

# Predicción de ventas para una empresa de Hardware Business-To-Business

Hernán Felipe Sánchez Cárdenas, ✉ [hfsanchezc@eafit.edu.co](mailto:hfsanchezc@eafit.edu.co)

Directora:

**Paula María Almonacid Hurtado**  
Área de macroeconomía y sistemas financieros  
[palmona1@eafit.edu.co](mailto:palmona1@eafit.edu.co)  
Investigadora Junior (Colciencias)  
Profesora asociada (EAFIT)



Universidad EAFIT  
Maestría en Ciencia de Datos y Analítica  
Medellín  
2024

## RESUMEN

Es bien sabido que las empresas de hardware desempeñan un papel crucial en la sociedad actual, ya que ofrecen la infraestructura y recursos tecnológicos necesarios para el correcto funcionamiento de los diferentes ámbitos que la componen. Aquellas pertenecientes al modelo business-to-business (B2B) no son la excepción, y uno de sus desafíos más importantes es la predicción de ventas debido a la impredecibilidad intrínseca de los negocios. Teniendo esto en cuenta, este trabajo está enfocado en el desarrollo e implementación de modelos estadísticos tradicionales de series de tiempo y de machine learning que permitan predecir el volumen de ventas de una empresa perteneciente a este rubro.

**Palabras clave:** *Series de Tiempo; predicción de ventas; Business-to-business; supply chain; machine learning; hardware.*

## TABLA DE CONTENIDOS

<b>I.</b>	<b>DESCRIPCIÓN DEL PROYECTO</b>	<b>2</b>
I-A.	Planteamiento del problema . . . . .	2
I-B.	Justificación . . . . .	2
I-C.	Objetivos . . . . .	3
<b>II.</b>	<b>ESTADO DEL ARTE</b>	<b>3</b>
<b>III.</b>	<b>MARCO TEÓRICO</b>	<b>5</b>
III-A.	Modelos Clásicos . . . . .	6
III-B.	Modelos de Aprendizaje Automático . . . . .	10
III-C.	Métricas de Evaluación . . . . .	12
<b>IV.</b>	<b>METODOLOGÍA</b>	<b>13</b>
IV-A.	Comprensión del Negocio . . . . .	13
IV-B.	Entendimiento de los datos . . . . .	13
IV-C.	Preparación de los datos . . . . .	14
IV-D.	Modelado . . . . .	14
IV-E.	Evaluación . . . . .	14
IV-F.	Despliegue . . . . .	14
<b>V.</b>	<b>PLAN DE GESTIÓN DE DATOS</b>	<b>14</b>
<b>VI.</b>	<b>ASPECTOS ÉTICOS</b>	<b>15</b>
	<b>REFERENCIAS</b>	<b>16</b>
<b>VII.</b>	<b>ANEXOS</b>	<b>18</b>
VII-A.	Carta . . . . .	18

## I. DESCRIPCIÓN DEL PROYECTO

### *I-A. Planteamiento del problema*

La industria tecnológica es uno de los ejes fundamentales de la sociedad actual, en el sentido en que está presente cada vez más en cualquier ámbito económico, político, social y cultural. En este contexto, las empresas de hardware desempeñan un papel crucial, ya que ofrecen la infraestructura y recursos tecnológicos necesarios para que estos se puedan desarrollar. Esto es congruente con el hecho de que algunas de las empresas más grandes del mundo actual, como Apple, Nvidia o Intel, por citar algunas, pertenezcan a este rubro [1]. Sin embargo, en los últimos cinco años, la industria ha enfrentado desafíos significativos en la cadena de suministros, principalmente debido a factores macroeconómicos. Entre estos, destacan la pandemia de COVID-19 en 2020 y la consecuente escasez de chips que ha perturbado severamente la producción y distribución de hardware [2].

Es en este contexto, que para una empresa que se dedica a la venta de hardware en un modelo de negocio business-to-business (B2B), la planeación y predicción de ventas es algo crítico. El mercado es cambiante y volátil, por lo cual se hace necesario utilizar estrategias de mercadeo y planeación que permitan asegurar la mayor satisfacción posible de los clientes en el proceso de compra, además de una adecuada gestión de inventarios y logística.

### *I-B. Justificación*

Previamente, se mencionó que la predicción de ventas es algo crucial para la gestión estratégica empresarial. Esta afirmación va encaminada específicamente a las decisiones relacionadas con la cadena de suministros: los pronósticos ayudan a tomar decisiones informadas sobre la adquisición, producción, almacenamiento y distribución de productos, buscando satisfacer la demanda de los clientes de manera eficiente [3]. En el mercado del hardware, esto es especialmente importante por los múltiples factores que influyen en la demanda, como pueden ser la cadena de suministro global (lo cual se puede evidenciar con la escasez de chips de años recientes), e incluso, con el ciclo de vida de los productos, que puede ser relativamente corto debido a la rápida obsolescencia de los productos [4].

Por otra parte, en la actualidad se cuenta, a diferencia de épocas anteriores, con grandes volúmenes de información. Gracias a la tecnología moderna que facilita su procesamiento y a modelos analíticos eficaces que ayudan a comprenderla, ahora se tiene una capacidad sin precedentes para aprovechar estos datos.

