Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física
Maestría en Ciencia de Datos

Modelo de Predicción para Empresa de Logística y Paquetería

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestro en Ciencia de Datos

Presenta:
Alejandro Manuel Aguilar Casillas

Director:
Mtro. Byron Michael Motta Bonilla

Tlaquepaque, Jalisco, 21 de mayo de 2025

Modelo de Predicción para Empresa de logística y paquetería

Alejandro Manuel Aguilar Casillas

Resumen

Este trabajo presenta una breve introducción al problema de una predicción de puntualidad en la entrega de envíos en una empresa transnacional de logística. El objetivo principal es desarrollar un modelo de análisis que permita optimizar la asignación de recursos y mejorar la precisión en los tiempos de entrega, aumentando así la satisfacción del cliente y reduciendo costos operativos. En particular, se aborda la relación entre el peso de los envíos, la frecuencia de operaciones y los costos logísticos asociados, utilizando técnicas de ciencia de datos y aprendizaje automático. El desarrollo del trabajo incluye la implementación de cuatro modelos: Máquina de Vectores de Soporte (por sus siglas en inglés SVM), Red Neuronal (Perceptrón Multicapa), Regresión Logística y XGBoost, evaluados según su capacidad para predecir la puntualidad de los envíos. Los resultados obtenidos muestran que la Regresión Logística es el modelo más efectivo, logrando un balance entre precisión, recuperación y exactitud, lo que lo convierte en la mejor opción para optimizar la operación logística en términos de confiabilidad y costos. Finalmente, se presentan las conclusiones, resaltando que este modelo permite a la empresa prever problemas y ajustar su estrategia operativa para cumplir con los tiempos de entrega establecidos en los contratos. La contribución principal de este trabajo radica en proporcionar un modelo analítico que optimiza la gestión de entregas y fortalece la competitividad de la empresa en el mercado global.

Tabla de Contenidos

Índice de figuras

Página

Índice de tablas

Página

Dedicado a mis padres, por su amor incondicional, sacrificios y por haberme enseñado la importancia de la perseverancia y el esfuerzo. Gracias por ser mi ejemplo y mi mayor inspiración, por estar siempre a mi lado en los momentos difíciles y por celebrar conmigo cada logro. Este trabajo es el resultado de todo lo que he aprendido de ustedes y del inmenso apoyo que me han brindado en cada paso de este camino.

A mi hermana, por su apoyo y por ser un pilar fundamental en mi vida, siempre con palabras de aliento en los momentos más difciles.

A mi asesor, Maestro Byron Michael Motta Bonilla, por su invaluable orientación,

paciencia y por siempre brindarme su apoyo en cada etapa de este proyecto, aportando su experiencia y conocimiento.

A mi coordinadora de maestría, Rocío Carrasco Navarro, por su constante motivación y por siempre estar disponible para resolver cualquier duda o desafío que surgiera.

A mis amigos Alejandro Díaz y Rodo Slay, quienes han estado presentes en cada paso de este camino, ofreciéndome su amistad, comprensión y apoyo incondicional, además de ser grandes fuentes de inspiración y motivación.

A mi maestro de apoyo, Gaddiel Desirena Lopez, por su constante disposición para compartir su conocimiento y por su

generoso acompañamiento durante todo el proceso, siempre con una actitud positiva y constructiva.

Al Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) , por brindarme la oportunidad de continuar mi formación académica y por ser un espacio que fomenta el desarrollo profesional y personal, permitiéndome alcanzar este logro.

Finalmente, a todas las personas que, de alguna manera, me acompañaron en este proceso, ya sea con un consejo, una palabra de aliento o su presencia, su apoyo ha sido esencial para poder alcanzar este logro. A todos ustedes, mi más sincero agradecimiento.

1 Introducción

La globalización en aumento y la exigencia de servicios de alta calidad han motivado a la industria logística a explorar soluciones cada vez más eficaces para la distribución de mercancías. La expansión del comercio electrónico y las expectativas de los consumidores han incrementado la presión sobre las empresas para adherirse a plazos de entrega rigurosos, preservando simultáneamente la rentabilidad y competitividad. Sin embargo, factores tales como la saturación de rutas, la disponibilidad de recursos y la incertidumbre de la demanda pueden provocar demoras considerables en los envíos, impactando tanto la percepción del cliente como los resultados financieros de la empresa.

Frente a esta problemática, el objetivo de este estudio es desarrollar un modelo predictivo para prever entregas tardías en una compañía transnacional de logística. Este análisis se basa en la información histórica de envíos, que incluye datos pertinentes como costos, pesos, destinos y plazos de entrega, con el propósito de identificar patrones y factores críticos asociados a los retrasos. Mediante la implementación de técnicas de ciencia de datos y algoritmos de aprendizaje automático, se aspira no solo a disminuir la incidencia de entregas tardías, sino también a optimizar la eficiencia operativa, reducir costos adicionales y robustecer la confianza de los clientes en la organización.

La propuesta incorpora un Análisis Exploratorio de Datos, también denominado (EDA) para entender la naturaleza de los envíos y los indicadores de rendimiento principales. Posteriormente, se llevará a cabo la implementación de diversos modelos de aprendizaje automático (tales como Redes Neuronales, XGBoost, Regresión Logística y otros), los cuales serán comparados y evaluados con métricas de clasificación. Así, se busca proporcionar una solución integral que facilite a la empresa la toma de decisiones fundamentadas, la optimización de sus procesos logísticos y, en última instancia, el incremento de su competitividad en el mercado global.

1.1 Contexto

El ámbito de la logística se halla ante retos continuos para asegurar la puntualidad y eficiencia en la distribución de mercancías, particularmente en mercados globales donde la competitividad está intrínsecamente vinculada con la satisfacción del cliente y la optimización de los costos operacionales. Dentro de este contexto, las organizaciones deben gestionar una intrincada red de distribución que engloba diversas regiones, distintas zonas horarias y una diversidad de servicios. Adicionalmente, elementos como la magnitud de los envíos, la elección de rutas y la distribución de recursos contribuyen a incrementar los niveles de complejidad operativa. En una entidad de logística transnacional, donde estas variables convergen, la habilidad para prever y gestionar problemas se vuelve esencial para satisfacer las expectativas del cliente y asegurar la viabilidad económica.

Este estudio emerge como una solución a una problemática frecuente: las entregas tardías. Estas circunstancias no solo han suscitado reclamaciones recurrentes por parte de los clientes, sino que también han conducido a la emisión de más de 2 millones de dólares en títulos de crédito durante el año 2023, lo que constituye una carga financiera considerable para la organización. Además, la ausencia de puntualidad ha generado una pérdida de confianza en mercados cruciales, como Asia-Pacífico (APAC), donde la corporación únicamente abastece el 10 % del mercado actual, a pesar de las proyecciones de alcanzar el 40 % para el año 2028. Este retraso en la expansión, junto con la presión operativa derivada de la administración de reclamaciones y notas de crédito, pone de manifiesto una necesidad imperante de perfeccionar los procedimientos logísticos y mejorar la puntualidad en las entregas.

La evaluación de los datos históricos de la empresa indicó que, de un total de 5,246 envíos efectuados, únicamente 2,628 llegaron a tiempo, lo que representa el 50.1 %, mientras que 2,618 envíos, que constituyen el 49.9 %, llegaron tarde. Este balance crítico enfatiza la relevancia de detectar patrones y elementos que inciden en la puntualidad de las entregas. La Figura 1.1 ofrece una representación gráfica de estos datos, ilustrando la distribución de los envíos a tiempo y tardíos, representados por los colores azul y morado, respectivamente.

En respuesta a esta problemática, emerge la propuesta de elaborar soluciones analíticas que faciliten la anticipación de posibles demoras y, consecuentemente, optimicen la utilización de recursos en el ámbito financiero. Este estudio plantea el desarrollo e implementación de un modelo predictivo fundamentado en técnicas de aprendizaje automático avanzado para examinar patrones históricos en las entregas, ofrecer instrumentos de apoyo para la formulación de decisiones estratégicas

más acertadas y identificar las causas primordiales de los retrasos.

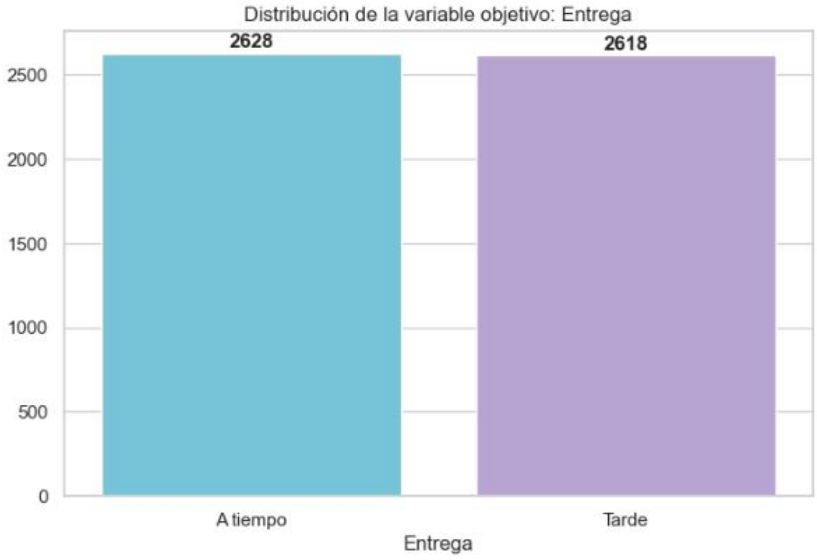


Figura 1.1: Distribución de envíos a tiempo y tardíos

1.1.1 *Estructura del Documento*

Este trabajo se organiza en cuatro capítulos principales que abordan desde el contexto y justificación hasta los resultados y conclusiones obtenidos. A continuación, se presenta una descripción de cada sección del documento:

1. **Introducción:** Este capítulo introduce el marco del problema de puntualidad en las entregas en una entidad logística transnacional y expone los antecedentes, tales como la pérdida de clientes, la presión interna sobre el departamento de ventas y la emisión de notas de crédito como consecuencia de demoras en las entregas. La sección 1.1 ofrece un análisis exhaustivo del trabajo, mientras que las secciones 1.2 a 1.4 abarcan la fundamentación del estudio, la formulación del problema práctico y científico, y la delineación de los objetivos generales y específicos del proyecto.

2. **Metodología:** Este capítulo detalla meticulosamente la metodología implementada para la elaboración del modelo predictivo. La sección 2.1 expone el conjunto de datos suministrado por Panjiva, subrayando su importancia, las técnicas de procesamiento y limpieza implementadas, junto con una descripción detallada de las variables contempladas. La sección 2.2 incorpora un análisis exploratorio inicial de los datos, con el objetivo de identificar patrones pertinentes, tales como la frecuencia de envíos por compañía y región. Las

secciones 2.3 y 2.4 proporcionan una explicación detallada de los modelos predictivos implementados (SVM, Redes Neuronales, Regresión Logística y XGBoost) y las métricas utilizadas para evaluar su desempeño. En la sección 2.5, se especifican los experimentos y simulaciones llevados a cabo, proporcionando una fundamentación robusta para los hallazgos obtenidos.

3. **Resultados y discusión:** En este capítulo se presentan y analizan los resultados obtenidos tras la implementación de los modelos predictivos. Se comparan sus desempeños en términos de métricas clave, como Accuracy, Recall, F1-Score y AUC-ROC. La discusión se centra en el modelo de Regresión Logística, que se identificó como el más adecuado para abordar el problema de las entregas tardías, destacando sus ventajas en el contexto de optimización logística. También se discuten las implicaciones prácticas de estos resultados en las operaciones de la empresa.

4. **Conclusiones y trabajo futuro:** Este último capítulo recoge las conclusiones principales del trabajo, destacando cómo el modelo predictivo contribuye a mejorar las operaciones logísticas, reducir los costos asociados a las entregas tardías y fortalecer la confianza del cliente. En la sección 4.1 se resumen los hallazgos más importantes, mientras que en la sección 4.2 se proponen líneas de trabajo futuro, como la inclusión de nuevas variables, la exploración de datos adicionales y la evaluación del modelo en diferentes contextos operativos para robustecer su aplicabilidad.

1.2 *Justificación*

Este trabajo ofrece una contribución al campo de la ciencia de datos, aplicando modelos modernos de machine learning para abordar un desafío crítico en la industria logística: la mejora en los tiempos de entrega. Al emplear diversas técnicas de machine learning con la información histórica de envíos, no solo ayuda a identificar patrones de comportamiento en las entregas tardías, sino que también permite desarrollar sistemas predictivos capaces de optimizar de manera continua las operaciones logísticas.

Este enfoque innovador tiene el potencial de transformar la toma de decisiones en la logística, reduciendo considerablemente los márgenes de error en las estimaciones de tiempo de entrega y mejorando la planificación de recursos. A nivel económico, busca fortalecer la relación con los clientes y mejorar la reputación corporativa. Al minimizar las entregas tardías, se reducen las reclamaciones y se refuerza la confianza en el servicio.

Además, este trabajo beneficiará a los ejecutivos de cuenta, ya que, al ser un modelo operativo más proactivo, ayudará a liberar recursos, optimizar la gestión del tiempo y priorizar estrategias de crecimiento y expansión del mercado.

Adicionalmente, esta iniciativa no se limita únicamente a optimizar los procesos internos de la empresa, sino que también impulsa la colaboración con proveedores y socios estratégicos. Al identificar las causas fundamentales de los retrasos, surgen oportunidades para renegociar acuerdos de nivel de servicio y alinear las expectativas de los clientes con las capacidades reales de la organización. De esta forma, la gestión basada en datos no solo disminuye la incertidumbre operativa, sino que también fomenta relaciones comerciales más sólidas y transparentes.

A largo plazo, estas mejoras contribuyen a consolidar la reputación de la empresa en el mercado, abriendo paso a nuevas alianzas y a la diversificación de su portafolio de servicios. Asimismo, la eficiencia lograda al reducir tiempos y costos genera un impacto positivo en la sostenibilidad de las operaciones, al disminuir la necesidad de reenvíos o de métodos de transporte urgentes y contaminantes.

1.3 Problema

1.3.1 Problema Práctico

Uno de los constantes desafíos dentro de una compañía transnacional de logística es el poder garantizar la entrega puntual,la eficiencia de los envíos y el cumplimiento de los acuerdos pactados con cada cliente. Al ser este un negocio tan competitivo hace que las compañías optimicen sus operaciones, primordialmente en la gestión de costos, así como la selección de destinos, el control del peso de las cargas, la cual influye de manera directa en el costo del transporte, combustible y el tiempo de entrega.

El peso y la naturaleza del envió suelen afectar de manera significativa en los costos totales y en el tiempo de entrega, así como en la eficiencia de todo lo que engloba el proceso logístico. Los costos que se tienen adicionales, como lo son los cargos por combustible y las dimensiones cubicas, estas pueden aumentar dependiendo de cada uno de los envíos, lo que resulta afectando la rentabilidad de las operaciones. Además de la amplia variedad de servicios y destinos hacen que se introduzca un grado considerable de complejidad en lo que son la toma de decisiones operativas, lo que hace necesario un análisis más profundo de las variables involucradas.

Ante esta situación, nace la necesidad de analizar la relación que se tiene entre el peso total de los envíos, las compañías que usan los

servicios logísticos y las regiones involucradas en todo esto. El objetivo es identificar patrones de comportamiento que permitan optimizar el uso de recursos y reducir costos. De esta manera, se busca garantizar la satisfacción del cliente mediante entregas puntuales y eficientes, mejorando la competitividad y sostenibilidad de la empresa en un entorno de mercado dinámico y exigente.

1.3.2 *Problema Científico*

El problema científico consiste en desarrollar un modelo analítico el cual esta basado en datos de una empresa transnacional de logística lo cual busca comprender como es que diferentes variables, como lo son el peso de los envíos, la compañía que usa el servicio y el destino de los envíos, influyen en los costos y en la eficiencia de los servicios. El enfoque primordial busca responder si es que el envío cumplirá con la fecha de entrega o llegará al destino tarde.

El objetivo es usar la ciencia de datos para lograr obtener un conocimiento amplio a partir del análisis de grandes volúmenes de datos logísticos, permitiendo predecir comportamientos y con esto poder tomar decisiones estratégicas informadas. Lo que se traduce en tener un modelo el cual ayuda a optimizar la asignación de recursos, ayuda a tener una eficiencia en el área operativa y genera una fortalece en la capacidad de la empresa al momento de tener que negociar las tarifas con los clientes, garantiza la puntualidad de los servicios, y genera nuevos clientes en un entorno el cual es sumamente competitivo.

Este trabajo utiliza técnicas de regresión y validación cruzada para medir con precisión la relación que se tiene de cada variable en los resultados operativos, los resultados serán clave para poder diseñar estrategias las cuales están basada en datos que ayudan a generar ahorros en costos, ayuda a tener mejores tiempos de transito y hace que el cliente tenga una experiencia satisfactoria del servicio brindado.

1.4 *Objetivos*

1.4.1 *Objetivo General*

El objetivo general de este trabajo es desarrollar un modelo analítico que permita predecir la puntualidad en la entrega de envíos de una empresa transnacional de logística, con el fin de optimizar el uso de recursos, mejorar la eficiencia operativa y aumentar la satisfacción del cliente mediante la entrega oportuna y confiable.

1.4.2 *Objetivos Específicos*

1. Analizar los datos históricos de envíos para identificar patrones y tendencias que afectan la puntualidad de las entregas.
2. Evaluar el impacto de variables como peso, frecuencia de envíos y destino en los tiempos de entrega y costos logísticos.
3. Implementar y comparar modelos de machine learning que permitan predecir con precisión la puntualidad de los envíos.
4. Determinar el modelo que ofrezca el mejor balance entre precisión y capacidad de generalización para su implementación en la operación logística de la empresa.
5. Proporcionar recomendaciones basadas en los resultados obtenidos para optimizar la gestión de recursos y mejorar la competitividad en el mercado.

