Medellín Seguro: Predicción Inteligente del número de hurtos a personas con algoritmos basados en Series Temporales.

Safe Medellin: Smart prediction of the number of thefts from people with algorithms based on time series

Cindy Paola Guerra Medina
cpguerram@eafit.edu.co

TRABAJO DE GRADO

Asesor:
Nicolas Alberto Moreno Reyes
namorenor@eafit.edu.co

UNIVERSIDAD EAFIT
Escuela de Administración
Maestría en ciencias de los datos y la analítica
Medellín
2024

Tabla de contenido

1 INTRODUCCIÓN 7

1.1 Planteamiento del problema 7
1.2 Justificación 8
1.3 Objetivos 9
1.3.1 Objetivo general 9
1.3.2 Objetivos específicos 9

2 MARCO TEÓRICO Y ESTADO DEL ARTE 10

2.1 Hurto 10
2.2 Forecasting 11
2.3 Algoritmos de series temporales 11
2.3.1 Suavizado exponencial 12
2.3.2 Técnica de series temporales ARIMA 14
2.3.3 Técnica de series temporales SARIMA 17
2.4 Técnica de escalamiento lineal (LST) 18
2.5 Medata (Estrategia de datos de Medellín) 19
2.6 IBM SPSS Modeler 19

3 DATOS 20

3.1 Plan de Gestión de Datos 20
3.2 Adquisición de datos 20
3.3 Descripción y análisis preliminar de los datos 21
3.4 Preprocesamiento de los datos 23
3.5 Aspectos éticos 23

4 DESARROLLO DE MODELOS 23

Fase I. Comprensión del negocio 24
Fase II. Entendimiento de los datos 27
Fase III. Preparación de los datos 28
Fase IV. Modelado 32
Fase V. Evaluación de resultados 35
Score del Modelo para selección de mejor modelo 36
Fase VI. Despliegue de resultados 38

5 ANÁLISIS DE RESULTADOS 39

6 CONCLUSIONES Y TRABAJO FUTURO 46

7 REFERENCIAS 49

Ilustraciones y Tablas

Ilustración 1 IBM SPSS Modeler 19
Ilustración 2 Vista de los 10 primeros registros 21
Ilustración 3 Datos Portal Metada 21
Ilustración 4 Auditoria de datos 22
Ilustración 5 Estadísticas descriptivas de la información 22
Ilustración 6 Ciclo de vida de minería de datos (IBM, 2020) 24
Ilustración 7 Flujo de automatización en IBM SPSS Modeler 26
Ilustración 8 Auditoria de datos 27
Ilustración 9 Densidad histórica de hurtos en Medellín 28
Ilustración 10 Hurtos sin dato del campo barrio y comuna 29
Ilustración 11 Cantidad de hurtos de personas por comuna 29
Ilustración 12 Activo analítico de preparación de datos IBM SPSS Modeler 32
Ilustración 13 Almacenamiento temporal de Modelos 33
Ilustración 14 Supernodo modelos para series de 1 desviación 33
Ilustración 15 Ejecución de competencia de modelos 34
Ilustración 16 Almacenamiento de parámetros de cada Modelos 35
Ilustración 17 . Ruta para la selección del modelo 38
Ilustración 18 Cálculo de score para cada modelo..... 38
Ilustración 19 . Resultado pronostico por código de barrio 39
Ilustración 20 Pronóstico para la ciudad de Medellín 40
Ilustración 21 Diagrama de cajas con atipicidades de dinero por medio de trasporte realizado el hurto 41
Ilustración 22 Mapas de calor del valor hurtado por las comunas de la ciudad 41
Ilustración 23 Valor total y promedio de dinero Hurtado a personas 42
Ilustración 24 Pronóstico de hurtos poblado 43
Ilustración 25 Ubicación geográfica del barrio poblado 43
Ilustración 26 Pronóstico de hurtos barrio poblado. 44
Ilustración 27 Pronóstico para el barrio candelaria 45
Ilustración 28 Pronóstico por barrios de Medellín mes diciembre 2024 46
Ilustración 29 Pronóstico por comunas de Medellín 47
Tabla 1 Cantidad de series originales 30
Tabla 2 Cantidad de series con desviaciones..... 30
Tabla 3 Valores Back-Testing 35
Tabla 4 Interpretación de parámetros de validación 36
Tabla 5 Importancia o peso del parámetro..... 37
Tabla 6 Pronóstico comuna Poblado 43
Tabla 7 Pronóstico barrio Poblado 44

Tabla 8 Pronóstico barrio la Candelaria 44
Tabla 9 Top 10 de los barrios con mayor predicción para diciembre 2024..... 45

RESUMEN

En la actualidad, estamos inmersos en la revolución de los datos, una era caracterizada por la importancia de comprender los sucesos pasados para predecir el futuro, y a partir de estos, apoyar las estrategias que faciliten la toma de decisiones de forma anticipada. En este contexto, Colombia enfrenta importantes retos en materia de seguridad y convivencia, desafíos que pueden ser abordados o estimados mediante el análisis de datos; en Medellín, el portal de datos abiertos Medata (medata.gov.co), permite el acceso a estadísticas históricas y descriptivas sobre la incidencia de delitos a personas como el hurto; el cual es un delito recurrente que afecta la seguridad, calidad de vida y economía de los ciudadanos.

Este proyecto plantea la utilización de algoritmos de series temporales incorporados en la plataforma IBM SPSS Modeler, una herramienta robusta y flexible que facilita la programación de competencia de modelos predictivos (IBM, 2023, IBM SPSS Modeler). A través de su capacidad para identificar patrones, tendencias y estacionalidad en los datos históricos, se busca estimar la incidencia futura de hurto a personas en la ciudad de Medellín, desagregando los análisis a nivel de comunas y barrios. Las proyecciones se realizarán de forma mensual para los meses de octubre, noviembre y diciembre del 2024, los cuales servirán como insumo para la planificación de estrategias preventivas de seguridad que contribuyan a la priorización de las zonas que requieren mayor atención y optimizar los recursos disponibles que minimicen los impactos negativos del delito y generen una mayor sensación de tranquilidad y confianza en la ciudadanía.

Palabras clave: Pronostico, Series de tiempo, automático, recalibración, hurtos, Medellín y SPSS.

ABSTRACT

Today, we are immersed in the data revolution, an era characterized by the importance of understanding past events to predict the future, and from these, support strategies that facilitate decision-making in advance. In this context, Colombia faces important challenges in terms of security and coexistence, challenges that can be addressed or estimated through data analysis; in Medellín, the open data portal Medata (medata.gov.co), allows access to historical and descriptive statistics on the incidence of crimes against persons such as theft; which is a recurring crime that affects the security, quality of life and economy of citizens.

This project proposes the use of time series algorithms implemented in the IBM SPSS Modeler platform, a robust and flexible tool that facilitates the programming of predictive model competition (IBM, 2023, SPSS Modeler. Through its ability to identify patterns, trends and seasonality in historical data, it seeks to estimate the future incidence of theft from persons in the city of Medellín, disaggregating the analysis at the level of communes and neighborhoods. The projections will be made on a monthly basis for the months of October, November and December 2024, which will serve as input for the planning of preventive security strategies that contribute to the prioritization of areas that require greater attention and optimize available resources that minimize the negative impacts of crime and generate a greater sense of tranquility and confidence in citizens.

Keywords: Forecast, Time series, automatic, recalibration, thefts, Medellín and SPSS.

1 INTRODUCCIÓN

La seguridad ciudadana es un factor clave en la calidad de vida y bienestar de una sociedad, ya que influye en la confianza de las personas y en el desarrollo económico de una ciudad. En Medellín, el hurto a personas representa un desafío constante que afecta no solo a las víctimas directas, sino también a la percepción general de realizar actividades con tranquilidad y confianza en el espacio público o fuera de sus hogares. Para enfrentar este problema, es fundamental contar con herramientas que permitan prever su evolución y tomar medidas preventivas que mitiguen el comportamiento delictivo.

En la era de los datos, la capacidad de analizar y pronosticar tendencias delictivas es cada vez más accesible, a través de modelos de machine learning como las series temporales, es posible identificar patrones y proyectar la evolución de los delitos, permitiendo a las autoridades anticiparse a los focos de inseguridad y optimizar el uso de recursos en la prevención del crimen. Este estudio propone el uso de modelos analíticos disponibles en la plataforma de minería de datos IBM SPSS Modeler para estimar el comportamiento futuro para los meses de octubre, noviembre y diciembre de 2024 del hurto a personas en Medellín, tomando como base los datos históricos disponibles en fuentes oficiales de la ciudad Medata.

Este trabajo se enmarca en prever información clave para la planificación de estrategias preventivas en seguridad, facilitar la priorización de zonas vulnerables de la ciudad y contribuir a la reducción de este delito; para la generación de un impacto positivo en la calidad de vida de los ciudadanos, a través de un enfoque innovador basado en datos.

1.1 Planteamiento del problema

La página de “Medellín en Cifras” presenta el histórico de hurtos ocurridos en la ciudad mediante tableros de control interactivos, ofreciendo datos anuales y mensuales que permiten un análisis descriptivo de la problemática; la administración distrital y los entes de seguridad informan una reducción histórica del 20% del número de hurtos reportados para el 2024 con relación al 2023. (Dúber Cano Aguirre, 2024, La lucha contra el hurto en Medellín da resultados contundentes con reducción histórica del 20 %)

A través de la herramienta de IBM SPSS Modeler se plantea el uso de modelos analítico con técnicas de machine learning, que permitirán la estimación mensual de la cantidad de hurtos para los meses de octubre noviembre y diciembre del 2024; donde los resultados proporcionan información clave para la fuerza pública y administrativa de la ciudad de Medellin, los valores pronosticados podrán apalancar estrategias de mitigación, contribuyendo a la reducción del número de casos y fortalecer las acciones preventivas en las zonas más vulnerables de la ciudad, comunas y barrios de Medellín.

En la política pública de seguridad y convivencia del municipio de Medellín acuerdo 21 del 2015, se destaca la frase seguridad ciudadana como el pilar de la seguridad humana para la protección de los derechos humanos. Dentro de este marco, se establecen cuatro categorías de análisis en el programa de Desarrollo Sostenible, enfocadas en indicadores clave de seguridad como denuncias de hurtos en vía pública, hurtos de motos y vehículos, robos en viviendas, y hurtos a establecimientos comerciales y financieros, los cuales serán el centro de seguimiento y monitoreo para las administraciones distrital. (Política Pública de Seguridad y Convivencia del Municipio de Medellín, 2015)

Dada la necesidad de prever la evolución de estos delitos, resulta fundamental contar con la estimación anticipada a través de técnicas de analítica predictiva con un enfoque de competencia de modelos de series temporales, el cual utilizará una automatización de técnicas de suavizamiento exponencial y autorregresivas de medias móviles ARIMA y SARIMA, que garanticen la confiabilidad en las proyecciones.

La automatización de la competencia de modelos se desarrollará a través de un flujo programado que selecciona en IBM SPSS Modeler la técnica adecuada para cada estimación después de validar los parámetros automatizados para la selección de la técnica. Lo anterior permite de forma rápida la recalibración de la técnica o modelo de series temporales para prever los cambios significativos en la eficiencia del pronóstico de las tendencias del delito y así anticipar el comportamiento del número de hurtos mensual en la ciudad de Medellín, sus comunas y barrios.

1.2 Justificación

Con una inversión superior a los 3.000 millones de pesos provenientes del Fondo de Ciencia, Tecnología e Innovación (FCTeI) del Sistema General de Regalías (SGR), se impulsa el desarrollo de proyectos analíticos enfocados en la predicción de crímenes en Bogotá. Este plan tiene como objetivo principal reducir los niveles de inseguridad en la capital, abordando los delitos de mayor impacto para la ciudadanía, entre los que se destacan el homicidio, las riñas con lesiones personales y los hurtos con uso de violencia. (Modelos de predicción: La seguridad, un nuevo reto de la tecnología, 2024)

A través del análisis de datos se traza una solución que apoye las estrategias para mitigar la inseguridad de una ciudad o país, los hurtos a personas es uno de los problemas críticos que afecta la confianza ciudadana, por lo que este proyecto utilizará herramientas de vanguardia que permite el uso de algoritmos analíticos con la flexibilidad de automatizar la competencia de modelos de series temporales. Una de las experiencias más representativas del uso de las técnicas de series temporales seleccionadas para el proyecto se utilizó en conjunto con coordenadas espaciales para aislar futuros delitos en la ciudad de Chicago de EEUU, donde se implementaron sistemas predictivos para dividir la ciudad en mosaicos espaciales de aproximadamente 300 metros de ancho y predecir el crimen dentro de estas áreas,

identificando positivamente patrones que permitieron una adecuada distribución de la fuerza policial en las zonas proyectas que mitigaron el aumento del número de crímenes. (Sara Sendino, 2022, Un algoritmo predice los crímenes que van a suceder con una semana de antelación).

Para la implementación de la solución se plantea la creación de un activo analítico (ruta) en la herramienta IBM SPSS Modeler que incorpore competencia de modelos disponibles para suavizamiento exponencial y autorregresivas de medias móviles ARIMA y SARIMA (IBM, 2021, documento de modelos de series temporales), el cual permitirá la creación de un proceso replicable de análisis de datos para pronóstico.
Los resultados de las proyecciones se visualizarán a nivel geográfico, permitiendo la priorización de las zonas más vulnerables o con mayor riesgo, diseñando estrategias que promuevan la optimización de recursos para una cobertura adecuada. La solución analítica permitirá disminuir la subjetividad e Incorporar a las políticas públicas los pronósticos para la gestión de la seguridad ciudadana de Medellín; permitiendo medir la efectividad de las estrategias implementadas, que impulsen una transformación de las autoridades para minimizar riesgos y mejorar la percepción de seguridad y fortalece la confianza de los ciudadanos.

La solución analítica o activo analítico a construir, busca implementar un flujo de proceso automático que permita parametrizar y replicar el desarrollo a diversas industrias o campos como pronóstico de la demanda, personal, número de quejas, cantidad de soportes, etc, cuando la variable a estimar sea numérica y univariada.

1.3 Objetivos

1.3.1 Objetivo general

Implementar un activo analítico o ruta que pronostique el número de hurtos a personas en la ciudad de Medellín para tres meses (octubre, noviembre y diciembre 2024) futuros, con un nivel de detalle que abarque total de ciudad, comunas y barrios principales; utilizando competencia de modelos de series temporales que permitan la recalibración automática en IBM SPSS Modeler, ante posibles cambios en la tendencia de los hurtos.

1.3.2 Objetivos específicos

- Construir rutas o activos analíticos en IBM SPSS Modeler que contenga la lógica de la metodología CRISP DM.
- Descomponer cada serie de tiempo o granularidad que se genere en: 1 desviación estándar, 2 desviaciones y tres.
- Automatizar la competencia de modelos para los algoritmos de suavizamiento exponencial, ARIMA y SARIMA para cada serie temporal.

- Crear un score de validación y selección del mejor modelo predictivo para cada serie de tiempo.
- Construir un indicador de backtesting para cada modelo seleccionado como el ganador.
- Generar una visual con los resultados de predicción que facilite el análisis para la toma de decisiones.

2 MARCO TEÓRICO Y ESTADO DEL ARTE

2.1 Hurto

El Artículo 239 del Código Penal colombiano establece que el delito de hurto se da cuando una persona se apodera de una cosa/mueble ajeno sin el consentimiento de su dueño y con la intención de obtener un provecho para sí o para un tercero. Este delito abarca desde la sustracción de objetos de poco valor, hasta casos más complejos como el hurto calificado, donde intervienen circunstancias agravantes que incrementan la pena. El hurto a personas es uno de los delitos más frecuentes en contextos urbanos y puede ser un indicador clave de la percepción de inseguridad en una ciudad. Analizar su comportamiento permite diseñar estrategias de prevención que reduzcan su incidencia y mejoren la confianza ciudadana. (LEY 2197 de 2022, Poder público – Rama legislativa, 25 de enero de 2022)

En Londres, se implementó un proyecto liderado por el departamento de criminología de la universidad de Cambridge, donde a través de métodos de series temporales como Clustering de serie temporal, identifica las ubicaciones de un cubo de espacio-tiempo que son más similares y las divide en clústeres distintos; se utilizaron para predecir delitos menores como hurtos en áreas comerciales; identificando periodos de tiempo críticos para la implementación de refuerzo de seguridad en áreas comerciales. (Brit. J. Criminol, 2004)

Un caso cercano de este estudio se llevó a cabo en la ciudad de Bogotá, donde la universidad de los Andes, a través de la facultad de economía, presentó avances del proyecto “Modelos Matemáticos, la clave para predecir el crimen en Bogotá” Este trabajo utiliza técnicas de analítica predictiva e inteligencia artificial para estimar la probabilidad de ocurrencia de un crimen en general, aunque no aborda directamente la cantidad o número de casos que se pueden presentar. Durante el desarrollo del proyecto se destacó la necesidad de complementar los análisis con modelos que incorporen la variable temporal o tiempo, con el fin de tener una visual a nivel geográfica. Durante la etapa de modelamiento identificaron correlaciones con factores como el día de la semana, zonas específicas, estrato socioeconómico y características del entorno geográfico. Además, se identificó que el evento tipificado como riña, tienden a ser un factor recurrente que afecta la ocurrencia de un crimen. Este enfoque permitió localizar puntos calientes o zonas prioritarias para atención y seguimiento

policial, optimizando el análisis de cámaras de seguridad y planificar estratégicamente la ubicación de CAI en lugares clave. (Prediciendo el crimen en Bogotá, 2020)

Un enfoque similar con componentes espacio temporal, fue utilizado para la predicción de delitos en la ciudad de Buenos Aires (CABA), en el estudio se realizó una comparación de modelos predictivos, los modelos de series de tiempo ARIMA que incluyen únicamente el recuento histórico de delitos y la técnica de aprendizaje automático XGBoost; el XGBoost mostró un rendimiento superior en la eficiencia de las predicciones con relación a los modelos de series temporales, pero el modelo no permitió verificar las contribuciones de factores espaciales existentes, que si permitió los modelos de series temporales con las variables de intervención. (Rafael zambrano, 2021)

2.2 Forecasting

Consiste en la estimación de la demanda futura de un producto, servicio o dato, que se desee conocer de forma anticipada o futura. Para ello se utilizarán los históricos de la variable numérica recolectada en una periodicidad o tiempo especifica, sobre este histórico se aplican diferentes metodologías o técnicas de series de tiempo para entender el comportamiento pasado y con este estimar el futuro. Para la implementación de técnicas de forecasting, es importante entender que es una serie de tiempo, la cual consiste en un conjunto de datos provenientes de realizaciones de una variable aleatoria que se han recolectado sucesivamente en el tiempo (Peña, 1990; Peña, 2010).

$$Y_t = \{Y_1, Y_2, Y_3, \ldots, Y_t, \ldots\}$$

Y_t es la serie de tiempo que contiene el conjunto de observaciones que toma la variable (cuantitativa) en diferentes momentos del tiempo. Los valores de la variable se deben recolectar en espacios de tiempo o periodos iguales y los datos deben tener un orden secuencial o cronológico, lo anterior significa que en la herramienta analítica los datos de cada serie de tiempo estarán en columnas (Peña, 1990; Peña, 2010).

