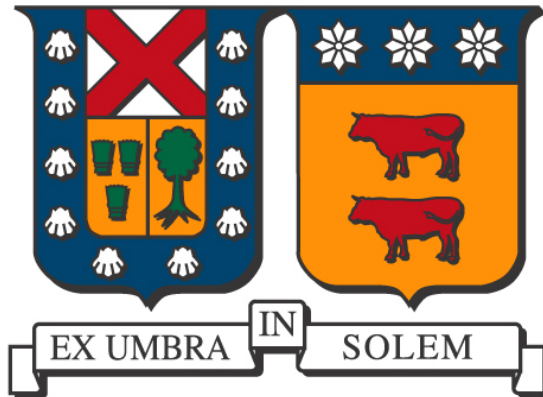


**UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA**  
**DEPARTAMENTO DE ELECTRÓNICA**  
**SANTIAGO - CHILE**



**“Procesamiento de Datos y Modelado Predictivo para la Seguridad Ciudadana”**

**FERNANDO ELÍAS AMTHAUER AROS**

**MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL  
TELEMÁTICO**

**PROFESOR GUÍA:**

**NICOLÁS TORRES RUDLOFF**

**PROFESOR CORREFERENTE:**

**PATRICIO OLIVARES**

**DICIEMBRE - 2025**



## CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

### 1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

**Tipo de monografía (marcar una opción):** ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

**Título del trabajo:** Procesamiento de Datos y Modelado Predictivo para la Seguridad Ciudadana

**Nombre del candidato(a):** Fernando Amthauer Aros

**Carrera / Grado:** Grado de Licenciado en Ciencias de la Ingeniería, título profesional de Ingeniero Civil Telemático

**Campus:** San Joaquín

**Departamento:** Departamento de Electrónica

### 2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Nicolás Torres Rudloff, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

### 3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

**Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):**

---

---

---

### 4.- FIRMAS

**Profesor(a) guía o director(a) de memoria o tesis:**

**Fecha:** 09/12/2025

**Firma:** 

**Estudiante o Candidato(a):**

**Fecha:** 09-12-2025

**Firma:** 

*Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.*

## Agradecimientos

El tan ansiado fin de esta etapa me toma por sorpresa; hace tres años me debatía ferozmente entre la idea de retirarme de la universidad para dedicarme por completo al trabajo o continuar para llegar hasta este punto. Hoy, escribiendo las últimas líneas del último escrito que haré durante mi carrera, quisiera agradecer a personas sin las cuales este trabajo no hubiese sido posible.

A Vania García, por su amor, comprensión y apoyo durante casi cuatro años de este proceso. Por cada conversación, viaje, comida, y reflexión que compartimos. Por tu alma, que brilla con luz propia en un mundo donde todo tiende a la uniformidad.

A Esteban Naranjo, por las incontables horas entre estudio, trabajo y amistad. Por cada hora dedicada a Factorio, Stellaris, código, conversaciones y a este proyecto mismo. Gracias por tu mirada objetiva, por tu voz sin miedo a opinar lo que cree correcto.

A Nicolás Ramírez, por tu amistad sincera y por recordarme cada prueba, trabajo o proyecto a entregar. Sin ti no podría haber llegado ni a segundo año.

Al profesor Nicolás Torres, quien dictó las pocas asignaturas que me mantenían interesado en la universidad y apoyó este proyecto desde el primer momento. Gracias por ser mi profesor guía y por responder con celeridad cada vez que pedimos su ayuda.

A mi familia, que aún sin entender por completo lo que conlleva este proceso buscaron alentarme a terminar.

Y para todo aquel que haya aportado de alguna forma durante mi formación escolar y universitaria.

Sepan que, mientras viajamos hacia un mundo en el que casi todo texto es producido por redes neuronales que predicen los tokens más probables, hoy una red neuronal “orgánica”, limitada y sintiente escribe estas palabras para ustedes.

Para todos los autodidactas que no puedan dormir  
concibiendo un proyecto.

## Resumen

La seguridad del entorno es un factor determinante en el mercado inmobiliario, influyendo directamente en la calidad de vida y la inversión. Sin embargo, evaluar dicho aspecto es difícil dado que la información disponible suele estar fragmentada y ser meramente descriptiva. El presente proyecto soluciona el problema mediante una plataforma unificada, y este informe detalla específicamente el desarrollo de su núcleo de procesamiento y análisis predictivo.

El objetivo principal fue construir un sistema capaz de transformar datos crudos en indicadores útiles. Para ello, se diseñó un flujo de procesamiento que estructura registros heterogéneos mediante ingeniería de atributos y agregación temporal. Además, se implementó un Índice de Crimen Objetivo, una métrica que pondera los delitos según su pena legal y ajusta los valores por población, permitiendo comparar objetivamente la seguridad entre distintos sectores.

La metodología siguió el estándar CRISP-DM, utilizando modelos de *Deep Learning* con la arquitectura Time Series Transformer. Se aplicó un entrenamiento global con salida probabilística (distribución t de Student), lo que permite al modelo aprender patrones generales y estimar la incertidumbre de las predicciones, superando a los métodos estadísticos tradicionales.

Los resultados validan la propuesta, obteniendo un modelo predictivo con un error sMAPE del 26.7 % y un MASE de 0.66. Es por ello que este trabajo aporta la inteligencia del sistema, permitiendo pasar de la simple visualización de datos históricos a la predicción de tendencias de seguridad en los barrios.

**Palabras Clave:** Deep Learning, Series temporales, Time Series Transformer, Pronóstico Probabilístico, Índice de Crimen, Ingeniería de Atributos, CRISP-DM.

# Abstract

Neighborhood safety is a determining factor in the real estate market, directly influencing both quality of life and investment. However, assessing this aspect is challenging given that available information is often fragmented and merely descriptive. The present project addresses this problem through a unified platform, and this report specifically details the development of its processing and predictive analytics core.

The main objective was to build a system capable of transforming raw data into useful indicators. To this end, a processing pipeline was designed to structure heterogeneous records using feature engineering and temporal aggregation. Furthermore, an Objective Crime Index was implemented—a metric that weights crimes according to their legal penalty and adjusts values by population, allowing for an objective comparison of safety across different sectors.

The methodology followed the CRISP-DM standard, using Deep Learning models based on the Time Series Transformer architecture. A global training approach with probabilistic output (Student-t distribution) was applied, enabling the model to learn general patterns and estimate prediction uncertainty, outperforming traditional statistical methods.

The results validate the proposal, achieving a predictive model with a sMAPE error of 26.7 % and a MASE of 0.66. In conclusion, this work provides the system's intelligence, enabling a shift from simple historical data visualization to the prediction of neighborhood safety trends.

**Keywords:** Deep Learning, Time Series, Time Series Transformer, Probabilistic Forecasting, Crime Index, Feature Engineering, CRISP-DM.

## Glosario

**Distribución t de Student** Distribución de probabilidad utilizada en la capa de salida del Transformer. Es útil para modelar colas pesadas ya que permite asignar mayor probabilidad a eventos extremos, a diferencia de la distribución Normal.

**Embeddings** Representaciones vectoriales de baja dimensión aprendidas por el modelo para codificar variables categóricas estáticas, permitiendo la transferencia de patrones y la distinción entre series geográficas.

**Sparsity** Fenómeno donde la mayoría de las series temporales tienen una baja densidad de eventos, lo que requiere técnicas de Transfer Learning para el modelado.

**Teacher Forcing** Técnica de entrenamiento para modelos secuenciales donde el decodificador recibe los valores objetivo reales del paso de tiempo anterior, en lugar de su propia predicción, para acelerar la eficiencia del aprendizaje de secuencia.

**Índice de Crimen** Métrica objetivo calculada en el Módulo PD que pondera la gravedad de los delitos por su pena mínima legal en días de presidio y normaliza el resultado por la densidad poblacional, creando la serie objetivo para el pronóstico.

## Siglas

**CRISP-DM** Cross-Industry Standard Process for Data Mining.

**DL** Deep Learning.

**ETL** Extract, Transform, Load.

**MA** Motor Analítico.

**MASE** Mean Absolute Scaled Error.

**MDA** Mean Directional Accuracy.

**NLL** Negative Log-Likelihood.

**PD** Procesamiento de Datos.

**PICP** Prediction Interval Coverage Probability.

**S.T.O.P.** Sistema Táctico de Operación Policial.

**sMAPE** Symmetric Mean Absolute Percentage Error.

## Tabla de Contenido

|          |                                                                  |           |
|----------|------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introducción</b>                                              | <b>1</b>  |
| 1.1      | Objetivos . . . . .                                              | 2         |
| 1.1.1    | Objetivo Principal . . . . .                                     | 2         |
| 1.1.2    | Objetivos Específicos . . . . .                                  | 2         |
| <b>2</b> | <b>Estado del Arte</b>                                           | <b>3</b>  |
| <b>3</b> | <b>Propuesta de Solución</b>                                     | <b>6</b>  |
| 3.1      | Estrategia de Procesamiento (Módulo PD) . . . . .                | 7         |
| 3.1.1    | Agregación y Suavizado Temporal . . . . .                        | 8         |
| 3.1.2    | Codificación de Variables Estáticas . . . . .                    | 8         |
| 3.2      | Estrategia Analítica y Modelado Predictivo (Módulo MA) . . . . . | 8         |
| 3.2.1    | Formulación e Implementación del Índice de Crimen . . . . .      | 9         |
| 3.2.2    | Arquitectura de Red Neuronal para Series Temporales . . . . .    | 10        |
| <b>4</b> | <b>Requisitos del Sistema</b>                                    | <b>10</b> |
| <b>5</b> | <b>Diseño de Arquitectura del Sistema</b>                        | <b>12</b> |
| 5.1      | Diseño del Módulo Procesador de Datos (PD) . . . . .             | 14        |
| 5.2      | Diseño del Motor Analítico (MA) . . . . .                        | 15        |
| 5.2.1    | Arquitectura del Índice de Crimen . . . . .                      | 15        |
| 5.2.2    | Arquitectura del Modelo Predictivo . . . . .                     | 15        |
| 5.3      | Infraestructura Tecnológica . . . . .                            | 16        |
| <b>6</b> | <b>Metodología</b>                                               | <b>18</b> |
| 6.1      | Análisis de datos . . . . .                                      | 18        |
| 6.1.1    | Caracterización Cuantitativa y Cobertura . . . . .               | 18        |
| 6.1.2    | Distribución Espacial y Concentración . . . . .                  | 21        |
| 6.1.3    | Análisis Temporal: Estacionalidad y fenómenos históricos . . .   | 22        |

