



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

**"EVALUACIÓN ENTRE METODOLOGÍAS DE APRENDIZAJE DE
MÁQUINAS Y MÉTODO TRADICIONAL DE PROGRAMACIÓN PARA LA
NORMALIZACIÓN DE DIRECCIONES EN UNA EMPRESA DE DESPACHO"**

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

FRANCISCA GABRIELA PALANECK QUEZADA

PROFESOR GUÍA:
JUAN ROMERO GODOY

MIEMBROS DE LA COMISIÓN:
EDGARDO SANTIBAÑEZ VIANI
MARÍA FERNANDA VARGAS COURBIS

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL
POR: FRANCISCA GABRIELA PALANECK QUEZADA
FECHA: 2024
PROF. GUÍA: Juan Pablo Romero Godoy

"EVALUACIÓN ENTRE METODOLOGÍAS DE APRENDIZAJE DE MÁQUINAS Y MÉTODO TRADICIONAL DE PROGRAMACIÓN PARA LA NORMALIZACIÓN DE DIRECCIONES EN UNA EMPRESA DE DESPACHO"

Históricamente, el servicio de despacho de paquetes ha sido esencial para el desarrollo y progreso de la sociedad, evolucionando desde mensajeros hasta drones. La normalización de direcciones es clave para su eficiencia y calidad, aunque presenta desafíos significativos para las empresas de despacho.

El presente trabajo tiene como objetivo principal proponer y evaluar mejoras en el sistema de normalización de direcciones de una empresa de despachos en Chile, a través del uso del aprendizaje automático. Se desarrolló un modelo supervisado que integra técnicas de Procesamiento de Lenguaje Natural (NLP) con un algoritmo de clasificación de Máquinas de Soporte Vectorial (SVM), específicamente diseñado para optimizar la normalización de direcciones.

Al comparar el modelo de aprendizaje supervisado propuesto con el sistema de programación tradicional actualmente en uso, se emplearon dos conjuntos de datos: uno ficticio y otro real. La precisión de ambos modelos se evaluó con una métrica estándar para el modelo de aprendizaje supervisado y una métrica específica desarrollada para este estudio en el caso del modelo tradicional. Los resultados indicaron que el modelo supervisado alcanzó una precisión del 97 % con la base de datos ficticia, superando significativamente al 59 % del modelo tradicional. Sin embargo, al aplicarlo a una base de datos real, la precisión del modelo supervisado disminuyó a 63 %, superando solo ligeramente al 61 % del modelo tradicional. Esta disminución sustancial en la precisión sugiere la posibilidad de sobreajuste en el modelo supervisado, limitando su efectividad en entornos reales.

Actualmente, la persona encargada de la corrección manual de direcciones en la empresa logra abarcar, cómo máximo, el 30 % de las correcciones necesarias diariamente, lo que subraya la urgencia de mejorar este proceso. Al final del estudio, se revela que una reducción de tan solo el 2 % en el error de normalización de nuevas direcciones podría resultar en un ahorro de 129 millones para la empresa.

Para el futuro, se propone la exploración de tecnologías avanzadas, como la implementación de chatbots de IA a través de una API específica para la normalización de direcciones, y la adopción de técnicas de aprendizaje por refuerzo para perfeccionar los modelos de aprendizaje supervisado. Estas iniciativas apuntan a abordar las limitaciones actuales, optimizando la normalización de direcciones para satisfacer el objetivo de la empresa de liderar en calidad de servicio. Este estudio sienta las bases para futuras mejoras que prometen no solo aumentar la precisión en la normalización de direcciones sino también contribuir significativamente a la eficiencia, satisfacción del cliente y éxito económico de la empresa.

*“ Todo lo que amas probablemente se perderá, pero al final,
el amor volverá de otra manera.”*
F. Kafka

Agradecimientos

En primer lugar, agradezco a Dios por darme la oportunidad de tener una familia hermosa, a la cual honro y amo inmensamente.