El resultado de una serie temporal se llama pronostico, el cual es la estimación a una situación de incertidumbre. El término predicción es similar, pero más general, y usualmente se refiere al resultado o estimación de los algoritmos estadísticos de series temporales, adicional a este resultado se estiman los errores del pronóstico, el cual es el valor absoluto o porcentual del pronóstico con respecto al valor real en cada periodo de evaluación o entrenado (Peña, 1990; Peña, 2010).

2.3 Algoritmos de series temporales

El objetivo del análisis de series temporal es elaborar un modelo estadístico que identifique la ecuación que describa el comportamiento histórico de la variable a analizar en este caso el número de hurtos a personas, a continuación se describen las técnicas de suavizamiento exponencial y autorregresivas de medias móviles ARIMA y

SARIMA, que se utilizaran durante la autorización de la competencia de modelos en IBM SPSS Modeler (IBM, 2021, documento de modelos de series temporales):

2.3.1 Suavizado exponencial

Son métodos o algoritmos que utiliza los valores ponderados de las observaciones anteriores de la serie para predecir los valores futuros, corrigiendo las predicciones a medida que entran nuevos datos. Los suavizado exponenciales son útiles para datos que muestran una tendencia o estacionalidad en su histórico de datos; a continuación se relacionan los modelos más representativos de la herramienta IBM SPSS Modeler, y los cuales serán usados en la competencia de modelos del proyecto.

Simples: Este modelo es adecuado para las series sin tendencia ni estacionalidad; y su principal característica es asignar más peso a los datos recientes que a los pasados, logrando una representación más ajustada a cambios recientes en la serie. En este ca so los pesos que se le asignan a las observaciones van decayendo exponencial mente a medida que las observaciones viejas se alejan en el tiempo. (Julio Alonso, 2020)

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t.$$

$\hat{y}_{t+1}$**:** Pronóstico para el siguiente periodo $t+1$
y_t**:** Valor observado
$\hat{y}_t$**:** Pronóstico del periodo actual
α**:** Coeficiente de suavizamiento $(0 \leq \alpha: \leq 1)$

Tendencia lineal de Holt: Este modelo es adecuado para las series con una tendencia lineal y sin estacionalidad y es una extensión del modelo de suavizamiento exponencial simple, Este modelo es útil cuando los datos muestran una tendencia creciente o decreciente. La diferencia es que se tiene dos ecuaciones, una para el nivel L y otra para la tendencia T. (Julio Alonso, 2020)

Nivel: $L_t = \alpha y_t + (1-\alpha)\,(L_{t-1} + T_{t-1})$
Tendencia: $T_t = \beta(L_t - L_{t-1}) + \,(1-\,\beta)\,T_{t-1}$
Pronostico: $\hat{y}_t = \,L_t + \,\hbar T_t$

L_t: Nivel en el tiempo t según datos actual y la tendencia.

T_t :Tendencia en el tiempo t, que refleja el cambio promedio entre los periodos.
$\hat{y}_t$: Pronóstico para h periodos.
y_t: Valor observado en el tiempo t.
α: Parámetro de suavizamiento para el nivel $(0 \leq \alpha: \leq 1)$
β: Parámetro de suavizamiento para la tendencia $(0 \leq \beta \leq 1)$

Tendencia lineal de Brown: Metodología adecuada para las series con una tendencia lineal y sin estacionalidad; Este modelo utiliza dos niveles de suavizamiento exponencial para capturar tanto el nivel como la tendencia de la serie. El modelo de Brown es una extensión del suavizamiento exponencial simple que utiliza un segundo nivel de suavizamiento para estimar la tendencia. Esto permite que el modelo capture el nivel general como la dirección del cambio en la tendencia constante. (Miller Jimmy Alarcón, 2009)

Para este modelo se obtiene 5 ecuaciones que se describen a continuación:

Suavizamiento General: $S1 = \alpha y_t + (1-\alpha)\, S1_{t-1}$
Suavizamiento Tendencia: $S2 = S1_t + (1-\alpha)\, S2_{t-1}$
Nivel: $L_t = 2\, S1_t - S2_t$
Tendencia: $T_t = \frac{\alpha}{1-\alpha}(S1_t - S2_t)$
Pronóstico para $\hbar$ periodos: $\hat{y}_t = L_t + \hbar T_t$

Tendencia amortiguada: Este modelo es adecuado para las series con una tendencia lineal que va desapareciendo (decreciente) o cayendo y sin estacionalidad, los parámetros de suavizado relevantes son el nivel, la tendencia y la tendencia de amortiguación. El suavizado exponencial amortiguado es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación y dos órdenes de media móvil. (José Alberto Mauricio, 2007)

Nivel: $L_t = \alpha y_t + (1-\alpha)\,(L_{t-1} + \phi T_{t-1})$
Tendencia: $T_t = \beta(L_t - L_{t-1}) + (1-\beta)\phi\, T_{t-1}$
Pronostico: $\hat{y}_t = L_t + \frac{1-\phi^h}{1-\phi} T_t$

Estacional simple: Este modelo es adecuado para las series sin una tendencia y un efecto estacional constante a lo largo del tiempo. Esta metodología captura la estacionalidad sin requerir componentes adicionales como tendencias lineales. se ajusta a los cambios de temporada o ciclos recurrentes y realiza pronósticos basados en la media móvil ponderado de las observaciones más recientes. (Julio Alonso, 2020)

Nivel: $L_t = \alpha(y_t - S_{t-m}) + (1-\alpha)\,(L_{t-1} + S_{t-m})$
Estacional: $S_t = \lambda(y_t - L_t) + (1-\lambda)\, S_{t-m}$
Pronostico: $\hat{y}_t = L_t + S_{t+h-m}$

L_t: Nivel en el tiempo t.
S_t :Estimacion de estacionalidad para el periodo t.
$\hat{y}_t$: Pronóstico para h periodos.
y_t: Valor observado en el tiempo t.
λ: Coeficiente de estacionalidad $(0 \leq \lambda : \leq 1)$
m: Periodo estacional

Aditivo de Winters: Este modelo es adecuado para las series con una tendencia lineal y un efecto estacional constante a lo largo del tiempo. es una extensión del suavizamiento exponencial simple que incluye no solo el nivel, sino también la tendencia y la estacionalidad. (Julio Alonso, 2020)

$$\text{Nivel: } L_t = \alpha(y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$\text{Tendencia: } T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$
$$\text{Estacional: } S_t = \lambda(y_t - L_t) + (1-\lambda)S_{t-m}$$
$$\text{Pronostico: } \hat{y}_t = L_t + hT_t + S_{t+h-m}$$

Multiplicativo de Winters: Este modelo es adecuado para series en las que haya una tendencia lineal y con un efecto estacional que cambie en función de la magnitud de las series, se supone que los efectos estacionales no son constantes, sino que varían. La técnica de **multiplicativo** es más adecuada cuando las fluctuaciones estacionales son proporcionales al nivel de la serie, cuando la magnitud de la estacionalidad aumenta o disminuye conforme lo hace el nivel de la serie.
Las ecuaciones de este modelo incluyen tres componentes que se ajustan a los datos a través de parámetros de suavizamiento: α para el nivel, β la tendencia y λ la estacionalidad. (Miller Jimmy Alarcón, 2009)

$$\text{Nivel: } L_t = \alpha + \frac{y_t}{S_{t-m}} + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$\text{Tendencia: } T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$
$$\text{Estacional: } S_t = \lambda\frac{y_t}{L_t} + (1-\lambda)S_{t-m}$$
$$\text{Pronostico: } \hat{y}_t = (L_t + hT_t)S_{t+m}$$

L_t: Nivel en el tiempo t.
S_t :Estimacion de estacionalidad para el periodo t.
$\hat{y}_t$: Pronóstico para h periodos.
y_t: Valor observado en el tiempo t.
λ: Coeficiente de estacionalidad $(0 \leq \lambda: \leq 1)$
β: Parámetro de suavizamiento para la tendencia $(0 \leq \beta \leq 1)$

2.3.2 Técnica de series temporales ARIMA

Son técnica robusta y sofisticada que se usan para descomponer la tendencia y la estacionalidad. Esto implica la especificación del orden autorregresivo y la media móvil, adicional el grado de diferenciación, con lo anterior se pude obtener modelos que contengan componente AR, MA, ARMA y ARIMA. Los modelos autorregresivos Integrados de medias móviles (ARIMA) son una técnica robusta que combina dependencia lineal entre una observación y sus valores pasados, elimina la tendencias o estacionalidad y modela la relación entre una observación y los errores pasados. (Hyndman & Athanasopoulos, 2018)
Los modelos ARIMA proporcionan un buen enfoque para encontrar el pronóstico de

una serie de tiempo. El suavizado exponencial y los modelos ARIMA son las dos técnicas más utilizadas para el pronóstico de series de tiempo. Mientras que los modelos de suavizado exponencial se basan en una descripción de la tendencia y la estacionalidad de los datos, los modelos ARIMA pretenden describir las autocorrelaciones de los datos. (Hyndman & Athanasopoulos, 2018).

Para continuar debemos comprender el concepto de estacionariedad y la técnica de diferenciación de una serie de tiempo; una serie de tiempo estacionaria es aquella cuyas propiedades no dependen del tiempo en el que se observa la serie. Por lo anterior, las series de tiempo con tendencias, o con estacionalidad, no son estacionarias: la tendencia y la estacionalidad afectarán el valor de la serie de tiempo en diferentes momentos. Por otro lado, una serie de ruido blanco es estacionaria: no importa cuándo la observes, debería verse igual en cualquier momento (Hyndman & Athanasopoulos, 2018).

Algunos casos pueden ser confusos: una serie temporal con un comportamiento cíclico (pero sin tendencia ni estacionalidad) es estacionaria. Esto se debe a que los ciclos no tienen una duración fija, la serie no puede tener los picos y valles de los ciclos. (Hyndman & Athanasopoulos, 2018)

La serie diferenciada es el cambio entre observaciones consecutivas en la serie original y se puede escribir como $y't = yt - yt - 1,$ La serie diferenciada tendrá sólo $T - 1$ valores, ya que no es posible calcular una diferencia $y'1$ para la primera observación. Cuando la serie diferenciada es ruido blanco, el modelo de la serie original se puede escribir como $yt - yt - 1 = \varepsilon t,$ dónde εt denota ruido blanco. Reorganizar esto conduce al modelo de "caminata aleatoria"

$$yt = yt - 1 + \varepsilon t.$$

Los pronósticos de un modelo de caminata aleatoria son iguales a la última observación, debido a que los movimientos futuros son impredecibles y es igualmente probable que sean hacia arriba o hacia abajo. Un modelo estrechamente relacionado permite que las diferencias tengan una media distinta de cero. $yt = C + yt - 1 + \varepsilon t.$

El valor de C es el promedio de los cambios entre observaciones consecutivas. Si C es positivo, entonces el cambio promedio es un aumento en el valor de yt. Por lo tanto, yt tenderá a desviarse hacia arriba. Sin embargo, si C es negativo, yt tiende a descender.

Una forma de determinar más objetivamente si se requiere la diferenciación es usar una prueba de raíz unitaria. Estas son pruebas de hipótesis estadísticas de estacionariedad que están diseñadas para determinar si se requiere diferenciación. La Prueba de Dickey-Fuller busca determinar la existencia o no de raíces unitarias en una serie de tiempo. La hipótesis nula de esta prueba es que existe una raíz unitaria en la serie, lo que significa que en un simple modelo auto regresivo de orden (1) $yt = pyt - 1 + \varepsilon t.$ (Gujarati and Porter, 2010)

Donde yt es la variable de interés, t es el índice de tiempo, ρ es un coeficiente, y εt es el termino de error. La raíz unitaria está presente si $\rho = 1$. En este caso, el modelo no sería estacionario. A continuación se describe el modelo de regresión:

$$\nabla yt = (p-1)yt - 1 + \varepsilon t = \delta yt - 1 + \varepsilon t$$

Donde ∇ es el operador de primera diferencia. Este modelo puede ser estimado y las pruebas para una raíz unitaria son equivalentes a pruebas $\delta = 0$ (donde $\delta = \rho - 1$). Dado que la prueba se realiza con los datos residuales en lugar de los datos en bruto, no es posible utilizar una distribución estándar para proporcionar valores críticos. Por lo tanto, esta estadística tiene una determinada distribución conocida como la tabla de Dickey-Fuller. (Gujarati and Porter, 2010)

Los modelos autorregresivos pronosticamos la variable de interés usando una combinación lineal de predictores. En un modelo de autorregresión, pronostica la variable de interés usando una combinación lineal de valores pasados de la variable. El término autoregresión indica que es un concepto de regresión de la variable contra sí misma. Así, un modelo autorregresivo de orden p Se puede escribir como

$$yt = C + \phi 1yt - 1 + \phi 2yt - 2 + \cdots + \phi pyt - p + \varepsilon t,$$

Dónde εt es ruido blanco. Esto es como una regresión múltiple, pero con valores rezagados de yt como predictores. Nos referimos como un $AR(p)$ modelo, un modelo autorregresivo de orden p.

Los modelos de promedio móvil usan valores pasados de la variable pronóstico, en una regresión, usa los errores de pronóstico pasados para un modelo parecido a una regresión.

$$yt = C + \varepsilon t + \theta 1\varepsilon t - 1 + \theta 2\varepsilon t - 2 + \cdots + \theta q\varepsilon t - q,$$

Dónde εt es ruido blanco. Nos referimos a esto como un $MA(q)$ modelo, un modelo de orden de media móvil q. Por supuesto, no *observamos* los valores de εt, por lo que no es una regresión. cada valor de yt se puede considerar como un promedio móvil ponderado de los últimos errores de pronóstico. (Hyndman & Athanasopoulos, 2018)

La diferenciación con autorregresión y un modelo de media móvil, obtenemos un modelo ARIMA no estacional. ARIMA es un acrónimo de AutoRegressive Integrated Moving Average (en este contexto, "integración" es lo contrario de diferenciación). El modelo completo se puede escribir como:

$$y't = C + \phi 1y't - 1 + \cdots + \phi py't - p + \theta 1\varepsilon t - 1 + \cdots + \theta q\varepsilon t - q + \varepsilon t,$$

Dónde $y't$ es la serie diferenciada (puede haber sido diferenciada más de una vez). Los "predictores" del lado derecho incluyen valores rezagados de yt y errores rezagados. A esto lo llamamos un $ARIMA(p, d, q)$ modelo, donde:

p: orden autorregresivo
d: grado de primera diferenciación
q orden de la media móvil.

2.3.3 Técnica de series temporales SARIMA

El modelo SARIMA, que significa "Seasonal Autoregressive Intergrated Moving Average," es una extensión del modelo ARIMA, que incluye componente estacional y son utilizados para series de tiempo que presentan comportamientos repetitivos en una periodicidad especifica, como mensual, trimestral, semestral o anual, pero no es adecuado para cambios abruptos o eventos inesperados. (R Adhikari , 2013)

En un modelo ARIMA estacional o SARIMA, los términos AR y MA estacionales se predicen utilizando valores de datos y errores en momentos que son m (el intervalo de la estacionalidad) cambios que se repite. (R Adhikari, 2013)

$$SARIMA\ (p, d, q)(P, D, m)$$

En IBM SPSS Modeler podemos encontrar los modelos disponibles para su uso dentro del nodo de Serie temporal, y se resumen como: Autorregresivo (AR): Es el número de órdenes autorregresivos del modelo. Los órdenes autorregresivos especifican los valores previos de la serie utilizados para predecir los valores actuales. Diferencia (d). Especifica el orden de diferenciación aplicado a la serie antes de estimar los modelos. El orden de la diferenciación corresponde al grado de tendencia de la serie y Media móvil (MA). Es el número de órdenes de media móvil presentes en el modelo. Los órdenes de media móvil especifican el modo en que se utilizan las desviaciones de la media de la serie para los valores previos con el fin de predecir los valores actuales. (IBM, 2021, documento de modelos de series temporales)

Al utilizar las técnicas descritas, se debe enfatizar que cada modelo genera un error de predicción, el cual se puede obtener como la diferencia en valor absoluto del valor pronosticado con respecto al valor real. Estos errores pueden deberse a la presentación de atipicidades como valores muy grandes o muy pequeño, los cuales generan cambios de comportamiento en la tendencia de la serie a modelar, para tratar de corregir esta problemática, el flujo en IBM SPSS Modeler identificará aquellos valores que se encuentre a 1, 2 y 3 desviación estándar con relación a la media en cada una de las series temporales a analizar. (IBM, 2021, documento de modelos de series temporales)

- **1 desviación estándar:** es el rango de datos dentro de 1 desviación estándar por encima y por debajo de la media *($\mu \pm \sigma$);* en una distribución normal, aproximadamente el 68.27% de los datos caen en este rango.
- **2 desviaciones estándar:** el rango de datos dentro de 2 desviaciones estándar por encima y por debajo de la media *($\mu \pm 2\sigma$);* aproximadamente 95.45% de los datos están dentro de este rango.
- **3 desviaciones estándar:** es el rango de datos con una cobertura aproximadamente 99.73% de los datos, por encima y por debajo de la media *($\mu \pm 3\sigma$).*

Al identificar un punto fuera de los rangos, será clasificado como atípico y este valor será remplazado por el valor de la banda o límite del rango donde se encuentre. Al tener una competencia de modelos se debe identificar los parámetros claves a analizar, para seleccionar la técnica que cumpla con los criterios de eficiencia, uno de los parámetros a utilizar es la medida de bondad de ajuste de un modelo lineal llamado R cuadrado; en ocasiones recibe el nombre de coeficiente de determinación. Puede tomar un valor entre 0 y 1. Un valor pequeño indica que el modelo no se ajusta bien a los datos.

Error Absoluto Medio (MAE) es una métrica estadística utilizada para medir la precisión de un modelo predictivo al calcular la media de las diferencias absolutas entre los valores reales y_t y los valores predichos $\hat{y}_t$ en una serie temporal u otros contextos de predicción. Es una métrica que indica cuánto en promedio se desvía la predicción del valor real, lo que significa que entre más cercano a 0 existe menos error entre la serie real y la ecuación pronosticada a utilizar. (IBM, 2021, documento de modelos de series temporales),

$$MAE = \frac{1}{n} \sum_{t=1}^{n} | y_t - \hat{y}_t |$$

Como otro criterio de evaluación para la selección del modelo se tiene el Back-Testing, será el proceso de comprobación y verificación de los tres últimos valores conocidos de la serie histórica. Es decir si y_t es el valor real de la serie para cualquier periodo t y $\hat{y}_t$ es el valor pronosticado para el periodo t, entonces el error de pronóstico conocido será $e_t = y_t - \hat{y}_t$.