## ÍNDICE DE FIGURAS

---

|          |                                                           |           |
|----------|-----------------------------------------------------------|-----------|
| 6.1.4    | Implicancias para el Modelado . . . . .                   | 24        |
| 6.2      | Preparación de los Datos . . . . .                        | 25        |
| 6.3      | Estrategia de Modelado . . . . .                          | 26        |
| 6.3.1    | Benchmark: Modelo SARIMA . . . . .                        | 26        |
| 6.3.2    | Modelo Propuesto: Time Series Transformer . . . . .       | 27        |
| 6.4      | Estrategia de Evaluación e Inferencia . . . . .           | 28        |
| <b>7</b> | <b>Resultados</b>                                         | <b>29</b> |
| 7.1      | Desempeño Comparativo: Precisión vs. Eficiencia . . . . . | 29        |
| 7.2      | Análisis de Sensibilidad por Volumen de Datos . . . . .   | 31        |
| 7.3      | Evaluación de Utilidad Operativa en Seguridad . . . . .   | 32        |
| 7.3.1    | Impacto en la Gestión de Riesgos . . . . .                | 33        |
| 7.4      | Calidad probabilística de ambos modelos . . . . .         | 33        |
| 7.4.1    | Discusión Estadística . . . . .                           | 34        |
| 7.5      | Limitaciones Encontradas . . . . .                        | 35        |
| <b>8</b> | <b>Conclusiones</b>                                       | <b>36</b> |

## Referencias

## Índice de figuras

|     |                                                                                                                                                                                                                                                                        |    |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Arquitectura General del Sistema. Los módulos PD y MA constituyen el foco de desarrollo de esta memoria. . . . .                                                                                                                                                       | 7  |
| 5.1 | Diagrama de flujo detallado del pipeline analítico, mostrando la conversión de los datos crudos en la serie objetivo ( $IC_{norm}$ ) dentro del Módulo de Procesamiento (PD) y su consumo como entrada paralela de <i>features</i> en el Motor Analítico (MA). . . . . | 13 |
| 6.1 | Distribución de Frecuencia por Tipología Delictual. Se evidencia la asimetría de los datos. . . . .                                                                                                                                                                    | 20 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                              |    |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 6.2 | Curva de concentración espacial de delitos. Se observa que una pequeña fracción de cuadrantes acumula gran parte de la historia delictual, dejando una extensa “zona de sparsity”. . . . .                                                                                                                                                                                                                                   | 21 |
| 6.3 | Descomposición de la serie temporal de delitos (2018-2025). El panel de “Estacionalidad” (verde) confirma un ciclo de 12 meses persistente. El panel de “Tendencia” (azul) evidencia los quiebres estructurales generados por el Estallido Social y la Pandemia. . . . .                                                                                                                                                     | 23 |
| 7.1 | Dispersión de Error Punto a Punto (MASE y sMAPE). Cada punto representa un cuadrante. La línea diagonal roja indica paridad de desempeño ( $y = x$ ). Los puntos situados en la zona verde (debajo de la diagonal) corresponden a series donde el Transformer obtuvo un menor error que el benchmark, evidenciando que la ventaja de SARIMA no es universal, sino dependiente de la dinámica específica de cada serie. . . . | 31 |
| 7.2 | Desempeño MASE según disponibilidad de datos. La brecha de error se reduce significativamente en las zonas de alta densidad, demostrando que la red neuronal converge al desempeño del experto local cuando dispone de información suficiente. . . . .                                                                                                                                                                       | 32 |

## Índice de tablas

|     |                                                                                                                                                                                             |    |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Comparativa de Capacidades Analíticas y Predictivas . . . . .                                                                                                                               | 5  |
| 5.1 | Stack Tecnológico para Procesamiento y Modelado . . . . .                                                                                                                                   | 17 |
| 6.1 | Resumen Técnico del Dataset S.T.O.P . . . . .                                                                                                                                               | 19 |
| 6.2 | Disparidad en la Frecuencia Delictual. La tabla presenta las cuatro categorías de mayor incidencia, junto con el delito de Homicidio, para ilustrar el severo desbalance de clases. . . . . | 20 |
| 6.3 | 5 Cuadrantes con Mayor Concentración Delictual (2018-2025) . . . . .                                                                                                                        | 22 |
| 7.1 | Comparativa de Desempeño y Eficiencia . . . . .                                                                                                                                             | 30 |
| 7.2 | Métricas de Dinámica y Seguridad Preventiva . . . . .                                                                                                                                       | 33 |

|     |                                                                        |    |
|-----|------------------------------------------------------------------------|----|
| 7.3 | Comparación de Calibración Probabilística y Detección de Eventos . . . | 34 |
|-----|------------------------------------------------------------------------|----|

### 1. Introducción

La adquisición de una propiedad constituye una de las decisiones financieras y personales más trascendentales en la vida de las personas. Tal como se establece en la literatura especializada, factores como el costo y la ubicación han sido históricamente los determinantes en este proceso Barreto et al., 2024. Sin embargo, la seguridad del entorno ha emergido como una variable crítica que incide directamente tanto en la calidad de vida de los residentes como en la plusvalía de los inmuebles a largo plazo. A pesar de esta relevancia, la evaluación objetiva de la seguridad barrial sigue siendo un desafío complejo debido a la dispersión, la heterogeneidad y la naturaleza técnica de la información disponible.

En el escenario actual, aunque existen fuentes de datos gubernamentales que proveen registros detallados de denuncias y casos policiales (como S.T.O.P. o CEAD), y portales privados como PortalInmobiliario o TocToc que consolidan oferta, precios y características de las viviendas, estos operan como silos de información aislados. La mera disponibilidad de estos registros en bruto no genera por sí misma un conocimiento accionable para el usuario final. Los datos crudos suelen contener ruido, presentan formatos inconsistentes y carecen de una estructura que permita su correlación directa. Más aún, la visualización de estadísticas pasadas no resuelve la necesidad fundamental de anticipar el comportamiento futuro de un sector. Por lo tanto, para que esta información aporte valor real a la toma de decisiones, es imperativo someterla a rigurosos procesos de limpieza, normalización y modelado matemático.

Este documento se enfoca en la formulación de la inteligencia lógica y el desarrollo de los mecanismos de procesamiento analítico, elementos indispensables para dotar de valor interpretativo al proyecto global. Sin una capacidad sistemática para depurar, normalizar y modelar la información heterogénea, la mera acumulación de registros resultaría en un repositorio de datos estéril, incapaz de orientar efectivamente la toma de decisiones. Por consiguiente, el núcleo de este trabajo consistió en diseñar y materializar los algoritmos de transformación y predicción necesarios para convertir datos

crudos en conocimiento estratégico.

### **1.1. Objetivos**

#### **1.1.1. Objetivo Principal**

Generar indicadores de seguridad objetivos y proyecciones de tendencias confiables para la toma de decisiones, mediante la implementación de algoritmos de procesamiento de datos y modelos de aprendizaje profundo.

#### **1.1.2. Objetivos Específicos**

- Desarrollar algoritmos de limpieza y normalización para estandarizar datos provenientes de fuentes policiales e inmobiliarias, asegurando la integridad de la base de datos.
- Implementar una métrica objetiva de seguridad que pondere la gravedad de los delitos y la densidad poblacional, permitiendo comparaciones equitativas.
- Entrenar modelos de Deep Learning basados en análisis de series temporales para predecir la evolución de la criminalidad a nivel de cuadrante.
- Validar la precisión de los modelos predictivos utilizando métricas robustas como sMAPE y MASE, asegurando que el sistema aporte valor real.
- Integrar los resultados del procesamiento y análisis en la base de datos central para su consumo por la interfaz de usuario.

### 2. Estado del Arte

En la actualidad, las soluciones disponibles operan bajo lógicas segmentadas que dificultan una evaluación integral del riesgo y su impacto en la plusvalía inmobiliaria. Para ilustrar esta brecha tecnológica desde una perspectiva analítica, la Tabla 2.1 contrasta las capacidades de procesamiento y modelado de las plataformas más relevantes frente a la solución desarrollada en este trabajo.

Como se desprende del análisis, las plataformas gubernamentales chilenas desempeñan un rol fundamental como repositorios de datos. Entre ellas destaca el **Centro de Estudios y Análisis del Delito (CEAD)**, organismo técnico dependiente de la Subsecretaría de Prevención del Delito del Ministerio del Interior y Seguridad Pública, 2025, y el **Sistema Táctico de Operación Policial (S.T.O.P)**, plataforma administrada por Carabineros de Chile, 2025. Si bien estos sistemas cumplen eficazmente su función de registro histórico, su enfoque se limita a presentar conteos brutos de delitos, careciendo de herramientas que ponderen la gravedad de los incidentes o proyecten escenarios futuros para el ciudadano común.

En el ámbito internacional, soluciones como **CrimeMapping** The Omega Group, 2025 agregan datos de múltiples agencias policiales para visualización, pero al igual que sus contrapartes locales, operan sin vincular la seguridad con variables económicas.

Por otra parte, el mercado inmobiliario nacional cuenta con una oferta digital robusta liderada por plataformas como **Portal Inmobiliario**, **TocToc** y **Yapo.cl** Portal Inmobiliario, 2025. A nivel global, referentes como **Zillow** o **Trulia** marcan la pauta en la digitalización del sector. Sin embargo, en la mayoría de estos portales, la variable *entorno* suele abordarse de manera descriptiva, como la cercanía a servicios, o anecdótica, sin métricas objetivas de riesgo delictivo que respalden la valoración financiera del sector.

Desde la perspectiva académica, la necesidad de correlacionar estas variables sigue siendo un campo de investigación activo. Investigaciones como la de Chisari et al., 2019

## **CAPÍTULO 2 : ESTADO DEL ARTE**

---

han cuantificado el impacto económico, demostrando que un incremento del 10 % en la tasa de criminalidad puede depreciar el valor de las propiedades hasta un 4 % en ciertos segmentos. En el contexto local, Vergara-Perucich, 2021 identificó la seguridad pública como un determinante urbano crítico en la formación de precios de vivienda en Chile.

Literatura más reciente refuerza y expande estos hallazgos. Estudios como los de Cuevas Ahumada, 2025 en el contexto latinoamericano han validado la relación negativa entre la percepción de inseguridad y los precios de vivienda. Asimismo, investigaciones en Europa Ceccato y Wilhelmsson, 2020 han demostrado la influencia crítica de la proximidad a focos delictuales en la valoración inmobiliaria. Estos antecedentes confirman que, si bien la correlación teórica está probada, existe un vacío tecnológico en la implementación de modelos automatizados que operacionalicen estas relaciones para el usuario final.

