PREDICCIÓN DEL MONTO TOTAL QUE SE VA A PAGAR POR REMESAS EN DOLARES QUE SE ORIGINAN EN UN DÍA

William Contreras Fuentes Cod 8992942.
Camilo Espinoza Guarnizo Cod 8992708.
Jorge Agredo Chávez Cod 8993411.

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Directora
María Constanza Pabón

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 09 DE 2025

TABLA DE CONTENIDO

Contenido

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

INTRODUCCIÓN

Las remesas son transferencias de dinero realizadas principalmente por emigrantes a sus familiares o relacionados en su país de origen [1]. Para 2024, representaron una parte significativa de la balanza de pagos de Colombia, aportando al crecimiento económico y alcanzando un máximo histórico del 2.8% del Producto Interno Bruto (PIB), año en el que superaron los USD $11.848 millones [2].

En el proceso operativo de las remesas participan varios actores: el remitente, el agente en el país de origen (APO), la empresa de transferencia de dinero (ETD), el agente pagador en el país de destino (APPD) y el receptor. Las transacciones entre la ETD y los agentes pagadores en Colombia se realizan en dólares, y se utiliza una tasa de cambio fija definida en la fecha de origen de cada transacción. Esto implica que todas las remesas iniciadas en un mismo día se liquidan bajo la misma tasa, sin importar cuándo se realice el cobro efectivo por parte del beneficiario.

En Colombia, la empresa de transferencia opera con dos agentes pagadores. Esta dinámica genera una doble incertidumbre operativa: por un lado, la exposición a variaciones en la tasa de cambio entre la fecha de origen y la fecha de pago, y por otro, la imposibilidad de prever con precisión qué porcentaje del monto originado será efectivamente pagado por cada APPD. Dado que el receptor puede acudir a cualquiera de ellos, los montos liquidados pueden variar significativamente, afectando los márgenes financieros de los agentes.

En este contexto, el presente trabajo se propuso desarrollar modelos predictivos que permitan estimar con antelación el monto total en dólares de remesas originadas en un día determinado, utilizando datos históricos operativos y variables adicionales como tasa de cambio, país de origen, características temporales (día, semana, festivos), entre otras. La metodología se basó en el enfoque CRISP-DM y exploró técnicas como modelos estadísticos clásicos, redes neuronales recurrentes y algoritmos de aprendizaje automático.

Como adelanto a los principales hallazgos, se destaca que los modelos de redes neuronales recurrentes y árboles de decisión optimizados ofrecieron los mejores niveles de precisión predictiva, alcanzando R^2 de hasta 87.32 % y 99.58 % respectivamente. En contraste, modelos más tradicionales como ARIMA o SVM mostraron un desempeño limitado. Estos resultados confirman que la incorporación de arquitecturas capaces de capturar relaciones no lineales y patrones secuenciales mejora sustancialmente la capacidad de pronóstico en contextos operativos complejos.

Los modelos desarrollados tienen un alto potencial para ser utilizados por el APPD en la gestión del riesgo cambiario, la planificación de liquidez y la optimización de procesos de monetización, lo que justifica la relevancia y aplicabilidad de este estudio.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

Las remesas son transferencias de dinero realizadas por determinados grupos de personas, entre los cuales se encuentran, principalmente, emigrantes que envían fondos a sus familiares en su país de origen. En la actualidad, estas transferencias representan una parte importante de la balanza de pagos de Colombia, alcanzando en 2024 un 2.8% del PIB [3].

Dado el creciente peso de las remesas en la economía colombiana, su estudio no solo es relevante desde una perspectiva macroeconómica, sino también desde el entendimiento de cómo operan estos flujos. Comprender en detalle cómo se ejecuta el proceso de envío, recepción y pago de las remesas permite identificar puntos críticos que afectan la eficiencia y sostenibilidad del sistema. En este proyecto, el interés se centró en el análisis de las remesas que llegan a Colombia y en el diseño de herramientas analíticas que permitan gestionar el riesgo cambiario derivados de este proceso.

Proceso operativo de remesas: En el proceso operativo de las remesas intervienen cinco actores principales.

- **Remitente:** Persona que origina la remesa y a quien se le informa el valor en moneda local que recibirá el receptor.
- **Agente en país de origen (APO):** Entidad encargada de recibir los fondos del remitente, informar las condiciones de la operación (costos y tasas), generar y comunicar el código de la transacción.
- **Empresa de transferencia de dinero (ETD):** Compañía que intermedia entre los agentes de ambos países y administra el flujo de fondos, incluyendo la fijación de tasas de cambio.
- **Agente pagador en el país destino (APPD):** Entidad en el país de destino (Colombia) responsable de entregar los fondos al receptor en moneda local.
- **Receptor:** Persona beneficiaria de la remesa, quien recibe el dinero en la moneda local.

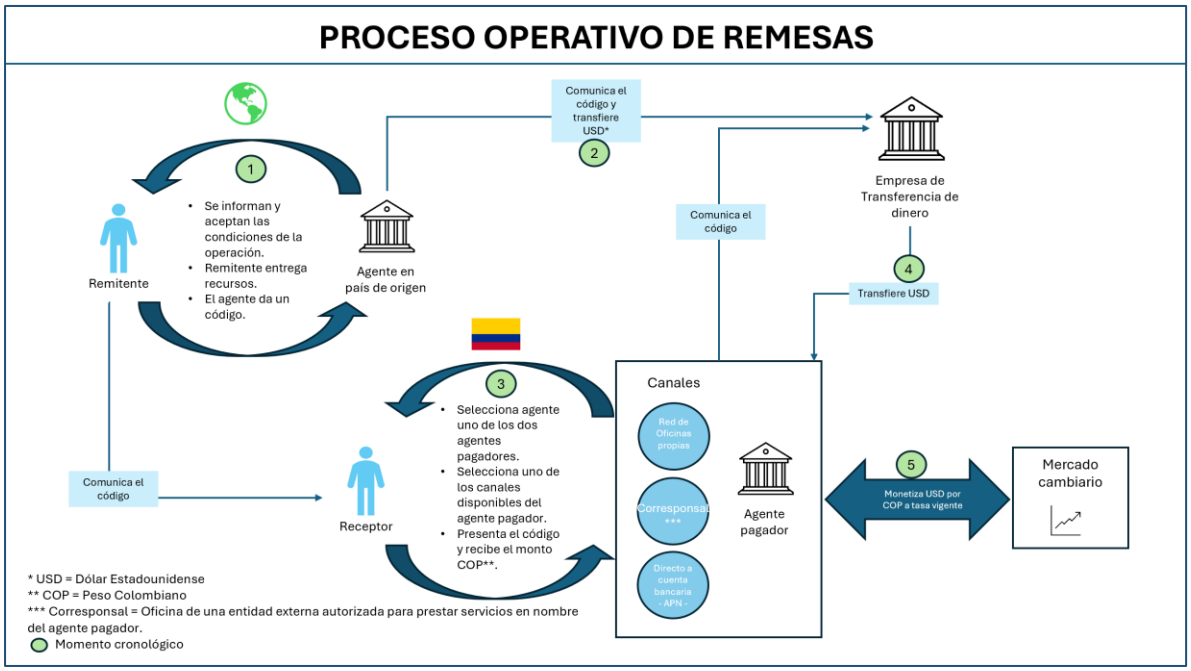

Ilustración 1 Proceso de operativo de remesas

El proceso descrito en la ilustración 1 puede dividirse cronológicamente en cinco momentos:

1. **Momento inicial:** El remitente acuerda con el APO las condiciones de la transacción, entrega los fondos y recibe un código que deberá comunicar al receptor.
2. **Liquidación en el país de origen:** El APO transfiere los fondos en dólares a la ETD.
3. **Cobro en el país de destino:** El receptor escoge uno de los dos APPD en Colombia, presenta el código recibido y cobra la remesa.
4. **Liquidación en el país de destino:** El APPD informa a la ETD que la remesa ha sido pagada, y esta transfiere los dólares correspondientes al agente.
5. **Monetización:** El APPD convierte los dólares de la remesa en pesos colombianos en el mercado cambiario.

Es importante destacar que los momentos 1 y 2 ocurren el mismo día, mientras que los momentos 3 y 4 dependen de la fecha en la que el receptor decida cobrar la remesa. Actualmente, existen dos APPD en Colombia, por lo que la ETD no puede determinar con certeza qué agente efectuará cada pago. Este esquema genera un desafío para los APPD en Colombia, ya que la tasa de cambio aplicada a la operación se fija con base en la tasa vigente en la fecha de origen y permanece inalterada hasta el momento del cobro. En consecuencia, existe una exposición a la volatilidad del tipo de cambio entre la fecha de origen y la fecha de pago de la remesa.

La tasa de mercado a la que el APPD puede monetizar los dólares puede diferir de la tasa pactada inicialmente en la operación. Es decir, puede haber transacciones pagadas en el día t con una tasa de conversión correspondiente a días anteriores, por lo que en escenarios de revaluación del peso colombiano (disminución en la tasa de cambio), esta situación puede traducirse en menores ingresos o incluso en pérdidas para el APPD objeto de este estudio, materializando el riesgo cambiario.

1.2 FORMULACIÓN DEL PROBLEMA

El problema que se aborda en este proyecto consiste en proyectar el monto total en dólares originado en un día determinado por remesas que serán cobradas en Colombia a través del APPD objeto de estudio. Esta estimación tiene como propósito proporcionar información clave para la toma de decisiones y el diseño de estrategias orientadas a mitigar el riesgo cambiario asociado al proceso de pago. El desarrollo de los modelos se basa en datos históricos suministrados por dicho agente, lo que permite caracterizar el comportamiento de sus flujos y anticipar su exposición frente a la volatilidad de la tasa de cambio.

En este trabajo se propuso el uso de técnicas de ciencia de datos para desarrollar modelos predictivos basados en datos históricos del APPD objeto de estudio. Para ello, se utilizaron variables derivadas de la propia serie, así como características contextuales relevantes, tales como el país de origen de la remesa, el canal de pago utilizado (Abono a cuenta (APN), Corresponsales o Red Propia), el día de la semana, la presencia de festivos y fechas especiales, entre otras. La combinación de estas variables permitió capturar patrones de comportamiento en los flujos de remesas y construir modelos capaces de anticipar los montos diarios con diferentes niveles de precisión, que podrían ser útiles para la gestión operativa y cambiaria del agente pagador. El análisis exploratorio permitió identificar patrones cíclicos semanales, así como variaciones asociadas a fechas especiales y días festivos. Estos hallazgos fueron validados mediante entrevistas con expertos directamente involucrados en el proceso operativo dentro del APPD, lo que aportó un entendimiento más profundo del comportamiento observado en los datos.

Durante el proceso de preparación de los datos, se aplicaron técnicas de transformación y validación orientadas a garantizar la integridad y consistencia del conjunto de información. Para la detección de desviaciones y valores atípicos se utilizaron métodos estadísticos como el análisis de percentiles y representaciones gráficas mediante diagramas de caja. Según el contexto y la naturaleza del valor atípico, se optó por su análisis individual y, en los casos pertinentes, por su exclusión del conjunto final. Es importante destacar que los datos presentaron un alto nivel de calidad, sin registros inconsistentes o incompletos, lo que facilitó su tratamiento y posterior modelado.

La selección de los modelos predictivos se basó en criterios técnicos y prácticos, priorizando aquellos enfoques que ofrecieran un buen equilibrio entre capacidad de generalización, interpretabilidad y desempeño en series temporales. Inicialmente se consideraron modelos clásicos de pronóstico como regresión lineal y autorregresivos

(ARIMA), así como enfoques más avanzados basados en aprendizaje automático, como árboles de decisión y redes neuronales artificiales.

Para el entrenamiento y validación de los modelos se implementó una estrategia de división temporal del conjunto de datos, utilizando la variable fecha de origen como criterio de corte. Esta división permitió separar un período inicial para entrenamiento y otro posterior para prueba, asegurando la consistencia cronológica del proceso y replicando un escenario real de predicción futura. Cada modelo fue entrenado utilizando los registros correspondientes a los primeros meses del histórico y evaluado con datos más recientes.

Durante el proceso, se aplicaron técnicas de optimización de hiperparámetros específicas para cada tipo de modelo, con el objetivo de maximizar el desempeño sin comprometer la generalización. Para la comparación de resultados se utilizaron métricas de error estándar como el Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) y Mean Absolute Percentage Error (MAPE). Estas métricas permitieron evaluar tanto la precisión como la robustez de cada modelo y facilitaron la selección de la alternativa más adecuada para el problema planteado.

El APPD objeto de estudio tiene como principal interés contar con modelos que puedan ser implementados operativamente dentro de sus procesos. Por esta razón, se contempla la posibilidad de mantener múltiples modelos en producción y evaluarlos de manera continua a través de esquemas de backtesting periódicos. Esta estrategia permite monitorear el desempeño predictivo en el tiempo, identificar posibles deterioros en la calidad del pronóstico y ajustar los modelos según sea necesario, garantizando así su utilidad práctica y sostenibilidad en un entorno cambiante.

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Implementar modelos predictivos que permitan estimar el monto total que se va a generar por remesas en dólares en un día.

2.2 OBJETIVOS ESPECÍFICOS

- ✓ Recolectar y preparar datos históricos identificando patrones, desviaciones y datos atípicos de remesas pagadas por el APPD en Colombia.
- ✓ Identificar variables significativas para el modelo y que aporten en la estimación del resultado.
- ✓ Seleccionar, probar y entrenar modelos pertinentes para la solución del problema como pueden ser regresión, árboles de decisión y redes neuronales.
- ✓ Optimizar y validar los modelos a partir de métricas objetivas.
- ✓ Implementar los modelos y realizar pruebas en escenarios reales.

3 MARCO TEÓRICO Y ANTECEDENTES

Para comprender adecuadamente la importancia y el contexto del proyecto, es esencial revisar el marco teórico y los antecedentes que sustentan esta investigación. En esta sección, se abordan los conceptos clave relacionados con las remesas, su impacto económico, y el uso de la ciencia de datos en la predicción de fenómenos financieros similares. Además, se analizarán estudios previos y metodologías utilizadas en la predicción de series temporales, para proporcionar una base sólida para el proyecto.

3.1 DEFINICIÓN DE REMESA

Una remesa es una transferencia de dinero realizada principalmente por emigrantes, trabajadores locales, inversionistas, etc. a sus relacionados en su país de destino. Estas transacciones hacen parte de las transferencias corrientes que se registran en la balanza de pagos[1] del país [3].

Algunas de las características que presentan las remesas son:

- **Transaccionalidad**: La mayoría de las remesas son enviadas desde países con economías más desarrolladas a países en vías de desarrollo.
- **Frecuencia y Monto**: Las remesas pueden ser enviadas de manera regular (diaria, semanal, mensual, trimestral) y los montos pueden variar según la capacidad económica del remitente y las necesidades del receptor.
- **Canales de Envío**: Las remesas pueden ser enviadas a través de diversos canales, como bancos, ETD, servicios en línea, e incluso aplicaciones móviles.

3.2 REGULACIÓN

La regulación de las remesas en Colombia se encuentra principalmente en las siguientes normativas y disposiciones legales:

1. Estatuto Orgánico del Sistema Financiero (EOSF): Este estatuto, contenido en el Decreto 663 de 1993 [4] y sus modificaciones posteriores, regula el sistema financiero colombiano en su totalidad, incluyendo las actividades relacionadas con las remesas.
2. Circular Básica Jurídica [5]: Emitida por la Superintendencia Financiera de Colombia, esta circular compila la normatividad y directrices que deben seguir las entidades financieras autorizadas para operar en el país, incluyendo la gestión de remesas.
3. Resoluciones del Banco de la República: Este banco emite resoluciones y circulares que regulan aspectos específicos de la política cambiaria y monetaria, las cuales también afectan el manejo de remesas. Una de las más relevantes es la Circular Reglamentaria Externa DCIN-83, que contiene las normas sobre operaciones cambiarias.

[1] La balanza de pagos para Colombia es donde se registran los flujos reales y financieros que el país intercambia con el resto de las economías del mundo. Presenta dos grandes cuentas: la cuenta corriente y la cuenta financiera.

4. Ley 190 de 1995 (Estatuto Anticorrupción) y la Ley 1121 de 2006 (Ley de Extinción de Dominio): Estas leyes contienen disposiciones relacionadas con la prevención del lavado de activos y la financiación del terrorismo, áreas en las cuales la supervisión de las remesas también es crucial.
5. Decretos y Resoluciones del Ministerio de Hacienda y Crédito Público: Este ministerio también puede emitir decretos y resoluciones que afecten la regulación de las remesas y el sistema financiero en general.

Estas normativas, junto con la supervisión y vigilancia constante de la Superintendencia Financiera de Colombia, aseguran que las actividades relacionadas con las remesas se realicen de manera legal, segura y transparente.

3.3 METODOLOGÍA CRISP-DM

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un marco estandarizado y ampliamente utilizado en el campo de la minería de datos y la ciencia de datos. Desarrollada en 1996, se ha consolidado como un enfoque fundamental para llevar a cabo proyectos de análisis de datos, proporcionando una estructura clara y flexible que se adapta a diferentes industrias y tipos de proyectos. Al seguir esta metodología, los equipos pueden trabajar de manera más efectiva y garantizar que sus esfuerzos en ciencia de datos estén alineados con los objetivos estratégicos de la organización.

CRISP-DM se compone de seis fases principales, cada una de las cuales se retroalimenta entre sí, formando un ciclo continuo de mejora y adaptación.

La primera fase, **Comprensión del negocio**, tiene como objetivo entender los requisitos y el contexto del proyecto desde la perspectiva del negocio. Esto implica identificar problemas específicos a resolver, definir los objetivos del proyecto y establecer métricas para medir el éxito. Esta alineación es crucial para asegurarse de que el análisis de datos genere resultados que realmente agreguen valor a la organización.

La segunda fase, **Comprensión de los datos**, implica recopilar y explorar los datos relevantes para el proyecto. Durante esta etapa, se identifican las fuentes de datos disponibles, se realiza una exploración exhaustiva de los mismos y se evalúa su calidad. La identificación de patrones, tendencias y anomalías en los datos es vital para comprender el contexto de estos y para las decisiones que se tomarán más adelante.

En la fase de **Preparación de los datos**, se realizan diversas tareas para limpiar y transformar los datos, asegurando que estén listos para el análisis. Esto puede incluir la eliminación de valores atípicos, el manejo de datos faltantes y la selección de características relevantes. La calidad y el rigor en esta etapa son cruciales, ya que influirán directamente en los resultados obtenidos en fases posteriores.

La cuarta fase, **Modelado**, se centra en seleccionar y aplicar técnicas adecuadas para construir modelos de datos. En este momento el equipo elige los algoritmos adecuados y ajusta los parámetros de los modelos construidos. La validación de los modelos es esencial durante esta etapa, ya que se pueden probar múltiples enfoques para determinar cuál ofrece el mejor rendimiento en relación con los objetivos establecidos.

Posteriormente, en la fase de **Evaluación**, se analizan los resultados de los modelos construidos para confirmarse que cumplen con los criterios de éxito definidos anteriormente. Esta etapa implica revisar si es necesario realizar ajustes o mejoras en el modelo. Puede incluir la validación con conjuntos de datos independientes para asegurar

que los resultados sean fiables y representativos.

Finalmente, la última fase es el **Despliegue** del modelo, que implica implementar el modelo en un entorno de producción. Esto puede abarcar desde la creación de informes hasta la integración del modelo en sistemas existentes. Durante esta etapa, también se elabora la documentación del proceso y se prepara a los usuarios finales para utilizar los resultados de manera efectiva [6].

3.4 APRENDIZAJE AUTOMÁTICO

Tan intuitivo como suena por su nombre, el aprendizaje automático consiste en utilizar técnicas estadísticas para desarrollar modelos que puedan aprender de los datos y hacer predicciones para lograr realizar la tomar decisiones informada. En esencia, fusiona la eficiencia computacional y la adaptabilidad de los algoritmos de aprendizaje automático con las capacidades de inferencia estadística y modelización.

El aprendizaje automático es una rama de la inteligencia artificial basada en el desarrollo de modelos estadísticos que permiten a las maquinas aprender de datos. Se trabaja con conceptos de la estadística y la teoría de la probabilidad aplicados mediante métodos computacionales buscando identificar patrones y hacer predicciones a partir de conjuntos de datos.

En este aprendizaje se trabaja con:

Datos: Esta basado en datos de entrada, que pueden ser de diversos tipos, tales como numéricos, categóricos, imágenes, videos o texto. Generalmente estos datos se dividen típicamente en dos conjuntos: entrenamiento y prueba; buscando el poder hacer un contraste entre lo aprendido en el proceso de entrenamiento y los resultados logrados en el proceso de pruebas.

Modelos: Se utilizan diferentes tipos de modelos, como la regresión lineal, árboles de decisión, redes neuronales y modelos de clasificación, dependiendo de la naturaleza del problema y los datos disponibles.

Proceso de Entrenamiento: Durante la fase de entrenamiento, se ajusta el modelo a los datos utilizando algoritmos de optimización. Este proceso implica minimizar la diferencia entre las predicciones del modelo y los valores reales (error).

Evaluación: Una vez finalizado el entrenado, el modelo se evalúa utilizando el conjunto de datos de prueba para medir su rendimiento.

Generalización: Un aspecto importante del aprendizaje automático es la capacidad de un modelo para generalizar a nuevos datos. Es crucial prevenir el sobreajuste, que ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad predictiva en datos no vistos. Estos valores son modificados en el proceso de evaluación.

3.5 MODELO PREDICTIVO

Un modelo predictivo es una herramienta estadística utilizada para pronosticar resultados futuros basándose en datos históricos y patrones subyacentes. Este enfoque se ha convertido en un elemento esencial en diversas disciplinas como la economía, la salud, el marketing y la ingeniería, las finanzas, entre otros. Su propósito principal es proporcionar una estimación o predicción de un fenómeno a partir de un conjunto de variables

independientes. Los modelos predictivos pueden clasificarse en varias categorías, dependiendo de la naturaleza de los datos y del enfoque utilizado. Entre las clasificaciones más comunes se encuentran los modelos estadísticos, que incluyen regresiones lineales y logísticas, análisis discriminante y modelos de series temporales, y los modelos de aprendizaje automático (Machine Learning), que incorporan algoritmos que permiten a las máquinas aprender patrones a partir de los datos, como árboles de decisión, redes neuronales y máquinas de soporte vectorial. El desarrollo de un modelo predictivo generalmente sigue una serie de pasos: recolección de datos relevantes y de calidad, preprocesamiento para limpiar y preparar los datos, selección del modelo más adecuado para el problema, entrenamiento y validación del modelo utilizando métricas como precisión y exactitud, y finalmente, implementación y monitoreo del modelo en funcionamiento. Los modelos predictivos tienen un amplio rango de aplicaciones que incluyen la identificación de clientes potenciales y personalización de ofertas en marketing, la predicción de enfermedades en salud, la evaluación de riesgo crediticio y pronóstico de tendencias del mercado en finanzas, y la optimización de la producción en manufactura. La implementación de modelos predictivos representa una ventaja competitiva significativa en la toma de decisiones informadas, ya que, aunque su construcción puede ser compleja, la capacidad de anticipar eventos futuros ofrece a las organizaciones la oportunidad de adaptarse y prosperar en entornos dinámicos [7].

Es crucial realizar una adecuada exploración y limpieza de los datos, así como la selección de variables relevantes para construir modelos precisos. Además, se deben realizar validaciones adecuadas utilizando técnicas como la validación cruzada o la separación de conjuntos de entrenamiento y prueba para evaluar el rendimiento de los modelos y evitar el sobreajuste.

3.6 CIENCIA DE DATOS EN FINANZAS

Predicción de series temporales basada en Machine Learning: aplicaciones económicas y financieras [8]. En este capítulo muestra casos de éxito donde se ha utilizado Machine Learning para la predicción en series temporales económicas y financieras. Particularmente se encuentran los procedimientos basados en árboles de decisión y redes neuronales los cuales se han mostrado especialmente ventajosos en el caso de la predicción de variables con dependencias no-lineales desconocidas.

3.7 MODELO ARIMA

El modelo ARIMA (AutoRegressive Integrated Moving Average) es uno de los enfoques más utilizados en el análisis de series temporales debido a su flexibilidad y capacidad para modelar datos complejos con patrones temporales. La popularidad de ARIMA radica en su enfoque estadístico, combinando componentes autoregresivos, de diferenciación y de media móvil, permitiendo capturar diferentes aspectos de las dinámicas de la serie temporal. Este modelo es especialmente útil, ya que las flujos de remesas a menudo muestran tendencias, estacionalidades y patrones impredecibles que requieren un análisis robusto y adaptativo.

La componente autorregresiva (AR) del modelo refleja la dependencia de una observación con respecto a sus valores pasados, lo que en términos de remesas implica que el ingreso de remesas en un día o semana puede estar influenciado por los periodos anteriores. Esta relación permite captar patrones persistentes en los datos, asociados a comportamientos históricos o fenómenos recurrentes, como campañas de remesas o eventos

económicos específicos. La parte integrada (I), por su parte, busca transformar la serie original en una serie estacionaria mediante diferenciaciones, eliminando tendencias o efectos de progresión que puedan sesgar los modelos predictivos. La estacionalidad es una condición clave para la implementación efectiva de ARIMA, dado que muchos de sus métodos estadísticos se basan en suposiciones de estabilidad en las propiedades de la serie a lo largo del tiempo.

Por último, la componente de media móvil (MA) del modelo se enfoca en el ajuste de los errores de predicción, considerando que las fluctuaciones imprevistas en la serie pueden ser explicadas mediante las perturbaciones pasadas. Este aspecto resulta especialmente valioso en series donde fenómenos externos o eventos repentinos afectan los flujos, como cambios en políticas migratorias, crisis económicas, factores climatológicos o festividades nacionales o de tradición. La combinación de estas tres componentes permite a ARIMA modelar de manera efectiva no solo los valores históricos, sino también las correlaciones y patrones residuales de la serie, lo que resulta en predicciones más precisas y confiables.

En cuanto a su aplicación en la generación de remesas, el modelo ARIMA puede ser utilizado para pronosticar flujos de remesas futuros, basándose en datos históricos. Este tipo de proyección es fundamental para la planificación económica, tanto a nivel nacional como para las familias que dependen de las remesas para su sustento. Además, el modelo permite identificar patrones estacionales, ya que las remesas pueden variar considerablemente durante ciertas festividades o épocas del año, tanto en el país de origen como en el país destino.

Para la correcta implementación de un modelo ARIMA, es imprescindible realizar un análisis exploratorio previo. Este incluye pruebas de estacionalidad, como la prueba de Dickey-Fuller, que determinan si la serie requiere diferenciación para estabilizar sus propiedades. Además, la identificación de los parámetros p, d y q (órdenes de los componentes AR, I y MA, respectivamente) suele realizarse mediante criterios estadísticos como el AIC (Criterio de Información de Akaike) o el BIC (Criterio de Información Bayesiano), que ayudan a seleccionar la configuración más eficiente y ajustada a los datos. La explotación de estos modelos requiere herramientas estadísticas y software especializados, como R o Python, que facilitan la automatización y validación del proceso de modelado.

Para implementar un modelo ARIMA, es esencial realizar un análisis preliminar de la serie temporal que incluye la evaluación de su estacionariedad y la determinación de los parámetros óptimos que definirán el modelo. Esto puede requerir el uso de herramientas estadísticas y software especializado, así como una interpretación cuidadosa de los datos [9].

3.8 MODELO SARIMA

El modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average) extiende el enfoque del ARIMA al abordar series temporales que exhiben patrones estacionales pronunciados, que se repiten en intervalos específicos como días, meses o años. En el análisis de remesas internacionales, tales patrones estacionales pueden ser evidentes debido a fenómenos como festividades, ciclos agrícolas, cambios en políticas migratorias o ciclos económicos globales, que afectan la cantidad de dinero enviada en determinados períodos del año. La incorporación de componentes estacionales en el modelo permite capturar estos efectos recurrentes y mejorar

la precisión de las predicciones.

SARIMA combina los componentes autoregresivos (AR), de media móvil (MA), diferenciaciones (I) y además integra componentes adicionales específicos para modelar la estacionalidad: la orden de la parte autoregresiva estacional (P), la orden de diferenciación estacional (D) y la orden de media móvil estacional (Q). Estos componentes trabajan conjuntamente para modelar tanto las dinámicas no estacionales como los patrones repetitivos propios de la serie temporal. El modelo se expresa mediante una formulación que incluye una parte no estacional (p, d, q) y otra estacional (P, D, Q), así como el período estacional s, que puede corresponder a la cantidad de observaciones en un ciclo completo.

En el contexto de la predicción de remesas, el uso de SARIMA resulta particularmente útil porque permite modelar no solo las tendencias generales y fluctuaciones aleatorias, sino también los picos o valles recurrentes asociados a períodos específicos del año. Esto es fundamental para las decisiones económicas y gubernamentales, ya que entender la estacionalidad puede mejorar la asignación de recursos, planificación de políticas migratorias y programas de asistencia social. Además, el reconocimiento de estos patrones estacionales ayuda a reducir errores en los pronósticos y a entender la naturaleza cíclica de los flujos de remesas, que puede estar influenciada por razones culturales, económicas o sociales [10].

3.9 MODELO HOLT-WINTERS

El método Holt-Winters, también conocido como suavizamiento exponencial triple, es una técnica estadística ampliamente utilizada para pronosticar series temporales que presentan patrones de tendencia y estacionalidad. Este método extiende la técnica de suavizamiento exponencial simple para incluir componentes de tendencia y estacionalidad, lo que lo hace especialmente adecuado para series de datos donde estos efectos son evidentes y predominantes. En el contexto de la predicción de remesas internacionales hacia Colombia, el modelo Holt-Winters resulta útil debido a que los flujos de remesas a menudo muestran patrones recurrentes en periodos específicos del año, como festividades, temporadas agrícolas o eventos económicos recurrentes, además de tendencias a largo plazo influenciadas por cambios económicos globales y migratorios.

El modelo Holt-Winters se desarrolla en versiones aditivas o multiplicativas, dependiendo de la naturaleza de la estacionalidad en los datos. La versión aditiva es apropiada cuando los componentes estacionales tienen una magnitud constante a lo largo del tiempo, mientras que la multiplicativa se emplea en casos donde la estacionalidad varía proporcionalmente con el nivel de la serie.

Este método optimiza la predicción mediante la actualización iterativa de tres componentes: nivel, tendencia y estacionalidad. La técnica ajusta estos componentes en cada período mediante funciones de suavizamiento con parámetros (alfa, beta, gamma) que controlan la sensibilidad del modelo a cambios recientes. La combinación de estos componentes proporciona pronósticos que reflejan la tendencia general y las variaciones estacionales, permitiendo una proyección más ajustada a los patrones históricos de las remesas. La capacidad de Holt-Winters para adaptarse rápidamente a cambios en la serie, gracias a sus parámetros de suavizamiento, contribuye a mejorar la precisión en entornos donde la dinámica de las remesas puede experimentar fluctuaciones abruptas por eventos externos o cambios en la economía global.

En la aplicación de Holt-Winters a datos de remesas, es fundamental un análisis previo que identifique la naturaleza de la estacionalidad y la tendencia en la serie. La selección entre modelos aditivos o multiplicativos depende de la forma en que la estacionalidad se manifiesta en los datos, constatando si la magnitud de las variaciones estacionales permanece constante o varía en proporción con el nivel de la serie. Además, la evaluación mediante métricas como el error cuadrático medio (MSE) o el error absoluto medio (MAE) permite validar y ajustar los parámetros del modelo para obtener pronósticos confiables.

La utilidad del método Holt-Winters radica en su simplicidad, eficiencia computacional y capacidad para capturar componentes de tendencia y estacionalidad en series temporales. En el contexto de las remesas, utilizar Holt-Winters permite a las instituciones y responsables políticos anticipar picos y valles en los flujos financieros, facilitando una mejor planificación económica, gestión de recursos y formulación de políticas migratorias [11].

3.10 LA METODOLOGÍA BOX-JENKINS

Es una técnica ampliamente utilizada para el análisis y la predicción de series temporales. Este enfoque resulta particularmente útil para entender patrones en datos históricos y realizar pronósticos sobre comportamientos futuros. Su aplicación puede ser altamente relevante en contextos económicos, como el estudio de la generación de remesas.

El primer paso en la aplicación de la metodología Box-Jenkins es la identificación de la serie temporal. Esto implica analizar los datos recopilados para detectar patrones, tendencias y estacionalidades. Técnicas como los gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF) son herramientas fundamentales en esta etapa, ya que permiten visualizar la relación entre los distintos valores de la serie a lo largo del tiempo.

Después de identificar las características de la serie, se procede a la especificación del modelo. Este paso es crucial, ya que en él se seleccionan las características adecuadas del modelo ARIMA, que incluye parámetros que reflejan componentes de auto-regresión (AR), integración (I) y media móvil (MA). Una vez establecido el modelo, se avanza hacia la estimación de parámetros, donde se utilizan técnicas como el método de máxima verosimilitud para ajustar el modelo a los datos observados.

Posteriormente, se hace el diagnóstico del modelo, que implica verificar que el ajuste sea adecuado. Esto puede incluir la evaluación de los residuos del modelo para asegurarse de que no se han dejado patrones no capturados que podrían afectar la precisión del pronóstico. Un modelo bien diagnosticado proporciona una base sólida para las predicciones futuras.

Finalmente, la metodología Box-Jenkins permite realizar pronósticos sobre la serie temporal, en este caso, sobre el flujo de remesas hacia Colombia. Al aplicar este enfoque, se obtienen datos históricos sobre los montos de remesas, los cuales pueden ser analizados para prever futuras tendencias [12].

3.11 MODELO ARIMAX

El modelo ARIMAX (AutoRegressive Integrated Moving Average with Exogenous variables) constituye una extensión del clásico ARIMA, mencionado anteriormente; el cual, integrando en su formulación variables externas

o explicativas que pueden influir en la serie temporal principal. En el análisis de series temporales como los flujos de remesas hacia Colombia, el ARIMAX resulta especialmente valioso, ya que permite incorporar variables externas que potencialmente afectan los montos de remesas enviados y recibidos, como días de la semana, eventos feriados, etc.

Este modelo combina los componentes autoregresivos (AR), de diferenciado (I) y de media móvil (MA), con un conjunto de variables exógenas (X), que pueden ser series temporales adicionales o indicadores económicos relevantes. La inclusión de estas variables permite mejorar la precisión del pronóstico al incorporar información adicional que no está contenida únicamente en la serie histórica de remesas, sino que influye en su comportamiento. Es decir, ARIMAX no solo modela la dependencia temporal de las remesas, sino también cómo estas están relacionadas con factores externos que varían en el tiempo.

