**Logotipo

Descripción generada automáticamente**

**Predicción de Tiempos de Entrega en la Cadena de Suministro: Una Solución de Machine Learning para la Gestión de los tiempos de entrega.**

Andres Esteban Fonseca Galviz

Juan Felipe Quinto Rios

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Fonseca Galviz & Quinto Rios, 2023) |
| **Referencia**  **Estilo APA 7 (2020)** | Fonseca Galviz, A. F, & Quinto Rios, J. F. (2023). *Predicción de Tiempos de Entrega en la Cadena de Suministro: Una Solución de Machine Learning para la Gestión de los tiempos de entrega.* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteIV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Tabla de contenido**

[Resumen 8](#_heading=h.3rdcrjn)

[Abstract 9](#_heading=h.26in1rg)

[1. Descripción del problema 10](#_heading=h.35nkun2)

[1.1. Problema de negocio 10](#_heading=h.1ksv4uv)

[1.2. Aproximación desde la analítica de datos 11](#_heading=h.44sinio)

[1.3. Origen de los datos 11](#_heading=h.z337ya)

[1.4. Métricas de desempeño 12](#_heading=h.3j2qqm3)

[2. Objetivos 13](#_heading=h.4i7ojhp)

[2.1. Objetivo general 13](#_heading=h.2xcytpi)

[2.2. Objetivos específicos 13](#_heading=h.3whwml4)

[3. Datos 14](#_heading=h.2bn6wsx)

[3.1. Datos originales 14](#_heading=h.qsh70q)

[3.2. Datsets 16](#_heading=h.3as4poj)

[3.3. Analítica descriptiva 17](#_heading=h.1pxezwc)

[4. Referencias 30](#_heading=h.49x2ik5)

**Siglas, acrónimos y abreviaturas**

**ERP** Enterprise Resource Planning

**I+D** Investigación y Desarrollo

**MAE** Mean Absolute Error (Error Absoluto Medio)

**MSE** Mean Squared Error (Error Cuadrático Medio)

**R²** Coeficiente de Determinación

**SQL** Structured Query Language

**CSV** Comma-Separated Values

**IQR** Interquartile Range (Rango Intercuartílico)

**XGBoost** Extreme Gradient Boosting

# Resumen

El presente trabajo aborda el desarrollo de un modelo predictivo para estimar los tiempos de entrega de materias primas importadas, con el objetivo de reducir los paros de producción ocasionados por desabastecimientos. Se utilizó un conjunto de datos históricos provenientes del sistema ERP de la empresa, que incluye información detallada sobre órdenes de compra, tiempos de entrega y características logísticas. Los datos fueron procesados, transformados y filtrados para asegurar su calidad y adecuación al análisis.

La estrategia metodológica incluyó el uso de modelos de aprendizaje automático, como Random Forest y XGBoost, los cuales destacan por su capacidad de manejar relaciones complejas y datos no distribuidos de manera normal. Se realizó un análisis exploratorio para identificar patrones, estacionalidades y variables clave, como país de origen, método de envío y categoría de proveedor.

Entre los principales obstáculos encontrados se destacan la presencia de valores extremos y la falta de normalidad en los datos, los cuales fueron abordados mediante técnicas de manejo de outliers y transformaciones específicas. Los resultados evidenciaron la efectividad del modelo para predecir los tiempos de entrega con una precisión significativa, lo que permitirá al área de compras anticipar retrasos y optimizar la planificación de la producción.

*Palabras clave*: cadena de suministro, predicción, machine learning, tiempos de entrega.

# Abstract

This study focuses on developing a predictive model to estimate delivery times for imported raw materials, aiming to reduce production stoppages caused by supply shortages. Historical data from the company’s ERP system, including detailed information on purchase orders, delivery times, and logistical characteristics, was used. The data was processed, transformed, and filtered to ensure its quality and relevance for analysis.

The methodological approach involved using machine learning models such as Random Forest and XGBoost, known for their ability to handle complex relationships and non-normally distributed data. Exploratory analysis was conducted to identify patterns, seasonality, and key variables such as country of origin, shipping method, and supplier category.

Key challenges included the presence of outliers and the lack of normality in the data, which were addressed through outlier handling techniques and specific transformations. The results demonstrated the model's effectiveness in accurately predicting delivery times, enabling the procurement department to anticipate delays and optimize production planning.

*Keywords***:** supply chain, prediction, machine learning, delivery times.

# Descripción del problema

La cadena de suministro es un pilar fundamental en los procesos de manufactura, ya que garantiza la disponibilidad continua de materias primas. La correcta estimación de los tiempos de entrega de las materias primas puede prevenir problemas críticos, como paros en la producción debido a la falta de abastecimiento. Por ello, se plantea un proyecto cuyo objetivo es desarrollar modelos predictivos para estimar los tiempos de entrega. Estos modelos podrían ayudar a mitigar las interrupciones en el proceso de fabricación provocadas por la escasez de materias primas, especialmente las importadas, que suelen tener tiempos de entrega más largos y afectan significativamente la cadena de suministro.

Este problema no es único para esta empresa, ya que situaciones similares se presentan en otros sectores. Por ejemplo, el estudio de [1] desarrolló modelos para predecir los tiempos de entrega en la cadena de suministro de piezas aeronáuticas, mejorando la precisión y reduciendo la incertidumbre en la disponibilidad de los materiales. De esta forma, se evidencia la efectividad de los modelos predictivos para enfrentar los desafíos en la cadena de suministro.

Un ejemplo adicional es el desarrollo de modelos de machine learning para abordar el incumplimiento en la entrega de pedidos farmacéuticos a pacientes, un aspecto crítico que afecta la adherencia terapéutica. Estos modelos clasifican los pedidos según su probabilidad de cumplimiento, mejorando la eficiencia en la entrega de medicamentos. Los modelos empleados incluyen regresión logística, Random Forest y Gradient Boosting, logrando resultados significativos al identificar factores clave que influyen en el cumplimiento de los pedidos [2]

## Problema de negocio

Actualmente, la empresa enfrenta desafíos significativos relacionados con la cadena de abastecimiento de materias primas importadas. Esta situación ha provocado interrupciones recurrentes en los procesos de producción. Según el informe de rendimientos y paros de producción de la compañía [1], la causa de paro más frecuente ha sido la falta de disponibilidad de materias primas en el momento necesario. Durante 2022, se registraron 724 eventos de baja disponibilidad de materias primas, lo que representó el 20% de los eventos de bajo rendimiento, y esta cifra aumentó a 768 en 2023, constituyendo el 31% del total de eventos. Estas dificultades han afectado de manera significativa la eficiencia operativa de la empresa, limitando su capacidad para mantener una producción continua y eficiente.

De acuerdo con la información extraída de la base de datos del maestro de piezas de la empresa, una parte significativa de la materia prima comprada [3] proviene de importaciones. Esta dependencia de proveedores externos incrementa la incertidumbre en los tiempos de entrega y en la disponibilidad de los materiales, lo cual subraya la necesidad de una intervención efectiva para minimizar riesgos y mejorar la planificación. Ante esta situación, el análisis se enfoca en las materias primas importadas, ya que son esenciales para el negocio debido a su naturaleza especializada, necesaria para la elaboración de ciertos productos. Además, estas materias primas presentan un costo elevado y generan un mayor impacto debido a sus tiempos de entrega, que suelen ser considerablemente mayores en comparación con las materias primas nacionales.

