGRESO INDEBIDO DE EQUIPOS O SISTEMAS DE COMUNICACION, FOTOGRAFIA Y/O FILMACION EN CENTROS DE DETEN
5	1	ADMINISTRACION PUBLICA (DELITO)	COMETIDOS POR PARTICULARES	PERTURBACION DEL ORDEN DONDE LA AUTORIDAD EJERCE SU FUNCION
6	3	ADMINISTRACION PUBLICA (DELITO)	COMETIDOS POR PARTICULARES	POSESION INDEBIDA DE TELEFONOS CELULARES O ARMAS,MUNICIONES O MATERIALES EXPLOSIVOS,INFLAMABLES,A SFI
7	3	ADMINISTRACION PUBLICA (DELITO)	COMETIDOS POR PARTICULARES	VIOLENCIA CONTRA LA AUTORIDAD PARA IMPEDIR EL EJERCICIO DE SUS FUNCIONES
...	...	...	...	...
6745	5	VIDA, EL CUERPO Y LA SALUD (DELITO)	LESIONES	LESIONES LEVES

Figura 13: 6745 registros de la cantidad de denuncias con tipo, sub tipo y modalidad en Lima Metropolitana.

Obtener archivos en formato PDF.

Figura 14: Imagen de archivos del INEI en formato PDF.

Extraer la información del INEI.

```
[ ] !pip install tika

[ ] from tika import parser

[ ] !pip install pdf_extractor

[ ] from google.colab import files
    data1 = files.upload()

    Elegir archivos  Ningún archivo seleccionado  Upload widget is only available when the cell has been executed in the
    Saving 107_parte_IV tasa migracion Lima 2007.pdf to 107_parte_IV tasa migracion Lima 2007.pdf

    raw = parser.from_file('107_parte_IV tasa migracion Lima 2007.pdf')
    print(raw['content'])
    algunas universidades, y hasta sedes de nuevas
    universidades privadas.

    Esta reconfiguración de la ciudad de Lima, que combina
    la lógica del gran capital y con la de los pequeños
    productores y comerciantes, es la nueva Lima que nace
    desde sus zonas de expansión, que sin dejar de mantener
    una relativa dependencia de los distritos tradicionales
    y modernos de Lima Centro, van construyendo nuevos
    "centros de la ciudad" de menor jerarquía, pero que
    se encuentran en continuo crecimiento, ganando así
    relativa autonomía. La ciudad de Lima pasa de tener
    un solo centro a ser multicéntrica y los conos dejan de
    ser solo ciudades dormitorio.
```

Figura 15: Codigo python para extraer información de formato PDF.

Codigo y Cuadro de obtener información estructurada sobre las variables x4, x5.

```
''''''''''''''''''''''''''''''''''''''''''''
'ÁREA Y POBLACIÓN DE LIMA METROPOLITANA
''''''''''''''''''''''''''''''''''''''''''''
Dim i, j As Integer
i = 8
For i = 7 To 49
c = 0
For j = 1 To 30
If IsNumeric(Mid(Hoja2.Cells(i, 11), j, 1)) Then
    'Hoja2.Cells(i, 13) = Mid(Hoja2.Cells(i, 11), j, 1)
 j = 31
End If
c = c + 1
Next
'MsgBox (c)
Hoja2.Cells(i, 14) = Mid(Hoja2.Cells(i, 11), 1, c - 1)
''''''''''''''''''''''''''''''''''''''''''''
c2 = InStr(Mid(Hoja2.Cells(i, 11), c, 20), " ")
Hoja2.Cells(i, 15) = Mid(Hoja2.Cells(i, 11), c, c2 - 1)
c3 = InStr(Mid(Hoja2.Cells(i, 11), c + c2, 20), " ")
Hoja2.Cells(i, 16) = Mid(Hoja2.Cells(i, 11), c + c2, c3 - 1)
c4 = InStr(Mid(Hoja2.Cells(i, 11), c + c2 + c3, 20), " ")
Hoja2.Cells(i, 17) = Mid(Hoja2.Cells(i, 11), c + c2 + c3, c4 - 1)
c5 = InStr(Mid(Hoja2.Cells(i, 11), c + c2 + c3 + c4, 20), " ")
Hoja2.Cells(i, 18) = Mid(Hoja2.Cells(i, 11), c + c2 + c3 + c4, c5 - 1)
c6 = InStr(Mid(Hoja2.Cells(i, 11), c + c2 + c3 + c4 + c5, 20), " ")
Hoja2.Cells(i, 19) = Mid(Hoja2.Cells(i, 11), c + c2 + c3 + c4 + c5, c6 - 1)
c7 = InStr(Mid(Hoja2.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6, 20), " ")
Hoja2.Cells(i, 20) = Mid(Hoja2.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6, c7 - 1)
Next
```

