PREDICCIÓN DE VENTAS POR IDENTIFICACIÓN DISPERSA DE UN SISTEMA ERP A PARTIR DE DATOS DE UN MÓDULO POS

PROGRAMA DE MAESTRÍA EN INGENIERÍA

CESAR DANIEL RINCÓN BRITO

ORC ID: https://orcid.org/0009-0007-9999-6889

Director

Dr. Luis Eduardo Tobón Llano

Pontificia Universidad Javeriana Cali

Facultad de Ingeniería y Ciencias

Enero 14 de 2026

Contenido

LISTA DE FIGURAS

LISTA DE TABLAS

Agradecimientos

En primer lugar, agradecer a mi director de proyecto, el Dr. Luis Eduardo Tobón Llano, por su guía, conocimiento y apoyo en la finalización de este proyecto.

A mis compañeros de especializaciones, maestrías y doctorados por sus aportes y apoyo en el transcurso de cada semestre.

A las entidades que prestaron sus recursos y su información para el desarrollo de esta investigación.

"El azar no es más que la medida de la ignorancia del hombre"

Henri Poincaré

Resumen

La eficiencia en la gestión de los recursos empresariales se ha convertido en un factor determinante para la competitividad organizacional. En este contexto, los sistemas de planificación de recursos empresariales (ERP) integran múltiples procesos empresariales en una plataforma centralizada, permitiendo el acceso y análisis de datos en tiempo real. Sin embargo, para aprovechar al máximo estos sistemas, es necesario aplicar herramientas analíticas que conviertan los datos disponibles en información estratégica.

Esta investigación se enfoca en la implementación de modelos predictivos dentro del entorno de un ERP, con el objetivo de anticipar comportamientos y optimizar procesos en áreas como finanzas, logística, ventas y recursos humanos. Entre las técnicas estudiadas se encuentra el algoritmo SINDy (Sparse Identification of Nonlinear Dynamical Systems), que permite descubrir modelos dinámicos a partir de datos mediante regresión dispersa, seleccionando solo los términos más relevantes de una biblioteca de funciones.

El estudio busca demostrar cómo, a través del análisis predictivo y la aplicación de modelos de identificación de sistemas dinámicos, es posible mejorar la toma de decisiones, reducir tiempos de espera y aumentar la precisión en operaciones críticas de la organización, especialmente cuando se dispone de una base de datos abundante y confiable dentro del sistema ERP.

Palabras clave: Keywords: Enterprise Resource Planning (ERP), Predictive Models, Dynamic Systems, System Identification, SINDy Regression, Data-Driven Modeling, Finance Analytics, Logistics Optimization, Business Intelligence.

Abstract

Efficiency in enterprise resource management has become a key factor for organizational competitiveness. In this context, Enterprise Resource Planning (ERP) systems integrate multiple business processes into a centralized platform, enabling real-time access to and analysis of data. However, to fully leverage these systems, it is necessary to apply analytical tools that transform available data into strategic information.

This research focuses on the implementation of predictive models within an ERP environment, with the aim of anticipating behaviors and optimizing processes in areas such as finance, logistics, sales, and human resources. Among the techniques studied is the SINDy algorithm (Sparse Identification of Nonlinear Dynamical Systems), which enables the discovery of dynamic models from data through sparse regression by selecting only the most relevant terms from a library of candidate functions.

The study aims to demonstrate how, through predictive analysis and the application of dynamic system identification models, it is possible to improve decision-making, reduce waiting times, and increase accuracy in critical organizational operations—especially when a rich and reliable database is available within the ERP system.

Keywords: Enterprise Resource Planning (ERP), Predictive Models, Dynamic Systems, System Identification,SINDy Regression, Data-Driven Modeling, Finance Analytics, Logistics Optimization, Business Intelligence.

1. Introducción

En la actualidad, la gestión eficiente de los recursos empresariales es esencial para el éxito y la competitividad de cualquier organización. Para lograrlo, muchas empresas han recurrido a la implementación de sistemas de planificación de recursos empresariales (ERP, por sus siglas en inglés). Un sistema ERP integra diversas funciones y procesos de una empresa, como finanzas, compras, inventario, ventas y recursos humanos, en una sola plataforma de software centralizada.

El sistema ERP proporciona una visión en conjunto de las operaciones empresariales al permitir la recopilación, el almacenamiento y el procesamiento de datos en tiempo real. Esta capacidad de recopilación y procesamiento de datos a gran escala brinda a las empresas una valiosa oportunidad para extraer información estratégica y tomar decisiones informadas. Sin embargo, el simple almacenamiento y procesamiento de datos no es suficiente. Es necesario aprovechar al máximo estos datos para obtener conocimientos predictivos que ayuden a las empresas a anticiparse a las tendencias y desafíos futuros. [3]

Es en este punto donde entran en juego los modelos predictivos. Los modelos predictivos son algoritmos y técnicas analíticas que permiten predecir eventos futuros o comportamientos basados en datos históricos. Estos modelos utilizan patrones y tendencias identificados en los datos para hacer inferencias y estimaciones precisas sobre eventos que aún no han ocurrido. En el contexto de un sistema ERP [4], los modelos predictivos pueden aplicarse a una amplia gama de áreas, como la demanda de productos, la gestión del inventario, la planificación de la producción, el análisis financiero y la gestión de recursos humanos, entre otros [5].

Consultar información y realizar un análisis predictivo para un determinado proceso es importante en los negocios, en especial para las áreas de finanzas, logística, ventas, atención al cliente, márquetin o recursos humanos, logrando ofrecer oportunas respuestas para determinar fallos, tendencias o lentitud en procesos por medio de herramientas tecnológicas y soporte continuo que permiten decidir cómo manipular estos datos [15].

Descubrir modelos de sistemas dinámicos a partir de datos es de vital importancia en la ciencia y la ingeniería [1]. Tradicionalmente, los modelos se derivan de los primeros principios, aunque este enfoque puede ser un desafío en muchos campos, como la ciencia del clima, las finanzas y la biología. Afortunadamente, el descubrimiento de modelos basados en datos, es decir, la

identificación de sistemas [6], es un campo en rápido desarrollo, con una gama de técnicas que incluyen enfoques lineales clásicos, tales como la descomposición de modo dinámico, modelos auto regresivos no lineales, redes neuronales, proceso de regresión gaussiana, análisis espectral, mapas de difusión y regresión dispersa son algunos de los desarrollos recientes.

La aplicación del algoritmo SINDy (Sparse Identification of Nonlinear Dynamical Systems) [2] en el pronóstico de ventas representa un enfoque innovador para comprender y modelar el comportamiento dinámico de sistemas discretos no lineales. A diferencia de los métodos estadísticos tradicionales, SINDy permite identificar las ecuaciones que gobiernan la evolución temporal de las variables del sistema directamente a partir de los datos [19], evidenciando relaciones y dependencias no lineales entre factores de venta. En el contexto financiero o comercial, este método facilita capturar la naturaleza variable de las dinámicas del mercado, donde las ventas pueden verse influenciadas por múltiples factores internos y externos que varían con el tiempo. Al generar modelos parciales de estas estructuras de datos, SINDy ofrece una herramienta potente para predecir tendencias futuras, evaluar escenarios y optimizar estrategias comerciales, contribuyendo a una toma de decisiones más informada y basada en evidencia cuantitativa.

1.1. Contribución a los ODS

1.1.1 ODS 8: Trabajo decente y crecimiento económico

La aplicación del modelo SINDy para la predicción del comportamiento de ventas en un sistema ERP permite mejorar la planificación y toma de decisiones en las organizaciones, impulsando la eficiencia y productividad. Al contar con herramientas analíticas que anticipan comportamientos de mercado, se generan condiciones favorables para un crecimiento económico sostenido, optimizando recursos y fortaleciendo la estabilidad laboral en entornos comerciales.

1.1.2 ODS 9: Industria, innovación e infraestructura

Esta investigación promueve el uso de metodologías innovadoras como SINDy para identificar modelos dinámicos en datos reales de sistemas ERP. El desarrollo de esta solución contribuye al fortalecimiento de la infraestructura tecnológica de las organizaciones y fomenta la innovación en la industria mediante la incorporación de

técnicas avanzadas de análisis de datos y predicción, facilitando la digitalización de procesos operativos [13].

1.1.3 ODS 12: Producción y consumo responsables

Al anticipar la demanda y el comportamiento de las ventas, el modelo permite una gestión más eficiente del inventario y los recursos, lo que reduce desperdicios, sobreproducción y pérdidas innecesarias. Esto facilita una producción ajustada a la demanda real, promoviendo prácticas de consumo más responsables y sostenibles dentro de la cadena de suministro empresarial.

1.1.4 ODS 17: Alianzas para lograr los objetivos

El desarrollo del proyecto se apoya en el trabajo colaborativo entre actores académicos y empresariales, destacando la importancia de integrar conocimiento científico con necesidades reales del entorno productivo. Este enfoque fomenta alianzas estratégicas que potencian el intercambio de datos, la validación de modelos y la implementación de soluciones tecnológicas en contextos empresariales reales.

2. Definición del problema de investigación

El estudio se centra en la aplicación de técnicas de análisis predictivo para aprovechar los datos dispersos recopilados de un módulo de punto de venta (POS) en un sistema de planificación de recursos empresariales (ERP) desarrollando un modelo que buscará utilizar datos históricos de ventas, junto con otros factores relevantes como la temporada, promociones, y características específicas del producto, para prever con precisión las futuras ventas. El objetivo es mejorar la toma de decisiones empresariales al anticipar las tendencias de ventas y optimizar la gestión de inventario y recursos en el contexto del comercio minorista o empresarial.

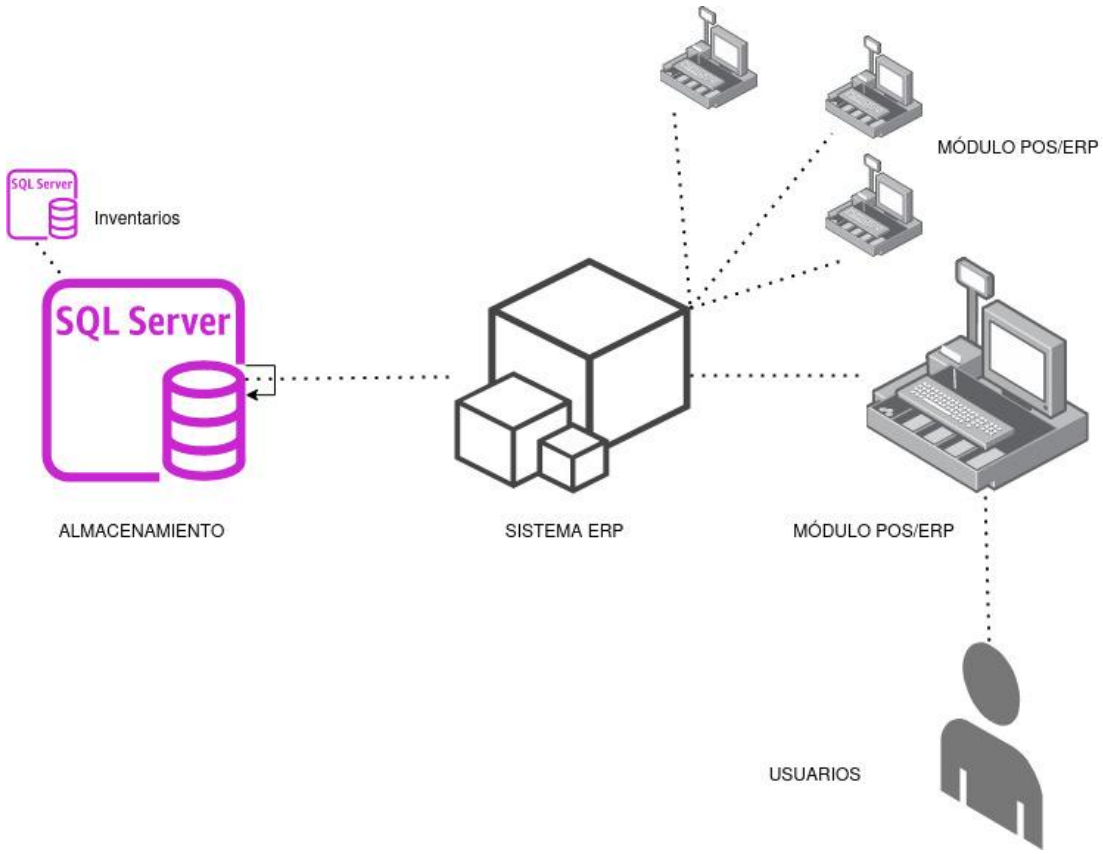


Figura 1. Sistema ERP y módulo POS (Punto de venta)

Fuente: El autor.

A partir de la Figura 1, la compañía cuenta con 98 tiendas a nivel nacional con un sistema ERP implementado donde las ventas se hacen mediante el módulo POS (Punto de venta), la cual ofrece diferentes tipos de producto en cosmética, cuenta con un registro de 65 empleados en la ciudad Santiago de Cali y 35 a nivel nacional, catalogada como empresa mediana, cada centro de

operación cuenta con ciertas características, patrones en ventas, tipos de producto vendidos, cantidades en gramos vendidos, promociones, descuentos, ubicación de local y cantidad de personal a cargo.

El sistema almacena todos los movimientos de cada venta, el módulo comercial, realiza las entradas y salidas de ítems, estas salidas se presentan en diferentes tipos, como kits, salidas en gramos, en onzas o presentaciones.

Cada producto es ofrecido en las siguientes presentaciones:

Unidad de medida	Gramos
1 onza	10
2 onzas	20
3.4 onzas	34
4.2 onzas	42
Cantidad en gramos	x

Tabla 1. Unidades de medida

A pesar de contar con datos históricos, la compañía carece de mecanismos analíticos robustos que permitan capturar la dinámica no lineal del comportamiento de ventas. Esto se traduce en previsiones imprecisas y pérdidas económicas asociadas al sobre inventario o desabastecimiento de inventario. El problema se agrava en contextos con patrones de ventas dispersos, discretos, estacionales o no lineales, donde los métodos tradicionales de predicción no ofrecen resultados satisfactorios.

2.1.Planteamiento del problema

El problema radica en que los sistemas ERP recopilan una gran cantidad de información, pero a menudo los datos relacionados con las transacciones de punto de venta están fragmentados o no estructurados. Esto dificulta su análisis y aprovechamiento efectivo para la toma de decisiones predictivas. Por lo tanto, es necesario encontrar una manera de identificar, clasificar y utilizar estos datos dispersos del módulo POS para generar modelos predictivos precisos que puedan anticipar eventos futuros y optimizar las operaciones del sistema ERP.

En efecto, una predicción es capaz de entregar una apreciación que busca estimar las tareas y el desempeño del trabajo, como ejemplo de ello el éxito de un estudio de mercado o la regulación del flujo de existencias en un punto de venta, no obstante, es posible evaluar diferentes áreas como la logística permitiendo observar la eficiencia de los procesos o los posibles fallos en el tiempo, en las ventas y el servicio al cliente se ayuda a establecer precios a los productos o servicios acordes a los cambios en los indicadores económicos, en consecuencia ofrecer novedades, precios especiales u obtener información de comportamiento de clientes en interacción con los nuevos productos.

Fecha	Valor	Unidades	Gramos	Pico
2/01/2022	490600	7	70	0
3/01/2022	813900	10	100	0
4/01/2022	1088500	17	170	0
5/01/2022	1048600	10	100	0
6/01/2022	2562770	65	650	1

Tabla 2. Salidas de producto en centros de operación

Fuente: El autor

El producto estrella de la compañía es la perfumería fina, este producto es distribuido en gramos, cada uno de sus proveedores ofrece cantidades a partir de 300 gramos y que pueden llegar a 10.000 gramos por referencia, de la Tabla 2 se obtienen datos del producto seleccionado, al valor

en ventas en el día, las unidades vendidas en presentación, la cantidad de gramos y si la fecha se cataloga como fecha especial, como fechas de pago, promociones o descuentos.

Conocer la dinámica de cada referencia es esencial para realizar el abastecimiento de cada centro de operación, de cada franquiciado y conocer la proyección que pueda tener un centro de operación.

En las finanzas se pretende controlar el flujo de compras, para conocer el estado actual que tiene una compañía, permitiendo conocer qué entidades realizan un mejor desempeño asegurando la disponibilidad que es fundamental en el manejo de la información.

En este contexto, surge la necesidad de aplicar técnicas más sofisticadas que permitan modelar la dinámica subyacente en los datos de ventas. El algoritmo SINDy, basado en identificación dispersa de sistemas dinámicos no lineales, representa una alternativa para modelar dichos comportamientos. No obstante, su implementación en el ámbito de predicción de ventas dentro de un ERP aún es incipiente y requiere ser evaluada en términos de precisión, aplicabilidad y beneficio operativo.

Por lo tanto, el problema central de esta investigación se puede formular en la siguiente pregunta:

¿Es posible predecir el comportamiento de las ventas de un sistema ERP POS empleando la identificación dispersa de los sistemas dinámicos no lineales para la planificación de los recursos empresariales?

2.2. Alcance del trabajo de grado

El presente documento de proyecto de grado tiene como objetivo principal presentar análisis predictivo del comportamiento de un sistema ERP para la planificación de recursos empresariales a partir de datos.

Al predecir posibles cuellos de botella o ineficiencias, se pueden tomar medidas preventivas para mejorar la eficiencia operativa y reducir costos, puede ayudar a prever la demanda futura de productos y materias primas. Esto permite mantener un nivel de inventario óptimo, evitando costosos excedentes o agotamientos de existencias, lo que mejora la eficiencia y la satisfacción del

cliente, como también ofrecer ofertas y servicios personalizados. Esto mejora la experiencia del cliente y aumenta la lealtad hacia la marca.

Un análisis predictivo es de gran ayuda para el desarrollo de herramientas de inteligencia artificial o el Big Data creadas por proveedores tecnológicos, servicios energéticos, compañías de seguros, la investigación médica, las finanzas, la fabricación, las cadenas de suministro, entre otros, pero una de las aplicaciones más beneficiosas de esta tecnología es analizar y predecir el comportamiento del cliente, logrando posicionarse frente al a competencia, desarrollar estrategias de márquetin, crear procesos eficientes, y alinear esfuerzos hacia un objetivo.

El proyecto de tesis tiene como objetivo investigar y desarrollar un sistema de predicción de ventas en un punto de venta utilizando datos del sistema POS. La tesis se centrará en evaluar la precisión, la robustez y la aplicabilidad práctica de estos métodos en el contexto de las ventas minoristas.

Se llevará a cabo una revisión exhaustiva de la literatura relacionada con la predicción de ventas en puntos de venta, incluyendo los avances más recientes en métodos de predicción y algoritmos. Se analizarán estudios previos que hayan utilizado algoritmos SINDY y se compararán con otras técnicas de predicción existentes.

Se recopilarán datos históricos detallados del sistema POS, que incluirán información sobre productos, precios, promociones, horarios, datos de clientes y otros factores relevantes para las ventas.

Se espera alojar un repositorio con una base de datos depurada y etiquetada como también las implementaciones necesarias de código en el lenguaje Python para la manipulación y ejecución de los datos que generan el modelo para la predicción de ventas a partir de los datos generados por el sistema ERP.

La tesis concluirá con un resumen de los hallazgos, destacando las fortalezas y debilidades de cada método evaluado. Se proporcionarán recomendaciones para futuras investigaciones en el campo de la predicción de ventas en puntos de venta y posibles aplicaciones prácticas para las empresas minoristas.

Este proyecto de tesis proporcionará una contribución significativa al campo de la predicción de ventas y otras variables de salida en puntos de venta al evaluar críticamente la eficacia del

algoritmo SINDY en comparación con métodos tradicionales. Al hacerlo, ayudará a los profesionales a comprender mejor qué técnicas son más adecuadas para predecir ventas u otras variables en entornos específicos, ofreciendo una base sólida para futuras investigaciones y aplicaciones prácticas en la industria.

3. Objetivos del proyecto

3.1. Objetivo general

Adaptar un método de identificación de la dinámica de sistemas no lineales dispersos para predecir el comportamiento de las ventas del módulo POS de un sistema ERP.