En el presente trabajo, se propone implementar y comparar modelos de series de tiempo y de machine learning para predecir el volumen de ventas de una empresa de hardware en un modelo B2B. Mediante la recolección y preparación de datos históricos de ventas, así

como variables relacionadas, se busca no solo garantizar la calidad del modelado predictivo, sino también analizar la variabilidad en la demanda y las variables clave que influyen en la precisión de las previsiones de ventas en este sector. Se incluirán mecanismos de ajuste flexibles para adaptarse a la naturaleza de los datos, con el objetivo de identificar el modelo más efectivo. Al hacerlo, el estudio aspira a aportar tanto a la práctica empresarial, mejorando la capacidad de las empresas de hardware para anticipar cambios en el mercado, como al conocimiento académico en el ámbito del análisis predictivo y el machine learning aplicado a contextos reales.

### *I-C. Objetivos*

#### *Objetivo General:*

- Desarrollar e implementar modelos de series de tiempo a través de aproximaciones tradicionales y de machine learning que permitan predecir el volumen de ventas de una empresa de hardware en el marco de un modelo business-to-business.

#### *Objetivos Específicos:*

- Recopilar y preparar datos históricos de ventas y variables relacionadas para asegurar su calidad en el modelamiento predictivo.
- Examinar la variabilidad en la demanda de productos de hardware en un modelo business-to-business, identificando las variables clave y posibles factores que influyen significativamente en la precisión de las previsiones de ventas para este sector.
- Comparar modelos de predicción para la demanda que utilicen técnicas tradicionales y de aprendizaje automático, incluyendo variables y mecanismos de ajuste flexibles que permitan adaptarse a la naturaleza de los datos, para identificar cuál se comporta mejor para estos.

## II. ESTADO DEL ARTE

Cuando se habla de predicción de ventas, en realidad se busca predecir la demanda por parte de los clientes con el objetivo de generar estrategias para satisfacerla. Mentzer y Moon definen esta actividad como “una proyección en el futuro de la demanda esperada dado un conjunto de factores ambientales mediante el uso de técnicas cualitativas o cuantitativas” [5]. Sin importar el enfoque o nombre que se le quiera dar, lo cierto es que esta actividad siempre ha sido un elemento fundamental para las empresas y su gestión de la cadena de suministro.

Históricamente, antes de la popularización de los métodos estadísticos tradicionales para pronósticos en series de tiempo, las grandes empresas solían depender de métodos de extrapolación, modelos de regresión lineal simple, medias móviles y, en gran medida, del juicio de expertos [6]. Sin embargo, en 1960, el método Holt-Winters surgió como una de las primeras y más importantes aproximaciones a modelos de predicción que no se basen únicamente en captar la tendencia de los datos, siendo una extensión del método de suavizado exponencial para capturar también la estacionalidad en los datos de series temporales [7]. Este enfoque abrió el camino para considerar no solo la tendencia sino también otros factores importantes de la naturaleza de los datos que son cruciales en muchas aplicaciones prácticas. Siguiendo esta innovación en el análisis de series temporales, en 1976 Box y Jenkins desarrollaron el modelo ARIMA cuyo enfoque se acerca más a describir los efectos de autocorrelación en los datos [7]. Gracias al trabajo de otros autores, fue posible agregar componentes estacionales y variables exógenas a este modelo con el fin de mejorar las predicciones captando la mayor cantidad de información posible de los datos: modelo SARIMA y SARIMAX [8]. Aún hoy, estos son algunos de los modelos más efectivos en el campo de la predicción de ventas, como se puede ver a continuación:

- **Machine-learning models for sales time series forecasting:** en este trabajo, Pavlyshenko utilizó datos históricos de ventas de la competición de Kaggle “Rossman Store Sales”, utilizando como variables el día de la semana, las promociones, la distancia a competidores, entre otros. Se comparó el modelo ARIMA con métodos de aprendizaje automático, superando al *Extra tree*, y teniendo un MAE de 13.8, muy similar al 13.6 obtenido con Random Forest, Redes Neuronales y el 12.6 obtenido con métodos de stacking [9].
- **Forecasting Seasonal and Trend-Driven Data: A Comparative Analysis of Classical Techniques:** el objetivo de este estudio fue comparar el desempeño de varios métodos cuantitativos clásicos de pronóstico para predecir las ventas semanales de un producto. Se compararon los modelos MA, suavizado exponencial simple, Holt-Winters, ARIMA, ARIMAX, SARIMA, SARIMAX y regresión lineal múltiple. Además de la información sobre las ventas semanales, se contaba con información de festivos y promociones especiales en un período de 2 años y medio. En este caso, el mejor modelo fue el SARIMAX, con un MAPE en el conjunto de prueba de 5.8 %, el cual es un excelente desempeño [10].

Con el desarrollo de las tecnologías de computación, los métodos de aprendizaje automático han emergido como un conjunto de herramientas útiles para la realización de predicciones en series de tiempo desde hace más de una década. En el contexto de la predicción de ventas, y predicción de ventas en modelo B2B, algunos estudios destacados son:

- **Hardware sales forecasting using clustering and machine learning approach:** en este trabajo realizado en una empresa de hardware, se utilizaron métodos de

clustering (KNN) para agrupar los datos, y sobre los clústers, métodos de predicción como ARIMA y LSTM (como modelo de machine learning). Las variables utilizadas fueron fecha, cantidad y stock. La métrica utilizada fue el costo ahorrado con la implementación de cada método (obtenido a partir del RMSE), obteniendo como mejor modelo la red neuronal LSTM [11].