Figura 16: Codigo VB para la obtención del área y población de Lima Metropolitana.

Distrito	Ubigeo	Area	Censo 2019
Ancón	150102	299,22	45999
Ate Vitarte	150103	77,72	666650
Barranco	150104	3,33	34270
Breña	150105	3,22	85343
Carabayllo	150106	346,88	320392
Chaclacayo	150107	39,5	47643
Chorrillos	150108	38,94	351582
Cieneguilla	150109	240,33	50348
Comas	150110	48,75	566852
El Agustino	150111	12,54	205302
Independencia	150112	14,56	234579
Jesús María	150113	4,57	82336
La Molina	150114	65,75	190640
La Victoria	150115	8,74	188913
Lima	150101	21,88	30327
Lince	150116	3,03	57276
Los Olivos	150117	18,25	400371
Lurigancho	150118	236,47	201998
Lurín	150119	181,12	89818
Magdalena del Mar	150120	3,61	62239
Miraflores	150122	9,62	95500
Pachacamac	150123	160,23	135250
Pucusana	150124	37,83	1810
Pueblo Libre	150121	4,38	87434
Puente Piedra	150125	71,18	370356
Punta Hermosa	150126	119,5	8201
Punta Negra	150127	130,5	8543
Rímac	150128	11,87	181718
San Bartolo	150129	45,01	8196
San Borja	150130	9,96	127402
San Isidro	150131	11,1	63484
San Juan de Lurigancho	150132	131,25	99999
San Juan de Miraflores	150133	23,98	435440
San Luis	150134	3,49	63284
San Martín de Porres	150135	36,91	759561
San Miguel	150136	10,72	152780
Santa Anita	150137	10,69	243327
Santa María del Mar	150138	9,81	89616
Santa Rosa	150139	21,5	19982
Santiago de Surco	150140	34,75	385952
Surquillo	150141	3,46	102534
Villa El Salvador	150142	35,46	489391
Villa María del Triunfo	150143	70,57	478785

Figura 17: Cuadro sobre área y población de Lima Metropolitana.

Codigo y Cuadro de obtener información estructurada sobre la variable x6.

```
'''''''''''''''''''''''''''''''''''''
'CANTIDAD DE MERCADOS DE ABASTOS
'''''''''''''''''''''''''''''''''''''
x = ""
For s = 3 To 58
  x = x & Hoja10.Cells(s, 2)
Next
i = 3
cantidad = Len(x)
j = 4
 For c = 1 To cantidad
  posicion = InStr(c, x, "MERCADO")
  Hoja11.Cells(j, 4) = posicion
  Hoja11.Cells(j, 5) = Mid(x, posicion, 30)
  c = posicion + 1
  j = j + 1
  If j = 1000 Then
   c = cantidad + 1
  End If
 Next
```

Figura 18: Codigo VB para la cantidad de mercados en Lima Metropolitana.