2 Metodología

2.1 Descripción de los Datos

2.1.1 Justificación y Relevancia del Conjunto de Datos

En el mundo de la logística y paquetería, la manera que se tiene para predecir con precisión el que un envío llegue a tiempo o tarde es fundamental para tener un servicio de calidad y tener al cliente contento. El conjunto de datos utilizado para este trabajo se seleccionó con base a la relación directa que se tiene con el problema que se busca resolver (la predicción de la puntualidad de las entregas). Este conjunto de datos incluye un extenso rango de características operacionales, financieros y de servicio las cuales permiten modelar la complejidad del proceso de envío, así como las condiciones que provocan que el envío llegue tarde.

Este conjunto de datos es la representación de las operaciones diarias de una empresa de logística y paquetería internacional. Se empleo información clave de las transacciones y operaciones de entrega de paquetes las cuales se llevaron a cabo durante un periodo de tiempo determinado, lo que permite tener una visión más amplia de como se comportaron las entregas y las variables que influyen en el comportamiento.

Las variables seleccionadas para el análisis proporcionan una visión integral de los factores que influyen en los tiempos de entrega de los paquetes. La diversidad de las variables, que abarcan características tanto operativas como financieras, permite capturar la complejidad del proceso de envío. Esta amplitud en los datos facilita la creación de un modelo predictivo robusto y generalizable que puede ayudar a la empresa a identificar paquetes que corren el riesgo de no ser entregados a tiempo.

2.1.2 Fuente de los Datos

Los datos para este trabajo vienen directamente de los registros operacionales de la compañía, los cuales abarcan transacciones realizadas de enero 2023 a diciembre 2023. Esto asegura que los datos

sean verídicos y reflejen las acciones auténticas de la empresa. Entre las cifras se encuentran tanto datos financieros como operativos, tales como el valor declarado de las mercancías, los gastos vinculados al transporte, y variables vinculadas a la ruta, el origen y los tiempos de viaje.

2.1.3 *Procesamiento y Limpieza de los Datos*

Antes de iniciar el análisis, se realizó un proceso exhaustivo de limpieza y preprocesamiento de los datos, todas las observaciones cuentan con información completa, ya que no se detectaron valores faltantes. Sin embargo, fue necesario realizar transformaciones específicas para asegurarse de que los datos estuvieran en el formato adecuado para el análisis, se eliminaron duplicados y se verificó la integridad de cada columna. Las variables de tipo fecha (como la fecha de envío y de entrega) fueron convertidas a un formato adecuado para facilitar su manipulación.

Además, se analizaron los tipos de datos de cada columna para asegurar que estuvieran alineados con los requerimientos de los algoritmos de modelado. Se realizó un análisis sumamente detallado de la distribución de cada una de las variables, buscando posibles valores atípicos los cuales pudieran afectar el rendimiento del modelo. Al tener un conjunto de datos tan amplio y con tanto nivel de detalle, esto nos beneficia ya que embona perfecto para realizar técnicas de machine learning las están orientadas a predecir los retrasos en la entrega de paquetes y el procesamiento cuidadoso asegura la calidad en el análisis realizado.

El conjunto de datos, como se planteó originalmente requería de distintos pasos de preprocesamiento de datos antes de que pudiera ser utilizado en el modelo predictivo. Los principales pasos que se realizaron fueron:

- **Verificación de datos faltantes:** El conjunto de datos ya no contiene datos faltantes ya que se realizó una limpieza de datos que contenían datos faltantes, después de realizar este proceso se hizo una validación para asegurarse de que no existan inconsistencias o missing values en las variables más importantes.
- **Codificación de variables categóricas:** Algunas de las variables del conjunto de datos eran categóricas, como lo son “Compañía", “Moneda", “País Destino", entre otras. Para poder ser utilizadas en modelos de machine learning, estas variables fueron codificadas utilizando técnicas como One-Hot Encoding o Label Encoding, según fuera necesario.
- **Escalado de variables numéricas:** Algunas de las variables del

conjunto de datos eran numéricas, como el valor de la mercancía y el peso total, presentaban rangos muy diferentes, lo que puede impactar en el rendimiento de ciertos modelos de machine learning. Para solventar este problema, se aplicó escalado utilizando técnicas como StandardScaler para normalizar las variables numéricas.

- **Transformación de fechas:** Las variables de fecha de envío y fecha de entrega se transformaron en características adicionales como el número de días de tránsito o la diferencia entre la fecha de envío y entrega, facilitando el análisis y la creación de modelos predictivos.

2.1.4 *Descripción de las Variables*

El conjunto de datos cuenta con 24 columnas, las cuales se segmentan entre lo que son variables categóricas y numéricas. A continuación se describe cada variable en el contexto del problema:

1. **Compañía (Categórica)**: Indica la compañía que realiza el envío. Es una variable que presenta 5 valores únicos (x1, x2, x3, x4 y x5).
2. **Moneda (Categórica)**: Se refiere a la moneda en la que se realiza el pago del envío. Contiene 12 valores únicos, lo que refleja la operación multinacional de la empresa.
3. **País Destino (Categórica)**: Esta variable indica el país destino del envío, tiene 16 valores únicos, es una variable muy importante ya que el destino del envío influye en el tiempo de entrega. Dentro de los países que se encuentran en el conjunto de datos son Argentina, Estados Únicos, Canadá, entre otros.
4. **Estado Destino y Estado Origen (Categórica)**: Estas variables indican el estado de destino y origen del paquete, con 105 y 13 valores únicos respectivamente.
5. **Ciudad Origen (Categórica)**: Esta variable representa la ciudad desde la que se origina el envío, tiene 61 valores únicos como Zapopan y Aguascalientes.
6. **Cargo Vuelo y Cargos Extras (Numérica)**: Representan los costos asociados a los vuelos y otros cargos adicionales (lo que puede incluir costos por servicio express u otros servicios adicionales).
7. **Costo Gas (Numérica)**: Refleja los costos de combustible relacionados a cada envío.
8. **Tipo Peso (Categórica)**: Describe la unidad de medida del peso del envío, ya sea en libras (P) o kilogramos (K).

9. **Peso Total (Numérica)**: Indica el peso total del envío.
10. **Fecha Envío (Categórica)**: Menciona la fecha de envío de los paquetes, se encontraron 185 fechas únicas entre febrero y diciembre del 2023.
11. **Fecha Entrega (Categórica)**: Menciona la fecha de entrega de los paquetes, se encontraron 121 fechas únicas entre enero y diciembre del 2023.
12. **Medidas Cúbicas (Numérica)**: Esta variable representa las dimensiones del paquete en centímetros cúbicos.
13. **Valor Mercancía (Numérica)**: Refleja el valor declarado de la mercancía.
14. **Total (Numérica)**: Monto total pagado por el envío.
15. **Días Tránsito (Numérica)**: Refleja el número de días que tarda el paquete en ser entregado. Es una variable numérica que varía entre 1 y 8 días, lo que refleja las diferencias en los tiempos de entrega dependiendo de las condiciones del envío.
16. **Business Days (Numérica)**: Representa los días hábiles involucrados en el envío.
17. **Servicio (Categórica)**: Describe el tipo de servicio utilizado para el envío, con valores como "Standard Overnightz "1Day Freight", tiene 9 valores únicos.
18. **Peso Final (Numérica)**: Indica el peso final del envío tras posibles ajustes.
19. **Región (Categórica)**: Representa la región de destino del envío, con 5 valores únicos: US, CAN, ASIA, LAC, EU.
20. **Entrega (Numérica)**: Es la variable objetivo que indica si el paquete fue entregado a tiempo o no. Esta variable tiene un valor de 1 para paquetes que llegaron tarde y 0 para los que llegaron a tiempo.

2.2 *Análisis Exploratorio*

2.2.1 *Renombrado y Limpieza de Datos*

El conjunto de datos original contenía múltiples columnas con nombres técnicos y poco descriptivos, lo que dificultaba su interpretación. Por ello, se realizó un proceso de renombrado para garantizar que las variables fueran intuitivas y reflejaran su contenido de manera clara. Por ejemplo:

- La variable *shp_co_nm* se renombró a *Compañía*.
- La variable *shp_rate_wgt* se renombró a *Peso_total*.

- La variable *dim_vol_qty* se renombró a *Medidas_cúbicas*.

Adicionalmente, se eliminaron columnas irrelevantes para el análisis, como *bill_to_tot* y *tot_wgt*, que no aportaban valor al problema planteado. Este proceso redujo el ruido en los datos y permitió centrar el análisis en las variables más relevantes.

Otro paso clave en la limpieza fue la eliminación de espacios en blanco innecesarios en variables categóricas, como *Tipo_peso* y *País_destino*, utilizando una función personalizada para garantizar la consistencia en los valores. Esto evitó problemas posteriores al aplicar transformaciones como codificación categórica.

2.2.2 *Transformación de Variables*

Para enriquecer el análisis y asegurar la coherencia en las unidades, se realizaron varias transformaciones de variables clave:

- **Conversión de peso:** La variable *Peso_total* contenía valores en diferentes unidades (kilogramos y libras). Se implementó una función personalizada para convertir todos los valores a libras, utilizando un factor de conversión de 2.2046 en aquellos casos donde la unidad era kilogramos.
- **Creación de Peso Final:** Tras la conversión, se creó una nueva variable llamada *Peso_Final*, que representa el peso de cada envío en libras. Esta transformación aseguró la uniformidad de los datos, facilitando su análisis y modelado.
- **Agrupación por región:** Se creó la variable *Región* a partir de la variable *País_destino*, asignando cada país a una región principal (*US*, *LAC*, *EU*, *CAN* y *ASIA*) mediante un mapeo predefinido. Esto permitió simplificar el análisis geográfico y detectar patrones regionales.
- **Cálculo de estado de entrega:** Se generó la variable binaria *Entrega* comparando los días reales de tránsito (*B_days*) con los días estimados (*Días_transito*). Los valores se definieron como:
  - *0:* Entrega a tiempo.
  - *1:* Entrega tardía.

Estas transformaciones no solo garantizaron la calidad y consistencia de los datos, sino que también permitieron crear nuevas características que capturan información clave para el análisis y modelado.

2.2.3 *Preparación Final para el Modelado*

La preparación de los datos es una etapa crítica para garantizar que los modelos predictivos puedan operar de manera eficiente y producir resultados confiables. En este proyecto, se aplicaron diversas técnicas

y transformaciones para asegurar que el conjunto de datos estuviera limpio, completo y optimizado para el análisis.

2.2.4 *Manejo de Variables Categóricas*

Las variables categóricas presentes en el conjunto de datos, como *Compañía*, *Región* y *Servicio*, requerían ser transformadas en un formato que los modelos pudieran interpretar. Para ello, se utilizó la técnica de *One-Hot Encoding*, que consiste en convertir cada categoría en una columna binaria que indica la presencia (*1*) o ausencia (*0*) de dicha categoría. Por ejemplo:

- La variable *Compañía*, con categorías *x1*, *x2*, *x3*, *x4* y *x5*, se transformó en cinco columnas independientes: *Compañía_x1*, *Compañía_x2*, etc.
- De manera similar, la variable *Región* se transformó en columnas como *Región_US*, *Región_LAC*, *Región_ASIA*, etc.

Esta transformación permitió que las relaciones entre categorías se procesaran como entradas numéricas en los modelos, evitando problemas relacionados con su interpretación ordinal.

2.2.5 *Normalización de Variables Numéricas*

Las variables numéricas, como *Peso_Final*, *Costo_gas*, *Medidas_cúbicas* y *Días_transito*, presentaban rangos de valores muy distintos. Por ejemplo:

- El *Peso_Final* oscilaba entre valores tan bajos como 2 libras y tan altos como 10,000 libras.
- El *Costo_gas* tenía valores que iban desde 0 hasta un máximo de 43 dólares.

Para evitar que estas diferencias en magnitud influyeran desproporcionadamente en los resultados de los modelos, se aplicó un proceso de normalización utilizando la fórmula de estandarización:

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

donde x es el valor de la variable, μ es la media, y σ es la desviación estándar. Esto permitió que todas las variables numéricas tuvieran una distribución centrada en 0 con una desviación estándar de 1, lo que mejora el rendimiento de modelos sensibles a la escala, como la Regresión Logística y las Redes Neuronales.

2.2.6 *División en Conjuntos de Entrenamiento y Prueba*

El conjunto de datos transformado fue dividido en dos subconjuntos:

- **Conjunto de entrenamiento (70 %):** Utilizado para ajustar los parámetros del modelo y aprender las relaciones entre las variables independientes (X) y la variable objetivo (y).
- **Conjunto de prueba (30 %):** Reservado para evaluar el rendimiento del modelo en datos no vistos, asegurando que la solución generalice correctamente.

Esta división se realizó de manera estratificada para garantizar que la proporción de entregas a tiempo (0) y tardías (1) se mantuviera en ambos subconjuntos.

2.2.7 *Selección de Variables*

Tras la limpieza y transformación de los datos, se seleccionaron las siguientes variables clave para el modelado:

- **Variables numéricas:** *Peso_Final*, *Costo_gas*, *Medidas_cúbicas*, *Días_transito*, *B_days*.
- **Variables categóricas transformadas:** *Compañía*, *Región*, *Servicio*.

Esta selección se realizó con base en su relevancia para predecir el estado de entrega, asegurando que los modelos dispusieran de información suficiente para identificar patrones y tendencias en los datos.

2.2.8 *Conclusión de la Preparación para el Modelado*

El proceso de preparación de los datos incluyó técnicas avanzadas de limpieza, transformación y selección de variables, lo que resultó en un conjunto de datos optimizado y listo para el desarrollo de modelos predictivos. Estas acciones garantizaron la calidad y consistencia de los datos, permitiendo construir modelos más robustos y confiables para abordar el problema de las entregas tardías en la empresa logística.

2.2.9 *Análisis de Entregas*

En el análisis inicial, se encontró que de los 5,246 envíos realizados:

- 2,628 envíos (50.1 %) llegaron a tiempo.
- 2,618 envíos (49.9 %) llegaron tarde.

La Figura 2.1 , presenta un desglose visual de los envíos a tiempo y tardíos, destacando la importancia de abordar los retrasos mediante modelos predictivos.

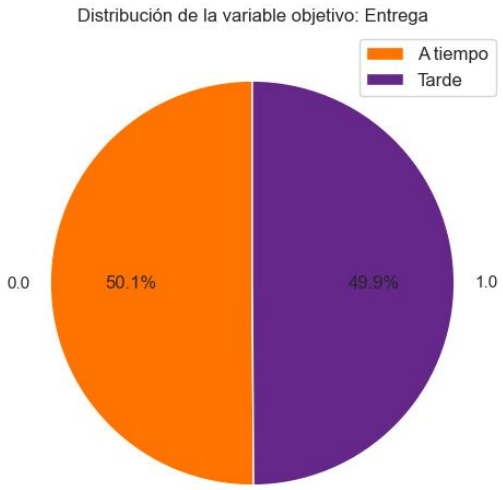


Figura 2.1: Distribucion de la variable objetivo: Entrega

2.2.10 *Frecuencia de Envíos por Compañía y Región*

En la Tabla 2.1 se muestra que la empresa x1 presenta una perspectiva variada con predominio en Estados Unidos y Latinoamérica (LAC). Respecto al peso total, EEUU tiene el mayor volumen de envíos con 474,991 libras, siendo LAC el siguiente con 347,849 libras. En comparación, tanto Canadá (CAN) como Europa (EU) reciben cantidades marginales, con 15 y 6 envíos respectivamente. La frecuencia muestra un total de 1,389 envíos a EEUU, corroborando su importancia considerable en este mercado, y 358 envíos a LAC, donde, a pesar de que se llevan a cabo menos operaciones, se gestionan cantidades considerables, lo que indica una tendencia a administrar envíos pesados. Por lo tanto, la empresa x1 sostiene una sólida estrategia en EEUU, mientras que en LAC constituye un importante mercado secundario con cargas considerables. Su actividad en CAN y EU podría crecer para capitalizar nuevas posibilidades de mercado.

Compañía	ASIA	CAN	EU	LAC	US
x1	0	2,078	307	347,850	474,991
x2	66	24,476	5,630	1,479	350,876
x3	0	0	0	10,371	264,424
x4	418	731	518	18,001	52,360
x5	261	383	5,902	14,268	268,250

Tabla 2.1: Peso total final (sin decimales) por Compañía y Región

La compañía x2 sostiene una sólida presencia en Estados Unidos y Canadá, enviando un total de 350,876 libras a Estados Unidos y 24,476 libras a Canadá, demostrando de esta manera su concentración en estos dos mercados de Norteamérica. Los envíos en Europa y Asia son

más livianos, con 5,630 y 66 libras, respectivamente. Su frecuencia de envíos llega a 1,006 a US y a 129 a CAN, lo que reafirma su orientación hacia Norteamérica, mientras que su escasa frecuencia en LAC y Asia podría señalar operaciones esporádicas o poco estructuradas. Gracias a su infraestructura consolidada en América, la compañía x2 posee la capacidad de explorar nuevas posibilidades en Europa y Asia.

A pesar de sus restricciones en sus operaciones, la compañía x3 pone un gran énfasis en US, realizando envíos pesados. Su operación más destacada, con 264,424 libras en US, es LAC con 10,371 libras, aunque en un volumen inferior. En cuanto a la frecuencia, x3 documenta 638 envíos a US y 87 a LAC, lo que indica que, a pesar de que su frecuencia no es tan elevada como en otras compañías, sus envíos son considerables. Esta experticia en envíos pesados sugiere un enfoque estratégico en Estados Unidos y probablemente hacia cargas especializadas o menos apremiantes.

La compañía x4 muestra una diversificación en expansión, con posible desarrollo en Asia. En peso total, US encabeza con 52,360 libras, LAC le sigue con 18,001 libras, mientras que Asia obtiene 418 libras, lo que indica que la compañía está indagando en este mercado. La frecuencia de envíos indica 518 a US y 211 a LAC, lo que evidencia una distribución considerable entre estas dos áreas fundamentales. Asia, que solo ha realizado 3 envíos, evidencia un potencial de expansión y una posibilidad para futuras ampliaciones. Por lo tanto, x4 se distingue por su estrategia variada y su concentración en diversas regiones, considerando a Asia como una región en auge que podría potenciar su crecimiento a nivel mundial.

