



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado

Facultad de Ingeniería de Sistemas e Informática

Unidad de Posgrado

**Modelo híbrido optimizado ARIMA–LSTM basado en
aprendizaje automático para el pronóstico de
denuncias por ciberdelitos contra el patrimonio en el
Perú**

TESIS

Para optar el Grado Académico de Doctor en Ingeniería de
Sistemas e Informática

AUTOR

Manuel Martin MORALES BARRENECHEA

ASESOR

Dr. Ciro RODRIGUEZ RODRIGUEZ

Lima, Perú

2026



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

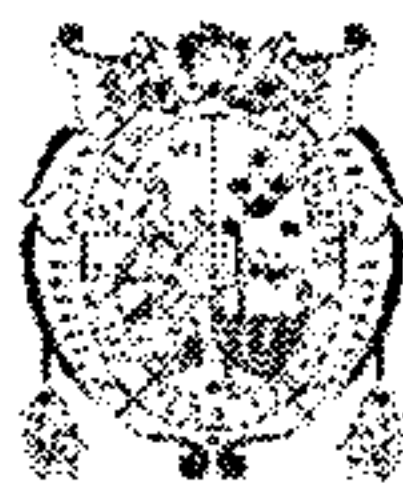
Referencia bibliográfica

Morales, M. (2026). *Modelo híbrido optimizado ARIMA–LSTM basado en aprendizaje automático para el pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú*. [Tesis de doctorado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Manuel Martin Morales Barrenechea
Tipo de documento de identidad	DNI
Número de documento de identidad	09774836
URL de ORCID	https://orcid.org/0009-0000-9482-3670
Datos de asesor	
Nombres y apellidos	Ciro Rodriguez Rodriguez
Tipo de documento de identidad	DNI
Número de documento de identidad	06020241
URL de ORCID	https://orcid.org/0000-0003-2112-1349
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Hugo Froilán Vega Huerta
Tipo de documento	DNI
Número de documento de identidad	06147737
Miembro del jurado 1	
Nombres y apellidos	Pedro Martin Lezama Gonzales
Tipo de documento	DNI
Número de documento de identidad	09656793
Miembro del jurado 2	
Nombres y apellidos	Ivan Carlo Petrlik Azabache
Tipo de documento	DNI
Número de documento de identidad	10140461
Miembro del jurado 3	
Nombres y apellidos	Ciro Rodriguez Rodriguez
Tipo de documento	DNI

Número de documento de identidad	06020241
Datos de investigación	
Línea de investigación	C.0.5.8. Inteligencia Artificial
Grupo de investigación	Tecnologías de la Información aplicados a la ciencia de los datos - ITDATA
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Edificio: Facultad de Ingeniería de Sistemas e Informática País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Calle: Av. Amezaga Latitud: -12.05642315 Longitud: -77.08433269
Año o rango de años en que se realizó la investigación	2024 - 2025
URL de disciplinas OCDE	Ingeniería de Sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

**ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE
DOCTOR EN INGENIERÍA DE SISTEMAS E INFORMÁTICA**

A los veinticuatro (24) días del mes de abril de 2026, siendo las 11:00 am., se reunieron en el Auditorio Alfredo Celso Alva Bravo, el Jurado de Tesis conformado por los siguientes docentes:

Dr. Hugo Froilán Vega Huerta (Presidente)
Dr. Pedro Martin Lezama Gonzales (Miembro)
Dr. Ivan Carlo Petrlik Azabache (Miembro)
Dr. Ciro Rodriguez Rodriguez (Miembro Asesor)

Se inició la Sustentación invitando al candidato a Doctor **MANUEL MARTIN MORALES BARRENECHEA**, para que realice la exposición oral de la tesis para optar el Grado Académico de Doctor en Ingeniería de Sistemas e Informática, siendo la Tesis intitulada:

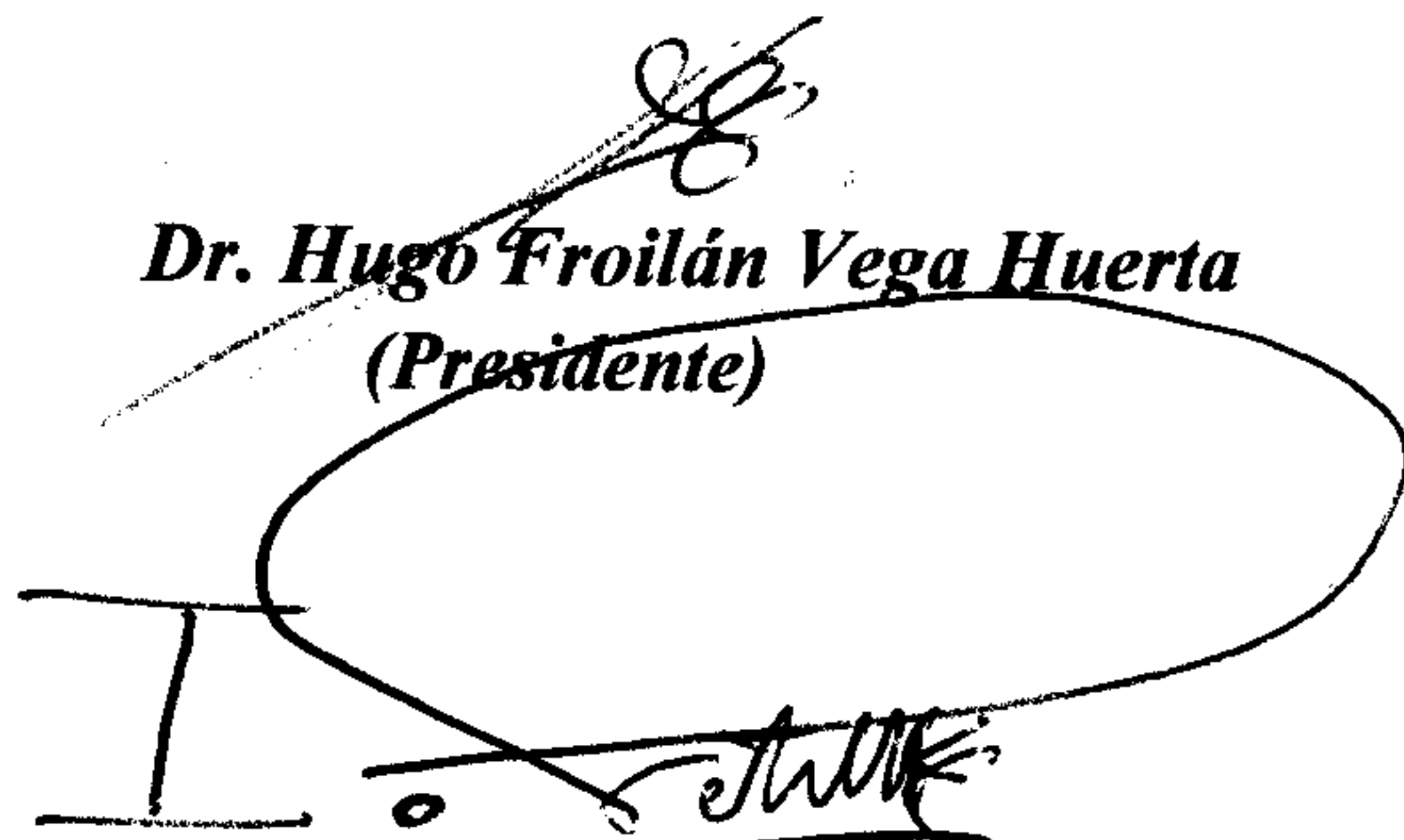
**“MODELO HÍBRIDO OPTIMIZADO ARIMA-LSTM BASADO EN APRENDIZAJE
AUTOMÁTICO PARA EL PRONÓSTICO DE DENUNCIAS POR CIBERDELITOS
CONTRA EL PATRIMONIO EN EL PERÚ”**

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

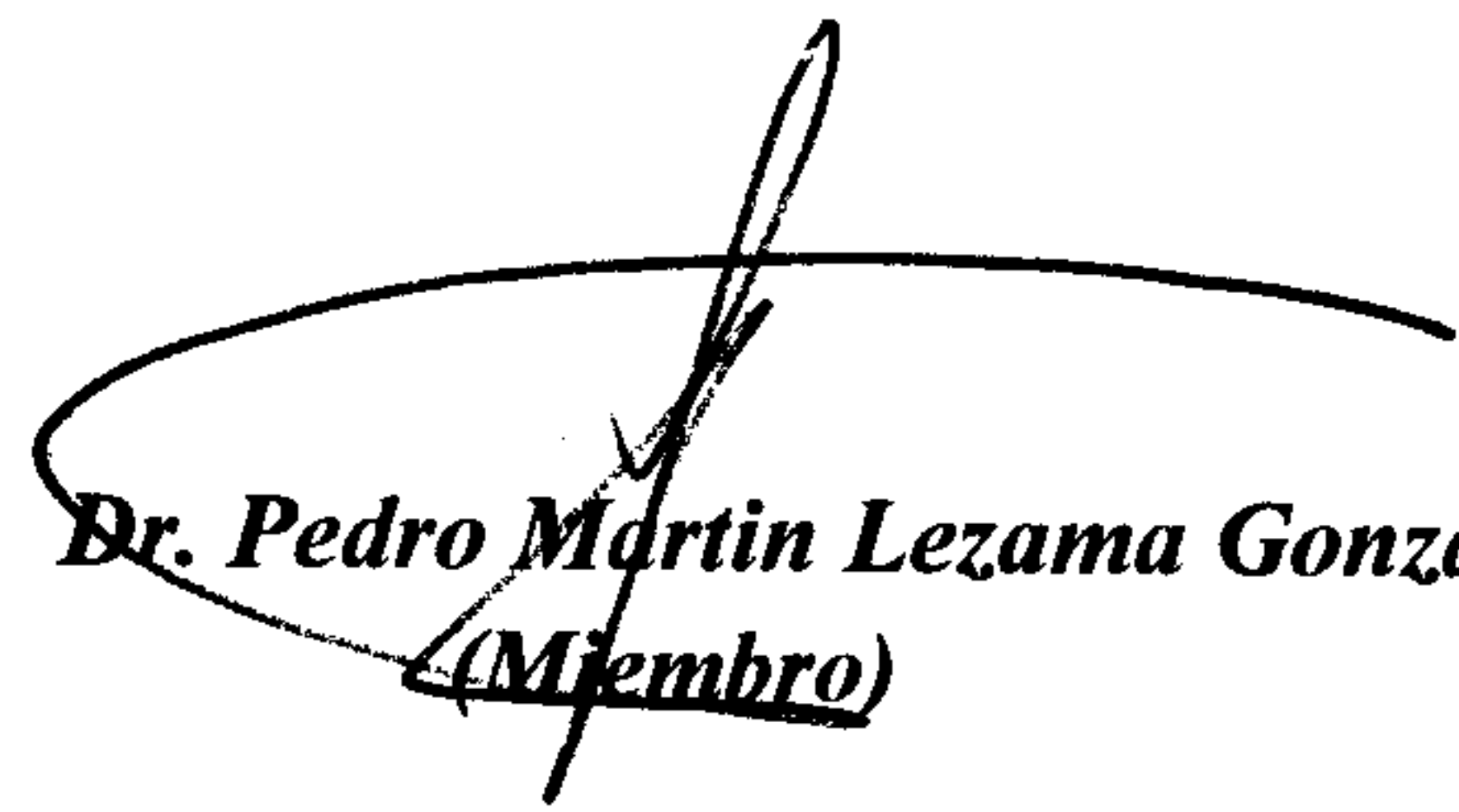
DIECINUEVE (19) EXCELENTE

Por tanto, el presidente del Jurado, de acuerdo con el Reglamento General de Estudios de Posgrado, otorga al Magíster **MANUEL MARTIN MORALES BARRENECHEA** el Grado Académico de Doctor en Ingeniería de Sistemas e Informática.

Siendo las 12 horas, el presidente del Jurado de Tesis, da por concluido el acto académico de Sustentación de Tesis.


Dr. Hugo Froilán Vega Huerta
(Presidente)


Dr. Ivan Carlo Petrlik Azabache
(Miembro)


Dr. Pedro Martin Lezama Gonzales
(Miembro)


Dr. Ciro Rodriguez Rodriguez
(Miembro Asesor)



Anexo 1

CERTIFICADO DE SIMILITUD

Yo, **Ciro Rodriguez Rodriguez** en mi condición de asesor acreditado con Dictamen N°000017-2024-UPG-VDIP-FISI/UNMSM de la tesis, cuyo título es “**Modelo híbrido optimizado ARIMA–LSTM basado en aprendizaje automático para el pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú**”, presentado por el magíster **Manuel Martin Morales Barrenechea** para optar el Grado Académico de Doctor en Ingeniería de Sistemas e Informática,

CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el **5 %** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del Grado Académico de Doctor en Ingeniería de Sistemas e Informática correspondiente.

Lima, 20 de marzo de 2026

Firma del Asesor

Nombres y Apellidos del asesor: **Ciro Rodriguez Rodriguez**

DNI: 06020241

* Según corresponda: bachiller, magíster, doctor; título profesional o título de segunda especialidad

La vigencia del presente certificado corresponde a un período de noventa (90) días desde la fecha de su emisión; siempre que no haya observaciones en el texto durante el proceso de revisión por el jurado informante o surja alguna contingencia.

Índice general

Índice general.....	i
Lista de tablas	iv
Lista de figuras.....	v
Dedicatoria.....	vi
Agradecimiento.....	vii
Resumen.....	viii
Abstract.....	ix
CAPÍTULO 1. INTRODUCCIÓN.....	1
1.1. Situación Problemática	1
1.2. Formulación del Problema.....	6
1.2.1. <i>Problema General</i>	6
1.2.2. <i>Problemas Específicos</i>	6
1.3. Justificación de la Investigación	6
1.3.1. <i>Justificación Teórica</i>	6
1.3.2. <i>Justificación Práctica</i>	7
1.3.3. <i>Justificación Social</i>	8
1.3.4. <i>Justificación Económica</i>	8
1.4. Objetivos de la Investigación.....	9
1.4.1. <i>Objetivo General</i>	9
1.4.2. <i>Objetivos Específicos</i>	9
CAPÍTULO 2. MARCO TEORICO	10
2.1. Marco filosófico o epistemológico de la investigación	10
2.1.1. <i>Marco filosófico</i>	10
2.1.2. <i>Marco epistemológico</i>	11
2.2. Antecedentes de la investigación	12
2.3. Marcos Conceptuales.....	17
2.3.1. <i>Marco Emprendedor</i>	17
2.3.2. <i>Marco Ético</i>	18
2.3.3. <i>Marco Sistémico</i>	19
2.4. Bases Teóricas	20
2.4.1. <i>Ley N° 30096, Ley de Ciberdelitos</i>	20
2.4.2. <i>Modalidades de fraude informático</i>	21
2.4.3. <i>Aprendizaje automático para la predicción del cibercrimen</i>	22
2.4.4. <i>Modelos de aprendizaje automático</i>	23
2.4.4.1. <i>Auto Regressive Integrated Moving Average (ARIMA)</i>	23
2.4.4.2. <i>Long Short-Term Memory (LSTM)</i>	24

CAPÍTULO 3. ESTADO DEL ARTE	27
3.1. Introducción.....	27
3.2. Metodología.....	28
3.2.1. Preguntas de investigación	29
3.2.2. Estrategia de investigación	29
3.2.3. Criterio de elegibilidad.....	30
3.2.4. Fuentes de información.....	31
3.2.5. Proceso de selección de artículos	31
3.2.6. Agrupación automática de artículos	32
3.3. Resultados.....	34
3.4. Conclusiones.....	39
CAPÍTULO 4. METODOLOGIA.....	40
4.1. Hipótesis general.....	40
4.2. Hipótesis específicas.....	40
4.3. Identificación de variables	41
4.4. Operacionalización de variables	42
4.5. Matriz de consistencia.....	42
4.6. Tipo y diseño de la investigación.....	44
4.7. Unidad de análisis.....	44
4.8. Población de estudio	44
4.9. Tamaño de muestra.....	45
4.10. Recolección de datos.....	45
4.11. Confiabilidad de datos	45
4.12. Análisis e interpretación de la información.....	46
4.13. Conjunto de datos	47
4.14. Preprocesamiento de los datos	48
4.15. Optimización por hiperparámetros.....	49
4.16. Modelo propuesto híbrido optimizado ARIMA-LSTM	49
4.17. Métricas de Evaluación de resultados	50
4.18. Metodología CRISP-DM	51
CAPÍTULO 5. RESULTADOS Y DISCUSIÓN	53
5.1. Análisis, interpretación y discusión de los resultados	53
5.1.1. Comprensión de la organización.....	53
5.1.1.1. Policía Nacional del Perú.....	53
5.1.1.2. División de Investigación de Delitos de Alta Tecnología.....	54
5.1.1.3. Sistema Informático de Denuncias Policiales	54
5.1.2. Comprensión de los datos	55
5.1.2.1. Validación de la fiabilidad de los datos	55
5.1.2.2. Análisis exploratorio de los datos	56

5.1.3. Preparación de los datos.....	57
5.1.3.1. Limpieza de datos	57
5.1.3.2. Pruebas de consistencia de los datos	58
5.1.3.3. Estructuración de la serie temporal de datos.....	58
5.1.3.4. Prueba ADF para validar la estacionariedad de las series	59
5.1.3.5. Diferenciación de series.....	60
5.1.4. Modelado	60
5.1.4.1. Configuración de hiperparámetros	61
5.1.4.2. Modelo ARIMA	62
5.1.4.3. Modelo LSTM	63
5.1.4.4. Modelo híbrido ARIMA-LSTM optimizado	65
5.1.5. Evaluación de resultados.....	66
5.2. Prueba de Hipótesis.....	67
5.3. Presentación de resultados	70
CONCLUSIONES	81
RECOMENDACIONES.....	83
REFERENCIAS BIBLIOGRÁFICAS.....	85
ANEXOS	92

Lista de tablas

Tabla 1. Operacionalización de variables	29
Tabla 2. Términos de búsqueda	30
Table 3. Criterio de elegibilidad	31
Tabla 4. Resultados correspondientes a P1	35
Tabla 5. Resultados correspondientes a P2	36
Table 6. Resultados correspondientes a P3	37
Tabla 7. Resultados correspondientes a P4	38
Tabla 8. Operacionalización de variables	42
Tabla 9. Matriz de consistencia	43
Tabla 10. Campos, tipos y descripción del conjunto de datos	47
Tabla 11. Fórmulas de los criterios de evaluación	51
Tabla 12. Configuración de hiperparámetros	61
Tabla 13. Comparación entre el modelo optimizado ARIMA, LSTM y ARIMA-LSTM	62
Tabla 14. Comparación modelo ARIMA vs ARIMA optimizado	63
Tabla 15. Comparación modelo LSTM simple vs LSTM optimizado	64
Tabla 16. Comparación modelo híbrido ARIMA-LSTM vs ARIMA-LSTM optimizado	65
Tabla 17. Resultados modelos ARIMA, LSTM y optimizado ARIMA-LSTM	66
Tabla 18. Comprobación de las hipótesis específicas	69

Lista de figuras

Figura 1. Distribución denuncias de ciberdelitos PNP (2018–2023).....	3
Figura 2. Diagrama de flujo ARIMA.....	24
Figura 3. Diagrama de flujo LSTM	26
Figura 4. Cadena de búsqueda de artículos.....	30
Figura 5. Diagrama de flujo PRISMA de la revisión sistemática	32
Figura 6. Agrupación jerárquica aglomerativa de artículos	33
Figura 7. Proceso de análisis e interpretación de la información.....	46
Figura 8. Diagrama de flujo de preprocesamiento de datos.....	48
Figura 9. Diagrama de flujo modelo híbrido optimizado ARIMA-LSTM	50
Figura 10. Fases de la metodología CRISP-DM.....	52
Figura 11. Sistema de Denuncias Policiales (SIDPOL).....	55
Figura 12. Comparación denuncias por fraude informático PNP y MPFN	55
Figura 13. Datos de denuncias por delito informáticos contra el patrimonio	56
Figura 14. Estadísticos descriptivos del conjunto de datos.....	56
Figura 15. Descomposición de la serie temporal de datos	57
Figura 16. Pruebas de consistencia de los datos	58
Figura 17. Serie temporal denuncias ciberdelitos contra el patrimonio.....	59
Figura 18. Prueba de estacionariedad Augmented Dickey-Fuller.....	59
Figura 19. Gráfico ACF y PACF de la serie original (no estacionaria).....	59
Figura 20. Diferenciación para transformar la serie no estacionaria a estacionaria.....	60
Figura 21. Gráfico ACF y PACF de la serie diferenciada (estacionaria).....	60
Figura 22. Modelo ARIMA optimizado	63
Figura 23. Modelo LSTM optimizado.....	64
Figura 24. Modelo ARIMA-LSTM optimizado	66

Dedicatoria

Para Fatima y Diego, amados hijos.

Agradecimiento

Al Dr. Ciro Rodriguez por la motivación
en la senda de la investigación
y el generoso aporte en este trabajo.

A la Universidad Nacional Mayor de San
Marcos que representa el conocimiento y la tradición
universitaria en el Perú y América.

Resumen

El crecimiento sostenido de los ciberdelitos, en particular aquellos que afectan el patrimonio en el Perú, plantea retos significativos para la prevención y respuesta institucional. Esta investigación tiene como objetivo entrenar un modelo utilizando métodos de aprendizaje automático para maximizar la precisión de la estimación de denuncias con respecto a los delitos cibernéticos contra la propiedad, lo que podría permitir diseñar estrategias efectivas en la lucha contra el cibercrimen. Utilizando la metodología CRISP-DM, desarrollamos un híbrido basado en modelos ARIMA y LSTM. ARIMA fue entrenado para representar los elementos lineales de la serie temporal y LSTM para representar la complejidad de los patrones no lineales. Debido a que la serie original no era estacionaria según la prueba ADF, aplicamos diferenciación para transformar la serie inicial a estacionaria antes de utilizar ARIMA. Adicionalmente, se utilizaron técnicas de optimización de hiperparámetros en ambos modelos para mejorar su rendimiento predictivo. Los resultados muestran que el MAPE del modelo híbrido optimizado ARIMA-LSTM fue del 10.73% que es superior al rendimiento de los modelos individuales. La prueba estadística Diebold–Mariano evidenció diferencias significativas en favor del enfoque combinado. El modelo fue evaluado con datos de 60,378 incidencias reportadas por la Policía Nacional del Perú entre el 2018 y el 2023. Como conclusión, el modelo híbrido propuesto constituye una solución metodológicamente sólida y estadísticamente precisa para el pronóstico de denuncias, aportando una herramienta innovadora que puede ser utilizada por entidades responsables de la investigación criminal y el cumplimiento de la ley, como la Policía Nacional y Fiscalía de la Nación, para anticipar tendencias y fortalecer la lucha contra el cibercrimen.

Palabras clave: modelo, ARIMA, LSTM, híbrido, serie

Abstract

The sustained growth of cybercrimes, particularly those that affect heritage in Peru, poses significant challenges that test the capacity of institutions to prevent and respond effectively. In this context, the aim of this study is to develop a model grounded in machine learning techniques that optimizes the forecast of complaints for cybercrimes against property, contributing to the design of effective approaches to combating cybercrime. To do this, the CRISP-DM framework was applied, implementing a mixed architecture that unites ARIMA and LSTM models. ARIMA was utilized to capture the linear components of the time series, while LSTM allowed complex nonlinear patterns to be modeled. The ADF test confirmed that the original series was not stationary, so differentiation was applied to transform it into stationary before implementing ARIMA. Likewise, hyperparameter calibration was performed on both models to refine their prediction efficiency. Results reveal that the hybrid ARIMA-LSTM approach reached a MAPE of 10.73%, exceeding the accuracy of the separate models. The Diebold–Mariano statistical revealed statistically significant differences favoring the combined approach. The model was evaluated with data from 60,378 records submitted by the Peruvian police between 2018 and 2023. In conclusion, the proposed hybrid model constitutes a methodologically sound and statistically accurate solution for the forecasting of complaints, providing an innovative tool that can be used by entities responsible for criminal investigation and law enforcement, such as the national police and the nation's prosecutor's office, to anticipate trends and strengthen the fight against cybercrime.

Keywords: model, ARIMA, LSTM, hybrid, series

CAPÍTULO 1. INTRODUCCIÓN

1.1. Situación Problemática

Como parte del entendimiento del delito informático o ciberdelincuencia, la Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC, 2022) indica que el cibercrimen constituye un fenómeno complicado el cual abarca acciones ilícitas dirigidas contra las TIC y la comunicación a través de ellas como medio con el propósito de cometer delitos. El Convenio de Budapest (UNODC, 2022) establece cuatro categorías principales de ciberdelitos: el primero son los que afectan la protección, la consistencia y el acceso oportuno a los datos o plataformas informáticas; el segundo los ciberdelitos propiamente dichos; el tercero es aquellos relacionados con contenidos ilícitos; y el cuarto los que infringen la propiedad intelectual o los derechos afines.

La delincuencia es un problema socioeconómico cometido principalmente por delincuentes que afecta negativamente la calidad de vida de la población y el avance económico de una nación (Bhardwaj & Bawa, 2022). Con el incansable avance de la tecnología, el crimen ha ampliado su alcance digitalmente lo que se conoce como cibercrimen. Las denuncias por delitos han mostrado un aumento sostenido en los últimos años, particularmente los casos de ciberdelitos que han tenido un considerable crecimiento desde el 2018 y también la velocidad en la que aumenta la tasa de criminalidad en el mundo digital por lo que lo convierte en un tema de atención relevante (Bhardwaj & Bawa, 2022).

En esta era del mundo cibernético, el ciberdelito se viene extendiendo considerablemente y los países hacen denotados esfuerzos por adaptarse y garantizar la seguridad cibernética (Goel, 2020; Veena, Meena, Kuppusamy, et al., 2022a). El crecimiento exponencial del Internet y uso acelerado de las tecnologías han contribuido con la generación de nuevas modalidades de ciberdelincuencia. Naciones con alto nivel de desarrollo, como el caso de

Inglaterra y Estados Unidos, mantienen una lucha constante contra los ciberataques y los ciberdelitos a través del desarrollo de estrategias y diversos sistemas de seguridad (Reid & Van Niekerk, 2014).

El Foro Económico Mundial, señala que el mundo digital está relacionado con el cibercrimen y cuyo costo mundial fue de US\$ 575 mil millones en el 2016 (Deutscher, 2017). El costo mundial del cibercrimen es del 30% y el gasto en ciberseguridad en el Perú se ubica por debajo al promedio global estimado en 0.12% del Producto Bruto Interno. Esto significaría que el Perú destina menos de la mitad del promedio global para salvaguardar su economía digital. En el año 2016, el gasto del Perú en materia de ciberseguridad fue de UDS 132 millones, un monto considerablemente menor que el de Brasil que destinó 1,431 y México con 1,041 millones de dólares (Deutscher, 2017).

El Ministerio de Justicia y Derechos Humanos (MINJUSDH), a través de su informe “Ciberdelincuencia información estadística y recomendaciones para la prevención” (Ministerio de Justicia y Derechos Humanos, 2022) presenta datos relevantes y orientaciones destinadas a mitigar este tipo de delitos y reportó que la Policía Nacional del Perú (PNP) presentó un incremento considerable de denuncias por ciberdelitos. Según este reporte, en el 2021 se registraron 14,671 denuncias por ciberdelitos lo que refleja un crecimiento del 65 % sobre lo registrado en 2020.

Según la PNP, el número de casos contra la Ley N° 30096 Ciberdelitos registrados en la plataforma de denuncias policiales (SIDPOL) equivalen a un total de 75,440 denuncias a nivel nacional entre el año 2018 y 2023. Las denuncias por el delito informático contra el patrimonio o fraude informático corresponden a 60,378 que equivale al 80% del total de denuncias por ciberdelitos, siendo este tipo de ciberdelito el que se comete con mayor frecuencia en nuestro país (Gobierno del Perú, 2024) como puede observarse en la Figura 1.

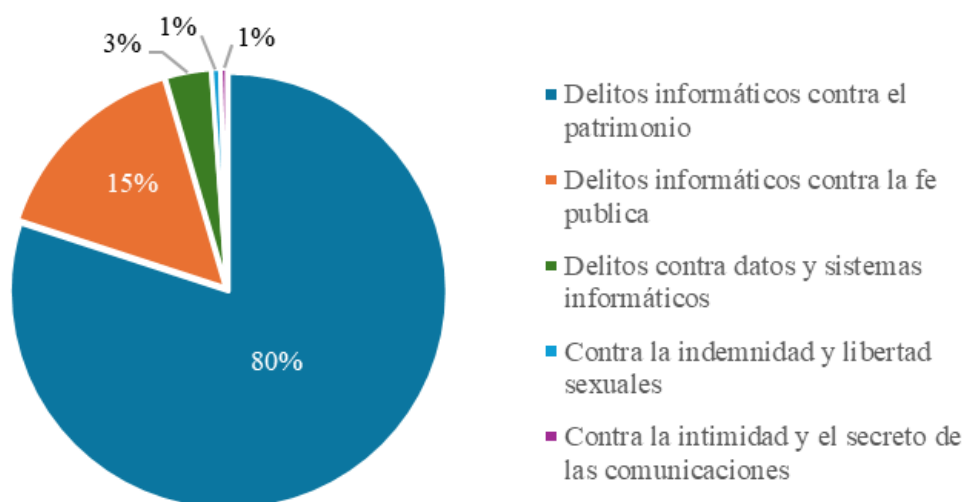


Figura 1. Distribución denuncias de ciberdelitos PNP (2018–2023)

Fuente. Datos tomados de la PNP

El fraude informático es el este es el tipo más frecuente de cibercrimen y de acuerdo con la PNP, el 75 % de los ciberdelitos registrados en el Perú se encuentran categorizados dentro de esta modalidad delictiva (Diario Oficial El Peruano, 2023). El titular de la División de Investigación de Delitos de Alta Tecnología (DIVINDAT), coronel PNP Luis Huamán Santamaría, indicó que las modalidades de engaño a las personas son el robo de información financiera o falsificación de identidad con el propósito de perpetrar distintos actos de ciberdelincuencia (Diario Oficial El Peruano, 2023).