## Aproximación desde la analítica de datos

Para abordar este desafío, se propone desarrollar un modelo predictivo basado en técnicas de machine learning que permita estimar los tiempos de entrega de los proveedores de materias primas. Este modelo ayudará al área de compras a anticipar posibles retrasos y planificar acciones correctivas, asegurando así la alineación entre los programas de producción y la disponibilidad de materiales. De esta manera, se busca reducir el impacto de los retrasos y optimizar la eficiencia operativa.

En el contexto colombiano, ya se han documentado casos exitosos en la implementación de tecnologías similares. Por ejemplo, el grupo Nutresa [4] ha aplicado modelos de machine learning en sus procesos de compras a proveedores y en la gestión de inventarios, demostrando la eficacia de estas tecnologías en mejorar la gestión de la cadena de abastecimiento. Esta referencia sirve como un ejemplo inspirador de cómo se puede aprovechar la analítica avanzada para optimizar el desempeño de la cadena de suministro.

## Origen de los datos

Los datos utilizados para desarrollar el modelo provienen de registros históricos internos de la empresa sobre compras de materias primas, particularmente importadas. Estos datos incluyen información detallada sobre las órdenes de compra, los tiempos de entrega y las características de los proveedores. La información se obtuvo de las bases de datos del sistema ERP, que está consolidado en la [5].Esta base de datos proporciona una visión completa y detallada del historial de transacciones, lo cual es fundamental para el desarrollo y entrenamiento del modelo predictivo.

## Métricas de desempeño

El desempeño del modelo será evaluado mediante métricas de machine learning y métricas de negocio, permitiendo así una evaluación integral del rendimiento técnico y operativo del sistema.

Las métricas de machine learning, como el MAE (Error Absoluto Medio), el R2 (Coeficiente de Determinación) y el Mean Squared Error (MSE), permitirán evaluar la precisión y capacidad del modelo para explicar la variabilidad en los tiempos de entrega. Se buscará un MAE inferior a cinco días y un R2 superior a 0.85, con el fin de lograr un nivel adecuado de precisión y utilidad para el contexto de la cadena de suministro.

Por otro lado, el impacto en el negocio se evaluará mediante métricas que midan la reducción de los paros de producción relacionados con la falta de materias primas. Se espera una disminución de al menos un 25% en los eventos de desabastecimiento, lo cual mejoraría la continuidad de la producción y la eficiencia operativa de la empresa. Además, se establecerá un retorno de inversión (ROI) mínimo para justificar la implementación del modelo, estimando un ROI de al menos un 10% anual para cubrir los costos de operación y mantenimiento del sistema.

Estos valores mínimos se justifican por la necesidad de asegurar un impacto positivo en la operación de la empresa. Un MAE inferior a cinco días permitiría realizar ajustes oportunos en la planificación, mientras que un R2 superior a 0.85 proporciona una capacidad explicativa adecuada. Asimismo, la reducción del 25% en los eventos de desabastecimiento tendría un efecto significativo en la eficiencia operativa, mejorando la productividad y reduciendo los costos asociados a paros de producción.

# Objetivos

## Objetivo general

El proyecto tiene como objetivo desarrollar un modelo de predicción utilizando datos históricos de compras importadas, con el propósito de reducir los paros de producción causados por desabastecimiento de materia prima importada. Este modelo permitirá estimar los tiempos de entrega de los proveedores de materia prima, facilitando la alineación estratégica de los programas de producción con los plazos previstos de entrega de material importado.

## Objetivos específicos

* Analizar los datos de entregas de proveedores para detectar patrones y tendencias, e identificar las variables que afectan el tiempo de entrega.
* Analizar el desempeño de diferentes modelos de predicción para identificar aquel que se ajuste mejor a las características de los datos de la cadena de suministro y filtros definidos en la fase de exploración
* Optimizar los parámetros del modelo para maximizar su precisión y minimizar el margen de error en las predicciones de tiempos de entrega

# Datos

Se realizó una recopilación de datos de la fuente de proveedores y tiempos de entrega, seguida de una exploración de datos con análisis descriptivo. Este proceso permitió identificar las variables que tienen un impacto significativo sobre la variable de respuesta, que en este caso es la diferencia en días en el tiempo de entrega. El objetivo de esta sección es describir estos datos históricos en el contexto de la predicción de tiempos de entrega, brindando una base detallada que permita identificar patrones y relaciones clave que optimicen la precisión en las proyecciones de tiempos de suministro.

## Datos originales

El conjunto de datos original se descargó en un archivo CSV de 1607 KB de tamaño, que contiene 9859 registros y 20 columnas, ubicado en la [5]. Este conjunto de datos proporciona información detallada sobre cada solicitud de compra realizada a los proveedores, incluyendo datos como la ubicación y el país del proveedor, la categoría del producto solicitado y el método de envío utilizado. Dicho conjunto de datos fue construido a partir de información histórica de proveedores y tiempos de entrega, extraída de las bases de datos del sistema ERP (Enterprise Resource Planning) de la empresa en estudio.

Para consolidar la información distribuida en distintas tablas, se identificaron las relaciones existentes y se seleccionó la información relevante con el objetivo de generar un archivo CSV que contuviera las características descriptivas del fenómeno a predecir junto con la variable objetivo. Para ello, se elaboró una consulta SQL que integrará los datos necesarios, en [6].

La consulta incluyó tablas como "order\_Master", la cual contiene información esencial sobre las órdenes de compra de abastecimiento. En esta tabla, se filtraron los estados correspondientes a aquellas órdenes que ya habían llegado y completado el proceso de la cadena de suministro.

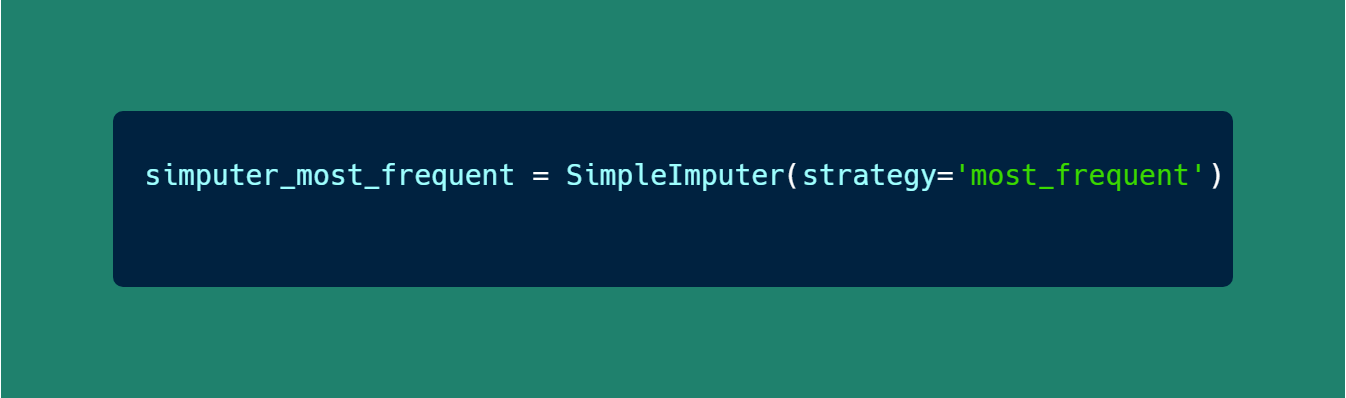
Posteriormente, se accedió a la tabla "Transaction\_History", en la que se registran las órdenes y las cantidades de ingreso. Con el propósito de evitar inconsistencias, se filtraron las órdenes con cantidad igual a cero, ya que podrían representar errores en el registro de la información.