La compañía x5 tiene una gran dependencia del mercado de Estados Unidos, destinando la mayor cantidad de su peso a US (268,250 libras), seguida por LAC (14,268 libras). Europa y Asia reciben las cantidades más bajas, lo que podría indicar oportunidades mal utilizadas. Con 716 envíos a Estados Unidos, su elevada frecuencia evidencia la dependencia de este mercado, mientras que las operaciones restringidas en Europa, Asia y CAN indican una ausencia de diversificación. En este escenario, x5 se encuentra con la oportunidad de diversificarse y expandirse en otras áreas.

El análisis comparativo revela que todas las empresas poseen un énfasis considerable en el mercado estadounidense, lo que evidencia la relevancia estratégica de dicha región. Las tácticas difieren en función del peso: x1 y x3 se distinguen por gestionar envíos de gran volumen en los Estados Unidos y las Américas Centrales y Orientales, mientras que x2 y x4 se desempeñan con envíos más ligeros pero diversificados en diversas regiones. Con respecto al potencial en Asia, a pesar de que el volumen actual es reducido, empresas como x2 y x4 ya han iniciado sus operaciones, lo que sugiere potenciales oportunidades futuras en dicha

región. Adicionalmente, x5 ejecuta una elevada frecuencia de envíos, a pesar de su peso relativamente reducido, mientras que x3 se concentra en una menor frecuencia de envíos, pero de volumen considerable. Lo mencionado anteriormente se puede observar en la Figura 2.2.

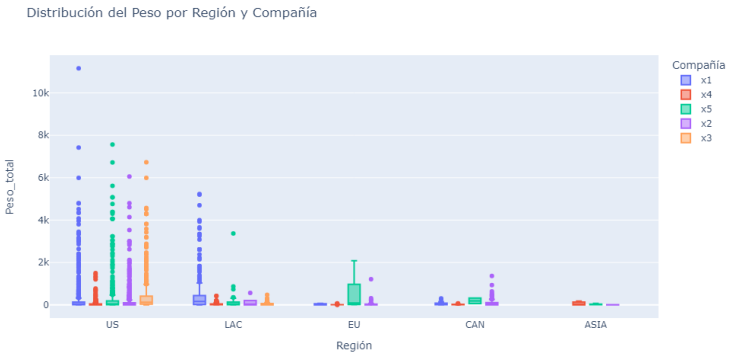


Figura 2.2: Distribución del Peso por Región y Compañía

2.2.11 *Entregas Tardías por Mes*

Los incrementos en las entregas tardías observados durante los meses de junio y julio pueden atribuirse principalmente al cierre del segundo trimestre en determinadas empresas de la data correspondiente. A lo largo de esta etapa, numerosas entidades experimentan un incremento considerable en sus operaciones comerciales, lo cual intensifica la cantidad de envíos y sobrecarga la red logística. Específicamente, el mes de julio, se distingue por la mayor cantidad de entregas tardías, registrando un total de **1,337 entregas tardías**, lo que representa el **51.07 %** de todas las entregas tardías del año. Esto es considerablemente superior en comparación con el mes siguiente con más demoras, junio (Mes 6), el cual tuvo **873 entregas tardías**, lo cual corresponde al **33.35 %**. Para una mejor comprensión, los resultados se pueden visualizar de manera más clara en la Tabla 2.2 y en la Figura 2.3.

Mes	Entregas Tardías	% del Total
Febrero	1	0.04 %
Abril	2	0.08 %
Mayo	12	0.46 %
Junio	873	33.35 %
Julio	1,337	51.07 %
Agosto	94	3.59 %
Septiembre	129	4.93 %
Octubre	170	6.49 %

Tabla 2.2: Entregas tardías por mes y su proporción respecto al total del periodo analizado.

En cambio, los meses con niveles reducidos de retrasos, tales como **febrero**, **abril** y **mayo**, exhiben cifras notablemente inferiores de entregas tardías. En el mes de febrero, se registraron únicamente **1 entrega tardía**, lo que representa apenas el **0.04 %** del total de entregas tardías. **Abril** registró **2 entregas tardías**, representando el **0.08 %**. En mayo, se presentaron **12 entregas tardías**, lo que representa el **0.46 %**. Estos meses suelen caracterizarse por una actividad comercial reducida, lo cual facilita un funcionamiento más eficiente de las operaciones logísticas. Además, durante estas fechas, numerosas compañías contratan personal adicional para gestionar el incremento de la carga laboral debido al volumen de envíos iniciales, lo cual contribuye a la disminución de los retrasos.

En términos generales, tras los picos de junio y julio, la cantidad de entregas tardías se reduce de manera gradual. Los meses de **agosto**, **septiembre** y **octubre** presentan cifras mucho más bajas de entregas tardías. En el mes de agosto, se registraron **94 entregas tardías**, lo cual representa el **3.59 %** del total de entregas tardías del año. En el mes de **septiembre**, se registraron **129 entregas tardías**, lo que representa el **4.93 %**. En el mes de **octubre**, se registraron **170 entregas tardías**, lo que representa el **6.49 %**. Esta conducta sugiere una reducción en la demanda de envíos o una mejora en la capacidad operativa y logística tras el pico estacional.

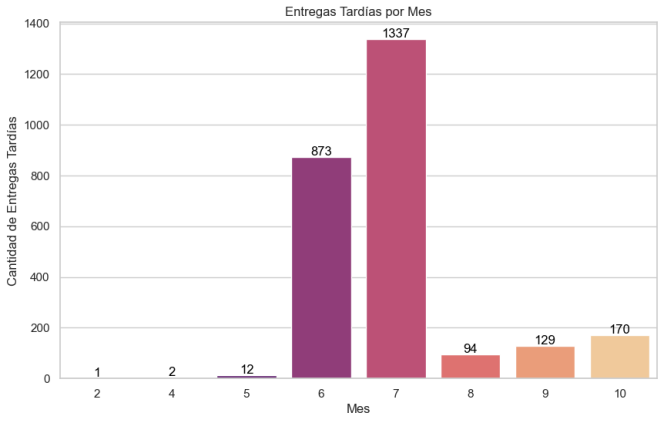


Figura 2.3: Entregas Tardías por Mes

2.2.12 *Entregas a Tiempo por Mes*

La evaluación del desempeño mensual de las entregas puntuales evidencia una progresión positiva en el segundo semestre del año, con un incremento gradual en los niveles de puntualidad y un pico operativo en el mes de **octubre**. En el presente mes, se documentaron 541 entregas puntuales, lo cual representa el **20.59 %** del total de entregas puntuales observadas entre enero y diciembre. Este

rendimiento en octubre es reconocido como el más destacado del periodo y podría atribuirse a una combinación de estabilidad operativa, optimización de recursos y una adecuada planificación logística. Además, este mes tiende a estar exento de fenómenos estacionales extremos, lo cual podría propiciar un ambiente más propicio para las operaciones. Para una mejor comprensión, los resultados se pueden visualizar de manera más clara en la Tabla 2.3 y en la Figura 2.4.

Mes	Entregas a Tiempo	% del Total
Junio	72	2.74 %
Julio	423	16.10 %
Agosto	384	14.61 %
Septiembre	458	17.43 %
Octubre	541	20.59 %
Noviembre	418	15.91 %
Diciembre	332	12.63 %

Tabla 2.3: Entregas a tiempo por mes y su proporción respecto al total semestral.

A lo largo de los meses de **julio**, **agosto** y **septiembre** se observaron elevados niveles de puntualidad. Específicamente, **julio** registró **423 entregas a tiempo** (16.10 %), lo que indica una recuperación considerable en comparación con junio. **Agosto** experimentó una ligera disminución con **384 entregas a tiempo** (14.61 %), mientras que **septiembre** experimentó una nueva crecimiento con **458 entregas a tiempo** (17.43 %). Estos hallazgos evidencian un período de mejora continua en el rendimiento operativo, lo cual indica una potencial estabilización de procesos o la ejecución exitosa de medidas correctivas tras los retos afrontados en el inicio del año.

Por otro lado, **junio** se destacó como el mes con la menor cantidad de entregas a tiempo, documentando únicamente **72 entregas** (2.74 %). Este valor atípicamente bajo podría atribuirse al inicio del periodo evaluado, modificaciones en los sistemas de gestión, modificaciones estratégicas o una disminución en la actividad comercial vinculada al cierre del trimestre fiscal. Esta cifra simboliza una oportunidad evidente de optimización para la organización, al indicar un periodo de vulnerabilidad en su cadena de suministro.

Además, **diciembre** evidenció una reducción en la puntualidad, con **332 entregas a tiempo** (12.63 %). Este descenso podría atribuirse a la elevada carga operativa característica de la temporada alta de fin de año, durante la cual el volumen de envíos se incrementa significativamente, lo que resulta en cuellos de botella y saturación de los recursos logísticos. En numerosas instancias, el aumento en la demanda sobrepasa la capacidad instalada, lo cual incide directamente en la puntualidad.

El mes de noviembre se mantuvo con un volumen sólido de

418 entregas realizadas a tiempo (15.91 %), aunque reflejó una ligera disminución en comparación con el pico alcanzado en el mes de octubre. No obstante, si se compara con el rendimiento de otros meses, sigue siendo un claro ejemplo de un excelente desempeño operativo en la empresa.

De manera generalizada, se registró una mejora notable desde junio hasta octubre, seguida por una reducción marginal hacia diciembre. Este comportamiento indica que, a pesar de que las capacidades logísticas experimentaron una mejora progresiva, también fueron desafiadas por los incrementos estacionales en la demanda de envíos hacia el final del año. La Figura 2.4 facilita la identificación de los meses que presentan prácticas operativas superiores, así como los que requieren la implementación de estrategias de mitigación para asegurar niveles constantes de rendimiento. Se podría consolidar los progresos alcanzados durante los meses de mayor demanda y mantener controles de calidad en los procesos durante los meses de mayor eficiencia.

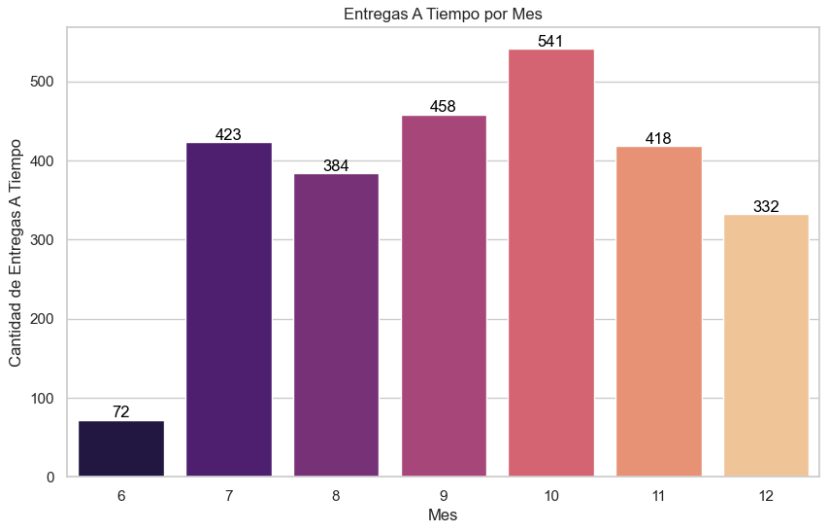


Figura 2.4: Entregas A Tiempo por Mes

2.2.13 *Análisis de Costos por Entrega a Tiempo vs. Tarde*

El análisis de los costos medios por envío de acuerdo a su puntualidad muestra una variación considerable en los gastos operativos entre los envíos que se entregaron en el plazo previsto y los que sufrieron demoras. Los envíos que se presentaron oportunamente registraron un costo promedio de **$838 USD**, mientras que aquellos que sufrieron demoras alcanzaron un costo promedio de **$961 USD**. Esto sugiere que los envíos postergados implican un incremento superior al **15 %** en el costo medio por paquete, lo que evidencia un impacto

financiero considerable derivado de las demoras logísticas.

Los costos adicionales pueden estar vinculados a una variedad de factores que inciden en la eficiencia operativa. Inicialmente, los retrasos pueden originar **sobrecostos por reprogramaciones o urgencias** debido a la necesidad de modificar las rutas de entrega, reasignar recursos o emplear métodos de transporte más costosos para reducir la repercusión del incumplimiento en los clientes. Adicionalmente, se acostumbra proporcionar **notas de crédito** a los clientes afectados, lo que incrementa el gasto total vinculado a las entregas tardías.

Un factor de importancia es el **incremento en los costos operativos** debido a la necesidad de gestionar reclamaciones, coordinar nuevas entregas y afrontar las incidencias ocasionadas por los retrasos. Estos procedimientos implican la utilización de personal adicional y recursos que podrían ser destinados a actividades de mayor estratégicidad. En última instancia, el efecto financiero de los retrasos se manifiesta en una disminución de la rentabilidad, dado que un incremento en el costo por envío sin un incremento proporcional en los ingresos afecta directamente el margen de ganancia de la empresa.

Para optimizar la eficiencia operativa y reducir los gastos logísticos, es crucial implementar herramientas analíticas que permitan anticipar y evitar retrasos. Optimizar la experiencia del cliente no solo eleva la calidad del servicio, sino que también reduce los gastos extras vinculados a la gestión de incidentes, fomentando una operación más eficiente y duradera. Para una mejor comprensión, los resultados se pueden visualizar de manera más clara en la Figura 2.5.

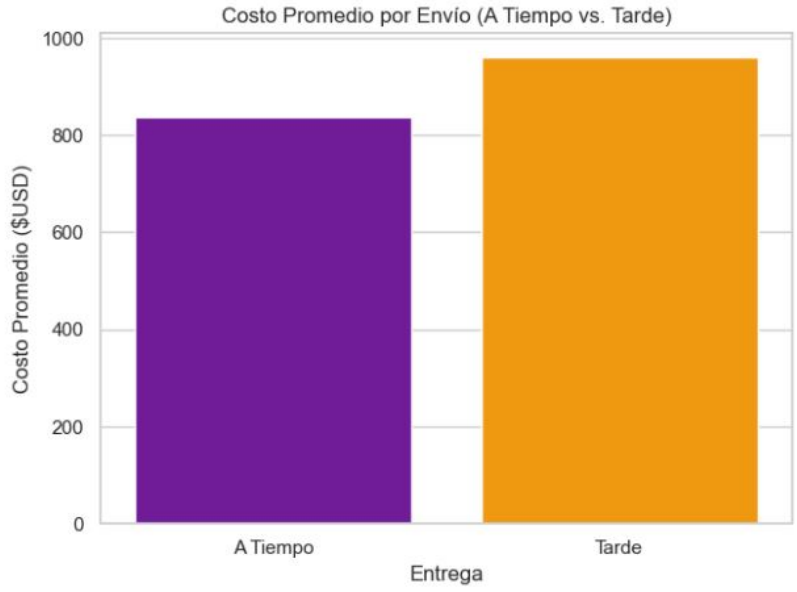


Figura 2.5: Análisis de Costos por Entrega a Tiempo vs. Tarde

2.2.14 *Mapa de Calor de Correlaciones*

La correlación extremadamente alta de 0.92 entre las variables Peso Total y Peso Final indica de manera contundente que ambos factores están estrechamente relacionados entre sí. Esto nos lleva a concluir que el Peso Total de los envíos ejerce una influencia directa significativa en el Peso Final, lo cual podría resultar de gran utilidad para mejorar la eficiencia de los procesos logísticos al anticipar el Peso Final a partir del Peso Total registrado inicialmente. La variable de Medidas Cúbicas muestra una correlación positiva significativa con la variable de Peso Total (0.63), lo cual sugiere que existe una relación moderadamente fuerte entre el Volumen y el peso de los envíos. Esto podría implicar que los costos asociados al transporte o almacenamiento de los productos dependen tanto del Volumen como del peso de los mismos, por lo que es fundamental e imprescindible considerar ambas dimensiones en el análisis logístico para tomar decisiones estratégicas informadas y eficaces. La variable Total presenta correlaciones moderadas con Cargo Vuelo (0.31) y Cargos Extras (0.29), lo cual indica que el Costo Total de un envío podría estar relacionado tanto con los cargos de vuelo como con los Cargos Extras. Este descubrimiento es sumamente relevante y significativo para el análisis detallado y exhaustivo de los costos, así como para la implementación efectiva de estrategias de optimización en los complejos y variados procesos de fijación de precios y tarifas de envío. La correlación negativa de -0.64 entre la variable Días Transito y la variable Entregas indica que a medida que transcurren más días de tránsito, la cantidad de entregas realizadas dentro del plazo esperado tiende a reducirse. Esto sugiere que los tiempos de tránsito más prolongados pueden tener un impacto negativo en la eficiencia de las entregas, lo cual podría implicar la imperiosa necesidad de mejorar y optimizar las rutas o los procesos logísticos existentes. El Mes del año muestra correlaciones generalmente bajas o incluso nulas con la gran mayoría de las variables analizadas, a excepción de Valor Mercancía, donde se observa una correlación significativa de -0.35. Esto nos lleva a considerar la posibilidad de que el valor de la mercancía pueda experimentar fluctuaciones dependiendo de la temporada, aunque estas variaciones no serían excesivamente significativas. Este descubrimiento podría tener importantes implicaciones para el desarrollo de estrategias de pricing y promoción de productos durante determinadas épocas del año. Para una mejor comprensión, los resultados se pueden visualizar de manera más clara en la Figura 2.6.