La presente memoria busca cubrir esta brecha mediante el desarrollo de algoritmos de procesamiento y modelos de aprendizaje automático capaces de transformar estos datos dispersos en predicciones prácticas.

## CAPÍTULO 2 : ESTADO DEL ARTE

---

*Tabla 2.1: Comparativa de Capacidades Analíticas y Predictivas*

| <b>Capacidad Analítica</b>          | <b>CEAD</b> | <b>S.T.O.P.</b> | <b>CrimeMapping</b> | <b>Portales Inmobiliarios</b> | <b>Propuesta</b> |
|-------------------------------------|-------------|-----------------|---------------------|-------------------------------|------------------|
| Estadística Descriptiva (Histórico) | ✓           | ✓               | ✓                   | ✓                             | ✓                |
| Índice de Seguridad Ponderado       | ✗           | ✗               | ✗                   | ✗                             | ✓                |
| Correlación Seguridad-Precio        | ✗           | ✗               | ✗                   | ✗                             | ✓                |
| Predicción de Tendencias            | ✗           | ✗               | ✗                   | ✗                             | ✓                |
| Normalización Poblacional           | ✓           | ✗               | ✗                   | ✗                             | ✓                |

*Fuente: Elaboración propia.*

### 3. Propuesta de Solución

El presente trabajo se enmarca dentro del desarrollo de un sistema integral de información para la seguridad ciudadana, cuyo objetivo es facilitar el acceso a datos de riesgo y mercado inmobiliario en una misma solución. Para comprender la propuesta de solución específica de esta memoria, es necesario primero visualizar la arquitectura general del sistema.

La solución global se estructura en un flujo secuencial de cinco módulos interconectados, tal como se ilustra en la Figura 3.1:

1. **Módulo Recolector de Datos (RD):** Encargado de la extracción automatizada de información desde fuentes oficiales, como S.T.O.P.; portales inmobiliarios, entre los que se incluyen TocToc, Portal Inmobiliario y Yapo; así como medios de prensa digital.
2. **Módulo Procesador de Datos (PD):** Responsable de la limpieza, normalización y estructuración de la información cruda.
3. **Motor Analítico (MA):** El núcleo de inteligencia que aplica algoritmos de Machine Learning para generar predicciones.
4. **Módulo de Almacenamiento (AD):** Repositorio central que actúa como fuente de verdad.
5. **Interfaz de Usuario (IU):** Plataforma web de visualización para el usuario final.

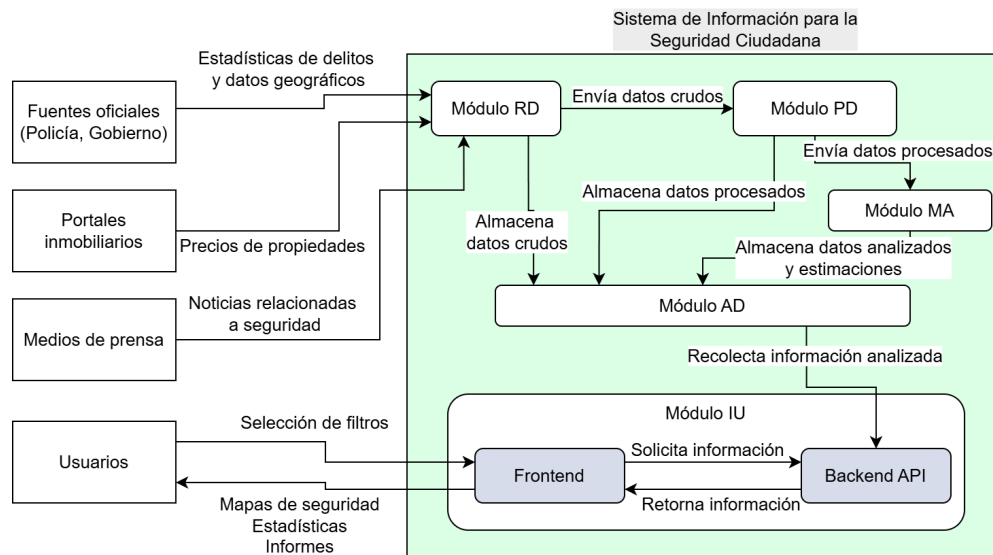


**Figura 3.1:** Arquitectura General del Sistema. Los módulos PD y MA constituyen el foco de desarrollo de esta memoria.

**Fuente:** Elaboración propia.

Dentro de este ecosistema, el núcleo de la contribución técnica de este trabajo de título reside en el diseño e implementación de los componentes de inteligencia: el **Módulo de Procesamiento de Datos (PD)** y el **Motor Analítico (MA)**.

Mientras que la recolección provee la materia prima y la interfaz gestiona la interacción, son estos módulos centrales los encargados de refinar, estructurar e interpretar dicha información. La propuesta de solución se estructura como un pipeline de ingeniería de datos que procesa la información bruta para generar métricas estandarizadas y estimaciones probabilísticas.

### 3.1. Estrategia de Procesamiento (Módulo PD)

Dado que los modelos de aprendizaje profundo requieren entradas altamente estructuradas, se desarrolló una estrategia de procesamiento enfocada en la *Ingeniería de Atributos* y la *Agregación Temporal*. El objetivo fue transformar los registros en bruto en series de tiempo consistentes:

### 3.1.1. Agregación y Suavizado Temporal

Los datos delictuales crudos presentan una alta varianza y ruido a nivel semanal, manifestándose en semanas con cero delitos seguidos de picos aleatorios. Para mitigar esto y facilitar la convergencia del modelo, se implementó una lógica de agregación en ventanas de cuatro semanas, el tamaño fue seleccionado para asegurar la consistencia estadística, ya que contiene exactamente cuatro ciclos semanales completos. Esto elimina el ruido de calendario y simplifica la agregación de datos de la fuente S.T.O.P. De esta forma, los registros históricos son agrupados en *time-steps*, sumando los incidentes dentro de cada ventana. Este proceso suaviza la curva de tendencias, permitiendo que el modelo capture patrones estacionales como aumentos en festividades o descenso en vacaciones de verano sin confundirse con el ruido estocástico semanal.

### 3.1.2. Codificación de Variables Estáticas

Para que el modelo global pueda distinguir entre diferentes zonas geográficas, se procesaron la comuna y región del cuadrante como variables categóricas. Estas se transformaron en formatos numéricos optimizados para ser utilizados como *Embeddings* (representaciones vectoriales) dentro del modelo de Machine Learning. Esto permite al sistema transferir aprendizaje: el modelo entiende que dos cuadrantes pertenecientes a la misma región o comuna pueden compartir características latentes, mejorando la predicción en zonas con pocos datos históricos.

## 3.2. Estrategia Analítica y Modelado Predictivo (Módulo MA)

El componente de inteligencia del sistema se diseñó para responder a dos necesidades: la evaluación objetiva del estado actual y la proyección futura.

### 3.2.1. Formulación e Implementación del Índice de Crimen

La evaluación de la seguridad no puede basarse únicamente en un conteo simple de eventos, ya que esto introduce sesgos críticos. Por un lado, los delitos de menor gravedad pero alta frecuencia (como hurtos) tienden a inflar desproporcionadamente las estadísticas, opacando eventos de mayor impacto social pero menor ocurrencia (como homicidios). Por otro lado, las zonas densamente pobladas presentan cifras absolutas naturalmente más altas. Para reducir el impacto de los dos hechos señalados anteriormente, se definió el *Índice de Crimen Normalizado*, denotado por  $IC_{norm}$ .

Su definición (Ecuación 1) consiste en una suma ponderada para corregir el desbalance de gravedad y normaliza el resultado según la demografía del sector:

$$IC_{norm} = \alpha \cdot \frac{\sum_{i=1}^n (C_i \times W_i)}{P_{com}} \quad (1)$$

Donde:

- $C_i$  representa la cantidad de casos del delito tipo  $i$ .
- $W_i$  es el factor de ponderación del delito  $i$ . Se calcula como la proporción entre su pena mínima en días  $p_i$  y la sumatoria de las penas mínimas de todos los delitos considerados  $W_i = \frac{p_i}{\sum_{j=1}^n p_j}$ . Esta ponderación asegura que los delitos graves tengan una incidencia mayor en el índice, evitando que sea dominado puramente por el volumen de delitos menores.
- $P_{com}$  corresponde a la población total de la comuna donde se ubica el cuadrante.
- $\alpha$  es el factor de escala. Para efectos de este proyecto, este factor toma el valor de 10.000. La justificación operativa de este escalamiento se detalla en la Sección 6.2

La denominación *norm* hace referencia a la normalización poblacional representada por el denominador  $P_{com}$ . Sin este ajuste, los cuadrantes con mayor densidad de

habitantes tendrían siempre índices elevados solo por su tamaño. Al dividir por la población y aplicar el factor de escala  $\alpha$ , la métrica permite comparar equitativamente la seguridad entre zonas geográficas de distinta envergadura y realidad delictual.

Para garantizar la eficiencia y la consistencia de los datos entre el entrenamiento de modelos y la visualización web, la lógica de este índice se implementó directamente en la base de datos mediante una Vista SQL.

### 3.2.2. Arquitectura de Red Neuronal para Series Temporales

Para la capacidad predictiva, se desestimaron los modelos estadísticos clásicos en favor de una arquitectura de tipo Transformer, adaptada específicamente para series temporales. Las características distintivas de esta implementación incluyen:

- **Modelo Global con Transferencia de Aprendizaje:** En lugar de entrenar modelos aislados para cada cuadrante, se entrenó un único modelo global con la totalidad de los datos históricos de Chile. Esto permite que el sistema aprenda patrones universales y los transfiera a zonas con poca historia de datos.
- **Manejo de Incertidumbre Probabilística:** Reconociendo la naturaleza estocástica del crimen, el modelo no predice un valor puntual determinista, sino que estima los parámetros de una distribución de probabilidad *t de Student* para cada paso de tiempo futuro. Esto es fundamental para la toma de decisiones, ya que permite entregar al usuario un intervalo de confianza que refleja la certidumbre del sistema sobre dicha predicción.

## 4. Requisitos del Sistema

Este trabajo tuvo la responsabilidad directa de satisfacer los requisitos funcionales de análisis y procesamiento, críticos para la inteligencia del sistema.