2.4 Técnica de escalamiento lineal (LST)

Esta técnica fue definida por Drewnowski y Scott (1966) y es una de las más utilizadas en la construcción de numerosos índices sintéticos sociales y económicos (Morris, 1979; Zárate Martín, 1988; PNUD, 1990-2011; Velázquez y Gómez Lende, 2005; Velázquez, 2008). El cálculo utiliza los valores máximos (Xmax) y mínimos (Xmin) de los indicadores y el rango en lugar de la media y/o desviación estándar. Estos valores pueden ser empíricos, históricos o bien ideales, dependiendo del objetivo de la medición

2.5 Medata (Estrategia de datos de Medellín)

El Portal web de datos abiertos continente publicaciones de información por las diferentes dependencias de la Alcaldía de Medellín; los datos son expuestos como información pública y abierta para su uso, permitiendo su uso sin restricciones legales para su aprovechamiento. (Alcaldía de Medellín, Medata estrategia de datos de Medellín)

2.6 IBM SPSS Modeler

Es una plataforma sólida, versátil y gráfica de ciencia de datos que permite elaborar modelos predictivos de Machine Learning de forma rápida e intuitiva, sin necesidad de programación. Permite descubrir patrones y tendencias en datos estructurados o no estructurados de manera sencilla, mediante una única interfaz visual soportada por análisis avanzado. La herramienta cuenta con un completo conjunto de funciones de integración con leguaje OPEN (R y Python) para preparación de datos, visualización y modelado predictivo, así como la lectura de información o bases de datos, hojas de cálculo y archivos sin formato, incluidos los archivos de IBM SPSS Statistics, SAS y Microsoft Excel. (IBM, 2023, IBM SPSS Modeler).

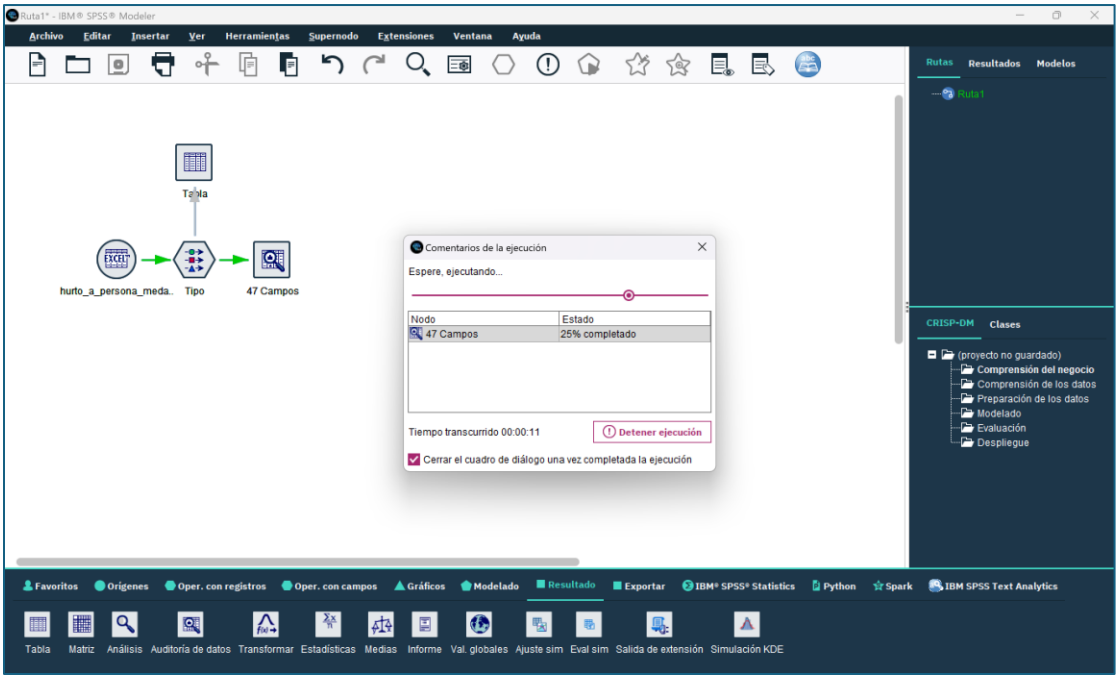


Ilustración 1 IBM SPSS Modeler

IBM SPSS Modeler ofrece una amplia gama de técnicas de minería de datos avanzadas diseñadas para cubrir las necesidades de las aplicaciones de Machine Learning, incluidos los siguientes algoritmos: Algoritmos de clasificación, Algoritmos de segmentación, Algoritmos de asociación, Extensibilidad: Integración con R y Python, etc. la herramienta permite calcular las predicciones o pronósticos basados en datos históricos a través de técnicas como árbol de decisión, redes neuronales, regresión

logística, series temporales, máquinas de vectores de soporte, regresión lineal/multiple, regresión de Cox, etc.

3 DATOS

3.1 Plan de Gestión de Datos

Medata es la estrategia de datos de la ciudad de Medellín, que busca la apropiación, apertura y uso de los datos como herramienta de gobierno, acción ciudadana y toma de decisiones. Los datos históricos de Hurtos de Medellín pertenecen al capítulo de Medellín en cifras, tema seguridad y set de datos: Casos de Hurto a Personas; los datos utilizados para este proyecto son públicos. (Alcaldía de Medellín, Medata estrategia de datos de Medellín)

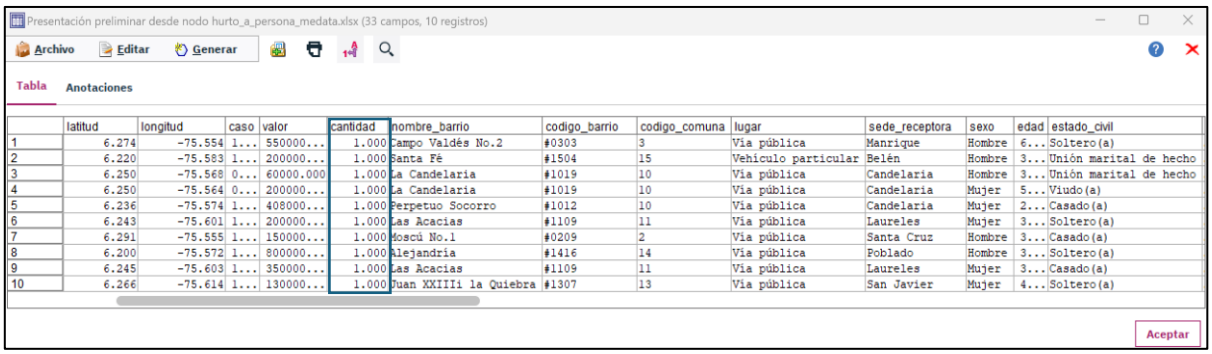
https://creativecommons.org/licenses/by-sa/4.0/legalcode es la licencia identificada dentro del portal de datos abiertos Medata que garantiza el aprovechamiento de la información del proyecto, compartir el resultado y conocimiento a través de instrumentos jurídicos libres y gratuitos.

Los resultados o pronósticos derivados de este proyecto son públicos y están documentados en este informe. Los activos analíticos o rutas construidas para el proyecto que incluye la automatización de la competencia de modelos no es publica, por lo anterior la propiedad intelectual de la ruta no deben ser publicados fuera de este proyecto. El activo analítico o ruta es propiedad del estudiante.

3.2 Adquisición de datos

El set de datos actualizado a septiembre del 2024, con el histórico de hurtos a personas fue entregado por medio de la respuesta a la radicación de PQR con el número de solicitud: 823475275843839398.

La fuente de datos se encuentra en un solo archivo con formato .csv, el cual contiene 33 variables, 351.056 registros y almacenados desde enero del 2017 hasta septiembre del 2024; la información fue descargada del correo electrónico y almacenado en el pc de trabajo local del estudiante. La variable objetivo (Y) a pronosticar es la cantidad de hurto a personas, la cual es una variable numérica que depende del tiempo y se identifica en el set de datos con el nombre de cantidad; para los modelos no se utilizará variables dependientes (X), solo se tendrán en cuenta los modelos univariado,

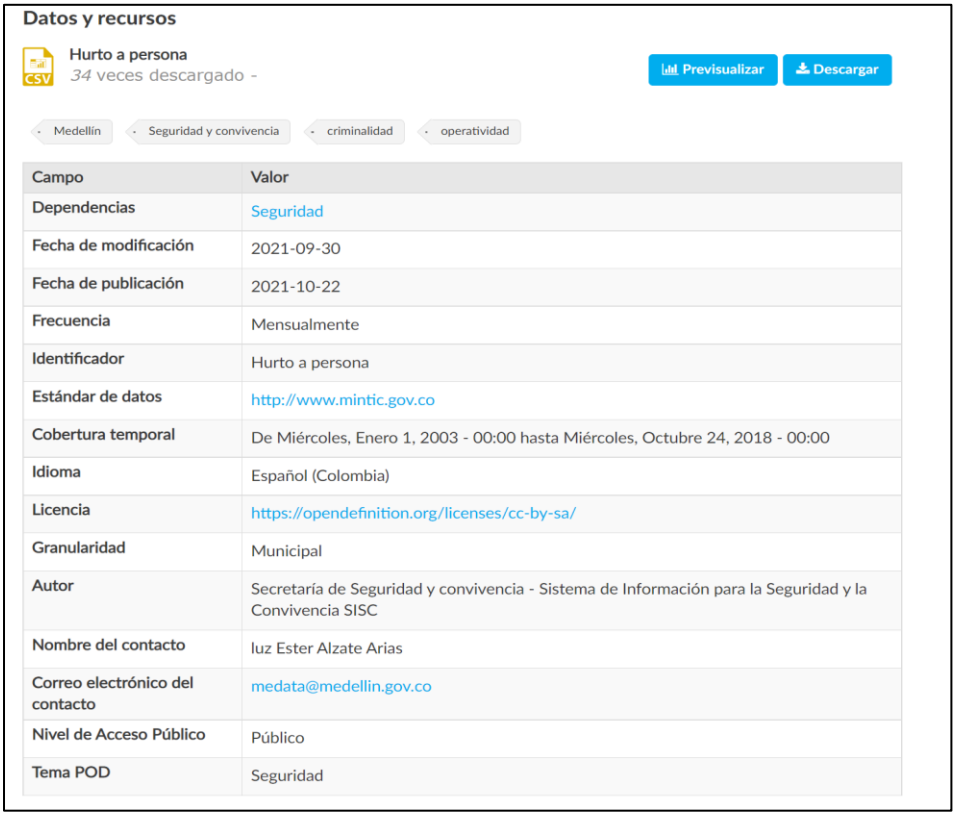


	latitud	longitud	caso	valor	cantidad	nombre_barrio	codigo_barrio	codigo_comuna	lugar	sede_receptora	sexo	edad	estado_civil
1	6.274	-75.554	1...	550000...	1.000	Campo Valdés No.2	#0303	3	Vía pública	Manrique	Hombre	6...	Soltero(a)
2	6.220	-75.583	1...	200000...	1.000	Santa Fé	#1504	15	Vehiculo particular	Belén	Hombre	3...	Unión marital de hecho
3	6.250	-75.568	0...	60000.000	1.000	La Candelaria	#1019	10	Vía pública	Candelaria	Hombre	3...	Unión marital de hecho
4	6.250	-75.564	0...	200000...	1.000	La Candelaria	#1019	10	Vía pública	Candelaria	Mujer	5...	Viudo(a)
5	6.236	-75.574	1...	408000...	1.000	Perpetuo Socorro	#1012	10	Vía pública	Candelaria	Mujer	2...	Casado(a)
6	6.243	-75.601	1...	200000...	1.000	Las Acacias	#1109	11	Vía pública	Laureles	Mujer	3...	Soltero(a)
7	6.291	-75.555	1...	150000...	1.000	Moscú No.1	#0209	2	Vía pública	Santa Cruz	Hombre	3...	Casado(a)
8	6.200	-75.572	1...	800000...	1.000	Alejandria	#1416	14	Vía pública	Poblado	Hombre	3...	Soltero(a)
9	6.245	-75.603	1...	350000...	1.000	Las Acacias	#1109	11	Vía pública	Laureles	Mujer	3...	Casado(a)
10	6.266	-75.614	1...	130000...	1.000	Juan XXIIIi la Quiebra	#1307	13	Vía pública	San Javier	Mujer	4...	Soltero(a)

Ilustración 2 Vista de los 10 primeros registros

3.3 Descripción y análisis preliminar de los datos

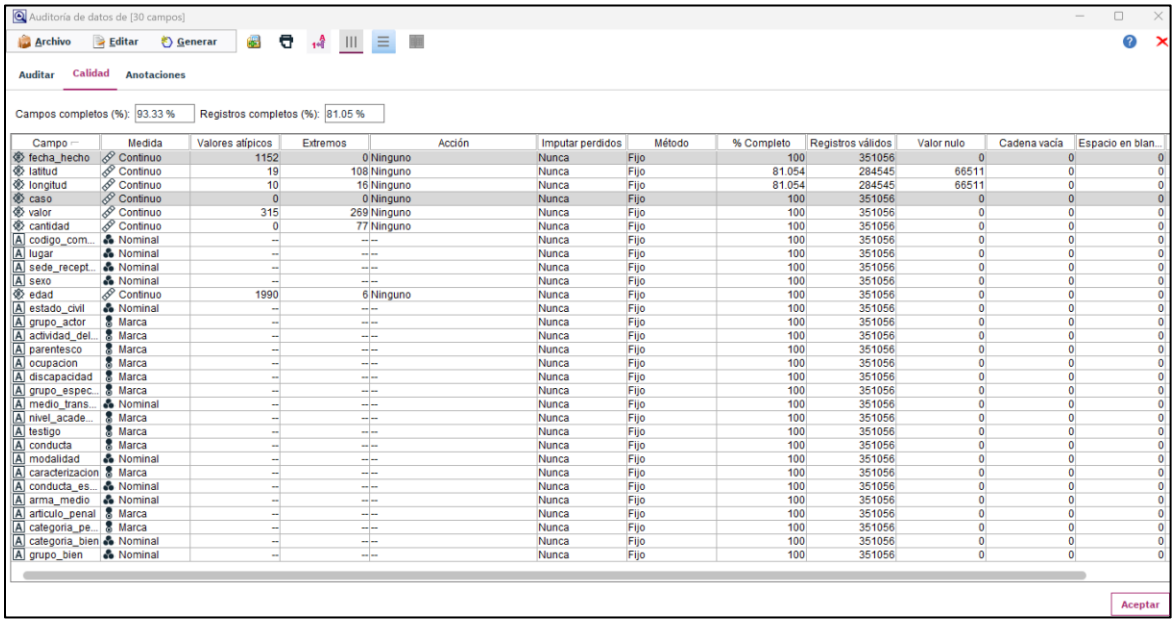
La información del número de casos de Hurtos a personas se encuentra almacenada en el módulo de seguridad y convivencia de Medata, se identifican hechos relacionados con la seguridad, convivencia, derechos humanos, ocurrido en la ciudad de Medellín y que han sido recopilados por el proyecto municipal Sistema de Información para la Seguridad y la Convivencia SISC; a continuación, se ilustra el detalle de la información abierta de los datos disponibles en la plataforma de Medata:



Campo	Valor
Dependencias	Seguridad
Fecha de modificación	2021-09-30
Fecha de publicación	2021-10-22
Frecuencia	Mensualmente
Identificador	Hurto a persona
Estándar de datos	http://www.mintic.gov.co
Cobertura temporal	De Miércoles, Enero 1, 2003 - 00:00 hasta Miércoles, Octubre 24, 2018 - 00:00
Idioma	Español (Colombia)
Licencia	https://opendefinition.org/licenses/cc-by-sa/
Granularidad	Municipal
Autor	Secretaría de Seguridad y convivencia - Sistema de Información para la Seguridad y la Convivencia SISC
Nombre del contacto	luz Ester Alzate Arias
Correo electrónico del contacto	medata@medellin.gov.co
Nivel de Acceso Público	Público
Tema POD	Seguridad

Ilustración 3 Detalle de Datos en el Portal Metada

El proyecto inicia con una exploración de la fuente de información de los hurtos de personas con el Nodo de auditoría de datos disponible es la herramienta IBM SPS Modeler, identificando que la variable cantidad y fecha del caso están al 100% de completitud, generando una confianza de completitud de información para el procesamiento de los modelos, sin necesitar etapa de imputación de datos faltantes:



Campo	Medida	Valores atípicos	Extremos	Acción	Imputar perdidos	Método	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio en blan...
fecha_hecho	Continuo	1152	0	Ninguno	Nunca	Fijo	100	351056	0	0	0
latitud	Continuo	19	108	Ninguno	Nunca	Fijo	81.054	284545	66511	0	0
longitud	Continuo	10	16	Ninguno	Nunca	Fijo	81.054	284545	66511	0	0
caso	Continuo	0	0	Ninguno	Nunca	Fijo	100	351056	0	0	0
valor	Continuo	315	269	Ninguno	Nunca	Fijo	100	351056	0	0	0
cantidad	Continuo	0	77	Ninguno	Nunca	Fijo	100	351056	0	0	0
codigo_com...	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
lugar	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
sede_recept...	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
sexo	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
edad	Continuo	1990	6	Ninguno	Nunca	Fijo	100	351056	0	0	0
estado_civil	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
grupo_actor	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
actividad_del...	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
parentesco	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
ocupacion	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
discapacidad	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
grupo_espec...	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
medio_trans...	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
nivel_acade...	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
testigo	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
conducta	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
modalidad	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
caracterizacion	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
conducta_es...	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
arma_medio	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
articulo_penal	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
categoria_pe...	Marca	--	--	--	Nunca	Fijo	100	351056	0	0	0
categoria_bien	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0
grupo_bien	Nominal	--	--	--	Nunca	Fijo	100	351056	0	0	0

Ilustración 4 Auditoria de datos

Se realiza una validación del comportamiento demográfico de las personas hurtadas y se identifica que la mayoría de los hurtos son registrados a Hombre con un promedio de edad de 34 años, siendo el bien o categoría que más se hurta elementos de tecnología (celular). El 16% de los casos son presentado en los domingos y sábados. A continuación se ilustra gráficas con los valores mencionados.