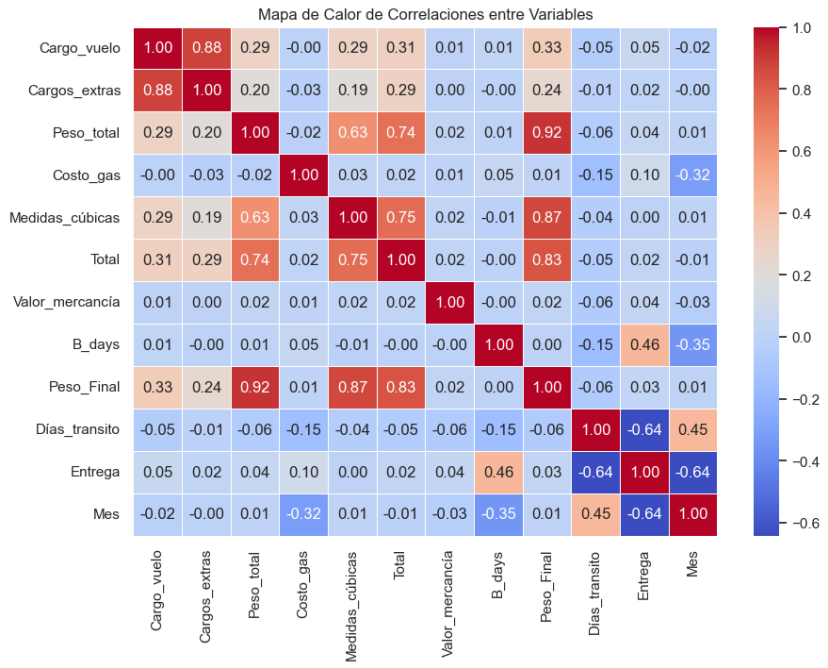


Figura 2.6: Mapa de Calor de Correlaciones entre Variables

2.2.15 *Análisis de la Distribución del Tiempo Real de Entrega por Compañía*

La Figura 2.7 ilustra la distribución de los días efectivos de entrega para diversas compañías, permitiendo así identificar las variaciones en la rápidez y consistencia de sus plazos de entrega. La mayoría de las empresas exhiben un elevado volumen de entregas en los primeros días, lo cual indica que la mayoría de las entregas se llevan a cabo en un plazo breve. Sin embargo, se puede apreciar una dispersión en la parte derecha del gráfico, lo cual indica que, aunque menos frecuentes, existen entregas que demandan varios días para su finalización.

Específicamente, la empresa x1 (representada en rojo) cuenta con una elevada frecuencia de entregas realizadas en días 0 y 1 con una disminución abrupta en la frecuencia conforme los días se incrementan. Esto indica que la empresa es altamente eficaz en sus entregas rápidas; sin embargo, también presenta algunos casos en los que los plazos de entrega exceden, aunque en menor medida. La prolongada cola en dirección derecha, que puede llegar hasta los 30 días, podría señalar ciertos retrasos logísticos o excepciones en el camino de entrega. No obstante, la mayoría de las entregas se finalizan en un plazo inferior a tres días.

En contraste, la compañía x4 (de color verde) exhibe una distribución más heterogénea, con una mayor variabilidad en los plazos de entrega. A pesar de que también presenta una cantidad considerable de entregas

en los primeros días, la frecuencia disminuye de manera gradual en comparación con x1, y la cola de la distribución se prolonga más allá de los 50 días, lo que indica una mayor cantidad de entregas en etapas tardías. Esto podría indicar que x4 experimenta una mayor variabilidad en sus procedimientos logísticos, lo que conduce a plazos de entrega más extensos y a una mayor inconsistencia.

La empresa x5 (color verde claro) presenta una distribución que se asemeja a la de x1, con una frecuencia notablemente elevada de entregas durante los primeros días (especialmente en 1 y 2 días). No obstante, su longitud de cola es considerablemente menor en contraste con la de x1, y no experimenta demoras en las entregas tan significativas. La gran mayoría de las entregas suelen completarse en un plazo de aproximadamente tres días hábiles, observándose una disminución significativa en la frecuencia a partir de ese momento. Esto indica que la empresa opera con una eficiencia notable en los plazos de entrega, aunque presenta menos situaciones excepcionales que x1.

Las compañías x2 (de color azul) y x3 (de color morado) exhiben conductas intermedias. x2 exhibe una concentración superior de entregas en los primeros cinco días, aunque la frecuencia disminuye de manera más gradual en comparación con x5 y x1, lo que indica una variabilidad moderada en los tiempos de entrega. En el caso de x3, la cola hacia la derecha no es tan extensa, lo cual indica una menor cantidad de entregas tardías. Sin embargo, aún exhibe una dispersión significativa que podría sugerir que existen factores impredecibles que influyen en sus plazos de entrega.

En términos generales, el análisis indica que las empresas x1 y x5 exhiben un rendimiento relativamente rápido y consistente, con escasas excepciones en los tiempos de entrega. En comparación, x4 parece enfrentar desafíos más significativos en términos de consistencia de sus plazos, con una mayor variación en los tiempos de entrega. Esto podría sugerir la necesidad de mejorar la optimización de sus rutas o la administración de los tiempos de tránsito. Respecto a las compañías x2 y x3, ambas exhiben comportamientos variados, con plazos de entrega más impredecibles, aunque no tan extensos como los de x4.

Es importante destacar que los datos demuestran una tendencia generalizada hacia la rápidez en las entregas, con un porcentaje elevado de las entregas efectuadas dentro de los primeros tres días, lo que evidencia un empeño por parte de la empresa en el cumplimiento de los plazos establecidos. No obstante, es evidente que se confrontan retos logísticos que resultan en un prolongado tiempo de tránsito, lo cual podría ser optimizado mediante la implementación de mejoras en los procedimientos operativos, tecnologías de seguimiento más sofisticadas o una gestión más eficiente de la cadena de suministro.

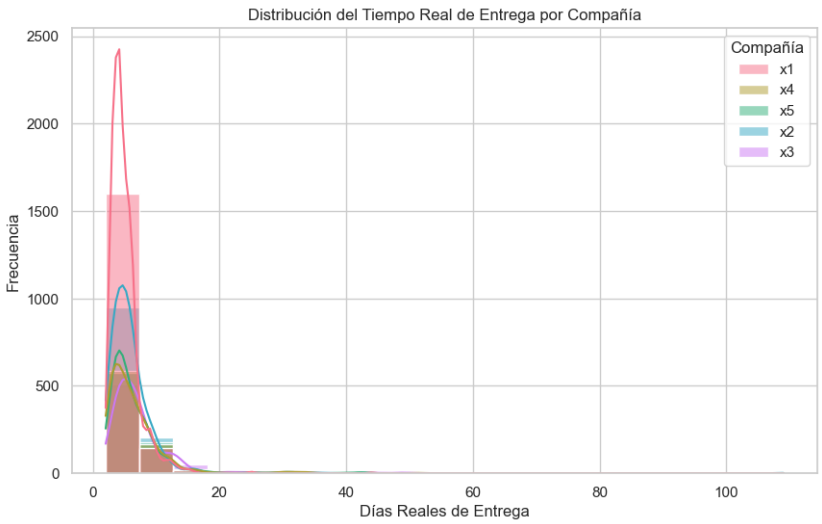


Figura 2.7: Distribución del Tiempo Real de Entrega por Compañía

2.2.16 *Análisis del Desempeño Logístico*

En la Figura 2.8 ofrece una comparativa exhaustiva del rendimiento logístico de cinco compañías distintas, considerando tanto el número total de envíos entregados a tiempo como aquellos considerados tardíos. La evaluación se llevó a cabo sobre un conjunto considerable de entregas, lo que posibilitó una apreciación precisa del rendimiento logístico individual.

Es evidente que la Compañía 1 exhibe el rendimiento más destacado, con un total de 1,224 entregas efectuadas a tiempo, lo que representa la cifra más elevada entre todas las compañías examinadas. A pesar de que también registró 544 envíos entregados de manera tardía, este número constituye únicamente el 30.8 % del total de envíos efectuados por la mencionada empresa (1,768 envíos totales).

La Compañía 2 presenta un rendimiento balanceado, aunque con una considerable capacidad de mejora, efectuando un total de 676 entregas puntuales en comparación con las 510 entregas tardías. La proporción de entregas tardías es notable, constituyendo aproximadamente el 43 % del total, lo que indica que, a pesar de la eficiencia operativa, existe un espacio considerable para la optimización de procesos.

La Compañía 3 muestra resultados sumamente alarmantes, registrando únicamente 202 entregas puntuales en comparación con 523 envíos atrasados. Esto sugiere que alrededor del 72 % de sus entregas se llevaron a cabo fuera del plazo previsto, lo que sugiere graves dificultades operativas o logísticas que pueden impactar severamente en la confianza y satisfacción del cliente, así como en la rentabilidad de

la compañía.

En relación con la Compañía 4, se registran 276 entregas puntuales, mientras que las entregas tardías se elevan a 510, lo que también evidencia una situación alarmante. En este contexto, la proporción de entregas puntuales es inferior al 40 %, lo que evidencia graves deficiencias operativas que requieren una intervención prioritaria.

En última instancia, la Compañía 5 exhibe un rendimiento igualmente negativo, registrando únicamente 250 entregas a tiempo en comparación con 531 entregas a tiempo, lo que representa un porcentaje de entregas a tiempo que supera el 68 %. Esta es una indicación clara de que se enfrenta a retos operativos cruciales y se requiere la implementación inmediata de estrategias correctivas.

Para concluir, estos hallazgos cuantitativos ponen de manifiesto una superioridad evidente de la Compañía 1 en lo que respecta a la eficiencia operativa y puntualidad. Sin embargo, en las demás empresas, es indispensable enfocarse en la revisión y optimización de los procesos logísticos y operativos para reducir de manera significativa el porcentaje de entregas tardías, garantizando de esta manera una mayor satisfacción del cliente y un mejor posicionamiento competitivo en el mercado.

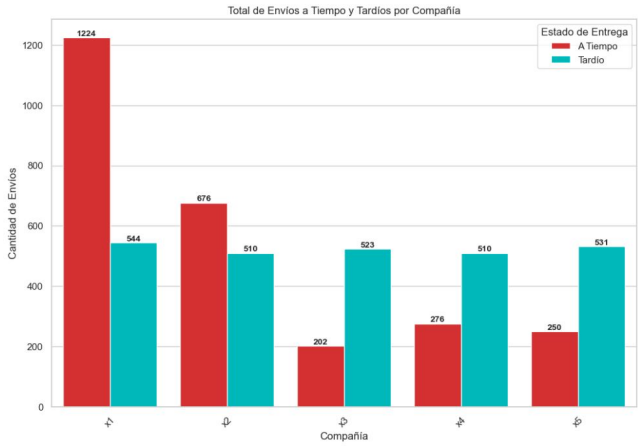


Figura 2.8: Análisis del desempeño logístico

2.2.17 *Análisis Comparativo: Entregas a Tiempo vs Entregas Tardías por Mes*

El nivel mensual de entregas muestra una variación notable entre los meses de junio a diciembre, particularmente al contrastar las entregas efectuadas en tiempo con aquellas que se llevaron a cabo con demora. Este contraste es particularmente evidente durante los meses de junio y julio, cuando se manifiestan los desafíos logísticos más significativos, mientras que los niveles de cumplimiento más elevados se logran entre septiembre y noviembre, lo que pone de manifiesto una evolución

positiva en la administración operativa con el transcurso del año.

El mes con mayor número de entregas atrasadas fue julio, registrando un total de 1,337 entregas, lo que constituye el 51.07 % del total anual de demoras. Este volumen significativo indica una potencial saturación operacional, posiblemente causada por un aumento estacional en la demanda, una insuficiencia en la capacidad instalada o modificaciones en los procedimientos de distribución. Sigue junio, el cual registró 873 entregas tardías, lo que representa un 33.35 %. En suma, estos dos meses registran un total de 2,210 entregas tardías, lo que equivale al 84 % de los retrasos documentados a lo largo del año. Este indicador pone de evidencia una concentración evidente de ineficiencia logística al concluir el segundo trimestre y al comienzo del tercer.

A partir del mes de agosto, se registra una disminución significativa en los retrasos, registrándose 94 entregas tardías (3.59 %), una reducción que podría estar asociada con las modificaciones operativas implementadas tras el desplome presentado en los meses previos. En septiembre se registraron 129 entregas tardías (4.93 %), mientras que en octubre se observó un incremento menor a 170 (6.49 %), lo que indica una recuperación sostenida, aunque no exenta de variabilidad. En última instancia, en los meses de noviembre y diciembre no se registraron entregas tardías, lo que sugiere que la operación alcanzó un nivel de estabilidad y eficiencia al concluir el año, posiblemente respaldado por una planificación de recursos más eficaz o una presión estacional reducida. Para facilitar el análisis, los resultados se ilustran en la Tabla 2.4 y en la Figura 2.9.

Mes	**Entregas a Tiempo**	**Entregas Tardías**
Enero	0	0
Febrero	0	1
Marzo	0	0
Abril	0	2
Mayo	0	12
Junio	72	873
Julio	423	1,337
Agosto	384	94
Septiembre	458	129
Octubre	541	170
Noviembre	418	0
Diciembre	332	0

Tabla 2.4: Comparación mensual de entregas a tiempo y entregas tardías

A diferencia de los datos anteriores, la ejecución puntual de las entregas demostró una evolución positiva a lo largo del semestre. El mes más destacado en términos de puntualidad fue octubre, registrando

541 entregas a tiempo, las cuales constituyen el 20.59 % del total del semestre. Este pico de cumplimiento indica que las acciones correctivas implementadas durante el tercer trimestre tuvieron un impacto significativo, consolidando un nivel operativo más sólido. Septiembre se encuentra a la vanguardia, con 458 entregas (17.43 %), julio con 423 entregas (16.10 %) y noviembre, con 418 entregas (15.91 %). Es importante mencionar que julio, a pesar de contar con un elevado número de entregas puntuales, también fue el mes con mayor número de demoras, lo que sugiere un volumen de operación elevado y potencialmente no debidamente dimensionado.

El rendimiento más eficaz en términos de puntualidad se registró en junio, con únicamente 72 entregas a tiempo (2.74 %). Este desempeño insatisfactorio, junto con el elevado volumen de entregas tardías del mismo mes, señala un punto de inflexión en la operación logística, probablemente asociado a deficiencias estructurales, insuficiente preparación ante la demanda o dificultades en la coordinación interna.

El contraste más notable se observa en los meses de junio y julio, periodo en el que los elevados índices de entregas tardías coinciden con un rendimiento insuficiente en puntualidad. En julio, por ejemplo, aunque se registran 423 entregas puntuales, también se presentan 1,337 entregas no cumplidas, lo que evidencia una saturación significativa en la infraestructura logística. Esta dualidad podría atribuirse a una distribución inequilibrada de recursos, procesos internos insuficientemente optimizados o fluctuaciones imprevistas en el volumen de envíos. Además, se observa una mejora sostenida tanto en la disminución de las entregas tardías como en el aumento de las entregas puntuales. Esta mejora puede atribuirse a la instauración de estrategias correctivas, un incremento en la anticipación en la distribución de recursos y un aprendizaje organizacional fundamentado en los errores cometidos durante los primeros meses del año.

En última instancia, los meses de noviembre y diciembre se distinguen por su estabilidad y su rendimiento excepcional. En noviembre, los niveles de cumplimiento se mantienen elevados con 418 entregas puntuales y cero demoras, mientras que en diciembre, a pesar de que el volumen total disminuye ligeramente a 332 entregas puntuales, no se registran entregas fuera de tiempo. Esta conducta indica una consolidación en los procesos logísticos, potencialmente respaldada por una disminución de la presión estacional o una coordinación más efectiva de la cadena de suministro.

Para concluir, se observa un efecto estacional evidente en el comportamiento logístico. Los picos de actividad en los meses de junio y julio indican una carga operativa incrementada y evidencian debilidades que requieren atención mediante una planificación y refuerzo de capacidad. En comparación, la segunda mitad del semestre

exhibe una progresión positiva, logrando niveles de eficiencia y cumplimiento logístico que se mantienen hasta la conclusión del año. Estos descubrimientos son cruciales para la anticipación de necesidades operativas futuras, la asignación estratégica de recursos y el fortalecimiento de la capacidad de respuesta durante los periodos de alta demanda. Además, facilitan la identificación de oportunidades para el mejoramiento constante en la administración logística fundamentada en evidencia empírica tangible.

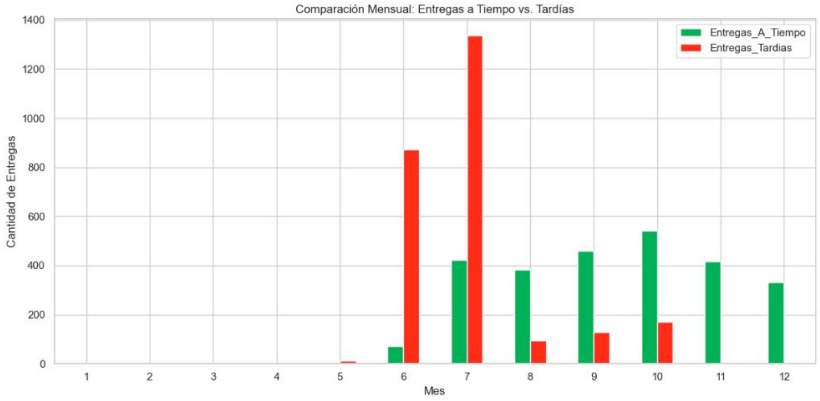


Figura 2.9: Análisis Comparativo Entregas a Tiempo vs Entregas Tardías por Mes

2.2.18 *Análisis Comparativo de Participación Regional en Exportaciones*

La comparación entre las exportaciones realizadas por la empresa en cuestión y el total de las exportaciones efectuadas por todas las compañías en diversas regiones facilita la identificación de áreas de liderazgo, posibilidades de crecimiento y posibles desequilibrios operativos. Al analizar los porcentajes de participación por región y por empresa, se logran diversos descubrimientos significativos que facilitan la evaluación del posicionamiento estratégico de la empresa en comparación con su competencia en el mercado logístico global.

Inicialmente, se resalta la región de **Latinoamérica (LAC)**, en la que la compañía muestra una participación particularmente significativa. Para la compañía x1, se constata que el 98.01 % de las exportaciones documentadas en esta región se atribuyen a sus operaciones, es decir, 347,850 libras, en comparación con un total de 354,904 libras. Este dominio es igualmente relevante para las compañías x4 y x5, las cuales aportan con el 92.02 % (18,001 de 19,561 libras) y el 84.01 % (14,268 de 16,983 libras) respectivamente. Este comportamiento puede ser atribuido a la solidez de las relaciones comerciales en América Latina, la minimización de obstáculos logísticos y una potencial ventaja competitiva derivada de tiempos de tránsito o tarifas. Una

representación más clara de estos resultados se encuentran en las Tablas 2.5.