DISTRITO	CANTIDAD DE MERCADOS
ANCON	8
ATE	44
BARRANCO	4
BREÑA	20
CARABAYLLO	24
CHACLACAYO	7
CHORRILLOS	35
CIENEGUILLA	4
COMAS	40
EL AGUSTINO	13
INDEPENDENCIA	16
JESUS MARIA	2
LA MOLINA	5
LA VICTORIA	34
LIMA	50
LINCE	6
LOS OLIVOS	48
LURIGANCHO	15
LURIN	9
MAGDALENA DEL MAR	6
MIRAFLORES	3
PACHACAMAC	9
PUCUSANA	2
PUEBLO LIBRE	10
PUENTE PIEDRA	28
PUNTA NEGRA	1
RIMAC	27
SAN BARTOLO	1
SAN BORJA	9
SAN ISIDRO	2
SAN JUAN DE LURIGANCHO	127
SAN JUAN DE MIRAFLORES	28
SAN LUIS	8
SAN MARTIN DE PORRES	75
SAN MIGUEL	19
SANTA ANITA	20
SANTA ROSA	2
SANTIAGO DE SURCO	6
SURQUILLO	13
VILLA EL SALVADOR	35
VILLA MARIA DEL TRIUNFO	41

Figura 19: Cuadro de la cantidad de mercados en Lima Metropolitana.

Codigo y Cuadro de obtener información estructurada sobre las variables x7, x8.

```
'''''''''''''''''''''''''''''''
'PEA OCUPADA Y PEA NO OCUPADA
'''''''''''''''''''''''''''''''
Dim i, j As Integer
i = 11
For i = 11 To 181
c = 0
For j = 1 To 50
If IsNumeric(Mid(Hoja7.Cells(i, 11), j, 1)) Then
    'Hoja2.Cells(i, 13) = Mid(Hoja2.Cells(i, 11), j, 1)
 j = 51
End If
c = c + 1
Next
Hoja7.Cells(i, 14) = Mid(Hoja7.Cells(i, 11), 1, c - 1)
c2 = InStr(Mid(Hoja7.Cells(i, 11), c, 20), " ")
Hoja7.Cells(i, 15) = Mid(Hoja7.Cells(i, 11), c, c2 - 1)
c3 = InStr(Mid(Hoja7.Cells(i, 11), c + c2, 20), " ")
Hoja7.Cells(i, 16) = Mid(Hoja7.Cells(i, 11), c + c2, c3 - 1)
c4 = InStr(Mid(Hoja7.Cells(i, 11), c + c2 + c3, 20), " ")
Hoja7.Cells(i, 17) = Mid(Hoja7.Cells(i, 11), c + c2 + c3, c4 - 1)
c5 = InStr(Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4, 20), " ")
Hoja7.Cells(i, 18) = Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4, c5 - 1)
c6 = InStr(Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5, 20), " ")
Hoja7.Cells(i, 19) = Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5, c6 - 1)
c7 = InStr(Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6, 20), " ")
Hoja7.Cells(i, 20) = Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6, c7 - 1)
c8 = InStr(Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6 + c7, 20), " ")
Hoja7.Cells(i, 21) = Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6 + c7, c8 - 1)
c9 = InStr(Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6 + c7 + c8, 20), " ")
Hoja7.Cells(i, 22) = Mid(Hoja7.Cells(i, 11), c + c2 + c3 + c4 + c5 + c6 + c7 + c8, c9 - 1)
Next
```

Figura 20: Codigo VB para la PEA Ocupada y No Ocupada en Lima Metropolitana.