Valor	Proporción	%	Recuento
Domingo		16.83	15384
Sábado		16.14	14751
Viernes		14.65	13395
Jueves		14.53	13280
Miércoles		13.7	12523
Martes		12.82	11717
Lunes		11.34	10362

Valor	Proporción	%	Recuento
Tecnología		41.91	16328
Dinero, joyas, piedras preciosas y título valor		26.12	10174
Documentos		10.17	3963
Prendas de vestir y accesorios		9.01	3510
Electrodomésticos		2.71	1055

Ilustración 5 Estadísticas descriptivas de la información

3.4 Preprocesamiento de los datos

El análisis preliminar identifico que se debe seleccionar el periodo de análisis de la historia de datos, debido a se debe totalizar todos los hurtos presentados en una periodicidad mensual, se decide tomar el periodo de tiempo de los tres últimos años 2022 al 2024, periodo que no se ve afectado por la pandemia COVID 19. Se realiza una selección de los casos con una mayor fecha de 2022-01-01.

Como base de entrenamiento para los modelos predictivos de los hurtos a personas, se trabaja con un total de 91.412 hurtos reportados desde el 2022, adicional en el preprocesamiento se eliminaron los hurtos presentados en veredas (rurales oficiales): 52. Áreas Institucionales (urbanos): 20. Áreas de expansión (urbano – rural): 7, los cuales no hacen parte de la granularidad del proyecto.

3.5 Aspectos éticos

Los datos presentados en Medata están bajo licencia "Database Contents License (DbCL) v1.0" la cual señala las normas del uso de los datos como abiertos. La licencia de IBM SPSS Modeler está bajo descarga gratuita del portal del fabricante IBM como temporal por 2 meses, para su prueba de funcionalidades.

Los datos de hurtos a personas son cifras históricas agregadas en toda la ciudad y, al ser de acceso público, no incluyen información sensible o que identifique las personas; el proyecto no intentará identificar a individuos específicos.

Aunque los datos son públicos, se utilizarán únicamente para cumplir con los objetivos del proyecto, los resultados son georreferenciado con el fin de identificar las zonas con proyección de priorización de estrategias de seguridad y no generar una percepción negativa injustificada sobre algún barrio o comuna en particular.

En caso de un cambio de tendencia en los hurtos de personas, el modelo podrá recalibrarse y seleccionar otra técnica que ajuste una eficiencia esperada del 80% , el cual estará automatizado como parámetro en el flujo de la herramienta IBM SPSS Modeler, este valor será tenido en cuenta para la selección del modelo o técnica de suavizamiento a utilizar para la proyección de cada serie; en caso de que la serie no cuente con los periodos idóneos para ingresar al proceso analítico, las proyecciones se realizarán con el promedio móvil de los tres últimos meses reportados de la historia.

4 DESARROLLO DE MODELOS

IBM SPSS Modeler permite la creación de rutas automáticas y facilita la programación de lenguajes OPEN, la flexibilidad de la herramienta permite la automatización de un flujo de proceso analítico que garantiza la competencia de modelos de series

temporales univariados disponibles en la herramienta; logrando para casa ejecución una recalibración automática del análisis o si se detectan cambios significativos en la eficiencia de las proyecciones o cambios en las tendencias de los datos. Cada serie temporal se ejecuta y almacena temporalmente con sus parámetros estimados para su validación posterior en la etapa de evaluación del modelo.

La metodología seleccionada para el proyecto es CRISP-DM (Cross-Industry Standard Process for Data Mining), reconocida por su ciclo de vida estructurado en seis fases, conocimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue.

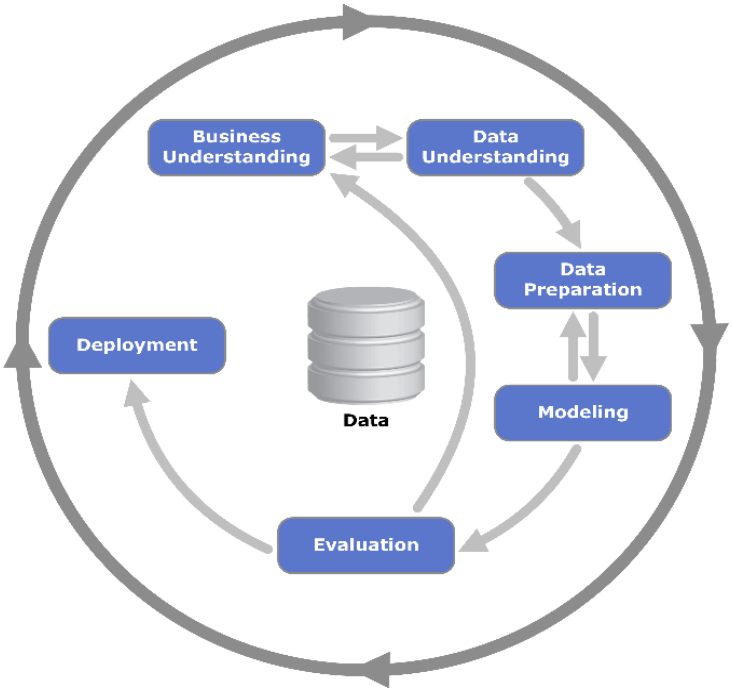


Ilustración 6 Ciclo de minería de datos

La secuencia de las fases no es estricta, es flexible permite adaptarse a las particularidades del proyecto, ajustando el modelo predictivo y recalibrándolo cuando se detecten cambios en los datos, optimizando así la precisión y relevancia del pronóstico de hurtos a persona. A continuación, se explica el desarrollo de las actividades realizadas en las diferentes fases del ciclo CRISP-DM. (IBM, 2020, CRISP-DM).

Fase I. Comprensión del negocio

El hurto a personas en la ciudad de Medellín es un delito recurrente que genera un ambiente de desconfianza entre los ciudadanos, este delito puede ocurrir en diferentes periodos de tiempo, lo que confirma que las técnicas a trabajar son las

series temporales, ya que estos modelos permiten estimar los valores futuros que pueden usarse para anticipar la planificación de las estrategias de prevención y seguridad de la ciudad.

El desarrollo del activo analítico o ruta en IBM SPSS Modeler permitirá la creación de una metodología que contenga la automatización de los pasos que garanticen la competencia de modelos de suavizamiento exponencial y autorregresivos ARIMA y SARIMA. Este enfoque permitirá identificar áreas de mayor riesgo que apoye una estrategia de asignación de recursos de la fuerza pública de manera eficaz y anticipada. Lograr la implementación de la automatización en la ruta, permitirá la reutilización de este activo para analizar otro tipo de fenómenos de manera rápida.

Como actividades principales de la fase se delimitó los objetivos de análisis, el rango histórico de la información para el entrenamiento de los modelos analíticos y el bosquejó de los pasos que debe cumplir en la automatización de la herramienta IBM SPSS Modeler; los pasos deben asegurar una recalibración del modelo predictivo, mediante la selección de la técnica de forma automática sin intervención de una persona. A continuación se describe el flujo del proceso que se automatiza en la ruta o activo analítico de IBM SPSS Modeler:

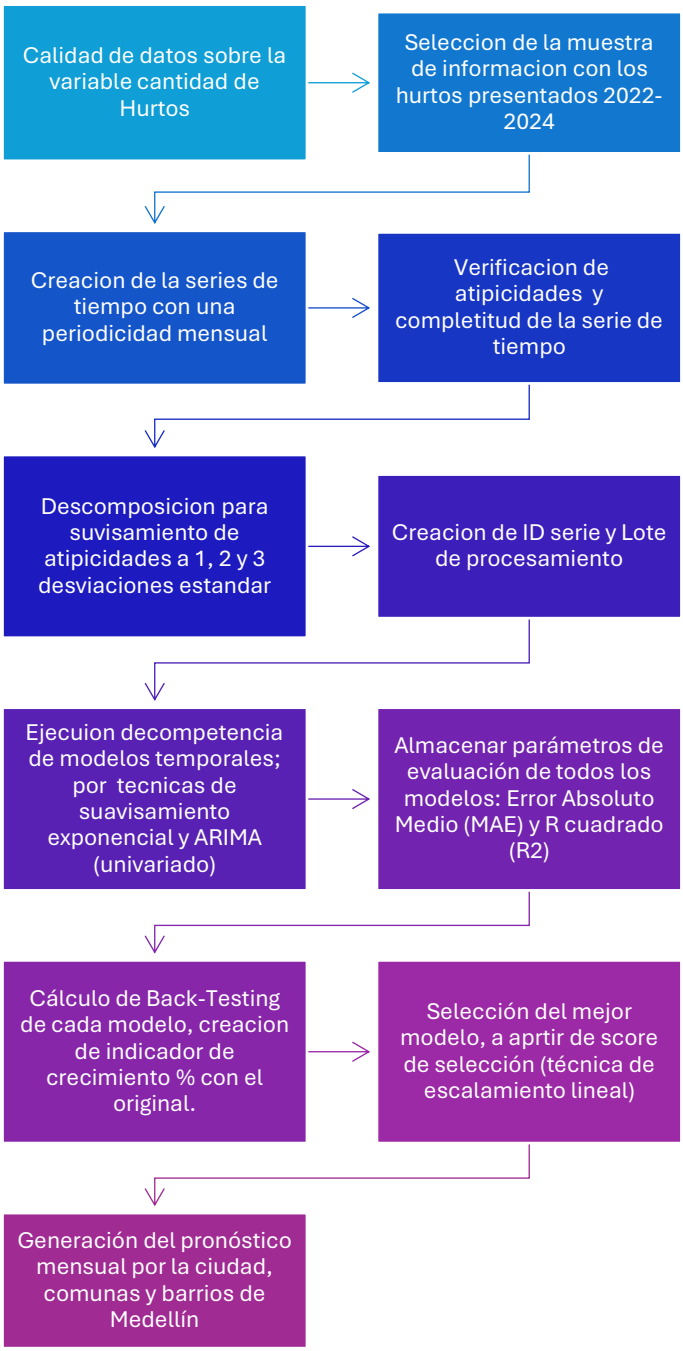


Ilustración 7 Flujo de automatización en IBM SPSS Modeler

Fase II. Entendimiento de los datos

Después de seleccionar el periodo de 2022 al 2024 con un total de hurtos de 91.412 con 30 campos complementarios o con información del caso, a continuación se presenta una evaluación de completitud y estadísticas descriptivas:

	Campo	Medida	Mín.	Máx.	Media	Desv. estándar	Sesgo	Mediana	Modo	Exclusivo	Válido
1	C1	Continuo	3635.000	586882.000	318509.650	170189.484	-0.173	330785.500	3635.000[1]	--	91412
2	fecha_hecho	Continuo	2022-01-01	2024-09-30	--	--	--	2023-03-19	2022-11-19	--	91412
3	hora_hecho	Continuo	00:00:00	23:59:00	--	--	--	12:00:00	18:00:00	--	91412
4	latitud	Continuo	5.605	6.372	6.247	0.028	-3.095	6.249	6.246	--	39132
5	longitud	Continuo	-75.709	-0.001	-75.556	1.267	59.600	-75.573	-75.575	--	39132
6	caso	Continuo	0.000	1.000	0.880	0.325	-2.334	1.000	1.000	--	91412
7	valor	Continuo	0.000	61472984.000	717951.335	4589490.069	58.841	0.000	0.000	--	91412
8	cantidad	Continuo	1.000	3.000	1.001	0.030	35.452	1.000	1.000	--	91412
9	nombre_barrio	Nominal	--	--	--	--	--	--	Sin dato	343	91412
10	codigo_comuna	Nominal	--	--	--	--	--	--	10	22	91412
11	lugar	Nominal	--	--	--	--	--	--	Vía pública	91	91412
12	sede_receptora	Nominal	--	--	--	--	--	--	Candelaria	25	91412
13	sexo	Nominal	--	--	--	--	--	--	Hombre	3	91412
14	edad	Continuo	-1.000	825.000	34.244	14.488	3.062	32.000	30.000	--	91412
15	estado_civil	Nominal	--	--	--	--	--	--	Sin dato	6	91412
16	grupo_actor	Marca	--	--	--	--	--	--	Sin dato	1	91412
17	actividad_delictiva	Marca	--	--	--	--	--	--	Sin dato	1	91412
18	parentesco	Marca	--	--	--	--	--	--	Sin dato	1	91412
19	ocupacion	Marca	--	--	--	--	--	--	Sin dato	1	91412
20	discapacidad	Marca	--	--	--	--	--	--	Sin dato	1	91412
21	grupo_especial	Marca	--	--	--	--	--	--	Sin dato	1	91412
22	medio_transporte	Nominal	--	--	--	--	--	--	Caminata	10	91412
23	nivel_academico	Marca	--	--	--	--	--	--	Sin dato	1	91412
24	testigo	Marca	--	--	--	--	--	--	Sin dato	1	91412
25	conducta	Marca	--	--	--	--	--	--	Hurto a persona	1	91412
26	modalidad	Nominal	--	--	--	--	--	--	Atraco	21	91412
27	caracterizacion	Marca	--	--	--	--	--	--	Sin dato	1	91412
28	conducta_especial	Nominal	--	--	--	--	--	--	Sin dato	16	91412
29	arma_medio	Nominal	--	--	--	--	--	--	No	7	91412
30	articulo_penal	Marca	--	--	--	--	--	--	Sin dato	1	91412
31	categoria_penal	Marca	--	--	--	--	--	--	Sin dato	1	91412
32	categoria_bien	Nominal	--	--	--	--	--	--	Sin dato	44	91412
33	grupo_bien	Nominal	--	--	--	--	--	--	Sin dato	5	91412

Ilustración 8 Auditoria de datos

Se identifica que las variables cantidad y fechas se encuentra con una completitud del 100% de información, las cuales son las principales variables para la creación de las series con una periodicidad mensual; 10 casos de hurtos a personas no tienen asignación de comuna, identificándolas como valores en blanco, a través de la completitud del nombre del barrio se logra rellenar la variable dejándola con 100% completa. Se identifica que el barrio con más reportes de casos de hurtos a personas es candelaria con el 12%, se verifica al graficar los hurtos por las coordenadas registradas:

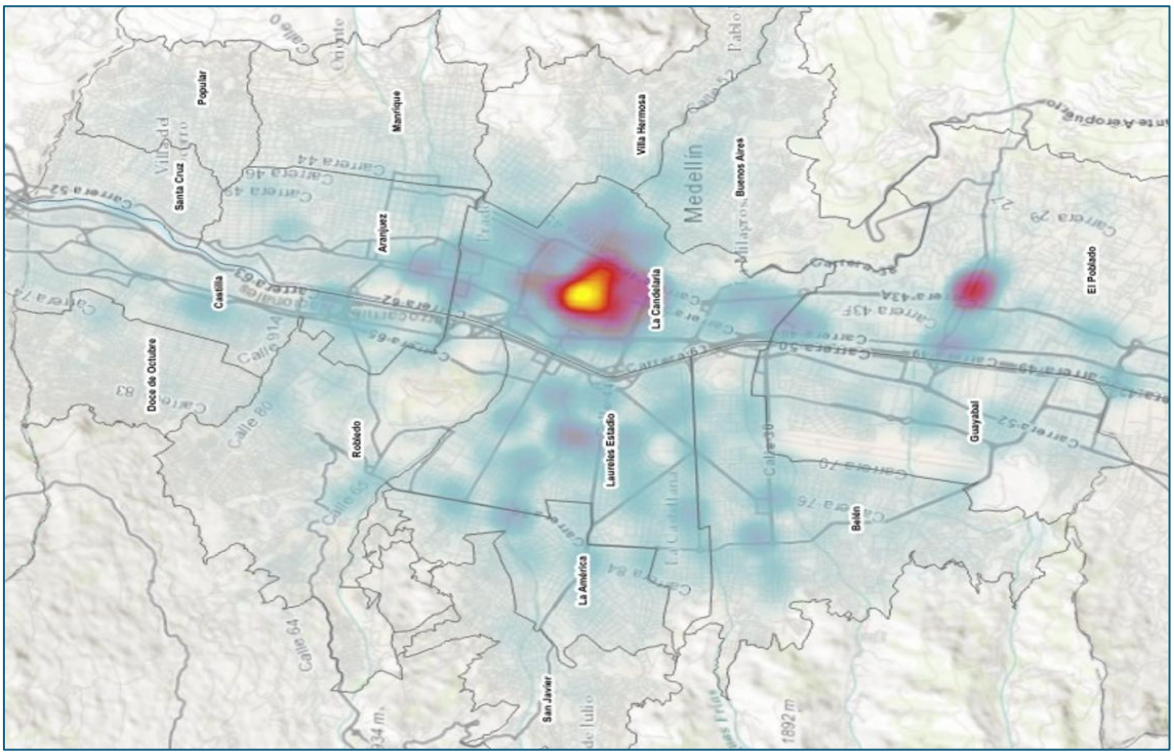
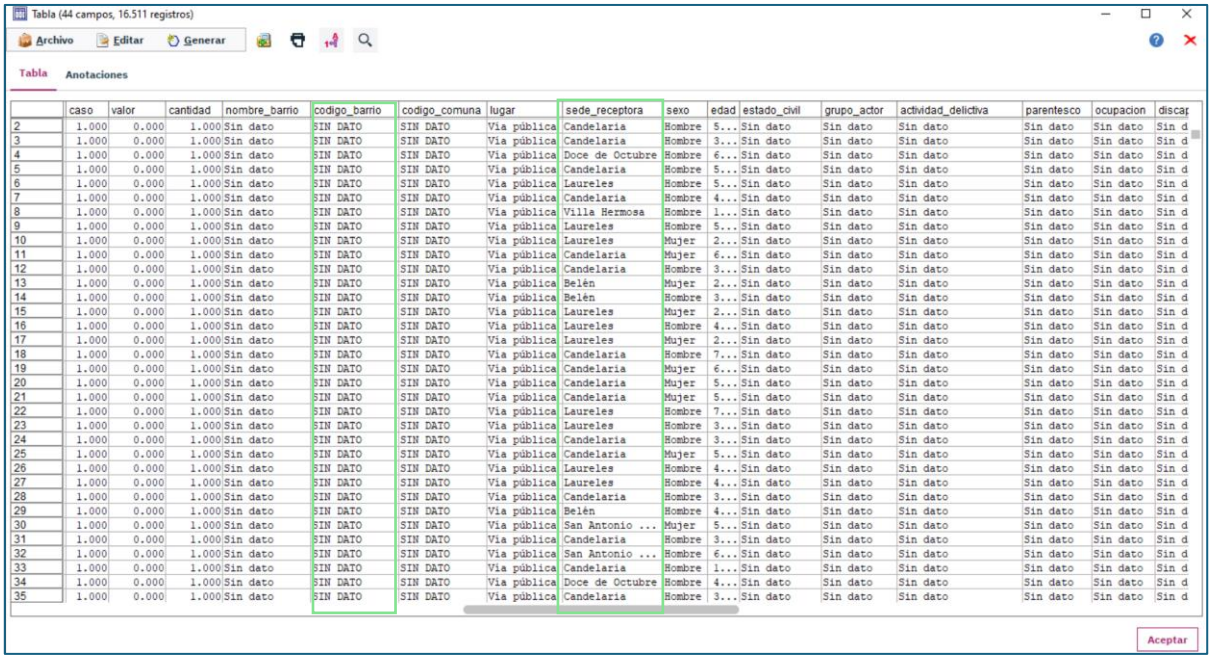


Ilustración 9 Densidad histórica de hurtos en Medellín

Fase III. Preparación de los datos

En la etapa de limpieza se identifica que las variables código de barrio y comuna, deben estar homologadas y con registros válidos para la creación de los totales de forma mensual; se evidencia en la fuente de información que 16.511 casos contiene la palabra sin dato, en el campo comuna y barrio, para los cuales se decide completar la información remplazando el valor por el reportado en la columna sede receptora de la denuncian del hurto, y a partir de tener diligenciada el barrio se completa la comuna al que pertenece.