La fecha de ingreso real del pedido se encuentra en la tabla "order\_Master", donde se registran las fechas de ingreso que corresponden a la llegada del material a la planta, y no al ingreso al inventario. Esto se debe a que los productos, al llegar a la planta, primero se ingresan a bodegas especiales donde se realiza un proceso de revisión de calidad. Además, se extrajo información del proveedor desde la tabla "vendor\_master", incluyendo detalles como el nombre del proveedor, país de origen y NIT.

Adicionalmente, de las tablas "Purchase\_Order\_Code", "Celsa\_Categorias", "Part\_Master" y "Part\_Master\_Ext" se extrajo información sobre la forma de envío, la moneda en la que se realizó el pago, la "unidad\_de\_medida/U" y la categoría del producto.

El dataframe requirió transformaciones adicionales, tales como la conversión de las columnas 'fecha\_ingreso\_inventario', 'fecha\_aproximada\_necesidad', 'fecha\_entrega\_real' y 'CreationDate\_proveedor' a tipo fecha. Estas transformaciones facilitaron el análisis temporal. Asimismo, se llevó a cabo un tratamiento de valores duplicados, identificando y eliminando 78 registros redundantes. Por otro lado, se identificaron 23 valores en blanco en la variable "unidad\_de\_medida/U", los cuales se transformaron a valores nulos para facilitar su identificación y tratamiento.

En el tratamiento de valores nulos, se encontraron los siguientes casos: 'forma\_envio' con 1 valor nulo, 'unidad\_de\_medida/U' con 23 valores nulos y 'terminos\_de\_pago' con 1 valor nulo. Al analizar los valores nulos en cada registro, se identificó que solo las columnas mencionadas presentaban valores faltantes, por lo que no se eliminaron dichos registros, ya que las demás columnas contenían información relevante y completa. Para manejar estos valores nulos se utilizó la técnica de "Simple Imputer" con la estrategia 'most\_frequent', que reemplaza el valor faltante de cada columna con el valor que más se repite [7].

**Figura 1** *Simpler imputer*

El acceso a los datos implicó una validación rigurosa de la información que sería de utilidad para el análisis. Se seleccionaron y verificaron los datos relevantes para garantizar su calidad y adecuación a los objetivos del estudio, eliminando aquellos elementos innecesarios o inconsistentes. Además, se debe considerar que los datos contienen nombres y ubicaciones que podrían ser potencialmente sensibles, así como cifras económicas y operacionales, por lo que se requiere un manejo cuidadoso para proteger la confidencialidad de dicha información.

## Datsets

Para la elaboración de este estudio, fue fundamental preparar de manera rigurosa los datos antes de realizar cualquier experimento de Machine Learning. A continuación, se describen en detalle los pasos realizados para el procesamiento y preparación del dataset, incluyendo la división en conjuntos de entrenamiento y validación.

En primer lugar, se procedió a filtrar las columnas irrelevantes o redundantes del dataset, con el propósito de mantener únicamente las variables que resultaran pertinentes para el análisis predictivo. Esto incluyó la eliminación de las columnas relacionadas con la fecha aproximada de necesidad, el día de la semana, el año aproximado y otros atributos que no aportaban valor significativo desde la perspectiva del modelo. Tras este filtrado, el dataset consistió en 735 registros y 8 variables relevantes, como se muestra en la siguiente descripción del dataset:

* **mes\_aproximado**: se transformó en un valor numérico (del 1 al 12) correspondiente al mes.
* **cantidad**: cantidad de unidades solicitadas, expresada como un valor numérico.
* **categoría** y **unidad\_de\_medida/U**: variables categóricas que se transformaron utilizando *One Hot Encoding*, lo cual permitió evitar problemas de multicolinealidad y trabajar con variables binarias.

Posteriormente, se aplicó una estandarización a las variables numéricas mediante la técnica de *Standard Scaling*. Este proceso fue crucial para asegurar que todas las variables numéricas estuvieran en una escala comparable, mejorando así el desempeño del modelo y evitando sesgos derivados de diferencias en las magnitudes de los datos.

Una vez preparado el dataset, se procedió a dividirlo en dos conjuntos de datos: uno para entrenamiento y otro para prueba, utilizando una proporción de 80% y 20%, respectivamente. Esta proporción es una práctica comúnmente aceptada en el ámbito del Machine Learning, ya que permite disponer de suficientes datos para entrenar al modelo sin comprometer la calidad de la evaluación. En total, el conjunto de entrenamiento comprendió 588 registros, mientras que el conjunto de prueba contó con 147 registros.

Este proceso de preparación fue esencial para garantizar que el modelo pudiera entrenarse de manera eficiente y minimizar los riesgos de sobreajuste, dado que los datos de prueba eran completamente independientes de los datos de entrenamiento. De esta manera, se logró asegurar una evaluación justa del modelo

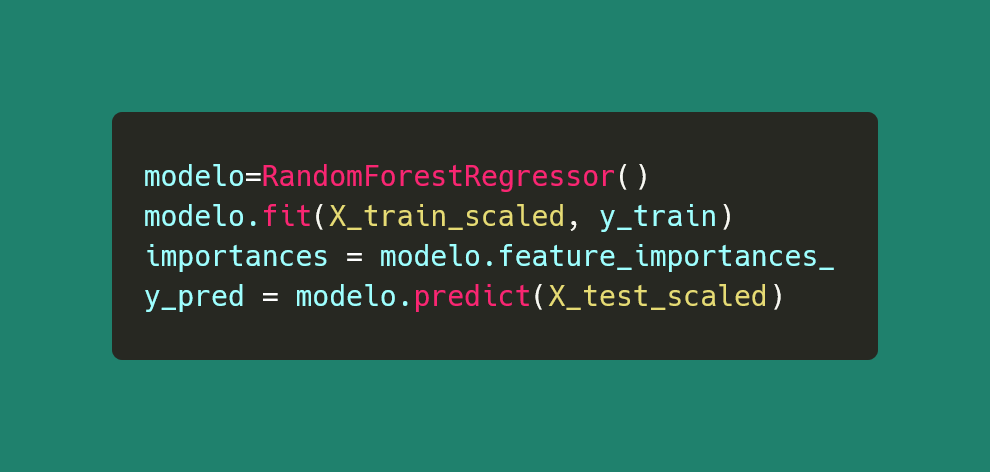
## Analítica descriptiva

El análisis de este conjunto de datos se centra en comprender las características principales del data set con el fin de identificar tendencias, irregularidades y posibles ajustes necesarios para el ajuste al modelo.

Inicialmente, se realizó un análisis preliminar para identificar las variables más importantes del conjunto de datos, enfocando el análisis exploratorio en aquellas que tienen un mayor impacto. Este enfoque se fundamenta en la literatura técnica orientada a enfoques similares al propuesto, como los trabajos de Steinberg [3] y GAM Escobar [2]. Estos autores emplearon modelos como Random Forest y técnicas de boosting, debido a su capacidad para predecir variables en entornos cambiantes, además de mejorar y ampliar la comprensión de las relaciones entre las variables.