Distrito	Total	Población Económicamente Activa - PEA Ocupada	Población Económicamente Inactiva -PEA No ocupada
Lima, Lima, distrito de Ancón	50762	33658	17104
Lima, Lima, distrito de Ate	479614	322845	156769
Lima, Lima, distrito de Barranco	30311	21453	8858
Lima, Lima, distrito de Breña	75516	52984	22532
Lima, Lima, distrito de Carabayllo	261790	168530	93260
Lima, Lima, distrito de Chaclacayo	34704	2293	32411
Lima, Lima, distrito de Chorrillos	271691	189686	82005
Lima, Lima, distrito de Cieneguilla	26620	18314	8306
Lima, Lima, distrito de Comas	434785	276464	158321
Lima, Lima, distrito de El Agustino	166350	11389	154961
Lima, Lima, distrito de Independencia	173379	110286	63093
Lima, Lima, distrito de Jesús María	67437	45697	21740
Lima, Lima, distrito de La Molina	129494	82529	46965
Lima, Lima, distrito de La Victoria	154164	111604	42560
Lima, Lima, distrito de Lima	226383	157272	69111
Lima, Lima, distrito de Lince	50484	35175	15309
Lima, Lima, distrito de Los Olivos	277903	181992	95911
Lima, Lima, distrito de Lurigancho	193245	126239	67006
Lima, Lima, distrito de Lurin	73006	53065	19941
Lima, Lima, distrito de Magdalena del Mar	52434	36441	15993
Lima, Lima, distrito de Miraflores	89993	62793	27200
Lima, Lima, distrito de Pachacamac	89154	62915	26239
Lima, Lima, distrito de Pucusana	11085	7650	3435
Lima, Lima, distrito de Pueblo Libre	76819	51626	25193
Lima, Lima, distrito de Puente Piedra	264916	172152	92764
Lima, Lima, distrito de Punta Hermosa	12821	9527	3294
Lima, Lima, distrito de Punta Negra	5797	4137	1660
Lima, Lima, distrito de Rímac	144376	99399	44977
Lima, Lima, distrito de San Bartolo	6322	4472	1850
Lima, Lima, distrito de San Borja	104969	70497	34472
Lima, Lima, distrito de San Isidro	56919	37535	19384
Lima, Lima, distrito de San Juan de Lurigancho	863228	583800	279428
Lima, Lima, distrito de San Juan de Miraflores	321576	227809	93767
Lima, Lima, distrito de San Luis	46265	31541	14724
Lima, Lima, distrito de San Martín de Porres	562866	364776	198090
Lima, Lima, distrito de San Miguel	136523	93804	42719
Lima, Lima, distrito de Santa Anita	168739	116289	52450
Lima, Lima, distrito de Santa María del Mar	847	624	223
Lima, Lima, distrito de Santa Rosa	21141	13796	7345
Lima, Lima, distrito de Santiago de Surco	318953	217544	101409
Lima, Lima, distrito de Surquillo	82552	58599	23953
Lima, Lima, distrito de Villa el Salvador	320957	230106	90851
Lima, Lima, distrito de Villa María del Triunfo	327685	230538	97147

Figura 21: Cuadro de la PEA Ocupada y No Ocupada en Lima Metropolitana.

Codigo y Cuadro de obtener información estructurada sobre las variables x9, x10.

```
''''''''''''''''''''''''''''''''''''''''''''''''
'INGRESO PER CÁPITA Y NIVEL SOCIO ECONÓMICO
''''''''''''''''''''''''''''''''''''''''''''''''
Dim i, j, c, c1 As Long
Dim x(1000000) As Long
j = 1
For i = 21363 To 151963
'c = ActiveSheet.Cells(Rows.Count, "J").End(xlUp).Row + 1
  If Hoja4.Cells(i, 1) = 1 Then
  x(j) = i
  j = j + 1
  End If
Next
For h = 1 To 45
  For ww = x(h) To x(h + 1) - 1
       If Hoja4.Cells(ww, 1) = 1 Then
         c = ActiveSheet.Cells(Rows.Count, "J").End(xlUp).Row + 1
         Hoja4.Cells(c, 10) = Hoja4.Cells(ww, 2)
       End If
       If Hoja4.Cells(ww, 3) = 1 Then
         c1 = ActiveSheet.Cells(Rows.Count, "K").End(xlUp).Row + 1
         Hoja4.Cells(c1, 11) = Hoja4.Cells(ww, 2)
       End If
       If Hoja4.Cells(ww, 3) = 2 Then
         c2 = ActiveSheet.Cells(Rows.Count, "L").End(xlUp).Row + 1
         Hoja4.Cells(c2, 12) = Hoja4.Cells(ww, 2)
       End If
       If Hoja4.Cells(ww, 3) = 3 Then
         c3 = ActiveSheet.Cells(Rows.Count, "M").End(xlUp).Row + 1
         Hoja4.Cells(c3, 13) = Hoja4.Cells(ww, 2)
       End If
       If Hoja4.Cells(ww, 3) = 4 Then
         c4 = ActiveSheet.Cells(Rows.Count, "N").End(xlUp).Row + 1
         Hoja4.Cells(c4, 14) = Hoja4.Cells(ww, 2)
       End If
       If Hoja4.Cells(ww, 3) = 5 Then
         c5 = ActiveSheet.Cells(Rows.Count, "O").End(xlUp).Row + 1
         Hoja4.Cells(c5, 15) = Hoja4.Cells(ww, 2)
       End If
  Next
Next
```

Figura 22: Codigo VB para el Nivel Socio Economico y el Ingreso per cápita por hogar en Lima Metropolitana.