3.2. Objetivos específicos

El objetivo general se pretende alcanzar cuando se desarrollen los siguientes objetivos específicos:

- Comprender algunas técnicas de identificación de modelos para predecir el comportamiento de los datos de un sistema ERP.
- Estructurar la base de datos real que servirán para entrenar y evaluar el modelo de predicción.
- Aplicar el algoritmo SINDy para la predicción de los datos de un ERP empleando las ecuaciones que rigen este tipo de sistemas dinámicos.
- Valorar el resultado de la predicción del algoritmo SINDy comparado con los datos reales.

3.3. Resultados esperados

En esta investigación se busca realizar un análisis predictivo en una entidad que cuente con bases de datos centralizadas a optimizar sus estadísticas de diferentes departamentos para definir sus puntos débiles, posibles fallos y estimar sus beneficios. Así mismo, se espera obtener información acerca de cómo anticiparse a los acontecimientos futuros y preparar los procesos para mejorar su rendimiento, para ser aplicadas en proyectos futuros en cualquier campo. De forma análoga se pretende obtener:

- Descubrir modelos de sistemas dinámicos a partir de datos.
- Mejorar el desempeño industrial, de innovación e infraestructura.
- Desarrollo económico inclusivo y sostenido para impulsar el progreso.
- Definir variables de entradas – salida e internas – externas

4. Justificación

El uso de modelos de sistemas dinámicos a partir de datos puede ser implementado por empresas para analizar clientes, los investigadores de la salud utilizarlo en enfermedades como el Alzheimer o la Epilepsia, las agencias de publicidad y márquetin para dirigirse a los consumidores, los bancos para prevenir pérdidas monetarias o para contrarrestar el fraude.

Se pretende conocer en una referencia de producto, el comportamiento donde se pueda lograr una acertada asignación de abastecimiento en cada uno de los 365 días del año, la falta de este conocimiento aumenta el valor en los pedidos, el transporte, la logística, y genera pérdidas de clientes y ventas bajas.

Centro de operación	ID	REGIONAL	VENTAS
UNICO CALI	002	OC	$ 59,061,461
PALMETTO PLAZA CALI	003	OC	$ 19,385,327
UNICO DOS QUEBRADAS	004	OC	$ 38,491,731
CHIPICHAPE	005	OC	$ 23,359,172
UNICO 3 CALI	006	OC	$ 11,278,685
PASARELA	016	OC	$ 11,533,959
CC UNICO PASTO	019	OC	$ 22,716,302
PALMIRA CENTRO	021	OC	$ 12,657,072
YUMBO	025	OC	$ 14,476,830
PALMIRA LLANOGRANDE	029	OC	$ 9,105,845
TERMINAL CALI	030	OC	$ 30,265,476
POPAYAN CAMPANARIO	032	OC	$ 29,357,909
POPAYAN CENTRO	033	OC	$ 19,332,702
PEREIRA CENTRO	034	OC	$ 24,303,289
ALAMEDA	035	OC	$ 14,407,048
CALIMA	036	OC	$ 37,320,435
COSMOCENTRO	037	OC	$ 16,767,981
14 DE PEREIRA	038	OC	$ 9,719,723
FLORESTA	039	OC	$ 15,424,706
CARTAGO	040	OC	$ 13,910,029
LIMONAR	041	OC	$ 7,325,176
BUENAVENTURA 1 ÉXITO	042	OC	$ 34,050,319
JAMUNDI	043	OC	$ 16,254,126
CUBA	044	OC	$ 22,120,939
PASOANCHO	045	OC	$ 9,543,234
SAN ANDRESITO PEREIRA LAGO	047	OC	$ 18,399,469
BUENAVENTURA CENTRO	046	OC	$ 23,816,719
SAN NICOLAS	048	OC	$ 15,260,017

PASAJE CALI	049	OC	$ 12,443,665
JARDIN PLAZA	050	OC	$ 27,430,153
UNICENTRO	051	OC	$ 15,559,184
14 DE LA QUINTA	052	OC	$ 12,320,119
EXITO PASTO	056	OC	$ 7,555,451
ZAMORACO CALI	072	OC	$ 8,957,548
POPULAR	087	OC	$ 4,732,321
LA ESTACION	088	OC	$ 7,638,630

Tabla 3. Centros de operación y salidas en gramos primer trimestre zona centro occidente

Fuente: El autor

De la Tabla 3 se observa la diferencia de salidas en gramos y valores en ventas bruto por centro de operación de la zona occidente de la referencia a observar, conocer la dinámica de cada referencia o presentación en las ventas diarias es fundamental porque permite identificar patrones de comportamiento, tendencias de consumo y variaciones estacionales que influyen directamente en la toma de decisiones estratégicas. Este conocimiento facilita una mejor planificación de inventarios, optimización de la cadena de suministro y diseño de campañas de márquetin más efectivas. Además, ayuda a anticipar cambios en la demanda, detectar posibles anomalías o caídas en el desempeño del producto, y ajustar las operaciones comerciales en tiempo real para maximizar la rentabilidad y minimizar pérdidas.

La infraestructura básica, como las vías, las tecnologías de la información, la energía eléctrica y el agua, sigue siendo escasa en muchos países en desarrollo, el campo de la industrialización, la innovación y la infraestructura, logra ofrecer una economía dinámica y competitiva que generan mayores ingresos, disminuyendo el porcentaje de desempleo. Este análisis desempeña un rol importante a la hora de introducir y promover nuevas tecnologías, facilitar el comercio y permitir el uso eficiente de los recursos.

5. Marco referencial

5.1.Marco conceptual

5.1.1 Identificación de sistemas dinámicos

La identificación de sistemas dinámicos es el proceso para obtener modelos matemáticos o conjunto de ecuaciones que describen como funciona o evoluciona un sistema en el tiempo o en el espacio.

En términos generales, un sistema es un objeto en el que variables de diferentes tipos interactúan y producen señales observables. Las señales observables que nos interesan suelen llamarse salidas (*outputs*).

El sistema también se ve afectado por estímulos externos. Las señales externas que pueden ser manipuladas por el observador se denominan entradas (*inputs*).

Otras señales se conocen como perturbaciones que son señales externas que afectan al sistema, pero que no están bajo control del observador y pueden dividirse en aquellas que se miden directamente y aquellas que solo se observan a través de su influencia en la salida. La distinción entre entradas y perturbaciones medidas suele ser menos importante para el proceso de modelado.

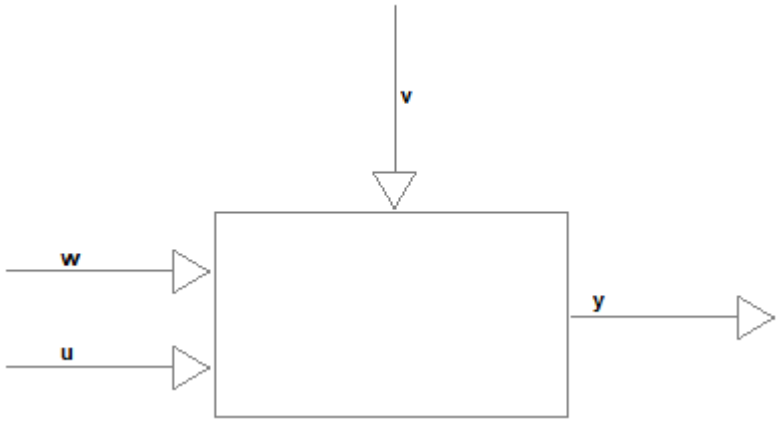


Figura 2. Sistema con salida y, entrada u, perturbación medida w y no medida v.

Fuente: Ljung L System Identification Theory for User-ed

5.1.2 Simulación en la identificación de sistemas

La simulación se usa para ver cómo respondería un sistema ante diferentes entradas. Es decir, se escoge una secuencia de entrada $u^*(t)$, y con base en un modelo del sistema, se calcula cómo sería la salida sin perturbaciones.

A. Salida sin perturbaciones

$$y^*(t) = G(q)u^*(t), t = 1,2, \dots, N \ (1)$$

$y^*(t)$: salida simulada del sistema, sin ruido o perturbaciones.

$u^*(t)$: secuencia de entrada elegida por el usuario.

$G(q)$: modelo del sistema, que puede ser una función de transferencia o un operador.

Esta ecuación muestra que, aplicando una entrada $u^*(t)$ el sistema genera una salida predicha $y^*(t)$ como si no existiera ninguna interferencia externa.

B. Evaluación de la perturbación (ruido)

Se introduce ruido o perturbación en el sistema para estudiar cómo afecta la salida:

1. Se genera una secuencia aleatoria $e^*(t)$ usando un generador de números aleatorios.
2. Luego se calcula la perturbación del sistema:

$$v^*(t) = H(q)e^*(t)$$

$v^*(t)$: perturbación simulada.

$H(q)$: modelo que transforma el ruido blanco en una perturbación estructurada.

Esto representa cómo las perturbaciones afectan realmente al sistema.

Al observar tanto la salida limpia $y^*(t)$ como la perturbación $v^*(t)$, el usuario puede entender mejor cómo responde el sistema ante una entrada determinada, tanto en condiciones ideales como en presencia de ruido.

5.1.3 Modelos para sistemas variables en el tiempo y no lineales

Si bien los modelos lineales e invariantes en el tiempo sin duda constituyen la forma más común de describir un sistema dinámico, también es bastante frecuente o necesario emplear otras descripciones como tratar los modelos con no linealidades en forma de elementos no lineales en la entrada y/o en la salida. También describir cómo manejar las no linealidades que pueden introducirse mediante transformaciones no lineales adecuadas de los datos.

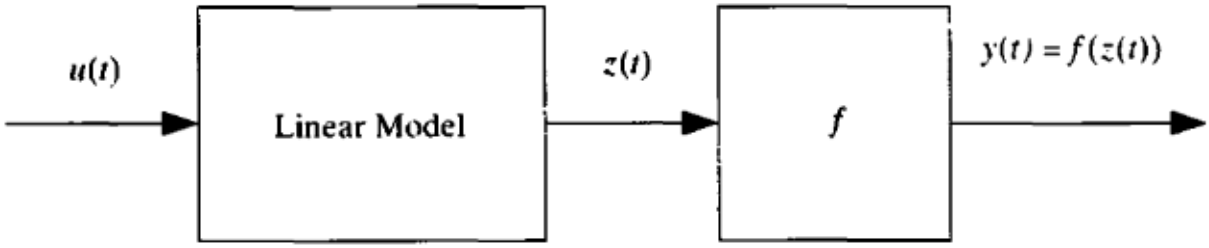


Figura 3. Modelo de Wiener.

Fuente: Ljung L System Identification Theory for User ed

La entrada $u(t)$ se procesa primero con un modelo lineal.

La salida intermedia $z(t)$ luego pasa por una función no lineal f.

Es común cuando el sistema es lineal pero la medición o percepción final es no lineal

5.1.4 Identificación de sistemas con SINDy

El método SINDy es una técnica de identificación de sistemas que busca descubrir modelos matemáticos a partir de datos obtenidos de sistemas dinámicos. Su enfoque se basa en la suposición de que, aunque la dinámica del sistema puede ser no lineal, ésta suele estar gobernada por un conjunto reducido de funciones. SINDy formula el problema como una regresión en la que se seleccionan únicamente aquellos términos funcionales que contribuyen significativamente a la evolución temporal del sistema. Este enfoque permite obtener ecuaciones diferenciales compactas y robustas, facilitando así la comprensión y simulación de sistemas complejos a partir de datos.

Debido a su capacidad para generar modelos, SINDy ha sido aplicado con éxito en diversas áreas, como sistemas biológicos, procesos físicos, ingeniería de control y análisis de series temporales no lineales.

5.1.5 Planeación de los recursos empresariales (ERP)

Un sistema ERP abarca una amplia gama de áreas funcionales, como finanzas, contabilidad, recursos humanos, gestión de la cadena de suministro, compras, ventas, inventario, producción, entre otras. Proporciona una base de datos centralizada y una arquitectura de software que permite a los diferentes departamentos y usuarios acceder y compartir información de manera eficiente.

Los sistemas ERP suelen tener módulos o componentes especializados que se adaptan a las necesidades específicas de cada organización. Estos módulos pueden incluir gestión financiera, gestión de recursos humanos, gestión de inventario, gestión de relaciones con los clientes (CRM), gestión de la cadena de suministro, entre otros.

Al utilizar un ERP, las organizaciones pueden mejorar la toma de decisiones, optimizar los flujos de trabajo, aumentar la productividad, reducir costos y mejorar la colaboración entre los diferentes departamentos. Además, un sistema ERP proporciona una base sólida para el análisis de datos y la generación de informes, lo que permite obtener información estratégica para el desarrollo y el crecimiento de la organización.

5.1.6 Punto de venta (POS)

El sistema POS consta de varios componentes, que incluyen una terminal de punto de venta (generalmente una computadora o una tableta), un lector de códigos de barras para escanear los productos, una impresora de recibos o tickets, un cajón de efectivo y, en algunos casos, un lector de tarjetas de crédito o débito.

El software del sistema POS se encarga de registrar y procesar las transacciones de venta, calcular el total de la compra, gestionar el inventario, generar recibos o tickets para los clientes y, en algunos casos, administrar otros aspectos del negocio, como el seguimiento de clientes y la generación de informes de ventas.

Los sistemas POS son ampliamente utilizados en diversos tipos de negocios, como tiendas minoristas, restaurantes, hoteles, supermercados y comercios en línea. Estos sistemas agilizan el proceso de venta, mejoran la precisión en los registros y facilitan la gestión de inventario, lo que permite a los comerciantes realizar un seguimiento de las ventas y los productos de manera eficiente.

Figura 4. Punto de venta ERP Sistemas de información SAS.

Fuente: https://www.s-i-e-s-a.com/gestion-para-el-sector-retail/

5.1.7 Almacenamiento de datos SQL Server

Una base de datos de SQL Server consta de una colección de tablas en las que se almacena un conjunto específico de datos estructurados. Una tabla contiene una colección de filas, también denominadas tuplas o registros, y columnas, también denominadas atributos. Cada columna de la tabla se ha diseñado para almacenar un determinado tipo de información; por ejemplo, fechas, nombres, importes en moneda o números [10].

5.1.8 Sector cosmético Senthia.

El posicionamiento de Senthia en Colombia se basa en su concepto de "Hágalo usted mismo" en perfumería y cosmética, ofreciendo productos de calidad para el cuidado personal y aromatización de espacios, con presencia en 67 ciudades y 100 puntos franquiciados. Senthia se destaca por su enfoque innovador y sensorial, transformando la experiencia cotidiana en momentos especiales [11]

5.2.Marco contextual

La compañía cuenta con más de 100 tiendas a nivel nacional, distribuidos en 3 zonas como se observa en la Figura 5, con más de 20 años de experiencia y procesos de alta calidad consolidándola como expertos en fragancias.

Figura 5. Zonas de puntos de venta a nivel nacional

Fuente: El autor

Se ha implementado el software ERP de la compañía multinacional Sistemas de información SAS desde el año 2017, el sistema para sector retail optimiza los procesos de compra, inventarios y ventas, el cual gestiona y controla los puntos de venta del negocio integrando actualizaciones de productos, precios, promociones y descuentos, y soportando la toma de decisiones gerenciales en tiempo real.

5.3.Marco tecnológico

El sistema ERP utilizado por la organización cuenta con un módulo POS que permite registrar ventas en tiempo real desde cada tienda. Esta información se almacena en una base de datos

relacional MS SQL [18], permitiendo acceder al histórico de ventas por producto, tienda y zona geográfica del sistema ERP y POS.

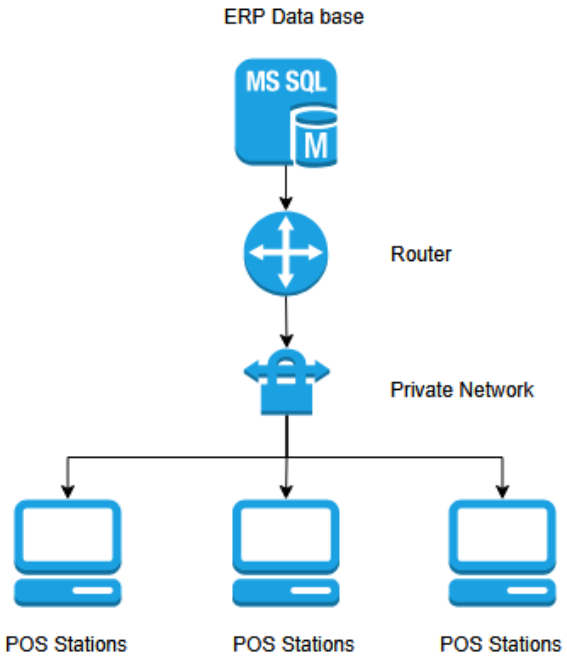


Figura 6. Almacenamiento de datos SQL para todos los puntos POS del ERP

Fuente: el autor

Para el procesamiento, modelado y simulación de datos se utilizó el lenguaje de programación Python 3.9 junto con bibliotecas científicas como Pandas, NumPy, y scikit-learn. La implementación del algoritmo SINDy se realizó a través de la librería PySINDy, la cual permite identificar ecuaciones diferenciales dispersas a partir de datos discretos.

La arquitectura de datos inicia con la recolección de transacciones desde los puntos de venta, las cuales son replicadas en un servidor central. A partir de esta base, se realiza la extracción y limpieza de datos utilizando scripts en Python [16], antes de aplicar el modelo de identificación SINDy.

A diferencia de modelos clásicos como las redes neuronales, SINDy permite descubrir las ecuaciones diferenciales que gobiernan el sistema de ventas, lo cual es útil cuando se analiza una dinámica no lineal dispersa entre regiones. Esta capacidad lo convierte en un modelo explicativo además de predictivo.

5.4.Marco teórico

5.4.1 SINDY

Comprender las restricciones dinámicas y los equilibrios en la naturaleza ha facilitado el rápido desarrollo del conocimiento y la tecnología, incluidos los aviones, los motores de combustión, los satélites y la energía eléctrica.

Debido a su capacidad para generar modelos, SINDy ha sido aplicado con éxito en diversas áreas, como sistemas biológicos, procesos físicos, ingeniería de control y análisis de series temporales no lineales.

Supongamos que tenemos un conjunto de mediciones x(t)e Rn de algún sistema físico en diferentes momentos **t**. SINDy busca representar la evolución temporal de x(t) en términos de una función no lineal f:

$$\frac{d}{dt}\, x(t) = f(x(t))$$

Esta ecuación constituye un sistema dinámico para las mediciones $x(t)$. El vector $x(t) = [x_1(t), x_2(t), \dots x_n(t)]^T$ da el estado del sistema físico en el tiempo t. La función $f\big(x(t)\big)$ limita cómo evoluciona el sistema en el tiempo.

La idea clave detrás de SINDy es que la función f a menudo es dispersa en el espacio de un conjunto apropiado de funciones de base. Por ejemplo, la función:

$$\frac{d}{dt}x = f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} 1 - x_1 + 3x_1x_2 \\ x_2^2 - 5x_1^3 \end{bmatrix}$$

es dispersa con respecto al conjunto de polinomios de dos variables en el sentido de que si escribiéramos una expansión de las funciones componentes f.

SINDy emplea una regresión dispersa para encontrar una combinación lineal de funciones base que capturen mejor el comportamiento dinámico del sistema físico, se desarrolla un marco novedoso para descubrir las ecuaciones gobernantes que subyacen a un sistema dinámico simplemente a partir de datos experimentales, aprovechando los avances en técnicas de dispersión y aprendizaje automático. Los modelos resultantes equilibran la complejidad del modelo con la capacidad descriptiva y evitan el sobreajuste. Hay muchos problemas críticos basados en datos,

como comprender la forma en que actúa el cerebro a partir de registros neuronales, deducir patrones climáticos, determinar la estabilidad de los mercados financieros, predecir y suprimir la propagación de enfermedades y controlar la turbulencia para un transporte y una energía renovable derivadas de fuentes naturales. Con grandes cantidades de datos, el descubrimiento de dinámicas basado en datos seguirá desempeñando un papel importante en estos esfuerzos.