En el contexto de las remesas internacionales, la incorporación de variables exógenas puede significar, por ejemplo, el comportamiento del índice de tasa de cambio, tasas de interés internacionales, políticas migratorias o eventos económicos globales que repercuten en los flujos migratorios y en las cantidades de dinero enviadas por las remesas. La modelación con ARIMAX, por tanto, permite entender y cuantificar estas relaciones, ayudando a realizar predicciones más ajustadas y a identificar cuáles factores externos son determinantes importantes en la variabilidad de las remesas [11].

3.12 REGRESIÓN CON SVM

La regresión mediante máquinas de vectores de soporte (Support Vector Regression, SVR) es una extensión de las máquinas de vectores de soporte (SVM) diseñada para problemas de predicción de variables continuas.

A diferencia de la clasificación, donde el objetivo es encontrar un hiperplano que separe las clases, en SVR se busca encontrar una función que aproxime los datos dentro de un margen de tolerancia ε.

El principio fundamental de SVR consiste en minimizar la complejidad del modelo, expresada como la norma del vector de pesos, mientras se permite una desviación máxima de ε entre las predicciones y los valores reales.

Los errores que superan este margen son penalizados de forma controlada mediante una función de pérdida epsilon-insensible. Además, SVR puede emplear kernels para manejar relaciones no lineales, proyectando los datos en espacios de mayor dimensionalidad donde la regresión lineal se vuelve viable.

El modelo final de SVR depende únicamente de un subconjunto de datos de entrenamiento conocidos como vectores de soporte, lo cual contribuye a su eficiencia computacional y a su capacidad de generalización. Gracias a estas propiedades, SVR se ha consolidado como una herramienta robusta para tareas de regresión en diversas áreas, como bioinformática, econometría, e ingeniería [13].

3.13 MODELO DE ÁRBOLES DE DECISIÓN (DECISION TREE REGRESSOR)

Es un algoritmo de aprendizaje supervisado que se utiliza para predecir valores continuos. Funciona dividiendo el espacio de características en subregiones mediante decisiones basadas en las características de los datos. El modelo toma decisiones en forma de árbol, donde cada nodo interno representa una prueba en una característica,

cada rama representa el resultado de la prueba y cada hoja representa un valor objetivo, en este caso, el valor continuo que se desea predecir es el monto en dólares que se generan en un día.

Un árbol de decisión se compone de nodos, ramas y hojas:

- Nodos Internos: Cada nodo interno representa una prueba o decisión sobre una característica del conjunto de datos.
- Ramas: Las ramas son las conexiones entre los nodos. Cada rama define el flujo de decisión basado en el resultado de la prueba en el nodo correspondiente.
- Hojas: Las hojas son los nodos terminales del árbol que contienen la predicción final.

El árbol comienza en un nodo raíz que representa todo el conjunto de datos. Se evalúa una característica y se divide el nodo en función del resultado de la evaluación. La división se hace buscando el punto óptimo que minimiza la varianza entre los valores de objetivo en las distintas ramas resultantes.

Para el Decision Tree Regressor, el criterio de división más común es el error cuadrático medio (MSE). Este mide la diferencia cuadrática promedio entre los valores reales y los valores predichos, y el objetivo es minimizar esta diferencia.

El proceso de división se repite de forma recursiva para cada subgrupo resultante, creando nodos adicionales y continuando hasta que se alcanza algún criterio de parada. Los criterios de parada incluyen: profundidad máxima del árbol, mínimo número de muestras requeridas para dividir un nodo, mínimo número de muestras en una hoja.

Las principales ventajas son:

- Simplicidad e Interpretabilidad: Los árboles de decisión se pueden visualizar fácilmente, permitiendo que los usuarios entiendan cómo se toman las decisiones.
- No Requieren Preprocesamiento: No es necesario escalar o normalizar los datos, ya que son invariables bajo transformaciones lineales.
- Versatilidad: Pueden manejar tanto características numéricas como categóricas y son útiles en diferentes problemas de regresión.
- Maniobrabilidad de Datos Faltantes: Pueden manejar datos faltantes de maneras específicas, tratando las instancias con datos faltantes como un subgrupo separado.

Algunas desventajas son:

- Sobreajuste: Los árboles de decisión son propensos al sobreajuste, especialmente cuando permiten una profundidad excesiva. Se ajustan demasiado a los datos de entrenamiento, lo que puede causar un rendimiento deficiente en datos no vistos.
- Inestabilidad: Pequeños cambios en los datos pueden llevar a grandes cambios en la estructura del árbol.

Los hiperparámetros que se pueden ajustar en el Decision Tree Regressor son:

- max_depth: Define la profundidad máxima del árbol. Controla la expansión vertical del árbol y ayuda a prevenir el sobreajuste.
- min_samples_leaf: Define el número mínimo de muestras que debe tener un nodo hoja.
- max_features: Indica el número de características a considerar al buscar la mejor división. Puede tomar uno de dos valores: 'sqrt' (que usa la raíz cuadrada del número total de características) o 'log2' (que usa el logaritmo en base 2 del número total de características), seleccionado aleatoriamente.
- criterion: Define la función para medir la calidad de una división. En el caso de regresión, 'squared_error' (también conocido como error cuadrático medio) es una medida común que busca minimizar la suma de los errores al cuadrado entre los valores predichos y reales.

Al generar el modelo se determinan los siguientes datos de evaluación del modelo:

- MAE (Error Absoluto Medio): es la media de las diferencias absolutas entre los valores predichos y los valores reales. Se puede interpretar como el promedio de los errores en magnitud sin considerar la dirección.
  Fórmula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\|(y_i - \widehat{y_i})\|$$

  y_i: es el valor real u observado.
  $\widehat{y_i}$: valor predicho
  n: número total de observaciones
  Un MAE menor indica un mejor ajuste del modelo.
  Si el MAE es 0, significa que las predicciones son perfectas.
- MSE (Error Cuadrático Medio): mide la media de las diferencias al cuadrado entre los valores predichos y los valores reales. Esta métrica penaliza errores más grandes debido al cuadrado de las diferencias.
  Formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

  y_i: es el valor real u observado.
  $\widehat{y_i}$: valor predicho
  n: número total de observaciones
  Un MSE menor indica un mejor ajuste del modelo. A diferencia del MAE, el MSE es más sensible a los errores grandes porque los mismos se amplifican al ser elevados al cuadrado.
- RMSE (Raíz del Error Cuadrático Medio): es la raíz cuadrada del MSE. Esta métrica proporciona una forma de medir el error en las mismas unidades que los datos originales.
  Fórmula:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Al igual que el MSE, un RMSE menor indica un mejor ajuste. El RMSE es útil porque está en las mismas unidades que la variable objetivo, lo que facilita la interpretación.

- R^2 (Coeficiente de Determinación): El R^2 es el coeficiente de determinación que mide la proporción de la varianza total en los valores reales que se explica por el modelo. Se calcula comparando el modelo con un modelo promedio que simplemente predice la media de los valores.

  Fórmula:

$$R^2 = 1 - \frac{RSS}{TSS}$$

  RSS: Suma de los cuadrados de los residuos

  TSS: Suma total de los cuadrados

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Un R^2 de 1 indica que el modelo explica completamente la varianza.

  Un R^2 de 0 indica que el modelo no explica nada de la varianza.

  Un R^2 negativo indica que el modelo es peor que simplemente predecir la media de los valores.

3.14 MODELO ELMAN

El modelo Elman, introducido por Jeffrey L. Elman en 1990, corresponde a una arquitectura de redes neuronales recurrentes diseñadas para el procesamiento de secuencias temporales y datos con dependencias en el tiempo. Esta red neuronal cuenta con una estructura simple pero poderosa, que incluye un capa de entrada, una capa oculta, una capa de contexto y una capa de salida. La capa de contexto recuerda automáticamente los estados anteriores de la capa oculta, permitiendo que la red modele relaciones temporales y patrones secuenciales en los datos.

En nuestra exposición, donde se busca establecer como objetivo principal, el monto de los giros realizados hacía Colombia generados en una fecha, el modelo Elman resulta particularmente útil debido a su capacidad para capturar dinámicas no lineales y dependencias temporales complejas que los modelos estadísticos tradicionales, como ARIMA o Holt-Winters, quizás no puedan captar completamente. El manejo de estas relaciones no lineales y secuenciales es una fortaleza principal de las redes neuronales recurrentes como Elman.

El proceso de entrenamiento de una red Elman implica ajustar los pesos de las conexiones para minimizar el error en las predicciones, mediante algoritmos de retro propagación del error adaptados a redes recurrentes (retro propagación a través del tiempo). Esta capacidad de aprender a partir de datos históricos, considerando

dependencias de estados pasados, permite que el modelo pueda proyectar con mayor precisión los futuros valores de las remesas, incluso en presencia de patrones complejos y variables.

Además, las redes Elman son flexibles y adaptables, lo que las hace aptas para incorporar variables externas o exógenas que puedan influir en los flujos de remesas, complementando los datos históricos. Por ejemplo, variables macroeconómicas, tasas de cambio, eventos políticos o migratorios que impactan en los valores de remesas pueden ser incluidos como entradas adicionales, mejorando la capacidad predictiva del modelo.

En términos de implementación, el entrenamiento de redes Elman requiere una cantidad significativa de datos históricos y una cuidadosa selección de hiperparámetros, como la cantidad de neuronas en la capa oculta, la tasa de aprendizaje y el número de iteraciones. Es importante también realizar validaciones con conjuntos de datos separados para evitar sobreajuste y asegurar la generalización del modelo. Por su capacidad para modelar relaciones no lineales y secuenciales, las redes Elman se convierten en una herramienta poderosa para pronosticar flujos de remesas en escenarios donde los patrones históricos son complejos y estacionales, y donde los modelos estadísticos tradicionales pueden presentar limitaciones [7].

3.15 MODELO JORDAN

El modelo Jordan, propuesto por Michael I. Jordan en 1986, es una arquitectura de red neuronal recurrente diseñada para procesar series temporales y datos secuenciales mediante conexiones hacia la capa de salida y, en algunas configuraciones, hacia la capa oculta o de contexto. La estructura del modelo incluye normalmente una capa de entrada, una capa oculta, una capa de contexto que mantiene un estado de memoria y una capa de salida. La capa de contexto recibe la salida de la capa oculta en el paso anterior, proporcionando a la red una forma de "recordar" información pasada y modelar dependencias temporales en los datos.

En el escenario de estimar el monto total de remesas enviadas hacia Colombia, el modelo Jordan resulta especialmente útil por su capacidad para captar relaciones no lineales, dependencias temporales y patrones históricos en los datos de flujo de remesas. Debido a que las remesas internacionales están influenciadas por múltiples factores como eventos económicos globales, tasas de cambio, temporadas migratorias, políticas migratorias y fenómenos sociales, un modelo que pueda recordar y aprender de las secuencias pasadas es fundamental para realizar predicciones precisas.

La arquitectura del modelo Jordan puede ser entrenada empleando algoritmos de retro propagación del error, en los que los pesos de la red se ajustan para minimizar la diferencia entre las predicciones y los valores reales de remesas en días anteriores. La capacidad del modelo para mantener la memoria de estados pasados permite que aprenda patrones recurrentes y dependencias a corto plazo, características inherentes a los flujos de remesas que suelen mostrar estacionalidad y tendencias variables en el tiempo.

Además, el modelo puede incluir variables externas o exógenas relevantes para mejorar la estimación del monto total de remesas en un día determinado, como el volumen de migrantes activos, tasas de cambio, indicadores económicos internacionales o eventos políticos. La incorporación de estos factores proporciona contexto adicional que facilita la predicción del flujo diario, considerando no solo la dependencia de los datos históricos, sino también las influencias externas que impactan en los montos de remesas.

Para implementar eficazmente el modelo Jordan, es necesario contar con una cantidad adecuada de datos

históricos, realizar una cuidadosa selección de hiperparámetros, como la cantidad de neuronas en la capa oculta y las tasas de aprendizaje, y validar el modelo usando datos no utilizados durante el entrenamiento. La estructura del modelo puede ajustarse para optimizar la capacidad de recordar información relevante, evitando problemas de sobreajuste y garantizando buenas capacidades de generalización [7].

3.16 MODELOS LSTM

El modelo LSTM (Long Short-Term Memory) fue introducido por Hochreiter y Schmidhuber en 1999 como una mejora de las redes neuronales recurrentes tradicionales, diseñada para abordar el problema del olvido de información en secuencias largas y mejorar la capacidad de aprendizaje de dependencias a largo plazo. La arquitectura LSTM incorpora unidades llamadas celdas, que contienen mecanismos de puerta —la puerta de entrada, la puerta de olvido y la puerta de salida— que controlan el flujo de información hacia adentro y hacia afuera de la celda. Este sistema permite que la red decida qué información mantener, olvidar o transmitir en cada paso temporal, facilitando el aprendizaje de patrones que se extienden a través de largos períodos.

Para nuestro ejercicio, el modelo LSTM es particularmente útil porque puede captar relaciones complejas y no lineales en los datos secuenciales con dependencias tanto a corto como a largo plazo. Los LSTM, con su capacidad para aprender y mantener información relevante a lo largo de largos períodos, resulta eficaz para modelar estos patrones temporales complejos y predecir con mayor precisión.

El entrenamiento del modelo LSTM implica ajustar sus pesos mediante algoritmos de retropropagación del error, utilizando variantes de optimización como Adam o RMSprop que facilitan la convergencia. Además, los modelos LSTM pueden ser enriquecidos con variables exógenas, para mejorar la estimación del monto total de remesas en un día específico.

Implementar un modelo LSTM requiere una cuidadosa selección de hiperparámetros, como el número de unidades ocultas, las tasas de aprendizaje, la cantidad de capas y la regularización para evitar el sobreajuste. La validación mediante conjuntos independientes de entrenamiento y prueba, además de técnicas de ajuste temprano, ayuda a garantizar una buena generalización del modelo a datos futuros. La ventaja principal del LSTM en este escenario radica en su capacidad para aprender y recordar patrones relevantes a largo plazo, capturando tendencias y estacionalidades que otros modelos estadísticos tradicionales podrían no detectar [14].

3.17 ANTECEDENTES

3.17.1 Estudios previos en predicción de remesas

- Forecasting remittances to Mexico with a Multi-State Markov–Switching model applied to the trend with controlled smoothness [15].

  En una primera instancia, se hace un análisis del porqué se está generando un aumento en el valor, número y frecuencia del envío de remesas hacia México, especialmente desde EEUU. Esto con el objetivo de proveer información para el diseño de políticas que puedan aumentar las mismas y mejorar su utilización.

El periodo de datos que se dispuso para el ejercicio son los datos de remesas de trabajadores que recibieron en dólares en México y registradas por BANXICO desde 1995 hasta 2016.

El Markov-Switching multiestatal que se menciona en el documento, está caracterizado como un modelo estadístico de comparación, con el que se busca identificar patrones en el comportamiento de las remesas, permitiendo que se apliquen ajustes a las predicciones basadas en los diferentes días del año. Para nuestro estudio, se utiliza el concepto de identificación de características de los días para poder hacer una predicción mucho más acertada. Se diferencia el método empleado, ya que en nuestro caso utilizamos Box-Jenkins. Este modelo, incluyen el enfoque ARIMA (autorregresivo integrado de media móvil), se centran en la identificación de patrones y en la modelización de datos estacionarios o transformados a estacionariedad. Por otro lado, el modelo Markov-Switching es útil para series con cambios estructurales y no estacionarios.

Un desafío común en el análisis de series de tiempo es la estacionariedad. El modelo Box-Jenkins ayudan a tratar la no estacionariedad a través de diferenciación, mientras que el modelo Markov-Switching pueden ayudar a identificar cuándo y cómo los patrones cambian de estado. Juntos, pueden proporcionar un enfoque más robusto para analizar datos complicados.

Por practicidad del trabajo realizado se aplicó únicamente el modelo Box-Jenkins buscando lograr de manera amplia la caracterización suficiente, tomando como experiencia la parametrización y las consideraciones generales mencionadas en el trabajo.

- Propuesta de optimización del costo logístico en las remesas de efectivo de las sucursales de una entidad bancaria mediante la aplicación de herramientas de analítica.

  El objetivo de este estudio es pronosticar montos de efectivo requeridos en sucursales que serán demandados para el pago y envío de remesas [7]. demostrando que se puede mejorar los niveles de efectivo en las sucursales sin disminuir el nivel de servicio al cliente al proporcionar la cantidad adecuada de efectivo en la ubicación y tiempo correctos a través de análisis de clústeres los cuales se utilizan en casi todos los campos donde existe una gran variedad de transacciones. Por lo tanto, puede ayudar a identificar agrupaciones naturales de clientes, productos, pacientes, entre otros. Se destacan algoritmos de agrupación en clústeres como K-means, K-menoides.

  Para poder utilizar la data que nos fue proporcionada por la entidad bancaria, ésta debió ser tratada con procesos de anonimización buscando el claro cumplimiento de las regulaciones existentes de habeas data. Como se menciona anteriormente todos los algoritmos utilizados para este proceso se realizaron mediante la herramienta K-nime haciendo agrupamientos utilizando la metodología k-means.

En este ejercicio se trabajó todo con una sucursal única nacional; ya que el trabajo de poder hacer la predicción en cada una de las oficinas es una labor que se debe realizar posterior a la completitud de este ejercicio directamente en la entidad bancaria.

En el trabajo se incorpora el modelo ARIMA para realizar los pronósticos de depósito de remesas en ventanilla de las oficinas. Esto indica que este modelo tiene una parte no estacional de tercer orden autorregresivo, sin diferencias y media móvil de segundo orden. Asimismo, presenta una parte estacional de primer orden, sin diferencias y media móvil de segundo orden con datos diarios.

- Comparing the forecasting performance of ARIMA and Neural Network Model by using the remittances of Bangladesh [8].

  El objetivo de este estudio es comparar la eficacia de dos metodologías de pronóstico ampliamente utilizadas, el modelo ARIMA (AutoRegressive Integrated Moving Average) y el modelo de Red Neuronal Artificial (ANN, por sus siglas en inglés), en la predicción de las remesas de Bangladesh. El modelo de red neuronal presentó una mejor capacidad para capturar la complejidad y no linealidad de las remesas. Aunque inicialmente requirió más tiempo y recursos para el entrenamiento, su rendimiento en la predicción a corto y largo plazo superó al modelo ARIMA en la mayoría de las métricas evaluadas.

  Este trabajo refuerza la idea del trabajo inicialmente planteado del modelo ARIMA para el trabajo de las series de tiempo expuestas, y aunque se contrastan con el modelo de redes neuronales ANN y se logra un mayor beneficio, se advierten las diferencias presentadas para el modelamiento de los modelos.

3.18 METODOLOGÍA

En este trabajo, se adopta la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) para guiar el desarrollo y ejecución del proyecto de análisis de datos. CRISP-DM es una metodología estándar de la industria que proporciona un enfoque sistemático para la planificación, ejecución y evaluación de proyectos relacionados con ciencia de datos. Tomando en cuenta el estándar de la metodología CRIPS-DM expuesta anteriormente en el marco de referencia, se explica cómo se va a hacer la apropiación para este proyecto:

Comprensión del Negocio: En este proyecto trabajaron tres personas, uno de los cuales labora en la entidad que autorizó el uso de los datos, teniendo en cuenta esta circunstancia, se ha estudiado en el grupo completo el proceso general del negocio de remesas.
A partir del entendimiento se identificó la problemática a abordar y que modelos podrían ayudar a gestionar.

Adicional, se hace la gestión interna para lograr las adecuadas autorizaciones por parte de la entidad para tener acceso a la información pertinente para el desarrollo del proyecto.

En esta etapa, se establecen los objetivos, se identifican los problemas clave y se formulan preguntas de negocio específicas, así como el marco teórico y la metodología que se va a emplear.

Durante la realización del proceso se profundizó en el conocimiento del negocio y las necesidades específicas que se abordaron en este proyecto.

1. Comprensión de los Datos: Se evaluó la estructura inicial de los datos, los cuales fueron transformados y recopilados teniendo en cuenta el objetivo del proyecto, en este caso, se revisó la totalidad de operaciones de remesas pagadas entre el primero de enero de 2023 y el 31 de diciembre de 2024.

   Una vez entregados al proyecto estos datos se analizaron para comprender sus características. La información otorgada por la entidad debió cumplir con ciertas características básicas de seguridad, como son: firma de acuerdos de confidencialidad con los miembros del equipo, anonimizar la información, garantizar el proceso de entrega, manejar el control de las copias de la data del proyecto, etc.

2. Preparación de los datos: En esta etapa se seleccionó, limpió y transformó los datos para el modelado considerando la sensibilidad de la información exigida por la empresa que autorizó el uso de los datos.

3. Modelado: En esta fase se aplicaron varias técnicas de modelado que los integrantes del proyecto aprendieron en las materias de la maestría:

   - Modelos de series de tiempo
   - Árboles de Decisión
   - Regresión con SVM
   - Redes Neuronales Artificiales (ANN)

4. Evaluación: Los modelos desarrollados se evaluaron en función de los objetivos del proyecto. Se utilizaron métricas de evaluación como MAE, MSE, RMSE y R^2 para validar el desempeño de los modelos.

5. Despliegue: En la fase final, los modelos y resultados se implementan en el entorno operativo del negocio. Se preparará el entorno, se monitoreará el modelo en producción.

Para este proyecto, el modelo no implica instalarlo en la infraestructura de la entidad, ya que por procesos internos, algunos de ellos definidos por la Superintendencia Financiera de Colombia, se deben realizar por personas vinculadas con la entidad y con niveles de confianza; tal como establece la Ley 1328 de 2009, donde se establece un marco regulatorio integral que busca proteger los derechos de los consumidores financieros y asegurar la transparencia y la responsabilidad en la prestación de servicios financieros en Colombia.

4 ANALISIS EXPLORATORIO DE DATOS

4.1 PROCESAMIENTO INICIAL

El APPD objeto de estudio de este proyecto facilitó el acceso a archivos planos que contienen el detalle de transacciones de remesas pagadas entre el 1 de enero de 2023 y el 31 de diciembre de 2024. La información se encontraba estructurada en archivos diarios separados por canal de pago: Abono a Cuenta (APN), Red Propia y Corresponsales.

Es importante aclarar que los canales de pago mencionados anteriormente son propios del APPD objeto de este estudio.

En cumplimiento de la normatividad vigente y del acuerdo de confidencialidad suscrito, fue necesario realizar un proceso de consolidación que permitiera unificar la información bajo criterios de anonimización y pertinencia analítica. Para ello, se empleó la herramienta KNIME, con la cual se integraron los archivos conservando únicamente las variables autorizadas para su análisis: monto total en dólares, canal de pago, fecha de origen, fecha de pago y país de origen. La selección de estas variables fue definida con base en las recomendaciones de expertos del área de negocio del APPD, quienes indicaron su relevancia para el entendimiento del comportamiento transaccional.

Como resultado de este proceso, se construyó una base de datos consolidada que sirvió como insumo para la fase de análisis exploratorio y la posterior implementación de modelos predictivos.

4.2 DESCRIPCIÓN GENERAL

Los datos otorgados corresponden a un total de 20.767.957 operaciones registradas como pagos de remesas realizados entre 1ro de enero de 2023 y el 31 de diciembre de 2024, que sumaron un total de USD $4.960 millones. Estos datos fueron agrupados y totalizados de la forma como se muestra en la tabla 1:

Nombre de la columna	*Descripción*
Fecha_Pago	Fecha en que se realiza el pago de la remesa.
Fecha_Origen	Fecha de generación de la remesa.
País_Origen	País de origen de la remesa
Canal	Medio del APPD objeto de este estudio donde se realiza el pago de la remesa al receptor.
Numero_transacciones	Número de operaciones.
Monto_USD	Monto de la remesa en dólares.

Tabla 1 Descripción variables iniciales recibidas

A continuación, según la tabla 2 se muestran filas aleatorias del conjunto de datos.

Fecha de Pago	Fecha de Origen	Pais Origen	Canal	Numero_transacciones	Monto_USD
25/01/2024	25/01/2024	PA	APN	320	159,495.32
11/04/2024	11/04/2024	KY	APN	1	250.83
03/07/2024	21/06/2024	US	Corresponsales	14	1,742.76
04/02/2023	01/02/2023	ES	Corresponsales	61	8,294.28
03/01/2024	03/01/2024	PT	RedPropia	6	3,840.60

Tabla 2 Muestra de registros del conjunto inicial de datos

Como ejemplo, en el cuadro anterior se muestra que el 3 de julio de 2024, se realizó el pago de 14 transacciones a través de corresponsales, provenientes de Estados Unidos y que fueron originados el 21 de junio de 2024.

4.3 TRATAMIENTO DE VALORES NULOS

En esta etapa del proceso de análisis de datos, se aborda el tratamiento de los valores nulos presentes en el conjunto de datos, ya que su presencia puede afectar significativamente tanto la calidad de los modelos predictivos como la validez de las conclusiones extraídas. A continuación, se presentan las ilustraciones 2 y 3 que muestran las variables con valores nulos y el tratamiento aplicado para gestionarlos adecuadamente.

```
Fecha de Pago            0
Fecha de Origen          0
Pais Origen             14
Canal                    0
Numero_transacciones     0
Monto_USD                0
dtype: int64
```

Ilustración 2 Revisión de valores nulos

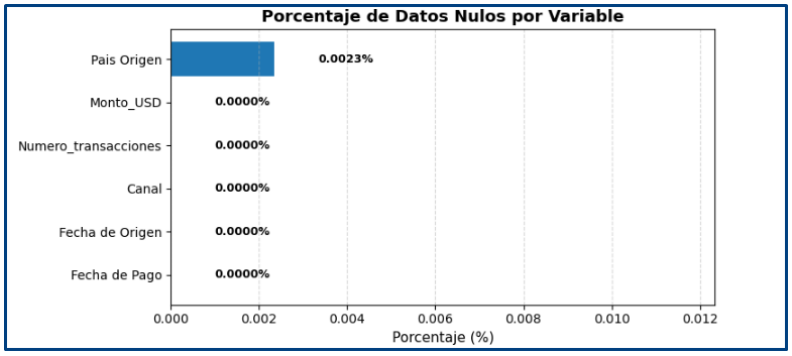

Ilustración 3 Verificación gráfica de valores nulos

No se encuentran valores nulos en el set de datos a excepción de la variable Pais_Origen variable que tiene un total de 14 operaciones sin información 0.002%, al ser un valor poco significativo se procede con la imputación de los datos usando la moda que es “US”.

4.4 TRANSFORMACIÓN DEL CONJUNTO DE DATOS

Dado el objetivo del proyecto, se requirió transformar estos datos transaccionales a un nivel de agregación diario por fecha de origen, generando una nueva base estructurada con las siguientes características:

- Fecha de Origen como índice principal de análisis.
- Monto_USD, correspondiente a la suma diaria de todas las remesas originadas en dólares.
- Montos diarios por canal de pago: APN, Corresponsales y RedPropia.
- Montos diarios por país de origen: desagregados por código ISO (ej. US, CA, CL, etc.).

El resultado fue una base tabular con una sola fila por cada fecha de origen y múltiples columnas numéricas, lo que permitió su uso directo como insumo para los modelos de series de tiempo y aprendizaje automático.

Fecha de Origen	Monto_USD	APN	Corresponsales	RedPropia	AU	BR	CA	CL	EC	ES	MX	PA	PE	US	Otros
2023-01-01	916,300.90	403,280.26	127,379.49	385,641.15	30,651.06	29,932.68	19,150.52	4,366.51	4,816.94	72,645.08	18,119.55	23,133.71	831.48	461,445.15	251,208.22
2023-01-02	4,381,038.20	1,845,134.00	891,001.69	1,644,902.51	39,553.29	240,818.06	76,224.61	47,204.26	202,190.79	813,046.11	146,267.18	164,118.58	462,802.58	1,655,811.83	533,000.91
2023-01-03	6,163,327.36	2,247,322.58	1,504,633.54	2,411,371.24	52,705.15	234,247.26	85,722.18	468,826.69	528,215.73	944,753.06	174,730.18	440,716.22	671,774.51	1,783,667.79	777,968.59
2023-01-04	6,230,153.05	2,379,079.07	1,457,531.13	2,393,542.85	97,793.76	249,985.99	86,896.62	408,663.76	519,023.89	924,372.80	154,202.93	380,269.10	695,298.18	1,934,012.36	779,633.66
2023-01-05	6,749,396.11	2,661,193.54	1,483,822.37	2,604,380.20	71,312.18	362,911.59	104,571.14	393,955.20	550,034.86	890,721.76	174,292.30	401,863.27	834,621.83	2,087,273.20	877,838.78

Tabla 3 Muestra de registros del conjunto de datos a trabajar

La tabla 3 muestra la estructura de la base de datos transformada, consolidada a nivel diario por fecha de origen.

Para facilitar el análisis y reducir la dimensionalidad, se seleccionaron los 10 países con mayor participación en el volumen total de remesas, representados individualmente en columnas específicas (por ejemplo: Estados Unidos – US, Canadá – CA, Brasil – BR, entre otros). Los países con menor participación fueron agrupados en una única categoría denominada “Otros”, permitiendo conservar la representatividad global sin introducir ruido por valores marginales o esporádicos.

4.5 VARIABLES COMPLEMENTARIAS.

A partir de variable fecha de origen se obtuvieron las siguientes variables.

- Es_Fin_de_Semana: Identifica mediante valores booleanos si la fecha de origen corresponde a un fin de semana (sábado y domingo) o no.
- Feriado_USA: Identifica mediante valores booleanos si la fecha de origen corresponde a un día feriado en Estados Unidos.
- Feriados Colombia: Identifica mediante valores booleanos si la fecha de origen corresponde a un día feriado en Colombia.
- Mes: Identifica el mes de la fecha en que se origina la remesa.

- Semana del Mes: Proporciona la información correspondiente a la semana del mes en que se origina la remesa.
- Dia de la semana: Proporciona la información correspondiente al día de la semana en que se origina la remesa.
- TRM: Los valores correspondientes a la TRM se obtuvieron de la página del banco de la república y se cruzaron con el set de datos a través de la fecha de origen.

4.6 DESCRIPCIÓN DE VARIABLES

En esta sección se presenta la descripción del conjunto de datos antes de aplicar el proceso de limpieza. Se detallan las variables que componen el set original, proporcionando una visión general de su contenido y estructura. En las secciones siguientes, se explicará qué registros fueron eliminados y las razones que justifican dichas eliminaciones, con el objetivo de mejorar la calidad y la relevancia del conjunto de datos para el análisis posterior.

1. Fechas:
    - Fecha de Pago: El rango es desde el 1 de enero de 2023 hasta el 31 de diciembre de 2024.
    - Fecha de Origen: El rango es desde el 20 de noviembre de 2018 hasta el 31 de diciembre de 2024, con una distribución centrada en fechas recientes (2023 y 2024).
2. Montos:
    - Monto_USD: El rango de esta variable agrupada por fecha de origen es de USD $0.03 hasta USD $13.6 millones. La mediana es de USD $6.2 millones y la media de USD $5.1 millones.
3. Canal
    - La variable Canal representa el medio del APPD objeto de estudio a través del cual se realiza la transacción de remesas. Los diferentes canales son: "RedPropia" que hace referencia las oficinas propias de la entidad, "APN" que se refiere a remesas que se realizaron directamente con abono a cuenta bancaria y "Corresponsales" que son empresas aliadas que prestan la red de oficinas.
4. País de Origen
    - La variable Pais_Origen indica el país de procedencia de la remesa, se representa mediante códigos abreviados según ISO 3166-1 alfa-3 (por ejemplo, "US", "ES", "EC", etc.), lo que permite segmentar y analizar el flujo de remesas desde distintos contextos geográficos.

5. Monto en dólares por fecha de origen (variable objetivo)

La ilustración 4 muestra la serie original de montos en dólares de remesas por fecha de origen.

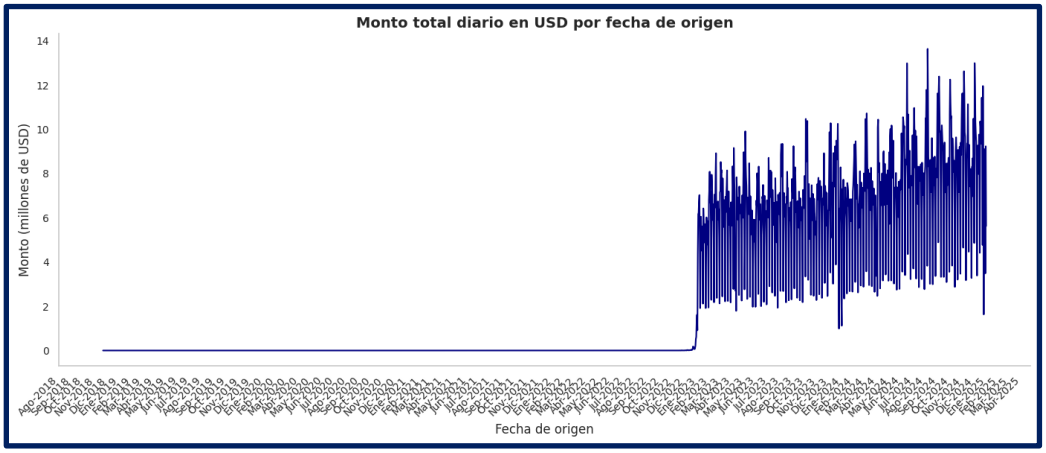

Ilustración 4 Monto total diario en dólares

Dado que el sistema de información del APPD registra las operaciones en la fecha en que son efectivamente pagadas al receptor —y no en la fecha en que fueron originadas por el remitente—, se generan distorsiones en los extremos de la serie temporal.

En el tramo inicial, se presentan pagos correspondientes a remesas originadas antes del inicio del periodo de análisis (1ro de enero de 2023), lo cual introduce un volumen atípicamente bajo y no representativo del comportamiento regular de la serie. Por esta razón, estas operaciones fueron excluidas del conjunto de datos. En total, se eliminaron 30.698 transacciones (equivalentes al 0,15% del total) por un valor de 5,5 millones de dólares (0,11%).

En el extremo final, debido a que algunas remesas aún no han sido cobradas, se presenta una disminución en los valores registrados, esto genera una representación incompleta del volumen real de operaciones. Esta situación se presenta a pesar de que la ETD cuenta con una política de inactivación de operaciones que no han sido cobradas en un plazo de 30 días. Si bien esta política obliga al remitente a realizar una reactivación de la operación, la operación conserva las mismas condiciones iniciales (incluyendo la tasa de cambio y los costos asociados), lo que genera que algunas transacciones puedan ser cobradas durante un periodo extendido.

A continuación, se presenta tabla 4 que muestra el comportamiento del pago total de remesas originadas en un día, gestionadas por el APPD. La columna "Días de diferencia" representa número de días transcurrido desde el momento de originación de la remesa (día 0) hasta su pago efectivo. Las columnas siguientes

detallan el número de transacciones (en millones), el porcentaje correspondiente sobre el total, y los montos pagados en millones de dólares, junto con su porcentaje acumulado.