DISTRITO	Nivel socio economico	Ingreso per cápita por hogar
DISTRITO LIMA	Medio alto	1579
DISTRITO ANCON	Alto	1031
DISTRITO ATE	Alto	829
DISTRITO BARRANCO	Medio	1760
DISTRITO BREÑA	Alto	1975
DISTRITO CARABAYLLO	Alto	1030
DISTRITO CHACLACAYO	Medio bajo	1450
DISTRITO CHORRILLOS	Medio bajo	1315
DISTRITO CIENEGUILLA	Medio bajo	1260
DISTRITO COMAS	Alto	1253
DISTRITO EL AGUSTINO	Bajo	1214
DISTRITO INDEPENDENCIA	Alto	912
DISTRITO JESUS MARIA	Alto	5000
DISTRITO LA MOLINA	Bajo	300
DISTRITO LA VICTORIA	Alto	1620
DISTRITO LINCE	Medio	1460
DISTRITO LOS OLIVOS	Medio bajo	1637
DISTRITO LURIGANCHO	Bajo	934
DISTRITO LURIN	Alto	1147
DISTRITO MAGDALENA DEL MAR	Alto	5000
DISTRITO PUEBLO LIBRE	Medio bajo	1135
DISTRITO MIRAFLORES	Alto	5000
DISTRITO PACHACAMAC	Medio alto	1056
DISTRITO PUCUSANA	Alto	768
DISTRITO PUENTE PIEDRA	Alto	945
DISTRITO PUNTA HERMOSA	Bajo	748
DISTRITO PUNTA NEGRA	Alto	1789
DISTRITO RIMAC	Alto	1338
DISTRITO SAN BARTOLO	Medio alto	1513
DISTRITO SAN BORJA	Medio bajo	80
DISTRITO SAN ISIDRO	Alto	5000
DISTRITO SAN JUAN DE LURIGANCHO	Bajo	821
DISTRITO SAN JUAN DE MIRAFLORES	Alto	1204
DISTRITO SAN LUIS	Alto	1813
DISTRITO SAN MARTIN DE PORRES	Alto	1426
DISTRITO SAN MIGUEL	Medio	680
DISTRITO SANTA ANITA	Alto	1536
DISTRITO SANTA MARIA DEL MAR	Medio alto	405
DISTRITO SANTA ROSA	Alto	1012
DISTRITO SANTIAGO DE SURCO	Bajo	617
DISTRITO SURQUILLO	Medio bajo	805
DISTRITO VILLA EL SALVADOR	Alto	1137
DISTRITO VILLA MARIA DEL TRIUNFO	Alto	1035

Figura 23: Cuadro del Nivel Socio Economico y el Ingreso per cápita por hogar en Lima Metropolitana.

Codigo y Cuadro de obtener información estructurada sobre la variable x11.

```
'''''''''''''''''''''''''''''''''
'NÚMERO DE HABITANTES POR SERENO
'''''''''''''''''''''''''''''''''
Dim i, c As Integer
For i = 11946 To 12070
c = ActiveSheet.Cells(Rows.Count, "R").End(xlUp).Row + 1
  If Hoja5.Cells(i, 1) = 2 And Hoja5.Cells(i, 2) <> "" Then
    Hoja5.Cells(c, 18) = Hoja5.Cells(i, 2)
  End If
Next
```

Figura 24: Codigo VB para el Numero de habitantes por sereno en Lima Metropolitana.