Brunton. propuso un algoritmo para SINDy, que abordó el problema de descubrimiento de un sistema como una regresión dispersa y un problema de detección comprimida. Bajtiarnia, propuso una versión modificada para descubrir la dinámica de las redes sociales, eliminando la necesidad de ruido en el sistema para su identificación. Más tarde, el acercamiento se amplió a la Identificación de Parámetros de la Dinámica de Redes Complejas que combina el algoritmo SINDy modificado anterior como un acercamiento a los algoritmos genéticos.

Aprovechamos el hecho de que la mayoría de los sistemas físicos tienen solo unos pocos términos relevantes que definen la dinámica, lo que hace que las ecuaciones gobernantes sean escasas en un espacio de funciones no lineales de alta dimensión. La combinación de métodos de dispersión en sistemas dinámicos es bastante reciente. Aquí, consideramos sistemas dinámicos mediante la ecuación (1):

$$\frac{d}{dt}x(t) = f\big(x(t)\big) \tag{1}$$

Donde el vector $x(t) \in$ denota el estado de un sistema en el tiempo t, y la función $f(x(t))$ representa las restricciones dinámicas que definen las ecuaciones de movimiento del sistema, como la segunda ley de Newton. Posteriormente, la dinámica se generalizará para incluir la parametrización, la dependencia del tiempo y el forzamiento.

La observación clave es que, la función f consta de solo unos pocos términos, lo que la hace escasa en el espacio de funciones posibles. Los avances recientes en la regresión dispersa hacen que este punto de vista de la insuficiencia sea favorable, porque ahora es posible determinar qué términos son distintos de cero sin realizar una búsqueda de fuerza bruta, estos tipos de simulación estadística son técnicas cada vez más comunes para desarrollar modelos aproximados rápidos de manera que conservan la precisión, proporcionando límites de incertidumbre completos para la aproximación [7].

Para determinar la función f a partir de los datos, tomado de esta una recolección, una línea de tiempo x(t) y se mide la derivada x˙(t) o la aproximación numérica a partir de x(t). Los datos se muestrean varias veces *t1, t2…, tm* y se organizan como se muestra en las siguientes dos matrices:

$$X = \begin{bmatrix} x^T & t_1 \\ X^T & t_m \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_n(t_1) \\ x_1(t_m) & x_n(t_m) \end{bmatrix}$$

$$\dot{X} = \begin{bmatrix} \dot{x}^T & t_1 \\ \dot{x}^T & t_m \end{bmatrix} = \begin{bmatrix} \dot{x}_1(t_1) & x_n(t_1) \\ x_1(t_m) & x_n(t_m) \end{bmatrix}$$

A continuación, se procede a construir una biblioteca $\theta(X)$ que consta de funciones candidatas no lineales de las columnas de X. Por ejemplo, $\theta(X)$ puede constar de términos constantes, polinómicos y trigonométricos:

$$\theta(X) = [1 \quad X \quad X^{P_2} \quad X^{P_3} \quad \dots \quad \sin x \quad \cos x \quad \dots]$$

los polinomios superiores se denotan como, etc., donde denota las no linealidades cuadráticas en el estado x:

$$x^{P_2} = \begin{bmatrix} x_1^2(t_1) & x_1(t_1)x_2(t_1) & \dots & x_2^2(t_1) & \dots & x_n^2(t_1) \\ x_1^2(t_m) & x_1(t_m)x_2(t_m) & \dots & x_2^2(t_m) & \dots & x_n^2(t_m) \end{bmatrix}$$

Cada columna de $\theta(X)$ representa una función candidata y existe una flexibilidad para elegir las entradas en esta matriz no lineal. Debido a que observamos que solo unas pocas de estas no linealidades están activas en cada fila de, podemos establecer un problema de regresión dispersa para determinar los vectores de coeficientes.

$$\Xi = [\, \varepsilon_1 \;\; \varepsilon_2 \;\; \dots \;\; \varepsilon_n] \tag{2}$$

Ahora bien, se determinan qué no linealidades están activas:

$$\dot{X} = \theta(X)\Xi$$

Una vez que se ha determinado Ξ, se puede construir un modelo de cada fila de las ecuaciones gobernantes de la siguiente manera:

$$\dot{X_k} = f_k(x) = \Theta(x^t)\varepsilon_k$$

Tenga en cuenta que $\theta(x^t)$ es un vector de funciones simbólicas de elementos de X, a diferencia de $\theta(x)$, que es una matriz de datos. De este modo:

$$\dot{x} = f(x) = \Xi^t(\Theta(x^t))^t$$

El sistema de la Figura 7 representa el sistema de Lorenz que evoluciona de acuerdo a las ecuaciones subyacentes, tomando medidas para X, Y y Z para luego ser ensamblados en una matriz de datos X a partir de los datos de Z, Y, y Z del sistema de Lorenz.

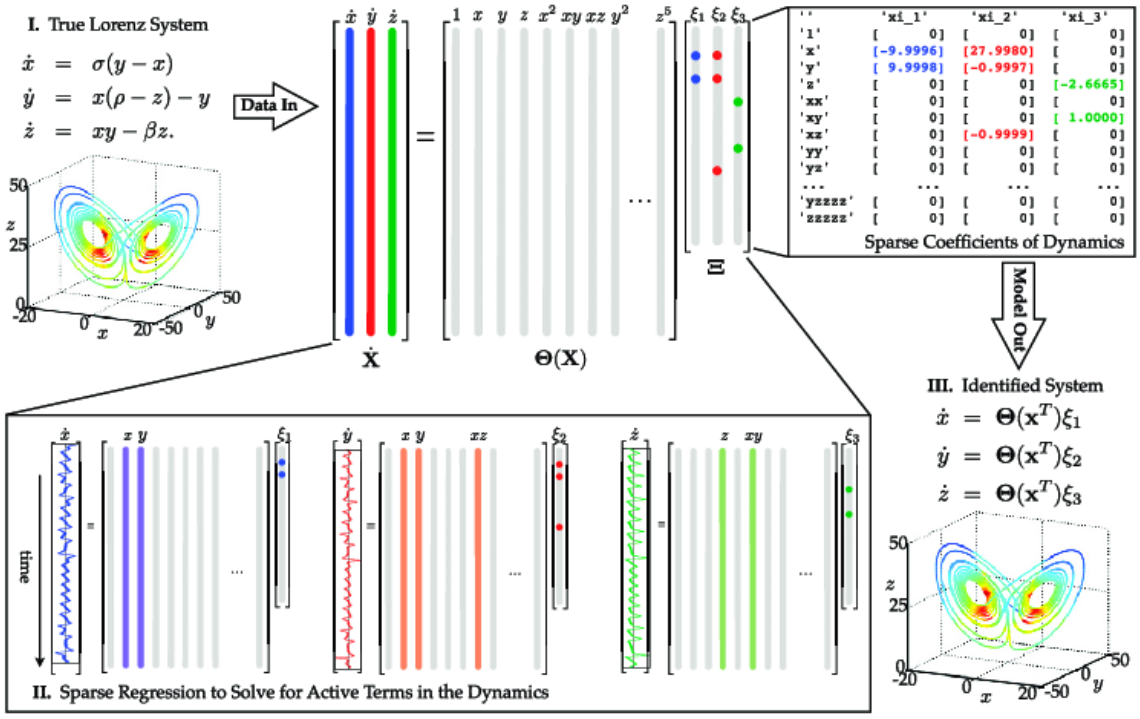


Figura 7. Esquema del algoritmo SINDy, demostrado en las ecuaciones de Lorenz

Fuente: Matos, Diego & Cunha Jr, Americo. (2019). A data-driven approach for inference of the evolution equation of a Duffing oscillator. 10.13140/RG.2.2.11999.20645.

Ahora bien, en la matriz θ(X) corresponde a la lista de términos candidato y la matriz Ξ una matriz de coeficientes que crean una cantidad de modelos posibles que pueden describir cómo evolucionan los datos en el tiempo y luego realizar una regresión dispersa para descubrir qué modelo se ajusta mejor a los datos como lo muestra la Figura 3. La ecuación 2 muestra el problema de optimización para ajustar los datos, agregando una matriz dispersa λΞ que intenta poner en cero

la mayor parte de coeficientes, siendo útil para obtener modelos dispersos y contrarrestar el sobre ajuste de datos.

$$\dot{x} = \theta(x)\Xi + nZ$$

donde Z se modela como una matriz de entradas gaussianas independientes idénticamente distribuidas con media cero y magnitud de ruido η. Así, buscamos una solución dispersa a un sistema sobre determinado con ruido. Con esta ecuación se resuelve un problema de optimización que solicita un conjunto de coeficientes que brinda el mejor modelo describiendo cómo evoluciona dinámicamente en el tiempo.

5.4.2 Aprendizaje automático

Para el análisis de los datos se utilizarán los módulos Scikit-learn de la biblioteca de Python para hacer aprendizaje automático.

Scikit-learn incluye algoritmos para la regresión, clasificación, división de datos de entrenamiento y prueba, métodos para evaluación y funciones de pre procesamiento de datos.

5.4.3 Técnicas de pre procesamiento

Existe un gran abanico de herramientas comunes para transformar datos crudos en representaciones más útiles para los modelos de aprendizaje automático. Muchos algoritmos se benefician de la estandarización como dejar los datos con media 0 y desviación estándar 1

Si hay outliers, se recomienda usar escaladores robustos u otros transformadores. Se pueden ver comparaciones gráficas del comportamiento de los distintos escaladores en situaciones con valores atípicos.

5.4.4 Clasificación

La clasificación es una de las tareas fundamentales en el aprendizaje automático supervisado, cuyo objetivo es asignar una etiqueta o clase a nuevas observaciones basándose en un conjunto de datos de entrenamiento previamente etiquetado. Este tipo de problemas se encuentra en una amplia gama de aplicaciones, como el reconocimiento de imágenes, la detección de fraudes, el diagnóstico médico o la segmentación de clientes. Para resolver estos problemas, se emplean algoritmos como árboles de decisión, máquinas de soporte vectorial (SVM), redes neuronales, entre otros, los cuales

construyen un modelo capaz de aprender los patrones que separan las distintas clases. La eficacia de un clasificador se evalúa comúnmente a través de métricas como la precisión, la sensibilidad o la matriz de confusión [8].

5.4.5 Regresión

La regresión es una técnica del aprendizaje automático supervisado que se utiliza para modelar la relación entre una variable dependiente continua o discreta y una o más variables independientes. Su propósito principal es predecir valores numéricos a partir de nuevas observaciones, basándose en un conjunto de datos de entrenamiento. Esta técnica es ampliamente usada en campos como la economía, la ingeniería y la ciencia de datos para estimar tendencias, costos, ventas o cualquier variable cuantitativa. Entre los modelos de regresión más conocidos se encuentran la regresión lineal, la regresión polinómica y los modelos de regresión basados en algoritmos más complejos como redes neuronales o bosques aleatorios. Una parte esencial del proceso de regresión es evaluar el desempeño del modelo usando métricas como el error cuadrático medio (MSE) o el coeficiente de determinación (R^2) [9].

5.4.6 Clustering

El clustering, o agrupamiento, es una técnica de aprendizaje automático no supervisado que tiene como objetivo organizar un conjunto de datos en grupos o clusters de manera que los objetos dentro de un mismo grupo sean más similares entre sí que con los de otros grupos. A diferencia del aprendizaje supervisado, el clustering no requiere etiquetas previas, lo que lo hace útil para descubrir estructuras ocultas, patrones o segmentos naturales en los datos. Este enfoque se emplea comúnmente en áreas como análisis de clientes, segmentación de mercado, biología computacional y detección de anomalías. Algunos de los algoritmos más conocidos para clustering incluyen K-means, DBSCAN y algoritmos jerárquicos. La elección del algoritmo y el número de grupos influyen significativamente en la calidad del agrupamiento [12].

5.4.7 Reducción de la dimensionalidad

La reducción de la dimensionalidad es una técnica fundamental en el aprendizaje automático y en el análisis de datos que consiste en transformar un conjunto de datos con muchas variables en una representación de menor dimensión que conserve la mayor cantidad posible de información

relevante. Esta técnica es especialmente útil cuando se trabaja con conjuntos de datos de alta dimensión, ya que ayuda a mitigar el efecto conocido como la «maldición de la dimensionalidad», mejora el rendimiento de los algoritmos, reduce el ruido y facilita la visualización de los datos.

5.4.8 Aprendizaje supervisado

El aprendizaje supervisado es una de las principales categorías del aprendizaje automático, y se basa en entrenar modelos utilizando un conjunto de datos etiquetado, es decir, datos que incluyen tanto las entradas como las salidas esperadas. Durante el proceso de entrenamiento, el modelo aprende a establecer una relación entre las características de entrada y las etiquetas de salida, con el fin de poder predecir correctamente nuevas instancias. Este tipo de aprendizaje es ampliamente utilizado en tareas de clasificación, donde se predicen categorías discretas, y de regresión, donde se predicen valores continuos. Su efectividad depende en gran medida de la calidad y cantidad de los datos de entrenamiento, así como de la correcta selección del modelo y sus hiperparámetros [14].

5.4.9 Aprendizaje no supervisado

El aprendizaje no supervisado es una rama del aprendizaje automático que se ocupa de descubrir patrones ocultos o estructuras subyacentes en datos que no están etiquetados, es decir, sin una salida conocida asociada a las entradas. A diferencia del aprendizaje supervisado, en este enfoque el algoritmo intenta organizar o segmentar los datos en grupos basados en similitudes, como ocurre en técnicas de *clustering* o análisis de componentes principales. El aprendizaje no supervisado es especialmente útil en exploración de datos, reducción de la dimensionalidad y compresión, y se emplea en áreas como la segmentación de clientes, la detección de anomalías y la minería de datos.

5.4.10 Métricas de desempeño

Las métricas de desempeño son fundamentales para evaluar la calidad de los modelos de aprendizaje automático, ya que permiten cuantificar su capacidad predictiva y orientar mejoras en su entrenamiento. La elección de la métrica adecuada depende del tipo de problema (clasificación, regresión, clustering, etc.). En clasificación, métricas comunes incluyen la precisión (*accuracy*), la exactitud (*precision*), la exhaustividad (*recall*) y la medida F1, mientras que en regresión se

emplean el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2). Estas métricas no solo permiten comparar modelos entre sí, sino también ajustar hiperparámetros y prevenir el sobreajuste (*overfitting*).

5.5. Estado del arte

El análisis predictivo se ha convertido en una herramienta poderosa en el ámbito empresarial para tomar decisiones informadas y anticiparse a eventos futuros. En particular, su aplicación en el contexto de sistemas de planificación de recursos empresariales (ERP) ha ganado importancia debido a la gran cantidad de datos generados por estos sistemas y la necesidad de utilizarlos de manera eficiente.

La identificación dispersa de datos del módulo de punto de venta (POS) es un enfoque novedoso para aprovechar los datos fragmentados o no estructurados provenientes de las transacciones de venta, esta identificación dinámica ha pasado por diferentes contribuciones donde existe un interés creciente por desarrollar modelos basados en la física y en los datos para establecer predicciones de propiedades, estructuras y procesos en materiales o actividades. Aunque los enfoques físicos disponibles pueden proporcionar predicciones razonablemente precisas, son computacionalmente costosos, lentos, complejos y requieren un alto conocimiento de campo para desarrollarlos y usarlos. Esto limita la aplicabilidad de dichos modelos en la industria y para fines como la optimización y el modelado en vivo del proceso. Los avances recientes en los enfoques de inteligencia artificial (IA) proporcionan una base para crear modelos basados en datos en este campo.

En muchos casos, existen bases naturales derivadas de las ecuaciones diferenciales, que promueven la escasez. Encontramos que nuestro método reduce con éxito la dinámica de las ecuaciones. Con el desarrollo de técnicas de aprendizaje automático, las metodologías de modelado se han convertido en una herramienta prometedora.

En este documento se observa la dinámica aproximada de varias ecuaciones diferenciales cuyas soluciones muestran comportamientos en múltiples escalas espaciales. Estas escalas pueden interactuar entre sí de manera no lineal a medida que evolucionan.[20]

El pronóstico de series de tiempo multivariadas se ha estudiado ampliamente a lo largo de los años con aplicaciones ubicuas en áreas como finanzas, tráfico o medio ambiente.

En la actualidad, el análisis financiero ha trascendido el uso tradicional de indicadores contables estáticos para incorporar técnicas avanzadas de predicción apoyadas en el aprendizaje automático. Estas herramientas permiten modelar patrones complejos y no lineales presentes en los datos financieros, mejorando la precisión en la estimación de riesgos, la proyección de ingresos y la toma de decisiones estratégicas. El aprendizaje automático se ha convertido en una herramienta esencial para procesar grandes volúmenes de datos, identificar relaciones ocultas y anticipar comportamientos del mercado que serían difíciles de detectar con métodos convencionales. Esto es especialmente relevante en entornos económicos volátiles, donde la capacidad de adaptación y predicción es crítica para la sostenibilidad financiera. Diversas investigaciones han demostrado que algoritmos como redes neuronales, máquinas de soporte vectorial y bosques aleatorios son efectivos en la predicción de variables financieras como precios de acciones, niveles de ventas y morosidad crediticia.

La investigación compara los resultados de predicción y los valores de error con la técnica sugerida, regresión multivariable, SVM y un método de identificación dispersa SINDy en conjuntos de datos sintéticos y del mundo real. La precisión de los pronósticos utilizando la técnica sugerida supera a la de los métodos a comparar. Por lo tanto, el enfoque propuesto puede utilizarse para resolver cuestiones prácticas que involucran la predicción del comportamiento de diversos procesos para los cuales se desconoce el modelo matemático pero los datos son accesibles en períodos discretos.

Aunque el análisis multivariable ha mejorado significativamente la precisión de las predicciones de ventas, sigue habiendo desafíos. La calidad y la disponibilidad de los datos, así como la interpretación de modelos complejos, son áreas que requieren atención continua. Además, las futuras investigaciones podrían explorar la integración de técnicas de inteligencia artificial, como el aprendizaje profundo, para mejorar aún más la precisión de las predicciones de ventas en puntos de venta.

La predicción de ventas en un punto de venta mediante análisis multivariable es un campo en constante evolución. Con el avance de las técnicas de análisis de datos y el desarrollo de nuevas herramientas tecnológicas, las empresas tienen a su disposición métodos cada vez más sofisticados

para prever y planificar sus ventas. La comprensión profunda de las complejas interrelaciones entre las variables es esencial para implementar modelos precisos y efectivos que impulsen el éxito comercial en el entorno competitivo actual.

6. Revisión de la literatura

La predicción de ventas ha sido una de las aplicaciones más exploradas en el campo de la inteligencia artificial (IA) y el aprendizaje automático, debido a su impacto directo en la planificación empresarial, la gestión de inventarios y la toma de decisiones estratégicas. Tradicionalmente, los métodos estadísticos como ARIMA, regresión lineal o modelos de suavizado exponencial han sido utilizados para predecir tendencias de ventas. Sin embargo, el auge de los enfoques basados en datos ha permitido el uso de algoritmos más sofisticados capaces de capturar relaciones no lineales y patrones complejos en grandes volúmenes de datos .

En el campo del aprendizaje automático, técnicas como árboles de decisión, máquinas de soporte vectorial (SVM), redes neuronales artificiales (ANN) y modelos basados en boosting (e.g., XGBoost) han demostrado ser eficaces para capturar la dinámica del comportamiento del consumidor y las fluctuaciones del mercado[18]. Estos modelos han sido aplicados a series temporales, datos de punto de venta, información macroeconómica y variables contextuales para generar predicciones con mayor precisión que los modelos tradicionales.

En años recientes, la atención se ha ampliado hacia el uso de algoritmos que permiten una identificación explícita de las ecuaciones dinámicas que rigen el sistema observado. En este contexto, el método SINDy (Sparse Identification of Nonlinear Dynamical Systems) ha emergido como una alternativa poderosa para identificar sistemas dinámicos dispersos a partir de datos observacionales. SINDy utiliza técnicas de regresión regularizada, como LASSO o STLSQ, para descubrir ecuaciones diferenciales que modelan la evolución de un sistema con un número reducido de términos, lo que resulta especialmente útil en escenarios donde se busca interpretabilidad y generalización.

Aplicado al campo de las ventas, el uso de SINDy permite no solo predecir el comportamiento futuro del sistema sino también entender las relaciones dinámicas entre variables como unidades vendidas, precios, estacionalidad, canales de distribución y comportamiento del consumidor. A diferencia de modelos puramente predictivos, SINDy aporta un marco híbrido entre identificación de sistemas y modelado basado en datos, lo que facilita la simulación de escenarios y la validación de hipótesis estructurales [17].