- **Sales Analysis and Forecasting using Machine Learning Approach:** en este caso, los investigadores compararon los modelos regresión lineal, random forest y XGBoost. Los mejores resultados, con una diferencia notable, se obtuvieron utilizando XGBoost, y mejorados con técnicas de ensamble [12].
- **Explaining machine learning models in sales predictions:** el objetivo de este trabajo fue aplicar modelos de machine learning para la predicción de ventas B2B. Se utilizaron como modelos Random Forest, Naive Bayes, Decision Trees, redes neuronales y SVM. En este caso, el proyecto se planteó como un problema de clasificación, con lo cual se usaron métricas como el área bajo la curva ROC y el accuracy. En este caso, el mejor modelo fue Random Forest [13].
- **Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm:** en este estudio se utilizaron datos de ventas B2B de tres años consecutivos (2016, 2017 y 2018). Estos incluyen las características categoría, región, tipo de artículo, ID de oportunidad, trimestre, nombre del producto, subcomponente del producto y ingresos por ventas. Se utilizó un modelo de gradient boosting, obteniendo un MAPE de 0.18, que en el contexto del negocio se consideró un buen resultado [14].
- **Sales Forecasting Model Based on Ensemble Learning and Its Application in Anomaly Detection:** para predecir ventas, se utilizaron modelos de ensamble basados en SVM, LSTM, ARIMA y Holt Exponential Smoother. En este caso, se emplearon los modelos de pronóstico de forma individual y conjunta, para mejorar la robustez. Se evaluaron los rendimientos usando RMSE, MAE y correlación de Pearson, concluyendo que hubo mejores resultados utilizando métodos de ensamble [15].

Esto permitió identificar que algunos de los métodos de aprendizaje automático más utilizados en el problema de la predicción de ventas, incluso en un modelo B2B, son random forest, gradient boosting, LSTM, métodos de ensamble y XGBoost.

### III. MARCO TEÓRICO

Para comenzar, es importante definir el modelo de negocio business-to-business (B2B). Este se puede entender como aquel en el cual las transacciones comerciales se realizan directamente entre empresas [16]. A diferencia de un modelo de ventas directas, también llamado Business-to-Customer (B2C), se cuenta con un elemento adicional que aún hoy

en día es parte fundamental de la planeación en la cadena de suministro: las negociaciones con los clientes. Esto tiene las siguientes implicaciones:

- **Tipo de clientes:** a diferencia del modelo B2C, no se le vende a personas sino a negocios, que pueden ser grandes multinacionales, pequeños y medianos negocios, instituciones de salud y educativas, o gubernamentales [17]. Es necesario adaptar los procesos a ellas y sus regulaciones.
- **Procesos de venta complejos:** de la mano del ítem anterior, para las empresas el proceso de venta suele ser más complejo y personalizado, ya que normalmente, del lado del cliente se requieren decisiones y aprobaciones internas. Se requiere un equipo de ventas capacitado que pueda asesorar y adaptar las ofertas a las necesidades específicas de cada cliente empresarial [18].
- **Relaciones a largo plazo:** al estar basadas en la confianza y el compromiso mutuo, las relaciones suelen ser más duraderas que las de un modelo de venta directa. Es particularmente importante por este motivo, dar un buen servicio postventa [17].
- **Precios negociables:** a diferencia del modelo de venta directa donde los precios suelen ser fijos para todos los clientes, los precios pueden negociarse para los clientes teniendo en cuenta factores como el volumen de compra, la relación comercial, entre otros [18].
- **Compras a gran escala:** las empresas suelen realizar compras a gran escala. Esto va de la mano con las relaciones a largo plazo, ya que a veces, se acuerda un alto volumen de compras distribuidas a lo largo de un período de tiempo pactado [17].
- **Personalización de productos y servicios:** dadas las necesidades de las empresas clientes, es posible ofrecer productos adaptados o personalizados para satisfacerlas.

Estas consideraciones implican que, por lo general, tras las negociaciones con los clientes, se tiene idea de las cantidades y productos que podrían comprar los clientes, lo cual a priori, permite facilitar el pronóstico de las ventas. Sin embargo, en este estudio no se contará con esta variable debido a las políticas de confidencialidad de la empresa que suministró los datos. Ahora bien, en este trabajo se busca abordar dos enfoques para predecir la demanda: modelos clásicos y de aprendizaje automático. Estos serán expuestos a continuación:

### *III-A. Modelos Clásicos*

- **Regresión lineal:** este modelo supone que la relación entre las variables es lineal o linealizable. Tiene una forma simple en la cual se emparejan 2 variables, y una múltiple en la cual influyen más de 2 variables y suele captar fenómenos más complejos [19]. Esta versión, se puede expresar como:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad (1)$$



donde  $y$  es la variable dependiente,  $X_1, X_2, \dots, X_p$  son las variables independientes,  $\beta_0$  es el intercepto,  $\beta_1, \beta_2, \dots, \beta_p$  son los coeficientes de regresión y  $\epsilon$  es el término de error. Por lo general, los coeficientes de regresión se estiman mediante el método de mínimos cuadrados (OLS) que minimiza la suma de los cuadrados de los residuos:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \dots - \beta_p X_{ip})^2.$$

Los métodos basados en este modelo suelen ser fáciles de implementar y a menudo logran hacer buenas estimaciones en series de tiempo bajo sus supuestos.

- **Holt-Winters:** este es un método de suavizado exponencial que se utiliza ampliamente para la predicción de series temporales con tendencia y estacionalidad. Este tiene dos variaciones: “el método aditivo es preferido cuando las variaciones estacionales son aproximadamente constantes a lo largo de la serie, mientras que el método multiplicativo es preferido cuando las variaciones estacionales cambian proporcionalmente al nivel de la serie” [7]. Tiene los siguientes componentes:

- Componente de nivel ( $l_t$ ): representa el nivel promedio de la serie en cada punto de tiempo y se ajusta conforme a nuevos datos.

$$L_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + T_{t-1})$$

donde  $\alpha$  es el parámetro de suavizado para el nivel.