Por esta razón, se han aplicado los modelos RandomForestRegressor y XGBRegressor utilizando la técnica de important feature el cual es una técnica que permite identificar las variables más relevantes en un modelo de aprendizaje automático, lo cual facilita determinar qué variables tienen un mayor impacto en el desempeño del modelo y así optimizar la selección de variables que puedan reducir la complejidad del conjunto de datos [8]. En la Figura 2 se presenta la implementación en Python de la técnica de feature importance, utilizada para identificar las variables más relevantes en los modelos de aprendizaje automático.

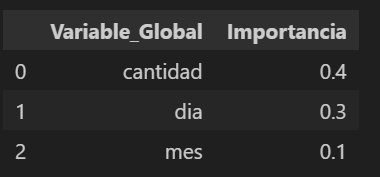
**Figura 2** *feature importance*



En el análisis de importancia de características, realizado mediante los modelos XGBRegressor y RandomForestRegressor, como se observa en la Figura 2 las variables país, Més, Dia, Cantidad, forma de envío y categoría de proveedor son las más influyentes y comunes en ambos modelos. Sin embargo, por conocimiento del negocio variables país, forma de envío y categoría, desempeñan un papel crítico en la predicción y el análisis de tiempos de entrega, ya que están directamente relacionadas con los aspectos logísticos y operativos que determinan la rapidez y fiabilidad de las entregas.

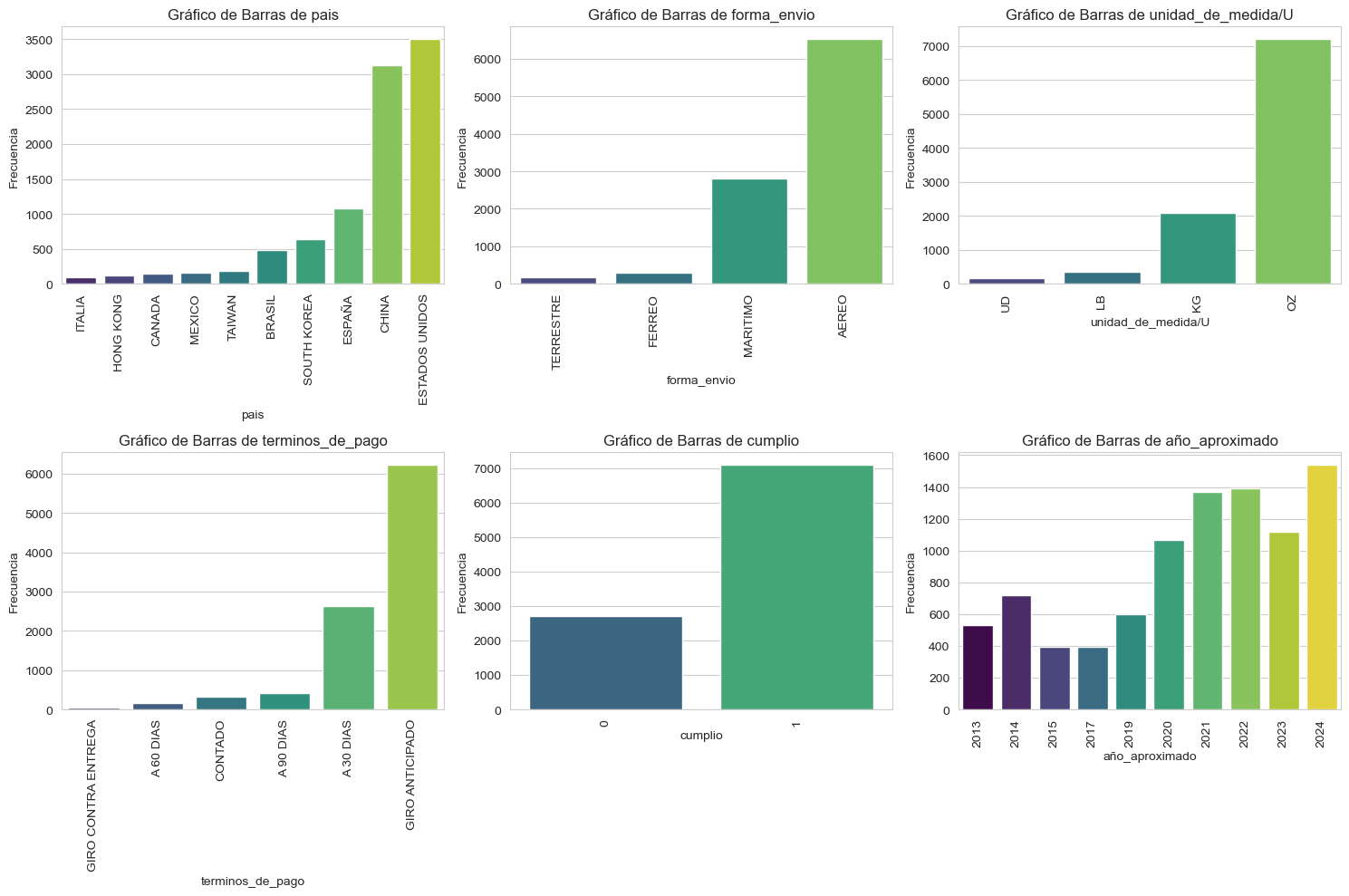
**Figura 3** *Tabla de Importancia Variables*

Modelo XGBRegressor Modelo RandomForestRegressor



Además del análisis de feature importance, se realiza un análisis de frecuencia de las principales variables identificadas como relevantes tales como pais, forma\_envio, unidad\_de\_medida/U, terminos\_de\_pago, cumplio, y año\_aproximado. En la figura 3, se presentan gráficos de barras que ilustran la distribución de estas variables en el dataset.

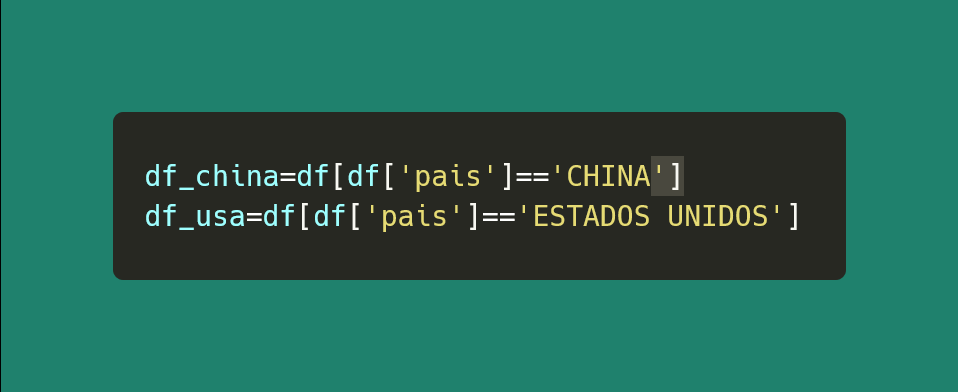
**Figura 3** *Análisis de Frecuencia*



Este análisis exploratorio revela patrones en las operaciones logísticas. El predominio de Estados Unidos y China como principales países, el uso prioritario del transporte aéreo, y la alta frecuencia de la unidad de medida OZ destacan como factores clave a considerar para optimizar la planificación logística y los tiempos de entrega, también, la alta frecuencia de Giro Anticipado se observa una preferencia por términos de pago que minimicen el riesgo. Además, el alto nivel de cumplimiento evidencia un desempeño eficiente en las entregas. Por último, el crecimiento en las operaciones anuales sugiere una expansión en el volumen logístico.