### Requisitos Funcionales Atendidos:

## CAPÍTULO 4 : REQUISITOS DEL SISTEMA

---

- **RF4: Normalización y estructuración de datos:** Implementado mediante scripts de Python que estandarizan fechas, clasifican tipos de delitos y vinculan espacialmente los datos.
- **RF5: Clasificación y categorización:** Desarrollo de lógica para categorizar la gravedad de los incidentes, base para el cálculo de índices.
- **RF6: Análisis estadístico:** Cumplido a través de la implementación del cálculo del Índice de Crimen y agregaciones estadísticas por zona (promedios, variaciones porcentuales).
- **RF7: Modelado predictivo:** Este fue el foco principal del Módulo MA. Se satisfizo mediante el desarrollo, entrenamiento y validación del modelo de series temporales para seguridad. (Nota: El modelado de precios inmobiliarios quedó limitado por la disponibilidad de datos históricos, como se detalla en la sección de Riesgos del informe grupal).

### Requisitos No Funcionales Clave:

- **RNF1: Rendimiento:** El procesamiento de datos y las predicciones se realizan en procesos en segundo plano y se almacenan pre-calculados en la base de datos. Esto asegura que, aunque el modelo de IA sea complejo, la consulta del usuario final esté en el orden de milisegundos.

### 5. Diseño de Arquitectura del Sistema

La arquitectura del sistema se concibe como una estructura modular y secuencial, diseñada para gestionar el ciclo de vida completo del dato, desde su recolección en fuentes heterogéneas hasta su transformación en información estratégica para la toma de decisiones. Si bien la plataforma global integra componentes de ingesta y visualización, este capítulo se centra en la especificación técnica del núcleo de inteligencia del proyecto: el Módulo de Procesamiento de Datos (PD) y el Motor Analítico (MA).

Estos componentes tienen como función estructurar y refinar los datos crudos, preparándolos para el modelo de aprendizaje profundo. Este proceso garantiza la integridad, consistencia y valor analítico de la información. El pipeline detallado de datos, que ilustra la conversión de los registros en las series objetivo y su integración con variables temporales y estáticas para alimentar el modelo predictivo, se presenta en la Figura 5.1.

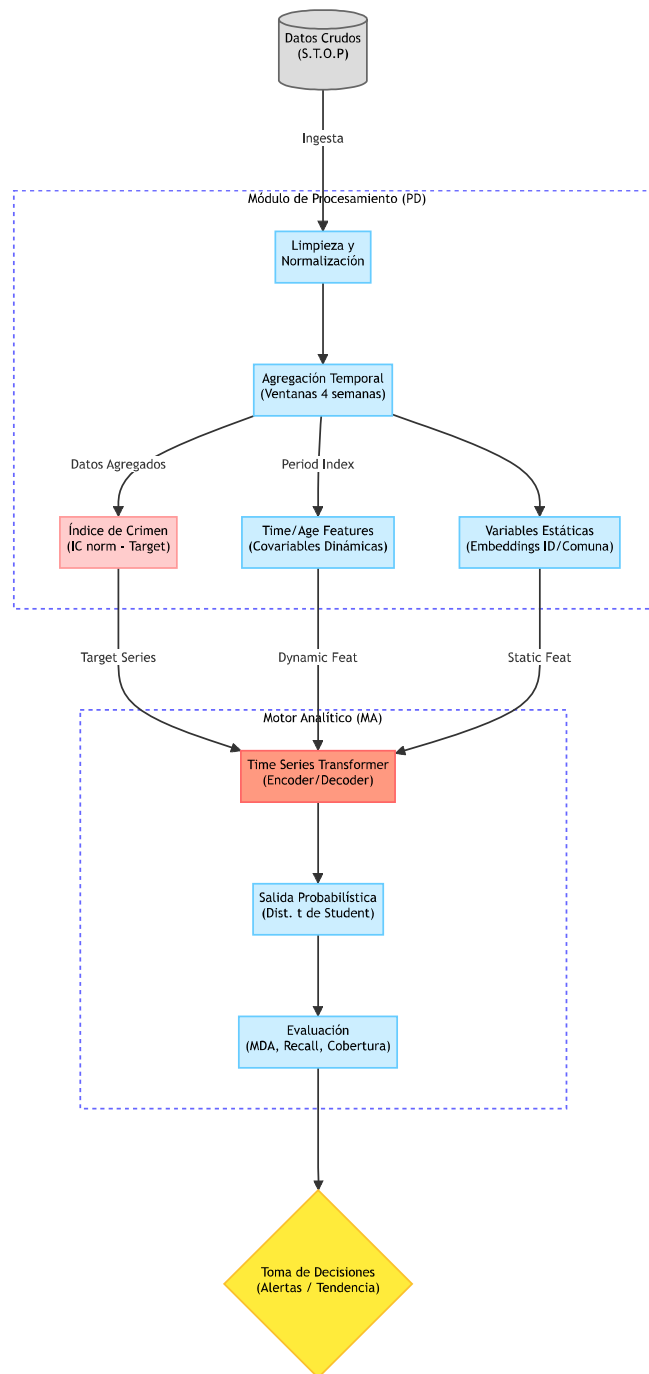


**Figura 5.1:** Diagrama de flujo detallado del pipeline analítico, mostrando la conversión de los datos crudos en la serie objetivo ( $IC_{norm}$ ) dentro del Módulo de Procesamiento (PD) y su consumo como entrada paralela de features en el Motor Analítico (MA).

**Fuente:** Elaboración propia.

### 5.1. Diseño del Módulo Procesador de Datos (PD)

Este módulo actúa como un filtro inteligente de pre-procesamiento entre la recolección de datos crudos y los modelos analíticos. Su diseño se centra en la transformación de datos para el aprendizaje automático, estructurándose en tres etapas lógicas:

- **Normalización Estructural:** La primera etapa del diseño aborda la naturaleza semi-estructurada de la fuente. El módulo implementa rutinas de *parsing* que transforman la estructura jerárquica de los reportes semanales crudos (en formato JSON anidado) en registros relacionales normalizados. Este proceso incluye el mapeo de categorías textuales a identificadores numéricos y la validación de integridad antes de la persistencia en base de datos.
- **Estrategia de Agregación Temporal:** Dado que los datos de entrada poseen una frecuencia semanal con alta varianza, se diseñó una estrategia de agregación en ventanas de cuatro semanas ( $N = 28$  días). Esta decisión de diseño se explica en la Sección 6.2
- **Codificación de Variables Estáticas:** Se implementó un mecanismo para procesar las variables categóricas permitiendo que el modelo aprenda de ellas.

El proceso se realiza en dos pasos. Primero, se transforman las etiquetas de Cuadrante, Comuna y Región en índices numéricos únicos. Esto se logra mediante la creación de diccionarios de mapeo independientes para cada variable, los cuales asignan un número entero secuencial (de 0 a  $N - 1$ ) a cada etiqueta distinta presente en el conjunto de datos, garantizando una identificación unívoca para cada entidad geográfica.

Posteriormente, siguiendo la documentación de la arquitectura *Time Series Transformer*, explicada (Ver sección 5.2.2), estos índices pasan por una capa de *Embedding* que les asigna un vector de números reales que el modelo aprende y ajusta durante el entrenamiento, permitiéndole generar representaciones matemáticas que capturan las similitudes entre las distintas zonas.

### 5.2. Diseño del Motor Analítico (MA)

El diseño del motor analítico se divide en una lógica determinista para la evaluación actual y una lógica estocástica para la proyección futura.

#### 5.2.1. Arquitectura del Índice de Crimen

Para garantizar la eficiencia del sistema y evitar el recálculo constante de métricas complejas, la lógica del Índice de Crimen (ecuación 1) se implementó como una vista materializada en PostgreSQL. Esto desacopla la lógica matemática del código de la aplicación, permitiendo consultas de alta velocidad.

#### 5.2.2. Arquitectura del Modelo Predictivo

Se seleccionó una arquitectura basada en Transformers, específicamente una topología Encoder-Decoder estándar adaptada para el pronóstico de series temporales. Esta elección se debe a su capacidad de atención (Self-Attention), superior a las redes recurrentes tradicionales para capturar dependencias a largo plazo. Los detalles específicos sobre la implementación de la red, la configuración de hiperparámetros y el entrenamiento se abordan en profundidad en la sección 6.3.2.

- **Enfoque de Modelo Global:** El diseño arquitectónico contempla un único modelo entrenado con los datos de todos los cuadrantes simultáneamente. Esto permite la transferencia de aprendizaje: el modelo aprende patrones universales de los cuadrantes con muchos datos y aplica ese conocimiento a cuadrantes con historia limitada.
- **Salida Probabilística:** En lugar de una regresión puntual, la capa de salida del modelo proyecta los parámetros de una distribución t de Student. Esto permite generar múltiples escenarios (pesimista, optimista, mediano) y entregar al usuario una banda de confianza, vital para la gestión de incertidumbre en seguridad.

### **5.3. Infraestructura Tecnológica**

El desarrollo de los módulos de procesamiento y análisis se basó en una arquitectura que prioriza la paralelización de tareas y la eficiencia en el manejo de series temporales. A continuación, se detallan las tecnologías críticas utilizadas en el Módulo de Procesamiento de Datos (PD) y el Motor Analítico (MA), resumidas en la Tabla 5.1.

## CAPÍTULO 5 : DISEÑO DE ARQUITECTURA DEL SISTEMA

**Tabla 5.1:** Stack Tecnológico para Procesamiento y Modelado

| Categoría                                       | Herramienta            | Rol Metodológico                | Justificación                                            |
|-------------------------------------------------|------------------------|---------------------------------|----------------------------------------------------------|
| <i>Infraestructura y ETL</i>                    |                        |                                 |                                                          |
| Lenguaje                                        | <b>Python</b> (v3.10+) | Programación del pipeline ETL.  | Estándar en Data Science y control de flujo.             |
| B. de Datos                                     | <b>PostgreSQL</b>      | Almacenamiento primario.        | Motor relacional robusto y transaccional.                |
| Geospatial                                      | <b>PostGIS</b>         | Manejo de geometrías.           | Consultas espaciales nativas y eficientes.               |
| <i>Núcleo de Inteligencia (Motor Analítico)</i> |                        |                                 |                                                          |
| Arq. IA                                         | <b>Hugging Face</b>    | Implementación del Transformer. | Arquitectura TimesSeriesTransformer lista para entrenar. |
| Framework                                       | <b>PyTorch</b>         | Motor de cómputo GPU.           | Backend de cálculo tensorial en GPU.                     |
| Benchmark                                       | <b>pmdarima</b>        | Línea base SARIMA.              | Selección automática de parámetros (AutoARIMA).          |
| Features                                        | <b>GluonTS</b>         | Ingeniería de atributos.        | Transformación de series y generación de features.       |
| Métricas                                        | <b>Evaluate</b>        | Cálculo de métricas.            | Cómputo validado de MASE y sMAPE.                        |

**Fuente:** Elaboración propia.

### 6. Metodología

Para el desarrollo de los componentes analíticos y de procesamiento, se adoptó la metodología **CRISP-DM (Cross-Industry Standard Process for Data Mining)** Schröer et al., 2021. La elección de este marco de trabajo se fundamenta en su naturaleza cíclica e iterativa, la cual se ajusta mejor a la incertidumbre inherente a los proyectos de aprendizaje automático que las metodologías de desarrollo de software tradicionales.