- Componente de tendencia ( $b_t$ ): captura la tendencia de crecimiento o decrecimiento en los datos a lo largo del tiempo. Este componente se actualiza en cada paso temporal para reflejar cambios en la tendencia.

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$$

donde  $\beta$  es el parámetro de suavizado para la tendencia.

- Componente estacional ( $s_t$ ): ajusta los patrones que se repiten en intervalos regulares, permitiendo capturar la estacionalidad en los datos. Dependiendo de la naturaleza de la estacionalidad, se puede elegir entre el método aditivo o multiplicativo.

$$S_t = \gamma(y_t - L_t) + (1 - \gamma)S_{t-s}$$

para el método aditivo, o

$$S_t = \gamma \frac{y_t}{L_t} + (1 - \gamma)S_{t-s}$$

para el método multiplicativo, donde  $\gamma$  es el parámetro de suavizado para la estacionalidad y  $s$  es la longitud del período estacional.

Teniendo en cuenta lo anterior, un modelo Holt-Winters se puede describir como:

Holt-Winters ( $L_t, T_t, S_t$ )

Dado por la siguiente ecuación para el método aditivo:

$$\hat{y}_{t+h} = (L_t + hT_t) + S_{t-s+h} \quad (2)$$

y la siguiente para el método multiplicativo:

$$\hat{y}_{t+h} = (L_t + hT_t) \times S_{t-s+h} \quad (3)$$

donde  $h$  es el horizonte de predicción.

Para ajustar este modelo, se deben determinar los parámetros de suavizado para el nivel ( $\alpha$ ), la tendencia ( $\beta$ ), y la estacionalidad ( $\gamma$ ). Estos parámetros controlan la tasa a la cual los componentes se actualizan con cada nueva observación. El método de ajuste puede realizarse utilizando técnicas como la minimización del error cuadrático medio (MSE) entre los valores predichos y observados [7].

Este método es especialmente útil en situaciones donde se espera que tanto la tendencia como los patrones estacionales sean importantes para la precisión de las predicciones.

- **SARIMAX:** se puede definir como un modelo autorregresivo integrado de media móvil que incorpora estacionalidad y variables exógenas. Está conformado por los siguientes componentes [7],[20]:

- Componente autorregresivo (AR): en este, se predice la variable de interés utilizando una combinación lineal de valores históricos de la variable. Se representa por el parámetro  $p$  que, generalmente, puede obtenerse a partir del análisis de la gráfica PACF (Función de autocorrelación parcial). Está dado por:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t$$

donde  $\phi_i$  son los coeficientes autoregresivos y  $\epsilon_t$  es el error en  $t$ .

- Componente Media Móvil (MA): captura la relación entre el valor actual de la serie de tiempo y los términos de error pasados. Esta se representa por el parámetro  $q$ , que por lo general se obtiene a partir de un análisis de la gráfica ACF (Función de autocorrelación). Su ecuación es:

$$\epsilon_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} + \nu_t$$

donde  $\theta_i$  son los coeficientes de la media móvil y  $\nu_t$  es un error blanco.

- Componente integrado (I): el análisis de series de tiempo supone que las series son estacionarias, es decir, su media y varianza son constantes a lo largo del tiempo. Para convertir una serie temporal no estacionaria en una serie estacionaria, se debe realizar un proceso de diferenciación y aplicar una prueba de hipótesis para comprobar que

la serie obtenida es estacionaria. El parámetro  $d$ , representa la cantidad de veces que hay que diferenciar la serie para que sea no estacionaria.

- Componente estacional (S): ajusta los patrones que se repiten en intervalos regulares más largos que el típico ciclo ARIMA. Cuenta con parámetros  $(P, D, Q)_s$ , donde de manera análoga,  $P$  es el orden del componente autorregresivo estacional (SAR),  $Q$  es el orden del componente de medias móviles (SMA),  $D$  es el número de diferenciaciones estacionales necesarias para hacer la serie temporal estacionalmente estacionaria, y  $s$  es la longitud del período estacional. Sus ecuaciones son:

$$y_t = \Phi_1 y_{t-s} + \Phi_2 y_{t-2s} + \dots + \Phi_P y_{t-Ps} + \epsilon_t$$

para la parte autoregresiva estacional, y

$$\epsilon_t = \Theta_1 \epsilon_{t-s} + \Theta_2 \epsilon_{t-2s} + \dots + \Theta_Q \epsilon_{t-Qs} + \nu_t$$

para la parte de media móvil estacional, donde  $\Phi_i$  y  $\Theta_i$  son los coeficientes estacionales.

- Variables exógenas (X): permite la inclusión de variables adicionales que pueden influir en la variable dependiente pero que no son modeladas directamente por el componente temporal del modelo [7]. Unas posibles variables exógenas podrían ser las fechas de lanzamiento de nuevas líneas/cambios generacionales de productos, precio de ciertos componentes clave en el mercado, entre otros.

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \dots + \beta_k x_{k,t} + \epsilon_t$$

donde  $x_{i,t}$  son las variables exógenas y  $\beta_i$  sus coeficientes.