Al complementar este análisis con los gráficos exploratorios, se observa que los países Estados Unidos y China concentran la mayor cantidad de transacciones logísticas, lo que sugiere su importancia estratégica en el dataset. Por otro lado, la predominancia del transporte aéreo y las características específicas de las categorías de proveedores resaltan cómo las dinámicas logísticas varían significativamente entre regiones. Dado el análisis de feature importance donde se destacó las variables país, forma de envío y categoría de proveedor son las más influyentes se decide filtrar el dataset por países Estados Unidos y China ya que son los que concentran la mayor cantidad de transacciones, lo que sugiere que son los principales actores en las operaciones logísticas analizadas.

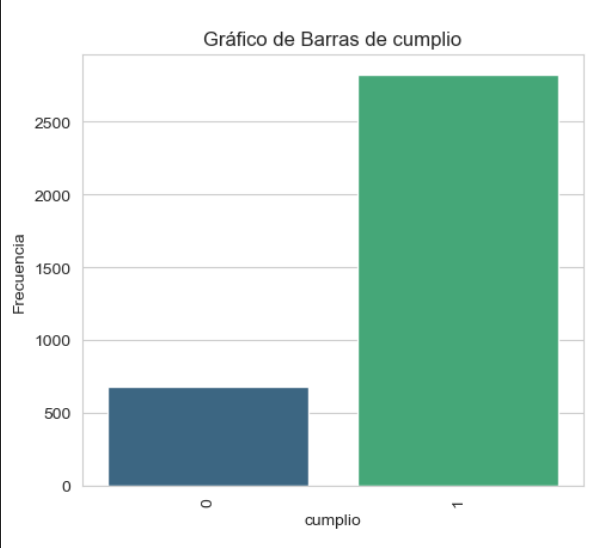
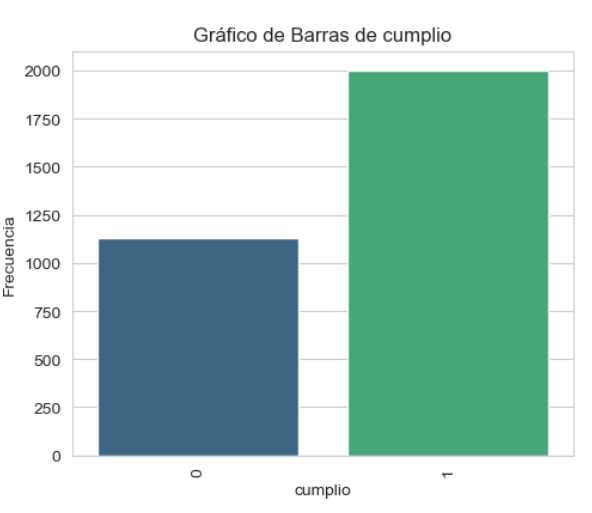
**Figura 5** *Filtro Por Países*



Por otro lado, se analiza el nivel de cumplimiento de las transacciones en ambos países filtrados. Este análisis permite identificar el país con mayores problemas logísticos, lo que facilita priorizar los esfuerzos de mejora en el lugar donde el impacto será más significativo.

**Figura 6** *Análisis de Cumplimiento*

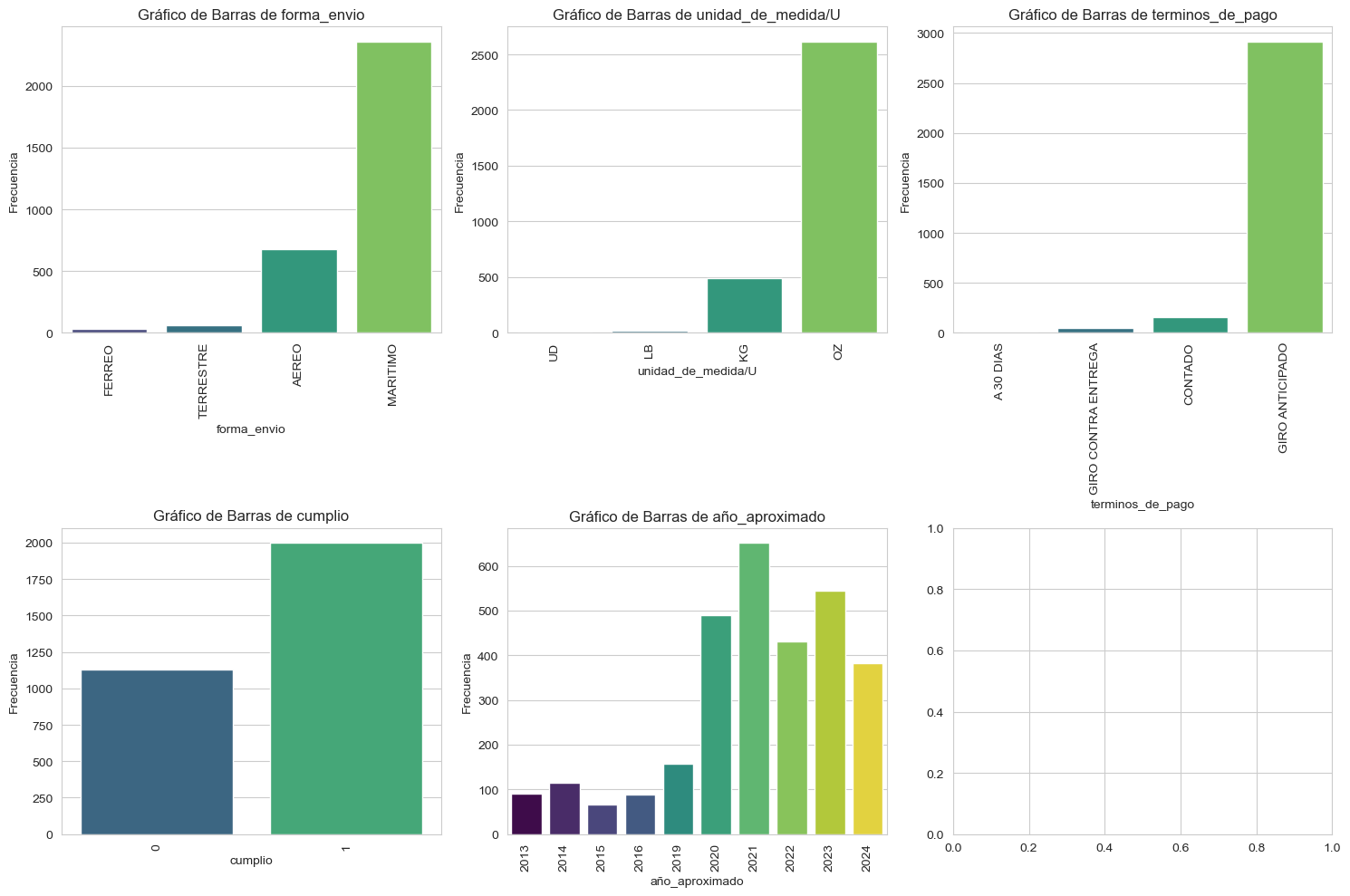
cumplimiento China cumplimiento Estados Unidos



Dado que China presenta un porcentaje más alto de incumplimientos en relación con su total de transacciones, sería más estratégico enfocar inicialmente el análisis y los esfuerzos de mejora en este país.