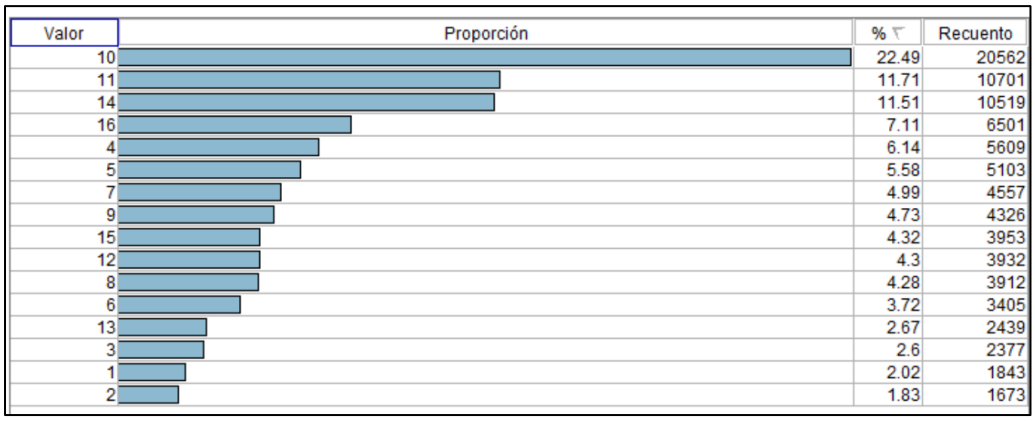


caso | valor | cantidad | nombre_barrio | codigo_barrio | codigo_comuna | lugar | sede_receptora | sexo | edad | estado_civil | grupo_actor | actividad_delictiva | parentesco | ocupacion | discap

2	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
3	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
4	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Doce de Octubre	Hombre	6...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
5	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
6	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Hombre	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
7	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	4...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
8	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Villa Hermosa	Hombre	1...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
9	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Hombre	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
10	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Mujer	2...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
11	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Mujer	6...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
12	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
13	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Belén	Mujer	2...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
14	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Belén	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
15	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Mujer	2...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
16	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Hombre	4...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
17	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Mujer	2...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
18	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	7...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
19	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Mujer	6...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
20	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Mujer	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
21	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Mujer	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
22	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Hombre	7...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
23	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
24	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
25	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Mujer	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
26	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Hombre	4...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
27	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Laureles	Hombre	4...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
28	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
29	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Belén	Hombre	4...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
30	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	San Antonio ...	Mujer	5...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
31	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
32	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	San Antonio ...	Hombre	6...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
33	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	1...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
34	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Doce de Octubre	Hombre	4...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d
35	1.000	0.000	1.000	Sin dato	SIN DATO	SIN DATO	Vía pública	Candelaria	Hombre	3...	Sin dato	Sin dato	Sin dato	Sin dato	Sin dato	Sin d

Ilustración 10 Registro de hurtos sin dato para el campo barrio y comuna

Con esta transformación de datos, se logra asignar a cada hurto el código correspondiente a la comuna, identificando a las comunas 10 (la Candelaria), 11 (laureles estadios) y 14 (el poblado) como las áreas con mayor incidencia de hurtos a personas.



Valor	Proporción	%	Recuento
10		22.49	20562
11		11.71	10701
14		11.51	10519
16		7.11	6501
4		6.14	5609
5		5.58	5103
7		4.99	4557
9		4.73	4326
15		4.32	3953
12		4.3	3932
8		4.28	3912
6		3.72	3405
13		2.67	2439
3		2.6	2377
1		2.02	1843
2		1.83	1673

Ilustración 11 Cantidad de hurtos de personas por comuna

Después de garantizar que las variables comuna y barrio contenga la información, se realiza la construcción de las series de tiempo, totalizando la cantidad de hurtos de forma mensual, construyendo 266 series.

Ítem	***Cantidad***
Medellín	1
Comunas	16

Barrio	249
Total serie	**266**

Tabla 1 Cantidad de series originales

Se crearán 4 nuevas series, para un total de 798, donde cada serie de tiempo es derivada al intervenir los atípicos que se encuentre a 1 desviación estándar, 2 deviaciones y 3 desviaciones; el propósito de este paso es permitir al flujo en IBM SPSS Modeler el modelamiento de las atipicidades o outliers de forma automática, antes de calcular los modelos predictivos; estas nuevas series derivadas de la variable original permite que por cada combinatoria se procese todos los modelos de suavizamiento exponencial, ARIMA y SARIMA disponibles en la herramienta analítica, logrando generar una competencia para la selección del mejor modelo. A continuación, se describe el incremento de las series de tiempo:

Ítem	*Cantidad*	*Desviaciones (3)*
Medellín	1	3
Comunas	16	48
Barrio	249	747
Total series	**266**	**798**

Tabla 2 Cantidad de series con desviaciones

Como siguiente paso, se crea todos los ID o nomenclatura que diferencia cada serie, para identificar si es ciudad total, barrios o comunas; adicional se le asigna un lote de procesamiento aleatorio para la ejecución de los modelos y los tiempos de procesamiento sea más rápidos a nivel de cómputo. Dentro de la base de datos se eliminas el resto de las variables complementarias, solo se utilizarán para estadísticas descriptivas que apoyarán los análisis. A continuación se describe un ejemplo de la estructura del ID o identificador de la serie de tiempo:

ID1= Medellin_original (Series original)
ID2= Medellin_original_dvs1 (Series temporal ajustada sus outliers que estén a 1 desviación)
ID3= Medellin_original_dvs2(Series temporal ajustada sus outliers que estén a 2 desviación)
ID4= Medellin_original_dvs3(Series temporal ajustada sus outliers que estén a 3 desviación)

El análisis de atípicos consiste en identificar aquellas observaciones que se están desviando cierto número de desviaciones estándar hacia arriba como hacia abajo de la media aritmética de cada serie de tiempo. Es decir, son aquellos datos que se salen del comportamiento normal de los registros de cada serie de tiempo construida.
Lo anterior quiere decir que se tendrán 3 escenarios adicionales al escenario original, el primero con 1 desviación estándar, el segundo con 2 desviaciones estándar y el

tercero con 3 desviaciones estándar. Las siguientes ecuaciones, muestran las fórmulas utilizadas dentro del flujo de automático de IBM SPSS Modeler:

1 atipicidades o outliers superior $O_i^{+k} = \theta_i + k * \sigma_i \quad \forall\, k \in (1{,}2\ y\ 3)$

2 atipicidades o outliers inferior $O_i^{-k} = \theta_i - k * \sigma_i \quad \forall\, k \in (1{,}2\ y\ 3)$

Donde:

O_i^{+k}: Valor máximo permitido de la desviación estándar.

O_i^{-k}: Valor mínimo permitido de la desviación estándar

θ_i: Media de la Serie i.

σ_i: Desviación estándar de la serie i.

k: El Número de desviación.

Los outliers o atípicos son los valores que están por fuera de un rango definido por la media θ_i y la desviación estándar σ_i de la serie. El valor de k define cuántas desviaciones estándar se permiten antes de considerar un valor como un outlier. Si un valor supera el rango, se ajusta al valor máximo o mínimo permitido.

Corrección de series:

Si el valor α_{ij} es mayor que el límite superior, se ajusta o remplaza por O_i^{+k}

Si el valor α_{ij} es menor que el límite superior, se ajusta o remplaza por O_i^{-k}

Donde, α_{ij} es el dato u observación en el periodo j de la serie i.

Una vez realizada la corrección de outliers, al final del ejercicio se tienen 4 escenarios posibles: escenario base (original) y escenarios con los ajustes de los atípicos a 1, 2 y 3 desviaciones. Es decir que el total de series de cada Modelo se multiplica por 3; como siguiente paso del flujo analítico, se realiza la validación de que las series construidas estén competas, que todos sus valores sean validos o tenga un valor diferente de cero. A continuación se visualiza la ruta o activo analítico que ejecutara los pasos descritos:

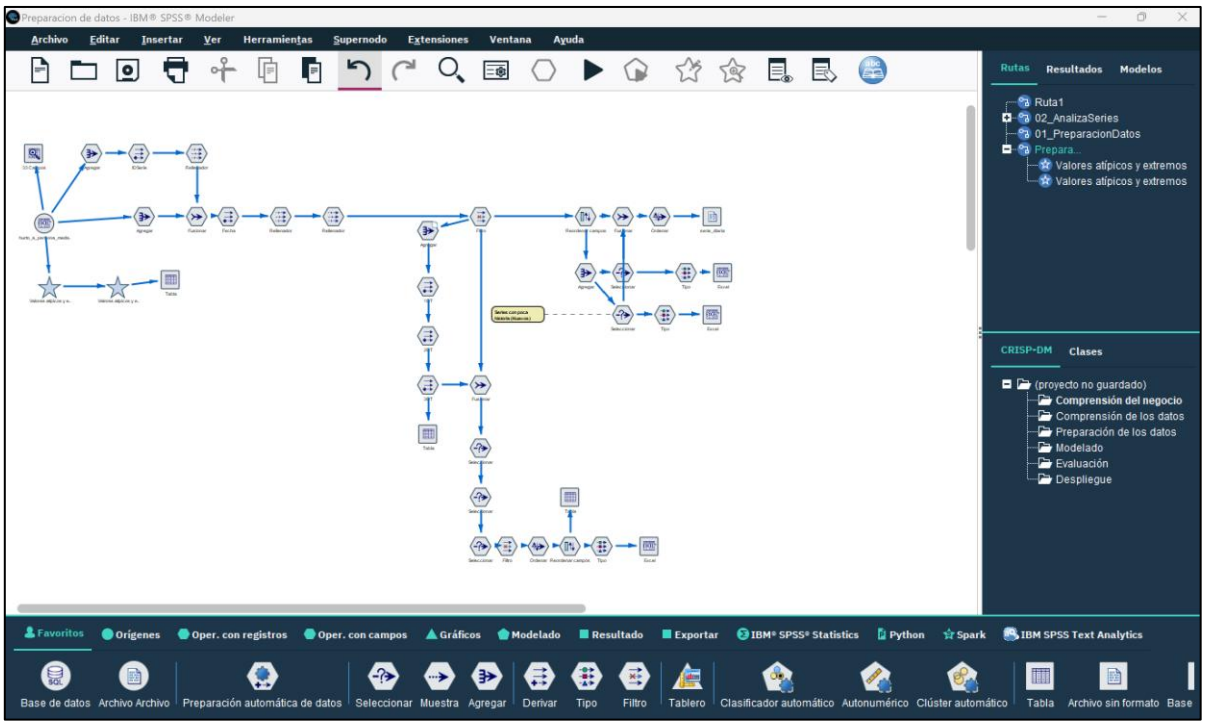


Ilustración 12 Activo analítico de preparación de datos en IBM SPSS Modeler

Fase IV. Modelado

Las técnicas de modelado univariado a trabajar en la herramienta de IBM SPSS Modeler son Suavizado exponencial y los modelos autorregresivos ARIMA y SARIMA, cada serie de tiempo creada deberá almacenar temporalmente todos sus indicadores; esto se debe a su utilización en la etapa de evaluación, donde después de su validación se selecciona la técnica indicada para cada ID creado. Este proceso garantiza que si hay cambios en la tendencia de los nuevos datos, el modelo cambie por otro que tenga mejor ajuste en los parámetros de eficiencia creados para el proceso de validación, de forma automática.

A continuación se muestra el almacenamiento que se genera al ejecutar el proceso de modelamiento, donde se ven las combinaciones de modelos utilizados y los cuales todos compite por ser seleccionado como el ganador para generar la proyección; los modelos quedan guardados en Nuggets, nodos generados por IBM SPSS Modeler con la configuración de sus parámetros; los Nuggets quedan guardados en una carpeta para su posterior etapa de verificación de los supuestos.

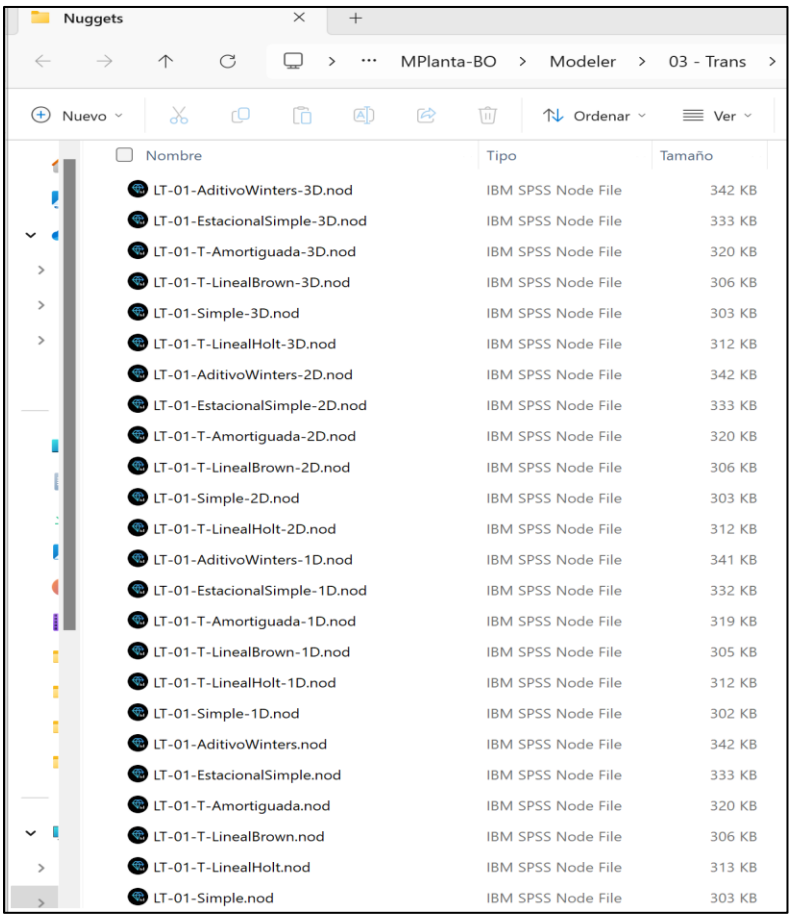


Ilustración 13 Almacenamiento temporal de Modelos

Para el almacenamiento y ejecución de cada combinación de modelo, se realiza la automatización en super nodos, el cual se evidencia en el flujo de la herramienta como un diamante de color amarillo, se asegura la automatización para los 4 caminos por separado. En la siguiente imagen se ilustra la automatización del flujo que garantiza que cada serie temporal debe ingresar por cada combinación:

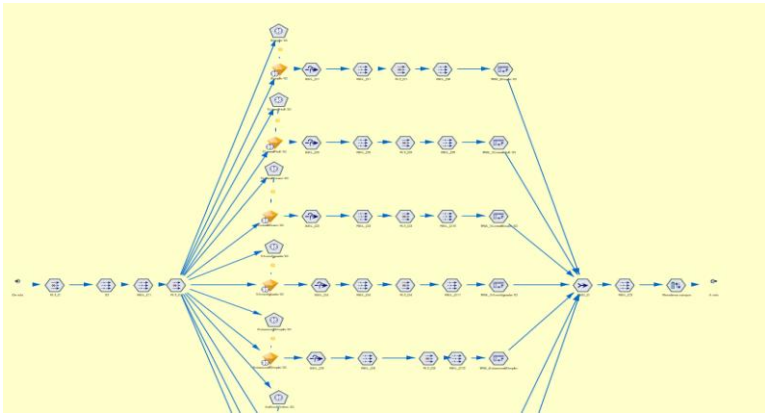


Ilustración 14 Supernodo para modelos con series de 1 desviación

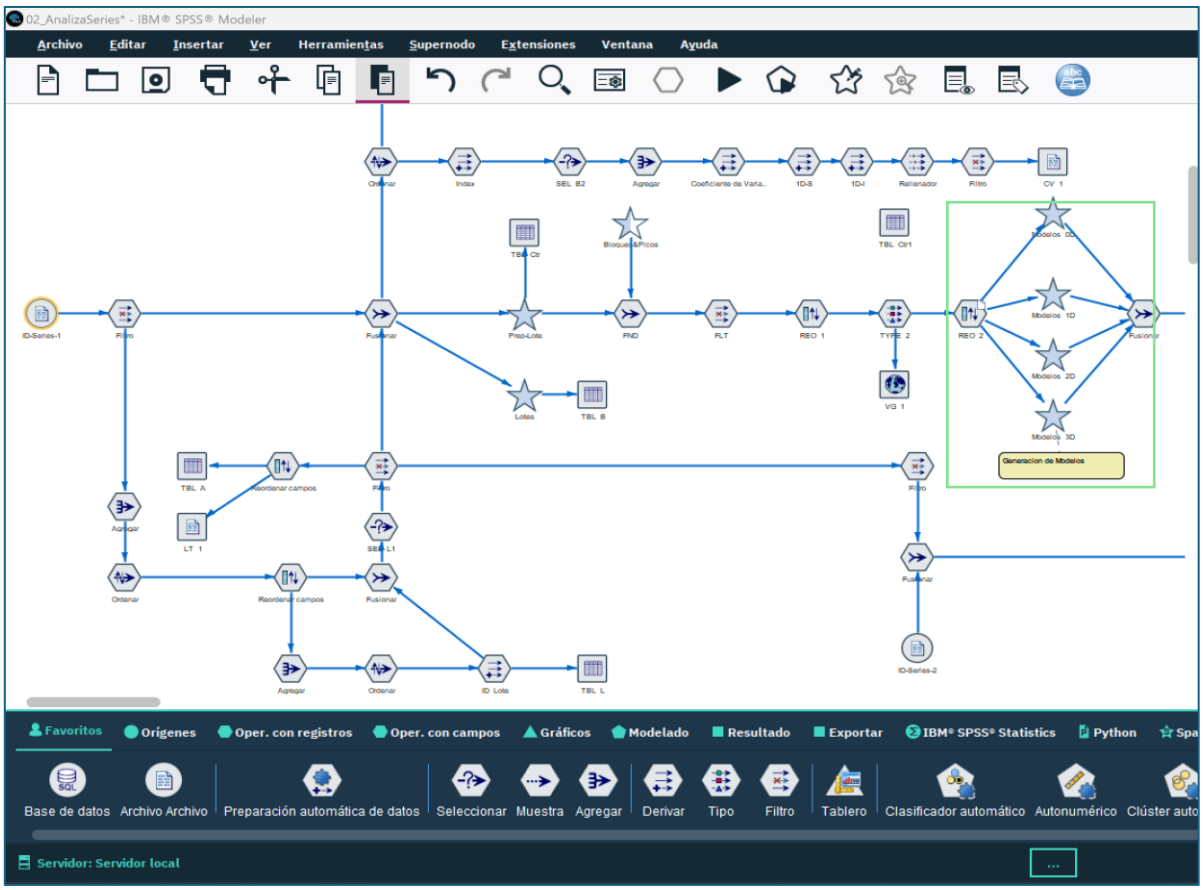


Ilustración 15 Ejecución de competencia de modelos

Cuando se tiene los resultados de cada modelo, se almacena los parámetros de evaluación, los cuales entraran a competir para la selección del modelo ganador. se extrae los valores correspondientes al indicador R2, el cual puede tomar un valor entre 0 y 1, el valor pequeño indica que el modelo no se ajusta bien a los datos. Error Absoluto Medio (MAE, siglas en inglés) el cual mide la desviación de la serie del nivel pronosticado por el modelo, entre más cercano a 0 significa que existe menos error entre la serie real y la serie o ecuación pronosticada a través de los parámetros, a continuación se visualiza el almacenamiento de los parámetros en formato .txt.