Días de diferencia	Número de transacciones (millones)	% transacciones	% transacciones acumulado	Monto (millones de USD)	% monto	% monto acumulado
0	12.2	58.95%	58.95%	3,209.6	64.71%	64.71%
1	5.0	24.30%	83.25%	1,049.1	21.15%	85.86%
2	1.5	7.00%	90.25%	280.0	5.64%	91.51%
3	0.7	3.56%	93.81%	149.3	3.01%	94.52%
4	0.4	1.95%	95.76%	84.0	1.69%	96.21%
5	0.2	1.18%	96.95%	56.4	1.14%	97.35%
6	0.2	0.78%	97.72%	40.6	0.82%	98.17%
7	0.1	0.52%	98.24%	23.6	0.48%	98.65%
8+	0.4	1.76%	100.00%	67.2	1.35%	100.00%

Tabla 4 Validación de pagos de acuerdo con el día generado

Esta tabla permite observar en cuántos días, a partir del momento de originación de las remesas, se completa aproximadamente el 100% del pago total de éstas, lo que brinda una visión despejada del comportamiento temporal del flujo de pagos.

Dicho esto, se tiene que luego de 7 días de originadas las remesas, el 98,2% de las transacciones y el 98,6% del monto total en dólares son pagados. Por esta razón, se estableció como fecha máxima de análisis por fecha de origen el 24 de diciembre de 2024, es decir, siete días antes de la fecha final del conjunto de datos (31 de diciembre de 2024). Esta decisión garantiza una representación estadísticamente confiable del comportamiento de las remesas, evitando la inclusión de datos aún no consolidados.

El ajuste aplicado a la base de datos implicó la exclusión de 167.329 operaciones (0,81% del total), correspondientes a un monto de 43,4 millones de dólares (0,87%). Dada la baja representatividad de dichas operaciones en términos relativos, su impacto sobre la gestión del riesgo cambiario es marginal y no afecta la validez ni la utilidad operativa del modelo desarrollado.

A continuación, en la tabla 5 se presenta la serie diaria de montos en dólares, ajustada con las exclusiones mencionadas en este apartado.

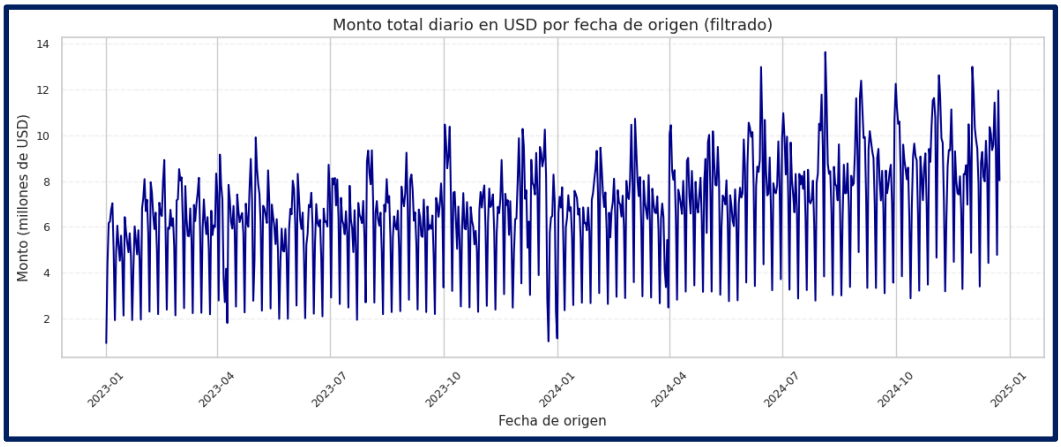

Tabla 5 Serie diaria de montos en dólares

4.7 DISTRIBUCIÓN DE LAS VARIABLES

Pais_Origen y Canal son variables categóricas que se pueden explorar para identificar si aportan en los diferentes modelos propuestos.

País de Origen

Para facilitar la exploración de los datos se seleccionaron los 10 países con mayor monto en dólares y el resto fue agrupado en la categoría “Otros”.

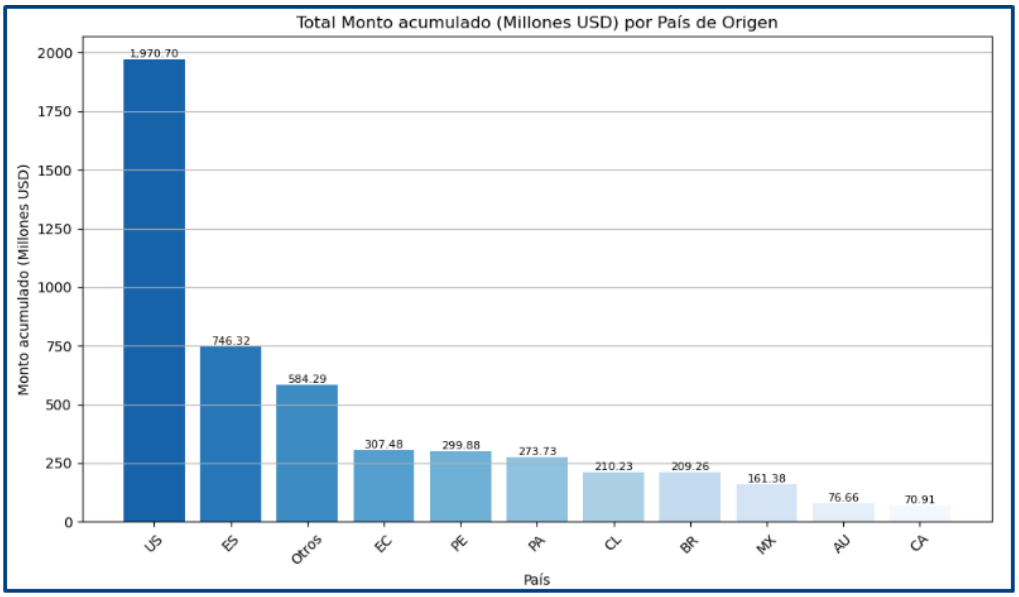

Ilustración 5 Top 10 monto acumulado en USD por país.

Los resultados de la ilustración 5 muestran que los montos acumulados de remesas tienen a Estados Unidos (US) como el país líder, seguido de España (ES), esto puede deberse a que estos países tienen una alta población de

migrantes. Países como Perú (PE), Ecuador (EC) y Panamá (PA) también destacan, posiblemente por su conexión económica o demográfica con Colombia. Finalmente, México (MX), Australia (AU) y Canadá (CA) completan el top 10, reflejando su relevancia como países remisores.

Canal

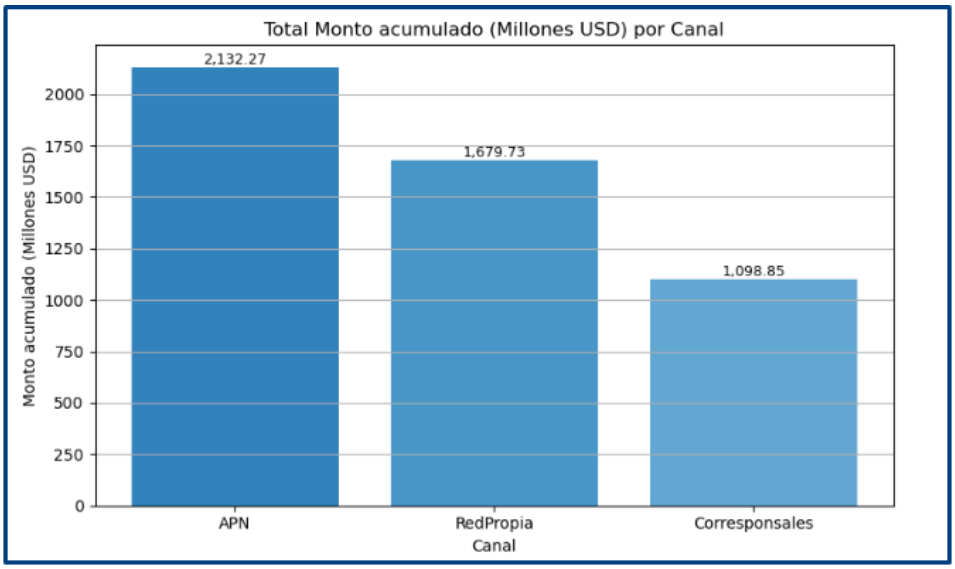

Ilustración 6 Monto acumulado en USD por canal.

La ilustración 6 de Monto acumulado en USD por canal revela que el monto mayor en USD acumulado proviene del canal APN, con un volumen acumulado superior a las USD $ 2.000 millones, lo que indica que este canal es el más utilizado y probablemente el más consolidado. Le sigue el canal RedPropia con un volumen acumulado superior a USD $ 1.600 millones, lo que muestra una relevancia significativa, posiblemente debido a su accesibilidad y cobertura en diferentes regiones. Finalmente, el canal Corresponsales tiene un volumen acumulado cercano a US $ 1.100 millones, lo que sugiere que es un canal menos preferido o utilizado, quizá por limitaciones en alcance o características específicas del servicio.

La ilustración 7 muestra la evolución mensual del monto acumulado en millones de dólares (USD) por canal de pago (APN, Corresponsales y Red Propia) durante el periodo analizado.

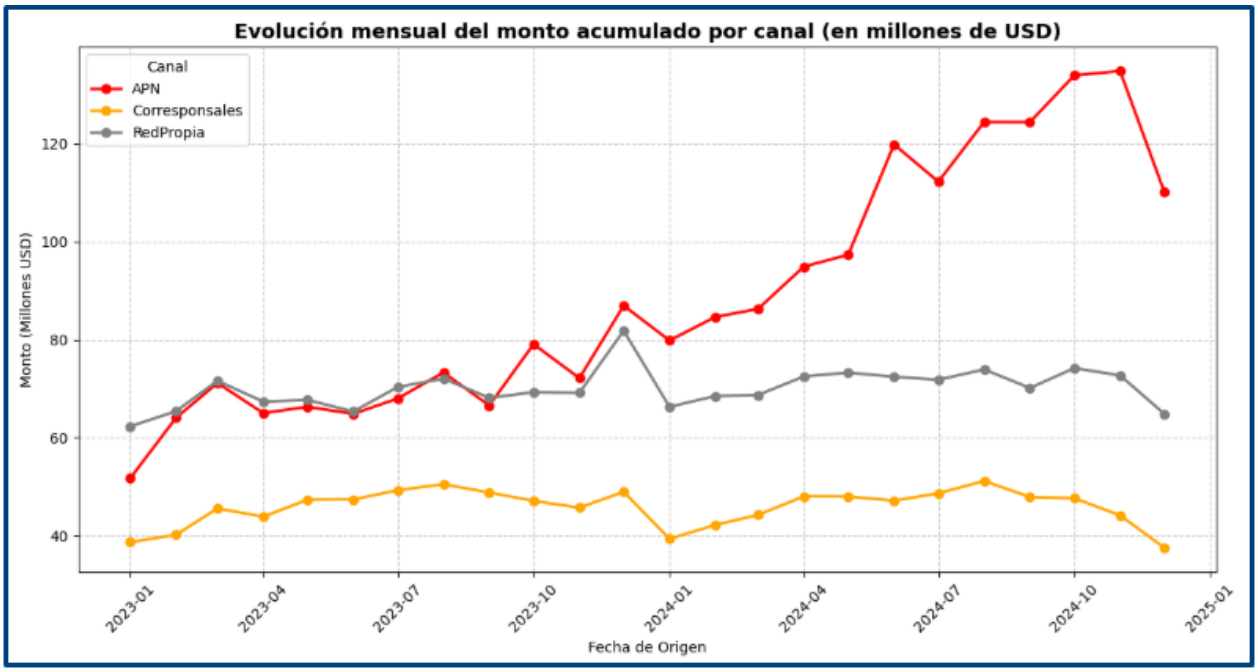

Ilustración 7 Distribución del monto en USD por canal y fecha

Se observa una tendencia relativamente estable con algunas fluctuaciones entre los canales.

La disminución en los montos del mes de diciembre de 2024 se debe a que los datos de ese mes solo registran 24 días. Por tanto, esta caída no representa una disminución real en la actividad, sino un efecto del corte temporal en la fuente de datos según el análisis del capítulo 4 numeral 4.6.

4.8 DATOS ATIPICOS

4.8.1 Distribución del monto diario en USD

A continuación, mediante la ilustración 8 se presenta un gráfico de cajas y bigotes que permite visualizar la distribución diaria del monto total en USD de las remesas originadas en un día. Esta representación gráfica y la tabla 6 facilitan la identificación de la mediana, la dispersión de los datos, así como la presencia de valores atípicos.

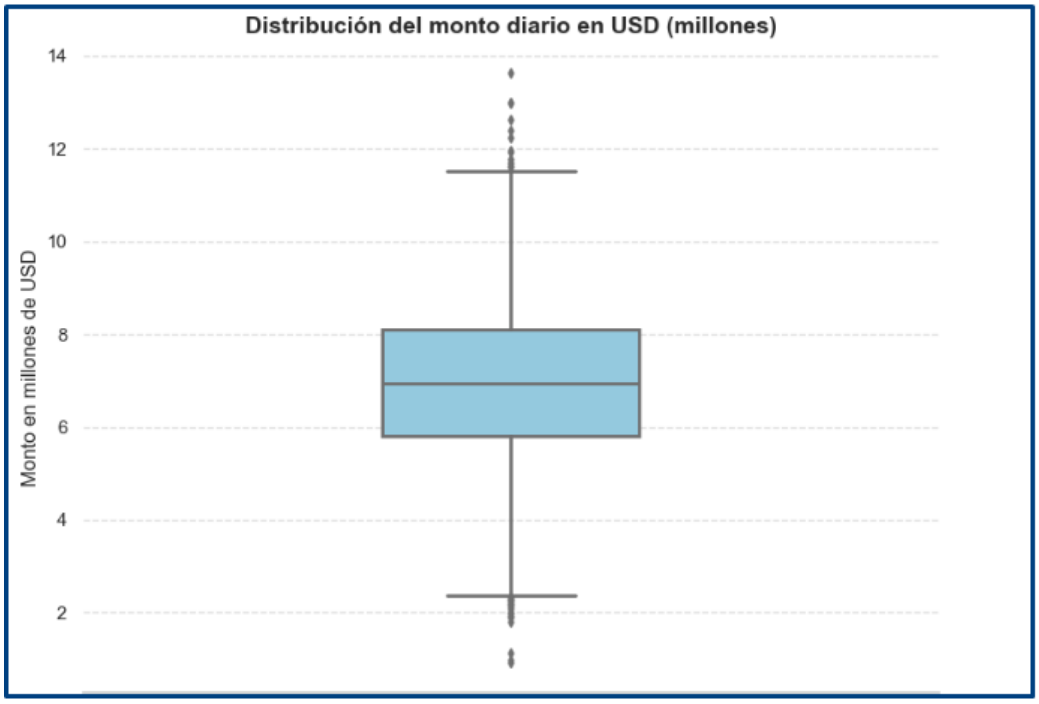

Ilustración 8 Distribución del monto diario en USD

Estadísticas descriptivas del monto diario en USD (millones)

Estadística	Monto_USD (Millones)
Mínimo	0.92
Q1	5.81
Mediana	6.95
Promedio	6.78
Q3	8.11
Máximo	13.63
IQR	2.3

Tabla 6 Datos estadísticos generales

Como se puede establecer en las ilustraciones anteriores, el promedio diario durante el periodo observado es de USD $6.78 millones con mediana USD $6.95 millones con mínimo en USD 920 mil y máximo en $13.63 millones.

Para el entendimiento de estos datos es relevante profundizar en el comportamiento por día de semana, la influencia de días festivos, así como su tendencia y ciclicidad.

4.8.2 Distribución en Monto_USD según el Canal

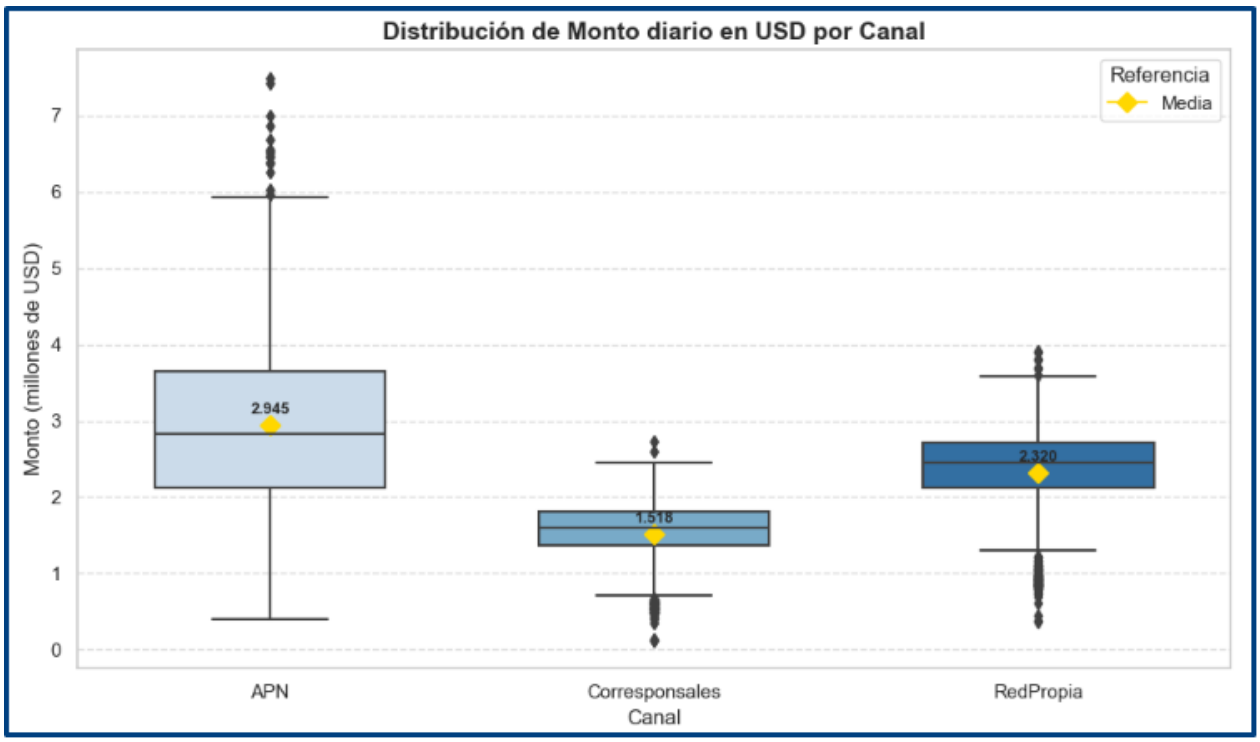

Ilustración 9 Distribución de monto diario en dólares por canal

La gráfica de cajas representada en la ilustración 9, muestra la distribución del monto diario en USD según el canal utilizado: APN, Corresponsales y RedPropia. Se observa que el canal APN presenta la mayor dispersión en los valores, con una media aproximada de 2.95 millones de USD y una cantidad considerable de valores atípicos por encima de los 6 millones, lo que sugiere una alta variabilidad en los montos procesados a través de este canal. Por su parte, el canal RedPropia presenta una distribución más concentrada, con una media cercana a los 2.32 millones de USD y menos presencia de valores extremos. Finalmente, el canal Corresponsales exhibe la media más baja, de aproximadamente 1.51 millones de USD, junto con una menor dispersión y pocos valores atípicos. En conjunto, estos resultados indican que APN es el canal con mayores volúmenes y variabilidad en las transacciones, mientras que Corresponsales y RedPropia muestran un comportamiento más estable y acotado en los montos diarios.

4.8.3 Distribución en Monto_USD según el país de origen

La visualización de la ilustración 10 muestra los códigos de los 10 países con montos en dólares más significativos y los demás quedarán en un grupo llamados “Otros”

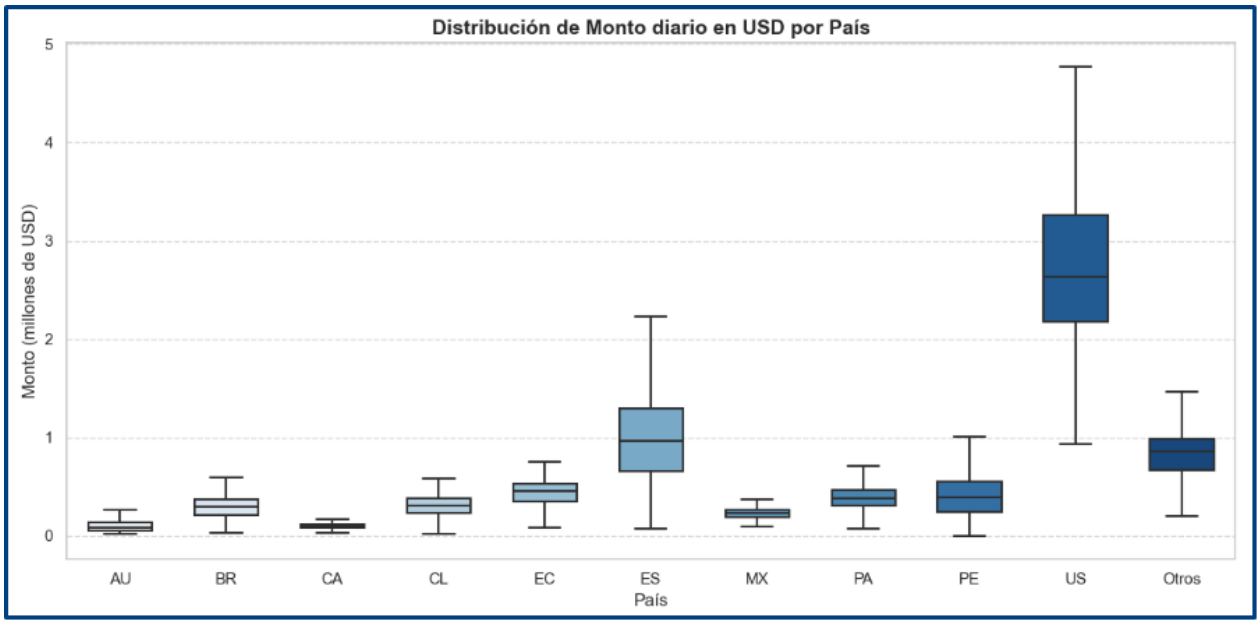

Ilustración 10 Distribución de monto diario en dólares por país de origen

La gráfica de cajas presenta la distribución diaria del monto en USD por país de origen de las remesas. Se incluyen los diez países con mayores montos acumulados y una categoría adicional denominada "Otros", que agrupa al resto. Se observa que Estados Unidos (US) concentra los montos más elevados y con mayor variabilidad, presentando una mediana notablemente más alta y una gran cantidad de valores atípicos, algunos superiores a los 7 millones de dólares diarios. España (ES) y Ecuador (EC) también destacan por registrar montos relativamente altos y con una dispersión considerable. Por otro lado, países como Australia (AU), Canadá (CA), Chile (CL), Panamá (PA) y Brasil (BR) muestran distribuciones más compactas, con montos diarios significativamente más bajos y menos variabilidad. La categoría "Otros" presenta valores más reducidos y estables, lo podría indicar que la mayor parte del volumen de las transacciones se concentra en unos pocos países.

4.8.4 Distribución en Monto_USD según el mes

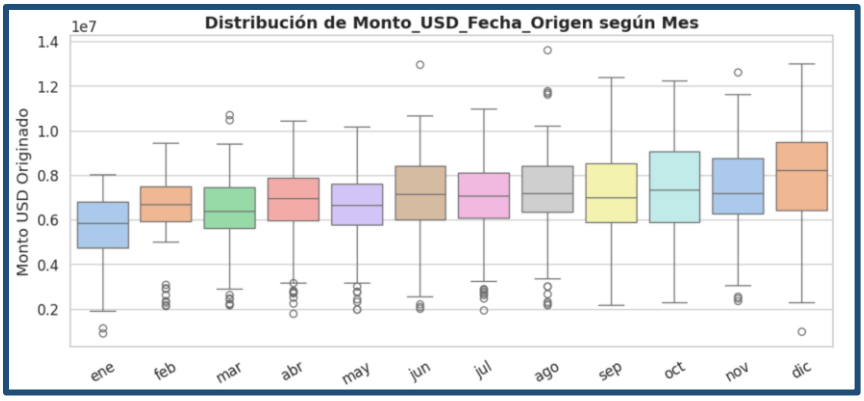

Ilustración 11 Distribución de monto en dólares por mes

La gráfica presentada a través de la ilustración 11 muestra la distribución del Monto_USD según el mes del año. A partir de esta visualización, se pueden identificar patrones estacionales y variabilidad en el volumen de remesas originadas mes a mes. Se observa que, si bien la mediana del monto mensual se mantiene relativamente estable a lo largo del año, hay algunos meses que presentan una mayor dispersión y presencia de valores atípicos.

En particular, los meses de diciembre, octubre y septiembre muestran una amplitud intercuartílica más alta, así como montos extremos superiores más pronunciados, lo cual sugiere que en estos periodos se generan envíos significativamente mayores en comparación con otros meses. Esto podría estar asociado a dinámicas de consumo estacional, como festividades de fin de año o actividades económicas puntuales.

Por el contrario, los primeros meses del año, como enero y febrero, exhiben distribuciones más contenidas, con medianas ligeramente más bajas y menor variabilidad. Esto podría indicar una desaceleración en el envío de remesas luego del pico de diciembre.

El análisis exploratorio realizado en este capítulo permitió evidenciar que existe cierta estacionalidad en los montos de remesas originadas, con tendencias de mayor volumen en el último trimestre del año, lo que resulta relevante para el diseño de modelos predictivos, ya que sugiere que incorporar la variable mes podría aportar valor explicativo en la estimación de remesas.

4.8.5 Distribución en Monto_USD según semana del mes

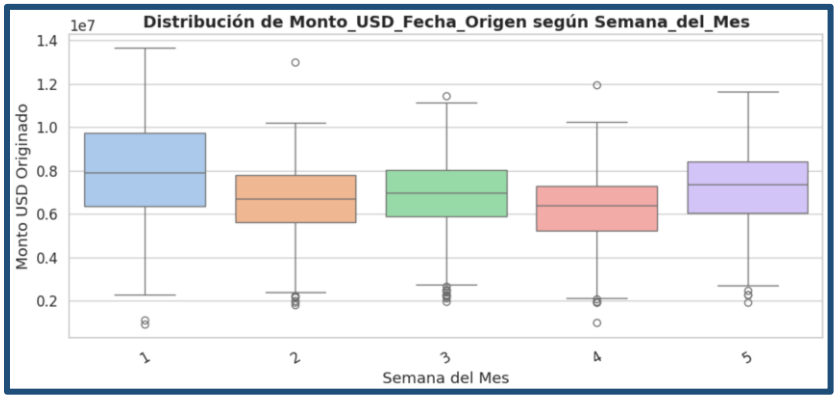

Ilustración 12 Distribución en Monto_USD según semana del mes

En la ilustración 12 se presenta la distribución del Monto_USD según la semana del mes en que se origina la remesa. El análisis permite observar ciertos patrones relevantes que podrían relacionarse con dinámicas operativas o comportamientos financieros recurrentes dentro del mes.

Destaca el hecho de que la primera semana del mes presenta no solo la mediana más alta, sino también una mayor dispersión y presencia de valores atípicos elevados. Este comportamiento sugiere que durante el inicio del mes se concentra un mayor volumen de envíos, posiblemente asociado al pago de obligaciones recurrentes o transferencias regulares hacia los hogares receptores.

En contraste, las semanas segunda, tercera y cuarta muestran medianas más bajas y distribuciones más contenidas. Esto podría indicar una menor actividad remesadora en el centro del mes. La quinta semana, que solo está presente en ciertos meses, presenta una mediana intermedia, aunque con menor dispersión en comparación con la primera semana.

En conjunto, estos resultados apoyan la hipótesis de que el comportamiento de las remesas tiene una dimensión intra-mensual significativa. Esta evidencia puede ser relevante al momento de modelar la variable temporal, ya que sugiere que incluir la posición de la semana dentro del mes podría mejorar la precisión predictiva del modelo.

4.8.6 Distribución en Monto_USD según día de la semana

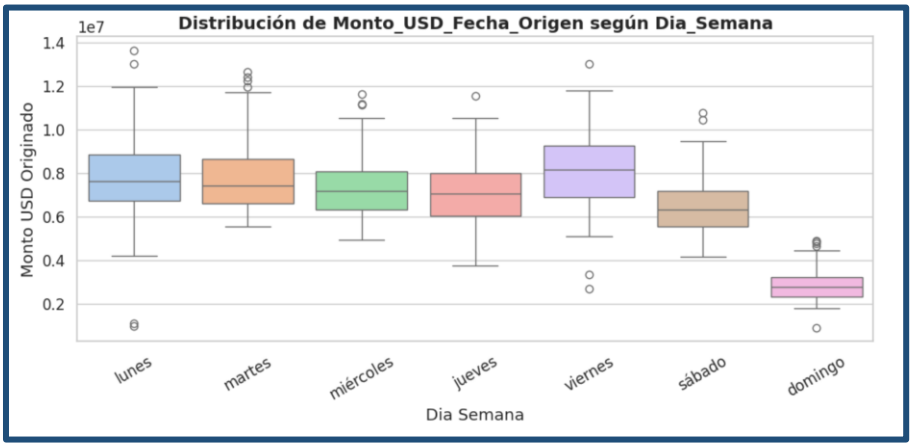

Ilustración 13 Distribución en Monto_USD según día de la semana

La gráfica ilustra la distribución del monto originado en USD a lo largo de los diferentes días de la semana. Se evidencia un patrón operativo claramente definido: los valores más altos tienden a concentrarse entre lunes y viernes, lo cual indica que el grueso de las remesas se origina durante los días hábiles.

En particular, lunes y viernes presentan tanto medianas elevadas como una mayor dispersión en los montos, lo cual puede estar vinculado a la programación de pagos institucionales o decisiones financieras al inicio y cierre de la semana. Los martes, miércoles y jueves muestran distribuciones relativamente similares, con medianas ligeramente menores, aunque aún dentro del rango de actividad regular.

Por otro lado, los fines de semana (sábado y domingo) presentan una clara disminución en la mediana de los montos originados, siendo el domingo el día con la menor actividad. Esta reducción puede deberse a restricciones operativas o una menor disponibilidad de canales activos para la emisión de remesas durante estos días.

Estos hallazgos reafirman la existencia de un comportamiento cíclico semanal en el envío de remesas y subrayan la importancia de incorporar la variable día de la semana como factor explicativo en modelos predictivos, en especial si se trabaja con datos de alta frecuencia como registros diarios.

A pesar de identificar datos atípicos en varias de las distribuciones analizadas, se optó por no eliminarlos ni transformarlos, ya que estos valores extremos pueden representar comportamientos reales del fenómeno bajo estudio, como picos de envío asociados a eventos específicos, estacionalidades marcadas o transacciones de alto valor no recurrentes. Excluir estos datos podría distorsionar la comprensión del comportamiento auténtico de las remesas, especialmente en el contexto económico y social actual.

4.9 ANÁLISIS TEMPORAL

Dado que la variable de interés es el monto diario en dólares organizado por la variable Fecha de Origen, es importante analizar tendencias y ciclos en el tiempo.

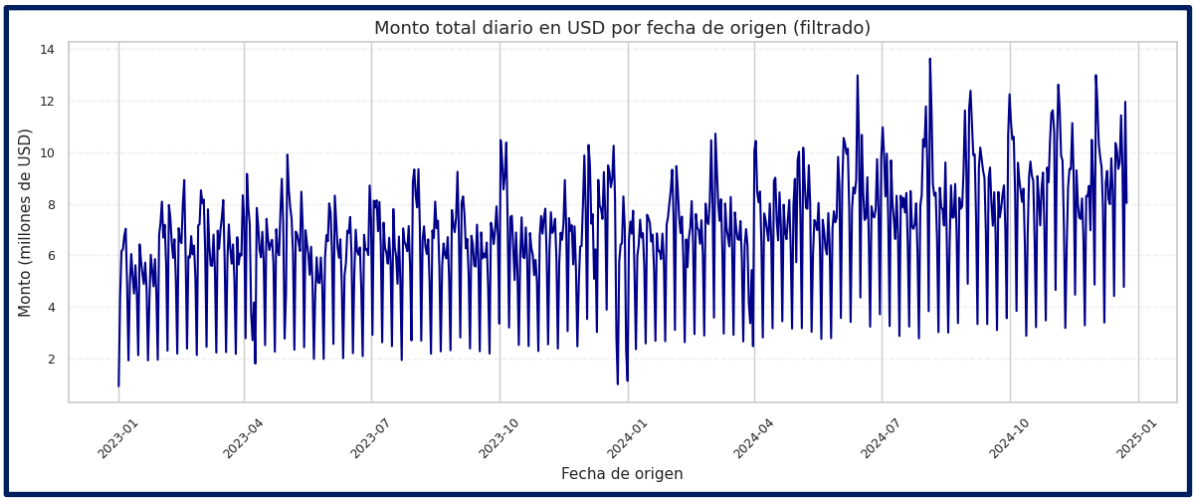

Ilustración 14 Serie de Tiempo: Monto en dólares por fecha de origen

En la ilustración 14 se presenta la evolución del monto total diario en dólares (USD) de las remesas originadas y pagadas por el APPD objeto de estudio, para el periodo comprendido entre el 1 de enero de 2023 y el 24 de diciembre de 2024.

El análisis de los datos indica que los montos de las remesas tienden a disminuir durante los fines de semana, seguidos de aumentos sostenidos en los días hábiles, lo que insinúa una mayor actividad de origen de remesas durante la semana laboral. A lo largo del periodo analizado, especialmente en 2024, se destaca una tendencia creciente en los montos promedio diarios.

4.9.1 Rezagos

La ilustración 15 presenta el gráfico de rezagos de la serie de tiempo correspondiente al monto total diario en dólares (USD) de las remesas del APPD objeto de estudio organizada por fecha de origen.