En Perú durante el 2023, el phishing se posicionó como una de las formas de estafa informática más investigadas con 294 denuncias, el carding con 262, la suplantación de identidad con 240, las compras en línea con 91, el thief transfer con 87, el robo de WhatsApp con 84, el SIM swapping con 65, el fake app con 32 y el vishing con 31 denuncias (Agencia Peruana de Noticias, 2024). El porcentaje de personas detenidas en relación con las denuncias por delitos informáticos en el Perú fue de 5.4% en el 2020, 2.4% en el 2021 y 5.8% en 2022 (Diario Oficial El Peruano, 2023).

Conforme al Instituto Nacional de Estadística e Informática en la encuesta nacional sobre victimización la cifra negra del delito, que referencia los incidentes delictivos que perjudicaron a la población pero no llegaron a ser denunciados a las

autoridades y no aparecen en las estadísticas criminales donde el 87.3% de hechos delictivos cometidos no se denunciaron porque las víctimas aseguran que no pasa nada o nada ocurre al denunciar, por considerarlo como hechos de poca importancia, por desconfianza en la policía y por falta de pruebas o falta de tiempo entre los principales motivos (INEI, 2017).

(Defensoría del Pueblo, 2023) en su informe sobre la ciberdelincuencia en el Perú, señala que la DIVINDAT no cuenta con especialistas en el manejo de las TIC y una adecuada infraestructura informática. En ese sentido, solo el 16% del personal está capacitado en ciberseguridad y no más del 5% cuenta con equipos de cómputo de alto rendimiento. Además, el 51% no cuenta con acceso al sistema de denuncias policiales y no se cuenta con presupuesto para implementar tecnologías. A esto se añade la carencia de datos precisos que permitan conocer la verdadera dimensión de las denuncias que permitan establecer acciones acordes a las capacidades policiales para combatir el fraude informático (Defensoría del Pueblo, 2023).

Es responsabilidad de los gobiernos controlar los incidentes delictivos a fin de salvaguardar a los ciudadanos y promover su bienestar (Butt et al., 2020). Los hechos delictivos cometidos por ciberdelincuentes se consideran de ámbito transnacional por lo que las organizaciones públicas y privadas están llamadas a cooperar y contrarrestar estos actos ilícitos en la base de sus leyes regulatorias (Parisi, 2019). La inseguridad ciudadana, en la que se incluye el delito informático, es un tema de relevante atención ya que, de acuerdo con Constitución Política del Perú, el Estado tiene la responsabilidad de salvaguardar a la población frente a cualquier amenaza que ponga en riesgo su seguridad.

El Perú se encuentra adherido al Convenio de Budapest sobre ciberdelincuencia y en marzo del 2019 mediante el Decreto Supremo N° 010-2019-RE, asume el compromiso de lucha decidida contra la ciberdelincuencia en un marco de cooperación internacional (UNODC, 2022). La Ley N° 30096, también llamada Ley de ciberdelitos, se promulgó en octubre del año 2013 y busca prevenir y sancionar los delitos utilizando las TIC que vulneran los sistemas y datos informáticos siendo el principal ciberdelito contra el patrimonio (Ley N° 30096, 2013). Esta norma fue actualizada mediante la Ley N° 30171, publicada el 10 de marzo de 2014.

En el aspecto social, los ciberdelitos generan un efecto considerable sobre los derechos de los ciudadanos, la economía nacional y la protección tanto de los ciudadanos como de las empresas y entidades del Estado según (Defensoría del Pueblo, 2023). En el informe defensorial sobre la delincuencia informática en el Perú donde, según información de la PNP, las denuncias por ciberdelitos contemplados en la Ley N° 30096 se cuadriplicaron entre 2018 y 2021 pasando de 3,031 a 12,827 (Defensoría del Pueblo, 2023). En ese mismo reporte, el 53% del total se concentró en Lima Metropolitana y las provincias de Lima, y el fraude informático con 12,827 denuncias en el 2021 fue el más recurrente con 72%.

Dentro del campo científico, la adopción de la inteligencia artificial para combatir la ciberdelincuencia es cada vez mayor. Las técnicas tradicionales han sentado las bases para la predicción de la ciberdelincuencia (Deepak et al., 2023a). Sin embargo, los modelos tradicionales suelen tener dificultades ante la dinámica cambiante de las ciberamenazas. (Aljarboua et al., 2022a). Por ello, los investigadores sugieren el uso de modelos híbridos para abordar estas limitantes (Deng et al., 2023; Xu et al., 2022). El enfoque híbrido ARIMA–LSTM es una estrategia utilizada para mejorar la precisión de la predicción, brindar información sobre los patrones y tendencias de la delincuencia (Kasemset et al., 2023; Sakshi & A, 2020).

En contraste, en el Perú las investigaciones predictivas en delitos se encuentran en una fase incipiente. Los trabajos existentes aplican técnicas tradicionales de aprendizaje automático (Random Forest, Decision Tree, SVM) principalmente a delitos comunes, como robos o fraudes menores (Chipa Sierra, 2023; Sueldo Alarcón, 2022), sin abordar la dinámica temporal de los ciberdelitos ni utilizar modelos híbridos optimizados. Bajo esa premisa, existen brechas sobre el uso del modelo híbrido ARIMA-LSTM en el campo de la predicción de cibercrimen.

En resumen, en el Perú se evidencia un considerable incremento de denuncias por ciberdelitos con el patrimonio, limitada capacidad operativa de la policía especializada y la ausencia de tecnologías basadas en modelo predictivos en ciberdelincuencia que permitan anticipar su comportamiento. Sin duda, estas brechas limitan a las autoridades peruanas en la planificación de estrategias preventivas basadas en datos para mejorar la ciberseguridad. Ante esta realidad,

resulta necesario construir un modelo híbrido optimizado ARIMA–LSTM que combine la fortaleza del modelado estadístico y del aprendizaje profundo para mejorar la precisión del pronóstico de denuncias de ciberdelitos contra el patrimonio.

1.2. Formulación del Problema

1.2.1. Problema General

PG: ¿De qué manera el desarrollo de un modelo híbrido optimizado ARIMA–LSTM puede mejorar la precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú?

1.2.2. Problemas Específicos

- PE1: ¿Cómo influye la estacionariedad de la serie temporal de denuncias por ciberdelitos contra el patrimonio en la eficiencia y precisión del modelo ARIMA en el pronóstico?
- PE2: ¿En qué medida la integración del modelo ARIMA con el modelo LSTM en un enfoque híbrido optimizado mejora significativamente la capacidad de pronóstico en comparación con los modelos individuales ARIMA y LSTM?
- PE3: ¿Cuál es el nivel de precisión que alcanza el modelo híbrido optimizado ARIMA–LSTM frente a los modelos individuales ARIMA y LSTM a través de la métrica MAPE como indicador de desempeño?

1.3. Justificación de la Investigación

1.3.1. Justificación Teórica

Este estudio investiga la construcción de un modelo basado en aprendizaje

automático para predecir quejas de ciberdelitos dirigidos a la propiedad. Idealmente, el estudio adopta la lógica del aprendizaje automático aplicada al análisis de series temporales, ya que permite la detección de comportamientos complejos en los datos. Esto nos permite trascender las limitaciones de las técnicas estadísticas clásicas al proporcionar una adaptabilidad mucho mayor a las características evolutivas de las acciones delictivas.

El modelo predictivo híbrido propuesto contribuye a nuestra comprensión teórica de cómo se puede utilizar la inteligencia artificial en la seguridad digital. La contribución es demostrar que el uso de modelos estadísticos y neuronales aumenta la precisión en la predicción de ciberdelitos y avanzar en el progreso de la analítica predictiva y el refuerzo de la base conceptual para futuras investigaciones sobre la prevención del delito con tecnologías inteligentes.

1.3.2. Justificación Práctica

Desde un punto de vista práctico, este estudio sirve como un recurso analítico útil para las instituciones encargadas de asegurar y administrar justicia, incluyendo la PNP y el MPFN. Este modelo proporciona la capacidad de anticipar denuncias relacionadas con delitos cibernéticos contra la propiedad, aumentando la capacidad institucional en términos de prevención y gestión de este tipo de delitos y en términos de acelerar la respuesta a las amenazas digitales para los ciudadanos, los sectores empresariales y las entidades estatales.

Además, la versatilidad del modelo permite que se utilice para otras clases de delitos cibernéticos, incluyendo aquellos que involucran la confidencialidad, integridad y disponibilidad de los sistemas informáticos (o datos) y aquellos relacionados con la propiedad intelectual e incluso todo tipo de delitos comunes que ponen en peligro la vida, la libertad y la seguridad pública. Así, el estudio es una contribución práctica replicable para el fortalecimiento de estrategias preventivas y la optimización de la inversión de recursos en el sistema de justicia.

1.3.3. Justificación Social

A nivel social, los delitos cibernéticos representan una amenaza creciente que afecta directamente a individuos, empresas e instituciones. Sus efectos se expresan en la pérdida o manipulación de datos, el robo de información personal y financiera, el fraude económico y la violación de la propiedad intelectual, todo lo cual obstaculiza la confianza digital y la seguridad social.

Estos delitos que generan altos costos sociales y económicos se relacionan con la recuperación de sistemas, la investigación forense y la interrupción de actividades productivas tras un ciberataque. Bajo esa premisa, los modelos predictivos impulsados por técnicas de aprendizaje automático, como el propuesto en esta investigación, contribuyen a mejorar la seguridad digital, ayudando a proteger el bienestar de los ciudadanos mediante la provisión de un entorno tecnológico confiable frente a las amenazas cibernéticas.

1.3.4. Justificación Económica

La seguridad digital se convierte en un activo para el surgimiento y la estabilidad de un país desde la perspectiva económica, ya que el cibercrimen genera una pérdida inmensa que influye tanto en los sectores público como privado. Prever y prevenir tales ataques contribuye directamente a reducir las pérdidas por fraude, interferencias operativas y recuperación de datos, fomentando así la confianza en los entornos digitales y aumentando la competitividad financiera de una economía (Rumi et al., 2018; ToppiReddy et al., 2018).

Según el Banco Interamericano de Desarrollo, las pérdidas económicas del cibercrimen ascienden a aproximadamente 575 mil millones de dólares anuales, lo que equivale al 0,5% del PIB mundial (Barrett et al., 2016). En América Latina, las pérdidas rondan los US\$ 90 mil millones cada año, y se estima que este costo crezca un 15% anual hasta alcanzar los US\$ 10.5 billones en 2025 (Morgan, 2020). En este contexto, el modelo predictivo propuesto en esta investigación

aporta una herramienta económica estratégica, al permitir anticipar patrones delictivos y orientar políticas de inversión en ciberseguridad de manera más eficiente y rentable.

1.4. Objetivos de la Investigación

1.4.1. Objetivo General

OG: Desarrollar un modelo híbrido optimizado ARIMA–LSTM que mejore la precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú.

1.4.2. Objetivos Específicos

- OE1: Lograr la estacionariedad de la serie temporal de denuncias por ciberdelitos contra el patrimonio, con el propósito de mejorar la eficiencia y precisión del modelo ARIMA en el pronóstico.
- OE2: Diseñar y evaluar un modelo híbrido optimizado ARIMA–LSTM que integre las ventajas de ambos enfoques para determinar si su combinación mejora significativamente la capacidad de pronóstico en comparación con los modelos individuales ARIMA y LSTM.
- OE3: Validar el rendimiento del modelo híbrido optimizado ARIMA–LSTM mediante la métrica MAPE, a fin de determinar el nivel de precisión alcanzado respecto a los modelos individuales ARIMA y LSTM en el pronóstico de denuncias por ciberdelitos contra el patrimonio.

A partir de los objetivos planteados, se establecen las hipótesis de la investigación, empleando métodos de estadística descriptiva. Esta disciplina es utilizada en el ámbito científico como base para el análisis y contraste de hipótesis dentro del marco de la estadística inferencial (Ritchey, 2014).

CAPÍTULO 2. MARCO TEORICO

El marco teórico representa el sustento conceptual que guía, fundamenta y delimita el desarrollo de una investigación científica, al permitir vincular el problema de estudio con teorías y modelos previamente desarrollados (R. Hernández et al., 2014). Para este estudio, resulta indispensable comprender el marco legal, las modalidades más recurrentes de ciberdelitos y los avances tecnológicos aplicados al pronóstico de ciberdelitos (Wall, 2024).

2.1. Marco filosófico o epistemológico de la investigación

Las bases filosóficas sustentan que la realidad objetiva puede conocerse a través de los datos. Por ello, el uso del aprendizaje automático se justifica para predecir patrones en denuncias de ciberdelitos. El marco epistemológico define cómo se construye y valida el conocimiento en la investigación. Ambos marcos se complementan facilitando la aplicación de la tecnología como medio para abordar problemas complejos, producir conocimiento útil y promover soluciones replicables.

2.1.1. *Marco filosófico*

Las bases filosóficas de la realidad o fundamentación ontológica proponen la idea de que la realidad objetiva puede conocerse a través de datos. Según (X. Wang et al., 2024), para ofrecer soluciones cercanas a la realidad objetiva es necesario emplear herramientas adecuadas. Entonces, se justifica científicamente el uso del aprendizaje automático en el pronóstico de ciberdelitos puesto que los

datos de las denuncias capturan los patrones de los ciberdelitos.

Esta investigación se basa en una corriente filosófica de tipo positivista, al asumir que la realidad puede ser observada, medida y modelada objetivamente (Deichmann, 2024). Desde esta perspectiva, se considera que los fenómenos sociales, como los ciberdelitos, presentan patrones observables en el tiempo y pueden ser analizados estadísticamente y pronosticados mediante modelos de aprendizaje automático.

Si bien el delito informático es un fenómeno social complejo, porque tiene impacto a nivel personal, empresarial y social, presenta comportamientos repetitivos que pueden ser detectados, modelados y pronosticados basados en la evidencia empírica disponible de las denuncias de la PNP como organismo encargado del combate contra la ciberdelincuencia.

2.1.2. Marco epistemológico

El marco epistemológico sustenta la generación del conocimiento derivado del proceso investigativo, al definir cómo este contribuye a la solución del problema y asegura la confiabilidad y rigurosidad de los hallazgos alcanzados. Se resalta la necesidad de contar con datos consistentes y variables bien definidas y alineadas al problema investigado para orientar el entrenamiento del modelo hacia la precisión del pronóstico de denuncias de ciberdelitos (Pérez-Escobar & Sarikaya, 2024).

Desde el enfoque epistemológico, esta investigación adopta un enfoque empírico cuantitativo, sustentada en el empleo de algoritmos de aprendizaje automático y métodos estadísticos aplicados a una serie temporal histórica de denuncias de ciberdelitos. El conocimiento válido sobre el comportamiento del delito cibernético puede generarse mediante el análisis sistemático de datos históricos estructurados y normalizados.

El modelo híbrido ARIMA-LSTM optimizado refleja una construcción de conocimiento basado en la objetividad, la verificabilidad y la validación empírica mediante métricas como MAE, RMSE y MAPE. La evolución del conocimiento científico en contextos sociales justifica el uso de técnicas de optimización y

pruebas estadísticas (ADF y Diebold-Mariano) asegurando la solidez metodológica, la fiabilidad del modelo y la reproducibilidad de este estudio.

2.2. Antecedentes de la investigación

Las técnicas tradicionales del aprendizaje automático han servido como base para predecir ciberdelitos (Deepak et al., 2023a). Modelos tradicionales como Decision Tree, Random Forest, Support Vector Machine han sido ampliamente empleadas en la predicción delictiva por su rendimiento en el procesamiento y análisis de datos estructurados de forma efectiva y proporcionar resultados interpretables (Bhardwaj & R.K., 2022a; Vijayalakshmi & Norbu, 2023a). No obstante, estos modelos suelen presentar limitaciones (Aljarboua et al., 2022a) por lo que diversos autores proponen emplear modelos híbridos para superarlas (Deng et al., 2023; Xu et al., 2022).

La combinación de los modelos ARIMA y LSTM mejora la precisión del pronóstico al aprovechar sus puntos fuertes en el manejo de patrones lineales y no lineales en datos de series temporales (Kasemset et al., 2023; Sakshi & A, 2020). Este modelo combinado es recientemente utilizado como una opción superior en muchas tareas de pronóstico debido a su capacidad para reducir el error y la varianza en comparación a los modelos tradicionales e individuales (Kasemset et al., 2023; Soy Temür & Yıldız, 2021) que se han aplicado en diversos sectores.

En el frente de la criminalidad, el modelo híbrido ARIMA–LSTM se ha explorado en la predicción de delitos que demostraron que esta combinación daba como resultado predicciones de la tasa de criminalidad más precisas en comparación con los modelos tradicionales (Natarajan, 2024; Vijayalakshmi & Norbu, 2023b). Con el propósito de identificar antecedentes relevantes en este estudio, se llevó a cabo un análisis sistemático de investigaciones previas ligadas al uso práctico de la IA para pronosticar la tasa de ciberdelitos (Barrenechea & Lengua, 2025). A partir de este proceso, se identificaron los siguientes hallazgos:

(Hussain et al., 2021) en su estudio **“Aplicación del aprendizaje automático en la predicción de delitos”** aplicó modelos tradicionales de aprendizaje automático con un enfoque clasificatorio para predecir zonas de alta

probabilidad de criminalidad en la India. Se emplearon modelos como K-Nearest Neighbour, Naïve Bayes, Decision Tree y Boosted Decision Tree, sobre un conjunto de datos delictivos históricos. El preprocesamiento, el análisis de los datos, el modelo y la evaluación de desempeño de los modelos fueron parte del desarrollo de la metodología del estudio. Los resultados mostraron que el algoritmo Boosted Decision Tree alcanzó la mayor precisión con un 96.0%, seguido por KNN (84.1%), Decision Tree (83.6%) y Naïve Bayes (77.8%). Se concluyó que los algoritmos de ensemble como Boosted Decision Tree son más eficaces para identificar patrones delictivos en comparación con técnicas individuales, aportando valor significativo para la planificación de políticas de seguridad pública.

(Zhang et al., 2022) en su estudio **“Comparación de técnicas de aprendizaje automático para pronosticar zonas delictivas”** analizó la eficacia mostrada por diversos modelos de aprendizaje automático para identificar áreas con alta incidencia delictiva en la ciudad estadounidense de San Francisco, utilizando un conjunto de datos públicos de delitos. Se analizaron 878,049 registros geolocalizados. Se utilizaron diversos enfoques de modelado clásico, entre ellos Decision Trees, Logistic Regression, Random Forests, Naive Bayes y XGBoost. Entre los resultados más destacados, XGBoost obtuvo el mejor desempeño con un accuracy de 87.5% y un F1-Score de 0.869, superando ampliamente a modelos más simples como Naive Bayes con un accuracy de 64.3%. La principal conclusión es que los modelos basados en técnicas de ensemble, como XGBoost y Random Forest, ofrecen predicciones más precisas en escenarios complejos como la predicción de zonas delictivas, lo que justifica explorar modelos avanzados como ARIMA-LSTM para captar componentes temporales y no lineales.

(Xu et al., 2022) dentro de su estudio titulado **“Aplicación de un modelo híbrido ARIMA-LSTM basado en el SPEI para la predicción de sequías”** mejoró la precisión del pronóstico de sequías a corto plazo mediante la aplicación de un modelo híbrido ARIMA-LSTM optimizado, utilizando como variable el indicador SPE, que combina precipitación y evapotranspiración. Se compararon los modelos individuales ARIMA, LSTM y el modelo híbrido propuesto, evaluando su desempeño con series temporales mensuales entre 2001 y 2020. El modelo híbrido

obtuvo valores de error de $MAE = 0.1869$, $RMSE = 0.2331$ y $MAPE = 1.4728$, superando ampliamente al modelo ARIMA (MAE de 0.2923, $RMSE$ de 0.3789 y $MAPE$ de 2.6312) y al LSTM (MAE de 0.2225, $RMSE$ de 0.2717 y $MAPE$ de 2.0262). Se concluyó que el modelo híbrido ARIMA-LSTM presenta una mayor capacidad para capturar tanto patrones lineales como no lineales de la serie, logrando una mejora cuantitativa significativa en los errores de predicción, por lo que resulta recomendable para escenarios complejos y con variabilidad climática o estructural.

(Soy Temür & Yıldız, 2021) en su estudio **“Comparación del rendimiento de pronóstico de los modelos ARIMA, LSTM e híbrido para el presupuesto de volumen de ventas de una empresa manufacturera”** evaluó el desempeño de los modelos ARIMA, LSTM y el modelo híbrido ARIMA-LSTM con el objetivo de predecir las ventas con un nivel de error mínimo para la planificación futura de la producción y las ventas. El MAE obtenido para el modelo híbrido fue aproximadamente un 40.5% mejor que el modelo ARIMA y un 52% mejor que el modelo LSTM. En cuanto a $RMSE$, para el conjunto de entrenamiento del 67%, los valores del $RMSE$ fueron de 825.742 para el modelo ARIMA, 1,042.470 para el LSTM y 746.022 para el modelo híbrido. Según el criterio de éxito de $MAPE$, el modelo híbrido alcanzó un valor 0.133, que fue significativamente inferior al 0.254 de ARIMA y al 0.245 de LSTM. Esto se traduce en un resultado de 43.2% mejor que los de ARIMA y un 45.7% mejores que los de LSTM. En resumen, el modelo híbrido siempre proporcionó previsiones de ventas más precisas y fiables, con tasas de error más bajas en los modelos MAE , $RMSE$ y $MAPE$, lo que demuestra su eficacia a la hora de modelar datos de series temporales complejas mediante la combinación de componentes lineales y no lineales.

(Kasemset et al., 2023) en su estudio **“Comparación de modelos de pronóstico del precio del chalote entre diversas técnicas”** comparó el rendimiento de varios modelos predictivos, incluyendo ARIMA, LSTM y su combinación híbrida ARIMA-LSTM, para la predicción de precios del chalote en Indonesia. Se utilizó una serie temporal de precios mensuales y se realizó la partición del dataset asignando el 80 % para el entrenamiento y el 20 % restante para las pruebas. Se obtuvo que el modelo híbrido ARIMA-LSTM superó a los modelos individuales, obteniendo un MAE de 1.3117, un $RMSE$ de 1.5646 y un

MAPE de 2.78%. En comparación, el modelo ARIMA alcanzó un MAPE de 7.25% y el LSTM un MAPE de 5.72%. La principal conclusión fue que la combinación de modelos permite capturar tanto los patrones lineales como no lineales de la serie, logrando así una mayor precisión en el pronóstico.

(Sakshi & A, 2020) en su estudio **“Un modelo híbrido ARIMA-LSTM para la predicción del mercado de valores utilizando datos en vivo”** plantea un modelo híbrido con el objetivo de incrementar la exactitud en la predicción de datos bursátiles en tiempo real. Para el experimento, se utilizó un conjunto de datos en vivo del mercado bursátil del S&P BSE Sensex (India) con una frecuencia diaria. Se compararon los modelos individuales ARIMA y LSTM frente al modelo híbrido. Los hallazgos indicaron que el modelo ARIMA-LSTM obtuvo un desempeño superior en comparación con los modelos individuales en todas las métricas evaluadas, el MAE se redujo a 0.0061, el RMSE a 0.0101 y el MAPE a 0.074%, evidenciando mejoras significativas en precisión. La conclusión principal es que la combinación secuencial de ARIMA para modelar la tendencia y LSTM para predecir los residuos permite capturar mejor la dinámica del comportamiento temporal. Este enfoque híbrido resulta relevante como antecedente directo para investigaciones sobre pronóstico de eventos complejos como los ciberdelitos, donde conviven patrones lineales y no lineales.

(Deng et al., 2023) en su estudio **“Un modelo híbrido ARIMA-LSTM optimizado por retropropagación en el pronóstico de visitas ambulatorias”** con el propósito de incrementar la exactitud en la predicción de visitas ambulatorias mediante un modelo híbrido optimizado ARIMA-LSTM con retropropagación, que combina la capacidad de ARIMA para capturar patrones lineales, LSTM para componentes no lineales y la optimización de parámetros mediante retropropagación. El experimento utilizó datos mensuales de visitas ambulatorias entre enero de 2017 y diciembre de 2019 en un hospital en China. Se compararon tres modelos: ARIMA, LSTM y ARIMA-LSTM-BP. Los resultados mostraron que el modelo híbrido obtuvo $MAE = 99.59$, $RMSE = 120.86$ y $MAPE = 7.51\%$, frente a valores más altos en los modelos individuales: el modelo LSTM tuvo un MAPE de 8.85% y el ARIMA de 10.79%. En conclusión, los hallazgos demuestran que la integración de métodos lineales, no lineales y técnicas de optimización permite mejorar significativamente la precisión del pronóstico en series temporales del

ámbito de la salud, lo que sustenta su aplicabilidad en otros dominios como la predicción de ciberdelitos.

(Jin et al., 2022) en su estudio **“Predicción de datos de COVID-19 mediante un modelo de pronóstico híbrido ARIMA-LSTM”** propusieron un modelo combinado para la predicción de la propagación y control del COVID-19 comparado con el modelo tradicional de regresión vectorial de soporte (SVR). Los resultados experimentales, sobre un conjunto de datos de casos confirmados en China del 1 de enero de 2021 al 10 de octubre de 2022 obtenidos de la Universidad de Hopkins, demostró para el modelo ARIMA-LSTM una precisión de predicción superior con un MSE de 4,049.913, un RMSE de 63,639, un MAPE de 0.205 y un MAE de 44,320 comparado con los resultados del modelo SVR con un MSE de 4,380.089, un RMSE de 66,188, un MAPE de 0.229 y un MAE de 45,744. En conclusión, esta investigación desarrolló con éxito un modelo híbrido sólido para la predicción de la COVID-19 que ofrece una mayor precisión para respaldar la toma de decisiones dentro del ámbito de la salud pública

(Natarajan, 2024) en su estudio **“Predicción de la tasa de delincuencia utilizando ARIMA y LSTM combinados”** desarrollaron un modelo híbrido ARIMA-LSTM con el objetivo de incrementar la exactitud en la estimación de las tasas de criminalidad. Se empleó un conjunto de datos históricos sobre delitos en distintas ciudades de la India y considerando 13 tipos de delitos desde el 2001 al 2013 obtenidos de las oficinas estatales de registro de delitos. Se aplicó el modelo ARIMA para capturar patrones lineales, y usar los residuos resultantes como entrada para el modelo LSTM, especializado en aprender componentes no lineales. Las métricas utilizadas para evaluar el desempeño fueron MAE, RMSE y MAPE. Esta investigación indicó que el modelo híbrido logró una precisión predictiva significativamente mayor que los modelos individuales. La combinación secuencial ARIMA y LSTM permitió una mejor captura de la estructura compleja de las series temporales de delitos, lo que respalda aún más su uso en un sistema de pronóstico más robusto en el campo de la seguridad pública.

(Sharmin et al., 2022) en su estudio **“Una investigación sobre la predicción de delitos mediante ARIMA automático y LSTM apilado”** compararon el rendimiento de modelos de predicción delictiva utilizando el

modelo ARIMA y el modelo LSTM apilado con datos mensuales de delitos registrados en Chicago entre los años 2001 y 2019. Con el modelo ARIMA se capturó los componentes lineales mientras que con el modelo LSTM apilado los patrones no lineales y secuenciales. En los resultados, el modelo Stacked LSTM superó al ARIMA, obteniendo un RMSE de 37.15 frente a 39.73, MAE de 29.51 frente a 31.32 y MAPE de 9.4% frente a 10.1%, respectivamente. La investigación concluye que los modelos basados en aprendizaje profundo pueden mejorar la precisión del pronóstico criminal en comparación con enfoques estadísticos clásicos. Este estudio refuerza la utilidad de modelos híbridos que combinan lo mejor de ambos enfoques para mejorar las predicciones de series temporales delictivas.

2.3. Marcos Conceptuales

2.3.1. *Marco Emprendedor*

Las oportunidades son la clave para el emprendimiento y donde reconocer y explotar estas oportunidades es parte importante del proceso emprendedor (Shane & Venkataraman, 2000), el rol del emprendedor es relevante para desarrollar la práctica y la competitividad de nuevos emprendimientos donde existe una estrecha relación de oportunidades y recursos que interactúan en un entorno cambiante y difícil que ofrece diversas oportunidades, pero se tienen limitaciones de recursos (Cai et al., 2018).

El trabajo de investigación realizado se considera un emprendimiento dado que recoge una oportunidad innovadora para desarrollar una arquitectura de aprendizaje automático para pronosticar denuncias de ciberdelitos contra el patrimonio. Se aborda la situación problemática planteada como una oportunidad de contribuir en la lucha del fraude informático contra el patrimonio que aqueja a la sociedad considerando las limitaciones de recursos informáticos y tecnológicos de las entidades públicas que gestionan el orden interno como la PNP y el MPFN.