Dado que el modelado de crimen requiere refinar constantemente la limpieza de datos en función de los resultados preliminares del modelo, la flexibilidad de CRISP-DM permitió iterar entre la comprensión de los datos y el modelado sin perder la estructura del proyecto.

#### 6.1. Análisis de datos

Se realizó un análisis exploratorio exhaustivo de los datos históricos provenientes de Carabineros. Este proceso permitió no solo dimensionar el volumen de información, sino también auditar la consistencia de los registros y establecer las propiedades estadísticas clave que eran esenciales para definir la arquitectura de salida y la estrategia de normalización requeridas por el modelo predictivo.

##### 6.1.1. Caracterización Cuantitativa y Cobertura

En términos de volumen, el conjunto de datos consolidado abarca un rango temporal de 7 años, desde el *1 de enero de 2018* hasta el *31 de octubre de 2025*.

La base de datos contiene un total de 3.716.704 registros transaccionales. Si bien el sistema define administrativamente 1.147 códigos de cuadrantes, el cruce de información reveló que solo un subconjunto de estos posee historial delictual efectivo. Aproximadamente un 10 % de los IDs definidos no presentan registros en la tabla de casos, correspondiendo a códigos administrativos obsoletos o zonas sin actividad reportada,

## CAPÍTULO 6 : METODOLOGÍA

---

los cuales fueron excluidos del análisis de concentración para evitar sesgos estadísticos.

Las características técnicas del dataset consolidado se resumen en la Tabla 6.1.

**Tabla 6.1:** *Resumen Técnico del Dataset S.T.O.P*

| <b>Característica</b>            | <b>Estadística</b>      |
|----------------------------------|-------------------------|
| Total de Registros (N)           | 3.716.704               |
| Total IDs Definidos              | 1.147                   |
| Cuadrantes con Historia Efectiva | 1.040                   |
| Rango Temporal                   | 01/01/2018 – 31/10/2025 |

**Fuente:** *Elaboración propia.*

Respecto a la tipología delictual, se observó un severo desbalance de clases (ver Figura 6.1). Los delitos de alta frecuencia (Hurtos) superan en órdenes de magnitud a los delitos críticos (Homicidios). Para observar numéricamente esta disparidad, la tabla 6.2 presenta la frecuencia detallada de los cuatro delitos con mayor incidencia junto al delito de Homicidio.

## CAPÍTULO 6 : METODOLOGÍA

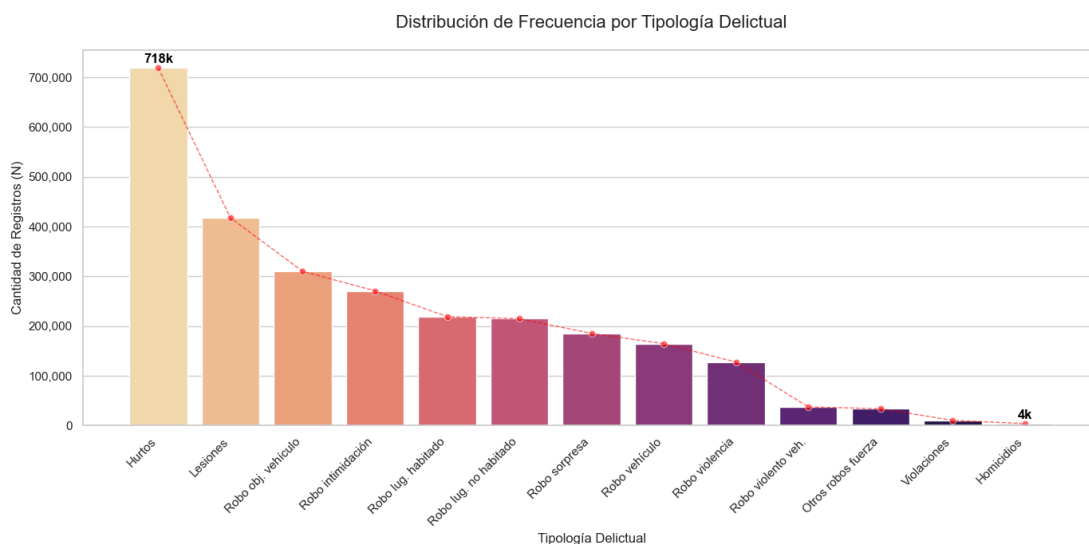


**Figura 6.1:** Distribución de Frecuencia por Tipología Delictual. Se evidencia la asimetría de los datos.

**Fuente:** Elaboración propia.

**Tabla 6.2:** Disparidad en la Frecuencia Delictual. La tabla presenta las cuatro categorías de mayor incidencia, junto con el delito de Homicidio, para ilustrar el severo desbalance de clases.

| Tipo de Delito                 | Cantidad (N°)    | Frecuencia ( %) |
|--------------------------------|------------------|-----------------|
| Hurtos                         | 718.767          | 26,47 %         |
| Lesiones                       | 417.435          | 15,37 %         |
| Robo objetos de/desde vehículo | 309.923          | 11,41 %         |
| Robo con intimidación          | 270.899          | 9,98 %          |
| Homicidios                     | 4.079            | 0,15 %          |
| <b>Total Muestra</b>           | <b>3.716.704</b> | <b>100,00 %</b> |

**Fuente:** Elaboración propia.

### 6.1.2. Distribución Espacial y Concentración

El análisis geoespacial expuso una concentración de la actividad criminal extremadamente desigual. Al examinar el universo efectivo de cuadrantes con actividad registrada ( $N = 1,040$ ), se observa una marcada disparidad entre zonas de alta densidad de eventos y zonas con actividad esporádica.

Tal como ilustra la Figura 6.2, existe un fenómeno de hiper-concentración: un subconjunto crítico compuesto por apenas el 10 % de los cuadrantes operativos (las 104 zonas críticas marcadas en rojo) acumula el 40,1 % de todo el volumen histórico de datos. Estas áreas, correspondientes a comunas demográficamente densas como La Florida o Concepción (detalladas en la Tabla 6.3), generan series temporales continuas que constituyen la fuente principal de aprendizaje de patrones para el sistema.

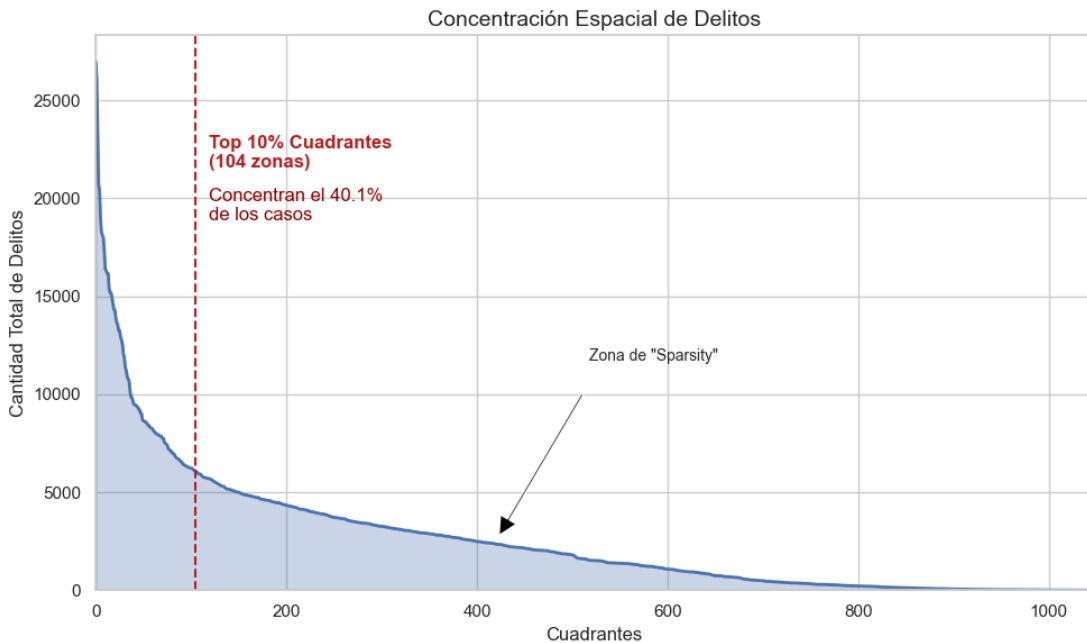


**Figura 6.2:** Curva de concentración espacial de delitos. Se observa que una pequeña fracción de cuadrantes acumula gran parte de la historia delictual, dejando una extensa “zona de sparsity”.

**Fuente:** Elaboración propia.

En contraposición, el 90 % restante de los cuadrantes se reparte el volumen rema-

## CAPÍTULO 6 : METODOLOGÍA

---

nente. La curva de distribución decae rápidamente hacia una *cola larga*, lo que indica una sparsity extrema en la gran mayoría del territorio. Estas zonas presentan registros esporádicos o discontinuos, careciendo de la densidad de información necesaria para el entrenamiento de modelos locales aislados sin caer en sobreajuste.

**Tabla 6.3:** 5 Cuadrantes con Mayor Concentración Delictual (2018-2025)

| ID Cuadrante | Comuna           | Región        | Total Delitos |
|--------------|------------------|---------------|---------------|
| 3324162      | La Florida       | Metropolitana | 26.964        |
| 1527001      | Concepción       | Biobío        | 26.098        |
| 3295123      | Providencia      | Metropolitana | 23.432        |
| 1730001      | Temuco           | La Araucanía  | 20.711        |
| 3919191      | Estación Central | Metropolitana | 20.366        |

**Fuente:** Elaboración propia.

### 6.1.3. Análisis Temporal: Estacionalidad y fenómenos históricos

Finalmente, se analizó el comportamiento de la serie temporal agregada a nivel nacional para detectar patrones cíclicos y evaluar el impacto de eventos históricos recientes. Mediante una descomposición aditiva de la serie mensual (Figura 6.3), se aislaron los componentes de tendencia y estacionalidad.