Estudios recientes han explorado la integración de SINDy con redes neuronales recurrentes (RNN) y técnicas de aprendizaje profundo para abordar dinámicas altamente no lineales en series temporales de ventas. También se han evaluado extensiones como SINDy-PI, que incorporan funciones implícitas y operadores de retardo para modelar sistemas con efectos diferidos, una característica común en contextos comerciales donde las acciones de marketing o cambios en el precio tienen efectos que no son inmediatos.

En resumen, la literatura indica una evolución desde métodos lineales tradicionales hacia enfoques híbridos que combinan IA, aprendizaje automático y modelos de identificación dinámica dispersa, destacando el potencial de herramientas como SINDy para ofrecer no solo precisión en la predicción, sino también transparencia y comprensión del sistema de ventas.

7. Antecedentes

En entornos comerciales, algunos estudios han comenzado a aplicar SINDy y otras técnicas de identificación dispersa para modelar dinámicas de comportamiento de consumidores, flujos de caja y decisiones de compra. Sin embargo, aún existe una escasa implementación en el ámbito específico de la predicción de ventas a partir de datos transaccionales reales, lo cual plantea una oportunidad relevante para explorar el potencial de estos métodos en contextos empresariales con alta complejidad no lineal.

Estos antecedentes evidencian una evolución progresiva desde enfoques puramente estadísticos hacia modelos híbridos basados en datos, que integran la interoperabilidad de los sistemas dinámicos con el poder predictivo de la inteligencia artificial. Así, el presente trabajo se inserta dentro de esta línea emergente de investigación, proponiendo el uso del algoritmo SINDy como una herramienta para modelar y predecir dinámicamente el comportamiento de las ventas en sistemas empresariales reales.

En el campo financiero, los antecedentes en la predicción de datos como ventas o comportamientos dinámicos se basan en la evolución desde los modelos clásicos de series temporales hacia enfoques más avanzados basados en inteligencia artificial. La gestión del riesgo financiero ha incorporado técnicas de aprendizaje automático y profundo en Python para evaluar riesgos de mercado, crédito y liquidez, mejorando la precisión frente a los métodos tradicionales. Estas herramientas permiten modelar la volatilidad, estimar riesgos mediante enfoques bayesianos y detectar fraudes mediante modelos de aprendizaje, reflejando una tendencia hacia la automatización y el uso de algoritmos para comprender patrones complejos en los datos financieros.

Con la llegada del aprendizaje automático y el aprendizaje profundo, los analistas pueden modelar comportamientos no lineales, identificar patrones ocultos y anticipar tendencias de manera más precisa. Libros como Foundations of Artificial Intelligence in Finance: Insights for Practitioners with Applications and Case Studies [21] destacan cómo la IA permite analizar grandes volúmenes de información, desde documentos financieros hasta movimientos bursátiles, mediante técnicas de análisis predictivo y procesamiento de lenguaje natural. Además, abordan la importancia de la ética, la sostenibilidad y la regulación en la aplicación de estos modelos, factores esenciales para garantizar decisiones financieras responsables.

8. Metodología

8.1. Metodología de trabajo propuesta

Para aplicar el algoritmo de predicción SINDy (Sparse Identification of Nonlinear Dynamical Systems) utilizando datos de una base de datos SQL Server proveniente de un sistema ERP POS, se considera estructurar el trabajo bajo una metodología de desarrollo basada en CRISP-DM que son las siglas de Cross-Industry Standard Process for Data Mining**,** consiste principalmente en la recolección de datos que posteriormente deben pasar por 6 fases que comprenden el análisis, organización y clasificación.

8.2. Metodología CRISP-DM

La metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) es uno de los marcos más ampliamente adoptados para llevar a cabo proyectos de minería de datos y ciencia de datos.

CRISP-DM se compone de seis fases principales mencionadas a continuación, donde cada una orienta desde la comprensión del problema hasta la entrega del modelo final.

8.3. Modelo de negocio

En esta fase se define los objetivos del negocio y los requisitos empresariales para la compañía en el manejo de inventarios, flujos y ventas de cada punto de venta en los centros de operación a nivel nacional.

A. Estudiar y definir los centros de operación donde se extraerán los datos
B. Seleccionar cada uno de los productos a ser evaluados tomando la referencia, unidades vendidas, gramos y presentaciones disponibles.

8.4. Clasificación de los datos

A. Clasificación y análisis de técnicas

Organizar las técnicas identificadas en categorías que permitan compararlas y entender sus diferencias como las basadas en datos (data-driven) y la capacidad para manejar datos dispersos o con ruido.

B. Profundización en SINDy (Sparse Identification of Nonlinear Dynamical Systems).

Con la documentación oficial, realizar la consulta a los fundamentos teóricos del método, cómo identifica dinámicas a partir de datos dispersos, verificar algunos ejemplos de aplicación en otras áreas como la física, biología y economía, finalmente, observar las limitaciones y condiciones de aplicabilidad.

C. Evaluación de la viabilidad de aplicación al ERP

Se analizan si la técnica estudiada SINDy, es viables para aplicarse a los datos del sistema ERP, en este aspecto se evalúa si son suficientes los datos disponibles, qué características tienen (frecuencia, ruido, faltantes), se requieren transformación o pre procesamiento especial lo cual permitirá justificar la elección final de este método en la tesis, documentando todo el proceso y realizando una reflexión crítica que justifique las técnicas a utilizar.

Recolección inicial de datos, familiarización, identificación de problemas de calidad y descubrimiento de primeros patrones

8.5. Preparación de los datos

D. Clasificación y análisis de técnicas

Organizar las técnicas identificadas en categorías que permitan compararlas y entender sus diferencias como las basadas en datos (Data-Driven) y la capacidad para manejar datos dispersos o con ruido.

E. Profundización en SINDy (Sparse Identification of Nonlinear Dynamical Systems).

Con la documentación oficial, realizar la consulta a los fundamentos teóricos del método, cómo identifica dinámicas a partir de datos dispersos, verificar algunos ejemplos de aplicación en otras áreas como la física, biología y economía, finalmente, observar las limitaciones y condiciones de aplicabilidad.

F. Evaluación de la viabilidad de aplicación al ERP

Se analizan si la técnica estudiada SINDy, es viables para aplicarse a los datos del sistema ERP, en este aspecto se evalúa si son suficientes los datos disponibles, qué características tienen (frecuencia, ruido, faltantes), se requieren transformación o pre-procesamiento especial lo cual

permitirá justificar la elección final de este método en la tesis, documentando todo el proceso y realizando una reflexión crítica que justifique las técnicas a utilizar.

Recolección inicial de datos, familiarización, identificación de problemas de calidad y descubrimiento de primeros patrones

8.6. Modelado

Aplicar el algoritmo SINDy en un entorno ERP permite identificar las ecuaciones que rigen la dinámica interna de los datos transaccionales del sistema, facilitando así la predicción de comportamientos futuros. Este enfoque combina el análisis basado en datos con principios físicos, extrayendo modelos dispersos que describen con precisión la evolución de variables críticas como ventas, inventario o demanda. Al modelar estas dinámicas mediante ecuaciones diferenciales, SINDy ofrece una herramienta robusta para la toma de decisiones informadas en tiempo real. Su capacidad para evitar el sobreajuste lo hace ideal para entornos complejos y variables como los ERP.

8.7. Evaluación

Valorar el resultado de la predicción del algoritmo SINDy implica comparar sus salidas con los datos reales del sistema ERP para evaluar la precisión del modelo y que este cumple con los objetivos del negocio. Esta comparación permite identificar qué tan bien las ecuaciones descubiertas representan la dinámica del proceso. Al analizar métricas de error como el RMSE o el MAE, se puede cuantificar el grado de ajuste entre la predicción y la realidad. Un buen desempeño validaría la capacidad del modelo para generalizar. En caso contrario, se podrían ajustar los términos no lineales o mejorar la calidad de los datos

A. Definición de Métricas de Evaluación

Se deben seleccionar métricas cuantitativas que permitan comparar las predicciones generadas por el modelo SINDy con los valores reales de ventas obtenidos del sistema ERP, en este caso se implementará l métrica de coeficiente de Determinación *R2* para evaluar el ajuste del modelo, estas métricas se aplican tanto al conjunto de entrenamiento como al de validación/prueba.

B. Visualización de Resultados

El análisis gráfico también es fundamental para evaluar el rendimiento del modelo. Se recomienda realizar gráficas de línea temporal donde se comparen visualmente las ventas reales con las predichas por el modelo, facilitando la detección de desviaciones significativas. Asimismo, las gráficas de dispersión son útiles para identificar qué tan cerca están las predicciones del comportamiento ideal. Este tipo de análisis va a permitir diagnosticar errores o sesgos en el ajuste.

Para ello se realizan las gráficas de línea temporal comparando las ventas reales y las predichas, estas visualizaciones, se pueden realizar mediante las librerías del extendido abanico de Python como Matplotlib, Seaborn y Plotly.

Se emplea la librería scikit-learn para obtener funciones auxiliares que ayuden a generar visualizaciones asociadas a el modelo de regresión y predicción.

C. Análisis cuantitativo

Se procede a realizar un análisis del comportamiento del modelo en distintos escenarios de negocio evaluando si el modelo predice con un gran porcentaje de aceptación en temporadas altas o solo en patrones regulares, se evalúa cómo se comporta con promociones, fechas especiales o eventos aleatorios no predefinidos y qué variables del sistema afectan más la predicción.

Con el análisis de inicia una etapa de interpretar los resultados donde se observa qué tan precisas son las predicciones en términos de utilidad empresarial, qué limitaciones tiene el modelo SINDy para este tipo de información extraída de un sistema ERP, y cómo se compararía con otros métodos más comunes o de mayor uso en el ámbito financiero

9. Implementación

9.1. Comprensión modelo de negocio

9.1.1 Compañía distribuidora de cosmética Senthia

Centro de operación	Regional	Centro de operación	Regional
ALSACIA	CO	MERCURIO	CO
PIEDECUESTA	CO	UIS BUCARAMANGA	CO
SOACHA	CO	ESPINAL	CO
IBAGUE	CO	PLAZA IMPERIAL	CO
PASTO 3 CENTRO	CO	AV CHILE	CO
KENNEDY 2	CO	CENTRO CRA 7	CO
TUNJUELITO	CO	SALITRE	CO
SANTA ISABEL	CO	CASTILLA	CO
VILL PUERTAS DEL SOL	CO	SANTA LUCIA NEIVA	CO
VILL LLANOCENTRO	CO	PORTAL DE LA 80	CO
CANAVERAL F.BLANCA	CO	MONACO 2	CO
CACIQUE BUCARAMANGA	CO	VENECIA 2	CO
UNICO NEIVA	CO	ANTARES	CO
CENTRO MONACO	CO	GUATAPURI	CO
TITAN	CO		

Tabla 4. Centros de operación zona centro oriente

Fuente: El autor

Centro de operación	Regional	Centro de operación	Regional
PRINCIPAL	OC	FLORESTA	OC
UNICO CALI	OC	CARTAGO	OC
PALMETTO PLAZA CALI	OC	LIMONAR	OC
UNICO DOS QUEBRA-	OC	BUENAVENTURA 1	OC
CHIPICHAPE	OC	JAMUNDI	OC
UNICO 3 CALI	OC	CUBA	OC
PASARELA	OC	PASOANCHO	OC
CC UNICO PASTO	OC	SAN ANDRESITO PE-	OC
PALMIRA CENTRO	OC	BUENAVENTURA CEN-	OC
YUMBO	OC	SAN NICOLAS	OC
PALMIRA LLANO-	OC	PASAJE CALI	OC
TERMINAL CALI	OC	JARDIN PLAZA	OC
POPAYAN CAMPANA-	OC	UNICENTRO	OC
POPAYAN CENTRO	OC	14 DE LA QUINTA	OC
PEREIRA CENTRO	OC	EXITO PASTO	OC
ALAMEDA	OC	ZAMORACO CALI	OC
CALIMA	OC	POPULAR	OC
COSMOCENTRO	OC	LA ESTACION	OC
14 DE PEREIRA	OC		

Tabla 5. Centros de operación zona occidente

Fuente: El autor

Centro de operación	Regional	Centro de operación	Regional
UNICO BARRANQUILLA	NT	SANTA MARTA OCEAN	NT
BARANQUILLA 2 FLO-	NT	MONTERIA 2 CC BUENA-	NT
BARRANQUILLA 3	NT	MAYALES VALLEDUPAR	NT
MONTERIA 1 CENTRO	NT	BUENAVISTA S. MARTA	NT
CARTAGENA	NT	ALAMEDAS	NT

Tabla 6. Centros de operación zona norte

Fuente: El autor

Cuenta con 196 productos con las siguientes categorías

Categorías	Categorías
AMBARADAS FEMENINAS	KITS
AMBARADAS MASCULINAS	MADEROSAS MASCULINAS
CÍTRICAS Y ACUÁTICAS	FRESCAS Y CÍTRICAS
FLORALES	FRUTALES
FRESCAS	GOURMAND

Tabla 7. Categorías de productos

Fuente: El autor

9.1.2 Centros de operación y puntos de venta

El desarrollo del proyecto se llevará a cabo en las bodegas de la compañía Essen-sale S.A.S ubicada en la Calle 41 # 6-16, Bodega 8 Parque Industrial La Esmeralda Barrio Las Delicias, Cali, Valle del Cauca.

Figura 8. Essen-sale, Sen-thia. (2024). Fotografía de Bodega principal

Recuperado de https://www.google.com/maps/

9.1.3 Productos para estudio

La empresa Essen-sale S.A.S dedicada al comercio al por menor de productos farmacéuticos y medicinales cosméticos y artículos de tocador en establecimientos especializados, con sede en Santiago de Cali incluida en las 500 empresas más importantes del Valle del cauca, contando con una amplia gama de productos que van desde el cuidado personal como la perfumería, cremas, splash, hasta elementos para el hogar como ambientadores y mejoras de interiores.

Producto	Presentación	Zonas	Periodos
Carmino	1 onza	OC	2022
	2 onzas	CO	2023
	3.4 onzas	NT	2024
	4.2 onzas		

Tabla 8. Selección de producto para modelo

Fuente: autor

9.2 Clasificación de los datos

A partir de la visualización del conjunto de datos cargado, donde se observan variables como la fecha del documento, el valor neto de la venta, las unidades vendidas, Cantidad en gramos y una variable categórica que representa umbrales comerciales, se pretende realizar una clasificación de los datos orientada a su uso en modelos de predicción de ventas mediante el algoritmo SINDy (Sparse Identification of Nonlinear Dynamical Systems)**.**

Variable	Cualitativa	Cuantitativa	Indep.	Dep.	Binaria
Venta neta		✓ Cont.		✓	
Gramos		✓ Disc.		✓	
Unidades		✓ Disc.		✓	
Ventas NT		✓ Disc.		✓	
Ventas OC		✓ Disc.		✓	
Ventas CO		✓ Disc.		✓	
Inventario		✓ Disc.		✓	
Fecha Esp.			✓		✓
Fin Semana	✓		✓		✓
CO	✓		✓		
Publicidad	✓		✓		✓
Producto			✓		
IPC		✓ Cont.		✓	
Fecha Doct.	✓		✓		
Zona	✓		✓		

Tabla 9. Identificación de datos

Fuente: El autor.

En el proceso de clasificación de datos para la predicción de ventas utilizando el algoritmo SINDy es fundamental identificar variables que actúen como disparadores o moduladores de comportamiento no lineal en el sistema. En este conjunto de datos, la variable binaria ha sido definida como una variable binaria que indica la presencia de fechas especiales o eventos comerciales significativos**,** tales como días sin IVA, fechas de pago, promociones, y celebraciones como el Día del Niño, de la Madre, del Padre o la temporada navideña. Estas fechas suelen generar picos en las ventas, rompiendo la linealidad y estacionalidad habitual de los datos. Al clasificar las observaciones según esta variable, se facilita la segmentación del sistema dinámico en condiciones normales y condiciones especiales, permitiendo que SINDy identifique de forma más precisa las ecuaciones que rigen la evolución del sistema en cada régimen. Esta clasificación no solo mejora la capacidad predictiva del modelo, sino que también aporta comprensión, ya que permite simular y analizar cómo los eventos externos o campañas específicas afectan dinámicamente las ventas a lo largo del tiempo.

9.2.1 Categorización de las tablas de datos y registros.

Se identifica el rol de cada tabla o campo, donde se procede a identificar tablas que contienen productos, ventas, inventarios, descuentos y reconocer si un campo es variable categórica, continua, discreta, booleana, cualitativa y cuantitativa, como lo muestra la tabla19

Se realizan agrupaciones por zona, producto, y las presentaciones, tomando un solo valor respectivamente.

La agrupación de la fecha del documento se realiza por día y se crea la variable dummy fechas especiales para los picos en las ventas.

Se define la variable de salida dependiente como el total de las ventas netas del producto en todas las zonas.

Campo	Categoría	Tipo
Producto	Producto	Categórica nominal
Género	Identificador	Categórica nominal
Zona	Geográfica	Categórica nominal
Unidades vendidas	Métrica de negocio	Numérica discreta
Vena neta	Métrica de negocio	Numérica continua
Fechas Especiales	Estado comercial	Booleana (binaria)
Picos	Estado comercial	Categórica ordinal (alto/medio/bajo)

Tabla 10. Categorización de los datos

Fuente: El autor

9.2.2 Selección de referencias y presentaciones.

Se selecciona de la base de datos del ERP – POS el producto Carmino de 4 centros de operación en las zonas centro oriente, occidente y norte.

9.2.3 Reconocimiento de patrones, anomalías o errores.

Se detectar tendencias, repeticiones o comportamientos frecuentes en los datos, para el caso de las ventas se detectan datos atípicos con valores muy altos fuera del promedio generados por facturas realizadas manualmente en procesos de migración o caídas de sistema, se considera no incluir estos datos

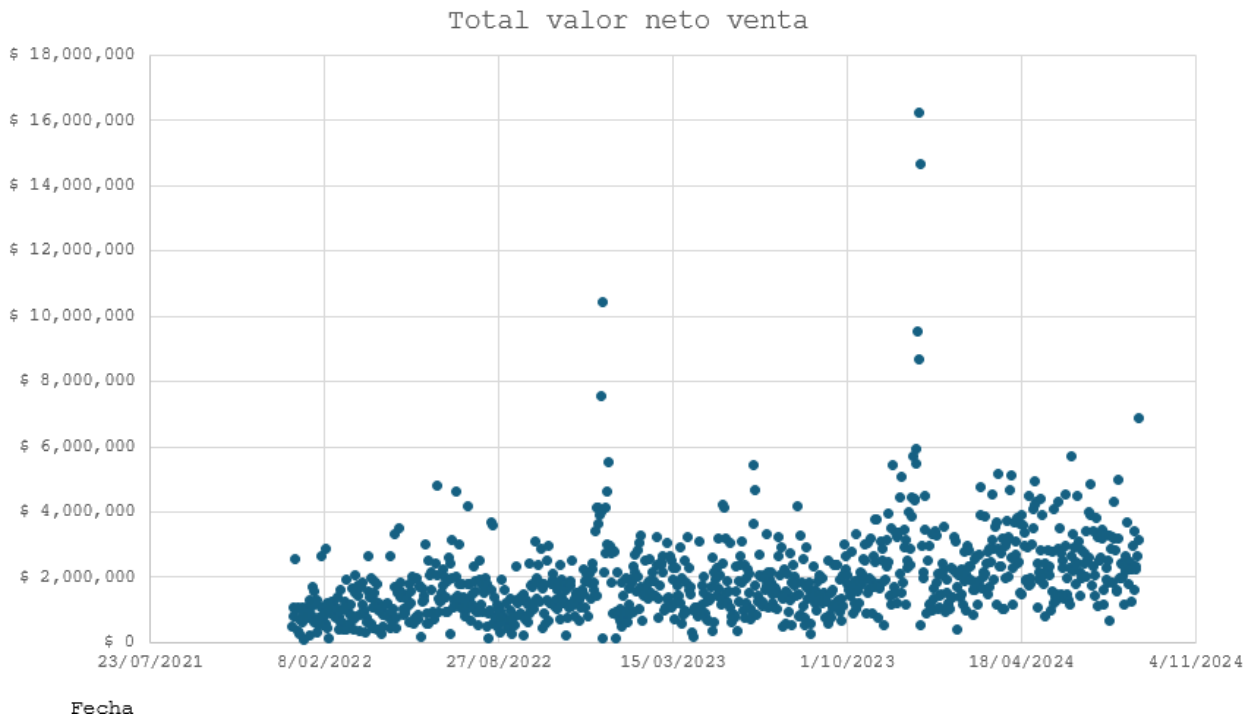


Figura 9. Datos exploratorios Venta neta Vs Fecha diaria

Fuente: El autor.