2.3.2. *Marco Ético*

La manera en que se entiende la ética aplicada a la inteligencia artificial sigue siendo motivo de debate a pesar de que, durante los últimos años, distintas entidades internacionales han difundido diversos lineamientos y marcos de principios (Jobin et al., 2019). Los principios éticos brindan una orientación sobre lo que debe o no hacerse tanto en la construcción como en la implementación de sistemas basados en IA por parte de todos aquellos que participan activamente en el ciclo de vida de un sistema de IA. Los principios éticos (OECD, 2019) adoptados por los países, incluido el Perú, siguen los siguientes principios:

- Desarrollo equitativo, sostenibilidad y bienestar social: la IA debe ser administrada de manera responsable, procurando generar beneficios tanto para las personas como para el entorno natural. Su implementación debe motivar la creatividad y el potencial de las personas, favorecer la integración de grupos minoritarios y reducir las desigualdades económicas y sociales (OECD, 2019).
- Principios centrados en el ser humano y orientados a la justicia social: La IA debe desarrollarse con un compromiso con el estado de derecho y los derechos humanos, así como con principios democráticos como la dignidad, la autonomía, la libertad, la privacidad, la protección de datos, la diversidad, la equidad, la justicia social y los derechos de los trabajadores. Además, las personas que diseñan sistemas inteligentes necesitan establecer mecanismos que garanticen y salvaguarden los derechos fundamentales, incluida la autodeterminación individual (OECD, 2019).
- Transparencia y explicabilidad: este principio se refiere a un esfuerzo por hacer que los sistemas inteligentes sean comprensibles y accesibles para todos los involucrados. Necesita describirse en términos muy claros del contexto, la función y la base teórica, con el fin de proporcionar una imagen general de su utilización/alcance. Así, apunta a (i) explicar cómo operan los sistemas de IA, (ii) iluminar a los usuarios y partes interesadas sobre cómo están interactuando con estos sistemas; (iii) asegurar que se conozcan los posibles impactos del resultado generado, y (iv) permitirles desafiar o refutar

conclusiones o pronósticos, para que podamos acceder a los factores y la lógica subyacente de una manera más comprensible (OECD, 2019).

- **Robustez, seguridad y protección:** estos son pilares fundamentales de cualquier sistema de IA por muchas razones: (i) los sistemas deben ser capaces de funcionar de manera segura, confiable y segura, incluso cuando se les desafíe o explote, evitando que representen un riesgo para las personas o el medio ambiente, (ii) la trazabilidad de los datos, procesos y decisiones debe mantenerse a lo largo del ciclo de vida del sistema (de modo que el resultado coincida con las mejores prácticas de vanguardia, los resultados recibidos y las preguntas planteadas sean evaluadas) y (iii) la gestión de riesgos debe aplicarse de manera contextualizada a las etapas individuales del desarrollo y uso del sistema para reducir las posibles vulnerabilidades en cuanto a privacidad, sesgos y seguridad digital (OECD, 2019).
- **Responsabilidad de todos los involucrados en el desarrollo, implementación y uso de sistemas inteligentes:** los involucrados deben asegurarse de que los sistemas funcionen correctamente y que lo hagan de manera ética. Esta responsabilidad también debe cumplirse en alineación consistente con sus responsabilidades, contexto y el estado del arte, para que las decisiones y resultados de la IA puedan ser rastreados, justificados y alineados con el interés público (OECD, 2019).

2.3.3. *Marco Sistémico*

El marco teórico sistémico plantea como un enfoque interdisciplinario que integra las ciencias naturales y sociales, ofreciendo un marco común para analizar la complejidad de los sistemas y formar científicos capaces de comprender sus múltiples dimensiones (Bertalanffy & Almela, 1976). También es una forma científica y sistemática de representación de la realidad y una práctica al trabajo interdisciplinario con los objetivos (Arnold Cathalifaud, Marcelo; Osorio, 1998): (i) desarrollar una terminología para describir características, funciones y comportamientos sistémicos, (ii) desarrollar leyes para estos comportamientos e (iii) impulsar la formalización de estas leyes.

Las aplicaciones de la TGS son ilimitadas y aplicadas en situaciones humanas, sociales y culturales, provenientes de los sistemas biológicos (seres vivos) y los sistemas creados por el ser humano (máquinas o artefactos tecnológicos) y los sistemas generales que se clasifican: (i) según su entitatividad se agrupan en reales (presumen existencia), ideales (construcciones simbólicas) y modelos (abstracciones de la realidad), (ii) según su origen pueden ser naturales o artificiales y (iii) según el grado de aislamiento pueden ser abiertos o cerrados objetivos (Arnold Cathalifaud, Marcelo; Osorio, 1998).

2.4. Bases Teóricas

2.4.1. *Ley N° 30096, Ley de Ciberdelitos*

La Ley N° 30096, vigente desde el 22 de octubre de 2013, fue creada con la finalidad de enfrentar los delitos realizados a través de las tecnologías de la información y comunicación, los cuales afectan sistemas y datos informáticos, siendo el principal ciberdelito aquel que atenta contra el patrimonio. Posteriormente, dicha norma fue actualizada por la Ley N° 30171 el 10 de marzo de 2014, con el propósito de adecuar su contenido a las diversas conductas delictivas contempladas en el Convenio de Budapest sobre Ciberdelincuencia, al que el Perú se sumó el 9 de marzo de 2019 mediante el Decreto Supremo N° 010-2019-RE.

Según lo establecido en la Ley N° 30096, modificada por la Ley N° 30171, se sanciona a toda persona que utilice medios tecnológicos o los sistemas informáticos con la intención de obtener algún tipo de ventaja indebida o ganancia ilegal en perjuicio de terceros. Esta disposición incluye conductas como la modificación, el borrado, la duplicación o cualquier forma de manipulación de información, además de acciones que perturben o dañen el funcionamiento normal de un sistema informático. Además, contempla penas más severas cuando el daño recae sobre bienes del Estado, especialmente aquellos destinados a fines sociales o asistenciales (Ley N° 30096, 2013).

2.4.2. *Modalidades de fraude informático*

El jefe de la División de Investigación de Delitos de Alta Tecnología (DIVINDAT), el coronel PNP Luis Huamán Santamaría, señaló las modalidades del fraude informático más investigadas en el territorio peruano (Diario Oficial El Peruano, 2023):

- El *phishing* es la forma de ingeniería social (truco psicológico para convencer a una persona de hacer algo que no debería) más conocida (Muncaster, 2022). Los cibercriminales se hacen pasar por entidades legítimas (amigos, organizaciones, empresas, etc.) a través del envío de emails o mensajes de texto que contienen un enlace o archivo con contenido malicioso buscan obtener información privada para acceder a cuentas personales de las víctimas para suplantar identidades y generar pérdidas económicas. Según el Instituto Nacional de Ciberseguridad, los atacantes informáticos envían mensajes a los usuarios simulando provenir de una fuente legítima, con el propósito de engañarlos y persuadirlos a ejecutar acciones que comprometan su información personal (INCIBE, 2023).
- El *carding* es un tipo de fraude informático que busca obtener datos de tarjetas bancarias de débito o crédito de manera ilícita para que luego los ciberdelincuentes realicen compras virtuales por internet. Previamente los ciberdelincuentes usan otras técnicas de fraude como el *phishing* y *vishing* (modalidad de phishing con uso de la voz) para obtener los datos de las tarjetas (BBVA, 2023). Según el Instituto Nacional de Ciberseguridad (INCIBE, 2023), los delitos de *carding* se incrementan en temporadas de mayor compra por Internet como navidad o campañas específicas con la finalidad que los cargos fraudulentos en las tarjetas bancarias pasen desapercibidos por la alta cantidad de compras que realizan los usuarios.
- El *SIM swapping* que, según el Instituto Nacional de Ciberseguridad, el SIM swapping o fraude de la SIM duplicada, los ciberdelincuentes llevan a cabo la duplicación fraudulenta de la tarjeta SIM de un dispositivo móvil mediante

la suplantación de identidad del titular. Una vez que la víctima pierde el servicio telefónico, los atacantes obtienen acceso a su información personal y asumen el control de diversas aplicaciones, redes sociales, correos electrónicos o servicios de banca digital. De esta manera, aprovechan los mensajes SMS de verificación para confirmar contraseñas y ejecutar delitos informáticos relacionados con operaciones bancarias o suplantaciones de identidad (INCIBE, 2023).

- El *thief transfer* que según el Instituto Nacional de Ciberseguridad (INCIBE, 2023), *el thief transfer* es una modalidad de delito informático y estafa a través de teléfonos móviles robados donde los delincuentes utilizan el móvil de la víctima para realizar transferencias de dinero usando las apps de banca móvil. Además, los delincuentes establecen comunicación con los contactos registrados en el móvil para solicitar con engaños depósitos de dinero o apoyo de emergencia.

2.4.3. *Aprendizaje automático para la predicción del cibercrimen*

El uso de la inteligencia artificial para combatir la ciberdelincuencia es cada vez mayor. Los investigadores emplean, cada vez más, técnicas de aprendizaje automático orientadas a examinar y anticipar patrones de comportamiento asociados al cibercrimen (Natarajan, 2024), alcanzando elevados niveles de exactitud en los procesos de detección y pronóstico (Bhardwaj & R.K., 2022a). Las técnicas tradicionales han sentado las bases para la predicción de la ciberdelincuencia (Deepak et al., 2023a). Los modelos tradicionales como Decision Tree, Random Forest, Support Vector Machine han sido muy empleados en la predicción de delitos gracias a su capacidad para procesar y analizar datos estructurados y proporcionar resultados interpretables (Bhardwaj & R.K., 2022a; Vijayalakshmi & Norbu, 2023a). Sin embargo, los modelos tradicionales suelen tener dificultades ante la dinámica cambiante de las ciberamenazas. (Aljarboua et al., 2022a). Por ello, los autores sugieren el uso de modelos híbridos para abordar estas limitantes (Deng et al., 2023; Xu et al., 2022).

Al mejorar la precisión del pronóstico utilizando ARIMA junto con

LSTM, encontramos que ambos modelos funcionan bien al tratar con los patrones lineales y no lineales de una serie temporal y, por lo tanto, pueden ayudarnos a hacer mejores predicciones (Kasemset et al., 2023; Sakshi & A, 2020). El modelo ARIMA trabaja con éxito las tendencias lineales y la estacionalidad (He, 2023; Sakshi & A, 2020), mientras que el modelo LSTM sobresale por modelar relaciones no lineales y dependencias a largo plazo (Kasemset et al., 2023; Soy Temür & Yıldız, 2021). La fusión de ambos modelos es más informativa para el análisis de series temporales, lo que puede mitigar la debilidad del modelado cuando se utilizan de forma independiente (He, 2023; Liu et al., 2022). En conjunto, en cuanto al enfoque híbrido ARIMA-LSTM, el método híbrido ARIMA-LSTM se considera como el marco atractivo para obtener predicciones más precisas en todos los campos (Kasemset et al., 2023; Sakshi & A, 2020) y ofrece información sobre modelos de crimen, patrones de crimen y tendencias de crimen.

2.4.4. Modelos de aprendizaje automático

2.4.4.1. Auto Regressive Integrated Moving Average (ARIMA)

El modelo ARIMA constituye una herramienta estadística muy utilizado en el análisis y la predicción de datos pertenecientes a series temporales (Liu et al., 2022) y es también conocido como el modelo Box Jenkins (H. Wang & Zhang, 2023). El modelo ARIMA(p, d, q) es un modelo completo que incluye tanto el modelo autorregresivo AR (Auto Regressive) como el modelo de promedio móvil MA (Moving Average) en el modelado de series temporales (Liu et al., 2022). En los parámetros, p es el término autorregresivo (AR), d es el orden de diferencia para transformar la serie en estacionaria y q corresponde al componente de promedio móvil del modelo (MA) correspondiente. La fórmula básica de ARIMA se representa por la fórmula siguiente:

$$L_t = \sum_{i=1}^p \phi_i L_{t-i} + \varepsilon_t - \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

Donde L_t es el valor de carga del clúster en el tiempo t ; ϕ y θ son el parámetro autorregresivo (AR) y el parámetro de promedio móvil (MA) respectivamente, p corresponde al orden del componente autorregresivo, q al orden del componente de promedio móvil, y ε representa la secuencia de errores o residuos de ARIMA.

La construcción del modelo ARIMA se lleva a cabo siguiendo los pasos: (i) verificar la estacionariedad de la serie temporal, si no es estacionaria se realiza la diferenciación utilizando la prueba ADF donde se determina d , (ii) identificar p y q mediante los gráficos ACF y PACF, (iii) comparar el AIC (Akaike Information Criterion) y el BIC (Bayesian Information Criterion) y (iv) la estimación de parámetros de ARIMA. Esta investigación propone y considera en el preprocesamiento de los datos, la verificación de la estacionalidad y la diferenciación de la serie temporal.

El diagrama de flujo ARIMA (Sinu et al., 2024; W. Wang et al., 2024) según muestra la Figura 2.

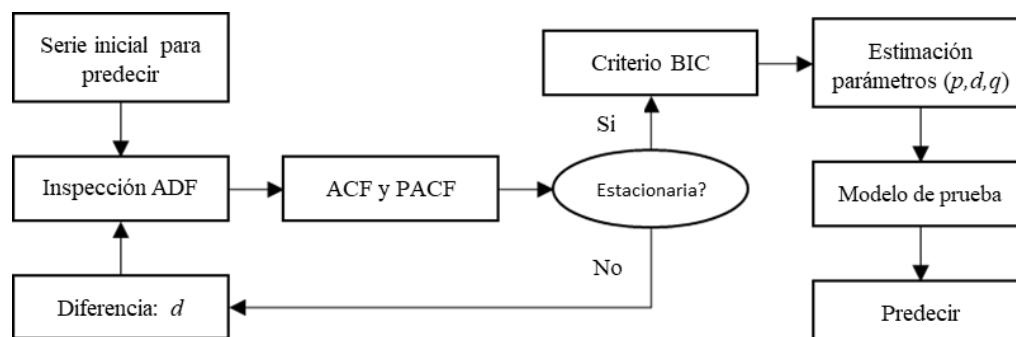


Figura 2. Diagrama de flujo ARIMA

Fuente. Datos tomados de Wang et al. (2024)

2.4.4.2. Long Short-Term Memory (LSTM)

El modelo LSTM es variante de arquitecturas neuronales recurrentes basado en el estudio de Graves y Schmidhuber (J. Hernández et al., 2018). El propósito del modelo es retener información durante periodos prolongados, lo que permite la identificación de patrones a largo plazo dentro de las tendencias de las series temporales. Las fórmulas de cálculo de tres compuertas, estados de memoria

de celda y capas ocultas (Liu et al., 2022) son los siguientes:

$$\begin{aligned}
 i_t &= \sigma(\omega x_t + u_i h_{t-1} + b_i) \\
 f_t &= \sigma(\omega x_t + u_f h_{t-1} + b_f) \\
 o_t &= \sigma(\omega x_t + u_o h_{t-1} + b_o) \\
 c_t &= f_t \odot c_{\{t-1\}} + i_t \odot \tanh(\omega x_t + u_c h_{t-1} + b_c) \\
 h_t &= o_t \odot \tanh(c_t)
 \end{aligned}$$

Los componentes del LSTM son la etapa de celda y la estructura de compuertas, que incluye el mecanismo de entrada i_t , el mecanismo de olvido f_t , y el mecanismo de salida o_t . La celda se encarga de almacenar valores a intervalos específicos, mientras que las tres compuertas regulan la transferencia de datos hacia y desde la celda en función de los valores de ponderación. Además, $\sigma()$ es la función logística, su intervalo de salida es (0,1), c_t es la celda de memoria, $\odot$ es el producto de los elementos del vector, $\tanh()$ es la función de activación y x_t es la entrada en el momento actual, h_{t-1} y h_t son el estado externo en el momento anterior y el momento actual, w^* , u^* y b^* son los parámetros de red aprendidos.

El proceso de cálculo de LSTM sigue los pasos: (i) utilizando el estado externo h_{t-1} en el momento anterior y la entrada x_t en el momento actual, los estados de las tres puertas y los estados candidatos se pueden calcular de acuerdo con las fórmulas, (ii) la celda de memoria c_t se actualiza combinando la puerta de olvido f_t y la puerta de entrada i_t y (iii) se combina con la puerta de salida o_t y la información del estado interno se pasa al estado externo h_t . El diagrama de flujo LSTM (Li et al., 2024) como se detalla en la Figura 3.

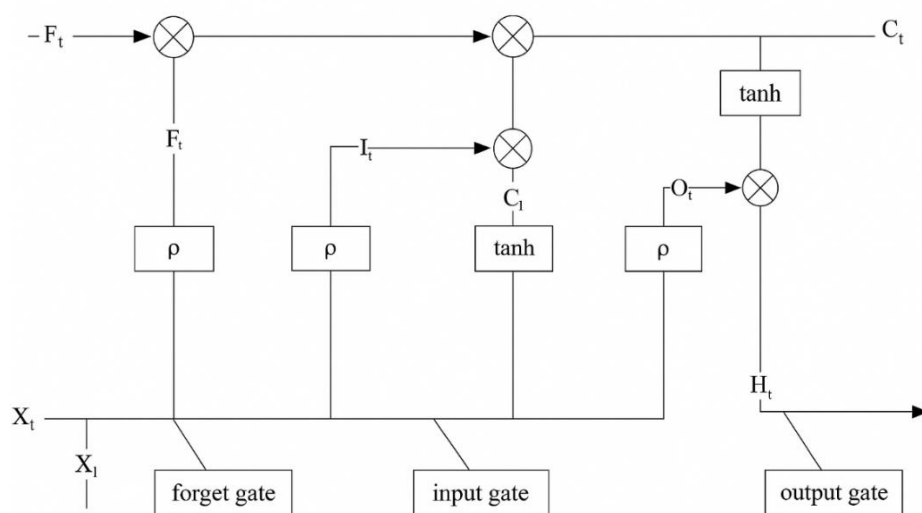


Figura 3. Diagrama de flujo LSTM
Fuente. Datos tomados de Li et al. (2024)

CAPÍTULO 3. ESTADO DEL ARTE

El estado del arte es fundamental en esta investigación porque permite identificar avances, brechas y tendencias en el campo de estudio (Kitchenham, 2007). La revisión exhaustiva y metódica de las publicaciones existentes, asegura la aplicación metodológica y la síntesis de conocimientos previos (Page et al., 2021). Mediante un examen sistemático de los estudios existentes respecto a la utilización de algoritmos inteligentes para proyectar el comportamiento de las tasas de ciberdelitos (Barrenechea & Lengua, 2025) se justifica la originalidad y relevancia de este estudio.

3.1. Introducción

La utilización de la inteligencia artificial en el combate contra la ciberdelincuencia ha ido en aumento, permitiendo automatizar el análisis de información y optimizar los procesos de decisión (Sharko et al., 2024). Los algoritmos de inteligencia artificial son elementos de prueba en el campo del análisis predictivo para la detección de fraudes (Parisi, 2019) donde los investigadores han demostrado que pueden trabajar eficazmente para los ciberdelitos con una precisión entre el 70% y el 90% (Aljarboua et al., 2022b).

La inteligencia artificial plantea grandes desafíos. Diversos métodos estadísticos y de aprendizaje automático se han aplicado para estimar las tasas de ciberdelincuencia, entre ellos los modelos SVM y KNN, que han mostrado tasas de precisión prometedoras de hasta el 89% y el 92.34% (Boonyopakorn et al., 2023; Veena, Meena, Teekaraman, et al., 2022). Se han desarrollado enfoques más sofisticados, como las arquitecturas de redes recurrentes de doble dirección basadas en unidades de memoria de corto plazo (BRNN-LSTM), con el fin de incrementar la exactitud del pronóstico al adaptarse adecuadamente a las características estadísticas de las series temporales asociadas a los ciberataques

(Fang et al., 2019). La integración de estos modelos predictivos mejora la asignación de recursos para la ciberdefensa, con el propósito de reducir la incidencia de los ciberdelitos (Ch et al., 2020). A medida que evolucionan las amenazas cibernéticas, la mejora continua en las metodologías de pronóstico sigue siendo crucial para estrategias de ciberseguridad efectivas (Veena, Meena, Kuppusamy, et al., 2022b).

Las brechas en el pronóstico de la tasa de delitos cibernéticos incluyen un alcance limitado. Los modelos existentes no consideran un enfoque integral para todos los tipos de ciberdelincuencia (Ch et al., 2020). Muchos estudios enfatizan el uso de técnicas de aprendizaje automático, pero no abordan adecuadamente la integración de diversos conjuntos de datos o los desafíos de preprocesamiento que pueden afectar la precisión del modelo (Duc M Cao et al., 2024; Veena, Meena, Kuppusamy, et al., 2022b). Si bien algunos modelos logran altas tasas de precisión, como el 91% con SVM, sigue siendo necesario mejorar las metodologías que puedan adaptarse a la naturaleza cambiante de las amenazas cibernéticas e incorporar datos en tiempo real de manera efectiva (Duc M Cao et al., 2024). Por último, la dependencia de datos históricos sin considerar las tendencias emergentes en el cibercrimen presenta una limitación significativa en las capacidades predictivas (Fang et al., 2019).

3.2. Metodología

Se utilizó la metodología PRISMA. Se justifica porque realiza un análisis estructurado de la producción científica (Falade et al., 2019; van Dinter et al., 2021) identificando, evaluando y sintetizando los estudios existentes que destacan las tendencias, metodologías y brechas (Harie et al., 2023) responder a las preguntas identificadas a partir de la revisión exhaustiva de estudios previos (Kumar, 2023). Para formular las preguntas que guiaron el estudio se recurrió al método PICO (Problema, Intervenciones, Comparadores y Resultados), que permitió precisar las preguntas y criterios de elegibilidad de los artículos científicos de esta investigación (Cumpston et al., 2023).

Los estudios seleccionados para la revisión sistemática fueron organizados

utilizando el método estadístico de agrupación jerárquica. Se utilizó para agrupar objetos similares en grupos según sus características (Shui et al., 2024). La agrupación jerárquica permitió clasificar los 30 artículos incluidos en cuatro grupos según la modalidad de ciberdelito y las técnicas de inteligencia artificial empleadas en la predicción. Esta agrupación jerárquica como instrumento facilitó el desarrollo de las respuestas (Shui et al., 2024) a las preguntas de este estudio.

3.2.1. Preguntas de investigación

Como parte del desarrollo del estudio, se empleó el método PICO para formular la pregunta principal y cuatro preguntas de investigación específicas (Problema, Intervenciones, Comparadores y Resultados) con el objetivo de extraer y resumir el conocimiento a partir de los artículos considerados durante el análisis de la literatura disponible. Estas interrogantes se presentan en la Tabla 1 que se elaboró a partir de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 1. Operacionalización de variables

Factor	ID	Pregunta
Pregunta Principal	PP	¿Qué modelos que usan métodos de inteligencia artificial para estimar la tasa de ciberdelincuencia de alto rendimiento se han propuesto en los últimos años?
Problema	P1	¿Qué cuestiones relacionadas con la ciberdelincuencia se han abordado con la previsión de la tasa de ciberdelincuencia con inteligencia artificial?
Intervención	P2	¿Qué métodos de inteligencia artificial se han utilizado para pronosticar la tasa de ciberdelincuencia?
Comparación	P3	¿Qué métricas de evaluación se han utilizado para medir el rendimiento de los métodos para pronosticar la tasa de ciberdelincuencia?
Resultados	P4	¿Qué mejoras se han incorporado a los modelos para pronosticar la tasa de ciberdelincuencia?

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

3.2.2. Estrategia de investigación

Siguiendo el proceso, se construyó la cadena de búsqueda siguiendo el método PICO, con cada uno de los cuatro factores definidos (Problema, Intervención, Comparación y Resultados) con su respectiva descripción, los términos de búsqueda y un conjunto de sinónimos utilizados para la investigación.

La Tabla 2 con las expresiones utilizadas para la búsqueda se elaboró de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 2. Términos de búsqueda

Factor	Descripción	Términos de búsqueda	Sinónimos
Problema	Pronóstico de ciberdelincuencia	"forecasting of cybercrime incidence"	"forecasting of cybercrime incidence", "prediction models for cybercrime occurrence", "projection of computer-related offense rates", "estimation of trends in computer crime cases", "forecasting schemes for cyberfraud incidence" y "predictive analysis of cyberfraud trends"
Intervención	Uso de inteligencia artificial	"intelligent computational systems"	"learning methods based on machines", "approaches rooted in deep hierarchical learning", "knowledge adaptation through transfer-based learning", "policy-driven reinforcement learning frameworks" y "artificial neural-network models".
Comparación	Precisión del pronóstico	"classification precision"	"measurement precision", "error magnitude", "classification accuracy", "evaluation score", "system performance", "model efficacy", "assessment metric" y "statistical properties"
Resultados	Identificación del modelo	"models"	"computational model", "data-driven pattern", "analytical method", "learning algorithm" y "processing technique".

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

La cadena de búsqueda (Figura 4) corresponde al utilizado en mi publicación previa (Barrenechea & Lengua, 2025):

((cybercrime AND rate AND forecast) OR (cybercrime AND rate AND prediction) OR ("computer crime" AND rate AND forecast) OR ("computer crime" AND rate AND prediction) OR (cyberfraud AND rate AND forecast) OR (cyberfraud AND rate AND prediction)) AND ("machine learning" OR "deep learning" OR "transfer learning" OR "reinforcement learning" OR "neural network") AND (precision OR error OR accuracy OR score OR performance OR efficacy OR metric OR statistical) AND (model OR pattern OR method OR algorithm OR technique)

Figura 4. Cadena de búsqueda de artículos

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

3.2.3. Criterio de elegibilidad

Para acotar el alcance de la exploración en los repositorios y seleccionar únicamente estudios vinculados al pronóstico de tasas de ciberdelincuencia mediante técnicas de inteligencia artificial, se definieron criterios específicos de elegibilidad. Se establecieron los criterios detallados en la Tabla 3 que se elaboró

a partir de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 3. Criterio de elegibilidad

Criterio	ID	Interrogante
Inclusión	I1	Artículos relacionados con la ciberdelincuencia
	I2	Artículos que aplican inteligencia artificial en modelos de conjuntos de datos para pronosticar la tasa de ciberdelincuencia
	I3	Revistas y artículos de congresos
Exclusión	E1	Artículos escritos en idiomas diferentes al español o al inglés
	E2	Artículos publicados antes de 2018
	E3	Artículos sin disponibilidad de texto completo

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

3.2.4. Fuentes de información

Se consultaron los repositorios científicos IEEE Xplore, Scopus, y Web of Science por ser consideradas fuentes sólidas y reconocidas en la comunidad académica. En las tres plataformas se aplicó la misma estrategia de búsqueda, lo que permitió reunir 229 artículos publicados entre 2018 y 2024. De ellos, 140 trabajos salieron de Scopus, 83 de IEEE Xplore y solo 6 de Web of Science.

3.2.5. Proceso de selección de artículos

La elección de estudios, representado mediante el diagrama PRISMA, se desarrolló en tres fases. En primer lugar, durante la fase de identificación, se ejecutó la estrategia de búsqueda aplicada en las bases de datos escogidas, lo que permitió reunir el conjunto inicial de artículos que cumplieran con los parámetros establecidos. Posteriormente, en la fase de selección, se evaluaron los títulos y resúmenes de los artículos tomando en cuenta los criterios establecidos para incluir o descartar estudios. Finalmente, en la fase de incorporación, los artículos preseleccionados fueron revisados en sus secciones de introducción, metodología y conclusiones, con el fin de determinar cuáles serían integrados en la síntesis cualitativa.

Durante la fase de identificación, la cadena de búsqueda se ejecutó en

Scopus, IEEE Xplore y Web of Science, obteniéndose 229 publicaciones, tal como se presenta en la Figura 5. Tras depurar los registros duplicados, se descartaron 19 documentos, quedando 210 estudios únicos. En la fase de selección, se revisaron de los artículos tomando en cuenta los criterios establecidos, lo que llevó a retirar 24 trabajos y conservar 186. Finalmente, en la fase de inclusión se descartaron 156 estudios que no cumplían los criterios definidos, resultando en un conjunto definitivo de 30 estudios para la síntesis cualitativa.

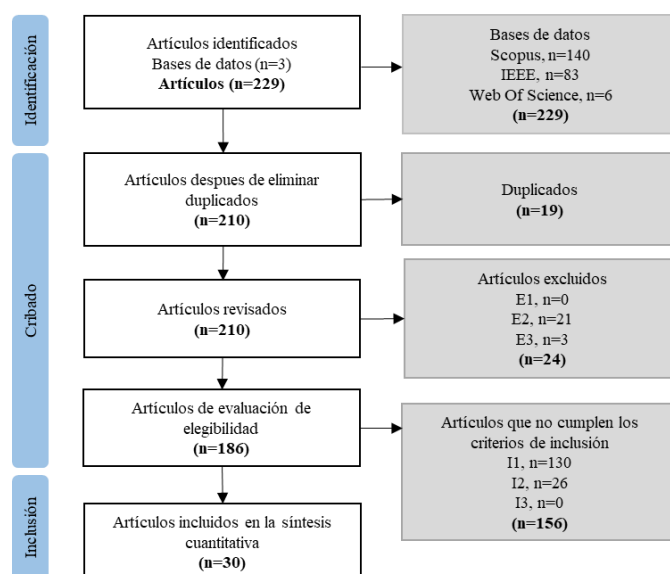


Figura 5. Diagrama de flujo PRISMA de la revisión sistemática
Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

3.2.6. Agrupación automática de artículos

La agrupación de los 30 artículos incluidos en la síntesis cualitativa se realizó mediante el método de agrupación jerárquica. Se aseguró la objetividad del análisis y se evitaron sesgos en la clasificación de los estudios (Villar & Lengua, 2024). Para la clasificación, los artículos fueron etiquetados por características como la modalidad de ciberdelito, el método de pronóstico utilizado, el tipo de datos empleados, los métodos de escalado y desequilibrio aplicadas a los datos, la técnica de ensamblaje y la optimización de los parámetros de los métodos utilizados en los modelos. Se utilizó el software DATAtab para la

agrupación jerárquica con la distancia euclidiana y su representación gráfica a través de un Dendograma según muestra la Figura 6.