A mi mamá, Lucía, por ser un ejemplo de amor, altruismo y generosidad. A mi hermana Lucita, por su ternura, comprensión y cariño inmensos. Gracias por su cariño y motivación para ser paciente y persistir durante toda la carrera

A Felipe, por ser un gran compañero.

A mi prima, Javiera, por ser como otra hermana y por estar, sobre todo, en los momentos difíciles.

Con profundo afecto, agradezco a aquellas amigas y amigos que han estado a mi lado, brindándome apoyo y aliento a lo largo de la carrera, sobre todo en los momentos más difíciles de mi vida. Su presencia y consuelo ha sido guía, recordándome la importancia de mantener el corazón firme y la mente clara.

Especialmente, dedico este esfuerzo y logro a mi padre, cuya memoria y enseñanzas continúan siendo una fuente de inspiración y fortaleza en mi vida.

Tabla de Contenido

1. Antecedentes generales	1
1.1. Caracterización de la industria	1
1.1.1. Rubro e industria del despacho en Chile	1
1.1.2. Servicios	2
1.1.3. Clientes	2
1.1.4. Dimensionamiento de actividad realizada por la empresa	3
1.2. Mercado y/o marco institucional	6
1.2.1. Actores	6
1.2.2. Tendencias del mercado	7
1.2.3. Adquisiciones, alianzas y ampliaciones	8
2. Descripción del Proyecto	9
2.1. Información general de la empresa de despacho	9
2.1.1. Proceso general de despacho	9
2.1.2. Productos y servicios	10
2.1.3. Clientes	12
2.1.4. Orden de servicio	12
2.1.5. Proceso general de una dirección asociada a una orden de servicio	13
2.2. Identificación del problema y oportunidad	15
2.2.1. Descripción del problema	15
2.2.2. Oportunidad de mejora	16
2.2.3. Causas	16
2.2.4. Consecuencias	16
2.3. Objetivos del proyecto	17
2.3.1. Objetivo general	17
2.3.2. Objetivos específicos	17
2.4. Alcances	17
2.4.1. Tipos de errores a abordar	17
2.4.2. Validación de direcciones	18
2.4.3. Avenidas y calles populares	18
2.4.4. Aplicabilidad del modelo	19
3. Marco Conceptual y Metodología	20
3.1. Estado del Arte	20
3.2. Normalización de direcciones	21
3.3. Ciencia de Datos	22
3.3.1. Aprendizaje Automático	22

3.3.1.1.	Aprendizaje supervisado	22
3.3.1.2.	Aprendizaje no supervisado	23
3.3.1.3.	Aprendizaje semi supervisado	23
3.3.1.4.	Aprendizaje por refuerzo	24
3.3.2.	Preparación de datos: conjuntos de entrenamiento y prueba	24
3.3.3.	Support Vector Machine	24
3.3.4.	Procesamiento de Lenguaje Natural	26
3.3.5.	Otros modelos de aprendizaje supervisado	26
3.3.6.	Métricas de evaluación en machine learning	29
3.4.	Interfaz de Programación de Aplicaciones	31
3.4.1.	API REST	32
3.5.	Metodología CRISP-DM	33
4.	Desarrollo Metodológico	34
4.1.	Comprensión del negocio	34
4.1.1.	Contexto del negocio	34
4.1.2.	Estandarización de direcciones	35
4.1.3.	Mercado y normalización de direcciones	35
4.1.4.	Fecha de promesa	36
4.1.5.	Nivel de servicio	36
4.1.6.	Impacto de las excepciones en el nivel de servicio	36
4.2.	Comprensión de los datos	38
4.2.1.	Exploración inicial de los datos	38
4.2.2.	Errores identificados en la declaración de dirección	40
4.2.3.	Problemas asociados a la normalización	41
4.3.	Preparación de los datos	41
4.3.1.	Construcción de la base de datos	42
4.4.	Modelamiento	43
4.4.1.	Justificación de la elección del algoritmo	44
4.4.2.	Modelo propuesto	46
4.4.2.1.	Descripción general del modelo propuesto	46
4.4.2.2.	Bases de datos	46
4.4.2.3.	Código	47
4.5.	Evaluación	50
4.5.1.	Modelo de normalización actualmente utilizado por la empresa	50
4.5.1.1.	Procesos relevantes modelo actual de normalización	50
4.5.1.2.	Descripción del código de programación tradicional	51
4.5.1.3.	Métricas	51
4.5.2.	Modelo de normalización propuesto	53
4.5.3.	Evaluación entre modelos	54
4.6.	Despliegue	56
4.6.1.	Recomendaciones para la mejora e implementación del modelo propuesto	56
4.6.1.1.	Carta Gantt	57
4.6.2.	Recomendaciones exploración chatbot IA	59
4.6.2.1.	Experimento realizado	60
5.	Resultados	64