Compañía	ASIA	CAN	EU	LAC	US
x1	0	2,078	307	347,850	474,991
x2	66	24,476	5,630	1,479	350,876
x3	0	0	0	10,371	264,424
x4	418	731	518	18,001	52,360
x5	261	383	5,902	14,268	268,250

Tabla 2.5: Exportaciones por región realizadas por la empresa

Compañía	ASIA	CAN	EU	LAC	US
x1	25,284	7,000	58,742	354,904	555,474
x2	18,576	30,960	89,564	28,469	381,034
x3	11,352	7,740	49,074	14,986	328,434
x4	1,032	1,548	3,637	19,561	56,145
x5	774	489	33,282	16,983	297,370

Tabla 2.6: Exportaciones totales por región considerando todas las paqueterías

En la región de **Estados Unidos (US)**, la compañía conserva una posición dominante. La compañía x1 constituye el 85.51 % del total de exportaciones hacia dicha lugar, enviando 474,991 libras en comparación con las 555,474 libras del total. Además, x2 alcanza un significativo 92.09 % (equivalente a 350,876 libras), mientras que x5 contribuye con el 90.21 % (equivalente a 268,250 libras). Estos niveles evidencian una sólida consolidación operativa en un mercado de alta demanda logística y de gran volumen. La consistencia en volúmenes elevados establece a estas compañías como líderes confiables para la ruta México-Estados Unidos.

En contraste, en las regiones de **Asia** y **Canadá**, se observan patrones de participación mucho más variables y fragmentados. Por ejemplo, la empresa x1 no presenta exportaciones a Asia, lo cual representa una **oportunidad clara de expansión**. En esta región, x4 y x5 tienen una participación del 40.50 % (418 de 1,032 libras) y 33.72 % (261 de 774 libras), respectivamente, lo que indica que han logrado establecer rutas de exportación más estables hacia dicha región. En el caso de Canadá, x2 domina con el 79.06 % del total (24,476 de 30,960 libras), seguido por x5 con el 78.32 % (383 de 489 libras). En contraste, x1 sólo exportó 2,078 libras frente a 7,000 en total, representando apenas el 29.69 %.

En lo que respecta a la **Unión Europea (EU)**, los niveles de participación son considerablemente bajos para la mayoría de las compañías. La empresa x1 apenas representa el 0.52 % del total exportado, con solo 307 libras de un total de 58,742. Otras compañías como x4 y x3 también mantienen participaciones mínimas, con 518 libras (14.24 %) y 0 libras (0.00 %) respectivamente. Sin embargo, x5

presenta una participación más significativa, con el 17.73 % (5,902 de 33,282 libras), lo cual puede indicar una estrategia más agresiva de inserción en el mercado europeo.

Desde una perspectiva general, los datos evidencian que la compañía en estudio ha conseguido un posicionamiento robusto en mercados clave como Estados Unidos y América Latina, demostrando fortalezas en términos de volumen y cobertura regional. Por ejemplo, de las entidades examinadas, x1 exportó en total 825,226 libras a LAC y US, lo cual representa el 91.7 % de su volumen total (900,304 libras). No obstante, también revelan áreas de mejora estratégica, especialmente en Asia (0 de 25,284 libras) y Europa (307 de 58,742 libras), donde la presencia es limitada o inexistente. Estos descubrimientos posibilitan la formulación de recomendaciones concretas dirigidas al crecimiento global: fortalecer las capacidades logísticas en regiones de escasa penetración, diversificar las rutas comerciales y explorar alianzas estratégicas con operadores locales para optimizar el acceso a mercados aún inexplorados.

En conclusión, el análisis comparativo no solo pone de evidencia la repercusión de las operaciones de la empresa en el sector logístico regional, sino que también ofrece una perspectiva conjunta del potencial de expansión en mercados emergentes. La representación de estos porcentajes facilita la priorización de decisiones estratégicas fundamentadas en pruebas cuantitativas, dirigidas a la optimización del alcance geográfico y la eficiencia en las exportaciones.

2.3 *Descripción de los Modelos*

El objetivo principal de esta sección es describir los modelos de machine learning utilizados para abordar el problema de predecir la puntualidad en las entregas. Se utilizaron cuatro modelos ampliamente reconocidos por su desempeño en problemas de clasificación: *Support Vector Machines (SVM)*, *Red Neuronal (Perceptrón)*, *Regresión Logística* y *XGBoost*. Cada modelo fue seleccionado por su capacidad para manejar conjuntos de datos con características categóricas y numéricas, además de su idoneidad para abordar problemas con clases balanceadas, como el caso de estudio.

2.3.1 *Red Neuronal (Perceptrón)*

"Una red neuronal es un método de la inteligencia artificial (IA) que enseña a las computadoras a procesar datos de una manera similar a como lo hace el cerebro humano. Se trata de un tipo de proceso de machine learning (ML) llamado aprendizaje profundo, el cual utiliza los nodos o las neuronas interconectados en una estructura de capas

que se parece al cerebro humano. Crea un sistema adaptable que las computadoras utilizan para aprender de sus errores y mejorar continuamente. De esta forma, las redes neuronales artificiales intentan resolver problemas complicados, como la realización de resúmenes de documentos o el reconocimiento de rostros, con mayor precisión".[1]

"Las redes neuronales pueden ayudar a las computadoras a tomar decisiones inteligentes con asistencia humana limitada. Esto se debe a que pueden aprender y modelar las relaciones entre los datos de entrada y salida que no son lineales y que son complejos [1]."

Importancia de las Redes Neuronales

El valor principal de las redes neuronales radica en su capacidad para adaptarse, aprender de los datos históricos y mejorar el rendimiento del sistema mediante la retropropagación del error. A medida que reciben más información, las redes ajustan internamente sus pesos sin intervención humana directa, optimizando su capacidad de predicción.

Gracias a su flexibilidad y escalabilidad, las redes neuronales han sido adoptadas en diversos sectores, como la logística, las finanzas, la salud y el comercio electrónico, donde permiten automatizar procesos complejos con alta precisión y eficiencia.

La red neuronal utilizada en este proyecto es un Perceptrón Multicapa (*MLP, Multilayer Perceptron*), un modelo basado en capas de neuronas conectadas diseñado para abordar problemas de clasificación binaria, como predecir si un envío llegará a tiempo. Su capacidad para aprender patrones complejos y no lineales lo hace ideal para datos con múltiples interacciones entre variables. Configurado con una capa oculta, una función de activación *ReLU* [2] y el optimizador *Adam* [1], este modelo fue seleccionado por su flexibilidad y capacidad para adaptarse a datos complejos.

Estructura del Modelo

La arquitectura de la red neuronal consta de las siguientes capas:

- **Capa de entrada (*Input Layer*):** Esta capa recibe un vector de entrada de 129 características ($x \in \mathbb{R}^{129}$), que representan las variables seleccionadas tras el proceso de ingeniería de características. Su función principal es distribuir las entradas hacia las siguientes capas de la red.
- **Primera capa oculta (*Hidden Layer*):** Esta capa contiene 2 neuronas y utiliza la función de activación **Unidad Lineal Rectificada (ReLU)** una de las funciones de activación más utilizadas en redes neuronales, particularmente en modelos de aprendizaje profundo. Se ha convertido en una elección predeterminada dentro de muchas arquitecturas modernas debido a su simplicidad y eficiencia computacional. La función

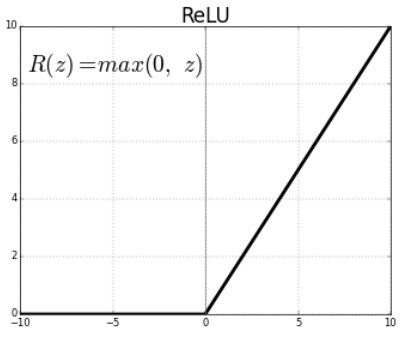
ReLU es una función lineal a tramos que devuelve el valor de entrada si éste es positivo; de lo contrario, devuelve cero. En otras palabras, permite que los valores positivos pasen sin alteración y anula todos los negativos. Esta propiedad ayuda a mantener la complejidad necesaria para que las redes neuronales puedan aprender patrones sin caer en problemas asociados con otras funciones de activación, como el *vanishing gradient*. Matemáticamente, ReLU se expresa como:

$$f(x) = \text{máx}(0, x) \tag{2.2}$$

o bien, de forma equivalente:

$$f(x) = \begin{cases} x, & \text{si } x > 0 \\ 0, & \text{si } x \leq 0 \end{cases} \tag{2.3}$$

Esta simplicidad es justamente lo que hace a ReLU tan eficaz durante el entrenamiento de redes neuronales profundas, ya que permite mantener la no linealidad necesaria sin requerir transformaciones complejas, facilitando así el aprendizaje eficiente del modelo [2]. Dicha representación se muestra en la Figura 2.10.



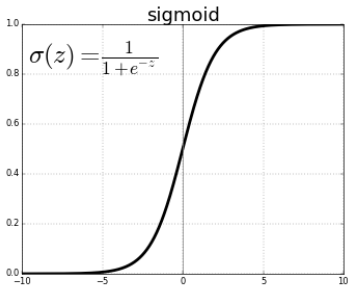
[2]

Figura 2.10: ReLU Activation Function

- **Capa de salida (*Output Layer*):** La salida de la red es una única neurona, que emplea una función de activación **sigmoide**, apropiada para problemas de clasificación binaria. La función sigmoide se define como:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.4}$$

donde x es el resultado de la combinación lineal de las entradas. Esta función transforma la salida en un valor entre 0 y 1, el cual se interpreta como la probabilidad de que el envío sea tardío (es decir, cuando $y = 1$) [3]. Lo anterior se ilustra de forma visual en la Figura 2.11.



[3]

Figura 2.11: Sigmoide

Función de Pérdida

Para optimizar los pesos de la red, se utilizó la función de pérdida de *error cuadrático medio* (**MSE, Mean Squared Error**). El Error Cuadrático Medio (MSE, por sus siglas en inglés) es una métrica ampliamente utilizada tanto en estadística como en aprendizaje automático, ya que permite evaluar el nivel de precisión de los modelos predictivos. Esta medida cuantifica el promedio de las diferencias al cuadrado entre los valores reales y los valores estimados por el modelo [4, 5].

El MSE representa una herramienta para determinar qué tan cercanas están las predicciones realizadas por un modelo respecto a los valores observados. Se calcula promediando los cuadrados de los errores individuales, entendiendo por error la diferencia entre el valor real y el valor pronosticado para cada observación del conjunto de datos. Cuanto menor sea el valor del MSE, mayor será la precisión del modelo evaluado[4, 5].

Se encuentra definida como:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 \tag{2.5}$$

donde:

- Y_i es el valor real de la variable objetivo para el ejemplo i,
- $\hat{Y}_i$ es la predicción de la red para el ejemplo i,
- N es el número total de ejemplos.

La elección de esta función de pérdida se debe a su simplicidad y su capacidad para medir el error promedio entre las predicciones y los valores reales[4][5].

Optimizador

El modelo usa el optimizador **Adam** (acrónimo de *Adaptive Moment Estimation*) es un método que combina las ventajas de dos técnicas ampliamente utilizadas en el entrenamiento de redes neuronales: ***Momentum*** y ***RMSprop***. Esta combinación permite ajustar de manera eficiente y adaptativa las tasas de aprendizaje de cada uno de los parámetros del modelo durante el proceso de entrenamiento.

Una de las principales fortalezas de Adam radica en su capacidad para adaptarse automáticamente a las características de cada parámetro, lo cual lo convierte en una opción altamente eficaz para modelos complejos y conjuntos de datos de gran tamaño. Además, su bajo consumo de memoria y su comportamiento robusto ante gradientes ruidosos lo hacen especialmente útil en contextos prácticos donde la estabilidad y la eficiencia son esenciales.

El método ***Momentum*** se utiliza para acelerar el proceso de descenso por gradiente al incorporar un promedio móvil ponderado exponencialmente de los gradientes pasados. Esto permite suavizar la trayectoria del proceso de optimización, ayudando al algoritmo a converger más rápido al reducir las oscilaciones.

La regla de actualización de los pesos con *momentum* se expresa como:

$$w_{t+1} = w_t - \alpha m_t \tag{2.6}$$

donde:

- m_t es el promedio móvil de los gradientes en el instante t,
- α es la tasa de aprendizaje,
- w_t y w_{t+1} son los pesos en los tiempos t y $t+1$, respectivamente.

El término de ***momentum*** se actualiza de forma recursiva como:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{\partial L}{\partial w_t} \tag{2.7}$$

donde:

- β_1 es el parámetro de momentum, comúnmente establecido en 0.9,
- $\frac{\partial L}{\partial w_t}$ es el gradiente de la función de pérdida respecto a los pesos en el instante t.

RMSprop es un método adaptativo para ajustar la tasa de aprendizaje, diseñado como una mejora de ***AdaGrad***. Mientras que AdaGrad acumula gradientes al cuadrado, RMSprop emplea un promedio móvil ponderado exponencialmente de los gradientes al cuadrado, lo cual ayuda a superar el problema de la disminución excesiva de la tasa de aprendizaje.

La fórmula de actualización de los pesos con RMSprop es:

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + \epsilon}} \frac{\partial L}{\partial w_t} \tag{2.8}$$

donde:

- v_t es el promedio ponderado exponencial de los gradientes al cuadrado:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \left(\frac{\partial L}{\partial w_t} \right)^2 \tag{2.9}$$

- ϵ es una constante pequeña (por ejemplo, 10^{-8}) utilizada para evitar divisiones por cero.

El optimizador Adam (*Adaptive Moment Estimation*) combina los enfoques de Momentum y RMSprop para ofrecer un proceso de optimización más equilibrado y eficaz. Los componentes principales de Adam son:

- **Estimación del primer momento (media):**

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{\partial L}{\partial w_t} \tag{2.10}$$

- **Estimación del segundo momento (varianza):**

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \left(\frac{\partial L}{\partial w_t} \right)^2 \tag{2.11}$$

- **Corrección de sesgo:** Como m_t y v_t se inicializan en cero, es necesario aplicar una corrección de sesgo para evitar que estén sesgados hacia valores pequeños en las primeras iteraciones:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{2.12}$$

- **Actualización final de los pesos:**

$$w_{t+1} = w_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{2.13}$$

Parámetros comunes en Adam:

- α: tasa de aprendizaje o tamaño del paso (por defecto: 0.001),
- β_1: factor de decaimiento del promedio de gradientes (por defecto: 0.9),
- β_2: factor de decaimiento del promedio de gradientes al cuadrado (por defecto: 0.999),
- ϵ: constante pequeña para evitar división por cero (por defecto: 10^{-8}).

La fuente usada para toda la sección fue: [6].

Motivación de la Elección del Modelo

La red neuronal fue seleccionada debido a su capacidad para capturar relaciones complejas entre las variables predictoras y la variable objetivo, algo que no siempre es posible con modelos lineales como la regresión logística. Además, la configuración del modelo, con una capa oculta y una función de activación no lineal, permite modelar interacciones no triviales en los datos. El uso de Adam como optimizador asegura una convergencia eficiente y estable durante el entrenamiento, incluso en presencia de datos heterogéneos.

2.3.2 *Gradient Boosting XGBoost*

El modelo **XGBoost** (*Extreme Gradient Boosting*) utilizado en este trabajo se diseñó para abordar un problema de clasificación binaria: predecir si un envío llegará a tiempo o no. Este modelo, basado en el método de *Gradient Boosting*, combina múltiples árboles de decisión en un ensamblado secuencial para optimizar la predicción final. A través de la reducción iterativa del error de predicción, XGBoost demuestra ser altamente eficiente y preciso en problemas complejos de clasificación.

XGBoost es un algoritmo de aprendizaje automático que forma parte del enfoque de aprendizaje por conjuntos, dentro del marco de *gradient boosting*. Este modelo utiliza árboles de decisión como estimadores base y aplica técnicas de regularización que permiten mejorar su capacidad de generalización frente a nuevos datos [7].

Una de sus principales ventajas es su alta eficiencia computacional, lo que lo convierte en una herramienta ideal para manejar grandes volúmenes de información. Además, ofrece mecanismos integrados para tratar valores faltantes, así como funcionalidades para evaluar la importancia de las variables predictoras. Gracias a estas características, XGBoost es ampliamente utilizado en tareas de regresión, clasificación y ordenamiento [7].

Configuración del Modelo

El modelo se configuró con los siguientes hiperparámetros, ajustados para garantizar un equilibrio entre la complejidad del modelo y su capacidad de generalización:

- **learning_rate:** Configurado inicialmente en 0.2 y posteriormente ajustado a 0.15 para reducir el tamaño de los pasos en la minimización de la función de pérdida. Una tasa de aprendizaje más baja permite que el modelo construya árboles más precisos, aunque a costa de un mayor tiempo de entrenamiento.“La tasa de aprendizaje (también conocida como “tamaño de paso” o “contracción”), es el hiperparámetro de impulso de gradiente más importante. En la biblioteca XGBoost, se conoce como “eta”, debe ser un número entre 0 y 1 y el predeterminado es 0.3. La tasa de aprendizaje determina la velocidad a la que el algoritmo de impulso aprende de cada iteración. Un menor valor de eta significa un aprendizaje más lento, ya que reduce la contribución de cada árbol en el conjunto, ayudando así a evitar el sobreajuste. Por el contrario, un valor más alto de eta acelera el aprendizaje, pero puede conducir a un sobreajuste si no se ajusta cuidadosamente"[8].
- **max_depth:** Se estableció en 3 y 4 para limitar la profundidad máxima de los árboles. Este valor controla

la complejidad del modelo, evitando el sobreajuste a los datos de entrenamiento."“Profundidad” o número de nodos de bifurcación de los árboles de decisión usados en el entrenamiento. Aunque una mayor profundidad puede devolver mejores resultados, también puede resultar en overfitting (sobre ajuste)"[9].