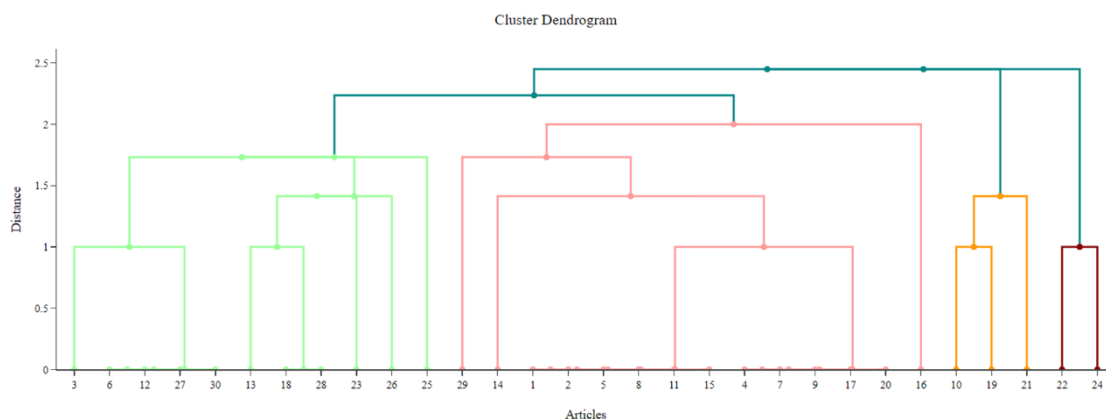


Figura 6. Agrupación jerárquica aglomerativa de artículos

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

Grupo 1: compuesto por once documentos que abordan estudios sobre ciberdelincuencia, ciberfraude y ciberataque que aplican métodos de aprendizaje automático y redes neuronales. Utilizan conjuntos de datos de series temporales, técnicas de escalado para estandarizar las características de los datos y técnicas de ensamblaje de algoritmos (Guan et al., 2022; Bardaje & R.K., 2022b; Bilen & Özer, 2021; Swaminathan et al., 2022; Singh et al., 2024), optimización del rendimiento con hiperparámetros (Al-Ghamdi & Alsubait, 2022; Verma & Thakur, 2023) y técnicas para manejar datos desequilibrados (Jadhav et al., 2022). Otros artículos utilizan datos sin una serie temporal y aplican técnicas de ensamblaje de algoritmos, técnicas para manejar datos desequilibrados y optimización con hiperparámetros (Aljarboua et al., 2022b; Amen et al., 2021; A. Sharma & Babbar, 2023).

Grupo 2: compuesto por catorce documentos que abordan estudios sobre ciberataque, cibercrimen y ciberacoso que aplican algoritmos de aprendizaje automático de tipo supervisado y no supervisado junto con arquitecturas neuronales. Utilizan series temporales (Veena, Meena, Kuppusamy, et al., 2022b) con técnicas de ensamblaje de algoritmos (Arora et al., 2019; Deepak et al., 2023b; Dipakkumar Pandya et al., 2022; Naaz, 2021; Omar et al., 2023; Almarshadani et al., 2019). El resto de los artículos utilizan datos que no son de series temporales

con técnicas de ensamblaje de algoritmos (Amir Latif et al., 2019; bin Asad et al., 2020; Matias et al., 2022; Philomina et al., 2022; Devadarshini et al., 2023), con técnicas de escalado de datos y optimización con hiperparámetros (Kotian & Sonkusare, 2021; Varsha et al., 2023).

Grupo 3: este grupo está compuesto por tres documentos que abordan estudios sobre ciberataques, ciberamenazas y ciberacoso que utilizan conjuntos de datos sin series temporales en nuevos modelos híbridos (Balogun et al., 2021; Lilhore et al., 2023). Otro artículo utiliza conjuntos de datos sin series temporales, pero aplica la optimización de hiperparámetros en los modelos de pronóstico para estas modalidades de ciberdelincuencia (Elmezain et al., 2022).

Grupo 4: este grupo está compuesto por dos artículos que abordan estudios de ciberdelincuencia y ciberfraude utilizando conjuntos de datos de series temporales, técnicas de escalado para estandarizar las características de los datos, optimización de hiperparámetros y ensamblaje de algoritmos (Ghofran Mohamed et al., 2023). Otro artículo considera estas técnicas de optimización, pero aplicadas a nuevos modelos híbridos (D. Sharma & Singh Kang, 2022).

3.3. Resultados

Con base en la revisión de los 30 estudios seleccionados para la síntesis cualitativa, se elaboraron las respuestas a las preguntas de investigación con un sentido claro y coherente con lo analizado. Para una mejor organización, comprensión y apoyo, las respuestas a todas las preguntas se sintetizan en tablas.

- PP: ¿Qué modelos que usan métodos de inteligencia artificial para estimar la tasa de ciberdelitos de alto rendimiento se han propuesto en los últimos años?

Los modelos de inteligencia artificial que se han propuesto en los últimos años para pronosticar, predecir y/o detectar diferentes tipos de ciberdelitos se clasificaron en cuatro grupos utilizando el método de agrupación jerárquica. Esta respuesta se detalla en la sección sobre agrupación automática de artículos. A continuación, se explorará más a fondo la revisión a través de las cuatro preguntas específicas.

- P1: ¿Qué problemas relacionados con la ciberdelincuencia se han abordado con el pronóstico de la tasa de ciberdelitos con inteligencia artificial?

De los artículos analizados, se encontró que se han abordado diversas problemáticas relacionadas con diferentes modalidades como el cibercrimen, el ciberataque, el ciberfraude y el ciberacoso. La inteligencia artificial se utilizó a través de diversas técnicas de aprendizaje automático y modelos neuronales con el fin de calcular la tasa de ciberdelincuencia. La explicación completa se presenta en la Tabla 4 que se elaboró a partir de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 4. Resultados correspondientes a P1

Palabra clave	Entrada
Predicción del cibercrimen	Modelos para la predicción de eventos y tasas de ciberdelincuencia en la sociedad y el sector financiero utilizando varias técnicas de aprendizaje automatizado y modelos de redes neuronales. Con supervisión (Aljarboua et al., 2022b; Singh et al., 2024; Bhardwaj & R.K., 2022b; Bilen & Özer, 2021; Deepak et al., 2023b; Dipakkumar Pandya et al., 2022; Swaminathan et al., 2022) y sin supervisión (Veena, Meena, Kuppusamy, et al., 2022b), aprendizaje automático y redes neuronales (Arora et al., 2019; Ghofran Mohamed et al., 2023). Se encuentra con una variedad de modelos de predicción y mitigación de amenazas que utilizan diversos conjuntos de datos y algoritmos para identificar patrones de ciberdelincuencia.
Predicción del ciberataque	Modelos que analizan la predicción de ciberataques como el phishing en sitios web, internet e IoT, la detección de Ransomware y la predicción de Malware mediante métodos de aprendizaje automático y redes neuronales. Utiliza (Varsha et al., 2023) aprendizaje automático con métodos de ensamblaje (Al-Ghamdi & Alsubait, 2022; Omar et al., 2023; Amen et al., 2021; Amir Latif et al., 2019; Balogun et al., 2021; Almashhadani et al., 2019; Naaz, 2021; A. Sharma & Babbar, 2023) y el uso de redes neuronales (bin Asad et al., 2020; Philomina et al., 2022; Devadarshini et al., 2023; Kotian & Sonkusare, 2021; Verma & Thakur, 2023). Para la predicción de ciberataques, destacan aprendizaje algorítmico y redes profundas que mejoran la eficiencia y la precisión.
Detección del ciberfraude	Modelos para la detección de ciberfraude generados por transacciones fraudulentas principalmente en el sector financiero con el uso de aprendizaje automático con métodos de ensamblaje (Guan et al., 2022; Jadhav et al., 2022; D. Sharma & Singh Kang, 2022; Al-Ghamdi & Alsubait, 2022). Estos modelos se utilizan para detectar y pronosticar el fraude cibernético que utilizan esquemas de aprendizaje automatizado y arquitectura profunda orientados a mejorar tanto la precisión como la eficiencia.
Previsión de ciberamenazas	Modelo eficiente de aprendizaje automático híbrido (EHML) con la aplicación de métodos de ensamblaje para la predicción de ciberamenazas (Lilhore et al., 2023). Este modelo híbrido es una propuesta diferente e innovadora en la predicción de ciberamenazas.
Predicción del ciberacoso	Modelos para la predicción del ciberacoso en redes sociales con la adopción de estrategias de aprendizaje automatizado y métodos asociados de ensamblaje con datos de texto de la red social X (antes Twitter) (Matias et al., 2022) y con imágenes maliciosas con la aplicación de redes neuronales (Elmezain et al., 2022). Estos modelos ensamblados con métodos de redes neuronales para predecir el ciberacoso sugieren, debido a su potencial, profundizar su aplicación en trabajos futuros.

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

- P2: ¿Qué métodos de inteligencia artificial se han utilizado para pronosticar la tasa de ciberdelitos?

De los artículos revisados, se encontró que, a través de la inteligencia artificial, se han utilizado diferentes métodos para predecir la tasa de ciberdelincuencia. Se utilizaron los métodos: aprendizaje automático, conjunto, impulso y redes neuronales. Además, la predicción y detección de eventos generados por los ciberdelincuentes. La respuesta se detalla en la Tabla 5 que se elaboró a partir de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 5. Resultados correspondientes a P2

Palabra clave	Entrada
Métodos de aprendizaje automático	Métodos simples de aprendizaje automático utilizados con otros métodos complejos en modelos para predecir el delito cibernético. K-nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), J48 entre otros se utilizaron (Aljarboua et al., 2022b; Almashhadani et al., 2019; Amen et al., 2021; Amir Latif et al., 2019; Al-Ghamdi & Alsubait, 2022; D. Sharma & Singh Kang, 2022; Arora et al., 2019; Balogun et al., 2021; Bhardwaj & R.K., 2022b; Bilen & Özer, 2021; Varsha et al., 2023; Deepak et al., 2023b; Dipakkumar Pandya et al., 2022; Elmezain et al., 2022; Ghofran Mohamed et al., 2023; Lilhore et al., 2023; Matias et al., 2022; Naaz, 2021; Omar et al., 2023; A. Sharma & Babbar, 2023; Devadarshini et al., 2023; Singh et al., 2024; Swaminathan et al., 2022; Verma & Thakur, 2023); y métodos no supervisados como K-Means y Gaussian Mixture Model (GMM) (Veena, Meena, Kuppusamy, et al., 2022b). SVM y LR destacan en su uso. Estos métodos tradicionales de aprendizaje automático son los más utilizados, pero no necesariamente los más efectivos para pronosticar las tasas de ciberdelincuencia.
Métodos de conjunto	Se encontraron métodos simples que se utilizan en paralelo para reducir la varianza de las estimaciones en los modelos de predicción de delitos cibernéticos, principalmente RF (Al-Ghamdi & Alsubait, 2022; Amen et al., 2021; Amir Latif et al., 2019; Bhardwaj & R.K., 2022b; Deepak et al., 2023b; Devadarshini et al., 2023; Dipakkumar Pandya et al., 2022; Ghofran Mohamed et al., 2023; Lilhore et al., 2023; Matias et al., 2022; Naaz, 2021; Almashhadani et al., 2019; A. Sharma & Babbar, 2023; D. Sharma & Singh Kang, 2022; Swaminathan et al., 2022; Omar et al., 2023; Verma & Thakur, 2023) que destaca en su aplicación, Logistic Model Tree (Amir Latif et al., 2019), Extra Tree (Dipakkumar Pandya et al., 2022) y Enhanced Decision Tree (Lilhore et al., 2023). Se ha descubierto que los métodos de aprendizaje por conjuntos se utilizan cada vez más para pronosticar las tasas de ciberdelincuencia debido a su capacidad para potenciar la exactitud en los procesos de predicción mediante la combinación de varios modelos.
Métodos de impulso	Las técnicas simples que se utilizan secuencialmente para mejorar el rendimiento del modelo de predicción de delitos cibernéticos fueron eXtreme Gradient Boosting (XGBoost) (Ghofran Mohamed et al., 2023; Bhardwaj & R.K., 2022b; Bilen & Özer, 2021; Jadhav et al., 2022; Omar et al., 2023; Singh et al., 2024), LightGBM (Bhardwaj & R.K., 2022b; bin Asad et al., 2020; Guan et al., 2022; Omar et al., 2023), Gradient Boosting (Bhardwaj & R.K., 2022b; Ghofran Mohamed et al., 2023; Philomina et al., 2022; D. Sharma & Singh Kang, 2022), Gradient Tree Boosting (Amen et al., 2021), CatBoost (Bhardwaj & R.K., 2022b) y AdaBoost (Ghofran Mohamed et al., 2023). Estos métodos de impulso son relevantes para pronosticar las tasas de ciberdelincuencia debido a su capacidad para administrar estructuras de datos complejas y mejorar la precisión predictiva.
Métodos de redes neuronales	Los métodos de redes neuronales se encontraron en modelos de predicción de ciberdelitos y ataques cibernéticos como Convolutional Neural Network (CNN) (bin Asad et al., 2020; Elmezain et al., 2022; Kotian & Sonkusare, 2021), Multilayer Perceptron (MLP) (Arora et al., 2019; Ghofran Mohamed et al., 2023), Artificial Neural Network (ANN) (Verma & Thakur, 2023) y Deep Neural Network (DNN) (Devadarshini et al., 2023). Se encuentra que los métodos de redes neuronales son cada vez más relevantes para pronosticar las tasas de ciberdelincuencia, ya que aprovechan su posibilidad de operar y analizar masivos conjuntos de datos para predecir amenazas potenciales. Se sugiere que su aplicación se profundice en trabajos futuros.

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

- P3: ¿Qué métricas de evaluación se han utilizado para medir el rendimiento de los métodos para pronosticar la tasa de ciberdelitos?

Las métricas son herramientas clave para evaluar cuantitativamente el rendimiento de los métodos utilizados. Esto contribuye a la selección, optimización y validación de los métodos de pronóstico de la tasa de ciberdelincuencia. Las principales métricas utilizadas que se encontraron en los artículos analizados corresponden a las métricas de clasificación y regresión. La respuesta se detalla en la Tabla 6 que se elaboró a partir de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 6. Resultados correspondientes a P3

Palabra clave	Entrada
Métricas de clasificación	Los criterios de evaluación basados en accuracy, precision, recall y F1-score se utilizaron para evaluar modelos de predicción y detección de ciberataques (Almashhadani et al., 2019; Amen et al., 2021; Amir Latif et al., 2019; Kotian & Sonkusare, 2021; Naaz, 2021; Omar et al., 2023; Philomina et al., 2022; A. Sharma & Babbar, 2023; Varsha et al., 2023), Ciberdelincuencia (Aljarboua et al., 2022b; Bilen & Özer, 2021; Devadarshini et al., 2023; Deepak et al., 2023b; Ghofran Mohamed et al., 2023; Singh et al., 2024; Swaminathan et al., 2022; Dipakkumar Pandya et al., 2022; Veena, Meena, Kuppusamy, et al., 2022b), ciberfraude (Al-Ghamdi & Alsubait, 2022; Guan et al., 2022; Jadhav et al., 2022), ciberacoso (Matias et al., 2022; Elmezzain et al., 2022) y ciberamenazas (Lilhore et al., 2023), la curva ROC (Receiver Operating Characteristic) y el valor AUC (Area Under the Curve) también se aplicaron en la evaluación para la predicción de ciberataques (Balogun et al., 2021; bin Asad et al., 2020), identificación de delitos cibernéticos (Arora et al., 2019) y detección de ciberfraude (D. Sharma & Singh Kang, 2022). Los estudios encuentran que estas métricas de clasificación se utilizan con frecuencia para evaluar la precisión y confiabilidad de los modelos utilizados en la detección y predicción del delito cibernético.
Métricas de regresión	MSE (Mean Square Error) y R2 (R-Squared) se utilizaron métricas de evaluación para el cibercrimen (Bhardwaj & R.K., 2022b) y predicción de ciberataques (Verma & Thakur, 2023). Estas métricas de regresión que permiten evaluar la eficacia y el desempeño de los modelos de pronóstico se usan principalmente en estudios que tienen como objetivo predecir cantidades continuas.

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

- P4: ¿Qué mejoras se han incorporado a los modelos para pronosticar la tasa de ciberdelitos?

Las mejoras que se han incorporado a los modelos a través de diferentes métodos y técnicas buscan optimizar la previsión de eventos y la tasa de ciberdelincuencia. Estas técnicas se utilizan para manejar datos desequilibrados, técnicas de escalado de características, técnicas de reducción de dimensionalidad, optimización de hiperparámetros e implementación de modelos fijos. La respuesta se detalla en el Tabla 7 que se elaboró a partir de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 7. Resultados correspondientes a P4

Palabra clave	Entrada
Técnicas para manejar datos desequilibrados	Los conjuntos de datos desequilibrados causados por la escasez, el sesgo de muestreo o los cambios recurrentes en el contexto cambiante de la ciberdelincuencia ponen en desventaja las clases de datos y tienen un impacto negativo en las métricas de evaluación en los modelos de ciberdelincuencia. Se aplicaron técnicas de submuestreo con la aplicación de Undersampling para la detección del ciberfraude (Guan et al., 2022) y Oversampling con la aplicación de SMOTE (Synthetic Minority Oversampling Technique) para la detección de phishing (Kotian & Sonkusare, 2021), Malware (D. Sharma & Singh Kang, 2022), transacciones fraudulentas con tarjeta de crédito (Jadhav et al., 2022; D. Sharma & Singh Kang, 2022), ciberdelincuencia (Ghofran Mohamed et al., 2023) y ciberataque (A. Sharma & Babbar, 2023). Estas técnicas se aplicaron eficazmente para casos de clases minoritarias como los ciberataques con la desventaja de una posible disminución de las tasas de detección al reducir los falsos positivos. Se recomienda una evaluación y un ajuste cuidadosos para garantizar un rendimiento óptimo.
Técnicas de escalado de características	Las técnicas de escalado de características se aplican a través del empleo de métodos de escala y homogeneización de los datos para mejorar el rendimiento y hacer que el modelo de predicción y detección de delitos cibernéticos sea más efectivo y preciso. Standard Scaler (Al-Ghamdi & Alsubait, 2022; Aljarboua et al., 2022b; Bilen & Özer, 2021; Singh et al., 2024) y Min Max Scaler (Amen et al., 2021; Verma & Thakur, 2023) se utilizan técnicas para llevar las características a una magnitud similar, haciéndolas comparables y evitando que cualquier característica domine el algoritmo porque tiene una escala mayor. El escalado de características debe ser solo una parte del proceso de ingeniería de características. Los estudios muestran que la codificación y la selección de características son clave con el fin de mejorar el rendimiento del modelo. Además de elegir la técnica para el escalado (normalización o estandarización) debe estar alineada con los requisitos específicos del algoritmo de aprendizaje automático utilizado.
Técnicas de reducción de dimensionalidad	Los métodos destinados a disminuir la cantidad de variables se emplean para eliminar características de los datos que reducen el rendimiento y lograr modelos predictivos óptimos y eficientes del cibercrimen. El PCA (Principal Component Analysis) (Bhardwaj & R.K., 2022b; Jadhav et al., 2022; Naaz, 2021) se utiliza para reducir las dimensiones de los conjuntos de datos en el análisis del ciberfraude para transacciones fraudulentas con tarjetas de crédito. Estos estudios muestran que estas técnicas simplifican grandes conjuntos de datos haciéndolos más manejables e interpretables. Sin embargo, se debe tener cuidado al elegir la técnica adecuada, según las particularidades de los datos analizados, para garantizar que no se pierda información crítica.
Optimización de hiperparámetros	La optimización de hiperparámetros beneficia la previsión de la tasa de ciberdelincuencia. Para conseguir los mejores resultados en la generación de modelos de predicción, la optimización de los hiperparámetros de los modelos de predicción de ciberataque (Kotian & Sonkusare, 2021; A. Sharma & Babbar, 2023; Varsha et al., 2023; Verma & Thakur, 2023), ciberfraude (Al-Ghamdi & Alsubait, 2022; Jadhav et al., 2022; D. Sharma & Singh Kang, 2022), ciberdelincuencia (Aljarboua et al., 2022b; Ghofran Mohamed et al., 2023) y ciberacoso (Elmezain et al., 2022). Sin embargo, para la optimización de hiperparámetros es necesario considerar la cantidad de recursos computacionales y la complejidad involucrada en el proceso de optimización.
Modelos mixtos	Los modelos mixtos ofrecen beneficios significativos en el pronóstico de la tasa de delitos cibernéticos al integrar múltiples técnicas analíticas para mejorar la precisión de la predicción. Combinando diferentes métodos con el fin de optimizar la eficacia y exactitud a través de modelos ensamblados para la predicción del cibercrimen (Aljarboua et al., 2022b; Arora et al., 2019; Bhardwaj & R.K., 2022b; Ghofran; Singh et al., 2024; Bilen & Özer, 2021; Deepak et al., 2023b; Dipakkumar Pandya et al., 2022; Swaminathan et al., 2022), ciberataque (Almashhadani et al., 2019; Amir Latif et al., 2019; bin Asad et al., 2020; Omar et al., 2023; Devadarshini et al., 2023; Naaz, 2021; Amen et al., 2021; Philomina et al., 2022; A. Sharma & Babbar, 2023; Varsha et al., 2023; Verma & Thakur, 2023), ciberfraude (Al-Ghamdi & Alsubait, 2022; Guan et al., 2022; Jadhav et al., 2022) y ciberacoso (Matias et al., 2022). Además, modelos híbridos para la detección del ciberfraude (D. Sharma & Singh Kang, 2022), ciberacoso (Elmezain et al., 2022), ciberataque (Balogun et al., 2021) y ciberamenaza (Lilhore et al., 2023). Los estudios coinciden en que la implementación de modelos mixtos demanda una cantidad elevada de recursos computacionales y la disponibilidad de grandes conjuntos de datos.

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

3.4. Conclusiones

Se han desarrollado varios enfoques y modelos de aprendizaje automático basados en las investigaciones revisadas para pronosticar las tasas de ciberdelincuencia. Algunas técnicas que se utilizan para estimar la frecuencia y el momento de los cibercrímenes incluyen SVM, LR, RF, XGBoost, MLP y CNN, utilizadas especialmente para anticipar diferentes tipos de ciberataques. Así mismo, la literatura resalta que el uso de datos reales es clave para aumentar la precisión de las predicciones y que el rendimiento de los modelos puede mejorarse mediante métodos apropiados.

Sin embargo, se reconoce la necesidad de mejorar los modelos existentes o de introducir una nueva arquitectura combinada para mejorar la capacidad predictiva. Con esto en mente, el modelo ARIMA-LSTM presentado en este estudio se considera un potencial candidato, ya que, al proporcionar una mayor precisión y eficiencia en la predicción de la actividad delictiva en el ámbito digital, representaría una estrategia de desarrollo alternativa.

CAPÍTULO 4. METODOLOGIA

Este capítulo presenta el enfoque metodológico utilizado a partir de las hipótesis como propuestas de solución tentativa al problema (Ñaupas et al., 2014). Se precisan las variables, su operacionalización y el tipo y diseño del estudio. Se identifica la población y muestra del estudio, así como los métodos de recolección y manejo de los datos. También se aborda el proceso de examinar los datos y su entendimiento, el dataset y las etapas de preprocesamiento. Se incluye la optimización por hiperparámetros y la construcción del modelo híbrido ARIMA–LSTM. Finalmente, se presentan las métricas empleadas para evaluar su desempeño.

4.1. Hipótesis general

HG: El desarrollo de un modelo híbrido optimizado ARIMA–LSTM mejora significativamente la precisión en el pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú, en comparación con los modelos individuales ARIMA y LSTM.

4.2. Hipótesis específicas

- HE1: La estacionariedad de la serie temporal de denuncias por ciberdelitos contra el patrimonio mejora significativamente la eficiencia y precisión del modelo ARIMA en el pronóstico.
- HE2: La integración del modelo ARIMA con el modelo LSTM en un enfoque híbrido optimizado mejora significativamente la capacidad de pronóstico en comparación con los modelos individuales ARIMA y LSTM.

- HE3: El modelo híbrido optimizado ARIMA–LSTM alcanza un valor de MAPE significativamente inferior al de los modelos individuales ARIMA y LSTM, lo que demuestra una mayor precisión en el pronóstico de denuncias por ciberdelitos contra el patrimonio.

4.3. Identificación de variables

Se identificaron en la investigación las variables:

- La variable independiente actúa sobre la variable dependiente y, según la hipótesis formulada, no está supeditada a ninguna otra (Ñaupas et al., 2014). La variable independiente de este estudio es el “Tipo de modelo de aprendizaje automático” que se define y operacionaliza en la Tabla 8.
- La variable dependiente de una hipótesis es el efecto, la consecuencia o vale decir el fenómeno objeto de análisis (Ñaupas et al., 2014). En cuanto a la variable dependiente, esta corresponde a la “Precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio” definido y operacionalizado en la Tabla 8.

4.4. Operacionalización de variables

Se detalla en la Tabla 8.

Tabla 8. Operacionalización de variables

Variable	Descripción conceptual	Descripción operacional	Dimensiones	Indicadores	Técnica
Independiente: Tipo de modelo de aprendizaje automático	Es el enfoque algorítmico utilizado para realizar el pronóstico de denuncias por ciberdelitos. Comprende modelos estadísticos, de redes neuronales y modelos híbridos que combinan ambos. (Hurwitz, Judith; Kirsch, 2018)	Implementación modelos de predicción del tipo ARIMA, LSTM y ARIMA–LSTM optimizado.	- Tipo de modelo - Técnicas de optimización	- ARIMA - LSTM - Híbrido optimizado ARIMA-LSTM - Auto Arima - Random Search - Configuración de hiperparámetros	Implementación de algoritmos y entrenamiento de los modelos en Python
Dependiente: Precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio	Grado con la exactitud en que se predicen los valores futuros de denuncias (Jenga et al., 2023)	Medida obtenida mediante métricas de error entre los valores pronosticados y los observados.	- Estacionariedad de la serie temporal - Evaluación de los modelos de pronóstico - Métricas de regresión	- Augmented Dickey Fuller (ADF) - Diebold–Mariano (DM) - Mean Absolute Error (MAE) - Root Mean Square Error (RMSE) - Mean Absolute Percentage Error (MAPE)	Pruebas estadísticas para análisis de serie de tiempos y evaluación de los modelos de pronóstico

Fuente. Elaboración propia.

4.5. Matriz de consistencia

Se presenta de manera completa en la Tabla 9.

Tabla 9. Matriz de consistencia

PROBLEMAS	OBJETIVOS	HIPÓTESIS	VARIABLE	DIMENSIONES	INDICADORES	DISEÑO METODOLÓGICO
PROBLEMA GENERAL	OBJETIVO GENERAL	HIPÓTESIS GENERAL				
PP: ¿De qué manera el desarrollo de un modelo híbrido optimizado ARIMA–LSTM puede mejorar la precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú?	OG: Desarrollar un modelo híbrido optimizado ARIMA–LSTM que mejore la precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú.	HG: El desarrollo de un modelo híbrido optimizado ARIMA–LSTM mejora significativamente la precisión en el pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú, en comparación con los modelos individuales ARIMA y LSTM.	Independiente: Tipo de modelo de aprendizaje automático	- Tipo de modelo - Técnicas de optimización	- ARIMA - LSTM - Híbrido optimizado ARIMA–LSTM - Auto Arima - Random Search - Hiperparámetros	La clase de investigación empleada es una investigación aplicada desde una perspectiva cuantitativa y de tipo cuasi experimental. Población: 60,378 denuncias por ciberdelitos contra el patrimonio.
PROBLEMAS ESPECIFICOS	OBJETIVOS ESPECIFICOS	HIPÓTESIS ESPECIFICAS	Dependiente: Precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio	- Estacionariedad de la serie temporal - Evaluación de los modelos de pronóstico - Métricas de regresión	- Augmented Dickey Fuller (ADF) - Diebold–Mariano (DM) - Mean Absolute Error (MAE) - Root Mean Square Error (RMSE) - Mean Absolute Percentage Error (MAPE)	Muestreo no probabilístico y por conveniencia se toma la totalidad de la población. Instrumento: fuentes gubernamentales (Policía Nacional del Perú)
PE1: ¿Cómo influye la estacionariedad de la serie temporal de denuncias por ciberdelitos contra el patrimonio en la eficiencia y precisión del modelo ARIMA en el pronóstico? PE2: ¿En qué medida la integración del modelo ARIMA con el modelo LSTM en un enfoque híbrido optimizado mejora significativamente la capacidad de pronóstico en comparación con los modelos individuales ARIMA y LSTM? PE3: ¿Cuál es el nivel de precisión que alcanza el modelo híbrido optimizado ARIMA–LSTM frente a los modelos individuales ARIMA y LSTM a través de la métrica MAPE como indicador de desempeño?	OE1: Lograr la estacionariedad de la serie temporal de denuncias por ciberdelitos contra el patrimonio, con el propósito de mejorar la eficiencia y precisión del modelo ARIMA en el pronóstico. OE2: Diseñar y evaluar un modelo híbrido optimizado ARIMA–LSTM que integre las ventajas de ambos enfoques para determinar si su combinación mejora significativamente la capacidad de pronóstico en comparación con los modelos individuales ARIMA y LSTM. OE3: Validar el rendimiento del modelo híbrido optimizado ARIMA–LSTM mediante la métrica MAPE, a fin de determinar el nivel de precisión alcanzado respecto a los modelos individuales ARIMA y LSTM en el pronóstico de denuncias por ciberdelitos contra el patrimonio.	HE1: La estacionariedad de la serie temporal de denuncias por ciberdelitos contra el patrimonio mejora significativamente la eficiencia y precisión del modelo ARIMA en el pronóstico. HE2: La integración del modelo ARIMA con el modelo LSTM en un enfoque híbrido optimizado mejora significativamente la capacidad de pronóstico en comparación con los modelos individuales ARIMA y LSTM. HE3: El modelo híbrido optimizado ARIMA–LSTM alcanza un valor de MAPE significativamente inferior al de los modelos individuales ARIMA y LSTM, lo que demuestra una mayor precisión en el pronóstico de denuncias por ciberdelitos contra el patrimonio.				

Fuente. Elaboración propia.

4.6. Tipo y diseño de la investigación

Este estudio adopta un enfoque cuantitativo basado en la recolección de datos numéricos que permiten verificar las hipótesis formuladas mediante procesos de medición y análisis estadístico que permitan la identificación de patrones de comportamiento para confirmar las hipótesis del estudio (R. Hernández et al., 2014). Se adopta un diseño cuasi experimental, ya que permite intervenir en la variable independiente con el propósito de identificar su influencia sobre la variable dependiente. Además, las unidades de estudio poseen una estructura previamente establecida que no puede ser alterada por el investigador. (R. Hernández et al., 2014) por las denuncias de ciberdelitos contra el patrimonio.