5.1.	Modelo de programación tradicional	64
5.2.	Modelo de machine learning propuesto	65
5.2.1.	Entrenamiento con base de datos ficticia	65
5.2.2.	Entrenamiento con base de datos real	66
5.3.	Comparación de resultados	67
5.4.	Limitaciones y áreas de mejora	68
6.	Conclusiones	70
6.1.	Costo promedio de un error	71
6.2.	Trabajos futuros	72
	Bibliografía	74
	Anexos	77
A.	Proceso de normalización actual	77
B.	Calles de la comuna de Cerrillos	78

Índice de Tablas

2.1.	Ejemplos de direcciones incorrectamente normalizadas y sus consecuencias en georreferenciación. Fuente: Elaboración propia.	15
2.2.	Comparación entre nombres populares y oficiales de calles y vías. Fuente: Elaboración propia.	19
3.1.	Resumen y comparación de algoritmos de clasificación de aprendizaje supervisado comúnmente empleados. Fuente: Elaboración propia.	28
4.1.	Órdenes de servicio con pinchazos de excepción de mal ruteo durante las semanas 1-19 (2023). Fuente: Elaboración propia.	37
4.2.	Proyección de fallo diario debido a excepción de mal ruteo en órdenes de servicio, considerando una demanda diaria estimada de 40.000 unidades. Fuente: Elaboración propia.	37
4.3.	Distribución de Incidencias MRH por Comunas de Santiago entre el 19/03/2023 y el 19/04/2023. Fuente: Elaboración propia.	39
4.4.	Ejemplos de direcciones declaradas de forma errónea. Fuente: Elaboración propia.	40
4.5.	Ejemplos de errores en la normalización de direcciones. Fuente: Elaboración propia.	41
4.6.	Extracto de la base de datos para entrenamiento del modelo de aprendizaje supervisado. Fuente: Elaboración propia.	43
4.7.	Ejemplos de resultados de normalización de direcciones según el modelo de programación tradicional actual. Fuente: Elaboración propia.	51
4.8.	Evaluación del modelo de programación tradicional utilizando una métrica propia. Fuente: Elaboración propia.	52
4.9.	Precisión obtenida para el modelo de programación tradicional, utilizando una base de datos ficticia y una base de datos real. Fuente: Elaboración propia. . .	53
4.10.	Métricas de rendimiento obtenidas para el modelo de normalización propuesto, con base de datos ficticia y real. Fuente: Elaboración propia.	53
4.11.	Análisis comparativo de normalización tradicional y normalización mediante chatbot de IA. Fuente: Elaboración propia.	61
5.1.	Métricas del modelo propuesto utilizando base de datos ficticia. Fuente: Elaboración propia.	65
5.2.	Métricas del modelo propuesto utilizando base de datos real. Fuente: Elaboración propia.	66
5.3.	Ejemplos de normalización de calles por modelos de programación tradicional y propuesto. Fuente: Elaboración propia.	67
5.4.	Comparación de precisión entre modelos de programación tradicional y propuesto según la base de datos aplicada. Fuente: Elaboración propia.	67
6.1.	Costo promedio de envíos de productos pequeños y medianos. Fuente: Diario La Tercera.	72