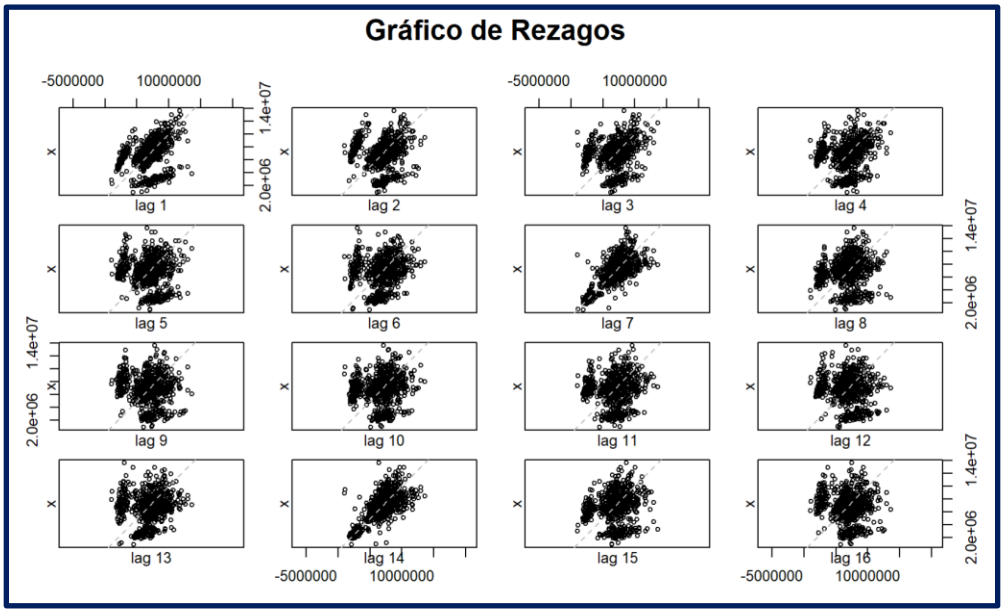

Ilustración 15 Rezagos

En este caso, se observan patrones estructurados en los primeros rezagos (lag 1 a lag 7), evidenciando una dependencia temporal en la serie. La forma alargada y agrupada de los puntos alrededor de una diagonal creciente en estos primeros rezagos indica una correlación positiva significativa, es decir, que los valores altos tienden a seguir a otros valores altos, y los valores bajos a otros bajos.

A medida que el número de rezagos aumenta por encima de 7, el patrón se torna más difuso, lo que sugiere una disminución de la dependencia temporal conforme aumenta la distancia entre observaciones. Estos resultados son consistentes con una estructura de autocorrelación de corto plazo y típica en procesos con estacionalidad semanal. Este diagnóstico refuerza la pertinencia de emplear modelos autorregresivos (como AR o ARIMA) o de recurrencia semanal para la predicción del comportamiento de la serie.

4.9.2 Estacionariedad

Con el objetivo de evaluar la estacionariedad de la serie de tiempo correspondiente al monto total diario en dólares (USD) de las remesas originadas, se aplicó la prueba de Dickey-Fuller aumentada (ADF). Esta prueba contrasta la hipótesis nula de presencia de raíz unitaria (es decir, que la serie no es estacionaria) frente a la hipótesis alternativa de estacionariedad.

El resultado obtenido fue un estadístico de prueba de -9.4892 con un valor-p de 0.01, utilizando un rezago de orden 8. Dado que el valor-p es inferior al umbral convencional de significancia ($\alpha = 0.05$), se rechaza la hipótesis nula de no estacionariedad, lo cual indica que la serie es estacionaria en niveles.

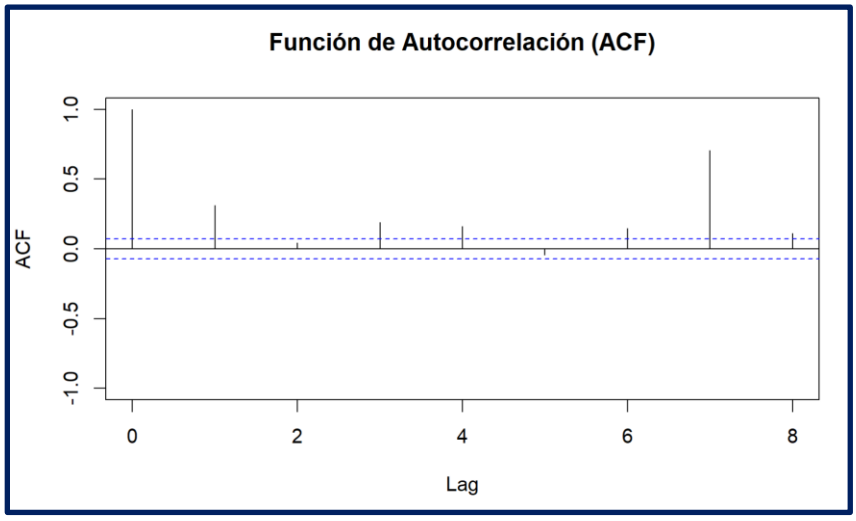

Ilustración 16 Función de Autocorrelación (ACF)

La ilustración 16 muestra la función de autocorrelación (ACF) calculada sobre la serie de tiempo correspondiente al monto diario en dólares (USD) de las remesas originadas y pagadas por el APPD analizado, considerando hasta ocho rezagos, donde se observa una autocorrelación significativamente positiva en el primer rezago (lag 1), lo cual indica que existe una dependencia temporal inmediata: los valores de un día están altamente relacionados con los del día anterior. El pico en el rezago 7 sugiere un patrón de periodicidad semanal.

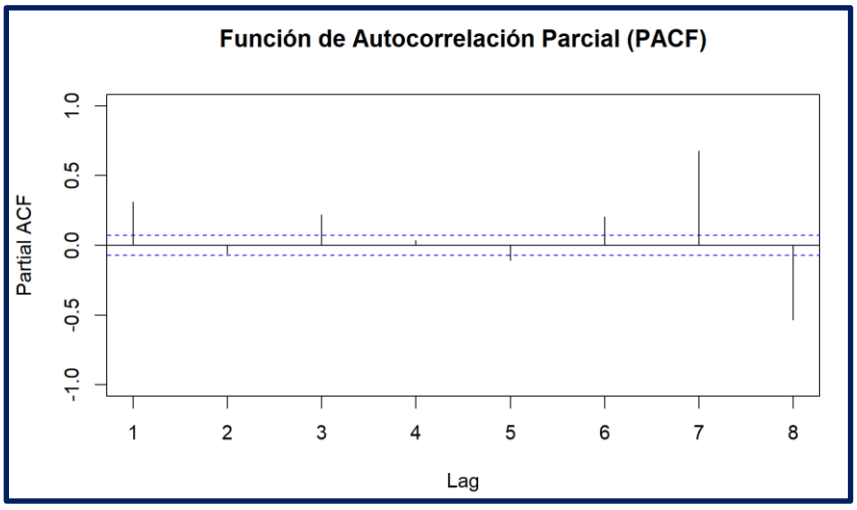

Ilustración 17 Función de Autocorrelación Parcial (PACF)

La función de autocorrelación parcial (PACF) que se visualiza en la ilustración 17 permite identificar de manera más precisa el número de rezagos significativos que deben ser considerados en la parte autorregresiva (AR) de un modelo de series de tiempo. En este caso, la PACF se calculó sobre la serie diaria

del monto total en dólares (USD) de las remesas originadas y pagadas por el APPD objeto de estudio, con un máximo de ocho rezagos.

El gráfico muestra que el primer rezago (lag 1) presenta una autocorrelación parcial positiva, aunque no significativamente distinta de cero. A partir de lag 2 hasta lag 6, las autocorrelaciones se mantienen próximas a cero y dentro del intervalo de confianza, lo que indica ausencia de efectos parciales relevantes en esos puntos. Sin embargo, en el rezago 7 se observa un pico positivo significativo, y en el rezago 8 una autocorrelación negativa también significativa.

Las pruebas aplicadas permiten concluir que la serie diaria del monto en dólares de remesas presenta evidencia estadística de estacionariedad. El test de Dickey-Fuller aumentado rechaza la hipótesis de raíz unitaria, mientras que los análisis de autocorrelación (ACF y PACF) muestran dependencia temporal de corto plazo y posibles patrones semanales. Estos resultados respaldan la aplicabilidad de modelos ARIMA sin diferenciación y refuerzan la necesidad de considerar componentes autorregresivos y estacionales en la modelación.

4.9.3 Descomposición de la serie temporal

La ilustración 18 presenta la descomposición de la serie de tiempo con el objetivo de desagregarla en sus tres componentes fundamentales: tendencia, estacionalidad y residuo.

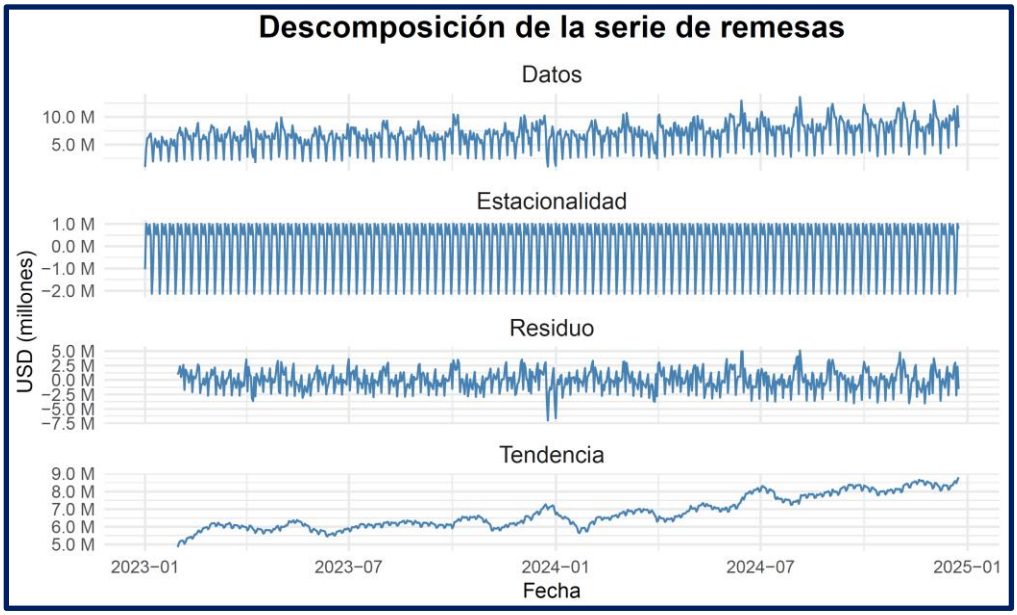

Ilustración 18 Descomposición de la serie de tiempo

La tendencia suavizada muestra un crecimiento progresivo en el monto diario promedio de remesas, especialmente a partir del segundo semestre de 2023 y durante todo 2024.

El componente estacional evidencia una clara estructura de periodicidad semanal, lo que indica que existen días con patrones repetitivos en los flujos de remesas. Este hallazgo es coherente con la operativa del negocio, en el cual la actividad se reduce durante los fines de semana y se concentra entre lunes y viernes.

El componente de ruido o residuo refleja las variaciones no explicadas por la tendencia ni la estacionalidad. Aunque la mayoría de los valores se mantienen dentro de un rango moderado, se observan picos puntuales que podrían corresponder a eventos atípicos, días festivos o interrupciones operativas.

En conjunto, esta descomposición confirma que la serie de remesas presenta una estructura estable y predecible en términos estacionales y de tendencia, lo cual habilita el uso de modelos de series de tiempo que exploten estas características para estimar el monto diario de remesas.

4.10 CORRELACIONES ENTRE VARIABLES

4.10.1 Variables numéricas

Se realizó un análisis de correlación con el objetivo de identificar qué variables numéricas están más asociadas al comportamiento del monto de remesas originadas. Para ello, se aplicó el coeficiente de correlación de Pearson entre esta variable objetivo y las demás variables continuas del conjunto de datos.

Se excluyeron del análisis las variables no numéricas, como las de tipo fecha o categóricas, y se ordenaron los resultados de mayor a menor correlación para facilitar su interpretación.

A continuación, en la ilustración 19, los resultados muestran que la variable Monto_USD_Fecha_Pago presenta la correlación más alta (0.95), lo cual es coherente, dado que ambas variables corresponden a montos financieros asociados dentro del mismo proceso operativo. Asimismo, se identificaron correlaciones muy altas con los canales de operación APN, RedPropia y Otros, lo que evidencia que estas variables representan factores determinantes en el volumen de remesas originadas. Por otro lado, varias variables asociadas a países emisores, como US (Estados Unidos), ES (España) y BR (Brasil), también mostraron correlaciones significativas, superiores a 0.88, lo cual es consistente con el rol que desempeñan estos países como principales fuentes de envío de remesas hacia Colombia.

Finalmente, el tipo de cambio (Valor_COP_TRM) evidenció una correlación negativa muy débil (-0.085), lo que sugiere que su relación con el monto originado no es lineal ni directa. Esto indica que, si bien podría influir en la dinámica de las remesas, su efecto no se manifiesta claramente a través de una simple correlación.

Variable	Correlación con Monto_USD_Fecha_Origen
Monto_USD_Fecha_Pago	0.95
APN	0.934
Otros	0.927
RedPropia	0.926
Corresponsales	0.893
US	0.893
ES	0.886
BR	0.883
CA	0.823
MX	0.81
PA	0.794
EC	0.764
CL	0.713
AU	0.627
PE	0.505
Valor_COP_TRM	-0.085

Ilustración 19 Correlación de variables numéricas

4.10.2 Valores booleanos

Se evaluó la relación entre la variable dependiente monto originado y un conjunto de variables booleanas, específicamente aquellas que indican si una fecha corresponde a un fin de semana o a un día festivo en Estados Unidos, Colombia o ambos. Para ello, se transformaron estas variables lógicas a valores binarios (0 y 1), y se aplicó nuevamente el coeficiente de correlación de Pearson con el fin de cuantificar la intensidad y dirección de la relación lineal.

Variable Booleana	Correlación con Monto_USD_Fecha_Or
Feriado_USA	−0,067
Feriados_Colombia	−0,182
Es_Fin_de_Semana	−0,594

Ilustración 20 Correlación variables booleanas

Los resultados de la ilustración 20 muestran una correlación negativa en todos los casos, lo cual sugiere que la presencia de estas condiciones tiende a estar asociada con una reducción en el monto de remesas

originadas. De forma particular, la variable Es_Fin_de_Semana presenta la correlación negativa más alta (-0.594), lo que indica que, durante los fines de semana, el monto originado de remesas tiende a disminuir de manera significativa.

Por otro lado, las variables relacionadas con festivos presentan correlaciones negativas más suaves: Feriados_Colombia (-0.182) y Feriado_USA (-0.067). Si bien estas asociaciones son más débiles, también apuntan hacia una menor actividad de remesas durante los días festivos, aunque con menor impacto que el observado en los fines de semana. Esto puede deberse a que algunos sistemas operan parcialmente en festivos, o que los usuarios anticipan o posponen sus transacciones cuando se aproxima una fecha no laborable.

En conjunto, estos hallazgos sugieren que las variables temporales asociadas al calendario deben ser tenidas en cuenta en los modelos predictivos, especialmente Es_Fin_de_Semana, dado su impacto sistemático y relevante sobre la dinámica de las remesas.

4.10.3 Variables Categóricas

A continuación, se presenta un análisis de varianza (ANOVA) para evaluar el impacto de variables categóricas relacionadas con el tiempo —como el mes, la semana del mes y el día de la semana— sobre el comportamiento del monto en USD. Este análisis permite identificar si existen diferencias significativas en los promedios según estas categorías temporales.

Resultados ANOVA - Variables Categóricas

Variable	Suma de cuadrados	Grados de libertad	Estadístico F	Valor p
Mes	255073468859357.0	11	4.841	0.0
Semana_del_Mes	233218554917436.9	4	12.214	0.0
Dia_Semana	2042493015317037.0	6	150.392	0.0

Ilustración 21 Correlación variables categóricas

La Ilustración 21 con los resultados ANOVA muestra que las variables categóricas Mes, Semana del Mes y Día de la Semana tienen un efecto estadísticamente significativo sobre la variable de interés (monto originado en USD), ya que en todos los casos el valor p es 0.000. Entre ellas, Día de la Semana presenta el mayor impacto, con un estadístico F de 150.392, lo que indica que existen diferencias sustanciales en los promedios según el día. Le siguen Semana del Mes (F = 12.214) y Mes (F = 4.841), que también presentan variaciones significativas, aunque en menor magnitud. Esto sugiere que el comportamiento del monto varía de forma importante según la estructura temporal, siendo especialmente relevante el día en que se realiza la transacción.

4.10.4 Matriz de Correlaciones - Variables Numéricas

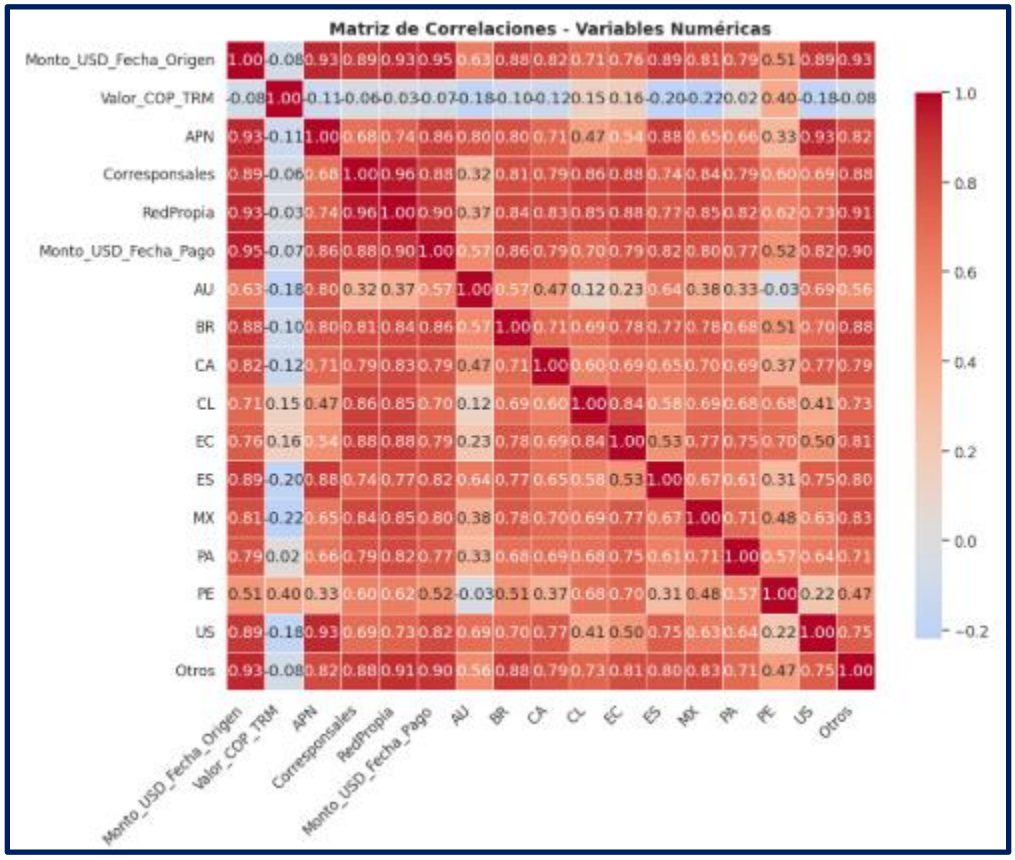

Ilustración 22 Matriz de correlación - Variables numéricas

La matriz de correlaciones de la ilustración 22 permite identificar la relación lineal entre variables numéricas del análisis, especialmente entre los diferentes canales, países y montos asociados a las fechas de origen. Se observa una alta correlación positiva entre Monto_USD_Fecha_Origen y variables como APN (0.93), RedPropia (0.90), y Monto_USD_Fecha_Pago (0.86), lo que sugiere que estas variables evolucionan de forma conjunta. Asimismo, varios países como EC, CA, CL y US muestran correlaciones altas con los montos, indicando su fuerte participación en los flujos. Por otro lado, la variable Valor_COP_TRM no presenta una relación lineal significativa con la mayoría de las variables (correlaciones cercanas a 0 o incluso negativas), lo que sugiere independencia con respecto a los montos transaccionados en USD. En general, la matriz resalta grupos de variables fuertemente interrelacionadas, útiles para la reducción de dimensiones o construcción de modelos predictivos.

5 IDENTIFICACIÓN DE VARIABLES PARA LOS MODELOS

La construcción de modelos predictivos requiere una cuidadosa selección y transformación de las variables explicativas. En este proyecto, el APPD objeto de estudio proporcionó un conjunto inicial de datos estructurados provenientes de archivos planos, con información detallada sobre las remesas pagadas entre el 1 de enero de 2023 y el 31 de diciembre de 2024. Esta base contenía variables clave como el monto en dólares, el canal de pago y el país de origen las cuales fueron consolidadas y preparadas para su análisis.

Con el fin de mejorar la capacidad explicativa del modelo, se incorporaron variables complementarias derivadas de la fecha de origen de cada transacción. Entre ellas se incluyeron indicadores temporales como el día de la semana, la semana del mes, el mes calendario y variables binarias que identifican la ocurrencia de feriados en Colombia, o en Estados Unidos. Adicionalmente, se integró la Tasa Representativa del Mercado (TRM) diaria, dada su relevancia para el comportamiento de los flujos de remesas en dólares.

5.1 SELECCIÓN DE VARIABLES

La identificación y validación de estas variables se realizó con la colaboración de expertos del APPD, quienes aportaron su conocimiento operativo del negocio. Esta retroalimentación fue determinante para reconocer variables que, si bien no mostraban una alta correlación lineal en los análisis estadísticos iniciales, sí reflejaban patrones significativos desde el punto de vista operativo, como la estacionalidad intra-mensual y el impacto de los días no laborables sobre la dinámica de origen de remesas.

5.2 VARIABLES REZAGADAS

A partir de ese proceso, se construyó un conjunto depurado de variables que combina elementos transaccionales, temporales y macroeconómicos. Teniendo como referencia el análisis exploratorio descrito en el capítulo anterior, se procedió a estructurar el conjunto definitivo de datos con enfoque predictivo, asegurando la coherencia temporal del modelo. Dado que el objetivo es estimar el monto total en dólares que se originará en un día determinado, se excluyó deliberadamente el uso de variables cuyo valor corresponde a la misma fecha de predicción, ya que en un escenario real estas aún no estarían disponibles al momento de generar la estimación. Por lo tanto, algunas variables explicativas fueron rezagadas, de modo que se utilice únicamente información observable antes del instante de la predicción, evitando así cualquier fuga de información desde el futuro hacia el pasado.

En este sentido, se aplicaron transformaciones de rezago (lags) a las siguientes variables:

- Monto en dólares por fecha de pago.
- Montos por canal de pago: APN, Red Propia y Corresponsales.
- Montos por país de origen: se seleccionaron los diez países con mayor volumen y se agruparon los restantes bajo la categoría “Otros”.

Los rezagos se establecieron en 7 días, en concordancia con los hallazgos del análisis temporal presentado en

el capítulo 4, el cual evidenció dependencias significativas en este intervalo. Estas transformaciones permiten capturar ciclos semanales, lo cual es fundamental para mejorar la precisión del modelo al anticipar el comportamiento del monto diario de remesas originadas.

5.3 VARIABLES TEMPORALES

Además de las variables rezagadas, se incluyeron variables temporales como el día de la semana, el mes y la semana del mes, que están estrechamente relacionadas con la dinámica del proceso de remesas. También se incluyeron variables que indican si la fecha de origen de la remesa corresponde a un fin de semana, así como variables booleanas que identifican si ese día es festivo en Estados Unidos o en Colombia.

Estas variables pueden tener un impacto significativo en la predicción de los montos en dólares de las remesas originadas en un día, ya que las remesas muestran una variación estacional dependiendo de si se generan en días de la semana específicos (como lunes o viernes) o en ciertos meses del año (como diciembre, cuando las remesas tienden a aumentar debido a las festividades) y por otro lado, las semanas del mes también afectan las remesas, ya que los ingresos de los trabajadores migrantes pueden estar ligados a ciclos semanales.

Por ejemplo, los días de la semana pueden capturar la variabilidad en el comportamiento de las remesas dependiendo de los días laborales o festivos. En cambio, los meses reflejan el impacto de las estacionalidades que afectan los flujos de remesas. Las semanas del mes permiten que el modelo ajuste aún más las predicciones a patrones repetitivos dentro del ciclo mensual.

5.4 VARIABLES CATEGÓRICAS CODIFICADAS CON ONE-HOT

Las variables temporales, como el día de la semana y el mes, son de naturaleza categórica, lo que significa que no pueden ser utilizadas directamente por modelos de aprendizaje automático sin una transformación adecuada. Para ello, se utilizó la codificación One-Hot, un enfoque estándar que convierte las categorías en variables binarias, lo que permite que el modelo pueda procesarlas eficazmente.

Por ejemplo, el día de la semana tiene 7 categorías posibles (lunes, martes, miércoles, etc.), y mediante la codificación One-Hot, se generan 7 columnas en el conjunto de datos, donde cada columna representa un día de la semana y toma el valor de 1 si el día corresponde a esa categoría o 0 en caso contrario. Este mismo proceso se aplicó a los meses y las semanas del mes.

El uso de One-Hot Encoding es fundamental, ya que permite que el modelo capture las relaciones no lineales entre los días, meses o semanas sin suponer un orden numérico entre ellos. Además, esta transformación facilita el aprendizaje de patrones complejos en las variables categóricas y mejora el rendimiento del modelo.

5.5 CONJUNTO FINAL DE DATOS

El conjunto de datos utilizado para entrenar los modelos predictivos fue construido siguiendo una estructura unificada para todos los enfoques, tanto de series de tiempo como de aprendizaje automático. A partir del análisis

exploratorio y del proceso de transformación descrito previamente, se conformó una tabla consolidada con una observación por cada día (fecha de origen), que incluye la variable objetivo (Monto_USD_Originado), las variables rezagadas de montos por canal y país, la TRM, y variables temporales codificadas (como día de la semana, mes, semana del mes, festivos en Colombia y Estados Unidos, etc.).

Esta estructura permitió aplicar distintos tipos de modelos sobre la misma base, garantizando comparabilidad y consistencia metodológica. Las variables categóricas fueron transformadas mediante codificación One-Hot, y las variables numéricas fueron tratadas para evitar fugas de información, incorporándolas únicamente de forma rezagada.

La siguiente tabla presenta la estructura final del conjunto de datos empleado para el entrenamiento y prueba de los modelos:

Variable	Tipo	Descripción
Fecha_Origen	object	Indice
Monto_USD_Originado	float64	Variable objetivo
Es_Fin_de_Semana	bool	Indica si la fecha es sabado o domingo
Feriado_USA	bool	Indica si fue feriado en Estados Unidos
Feriados_Colombia	bool	Indica si fue feriado en Colombia
Valor_COP_TRM_lag7	float64	Tasa representativa del mercado (TRM) rezagada 7 días
APN_lag7	float64	Monto diario por canal de pago (APN) rezagado 7 días
Corresponsales_lag7	float64	Monto diario por canal de pago (Corresponsales) rezagado 7 días
RedPropia_lag7	float64	Monto diario por canal de pago (RedPropia) rezagado 7 días
AU_lag7	float64	Monto diario originado desde AU rezagado 7 días
BR_lag7	float64	Monto diario originado desde BR rezagado 7 días
CA_lag7	float64	Monto diario originado desde CA rezagado 7 días
CL_lag7	float64	Monto diario originado desde CL rezagado 7 días
EC_lag7	float64	Monto diario originado desde EC rezagado 7 días
ES_lag7	float64	Monto diario originado desde ES rezagado 7 días
MX_lag7	float64	Monto diario originado desde MX rezagado 7 días
PA_lag7	float64	Monto diario originado desde PA rezagado 7 días
PE_lag7	float64	Monto diario originado desde PE rezagado 7 días
US_lag7	float64	Monto diario originado desde US rezagado 7 días
Otros_lag7	float64	Monto diario originado desde Otros rezagado 7 días
Mes_abr	bool	Variable dummy para el mes de Abr
Mes_ago	bool	Variable dummy para el mes de Ago
Mes_dic	bool	Variable dummy para el mes de Dic
Mes_ene	bool	Variable dummy para el mes de Ene
Mes_feb	bool	Variable dummy para el mes de Feb
Mes_jul	bool	Variable dummy para el mes de Jul
Mes_jun	bool	Variable dummy para el mes de Jun
Mes_mar	bool	Variable dummy para el mes de Mar
Mes_may	bool	Variable dummy para el mes de May
Mes_nov	bool	Variable dummy para el mes de Nov
Mes_oct	bool	Variable dummy para el mes de Oct
Mes_sep	bool	Variable dummy para el mes de Sep
Semana_del_Mes_1	bool	Variable dummy para la semana 1 del mes
Semana_del_Mes_2	bool	Variable dummy para la semana 2 del mes
Semana_del_Mes_3	bool	Variable dummy para la semana 3 del mes
Semana_del_Mes_4	bool	Variable dummy para la semana 4 del mes
Semana_del_Mes_5	bool	Variable dummy para la semana 5 del mes
Dia_Semana_Num_1	bool	Variable dummy para el día 1 de la semana (1=Lunes, ..., 7=Domingo)
Dia_Semana_Num_2	bool	Variable dummy para el días 2 de la semana
Dia_Semana_Num_3	bool	Variable dummy para el día 3 de la semana
Dia_Semana_Num_4	bool	Variable dummy para el día 4 de la semana
Dia_Semana_Num_5	bool	Variable dummy para el día 5 de la semana
Dia_Semana_Num_6	bool	Variable dummy para el día6 de la semana
Dia_Semana_Num_7	bool	Variable dummy para el día 7 de la semana

Tabla 7 Conjunto de entrenamiento

Esta estructura permitió que todos los modelos trabajaran con una base homogénea, donde la diferencia radicó en la forma como cada técnica explota la temporalidad: los modelos de series de tiempo utilizan principalmente la variable objetivo, mientras que los modelos de machine learning incorporan rezagos, variables adicionales y codificaciones categóricas para capturar relaciones no lineales.

5.6 DIVISIÓN DEL CONJUNTO DE DATOS: ENTRENAMIENTO Y PRUEBA

Con el objetivo de evaluar de manera realista el desempeño de los modelos predictivos, se definió una estrategia de división temporal del conjunto de datos. Esta decisión se fundamenta en la naturaleza secuencial del problema, donde se busca predecir montos futuros a partir de información históricamente disponible, evitando cualquier fuga de información desde el futuro hacia el pasado.

En este sentido, se estableció como conjunto de entrenamiento todos los registros cuya fecha de origen se encuentra comprendida entre el 1 de enero de 2023 y el 30 de septiembre de 2024. Este subconjunto representa aproximadamente el 85% del total de observaciones y se utilizó para ajustar y calibrar los modelos. Por su parte, el conjunto de prueba abarca las observaciones comprendidas entre el 1 de octubre y el 24 de diciembre de 2024, periodo que se mantuvo reservado para validar la capacidad de generalización de los modelos y simular un escenario de predicción futura real. Esta ventana temporal permite evaluar cómo se comporta cada enfoque ante datos no vistos y medir con objetividad su precisión mediante métricas estandarizadas.

La proporción utilizada para la división del conjunto de datos —85% para entrenamiento y 15% para prueba— responde a recomendaciones ampliamente aceptadas en la literatura sobre modelado de series temporales. De acuerdo con Hyndman y Athanasopoulos [11], en contextos donde la serie es suficientemente larga, destinar entre el 70% y 90% de los datos para el ajuste del modelo es una práctica válida, siempre que se preserve la estructura temporal y se garantice una muestra de prueba representativa para validar el desempeño. En este proyecto, al tratarse de una serie diaria con estacionalidad semanal y mensual, se priorizó una base amplia de entrenamiento que permitiera al modelo aprender adecuadamente estos patrones.

Por otro lado, conservar una fracción final del 15% como conjunto de prueba (equivalente a casi tres meses de datos) permite evaluar la capacidad de generalización del modelo sobre un horizonte temporal reciente y operacionalmente relevante. Esta estrategia es coherente con enfoques propuestos en trabajos aplicados en predicción financiera [16], donde se sugiere que, en escenarios no estacionarios o sujetos a cambios graduales, el uso de una ventana continua y cronológicamente posterior para la validación simula de manera más precisa el desempeño futuro del modelo en producción.

6 APLICACIÓN DE MODELOS

6.1 MODELO ARIMA

Como primer enfoque para la predicción del monto diario en dólares de remesas originadas, se implementó el modelo ARIMA (AutoRegressive Integrated Moving Average), una técnica clásica de series de tiempo que permite capturar patrones de autocorrelación y estacionalidad.

Durante el análisis descriptivo de la serie se realizaron diversas pruebas para determinar su estacionariedad, un requisito fundamental para la aplicación del modelo ARIMA. En particular, se empleó la prueba de Dickey-Fuller aumentada (ADF), lo que permitió identificar de manera estadística que la serie no presenta raíz unitaria, pues el p-valor obtenido fue inferior al umbral del 5%. Además, el análisis de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) reveló patrones consistentes en los primeros rezagos, especialmente en los lags 1 y 7, evidenciando la presencia de una estructura cíclica semanal. Estos hallazgos confirman que, a pesar de la existencia de estacionalidad, la serie resulta estacionaria en niveles, lo que valida el uso directo del modelo ARIMA sin requerir diferenciación adicional. Con base en estos resultados, se puede proceder con la construcción y ajuste del modelo ARIMA para la predicción de los montos diarios de remesas.

Pruebas y entrenamiento del modelo

Una vez estructurado el conjunto de datos y construida la serie temporal diaria del monto originado en dólares por remesas, se procedió a ajustar un modelo ARIMA como primer enfoque de predicción. Para ello, se utilizó la función auto.arima() del paquete forecast en R, la cual selecciona de forma automática la mejor combinación de parámetros (p, d, q) optimizando el criterio de información de Akaike (AIC).

Se utilizó únicamente la serie temporal diaria Monto_USD_Originado, sin variables exógenas ni codificaciones. No se aplicó escalamiento.

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,462,677 USD	El error promedio entre las predicciones y los valores reales es de aproximadamente 1.46 millones de USD.
RMSE (Raíz del Error Cuadrático Medio)	1,860,567 USD	La desviación promedio de las predicciones respecto a los valores reales es de aproximadamente 1.86 millones de USD.
MAPE (Error Porcentual Absoluto Medio)	30.40%	En promedio, el modelo comete un error del 30.4% respecto al valor real.
R^2 (Coeficiente de determinación)	-0,7%	El modelo no explica la variabilidad de la serie y su desempeño es peor que una predicción constante.

Tabla 8 Resultado aplicación modelo ARIMA

El modelo seleccionado fue un ARIMA (0,1,5), lo que indica que la serie fue diferenciada una vez para alcanzar estacionariedad (d=1), y que la estructura de dependencia temporal se explica mediante cinco componentes de media móvil (q = 5), sin incluir términos autoregresivos (p = 0). Los coeficientes estimados fueron todos estadísticamente significativos, y el modelo presentó un AIC de 20026.93.

Estas cifras sugieren un ajuste razonable para una primera aproximación.

Posteriormente, se evaluó el comportamiento de los residuos mediante para verificar si los errores del modelo cumplen con los supuestos de independencia y normalidad. Los residuos se distribuyen en torno a cero y no presentan una tendencia sistemática. Sin embargo, la función de autocorrelación (ACF) reveló picos significativos en múltiplos de 7 lags, lo cual sugiere la existencia de una estacionalidad semanal no capturada por el modelo.

Este hallazgo fue corroborado por la prueba de Ljung-Box, que arrojó un valor Q^*=508.87 con df = 9 y un p-valor < 2.2e-16, indicando que los residuos presentan autocorrelación significativa y no se comportan como ruido blanco. Por tanto, aunque el modelo ARIMA (0,1,5) representa adecuadamente la tendencia de la serie, no logra capturar del todo la estructura cíclica subyacente, lo cual plantea la necesidad de explorar modelos con componentes estacionales.