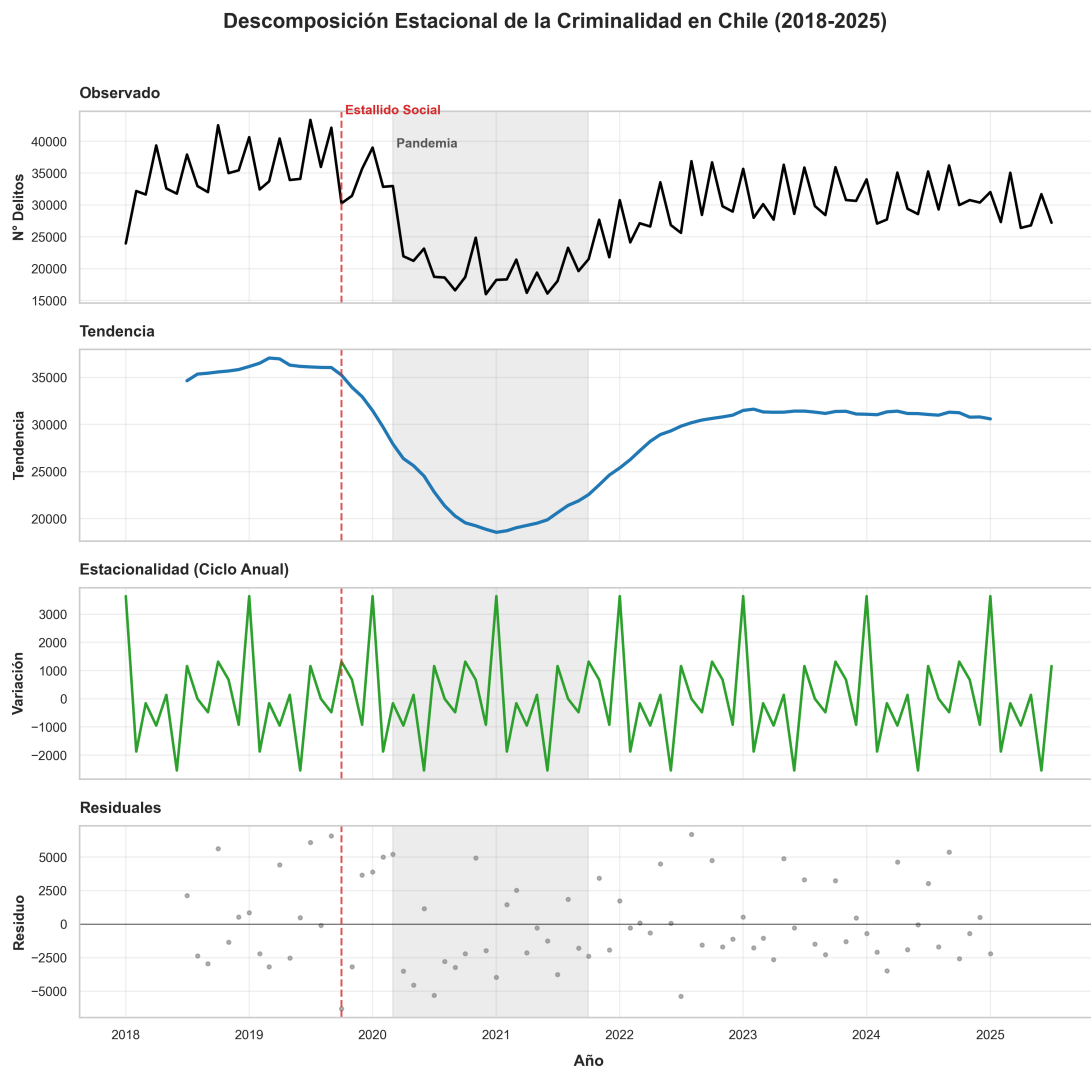


**Figura 6.3:** Descomposición de la serie temporal de delitos (2018-2025). El panel de “Estacionalidad” (verde) confirma un ciclo de 12 meses persistente. El panel de “Tendencia” (azul) evidencia los quiebres estructurales generados por el Estallido Social y la Pandemia.

**Fuente:** Elaboración propia.

El análisis visual de la Figura 6.3 arroja dos conclusiones determinantes:

1. **Persistencia de la Estacionalidad:** Existe un patrón oscilatorio robusto que se repite sistemáticamente cada 12 meses, con picos hacia fin de año y valles en febrero. Este ciclo se mantiene constante incluso durante la pandemia.

2. **No-Linealidad de la Tendencia:** La criminalidad no sigue una tendencia lineal simple; sufrió una caída abrupta de más de un 40 % durante las restricciones sanitarias (2020-2021) seguida de una recuperación progresiva.

### 6.1.4. Implicancias para el Modelado

La integración de los hallazgos cuantitativos fundamentó las decisiones críticas de diseño del Motor Analítico:

1. **Estrategia ante la Sparsity:** La existencia de una “cola larga” de cuadrantes con baja densidad de datos plantea un desafío para los modelos locales aislados, los cuales requieren historias densas para calibrar sus parámetros estacionales. Se optó por un enfoque global con el objetivo de explotar patrones compartidos: la arquitectura *Times Series Transformer* permite aprender la dinámica temporal, como ciclos y tendencias, de los cuadrantes con mayor información y generalizar este conocimiento hacia las zonas con historial escaso, ofreciendo una solución escalable que no requiere gestionar miles de modelos independientes.
2. **Necesidad de Ponderación:** La relación 175:1 entre delitos frecuentes y graves valida el diseño del *Índice de Crimen Normalizado*, evitando que la función de pérdida del modelo se optimice exclusivamente para minimizar el error en delitos menores de alto volumen.
3. **Captura de Dependencias Temporales:** La confirmación de la estacionalidad de la serie y los quiebres estructurales durante eventos como la pandemia hacen mandatorio el uso de arquitecturas con memoria a largo plazo. Se seleccionó el Transformer por su capacidad de atención para relacionar eventos distantes en el tiempo, utilizando **SARIMA** como benchmark estadístico para validar si esta complejidad arquitectónica se traduce en una mejora real de desempeño.

### 6.2. Preparación de los Datos

Esta fase fue crítica para adaptar los datos crudos y dispersos a los requisitos de entrada de una arquitectura de Deep Learning. Se diseñó un pipeline de transformación secuencial enfocado en mitigar el ruido y enriquecer la señal:

- **Normalización Demográfica y Agregación:** Las series temporales fueron normalizadas dividiendo el volumen de casos por la población específica de la comuna asociada  $P_{com}$  y aplicando el factor de escala  $\alpha = 10,000$ . La incorporación de este factor responde a un criterio práctico de operabilidad. Las tasas per cápita crudas resultan en valores decimales infinitesimales (del orden de  $10^{-4}$  o menores), lo cual dificulta su lectura e interpretación directa. Al escalar a una base de 10.000 habitantes, los índices se transforman a magnitudes numéricas manejables, facilitando tanto su análisis comparativo como su visualización en la interfaz final, sin alterar las propiedades estadísticas de la serie que procesa el modelo.

Posteriormente, los datos se agregaron en ventanas de cuatro semanas para suavizar la varianza y estabilizar la serie para el Transformer. La ventana de 28 días fue seleccionada bajo los siguientes criterios técnicos:

1. **Consistencia Estadística:** Una ventana fija de 28 días incluye exactamente cuatro ciclos semanales completos, eliminando el ruido de calendario que introducen los meses de 30 o 31 días.
  2. **Coherencia Operacional:** La fuente de datos STOP proporciona reportes con una frecuencia semanal. La agregación a 28 días permite sumar coherentemente cuatro observaciones semanales completas, manteniendo la integridad de la serie de tiempo sin la necesidad de imputar o truncar semanas parciales, lo que simplifica el *pipeline* de procesamiento.
- **Variables dinámicas:** Se enriquecieron los datos inyectando características temporales dinámicas como el “mes del año” y la “edad relativa” de la serie. Esto

permite al mecanismo de atención comprender la posición temporal absoluta de cada dato, facilitando la captura de ciclos anuales.

- **Variables Estáticas:** Las variables categóricas estáticas, como la comuna, la región y el cuadrante, se procesaron mediante la asignación de identificadores numéricos únicos. Estos se definieron secuencialmente entre 0 y  $N - 1$  (donde  $N$  representa el número total de categorías únicas para la característica en cuestión) y sirven como inputs para la capa de Embedding de la red neuronal, lo que posteriormente permite al modelo aprender similitudes latentes entre zonas geográficas y habilitar la transferencia de aprendizaje.

### 6.3. Estrategia de Modelado

Dada la naturaleza estocástica y estacional del fenómeno delictual identificada en la fase de exploración, se optó por un enfoque comparativo. Se implementó un modelo estadístico clásico como línea base para validar si la complejidad de una arquitectura de Deep Learning aporta un valor real en términos de precisión o utilidad operativa.

#### 6.3.1. Benchmark: Modelo SARIMA

Para establecer una referencia de desempeño robusta, se seleccionó el modelo **SARIMA (Seasonal AutoRegressive Integrated Moving Average)**. La elección de este modelo se fundamenta en el análisis temporal (ver Figura 6.3), que reveló un ciclo anual persistente ( $m = 12$ ) en la criminalidad.

La configuración del benchmark siguió los siguientes criterios:

- **Configuración Estacional:** Se forzó el parámetro de estacionalidad  $m = 12$  para capturar los patrones mensuales recurrentes.
- **Optimización de Hiperparámetros:** Se utilizó el algoritmo *AutoARIMA*, basado en la minimización del criterio AIC, para encontrar automáticamente la combi-

nación óptima de parámetros  $(p, d, q)(P, D, Q)_{12}$  específica para cada cuadrante.

- **Enfoque Local:** SARIMA opera como un modelo local; es decir, se entrena una instancia independiente para cada serie temporal, utilizando exclusivamente la historia de ese cuadrante sin transferencia de información externa.