Índice de Ilustraciones

1.1.	El pronóstico de la demanda de servicios de paquetería en Chile para el 2028 muestra un aumento casi al triple de la demanda actual. Fuente: McKinsey.	1
1.2.	Las ventas B2C a través del e-commerce en Chile han tenido un aumento exponencial entre el 2010 y el 2023. Fuente: Statista.	3
1.3.	Órdenes con salida a ruta según zonales de Chile (norte, centro, sur y Santiago) entre enero y junio del 2023. Fuente: Elaboración propia.	4
1.4.	Órdenes efectivamente entregadas entre enero y junio de 2023. Fuente: Elaboración Propia.	5
1.5.	Efectividad del servicio. Fuente: Elaboración Propia.	5
1.6.	Paquetes entregados por empresas pequeñas y medianas de última milla (millones). Fuente: KawesLab.	7
2.1.	Servicio de Fullfilment. Fuente: Elaboración propia.	11
2.2.	Servicio de envíos Same Day. Fuente: Elaboración propia.	11
2.3.	Servicio Puntos Pick-Up. Fuente: Elaboración propia.	12
2.4.	Servicio Logística Reversa. Fuente: Elaboración propia.	12
2.5.	Esquema de pinchazos que puede tener una orden de servicio. Fuente: Elaboración propia.	13
2.6.	Proceso general de una nueva dirección que se registra en el diccionario de la empresa. Fuente: Elaboración propia.	14
3.1.	Metodología CRISP-DM. Fuente: Health Data Miner.	33
4.1.	Fragmento del código para la normalización de direcciones. Fuente: Elaboración propia.	47
4.2.	Fragmento del código para la normalización de direcciones. Fuente: Elaboración propia.	48
4.3.	Fragmento del código en Python para la normalización de direcciones. Fuente: Elaboración propia.	48
4.4.	Fragmento del código en Python para la normalización de direcciones (métricas de evaluación). Fuente: Elaboración propia.	49
4.5.	Visualización de resultados del modelo en Python. Fuente: Elaboración propia.	49
4.6.	Propuesta de carta Gantt de doce meses de duración para el proyecto. Fuente: Elaboración propia.	59
4.7.	Captura de pantalla mostrando un mensaje enviado al chatbot Bing, en el que se solicita la normalización de direcciones con indicaciones específicas para incluir calle, numeración y comuna, entre otras. Fuente: Elaboración propia.	62

- 4.8. Ejemplos de interacción con el chatbot Bing: a la izquierda, las direcciones enviadas como input; a la derecha, las respuestas del chatbot con las direcciones normalizadas. La imagen ilustra la capacidad del chatbot para procesar y normalizar las direcciones según las especificaciones dadas. Fuente: Elaboración propia.

.....

Capítulo 1

Antecedentes generales

1.1. Caracterización de la industria

1.1.1. Rubro e industria del despacho en Chile

Las empresas de distribución han adquirido un papel fundamental en la economía global al facilitar el transporte y la entrega de diversos bienes a nivel local, nacional e internacional. Estas empresas despliegan una amplia gama de servicios diseñados para satisfacer las crecientes demandas de sus clientes, tales como entregas rápidas y servicios de seguimiento de envíos. Su relevancia se ha acrecentado en años recientes, especialmente con el aumento significativo de las compras en línea.

A medida que el comercio electrónico continúa expandiéndose, se proyecta que la demanda de servicios de paquetería en Chile alcance casi el triple de la demanda actual para el año 2028, subrayando la importancia estratégica y el potencial de crecimiento en este sector.