6.2 MODELO SARIMA

Dado que en el análisis exploratorio se identificaron patrones cíclicos con periodicidad semanal en los montos diarios de remesas originadas, se consideró pertinente ajustar un modelo SARIMA, que extiende la familia ARIMA al incorporar componentes estacionales. Este tipo de modelo permite capturar dinámicas periódicas propias del comportamiento operativo del sistema de remesas, como la disminución de transacciones durante fines de semana y el aumento durante días hábiles. El modelo seleccionado fue un "SARIMA(1,0,0)(0,1,1) [7]", el cual incluye una estructura autorregresiva de corto plazo, diferenciación estacional de primer orden y un término de media móvil estacional semanal. Esta configuración refleja de manera adecuada tanto la dependencia temporal inmediata como la estacionalidad de siete días identificada en la serie, permitiendo mejorar la capacidad predictiva frente al modelo ARIMA básico.

Se utilizó únicamente la serie temporal diaria Monto_USD_Originado, sin variables exógenas ni codificaciones. No se aplicó escalamiento.

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	619,828 USD	El error promedio entre las predicciones y los valores reales es de aproximadamente 620 mil USD.
RMSE (Raíz del Error Cuadrático Medio)	954,883 USD	La desviación promedio de las predicciones respecto a los valores reales se reduce a menos de 1 millón de USD.
MAPE (Error Porcentual Absoluto Medio)	11.10%	El modelo comete, en promedio, un error del 11.1% respecto al valor real.
R^2 (Coeficiente de determinación)	59,52%	Aproximadamente el 59.5% de la variabilidad en la serie es explicada por el modelo SARIMA.

Tabla 9 Resultado aplicación modelo SARIMA

Evaluación de supuestos del modelo SARIMA

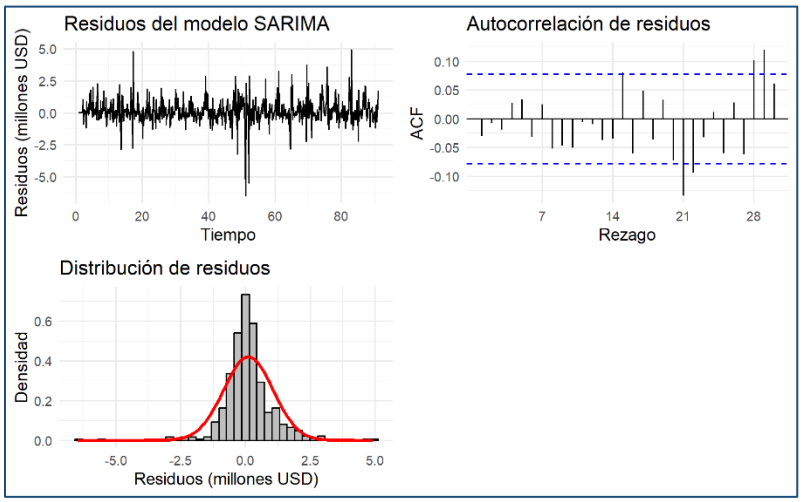

Ilustración 23 Errores del modelo SARIMA

Luego del ajuste del modelo "SARIMA(1,0,0)(0,1,1)[7]", se realizó el análisis de los residuos con el fin de verificar el cumplimiento de los supuestos estadísticos fundamentales. Los residuos presentan una distribución centrada alrededor de cero y no evidencian cambios sistemáticos en la varianza a lo largo del tiempo, lo que sugiere homocedasticidad. Además, el histograma de residuos muestra una forma aproximadamente normal, con ligera asimetría y colas algo pesadas.

En cuanto a la independencia de los errores, la función de autocorrelación (ACF) no muestra picos significativos, y los valores se mantienen dentro del intervalo de confianza, lo que indica la ausencia de autocorrelación estructural. Este hallazgo fue respaldado por la prueba de Ljung-Box, que arrojó un estadístico Q^* = 9.66, con df = 12 y un p-valor de 0.646, por lo que no se rechaza la hipótesis nula de que los residuos son ruido blanco. En contraste con el modelo ARIMA anterior, el SARIMA logra capturar adecuadamente la estacionalidad semanal y eliminar la dependencia serial en los errores. Estos

resultados validan el modelo como una representación más adecuada del comportamiento temporal de la serie de remesas.

Resultados

Es importante señalar que todas las métricas de error reportadas hasta este punto (6.1 y 6.2) —tales como MAE, RMSE, MAPE y el R^2— fueron calculadas utilizando únicamente el conjunto de entrenamiento. Por tanto, estos resultados reflejan la capacidad del modelo para ajustarse a los datos históricos que se utilizaron en su calibración.

A continuación, se ha realiza la validación sobre el conjunto de prueba, que corresponde a las observaciones comprendidas entre el 1 de octubre y el 24 de diciembre de 2024.

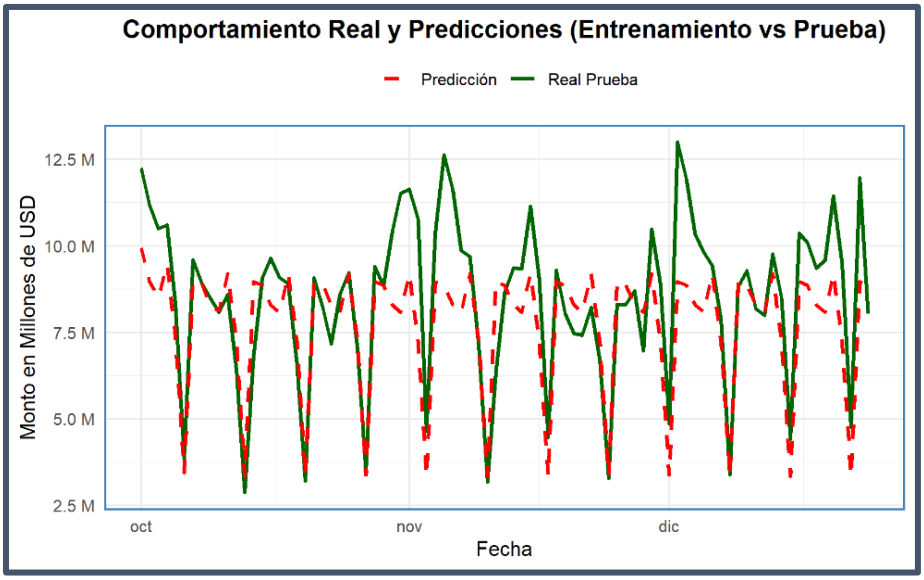

Ilustración 24 Predicciones del Modelo SARIMA

En la comparación entre el comportamiento real del monto diario de remesas originadas y las predicciones generadas por el modelo SARIMA, evidencia que el modelo logra capturar adecuadamente la estructura estacional semanal observada en la serie. A lo largo del gráfico que se muestra en la Ilustración 24 se aprecia que las predicciones siguen con coherencia la secuencia de repuntes y caídas del comportamiento real, manteniéndose dentro de rangos razonables. Sin embargo, en ciertos momentos, particularmente en presencia de picos abruptos, el modelo tiende a subestimar los valores máximos o sobreestimar los mínimos, lo que sugiere una respuesta algo suavizada frente a variaciones extremas.

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,137,196 USD	El error promedio entre las predicciones y los valores reales en el conjunto de prueba es de aproximadamente 1.14 millones de USD.
RMSE (Raíz del Error Cuadrático Medio)	1,501,492 USD	La desviación promedio de las predicciones respecto a los valores reales fue de aproximadamente 1.5 millones de USD.
MAPE (Error Porcentual Absoluto Medio)	12.9%	El modelo tuvo un error porcentual medio de 12.9% en datos no vistos, lo que confirma un buen desempeño fuera de muestra.
R^2 (Coeficiente de determinación)	59,52%	Aproximadamente el 59.5% de la variabilidad en la serie es explicada por el modelo SARIMA.

Tabla 10 Resultado aplicación modelo SARIMA - Conjunto de prueba

A pesar de estas diferencias puntuales, el desempeño general del modelo resulta satisfactorio, como lo reflejan las métricas de error obtenidas en el conjunto de prueba: un MAE de aproximadamente 1.14 millones de dólares y un MAPE del 12.9%. Estos resultados confirman que el modelo SARIMA es capaz de ofrecer estimaciones estables y razonablemente precisas para el monto diario de remesas originadas, dentro de un horizonte de corto plazo.

6.3 MODELO HOLT – WINTERS

Como complemento a los modelos autorregresivos, se implementó el método de Holt-Winters, una técnica de suavizamiento exponencial que permite capturar de manera efectiva componentes de nivel, tendencia y estacionalidad en series temporales. Este modelo es especialmente útil cuando se identifican patrones estacionales regulares y sostenidos en el tiempo, como ocurre en la serie diaria de remesas analizada, donde se observa una estructura semanal consistente. El enfoque Holt-Winters ofrece una alternativa computacionalmente eficiente y de fácil interpretación, al no requerir una formulación compleja de parámetros como en el caso de los modelos ARIMA o SARIMA. En esta aplicación, se empleó la versión aditiva del modelo, dado que la estacionalidad observada en la serie tiende a mantenerse constante en magnitud a lo largo del tiempo.

Dado que la serie de montos diarios de remesas presenta valores estrictamente positivos y una variabilidad relativa significativa entre días, se consideró pertinente evaluar el desempeño del modelo Holt-Winters bajo dos enfoques: utilizando la serie original sin transformar y aplicando una transformación logarítmica previa al ajuste. La transformación logarítmica es una práctica común en el análisis de series temporales financieras, ya que permite estabilizar la varianza, reducir el impacto de valores atípicos y tratar de manera más adecuada patrones multiplicativos mediante un modelo aditivo en escala log. De esta manera, se busca comparar si el ajuste directo sobre la serie o el ajuste sobre su

transformación logarítmica ofrece un mejor desempeño predictivo, tanto en términos absolutos como relativos, sobre el conjunto de prueba.

En resumen, se utilizó únicamente la serie temporal diaria Monto_USD_Originado, sin inclusión de variables exógenas ni codificaciones adicionales. El modelo fue ajustado directamente sobre esta serie, por lo que no se aplicaron procesos de escalamiento, normalización ni transformación de variables categóricas. Para el segundo enfoque (modelo transformado), se aplicó una transformación logarítmica a la serie antes del ajuste y se revirtió al final para la interpretación de resultados.

Serie original

Resultados

El modelo Holt-Winters ajustado sobre la serie original seleccionó un componente de tendencia y estacionalidad aditiva, con parámetros de suavizamiento α=0.606, β=0.0011 y γ=0.152. El valor relativamente alto de α sugiere que el modelo da mayor peso a los valores más recientes al actualizar el nivel de la serie, mientras que el valor extremadamente bajo de β indica una tendencia casi constante a lo largo del tiempo. Por su parte, el parámetro γ, que regula el peso asignado a la estacionalidad, refleja una actualización moderada del patrón estacional semanal.

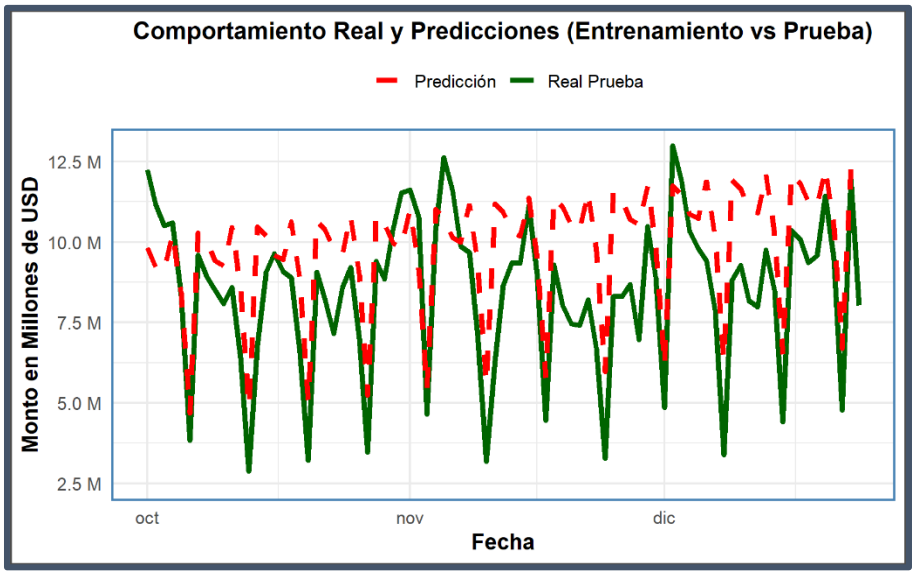

Ilustración 25 Predicciones del Modelo HW sin transformar

Según la ilustración 25, los coeficientes estacionales estimados para los siete días de la semana muestran una variabilidad considerable, con valores positivos en los días de mayor actividad remesadora y negativos en los días con menores montos, destacándose un efecto claramente decreciente los fines de semana, en particular el día 6, cuyo valor estacional negativo es el más pronunciado.

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,755,844 USD	El modelo incurre en un error absoluto promedio de aproximadamente 1.76 millones de dólares por día durante el periodo de prueba.
RMSE (Raíz del Error Cuadrático Medio)	2,013,179 USD	En promedio, las predicciones se desvían del valor real en aproximadamente 2.01 millones de dólares.
MAPE (Error Porcentual Absoluto Medio)	24.84%	El modelo presenta un error porcentual medio del 24.84%, lo cual indica un desempeño razonable, aunque inferior al alcanzado por modelos más complejos.
R^2 (Coeficiente de determinación)	27.24%	Solo el 27.2% de la variabilidad del monto originado es explicada por el modelo Holt-Winters.

Tabla 11 Aplicación resultados modelo Holt-Winters

El modelo Holt-Winters ajustado sobre la serie original sin transformación logarítmica permitió capturar adecuadamente la estacionalidad semanal y la tendencia subyacente de los datos de remesas. No obstante, al comparar las predicciones con los valores reales durante el periodo de prueba, se observa que el modelo tiende a suavizar los extremos, especialmente los picos pronunciados de ciertos días, lo que se traduce en una menor sensibilidad frente a las variaciones abruptas.

Serie con transformación logarítmica

Para este modelo Holt-Winters, se aplicó una transformación logarítmica natural a la serie temporal con el fin de estabilizar la varianza y mitigar el impacto de valores atípicos. Al transformar los datos a escala logarítmica, es posible modelar la estacionalidad mediante una estructura aditiva y obtener predicciones más robustas.

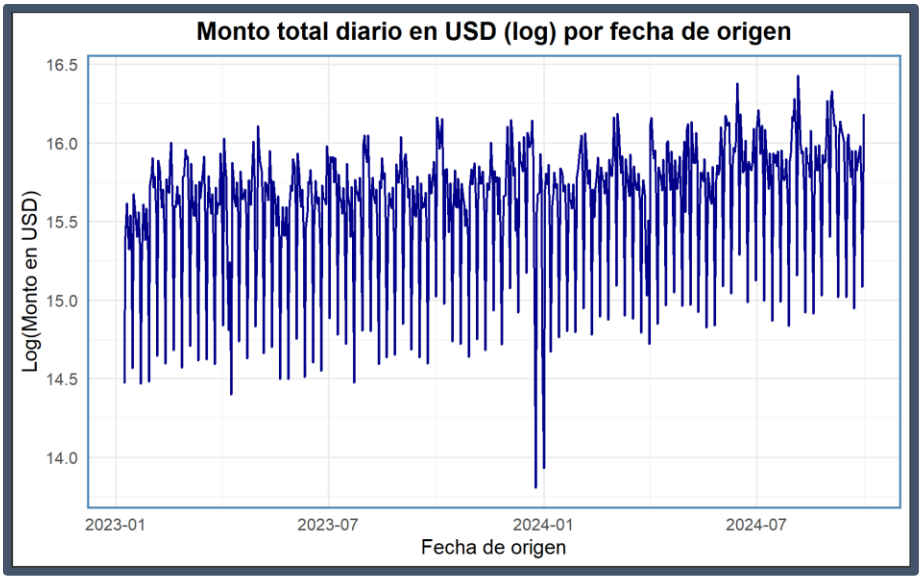

Ilustración 26 Serie transformada

Resultados

Una vez generadas las predicciones en la escala logarítmica, se aplicó el proceso inverso mediante la función exponencial, con el fin de obtener los valores en niveles originales y permitir su comparación directa con los datos reales. Este procedimiento también fue necesario para representar gráficamente la evolución del modelo sobre el periodo de prueba en unidades monetarias comprensibles.

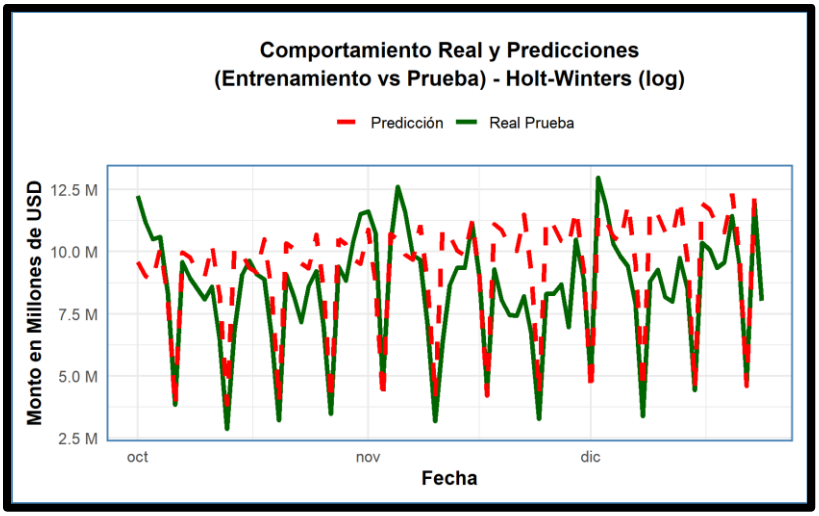

Ilustración 27 Predicciones del Modelo HW transformado

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,437,207 USD	El error promedio absoluto entre las predicciones y los valores reales fue de aproximadamente 1.44 millones de USD.
RMSE (Raíz del Error Cuadrático Medio)	1,749,792 USD	La desviación promedio de las predicciones fue inferior a la del modelo Holt-Winters sin transformación, mejorando su capacidad de ajuste global.
MAPE (Error Porcentual Absoluto Medio)	18.21%	El error porcentual promedio disminuyó significativamente, lo que refleja una mejora notable en términos relativos.
R^2 (Coeficiente de determinación)	45.03%	Solo el 45.03% de la variabilidad del monto originado es explicada por el modelo Holt-Winters.

Tabla 12 Aplicación resultados modelo Holt-Winters - Serie transformada

En cuanto al rendimiento, representado en la tabla 12, el modelo Holt-Winters transformado mostró mejoras significativas frente a la versión anterior. Las métricas fuera de muestra reflejan una reducción en el error absoluto medio (MAE), que pasó de 1.76 millones a 1.44 millones de USD, y una caída en el error porcentual absoluto medio (MAPE) de 24.84% a 18.21%. Estos resultados evidencian que la transformación logarítmica permitió al modelo adaptarse mejor a la dinámica de la serie, especialmente en los rangos de menor volumen, y mejorar la precisión relativa de las predicciones. A pesar de que el modelo mantiene una tendencia a suavizar algunos picos, su comportamiento general fue más ajustado, lo cual se aprecia en la visualización del comportamiento real frente a las predicciones generadas. En conjunto, la aplicación de la transformación logarítmica resulta una estrategia efectiva para optimizar modelos de suavizamiento exponencial en contextos con variabilidad estructural significativa.

6.4 MODELO BOX – JENKIS

El enfoque Box-Jenkins constituye una metodología sistemática para la identificación, estimación, validación y pronóstico de modelos ARIMA y sus extensiones. A diferencia de los enfoques automatizados, la metodología Box-Jenkins se apoya en un análisis detallado de la estructura de la serie temporal, utilizando herramientas como las funciones de autocorrelación (ACF), autocorrelación parcial (PACF), y pruebas de estacionariedad para determinar la forma funcional más adecuada del modelo. Esta técnica permite ajustar modelos personalizados, considerando de manera cuidadosa los órdenes de los componentes autoregresivos (AR), de diferenciación (I) y de media móvil (MA), así como sus posibles componentes estacionales. En esta sección se implementa un modelo siguiendo los lineamientos de Box y Jenkins, con el propósito de validar un enfoque más estructurado y guiado por el análisis realizado en la sección de descripción.

Para este modelo se utilizó únicamente la serie temporal diaria Monto_USD_Originado, sin

incorporación de variables exógenas ni codificaciones. La serie fue utilizada en su forma original, no se aplicaron escalamientos ni transformaciones adicionales. El análisis de ACF y PACF sobre esta serie permitió identificar la estructura de rezagos para la especificación del modelo ARIMA.

Preprocesamiento

Tal como se estableció en el análisis exploratorio (sección 4.9.2), la serie correspondiente al monto diario en dólares originado por remesas fue evaluada mediante la prueba de Dickey-Fuller aumentada, la cual arrojó un estadístico de -9.49 y un valor-p menor a 0.01, permitiendo concluir que la serie es estacionaria en niveles. Este hallazgo resulta clave para la aplicación del enfoque Box-Jenkins, ya que elimina la necesidad de diferenciación adicional y permite trabajar directamente con un modelo ARIMA sin componente integrado (d = 0). A partir de este punto, se procedió con el análisis de autocorrelación y autocorrelación parcial para identificar la estructura de rezagos más adecuada, y construir un modelo que represente correctamente la dinámica temporal de la serie.

Pruebas y entrenamiento del modelo

En la sección 4.9.3 del análisis exploratorio, la serie fue evaluada mediante las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) con el fin de identificar posibles estructuras de rezago. En esta etapa, como parte de la aplicación formal del enfoque Box-Jenkins, se retoman dichos gráficos para orientar la especificación del modelo ARIMA. En la ACF se observa un pico pronunciado en el primer rezago, así como una señal secundaria en el rezago 7, mientras que la PACF presenta valores significativos en los rezagos 1 y 7. Este comportamiento sugiere una combinación de efectos autorregresivos de corto plazo (AR (1)) y potencialmente estacionales (AR (7)), que reflejan la influencia de la dinámica semanal previamente identificada en la serie. En conjunto, estos patrones respaldan el ajuste inicial de un modelo ARIMA (7,0,1) o la consideración de una estructura SARIMA con estacionalidad semanal.

Resultados

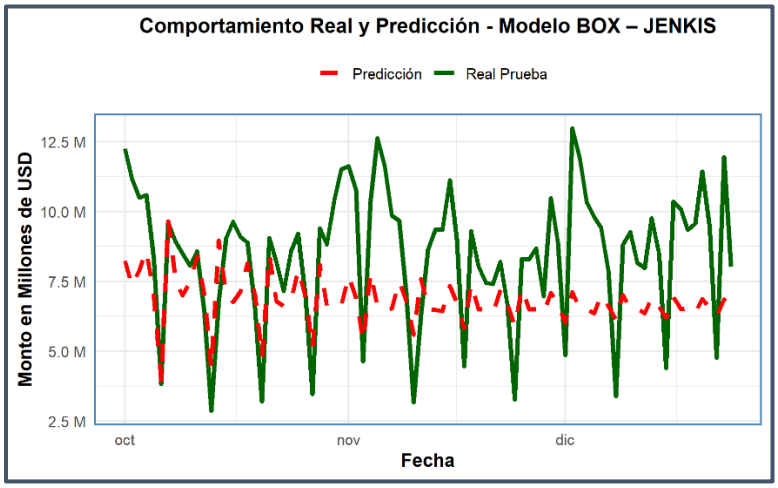

Ilustración 28 Predicciones del Modelo Box - Jenkis

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	936,224 USD	El error promedio diario entre las predicciones del modelo y los valores reales fue inferior al millón de dólares, lo que refleja una buena precisión absoluta.
RMSE (Raíz del Error Cuadrático Medio)	1,260,618 USD	La desviación promedio de las predicciones respecto a los valores reales es de 1.26 millones de dólares.
MAPE (Error Porcentual Absoluto Medio)	16.77%	El error relativo promedio del modelo es del 16.77%, lo que representa una mejora importante respecto a modelos sin estacionalidad explícita.
R^2 (Coeficiente de determinación)	21.74%	Solo el 21.74% de la variabilidad del monto originado es explicada por el modelo.

Tabla 13 Aplicación resultados modelo Box Jenkins

La comparación entre los valores reales y las predicciones del modelo ARIMA (7,0,1) durante el periodo de prueba evidencia que el modelo no logra capturar adecuadamente el patrón general de la serie, manteniendo una trayectoria estable en torno al promedio semanal. Sin embargo, al tratarse de un modelo sin componente estocástico estacional explícito (como en un SARIMA), su capacidad para ajustarse a las fluctuaciones extremas diarias es limitada. Como se observa en la gráfica, las predicciones tienden a suavizar los picos negativos de algunos fines de semana, y no alcanzan los niveles máximos registrados en días con mayor actividad. El modelo muestra un error absoluto medio (MAE) inferior a un millón de dólares y un MAPE del 16.77%.

6.5 MODELO SARIMAX

Con el objetivo de incorporar información exógena en la predicción del monto diario de remesas, se implementó un modelo SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors). Esta extensión del modelo ARIMA permite no solo capturar la dinámica temporal de la serie principal, sino también integrar variables externas que pueden influir en su comportamiento. En este caso, se consideraron como regresores exógenos las variables operativas relacionadas con canales de distribución, países de origen de las remesas, y condiciones de mercado como la tasa de cambio. A diferencia de modelos puramente univariados como ARIMA, el enfoque SARIMAX ofrece una ventaja clave: permite modelar el efecto directo de factores observables sobre la variable objetivo, facilitando la interpretación y el control de eventos específicos en la serie.

En el modelo SARIMAX utilizó como variable principal la serie temporal diaria Monto_USD_Originado, complementada con un conjunto de variables exógenas. Estas incluyeron variables rezagadas de 7 días para montos diarios por canal (APN, Corresponsales, RedPropia), por país (US, MX, etc.), la variable Monto_USD_Pago, así como variables de calendario (dummies de día de la semana, mes, semanas del mes y festivos en Colombia y Estados Unidos).

Preprocesamiento de los datos

Previo a la estimación del modelo SARIMAX, se llevó a cabo el preprocesamiento de los datos con el fin de garantizar el cumplimiento de los supuestos estadísticos necesarios para este tipo de modelos.

En primer lugar, se comprobó la estacionariedad de la serie objetivo (Monto_USD_Originado) mediante la prueba de Dickey-Fuller aumentada, como se describe en la sección 4.9.2 Estacionariedad. El resultado de dicha prueba permitió rechazar la hipótesis nula de raíz unitaria con un valor-p inferior al 5%, lo que indicó que la serie era estacionaria en niveles y, por tanto, no requería diferenciación adicional.

Posteriormente, se conformó el conjunto de variables exógenas que serían incluidas en el modelo, seleccionadas a partir del análisis presentado en la sección 5.1 Identificación de Variables para los Modelos. La segmentación del conjunto de datos en entrenamiento y prueba se realizó según la estructura temporal descrita en la sección 5.5 División del conjunto de datos: entrenamiento y prueba, utilizando como punto de corte el 30 de septiembre de 2024.

Modelo inicial

Para determinar la estructura más adecuada del componente autorregresivo del modelo SARIMAX, se utilizó la función auto_arima del paquete pmdarima, ampliamente reconocida en aplicaciones de series temporales. Este procedimiento permite identificar de manera automática la combinación óptima de los parámetros (p, d, q) del modelo ARIMA y, adicionalmente, los componentes estacionales (P, D, Q, s) cuando se activa la opción seasonal=True. En este caso, se optó por permitir la incorporación de una

estructura estacional explícita, reconociendo la existencia de patrones recurrentes en los montos de remesas a lo largo del tiempo, que podrían no ser capturados únicamente con variables exógenas.

Durante el proceso de búsqueda, se exploraron múltiples combinaciones de parámetros autorregresivos (p), diferenciaciones (d), y medias móviles (q), así como sus contrapartes estacionales (P, D, Q). Para el modelo SARIMAX, estos valores fueron evaluados dentro de los siguientes rangos:

- p, d, q ∈ {0, 1, 2, 3}
- P, D, Q ∈ {0, 1}
- s = 7, correspondiente a la estacionalidad semanal identificada en la serie

La función auto_arima() utilizó como criterio de selección el menor valor del AIC (Akaike Information Criterion), que permite evaluar el equilibrio entre ajuste y complejidad del modelo. La búsqueda se realizó de forma eficiente mediante el argumento stepwise=True, que activa un enfoque heurístico de exploración progresiva del espacio de parámetros, reduciendo así significativamente la carga computacional sin comprometer la calidad del resultado. El modelo obtenido fue luego utilizado como base para la estimación final del SARIMAX con las variables operativas seleccionadas.

Resultados

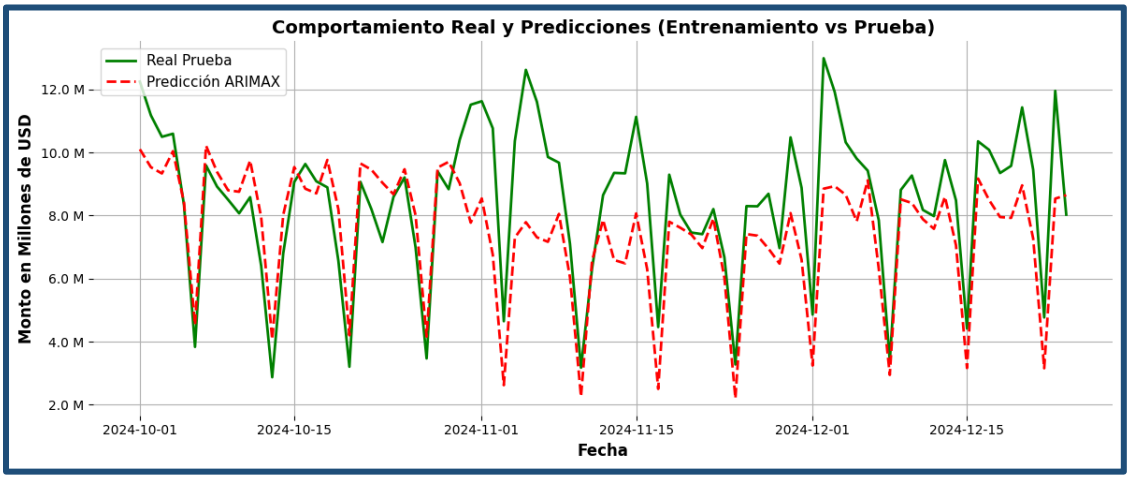

Ilustración 29 Predicciones del Modelo SARIMAX

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,437,243 USD	En promedio, las predicciones difieren del valor real en aproximadamente 1.43 millones de dólares.
RMSE (Raíz del Error Cuadrático Medio)	1,797,344 USD	El error típico de las predicciones se aproxima a 1.79 millones de dólares.
MAPE (Error Porcentual Absoluto Medio)	17.37 %	El error porcentual promedio fue del 17.37 %.
R^2 (Coeficiente de determinación)	42.00 %	El modelo explicó el 42.0 % de la variabilidad observada en los datos de prueba.

Tabla 14 Aplicación resultados modelo SARIMAX

El modelo SARIMAX estimado, que incorpora como regresores exógenos las variables operativas y rezagos asociados al comportamiento transaccional, presentó un desempeño moderado sobre el conjunto de prueba. El error absoluto medio (MAE) alcanzó un valor de 1.43 millones de dólares, mientras que la raíz del error cuadrático medio (RMSE) se situó en 1.79 millones de dólares, lo que indica la magnitud promedio de las desviaciones entre los valores predichos y los observados. El error porcentual absoluto medio (MAPE) fue de 17.37 %, reflejando un margen de error relativo relevante en el contexto operativo. Por su parte, el coeficiente de determinación (R^2) fue de 42.0 %, lo cual indica que el modelo logró explicar menos de la mitad de la variabilidad presente en los montos diarios originados.

Estos resultados muestran que, si bien el modelo logra capturar parcialmente la dinámica de la serie, existen componentes no explicados que podrían deberse a patrones más complejos, no lineales, o a efectos estacionales no capturados por completo en la estructura actual.

Variables significativas

Una vez ajustado el modelo SARIMAX con la totalidad de las variables exógenas disponibles, se procedió a analizar los coeficientes estimados con el fin de identificar cuáles variables presentan un efecto significativo sobre la variable objetivo. Para ello, se evaluaron tanto la magnitud de los coeficientes como sus valores p y los intervalos de confianza al 95 %.

Del conjunto total de variables exógenas incluidas, se identificaron como estadísticamente significativas (p-valor < 0.05) aquellas asociadas a días especiales o estacionales, así como ciertas variables estructurales de calendario. En particular, las variables Es_Fin_de_Semana, Feriado_USA y Feriados_Colombia mostraron coeficientes negativos y altamente significativos, lo que sugiere una disminución sistemática en los montos originados durante esos días. Igualmente, varias dummies de meses (como Mes_ene, Mes_feb, Mes_dic, Mes_oct, Mes_sep) y de días de la semana (Dia_Semana_Num_1, Dia_Semana_Num_2, Dia_Semana_Num_3, Dia_Semana_Num_4, Dia_Semana_Num_6, Dia_Semana_Num_7) presentaron efectos significativos, indicando patrones recurrentes de comportamiento temporal en la serie.

Por el contrario, la mayoría de los coeficientes asociados a los canales de distribución (APN, RedPropia, Corresponsales) y a los países de origen (AU, BR, CA, EC, etc.) no resultaron estadísticamente significativos, con valores p cercanos a 1 y amplios intervalos de confianza que incluyen el cero. Esto sugiere que, dentro del enfoque lineal del modelo SARIMAX, estas variables no presentan un aporte relevante en la explicación de la variabilidad de los montos diarios de remesas. Una excepción parcial fue el rezago de un día del monto pagado (Monto_USD_Pago_lag1), cuyo coeficiente fue negativo y significativo ($p \approx 0.02$), indicando un posible efecto de retroalimentación operativa de corto plazo.
En conjunto, los resultados permiten concluir que las variables temporales y estructurales del calendario capturan una porción significativa de la dinámica de la serie, mientras que los componentes operativos específicos tienen una menor capacidad explicativa dentro del marco de este modelo lineal.

Modelo con variables significativas

Con base en los resultados del análisis de significancia estadística de los coeficientes, se construyó un segundo modelo SARIMAX utilizando únicamente las variables exógenas que mostraron un efecto significativo sobre la variable objetivo. Esta selección incluyó variables asociadas al calendario, como dummies de días de la semana, meses y feriados. El objetivo fue reducir la complejidad del modelo manteniendo un nivel aceptable de desempeño predictivo y mejorar la interpretabilidad de los resultados.