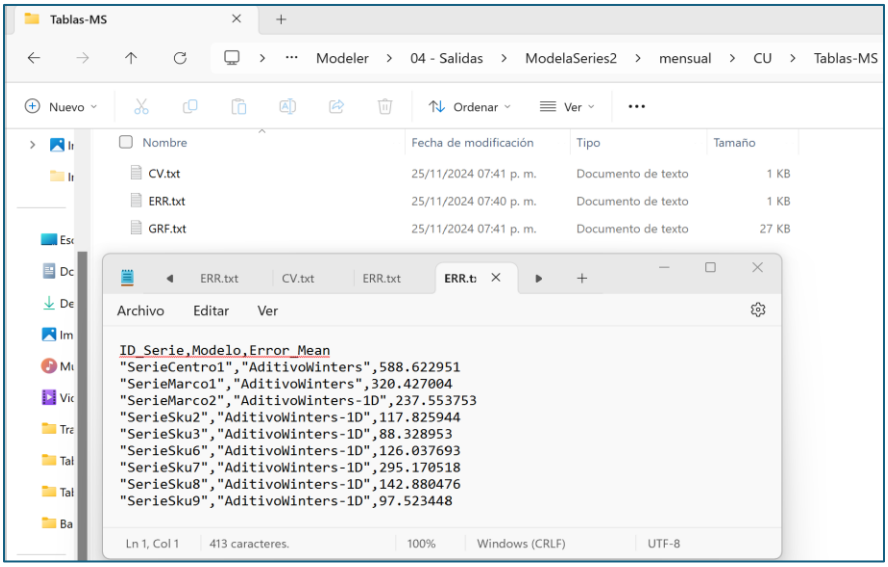


Ilustración 16 Almacenamiento de parámetros de cada Modelos

El almacenar la información permite automatizar un indicador o score que seleccione automáticamente el mejor modelo cuando cumpla con los parámetros que se requiere para cada serie de tiempo, y garantice una eficiencia igual o mayor del 80%.

Fase V. Evaluación de resultados

Esta etapa requiere de una programación idónea que garantice la evaluación automática de todos los modelos generados en el modelamiento; se complementa la verificación de los modelos con la medida de **Back-Testing,** el cual se llamara CR, notación de crecimiento. Este Indicador mide la variación porcentual de los valores estimados versus los resultados reales de cada uno de los modelos, lo que significa que cada serie se le quitan los tres últimos registros reales y se pronostican con todo el modelo realizado.

Con el fin de identificar si el modelo entiende los datos con los que fue entrenado, una vez predicho los tres valores a través del modelo se calcula un promedio de valores porcentuales. Entre más cercano a cero indica que la técnica es la que entiende el comportamiento histórico del número de hurtos cometidos a personas y sería un candidato para seleccionar como ganador.

A continuación se realiza un ejemplo:

ID-SERIE	MES REAL 1	MES REAL 2	MES REAL 3	MOD.SIMPLE MES 1	MOD.SIMPLE MES 2	MOD.SIMPLE MES 3
ID3	1690,713	2394,074	1392,764	1534,963	2302,444	2302,444

Tabla 3 Valores Back-Testing

$$CR_{Modelo\ Simple} = \frac{\frac{\text{Simple Mes 1} - \text{Mes Real 1}}{\text{Mes Real 1}} + \frac{\text{Simple Mes 2} - \text{Mes Real 2}}{\text{Mes Real 2}} + \frac{\text{Simple Mes 3} - \text{Mes Real 3}}{\text{Mes Real 3}}}{3}$$

En este ejemplo el resultado del CR:

$$CR_{Modelo\ Simple} = 0.261$$

Este valor es calculado para todos los modelos y almacenado como un nuevo parámetro de verificación para cada serie.

Hasta este punto de la automatización en IBM SPSS Modeler, se cuenta con información almacenada del R2, MAE y CR de cada modelo, a partir de este punto debemos seleccionar el mejor modelo entre todas las combinaciones realizadas, por lo anterior se crea un score que valide el cumplimiento de los tres indicadores.

Score del Modelo para selección de mejor modelo

Como el proceso o flujo debe validar y seleccionar automáticamente la técnica ganadora, se diseña un escore a partir de un puntaje ponderado para cada modelo. Para crear el score se deben de tener los resultados de los tres parámetros en la misma escala, para lo cual se decide llevar los valores a una escala de números entre 0 y 1.

Como los valores de R2, MAE y CR se encuentra en diferente escala y tienen diferente interpretación, es necesario estandarizar las medidas para utilizarlos dentro del score y crear un puntaje por cada técnica de serie de tiempo.

R^2	Puede tomar un valor entre 0 y 1. Un valor pequeño indica que el modelo no se ajusta bien a los datos.
Error Absoluto Medio	Mide la desviación de la serie del nivel pronosticado por el modelo. Entre más cercano a 0 significa que existe menos error entre la serie real y la pronosticada.
Back-Testing	Valor que mide la variación promedio porcentual; Entre más cercano a 0 significa que existe menos variación de los meses pronosticados y los meses reales.

Tabla 4 Interpretación de parámetros de validación

Para que los valores se encuentren con la misma escala se usa la técnica de escalamiento lineal (Actis di Pasquale & Balsa, 2017), con el fin de que las medidas queden estandarizadas de 0 a 1. Esta técnica utiliza los valores mínimos y máximos de una variable y calcular el valor con base en la siguiente formula:

$$I_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Donde L_i es el valor en la nueva escala; X_i es el valor en la escala original; X_{max} es el valor máximo de la escala; X_{min} es el valor mínimo de la escala.

El score o puntaje de selección será una medida que toma un valor ente 0 y 1, lo que significa que si el valor este más cercano a 1 o por encimas del 0,8 (se define como el parámetro de corte para la selección del modelo) será el modelo seleccionado o ganador de la competencia. Para la creación del puntaje se dará una importancia o peso a cada uno de los parámetros, lo que significa que las tres medidas deben de tener un factor de peso que al sumarlo den el 100%.

El R^2 recibe el mayor peso porque el objetivo principal al construir un modelo es que este se ajuste a los datos históricos y pueda capturar su comportamiento. El MAE tiene una importancia secundaria, luego su enfoque es de precisión de las predicciones individuales. El Back-Testing es crucial para garantizar que el modelo no solo sea bueno en comprender los datos históricos, sino que también pueda predecir eficazmente los datos futuros.

Los pesos asignados a cada métrica (50% a R2, 20% al MAE y 30% al Back-Testing) se fundamenta en la importancia relativa de cada métrica para el objetivo de predicción del proyecto. Estos pesos se distribuyen de manera lógica según la prioridad del ajuste y la capacidad de predicción en el futuro, y podrían ser ajustados con base en experimentación o análisis adicionales como el cálculo a través de técnicas como PCA. se realizar un análisis de sensibilidad o pruebas de simulación en los cuales se varían los pesos y no se evidencia un impacto significativo en la selección del modelo.

MEDIDA	PORCENTAJE PONDERADO
R^2	50%
ERROR ABSOLUTO MEDIO (MAE)	20%
BACK-TESTING (CR)	30%
TOTAL	100%

Tabla 5 Importancia o peso del parámetro

Por lo anterior se automatiza la siguiente ecuación para calcular el score de selección de la técnica ganadora:

$$Score\ Modelo = (R^2 * 0.5) + (MAE * 0.2) + (CR * 0.3)$$

A continuación se visualiza la automatización del proceso en IBM SPSS Modeler:

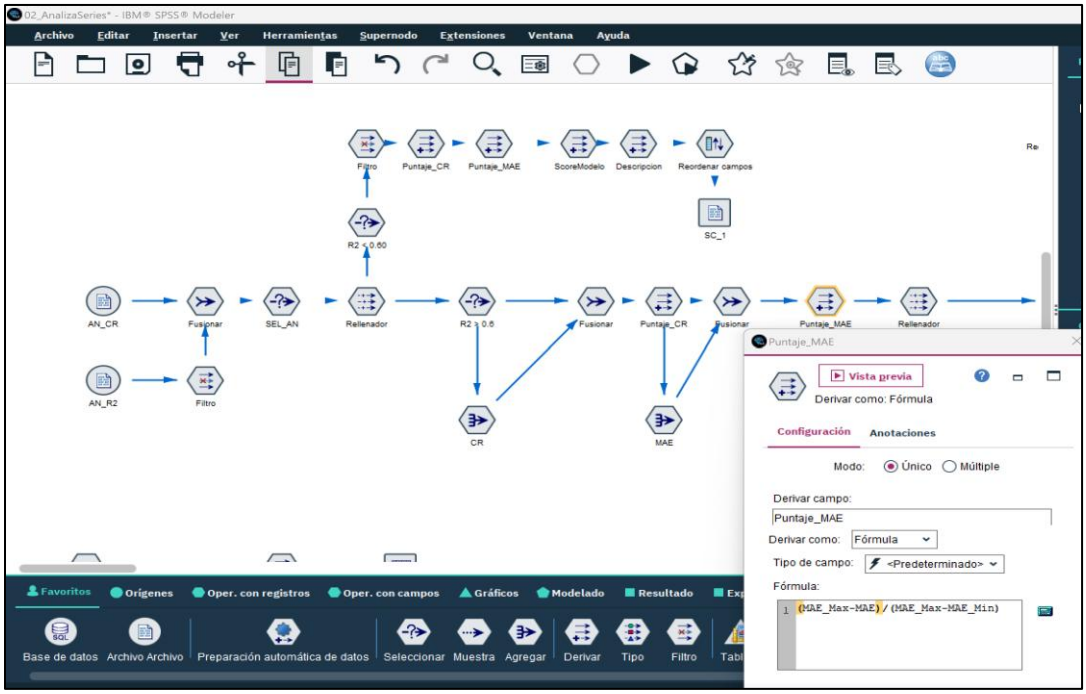


Ilustración 17 . Ruta para la selección del modelo

Dentro del flujo se muestra a continuación la fórmula de cálculo del score de selección del modelo.

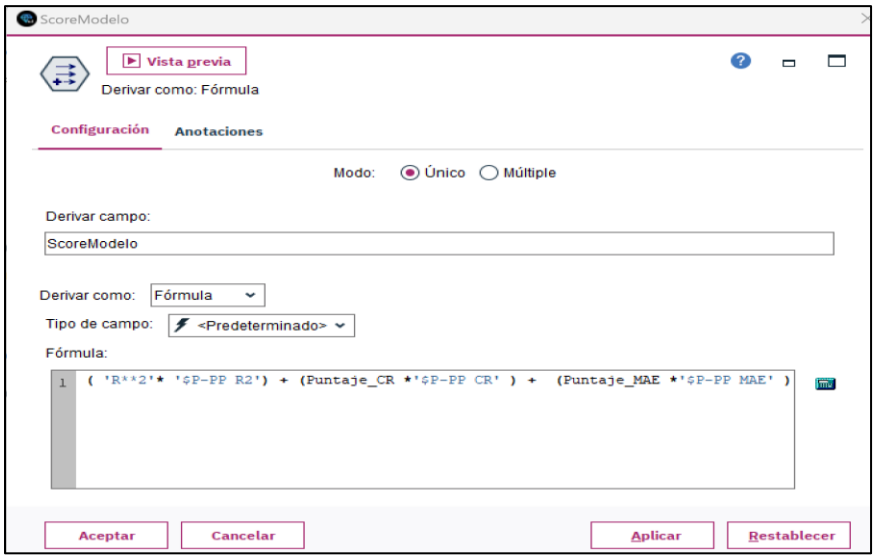
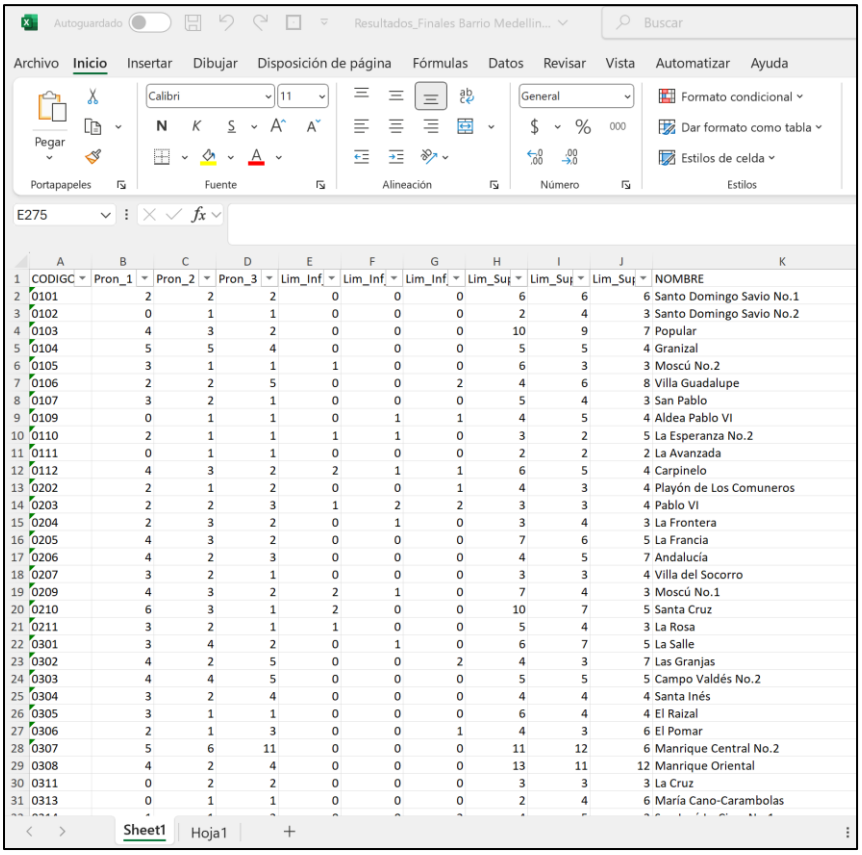


Ilustración 18 Cálculo del score para cada modelo

Fase VI. Despliegue de resultados

Una vez se tiene la técnica de modelamiento de serie de tiempo seleccionada, se debe generar la proyección para cada granularidad a nivel de ciudad, comuna y barrio para los meses de octubre, noviembre y diciembre del 2024, Los pronósticos quedan

almacenados en un archivo de Excel, el cual tiene el valor pronostico, el intervalo de confianza para el pronóstico.



CODIGO	Pron_1	Pron_2	Pron_3	Lim_Inf	Lim_Inf	Lim_Inf	Lim_Sup	Lim_Sup	Lim_Sup	NOMBRE
0101	2	2	2	0	0	0	6	6	6	Santo Domingo Savio No.1
0102	0	1	1	0	0	0	2	4	3	Santo Domingo Savio No.2
0103	4	3	2	0	0	0	10	9	7	Popular
0104	5	5	4	0	0	0	5	5	4	Granizal
0105	3	1	1	1	0	0	6	3	3	Moscú No.2
0106	2	2	5	0	0	2	4	6	8	Villa Guadalupe
0107	3	2	1	0	0	0	5	4	3	San Pablo
0109	0	1	1	0	1	1	4	5	4	Aldea Pablo VI
0110	2	1	1	1	1	0	3	2	5	La Esperanza No.2
0111	0	1	1	0	0	0	2	2	2	La Avanzada
0112	4	3	2	2	1	1	6	5	4	Carpinelo
0202	2	1	2	0	0	1	4	3	4	Playón de Los Comuneros
0203	2	2	3	1	2	2	3	3	4	Pablo VI
0204	2	3	2	0	1	0	3	4	3	La Frontera
0205	4	3	2	0	0	0	7	6	5	La Francia
0206	4	2	3	0	0	0	4	5	7	Andalucía
0207	3	2	1	0	0	0	3	3	4	Villa del Socorro
0209	4	3	2	2	1	0	7	4	3	Moscú No.1
0210	6	3	1	2	0	0	10	7	5	Santa Cruz
0211	3	2	1	1	0	0	5	4	3	La Rosa
0301	3	4	2	0	1	0	6	7	5	La Salle
0302	4	2	5	0	0	2	4	3	7	Las Granjas
0303	4	4	5	0	0	0	5	5	5	Campo Valdés No.2
0304	3	2	4	0	0	0	4	4	4	Santa Inés
0305	3	1	1	0	0	0	6	4	4	El Raizal
0306	2	1	3	0	0	1	4	3	6	El Pomar
0307	5	6	11	0	0	0	11	12	6	Manrique Central No.2
0308	4	2	4	0	0	0	13	11	12	Manrique Oriental
0311	0	2	2	0	0	0	3	3	3	La Cruz
0313	0	1	1	0	0	0	2	4	6	María Cano-Carambolas

Ilustración 19 . Resultado pronostico por código de barrio

5 ANÁLISIS DE RESULTADOS

El proyecto inició con un análisis exploratorio de los datos, con el fin de conocer y comprender el fenómeno del hurto a una persona, a continuación se analiza los resultados predictivos para algunas de las zonas identificadas como vulnerables y a nivel de toda la ciudad.