Como parte de este enfoque, se realizará un nuevo análisis de frecuencia sobre las principales variables, como forma de envío, categoría de proveedor, y unidad de medida. Este análisis permitirá, identificar patrones específicos en las transacciones de China, como la preferencia por ciertos métodos de envío o categorías de productos.

**Figura 7** *Análisis Frecuencia China*

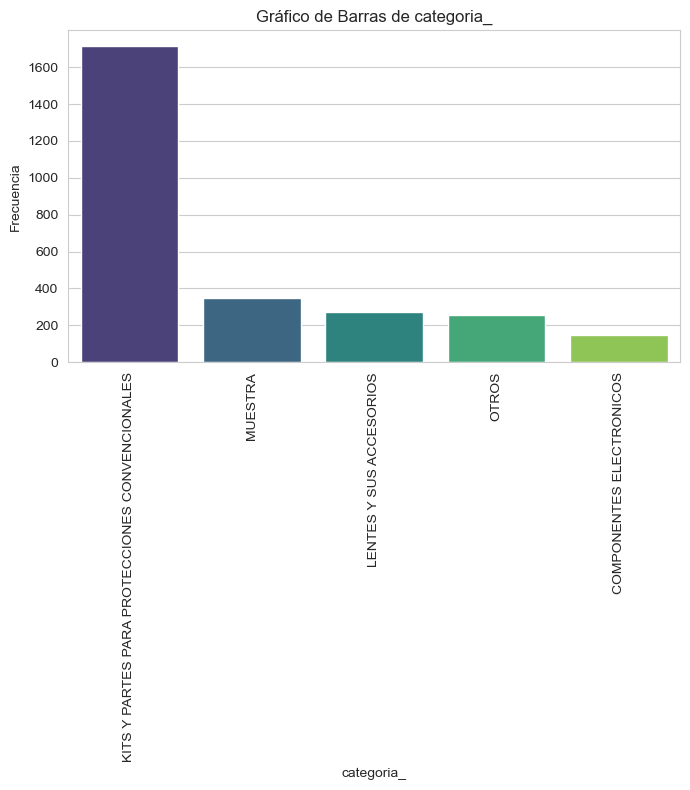


La preferencia por el envío marítimo sugiere que los costos son una consideración importante. Además, el uso de la unidad OZ indica un enfoque en productos de menor tamaño y peso. Por otro lado, el pago anticipado sugiere una política de bajo riesgo en términos financieros con los proveedores.

También, la mayoría de los envíos desde China cumplen con los tiempos o requisitos establecidos, con un 63.86% de cumplimiento. Este porcentaje sugiere que, en general, las operaciones con proveedores de China son confiables; sin embargo, el 36.14% de incumplimiento es relativamente alto, lo cual representa un área de oportunidad. Así mismo se observó un crecimiento significativo en las importaciones desde China a partir de 2020, seguido de una estabilización de los pedidos desde ese año.

Dado que las categorías de proveedor fueron identificadas como una de las variables más importantes en el análisis de feature importance, resulta pertinente hacer un análisis para comprender la distribución de categorías de productos, ya que tienen significancia en los tiempos de preparación, transporte y entrega.

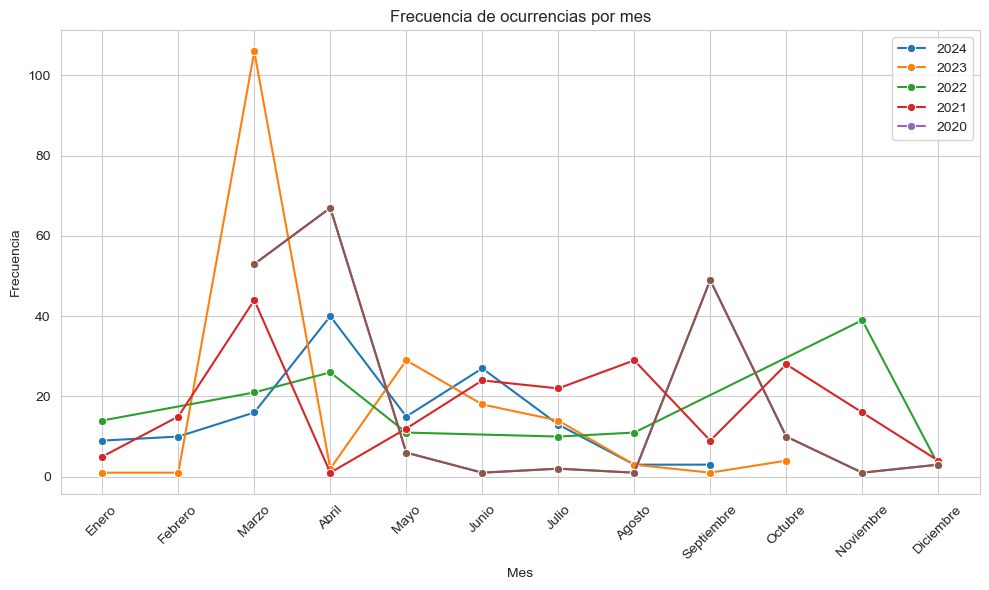
**Figura 8** *Análisis categorías productos China*



En el grupo de categorías "Kits y Partes para Protecciones Convencionales", "Muestra", "Lentes y sus Accesorios", "Otros" y "Electrónicos" se representa la mayor parte del total, por lo que el análisis debería priorizar el abastecimiento para estas categorías. Es importante hacer una salvedad con la categoría Muestra, ya que no representa material solicitado directamente para las órdenes de producción. Estas muestras son utilizadas por el área de I+D para experimentar en mejoras o nuevos productos.

El análisis estacional, basado en los meses del año, permitirá determinar si ciertos períodos tienen una mayor incidencia de incumplimientos.

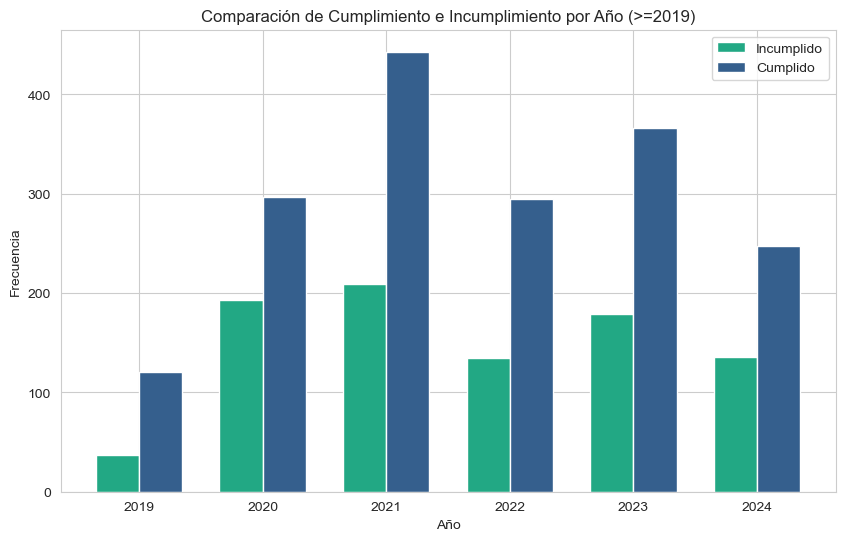
**Figura 9** *Tendencia Incumplimiento Meses*



En la figura 9 se observan picos notables de incumplimiento en ciertos meses, especialmente en marzo de 2023 y en abril de 2020. Sin embargo, cada año muestra una distribución diferente a lo largo de los meses. Por ejemplo, en 2020, se presenta otro pico en septiembre, mientras que en 2022 y 2024 la distribución es más equilibrada, con fluctuaciones menores a lo largo del año.