Resultados

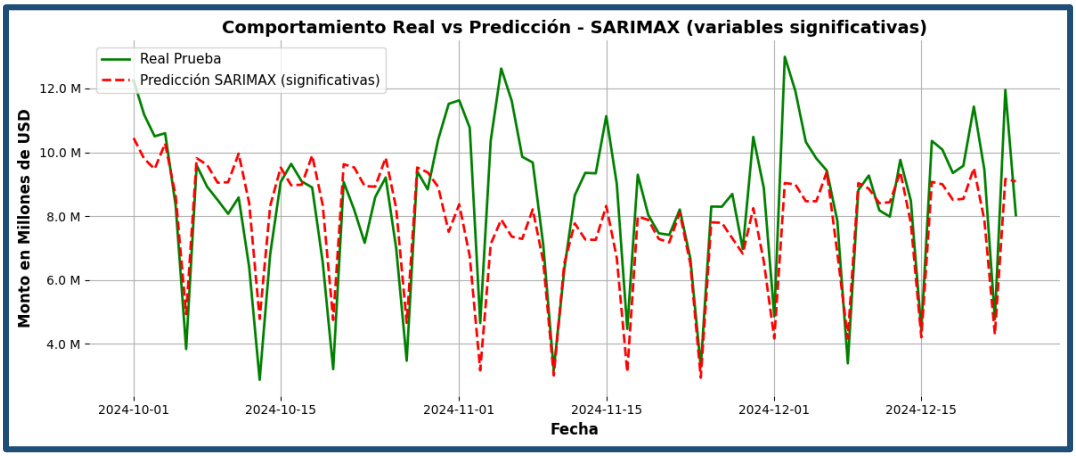

Ilustración 30 Predicciones del Modelo SARIMAX con variables significativas

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,272,700 USD	En promedio, las predicciones difieren del valor real en aproximadamente 1.27 millones de dólares.
RMSE (Raíz del Error Cuadrático Medio)	1,669,090 USD	El error típico de las predicciones fue de 1.66 millones de dólares.
MAPE (Error Porcentual Absoluto Medio)	15.33 %	El error relativo promedio fue del 15.33 %.
R^2 (Coeficiente de determinación)	49.98 %	El modelo explicó el 49.98% de la variabilidad observada en los datos de prueba.

Tabla 15 Aplicación resultados modelo SARIMAX - Variables significativas

El modelo SARIMAX ajustado con un subconjunto reducido de variables exógenas estadísticamente significativas presentó un desempeño intermedio en comparación con el modelo completo. El error absoluto medio (MAE) fue de 1.27 millones de dólares, con una raíz del error cuadrático medio (RMSE) de 1.66 millones y un error porcentual absoluto medio (MAPE) de 15.33%. El coeficiente de determinación (R^2) alcanzó un 49.98%, lo que indica que el modelo fue capaz de explicar cerca del 50% de la variabilidad de los datos de prueba.

6.6 MODELOS DE REDES NEURONALES

Las redes neuronales artificiales constituyen una familia de modelos computacionales inspirados en la estructura y funcionamiento del cerebro humano. Estas redes están conformadas por capas de nodos interconectados que simulan neuronas artificiales y que, a través de un proceso iterativo de aprendizaje, ajustan sus pesos internos para minimizar el error en tareas de predicción o clasificación. A diferencia de los modelos estadísticos tradicionales, las redes neuronales no requieren especificar una forma funcional entre las variables independientes y la variable dependiente, lo que les permite capturar relaciones complejas y no lineales en los datos. Esta flexibilidad ha hecho que su aplicación sea cada vez más común en entornos donde los patrones subyacentes no pueden ser representados adecuadamente mediante técnicas lineales, como ocurre en muchos problemas asociados a la predicción de series de tiempo financieras, económicas u operativas.

Dataset utilizado

Para los modelos de redes neuronales se utilizó como variable objetivo Monto_USD_Originado, y se aplicaron estructuras de datos en formato supervisado secuencial, con base en rezagos diarios.

En los modelos **Elman** y **Jordan**, el dataset fue transformado para generar secuencias con 7 u 8 rezagos consecutivos de la serie Monto_USD_Originado como entradas, y el valor del día siguiente como salida. No se incluyeron variables exógenas. La serie fue normalizada utilizando la técnica *Min-Max Scaling*, ajustándola al rango [0, 1], dado que las redes neuronales son sensibles a la escala de los datos.

Para el modelo **LSTM**, se incluyeron además variables exógenas booleanas y numéricas, y se organizaron en ventanas móviles de 5 días. Todos los datos numéricos fueron escalados con *Min-Max Scaling* antes del entrenamiento.

La división entre entrenamiento y prueba respetó la estructura cronológica del conjunto original: entrenamiento hasta el 30 de septiembre de 2024, prueba desde el 1 de octubre al 24 de diciembre de 2024.

6.6.1 ELMAN

Entre las distintas arquitecturas disponibles para modelar secuencias temporales, las redes neuronales recurrentes (RNN) han demostrado ser especialmente útiles debido a su capacidad para incorporar dependencias dinámicas a lo largo del tiempo. Dentro de esta categoría, la red neuronal tipo Elman constituye una variante clásica que introduce una capa de contexto capaz de almacenar temporalmente el estado anterior de la capa oculta, generando así una forma de "memoria corta" en el modelo. Esta retroalimentación interna le permite a la red Elman aprender patrones secuenciales que dependen no solo de la entrada actual, sino también del historial reciente, lo cual es fundamental en series con estructura temporal.

En el contexto de este proyecto, dicha arquitectura resulta apropiada para capturar la dinámica del monto diario originado por remesas, pues permite modelar tanto la variabilidad diaria como los ciclos y dependencias implícitas en la serie, sin necesidad de asumir linealidad ni estacionariedad.

Preprocesamiento

Para entrenar la red neuronal recurrente tipo Elman, fue necesario transformar la serie original en un formato compatible con aprendizaje supervisado. A partir de la variable objetivo: Monto_USD_Originado, se generaron variables de entrada a partir de rezagos de esta misma serie, específicamente desde el día 7 hasta el día 14, con el fin de capturar patrones temporales.

Dado que las redes neuronales son altamente sensibles a la escala de los datos, tanto la variable objetivo como las variables explicativas fueron normalizadas utilizando la técnica min-max scaling, transformando sus valores a un rango [0, 1].
Finalmente, la base de datos fue reestructurada en forma de secuencias de longitud fija, donde cada observación incluye múltiples días consecutivos de información histórica (inputs) y un valor futuro de Monto_USD_Originado como salida esperada, permitiendo así capturar dependencias dinámicas en el tiempo durante el entrenamiento del modelo Elman.

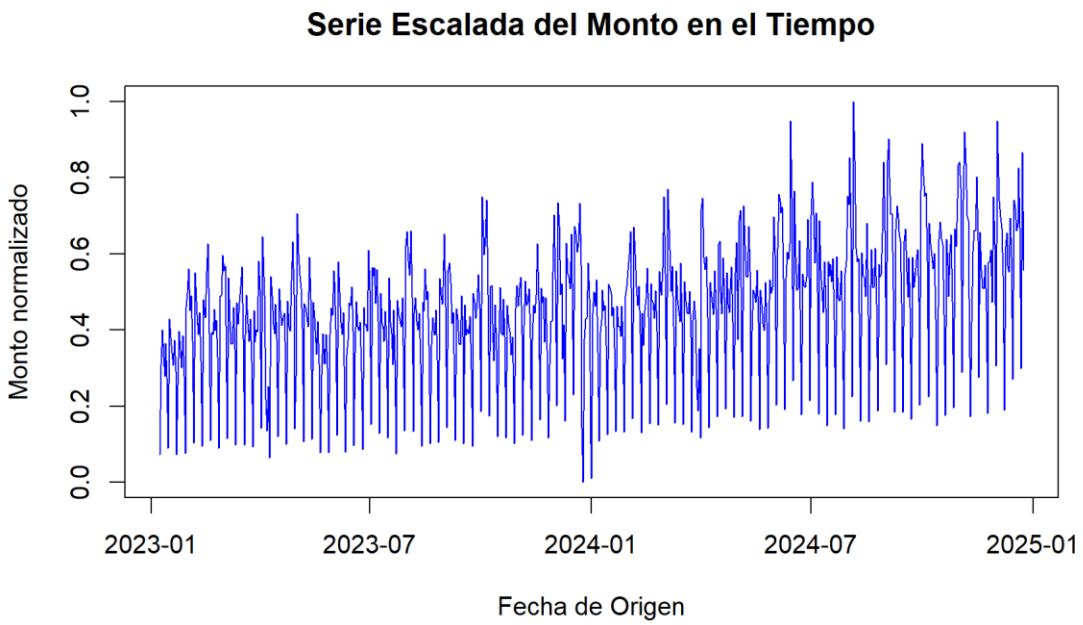

Ilustración 31 Transformación de la variable objetivo

Posteriormente, se reestructuraron los datos en formato secuencial para alimentar la red recurrente. Se definieron ventanas temporales deslizantes en las que cada instancia del modelo recibe como entrada una secuencia de observaciones correspondientes a rezagos diarios consecutivos del monto originado por remesas, específicamente desde el día 7 hasta el día 14 anterior a la fecha objetivo. Esta representación permite a la red capturar la dinámica interna de la serie sin requerir información exógena adicional.

Para garantizar una evaluación rigurosa y libre de sesgos por fuga de información, se implementó una división temporal estricta del conjunto de datos. Los registros anteriores al 1 de septiembre de 2024 fueron utilizados exclusivamente para el entrenamiento, mientras que aquellos desde esa fecha en adelante conformaron el conjunto de prueba. Esta estrategia respeta el principio de no anticipación y simula un escenario real de predicción futura, permitiendo así medir con mayor fidelidad la capacidad de generalización del modelo.

Modelo inicial

Resultados: En la ilustración 32 y en la tabla 16 se describen los resultados pertinentes al modelo Elman.

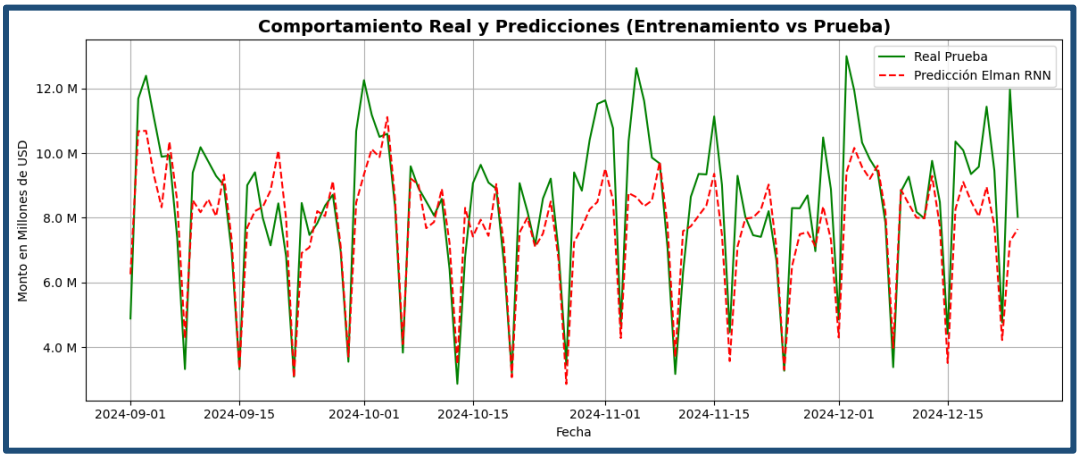

Ilustración 32 Predicción de la red neuronal recurrente Elman

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,059,740 USD	En promedio, las predicciones difieren del valor real en aproximadamente 1,05 millones de dólares por día.
RMSE (Raíz del Error Cuadrático Medio)	1,377,072 USD	El error típico del modelo es de 1,3 millones de dólares.
MAPE (Error Porcentual Absoluto Medio)	12,12%	El modelo comete, en promedio, un error de 12% respecto al valor real, indicando alta precisión.
R^2 (Coeficiente de determinación)	66.56%	El 66.56% de la variabilidad observada en el monto diario es explicada por el modelo.

Tabla 16 Aplicación resultados modelo red neuronal recurrente Elman

Este modelo de red neuronal recurrente tipo Elman fue implementado utilizando una arquitectura sencilla en PyTorch, compuesta por una única capa oculta con 32 neuronas y función de activación tangencial (tanh). La salida del modelo fue definida como una capa lineal. Se empleó una tasa de aprendizaje de 0.001 y se entrenó durante 50 épocas con el optimizador Adam, priorizando la estabilidad en la convergencia.

Una vez entrenado, el modelo fue evaluado sobre el conjunto de prueba correspondiente a los registros posteriores al 1 de septiembre de 2024. El desempeño obtenido fue moderado, con un error absoluto medio (MAE) de 1,059,740 USD, un RMSE de 1,377,072 USD, y un MAPE de 12.12%. El coeficiente de determinación R^2 fue de 66.56%, indicando que el modelo logra capturar parte importante de la varianza del monto originado por remesas, aunque aún existen oportunidades de mejora en precisión.

Tal como se observa en la figura, la serie predicha reproduce la forma general del comportamiento real, reflejando ciclos y tendencias, aunque con cierta subestimación en los picos de alto volumen. Esto sugiere que, si bien la red Elman ofrece una base sólida para modelar la secuencia, será necesario explorar ajustes de hiperparámetros y arquitecturas más profundas para optimizar su capacidad predictiva.

Optimización del modelo

La red neuronal Elman requiere la optimización de tres hiperparámetros principales: el tamaño de la capa oculta, la tasa de aprendizaje y el número máximo de iteraciones. Estos parámetros influyen directamente en la capacidad del modelo para aprender patrones secuenciales y generalizar sobre datos no vistos.

- Tamaño de la capa oculta

Con el fin de determinar el valor óptimo para el hiperparámetro hidden_size, correspondiente al número de neuronas en la capa oculta de la red Elman, se realizó un proceso sistemático de evaluación sobre siete configuraciones: 8, 16, 32, 64, 96, 128 y 160 neuronas. Para cada valor, se entrenó el modelo con una misma tasa de aprendizaje (0.001) y 50 épocas, repitiendo el entrenamiento tres veces con diferentes inicializaciones aleatorias. El desempeño se evaluó sobre el conjunto de prueba, utilizando las métricas RMSE, MAE, MAPE y R^2.

De acuerdo con los resultados, el valor de hidden_size = 96 ofreció el mejor desempeño general, al obtener el menor RMSE (1,279,101 USD) y el mayor coeficiente de determinación R^2 (71.14%), lo que indica una mayor capacidad para explicar la varianza del monto diario de remesas. Aunque la configuración con 32 neuronas presentó un MAPE ligeramente menor (11.32%), esta diferencia es marginal y no compensa la pérdida relativa en RMSE y R^2.

Por tanto, se seleccionó hidden_size = 96 como el valor óptimo para el modelo final, al representar el mejor compromiso entre precisión, capacidad explicativa y generalización.

- Tasa de aprendizaje

Para encontrar la tasa de aprendizaje más adecuada (learning_rate) para el modelo Elman, se evaluaron siete valores dentro de una escala logarítmica, variando desde 0.1 hasta 0.0001. En cada configuración, se mantuvieron constantes el número de neuronas ocultas (hidden_size = 96), el número de épocas (50) y el resto de los hiperparámetros. Cada experimento fue repetido tres veces y evaluado con las métricas RMSE, MAE, MAPE y R^2 sobre el conjunto de prueba.

Los resultados muestran que la tasa de aprendizaje de 0.001 ofrece el mejor equilibrio entre velocidad de convergencia y precisión del modelo, permitiendo optimizar el desempeño sin comprometer la estabilidad del entrenamiento. Por esta razón, se adoptó este valor como tasa de aprendizaje óptima para la red Elman en la fase final del modelado.

- Número máximo de iteraciones

La cantidad de iteraciones (o épocas) durante el entrenamiento de redes neuronales influye directamente en la calidad del modelo: pocas épocas pueden resultar en subajuste, mientras que un número excesivo puede causar sobreajuste. Para encontrar el valor óptimo, se evaluaron seis configuraciones (20, 40, 60, 80, 100 y 150 épocas), manteniendo fijos los hiperparámetros previamente seleccionados (hidden_size = 96, learning_rate = 0.001). Cada modelo fue entrenado tres veces y evaluado sobre el conjunto de prueba.

Los resultados indican que el mejor desempeño del modelo se alcanza con 40 épocas, logrando un equilibrio entre precisión y generalización. Este valor fue adoptado como configuración final para el modelo Elman.

Resultados

Con base en el proceso de optimización de hiperparámetros, se configuró la red Elman con 96 neuronas ocultas, una tasa de aprendizaje de 0.001 y 40 iteraciones. El modelo final fue entrenado sobre los datos previos al 1 de septiembre de 2024 y evaluado sobre el conjunto de prueba posterior a esa fecha. Los resultados muestran un MAE de USD 968 mil, un RMSE de 1.3 millones de USD, y un MAPE del 11.53%, lo que indica un nivel adecuado de precisión para fines de predicción operativa. Además, el modelo explicó el 70% de la varianza del monto de remesas diario ($R^2 = 0.7004$), lo que respalda su capacidad de generalización y robustez.

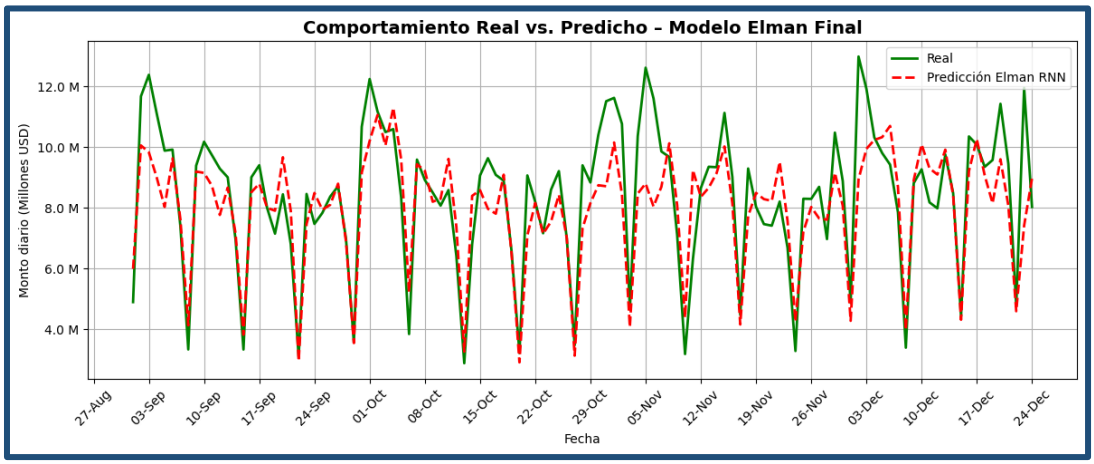

Ilustración 33 Predicción de la red neuronal recurrente Elman optimizada

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	968,270 USD	En promedio, las predicciones difieren del valor real en aproximadamente 968 mil dólares por día.
RMSE (Raíz del Error Cuadrático Medio)	1,303,275 USD	El error típico del modelo es de 1,3 millones de dólares.
MAPE (Error Porcentual Absoluto Medio)	11,53%	El modelo comete, en promedio, un error de 11.5% respecto al valor real, indicando alta precisión.
R^2 (Coeficiente de determinación)	70.04%	El 70.04% de la variabilidad observada en el monto diario es explicada por el modelo.

Tabla 17 Aplicación resultados modelo red neuronal recurrente Elman optimizada

6.6.2 JORDAN

Como parte del enfoque basado en redes neuronales recurrentes, se implementó un modelo Jordan, una variante clásica dentro de esta familia que se distingue por incorporar una capa de contexto alimentada desde la salida del modelo. Esta arquitectura le permite capturar dependencias temporales de corto plazo, lo cual resulta particularmente relevante en la predicción de series como los montos diarios de remesas, donde el comportamiento del día actual puede estar influenciado por los valores recientes. A diferencia del modelo Elman, cuya capa de contexto se retroalimenta desde la capa oculta, el modelo Jordan basa su memoria en la salida pasada, lo que le permite reforzar su capacidad de ajuste ante patrones de tipo operativos o de decisión acumulada. Esta característica lo convierte en una alternativa adecuada para evaluar el impacto de secuencias recientes de valores monetarios en la predicción del monto originado en un día específico. A continuación, se presenta el proceso de entrenamiento, optimización y evaluación del modelo Jordan, junto con sus resultados en el conjunto de prueba.

Preprocesamiento

Para el entrenamiento del modelo Jordan se utilizó el mismo conjunto de datos preparado en la sección del modelo Elman, a fin de garantizar la consistencia metodológica y facilitar la comparación de resultados entre arquitecturas. Como se describió anteriormente, la variable objetivo corresponde al monto total diario de remesas originadas, el cual fue normalizado mediante la técnica min-max para adecuarlo al rango [0,1], requerido por las redes neuronales. A partir de esta serie normalizada, se construyó una estructura supervisada utilizando ocho rezagos consecutivos como variables de entrada (lag1 a lag8) y el valor del día siguiente como variable objetivo (Target). Esta representación permite que la red capture patrones secuenciales y cíclicos observados en la dinámica de las remesas, como los efectos de arrastre semanal. El conjunto resultante fue posteriormente dividido en entrenamiento y prueba, utilizando como punto de corte la fecha del 30 de septiembre de 2024, en coherencia con la estrategia temporal definida para todos los modelos.

Resultados: En la ilustración 34 y en la tabla 18 se describen los resultados pertinentes al modelo Jordan.

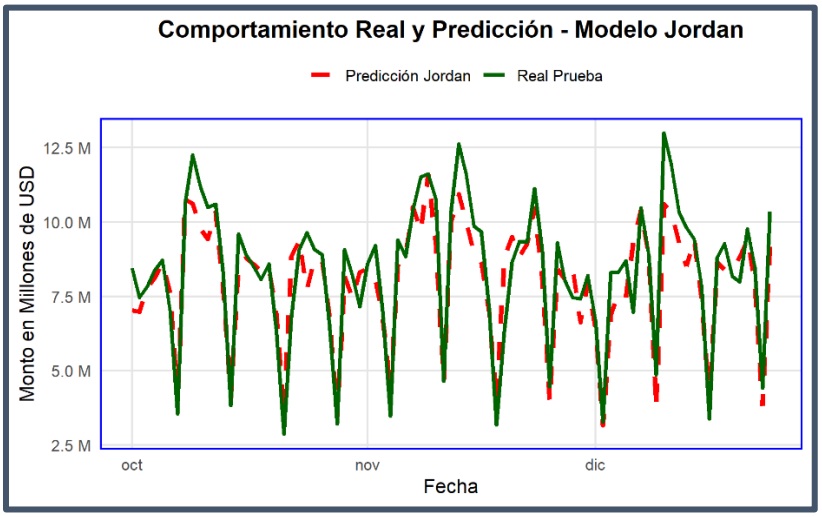

Ilustración 34 Predicción de la red neuronal recurrente Jordan

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	718,070 USD	En promedio, las predicciones se desvían 718 mil dólares del valor real.
RMSE (Raíz del Error Cuadrático Medio)	943,606 USD	El error típico de las predicciones es cercano a 944 mil dólares.
MAPE (Error Porcentual Absoluto Medio)	9.1 %	Las predicciones tienen un error porcentual promedio menor al 10%.
R^2 (Coeficiente de determinación)	83.69 %	El modelo explica el 83.7% de la variabilidad observada en los montos diarios.

Tabla 18 Aplicación resultados modelo red neuronal recurrente Jordan

El modelo Jordan inicial fue configurado con una única capa oculta de cinco neuronas (size = 5), una tasa de aprendizaje de 0.1 y un máximo de 1000 iteraciones. Esta configuración fue seleccionada como punto de partida por su simplicidad y por ser consistente con el modelo Elman base, permitiendo una comparación directa del desempeño entre ambas arquitecturas.

Tras el entrenamiento y evaluación sobre el conjunto de prueba, el modelo arrojó resultados alentadores: un error absoluto medio (MAE) de 718.070 USD, un RMSE de 943.606 USD y un error porcentual medio (MAPE) de 9.1%. El coeficiente de determinación (R^2) alcanzó un 83.69%, lo que indica que el modelo logró explicar más del 83% de la variabilidad observada en los montos diarios de remesas. Visualmente, el modelo capturó de forma adecuada la tendencia general y los ciclos semanales de la serie, aunque presentó ligeras desviaciones frente a picos extremos. Estos resultados sugirieron que la

arquitectura Jordan era viable para el problema planteado, y motivaron su posterior proceso de optimización.

Optimización del modelo

Con el objetivo de mejorar el desempeño del modelo Jordan, se realizó un proceso sistemático de optimización de hiperparámetros en tres fases: tamaño de la capa oculta, tasa de aprendizaje y análisis de convergencia. Esta estrategia permitió identificar la configuración que maximizaba la capacidad predictiva del modelo sin incurrir en sobreajuste, asegurando su robustez en escenarios de prueba.

- Tamaño de la capa oculta

En una primera etapa, se evaluaron distintas configuraciones del hiperparámetro size, que define la cantidad de neuronas en la capa oculta de la red. Se probaron valores de 3, 5, 10, 15 y 20, manteniendo constante la tasa de aprendizaje (learnFuncParams = 0.1) y el número máximo de iteraciones (maxit = 1000). Para cada configuración se entrenó un modelo independiente y se calculó el error cuadrático medio (MSE) sobre el conjunto de prueba. Los resultados indicaron que el modelo con 10 neuronas ocultas obtuvo el menor MSE (762.9 millones), mostrando un mejor equilibrio entre complejidad y capacidad de generalización.

- Tasa de aprendizaje

En la segunda fase, se exploraron distintos valores de tasa de aprendizaje (learning rate), utilizando el tamaño de capa oculta óptimo encontrado en la fase anterior (size = 10). Se evaluaron los valores 0.01, 0.05, 0.1, 0.2 y 0.5. El modelo entrenado con una tasa de aprendizaje de 0.5 alcanzó el menor MSE (695.9 millones), lo cual sugiere que una mayor velocidad de actualización permitió al modelo ajustar de manera más eficiente sus pesos, sin comprometer la estabilidad del proceso de entrenamiento.

- Número máximo de iteraciones

Finalmente, se examinó la evolución del error durante el proceso iterativo utilizando la función plotIterativeError(), lo que permitió verificar que el modelo convergía de forma estable sin evidencia de oscilaciones ni divergencias. La reducción progresiva del error ponderado durante las primeras iteraciones y su posterior estabilización confirmaron que los hiperparámetros seleccionados (10 neuronas ocultas y tasa de aprendizaje 0.5) eran adecuados para la arquitectura Jordan en este contexto.

Resultados

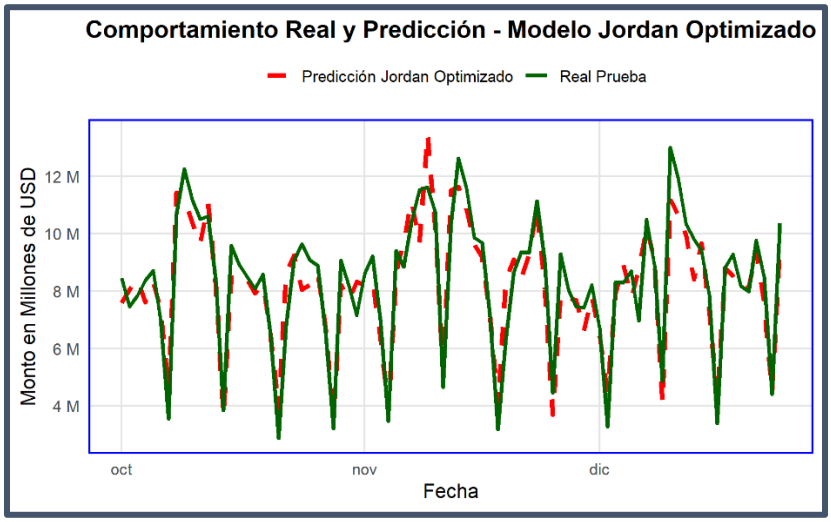

Ilustración 35 Predicción de la red neuronal recurrente Jordan optimizado

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	679,373 USD	En promedio, las predicciones difieren del valor real en 679 mil dólares.
RMSE (Raíz del Error Cuadrático Medio)	831,933 USD	El error típico en las predicciones fue de aproximadamente 832 mil dólares.
MAPE (Error Porcentual Absoluto Medio)	9.09 %	El error promedio relativo fue cercano al 9%.
R^2 (Coeficiente de determinación)	87.32 %	El modelo explicó el 87.32% de la variabilidad observada en el conjunto de prueba.

Tabla 19 Aplicación resultados modelo red neuronal recurrente Jordan optimizado

Una vez completado el proceso de optimización de hiperparámetros, el modelo Jordan fue entrenado con su configuración final: una capa oculta con 10 neuronas y una tasa de aprendizaje de 0.5. Esta arquitectura permitió mejorar de forma notable el desempeño predictivo frente al modelo inicial. En la evaluación sobre el conjunto de prueba, el modelo alcanzó un error absoluto medio (MAE) de 679.373 USD, una raíz del error cuadrático medio (RMSE) de 831.933 USD y un error porcentual absoluto medio (MAPE) de 9.09%. Asimismo, el coeficiente de determinación (R^2) se elevó hasta 87.32%, lo que indica que el modelo explica más del 87% de la variabilidad observada en los montos diarios de remesas originadas. Estos resultados consolidan a la red Jordan optimizada como una alternativa eficiente y robusta dentro del conjunto de modelos analizados, capaz de capturar con precisión los patrones secuenciales del comportamiento transaccional.

6.6.3 MODELO LSTM

Como parte del enfoque basado en redes neuronales, se implementó un modelo LSTM (Long Short-Term Memory), una arquitectura especializada en el procesamiento de secuencias temporales que permite capturar patrones de dependencia a corto y largo plazo en los datos. Esta técnica resulta especialmente adecuada para problemas de series de tiempo, como la predicción de montos diarios de remesas, donde las observaciones anteriores pueden influir de forma acumulativa sobre el valor futuro. A diferencia de los modelos estadísticos tradicionales, el LSTM no requiere supuestos de estacionariedad y tiene la capacidad de modelar relaciones no lineales complejas entre las variables. A continuación, se describe el proceso de preparación de los datos, la estructura del modelo inicial y los resultados obtenidos.

Preprocesamiento

Para la implementación del modelo LSTM se realizó un proceso de preprocesamiento específico orientado a adaptar los datos al formato requerido por las redes neuronales recurrentes. En primer lugar, se dividió el conjunto de datos en entrenamiento y prueba, utilizando como punto de corte el 30 de septiembre de 2024 con el fin de conservar la estructura cronológica y evitar cualquier fuga de información. A continuación, se seleccionaron como variables explicativas todas las columnas numéricas y booleanas distintas a la variable objetivo (Monto_USD_Originado) y a la columna de fecha (Fecha_Origen).

Tanto la variable objetivo como las variables explicativas fueron transformadas mediante la técnica de escalamiento min-max, ajustando sus valores al rango [0, 1]. Este paso fue fundamental para facilitar la convergencia del modelo y garantizar un entrenamiento numéricamente estable. Una vez escalados los datos, se construyeron secuencias temporales utilizando una ventana deslizante de tamaño cinco (time_steps = 5). Este procedimiento generó matrices tridimensionales donde cada entrada del modelo contenía cinco días consecutivos de información histórica como predictores, y el valor correspondiente al sexto día como variable objetivo. El resultado de este proceso fue la conformación de los tensores X_train, X_test, y_train y y_test, que alimentaron la arquitectura de red neuronal durante el entrenamiento y la evaluación.

Modelo inicial

El modelo LSTM base fue construido utilizando todas las variables exógenas disponibles en el conjunto de datos, con excepción de la variable objetivo (Monto_USD_Originado) y la fecha (Fecha_Origen).

Resultados

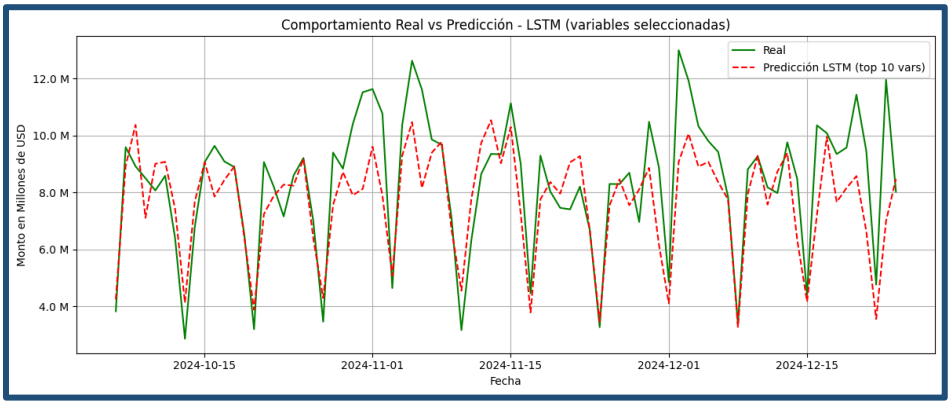

Ilustración 36 Predicción del modelo LSTM

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,478,192 USD	En promedio, las predicciones se desviaron del valor real en aproximadamente 1.47 millones de dólares.
RMSE (Raíz del Error Cuadrático Medio)	1,935,896 USD	El error típico de las predicciones fue de aproximadamente 1.93 millones de dólares.
MAPE (Error Porcentual Absoluto Medio)	16.86%	El error relativo promedio fue del 16.86 %.
R^2 (Coeficiente de determinación)	32.14%	El modelo explicó el 32.14% de la variabilidad observada en el conjunto de prueba.

Tabla 20 Aplicación resultados modelo LSTM

Entre estas se incluyeron tanto variables numéricas como booleanas, incluyendo los rezagos de indicadores operativos, así como codificaciones temporales como dummies de meses, semanas y días de la semana. Estas variables fueron normalizadas y estructuradas en secuencias temporales de cinco días (time_steps = 5), permitiendo al modelo capturar la dinámica a corto plazo en la serie de remesas.

El modelo fue configurado con una capa LSTM de 64 unidades, seguida de una capa Dropout del 20% para mitigar el riesgo de sobreajuste, y una capa densa final con salida lineal. Tras ser entrenado sobre el conjunto de entrenamiento y evaluado con datos posteriores al 30 de septiembre de 2024, el modelo alcanzó un MAE de 1.47 millones de dólares, un RMSE de 1.93 millones, un MAPE de 16.86 % y un coeficiente de determinación (R^2) de 32.12 %. Estos resultados indican que el LSTM no fue capaz de capturar parte relevante de la estructura temporal y estacional de la serie.

Modelo optimizado

Con el objetivo de mejorar el desempeño del modelo LSTM base, se llevó a cabo un proceso de

optimización de hiperparámetros utilizando la librería keras_tuner. Para ello, se implementó una búsqueda aleatoria (RandomSearch) sobre distintos valores de configuración, incluyendo el número de unidades LSTM, la tasa de dropout y la tasa de aprendizaje. Se mantuvo como estructura base una sola capa recurrente seguida de una capa densa de salida, considerando un problema de regresión.