El análisis exploratorio de la Figura 9 presenta un patrón con picos altos en fechas específicas, destacando particularmente los días entre el 20 de diciembre de 2023 y 20 de diciembre de 2024 hasta su respectivo fin de mes, que probablemente corresponden a incrementos estacionales vinculados a las compras navideñas. Además, se observan otros picos significativos a mitad de año, que podrían estar relacionados con eventos promocionales, temporadas escolares o días sin IVA.

La mayoría de los puntos de venta se concentran en rangos de ventas inferiores a 2 millones, lo que sugiere que el flujo de ventas diario es relativamente estable, con valores cambiantes solo en ciertas fechas clave. Estos picos atípicos representan variaciones que podrían ser explicadas por campañas comerciales, factores socioeconómicos o comportamiento de compra estacional, y deben ser incorporados en el modelado predictivo como variables de eventos especiales para mejorar la precisión.

Se detecta el patrón de fechas especiales y quincenas, para la elaboración de las variables de control binarias.

9.2.4 Análisis de variables objetivo y predictores.

En el proceso de examinar las variables objetivo se establece como salida Y el neto de las ventas, para el caso de los predictores se consideran las unidades vendidas, la variable binaria de control, fecha de pago y los picos en ventas de carácter ordinal, influencia de las unidades, picos y fechas especiales en la variable objetivo venta neta.

Matriz de correlación.

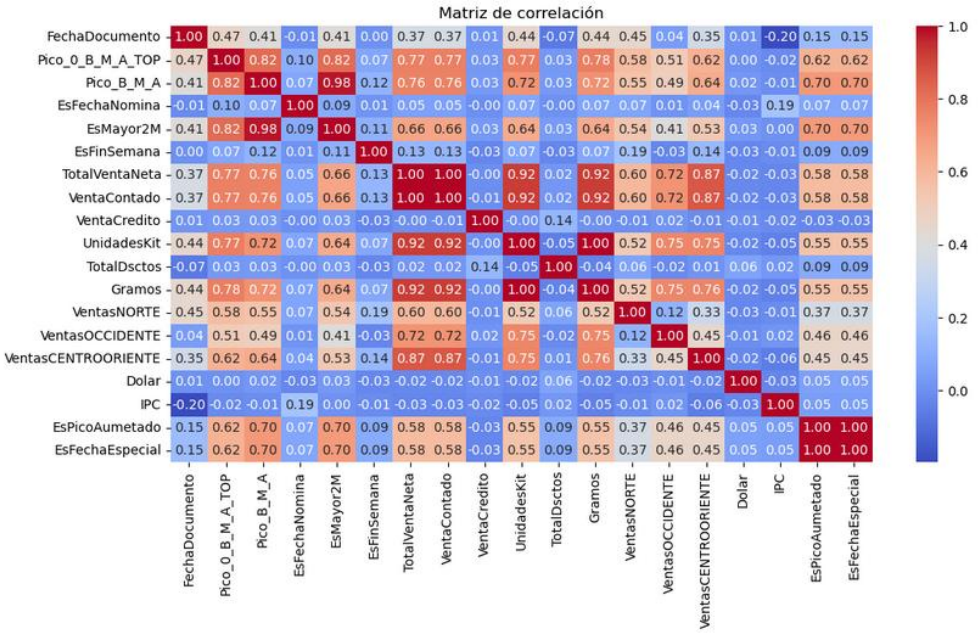


Figura 10. Matriz de correlación

Fuente: El autor

La Figura 10. Presenta las variables relacionadas con ventas diarias y condiciones externas, con relaciones fuertes y positivas en:

Variables	r	Relación
TotalVentaNeta		Variable predictora
UnidadesKit	0.92	A más kits vendidos, más ventas de contado
VentasOCCIDENTE	0.72	La venta oc a contribuye fuertemente a la venta neta
VentasCENTROORIENTE	0.87	La venta co a contribuye fuertemente a la venta neta
VentasNORTE	0.60	La venta nt a contribuye fuertemente a la venta neta
Pico_B_M_A	0.76	Variable temporal que representa los picos medios altos y bajos
EsFechaEspecial	0.58	Variable temporal que representa los picos fechas especiales

Tabla 11. Variables y correlación

Fuente: El autor

Graficas de Boxplot.

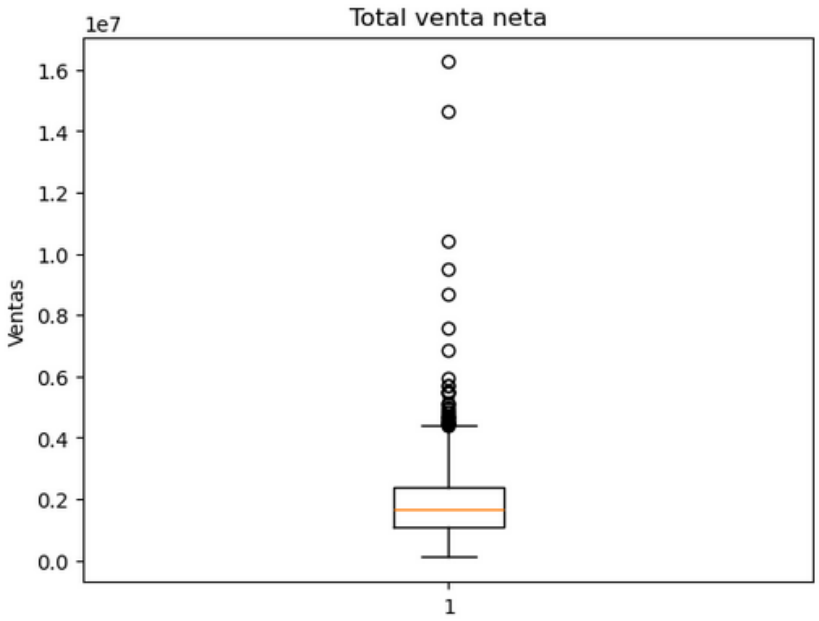


Figura 11. Boxplot Venta neta total

Fuente: El autor

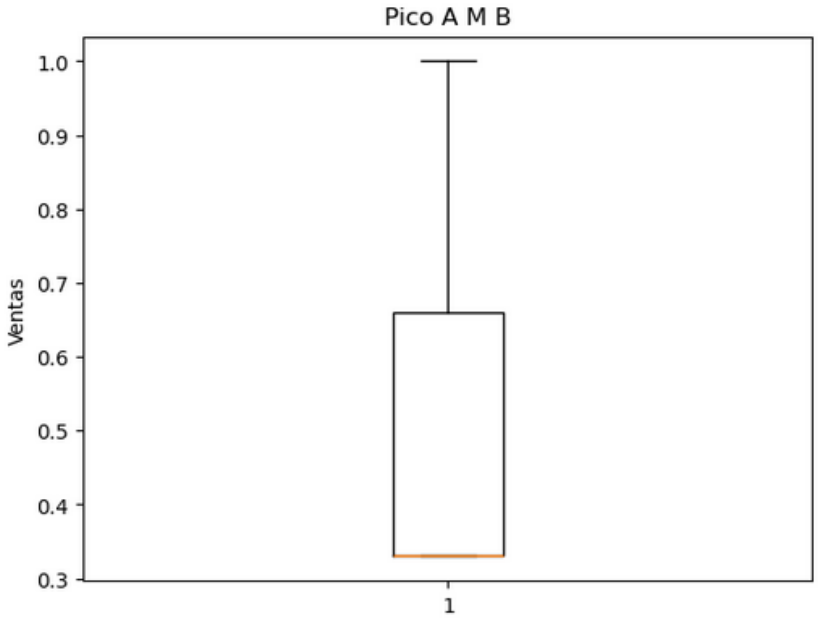


Figura 12. Boxplot Venta neta total

Fuente: El autor

De la Figura 12 Distribución de ventas la sección naranja representa el rango intercuartílico (IQR), es decir, qué tan esparcidos están los datos en torno a la mediana, donde está el 50 % central de las ventas.

La línea dentro de la caja es la mediana la cual gravita alrededor de 1.8 millones.

El mínimo típico (bigote inferior) está por encima de 0, y el máximo típico (bigote superior) ronda los 2.5 millones.

Outliers *(valores atípicos)*

Los círculos por encima del bigote superior son valores que se salen del rango normal.

Hay bastantes outliers y algunos son muy extremos (uno supera los 16 millones).

Esto indica que, aunque la mayoría de ventas netas están en un rango más bajo, ocasionalmente hay transacciones muy grandes.

Sesgo

El bigote superior es mucho más largo y con más puntos extremos que el inferior esto indica asimetría positiva indicado por la cola larga hacia la derecha.

Esto es común en datos de ventas, donde la mayoría de transacciones son pequeñas o medianas, pero unas pocas son muy grandes.

9.3 Preparación de los datos

9.3.1 Recolección de los datos

A partir de una consulta SQL Server se obtiene los datos registrados del sistema POS ERP, como herramienta se ejecuta el script en Microsoft SQL Server Management.

Consulta SQL para productos compañía Sen - thia

```
SELECT
        f9920_id_fecha_docto,
        sum(f9920_valor_neto) as totalValorNetoVenta,
        sum(f9920_ventas_contado) as f9920_ventas_contado,
        sum(f9920_ventas_credito) as f9920_ventas_credito,
        count(f120_referencia) as Unidades_Vendidas,
```

```sql
        sum(f9931_vlr_tot) as TotalDsctos,
        sum(f9930_cant_1 * f9933_cant_1) as gramos ,
        SUM(CASE WHEN f285_id_regional = 'NT' THEN f9920_valor_neto ELSE 0 END) AS VentasNORTE,
        SUM(CASE WHEN f285_id_regional = 'OC' THEN f9920_valor_neto ELSE 0 END) AS VentasOCCIDENTE,
        SUM(CASE WHEN f285_id_regional = 'CO' THEN f9920_valor_neto ELSE 0 END) AS VentasCENTROORIENTE
FROM
        t9910_pdv_a_control_tpv
INNER JOIN t9920_pdv_a_doctos ON
        t9910_pdv_a_control_tpv.f9910_guid = t9920_pdv_a_doctos.f9920_guid_control_tpv
INNER JOIN t285_co_centro_op on
        t285_co_centro_op.f285_id = t9920_pdv_a_doctos.f9920_id_co
        AND t9920_pdv_a_doctos.f9920_id_cia = 1
INNER JOIN t9930_pdv_a_movto_venta ON
        t9920_pdv_a_doctos.f9920_guid = t9930_pdv_a_movto_venta.f9930_guid_docto
INNER JOIN t9933_pdv_a_movto_venta_prep ON
        t9930_pdv_a_movto_venta.f9930_guid = t9933_pdv_a_movto_venta_prep.f9933_guid_movto
INNER JOIN  t9931_pdv_a_movto_venta_dscto on
        t9930_pdv_a_movto_venta.f9930_guid = t9931_pdv_a_movto_venta_dscto.f9931_guid_movto
INNER JOIN t121_mc_items_extensiones ON
        t9933_pdv_a_movto_venta_prep.f9933_rowid_item_ext = t121_mc_items_extensiones.f121_rowid
INNER JOIN t120_mc_items ON
        t121_mc_items_extensiones.f121_rowid_item = t120_mc_items.f120_rowid
WHERE
        --t9910_pdv_a_control_tpv.f9910_id_co = '002' and   /* Centro de operación */
        f120_referencia like 'MPPEF%'
        and /*EL PRODUCTO A EVALUAR*/
        --f120_referencia = 'MPPEF0000xxx' and /*EL PRODUCTO A EVALUAR*/
        f9920_id_fecha_docto BETWEEN '20220101' AND '20241231' /* Rango de fechas */
        AND t9920_pdv_a_doctos.f9920_id_cia = 1
        AND t285_co_centro_op.f285_id_regional in ('OC', 'CO', 'NT')
group by
        f9920_id_fecha_docto
ORDER BY

        f9920_id_fecha_docto
```

Figura 13. Fuente SQL ventas POS ERP

Fuente: Autor

Results | Messages

	fecha_docto	totalValorNetoVe...	f9920_ventas_contado	f9920_ventas_credito	Unidades_Vendidas	TotalDsctos	gramos	VentasNORTE	VentasOCCIDENTE	VentasCENTROORIENTE
10	2022-01-17 ...	35439655,00	34941975,00	497680,00	64	363980,00	1786.00000000	0,00	34994375,00	445280,00
11	2022-01-18 ...	11294640,00	9811040,00	1483600,00	71	195758,00	798.00000000	0,00	10620240,00	674400,00
12	2022-01-19 ...	3410185,00	3074185,00	336000,00	29	100349,00	390.00000000	2946185,00	315200,00	148800,00
13	2022-01-20 ...	8722450,00	8616850,00	105600,00	31	197014,00	1324.00000000	0,00	8703325,00	19125,00
14	2022-01-21 ...	17970875,00	17246875,00	724000,00	71	175163,00	1142.00000000	1621125,00	16349750,00	0,00
15	2022-01-22 ...	4034100,00	4034100,00	0,00	14	42730,00	250.00000000	0,00	4034100,00	0,00
16	2022-01-23 ...	15168655,00	14480255,00	688400,00	90	221149,00	795.00000000	8884130,00	5831325,00	453200,00
17	2022-01-24 ...	9703175,00	9524775,00	178400,00	30	140166,00	950.00000000	0,00	9560025,00	143150,00
18	2022-01-25 ...	9406350,00	9385150,00	21200,00	32	82626,00	586.00000000	11625,00	442875,00	8951850,00
19	2022-01-26 ...	5805575,00	5805575,00	0,00	37	116866,00	522.00000000	1515600,00	112175,00	4177800,00
20	2022-01-27 ...	48707525,00	48306725,00	400800,00	109	384913,00	2164.00000000	274750,00	10000575,00	38432200,00
21	2022-01-28 ...	11857950,00	11747150,00	110800,00	50	149875,00	1004.00000000	0,00	11738800,00	119150,00
22	2022-01-29 ...	6918260,00	6918260,00	0,00	28	66529,00	493.00000000	0,00	6918260,00	0,00

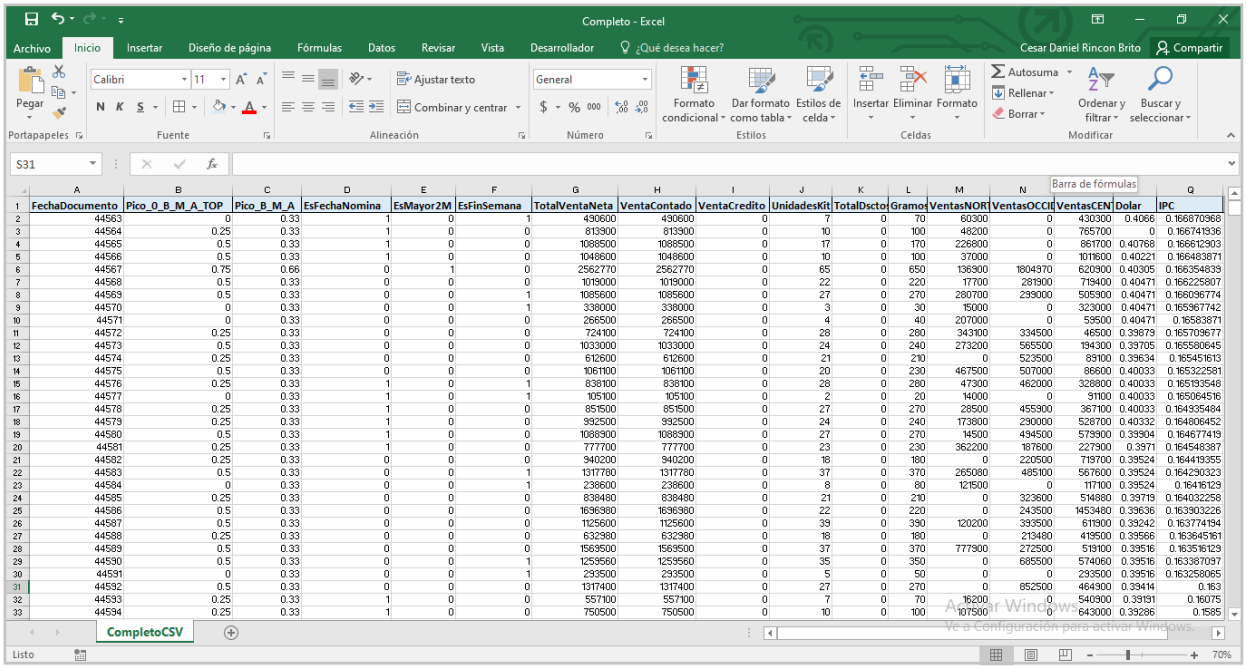
Query executed successfully. | P91B03A (12.0 RTM) | SIESA\cesar.rincon (55) | UnoEE | 00:00:00

Tabla 12. Datos consulta base de datos de pruebas

Fuente: El autor (Base de datos ERP)

9.3.2 Transformar y formatear los datos para que estén listos para el modelado.

Para obtener los datos formateados para el modelado y ser utilizado en los módulos de PySindy, se realiza una exportación a documento plano CVS separado por comas.



FechaDocumento	Pico_0_B_M_A_TOP	Pico_B_M_A	EsFechaNomina	EsMayor2M	EsFinSemana	TotalVentaNeta	VentaContado	VentaCredito	UnidadesKit	TotalDscto	Gramos	VentasNOR	VentasOCCI	VentasCEN	Dolar	IPC
44563	0	0.33	1	0	1	490600	490600	0	7	0	70	60300	0	430300	0.4066	0.166870968
44564	0.25	0.33	1	0	0	813900	813900	0	10	0	100	48200	0	765700	0	0.166741936
44565	0.5	0.33	1	0	0	1088500	1088500	0	17	0	170	226800	0	861700	0.40768	0.166612903
44566	0.5	0.33	1	0	0	1048600	1048600	0	10	0	100	37000	0	1011600	0.40221	0.166483871
44567	0.75	0.66	0	1	0	2562770	2562770	0	65	0	650	136900	1804970	620900	0.40305	0.166354839
44568	0.5	0.33	0	0	0	1019000	1019000	0	22	0	220	17700	281900	719400	0.40471	0.166225807
44569	0.5	0.33	0	0	1	1085600	1085600	0	27	0	270	280700	299000	505900	0.40471	0.166096774
44570	0	0.33	0	0	1	338000	338000	0	3	0	30	15000	0	323000	0.40471	0.165967742
44571	0	0.33	0	0	0	266500	266500	0	4	0	40	207000	0	59500	0.40471	0.16583871
44572	0.25	0.33	0	0	0	724100	724100	0	28	0	280	343100	334500	46500	0.39879	0.165709677
44573	0.5	0.33	0	0	0	1033000	1033000	0	24	0	240	273200	565500	194300	0.39705	0.165580645
44574	0.25	0.33	0	0	0	612600	612600	0	21	0	210	0	523500	89100	0.39634	0.165451613
44575	0.5	0.33	0	0	0	1061100	1061100	0	20	0	230	467500	507000	86600	0.40033	0.165322581
44576	0.25	0.33	1	0	1	838100	838100	0	28	0	280	47300	462000	328800	0.40033	0.165193548
44577	0	0.33	1	0	1	105100	105100	0	2	0	20	14000	0	91100	0.40033	0.165064516
44578	0.25	0.33	1	0	0	851500	851500	0	27	0	270	28500	455900	367100	0.40033	0.164935484
44579	0.25	0.33	1	0	0	992500	992500	0	24	0	240	173800	290000	528700	0.40332	0.164806452
44580	0.5	0.33	1	0	0	1088900	1088900	0	27	0	270	14500	494500	579900	0.39904	0.164677419
44581	0.25	0.33	1	0	0	777700	777700	0	23	0	230	362200	187600	227900	0.3971	0.164548387
44582	0.25	0.33	0	0	0	940200	940200	0	18	0	180	0	220500	719700	0.39524	0.164419355
44583	0.5	0.33	0	0	1	1317780	1317780	0	37	0	370	265080	485100	567600	0.39524	0.164290323
44584	0	0.33	0	0	1	238600	238600	0	8	0	80	121500	0	117100	0.39524	0.16416129
44585	0.25	0.33	0	0	0	838480	838480	0	21	0	210	0	323600	514880	0.39719	0.164032258
44586	0.5	0.33	0	0	0	1696980	1696980	0	22	0	220	0	243500	1453480	0.39636	0.163903226
44587	0.5	0.33	0	0	0	1125600	1125600	0	39	0	390	120200	393500	611900	0.39242	0.163774194
44588	0.25	0.33	0	0	0	632980	632980	0	18	0	180	0	213480	419500	0.39586	0.163645161
44589	0.5	0.33	0	0	0	1569500	1569500	0	37	0	370	777900	272500	519100	0.39516	0.163516129
44590	0.5	0.33	0	0	1	1259560	1259560	0	35	0	350	0	685500	574060	0.39516	0.163387097
44591	0	0.33	0	0	1	293500	293500	0	5	0	50	0	0	293500	0.39516	0.163258065
44592	0.5	0.33	0	0	0	1317400	1317400	0	27	0	270	0	852500	464900	0.39414	0.163
44593	0.25	0.33	1	0	0	557100	557100	0	7	0	70	16200	0	540900	0.39191	0.16075
44594	0.25	0.33	1	0	0	750500	750500	0	10	0	100	107500	0	643000	0.39286	0.1585

Tabla 13. Exportación de datos SQL server a CVS

Fuente: El autor

9.3.3 Codificación de variables categóricas

En este proceso se convierten las variables cualitativas (no numéricas) en valores numéricos, para que puedan ser utilizadas por modelos estadísticos o de machine learning, que solo procesan números.