Distritos	Habitantes por sereno
La Punta	5785
San Isidro	5204
Santa María del Mar	4787
Punta Hermosa	4550
San Bartolo	2824
Punta Negra	2805
Miraflores	2675
Carmen de La Legua Reynoso	2559
San Borja	2546
Jesús María	1696
Barranco	1657
Lima	1623
Lurín	1463
Magdalena del Mar	1419
Santiago de Surco	1387
Lince	1361
Lurigancho	1352
La Molina	1338
Pueblo Libre	1270
Surquillo	1135
San Miguel	1118
Bellavista	1041
La Perla	914
La Victoria	855
San Luis	756
Cieneguilla	752
Santa Rosa	684
Ventanilla	597
Puente Piedra	390
Breña	387
Pucusana	369
Pachacámac	367
Chaclacayo	344
El Agustino	332
Los Olivos	316
Santa Anita	281
Ancón	227
Ate	197
Callao	194
Rímac	182
San Juan de Miraflores	178
Carabayllo	167
Independencia	119
San Martín de Porres	115
San Juan de Lurigancho	111
Chorrillos	82
Villa El Salvador	64
Comas	57
Villa María del Triunfo	44

Figura 25: Cuadro del Numero de habitantes por sereno en Lima Metropolitana.

Codigo de Transformar la información de las variables x4, x5, x6, x7, x8, x9, x10, x11 de nivel distrito a nivel comisaria

Nª	DISTRITO	COMISARIA	PORCENTAJE
1	ANCON	ANCÓN	100%
2	ATE	HUAYCAN	30%
3	ATE	SALAMANCA	10%
4	ATE	SANTA CLARA	30%
5	ATE	VITARTE	30%
6	BARRANCO	BARRANCO	50%
7	BARRANCO	COMTUR LIMA SUR	50%
8	BREÑA	BREÑA	50%
9	BREÑA	CHACRA COLORADA	50%
10	CARABAYLLO	CARABAYLLO	32%
11	CARABAYLLO	PROGRESO	65%
12	CARABAYLLO	SANTA ISABEL	3%
13	CHACLACAYO	CHACLACAYO	100%
14	CHORRILLOS	CHORRILLOS	14%
15	CHORRILLOS	SAN GENARO	29%
16	CHORRILLOS	VILLA	57%
17	CIENEGUILLA	CIENEGUILLA	100%
18	COMAS	CIE -COMAS	16%
19	COMAS	COLLIQUE	22%
20	COMAS	COMISARIA DE MUJERES DE COLLIQUE	7%
21	COMAS	LA PASCANA	14%
22	COMAS	SANTA LUZMILA	22%
23	COMAS	TUPAC AMARU	16%
24	COMAS	UNIVERSITARIA	3%
25	EL AGUSTINO	COMISARIA MUJERES EL AGUSTINO	18%
26	EL AGUSTINO	EL AGUSTINO	45%
27	EL AGUSTINO	SAN CAYETANO	7%
28	EL AGUSTINO	SAN PEDRO	5%
29	EL AGUSTINO	SANTOYO	11%
30	EL AGUSTINO	VILLA HERMOSA	14%
31	INDEPENDENCIA	COMISARIA DE MUJERES DE INDEPENDENCIA	36%
32	INDEPENDENCIA	INDEPENDENCIA	15%
33	INDEPENDENCIA	LA UNIFICADA	6%
34	INDEPENDENCIA	PAYET	18%
35	INDEPENDENCIA	TAHUANTINSUYO	24%
36	JESUS MARIA	JESÚS MARÍA	100%
37	LA MOLINA	LA MOLINA	63%
38	LA MOLINA	LAS PRADERAS	25%
39	LA MOLINA	SANTA FELICIA	13%
40	LA VICTORIA	APOLO	37%
41	LA VICTORIA	LA VICTORIA	53%
...	...	...	...
143	VILLA MARIA DEL TRIUNFO	VMT	11%

Figura 26: Cuadro del área de jurisdiccion que posee la comisaria en un distrito.