Para la ciudad de Medellín se estima para los meses octubre de 2024 un total de 1.880 hurtos a personas, mes de noviembre 2024: 1.996 hurtos y para el cierre del año, diciembre 2024 con 1.957 casos a cubrir por los entes de seguridad de la ciudad, esta proyección estima un pequeño decrecimiento con relación al último trimestre.

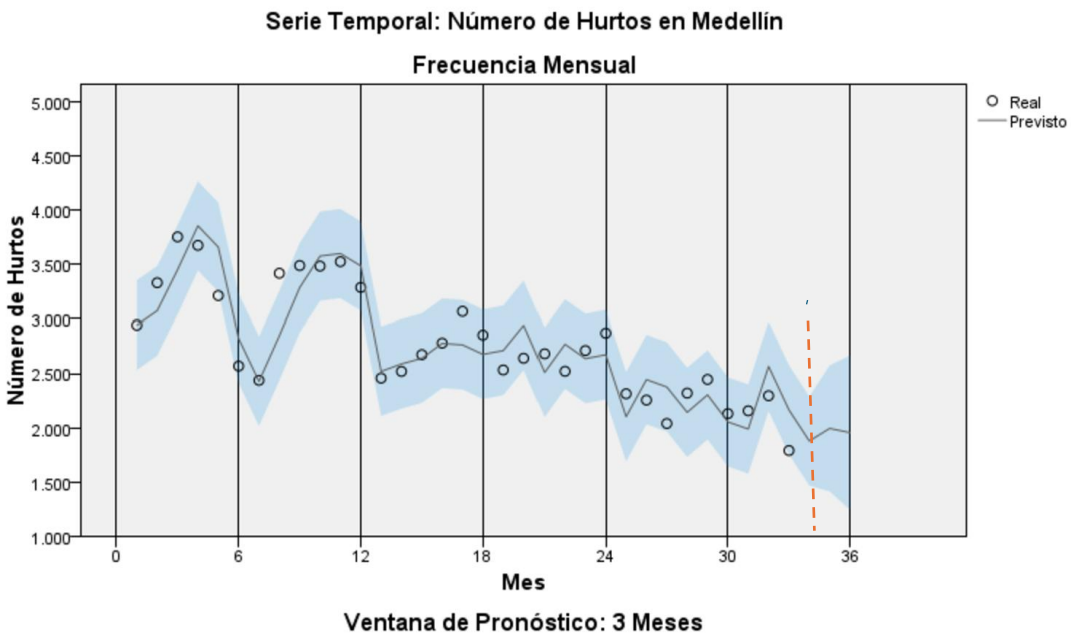


Ilustración 20 Pronóstico para la ciudad de Medellín

Información del modelo

Método de generación de modelos		Suavizado exponencial Aditivo Winters
Número de predictores		1
Ajuste del modelo	MSE	58.548,253
	RMSE	241,967
	RMSPE	8,679
	MAE	192,472
	MAPE	7,168
	MAXAE	571,013
	MAXAPE	20,878
	AIC	365,116
	BIC	369,605
	R cuadrado	0,792
	R cuadrado estacionario	0,670
Prueba de Ljung-Box (número)	Estadístico	31,649
	gl	15,0
	Significancia	0,0

El modelo predictivo ganador de la competencia de modelos de series temporales es la serie original sin descomposición de desviaciones, con la técnica de suavizamiento exponencial Aditivo de Winters, el cual genero un escore de selección del 85% de cumplimiento en sus parámetros. Para el ajuste del modelo se muestra que el valor medio de su error (MAPE) está al rededor del 7,2 %.

Las proyecciones para la ciudad de Medellín, con relación al trimestre anterior se visualiza un decrecimiento de la ocurrencia del delito; donde se recomienda una gestión diferente con la fuerza pública y fomentar una mayor cultura de precaución entre las personas; el valor mínimo del número de hurtos a personas para el trimestre se estima que sea 1.245 casos.

Es importante recalcar a la ciudadanía las recomendaciones de la Policía Nacional, como solicitar acompañamiento al realizar retiros o transporte de altas sumas de dinero en efectivo, es una de las calves que permite mitigar el riesgo de presentar un hurto, y que podría comprometer la vida de la persona. Según los datos de los últimos tres años, el medio de transporte más común en el momento de los hurtos es a pie o

caminando, siendo este escenario el principal para hurtar grandes cantidades de dinero.

A continuación se muestras las estadísticas descriptivas analizadas con las variables demográficas y monetarias.

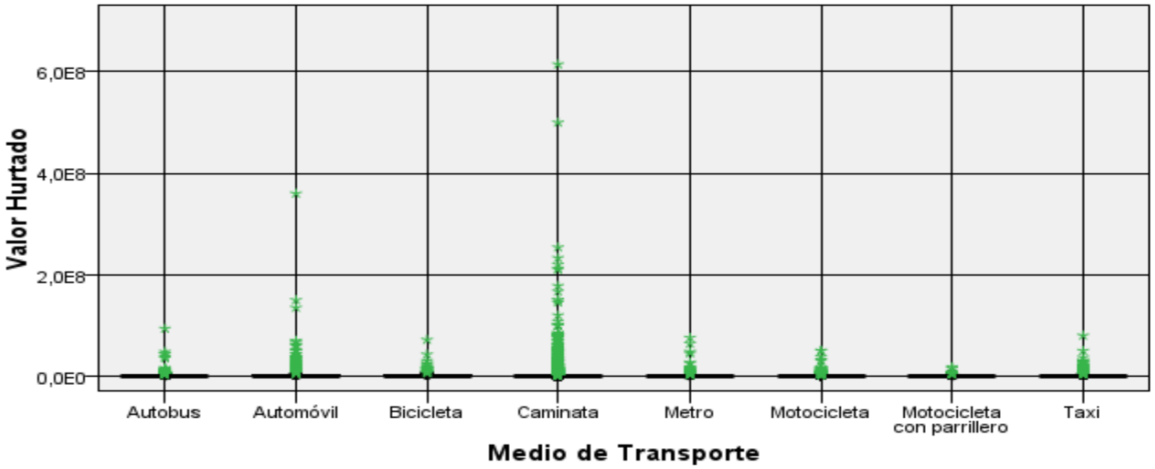


Ilustración 21 Diagrama de cajas con atipicidades de dinero por medio de trasporte del hurto

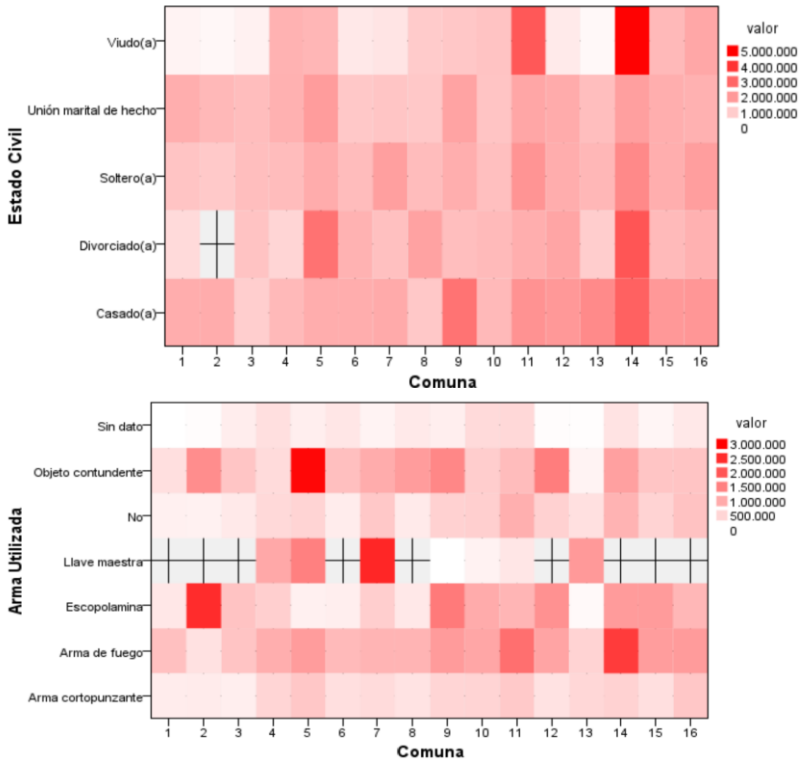


Ilustración 22 Mapas de calor del valor hurtado por las comunas de la ciudad

El Poblado es una de las comunas más representativas de la ciudad, se destaca como una de las comunas con mayor incidencia de hurtos a personas con armas de fuego, donde cada hurto puede estar representando una pérdida de alrededor de 3 millones de pesos para la víctima. Las festividades de diciembre representan una temporada ideal para llevar a cabo jornadas de sensibilización, enfocadas en promover medidas de seguridad como evitar la exhibición de objetos de valor en público, abstenerse de transportar altas sumas de dinero sin acompañamiento, preferir transitar por lugares concurridos y especialmente colaborar con las autoridades mediante la denuncia oportuna de cualquier irregularidad.

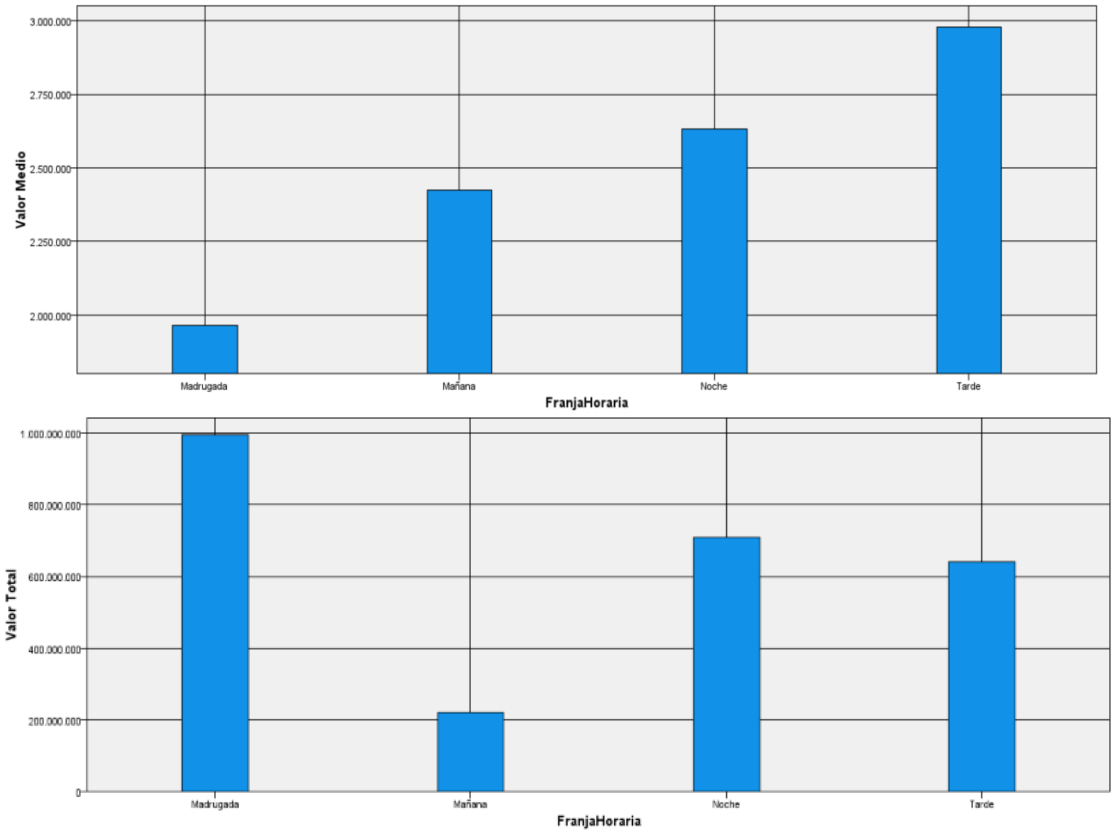


Ilustración 23Valor total y promedio de dinero Hurtado a personas

Al analizar los hurtos registrados en la comuna de El Poblado según jornadas horarias madrugada (00:00-06:00 a.m.), mañana (06:01-11:59 a.m.) y tarde (12:00-6:00 p.m.), se observa que la tarde es el periodo de tiempo con más hurtos, con un impacto económico, acumulando más de 600 mil millones de pesos hurtados en los últimos tres años. Cada hurto a persona en este horario a representado en promedio, una pérdida de 3 millones de pesos a cada víctima.

A continuación se visualiza el comportamiento histórico y proyección para la Comuna 14 el Poblado:

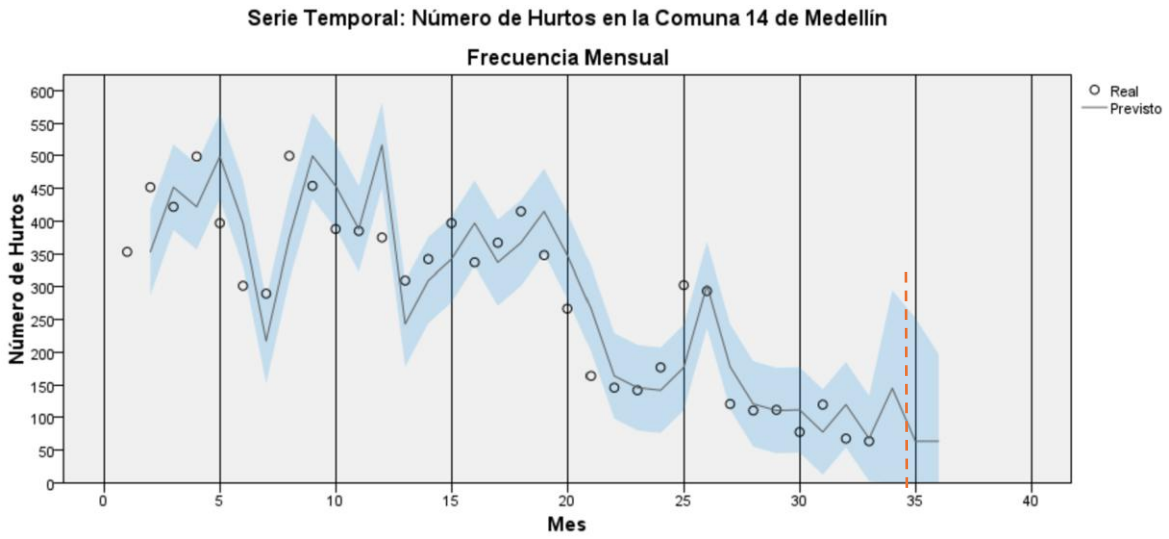


Ilustración 24 Pronóstico de hurtos poblado

145	64	65
Octubre	Noviembre	Diciembre

Tabla 6 Pronóstico comuna Poblado

Modelo generado por un ARIMA (1,1,2) y un score de selección 89% de eficiencia en la validación de los parámetros, las estimaciones proyectan un crecimiento en el número de hurtos para el mes de octubre, finalizando el año con un estimado de 60 casos, esto se puede asumir a las estrategias de incremento de la presencia policial durante la temporada navideña, se recomienda reforzar la vigilancia en áreas comerciales y residenciales, lo que conlleva a persuadir a los delincuentes y generar confianza en la ciudadanía. Cuando se analiza dentro de la comuna la zona rosa o espacio más popular de la comuna representada como el barrio (1418):

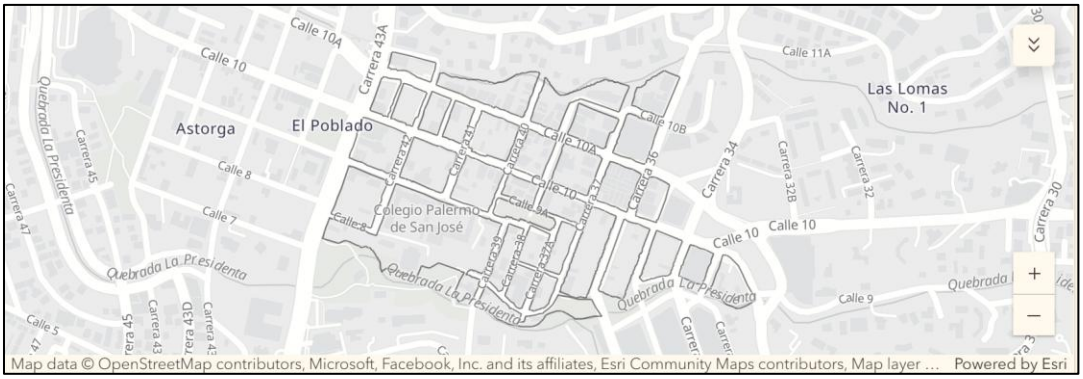


Ilustración 25 Ubicación geográfica del barrio poblado

El poblado es una zona con una percepción de mayor capacidad económica y alta afluencia de turistas; las proyecciones nos muestran un crecimiento para los próximos tres meses, se estima un máximo de números de hurtos a personas en los tres meses de 184 casos. La concentración de personas en centros turísticos facilita a los delincuentes mezclarse con la multitud y cometer hurtos sin ser detectados. Se recomienda a la ciudadanía estar muy atentos a las compras por festividades, pagos de salarios y primas; luego el mes de diciembre es la oportunidad para intensificar la vigilancia para incentivar la tranquilidad.

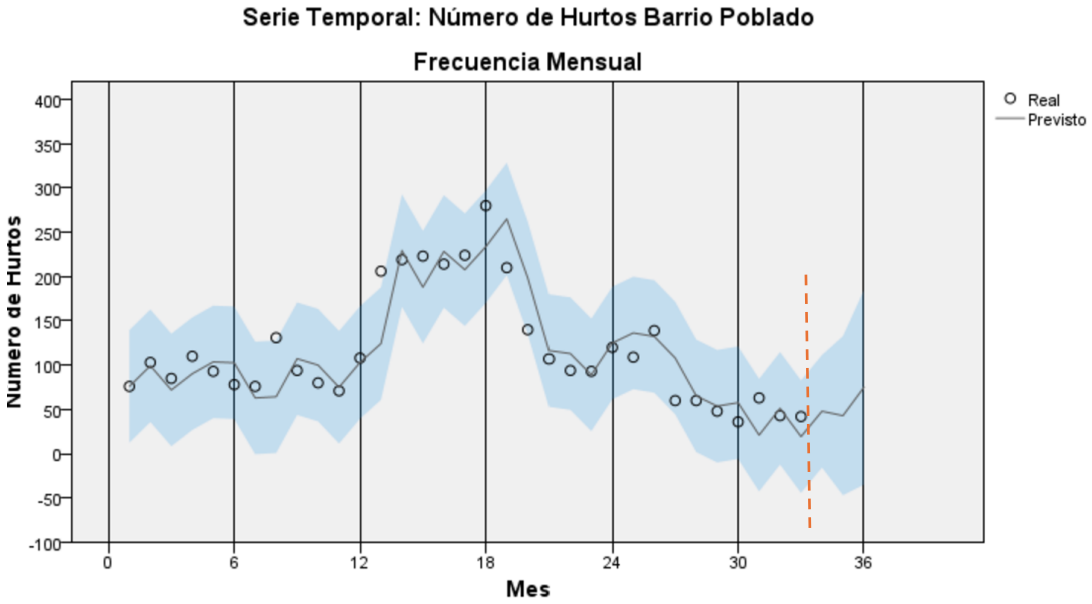


Ilustración 26 Pronóstico de hurtos barrio poblado.