Algunos meses, como junio y diciembre, presentan bajas frecuencias de incumplimiento en la mayoría de los años.

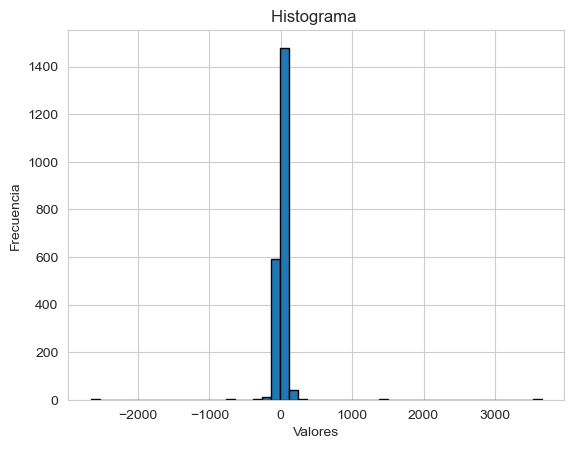
**Figura 10** *Cumplimiento por año*



Para el año 2021, se observa un pico de cumplimiento significativamente mayor en comparación con otros años. Sin embargo, en 2020 y 2021, la frecuencia de incumplimiento fue mayor en comparación con los demás años, mientras que en 2022 y 2024, las frecuencias de incumplimiento disminuyeron. A lo largo de los años, el incumplimiento se mantiene constante, lo que sugiere que existen entregas que no se logran completar a tiempo independientemente del año.

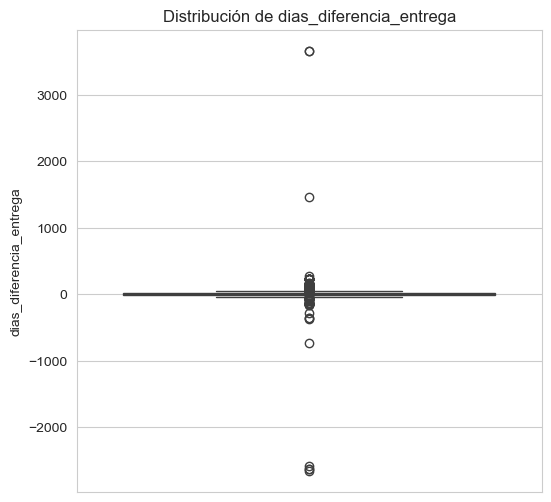
Analizar la variable días de diferencia de entrega con el fin de identificar si hay patrones recurrentes, y Analizar la dispersión de los valores y calcular métricas clave, como media, mediana.

**Figura 11** *distribución variable de días de cumplimiento*



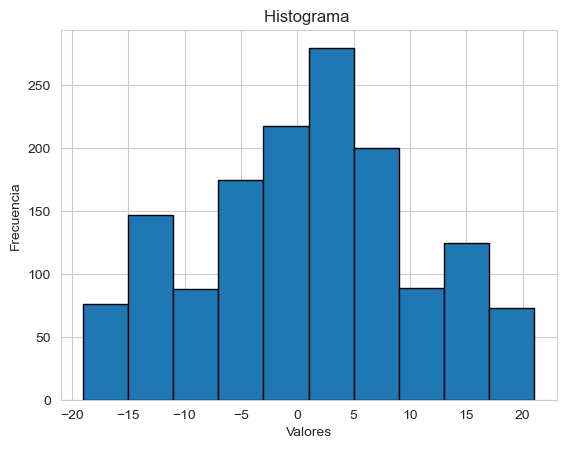
En la figura 11 la mayoría de los valores parecen estar cerca de cero, lo que sugiere que la gran mayoría de los tiempos de entrega están muy próximos a la fecha programada. Sin embargo, existen valores extremos, tanto negativos como positivos. Los valores negativos indican entregas adelantadas, mientras que los positivos representan retrasos. La frecuencia de estos valores extremos es baja, lo cual indica que son excepciones en los datos.

**Figura 12** *Box Plot variable de días de cumplimiento*



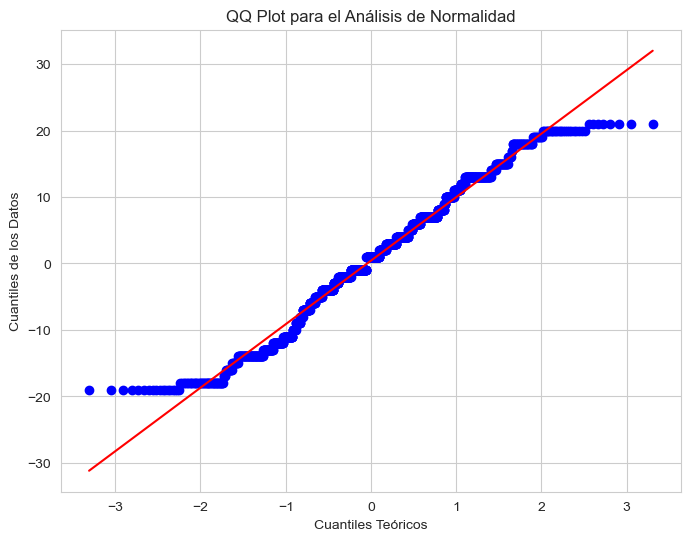
La mayor parte de los datos en la figura 12 se concentran alrededor de la mediana, que está cercana a cero, por lo que en general, las entregas se realizan cerca de la fecha programada. Sin embargo, la presencia de outliers extremos en ambas direcciones podría estar afectando la simetría de la distribución, lo cual podría requerir un tratamiento especial para el manejo de outliers.

**Figura 13** *distribución de días de cumplimiento*



Tras aplicar el manejo de outliers mediante el método de rango intercuartílico y filtrar los datos con valor cero (entregas que llegan a tiempo según la necesidad), los datos en la figura 13 muestran una distribución más centrada y simétrica, con una forma cercana a la distribución normal. Esto resulta en una distribución más controlada y centrada en la media.