### 6.3.2. Modelo Propuesto: Time Series Transformer

Como solución principal, se instanció la arquitectura **Time Series Transformer**, específicamente la implementación *TimeSeriesTransformerForPrediction* disponible en la biblioteca de código abierto *Hugging Face Transformers* Hugging Face, 2022. Esta arquitectura corresponde a un Transformer tipo *vanilla* (Encoder-Decoder) adaptado para el pronóstico probabilístico de series temporales.

La elección de esta implementación específica se fundamenta en sus capacidades nativas para manejar múltiples series temporales correlacionadas y su enfoque probabilístico. Sus características técnicas principales son:

- **Arquitectura Encoder-Decoder:** El modelo procesa una ventana de contexto histórica (*past\_values*) a través del codificador y genera predicciones futuras (*future\_values*) mediante el decodificador. Utiliza mecanismos de auto-atención para capturar dependencias a largo plazo entre los pasos de tiempo.
- **Manejo de Features:** El modelo integra nativamente distintos tipos de entradas mediante embeddings y concatenación:
  - *Time Features:* Variables temporales, como el mes del año y edad de la serie, que actúan como “positional encodings”.
  - *Static Categorical Features:* Embeddings aprendidos para identificadores estáticos (ID del cuadrante, ID de la comuna, ID de la región), permitiendo al modelo distinguir y transferir patrones entre diferentes zonas geográficas.

- **Salida Probabilística:** A diferencia de los modelos de regresión puntual, la cabecera del modelo no predice un valor único, sino que aprende los parámetros de una distribución de probabilidad. En este proyecto, se configuró una distribución **t de Student** para modelar eficazmente las colas pesadas y la incertidumbre inherente a los datos delictuales.
- **Estrategia de Entrenamiento:** El entrenamiento se realiza mediante la técnica de *Teacher Forcing*, donde el decodificador recibe los valores reales desplazados una posición como entrada durante la fase de aprendizaje, optimizando la Log-Verosimilitud Negativa.

### 6.4. Estrategia de Evaluación e Inferencia

Para validar la robustez del modelo propuesto, se diseñó un esquema de evaluación riguroso que simula un escenario de producción real, respetando la cronología de los datos, es decir, entrenamiento hasta 2023 y prueba en 2024-2025.

- **Inferencia Autoregresiva:** Durante la etapa de prueba, el modelo genera predicciones de manera secuencial: muestrea un valor de la distribución predicha y lo retroalimenta como entrada para el siguiente paso de tiempo, construyendo así la trayectoria futura completa sin conocer el futuro real.
- **Métricas de Precisión Estadística:** Se utilizaron **sMAPE (Symmetric Mean Absolute Percentage Error)** y **MASE (Mean Absolute Scaled Error)** para cuantificar la desviación de las predicciones respecto a la realidad y comparar el desempeño contra métodos ingenuos.
- **Métricas de Negocio y Seguridad:** Dado el contexto de toma de decisiones inmobiliarias, se incorporaron métricas orientadas a la utilidad operativa, calculadas sobre el conjunto de prueba:

- **MDA (Mean Directional Accuracy):** Para evaluar la capacidad de anticipar cambios de tendencia (subidas/bajadas), vital para determinar la evolución de la seguridad del sector y planificar medidas preventivas.
- **Recall de Picos:** Para medir la sensibilidad del sistema ante eventos críticos de alta delincuencia, es decir, que tan capaz es de emitir alertas tempranas.
- **Cobertura Probabilística:** Para verificar si los intervalos de confianza generados son honestos y capturan la realidad, evitando la sobreconfianza.

## 7. Resultados

A continuación se presentan los hallazgos cuantitativos obtenidos tras la implementación del pipeline de procesamiento y el entrenamiento de los modelos predictivos. La evaluación se centra en contrastar el desempeño del modelo propuesto frente al benchmark estadístico, analizando no solo la precisión del error, sino también la utilidad operativa para la toma de decisiones.

### 7.1. Desempeño Comparativo: Precisión vs. Eficiencia

Se evaluó el desempeño de ambos modelos bajo condiciones idénticas, utilizando el mismo subconjunto de entrenamiento y prueba de 250 cuadrantes representativos.

La Tabla 7.1 resume los resultados, evidenciando el *trade-off* entre la precisión local y la eficiencia global.

## CAPÍTULO 7 : RESULTADOS

---

**Tabla 7.1:** Comparativa de Desempeño y Eficiencia

| Modelo               | MASE  | sMAPE   | Tiempo Entrenamiento |
|----------------------|-------|---------|----------------------|
| SARIMA (Local)       | 0.608 | 23.45 % | 23 min               |
| Transformer (Global) | 0.660 | 26.70 % | 8 min                |

**Fuente:** Elaboración propia.

En primer lugar, respecto a la Precisión Absoluta, el modelo SARIMA obtuvo una leve ventaja en el error escalado, obteniendo un MASE 0.608 frente a 0.660 del modelo Transformer. Esta diferencia se atribuye a la naturaleza conservadora del algoritmo, ya que, ante series con alto ruido, el modelo estadístico tiende a la regresión a la media. Aunque esta estrategia minimiza el error cuadrático, conlleva un aplanamiento de la dinámica del pronóstico.

En lo referente a la Eficiencia Computacional, el resultado es inverso. El entrenamiento del modelo global fue casi tres veces más rápido que el ajuste secuencial de SARIMA. Esta marcada ventaja de eficiencia es crítica para la viabilidad de un sistema nacional en tiempo real, ya que el Transformer escala mucho mejor al universo completo de cuadrantes ( $N = 1147$ ) que el enfoque local de SARIMA.

Para complementar el análisis de promedios, la Figura 7.1 desagrega el desempeño punto a punto para 100 de las series evaluadas.

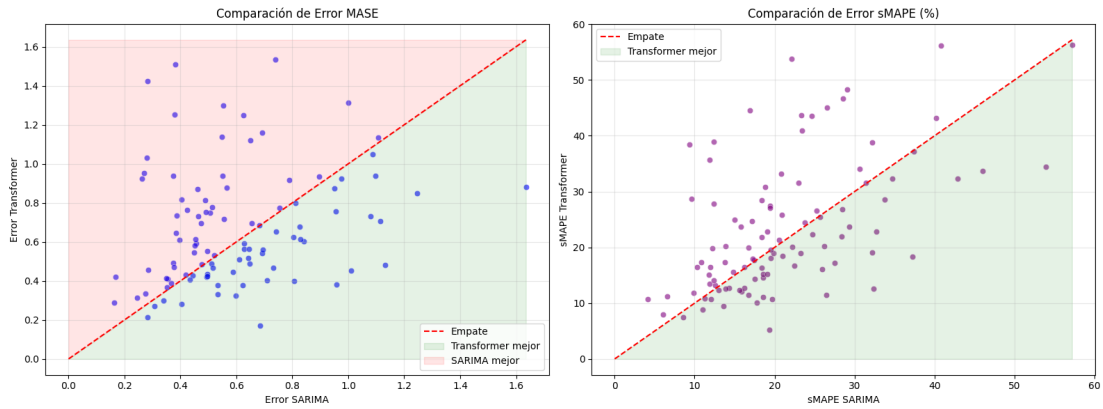


**Figura 7.1:** *Dispersión de Error Punto a Punto (MASE y sMAPE). Cada punto representa un cuadrante. La línea diagonal roja indica paridad de desempeño ( $y = x$ ). Los puntos situados en la zona verde (debajo de la diagonal) corresponden a series donde el Transformer obtuvo un menor error que el benchmark, evidenciando que la ventaja de SARIMA no es universal, sino dependiente de la dinámica específica de cada serie.*

**Fuente:** *Elaboración propia.*

La dispersión visual revela una heterogeneidad importante: si bien el centro de masa de la distribución favorece ligeramente a SARIMA, representado por la “zona roja” (donde el modelo estadístico entregó una precisión superior), existe una densidad significativa de cuadrantes en la “zona verde” donde la arquitectura de Deep Learning logró un mayor ajuste. Esto sugiere que la elección del “mejor modelo” no es binaria, lo que motiva el análisis de sensibilidad presentado a continuación.

### 7.2. Análisis de Sensibilidad por Volumen de Datos

Para verificar la hipótesis de transferencia de aprendizaje, se desagregó el desempeño según la densidad histórica de datos (Figura 7.2).

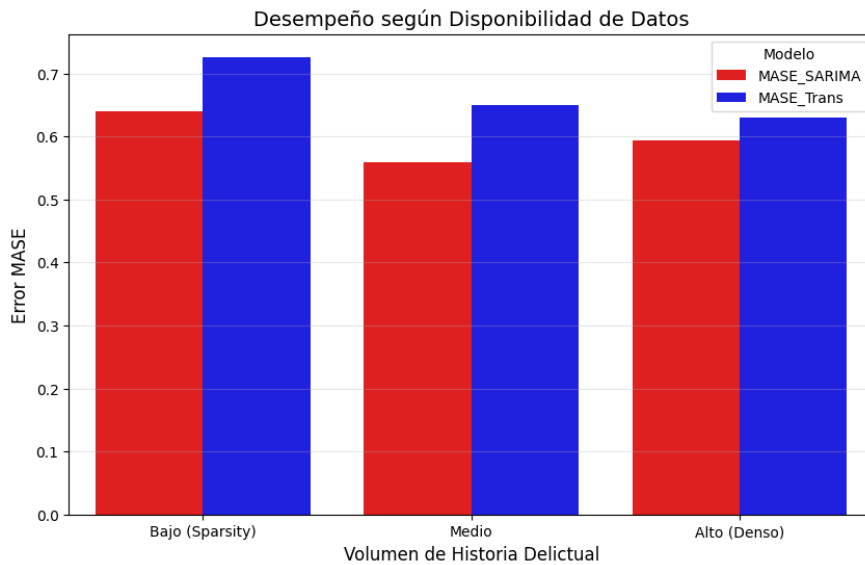


**Figura 7.2:** Desempeño MASE según disponibilidad de datos. La brecha de error se reduce significativamente en las zonas de alta densidad, demostrando que la red neuronal converge al desempeño del experto local cuando dispone de información suficiente.

**Fuente:** Elaboración propia.

Se observó que en zonas de *sparsity*, SARIMA minimiza el error prediciendo tendencias nulas. El Transformer, al intentar transferir patrones, genera falsos positivos que elevan su MASE pero mantienen al sistema en alerta.

### 7.3. Evaluación de Utilidad Operativa en Seguridad

Más allá del error estadístico, se evaluó la utilidad táctica del modelo mediante métricas orientadas a la gestión del riesgo y la seguridad ciudadana. En este contexto, la capacidad de anticipar correctamente la evolución de la criminalidad (si aumenta o disminuye) y la detección oportuna de eventos críticos son prioritarias frente a la precisión decimal del conteo.