Cada combinación generada por el tuner fue entrenada sobre el conjunto de entrenamiento y validada utilizando un 10% del conjunto como subconjunto de validación interna. La mejor configuración identificada correspondió a una red con 96 unidades LSTM, una tasa de aprendizaje de 0.001 y una tasa de dropout del 0.2. Esta arquitectura fue reentrenada desde cero y evaluada sobre el conjunto de prueba real, garantizando una estimación objetiva del rendimiento fuera de muestra.

Resultados

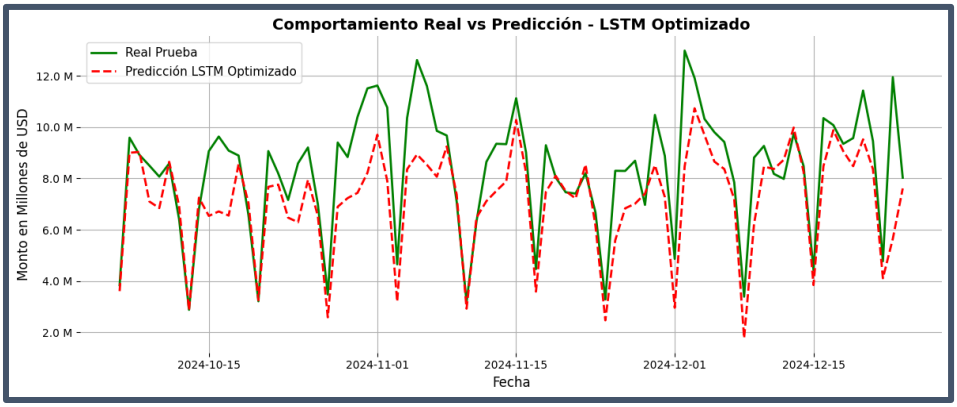

Ilustración 37 Predicción del modelo LSTM optimizado

Métrica	Valor	Interpretación
MAE (Error Absoluto Medio)	1,158,205 USD	Error medio absoluto de las predicciones respecto al valor real.
RMSE (Raíz del Error Cuadrático Medio)	1,532,329 USD	Error típico de las predicciones.
MAPE (Error Porcentual Absoluto Medio)	13.81 %	Error relativo promedio en porcentaje.
R^2 (Coeficiente de determinación)	57.48 %	El modelo explicó el 57.48% de la variabilidad en los datos de prueba.

Tabla 21 Aplicación resultados modelo LSTM optimizado

La optimización del modelo LSTM condujo a una mejora significativa respecto al modelo base. El modelo optimizado alcanzó un MAE de 1.16 millones de dólares, un RMSE de 1.53 millones, un MAPE de 13.81% y un R^2 de 57.48%, mejorando considerablemente en todas las métricas respecto a la versión inicial, que presentó un MAE de 1.48 millones, un RMSE de 1.94 millones, un MAPE de 16.86% y un R^2 de apenas 32.14%.

Estos resultados demuestran que la búsqueda de una configuración más adecuada para la arquitectura

de la red (en términos de número de capas, neuronas y tasa de aprendizaje) permitió al modelo capturar con mayor precisión las dinámicas de la serie temporal. Sin embargo, si bien se obtuvieron mejoras cuantitativas claras, el desempeño general sigue limitado.

6.7 MODELO DE REGRESIÓN CON SVM

En modelos de regresión como Support Vector Machine (SVM), es fundamental identificar las características que ayudan a predecir la variable de interés de manera eficiente. SVM utiliza una función de optimización para encontrar el hiperplano que mejor separa los datos en el espacio de características.

Se eligió el modelo de Support Vector Regression (SVR) debido a su capacidad para manejar datos no lineales. Este modelo se adapta muy bien cuando las relaciones entre las variables no son triviales o lineales.

Las características o variables que se utilizan en el modelo juegan un papel crucial en la determinación de la capacidad del modelo para generalizar en nuevos datos.

Los modelos de regresión con SVM utilizaron como variable objetivo Monto_USD_Originado, acompañado de un conjunto de variables explicativas compuestas por:

- Variables rezagadas (7 días) de montos diarios por canal (APN, Corresponsales, RedPropia) y por país (US, MX, ES, entre otros),
- La TRM (Valor_COP_TRM_lag7),
- Variables temporales codificadas mediante One-Hot Encoding (mes, día de la semana, semana del mes y variables booleanas de festivos).

Antes del entrenamiento, todas las variables numéricas fueron escaladas mediante StandardScaler, técnica necesaria para el correcto funcionamiento del algoritmo SVM.

La división del conjunto de datos respetó el criterio temporal: los registros hasta el 30 de septiembre de 2024 se usaron para entrenamiento, y los posteriores hasta el 24 de diciembre de 2024 para validación.

6.7.1 Pruebas y entrenamiento del modelo

El modelo SVM es una técnica de aprendizaje supervisado que busca encontrar un hiperplano en un espacio de alta dimensión que separe de manera óptima las clases o prediga una variable continua. Para problemas de regresión, Support Vector Regression (SVR) es el modelo utilizado, que adapta la idea de SVM para predicción continua.

A continuación, se describe el proceso realizado.

Preprocesamiento de los datos.

- Normalización de los datos: Se utilizó StandardScaler para escalar las características, dado que SVM es sensible a la escala de los datos.

Entrenamiento y validación

- El modelo fue entrenado utilizando el conjunto de datos de entrenamiento y se evaluó sobre el conjunto de prueba.
- Se utilizó GridSearchCV para encontrar los mejores hiperparámetros, como C, épsilon, gamma, y el kernel.

  La búsqueda de los mejores hiperparámetros para el modelo SVM fueron determinados a partir de un proceso de optimización utilizando una búsqueda en cuadrícula (Grid Search), donde se exploraron una variedad de configuraciones posibles para los hiperparámetros más relevantes del modelo.

A continuación, en la tabla 22, se detallan los valores probados para cada hiperparámetro.

Hiperparámetros	Rango
C	100, 1000, 5000, 10000, 50000, 100000
épsilon	0.01, 0.05, 0.1, 0.2, 0.5
gamma	'scale', 'auto', 0.01, 0.05, 0.1, 0.5
kernel	'rbf', 'poly', 'sigmoid'
degree	2, 3

Tabla 22 Hiperparámetros modelo inicial SVM

El modelo SVM se entrenó y evaluó utilizando validación cruzada de 5 pliegues (5-fold cross-validation), lo que asegura que el modelo generalice bien a datos no vistos.

Modelo Inicial. Se desarrolló un modelo inicial utilizando todas las variables disponibles en el conjunto de datos, las cuales fueron seleccionadas de acuerdo con los criterios expuestos en el capítulo 5 de este documento.

Resultados

Hiperparámetros	Métricas obtenidas		
C = 100000 epsilon = 0.01 gamma = auto kernel = sigmoid degree = 2	MAE (Error Absoluto Medio)	1,358,449.70 USD	El error promedio entre las predicciones y los valores reales es de aproximadamente 1.35 millones de USD.
	MSE (Error Cuadrático Medio)	3,037,836,313,748.03	Este valor (unidades cuadráticas de la variable objetivo) elevado indica que existen desviaciones significativas en algunas predicciones que afectan negativamente el rendimiento global.
	RMSE (Raíz del Error Cuadrático Medio)	1,742,938.98 USD	Significa que la desviación promedio de las predicciones respecto a los valores reales es de aproximadamente 1.74 millones de USD.
	R^2 (Coeficiente de Determinación)	0.4587	El modelo explica el 45.87% de la variabilidad en los datos. Es un valor bajo, e indica que más de la mitad de la variabilidad no está siendo explicada por el modelo.

Tabla 23 Resultado aplicación modelo inicial SVM

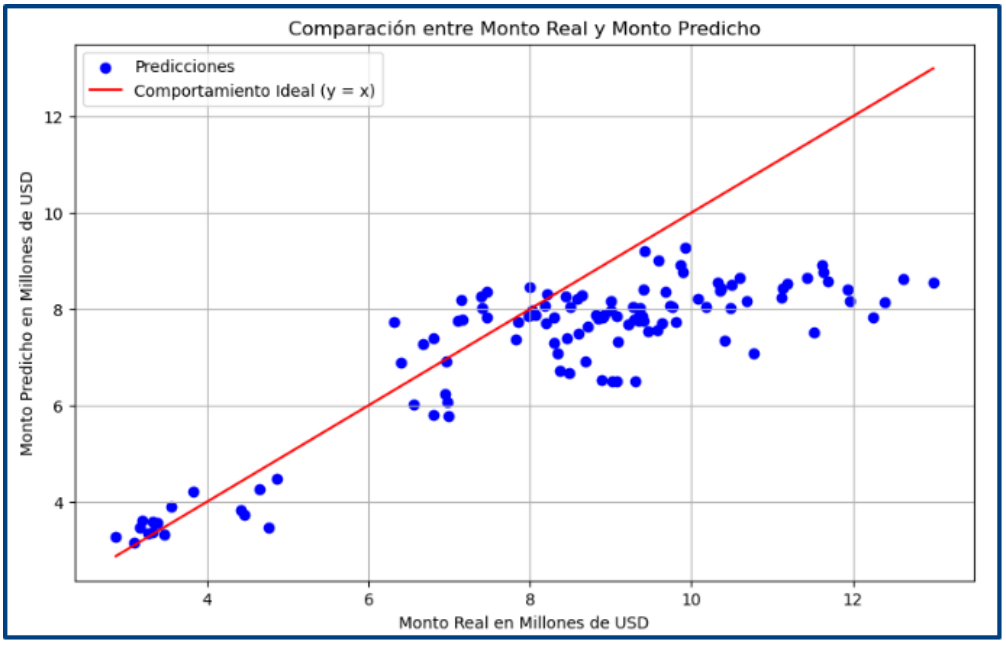

Ilustración 38 Comparación entre monto real y monto predicho

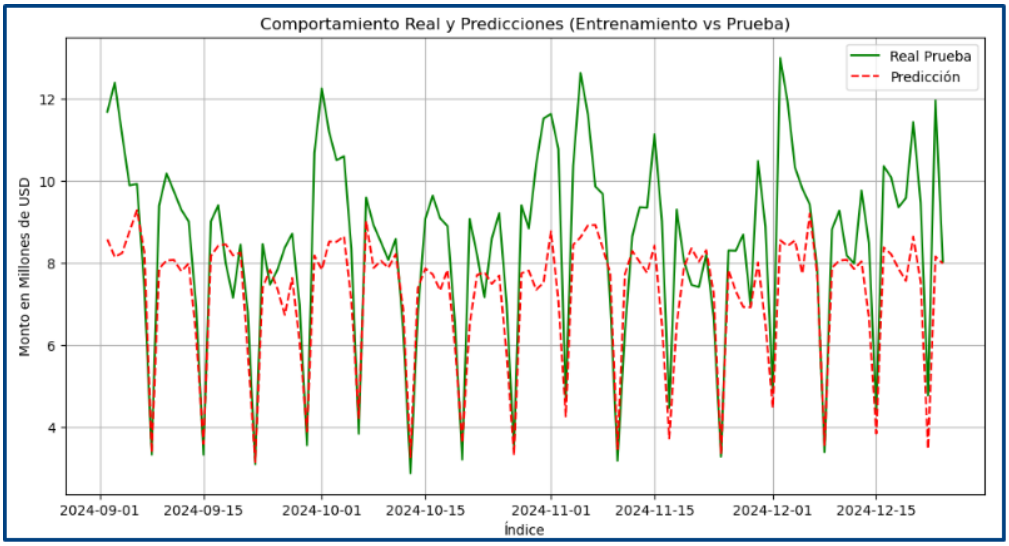

Ilustración 39 Comportamiento real y predicciones (Entrenamiento VS Pruebas)

6.7.2 Optimización y validación del modelo a partir de métricas objetivas

La optimización del modelo en SVM se basa en encontrar los mejores parámetros que minimicen el error en las predicciones.

Con el fin de obtener los mejores resultados posibles y optimizar el rendimiento del modelo predictivo, se llevaron a cabo diversas pruebas en el proceso de entrenamiento. Estas pruebas incluyeron una búsqueda de hiperparámetros para identificar las configuraciones más efectivas, así como la evaluación de diferentes enfoques en la selección de características

- **Modelo 1.** Para la mejora del rendimiento del modelo, se amplió el rango de hiperparámetros probados para el modelo inicial SVM con el objetivo de identificar la mejor combinación de parámetros que minimice el error de predicción. Los hiperparámetros explorados en esta optimización se muestran en la tabla 24.

Hiperparámetros	Rango
C	100, 1000, 5000, 10000, 50000, 100000, 200000
épsilon	0.01, 0.05, 0.1, 0.2, 0.5, 1.0
gamma	'scale', 'auto', 0.01, 0.05, 0.1, 0.2
kernel	'rbf', 'poly', 'sigmoid'
degree	2, 3, 4

Tabla 24 Hiperparámetros modelo SVM Modelo 1

Resultados:

Hiperparámetros	Métricas obtenidas		
C = 200000 epsilon = 1.0 gamma = 0.01 kernel = rbf degree = 2	MAE (Error Absoluto Medio)	1,568,194.19 USD	El modelo está, en promedio, 1.57 millones de USD alejado del valor real del monto total las remesas originadas en un día.
	MSE (Error Cuadrático Medio)	3,798,410,106,995.450	Este valor (unidades cuadráticas de la variable objetivo) elevado indica que algunas predicciones tienen grandes errores.
	RMSE (Raíz del Error Cuadrático Medio)	1,948,951.02 USD	1.95 millones de USD refleja un error significativo en las predicciones, especialmente para valores más extremos.
	R^2 (Coeficiente de Determinación)	0.3232	El modelo explica el 32.3% de la variabilidad en los montos de las remesas. Existe una capacidad limitada para ajustarse a los datos. 67.68% de la variabilidad en los montos de las remesas no está siendo explicada por el modelo

Tabla 25 Resultado aplicación modelo SVM Modelo 1

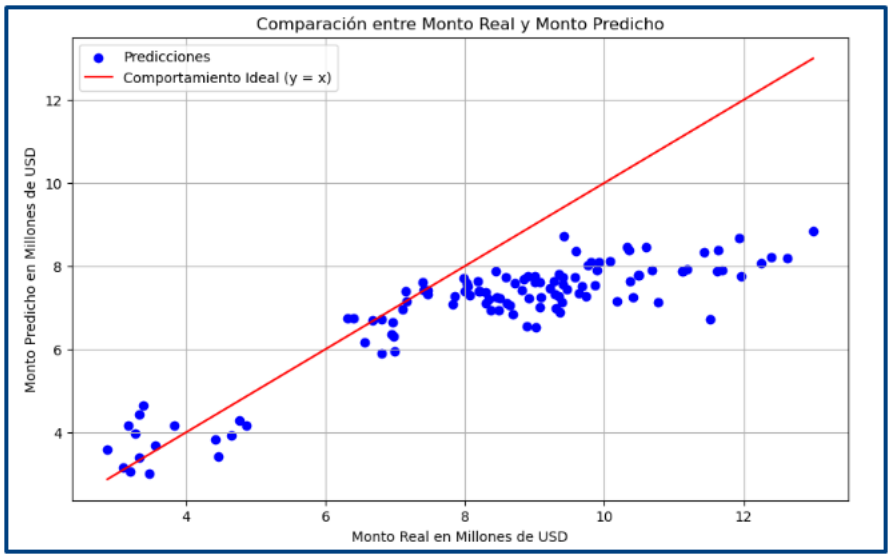

Ilustración 40 Comparación entre monto real y monto predicho Modelo 1

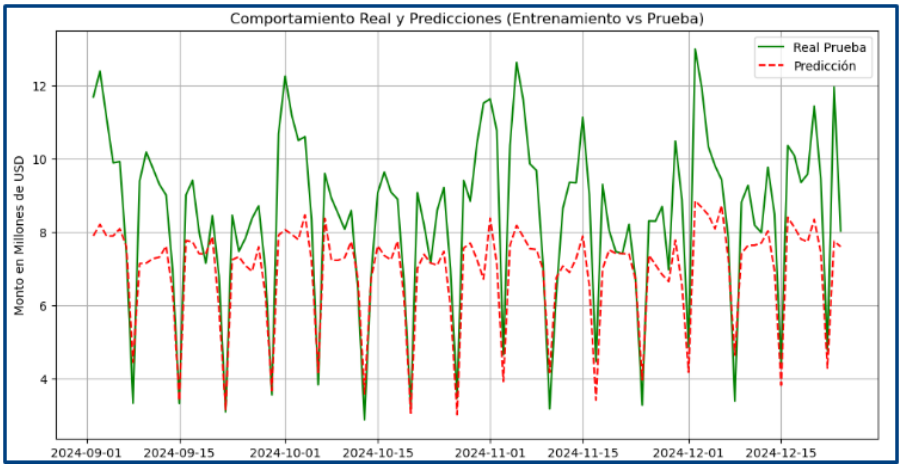

Ilustración 41 Comportamiento real y predicciones (Entrenamiento VS Pruebas) Modelo 1

- **Modelo 2.** En este modelo se amplió el rango de los hiperparámetros en comparación con el Modelo 1. Este ajuste se realizó con el objetivo de explorar un espectro más amplio de posibles configuraciones, permitiendo que el modelo se adaptara de manera más flexible a las características del conjunto de datos.

Hiperparámetros	Rango
C	100, 500, 1000, 5000, 10000, 20000, 50000, 100000, 200000, 500000
épsilon	0.01, 0.05, 0.1, 0.2, 0.5, 1.0
gamma	'scale', 'auto', 0.01, 0.05, 0.1, 0.2
kernel	'rbf', 'poly', 'sigmoid'
degree	2, 3, 4

Tabla 26 Hiperparámetros modelo SVM Modelo 2

Resultados:

Hiperparámetros	Métricas obtenidas		
C = 500000 epsilon = 1.0 gamma = 0.01 kernel = rbf degree = 2	MAE (Error Absoluto Medio)	1,336,015.16 USD	Este valor indica que, en promedio, el modelo se desvía 1.37 millones de USD de los valores reales de las remesas.
	MSE (Error Cuadrático Medio)	2,851,176,419,249.80	Este valor (unidades cuadráticas de la variable objetivo) elevado refleja la existencia de algunas predicciones con errores muy grandes, lo que impacta negativamente el desempeño general.
	RMSE (Raíz del Error Cuadrático Medio)	1,688,542.69 USD	El RMSE muestra una desviación promedio de 1.68 millones de USD en las predicciones del modelo indicando un error sustancial en las predicciones, especialmente en los valores más altos o más bajos de las remesas.
	R^2 (Coeficiente de Determinación)	0.4920	El modelo explica aproximadamente el 49% de la variabilidad en los montos de las remesas. Aunque es una explicación moderada, también indica que hay un 51% de variabilidad que el modelo no está capturando.

Tabla 27 Resultados aplicación modelo SVM Modelo 2

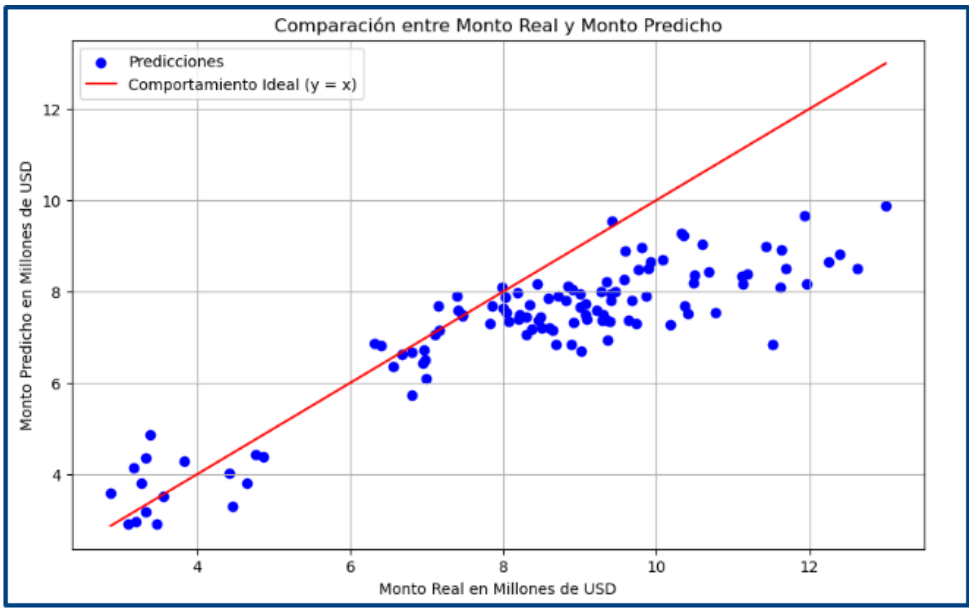

Ilustración 42 Comparación entre monto real y monto predicho Modelo 2

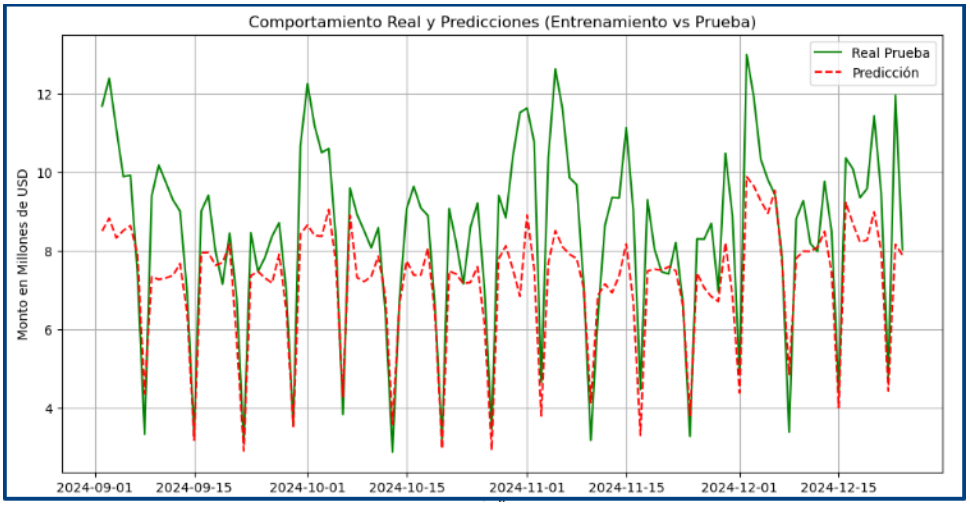

Ilustración 43 Comportamiento real y predicciones (Entrenamiento VS Pruebas) Modelo 2

- **Modelo 3**. En el Modelo 3, se realizó un ajuste al código con los valores de los parámetros clave, como C, épsilon, gamma, kernel y degree que se muestran en la tabla 28. El objetivo fue encontrar una configuración más afinada para mejorar la precisión del modelo en la predicción de los montos de las remesas.

Hiperparámetros	Rango
C	200000, 300000, 400000, 500000, 600000
épsilon	0.01, 0.05, 0.1, 0.2, 0.5
gamma	0.005, 0.01, 0.02, 0.05
kernel	sigmoid
degree	2, 3, 4

Tabla 28 Hiperparámetros modelo SVM Modelo 3

Resultados:

Hiperparámetros	Métricas obtenidas		
C = 600000 epsilon = 0.5 gamma = 0.01 kernel = sigmoid degree = 2	MAE (Error Absoluto Medio)	1,218,649.06 USD	El modelo tiene una desviación promedio de 1.2 millones de USD respecto a los valores reales.
	MSE (Error Cuadrático Medio)	2,468,535,646,161.09	Este valor (unidades cuadráticas de la variable objetivo) indica que, que el modelo presenta algunas predicciones con grandes errores.
	RMSE (Raíz del Error Cuadrático Medio)	1,571,157.42 USD	La desviación promedio de las predicciones es de aproximadamente 1.57 millones de USD. Las predicciones del modelo tienen un margen de error considerable, especialmente en valores extremos.
	R^2 (Coeficiente de Determinación)	0.5601	El modelo explica el 56% de la variabilidad en los montos de las remesas. Si bien este valor es relativamente mejor que en otros modelos, aún queda un 44% de la variabilidad no explicada.

Tabla 29 Resultados modelo SVM Modelo 3

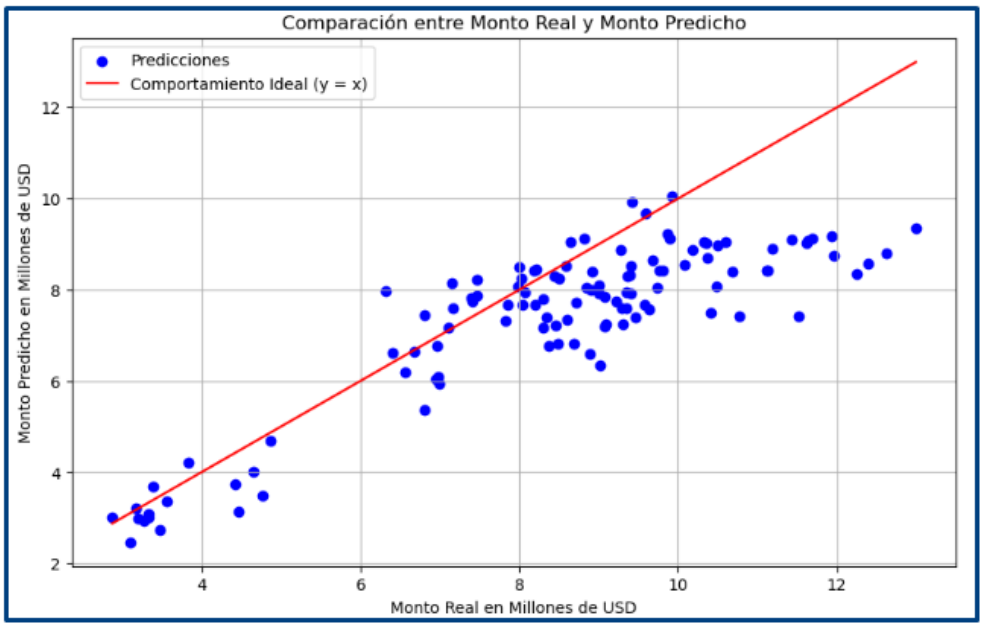

Ilustración 44 Comparación entre monto real y monto predicho Modelo 3

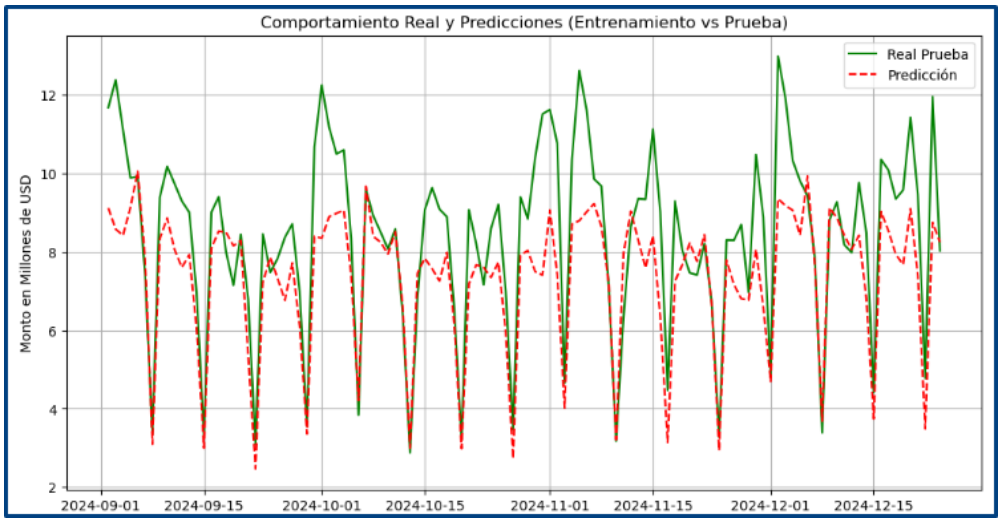

Ilustración 45 Comportamiento real y predicciones (Entrenamiento VS Pruebas) Modelo 3

- **Modelo 4**. En el Modelo 4, se implementó un enfoque adicional para mejorar la precisión de las predicciones mediante el filtrado de variables altamente correlacionadas. La correlación alta entre variables puede inducir a redundancias que afectan la capacidad del modelo para generalizar, lo que podría ocasionar un sobreajuste.

  Se calculó la matriz de correlación de las variables en el conjunto de entrenamiento y se identificaron las variables con correlación superior a 0.9. Las variables altamente correlacionadas que no aportaban nueva información son RedPropia_lag7, US_lag7 y Monto_USD_Pago_lag7.

Tras la identificación de las variables altamente correlacionadas, se eliminaron estas columnas en los conjuntos de entrenamiento y prueba, con el objetivo de reducir la redundancia y mejorar la capacidad del modelo para generalizar.

Luego de limpiar las variables redundantes, se procedió con el ajuste de los hiperparámetros del modelo utilizando GridSearchCV. En esta etapa, se exploraron varios rangos para los hiperparámetros clave:

Hiperparámetros	Rango
C	100, 500, 1000, 5000, 10000, 20000, 50000, 100000, 200000, 500000
épsilon	0.01, 0.05, 0.1, 0.2, 0.5, 1.0
gamma	'scale', 'auto', 0.01, 0.05, 0.1, 0.2
kernel	'rbf', 'poly', 'sigmoid'
degree	2, 3, 4

Tabla 30 Hiperparámetros modelo SVM Modelo 4

Resultados:

Hiperparámetros	Métricas obtenidas		
C = 500000 epsilon = 1.0 gamma = 0.01 kernel = rbf degree = 2	MAE (Error Absoluto Medio)	1,385,063.13 USD	El modelo presenta, en promedio, un error de 1.39 millones de USD en las predicciones con respecto a los valores reales
	MSE (Error Cuadrático Medio)	3,017,633,155,933.30	Este valor (unidades cuadráticas de la variable objetivo), es alto, lo que refleja que el modelo tiene predicciones con errores significativos.
	RMSE (Raíz del Error Cuadrático Medio)	1,737,133.60 USD	La desviación promedio de las predicciones es de aproximadamente 1.74 millones de USD. Confirma el alto error en las predicciones, especialmente cuando se presentan valores más extremos
	R^2 (Coeficiente de Determinación)	0.4623	El modelo explica solo el 46% de la variabilidad en los montos de las remesas.

Tabla 31 Resultados modelo SVM Modelo 4

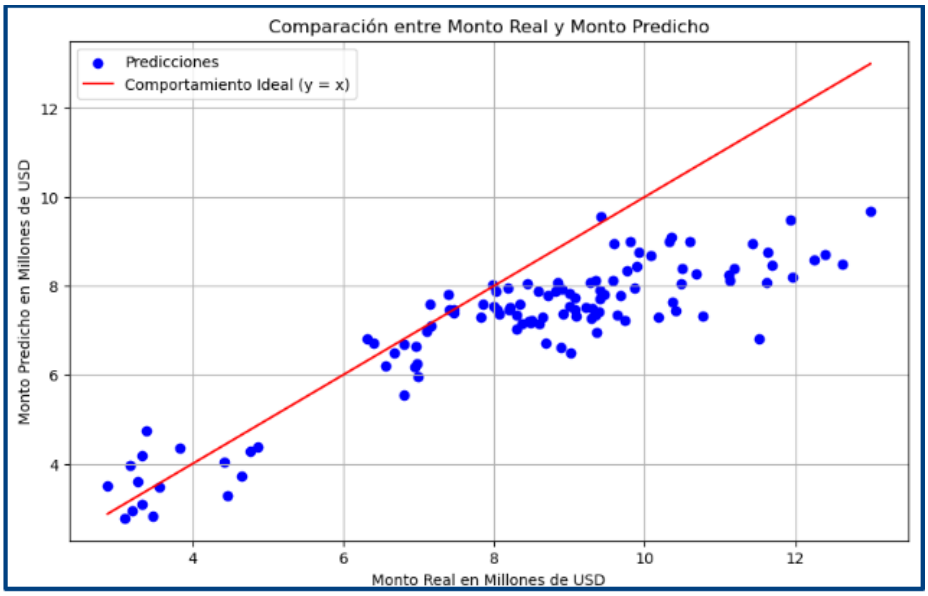

Ilustración 46 Comparación entre monto real y monto predicho Modelo 4

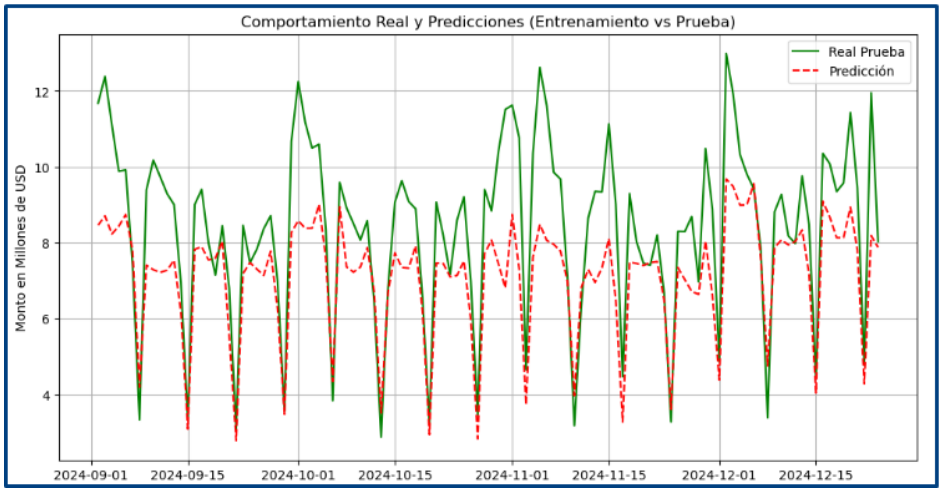

Ilustración 47 Comportamiento real y predicciones (Entrenamiento VS Pruebas) Modelo 4

6.7.3 Comparación de rendimiento de los modelos.

En esta sección se presenta la comparación de resultados obtenidos de los diferentes modelos desarrollados para predecir el monto en dólares de las remesas originadas en un día. A través de una serie de pruebas y ajustes de parámetros, se comparan los rendimientos de varios enfoques, evaluados mediante métricas estándar como el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2). Estas métricas proporcionan información sobre el desempeño y la capacidad predictiva de cada modelo, permitiendo identificar cuál de ellos es más adecuado para predecir con precisión los montos de las remesas.