Codigo de Generar una base de datos de la cantidad de denuncias en las comisarías de Lima Metropolitana con las variables x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11

```
Dim i, j, jj As Integer
' HABITANTES
For i = 2 To 8949
  'DISTRITO
  For j = 7 To 49
    If Hoja16.Cells(i, 3) = Hoja2.Cells(j, 34) Then
      com = Hoja2.Cells(j, 32)
      j = 50
    End If
  Next
 'PORCENTAJE, COMISARIA
  For jj = 4 To 146
    If Hoja16.Cells(i, 4) = Hoja15.Cells(jj, 26) Then
      porc = Hoja15.Cells(jj, 27)
      jj = 150
    End If
  Next
  Hoja16.Cells(i, 9) = com * porc

Next
```

Figura 27: Codigo de Generar una base de datos de la cantidad de denuncias en las comisarías de Lima Metropolitana con las variables x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11.

Codigo de Eliminar los datos atípicos a la base de datos con las variables w, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11

```
Range("K1").Select
ActiveWorkbook.Worksheets("Hoja18").AutoFilter.Sort.SortFields.Clear
ActiveWorkbook.Worksheets("Hoja18").AutoFilter.Sort.SortFields.Add Key:=Range _
    ("K1:K8949"), SortOn:=xlSortOnValues, Order:=xlDescending, DataOption:= _
    xlSortNormal
With ActiveWorkbook.Worksheets("Hoja18").AutoFilter.Sort
    .Header = xlYes
    .MatchCase = False
    .Orientation = xlTopToBottom
    .SortMethod = xlPinYin
    .Apply
End With
Dim i As Integer
For i = 2 To 1000
  Hoja18.Cells(i, 18) = 1
Next
For i = 7000 To 8590
  Hoja18.Cells(i, 18) = 1
Next
```

Figura 28: Codigo de Eliminar los datos atípicos.

Mostrar base de datos

N°	w	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
1	3	14	50	233	35871	31.7	3	14589	731	1025	4	476
2	2	11	54	199	66128	6.38	6	30902	1434	2571	1	321
3	52	14	30	130	92891	5.85	12	43176	2318	2621	1	354
4	31	17	42	52	17440	0.99	2	7706	354	1174	3	2734
5	1	18	41	80	105266	99.82	7	42307	2248	1102	4	2559
6	1	14	30	127	92891	5.85	12	43176	2318	2621	1	354
7	17	14	23	88	85670	4.87	8	37856	1737	1174	3	2734
8	1	19	29	121	66128	6.38	6	30902	1434	2571	1	321
9	1	14	30	126	92891	5.85	12	43176	2318	2621	1	354
10	1	14	30	125	92891	5.85	12	43176	2318	2621	1	354
11	9	14	47	193	85670	4.87	8	37856	1737	1174	3	2734
12	1	19	29	117	66128	6.38	6	30902	1434	2571	1	321
13	21	14	30	124	92891	5.85	12	43176	2318	2621	1	354
14	17	14	30	123	92891	5.85	12	43176	2318	2621	1	354
15	1	14	23	89	85670	4.87	8	37856	1737	1174	3	2734
16	1	10	17	30	59600	3.03	11	29310	1407	2143	2	318
17	6	10	18	112	94508	8	8	39160	2146	1137	3	2900
18	2	11	54	201	98325	6	16	43847	2399	1241	4	2362
19	8	14	50	232	35871	31.7	3	14589	731	1025	4	476
20	171	14	47	192	85670	4.87	8	37856	1737	1174	3	2734
21	2	10	19	206	14019	1.04	3	6378	319	1445	3	246
22	2	14	15	3	31133	1.07	7	14592	759	1920	2	1519
23	1	11	55	236	28038	2.08	5	12755	637	1445	3	246
...	...	...	...	...	...	...	...	...	...	...	...	...
5453	71	14	30	123	50751	3.13	7	21929	1217	1358	3	1832
5154	1	14	30	123	73000	22	8	30700	1670	1394	3	1632

Figura 29: Base de datos

Codigo Python para los modelos de Random Forest Regressor y Árbol de decisión regresión.

```
[ ] import matplotlib.pyplot as plt
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.metrics import roc_auc_score
    from sklearn.model_selection import train_test_split
```