La Tabla 7.2 presenta los indicadores de desempeño dinámico:

*Tabla 7.2: Métricas de Dinámica y Seguridad Preventiva*

| Indicador                        | SARIMA | Transformer |
|----------------------------------|--------|-------------|
| Acierto Tendencia (MDA)          | 38.0 % | 57.8 %      |
| Recall de Picos                  | 20.8 % | 36.6 %      |
| Volatilidad ( $\sigma$ predicha) | 0.003  | 0.008       |

*Fuente: Elaboración propia.*

### 7.3.1. Impacto en la Gestión de Riesgos

El análisis de estos indicadores revela diferencias funcionales críticas para la implementación del sistema:

1. **Anticipación de Tendencias:** El Transformer logra predecir la dirección de la evolución delictual en un 57.8 % de los casos, superando el azar. En contraste, SARIMA, con un 38 %, resulta ineficaz para determinar si un sector se volverá más seguro o inseguro en el futuro próximo, limitando su utilidad para la planificación preventiva.
2. **Sensibilidad ante Alertas:** El modelo propuesto es casi 2 veces más sensible detectando picos de actividad criminal en el 36.6 % de las ocasiones en contraposición al 20.8 % de SARIMA. Esta característica lo valida como una herramienta superior para sistemas de alerta temprana, donde el costo de un falso negativo puede afectar la seguridad de los usuarios o sus inversiones inmobiliarias.

### 7.4. Calidad probabilística de ambos modelos

Finalmente, se realizó una evaluación de la calidad probabilística de los modelos. En la gestión de riesgos, es crítico que el sistema no solo minimice el error, sino que

## CAPÍTULO 7 : RESULTADOS

---

entregue intervalos de confianza honestos y sea capaz de detectar anomalías.

Los resultados cuantitativos de las métricas avanzadas se presentan en la Tabla 7.3.

**Tabla 7.3:** Comparación de Calibración Probabilística y Detección de Eventos

| Métrica               | SARIMA | Transformer |
|-----------------------|--------|-------------|
| Cobertura 90 % (PICP) | 0.417  | 0.835       |
| Ancho Intervalo       | 0.052  | 0.141       |
| Recall de Picos       | 0.208  | 0.366       |

*Fuente:* Elaboración propia.

### 7.4.1. Discusión Estadística

El análisis de estas métricas revela deficiencias estructurales en el benchmark que no son visibles al utilizar como métrica MASE y sMAPE solamente:

- Sobreconfianza en SARIMA:** Aunque el modelo estadístico genera intervalos de confianza muy precisos, con un ancho estrecho de 0.052, su cobertura del 41.7 %, en lugar del 90 % teórico esperable indica un problema severo de sobreconfianza. En la práctica, esto significa que más de la mitad de las veces la realidad cae fuera de los márgenes de seguridad predichos, lo cual es inaceptable para una herramienta de gestión de riesgo.
- Robustez Probabilística del Transformer:** El modelo propuesto alcanza una cobertura del 83.5 %, demostrando una calibración superior. Al utilizar una distribución *t de Student*, el modelo reconoce su propia incertidumbre entregando intervalos más anchos cuando la situación lo amerita, garantizando que el usuario tenga una estimación realista del riesgo.

3. **Sensibilidad ante Amenazas:** Con un Recall de picos del 36.6 %, en comparación con el 20.8 % del benchmark, el Transformer demuestra ser una herramienta mucho más sensible para la detección de anomalías. Mientras SARIMA tiende a suavizar los picos, ya que su estrategia es la regresión a la media, el Transformer logra identificar y alertar sobre una mayor proporción de eventos de alta criminalidad.

Si bien SARIMA ofrece un error absoluto marginalmente menor al privilegiar la estabilidad en series tranquilas, su incapacidad para anticipar cambios de dirección y detectar picos limita severamente su utilidad operativa. En contraste, el Time Series Transformer se valida como el modelo más idóneo para el sistema, ya que equilibra una precisión competitiva con una sensibilidad crítica ante riesgos y una estimación de incertidumbre honesta. Estas capacidades maximizan la captura de tendencias y la alerta temprana, atributos indispensables para la toma de decisiones en el contexto de la seguridad ciudadana.

### 7.5. Limitaciones Encontradas

Es importante transparentar que el desarrollo de modelos de valoración inmobiliaria (predicción de precios) no pudo completarse al mismo nivel que los de seguridad. La falta de una base de datos histórica profunda de precios de venta (más allá de la oferta actual recolectada en tiempo real) impidió entrenar modelos de regresión robustos. Este componente quedó en fase de diseño conceptual y recolección de datos para trabajo futuro.

### 8. Conclusiones

El presente trabajo de título abordó el desafío de transformar datos heterogéneos y dispersos de seguridad ciudadana en inteligencia accionable, mediante el diseño e implementación de un Módulo de Procesamiento (PD) y un Motor Analítico (MA) basados en Deep Learning. La investigación permitió validar que la aplicación de arquitecturas globales de redes neuronales, específicamente el *Time Series Transformer*, constituye una solución superior en términos de escalabilidad y utilidad operativa frente a los métodos estadísticos tradicionales para este dominio.

En primer lugar, la caracterización cuantitativa de los datos reveló una complejidad estructural crítica: una distribución de “cola larga” donde el 10 % de los cuadrantes concentra el 40 % de la actividad delictual, y un desbalance de clases de 175:1 entre delitos frecuentes y graves. Estos hallazgos validaron la necesidad de implementar el *Índice de Crimen Normalizado*, una métrica que logró corregir el sesgo de frecuencia para visibilizar amenazas críticas en zonas de menor densidad, superando las limitaciones de los conteos simples utilizados en las plataformas actuales.

En segundo lugar, la evaluación comparativa del modelo predictivo arrojó resultados concluyentes sobre el *trade-off* entre precisión estática y dinámica. Si bien el benchmark SARIMA obtuvo un error absoluto marginalmente menor, con un MASE 0.608 contra 0.660 del Times Series Transformer, gracias a su tendencia conservadora, su incapacidad para anticipar cambios de tendencia, con un bajo MDA de 38 %, lo invalida como herramienta de alerta temprana. Por el contrario, el Transformer demostró superioridad táctica: Primero, en su capacidad de anticipación, logrando un acierto direccional del 57.8 %, lo que permite identificar correctamente la evolución futura de la seguridad barrial, un atributo fundamental para la toma de decisiones de inversión. Segundo, en su sensibilidad ante riesgos, donde su tasa de *Recall* de picos del 36.6 % casi duplicó la del benchmark, que obtuvo un 20.8 %, validando que la distribución de *salida t de Student* modela eficazmente los eventos extremos que los enfoques gaussianos tienden a ignorar. Finalmente, la escalabilidad operacional quedó demostrada mediante

## CAPÍTULO 8 : CONCLUSIONES

---

un entrenamiento global completado en solo 8 minutos, un proceso casi tres veces más rápido que el ajuste secuencial de SARIMA, confirmando la viabilidad técnica para un despliegue nacional con costos de mantenimiento sostenibles.

No obstante, el proyecto enfrentó limitaciones en la dimensión de valoración inmobiliaria. La falta de acceso a históricos de precios inmobiliarios impidió replicar la profundidad del análisis predictivo en la variable de precios.

Como trabajo futuro, se propone enriquecer el modelo Transformer incorporando variables dinámicas (indicadores macroeconómicos, eventos deportivos, etc) para reducir los falsos positivos en zonas de baja densidad. Asimismo, la integración de mecanismos de *Ensemble Learning* podría combinar la estabilidad de SARIMA en periodos calmos con la sensibilidad del Transformer ante anomalías, optimizando el desempeño global del sistema.

En definitiva, este trabajo aporta la lógica algorítmica necesaria para transitar desde la mera visualización de estadísticas pasadas hacia la anticipación de escenarios futuros, dotando al sistema de una capacidad analítica real para apoyar la toma de decisiones en el complejo entorno de la seguridad urbana.

## Referencias

- Barreto, U., Abarca, Y., & Pellicer, E. (2024). Key factors in the acquisition of residential properties: A comprehensive study of the global real estate market. *Revista De La Construcción. Journal of Construction*, 23(3), 623-638. <https://doi.org/10.7764/RDLC.23.3.623>
- Carabineros de Chile. (2025). Sistema Táctico de Operación Policial (S.T.O.P) [Consultado el 17 de noviembre de 2025]. <https://stop.carabineros.cl/>
- Ceccato, V., & Wilhelmsson, M. (2020). Do crime hot spots affect housing prices? *Nordic Journal of Criminology*, 21(1), 84-102. <https://doi.org/10.1080/2578983X.2019.1662595>
- Chisari, O. O., Ramos, M. P., & León, S. (2019). Crimen y configuración urbana: una evaluación de precios de propiedades y bienestar en equilibrio general para la ciudad de Buenos Aires. *El trimestre económico*, 86(342), 437-466. <https://doi.org/10.20430/ete.v86i342.774>
- Cuevas Ahumada, V. M. (2025). Perception of Public Insecurity and Housing Prices in Mexico. *Investigación Económica*, 84(331), 65-91. <https://doi.org/10.22201/fe.01851667p.2025.331.90270>
- Hugging Face. (2022). Time Series Transformer: Model Documentation [Versión consultada: Noviembre, 2025]. [https://huggingface.co/docs/transformers/model\\_doc/time\\_series\\_transformer](https://huggingface.co/docs/transformers/model_doc/time_series_transformer)
- Ministerio del Interior y Seguridad Pública. (2025). Estadísticas delictuales - Centro de Estudios y Análisis del Delito (CEAD) [Consultado el 17 de noviembre de 2025]. <https://cead.minsegpublica.gob.cl>
- Portal Inmobiliario. (2025). Análisis de mercado inmobiliario en Chile [Consultado el 17 de noviembre de 2025]. <https://www.portalinmobiliario.com/>
- Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM. *Procedia Computer Science*, 181, 526-534. <https://doi.org/10.1016/j.procs.2021.01.199>

The Omega Group. (2025). Interactive Crime Maps - CrimeMapping.com [Consultado el 17 de noviembre de 2025]. <https://www.crimemapping.com>

Vergara-Perucich, J. F. (2021). Determinantes urbanos del precio de la vivienda en Chile: una exploración estadística. *Urbano*, 24(43), 40-51. <https://doi.org/10.22320/07183607.2021.24.43.04>