Variable	Valores
Picos ventas	0.33, 0.66, 1 (Bajo, Medio, Alto)
Fecha especial	0,1

Tabla 14. Codificado con One-Hot.

Fuente: El autor

9.3.4 Creación de variables derivadas

En este proceso se generan nuevas variables a partir de datos existentes, aplicando transformaciones, cálculos, reglas de negocio o combinaciones.

A partir del promedio de ventas en los picos, se calcula el pico alto, medio y bajo.

Se calculan las fechas especiales basados en picos y días de alto flujo de ventas.

Selección de atributos

Se eligen las variables del dataset que se usarán para entrenar el modelo, para este estudio se han seleccionado las ventas segmentadas, las unidades, fechas especiales y los picos clasificados en alto, medio y bajo.

	A	B	C	D	E	F	G
1	Pico_B_M_A	TotalVentaNeta	UnidadesKit	VentasNORTE	VentasOCCIDENTE	VentasCENTROORIENTE	EsFechaEspecial
2	33	490600	7	60300	0	430300	0
3	33	813900	10	48200	0	765700	0
4	33	1088500	17	226800	0	861700	0
5	33	1048600	10	37000	0	1011600	0
6	66	2562770	65	136900	1804970	620900	1
7	33	1019000	22	17700	281900	719400	0
8	33	1085600	27	280700	299000	505900	0
9	33	338000	3	15000	0	323000	0

Tabla 15. Selección de variables CSV.

Fuente: El autor

Variable	Data set	Columna	Registros
Ventas (Pred)	y_trainDatos	TotalVentaNeta	1093
Pico categórica	r_trainDatos	Pico B_M_A	1093
Unidades	x_trainDatos	Unidades Kit	1093
Vta CO	n_trainDatos	Ventas centro oriente	1093

Vta OC	k_trainDatos	Ventas occidente	1093
Vta NT	l_trainDatos	Ventas norte	1093
Fecha especial	u_trainDatos	Pico 1/0	1093

Tabla 16. Selección de variables.

Fuente: El autor

9.4 Modelado

Selección de la técnica de modelado

Para la fase de modelado se seleccionó la técnica SINDy (Sparse Identification of Nonlinear Dynamics). Esta metodología permite identificar ecuaciones diferenciales no lineales dispersas que gobiernan la dinámica subyacente en los datos. A diferencia de otros enfoques puramente estadísticos o de aprendizaje automático, SINDy no solo predice, sino que también ofrece un modelo matemático interpretable del sistema. La elección de esta técnica se fundamenta en las siguientes razones:

Naturaleza dinámica de las ventas

Las series de ventas suelen presentar comportamientos no lineales, influenciados por múltiples factores (tendencias, estacionalidad, promociones, choques externos). Estos patrones no siempre son capturados adecuadamente por modelos lineales tradicionales.

Capacidad de descubrimiento de ecuaciones

SINDy permite aproximar el sistema de ventas mediante un conjunto reducido de funciones base (polinomiales, trigonométricas, interacciones, etc.), lo que facilita obtener una expresión matemática compacta que describe la evolución del fenómeno.

Interpretabilidad

A diferencia de modelos como las redes neuronales, los resultados de SINDy se expresan en términos de ecuaciones diferenciales parciales u ordinarias. Esto facilita comprender las relaciones entre variables de ventas y explicar los resultados a un público no técnico.

Flexibilidad y ajuste

La técnica permite explorar diferentes librerías de funciones (polinomios, funciones trigonométricas, exponenciales), así como ajustar el umbral de dispersión para balancear entre precisión y simplicidad del modelo.

Comparación con otros enfoques

Aunque podrían emplearse técnicas clásicas de predicción como ARIMA, Prophet o modelos de regresión, estas se centran más en la predicción puntual. Por ello, SINDy se considera más adecuado para este objetivo de investigación.

División de los datos en entrenamiento, validación y prueba

Para garantizar la robustez y generalización del modelo identificado con SINDy, los datos de ventas fueron divididos en tres subconjuntos:

Datos de entrenamiento

Utilizados para ajustar el modelo y estimar las ecuaciones dinámicas que describen el comportamiento de las ventas. En este conjunto, SINDy identifica las funciones candidatas y aplica técnicas de dispersión para obtener un sistema reducido de ecuaciones.

Datos de validación

Empleados para seleccionar los parámetros de regularización y el umbral de dispersión del algoritmo, evaluando la capacidad del modelo de generalizar más allá de los datos de entrenamiento. Esto evita el sobreajuste y ayuda a encontrar un balance entre precisión y simplicidad del modelo.

Datos de prueba

Reservados para la evaluación final del modelo. En este subconjunto se verifica el poder predictivo y la estabilidad de las ecuaciones identificadas por SINDy en datos nunca antes vistos, garantizando una medición objetiva del desempeño.

La división de los datos se realizó respetando la naturaleza temporal de las ventas (serie de tiempo). Por lo tanto, se optó por una partición secuencial (entrenamiento al inicio, validación en un bloque intermedio y prueba en el bloque más reciente), con el fin de evitar fugas de información y simular un escenario real de predicción futura.

Aunque el modelo SINDy eliminó el aporte de la variable *Unidades* al poner en cero sus coeficientes, la estructura interna del modelo sigue dependiendo de la posición y el número de entradas originales.

En SINDy, las entradas (`u`) se usan para construir la biblioteca de términos en un orden fijo. Incluso si una variable no contribuye al modelo final, su ausencia modificaría el orden y el tamaño de las columnas, alterando la correspondencia entre las variables y los términos generados.

Por este motivo, no es posible retirar la variable *Unidades* sin reentrenar el modelo desde cero, ya que su eliminación rompe la estructura de índices con la que el modelo fue entrenado y genera inconsistencias en la simulación o predicción.

En el modelo obtenido mediante Sparse Identification of Nonlinear Dynamics (SINDy), la variable *Unidades* presenta coeficientes nulos, lo que indica que no aporta de manera directa a la dinámica identificada. Sin embargo, su presencia en el conjunto de entradas es estructuralmente necesaria por dos razones:

Construcción de la biblioteca de funciones $\boldsymbol{\theta(X, U)}$

La biblioteca de términos se genera a partir de todas las combinaciones polinomiales y cruzadas de las variables de estado y las entradas externas. La posición de cada variable en $\boldsymbol{X}$ y $\boldsymbol{U}$ define el orden y la correspondencia de las columnas en θ. Eliminar una variable altera el mapeo entre términos y coeficientes, invalidando la interpretación y el uso del modelo entrenado.

Consistencia dimensional en la simulación

Durante la simulación y validación, SINDy espera recibir datos con la misma dimensionalidad de $\boldsymbol{X}$ y $\boldsymbol{U}$ usada en el entrenamiento. Si se retira una variable de control, la dimensión de $\boldsymbol{U}$ se reduce y la indexación de las entradas se desajusta, provocando errores en el cálculo de $\dot{X}$ o $X_{[K+1]}$.

En consecuencia, aunque *Unidades* no influya explícitamente en la ecuación final, su eliminación sin reentrenar el modelo genera una incompatibilidad estructural que impide la reutilización directa del modelo. Para excluirla de manera segura, sería necesario reconstruir θ y reestimar los coeficientes con la nueva configuración de variables.

División de los datos en entrenamiento y prueba

Variables definidas

Variable dependiente	y_trainDatos = df['TotalVentaNeta'].head(1093)
Variables independientes	x_trainDatos = df['UnidadesKit'].head(1093)
	r_trainDatos = df['Pico_B_M_A'].head(1093)
	n_trainDatos = df['VentasCENTROORIENTE'].head(1093)
	k_trainDatos = df['VentasOCCIDENTE'].head(1093)
	l_trainDatos = df['VentasNORTE'].head(1093)
	u_trainDatos = df['EsFechaEspecial'].head(1093)

Tabla 17. Variables definidas entradas y salidas

Estas series representan la dinámica del sistema que SINDy intentará modelar.

Para entrenar y validar el modelo se dividen los datos en proporciones estándar:

- 90% entrenamiento
- 10% validación

Se realiza la división del conjunto de datos con las librerías de Python, esto se logra con «train_test_split».

1. Separar el 90% entrenamiento y el 10% restante para validación.
2. Aplicar el mismo procedimiento se aplica a cada variable (y, r, x, n, k, l, u).

Como SINDy requiere representar la evolución del sistema en función del tiempo, se construyen las trayectorias combinando las variables de entrada:

u_train6	np.hstack((r_train[1:], x_train[:-1], n_train[:-1], k_train[:-1], l_train[:-1], u_train[:-1])
r_train[1:]	Corresponde a la serie desplazada de Pico_B_M_A.
x_train[:-1]	Corresponde a la serie desplazada de UnidadesKit.
n_train[:-1]	Corresponde a la serie desplazada de VentasCENTROORIENTE.
k_train[:-1]	Corresponde a la serie desplazada de VentasOCCIDENTE
l_train[:-1]	Corresponde a la serie desplazada de VentasNORTE
u_train[:-1]	Corresponde a la serie desplazada de Fecha especial (uPicoBin)

Tabla 18. Series de tiempo y trayectorias

El resultado es una matriz de características alineadas temporalmente para capturar la relación entre entradas y la variable objetivo.

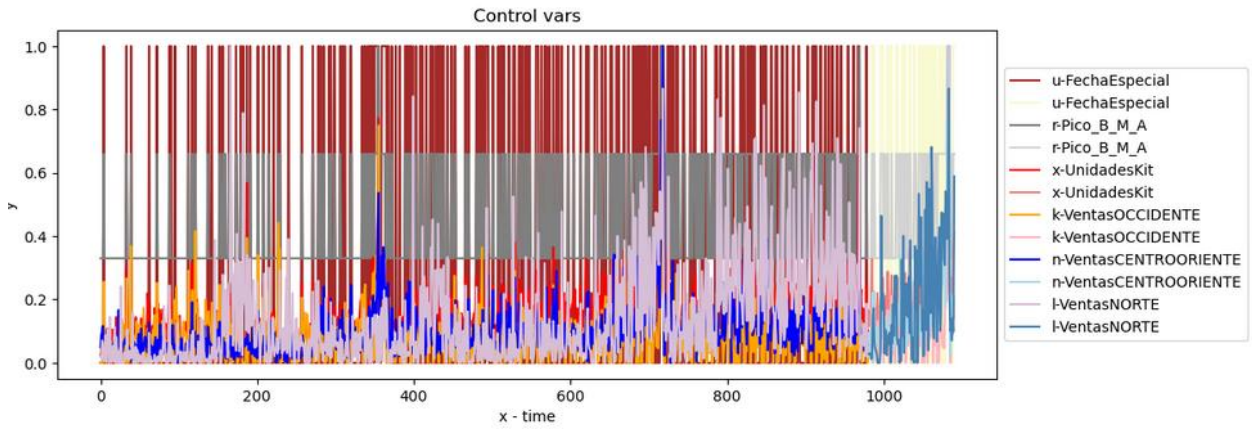


Figura 14 Datos escalados entrenamiento y validación.

Fuente: El autor

Construcción del modelo

En este caso, se plantea la construcción de un modelo para analizar y predecir el comportamiento de las ventas, tomando en cuenta variables de entrada asociadas a las unidades vendidas y a los picos de demanda.

Definición de las variables

Se define la lista de nombres de las variables que estarán en el modelo.

- feature_names = ['Ventas']
- input_features = ['rPicoBMA', 'xUnd', 'nVtaCO', 'kVtaOCC', 'lVtaNT', 'uPicoBin(FechaEspecial)']

En este caso, las ventas se establecen como la variable dependiente, mientras que las unidades y los picos son variables explicativas o, de entrada.

Selección del optimizador

El optimizador define cómo se identifican los coeficientes que mejor describen la dinámica: optimizer=ps.STLSQ(threshold=0.05)

STLSQ (Sequentially Thresholded Least Squares): este método aplica un ajuste por mínimos cuadrados secuenciales con umbral, eliminando términos que no aportan significativamente al modelo.

El parámetro threshold=0.05 controla el nivel de parsimonia, es decir, qué tan estricta es la eliminación de términos.

Biblioteca de funciones

Se define el uso de una librería polinómica con grado 2, lo que significa que el modelo tendrá en cuenta términos lineales, cuadráticos e interacciones entre las variables.

feature_library = ps.PolynomialLibrary (degree=2)

Tiempo discreto

El modelo trabaja con datos en tiempo discreto, lo cual es coherente con registros de ventas en periodos definidos (diarios, semanales o mensuales)

discrete_time=True

El modelo construido permite:

Identificar relaciones explícitas entre las ventas y las variables de entrada.

Obtener una ecuación dinámica simplificada que represente el sistema.

Realizar simulaciones para predecir la evolución futura de las ventas en función de cambios en las variables de control como las ventas segmentadas o fechas especiales.

Evaluación del modelo inicial

Realizado el entrenamiento con el conjunto de datos seleccionado y las variables de entrada y salida se obtiene:

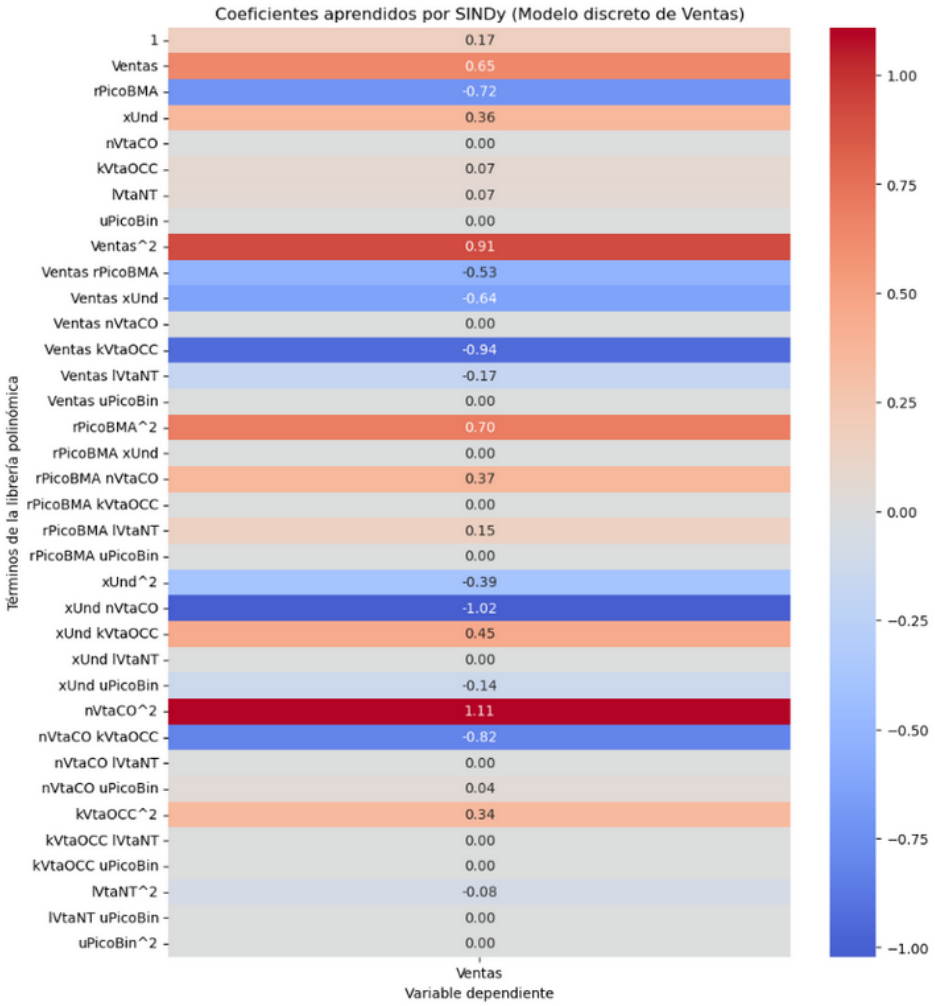


Tabla 19. Tabla Coeficientes aprendidos por SINDy

La Tabla 21 representa el mapa de calor de coeficientes aprendidos por SINDy para un modelo discreto de Ventas, cada fila es un término candidato (constante, lineal, cuadrático o producto cruzado). El color indica la magnitud y el signo del coeficiente:

- Rojo / positivo (relación directa con las ventas).
- Azul / negativo (relación inversa).
- Gris / coeficiente casi 0 (no aporta al modelo).

A partir de los coeficientes seleccionados SINDy genera el modelo discreto:

$$
\begin{aligned}
Ventas)[k+1] &= 0.167\,1 + 0.645\,Ventas[k] + -0.721\,rPicoBMA[k] + 0.361\,xUnd[k] \\
&+ 0.072\,kVtaOCC[k] + 0.067\,lVtaNT[k] + 0.914\,Ventas[k]\text{^}2 \\
&+ -0.528\,Ventas[k]\,rPicoBMA[k] + -0.635\,Ventas[k]\,xUnd[k] \\
&+ -0.943\,Ventas[k]\,kVtaOCC[k] + -0.173\,Ventas[k]\,lVtaNT[k] \\
&+ 0.699\,rPicoBMA[k]\text{^}2 + 0.368\,rPicoBMA[k]\,nVtaCO[k] \\
&+ 0.146\,rPicoBMA[k]\,lVtaNT[k] + -0.392\,xUnd[k]\text{^}2 \\
&+ -1.021\,xUnd[k]\,nVtaCO[k] + 0.452\,xUnd[k]\,kVtaOCC[k] \\
&+ -0.139\,xUnd[k]\,uPicoBin[k] + 1.109\,nVtaCO[k]\text{^}2 \\
&+ -0.820\,nVtaCO[k]\,kVtaOCC[k] + 0.036\,nVtaCO[k]\,uPicoBin[k] \\
&+ 0.344\,kVtaOCC[k]\text{^}2 + -0.078\,lVtaNT[k]\text{^}2
\end{aligned}
$$

Ecuación del modelo SINDy(1)

Constante (0.151)

El modelo predice que, independientemente de las demás variables, siempre habrá un aporte base de 0.151 a las ventas en el siguiente periodo. Esto puede interpretarse como una "demanda mínima" o una tendencia estructural de fondo.

Términos lineales (Ventas, rPicoBMA, xUnd, kVtaOCC, lVtaNT.)