- **n_estimators:** Ajustado inicialmente a 11 y luego reducido a 5, lo que define el número total de árboles en el ensamblado. Un menor número de árboles favorece la simplicidad del modelo y reduce el riesgo de sobreajuste.“Especifica el número de árboles que se construirán en el conjunto. Cada ronda de impulso agrega un nuevo árbol al conjunto y el modelo aprende lentamente a corregir los errores cometidos por los árboles anteriores. n_estimators dirige la complejidad del modelo e influye tanto en el tiempo de entrenamiento como en la capacidad del modelo para generalizar a datos no vistos. Aumentar el valor de n_estimators suele aumentar la complejidad del modelo, ya que permite que el modelo capture patrones más complejos en los datos. Sin embargo, agregar demasiados árboles puede provocar un sobreajuste. En términos generales, a medida que aumenta n_estimators, la tasa de aprendizaje debería disminuir"[8].
- **subsample:** Configurado en 0.8, lo que indica que cada árbol utiliza el 80 % de las observaciones disponibles. Esto introduce diversidad en el ensamblado y mejora la capacidad del modelo para generalizar.
- **colsample_bytree:** Establecido en 0.7, indicando que cada árbol usa el 70 % de las características disponibles. Este ajuste favorece la diversificación de los árboles individuales dentro del ensamblado.
- **random_state:** Se fijó en 42 para garantizar la reproducibilidad de los resultados.

Ventajas de XGBoost

- **Alta Precisión:** El enfoque de conjunto de XGBoost, que combina múltiples modelos, ofrece una precisión superior en comparación con modelos individuales como los árboles de decisión[10].
- **Escalabilidad:** Está optimizado para manejar grandes volúmenes de datos y puede ejecutarse eficientemente en sistemas con capacidades de procesamiento en paralelo[10].
- **Flexibilidad:** XGBoost es una herramienta versátil que puede utilizarse en tareas de regresión, clasificación y ranking[10].

- **Interpretabilidad:** Aunque no es tan fácilmente interpretable como los modelos más simples, XGBoost proporciona información sobre la importancia de las variables, lo cual ayuda a entender qué factores influyen significativamente en las predicciones[10].

Fundamentos Matemáticos de XGBoost

Función Objetivo

El entrenamiento de XGBoost implica minimizar una función objetivo que combina una función de pérdida con un término de regularización. Para problemas de regresión, esta se define como:

$$\text{Objective} = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \right] \tag{2.14}$$

donde:

- n: número de observaciones.
- y_i: etiqueta real.
- $\hat{y}_i$: predicción.
- $l(y_i, \hat{y}_i)$: función de pérdida.
- K: número de árboles.
- f_k: árbol k-ésimo.
- $\Omega(f_k)$: término de regularización que penaliza la complejidad.

Función de Pérdida

La función de pérdida $l(y_i, \hat{y}_i)$ es específica del problema. Para regresión, puede ser la pérdida cuadrática:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \tag{2.15}$$

Regularización

El término de regularización se expresa como:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|_2^2 \tag{2.16}$$

donde:

- γ: controla la complejidad del árbol.
- T: número de hojas.
- λ: parámetro de regularización.
- w: pesos de las hojas.

Adición de Árboles

El entrenamiento se realiza de forma iterativa. En cada iteración t, se añade un nuevo árbol para corregir los errores de los árboles anteriores:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + \eta f_t(x_i) \tag{2.17}$$

donde:

- $f_t(x_i)$: predicción del árbol en la iteración t.
- η: tasa de aprendizaje (learning rate).

XGBoost utiliza un algoritmo de descenso de gradiente para minimizar la función objetivo. Cada nuevo árbol minimiza el gradiente de la función de pérdida:

$$\text{Residual} = -\frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} \tag{2.18}$$

Estos residuos se utilizan como nuevas etiquetas para ajustar el siguiente árbol, mejorando progresivamente el modelo. Para ilustrar el funcionamiento del algoritmo de *gradient tree boosting*, considere la siguiente formulación concisa[11].

2.3.3 *Regresión Logística*

La regresión logística es un modelo lineal ampliamente utilizado en problemas de clasificación binaria. Este modelo estima la probabilidad de pertenencia a una clase específica basándose en una combinación lineal de las variables predictoras. Su simplicidad y capacidad interpretativa lo convierten en una excelente opción para este proyecto. Además, los resultados obtenidos con este modelo mostraron una alta precisión y generalización.

La regresión logística fue evaluada con diversas configuraciones de hiperparámetros, variando el tipo de *solver* y el valor del parámetro de regularización C. Se fijó el número máximo de iteraciones en **200** para asegurar la convergencia del modelo en todos los casos. A continuación, se presentan las combinaciones más representativas:

- **Solver:** Se probaron los métodos *newton-cg*, *lbfgs* y *liblinear*.
- **C (Regularización):** Se evaluaron tres valores distintos: 1.0, 0.5 y 0.1.
- **max_iter:** Se estableció en 200.

Fundamentos Matemáticos de la Regresión Logística

La regresión logística es un modelo estadístico comúnmente utilizado para modelar una variable dependiente binaria mediante

el uso de la función logística. Esta función también es conocida como la *función sigmoide*, y se define como:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \tag{2.19}$$

Esta función permite a la regresión logística transformar los valores de entrada en un rango comprendido entre 0 y 1, lo que facilita la interpretación probabilística de las predicciones. Aunque se utiliza principalmente para tareas de clasificación binaria, también puede aplicarse en problemas de clasificación multiclase.

Supongamos inicialmente que $p(x)$ es una función lineal. Sin embargo, este enfoque presenta un inconveniente: mientras que p representa una probabilidad que debe estar acotada entre 0 y 1, una función lineal como $p(x)$ no posee esta restricción, ya que puede tomar valores no acotados.

Para resolver este problema, se propone aplicar la transformación *logit*, la cual consiste en expresar la razón de probabilidades $\log\left(\frac{p(x)}{1-p(x)}\right)$ como una función lineal de las variables independientes:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \alpha_0 + \alpha \cdot x \tag{2.20}$$

Resolviendo para $p(x)$, se obtiene:

$$p(x) = \frac{e^{\alpha_0 + \alpha}}{e^{\alpha_0 + \alpha} + 1} \tag{2.21}$$

Para que la regresión logística actúe como un clasificador lineal, es posible definir un umbral específico, por ejemplo, 0.5. De este modo, se puede minimizar la tasa de error de clasificación prediciendo $y = 1$ cuando $p \geq 0.5$ y $y = 0$ cuando $p < 0.5$. En este caso, las clases posibles son 1 y 0.

Dado que la regresión logística predice probabilidades, es posible ajustar el modelo utilizando la función de verosimilitud. Por lo tanto, para cada observación de entrenamiento x, la clase predicha es y. La probabilidad de que y tome un valor determinado es p si $y = 1$, y $1 - p$ si $y = 0$. Así, la verosimilitud puede expresarse como:

$$L(\alpha_0, \alpha) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1 - y_i} \tag{2.22}$$

El producto puede transformarse en una suma aplicando el logaritmo:

$$l(\alpha_0, \alpha) = \sum_{i=1}^{n} \left[y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)) \right] \tag{2.23}$$

$$l(\alpha_0, \alpha) = \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1 - p(x_i)}\right) \tag{2.24}$$

Reemplazando la expresión de $p(x)$, se tiene:

$$l(\alpha_0, \alpha) = \sum_{i=0}^{n} \left[-\log\left(1 + e^{\alpha_0 + \alpha \cdot x_i}\right) + y_i(\alpha_0 + \alpha \cdot x_i)\right] \tag{2.25}$$

El siguiente paso es maximizar la función de verosimilitud anterior, ya que en el caso de la regresión logística se implementa el método de ascenso por gradiente (lo opuesto al descenso por gradiente).

Estimación por Máxima Verosimilitud (MLE: Maximum Likelihood Estimation)
Es un método para estimar los parámetros de una distribución de probabilidad mediante la maximización de una función de verosimilitud, con el objetivo de aumentar la probabilidad de observar los datos disponibles. La MLE se puede obtener derivando la ecuación de verosimilitud respecto a los diferentes parámetros y estableciendo la derivada igual a cero.

Por ejemplo, la derivada con respecto a uno de los componentes del parámetro α, es decir, a_j, está dada por:

$$\frac{\partial l}{\partial \alpha_j} = \sum_{i=0}^{n} \left(y_i - p(x_i; \alpha_0, \alpha)\right) x_{ij} \tag{2.26}$$

La fuente usada para toda la sección fue: [12, 13].

2.3.4 *Support Vector Machine (SVM)*

El modelo de SVM es una técnica supervisada utilizada principalmente para clasificación binaria. Su objetivo es encontrar un hiperplano óptimo que separe las clases, maximizando el margen entre ellas. En este proyecto, se utilizó un kernel *RBF* (Radial Basis Function), ya que permite manejar relaciones no lineales entre las variables predictoras y la variable objetivo. Este modelo es especialmente útil para capturar patrones complejos en los datos.

Para el modelo de *Máquinas de Vectores de Soporte (SVM)*, se evaluaron distintas configuraciones de hiperparámetros utilizando tres funciones kernel: `rbf`, `poly` y `sigmoid`. En total se realizaron nueve simulaciones variando el valor del parámetro de penalización C, el grado del polinomio (`degree`) para el kernel `poly`, y el parámetro `gamma` que controla la influencia de cada punto de entrenamiento. Los valores utilizados fueron los siguientes:

- Kernel RBF: $C = \{1, 10, 50\}$, `gamma` = {0.01, 0.1, 1.0}.
- Kernel Polinomial: $C = \{1, 10, 50\}$, `degree` = {2, 3, 4}, `gamma` = {0.01, 0.1, 1.0}.
- Kernel Sigmoide: $C = \{1, 10, 50\}$, `gamma` = {0.01, 0.1, 1.0}.

Fundamentos Matemáticos de Support Vector Machine (SVM)

Consideremos un problema de clasificación binaria con dos clases etiquetadas como $+1$ y -1. Se dispone de un conjunto de entrenamiento compuesto por vectores de características de entrada X y sus correspondientes etiquetas de clase Y.

La ecuación del hiperplano lineal que separa ambas clases se puede expresar como:

$$w^T x + b = 0 \tag{2.27}$$

donde:

- w es el vector normal al hiperplano,
- x es un vector de entrada del espacio de características,
- b es el sesgo (bias) o término independiente.

Distancia de un Punto al Hiperplano

La distancia entre un punto de datos x_i y el hiperplano de decisión puede calcularse como:

$$d_i = \frac{w^T x_i + b}{\|w\|} \tag{2.28}$$

donde $\|w\|$ representa la norma euclidiana del vector de pesos w.

Clasificador SVM Lineal

La predicción del modelo depende de la posición del punto con respecto al hiperplano:

$$\hat{y} = \begin{cases} 1 & \text{si } w^T x + b \geq 0 \\ 0 & \text{si } w^T x + b < 0 \end{cases} \tag{2.29}$$

donde $\hat{y}$ es la clase predicha para un punto de datos.

Problema de Optimización para SVM

Para un conjunto de datos linealmente separable, el objetivo es encontrar el hiperplano que maximice el margen entre las dos clases, garantizando al mismo tiempo que todos los puntos sean clasificados correctamente. Esto conduce al siguiente problema de optimización:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2.30}$$

Sujeto a la restricción:

$$y_i(w^T x_i + b) \geq 1 \quad \text{para } i = 1, 2, 3, \ldots, m \tag{2.31}$$

Donde:

- y_i es la etiqueta de clase (+1 o -1) para la instancia de entrenamiento i,
- x_i es el vector de características de la instancia i,
- m es el número total de instancias de entrenamiento.

La condición $y_i(w^T x_i + b) \geq 1$ asegura que cada punto esté correctamente clasificado y se encuentre fuera del margen.

Clasificador SVM Lineal con Margen Suave

En presencia de valores atípicos o datos no separables linealmente, SVM permite cierta cantidad de errores de clasificación mediante la introducción de variables de holgura ζ_i. El problema de optimización se modifica de la siguiente manera:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \zeta_i \tag{2.32}$$

Sujeto a las restricciones:

$$y_i(w^T x_i + b) \geq 1 - \zeta_i \quad \text{y} \quad \zeta_i \geq 0 \quad \text{para } i = 1, 2, \ldots, m \tag{2.33}$$

Donde:

- C es un parámetro de regularización que controla el equilibrio entre maximizar el margen y penalizar errores de clasificación.
- ζ_i son variables de holgura que representan el grado de violación del margen para cada punto de datos.

Problema Dual para SVM

El problema dual consiste en maximizar los multiplicadores de Lagrange asociados a los vectores de soporte. Esta transformación permite resolver el problema de optimización usando funciones núcleo (kernels) para clasificación no lineal. La función objetivo dual se define como:

$$\max_{\alpha} \left(-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j t_i t_j K(x_i, x_j) + \sum_{i=1}^{m} \alpha_i \right) \tag{2.34}$$

Donde:

- α_i son los multiplicadores de Lagrange asociados a la muestra i,
- t_i es la etiqueta de clase para la muestra i (ya sea +1 o -1),
- $K(x_i, x_j)$ es la función núcleo que calcula la similitud entre los puntos x_i y x_j. Esta función permite a SVM manejar problemas de clasificación no lineal al mapear los datos a un espacio de mayor dimensión.

La formulación dual optimiza los multiplicadores de Lagrange α_i, y los vectores de soporte son aquellas muestras de entrenamiento para las cuales $\alpha_i > 0$.

Frontera de Decisión de SVM

Una vez que se resuelve el problema dual, el límite de decisión se define como:

$$w = \sum_{i=1}^{m} \alpha_i t_i K(x_i, x) + b \tag{2.35}$$

Donde w es el vector de pesos, x es el punto de prueba y b es el término de sesgo (bias).

Finalmente, el término b se determina utilizando los vectores de soporte, los cuales satisfacen:

$$t_i(w^T x_i - b) = 1 \quad \Rightarrow \quad b = w^T x_i - t_i \tag{2.36}$$

Donde x_i es cualquier vector de soporte.

Este apartado concluye el marco matemático del algoritmo de *Máquinas de Vectores de Soporte* (SVM, por sus siglas en inglés), el cual permite realizar clasificación tanto lineal como no lineal mediante la formulación dual y el uso del *truco del núcleo* (*kernel trick*).

La fuente usada para toda la sección fue: [14].

2.3.5 *Proceso de Selección de Modelos*

La selección de estos modelos se basó en su pertinencia para abordar problemas de clasificación binaria y su historial de desempeño en proyectos similares. La inclusión de modelos tanto lineales (como la Regresión Logística) como no lineales (SVM, Red Neuronal y XGBoost) permitió comparar su capacidad para capturar patrones en los datos y evaluar su capacidad predictiva bajo diferentes enfoques.

Para respaldar esta selección, se consultaron diversas referencias en la literatura. Por ejemplo:

- Para **SVM**, se consideró su capacidad para manejar problemas de clasificación en datos con patrones complejos, como se describe en [14, 15].

- La **Red Neuronal** fue seleccionada siguiendo estudios que destacan su flexibilidad en problemas de clasificación no lineales, como se describe en [16].
- La **Regresión Logística** fue elegida por su simplicidad y capacidad de interpretación, como se detalla en [17].
- **XGBoost** fue seleccionado por su robustez y eficiencia, destacada en proyectos de clasificación, como se detalla en [11, 10, 9, 8, 7].

2.3.6 *Justificación de la Comparación*

La comparación de estos modelos se realizó para identificar el enfoque más adecuado según las métricas de desempeño, como la precisión, el *F1-Score* y el área bajo la curva ROC (AUC-ROC). Cada modelo aporta ventajas únicas:

- SVM y la Red Neuronal son ideales para capturar relaciones no lineales.
- La Regresión Logística proporciona una solución interpretable y eficiente.
- XGBoost ofrece un balance entre precisión y velocidad de entrenamiento, especialmente en datos estructurados.
- Este enfoque comparativo no solo permitió seleccionar el mejor modelo para el problema planteado, sino también explorar diferentes metodologías para abordar problemas de clasificación binaria en el contexto logístico.

2.4 *Descripción de las Métricas*

En este proyecto se utilizaron diversas métricas para la comparación y evaluación de los modelos de predicción utilizados para determinar si un paquete llegaría a tiempo a su destino o llegaría tarde. Las métricas elegidas fueron fundamentales para evaluar la precisión, la capacidad de generalización y el rendimiento general de cada modelo. A continuación, se describen las métricas utilizadas y el proceso de selección que justifica su pertinencia.

Las métricas clave seleccionadas fueron:

1. **Precisión (Accuracy)**
   La precisión, también conocida como *accuracy*, es una métrica fundamental en problemas de clasificación, ya que evalúa la proporción de predicciones correctas realizadas por un modelo respecto al total de predicciones. Su cálculo es sencillo: se divide el número de aciertos entre la cantidad total de casos evaluados, obteniendo un valor entre 0 y 1 o, de forma equivalente, un

porcentaje. Por ejemplo, si un modelo realiza 90 clasificaciones correctas sobre un total de 100, su precisión es del 90 %. Matemáticamente, se expresa como:

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.37}$$

donde TP representa los verdaderos positivos, TN los verdaderos negativos, FP los falsos positivos y FN los falsos negativos. Esta métrica refleja de manera intuitiva la capacidad general del modelo para clasificar correctamente los datos, por lo que suele ser la primera medida de rendimiento que se consulta. No obstante, su simplicidad también representa una limitación en escenarios con clases desbalanceadas. En estos casos, un modelo que predice mayoritariamente la clase dominante puede alcanzar una alta precisión, incluso si falla sistemáticamente en identificar la clase minoritaria. Además, cuando existen costos asimétricos entre los falsos positivos y los falsos negativos—como en aplicaciones médicas o de ciberseguridad—la precisión puede ser engañosa. Dado que esta métrica considera conjuntamente los cuatro componentes de la matriz de confusión, resulta adecuada cuando las clases están balanceadas y no existe un sesgo de importancia entre ellas. Sin embargo, para una evaluación más robusta, se recomienda complementarla con otras métricas como la sensibilidad, la especificidad o la F1-score, especialmente en entornos donde los errores tienen diferentes implicaciones prácticas [18, 19, 20, 21].