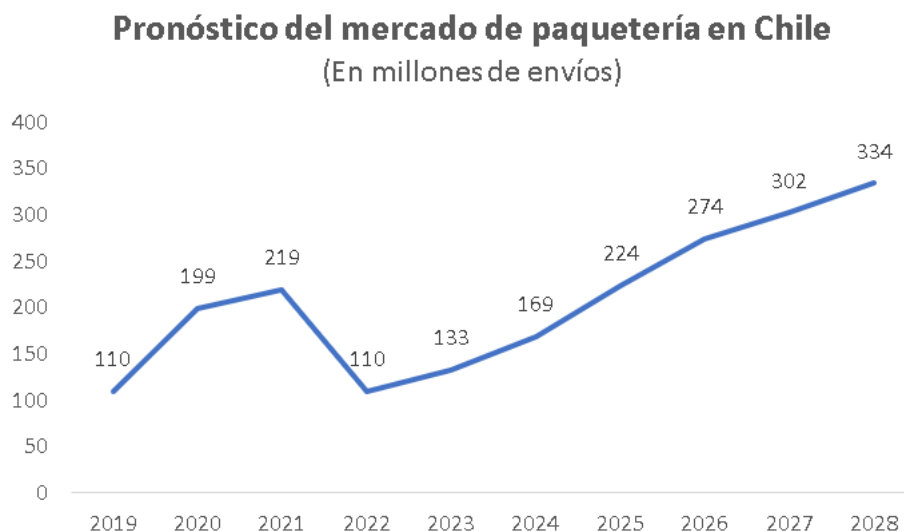


Figura 1.1: El pronóstico de la demanda de servicios de paquetería en Chile para el 2028 muestra un aumento casi al triple de la demanda actual. Fuente: McKinsey.

La diversidad geográfica y las considerables distancias presentes en el territorio chileno

han generado una imperiosa demanda de servicios de despacho eficientes y confiables. La industria de despachos ha desempeñado un papel trascendental en la facilitación de la conectividad y accesibilidad de bienes y servicios para los ciudadanos chilenos.

En la era digital, la industria de despachos en Chile ha adoptado nuevas tecnologías para mejorar sus operaciones. La implementación de sistemas de gestión de inventarios, rastreo en tiempo real y rutas optimizadas ha permitido una mayor eficiencia en la entrega de productos. Al igual que en muchas otras industrias, las empresas del rubro del despacho están explorando tecnologías de inteligencia artificial para mejorar la eficiencia operativa y la calidad del servicio.

En los últimos años, ha habido un notorio aumento en la expansión de la industria de envíos en Chile, particularmente en lo que respecta al transporte de paquetes. En el país operan diversas empresas destacadas que brindan servicios de entrega, como Blue Express, Chilexpress, Correos de Chile y Starken. Estas compañías proporcionan una diversidad de opciones para envíos y servicios de mensajería, siendo reconocidas por su atención al cliente y el nivel de servicio que entregan. Al momento de seleccionar una empresa de envíos, los consumidores valoran principalmente estos aspectos. Sin embargo, existe una distinción significativa entre los clientes finales, que adquieren productos en el sector retail, y los clientes corporativos, que son las propias empresas de retail, y que eligen servicios de envío para distribuir sus órdenes de compra.

1.1.2. Servicios

Los servicios de las empresas de despacho en Chile se pueden segmentar por tipo de envío. Algunos tipos de envío comunes son:

- **Envíos nacionales:** Son aquellos envíos que se realizan dentro del territorio nacional. Las empresas de despacho ofrecen diferentes opciones de tiempo de entrega, como envío estándar o envío urgente.
- **Envíos internacionales:** Son aquellos envíos que se realizan fuera del territorio nacional. Las empresas de despacho ofrecen diferentes opciones de tiempo de entrega y costos, dependiendo del destino y del tipo de producto que se envía.
- **Envíos urgentes:** Son aquellos envíos que requieren una entrega rápida. Las empresas de despacho ofrecen opciones de entrega en el mismo día o al día siguiente para este tipo de envíos.
- **Envíos estándar:** Son aquellos envíos que no requieren una entrega urgente. Las empresas de despacho ofrecen opciones de entrega con plazos más largos y a menudo a un costo más bajo para este tipo de envíos.