Ventas (0.65) significa que el valor de ventas pasadas influye positivamente en las futuras.

rPicoBMA (–0.72) impacta de manera negativa: cuando aumenta este indicador, las ventas tienden a bajar.

Cuadráticos (Ventas², nVtaCO², etc.)

Ventas² (0.91) y nVtaCO² (1.11) muestran que hay efectos no lineales fuertes, donde el crecimiento no es proporcional.

Estos términos indican que la relación entre variables y ventas no es lineal simple, sino que hay aceleraciones o saturaciones.

Términos cruzados (interacciones, ej. Ventas × rPicoBMA, xUnd × nVtaCO, etc.)

Ventas × kVtaOCC (–0.94) muestra que la combinación entre ventas previas y occidente reduce el valor futuro → interacción negativa.

xUnd × kVtaOCC (0.45) es positiva: cuando ambas crecen, se refuerzan.

Coeficientes generados por SINDy:

```
[[0.16731870832883014, 0.6453363292411931, -0.720568060318899,
0.3610725601307169, 0.0, 0.07246544886600681, 0.06651224282332525, 0.0,
0.9138453937284441, -0.5282876004341885, -0.6350143423605925, 0.0, -
0.9434389750368662, -0.17333593853134385, 0.0, 0.6992481208660893, 0.0,
0.3678796985714399, 0.0, 0.14620509510066781, 0.0, -0.39228872569553463, -
1.0214023374751535, 0.45181045233063555, 0.0, -0.1385456273860429,
1.1091809888180277, -0.8198005410768302, 0.0, 0.035789188996238376,
0.34360751941139744, 0.0, 0.0, -0.07805655764751561, 0.0, 0.0]]

Variables de coeficientes:

['1', 'Ventas', 'rPicoBMA', 'xUnd', 'nVtaCO', 'kVtaOCC', 'lVtaNT', 'uPico-
Bin', 'Ventas^2', 'Ventas rPicoBMA', 'Ventas xUnd', 'Ventas nVtaCO', 'Ventas
kVtaOCC', 'Ventas lVtaNT', 'Ventas uPicoBin', 'rPicoBMA^2', 'rPicoBMA xUnd',
'rPicoBMA nVtaCO', 'rPicoBMA kVtaOCC', 'rPicoBMA lVtaNT', 'rPicoBMA uPico-
Bin', 'xUnd^2', 'xUnd nVtaCO', 'xUnd kVtaOCC', 'xUnd lVtaNT', 'xUnd uPico-
Bin', 'nVtaCO^2', 'nVtaCO kVtaOCC', 'nVtaCO lVtaNT', 'nVtaCO uPicoBin',
'kVtaOCC^2', 'kVtaOCC lVtaNT', 'kVtaOCC uPicoBin', 'lVtaNT^2', 'lVtaNT uPico-
Bin', 'uPicoBin^2']
```

En la lista de coeficientes se observan en 0.0, tales como:

- `nVtaCO`, `uPicoBin`, `Ventas kVtaOCC`.

Esto significa que el modelo no encontró evidencia suficiente para incluirlos en la ecuación. La regularización de SINDy descartó esos términos por ser poco relevantes.

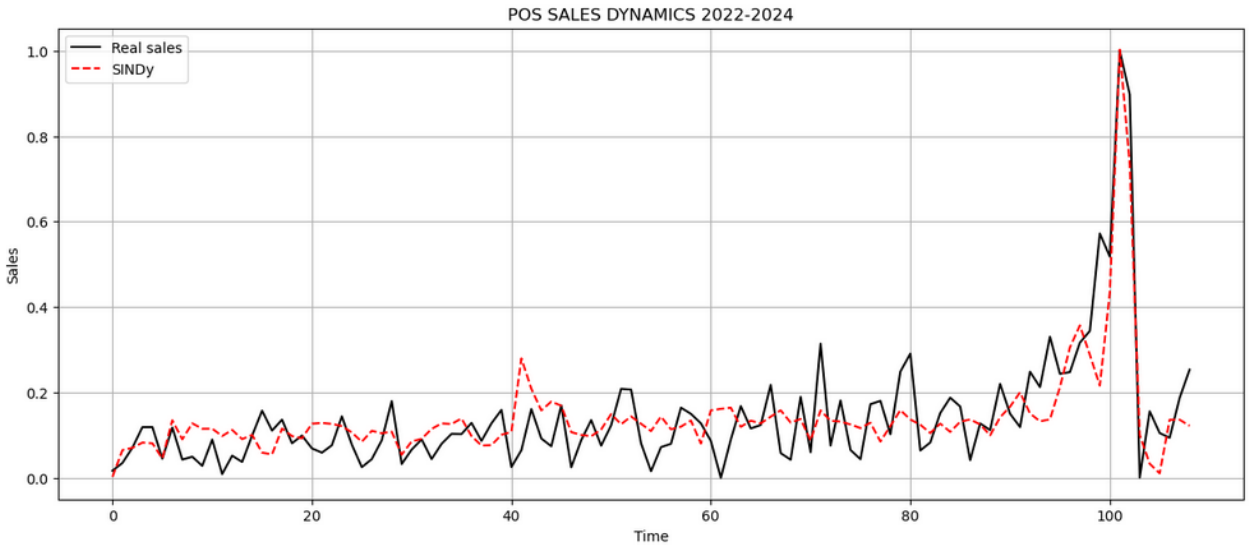


Figura 15. Simulación datos de validación

Fuente. El autor

De la Figura 15 el modelo captura adecuadamente la forma general de la serie, siguiendo de cerca los altibajos de los datos reales, aunque la curva simulada es más suavizada que la real, en intervalos sin picos, el modelo tiende a subestimar o sobreestimar levemente, mostrando menos variabilidad que los datos reales.

Documentación y comportamiento del modelo

El sistema de ventas está altamente influenciado por los picos, tanto de forma lineal como cuadrática.

Las ventas actuales, por sí solas, no aparecen de forma aislada, pero sí tienen un papel relevante cuando interactúan con los picos.

El modelo sugiere que los picos son la variable clave para explicar la dinámica futura de las ventas.

El término cuadrático muestra que el efecto de los picos no es proporcional, sino que se intensifica a medida que los picos son más grandes.

De la Figura 15 La tendencia global de mantener las ventas alrededor de valores bajos (0.05–0.2) también se refleja en la simulación.

En el sobre ajuste el hecho de que el modelo capture el pico mayor y la forma general indica que aprendió bien la dinámica principal. No obstante, el suavizado en valores pequeños sugiere que aún no capta toda la versatilidad de los datos, el término cuadrático en Picos ayuda a capturar los grandes saltos, pero no todos los detalles de la dinámica.

EL 10% de validación muestra que no está sobre ajustado, la similitud entre las curvas indica que la estructura descubierta por SINDy es estable y puede proyectarse fuera del entrenamiento. La diferencia en las pequeñas oscilaciones sugiere que aún pueden explorarse librerías de funciones más robustas como las trigonométricas o polinómicas de grado 3 o bien ajustar el threshold para refinar la parsimonia.

El modelo SINDy es estable, no obstante, pierde algo de precisión, mantiene un buen nivel predictivo en datos no vistos.

9.5 Evaluación

Evaluación del rendimiento del modelo

De la Figura 15 el resultado de la validación registra un R^2 de

$$\mathbf{R^2 = 0.7096}$$

El modelo explica aproximadamente el 71% de la variabilidad de las ventas en los datos de validación.

Existe un 29% de variabilidad no capturada, atribuida a ruido, factores no modelados o limitaciones en la estructura actual del modelo.

Dado que la evaluación se hizo con datos no usados en entrenamiento, se demuestra que el modelo generaliza bien y no está sobre ajustado.

El 29% de variabilidad no explicada puede deberse a:

Factores externos no modelados (promociones, estacionalidad, impulso por asesoras, clima, o nuevos mercados).

Limitaciones de la librería polinómica de grado 2 (se podría explorar grado 3 o trigonométrica).

Posible ruido en los datos históricos.

predictivo en datos no vistos.

9.6 Evaluación del conjunto prueba

El conjunto de datos de prueba se seleccionó de la siguiente manera:

Conjunto	Registros	Índices	Período
Train	976	0-975	2022-01-02 a 2024-09-06
Validación	109	976-1084	2024-09-07 a 2024-12-24
Test	109	1092-1200	2025-09-07 a 2025-12-24

Tabla 20. División del dataset para entrenamiento, validación y prueba.

Con la identificación del modelo SINDy se obtuvieron los siguientes resultados:

Conjunto	Registros	R^2
Validación	109	0.7006
Test	109	0.6691

Tabla 23. Métricas de rendimiento del conjunto de prueba y validación.

Correlación de Pearson:

- Validación: $r = 0.8370$ (p-value < 0.001)
- Test: $r = 0.8564$ (p-value < 0.001)

Resultados gráficos

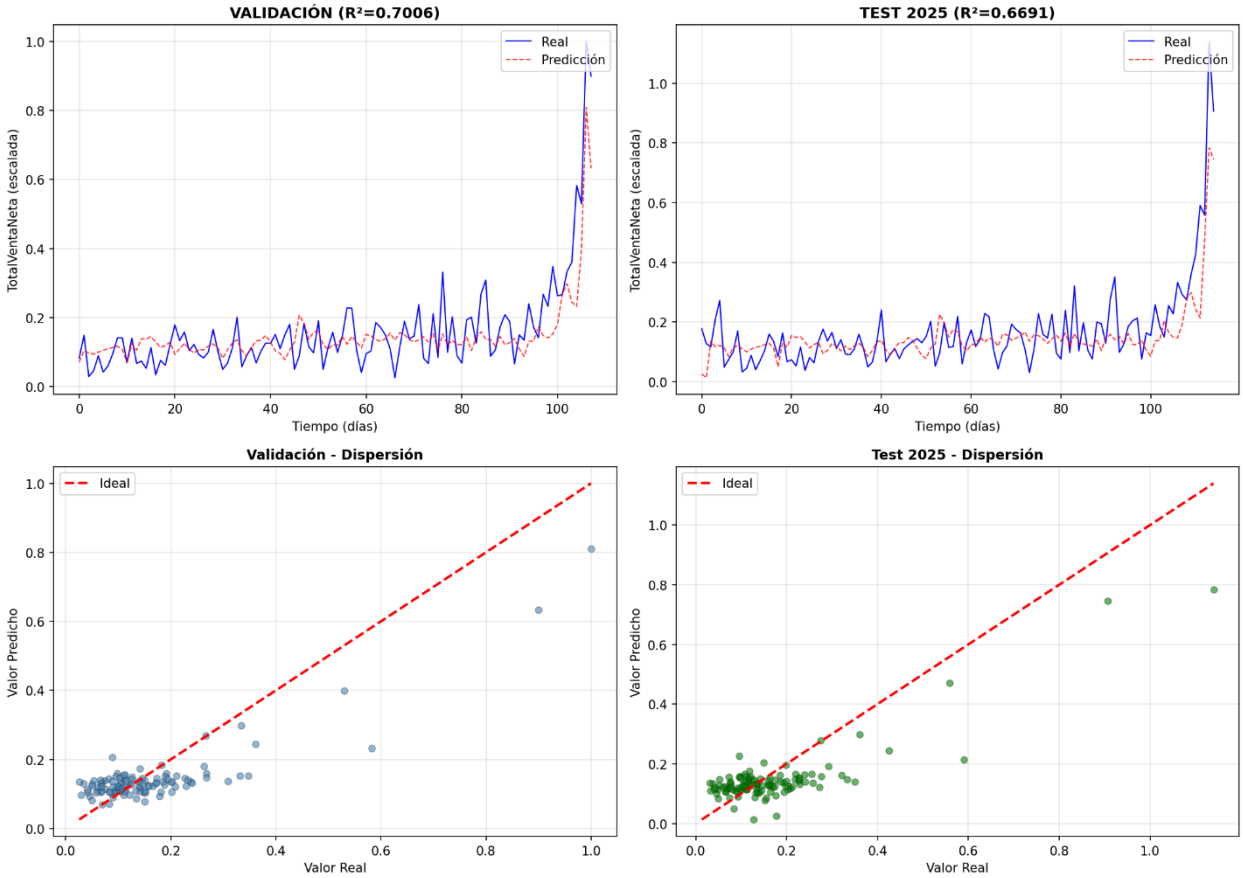


Figura 16. Series temporales y gráficos de dispersión para los conjuntos de validación (R²=0.7006) y test (R²=0.6691 Fuente. El autor

Predicción vs Real en Test

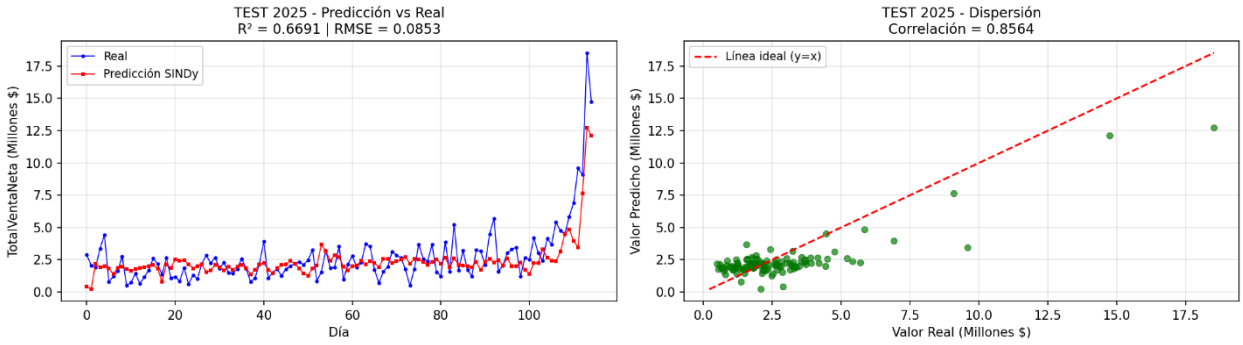


Figura 17. Comparación detallada entre valores reales y predicciones del modelo SINDy para el conjunto de test (período 2025). Izquierda: serie temporal. Derecha: gráfico de dispersión con correlación r=0.8564. Fuente. El autor

A partir de los datos obtenidos del conjunto test se pudo observar lo siguiente:

Capacidad de Generalización: El modelo SINDy demuestra buena capacidad de generalización, manteniendo un R^2 de 0.6691 en el conjunto de test (datos de 2025) que no fueron vistos durante el entrenamiento.

Correlación Significativa: Las correlaciones de Pearson superiores a 0.81 en validación y test indican una relación fuerte y estadísticamente significativa entre las predicciones y los valores reales.

Interpretabilidad: A diferencia de modelos de caja negra, SINDy proporciona una ecuación explícita que permite entender cómo las variables de control influyen en las ventas futuras.

Interacciones No Lineales: El modelo captura interacciones cuadráticas y cruzadas entre variables (como y·r, y·x, r·n), revelando relaciones complejas en la dinámica de ventas.

Aplicabilidad Práctica: Con un MAPE cercano al 48% en test, el modelo es útil para identificar tendencias y patrones, aunque las predicciones puntuales tienen margen de mejora debido a la naturaleza estocástica de las ventas.

Datos Sintéticos con correlaciones Temporales Preservadas

Se generó un segundo conjunto de datos sintético aplicando perturbaciones controladas al test original:

- Ruido AR(1) ruido de auto regresión de orden 1, cada error depende un poco del error anterior, correlacionado a variables continuas (preserva autocorrelación)
- Cambios aleatorios del 15% en variables categóricas
- Valores mantenidos en rango [0, 1]

Variable	Test Original	Test sintético aleatorio
TotalVentaNeta	0.7436	0.7547
UniadesKit	0.6320	0.6208
VentasCENTROORIENTE	0.6922	0.6582
VentasOCCIDENTE	0.5133	0.5351
VentasNorte	0.3278	0.3514

Tabla 24. Autocorrelaciones preservadas en el conjunto de datos sintético.

Variable	Test Original	Test sintético aleatorio	Diferencia
R^2	0.6289	0.6211	0.0078
RMSE	0.0792	0.0834	0.0042
Correlación	0.8011	0.8022	0.0011

Tabla 25. Comparación entre test original y sintético con correlaciones preservadas.

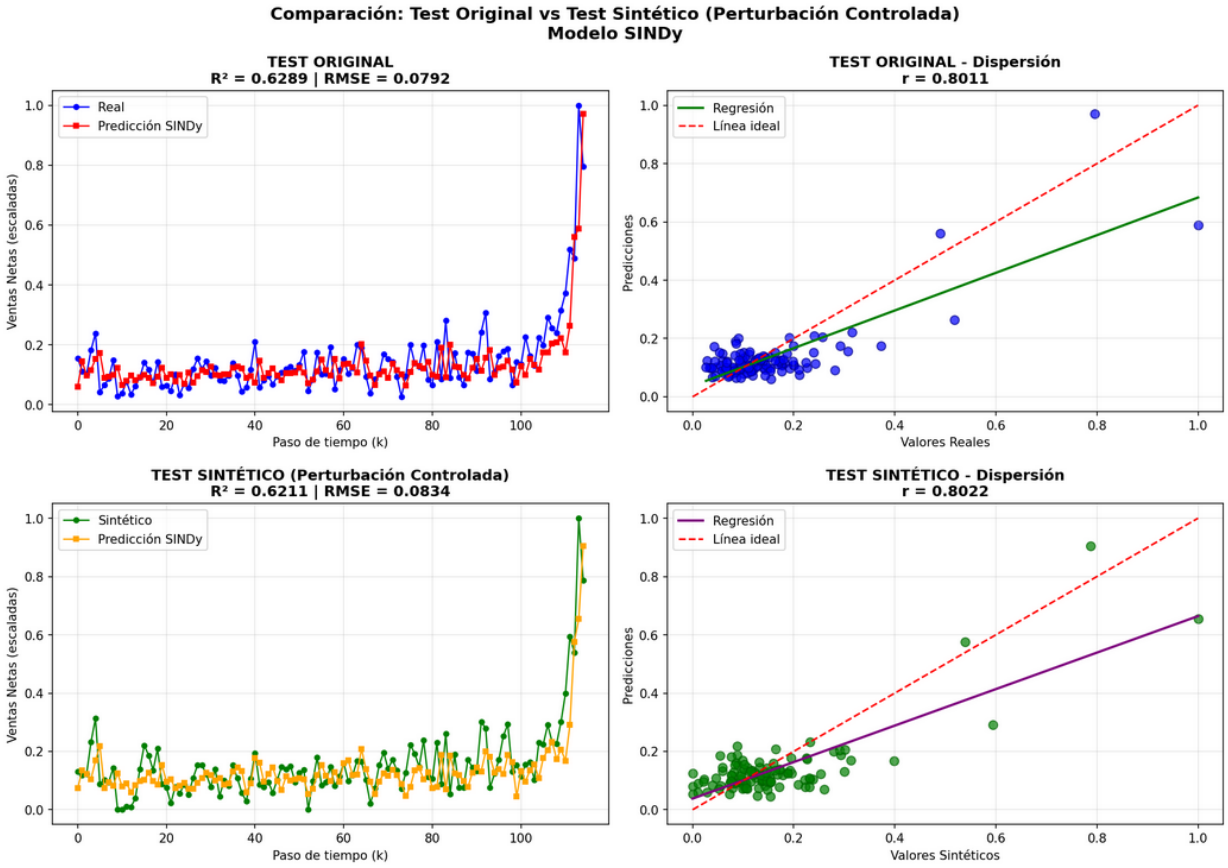


Figura 18. Comparación entre test original (arriba) y test sintético con correlaciones temporales preservadas (abajo). El modelo mantiene un rendimiento similar (diferencia de $R^2 < 1\%$). Fuente. El autor

El modelo generaliza correctamente con datos sintéticos que preservan la estructura temporal, el R^2 es prácticamente igual (diferencia de solo 0.78%).

Esto confirma que SINDy captura la dinámica real de las ventas y no está sobre ajustado a los datos de entrenamiento. El modelo es robusto ante pequeñas perturbaciones en los datos, manteniendo su capacidad predictiva.

Datos Sintéticos Aleatorios (Sin Correlaciones Temporales)

Se generó un conjunto de datos sintético con distribuciones estadísticas similares al test original, pero sin preservar correlaciones temporales (datos completamente aleatorios).