4.7. Unidad de análisis

Se define como el componente fundamental en el estudio de contenido (Ñaupas et al., 2014) y define a los sujetos, casos o elementos que serán evaluados mediante la aplicación del instrumento correspondiente (R. Hernández et al., 2014). En el caso de esta investigación, la unidad de observación se compone de denuncias relacionadas con ciberdelitos contra el patrimonio.

4.8. Población de estudio

Esta población corresponde al conjunto global de unidades que presentan los atributos establecidos en el estudio (R. Hernández et al., 2014). La población o universo para esta investigación está conformada por 60,378 denuncias según Ley N° 30096 - Ley de Ciberdelitos reportados por la PNP durante el periodo comprendido entre 2018 y 2023.

4.9. Tamaño de muestra

Se entiende como muestra a una fracción de la población que es seleccionada por presentar las mismas características que el conjunto total de estudio (Ñaupas et al., 2014). El muestreo por conveniencia se compone de los casos accesibles y disponibles para el investigador (R. Hernández et al., 2014). Para este estudio se utiliza un muestreo no probabilístico de tipo conveniencia; por ello, se considera la totalidad de la población, integrada por las 60,378 denuncias de ciberdelitos contra el patrimonio registradas por la PNP en el periodo 2018–2023.

4.10. Recolección de datos

El acopio de datos implica determinar las técnicas o fuentes que permitan obtener información con un fin concreto. Entre estas técnicas se encuentra el uso de datos secundarios, los cuales provienen de registros públicos o de fuentes abiertas disponibles (Hernandez Sampieri et al., 2013). En este caso, se tomarán fuentes oficiales del Estado que ponen a disposición del público la información sobre denuncias de ciberdelitos contra el patrimonio, provenientes de la PNP y MPFN. El acceso a dichos datos se realiza conforme a lo establecido en la Ley N° 27806 sobre Transparencia y Acceso a la Información Pública.

4.11. Confiabilidad de datos

Las bases de datos generadas por entidades gubernamentales son valoradas debido a su organización sistemática y su amplia cobertura (R. Hernández et al., 2014). La información empleada en este estudio proviene de las denuncias por ciberdelitos contra el patrimonio registradas a nivel nacional en el Sistema de Denuncias de la PNP, conforme la Ley N° 30096, Ley de Ciberdelitos. La información utilizada en este estudio fue entregada por la DIVINDAT, conforme a la hoja de trámite N° 20231743088–A–SIGE desde la Ley N° 27806 de

Transparencia y Acceso a la Información Pública. Estos datos fueron proporcionados específicamente para los fines de la presente investigación (ver Anexo 1).

4.12. Análisis e interpretación de la información

Se desarrolló el proceso mediante el cual se analiza y se interpreta la información (R. Hernández et al., 2014), tal como se detalla en la Figura 7. En primer lugar, a partir del conjunto de datos sobre ciberdelitos, se realizó una verificación de la fiabilidad de la información con el propósito de asegurar que los datos empleados representen de forma precisa y consistente el pronóstico de ciberdelitos, reforzando así la validez y solidez científica del estudio.

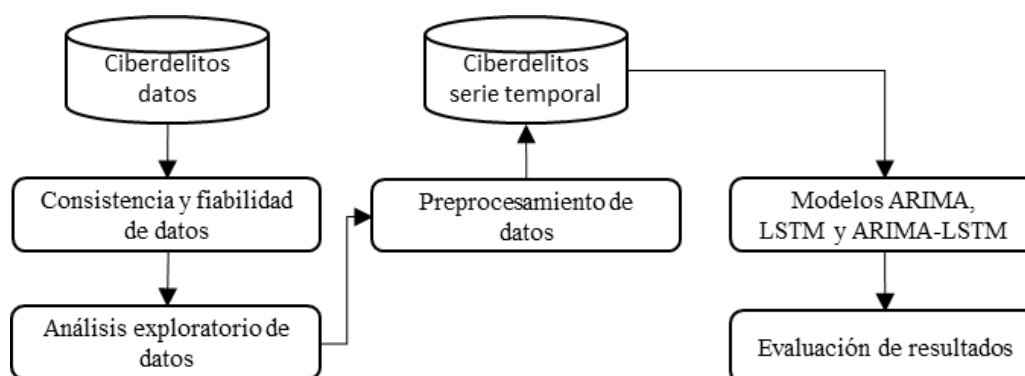


Figura 7. Proceso de análisis e interpretación de la información

Fuente. Elaboración propia.

Luego, se realizaron las pruebas de consistencia de los datos para asegurar que los datos no contengan contradicciones, errores o duplicidades que puedan afectar la validez del análisis. Posteriormente, se realizó un análisis exploratorio y estadístico para entender mejor cómo está organizada la información del dataset y determinar la forma de su distribución, tendencias y variabilidad que permitan confirmar si los datos cumplen condiciones necesarias para los modelos estadísticos o de pronóstico con los modelos de inteligencia artificial.

Seguido, se ejecutó el preprocesamiento de los datos que incluyó: la limpieza de datos (registros duplicados o fuera de rango), las pruebas de

consistencia de datos (registros con valores vacíos o nulos), la estructuración de la serie temporal de datos (agrupación mensual de los datos), la aplicación de la prueba ADF (Augmented Dickey-Fuller) con el propósito de verificar la estacionariedad de la serie, considerando un p-valor inferior al nivel de significancia de 0.05, además de realizar las transformaciones que la serie de datos requería (siempre y cuando la serie no sea estacionaria).

A continuación, la serie de datos estacionaria se utilizó como datos de entrada, en una división de 60% para el entrenamiento y 40% para la validación, para la construcción y optimización de los modelos de aprendizaje automático ARIMA, LSTM y ARIMA-LSTM. Por último, se analizan los desempeños obtenidos durante el entrenamiento de los modelos empleando métricas propias de regresión como MAE, RSME y MAPE.

4.13. Conjunto de datos

Está compuesto por 60,378 registros que corresponden a los eventos reportados en el SIDPOL. Los casos de ciberdelitos son delitos contra el patrimonio a nivel nacional registrados entre el año 2018 y 2023. En la Tabla 10 se identifican los campos, el tipo y la descripción de estos.

Tabla 10. Campos, tipos y descripción del conjunto de datos

Nro	Campo	Tipo	Descripción
1	DPTO_HECHO	Objecto	Departamento donde se registró el hecho
2	PROV_HECHO	Objecto	Provincia donde se reportó el hecho
3	DIST_HECHO	Objecto	Distrito donde se registró el hecho
4	COMISARIA	Objecto	Comisaria donde se reportó el hecho
5	TIPO	Objecto	Tipo del hecho delictivo
6	SUB_TIPO	Objecto	Subtipo de delito
7	MODALIDAD	Objecto	Modalidad del delito
8	SIT_PERSONA	Objecto	Situación del denunciante
9	SEXO	Objecto	Sexo del denunciante
10	FECHA_HORA_HECHO	datetime	Fecha y hora del delito

Fuente. Elaboración propia.

4.14. Preprocesamiento de los datos

El dataset de ciberdelitos no debe introducirse directamente en el modelo como datos de entrada. Es necesario preprocesar la serie temporal original para validar y obtener mejores resultados de entrenamiento del modelo (Xu et al., 2022). La estacionariedad de la serie temporal es un requisito para la aplicación del modelo ARIMA. Los datos normalizados eliminan el impacto en la red neuronal y facilitan el análisis mejorando la velocidad de entrenamiento del modelo (Xu et al., 2022). El ruido y valores atípicos es un fenómeno común en los datos sobre delitos (Sharmin et al., 2022). El preprocesamiento de la información mejora la organización de los datos y permite obtener conjuntos estadísticamente consistentes, lo que facilita la elaboración de predicciones más precisas sobre la incidencia de ciberdelitos (Safat et al., 2021).

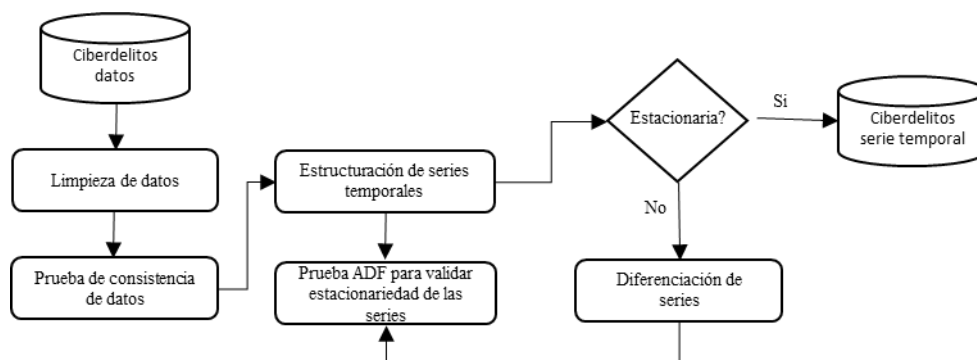


Figura 8. Diagrama de flujo de preprocesamiento de datos

Fuente. Elaboración propia a partir de Morales-Barrenechea et al. (2025)

Este flujo representa el paso a paso para el preprocesamiento de datos (Morales-Barrenechea et al., 2025) según muestra la Figura 8. Primero, la limpieza de datos elimina los registros fuera del rango de tiempo de análisis y los registros duplicados. Luego se realizan las pruebas de consistencia para identificar y eliminar los registros que tienen valores vacíos o nulos. A continuación, se estructura la serie temporal de datos a partir de la agrupación periódica de tiempo. Luego, se realiza la prueba ADF para validar la estacionariedad siempre y cuando p-valor sea menor al umbral de significancia de 0.05. Si no es estacionaria, se

aplica la diferenciación de la serie hasta conseguir la nueva serie transformada o estacionaria que es requisito clave para el modelo ARIMA.

4.15. Optimización por hiperparámetros

Es clave definir los hiperparámetros de los modelos ARIMA y LSTM dado que son componentes fundamentales de los modelos de aprendizaje automático. Es necesario realizar la búsqueda aleatoria en un rango de valores y seleccionar aquellos que proporcionan un rendimiento de regresión conveniente. (Sharmin et al., 2022). El modelo ARIMA es eficaz para identificar patrones de criminalidad incluso con ciertas fluctuaciones (Sharmin et al., 2022) y se usa la técnica Auto Arima para obtener los mejores parámetros p , d , q del modelo. En el caso de la red LSTM, se usa al método de Random Search para ajustar los hiperparámetros, que permite explorar distintas configuraciones de hiperparámetros, a través del número de épocas para el entrenamiento y la cantidad de neuronas que conforman la capa oculta, para identificar la configuración que ofrezca el mejor desempeño predictivo (Deng et al., 2023).

4.16. Modelo propuesto híbrido optimizado ARIMA-LSTM

Se diseñó un modelo híbrido optimizado ARIMA-LSTM para el pronóstico de ciberdelitos por fraude informático (Morales-Barrenechea et al., 2025). Según (Kasemset et al., 2023), los modelos híbridos ARIMA-LSTM han demostrado mejorar significativamente la precisión del pronóstico al combinar las fortalezas de ambos enfoques en el estudio de datos temporales. La combinación de estos modelos implicó seleccionar y optimizar los parámetros, comprobar la estacionalidad de la serie para el modelo ARIMA y luego entrenar al modelo LSTM sobre los residuos obtenidos de ARIMA para capturar los componentes no lineales y no explicados por el modelo lineal (He, 2023; Sakshi & A, 2020), mejorando la precisión del pronóstico mediante una arquitectura híbrida (Xu et al., 2022) que se representa por la fórmula siguiente:

$$Y_t = L_t + N_t$$

Donde Y_t es una combinación de partes lineales (L_t) y no lineales (N_t).

El modelo propuesto se detalla en la Figura 9. Primero, se considera el preprocesamiento del conjunto de datos en un flujo que se detalla en el apartado de preprocesamiento. Luego, el modelo ARIMA modela los componentes lineales. Posteriormente, los residuos de ARIMA son la entrada para el modelo LSTM que se utiliza para entrenar la tendencia no lineal mediante el modelado de la serie de residuos. En el desarrollo de los modelos, se aplicaron técnicas de optimización de hiperparámetros en ambos modelos para mejorar su rendimiento predictivo. En el último paso, el modelo híbrido optimizado ARIMA-LSTM genera los valores finales de pronóstico para su evaluación.

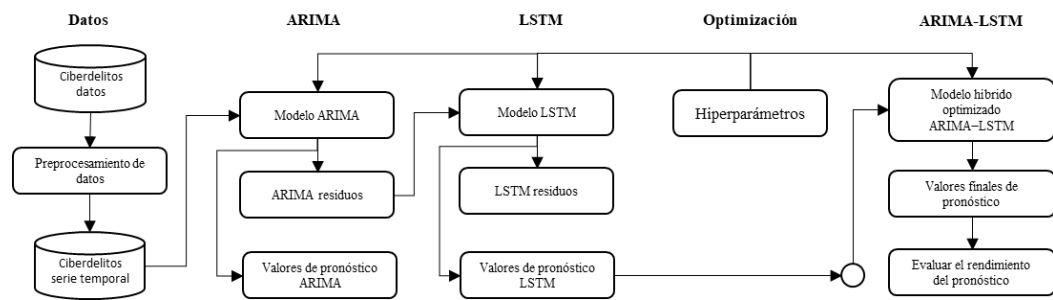


Figura 9. Diagrama de flujo modelo híbrido optimizado ARIMA-LSTM

Fuente. Elaboración propia a partir de Morales-Barrenechea et al. (2025)

4.17. Métricas de Evaluación de resultados

La exactitud del pronóstico de cada modelo de pronóstico se evaluó con los indicadores RMSE, MAE y MAPE. Los valores más bajos de estos criterios indican un mejor rendimiento del modelo híbrido ARIMA-LSTM optimizado. Las fórmulas utilizadas para calcular estos criterios (Borzemski & Wojtkiewicz, 2011; Chai & Draxler, 2014; Kasemset et al., 2023; Liu et al., 2022) como puede verse en la Tabla 11.

Tabla 11. Fórmulas de los criterios de evaluación

Criterio	Evaluación	Formula	Valores
Raíz del Error Cuadrático Medio (RMSE)	Es la raíz cuadrada del MSE y expresa el error en las mismas unidades que los datos originales, facilitando la interpretación de la precisión del modelo.	$RMS = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	Un valor 0 indica que los valores esperados y reales coinciden con precisión. Los valores reducidos de RMSE indican que el modelo genera estimaciones con un mayor grado de exactitud. Por el contrario, valores más elevados reflejan errores de mayor magnitud y, en consecuencia, predicciones menos confiables.
Error Absoluto Medio (MAE)	Cuantificar la magnitud media del error en las predicciones, prescindiendo del sentido de la desviación.	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	Un valor 0 significaría una predicción perfecta. Cuanto más pequeño sea el MAE, mejores serán las predicciones del modelo.
Promedio del error absoluto expresado en porcentaje (MAPE)	Medida porcentual para la precisión de la predicción, que es útil para comprender el error en relación con los valores reales.	$MAPE = \frac{1}{n} \sum_{i=1}^n \left \frac{y_i - \hat{y}_i}{y_i} \right $	Pronóstico muy preciso < 10% Buen pronóstico 10% - 20% Pronóstico razonable 20% - 50% Pronóstico débil e inexacto > 50%

Fuente. Elaboración propia.

Donde y_i representa el valor real, $\hat{y}_i$ el valor previsto calculado por el modelo y n representa el número de períodos previstos. Los valores bajos de estos indicadores indicaron una ligera variación entre los datos reales y los pronosticados. El modelo con los valores de criterio más bajos obtenidos se elige el modelo más adecuado (Soy Temür & Yıldız, 2021; Zou, 2025).

4.18. Metodología CRISP-DM

Es una guía metodológica diseñada para proyectos de minera de datos y también aplicada para desarrollar modelos de clasificación basado en aprendizaje automático (Pahuja & Kamal, 2023). La guía del modelo CRISP-DM describe la idea principal, las tareas y los resultados de cada una de sus fases (Maataoui et al., 2023). Se aplicó el marco metodológico CRISP-DM como procedimiento para desarrollar el modelo propuesto (Cunha et al., 2021; Gil Jauregui et al., 2024) en las seis fases según la Figura 10.

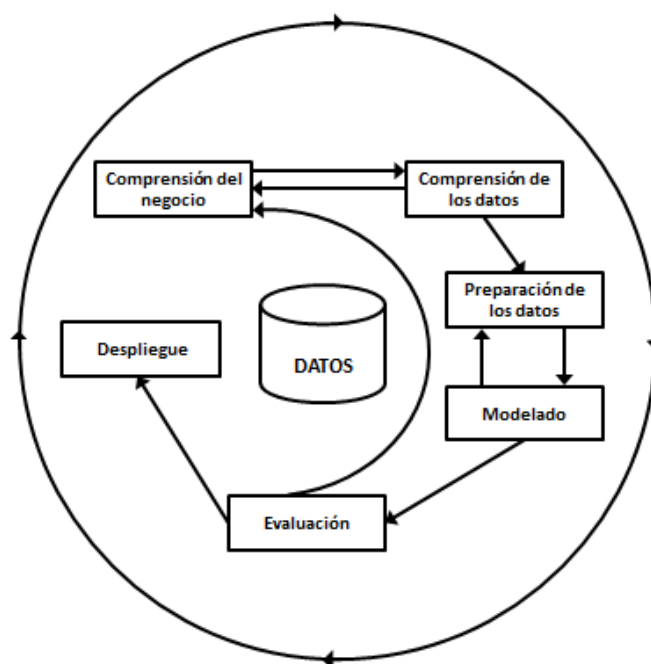


Figura 10. Fases de la metodología CRISP-DM

Fuente. Datos tomados de Maataoui et al. (2023)

El proceso metodológico comprende varias etapas interrelacionadas. En primer lugar, la comprensión del negocio busca reconocer los objetivos que persigue el proyecto desde el punto de vista de la organización, permitiendo definir el problema y planificar la estrategia para alcanzar las metas propuestas. Posteriormente, en la comprensión de los datos, se recolecta y explora la información disponible con la finalidad de detectar posibles deficiencias de calidad, subconjuntos relevantes y formular hipótesis iniciales. La fase de preparación de la data implica elaborar la versión definitiva del conjunto de datos que servirá para el modelado, seleccionando y depurando registros, atributos y tablas, además de transformar los datos para ajustarlos a las técnicas analíticas. En la etapa de modelado, se aplican distintos algoritmos y se ajustan sus parámetros para optimizar los resultados con la posibilidad de regresar a la fase anterior por algún cambio específico. En la evaluación se valida que el modelo cumpla con los objetivos propuestos por medio de métricas de desempeño. Finalmente, la fase de despliegue integra al modelo dentro de los procesos de toma de decisiones para hacer posible su aplicación práctica en el proyecto.

CAPÍTULO 5. RESULTADOS Y DISCUSIÓN

Se presenta los principales hallazgos del estudio, se describen y analizan los resultados para tener una interpretación clara del fenómeno investigado. Se desarrolló a través de las seis etapas que conforman el proceso CRISP-DM. Se exponen los hallazgos derivados del modelado y la evaluación del desempeño del enfoque propuesto, así como la validación estadística realizada mediante la prueba de hipótesis. Asimismo, se contrasta la hipótesis general junto con las hipótesis específicas, verificando su consistencia con la evidencia empírica obtenida. Finalmente, se exponen los hallazgos más significativos del estudio, destacando lo que realmente aporta al desarrollo de la investigación, resaltando su aporte metodológico y su coherencia con los objetivos planteados.

5.1. Análisis, interpretación y discusión de los resultados

El desarrollo del modelo para el pronóstico de la tasa de incidencias de ciberdelitos contra el patrimonio se aplicó la metodología CRISP-DM que es usada para desarrollar modelos de aprendizaje automático. Cada etapa del método CRISP-DM se describen en las siguientes secciones.

5.1.1. Comprensión de la organización

5.1.1.1. Policía Nacional del Perú

La PNP es la entidad peruana encargada de preservar, garantizar y restablecer el orden interno, además de ofrecer protección y asistencia a la población y a la comunidad en general. De igual forma, vela por el cumplimiento

de la normativa vigente y la salvaguarda del patrimonio público y privado. Entre sus funciones también se incluyen la prevención, investigación y lucha contra diversas manifestaciones delictivas, así como la supervisión y control de las zonas fronterizas. Todas estas acciones buscan resguardar a la ciudadanía y contribuir a su desarrollo integral dentro de un entorno de convivencia pacífica (Plataforma Digital Única del Estado Peruano, 2025).

5.1.1.2. División de Investigación de Delitos de Alta Tecnología

La DIVINDAT como órgano técnico de la PNP que desempeña labores de naturaleza técnica y operativa de la Dirección de Investigación Criminal (DIRINCRI). Esta área es responsable de ejecutar acciones orientadas a prevenir, combatir, investigar y reportar los ciberdelitos (Ley N° 30096), así como aquellos delitos perpetrados a través de la utilización de las TIC, realizado por delincuentes comunes o por organizaciones criminales (Policía Nacional del Perú, 2018).

5.1.1.3. Sistema Informático de Denuncias Policiales

El SIDPOL es la plataforma de inscripción y control de denuncias implementado por la PNP (Figura 11). Su finalidad es registrar, a nivel nacional, las denuncias policiales, clasificándolas según su modalidad, tipo, ubicación del hecho, así como los datos de las personas, armas, vehículos y bienes implicados en el hecho delictivo (Policía Nacional del Perú, 2025).

SISTEMA DE REGISTRO Y CONTROL DE DENUNCIAS

SI UD ES CAMBIADO DE COMISARÍA NOTIFIQUE INMEDIATAMENTE A LA DIRTIC PNP
PARA ACTUALIZAR SU ACCESO, CASO CONTRARIO, SUS DENUNCIAS
CONTINUARÁN SIENDO REGISTRADAS EN SU COMISARIA ANTERIOR



Iniciar sesión

CIP de usuario :

Contraseña :

☐ Deseo cambiar de contraseña

Validación Caracteres



Inicio de sesión

Figura 11. Sistema de Denuncias Policiales (SIDPOL)

Fuente. Datos tomados de la Policía Nacional del Perú

5.1.2. Comprensión de los datos

5.1.2.1. Validación de la fiabilidad de los datos

A continuación, se procedió a comparar con datos de ciberdelitos proporcionados por el MPFN (ver Anexo 2) con 56,568 denuncias registradas que también muestra una tendencia clara y constante de aumento en los ciberdelitos contra el patrimonio en el Perú (Figura 12) y la verificación de la estabilidad temporales.

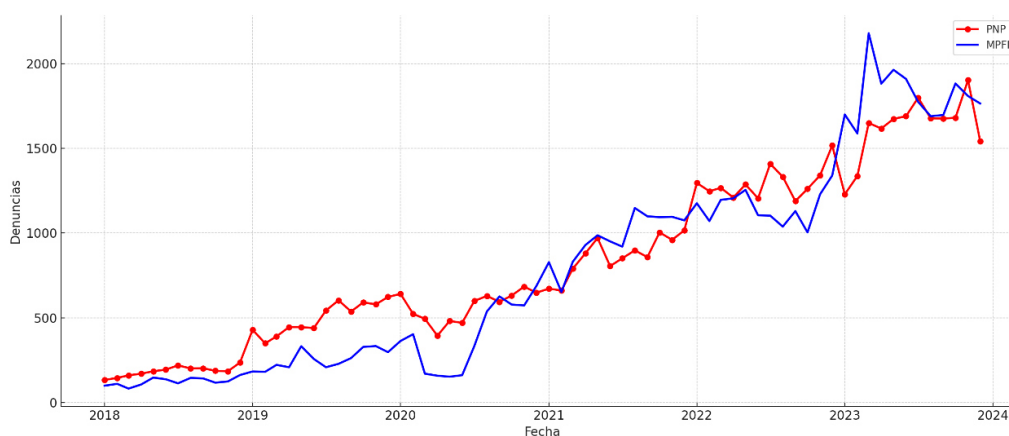


Figura 12. Comparación denuncias por fraude informático PNP y MPFN

Fuente. Elaboración propia.

5.1.2.2. Análisis exploratorio de los datos

Con la finalidad de iniciar la exploración preliminar de los datos, se empleó el lenguaje de programación Python mediante el entorno Jupyter Notebook. La Figura 13 presenta las variables del dataset de las denuncias con el subtipo de delito informático, el año y mes, y la cantidad de denuncias.

				SUB_TIPO	AÑO_MES	Q_DENUNCIAS
0	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2018-01	73
1	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2018-02	118
2	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2018-03	122
3	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2018-04	128
4	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2018-05	173
..				...	...	...
67	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2023-08	1789
68	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2023-09	1662
69	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2023-10	1743
70	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2023-11	2060
71	DELITOS	INFORMATICOS	CONTRA	EL PATRIMONIO	2023-12	1668

Figura 13. Datos de denuncias por delito informáticos contra el patrimonio
Fuente. Elaboración propia.

Tal como se aprecia en la Figura 14, se generan resultados estadísticos descriptivos incluyendo los valores correspondientes a la media, mediana, desviación estándar, mínimos, cuartiles y máximos para comprender la distribución de las denuncias por ciberdelitos contra el patrimonio.

```

Estadísticas descriptivas (numéricas):
count      72.000000
mean       838.583333
std         523.217181
min         73.000000
25%         400.500000
50%         726.000000
75%        1236.000000
max        2060.000000
Name: Q_DENUNCIAS, dtype: float64

```

Figura 14. Estadísticos descriptivos del conjunto de datos
Fuente. Elaboración propia.

Como muestra la Figura 15, la serie temporal de datos se desglosa en sus tres componentes principales: la tendencia, el patrón estacional y los residuos. Se

observa una tendencia sostenida al alza a lo largo del intervalo analizado, mientras que la estacionalidad presenta patrones repetitivos anuales con picos en diciembre y caídas en febrero, y los residuos alrededor de cero con algunos puntos atípicos (outliers) que corresponden a eventos excepcionales como la pandemia COVID-19.

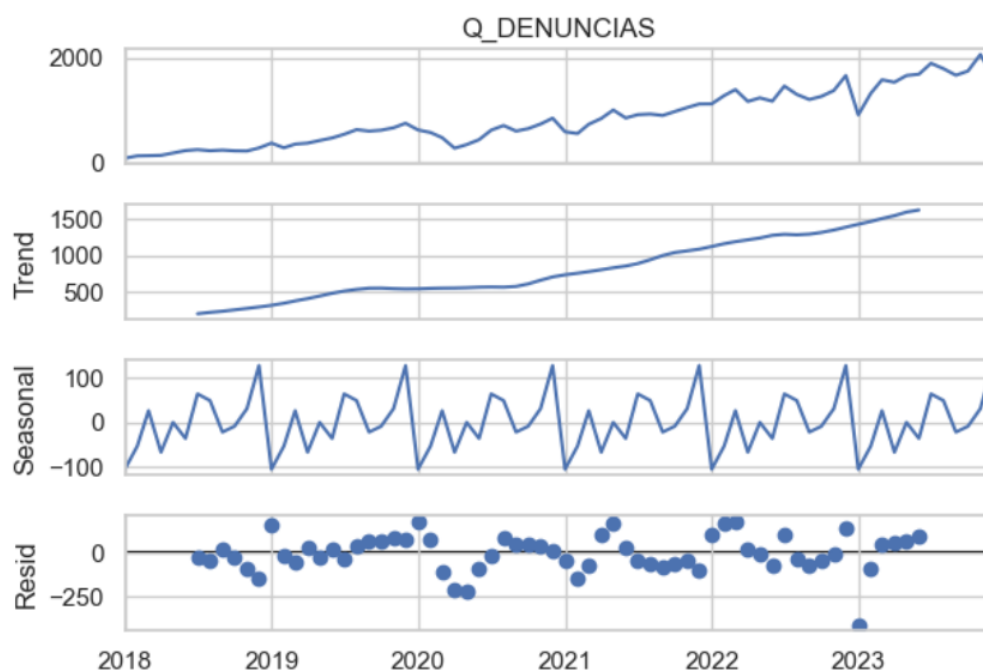


Figura 15. Descomposición de la serie temporal de datos

Fuente. Elaboración propia.

5.1.3. Preparación de los datos

Este proceso preparativo de la data se llevó a cabo utilizando el lenguaje de programación Python mediante el entorno interactivo Jupyter Notebook.

5.1.3.1. Limpieza de datos

En esta etapa, se eliminaron 8 registros que estaban fuera del rango 2018 y 2023 quedando un total de 60,378 registros, no se encontraron registros duplicados y los campos no contenían valores vacíos o nulos.

5.1.3.2. Pruebas de consistencia de los datos

Entre las pruebas de consistencia se realizó: (i) validación de fechas en el campo FECHA_HORA_HECHO que tenga el formato datetime y las fechas estén en orden cronológico, (ii) coherencia entre variables para verificar que los campos DPTO_HECHO, PROV_HECHO, DIST_HECHO, COMISARIA, TIPO, SUB_TIPO, MODALIDAD, SIT_PERSONA y SEXO no tengan valores nulos o vacíos, y la cantidad de denuncias no sean valores negativos ni inverosímiles (Figura 16; **Error! No se encuentra el origen de la referencia.**), y (iii) duplicados para detectar y eliminar filas duplicadas.

```
Validación de campos nulos:
DPTO_HECHO      0
PROV_HECHO      0
DIST_HECHO      0
COMISARIA       0
TIPO            0
SUB_TIPO        0
MODALIDAD       0
SIT_PERSONA     0
SEXO            0
FECHA_HORA_HECHO 0
AÑO            0
AÑO_MES        0
dtype: int64

Validación de campos vacíos (cadenas vacías o solo espacios):
DPTO_HECHO      0
PROV_HECHO      0
DIST_HECHO      0
COMISARIA       0
TIPO            0
SUB_TIPO        0
MODALIDAD       0
SIT_PERSONA     0
SEXO            0
FECHA_HORA_HECHO 0
AÑO            0
AÑO_MES        0
dtype: int64

Validación de valores negativos en Q_DENUNCIAS:
Cantidad de registros con Q_DENUNCIAS negativa: 0
```

Figura 16. Pruebas de consistencia de los datos

Fuente. Elaboración propia.