Teniendo en cuenta lo anterior, un modelo SARIMAX se puede describir como:

$$\text{SARIMAX } (p, d, q) (P, D, Q)_s$$

$$\hat{y}_{t+h} = \mathbf{x}_{t+h}^T \boldsymbol{\beta} + \sum_{i=1}^p \phi_i \hat{y}_{t+h-i} + \sum_{j=1}^q \theta_j \epsilon_{t+h-j} + \sum_{k=1}^P \Phi_k \hat{y}_{t+h-ks} + \sum_{l=1}^Q \Theta_l \epsilon_{t+h-ls} \quad (4)$$

Para encontrar los hiperparámetros del modelo, más allá del análisis de correlaciones, es posible utilizar el método de gridsearch [20]. Este consiste en realizar validación cruzada buscando la combinación de hiperparámetros que minimice una métrica como el criterio de información de Akaike promedio de los pliegues para cada combinación.

- Criterio de Información de Akaike (AIC): se trata de un indicador estadístico empleado para evaluar y comparar la eficacia de varios modelos. Un valor más bajo indica que el modelo tiene mayor calidad. Además, este indicador incluye una

penalización por el uso de un número elevado de parámetros, lo cual es bueno para prevenir el sobreajuste de los modelos [21].

### III-B. Modelos de Aprendizaje Automático

- **Random Forest:** antes de hablar de este modelo, es importante mencionar su base: los árboles de decisión. Estos son modelos no paramétricos que utilizan una estructura similar a la de un árbol, donde cada nodo interno representa una prueba sobre un atributo, cada rama representa el resultado de la prueba, y cada nodo hoja representa una clase o valor de la variable objetivo para tomar decisiones con base en los datos de entrada [22], como se puede ver en la figura 1 a modo de ejemplo.

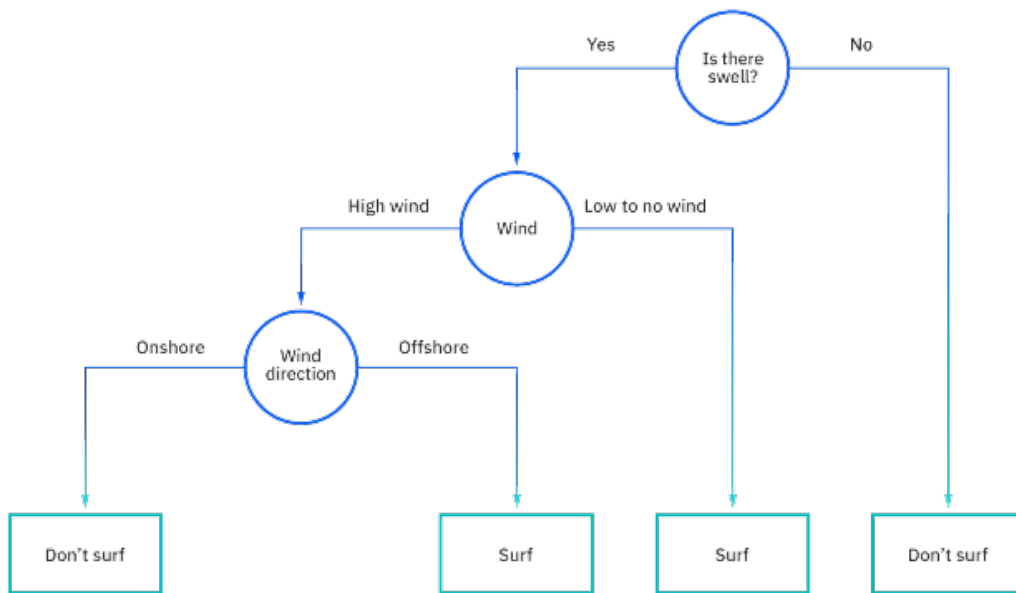


FIG. 1: Representación gráfica de un árbol de decisión [22].

El algoritmo random forest inicia creando múltiples subconjuntos de los datos, asignando un árbol de decisión a cada uno. Aleatoriamente, se selecciona un subconjunto de características en cada nodo, lo cual reduce el sobreajuste. A continuación, cada árbol se entrena, y en un contexto de regresión, se produce una predicción. Finalmente, se promedian las predicciones de todos los árboles para producir el resultado final, como se ilustra en la figura 2.

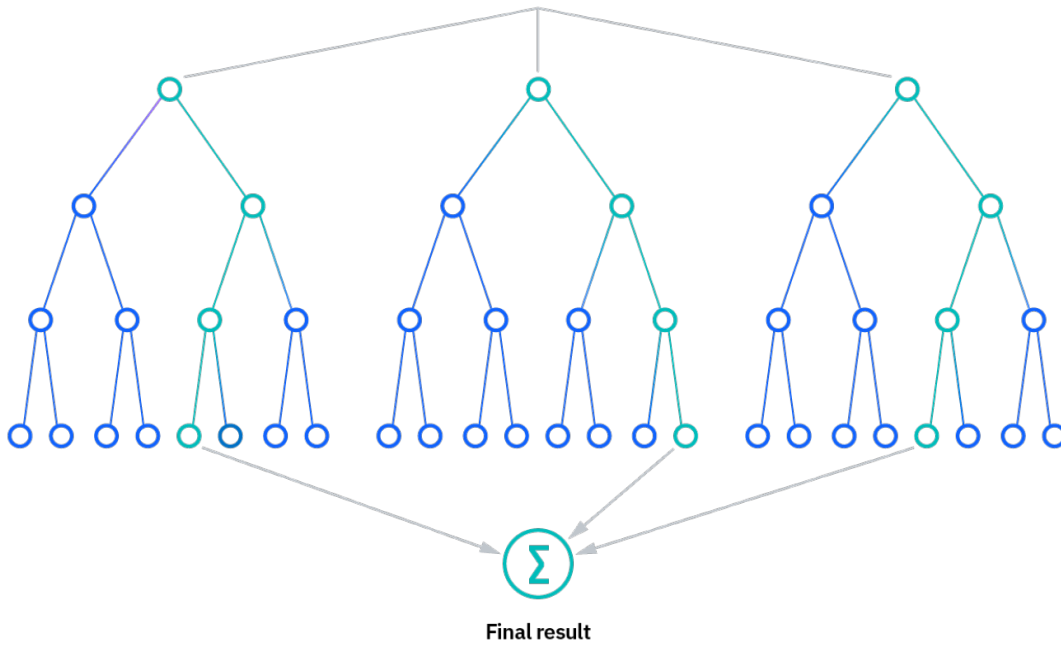


FIG. 2: Representación gráfica del modelo Random Forest [22].