1.1.3. Clientes

Para comprender mejor el mercado de envíos, es esencial distinguir entre los dos principales tipos de clientes que interactúan con las empresas de despacho. Por un lado, están los clientes finales, que compran productos a través de plataformas de retail y reciben sus paquetes mediante empresas que realizan servicios de despacho. Por otro lado, están los clientes corporativos (e-commerce), que en este contexto se refieren a las empresas de retail que

contratan los servicios de las empresas de despacho para la distribución de sus productos a los consumidores finales. Hay otro tipo de clientes que aplican a esta descripción, como las PYMEs y los emprendedores. Así, las empresas del rubro del despacho, son intermediarios cruciales en el proceso de entrega, asegurando que los productos lleguen de forma segura y puntual a su destino final.

El sustancial aumento en las ventas derivadas del comercio electrónico en Chile durante el período comprendido entre 2010 y 2023 (Figura 1.2), respalda la validez del pronóstico representado en la Figura 1.1.

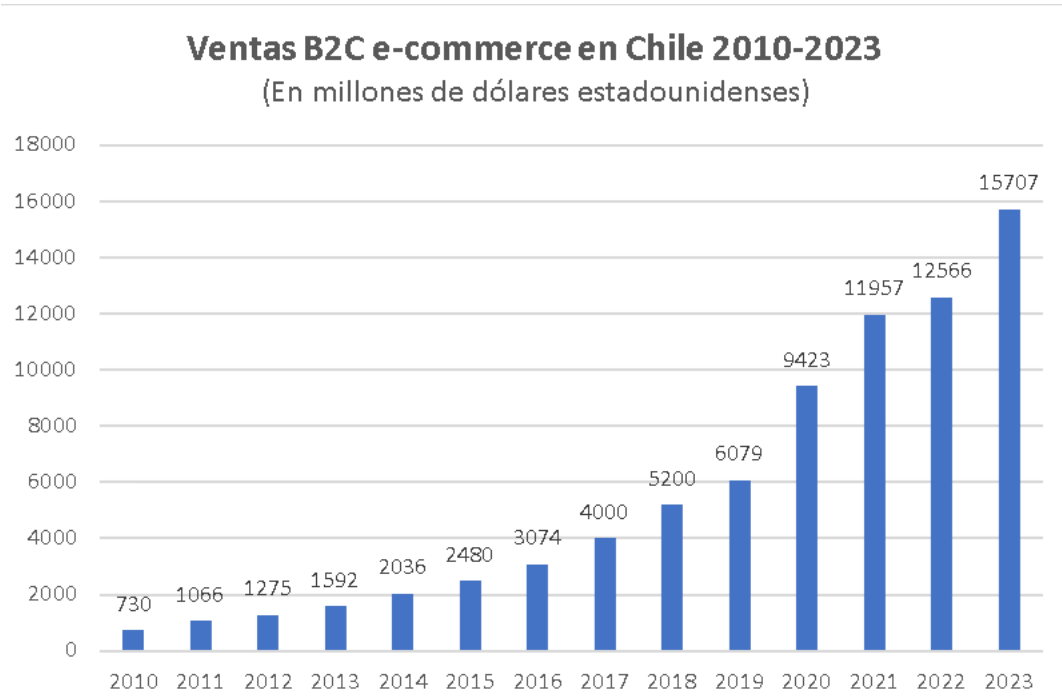


Figura 1.2: Las ventas B2C a través del e-commerce en Chile han tenido un aumento exponencial entre el 2010 y el 2023. Fuente: Statista.

El e-commerce cada vez cuenta con más presencia en Chile, experimentando un crecimiento exponencial de ventas en los últimos años. Además, la pandemia tuvo un impacto significativo en el aumento de las compras por internet. Con el distanciamiento social, el confinamiento y otras medidas tomadas en respuesta a la pandemia, los consumidores han aumentado las compras en línea y el uso de las redes sociales[1].

Según un estudio de la Conferencia de las Naciones Unidas sobre Comercio y Desarrollo (UNCTAD), más de la mitad de los encuestados compran en línea con mayor frecuencia tras la pandemia[2]. Esto ha llevado a un aumento en la demanda de servicios de empresas de despacho para entregar los productos comprados en línea a los clientes.