5.1.3.3. Estructuración de la serie temporal de datos

El sistema estructural de la serie temporal está organizado a partir de la agrupación mensual de la fecha del ciberdelito (FECHA_HORA_HECHO) donde

se obtuvo una serie temporal univariada con 72 observaciones de denuncias por ciberdelitos contra el patrimonio entre enero 2018 y diciembre 2023 según lo ilustrado en la Figura 17.

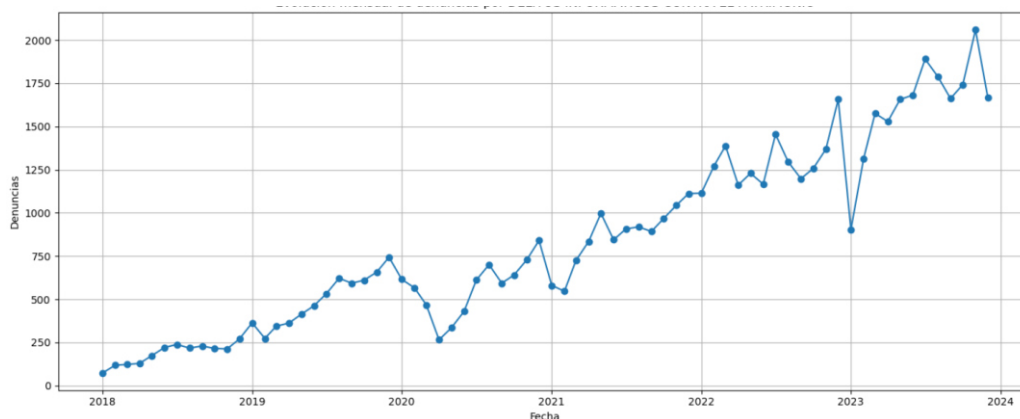


Figura 17. Serie temporal denuncias ciberdelitos contra el patrimonio

Fuente. Elaboración propia.

5.1.3.4. Prueba ADF para validar la estacionariedad de las series

A continuación, se realizó la prueba ADF para validar la estacionariedad de la serie y donde se obtuvo un $p\text{-valor}=0.991 > 0.05$ (Figura 18) por lo que la serie se consideró no estacionaria en su forma original y que se confirmó con los gráficos ACF y PACF de la Figura 19.

ADF Statistic: 0.7959713572124555
 p-value: 0.9915708590906905
 Critical Value (1%): -3.542412746661615
 Critical Value (5%): -2.910236235808284
 Critical Value (10%): -2.5927445767266866

Figura 18. Prueba de estacionariedad Augmented Dickey-Fuller

Fuente. Elaboración propia.

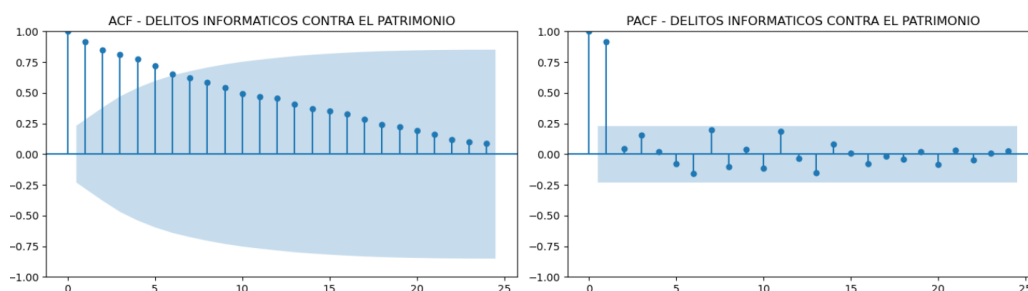


Figura 19. Gráfico ACF y PACF de la serie original (no estacionaria)

Fuente. Elaboración propia.

5.1.3.5. Diferenciación de series

Dado que la serie se consideró no estacionaria en su forma original, se aplicó la diferenciación con el método `serie.diff()` (Figura 20) obteniendo finalmente una serie transformada o estacionaria (Figura 21) como requisito clave para el modelo ARIMA.

```
serie_diff = serie.diff().dropna()
adf_diff = adfuller(serie_diff)
print("Resultados de la prueba ADF")
print("-----")
print(f"Serie diferenciada - ADF p-value: {adf_diff[1]:.5f}")
```

Resultados de la prueba ADF

Serie diferenciada - ADF p-value: 0.02940

Figura 20. Diferenciación para transformar la serie no estacionaria a estacionaria

Fuente. Elaboración propia.

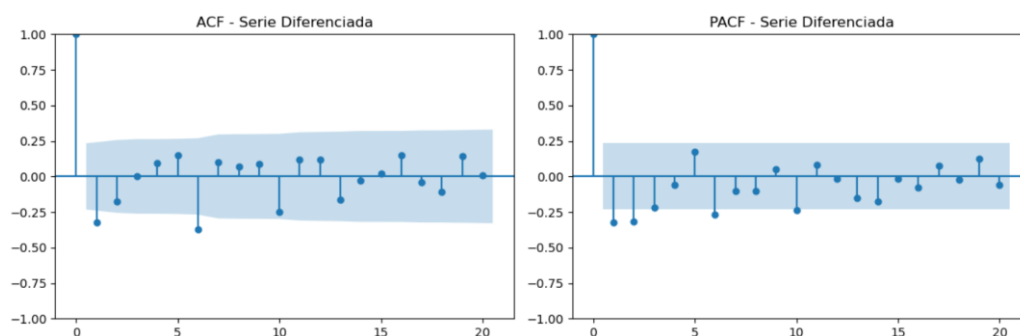


Figura 21. Gráfico ACF y PACF de la serie diferenciada (estacionaria)

Fuente. Elaboración propia.

5.1.4. Modelado

Para el entrenamiento de los modelos se recurrió al lenguaje Python y se ejecutó el trabajo desde la interfaz Jupyter Notebook, y las librerías de la inteligencia artificial `pandas`, `numpy`, `statsmodels`, `pdmarima`, `scikit-learn`, `Tensorflow`, `Keras` y `matplotlib`.

5.1.4.1. Configuración de hiperparámetros

La optimización del modelo híbrido ARIMA-LSTM se realizó a través de la configuración de los hiperparámetros que fue clave con el fin de optimizar el desempeño predictivo del modelo. Se utilizaron técnicas para automatizar el proceso de selección de las combinaciones de los parámetros con algunos ajustes manuales para conseguir un rendimiento de regresión óptimo. Las técnicas de selección y los hiperparámetros usados en los modelos de regresión tal como se puede observar en la Tabla 12 que se elaboró a partir de los resultados publicados previamente en (Barrenechea & Lengua, 2025).

Tabla 12. Configuración de hiperparámetros

Modelo	Hiperparámetros
ARIMA	- Hiperparámetros: seasonal=false, suppress warnings=true, max p=4, max q=4, trace=True, scoring=mean squared error. - Técnica: Auto Arima para obtener los parámetros p, d, q del modelo ARIMA.
LSTM	- Hiperparámetros: activation='relu', dropout rate=0.25, LSTM units=50, epochs=200, batch size=12, optimizer='adam', loss function=mean squared error. - Técnica: Random Search para seleccionar aleatoriamente combinaciones de hiperparámetros (hidden units, activation function, optimizer).

Fuente. Elaboración propia a partir de Barrenechea & Lengua (2025)

En la construcción del modelo híbrido ARIMA-LSTM, se seleccionó la configuración ARIMA(2,1,2) como resultado del proceso de ajuste. Posteriormente, se aplicó la prueba de Ljung–Box, obteniéndose un p-valor de 0.77, superior al umbral de significancia del 5%, lo que indicó que los residuos del modelo no presentan autocorrelación significativa y pueden ser considerados ruido blanco. Esta característica permitió emplearlos como entrada para el modelo LSTM, encargado de capturar los patrones no lineales remanentes.

Tanto el modelo ARIMA como el LSTM fueron optimizados mediante técnicas de ajuste de hiperparámetros como se puede apreciar en la Tabla 12. Para el método LSTM la división de la serie temporal se realizó 60-40 para los datos de entrenamiento y prueba respectivamente. Como resultado del proceso de optimización (Tabla 13), el modelo híbrido optimizado ARIMA-LSTM logró un

desempeño sobresaliente, resultando en valores de $MAE = 141.31$, $RMSE = 188.64$ y $MAPE = 10.73\%$.valor que según el criterio de evaluación establecido en la Tabla 11, se clasifica como de alta precisión predictiva.

Tabla 13. Comparación entre el modelo optimizado ARIMA, LSTM y ARIMA-LSTM

Modelo	Criterio		
	MAE	RMSE	MAPE %
ARIMA	104.39	143.95	15.43
LSTM	196.75	249.25	13.28
Híbrido optimizado ARIMA-LSTM	141.31	188.64	10.73

Fuente. Elaboración propia.

5.1.4.2. Modelo ARIMA

Se construyó un modelo ARIMA simple y un modelo ARIMA optimizado para comparar las métricas de rendimiento de ambos modelos. Dado que la serie original no es estacionaria, lo cual ya había sido indicado por la prueba ADF (Figura 18) se aplicó la diferenciación obteniendo un nuevo ADF con p-value de 0.03 (Figura 20) que es menor umbral común de significancia (0.05).

El modelo ARIMA simple obtuvo un AIC de 919.37 para el orden ARIMA (1,1,1). El modelo ARIMA fue optimizado por Grid search para el obtener el orden ARIMA(1,1,2) con un AIC de 916.58 como métrica de selección y donde los resultados de rendimiento del modelo optimizado para $MAE=104.39$, $RMSE=143.95$ y $MAPE=15.43\%$ (buena precisión) considerados menores comparadas al modelo simple según lo ilustrado en la Tabla 14.

Se efectuó la estimación del porcentaje de mejora porcentual a través de la siguiente formula:

$$\Delta MAPE = \left(\frac{(MAPE_{simple} - MAPE_{optimizado})}{MAPE_{simple}} \right) \times 100$$

Donde $\Delta MAPE > 0$ se considera una mejora y $\Delta MAPE < 0$ un deterioro del modelo.

La mejora porcentual obtenida $\Delta MAPE \approx 0.45\%$ al usar el modelo

ARIMA optimizado respecto al modelo ARIMA.

Tabla 14. Comparación modelo ARIMA vs ARIMA optimizado

ARIMA simple			ARIMA optimizado		
MAE	RMSE	MAPE %	MAE	RMSE	MAPE %
110.42	150.19	15.50	104.39	143.95	15.43

Fuente. Elaboración propia.

Según muestra la Figura 22, el modelo ARIMA optimizado que captura acertadamente los componentes autorregresivos y de media móvil han sido adecuadamente ajustados y la proyección futura demuestra un aumento sostenido en los ciberdelitos contra el patrimonio.

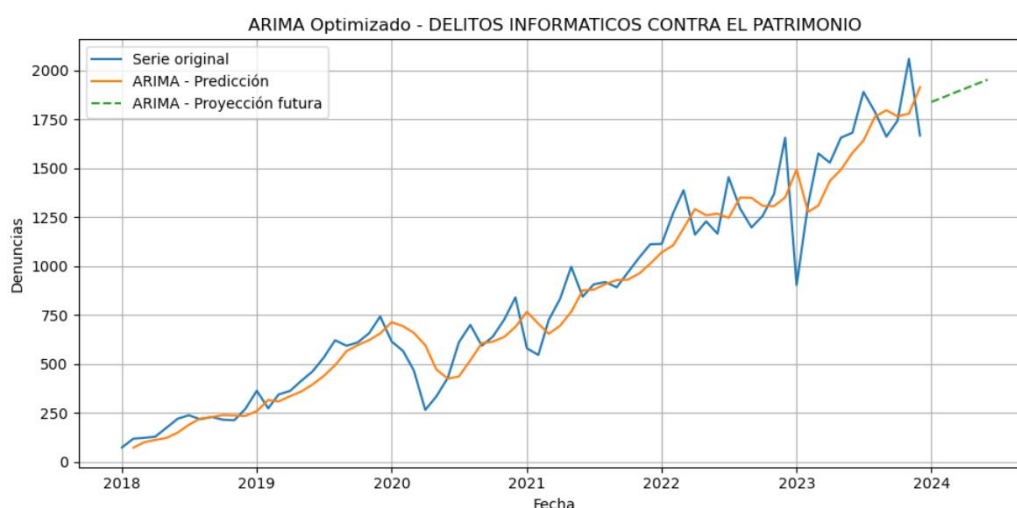


Figura 22. Modelo ARIMA optimizado

Fuente. Elaboración propia.

5.1.4.3. Modelo LSTM

Se construyó un modelo LSTM y un modelo LSTM optimizado para comparar las métricas de rendimiento de ambos modelos. Se utilizó la técnica MinMaxScaler para el escalamiento de datos, 6 pasos que es la cantidad de observaciones anteriores que se utilizaron como entrada para predecir el siguiente valor de la serie temporal. Para el entrenamiento del conjunto de datos se

utilizaron 50 épocas o pasadas.

El modelo LSTM optimizado se desarrolló mediante Random Search para varias combinaciones de hiperparámetros (unidades, activación, optimizador) usando Keras Tuner con validación interna y Early Stopping para evitar sobreajuste. Los resultados de métricas de rendimiento del modelo optimizado sobresalieron en RMSE=249.25 y MAPE=13.28% (buena precisión) que fueron menores comparadas al modelo simple como se evidencia en la Tabla 15.

Tabla 15. Comparación modelo LSTM simple vs LSTM optimizado

LSTM simple			LSTM optimizado		
MAE	RMSE	MAPE %	MAE	RMSE	MAPE %
186.08	257.01	13.82	196.75	249.25	13.28

Fuente. Elaboración propia.

Se cuantificó la $\Delta\text{MAPE} \approx 3.91\%$ de mejora en MAPE al usar el modelo LSTM optimizado frente al LSTM simple.

La Figura 23 muestra el modelo LSTM optimizado que logró ajustarse a la tendencia creciente de las denuncias por ciberdelitos contra el patrimonio. No obstante, la suavización excesiva y la falta de picos sugieren que, si bien el modelo es útil para su aplicación general, se sugiere un enfoque híbrido con ARIMA para mejorar la precisión ante cambios abruptos o eventos atípicos.

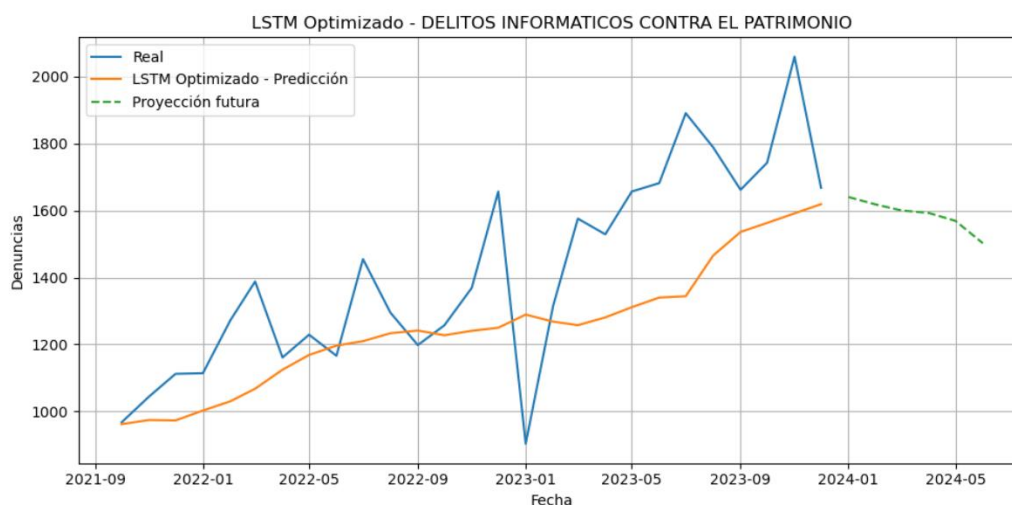


Figura 23. Modelo LSTM optimizado

Fuente. Elaboración propia

5.1.4.4. Modelo híbrido ARIMA-LSTM optimizado

El modelo híbrido se optimiza por la combinación de la predicción de ARIMA que captura la tendencia y linealidad más los residuos ajustados con LSTM que captura la no linealidad. El componente ARIMA se optimizó automáticamente con Auto Arima, buscando el modelo ARIMA (2,1,2) con el menor AIC de 919.20.

El componente LSTM se optimizó mediante la técnica MinMaxScaler para el escalamiento de datos, 5 como mejor número de pasos evaluando MAPE a través de *Grid* manual. El modelo se ajustó con 1a capa LSTM con 50 neuronas, la función de activación 'relu', el optimizador de pesos de red 'adam' y 200 épocas o pasadas por los datos de entrenamiento.

El modelo híbrido optimizado ARIMA-LSTM supera al modelo híbrido simple en MAE=141.31, RSME=188.64 y MAPE=10.73% (buena precisión) tal como se muestra en la Tabla 16. Se evidencia que la optimización de arquitectura y parámetros del componente LSTM mejora sustancialmente la capacidad predictiva de modelos híbridos aplicados a series temporales complejas como las denuncias por ciberdelitos.

Tabla 16. Comparación modelo híbrido ARIMA-LSTM vs ARIMA-LSTM optimizado

Híbrido ARIMA-LSTM			Híbrido optimizado ARIMA-LSTM		
MAE	RMSE	MAPE %	MAE	RMSE	MAPE %
167.87	217.07	12.45	141.31	188.64	10.73

Fuente. Elaboración propia

Se cuantificó la $\Delta\text{MAPE} \approx 14\%$ de mejora en MAPE al usar el modelo híbrido optimizado ARIMA-LSTM optimizado frente a los individuales ARIMA y LSTM.

En el Figura 24 el modelo híbrido optimizado ARIMA-LSTM demostró mayor capacidad de adaptación a los patrones temporales de la serie de denuncias superando a los modelos individuales. La predicción futura conserva el comportamiento de la serie reciente sin generar fluctuaciones, lo que refuerza su

utilidad para análisis prospectivos en ciberdelitos y prevención del delito.



Figura 24. Modelo ARIMA-LSTM optimizado

Fuente. Elaboración propia

5.1.5. Evaluación de resultados

Para evaluar los resultados de los modelos ARIMA, LSTM y ARIMA–LSTM optimizado se utilizaron las métricas MAE, RMSE y MAPE que se muestran en la Tabla 17. De acuerdo con los valores presentados, el modelo híbrido ARIMA-LSTM optimizado obtuvo el menor MAPE (10.73%), lo que indica un error porcentual absoluto medio inferior al de los modelos individuales ARIMA (15.43%) y LSTM (13.28%). Esta reducción en el error porcentual implica una mejora significativa en la precisión relativa del modelo híbrido optimizado.

Tabla 17. Resultados modelos ARIMA, LSTM y optimizado ARIMA-LSTM

Modelo	Simple			Optimizado		
	MAE	RMSE	MAPE %	MAE	RMSE	MAPE %
ARIMA	110.42	150.19	15.50	104.39	143.95	15.43
LSTM	186.08	257.01	13.82	196.75	249.25	13.28
ARIMA-LSTM	167.87	217.07	12.45	141.31	188.64	10.73

Fuente. Elaboración propia

5.2. Prueba de Hipótesis

Esta investigación consideró la comprobación de la hipótesis general (HG) y también de las tres hipótesis específicas (HE) por métodos estadísticos y métricas de rendimiento según lo planteado en cada objetivo. Las validaciones y comprobaciones se realizaron de la siguiente forma:

HE1: La estacionariedad de la serie temporal de denuncias por ciberdelitos contra el patrimonio mejora significativamente la eficiencia y precisión del modelo ARIMA en el pronóstico.

Para validar esta hipótesis, se aplicó la prueba estadística Augmented Dickey–Fuller (ADF), la cual permite determinar si una serie temporal es estacionaria. Las hipótesis estadísticas planteadas fueron:

H0: (p-valor >0.05), la serie temporal es no estacionaria.

H1: (p-valor <0.05), la serie temporal es estacionaria.

En el análisis realizado, Se obtuvo un p-valor = 0.03, el cual es inferior al umbral de significancia establecido de 0.05. En consecuencia, se descarta la hipótesis nula (H0) y se confirma la hipótesis alternativa (H1), lo que permite concluir que la serie temporal presenta estacionariedad.

HE2: La integración del modelo ARIMA con el modelo LSTM en un enfoque híbrido optimizado mejora significativamente la capacidad de pronóstico en comparación con los modelos individuales ARIMA y LSTM.

Para validar esta hipótesis, se aplicó la prueba estadística Diebold–Mariano (DM), utilizada para comparar la precisión de dos modelos de pronóstico. Se plantearon las siguientes hipótesis estadísticas siguientes:

H0: (p-valor >0.05) No hay diferencia significativa en la precisión entre el modelo híbrido ARIMA–LSTM optimizado y los modelos individuales ARIMA y LSTM.

H1: (p-valor <0.05) Existe una diferencia significativa en la precisión entre el modelo híbrido ARIMA–LSTM optimizado y los modelos individuales ARIMA

y LSTM. Se validó que el resultado de la prueba DM en ambas comparaciones fue:

El modelo híbrido ARIMA-LSTM optimizado comparado con el modelo individual ARIMA obtuvo un p-valor = 0.03, el cual es menor que 0.05. El modelo híbrido ARIMA-LSTM optimizado comparado con el modelo individual LSTM arrojó un p-valor = 0.03, al compararlo con el nivel de significancia de 0.05, el resultado indica que la hipótesis nula (H_0) no puede sostenerse. Por ello, se descarta y se da respaldo a la hipótesis alternativa (H_1), concluyéndose que el modelo híbrido ARIMA-LSTM optimizado presenta una mejora significativa en la precisión del pronóstico en comparación con los modelos individuales.

HE3: El modelo híbrido optimizado ARIMA–LSTM alcanza un valor de MAPE significativamente inferior al de los modelos individuales ARIMA y LSTM, lo que demuestra una mayor precisión en el pronóstico de denuncias por ciberdelitos contra el patrimonio.

Para validar esta hipótesis, se determinó la capacidad predictiva según el rendimiento MAPE para tres enfoques: modelo ARIMA, modelo LSTM y el modelo híbrido optimizado ARIMA-LSTM. Este índice se utiliza principalmente para determinar el error porcentual promedio de los modelos de series temporales al medir la precisión relativa entre el estado real y el predicho. Se plantearon las hipótesis estadísticas siguientes :

H_0 : Si el MAPE del modelo ARIMA-LSTM optimizado es menor que el MAPE de los modelos ARIMA y LSTM individuales, entonces el modelo híbrido presenta mayor precisión en el pronóstico.

H_1 : Si el MAPE del modelo ARIMA-LSTM optimizado es mayor que el MAPE de los modelos ARIMA y LSTM individuales, entonces el modelo híbrido no supera en precisión a los modelos individuales.

De acuerdo con los resultados que se presentan en la Tabla 17; **Error! No se encuentra el origen de la referencia.**, el modelo híbrido ARIMA–LSTM optimizado obtuvo el menor MAPE (10.73%), lo que indica un error porcentual absoluto medio inferior al de los modelos individuales ARIMA (15.43%) y LSTM (13.28%). Esta reducción en el error porcentual implica una mejora significativa

en la precisión relativa del modelo híbrido optimizado.

En consecuencia, se mantiene la validez de la hipótesis nula (H_0) y se descarta la hipótesis alternativa (H_1), determinándose que el modelo híbrido ARIMA-LSTM optimizado ofrece una mayor precisión en el pronóstico de denuncias de ciberdelitos contra el patrimonio en comparación con los modelos individuales.

HG: El desarrollo de un modelo híbrido optimizado ARIMA–LSTM mejora significativamente la precisión en el pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú, en comparación con los modelos individuales ARIMA y LSTM.

H_0 plantea que no hay una diferencia que pueda considerarse estadísticamente relevante en la precisión del pronóstico denuncias por ciberdelitos entre el modelo híbrido ARIMA–LSTM y los modelos individuales ARIMA y LSTM.

Para la validación de la hipótesis general (HG), se contrastaron las hipótesis específicas (HE1, HE2 y HE3) mediante pruebas estadísticas y métricas de desempeño del modelo (MAPE, RMSE, MAE). Los resultados, resumidos en la Tabla 18 evidencian que las tres hipótesis específicas fueron verificadas y aceptadas, lo que demuestra la validez del enfoque híbrido propuesto.

Tabla 18. Comprobación de las hipótesis específicas

Hipótesis	Conclusión	Estado
HE1	La serie temporal puede hacerse estacionaria, condición necesaria para el funcionamiento adecuado del modelo ARIMA.).	Comprobada
HE2	El modelo híbrido ARIMA–LSTM presenta un desempeño superior en capacidad de pronóstico frente a los modelos individuales ARIMA y LSTM.	Comprobada
HE3	El modelo híbrido ARIMA–LSTM optimizado alcanza un valor de MAPE significativamente menor, validando una mayor precisión de pronóstico.	Comprobada

Fuente. Elaboración propia

Por lo tanto, los resultados llevan a descartar la hipótesis nula (H_0) y dar respaldo a la hipótesis general (HG), concluyendo que el modelo híbrido optimizado ARIMA–LSTM mejora de manera significativa la precisión del pronóstico de denuncias por ciberdelitos contra el patrimonio en comparación con los modelos

individuales.

5.3. Presentación de resultados

El modelo de pronóstico de denuncias por ciberdelitos contra el patrimonio se desarrolló bajo la metodología CRISP-DM. Se propone el modelo híbrido ARIMA-LSTM, un modelo para capturar tanto los componentes lineales de los datos como los patrones no lineales. La estructura lineal de la serie, como las tendencias, autocorrelaciones y estacionalidad, se modeló utilizando el modelo estadístico ARIMA. La red LSTM se usó para detectar y aprender de las dependencias temporales no lineales que no pueden ser modeladas usando modelos clásicos. La optimización de hiperparámetros se aplica mediante Random Search para optimizar varias combinaciones de parámetros y el enfoque de Early Stopping para mantener la robustez del modelo y evitar el sobreajuste.

Ambos enfoques combinados operan bajo una lógica de complementariedad. También se informa por (Kasemset et al., 2023; Sakshi & A, 2020) que los modelos híbridos ARIMA-LSTM pueden mejorar efectivamente la precisión del pronóstico de los datos al combinar las ventajas de ambos enfoques en el caso del análisis de series temporales. Esta arquitectura mixta permite a los modelos trabajar con dinámicas lineales y no lineales, lo cual es el caso de los ciberdelitos, ya que pueden ser influenciados tanto por patrones constantes como por eventos ocasionales. En conclusión, nuestro modelo propuesto también es un marco híbrido estadístico-neuronal, diseñado para abordar la complejidad de las denuncias de ciberdelitos en Perú, lo que proporciona estimaciones más precisas y útiles para la formulación de políticas estratégicas en la lucha contra el ciberdelito.

El conjunto de datos utilizado para construir el modelo de pronóstico corresponde a incidentes específicos de ciberdelitos contra la propiedad reconocidos por la PNP en el sistema SIDPOL. El conjunto de datos desde enero del 2018 hasta diciembre del 2023 contiene un total de 60,378 registros válidos que han sido procesados, limpiados y organizados como una serie temporal de

frecuencia mensual. Cada registro representa una denuncia y contiene la ubicación en la que ocurrió el hecho, la comisaría, el tipo de ciberdelito y la fecha de la ocurrencia. Para fines de este estudio, se consolidaron los datos por mes, dando como resultado una serie temporal univariada con 72 observaciones mensuales de denuncias por ciberdelitos contra el patrimonio entre el 1 de enero de 2018 y el 1 de diciembre de 2023. Esta serie es la base sobre la cual se soporta el modelo de pronóstico de denuncias de ciberdelitos contra el patrimonio.

Para entrenar y evaluar adecuadamente el modelo híbrido ARIMA-LSTM, se aplicó la estrategia de división temporal de los datos 60-40 para el entrenamiento la prueba respectivamente y respetando el orden cronológico de la serie para mantener la estructura temporal. La división quedó establecida de la siguiente manera: la serie original de enero del 2018 hasta diciembre de 2023 equivalente a 72 meses de observaciones, para el periodo de entrenamiento desde enero del 2018 hasta setiembre del 2022 correspondiente a 57 meses (60% del total) sobre los cuales se refinaron los parámetros de los modelos, y en la etapa de prueba desde octubre del 2022 hasta diciembre del 2023 equivalente a 15 meses (40% del total) utilizado para validar el desempeño predictivo del modelo. La proyección futura, del modelo entrenado y validado, fue de 12 meses adicionales abarcando el periodo desde diciembre de 2023 hasta noviembre de 2024. Esta segmentación temporal permitió una evaluación rigurosa del modelo bajo condiciones realistas, asegurando que las predicciones se realizaran sobre datos no observados durante el entrenamiento, lo cual fortalece la validez del modelo híbrido propuesto.

Los resultados cuantitativos del modelo híbrido optimizado ARIMA-LSTM fueron evaluados utilizando métricas estándar de error MAE, RMSE y MAPE aplicadas al periodo de prueba. Como se indicó anteriormente, este modelo combina un modelo ARIMA (2,1,2) con una red neuronal LSTM cuyos hiperparámetros fueron optimizados mediante búsqueda aleatoria, validación interna y mecanismo de detención temprana. La comparación de resultados se realizó frente a los modelos base de ARIMA (1,1,1) y LSTM individuales con lag de 3 pasos. Se obtuvo un resultado significativamente menor de MAPE para el modelo híbrido optimizado ARIMA-LSTM (10.73%) respecto a los modelos individuales ARIMA (15.50%) y LSTM (13.82%) según muestra

la Tabla 17. Estos resultados evidencian una mejor capacidad para predecir la proporción del error respecto a los valores registrados.