Este modelo se caracteriza por desempeñarse bien en conjuntos de datos con alta dimensionalidad y con múltiples datos faltantes. En este caso de estudio, también es muy útil porque puede manejar datos de demanda histórica [13],[22].

- **Gradient Boosting:** este es un algoritmo basado en la combinación de múltiples modelos débiles, también conocidos como “weak learners”, que son modelos con un rendimiento que apenas supera el azar. Este proceso se realiza de forma iterativa, donde cada modelo corrige los errores de su predecesor, resultando en un modelo fuerte [23]. Este enfoque se conoce como “boosting”, y ha demostrado ser una herramienta altamente efectiva para realizar predicciones en series de tiempo o series de datos que no siguen un comportamiento lineal, aunque corre el riesgo de caer en sobreajuste si no está bien regularizado. Algunas técnicas de regularización comunes con la poda de árboles, L1 y L2, entre otras [13].
- **LSTM:** una red neuronal es un conjunto de algoritmos basados en el funcionamiento del cerebro humano que se han utilizado en problemas de clasificación, regresión, entre otros. Básicamente, está compuesta por capas de neuronas interconectadas que reciben entradas, las transforman y producen salidas. Un tipo de red neuronal es la RNN (Red Neuronal Recurrente), que es útil para procesar datos secuenciales debido a su capacidad para mantener una memoria de estados anteriores a través de bucles recurrentes. Sin embargo, estas tienen limitaciones cuando se busca capturar información a largo plazo. Por este motivo, surgió el modelo Long Short-Term Memory que incluye celdas de memoria y mecanismos de puertas que permiten superar estas limitaciones [24], lo cual lo hace viable para el problema de predicción de la demanda, ya que le hace capaz de aprender patrones temporales.

- **Métodos de ensamble:** son técnicas que combinan múltiples modelos para mejorar el rendimiento general de las predicciones. Los más comunes son:
  - Bagging: se basa en reducir la varianza a través de la generación de múltiples subconjuntos del conjunto de datos original mediante muestreo con reemplazo y entrenando un modelo en cada subconjunto. Finalmente, se obtienen predicciones a partir de agregar mediante promedios (en el caso de la regresión) las predicciones de cada modelo. Un ejemplo es el mismo Random Forest mencionado previamente.
  - Boosting: reduce el sesgo mediante mejora secuencial de los modelos, siendo el gradient boosting uno de los más famosos.
  - Stacking: se centra en reducir el error mediante la combinación de modelos base; aquí, las predicciones de los modelos base se utilizan como características de entrada para el modelo objetivo.

Al reducir los factores anteriores, los métodos de ensamble reducen el overfitting. Por lo general, estos métodos presentan mejores resultados que sus componentes de forma individual [25].

- **eXtreme Gradient Boosting (XBoost):** es un método de ensamble que implementa árboles de decisión con el método de gradient boosting. Está diseñado para aumentar la velocidad y el rendimiento implementando técnicas de regularización que ayudan a prevenir el sobreajuste y utiliza optimizaciones de hardware para acelerar el proceso de entrenamiento. Además, incluye características como la paralelización del árbol de decisiones y el manejo eficiente de valores faltantes, lo cual lo ha convertido en un modelo bastante destacado para realizar predicciones en series de tiempo al poder captar una gran cantidad de información de una forma computacionalmente eficiente [26].

### III-C. Métricas de Evaluación

- **Error Cuadrático Medio (MSE):** Es un indicador comúnmente usado para evaluar la precisión de las predicciones. Se calcula como la media de los cuadrados de los errores de predicción.

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

Entre menor sea el MSE, mejor es el ajuste del modelo.

- **MAPE:** el error porcentual medio absoluto es otra métrica utilizada para evaluar la precisión en las predicciones. Es una de las métricas más utilizadas en problemas de regresión aplicados a series de tiempo, especialmente por su facilidad a la hora de interpretar los resultados al ser presentada como un porcentaje de error.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Naturalmente, valores menores del MAPE indican un mayor ajuste del modelo.

- **Coefficiente de determinación  $R^2$ :** indica la proporción de la varianza de la variable dependiente que es explicada por el modelo. Es una métrica muy utilizada en problemas de regresión, y entre mayor es, mejor es el modelo:

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

## IV. METODOLOGÍA

Se utilizará una metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) adaptada al objetivo de predecir la demanda en una empresa B2B del sector de Hardware comparando modelos tradicionales de series de tiempo con un enfoque basado en aprendizaje automático. A continuación, se describen las fases que se seguirán en este proyecto:

### IV-A. *Comprensión del Negocio*

En el marco de este proyecto, se subraya la importancia de la precisión de las predicciones de ventas para optimizar la planificación de la cadena de suministro y mejorar la satisfacción de los clientes a la par de reducir costos de inventario. Por lo tanto, en esta fase inicial se buscará entender el contexto del mercado de hardware en un modelo B2B, buscando entender los factores y variables clave para cumplir el objetivo de la minería de datos de desarrollar de manera comparativa modelos predictivos de series de tiempo y aprendizaje automático para poder predecir las ventas.