1.1.4. Dimensionamiento de actividad realizada por la empresa

La empresa de despacho (objeto de análisis) representa una empresa especializada en almacenamiento, distribución y logística, que cuenta con una amplia cobertura nacional,

extendiendo su presencia a lo largo de todo Chile.

En cuanto a sus operaciones, la empresa exhibe un tiempo promedio de entrega de 1.1 días en el territorio chileno. En el caso del servicio que garantiza la entrega de pedidos al día siguiente, se logra una eficacia del 90 %, materializándose la entrega de dicho porcentaje de paquetes. La entidad cuenta con una red robusta compuesta por aproximadamente 1.300 socios couriers y dispone de más de 1.400 puntos estratégicos destinados tanto a la recepción como a la devolución de pedidos, consolidando así su infraestructura operativa.

El gráfico de la Figura 1.3 ilustra las órdenes de servicio que salieron a ruta en Santiago, así como en las regiones del sur, centro y norte del país, durante el período que abarca desde enero hasta junio del año 2023.

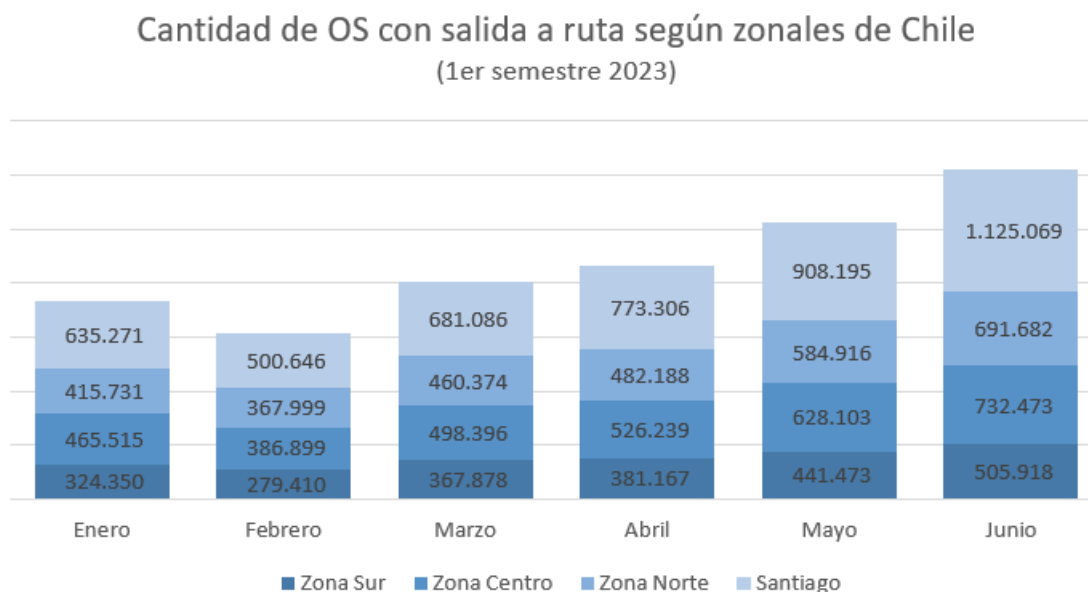


Figura 1.3: Órdenes con salida a ruta según zonales de Chile (norte, centro, sur y Santiago) entre enero y junio del 2023. Fuente: Elaboración propia.

Se evidencia un crecimiento constante en el número de órdenes de servicio con salida a despacho, con la única excepción del mes de febrero.

En la Figura 1.4 se observa el aumento de la cantidad de órdenes de servicio que fueron efectivamente despachadas entre los meses de enero y junio de 2023 según zonales.