En el análisis visual de los resultados del modelo híbrido optimizado ARIMA-LSTM (Figura 24) se destaca la coincidencia entre curvas con una alta correspondencia entre la serie original y la curva generada por el modelo híbrido optimizado en el periodo de entrenamiento y validación. Las líneas de la serie original y el modelo híbrido optimizado muestran un comportamiento paralelo, lo que indica que el modelo ha sido capaz de seguir con precisión la dinámica de las denuncias reales. La tendencia general muestra tendencia creciente en la cantidad de denuncias desde el 2018 hasta el 2023, con incrementos marcados a partir del año 2021. Esta tendencia es captada de manera consistente por el modelo, lo que refuerza su capacidad para incorporar estructuras a largo plazo.

En la estacionalidad y fluctuaciones de la curva, no se observa una estacionalidad perfectamente regular, pero se identifican ciclos con incrementos periódicos y caídas abruptas. El modelo logra adaptarse a dichas oscilaciones, aunque se evidencian pequeñas discrepancias en algunos picos extremos, atribuibles a eventos atípicos no capturados por la lógica secuencial del modelo. La proyección futura para el año 2024 de la línea verde punteada representa la proyección a 12 meses realizada con el modelo híbrido optimizado entrenado. Se mantiene una tendencia moderadamente creciente, con cierta estabilización en los niveles altos de denuncias, lo que podría sugerir cierta estabilización en la cantidad de denuncias de ciberdelitos.

Otro hallazgo relevante es la caída en la serie real a mediados de 2020, atribuible a las restricciones durante la pandemia COVID-19. El modelo de pronóstico reconoce dicha anomalía, aunque la suaviza, lo cual es esperable por su enfoque promedio y generalizador. En conclusión, los resultados logrados indican que el modelo híbrido optimizado ARIMA-LSTM ofrece una mejora significativa en la precisión predictiva respecto a los modelos individuales según la medición de MAPE. El desempeño cuantitativo (Tabla 16) visual en la gráfica (Figura 24) con el modelo híbrido optimizado confirma su idoneidad como herramienta de pronóstico en el ámbito de la criminalidad digital, siendo capaz de adaptarse a variaciones complejas en el comportamiento de la serie temporal analizada.

La robustez del modelo híbrido optimizado ARIMA-LSTM fue evaluada utilizando verificación estadística, empleando tres hipótesis individuales, con el objetivo de evaluar la fiabilidad del método desde una perspectiva metodológica y comparativa. Estas validaciones estaban dirigidas a validar la estacionariedad de la serie temporal (HE1), evaluar el rendimiento del modelo híbrido en comparación con los componentes individuales (HE2), y evaluar cuantitativamente su fiabilidad predictiva empleando MAE, RMSE y MAPE. Estas validaciones fueron útiles para verificar la validez del modelo propuesto. La estacionariedad de la serie temporal es un criterio importante en la validación de HE1 para validar la aplicación del modelo ARIMA, se realizó la prueba estadística de Dickey-Fuller Aumentada (ADF), que arrojó un valor p de 0.991 por encima del nivel de significancia de 0.05. Este resultado nos permitió concluir que la serie no era estacionaria en su estado original, como fue validado por los gráficos (Figura 19) de Autocorrelación (ACF) y Autocorrelación Parcial (PACF).

Para convertir la serie a estacionaria, se aplicó una diferenciación simple mediante el método `serie.diff`. Luego, se volvió a aplicar la prueba ADF a la serie transformada, obteniéndose un p-valor de 0.03, el cual es menor al umbral de 0.05. Este resultado indicó que la serie diferenciada cumplió con la condición de estacionariedad que fue corroborado con nuevos gráficos ACF y PACF (Figura 21) con lo que se validó la hipótesis HE1. En la validación de la HE2, se planteó que el modelo híbrido ARIMA-LSTM optimizado superaría significativamente, en términos de precisión de pronóstico, a los modelos individuales ARIMA y LSTM. Se realizó la evaluación estadística de Diebold–Mariano (DM) a fin de verificar la calidad de las predicciones de dos modelos aplicados sobre una misma serie temporal. Los hallazgos obtenidos al comparar el modelo híbrido ARIMA–LSTM optimizado y el modelo ARIMA individual se obtuvo un p-valor = 0.03 y en la comparación entre el modelo híbrido ARIMA-LSTM optimizado y el modelo LSTM individual se obtuvo un p-valor = 0.03. En ambos casos, los valores obtenidos son menores al umbral de significancia del 5%, por lo que el modelo híbrido ofrece una mejora estadísticamente significativa en su capacidad de pronóstico respecto a los modelos individuales con lo que validó la hipótesis HE2.

La validación de la HE3 se centró en el análisis cuantitativo sobre el

rendimiento del modelo híbrido respecto al error de pronóstico con MAPE que permitió evaluar el error relativo en función de los valores reales observados. Se obtuvo un resultado significativamente menor de MAPE para el modelo híbrido optimizado ARIMA-LSTM (10.73%) respecto a los modelos individuales ARIMA (15.50%) y LSTM (13.82%). Esta reducción en el error del modelo híbrido optimizado implica una mejora cuantitativa relevante en la precisión relativa del modelo, lo que confirma su fiabilidad como herramienta de pronóstico con lo que se validó la hipótesis HE3. En resumen, las tres hipótesis planteadas fueron validadas de manera estadística y los resultados permitieron afirmar la HG donde el modelo híbrido ARIMA-LSTM optimizado es estadísticamente confiable y predictivamente eficiente en el contexto del pronóstico de denuncias por ciberdelitos contra el patrimonio en el Perú.

La interpretación de los resultados obtenidos de la arquitectura híbrida optimizada ARIMA-LSTM presentan una clara tendencia creciente en las denuncias por ciberdelitos contra el patrimonio registradas a lo largo del intervalo que abarca desde enero de 2018 hasta diciembre de 2023, con una proyección del modelo que anticipa un crecimiento hacia finales de 2024. Esto se evidencia tanto en el comportamiento general de la serie como en la trayectoria de la curva de predicción del modelo. La precisión del modelo híbrido optimizado, reflejada en un MAPE del 10.73%, confirma su capacidad para ajustarse adecuadamente a las fluctuaciones de la serie y anticipar posibles trayectorias futuras. A lo largo del periodo de observación se pueden identificar momentos de incremento o disminución que responden a diversos factores de índole tecnológico y social.

La evaluación de la información registrada evidencia que las denuncias por ciberdelitos en el Perú experimentaron un incremento de aproximadamente cuatro veces entre los años 2018 y 2021, aumentando de 3,031 a más de 12,827 casos según informe (Defensoría del Pueblo, 2023). Este incremento sostenido está relacionado al crecimiento del acceso a internet, el aumento en la utilización de servicios digitales y la mayor vulnerabilidad de los usuarios frente a delitos como phishing, fraude digital y usurpación de la identificación de la persona, particularmente durante la pandemia por COVID-19 según (Diario Inforbae, 2020). Durante el año 2020 se identifica una caída transitoria en las denuncias en la segunda mitad del año que coincide con las restricciones implementadas por el

Estado dentro del periodo de emergencia nacional. A partir del año 2021, se observa un nuevo repunte que alcanza niveles históricos en 2022, superando las 22,000 denuncias anuales, según reportes periodísticos citados por (Diario Gestión, 2024).

El modelo híbrido optimizado ARIMA-LSTM también proyecta una continuidad en la tendencia ascendente para el año 2024 pero con una pendiente ligeramente más moderada. Esta posible estabilización en la tasa de crecimiento podría deberse al fortalecimiento de las capacidades de ciberseguridad del sector público y privado, y una ciudadanía más informada sobre los mecanismos de prevención y denuncia. Las políticas públicas implementadas por el Ministerio de Justicia, la PNP y otras instituciones han promovido la articulación de respuestas frente al delito cibernético, tal como se señala en el estudio sobre ciberdelitos en el país y sugerencias a la ciudadanía (Ministerio de Justicia y Derechos Humanos, 2022). Por otro lado, el pronóstico muestra que los delitos cibernéticos continúan siendo un nivel de riesgo permanente y creciente para la seguridad digital del país.

La capacidad del modelo para predecir el comportamiento de las denuncias permite a estos actores realizar una planificación más precisa para la distribución de recursos operativos, el diseño de medidas preventivas y la configuración de unidades especializadas. Por ejemplo, si hay un pronóstico de aumento de denuncias en algunos meses, las autoridades podrían mejorar la capacidad de trabajo de la DIVINDAT o crear campañas educativas dirigidas a grupos o áreas vulnerables. De manera similar, el MPFN puede aplicar estas suposiciones para mejorar su capacidad de acción fiscal y facilitar la gestión óptima de la carga procesal de los delitos cibernéticos. En este sentido, el modelo híbrido optimizado ARIMA-LSTM ya no es solo un instrumento técnico, sino que también es una plataforma práctica para guiar políticas públicas y aumentar la eficiencia institucional en la campaña contra el ciberdelito.

Los hallazgos del modelo híbrido optimizado ARIMA-LSTM no solo reproducen los patrones históricos observados, sino que también proporcionan la capacidad de contextualizar estos patrones en términos de factores sociales, políticos y tecnológicos. El modelo permite incluir la observación general y las anomalías, lo que indica su utilidad como herramienta de análisis predictivo. Este comportamiento va en línea con los eventos que tuvieron lugar en Perú en el

momento estudiado, y proporciona una base sólida para tomar mejores decisiones orientadas a la prevención del delito, el diseño de políticas públicas y la planificación de estrategias para la seguridad digital.

En comparación con estudios anteriores, la combinación ARIMA-LSTM mejora el MAPE en un 14% comparado con los modelos individuales ARIMA (15.50%) y LSTM (13.82%) en la predicción de incidentes de fraude cibernético dentro de Perú. Esta mejora se alinea con los hallazgos en (Sharmin et al., 2022) para la predicción criminal en Londres entre el 2014 y 2020, donde el modelo propuesto ARIMA-LSTM también superó de manera consistente a modelos individuales, alcanzando una precisión del 97% frente a ARIMA (89%) y el modelo de redes neuronales (87%). A diferencia de dicho estudio, que aplicó la función Auto Arima para la selección de parámetros y una arquitectura stacked para LSTM, en este estudio se optó por una optimización basada en hiperparámetros y adaptando el modelo a un conjunto de datos de 60,378 denuncias entre el 2018 y 2023. Sin embargo, se encontró que la mejora en precisión del modelo híbrido no solo depende de la estructura técnica, sino también del contexto delictivo y la calidad de los datos temporales disponibles. Mientras que la exploración de Londres abarcó 14 tipos de delitos comunes y series con mayor dispersión, el presente trabajo se centró exclusivamente en el ciberdelito contra el patrimonio, permitiendo una calibración más fina del modelo y resultados más representativos dentro del campo de la cibercriminalidad. Por lo tanto, ambos estudios refuerzan la solidez de la metodología bajo el modelo híbrido ARIMA-LSTM utilizado y permiten la aplicabilidad del enfoque basado en el modelo híbrido en varios escenarios donde el tipo de datos cubre elementos lineales y no lineales.

En contraste con el estudio realizado en India (Natarajan, 2024), con 9,840 registros de diferentes tipos de delitos desde 2001 hasta 2013, cuando la combinación ARIMA-LSTM alcanzó una precisión del 97%, el estudio actual añade evidencia de una métrica aún más rigurosa como MAPE y una categoría particular de ciberdelito. Sin embargo, concluimos que, si bien ambos estudios encontraron que el modelo híbrido era preferido a los constructos específicos de ARIMA y LSTM, el presente trabajo aplica más validación estadística, como la prueba de Diebold-Mariano y la optimización de hiperparámetros utilizando la

metodología Auto Arima y Random Search, lo que mejora su fiabilidad metodológica. No solo estas diferencias en contexto y tecnología verifican la robustez de nuestro enfoque ARIMA-LSTM, sino también que es capaz de responder de manera flexible a fenómenos criminales emergentes desde estas nuevas perspectivas como el ciberdelito.

El otro trabajo relevante es el estudio (Deng et al., 2023), que ha adoptado un modelo híbrido basado en un conjunto de datos de médicos con una frecuencia semanal y una ventana de tiempo de solo 24 semanas. En este caso, el modelo híbrido alcanzó un MAPE de 14.13%, sin superar el desempeño logrado en este estudio de 10.73%. A diferencia de ese análisis, el modelo propuesto fue entrenado sobre una base más extensa (72 meses) y con un volumen significativamente mayor de casos (60,378 denuncias), lo que indica una mayor capacidad de generalización y solidez ante patrones temporales complejos. Los hallazgos indican que, incluso en condiciones más exigentes, el modelo optimizado logra una mejora del MAPE superior al 32%, lo cual refuerza la efectividad de la optimización con hiperparámetros y la aplicación del modelo híbrido en dominios de alta volatilidad como la criminalidad digital.

El estudio (Kasemset et al., 2023) , emplea una serie temporal de 84 meses sobre precios agrícolas, el modelo ARIMA-LSTM obtuvo un MAPE de 13.62%, superando a los modelos individuales ARIMA (16.60%) y al LSTM (15.04%). De igual manera, en esta investigación aplicada a los ciberdelitos en el Perú, el modelo híbrido optimizado alcanzó un MAPE de 10.73%, también por debajo del modelo ARIMA (15.43%) y del modelo LSTM (13.28%). Estos hallazgos confirman que la combinación ARIMA-LSTM de ambos estudios mejoran considerablemente el MAPE respecto a los modelos individuales, lo cual refuerza la validez del enfoque híbrido en contextos de predicción temporal con alta variabilidad. Sin embargo, a diferencia de (Kasemset et al., 2023), este estudio incorporó en el modelo procesos de ajuste de hiperparámetros que contribuyeron a la mejora sustancial del desempeño predictivo en el contexto del cibercrimen.

Además, se desarrolló el modelo híbrido ARIMA-LSTM optimizado que capturó tanto las tendencias lineales como las variaciones no lineales con mayor precisión en este estudio en comparación con estudios anteriores. Esta capacidad de integración ha logrado, por lo tanto, un mejor rendimiento relativo y absoluto

en la predicción de informes de ciberdelitos. Aunque los resultados de este experimento confirman que la aplicación del enfoque híbrido a estos casos es útil, también sirven para extender el rendimiento propuesto a un subconjunto más amplio de problemas sociales o criminales con comportamientos temporales más complejos. De manera similar, la futura incorporación de variables exógenas, por ejemplo, indicadores de acceso a internet, eventos tecnológicos o políticas públicas, podría añadir más potencial para enriquecer la capacidad predictiva del modelo.

En cuanto a las limitaciones, aunque los resultados de este trabajo indican que el modelo híbrido ARIMA-LSTM optimizado ha predicho con éxito los informes de ciberdelitos reportados contra la propiedad, es importante tener en cuenta una serie de limitaciones metodológicas y específicas del contexto que deben considerarse al interpretar y utilizar los resultados. Una primera limitación tiene que ver con la naturaleza y el nivel de confiabilidad de la información que se analizó. El conjunto de datos fue construido a partir de registros administrativos proporcionados por la PNP, correspondientes al intervalo que abarca desde enero de 2018 hasta diciembre de 2023. Si bien se cuenta con un volumen considerable de 60,378 denuncias, estos datos pueden estar sujetos a sesgos derivados del subregistro, demoras en la denuncia, errores de codificación o variaciones en los procedimientos de registro policial a lo largo del tiempo que podrían haber afectado momentáneamente la representatividad de la serie. En segundo lugar, las restricciones propias del modelo propuesto ARIMA-LSTM optimizado que, si bien demostró un alto rendimiento predictivo para el caso específico de los ciberdelitos contra el patrimonio, la capacidad de generalizar los tipos de ciberdelitos, categorías delictivas o contextos institucionales aún no ha sido evaluada. Asimismo, el enfoque actual se limita a una serie temporal univariada, sin considerar posibles factores exógenos como indicadores tecnológicos, económicos, cambios legislativos o campañas de prevención, cuya inclusión podría enriquecer los modelos futuros.

Otro aspecto relevante es la dependencia del modelo ARIMA de supuestos estructurales relacionados con la estacionariedad de la serie temporal. La serie original debe diferenciarse para satisfacer este requisito, lo que lleva a una transformación que puede ser estadísticamente necesaria pero que podría afectar

la interpretación de los resultados. Además, el modelo híbrido presume que la estructura de la serie no cambia a un ritmo sustancial, lo que puede fallar en una crisis o alteración repentina de los patrones criminales. Cualquier variación importante en la dinámica estructural de la serie probablemente debilitará el poder predictivo del marco si no se actualiza periódicamente con nuevos datos. La conclusión general es que, aunque el modelo híbrido ARIMA-LSTM ha demostrado ser un método poderoso y eficiente para pronosticar ciberdelitos en el contexto peruano, necesita considerar cuidadosamente las suposiciones y la calidad de los datos y la actualización metodológica para poder adaptarse a cambios significativos en el campo criminal y el panorama técnico.

Como resultado, los hallazgos iniciales basados en los conocimientos obtenidos del modelo híbrido ARIMA-LSTM optimizado nos llevan a la conclusión de que la predicción proporcionada aquí es bastante precisa con respecto al desarrollo mensual de los ataques de ciberdelito en Perú contra la propiedad. Con un MAPE del 10.73%, LSTM es capaz de manejar eventos criminales complejos. La mejora muy significativa en el error porcentual medio y la coherencia visual de la serie observada con las predicciones generadas añade credibilidad a la fiabilidad del modelo como un modelo capaz de predecir ciberdelitos.

En cuanto a la aplicabilidad, los hallazgos indican que el sistema es técnicamente posible y, además, también muy valioso para permitir decisiones organizacionales. La PNP podría utilizar las predicciones proporcionadas para predecir picos de delitos digitales y guiar la disposición de todos los recursos humanos, técnicos y logísticos antes y a intervalos regulares en la respuesta. Asimismo, el MPFN puede usar estas previsiones para formular una planificación fiscal adecuada que se centre en la capacitación de los fiscales de ciberdelitos o en la educación ciudadana sobre previsiones de alta criminalidad. La capacidad de adoptar un modelo anticipatorio puede usarse en ambos casos para apoyar y reforzar el análisis de series temporales de denuncias por ciberdelitos y ayudar a pasar de tácticas reactivas a métodos proactivos basados en evidencia.

Refiriéndose al trabajo involucrado en el desarrollo de esta tesis doctoral, la fase de muestreo y validación de los hallazgos recopilados constituye una contribución metodológica y empírica destacada. Desde una perspectiva

metodológica, el enfoque se válida para el modelo híbrido utilizando la optimización de hiperparámetros en un área compleja o desatendida como el ciberdelito que se ha utilizado en Perú y América Latina. Específicamente, también se encuentran patrones temporales, se define el marco necesario y para prepararse para la futura aplicación de arquitectura mixta ya sea en la predicción de otros tipos de delitos o en la incorporación de otras variables contextuales que enriquezcan el análisis y la previsión.

En conclusión, el modelo híbrido ARIMA-LSTM optimizado proporciona un método confiable, preciso y práctico para predecir informes de ciberdelitos durante el curso del estudio de ciberdelitos en Perú. La implementación efectiva de sus funciones proporciona un desarrollo significativo en la configuración de los discursos de la ciencia de datos y la gestión de la seguridad pública juntos, y el apoyo empíricamente validado para su efectividad respalda materialmente los objetivos principales de este estudio doctoral.

CONCLUSIONES

Se demuestra que realmente es posible desarrollar un modelo de aprendizaje automático que permita realizar predicciones sobre el progreso mensual de las denuncias de ciberdelitos relacionados con la propiedad en Perú y cumplir con el objetivo general de esta investigación.

Para el **OE1/HE1**, la serie temporal original no es estacionaria, con un $p=0.991$ en la prueba ADF, lo que respalda el uso de la diferenciación. Encontramos que la serie transformada tenía un p-valor de 0.03 (inferior al umbral de significancia del 5%), lo que indica que la serie diferenciada es estacionaria. Este hallazgo sustenta la idoneidad metodológica del modelo ARIMA ya que modela la estructura lineal en construcción, satisfaciendo así la necesidad de establecer la condición adecuada de estacionariedad para los modelos autorregresivos para facilitar la precisión del modelo ARIMA en la previsión de la serie temporal de incidentes de ciberdelitos contra la propiedad.

Para el **OE2/HE2**, se encontró que el modelo híbrido optimizado ARIMA-LSTM supera significativamente a los otros modelos individuales ARIMA y LSTM. La prueba de Diebold-Mariano indicó p-valor de 0.03 en ambas comparaciones, sustentando las diferencias entre la precisión del modelo híbrido y los modelos ARIMA y LSTM de modo individual. Por lo tanto, el rendimiento del enfoque combinado ARIMA-LSTM para la precisión de la previsión valida aún más la hipótesis y proporciona evidencia para cumplir el objetivo.

En cuanto al **OE3/HE3**, el análisis de métricas de error muestra que el modelo híbrido logra un MAPE del 10.73% en comparación con la precisión del ARIMA (15.43%) y LSTM (13.28%). Es la diferencia del 14% en comparación con los mejores modelos individuales, por lo tanto, para validar la eficacia de la combinación metodológica en un escenario complejo como el contexto de ciberdelito. Esta confirmación de la hipótesis sustenta la idoneidad del enfoque híbrido como estrategia de modelado y previsión de serie temporal caracterizada por variabilidad y dinámicas lineales y no lineales.

Desde una perspectiva aplicada, estos hallazgos tienen un significado real. Instituciones como la PNP y el MPFN podrían aprovechar este modelo para

predecir la respuesta a los informes de ciberdelitos en el futuro, lo que llevaría al desarrollo eficiente de recursos, al diseño de estrategias de prevención y a la orientación de acciones operativas específicas. El valor predictivo del modelo puede cambiar la lógica a una lógica proactiva sobre reactiva en la lucha contra el ciberdelito.

Finalmente, el desarrollo y validación de un modelo híbrido optimizado ARIMA-LSTM es un paso metodológico hacia la mejora del análisis predictivo del crimen sobre la base del contexto peruano y latinoamericano. Adicionalmente, la implementación de técnicas de optimización de hiperparámetros, la validación empírica de las hipótesis y la adaptabilidad del modelo a las particularidades del ciberdelito en Perú refuerzan la naturaleza innovadora y relevante de esta propuesta. Sin duda, estas conclusiones respaldan el logro del propósito de esta tesis y también abren el camino para futuros trabajos exploren el análisis de ciberdelitos en modelos híbridos utilizando enfoques de aprendizaje automático y redes neuronales.

RECOMENDACIONES

El análisis de los hallazgos sugiere que la validación externa del modelo híbrido mejorado ARIMA-LSTM a otras instituciones, como el MPFN, puede expandirse. Esta expansión proporcionaría un punto de contraste para el modelo en términos de capacidad de generalización entre instituciones y, además, mejoraría la utilidad del modelo en un contexto nacional de gestión del crimen.

También se propone el modelo diseñado para pronosticar denuncias de otros delitos de alta incidencia prevalentes como el hurto, robo, extorsión o violencia doméstica, y debería replicarse para los datos. Esta aplicación cruzada proporcionaría una evaluación del modelo sobre si el rendimiento predictivo impulsado por series temporales del modelo es válido, fuerte y consistente, cuando los patrones estructurales y dinámicas son diferentes de los delitos cibernéticos.

Las variables externas relevantes como, por ejemplo, indicadores de conectividad digital, campañas de concientización ciudadana, cambios regulatorios o factores socioeconómicos, deben considerarse en el proceso de modelado. Estas variables mejorarían el análisis al incluir efectos exógenos para mejorar la precisión y la profundidad explicativa del modelo.

Además, se podría construir un sistema de pronóstico automatizado, que podría incorporar el modelo híbrido en sistemas institucionales como el Sistema de Denuncias Policiales. La herramienta permitiría predicciones en tiempo real, fortaleciendo así la capacidad de la DIVINDAT y otras unidades operativas para tomar decisiones tácticas, estratégicas y de planificación según la evidencia esperada sobre delitos cibernéticos.

Se recomienda un análisis comparativo de referencia del rendimiento del modelo optimizado ARIMA-LSTM con otros enfoques híbridos recientes (por ejemplo, ARIMA-XGBoost o Prophet-LSTM). Esta comparación podría permitir detectar mejoras potenciales en la calidad del pronóstico y elegir arquitecturas eficientes basadas en la naturaleza del delito y el comportamiento criminal temporal.

Por último, el modelo podría adaptarse para su uso en el monitoreo de sistemas institucionales o en observatorios del crimen donde se implementen

sistemas de alerta temprana. La inclusión del modelo en tales sitios ayudaría a señalar variaciones significativas del volumen de informes anticipado, permitiendo la prevención oportuna y fortaleciendo la gestión del conocimiento para combatir el delito cibernético en el Perú.

REFERENCIAS BIBLIOGRÁFICAS

- Agencia Peruana de Noticias. (2024, January 11). *Modalidades de delitos informáticos más denunciadas en el Perú*. <https://andina.pe/agencia/noticia-estas-son-las-modalidades-delitos-informaticos-mas-denunciadas-el-peru-969892.aspx>
- Al-Ghamdi, N., & Alsubait, T. (2022). Digital Forensics and Machine Learning to Fraudulent Email Prediction. *2022 Fifth National Conference of Saudi Computers Colleges (NCCC)*, 99–106. <https://doi.org/10.1109/NCCC57165.2022.10067685>
- Aljarboua, E. F., Bte Md. Din, M., & Bakar, A. A. (2022a). Cyber-Crime Detection: Experimental Techniques Comparison Analysis. *2022 International Visualization, Informatics and Technology Conference (IVIT)*, 124–129. <https://doi.org/10.1109/IVIT55443.2022.10033332>
- Aljarboua, E. F., Bte Md. Din, M., & Bakar, A. A. (2022b). Cyber-Crime Detection: Experimental Techniques Comparison Analysis. *2022 International Visualization, Informatics and Technology Conference (IVIT)*, 124–129. <https://doi.org/10.1109/IVIT55443.2022.10033332>
- Almashhadani, A. O., Kaiiali, M., Sezer, S., & O’Kane, P. (2019). A Multi-Classifer Network-Based Crypto Ransomware Detection System: A Case Study of Locky Ransomware. *IEEE Access*, 7, 47053–47067. <https://doi.org/10.1109/ACCESS.2019.2907485>
- Amen, K., Zohdy, M., & Mahmoud, M. (2021). Machine Learning for Multiple Stage Phishing URL Prediction. *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 794–800. <https://doi.org/10.1109/CSCI54926.2021.00049>
- Amir Latif, R. M., Umer, M., Tariq, T., Farhan, M., Rizwan, O., & Ali, G. (2019). A Smart Methodology for Analyzing Secure E-Banking and E-Commerce Websites. *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 589–596. <https://doi.org/10.1109/IBCAST.2019.8667255>
- Arnold Cathalifaud, Marcelo; Osorio, F. (1998). Introduccion a los conceptos basicos de la teoria general de sistemas. *Cinta de Moebio*, 3.
- Arora, T., Sharma, M., & Khatri, S. K. (2019). Detection of Cyber Crime on Social Media using Random Forest Algorithm. *2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC)*, 47–51. <https://doi.org/10.1109/PEEIC47157.2019.8976474>
- Balogun, A. O., Adewole, K. S., Bajeh, A. O., & Jimoh, R. G. (2021). Cascade Generalization Based Functional Tree for Website Phishing Detection. In *Communications in Computer and Information Science: 1487 CCIS* (pp. 288–306). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-16-8059-5_17
- Barrenechea, M. M. M., & Lengua, M. A. C. (2025). A systematic literature review on the use of artificial intelligence for cybercrime rate forecasting. *Bulletin of Electrical Engineering and Informatics*, 14, 2042–2054. <https://doi.org/10.11591/eei.v14i3.9213>
- Barrett, K.-A., Fain, R., García, C., García-Belenguer, G., Lillo, C., Marchiori, B., Pelletier, E., & Subero, D. (2016). *Ciberseguridad ¿Estamos preparados en América Latina y el Caribe? Mejorando vidas*. www.observatoriociberseguridad.com
- BBVA. (2023, July 10). *Ciberseguridad, ataques informáticos*. Banco Bilbao Vizcaya Argentaria (BBVA).
- Bertalanffy, L., & Almela, J. (1976). Teoría general de los sistemas: fundamentos, desarrollo, aplicaciones. In *Teoría general de los sistemas: fundamentos, desarrollo, aplicaciones*.
- Bhardwaj, G., & Bawa, R. K. (2022). Machine Learning techniques based exploration of various types of crimes in India. *Indian Journal of Computer Science and Engineering*, 13(4). <https://doi.org/10.21817/indjcse/2022/v13i4/221304142>
- Bhardwaj, G., & R.K., Dr. B. (2022a). MACHINE LEARNING TECHNIQUES BASED EXPLORATION OF VARIOUS TYPES OF CRIMES IN INDIA. *Indian Journal of Computer Science and Engineering*, 13(4), 1293–1307. <https://doi.org/10.21817/indjcse/2022/v13i4/221304142>
- Bhardwaj, G., & R.K., Dr. B. (2022b). MACHINE LEARNING TECHNIQUES BASED EXPLORATION OF VARIOUS TYPES OF CRIMES IN INDIA. *Indian Journal of Computer Science and Engineering*, 13(4), 1293–1307. <https://doi.org/10.21817/indjcse/2022/v13i4/221304142>
- Bilen, A., & Özer, A. B. (2021). Cyber-attack method and perpetrator prediction using machine learning algorithms. *PeerJ Computer Science*, 7, 1–21. <https://doi.org/10.7717/peerj-cs.475>
- bin Asad, A., Mansur, R., Zawad, S., Evan, N., & Hossain, M. I. (2020). Analysis of Malware