### IV-B. *Entendimiento de los datos*

Se contará con una serie de tiempo correspondiente a transacciones individuales de ventas entre 2021 y 2024 en México obtenidos a partir del sistema de información de la empresa, que permite descargar la información en formato CSV. Tras realizar el proceso de anonimización necesario, se realizará un análisis exploratorio con el objetivo de detectar patrones de tendencia y estacionalidad que puedan ser utilizados en el modelado para mejorar la precisión en las predicciones.

#### *IV-C. Preparación de los datos*

En esta etapa, se transformará los datos a la frecuencia más adecuada (semanal o mensual), se eliminarán datos atípicos y en general, se depurará el dataset para facilitar la implementación de los modelos. Así mismo, se verificará si la serie en cuestión cumple supuestos como la estacionariedad, que serán tenidos en cuenta en el modelado de algunos modelos y en ciertos casos pueden implicar la necesidad de llevar a cabo transformaciones al conjunto de datos. Finalmente, se realizará la partición adecuada del dataset en conjuntos de entrenamiento y prueba, de forma que no se pierdan las propiedades temporales de la serie.

#### *IV-D. Modelado*

En esta fase, se seleccionarán algunos de los modelos tradicionales y de aprendizaje automático de referencia, buscando cumplir el objetivo general del proyecto, entre los cuales se incluyen Holt-Winters, SARIMAX, random forest, XGBoost, métodos de ensamble, entre otros. Estos se entrenarán utilizando el conjunto de datos de entrenamiento de acuerdo con las condiciones de cada método.

#### *IV-E. Evaluación*

Se analizará el rendimiento de los modelos en el conjunto de prueba. Con esto, se contará con métricas de evaluación en ambos conjuntos, lo cual posteriormente servirá para el análisis comparativo del desempeño de los modelos. En este caso, se utilizarán métricas como el Error cuadrático medio (MSE), el error porcentual medio absoluto (MAPE) y el coeficiente de determinación ( $R^2$ ).

#### *IV-F. Despliegue*

En este trabajo no se tiene contemplado realizar un despliegue formal a producción. El alcance del mismo se limita a un ámbito académico que será documentado detalladamente.

### **V. PLAN DE GESTIÓN DE DATOS**

Para el desarrollo de este proyecto, se usarán datos históricos de ventas en una empresa de hardware que sigue el modelo Business-to-Business entre los años 2021 y 2024. Con el fin de proteger la privacidad tanto de la empresa, como de sus clientes, se anonimizará o eliminará todo tipo de información que pueda ser considerada sensible o confidencial desde el momento en que se reciban los datos. Estos, serán utilizados exclusivamente por el autor del proyecto, y a ellos podrá tener acceso la directora de este cuando lo requiera para el cumplimiento de sus funciones, manteniendo las políticas de confidencialidad de



la empresa en la cual se prohíbe su divulgación. Finalmente, se añade que estos sólo se usarán para el fin académico descrito en este documento.

## VI. ASPECTOS ÉTICOS

En el contexto de este proyecto, los datos suministrados por la empresa se emplearán únicamente para las tareas especificadas en este documento, necesarias para alcanzar los objetivos planteados. Asimismo, los modelos y los resultados derivados del desarrollo del proyecto se utilizarán exclusivamente para los propósitos definidos aquí. Se obtendrán los datos desde el sistema de información de la empresa, garantizando la anonimización de toda información sensible y confidencial.

## REFERENCIAS

- [1] A. Murphy y H. Tucker, *The Global 2000*, <https://www.forbes.com/lists/global2000/?sh=a595525ac042>, 2023.
- [2] D. Jawaid, *100 Biggest Technology Companies in the World*, <https://finance.yahoo.com/news/100-biggest-technology-companies-world-175211230.html>, 2023.
- [3] T. Boone, R. Ganeshan, A. Jain y N. Sanders, «Forecasting sales in the supply chain: Consumer analytics in the big data era,» *International Journal of Forecasting*, págs. 170-180, 2019.
- [4] Y. Liu, L. Feng y B. Jin, «Future-Aware Trend Alignment for Sales Predictions,» *Information*, vol. 11, 2020. DOI: 10.3390/info11120558.
- [5] J. T. Mentzer y M. A. Moon, *Sales forecasting management: a demand management approach*. Sage Publications, 2004.
- [6] N. Sanders y K. Manrodt, «Forecasting Practices in US Corporations: Survey Results,» *Interfaces*, págs. 92-100, 1994.
- [7] R. Hyndman y G. Athanasopoulos, *Forecasting: Principles and Practice*. Monash University, 2013, Accessed: April 2024.
- [8] G. Box, S. Hilmer y G. Tiao, *Seasonal Analysis of Economic Time Series*. National Bureau of Economic Research, 1978.
- [9] B. Pavlyshenko, «Machine-Learning Models for Sales Time Series Forecasting,» en *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, 2019, Lviv: Data Stream Mining & Processing (DSMP). DOI: 10.3390/data4010015.
- [10] Z. Marzak, R. Benabbou, S. Mouatassim y J. Benhra, «Forecasting Seasonal and Trend-Driven Data: A Comparative Analysis of Classical Techniques,» *Journal of Optimization in Industrial Engineering*, vol. 16, n.º 35, págs. 49-62, 2023, ISSN: 2251-9904. DOI: 10.22094/joie.2023.1984123.2057. eprint: sanad.iau.ir/fa/Article/Download/951128. dirección: sanad.iau.ir/fa/Article/951128.
- [11] R. Puspita y L. A. Wulandhari, «Hardware sales forecasting using clustering and machine learning approach,» *IAES International Journal of Artificial Intelligence*, vol. 11, n.º 3, págs. 1074-1084, 2022, Cited by: 2; All Open Access, Gold Open Access. DOI: 10.11591/ijai.v11.i3.pp1074-1084. dirección: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133124127&doi=10.11591%2fijai.v11.i3.pp1074-1084&partnerID=40&md5=7a630d771ebaf44c001ada0ad57c43e7>.
- [12] D. Rajpoot, B. Mittal, H. Dudani y U. Singhal, «Sales Analysis and Forecasting using Machine Learning Approach,» en *ACM International Conference Proceeding Series*, 2023. DOI: 10.1145/3607947.3608032.
- [13] M. Bohanec, M. Kljajić Borštnar y M. Robnik-Šikonja, «Explaining machine learning models in sales predictions,» *Expert Systems with Applications*, vol. 71, págs. 416-428,