2. **Precisión**

   La precisión, o *precision*, es una métrica empleada en clasificación binaria para evaluar la proporción de instancias clasificadas como positivas que realmente pertenecen a la clase positiva. Es decir, mide cuántas de las predicciones positivas realizadas por el modelo son correctas. Se calcula dividiendo los verdaderos positivos (TP) entre el total de instancias predichas como positivas, es decir, la suma de verdaderos positivos y falsos positivos ($TP + FP$). Matemáticamente, se expresa como:

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP} \tag{2.38}$$

   Esta métrica resulta especialmente útil en contextos donde los falsos positivos tienen un costo elevado, como en sistemas de detección de spam, fraudes o diagnósticos médicos. Una precisión alta indica que el modelo comete pocos errores al

identificar la clase positiva, es decir, que pocas de las instancias que clasifica como positivas son incorrectas. Sin embargo, su interpretación debe realizarse con cuidado cuando se trabaja con conjuntos de datos desequilibrados o cuando la clase positiva es poco frecuente, ya que en esos casos puede ser una métrica insuficiente por sí sola. Además, la precisión se relaciona inversamente con la métrica de *recall* o sensibilidad: al aumentar el umbral de clasificación, se reducen los falsos positivos (mejorando la precisión), pero se incrementan los falsos negativos (afectando el recall), y viceversa. Por esta razón, suele analizarse en conjunto con otras métricas como la sensibilidad y la F1-score, que ofrecen un panorama más completo del desempeño del modelo. En resumen, la precisión es clave cuando se prioriza minimizar las falsas alarmas en la clasificación, aunque debe ser interpretada en función del problema y del equilibrio entre clases [18, 19, 20, 21].

3. **Sensibilidad o Recall**

   La métrica *recall*, también conocida como sensibilidad o tasa de verdaderos positivos (*True Positive Rate, TPR*), mide la capacidad de un modelo de aprendizaje supervisado para identificar correctamente las instancias que realmente pertenecen a la clase positiva. En otras palabras, cuantifica qué fracción de los casos positivos reales fueron detectados por el modelo. Se calcula dividiendo el número de verdaderos positivos (TP) entre el total de positivos reales, es decir, la suma de verdaderos positivos y falsos negativos ($TP + FN$). Su expresión matemática es:

$$\text{Recall} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN} \tag{2.39}$$

   El *recall* adquiere gran relevancia en problemas donde es crítico minimizar los falsos negativos, como en detección de enfermedades, fraudes o amenazas de seguridad. Por ejemplo, en un sistema de clasificación de correos electrónicos, el *recall* indica la proporción de correos spam que fueron correctamente detectados como tal. Un modelo perfecto tendría un *recall* de 1.0, es decir, identificaría correctamente el 100 % de las instancias positivas, sin omitir ninguna. No obstante, en contextos donde la clase positiva es poco frecuente (datasets desequilibrados), esta métrica puede perder relevancia si se considera de forma aislada. Además, existe una relación inversa con la precisión: al ajustar el umbral de clasificación para detectar más positivos, se incrementa el *recall*, pero también los falsos positivos, lo que puede disminuir la precisión. Por esta razón, el *recall* suele

analizarse junto con otras métricas como la precisión y la F1-score, proporcionando así una evaluación más equilibrada del desempeño del modelo ante distintos tipos de errores [18, 21].

4. **F1-Score**
   La métrica *F1-score* es el promedio armónico entre la precisión (*precision*) y la sensibilidad (*recall*), y se utiliza frecuentemente para evaluar el rendimiento de modelos de clasificación, tanto binarios como multiclase. A diferencia de la precisión o la sensibilidad consideradas de forma aislada, el F1-score combina ambas métricas en un único valor, ofreciendo una visión más equilibrada del desempeño del modelo, especialmente en contextos con clases desbalanceadas. Su cálculo se realiza mediante la siguiente fórmula:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.40}$$

   El F1-score toma valores entre 0 y 1, donde 1 representa un modelo perfecto y 0 indica un desempeño deficiente. Esta métrica resulta especialmente útil en aplicaciones donde es importante encontrar un equilibrio entre los errores tipo I (falsos positivos) y tipo II (falsos negativos), como en detección de fraudes, mantenimiento predictivo o clasificación médica. Dado que utiliza el promedio armónico en lugar del promedio aritmético, penaliza fuertemente los valores extremos: un modelo con alta precisión pero baja sensibilidad, o viceversa, tendrá un F1-score bajo. No obstante, también tiene limitaciones, ya que asigna el mismo peso a ambas métricas, lo cual no siempre es apropiado. En escenarios donde uno de los dos tipos de error es mucho más costoso que el otro —por ejemplo, en detección de fallas críticas donde los falsos negativos son inaceptables— puede ser preferible optimizar directamente la sensibilidad. A pesar de esto, el F1-score sigue siendo una métrica sólida y ampliamente adoptada, ya que ofrece una síntesis clara del balance entre identificar correctamente los casos positivos y evitar clasificaciones erróneas [22, 23, 24, 25].

5. **Curva ROC y AUC (Área Bajo la Curva ROC)**
   La *curva ROC* (del inglés *Receiver Operating Characteristic*) es una herramienta gráfica utilizada para evaluar el desempeño de un modelo de clasificación binaria a través de distintos umbrales de decisión. Esta curva muestra la relación entre la tasa de verdaderos positivos (*True Positive Rate, TPR*) y la tasa de falsos positivos (*False Positive Rate, FPR*):

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}$$

Cada punto en la curva representa un umbral diferente aplicado sobre las probabilidades de predicción del modelo. De esta forma, la curva proporciona una representación visual de las compensaciones entre sensibilidad y especificidad. El análisis de la curva ROC permite identificar cómo se comporta el modelo al variar el umbral: a mayor TPR, mejor la capacidad del modelo para identificar correctamente casos positivos; a menor FPR, mejor su habilidad para evitar clasificar erróneamente casos negativos como positivos.

Una métrica derivada de esta curva es el *Área Bajo la Curva* (*Area Under the Curve, AUC*), que resume el rendimiento global del modelo en un único valor entre 0 y 1. Un AUC igual a 1.0 indica un modelo perfecto, capaz de clasificar correctamente todos los casos positivos y negativos. En cambio, un AUC de 0.5 refleja un modelo sin capacidad de discriminación, equivalente a realizar predicciones aleatorias. Por lo tanto, cuanto mayor sea el AUC, mejor será el rendimiento discriminativo del modelo.

$$\text{AUC} = \int_0^1 \text{TPR}(x)\, dx \tag{2.41}$$

Desde una perspectiva práctica, el AUC representa la probabilidad de que el modelo asigne una mayor puntuación de probabilidad a una instancia positiva seleccionada al azar que a una instancia negativa también aleatoria. Por esta razón, el AUC es frecuentemente utilizado como métrica comparativa entre diferentes modelos, especialmente cuando se dispone de un conjunto de datos equilibrado. No obstante, en situaciones con datos desequilibrados, se recomienda complementar este análisis con curvas de precisión-recall.

La curva ROC y el AUC son ampliamente utilizadas en aplicaciones de inteligencia artificial, aprendizaje profundo y redes neuronales convolucionales (CNN), ya que permiten evaluar modelos en función de todos los umbrales posibles y no solamente en un punto fijo. Esta flexibilidad permite, además, seleccionar umbrales óptimos según los costos relativos de errores: si los falsos positivos son costosos, es preferible un umbral que minimice la FPR, aun sacrificando algo de TPR; por el contrario, si los falsos negativos tienen un impacto mayor, conviene elegir umbrales que maximicen la TPR. Esta capacidad para visualizar y balancear las decisiones del modelo convierte a la curva ROC en una herramienta indispensable en problemas de clasificación binaria [26, 27, 28].

6. **Matriz de Confusión**

   La *matriz de confusión*, también conocida como matriz de error,

es una herramienta fundamental para evaluar el rendimiento de modelos de clasificación, ya sea binaria o multiclase. Se trata de una tabla que resume las predicciones realizadas por el modelo frente a los valores reales, proporcionando así una visión detallada de los aciertos y errores del clasificador. A través de esta representación, se pueden analizar de forma directa las verdaderas predicciones positivas y negativas, así como los errores de tipo I (falsos positivos) y tipo II (falsos negativos).

En el caso más común de clasificación binaria, la matriz se estructura en una tabla de 2×2, cuyos elementos se definen de la siguiente manera:

- **Verdadero Positivo (TP)**: el modelo predice una clase positiva y efectivamente la instancia pertenece a dicha clase.
- **Verdadero Negativo (TN)**: el modelo predice una clase negativa y la instancia es realmente negativa.
- **Falso Positivo (FP)**: el modelo predice positivo, pero la instancia es negativa (error tipo I).
- **Falso Negativo (FN)**: el modelo predice negativo, pero la instancia es positiva (error tipo II).

Esta estructura permite calcular múltiples métricas de rendimiento relevantes, como la precisión (*precision*), la sensibilidad (*recall*), la exactitud (*accuracy*) y el F1-score, cada una con una utilidad particular dependiendo del contexto del problema. Por ejemplo:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.42}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.43}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.44}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.45}$$

Una de las ventajas clave de la matriz de confusión es su capacidad para mostrar con claridad dónde y cómo se equivoca un modelo, lo que permite tomar decisiones más informadas sobre cómo ajustar los umbrales de decisión, modificar el conjunto de entrenamiento o emplear técnicas de balanceo de clases. También es útil para identificar sesgos en modelos entrenados con conjuntos de datos desbalanceados.

En problemas multiclase, la matriz de confusión se amplía para incluir todas las clases posibles. Por ejemplo, en un modelo

financiero que clasifica facturas como pagadas .ª tiempo", "tarde.º "muy tarde", la matriz tendrá dimensiones 3×3. Cada celda mostrará cuántas veces una clase real fue clasificada como una clase predicha específica, lo que permite detectar patrones de error y relaciones ordinales entre clases.

Por su capacidad para proporcionar información granular y práctica, la matriz de confusión se considera una herramienta esencial en el análisis de modelos de clasificación. Su interpretación adecuada permite evaluar no solo el rendimiento global, sino también la calidad de las decisiones que toma el modelo en contextos críticos como el diagnóstico médico, la detección de fraudes, la clasificación de correos electrónicos o la gestión de riesgo financiero [23].

La selección de estas métricas fue guiada por la necesidad de evaluar no solo la precisión global de los modelos, sino también su capacidad para manejar los desbalances en las clases y su efectividad en la identificación de los casos más importantes (paquetes que llegan tarde). Además, estas métricas permitieron comparar de manera justa los diferentes modelos, entre ellos la regresión logística, que resultó ser el más efectivo en este contexto.

2.5 *Descripción de los Experimentos o Simulaciones*

Para evaluar la efectividad de distintos modelos de aprendizaje automático en la predicción de la entrega o retraso de paquetes, se llevaron a cabo diversas simulaciones computacionales. Se implementaron y compararon cuatro modelos: **Regresión Logística, XGBoost, SVM y Red Neuronal Artificial (RNA)**. Cada modelo fue ajustado con diferentes configuraciones de hiperparámetros para optimizar su desempeño y determinar la mejor arquitectura posible.

2.5.1 *Red Neuronal (Perceptrón)*

Para evaluar la efectividad de un modelo basado en **redes neuronales artificiales (RNA)** en la predicción de la entrega o retraso de paquetes, se realizaron diversas simulaciones computacionales. Se probaron diferentes configuraciones de hiperparámetros con el objetivo de optimizar el desempeño del modelo y encontrar la mejor arquitectura posible.

Configuración del Modelo

Se implementó una red neuronal multicapa (MLP) utilizando la biblioteca **TensorFlow y Keras**. La estructura base del modelo fue la siguiente:

- Capa de entrada: 129 neuronas (correspondientes a las características del dataset).
- Capas ocultas: Se evaluaron arquitecturas con 10, 15, 20, 25 y 30 neuronas en una única capa oculta.
- Capa de salida: 1 neurona con función de activación *sigmoide*, dado que el problema es de clasificación binaria (entrega a tiempo vs. retraso).
- Función de pérdida: *Binary Cross-Entropy*, apropiada para problemas de clasificación.
- Optimizador: *Adam*, con variaciones en la tasa de aprendizaje para evaluar su impacto.
- Regularización: Se aplicó *Dropout* con valores entre 0.2 y 0.5 para mitigar el sobreajuste.

Conjunto de Datos y Preprocesamiento

El conjunto de datos utilizado proviene de información operativa sobre los envíos y contiene **129 variables predictoras**. Se aplicaron los siguientes pasos de preprocesamiento:

- Eliminación de columnas irrelevantes o con valores faltantes significativos.
- Codificación de variables categóricas mediante *One-Hot Encoding*.
- Estandarización de los valores numéricos usando *StandardScaler* de scikit-learn.
- División del conjunto de datos en **70 % entrenamiento y 30 % prueba** para evaluar la generalización del modelo.
- Balanceo de clases mediante *Random OverSampling*, garantizando una distribución equitativa de la variable objetivo.

Diseño de las Simulaciones

Se llevaron a cabo cinco simulaciones independientes con diferentes combinaciones de hiperparámetros. Cada modelo fue entrenado utilizando *Keras* y evaluado en un conjunto de prueba, registrando métricas de desempeño como **accuracy, precision, recall, F1-score y AUC-ROC**.

- **Número de neuronas en la capa oculta**: 10, 15, 20, 25 y 30.
- **Tasa de aprendizaje**: 0.001, 0.0005, 0.0003, 0.0001 y 0.00005.
- **Dropout**: 0.2, 0.3, 0.4, 0.5 y 0.2.
- **Número de épocas de entrenamiento**: 20, 25, 30, 35 y 40.
- **Tamaño del batch**: 32, 32, 64, 64 y 128.

Resultados de las Simulaciones

Los resultados de las simulaciones, que se presentan en la Tabla 2.7, demostraron que el modelo con la mejor configuración fue aquel con:

- 10 neuronas en la capa oculta.
- Tasa de aprendizaje de 0.001.
- Dropout de 0.2.
- 20 épocas de entrenamiento y batch size de 32.
- Este modelo obtuvo un **accuracy de 97.65 %**, una **precisión de 99.09 %** y un **AUC-ROC de 97.66 %**, indicando una alta capacidad de predicción y generalización.

Neurons	Learning Rate	Dropout	Epochs	Batch Size	Accuracy	Precision	Recall	F1-Score	ROC-AUC
10	0.0010	0.20	20	32	0.98	0.99	0.96	0.98	0.98
15	0.0005	0.30	25	32	0.96	0.98	0.94	0.96	0.96
20	0.0003	0.40	30	64	0.89	0.90	0.88	0.89	0.89
25	0.0001	0.50	35	64	0.84	0.82	0.87	0.85	0.84
30	0.00005	0.20	40	128	0.81	0.81	0.82	0.82	0.81

Tabla 2.7: Resultados de la red neuronal con diferentes combinaciones de hiperparámetros.

Conclusión

A partir de los experimentos realizados, se identificó la arquitectura óptima para la predicción de la puntualidad en la entrega de paquetes. Los resultados sugieren que un modelo **moderadamente complejo** (10 neuronas, tasa de aprendizaje de 0.001, dropout de 0.2) logra un equilibrio adecuado entre precisión y generalización.

2.5.2 *Gradient Boosting XGBoost*

Para evaluar el rendimiento de XGBoost en la predicción de la entrega de paquetes a tiempo, se llevaron a cabo diversas simulaciones con diferentes configuraciones de hiperparámetros. Se optimizaron valores como la tasa de aprendizaje, la profundidad de los árboles y el número de estimadores para encontrar el mejor balance entre precisión y generalización.

Configuración del Modelo

XGBoost es un algoritmo de boosting basado en árboles de decisión que mejora iterativamente el desempeño minimizando errores. Las configuraciones del modelo fueron:

- **Learning Rate**: 0.0075, 0.005, 0.0025, 0.001.
- **Profundidad Máxima de los Árboles (max_depth)**: 2 o 3.
- **Número de Estimadores (n_estimators)**: 100, 150, 200, 250, 300.
- **Subsample**: 0.6 - 0.8 para evitar sobreajuste.
- **Colsample_bytree**: 0.7 - 0.9, para diversificar la selección de características en cada árbol.
- **Early Stopping**: 10 iteraciones sin mejora en la validación.

Conjunto de Datos y Preprocesamiento
El conjunto de datos contiene **129 variables predictoras**. Se realizaron los siguientes pasos de preprocesamiento:

- Eliminación de columnas irrelevantes o con valores faltantes significativos.
- Codificación de variables categóricas mediante One-Hot Encoding.
- Normalización de los valores numéricos.
- División del conjunto de datos en 70 % entrenamiento y 30 % prueba.

Resultados de las Simulaciones
Se ejecutaron cinco simulaciones con diferentes combinaciones de hiperparámetros, variando la tasa de aprendizaje, la profundidad del árbol, el número de estimadores y los valores de subsample. Cada modelo fue entrenado y evaluado en el conjunto de prueba. Los resultados obtenidos para las simulaciones, que se muestran en la Tabla 2.8, fueron los siguientes:

Learning Rate	Max Depth	N Estimators	Subsample	Colsample Bytree	Accuracy	Precision	Recall	F1-Score	ROC-AUC
0.0075	2	100	0.7	0.8	0.94	1.00	0.89	0.94	0.94
0.0050	2	150	0.8	0.9	0.95	1.00	0.90	0.95	0.95
0.0025	3	200	0.7	0.8	0.99	1.00	0.98	0.99	0.99
0.0010	2	250	0.6	0.7	0.92	1.00	0.83	0.92	0.92
0.0010	3	300	0.7	0.9	0.99	1.00	0.99	0.99	0.99

Tabla 2.8: Resultados de XGBoost con diferentes combinaciones de hiperparámetros.

Conclusión
El modelo que presentó el mejor equilibrio entre precisión y generalización fue el que utilizó una tasa de aprendizaje de 0.005, una profundidad máxima de 2 y 150 estimadores. Este modelo obtuvo un accuracy de 94.9 % y un ROC-AUC de 94.9 %, indicando que logró una buena discriminación sin caer en sobreajuste. Comparado con la Red Neuronal, XGBoost mostró ser más estable y rápido en entrenamiento, pero con un menor recall, lo que significa que podría estar perdiendo algunos casos positivos.