- Prediction Based on Infection Rate Using Machine Learning Techniques. *2020 IEEE Region 10 Symposium (TENSYP)*, 706–709. <https://doi.org/10.1109/TENSYP50017.2020.9230624>
- Boonyopakorn, P., Wisitpongphan, N., & Changsan, U. (2023). Classifying Cybercrime and Threat on Thai Online News: A Comparison of Supervised Learning Algorithms. *2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, 1–6. <https://doi.org/10.1109/ITC-CSCC58803.2023.10212562>
- Borzemski, L., & Wojtkiewicz, M. (2011). *Evaluation of Chaotic Internet Traffic Predictor Using MAPE Accuracy Measure* (Vol. 1, pp. 173–182). https://doi.org/10.1007/978-3-642-21771-5_19
- Butt, U. M., Letchmunan, S., Hassan, F. H., Ali, M., Baqir, A., & Sherazi, H. H. R. (2020). Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review. *IEEE Access*, 8, 166553–166574. <https://doi.org/10.1109/ACCESS.2020.3022808>
- Cai, L., Peng, X., & Wang, L. (2018). The characteristics and influencing factors of entrepreneurial behaviour: The case of new state-owned firms in the new energy automobile industry in an emerging economy. *Technological Forecasting and Social Change*, 135. <https://doi.org/10.1016/j.techfore.2018.04.014>
- Ch, R., Gadekallu, T. R., Abidi, M. H., & Al-Ahmari, A. (2020). Computational System to Classify Cyber Crime Offenses using Machine Learning. *Sustainability*, 12(10), 4087. <https://doi.org/10.3390/su12104087>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chipa Sierra, R. (2023). *Modelo de regresión para el pronóstico de la cantidad de denuncias por delitos que se registran en las Comisarias de la Policía Nacional de Perú en Lima Metropolitana*. Pontificia Universidad Católica del Perú.
- Cumpston, M. S., McKenzie, J. E., Ryan, R., Flemyng, E., Thomas, J., & Brennan, S. E. (2023). Development of the InSynQ checklist: A tool for planning and reporting the synthesis questions in systematic reviews of interventions. *Cochrane Evidence Synthesis and Methods*, 1(10), 1–13. <https://doi.org/10.1002/cesm.12036>
- Cunha, A. F., Ferreira, D., Neto, C., Abelha, A., & Machado, J. (2021). *A CRISP-DM Approach for Predicting Liver Failure Cases: An Indian Case Study* (Vol. 1, pp. 156–164). https://doi.org/10.1007/978-3-030-80624-8_20
- Deepak, N. S., Hanitha, T., Tanniru, K., Kiran, L. R., Sai, N. R., & Kumar, M. J. (2023a). Analyze and Forecast the Cyber Attack Detection Process using Machine Learning Techniques. *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 1732–1738. <https://doi.org/10.1109/ICESC57686.2023.10193289>
- Deepak, N. S., Hanitha, T., Tanniru, K., Kiran, L. R., Sai, N. R., & Kumar, M. J. (2023b). Analyze and Forecast the Cyber Attack Detection Process using Machine Learning Techniques. *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 1732–1738. <https://doi.org/10.1109/ICESC57686.2023.10193289>
- Defensoría del Pueblo. (2023). *LA CIBERDELINCUENCIA EN EL PERÚ: ESTRATEGIAS Y RETOS DEL ESTADO*. <https://www.defensoria.gob.pe/wp-content/uploads/2023/05/INFORME-DEF-001-2023-DP-ADHPD-Ciberdelincuencia.pdf>
- Deichmann, U. (2024). Contrasting philosophical and scientific views in the long history of studying the generation of form in development. *BioSystems*, 242, 105260. <https://doi.org/https://doi.org/10.1016/j.biosystems.2024.105260>
- Deng, Y., Fan, H., & Wu, S. (2023). A hybrid ARIMA-LSTM model optimized by BP in the forecast of outpatient visits. *Journal of Ambient Intelligence and Humanized Computing*, 14(5), 5517–5527. <https://doi.org/10.1007/s12652-020-02602-x>
- Deutscher, S. (2017, January 25). *Perú tiene un gasto en ciberseguridad que está muy por debajo del promedio mundial*. Diario Gestión.
- Devadarshini, P. T., Chandrashekar, B., Pundir, S., Tiwari, M., Madala, R., & Indhuma, E. (2023). Cognitive Defense Cyber Attack Prediction and Security Design in Machine Learning Model. *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, 1361–1366. <https://doi.org/10.1109/IC3I59117.2023.10397602>
- Diario Gestión. (2024, March 22). *Delitos informáticos: los ciberdelitos con más denuncias y sus penas*. <https://www.gob.pe/25710-cuales-son-los-casos-de-ciberdelincuencia-mas-comunes-en-el-peru>
- Diario Inforbae. (2020, October 22). *Ciberdelito: desde el inicio de la cuarentena, los ataques informáticos crecieron cerca del 70 por ciento*. <https://www.inforbae.com/sociedad/2020/10/22/ciberdelito-desde-el-inicio-de-la-cuarentena-los->

- ataques-informaticos-crecieron-cerca-del-70-por-ciento/?utm_source=chatgpt.com
- Diario Oficial El Peruano. (2023, June 22). *¡Cuidado con los fraudes informáticos! Estas son las modalidades más denunciadas en Perú*. Diario Oficial El Peruano.
- Dipakkumar Pandya, D., Amarawat, G., Jadeja, A., Degadwala, S., & Vyas, D. (2022). Analysis and Prediction of Location based Criminal Behaviors Through Machine Learning. *2022 International Conference on Edge Computing and Applications (ICECAA)*, 1324–1332. <https://doi.org/10.1109/ICECAA55415.2022.9936498>
- Duc M Cao, Md Abu Sayed, Md Abu Sayed, Md Tuhin Mia, Eftekhar Hossain Ayon, Bishnu Padh Ghosh, Rejon Kumar Ray, Aqib Raihan, Aslima Akter, & Mamunur Rahman. (2024). Advanced Cybercrime Detection: A Comprehensive Study on Supervised and Unsupervised Machine Learning Approaches Using Real-world Datasets. *Journal of Computer Science and Technology Studies*, 6(1), 40–48. <https://doi.org/10.32996/jcsts.2024.6.1.5>
- Elmezain, M., Malki, A., Gad, I., & Atlam, E.-S. (2022). Hybrid deep learning model-based prediction of images related to cyberbullying. *International Journal of Applied Mathematics and Computer Science*, 32(2), 323–334. <https://doi.org/10.34768/amcs-2022-0024>
- Falade, A., Azeta, A., Oni, A., & Odun-ayo, I. (2019). Systematic Literature Review of Crime Prediction and Data Mining. *Review of Computer Engineering Studies*, 6(3), 56–63. <https://doi.org/10.18280/rces.060302>
- Fang, X., Xu, M., Xu, S., & Zhao, P. (2019). A deep learning framework for predicting cyber attacks rates. *EURASIP Journal on Information Security*, 2019(1), 5. <https://doi.org/10.1186/s13635-019-0090-6>
- Ghofran Mohamed, A., Amal Elsayed, & Aboutabl Galal, A. (2023). MACHINE LEARNING FOR DETECTING CYBERCRIME IN THE BANKING SECTOR. *Journal of Southwest Jiaotong University*, 58(5), 786–799. <https://doi.org/10.35741/issn.0258-2724.58.5.60>
- Gil Jauregui, J. O., Carmen Cruzatti, A. G., Cano Lengua, M. A., & Villaverde Medrano, H. (2024). Proposed Feature Selection Technique for Pattern Detection in Patients with Pneumonia Records. *International Journal of Online and Biomedical Engineering (IJOE)*, 20(07), 69–89. <https://doi.org/10.3991/ijoe.v20i07.47647>
- Gobierno del Perú. (2024, January 14). *¿Cuáles son los casos de ciberdelincuencia más comunes en el Perú?* Plataforma Digital Del Estado Peruano. <https://www.gob.pe/25710-cuales-son-los-casos-de-ciberdelincuencia-mas-comunes-en-el-peru>
- Goel, S. (2020). National Cyber Security Strategy and the Emergence of Strong Digital Borders. *Connections: The Quarterly Journal*, 19(1), 73–86. <https://doi.org/10.11610/Connections.19.1.07>
- Guan, M., Xue, R., Wu, Z., Yang, H., Song, D., & Zhang, Z. (2022). A high performance fraud detection strategy prediction model. *2022 2nd International Conference on Computer Science and Blockchain (CCSB)*, 107–110. <https://doi.org/10.1109/CCSB58128.2022.00026>
- Harie, Y., Gautam, B., & Wasaki, K. (2023). Computer Vision Techniques for Growth Prediction: A Prisma-Based Systematic Literature Review. *Applied Sciences*, 13, 5335. <https://doi.org/10.3390/app13095335>
- He, C. (2023). A Hybrid Model Based on Multi-LSTM and ARIMA for Time Series Forecasting. *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 612–616. <https://doi.org/10.1109/ICSP58490.2023.10248909>
- Hernández, J., López, D., & Vera, N. (2018). Primary user characterization for cognitive radio wireless networks using long short-term memory. *International Journal of Distributed Sensor Networks*, 14(11), 1–20. <https://doi.org/10.1177/1550147718811828>
- Hernández, R., Fernández, C., & Baptista, M. del P. (2014). Metodología de la Investigación Hernández Sampieri 6a Edición. In *Septiembre 2015*.
- Hernandez Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. del P. (2013). Metodología de la Investigación Hernández Sampieri 6a Edición. In *Journal of Chemical Information and Modeling* (Vol. 53, Number 9).
- Hurwitz, Judith; Kirsch, D. (2018). *¿Qué es Machine Learning? - Colombia | IBM*. In *IBM Limited Edition*.
- Hussain, S. K. S., Reddy, T. U., Kiran, D. S., & Samreen, S. K. (2021). Application of Machine Learning in Crime Prediction. *Proceedings of the 5th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2021*. <https://doi.org/10.1109/ICECA52323.2021.9675906>
- INCIBE. (2023). *Ingeniería social y fraudes online*. Instituto Nacional de Ciberseguridad.
- INEI. (2017). *Encuesta nacional especializada sobre victimización, cifra negra del delito y denuncia*. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1540/cap05.pdf

- Jadhav, P. A., Lalwani, U., Gour, A., Shayan, M., & Motwani, S. (2022). Identification of Fraudulent Credit Card transactions using Machine Learning Algorithms. *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, 1–4.
<https://doi.org/10.1109/I2CT54291.2022.9824165>
- Jenga, K., Catal, C., & Kar, G. (2023). Machine learning in crime prediction. *Journal of Ambient Intelligence and Humanized Computing*, 14(3), 2887–2913. <https://doi.org/10.1007/s12652-023-04530-y>
- Jin, Y., Wang, R., Zhuang, X., Wang, K., Wang, H., Wang, C., & Wang, X. (2022). Prediction of COVID-19 Data Using an ARIMA-LSTM Hybrid Forecast Model. *Mathematics*, 10(21).
<https://doi.org/10.3390/math10214001>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9). <https://doi.org/10.1038/s42256-019-0088-2>
- Kasemset, C., Phuruan, K., & Opassuwan, T. (2023). Shallot Price Forecasting Models: Comparison among Various Techniques. *Production Engineering Archives*, 29(4), 348–355.
<https://doi.org/10.30657/pea.2023.29.40>
- Kitchenham, B. (2007). Guidelines for performing systematic literature reviews in software engineering. *Technical Report, Ver. 2.3 EBSE Technical Report. EBSE*.
- Kotian, P., & Sonkusare, R. (2021). Detection of Malware in Cloud Environment using Deep Neural Network. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–5.
<https://doi.org/10.1109/I2CT51068.2021.9417901>
- Kumar, A. (2023). Systematic Literature Review (SLR). In *Meta-analysis in Clinical Research: Principles and Procedures* (Vol. 1, pp. 7–14). Springer Nature Singapore.
https://doi.org/10.1007/978-981-99-2370-0_2
- Ley N° 30096. (2013). *LEY DE DELITOS INFORMÁTICOS-LEY N° 30096*.
<https://www.gob.pe/institucion/mpfn/informes-publicaciones/1678028-ley-n-30096>
- Li, Y., Xu, Q., Wang, Y., & Liu, B. (2024). Genetic Algorithms Application for Pricing Optimization in Commodity Markets. *Mathematics*, 12(9), 1289. <https://doi.org/10.3390/math12091289>
- Lilhore, U. K., Simaiya, S., Sandhu, J. K., Baliyan, A., & Garg, A. (2023). EHML: An Efficient Hybrid Machine Learning Model for Cyber Threat Forecasting in CPS. *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, 1453–1458.
<https://doi.org/10.1109/AISC56616.2023.10084987>
- Liu, X., Xie, X., & Guo, Q. (2022). Research on Cloud Computing load forecasting based on LSTM-ARIMA combined model. *2022 Tenth International Conference on Advanced Cloud and Big Data (CBD)*, 19–23. <https://doi.org/10.1109/CBD58033.2022.00013>
- Maataoui, S., Bencheikh, G., & Bencheikh, G. (2023). Predictive Maintenance in the Industrial Sector: A CRISP-DM Approach for Developing Accurate Machine Failure Prediction Models. *2023 Fifth International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, 223–227. <https://doi.org/10.1109/ACTEA58025.2023.10193983>
- Matias, S. M. M., Costales, J. A., & De Los Santos, C. M. (2022). A Framework for Cybercrime Prediction on Twitter Tweets Using Text-Based Machine Learning Algorithm. *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 235–240.
<https://doi.org/10.1109/PRAI55851.2022.9904212>
- Ministerio de Justicia y Derechos Humanos. (2022). *CIBERDELINCUENCIA REPORTE DE INFORMACIÓN ESTADÍSTICA Y RECOMENDACIONES PARA LA PREVENCIÓN*.
<https://cdn.www.gob.pe/uploads/document/file/3562747/Reporte%20de%20Ciberdelincuencia.pdf?v=1661790352>
- Morales-Barrenechea, M. M., Rodriguez, C., Cancho-Rodriguez, E. D., & Navarro, R. R. H. (2025). Autoregressive integrated moving average-long short-term memory optimized hybrid model for cybercrime forecasting. *Bulletin of Electrical Engineering and Informatics*, 14(5), 3883–3894.
<https://doi.org/10.11591/eei.v14i5.9769>
- Morgan, S. (2020, November 13). *Cybercrime To Cost The World \$10.5 Trillion Annually By 2025*. Cybercrime Magazine. <https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/>
- Muncaster, P. (2022, January 11). *5 formas en que los cibercriminales roban contraseñas*. Welivesecurity.
- Naaz, S. (2021). Detection of Phishing in Internet of Things Using Machine Learning Approach. *International Journal of Digital Crime and Forensics*, 13(2), 1–15.
<https://doi.org/10.4018/IJDCF.2021030101>
- Natarajan, C., Priya, S. D. K. (2024). CRIME RATE PREDICTION USING COMBINED ARIMA AND LSTM. *International Research Journal of Modernization in Engineering Technology and*

- Science*, 06(01), 3694–3703. <https://doi.org/10.56726/IRJMETS48515>
- Ñaupas, H., Mejía, E., Novoa, E., & Villagómez, A. (2014). Metodología de la investigación Naupas Humberto. *Ediciones De La U*.
- OECD. (2019, November 13). *Recommendation of the council on artificial intelligence*. Artificial Intelligence in Society. <https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/>
- Omar, A. R., Taie, S. A., & E. Shaheen, M. (2023). From Phishing Behavior Analysis and Feature Selection to Enhance Prediction Rate in Phishing Detection. *International Journal of Advanced Computer Science and Applications*. <https://api.semanticscholar.org/CorpusID:259142036>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Alonso-Fernández, S. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología*, 74(9), 790–799. <https://doi.org/https://doi.org/10.1016/j.recesp.2021.06.016>
- Pahuja, L., & Kamal, A. (2023). EnLEFD-DM: Ensemble Learning based Ethereum Fraud Detection using CRISP-DM framework. *Expert Systems*, 40(9), 1–18. <https://doi.org/10.1111/exsy.13379>
- Parisi, A. (2019). *Hands-on artificial intelligence for cybersecurity : implement smart AI systems for preventing cyber attacks and detecting threats and network anomalies* (1st ed., Vol. 1). Packt Publishing.
- Pérez-Escobar, J. A., & Sarikaya, D. (2024). Philosophical Investigations into AI Alignment: A Wittgensteinian Framework. *Philosophy and Technology*, 37(3). <https://doi.org/10.1007/s13347-024-00761-9>
- Philomina, J., Fahim Fathima, K. A., Gayathri, S., Elias, G. E., & Menon, A. A. (2022). A comparative study of machine learning models for the detection of Phishing Websites. *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–7. <https://doi.org/10.1109/IC3SIS54991.2022.9885595>
- Plataforma Digital Única del Estado Peruano. (2025, March). *Policía Nacional del Perú*. <https://www.gob.pe/institucion/pnp/institucional>
- Policía Nacional del Perú. (2018). *División de Investigación de Delitos de Alta Tecnología*. https://desaparecidosenperu.policia.gob.pe/desaparecidos/division_investigacion_alta_tecnologia
- Policía Nacional del Perú. (2025). *Sistemas PNP*. <https://www.sistemasnpn.com/info/que-es-sidpol.php>
- Reid, R., & Van Niekerk, J. (2014). From information security to cyber security cultures. *2014 Information Security for South Africa*, 1–7. <https://doi.org/10.1109/ISSA.2014.6950492>
- Ritchey, F. (2014). Estadística para las ciencias sociales. In *Estadística para las ciencias sociales*. <https://doi.org/10.15332/dt.inv.2021.02408>
- Rumi, S. K., Deng, K., & Salim, F. D. (2018). Crime event prediction with dynamic features. *EPJ Data Science*, 7(1), 43. <https://doi.org/10.1140/epjds/s13688-018-0171-7>
- Safat, W., Asghar, S., & Gillani, S. A. (2021). Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *IEEE Access*, 9, 70080–70094. <https://doi.org/10.1109/ACCESS.2021.3078117>
- Sakshi, K., & A, V. (2020). An ARIMA- LSTM Hybrid Model for Stock Market Prediction Using Live Data. *Journal of Engineering Science and Technology Review*, 13(4), 117–123. <https://doi.org/10.25103/jestr.134.11>
- Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1). <https://doi.org/10.5465/AMR.2000.2791611>
- Sharko, A. D., Sharko, G., & Qose, S. (2024). Artificial Intelligence In Cybersecurity Applications. *2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES)*, 175–180. <https://doi.org/10.1109/INES63318.2024.10629129>
- Sharma, A., & Babbar, H. (2023). Machine Learning Solutions for Evolving Injection Attack Landscape. *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, 1–6. <https://doi.org/10.1109/INCOFT60753.2023.10425456>
- Sharma, D., & Singh Kang, S. (2022). Hybrid model for detection of frauds in credit cards. *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 70–77. <https://doi.org/10.1109/ICAC3N56670.2022.10074057>
- Sharmin, S., Alam, F. I., Das, A., & Uddin, R. (2022). An Investigation into Crime Forecast Using Auto ARIMA and Stacked LSTM. *2022 International Conference on Innovations in Science, Engineering and Technology (ICISSET)*, 415–420.

- <https://doi.org/10.1109/ICISSET54810.2022.9775862>
- Shui, B., Cai, Z., & Luo, X. (2024). Towards customized mitigation strategy in the transportation sector: An integrated analysis framework combining LMDI and hierarchical clustering method. *Sustainable Cities and Society*, 107, 105340. <https://doi.org/10.1016/j.scs.2024.105340>
- Singh, C., Singh, R., Shivaputra, Tiwari, M., & Hazela, B. (2024). Analyse and Predict the Detection of the Cyber - Attack Process by Using a Machine-Learning Approach. *EAI Endorsed Transactions on Internet of Things*, 10, 1–6. <https://doi.org/10.4108/eetiot.5345>
- Sinu, E. B., Kleden, M. A., & Atti, A. (2024). APPLICATION OF ARIMA MODEL FOR FORECASTING NATIONAL ECONOMIC GROWTH: A FOCUS ON GROSS DOMESTIC PRODUCT DATA. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 18(2), 1261–1272. <https://doi.org/10.30598/barekengvol18iss2pp1261-1272>
- Soy Temür, A., & Yıldız, Ş. (2021). Comparison of Forecasting Performance of ARIMA LSTM and HYBRID Models for The Sales Volume Budget of a Manufacturing Enterprise. *Istanbul Business Research*, 50(1), 15–46. <https://doi.org/10.26650/ibr.2021.51.0117>
- Sueldo Alarcón, B. H. P. C. E. D. (2022). *Servicio web para la predicción de los delitos de robo utilizando machine learning en la ciudad de Lima Metropolitana*. Universidad Peruana de Ciencias Aplicadas.
- Swaminathan, A., Ramakrishnan, B., M, K., & R, S. (2022). Prediction of Cyber-attacks and Criminality Using Machine Learning Algorithms. *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 547–552. <https://doi.org/10.1109/3ICT56508.2022.9990652>
- ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime Prediction & Monitoring Framework Based on Spatial Analysis. *Procedia Computer Science*, 132, 696–705. <https://doi.org/https://doi.org/10.1016/j.procs.2018.05.075>
- UNODC. (2022). *COMPENDIO DE CIBERDELINCUENCIA ORGANIZADA*. https://www.unodc.org/documents/Cybercrime/tools-and-resources/compendio_de_delincuencia_organizada_es.pdf
- van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136, 106589. <https://doi.org/10.1016/j.infsof.2021.106589>
- Varsha, N., Kumar, D. P., Akhil, B., & Mouli, K. (2023). Phishing sites detection using Machine Learning. *2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, 1–7. <https://doi.org/10.1109/ICIICS59993.2023.10421336>
- Veena, K., Meena, K., Kuppusamy, R., Teekaraman, Y., Angadi, R. V., & Thelkar, A. R. (2022a). Cybercrime: Identification and Prediction Using Machine Learning Techniques. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/8237421>
- Veena, K., Meena, K., Kuppusamy, R., Teekaraman, Y., Angadi, R. V., & Thelkar, A. R. (2022b). Cybercrime: Identification and Prediction Using Machine Learning Techniques. *Computational Intelligence and Neuroscience*, 2022, 1–10. <https://doi.org/10.1155/2022/8237421>
- Veena, K., Meena, K., Teekaraman, Y., Kuppusamy, R., & Radhakrishnan, A. (2022). SVM Classification and KNN Techniques for Cyber Crime Detection. *Wireless Communications and Mobile Computing*, 2022, 1–9. <https://doi.org/10.1155/2022/3640017>
- Verma, R., & Thakur, B. (2023). Machine Learning Techniques for the Prediction of Cyber-Attacks. *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 978–985. <https://doi.org/10.1109/ICCCIS60361.2023.10425542>
- Vijayalakshmi, M., & Norbu, R. (2023a). Smart Police: A Hybrid Deep Learning Model for Crime Proactivity Assessment. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–7. <https://doi.org/10.1109/ICCCNT56998.2023.10308231>
- Vijayalakshmi, M., & Norbu, R. (2023b). Smart Police: A Hybrid Deep Learning Model for Crime Proactivity Assessment. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–7. <https://doi.org/10.1109/ICCCNT56998.2023.10308231>
- Villar, J. R. N., & Lengua, M. A. C. (2024). A Systematic Review of the Literature on the Use of Artificial Intelligence in Forecasting the Demand for Products and Services in Various Sectors. *International Journal of Advanced Computer Science and Applications*, 15(3), 144–156. <https://doi.org/10.14569/IJACSA.2024.0150315>
- Wall, David. (2024). *Cybercrime : the transformation of crime in the information age*. Polity Press.
- Wang, H., & Zhang, B. (2023). Research on ARIMA Model for Short-Term Traffic Flow Prediction based on Time Series. *2023 8th International Conference on Intelligent Informatics and*

- Biomedical Sciences (ICIIBMS)*, 92–95. <https://doi.org/10.1109/ICIIBMS60103.2023.10347816>
- Wang, W., Ma, B., Guo, X., Chen, Y., & Xu, Y. (2024). A Hybrid ARIMA-LSTM Model for Short-Term Vehicle Speed Prediction. *Energies*, 17(15), 3736. <https://doi.org/10.3390/en17153736>
- Wang, X., Yang, J., Liu, Y., Wang, Y., Wang, F. Y., Kang, M., Tian, Y., Rudas, I., Li, L., Fanti, M. P., Alrifae, B., Deveci, M., Mishra, D., Khan, M. K., Chen, L., & Reffye, P. De. (2024). Parallel intelligence in three decades: a historical review and future perspective on ACP and cyber-physical-social systems. *Artificial Intelligence Review*, 57(9). <https://doi.org/10.1007/s10462-024-10861-9>
- Xu, D., Zhang, Q., Ding, Y., & Zhang, D. (2022). Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environmental Science and Pollution Research*, 29(3), 4128–4144. <https://doi.org/10.1007/s11356-021-15325-z>
- Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., & Chen, J. (2022). Interpret N° machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94. <https://doi.org/10.1016/j.compenvurbsys.2022.101789>
- Zou, Y. (2025). Mathematical Modeling and Algorithm Application The Prediction of Influenza Using the Hybrid ARIMA-LSTM Model. *Mathematical Modeling and Algorithm Application*, 4(2), 20–25. <https://doi.org/http://dx.doi.org/10.54097/vgfxm178>

ANEXOS

Anexo 1. Entrega base denuncias de la Policía Nacional del Perú

De: DIVIEST-DIRPEP-EMG-PNP <estadisticaemg@yahoo.com>

Fecha: 20 de marzo de 2024, 12:03:56 p. m. PET

Para: martin.moralesb@unmsm.edu.pe, martin.moralesb@outlook.com

Asunto: PEDIDO DE INFORMACION ESTADISTICA_DELITOS INFORMATICOS

Responder-Para: DIVIEST-DIRPEP-EMG-PNP <estadisticaemg@yahoo.com>

Sr.

Manuel Martin Morales Barrenechea

DNI. N° 09774836

REF. HT-20231743088-A-SIGE. 19MAR2024.

Es grato dirigirme a Usted a mérito del documento de la referencia, con la finalidad de remitirle información estadística relacionado con la cantidad de denuncias y personas detenidos contra La Ley 30096 delitos informáticos modificada por la Ley 30171, correspondiente a los años 2018 al 2023.

Fuente: SIDPOL-PNP (Administrada y desarrollado por la DIVINF/DIRTIC PNP).

Información preliminar

Atte.

-----FIRMADO-----

DIVISIÓN DE ESTADÍSTICA
DIRTIC-SECEJE-PNP

"Este envío de correo tiene validez legal y administrativa en toda su notificación, comunicación y extensión del tema que lo genera".

LEY No. 27419: LEY SOBRE NOTIFICACIÓN POR CORREO ELECTRÓNICO (publicada el 7 de Febrero del 2001)

AVISO DE CONFIDENCIALIDAD

La información contenida en este e-mail es reservada, privilegiada y se encuentra dirigida exclusivamente a su destinatario. Su revisión, difusión, distribución o copiado se encuentra prohibido. Si ha recibido este e-mail por error por favor bórrelo y envíe un mensaje al remitente.

Anexo 2. Entrega base denuncias del Ministerio Publico Fiscalía de la Nación



MINISTERIO PÚBLICO
REPÚBLICA DEL PERÚ

Decenio de la Igualdad de oportunidades para mujeres y hombres
Año del Bicentenario, de la consolidación de nuestra Independencia, y de la
conmemoración de las heroicas batallas de Junín y Ayacucho
OFICINA DE RACIONALIZACIÓN Y ESTADÍSTICA



Firma
Digital

Firmado digitalmente por SANDOVAL
CAMPOS Rosauro FAU 20131370301
v0.1
Gerente De La Oficina De
Racionalización Y Estadística
Motivo: Soy el autor del documento
Fecha: 14.03.2024 18:50:50 -05:00

Lima, 14 de Marzo del 2024

OFICIO N° 000112-2024-MP-FN-GG-OGPLAP-ORACE

Doctora
LOURDES BERNARDITA TELLEZ PEREZ DE VARGAS
Presidenta de la Junta de Fiscales Superiores
Distrito Fiscal de Lima Centro

Presente. -

Asunto : Remite registro de Delitos Informáticos (Ley N° 30096) denunciados en el Ministerio Público, a nivel nacional, 2018 - 2023.

Referencia : PROVEIDO N° 000149-2024-MP-FN-GG-OGPLAP-ORACE (11MAR2024)
OFICIO N° 000567-2023-MP-FN-GG-OCPLAP-ORACE (23OCT2023)
OFICIO 000973-2023-UFEC (25OCT2023)
MUPDFL20230008828

Expediente : MUPGG120240001601

Tengo el agrado de dirigirme a usted para expresarle mi cordial saludo y, en atención a los documentos de la referencia, mediante el cual al amparo de la Ley N° 27806 - Ley de Transparencia y Acceso a la Información Pública, el ciudadano Manuel Martín Morales Barrenechea solicita información de los registros de Delitos Informáticos - Ley N° 30096, denunciados en Fiscalías del Ministerio Público, a nivel nacional del 2018 al 2023, según delito genérico y específico, detallado de acuerdo a diferentes campos solicitados.

Sobre el particular, se remite a su Despacho información disponible, de registros de Delitos Informáticos denunciados en fiscalías del Ministerio Público a nivel nacional por año, distrito fiscal, fiscalía, fecha de la denuncia, delito genérico, específico, estado de la denuncia, delito genérico, específico, artículo, descripción del artículo, departamento, provincia y distrito geográfico, correspondiente al periodo 2018 – 2023, para que a través de su intermedio y siguiendo con las rutas de comunicación establecidas, se dé respuesta a la solicitud del ciudadano en mención.

Dicha información ha sido elaborada de acuerdo a los reportes extraídos del Sistema de Informático de Apoyo al Trabajo Fiscal – SIATF, Sistema de Gestión Fiscal – SGF y Sistema de Bandeja Fiscal.

Sin otro particular, hago propicia la ocasión para expresarle mi mayor consideración.

Atentamente,

ROSAURA SANDOVAL CAMPOS
OFICINA DE RACIONALIZACIÓN Y ESTADÍSTICA