- 2017, Cited by: 90. DOI: 10.1016/j.eswa.2016.11.010. dirección: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85006817279&doi=10.1016%2fj.eswa.2016.11.010&partnerID=40&md5=fb41510228cba5c608119b4c8b1cdd08>.
- [14] O. Wisesa, A. Andriansyah y O. Khalaf, «Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm,» *Majlesi Journal of Electrical Engineering*, vol. 14, págs. 145-153, 2020. DOI: 10.1109/BCWSP50066.2020.9249397.
  - [15] M. Wang, S. Tong, Q. Pei, J. Ge, Y. Liu y J. Sun, «Sales Forecasting Model Based on Ensemble Learning and Its Application in Anomaly Detection,» Cited by: 0, 2023, págs. 156-161. DOI: 10.1145/3585542.3585566. dirección: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163837018&doi=10.1145%2f3585542.3585566&partnerID=40&md5=9f72a663944a56aa219c5a839882bd46>.
  - [16] P. Turnbull, «Business-to-Business Marketing: Organizational Buying Behaviour,» en *The Marketing Book*, M. Baker, ed., Oxford, 1994, págs. 216-217.
  - [17] M. Hutt y T. Speh, *Business Marketing Management: B2B*. South Western CENGAGE Learning, 2012, págs. 11, 93-94.
  - [18] N. Ellis, *Business-to-Business Marketing: Relationships, Networks & Strategies*. Oxford University Press, 2010.
  - [19] R. Montero Granados, «Modelos de regresión lineal múltiple,» Universidad de Granada, España, Documentos de Trabajo en Economía Aplicada, 2016.
  - [20] J. Brownlee, *Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future*. Machine Learning Mastery, 2020, págs. 182-248, Accessed: April 2024.
  - [21] C. Miranda, *Modelización de Series Temporales modelos clásicos y SARIMA*, [https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM\\_MIRANDA\\_CHINLLI\\_CARLOS.pdf](https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_MIRANDA_CHINLLI_CARLOS.pdf), Accessed: May 2024, 2021.
  - [22] IBM, *Random Forest*, Accessed: May 2024, 2024. dirección: <https://www.ibm.com/es-es/topics/random-forest>.
  - [23] G. Biau y B. Cadre, «Optimization by gradient boosting,» *ArXiv*, 2017. DOI: 10.1007/978-3-030-73249-3\_2.
  - [24] C.-L. Hung, «Chapter 11 - Deep learning in biomedical informatics,» en *Intelligent Nanotechnology*, 2023, págs. 307-329. DOI: 10.1016/B978-0-323-85796-3.00011-1.
  - [25] Y. Aun, Y.-M. Khaw, M. Gan y D. Chern, «A Machine-Learning Ensemble Method for Temporal-aware Sales Forecasting,» en *2022 5th International Conference on Data Science and Information Technology, DSIT 2022 - Proceedings*, 2022. DOI: 10.1109/DSIT55514.2022.9943925.
  - [26] C. Bentéjac, A. Csörgő y G. Martínez-Muñoz, *A Comparative Analysis of XGBoost*, [https://www.researchgate.net/publication/337048557\\_A\\_Comparative\\_Analysis\\_of\\_XGBoost/citations](https://www.researchgate.net/publication/337048557_A_Comparative_Analysis_of_XGBoost/citations), Accessed: April 2024, 2019.

## VII. ANEXOS

### VII-A. Carta

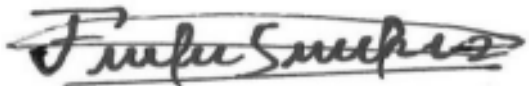
Medellín, jueves 30 de mayo de 2024

**Coordinador Maestría en Ciencias de los Datos y Analítica**  
La Universidad

Estimado coordinador,

A continuación, presento el proyecto titulado: “**Predicción de ventas para una empresa de Hardware Business-To-Business**” para que este sea considerado como trabajo de maestría en la modalidad de profundización. Este proyecto ha sido revisado por la directora **Paula María Almonacid Hurtado** y recibe su aval para ser presentado al comité. De antemano agradezco su colaboración y trámite respectivo ante el comité.

Atentamente,



---

Hernán Felipe Sánchez Cárdenas  
Código: 1000009318



---

Paula María Almonacid Hurtado  
palmona1@eafit.edu.co