2.5.3 *Regresión Logística*

Para evaluar el rendimiento de la regresión logística en la predicción de la entrega de paquetes a tiempo, se realizaron diversas simulaciones con distintas configuraciones de hiperparámetros. Se exploraron diferentes solvers y niveles de regularización para determinar la mejor combinación en términos de precisión y generalización.

Configuración del Modelo
La regresión logística se probó con los siguientes hiperparámetros:

- **Número máximo de iteraciones (max_iter)**: 200.
- **Solvers utilizados**: newton-cg, lbfgs, liblinear.
- **Parámetro de regularización (C)**: 1.0, 0.5, 0.1.

Conjunto de Datos y Preprocesamiento
El conjunto de datos contiene **129 variables predictoras**. Se aplicaron los siguientes pasos de preprocesamiento:

- Eliminación de columnas irrelevantes o con valores faltantes significativos.
- Codificación de variables categóricas mediante One-Hot Encoding.
- Normalización de los valores numéricos.
- División del conjunto de datos en 70 % entrenamiento y 30 % prueba.

Resultados de las Simulaciones
Se ejecutaron nueve simulaciones con diferentes combinaciones de solver y niveles de regularización. Cada modelo fue entrenado y evaluado en el conjunto de prueba. Los resultados obtenidos para las simulaciones, que se muestran en la Tabla 2.9, fueron los siguientes:

Max Iter	Solver	C	Accuracy	Precision	Recall	F1-Score	ROC-AUC
200	newton-cg	1.0	0.99	1.00	0.99	0.99	0.99
200	newton-cg	0.5	0.99	0.99	0.98	0.99	0.99
200	newton-cg	0.1	0.98	0.99	0.96	0.98	0.98
200	lbfgs	1.0	0.99	1.00	0.99	0.99	0.99
200	lbfgs	0.5	0.99	0.99	0.98	0.99	0.99
200	lbfgs	0.1	0.98	0.99	0.96	0.98	0.98
200	liblinear	1.0	0.99	1.00	0.99	0.99	0.99
200	liblinear	0.5	0.98	0.99	0.97	0.98	0.98
200	liblinear	0.1	0.98	0.99	0.96	0.97	0.98

Tabla 2.9: Resultados de regresión logística con diferentes solvers y valores de *C*.

Conclusión

El mejor modelo encontrado fue el que utilizó el solver **newton-cg** con un parámetro de regularización **C=1.0**, obteniendo un accuracy de 99.24 % y un ROC-AUC de 99.24 %. Esto indica que el modelo logró un excelente balance entre precisión y recall.

Se encontró que los valores más bajos de **C** (mayor regularización) reducen el desempeño del modelo, mientras que solvers como *liblinear* tuvieron un desempeño ligeramente menor en recall en comparación con *newton-cg* y *lbfgs*.

2.5.4 SVM

Para evaluar el rendimiento del modelo de Máquinas de Vectores de Soporte (SVM) en la predicción de la entrega de paquetes, se llevaron a cabo diversas simulaciones variando los hiperparámetros clave, tales como el kernel, el parámetro de regularización C, el parámetro γ y el grado en los modelos polinomiales.

Configuración del Modelo

Se realizaron pruebas con los siguientes valores de hiperparámetros:

- **Kernels evaluados**: RBF, Polinomial, Sigmoide.
- **Valores de C**: 1.0, 10.0, 50.0.
- **Valores de Gamma**: 0.01, 0.1, 1.0.
- **Grados probados en polinomiales**: 2, 3, 4.

Conjunto de Datos y Preprocesamiento

El conjunto de datos contiene **129 variables predictoras**. Se aplicaron los siguientes pasos de preprocesamiento:

- Eliminación de columnas irrelevantes o con valores faltantes significativos.
- Codificación de variables categóricas mediante One-Hot Encoding.
- Normalización de los valores numéricos.
- División del conjunto de datos en 70 % entrenamiento y 30 % prueba.

Resultados de las Simulaciones

Se realizaron nueve simulaciones variando los valores de C, γ y el grado en los modelos polinomiales. Cada modelo fue entrenado y evaluado en el conjunto de prueba.Los resultados obtenidos para las simulaciones, que se muestran en la Tabla 2.10, fueron los siguientes:

Kernel	C	Degree	Gamma	Accuracy	Precision	Recall	F1-Score	ROC-AUC
rbf	1.0	N/A	0.01	0.92	0.95	0.88	0.92	0.92
rbf	10.0	N/A	0.10	0.92	0.95	0.89	0.92	0.92
rbf	50.0	N/A	1.00	0.91	0.87	0.95	0.91	0.91
poly	1.0	2	0.01	0.81	0.80	0.84	0.82	0.81
poly	10.0	3	0.01	0.98	0.98	0.99	0.98	0.98
poly	50.0	4	0.01	0.97	0.96	0.97	0.97	0.97
sigmoid	1.0	N/A	0.01	0.95	0.95	0.92	0.95	0.95
sigmoid	10.0	N/A	0.10	0.78	0.78	0.79	0.78	0.78
sigmoid	50.0	N/A	1.00	0.69	0.70	0.67	0.68	0.69

Tabla 2.10: Resultados de SVM con diferentes kernels y combinaciones de hiperparámetros.

Conclusión

El modelo que presentó el mejor desempeño fue el que utilizó el kernel **Polinomial** con $C = 10.0$, $degree = 3$ y $\gamma = 0.1$, logrando un accuracy

de **98.16 %** y un ROC-AUC de **98.16 %**. Este modelo mostró el mejor balance entre precisión y recall.

El kernel **RBF** con $C = 10.0$ y $\gamma = 0.1$ fue la segunda mejor opción, con un accuracy de **92.25 %**, siendo una alternativa competitiva.

Por otro lado, el kernel **Sigmoide** tuvo un desempeño inferior, especialmente con $C = 50.0$, donde el accuracy cayó a **68.81 %**. Esto indica que no es un modelo adecuado para este problema.

3 Resultados y Discusión

En este capítulo se presentan los resultados obtenidos del desarrollo de este trabajo y una discusión sobre el objeto de estudio.

3.1 Resultados y Discusión

Resumen de los Modelos Evaluados

En este trabajo se evaluaron cuatro modelos de clasificación para predecir si un paquete llegaría a tiempo o con retraso. Los modelos probados fueron: Regresión Logística, XGBoost, Redes Neuronales y Máquinas de Vectores de Soporte (SVM). Cada modelo fue entrenado con múltiples configuraciones de hiperparámetros y validado en un conjunto de prueba para determinar su desempeño.

Resultados de las Simulaciones

Los resultados de las simulaciones realizadas con cada modelo se presentan en la Tabla 3.1. Se evaluaron métricas clave como la exactitud (accuracy), precisión, recall, F1-score y ROC-AUC, con el objetivo de determinar el mejor desempeño general.

Modelo	Accuracy	Precision	Recall	F1-score	ROC-AUC
Regresión Logística	0.99	1.00	0.99	0.99	0.99
XGBoost	0.95	1.00	0.90	0.95	0.95
Red Neuronal	0.98	0.99	0.96	0.98	0.98
SVM (*Polinomial*, C=10.0, deg=3)	0.98	0.98	0.99	0.98	0.98

Tabla 3.1: Comparación de los resultados obtenidos en las simulaciones de los cuatro modelos evaluados.

Análisis Comparativo de Resultados

El modelo con el mejor desempeño global fue la **Regresión Logística**, obteniendo un accuracy del **99.24 %**, con la mayor precisión y recall entre todos los modelos evaluados. Esto sugiere que la Regresión Logística fue capaz de clasificar correctamente la mayoría de los casos sin incurrir en sobreajuste.

Por otro lado, la **Red Neuronal** presentó un rendimiento muy competitivo, con un accuracy de **97.7 %** y un excelente balance entre precisión y recall. La ventaja de este modelo radica en su capacidad para aprender patrones complejos en los datos, aunque con un mayor

costo computacional y tiempo de entrenamiento comparado con los otros modelos.

El modelo de **XGBoost** obtuvo un accuracy del **94.9 %**, destacándose por su alta precisión (100 %), lo que significa que prácticamente no generó falsos positivos. Sin embargo, su recall fue menor (89.8 %), lo que indica que el modelo dejó de identificar algunos casos de paquetes retrasados. XGBoost es una opción sólida debido a su interpretabilidad y eficiencia en el entrenamiento.

Finalmente, el mejor modelo de **SVM**, utilizando un kernel polinomial con $C = 10.0$ y $degree = 3$, logró un accuracy de **98.16 %**, con un buen balance de precisión y recall. Este modelo mostró un buen desempeño general, aunque el costo computacional es considerablemente mayor en comparación con la Regresión Logística y XGBoost.

Desafíos en la Construcción del Modelo

Si bien el enfoque principal de este trabajo fue la evaluación de modelos de clasificación, es importante destacar que el proceso más complejo y que requirió mayor tiempo no fue la construcción del modelo en sí, sino la preparación y procesamiento de los datos. Para un científico de datos, el desafío más significativo no radica en la implementación de los modelos, ya que estos pueden ser importados fácilmente desde librerías especializadas, sino en la obtención de una buena base de datos y su correcta manipulación.

El análisis exploratorio de datos, la selección de variables relevantes y la limpieza de los datos fueron aspectos cruciales que demandaron una cantidad considerable de tiempo y esfuerzo. Aspectos aparentemente sencillos, como la identificación de las variables que realmente aportaban al modelo, resultaron ser un proceso iterativo y laborioso. La calidad de los datos impactó directamente en el desempeño de los modelos, lo que resalta la importancia de una fase de preprocesamiento rigurosa.

En este proyecto, una parte significativa del tiempo fue invertida en estos procesos previos a la construcción del modelo, destacando la necesidad de metodologías eficientes para la preparación de datos. Al final del día, la construcción de los modelos es una tarea relativamente sencilla en comparación con la complejidad de trabajar con datos en estado crudo y optimizar su calidad para obtener resultados confiables.

Conclusión

Los resultados obtenidos indican que la **Regresión Logística** fue el modelo con mejor desempeño global, seguido de cerca por la Red Neuronal y SVM. XGBoost también presentó buenos resultados, pero con menor recall. La elección del modelo adecuado dependerá

del balance entre precisión, interpretabilidad y costo computacional requerido para la implementación en un entorno real. No obstante, este estudio permitió evidenciar que el mayor reto en la ciencia de datos no es la construcción del modelo, sino la preparación, curación y selección adecuada de los datos, lo cual define en gran medida la calidad de los resultados obtenidos.

4 Conclusiones y Trabajo Futuro

4.1 Conclusiones

Este trabajo tuvo como objetivo evaluar diferentes modelos de clasificación para predecir si un paquete llegaría a tiempo o con retraso, utilizando técnicas de ciencia de datos y aprendizaje automático. Se analizaron cuatro modelos principales: Regresión Logística, XGBoost, Redes Neuronales y Máquinas de Vectores de Soporte (SVM), cada uno con múltiples configuraciones de hiperparámetros para optimizar su desempeño.

Los resultados obtenidos mostraron que la Regresión Logística fue el modelo con mejor desempeño global, alcanzando una precisión del 99.24 % y un balance óptimo entre recall y exactitud. La Red Neuronal y el modelo SVM con kernel polinomial también mostraron un desempeño sobresaliente, con una capacidad elevada para clasificar correctamente los datos. Por otro lado, el modelo de XGBoost destacó por su alta precisión, aunque su recall fue menor, lo que indica que es más conservador al detectar paquetes retrasados.

Uno de los hallazgos más importantes de este trabajo es que, más allá de la selección del modelo, el factor determinante para obtener buenos resultados fue la calidad del preprocesamiento de los datos. La limpieza, selección y transformación de las variables fueron procesos fundamentales que impactaron significativamente en el rendimiento de los modelos. Esto refuerza la idea de que el éxito en ciencia de datos no solo radica en la elección del algoritmo, sino en la preparación adecuada de la información utilizada.

Finalmente, se concluye que la combinación de técnicas de preprocesamiento, junto con una selección adecuada del modelo, permite mejorar la capacidad predictiva en problemas de clasificación logística. Dependiendo del caso de uso, modelos más sofisticados como las Redes Neuronales pueden ser beneficiosos, pero con un mayor costo computacional. Para aplicaciones en tiempo real, modelos más interpretables como la Regresión Logística o XGBoost pueden ser más adecuados.

4.2 Trabajo Futuro

A partir de los resultados obtenidos en este estudio, se identifican diversas oportunidades para mejorar y extender este trabajo en futuras investigaciones.

- **Optimización de hiperparámetros:** Se pueden explorar técnicas como Grid Search o Bayesian Optimization para encontrar configuraciones más óptimas de los modelos evaluados, con el objetivo de mejorar su desempeño.
- **Incorporación de más variables predictoras:** Un análisis más profundo de los datos podría permitir la identificación de nuevas variables que impacten en la predicción de entregas a tiempo, mejorando la robustez del modelo.
- **Validación con datos en producción:** Se recomienda probar estos modelos con datos en tiempo real dentro de un entorno de producción para evaluar su desempeño en condiciones operativas reales.
- **Exploración de modelos adicionales:** Además de los modelos evaluados, podría ser útil probar arquitecturas más avanzadas, como Redes Neuronales Recurrentes (RNNs) o Modelos de Bosques Aleatorios, para comparar su rendimiento.
- **Implementación en un sistema de decisión empresarial:** Los resultados de este estudio podrían aplicarse en una herramienta automatizada para la toma de decisiones en logística, permitiendo optimizar rutas y tiempos de entrega.
- **Explicabilidad de los modelos:** Se recomienda explorar técnicas de interpretabilidad, como SHAP o LIME, para comprender mejor cómo cada variable influye en las predicciones de los modelos y mejorar la transparencia en la toma de decisiones.

Este trabajo proporciona una base sólida para la predicción de entregas en logística, pero aún existen muchas oportunidades para mejorar y expandir la investigación. La aplicación de estos modelos en escenarios más amplios y la integración con herramientas avanzadas de optimización pueden potenciar significativamente su impacto en la industria.

Bibliografía

[1] I. Amazon Web Services, "¿Qué es una red neuronal?" [Online]. Available: https://aws.amazon.com/es/what-is/neural-network/

[2] GeeksforGeeks, "ReLU Activation Function in Deep Learning," Jan. 2025. [Online]. Available: https://www.geeksforgeeks.org/relu-activation-function-in-deep-learning/

[3] G. Kogan, "Neural networks," 2018. [Online]. Available: https://ml4a.github.io/ml4a/neural_networks/

[4] GeeksforGeeks, "geeksforgeeksMSE," May 2025. [Online]. Available: https://www.geeksforgeeks.org/mean-squared-error/

[5] Encord, "Mean Square Error (MSE)." [Online]. Available: https://encord.com/glossary/mean-square-error-mse/

[6] GeeksforGeeks, "What is Adam Optimizer?" Apr. 2025. [Online]. Available: https://www.geeksforgeeks.org/adam-optimizer/

[7] A. Tyagi, "What is XGBoost Algorithm?" Apr. 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[8] Eda Kavlakoglu and Erika Russi, "¿Qué es XGBoost?" May 2024. [Online]. Available: https://www.ibm.com/mx-es/think/topics/xgboost

[9] J. B. Mendoza Vega, "Tutorial: XGBoost en Python," Aug. 2020. [Online]. Available: https://medium.com/@jboscomendoza/tutorial-xgboost-en-python-53e48fc58f73

[10] F. Omarzai, "XGBoost Classification In Depth," Jul. 2024. [Online]. Available: https://medium.com/@fraidoonomarzai99/xgboost-classification-in-depth-979f11ef4bf9

[11] Thommaskevin, "TinyML — XGBoost (Classifier)," Dec. 2023. [Online]. Available: https://medium.com/@thommaskevin/tinyml-xgboost-classifier-795202285779

[12] S. N. Wood, *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press, May 2017.

[13] K. Rai, "The math behind Logistic Regression," Jun. 2020. [Online]. Available: https://medium.com/analytics-vidhya/the-math-behind-logistic-regression-c2f04ca27bca

[14] "Support Vector Machine (SVM) Algorithm," section: Machine Learning. [Online]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[15] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: https://link.springer.com/article/10.1007/BF00994018

[16] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press, 1995. [Online]. Available: https://books.google.com.mx/books?hl=es&lr=&id=ToSoBgAAQBAJ&oi=fnd&pg=PP1&dq=bishop+1995+neural&ots=jO6TtK3yri&sig=WfftZzU1_cUezeZVkA2rYpVSPfs#v=onepage&q=bishop%201995%20neural&f=false

[17] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, segunda ed. Springer, 2009. [Online]. Available: https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf

[18] "Accuracy vs. precision vs. recall in machine learning: what's the difference?" [Online]. Available: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

[19] "Accuracy." [Online]. Available: https://foqum.io/blog/termino/accuracy-exactitud/

[20] "¿Cuál es la diferencia entre exactitud y precisión? [2024] • Asana." [Online]. Available: https://asana.com/es/resources/accuracy-vs-precision

[21] "Clasificación: Exactitud, recuperación, precisión y métricas relacionadas | Machine Learning." [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419

[22] "Understanding and Applying F1 Score: AI Evaluation Essentials with Hands-On Coding Example." [Online]. Available: https://arize.com/blog-course/f1-score/

[23] P. Kashyap, "Understanding Precision, Recall, and F1 Score Metrics," Dec. 2024. [Online].

Available: https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093

[24] "Analyzing machine learning model performance | IBM Cloud Docs." [Online]. Available: https://cloud.ibm.com/docs/watson-knowledge-studio?topic=watson-knowledge-studio-evaluate-ml

[25] "What is F1 Score." [Online]. Available: https://dataheroes.ai/glossary/f1-score/

[26] "Clasificación: ROC y AUC | Machine Learning." [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419

[27] "AUC and the ROC Curve in Machine Learning | DataCamp." [Online]. Available: https://www.datacamp.com/tutorial/auc

[28] A. Bhandari, "Guide to AUC ROC Curve in Machine Learning," Jun. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/