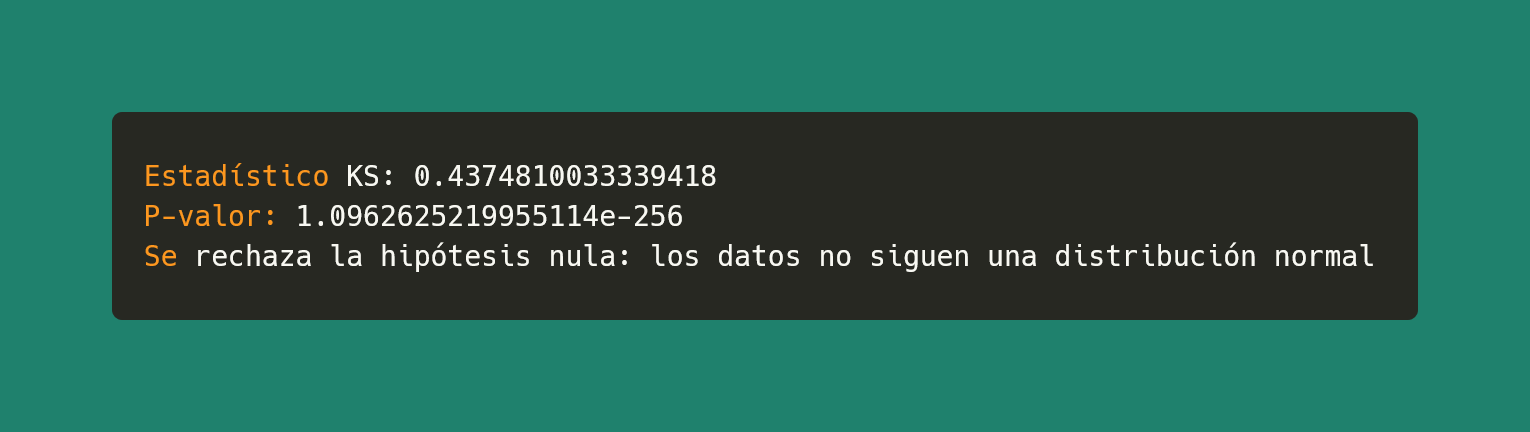
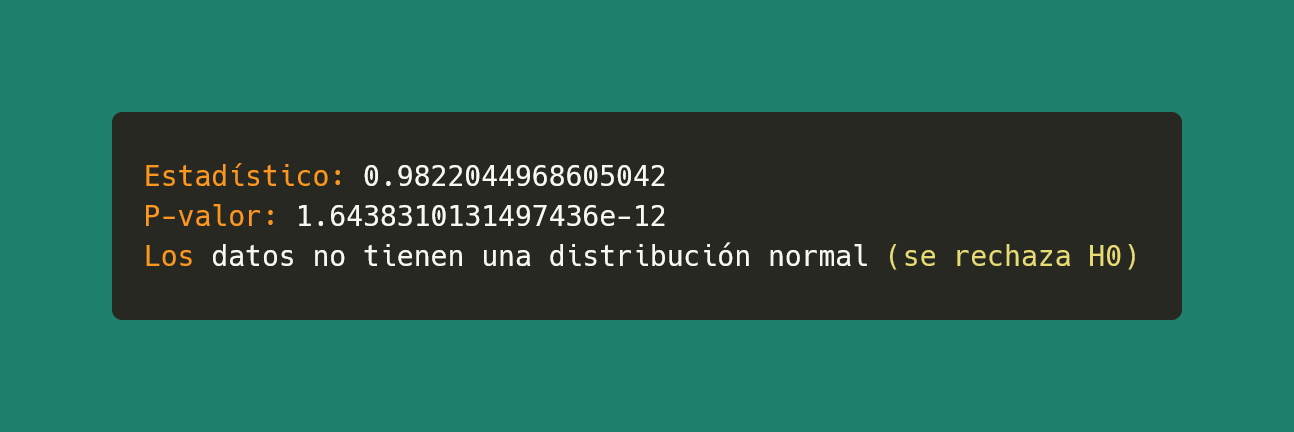
Para comprender mejor la distribución de la variable días de diferencia de entrega, se realizará un gráfico Q-Q (quantile-quantile). Este gráfico permite comparar la distribución de los datos observados con una distribución normal teórica

**Figura 14** *Análisis de normalidad Q-Q de la variable días de cumplimiento*

aplicando los métodos de IQR outliers y el filtro de los valores cero, los días de diferencia tienen una distribución normal en el centro de los datos siendo una distribución más natural en comparación que los datos son filtro de outliers, sin embargo, en las colas los puntos no siguen perfectamente la distribución normal, sugiriendo la posible presencia de sesgo o datos extremos en los datos ya depurados.

Para evaluar la distribución de la variable **"días de diferencia de entrega"**, además del gráfico Q-Q, se aplicarán las pruebas estadísticas de **Shapiro-Wilk** y **Kolmogorov-Smirnov**. Estas pruebas permiten determinar si la variable sigue una distribución normal. Estas pruebas permiten determinar si la variable sigue una distribución normal, lo cual es esencial para decidir los tipos de modelos que pueden ser usados según la distribución.

Test shapiro Test kolmogorov smirnov



Para ambas pruebas estadísticas de normalidad, el valor p es mucho menor al de 0.05 por lo que hay evidencia para rechazar la hipótesis nula de que los datos sigan una distribución normal. Se considera para un análisis posterior usar métodos que no dependen de la normalidad de los datos como Mann-Whitney u Wilcoxon. También otra alternativa es utilizar modelos que no requieran normalidad en los datos, como random forest o los métodos de boosting.

# Referencias

[1] A. G. Banerjee, W. Yund, D. Yang, P. Koudal, J. Carbone, y J. Salvo, “A Hybrid Statistical Method for Accurate Prediction of Supplier Delivery Times of Aircraft Engine Parts”, en *Volume 1B: 35th Computers and Information in Engineering Conference*, Boston, Massachusetts, USA: American Society of Mechanical Engineers, ago. 2015, p. V01BT02A037. doi: 10.1115/DETC2015-47605.

[2] G. A. M. Escobar, “Predicció de cumplimiento de entrega de pedidos farmacéuticos”, Monografia, Universidad de Antioquia, Antioquia Colombia.

[3] F. Steinberg, P. Burggräf, J. Wagner, B. Heinbach, T. Saßmannshausen, y A. Brintrup, “A novel machine learning model for predicting late supplier deliveries of low-volume-high-variety products with application in a German machinery industry”, *Supply Chain Anal.*, vol. 1, p. 100003, mar. 2023, doi: 10.1016/j.sca.2023.100003.

[4] X. Zhou, J. Wang, Y. Liu, X. Wu, Z. Shen, y C. Leung, “Inductive Graph Transformer for Delivery Time Estimation”, en *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, feb. 2023, pp. 679–687. doi: 10.1145/3539597.3570409.

[5] “Df\_proyecto/CSV\_Categorias/data\_all.csv at main · Andres-Fonseca/Df\_proyecto”, GitHub. Consultado: el 21 de noviembre de 2024. [En línea]. Disponible en: https://github.com/Andres-Fonseca/Df\_proyecto/blob/main/CSV\_Categorias/data\_all.csv

[6] “Df\_proyecto/Data\_base/consulta\_seminario\_all.sql at main · Andres-Fonseca/Df\_proyecto”, GitHub. Consultado: el 21 de noviembre de 2024. [En línea]. Disponible en: https://github.com/Andres-Fonseca/Df\_proyecto/blob/main/Data\_base/consulta\_seminario\_all.sql

[7] “SimpleImputer”, scikit-learn. Consultado: el 21 de noviembre de 2024. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.impute.SimpleImputer.html

[8] “Feature importances with a forest of trees”, scikit-learn. Consultado: el 21 de noviembre de 2024. [En línea]. Disponible en: https://scikit-learn/stable/auto\_examples/ensemble/plot\_forest\_importances.html

[1] O. Celsa, “Reporte Bi paros-rendimiento”, O. Celsa, Antioquia Colombia, Power BI 2024.

[2] O. Celsa, “Base de Datos del Maestro de Piezas”. Interno, 2024.

[3] Grupo Nutresa, “Informe\_integrado\_2022-Grupo\_Nutresa”. Grupo Nutresa.

[6] G. A. M. Escobar, “Predicción de cumplimiento de entrega de pedidos farmacéuticos”, Monografía, Universidad de Antioquia, Antioquia Colombia.