75	43	48
Octubre	Noviembre	Diciembre

Tabla 7 Pronóstico barrio Poblado

Al observar el barrio la candelaria se estima con un crecimiento sostenido de más de 100 caso de hurtos a personas a en cada mes; el cual se vuelve uno de los focos estratégicos de patrullaje constante por parte de la fuerza pública, frecuentar puntos de control y vigilancia en horarios específicos.

BARRIO	AÑO 2024	PRONOSTICO	LIMETI INFERIOR	LIMETI SUPERIOR
LA CANDELARIA	Octubre	155	1	350
	Noviembre	158	2	317
	Diciembre	124	12	237

Tabla 8 Pronóstico barrio la Candelaria

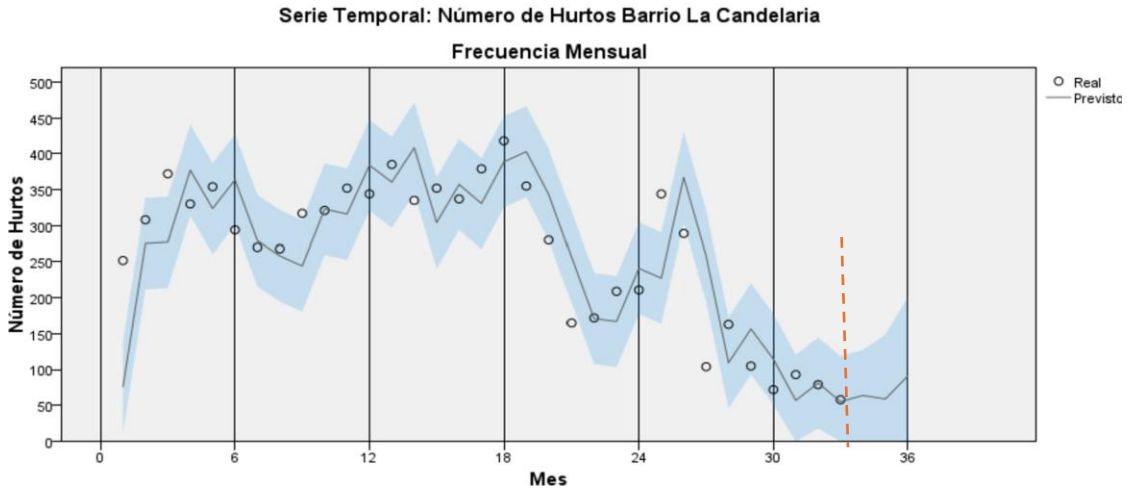


Ilustración 27 Pronóstico para el barrio candelaria

Es fundamental que tanto las autoridades como la comunidad mantengan una colaboración permanente para seguir disminuyendo la incidencia de hurtos y mejorar la percepción de seguridad en La Candelaria y en toda la ciudad. Así como se analizaron algunos puntos importantes de la ciudad que siempre han sido catalogados como peligrosos, en el análisis se identifican algunos barrios que se estiman con valores para cada mes de 2 hurtos a reportar, los culés son: Santo Domingo Savio No.2, Aldea Pablo VI, La Avanzada, María Cano-Carambolas, Belalcázar, Alfonso López, Santander, Villa Flora, Villatina, Villa Lilliam, Los Cerros El Vergel, Cataluña, Santa Rosa de Lima, Metropolitano, Juan XXIII La Quiebra, El Corazón, Nuevos Conquistadores, Lalinde, La Hondonada, La Palma, El Picacho, El Jardín y El Salado.

A continuación se identifica el top 10 de los barrios con mayor número pronosticados de hurtos a personas, para el mes de diciembre del 2024:

CODIGO	NOMBRE	PRON	LIM_INF	LIM_SUP	MODELO SERIE TEMPORAL
1019	LaCandelaria	124	12	237	Suavizado Exponencial
1418	El Poblado	48	0	111	Suavizado Exponencial
1001	Prado	30	8	52	Holt's Linear Trend Model
1108	Laureles	30	0	73	Holt-Winters Aditivo
1113	Estadio	26	0	67	Holt-Winters Aditivo
1603	Belén	26	0	62	Suavizado Exponencial Simple (SES)
413	Aranjuez	17	0	53	ARIMA (AutoRegressive Integrated Moving Average)
907	Buenos Aires	17	0	38	Suavizado Exponencial
717	Robledo	16	0	39	ARIMA (AutoRegressive Integrated Moving Average)
1007	Guayaquil	13	0	36	SARIMA (Seasonal ARIMA)

Tabla 9 Top 10 de los barrios con mayor predicción para diciembre 2024

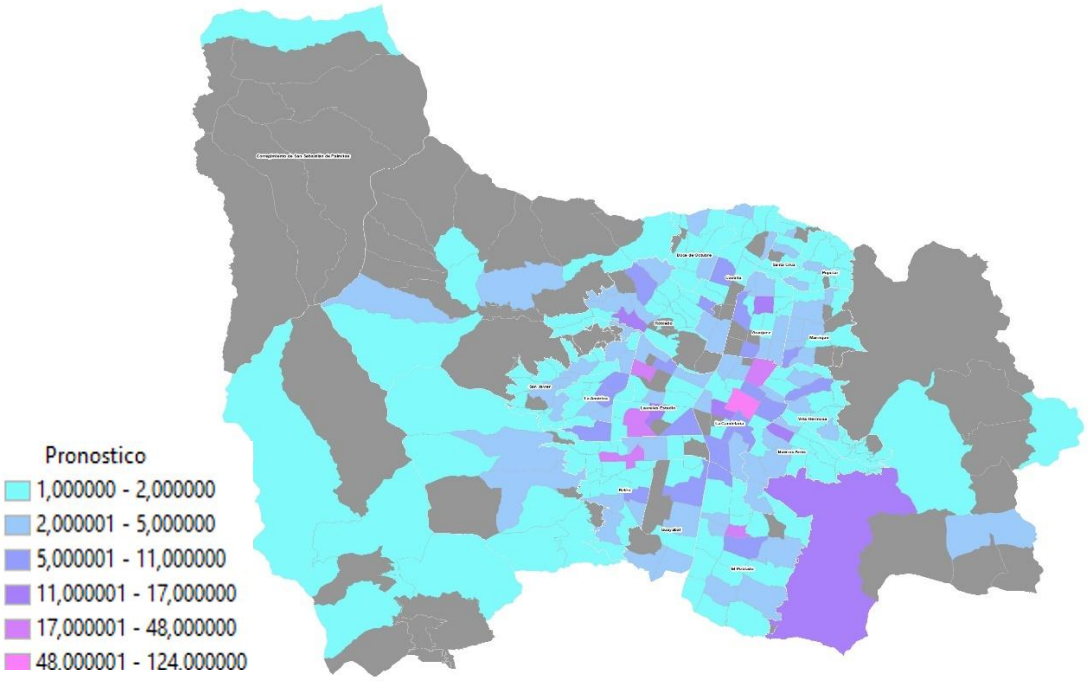


Ilustración 28 Pronóstico por barrios de Medellín mes diciembre 2024

6 CONCLUSIONES Y TRABAJO FUTURO

Para culminar el análisis, podemos evidenciar que el último trimestre del año las comunas con una mayor proyección de número de casos de hurtos a personas son Aranjuez, La candelaria, Belén y el poblado; dando una priorización a la fuerza policial y locales. La disminución del número de hurtos a personas requiere un enfoque colaborativo entre la ciudadanía y las autoridades, las personas deben adoptar medidas de seguridad como autoprotección, atención al entorno, reducción de la exposición de objetos de valor; por otro lado, las autoridades tienen la responsabilidad de implementar estrategias efectivas, como el incremento de la vigilancia, el uso de tecnología, el seguimiento a las alertas derivadas de las estrategias y la colaboración comunitaria para este fin de año.

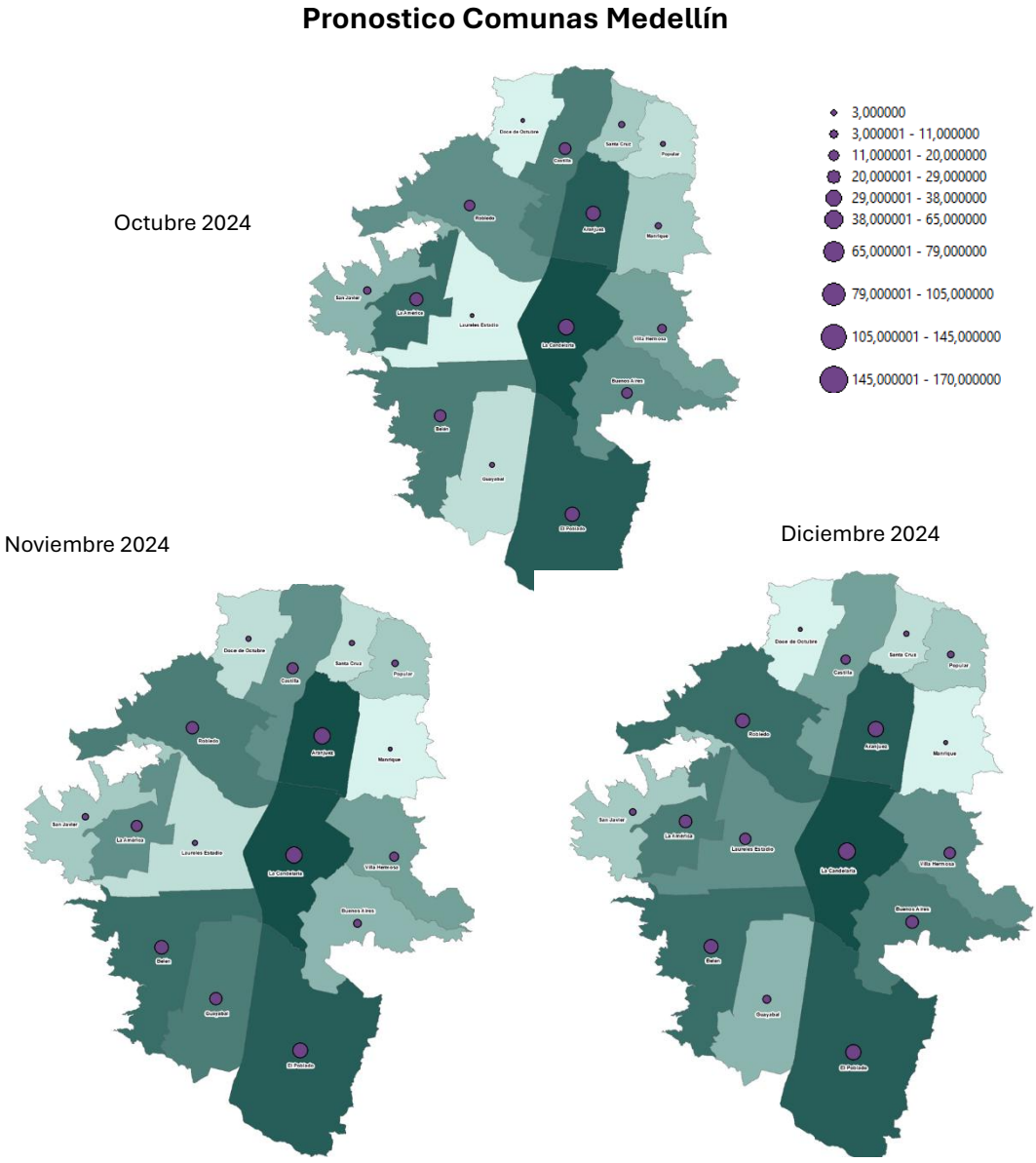


Ilustración 29 Pronóstico por comunas de Medellín

Para los meses de análisis las medidas adoptadas para mitigar el hurto en las comunas de Manrique, san Javier, Doce de octubre y Santa cruz, deben de continuar para generar un entorno seguro para los ciudadanos.

Como parte de este análisis se identifica que para una planeación preventiva que anticipe tendencias delictivas, es de vital importancia la implementación de técnicas de series temporales, el cual permite tener estadísticas futuras que apoyen a la fuerza policial y locales de Medellín en la optimización de los recursos disponibles, identificando por ejemplo la comuna de Guayabal como una zona con atipicidades no constantes pero si por ejemplo para el mes de noviembre generar grupos de vigilancia,

aumentar la presencia de policía, y aumentar el patrullaje en barrios como Santa Fe y Guayabal; lo anterior permitirá generar un impacto positivo en la percepción de seguridad y la calidad de vida de la ciudadanía.

El análisis realizado con las técnicas de series temporales se proyecta como una metodología adecuada para identificar los patrones delictivos, los cuales permitieron evaluar las tendencias históricas y proyectar el número de casos de hurtos a personas, con una granularidad mensual.

La construcción del activo analítico permite recalibración automática en cada ejecución, pero uno de los puntos a mejorar es el score de selección del mejor modelo a utilizar para la proyección de la serie, donde se propone como una nueva etapa o mejora del flujo. Se identifica como importante incluir las medidas de validación del BIC y el AIC para cada modelo de serie temporal, el cual genera mayor confianza en el momento de la selección; así como la validación de la significancia adecuada de los parámetros de la ecuación de la serie temporal. Lo anterior permite el ingreso de variables explicativas o exógenas al flujo metodológico creado, el cual será de gran utilidad para entender el comportamiento de los hurtos a personas con covariables, lo que significa que la automatización realizada va a permitir la ejecución del análisis con variables dependiente y dejar de ser univariado. IBM SPSS Modeler permitió construir un producto mínimo viable que estima una variable numérica, lo que significa la reutilización de la ruta o activo analítico para otro proyecto.

La herramienta no solo proporciono una solución avanzada para analizar los datos, sino que también aseguro un trabajo visualmente claro y escalable que se puede convertir en una solución del portafolio comercial de la organización en la que me encuentro laborando.

7 REFERENCIAS

Actis di Pasquale, E., & Balsa, J. (2017). La técnica de escalamiento lineal por intervalos. Revista de métodos cuantitativos para la economía y la empresa, 164-193.

Alcaldia de Medellin, GeoMedellin, https://www.medellin.gov.co/geomedellin.

Alcaldia de Medellin, Medellin como vamos, seguridad y Convivencia, https://www.medellincomovamos.org/sectores/seguridad-y-convivencia

Alcaldia de Medellin, Geo Medellin, estadisticas avanzadas, https://m-medellin.maps.arcgis.com/apps/dashboards/62f95753163348cf8bb537051cc3ed9b

Alcaldia de Medellin, Medata datos abiertos de la ciudad de Medellin, http://medata.gov.co/medell%C3%ADn-en-cifras/hurtos-y-capturas-2003-%E2%80%93-2018

Alcaldia de Medellin, Medata datos abiertos hurtos a personas reportado en la ciudad de Medellin, https://medata.gov.co/dataset/hurto-persona

Alcaldía de medellin, (Política Pública de Seguridad y Convivencia del Municipio de Medellín, 2015), https://drive.google.com/file/d/1nckqSrvSP75vdNIEWNtKAZRG8tV1Kz27/view

Brit. J. Criminol. (2004) 44, The Future of Crime Mapping 641–658 Advance Access publication 7 May 2004.

Colombia es ciencia, (Modelos de predicción: La seguridad, un nuevo reto de la tecnología, 2024) https://colombiaesciencia.minciencias.gov.co/content/modelos-de-predicci%C3%B3n-la-seguridad-un-nuevo-reto-de-la-tecnolog%C3%ADa

Dúber Cano Aguirre, 2024, La lucha contra el hurto en Medellín da resultados contundentes con reducción histórica del 20 %, https://www.medellin.gov.co/es/sala-de-prensa/noticias/la-lucha-contra-el-hurto-en-medellin-da-resultados-contundentes-con-reduccion-historica-del-20/

Gujarati and Porter, Fifth Edition 2010. Basic Econometrics, editorial McGraw-Hill

Hyndman, R., & Athanasopoulos, G. (2018, Abril). Pronóstico: principios y práctica. Retrieved from https://otexts.com/fpp2/

IBM, International Business Machines. (2016). Nodos de modelado de IBM SPSS, Modeler 17. Obtenido de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.0/es/ModelerModelingNodes.pdf.

IBM, Modelos de series de tiempo https://www.ibm.com/docs/es/spss-modeler/saas?topic=SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/timeseriesnode_general.htm.

Jose Alberto Mauricio, introduction al analysis temporal, marzo 2007, [En línea], https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf.

Julio Cesar Alonso (2020), Introducción a los pronósticos con modelos estadístico de series de tiempo para científico dedatos (en R).

LEY 2197 de 2022, Poder público – Rama legislativa, 25 de enero de 2022, http://www.secretariasenado.gov.co/senado/basedoc/ley_2197_2022.html#:~:text=Art%C3%ADculo%20239.,ciento%20ocho%20(108)%20meses.

Miller Jimmy Alarcón, Calificación del metodo de pronostico, segunda parte 2009 Poliantea.

Peña, 2025, Análisis de series temporales, editorial Alianza. https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf

Policía nacional de Colombia, información de criminalidad: https://www.policia.gov.co/grupo-informaci%C3%B3n-criminalidad/estadistica-delictiva.

Policía nacional de Colombia, normatividad hurtos: https://www.policia.gov.co/denuncia-virtual/normatividad-hurto

Rafael zambrano, Un enfoque espaciotemporal para la predicción de delitos en la ciudad de Buenos Aires, Revista de investigación en modelos matematicos aplicados a la gestion y la economia - año 7 volumen II (2020-II).

Ratnadip Adhikari, An Introductory Study on Time Series Modeling and Forecasting, https://arxiv.org/pdf/1302.6613

Sara Sendino, Un algoritmo predice los crímenes que van a suceder con una semana de antelación, publicacion Julio 2022. https://www.lasexta.com/tecnologia-tecnoxplora/ciencia/algoritmo-predice-crimenes-que-van-suceder-semana-antelacion_2022070662c5a4630ff4480001e467d4.html

Universidad de los Andes Prediciendo el crimen en Bogotá, 2020, https://www.uniandes.edu.co/es/noticias/economia-y-negocios/modelos-matematicos-para-predicir-el-crimen-en-bogota