Tesis Random Forest Regresor Prueba.

```
[ ] import pandas as pd
    import numpy as np
```

```
[ ] from google.colab import files
    uploaded = files.upload()
```

```
    Elegir archivos  Ningún archivo seleccionado  Upload widget is only available when the cell has been executed in the cu
    Saving datafinal2.xlsx to datafinal2.xlsx
```

```
[ ] import io
    data=pd.read_excel(io.BytesIO(uploaded['datafinal2.xlsx']))
    data
```

```
[ ] X=data
    y=X.pop("y")
```

```
[ ] len(y)
```

```
    5154
```

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=20)
```

```
[ ] model = RandomForestRegressor(n_estimators = 100,oob_score=True, random_state = 20)
```

```
[ ] model.fit(X_train, y_train)
```

```
    RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                          max_depth=None, max_features='auto', max_leaf_nodes=None,
                          max_samples=None, min_impurity_decrease=0.0,
                          min_impurity_split=None, min_samples_leaf=1,
                          min_samples_split=2, min_weight_fraction_leaf=0.0,
                          n_estimators=100, n_jobs=None, oob_score=True,
                          random_state=20, verbose=0, warm_start=False)
```

```
[ ] model.score(X_train, y_train)
```

```
    0.8841541187406778
```

```
[ ] model.score(X_test, y_test)
```

```
    0.8205943048154256
```

```
results=[]
n_estimator_options=[30,50,100,200,500,1000,2000]

for trees in n_estimator_options:
  model = RandomForestRegressor(trees,oob_score=True, random_state = 20)
  model.fit(X_train, y_train)
  print(trees,"trees")
  roc=model.score(X_train, y_train)
  print("C-star:", roc)
  results.append(roc)
  print("")
pd.Series(results,n_estimator_options).plot();
```

Figura 30: Codigo Python para el modelo de Random Forest Regressor.

```
[ ] %matplotlib inline
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns; sns.set()
```

```
[ ] from google.colab import files
    uploaded = files.upload()
```

Elegir archivos Ningún archivo seleccionado Upload widget is only available when the cell has bee

```
Saving datafinal2.xlsx to datafinal2.xlsx
```

```
[ ] print(uploaded)
```

```
{'datafinal2.xlsx': b'PK\x03\x04\x14\x00\x06\x00\x08\x00\x00\x00!\x00!\x8cF:s\x
```

Tesis, Prueba Arbol de Regresion, Prueba con data pequena 100 datos

```
[ ] import pandas as pd
    import numpy as np
```

```
[ ] from google.colab import files
    uploaded = files.upload()
```

Elegir archivos Ningún archivo seleccionado Upload widget is only available when the cell has bee

```
Saving datafinalprueba2.xlsx to datafinalprueba2.xlsx
```

```
    import io
    data=pd.read_excel(io.BytesIO(uploaded['datafinal2.xlsx']))
    data
```

```
[ ] X=data
    y=X.pop("y")
```

```
[ ] len(y)
```

```
5154
```

```
[ ] import matplotlib.pyplot as plt
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.metrics import roc_auc_score
    from sklearn.model_selection import train_test_split
```

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

```
[ ] from sklearn.tree import DecisionTreeRegressor
```

```
[ ] adr=DecisionTreeRegressor(max_depth=5)
```

```
[ ] adr.fit(X_train,y_train)
```

```
DecisionTreeRegressor(ccp_alpha=0.0, criterion='mse', max_depth=5,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort='deprecated',
                      random_state=None, splitter='best')
```

```
[ ] adr.score(X_train,y_train)
```

```
0.3989399662002473
```

```
[ ] adr.score(X_test, y_test)
```

```
0.4312117405922976
```

```
[ ] adr.feature_importances_
```

```
array([0.        , 0.00169295, 0.68569804, 0.        , 0.05784326,
       0.12268455, 0.11162593, 0.00266669, 0.        , 0.0169269 ,
       0.00086168])
```

Figura 31: Codigo Python para el modelo de Árbol de decisión regresión.

Extraer información del Sistema de denuncias policiales SIDPOL.

Figura 32: Sistema de denuncias policiales SIDPOL