En la tabla 32 se resumen las métricas que se obtuvieron para cada uno de los modelos SVM entrenados

	MAE	MSE	RMSE	R^2
Modelo Inicial	1,358,449.7 USD	3,037,836,313,748.03	1,742,938.98 USD	45.87%
Modelo 1	1,568,194.1 USD	3,798,410,106,995.450	1,948,951.02 USD	32.30%
Modelo 2	1,336,015.1 USD	2,851,176,419,249.80	1,688,542.69 USD	49.00%
Modelo 3	1,218,649.0 USD	2,468,535,646,161.09	1,571,157.42 USD	56.00%
Modelo 4	1,385,063.1 USD	3,017,633,155,933.30	1,737,133.60 USD	46.00%

Tabla 32 Comparación de rendimiento de los modelos SVM

Análisis de los resultados:

Métrica	Mejor Modelo
Error Absoluto Medio (MAE)	El Modelo 3 presenta el menor MAE (1,218,649.0 USD), lo que indica que, en promedio, las predicciones de este modelo tienen una desviación más pequeña con respecto a los valores reales.
Error Cuadrático Medio (MSE)	El Modelo 3 también tiene el MSE más bajo (2,468,535,646,161.09), lo que refleja que comete menos errores grandes en comparación con los demás, y, por lo tanto, es más preciso al penalizar los errores más significativos.
Raíz del Error Cuadrático Medio (RMSE)	El Modelo 3 presenta el RMSE más bajo (1,571,157.42 USD), confirmando que tiene una deviación promedio más pequeña en sus predicciones. Este resultado es consistente con los valores de MAE y MSE, lo que refuerza su rendimiento general.
Coeficiente de Determinación (R^2)	El Modelo 3 es el que tiene el mayor R^2 de 56%, lo que implica que es capaz de explicar el 56% de la variabilidad en los montos de remesas. Aunque no es un valor muy alto, es el más destacable entre los modelos evaluados

Tabla 33 Análisis de resultados

Del análisis de los resultados de la tabla 33 se tiene que el Modelo 3 ha demostrado ser el más preciso y robusto entre los modelos evaluados. Con el menor MAE, MSE y RMSE, así como el mayor R^2, este modelo ha mostrado una mejor capacidad predictiva, siendo capaz de explicar una mayor proporción de la variabilidad en los montos en dólares de las remesas originadas en un día.

6.8 MODELO DE ÁRBOLES DE DECISIÓN (DECISION TREE REGRESSOR)

Los modelos de árboles de decisión son una herramienta eficaz y ampliamente utilizada para tareas de predicción en datos complejos. En el contexto de la predicción del valor de las remesas, estos modelos permiten dividir el espacio de características en estructuras simples de decisiones que reflejan patrones inherentes en los datos históricos. Esta capacidad para captar relaciones no lineales hace que los árboles de decisión sean particularmente útiles para interpretar y explicar las influencias subyacentes en los flujos de remesas. A continuación, utilizamos el modelo de árbol de decisión para identificar y prever tendencias basándose en las características disponibles.

Los modelos de árboles de decisión (DecisionTreeRegressor) utilizaron como variable objetivo Monto_USD_Originado y como predictores un conjunto de variables explicativas, que incluyeron:

- Variables rezagadas (7 días) de montos por canal (APN, Corresponsales, RedPropia) y por país (US, MX, ES, etc.),
- TRM (Valor_COP_TRM_lag7), y
- Variables temporales codificadas mediante One-Hot Encoding (mes, día de la semana, semana del mes y festivos en Colombia y Estados Unidos).

A diferencia de otros modelos, no se aplicó escalamiento previo a las variables, ya que los árboles de decisión no son sensibles a la escala de los datos.

El conjunto de entrenamiento comprendió las fechas entre el 1 de enero de 2023 y el 30 de septiembre de 2024, mientras que el conjunto de prueba abarcó del 1 de octubre al 24 de diciembre de 2024.

6.8.1 Hiperparámetros por defecto

Tomando en cuenta este set, se utiliza el modelo con los hiperparámetros por defecto, se enlistan a continuación en la tabla 34:

Hiperparámetros	Rango
min_samples_split	Random entre 2 y 4
min_samples_leaf	Random entre 1 y 5
max_features	Random entre 0 y 2
max_depth	Random entre 5 y 7

Tabla 34 Definición de hiperparámetros iniciales – Decision Tree Regressor

Resultados:

Hiperparámetros	Métricas obtenidas		
	MAE (Error Absoluto Medio)	957,345.28 USD	El error promedio entre las predicciones y los valores reales es de aproximadamente 0.96 millones de USD.

min_samples_split >=2 y <= 4 min_samples_leaf >=1 y <= 5 max_features >=0 y <= 2 max_depth=np >=5 y <= 7	MSE (Error Cuadrático Medio)	1,622,196,181,024.84	Este valor (unidades cuadráticas de la variable objetivo) es relativamente alto, lo que indica que hay errores grandes en algunas predicciones, pero es importante considerar que las unidades de medida son muy grandes (en millones de USD).
	RMSE (Raíz del Error Cuadrático Medio)	1,273,654.65 USD	Significa que la desviación promedio de las predicciones respecto a los valores reales es de aproximadamente 1.27 millones de USD.
	R^2 (Coeficiente de Determinación)	0.7109	Indica que el modelo explica el 71.1% de la variabilidad en los datos.

Tabla 35 Resultados hiperparámetros iniciales – Decision Tree Regressor

Al graficar los valores de las predicciones vs los valores reales se puede apreciar:

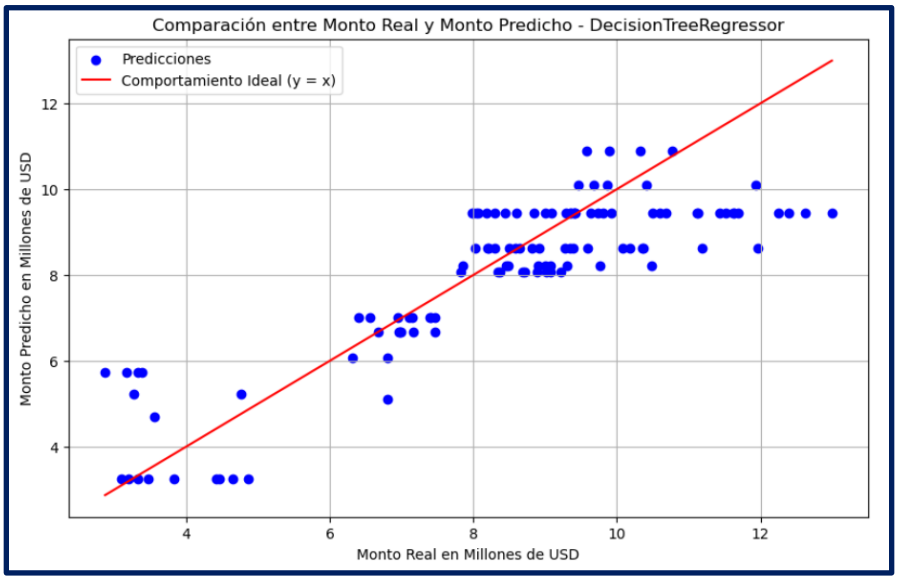

Ilustración 48 Comparación entre Monto Real y Monto Predicho – Decision Tree Regressor Valores por defecto

La dispersión y la alineación de los puntos sugieren que el modelo puede estar enfrentando problemas de sobreajuste. Los árboles de decisión son propensos a sobre ajustarse si no se podan adecuadamente.

Al realizar la superposición de los valores calculados contra los valores reales se evidencia:

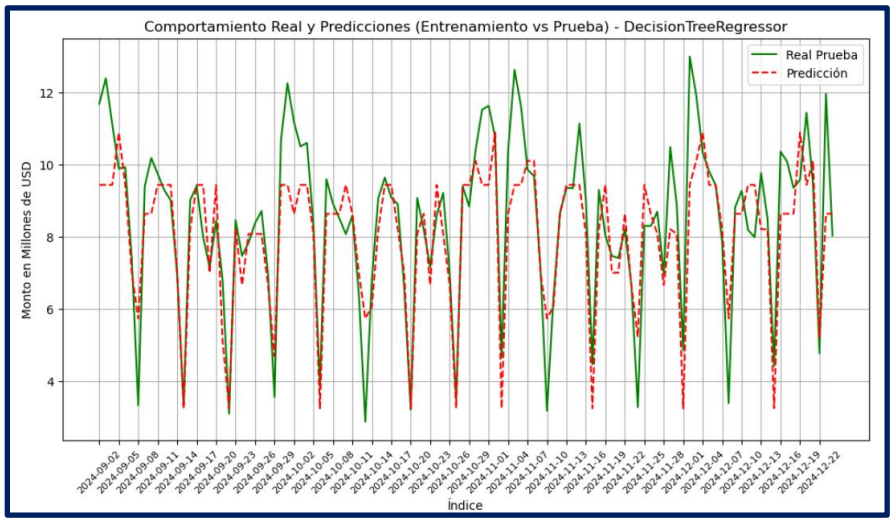

Ilustración 49 Comportamiento real y predicciones

El valor de asertividad es cercano al 65%. Hay diferencias notables entre las predicciones y los valores reales en ciertos puntos, especialmente si las líneas se separan significativamente en algunas áreas. Se puede observar cómo el modelo maneja los picos y valles

6.8.2 Optimización y validación del modelo utilizando GridSearchCV

Dado el bajo nivel de acierto que evidenciamos en el primer proceso, se procede a aplicar el algoritmo GridSearchCV, para ayudar a la identificación de los mejores hiperparámetros posibles para el modelo.

De esta labor se determinan los siguientes valores:

Hiperparámetros	Rango
min_samples_split	5
min_samples_leaf	1
max_features	None
max_depth	7

Tabla 36 Definición de hiperparámetros – Decision Tree Regressor aplicando GridSearchCV

Resultados:

Hiperparámetros	Métricas obtenidas		
min_samples_split = 5 min_samples_leaf = 1 max_features = None max_depth=7	MAE (Error Absoluto Medio)	136,325.57 USD	Se puede apreciar un notable cambio entre las dos mediciones del error medio. Se disminuyó el error en un 85.76% llegando a ser de apenas aproximadamente 136 mil dólares.
	MSE (Error Cuadrático Medio)	134,739,138,720.59	Este valor (unidades cuadráticas de la variable objetivo) es muy bajo comparado con el resultado anterior; llegando únicamente al 8.31%.
	RMSE (Raíz del Error Cuadrático Medio)	367,068.30 USD	La desviación promedio de las predicciones se redujo hasta quedar cercano a los 367 mil dólares con relación a los valores reales. .
	R^2 (Coeficiente de Determinación)	0.9759	En el modelo probando los hiperparámetros iniciales se logró un acierto cercano al 71%. Con este ajuste de parámetros se logra alcanzar un acierto del 97.59%. Este es un número muy significativo que se deberá tener en cuenta en el momento de la implementación final del proyecto.

Tabla 37 Resultados modelo Decision Tree Regressor aplicando GridSearchCV

Para constatar el resultado indicado, procedimos a graficar los valores reales con los que el modelo arrojó. El resultado se muestra a continuación:

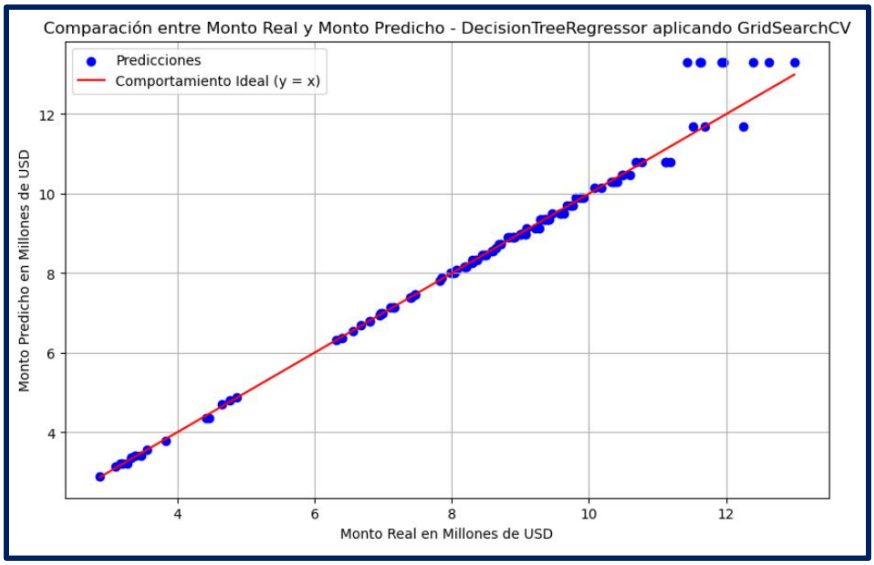

Ilustración 50 Comparación entre Monto Real y Monto Predicho - DecisionTreeRegressor Optimización GridSearchCV

Se evidencia un comportamiento muy estable en toda la recta de las predicciones, sin embargo, en los datos cuyo valor está muy alto, la predicción pierde un poco de exactitud.

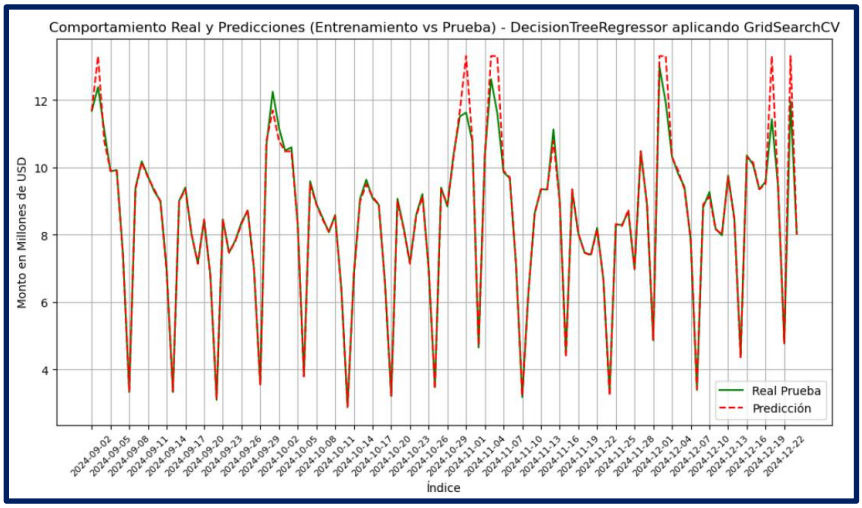

Ilustración 51 Comportamiento real y predicciones – Decision Tree Regressor Optimización GridSearchCV

Cuando se muestran los datos en el tiempo, se evidencia de igual manera el hallazgo que se indicó anteriormente, correspondiente al error cuando los valores se encuentran en los picos altos. En la ilustración 51 se aprecia que el valor para perder la precisión es sobre los 11 millones aproximadamente.

6.8.3 Optimización y validación del modelo utilizando RandomizedSearchCV

Tratando de reforzar la aplicación del modelo e intentando mejorar el comportamiento del modelo en los valores muy altos, se hace la búsqueda de hiperparámetros por el método de RandomizedSearchCV. De este se logra obtener los valores que se detallan en la tabla 37.

Hiperparámetros	Rango
min_samples_split	4
min_samples_leaf	2
max_features	None
max_depth	2

Tabla 38 Hiperparámetros encontrados con RandomizedSearchCV

Resultados:

Hiperparámetros	Métricas obtenidas		
min_samples_split = 4 min_samples_leaf = 2 max_features = None max_depth=2	MAE (Error Absoluto Medio)	900,575.20 USD	En este nuevo proceso, el error Absoluto medio disminuye aún más, llegando a niveles 900 mil dólares
	MSE (Error Cuadrático Medio)	1,599,101,440,503.78	El valor aumenta en el error cuadrado. Se incrementa desmejorando el resultado de la aplicación del modelo.
	RMSE (Raíz del Error Cuadrático Medio)	1,264,555.82 USD	En promedio, las predicciones actuales son menos precisas que las anteriores en esa métrica específica (RMSE).
	R^2 (Coeficiente de Determinación)	0.7150	La caída del R^2 de 0.9759 a 0.7150 refleja reducción en la capacidad del modelo para explicar la variabilidad en los datos,

Tabla 39 Resultados método – Decision Tree Regressor Optimización GridSearchCV

La presentación grafica de los resultados es la siguiente:

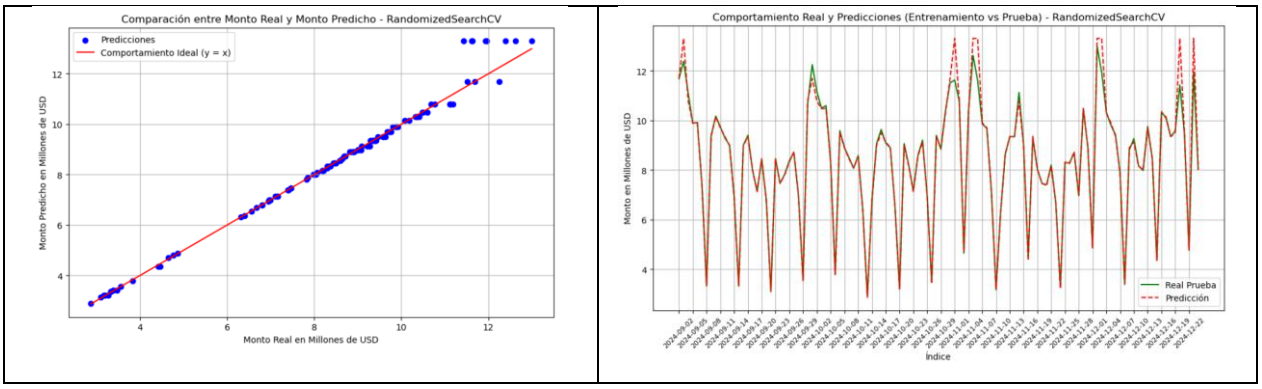

Ilustración 52 Comportamiento real y predicciones – Decision Tree Regressor Optimización RandomizedSearchC

6.8.4 Optimización y validación del modelo utilizando Bayesian

Para finalizar la determinación de los hiperparámetros, vamos a utilizar el método de Bayasian para determinar los valores más convenientes

Hiperparámetros	Rango
min_samples_split	2
min_samples_leaf	1
max_features	None
max_depth	10

Tabla 40 Hiperparámetros encontrados con Bayasian

Resultados:

Hiperparámetros	Métricas obtenidas		
min_samples_split = 2 min_samples_leaf = 1 max_features = None max_depth=10	MAE (Error Absoluto Medio)	135.232,44 USD	El nuevo ajuste no altera el resultado esperado.
	MSE (Error Cuadrático Medio)	137.581.020.806,79	El nuevo ajuste no altera el resultado esperado.
	RMSE (Raíz del Error Cuadrático Medio)	370.919,15 USD	El nuevo ajuste no altera el resultado esperado.
	R^2 (Coeficiente de Determinación)	0,9754	El nuevo ajuste no altera el resultado esperado.

Tabla 41 Resultados método – Decision Tree Regressor Optimización Bayasian

El resultado de este método para encontrar los hiperparámetros no varía; por esta razón no se hace más trabajo en ajuste de hiperparámetros por considerar que el resultado esperado está acorde a lo esperado y que el error presente no implica un mal funcionamiento del modelo en general. Se determina que este es una buena parametrización para tener en cuenta. El acierto es de 97.54%

A continuación, en la ilustración 53, se presentan de manera gráfica los resultados:

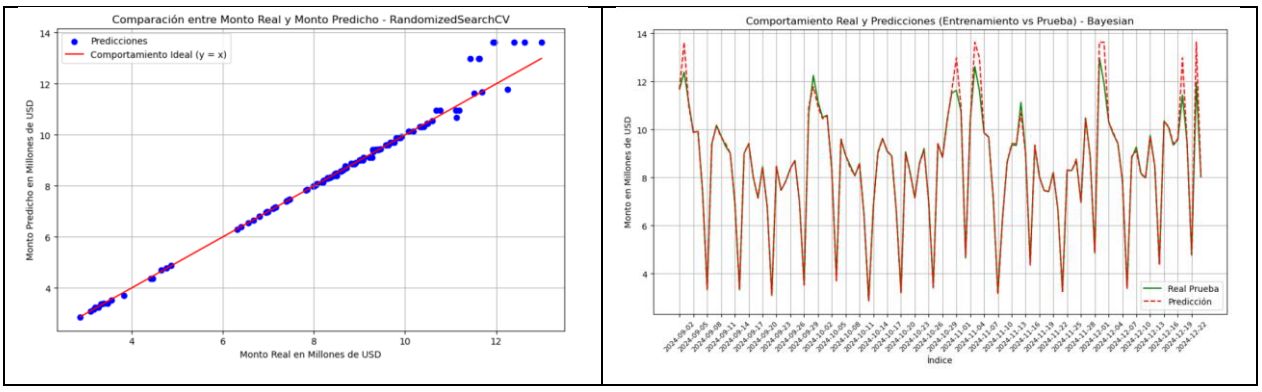

Ilustración 53 Comportamiento real y predicciones - DecisionTreeRegressor Optimización Bayesian

6.8.5 Comparación de rendimiento de los modelos

Habiendo trabajado en la búsqueda de los hiperparámetros y logrando unos grandes resultados, se muestra a continuación la tabla 42 donde se plasma el resumen logrado anteriormente.

	MAE	MSE	RMSE	R^2
Modelo Inicial	957,345.28 USD	1,622,196,181,024.84	1,273,654.65	71.09%
Modelo 1 – GridSearchCV	136,325.57 USD	134,739,138,720.59	367,068.30	97.60%
Modelo 2 - RandomizedSearchCV	900,575.20 USD	1,599,101,440,503.78	1,264,555.82	71.50%
Modelo 3 - BayesSearchCV	135,232.44 USD	137,581,020,806.79	370,919.15	97.54%

Tabla 42 Resultado de aplicar el método Decision Tree Regressor

Análisis de resultados:

Las métricas de rendimiento muestran una mejora significativa en el ajuste del modelo tras aplicar técnicas de optimización de hiperparámetros. El modelo inicial tenía un Error Absoluto Medio (MAE) de aproximadamente 957,345.28 USD y explicaba solo el 71.09% de la variabilidad de los datos, reflejando que su capacidad predictiva era moderada y con errores relativamente altos. En cambio, los modelos ajustados mediante búsqueda de hiperparámetros, como GridSearchCV, RandomizedSearchCV y BayesSearchCV, lograron reducir notablemente estos errores, alcanzando un MAE cercano a los 136 USD mil dólares y aproximadamente el 97.6% con un aumento en la precisión. Aunque ambos métodos de optimización (aleatorio y Randomized) arrojaron resultados muy similares, el modelo con GridSearchCV mostró ligeramente mejores métricas, indicando que la búsqueda exhaustiva permitió encontrar los hiperparámetros más adecuados. En conjunto, estos resultados sugieren que la afinación de hiperparámetros es crucial para mejorar significativamente la precisión del modelo predictivo, alcanzando un rendimiento óptimo con un error medio por debajo de los 137,000 USD.

6.9 DESEMPEÑO DE MODELOS: MEJORES RESULTADOS POR TÉCNICA

A continuación, se presentan los mejores resultados obtenidos para cada uno de los modelos evaluados durante el proceso de desarrollo. Los resultados se centran en el coeficiente de determinación R^2, que es una métrica clave utilizada para medir la capacidad de los modelos de explicar la variabilidad de los datos de manera efectiva. A través de la comparación de los diferentes enfoques, se busca identificar el modelo con el mejor desempeño y determinar cuál es el más adecuado para la predicción del monto en dólares de las remesas originadas en un día en función de los resultados obtenidos.

Modelo	Mejor resultado R^2
Modelo de regresión con SVM	56.00%
Series de Tiempo - Modelo SARIMA	59,52%
Red neuronal tipo JORDAN	87.30%
Modelo de árboles de decisión (DECISION TREE REGRESSOR)	97.60%.

Tabla 43 Mejores resultados de cada modelo implementado

En los resultados obtenidos, según la tabla 43, los Modelos de Red Neuronal Tipo JORDAN y Árboles de Decisión (Decision Tree Regressor) se destacaron por su capacidad para explicar la variabilidad de los datos, alcanzando valores de R^2 de superiores al 87% y 99% respectivamente. Estos modelos lograron un alto desempeño debido a las características inherentes de cada uno y su capacidad para capturar patrones complejos en los datos de las remesas.

- Red Neuronal Tipo JORDAN:

Las redes neuronales tipo Jordan, una variante de las redes neuronales recurrentes, destacan por su capacidad de modelar dependencias temporales mediante una estructura en la que las salidas anteriores se retroalimentan hacia las capas ocultas. Esta arquitectura permite capturar la dinámica de los datos secuenciales, lo que resulta especialmente útil en problemas donde la evolución de los eventos a lo largo del tiempo influye directamente en la predicción.

El modelo Jordan, al integrar estas conexiones recurrentes, ofrece una ventaja en la identificación de patrones temporales, facilitando la captura de tendencias y relaciones entre los datos históricos y las futuras predicciones. Esta capacidad de aprendizaje secuencial permitió que el modelo alcanzara un 87% de precisión, lo que indica una sólida capacidad para realizar pronósticos en entornos donde la dependencia de estados previos es clave.

- Modelo de Árboles de Decisión (DECISION TREERE GRESSOR):

El DECISION TREE REGRESSOR se destacó por su capacidad para modelar relaciones no lineales y por su flexibilidad en la segmentación de los datos. Este modelo no asume ninguna relación lineal entre las variables, lo que le permite ajustarse mejor a patrones complejos en los datos, como los de las remesas. El algoritmo de árbol de decisión divide el espacio de características en segmentos más pequeños, lo que facilita la identificación de reglas específicas que afectan las predicciones. Esta característica de "dividir y conquistar" fue fundamental en el contexto del proceso de remesas, ya que existen muchos factores que pueden cambiar de manera no lineal en función de las fechas, valores y otros factores económicos. Al alcanzar un R^2 de 97.60%, el modelo mostró una precisión notable al explicar las variabilidades en los montos de las remesas, lo que lo hace altamente confiable para predicciones en este dominio.

Con base en los resultados obtenidos y su capacidad para capturar dependencias temporales de manera efectiva, el modelo de Árboles de Decisión (DECISION TREE REGRESSOR) se presenta como la mejor opción para predecir el monto en dólares de las remesas originadas en un día.

7 CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

El presente trabajo de grado permitió analizar, construir y evaluar múltiples modelos de predicción aplicados a series de tiempo, con el objetivo de estimar el monto total en dólares de remesas originadas diariamente. El proceso metodológico integró enfoques estadísticos clásicos, modelos de redes neuronales recurrentes y técnicas de aprendizaje automático, lo que permitió explorar distintos niveles de complejidad, interpretabilidad y desempeño predictivo.

Los modelos estadísticos tradicionales ofrecieron una base sólida para entender la estructura temporal de la serie. El modelo ARIMA, aunque útil como punto de partida, presentó un rendimiento deficiente (R^2 negativo), confirmando que no es suficiente para capturar la dinámica de una serie con estacionalidad marcada. La incorporación de estacionalidad mediante el modelo SARIMA mejoró significativamente el desempeño (R^2: 59.52 %), y los modelos Holt-Winters –tanto en su versión original como con transformación logarítmica– lograron representar parcialmente los ciclos temporales. El modelo Box-Jenkins, con estructura ARIMA(7,0,1), alcanzó un desempeño limitado, inferior al de SARIMA, pero con menor error absoluto.

Posteriormente, se exploraron modelos con variables exógenas mediante SARIMAX. El modelo con todas las variables disponibles presentó menor precisión, mientras que el modelo reducido a variables estadísticamente significativas mejoró tanto el MAE como la capacidad explicativa (R^2: 49.98 %).

La mayor mejora en desempeño se obtuvo al aplicar modelos de redes neuronales recurrentes. El modelo Jordan fue el que alcanzó el mejor resultado global (R^2: 87.32 %), con un MAE bajo de USD 679,373. Los modelos Elman, tanto en su versión base como optimizada, ofrecieron desempeños aceptables, con R^2 entre el 66% y el 70%. Estos modelos demostraron una capacidad aceptable para capturar la estructura secuencial y no lineal de la serie de remesas, validando su aplicación en contextos con suficientes datos históricos estructurados.

Dentro del grupo de técnicas de aprendizaje automático, se destaca el modelo de árboles de decisión optimizado mediante búsqueda en malla (GridSearchCV), que logró un R^2 de 99.58 % con un MAE muy bajo, sin requerir estructuras de secuencia. En contraste, el modelo LSTM presentó un desempeño variable: su versión base obtuvo un R^2 de 32.14 %, mientras que la versión optimizada alcanzó R^2 de 57.48 %. A pesar de los ajustes arquitectónicos, no se lograron mejoras sustanciales, lo que sugiere que un incremento en la complejidad del modelo no siempre se traduce en mayor precisión.

Finalmente, el modelo de regresión basado en SVM mostró un desempeño limitado (R^2: 56 %), a pesar de los ajustes en los hiperparámetros, siendo la técnica con menor capacidad para explicar la variabilidad en los montos de remesas.

Desde una perspectiva de negocio, los resultados obtenidos tienen aplicaciones relevantes para la gestión operativa y estratégica del APPD. La capacidad de estimar con alta precisión el monto total en dólares que será

originado en un día determinado permite mejorar la planificación financiera, anticipar requerimientos de liquidez y ajustar posiciones en mercados cambiarios. En particular, disponer de predicciones confiables contribuye a una gestión más eficiente del riesgo de tasa de cambio, al permitir tomar decisiones con base en proyecciones técnicas y no únicamente en tendencias históricas o criterios subjetivos. Los modelos desarrollados, especialmente aquellos con alto R^2 y bajo error absoluto, podrían integrarse como herramientas complementarias en los procesos de pronóstico de flujo operativo del APPD y respaldar decisiones de cobertura, asignación de recursos y monitoreo de exposiciones cambiarias.

En conjunto, los resultados muestran que el desempeño predictivo mejora sustancialmente al incorporar arquitecturas que capturan no linealidades y secuencias, como redes recurrentes y árboles de decisión. Sin embargo, los modelos estadísticos siguen siendo útiles como referencia inicial y como herramientas interpretables. La selección adecuada de variables, la validación del desempeño en conjuntos separados y la combinación de enfoques clásicos y modernos constituyen elementos fundamentales para abordar de manera efectiva problemas de predicción en series temporales operativas.

7.2 TRABAJOS FUTUROS

A pesar de que en algunos de los modelos desarrollados se ha mostrado un rendimiento sólido en términos de precisión y capacidad explicativa, existen diversas líneas de trabajo que podrían ser exploradas en fases futuras para robustecer, extender o adaptar el enfoque actual. La implementación práctica de modelos predictivos en entornos dinámicos exige tanto un perfeccionamiento técnico como una integración estratégica con los procesos operativos del negocio. A continuación, se presentan algunas recomendaciones para trabajos futuros:

- Optimización continua y mantenimiento adaptativo

Una de las recomendaciones clave es establecer un proceso de optimización y reevaluación periódica del modelo, tanto en términos de hiperparámetros como en selección de variables. Esto resulta especialmente relevante ante cambios en la dinámica de los flujos de remesas, ajustes regulatorios o condiciones macroeconómicas distintas. Además, se sugiere aplicar procesos de validación cruzada y monitoreo de métricas en tiempo real, que permitan anticipar degradaciones en el desempeño y activar procesos de reentrenamiento de forma oportuna.

- Despliegue en entorno productivo

El valor de un modelo predictivo se materializa plenamente cuando puede ser utilizado de forma recurrente y automatizada dentro de los procesos del negocio. Por ello, una línea clave de trabajo futuro es el desarrollo de un entorno de producción que integre la actualización de datos, la generación de predicciones y el monitoreo de resultados. Este despliegue debería considerar aspectos de arquitectura como la modularidad, la escalabilidad y la trazabilidad de resultados. Asimismo, se recomienda incorporar mecanismos de alerta frente a desviaciones inesperadas y métricas de seguimiento que aseguren la estabilidad del modelo en el tiempo. En el caso específico del APPD, la implementación de un sistema de predicción diaria permitiría anticipar el monto de remesas a gestionar, mejorando la toma de decisiones en términos de asignación de liquidez, cobertura y administración del riesgo de tasa de cambio.

8 REFERENCIAS BIBLIOGRÁFICAS

[1] B. Mundial, «Migration and Remittances: Recent Developments and Outlook,» *Migration and Remittances: Recent Developments and Outlook,* 2018.

[2] Valora_Analitik, «Yahoo Finanzas,» 24 01 2025. [En línea]. Available: https://es-us.finanzas.yahoo.com/noticias/remesas-colombia-rompieron-r%C3%A9cord-2024-232000279.html. [Último acceso: 20 02 2025].

[3] B. d. l. R. d. Colombia, «banrep.gov.co,» 9 10 2024. [En línea]. Available: https://www.banrep.gov.co/es/blog/evolucion-reciente-ingresos-externos-remesas-hacia-colombia.

[4] Funcion_Publica, «Funcion Publica,» 25 10 2022. [En línea]. Available: https://www1.funcionpublica.gov.co/web/eva/curso-para-veedurias-ciudadanas.

[5] Superintendencia_Financera_de_Colombia, «superfinanciera.gov.co,» 31 01 2025. [En línea]. Available: https://www.superfinanciera.gov.co/publicaciones/10083443/normativanormativa-generalcircular-basica-juridica-ce-10083443/.

[6] R. &. H. J. Wirth, «CRISP-DM: Towards a standard process model for data mining. En Proceedings of the 4th International Workshop on Knowledge Discovery in Databases,» de *CRISP-DM: Towards a standard process model for data mining. En Proceedings of the 4th International Workshop on Knowledge Discovery in Databases*, 2000, pp. 29 - 39.

[7] C. M. Bishop, Pattern Recognition and Machine Learning. Springer., 2009.

[8] L. &. R. E. Pascual, Predicción de series temporales basada en Machine Learning: aplicaciones económicas y financieras, En F. Pérez (Ed.), 2021.

[9] G. J. G. &. R. G. Box, Time Series Analysis: Forecasting and Control. 5th Edition, Pearson, 2015.

[10] S. J. Taylor, Modelling Financial Time Series. World Scientific Publishing, 2003.

[11] R. &. A. G. Hyndman, Forecasting: Principles and Practice (2nd ed.), OTexts, 2018.

[12] D. &. R. J. Peña, «Modeling and Forecasting Remittances: An Application of Box-Jenkins Methodology.,» *Journal of Forecasting,* vol. 21, nº 1, pp. 33 - 45, 2002.

[13] A. J. &. S. B. Smola, «A tutorial on support vector regression. Statistics and Computing,» de *Statistics and computing*, vol. 14, Springer, 2004, p. 199–222.

[14] S. &. S. J. Hochreiter, Long short-term memory. Neural Computation, 1997.

[15] I. Islas, V. M. Guerrero y E. Silva, «Forecasting remittances to Mexico with a Multi-State Markov–Switching model applied to the trend with controlled smoothness.,» *Romanian Journal of Economic Forecasting,* vol. 1, nº 22, pp. 38 - 56, 2019.

[16] G. Bontempi, S. B. Taieb y Y. A. Le Borgne, «European Business Intelligence Summer School,» de *Machine Learning Strategies for Time Series Forecasting*, Springer, 2013, p. 62–77.

[17] IBM, «ibm.com,» [En línea]. Available: https://www.ibm.com/es-es/topics/machine-learning.