Métrica	Test Original	Test sintético aleatorio
R^2	0.6289	-1.1913
RMSE	0.0792	0.1784
Correlación	0.8011	-0.0508

Tabla 26. Comparación entre test original y sintético aleatorio.

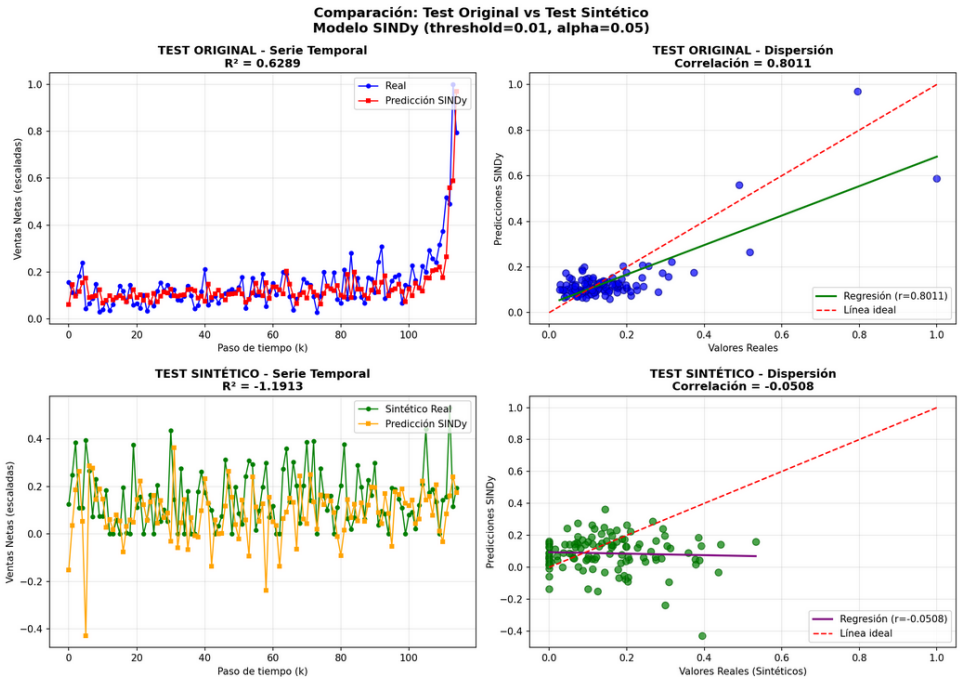


Figura 19. Comparación entre test original (arriba) y test sintético aleatorio (abajo). El modelo falla completamente con datos aleatorios (R^2 negativo). Fuente. El autor

Interpretación

El R^2 negativo (-1.19) en el test sintético aleatorio es un resultado esperado y positivo para validar el modelo:

Los datos sintéticos son aleatorios, no tienen estructura temporal ni correlaciones entre variables que existen en los datos reales de ventas.

SINDy captura patrones dinámicos reales, el modelo aprendió relaciones específicas como:

- Dependencia temporal entre ventas consecutivas ($y[k] \rightarrow y[k+1]$)
- Correlaciones entre unidades vendidas y ventas totales
- Efectos de las ventas regionales

Sin patrones, un R^2 negativo significa que predecir simplemente la media sería mejor que usar el modelo. Esto confirma que el modelo no memoriza, sino que aprende la dinámica real del negocio.

Esta validación demuestra que el modelo SINDy funciona correctamente con datos que siguen patrones reales de ventas, no funciona con datos puramente aleatorios lo cual es un comportamiento esperado y no está sobre ajustado, generaliza a nuevos datos con estructura similar

Evaluación de objetivos específicos

Objetivo específico 1.

Comprender algunas técnicas de identificación de modelos para predecir el comportamiento de los datos de un sistema ERP.

Se llevó a la práctica para identificar ecuaciones diferenciales que describen la dinámica de las variables en el sistema de ventas. Esto permitió observar cómo los datos pueden ser representados mediante un modelo matemático interpretable, lo que confirma que la técnica ayuda a predecir comportamientos futuros en el sistema.

Se implementó y validó una técnica de identificación de modelos (SINDy) aplicada a datos reales de ventas POS de un sistema ERP, logrando comprender cómo estas herramientas permiten anticipar la evolución de las variables.

Objetivo específico 2.

Estructurar la base de datos real que servirán para entrenar y evaluar el modelo de predicción.

Para cumplir con el objetivo de estructurar la base de datos real que servirá para entrenar y evaluar el modelo de predicción, se llevó a cabo un proceso de organización y preparación de los datos provenientes del sistema ERP. Este procedimiento garantizó que la información se encontrara en un formato adecuado, consistente y reproducible para su posterior análisis.

En primer lugar, se documentó un pipeline de preprocesamiento, en el cual se definieron los pasos realizados de forma clara y replicable. Se especificaron las fuentes de datos, incluyendo las tablas y campos utilizados del ERP, así como las uniones y agrupaciones efectuadas según dimensiones relevantes como producto, cliente y fecha. Además, se establecieron las variables empleadas en el modelo, tales como ventas, unidades y picos, junto con la creación de variables derivadas como rezagos (lags), y variables de calendario en formato categórico (dummies).

Posteriormente, se abordó el tratamiento de valores nulos y atípicos mediante técnicas de imputación, justificando cada decisión en función de la naturaleza de los datos y su impacto en la consistencia del modelo. Asimismo, se aplicaron procesos de normalización o estandarización cuando fue necesario, con el fin de mejorar la comparabilidad entre variables.

La división de los datos se realizó de forma predefinida con un enfoque temporal. Se generaron conjuntos de entrenamiento (90%), validación (10%) manteniendo la coherencia en la separación por bloques temporales. Para asegurar la transparencia del proceso, se documentaron los tamaños de cada muestra y la metodología de selección utilizada.

Objetivo específico 3.

Aplicar el algoritmo SINDy para la predicción de los datos de un ERP empleando las ecuaciones que rigen este tipo de sistemas dinámicos.

Para cumplir el objetivo de aplicar el algoritmo SINDy para la predicción de los datos de un ERP empleando las ecuaciones que rigen este tipo de sistemas dinámicos, se llevó a cabo un proceso estructurado que incluyó la preparación y organización de los datos, el entrenamiento y validación del modelo, así como el desarrollo de una aplicación en Python para su implementación práctica. Durante el proceso, se seleccionaron las variables más representativas del sistema, se ajustaron los parámetros del modelo y se evaluó su capacidad predictiva. Como resultado, se obtuvo un modelo con un coeficiente de determinación (R^2) cercano al 70 %, lo que evidencia una buena capacidad del algoritmo para capturar la dinámica del sistema y predecir su comportamiento con un nivel aceptable de precisión

Finalmente, se establecieron artefactos entregables que evidencian el cumplimiento de este objetivo. Entre ellos, se incluyen tablas con estadísticas descriptivas (valores mínimos y máximos) por variable y por cada subconjunto de datos, así como el código y notebooks empleados para generar los objetos `x_train`, `x_val` y `x_test`, con rutas y versiones claramente indicadas. Adicionalmente, se registraron versiones mediante hashes y archivos CSV de las muestras utilizadas, garantizando así la reproducibilidad del procedimiento completo.

El código desarrollado y los experimentos realizados se encuentran disponibles en el repositorio de GitHub [22].

Objetivo específico 4.

- Valorar el resultado de la predicción del algoritmo SINDy comparado con los datos reales.

Para cumplir el objetivo de valorar el resultado de la predicción del algoritmo SINDy comparado con los datos reales, se realizó una evaluación cuantitativa y visual del desempeño del

modelo. Se compararon las series predichas con los valores observados en el ERP, analizando métricas como el coeficiente de determinación (R^2) para determinar el grado de ajuste del modelo. Además, se generaron gráficas de comparación entre las predicciones y los datos reales, lo que permitió identificar el nivel de concordancia y las posibles desviaciones del modelo. En conjunto, este análisis permitió verificar la capacidad del algoritmo SINDy para representar de forma precisa la dinámica del sistema y predecir su comportamiento futuro.

10 Conclusiones

Se pudo observar si el modelo SINDy es una alternativa viable para sistemas ERP con datos dispersos obtenidos de las ventas, si las ecuaciones descubiertas permitieron interpretar el comportamiento del sistema y si existen posibilidades de mejora o integración con otros métodos.

Se verificó que el modelo SINDy es sólido para capturar la dinámica principal de las ventas.

La ecuación encontrada con dependencia fuerte de los picos, fechas especiales y ventas segmentadas está respaldada por un buen nivel de ajuste en validación.

El resultado R2=0.7096 muestra que el modelo tiene un muy buen desempeño predictivo y capacidad de generalización, explicando más del 70% de la variabilidad de las ventas en los datos de validación. Esto lo convierte en una base confiable para análisis y proyecciones, aunque aún queda un margen de mejora para capturar la variabilidad restante.

El conjunto de prueba final indica que el modelo generaliza adecuadamente, reproduciendo bien la dinámica principal y los picos grandes de ventas, aunque tiende a suavizar las variaciones más pequeñas. Esto lo convierte en un modelo útil para proyecciones generales, pero con margen de mejora si se busca capturar la variabilidad fina del sistema.

La validación del conjunto de datos de prueba demuestra que el modelo SINDy funciona correctamente con datos que siguen patrones reales de ventas, no funciona con datos puramente aleatorios lo cual es un comportamiento esperado y no está sobre ajustado.

10.1 Posibles mejoras

La diferencia respecto al 100% indica que aún puede mejorarse el modelo, lo cual se recomienda:

Incluir más variables explicativas como promociones, descuentos, presencia de campañas publicitarias realizadas por las asesoras de ventas o variables de regiones donde se pueda establecer tipo de lugar como sector comercial, local, centro comercial, o residencial.

Ampliar la biblioteca de funciones, actualmente usas una *PolynomialLibrary* grado 2, lo cual limita las relaciones a términos lineales, cuadráticos e interacciones, se podría considerar con más variables

Establecer nuevas librerías implementando *CustomLibrary* incluir funciones que tengan sentido en el negocio como las exponenciales si hay crecimientos acelerados, esto permitiría al modelo capturar patrones más sutiles sin perder precisión.

Hibridar el modelo, estrategias combinadas entre SINDy y ML, como ejemplo de ello, Random Forest o Redes Neuronales, usar SINDy para la base interpretable y un modelo más flexible para capturar el ruido residual.

Evaluar modelos alternativos en paralelo, no obstante, SINDy es el enfoque principal, siempre es buena práctica comparar con regresiones y otros modelos de aprendizaje automático.

11 Tendencias actuales

11.1 Instrucciones NPL para modelado a partir de datos con agentes de inteligencia artificial

Las instrucciones NPL para modelado consisten en expresiones en lenguaje natural que permiten especificar tareas de aprendizaje automático sobre un conjunto de datos, facilitando la creación de modelos de inteligencia artificial sin necesidad de codificación explícita

Instrucciones NPL: indicaciones escritas en lenguaje humano que son entendidas y procesadas mediante técnicas de procesamiento de lenguaje natural.

Para modelado: su propósito es definir qué tipo de modelo se debe generar.

A partir de datos: implica que el modelo no es puramente simbólico, sino que se construye o entrena usando un conjunto de datos.

Con inteligencia artificial: las instrucciones guían a un sistema de IA, por ejemplo, un LLM como OpenAI GPT, Copilot, LLaMA (Meta), Claude (Anthropic), Gemini, DeepSeek o una plataforma AutoML tales como Google AutoML, H2O AutoML, Auto-sklearn, TPOT para transformar datos en un modelo predictivo, explicativo o de simulación.

Se observó una prueba de cada modelo de inteligencia artificial para OpenAI, Claude y DeepSeek.

11.1.1. GPT OpenAI

OpenAI generó la información de la matriz de correlación, utilizando los métodos de regresión lineal y random forest. La regresión lineal obtuvo un R^2=1, Random Forest obtuvo un R^2=0.995.

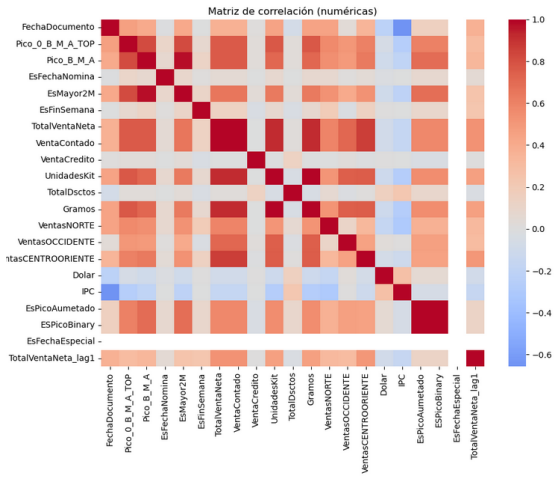


Tabla x Matriz de correlación OpenAI GPT

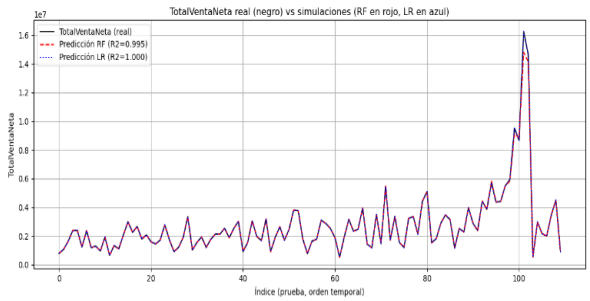


Gráfico de simulación OPEN AI GPT

```
Ecuación del modelo lineal (escala original):
TotalVentaNeta = -0.000000 + (-0.000000)*TotalVentaNeta_lag1 + (+1.000000)*TotalVentaNeta + (-0.000000)*VentaContado +
(-0.000000)*Gramos + (+0.000000)*UnidadesKit + (+0.000000)*VentasCENTROORIENTE + (+0.000000)*Pico_0_B_M_A_TOP + (+0.00
0000)*Pico_B_M_A
```

Esta ecuación es básicamente: $TotalVentaNeta_{[k+1]} = 1 * TotalVentaNeta[k]$

Esto significa que el sistema es una identidad, el valor en el siguiente paso es prácticamente igual al valor actual, por lo tanto, OPEN AI no ha generado un modelo óptimo.

11.1.2 Claude AI

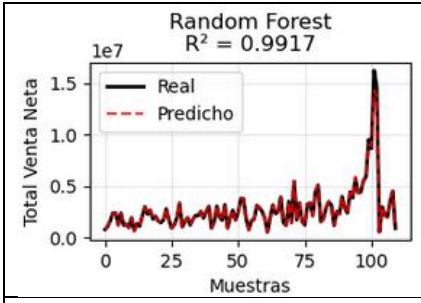
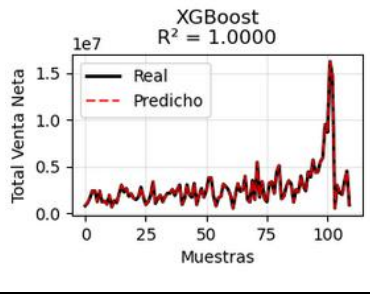
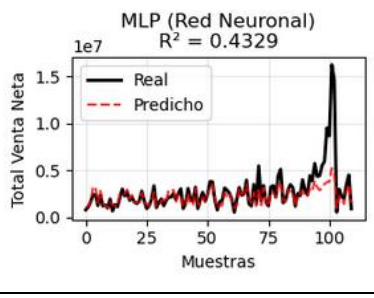
Claude utilizó la regresión polinomial no lineal de grado 2 incluyendo términos lineales, cuadráticos y de interacción, capturando relaciones no lineales y correlaciones entre variables.

Se puede observar que el modelo Claude no logra capturar bien la información, a partir de la entrada de texto predefinido no genera un modelo de predicción adecuado, de igual forma que OPENAI debe realizarse el proceso por partes y verificado por el experto en el negocio.

11.1.3. DEEPSEEK

El agnte de DeepSeek utilizó los algoritmos de Machine Learning No Lineales Random Forest**,** XGBoost y MLP (Perceptrón Multicapa) Como métricas de evaluación utilizó R^2 (Coeficiente de determinación).

Tomó las características del sistema no lineal como discreto, multivariable y predictivo, generando la matriz de correlación, graficas comparativas de todos los modelos, detalles de métricas, análisis de características.

		
Random Forest R²: 0.9917	XGBoost R²: 1.0000	MLP R²: 0.4329

A pesar de que DeepSeek generó un reporte muy completo con 3 métodos, su mejor resultado se obtuvo con Random Forest, pero se generó una ecuación lineal, donde utilizó solo los valores segmentados de las ventas omitiendo fechas especiales y picos.

Se pudo evidenciar que la forma de utilizar estas herramientas es muy óptima para la predicción y la analítica de datos, pero debe ser correctamente estructurada para que el agente pueda generar la información precisa.

El uso del lenguaje natural y la inteligencia artificial en la predicción de ventas automatizada ofrece grandes ventajas al anticipar tendencias y optimizar decisiones empresariales. Sin embargo, es fundamental conocer a fondo el negocio, ya que un modelo mal planteado puede generar resultados que no reflejen la realidad comercial y para ello se establecen metodologías configurables con agentes creando un equipo de análisis de datos impulsado por IA que incluyen un científico de datos para el análisis estadístico y selección de modelos de ML, un ingeniero de datos para el diseño de pipelines y procesos ETL, un ingeniero de ML para la implementación de modelos y pruebas, un experto en visualización para el diseño de paneles de datos y un asesor de ética para detección de sesgos y evaluación de equidad, todo controlado de manera automática por los agentes siguiendo la secuencia lógica del diagrama de flujo del proceso.

Bibliografía

[1] Klaus, H., Rosemann, M., & Gable, G. G. (2000). What is ERP? Information systems frontiers, 2, 141-162.

[2] Rolia, J., Casale, G., Krishnamurthy, D., Dawson, S., & Kraft, S. (2009, October). Predictive modelling of SAP ERP applications: challenges and solutions. In Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools (pp. 1-9).

[3] Babu, M. P., & Sastry, S. H. (2014, June). Big data and predictive analytics in ERP systems for automating decision-making process. In 2014 IEEE 5th international conference on software engineering and service science (pp. 259-262). IEEE.

[4] Tan, P.-N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.

[5] S. Ebrahim (2022). Data-driven models for structure-property prediction in additively manufactured steels. Computational Materials Science.

[6] Chang, Yen-Yu & Sun, Fan-Yun & Wu, Yuen-Hua & Lin, Shou-De. (2018). A Memory-Network Based Solution for Multivariate Time-Series Forecasting.

[7] S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Discovering governing equations from data by sparse identification of nonlinear dynamical systems,” Proceedings of the National Academy of Sciences, vol. 113, no. 15, pp. 3932–3937, 2016.

[8] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Lear-ning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3): e0194889.

[9] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

[10] Microsoft. (2023). *Bases de datos - SQL Server*. Microsoft Learn. https://learn.microsoft.com/es-es/sql/relational-databases/databases/databases?view=sql-server-ver17

[11] Senthia. (2017). *Hágalo usted mismo*. Senthia. https://www.senthia.com/

[12] Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big Data Analytics in Operations Management. Production and Operations Management, 27(10), 1868-1889.

[13] Karasan, A. (2021). Machine Learning for Financial Risk Management with Python.

[14] Galicia, A., Talavera-Llames, R., Troncoso, A., Koprinska, I., & Martínez-Álvarez, F. (2019). Multi-step forecasting for big data time series based on ensemble learning. Knowledge-Based Systems, 163, 830-841.

[15] Sapankevych, N. I., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. IEEE computational intelligence magazine, 4(2), 24-38.

[16] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[17] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[18] Russell, S. J., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

[19] Lusch, B., Kutz, J. N., & Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. Nature Communications, 9(1), 4950.

[20] Champion, K., Lusch, B., Kutz, J. N., & Brunton, S. L. (2019). Data-driven discovery of coordinates and governing equations. Proceedings of the National Academy of Sciences, 116(45), 22445–22451.

[21] Jafar, S., Hemachandran K., Nawaah D., Rodriguez R.(2025). Foundations of Artificial Intelligence in Finance: Insights for Practitioners with Applications and Case Studies

[22] Rincon, D. (2025). Predicción de ventas por identificación dispersa de un sistema ERP a partir de datos de un módulo POS [Repositorio GitHub]. GitHub. https://github.com/danielrincon302/SINDy-ERP-POS-Dynamics