Cantidad de OS despachadas según zonales de Chile (1er semestre 2023)

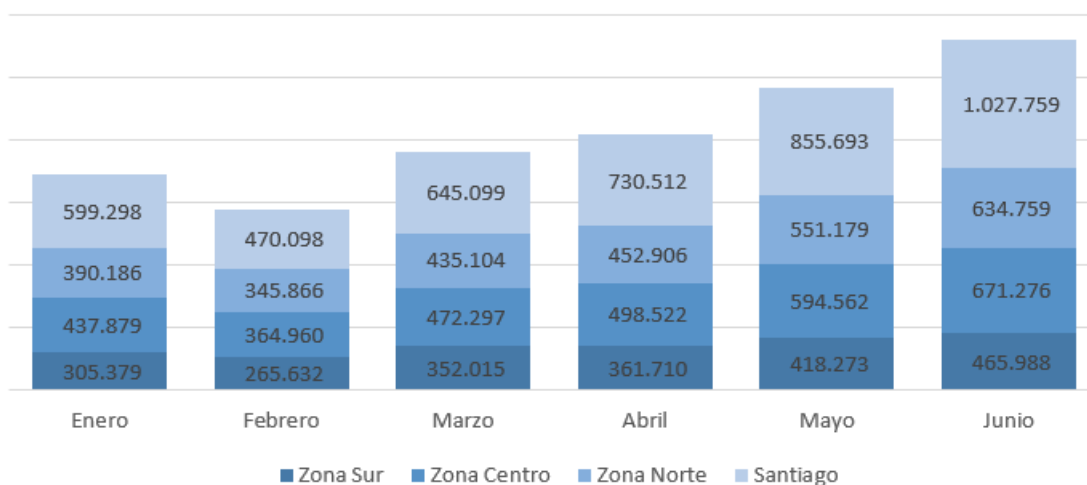


Figura 1.4: Órdenes efectivamente entregadas entre enero y junio de 2023.
Fuente: Elaboración Propia.

Considerando los datos de las Figuras 1.4 y 1.3, se construye el gráfico de la Figura 1.5, que muestra el total de órdenes de servicio que salieron a ruta en Chile, durante el primer semestre del 2023, y el porcentaje de efectividad de entrega de las mismas.

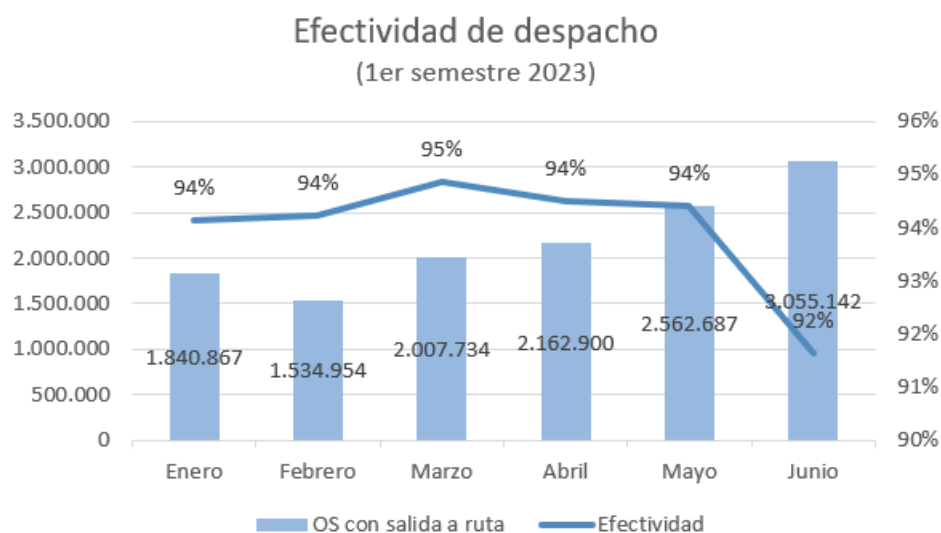


Figura 1.5: Efectividad del servicio. Fuente: Elaboración Propia.

La disminución registrada en febrero, tanto en las órdenes de servicio programadas para salir a ruta como en las órdenes de servicio efectivamente despachadas, es atribuida a la reducción de la demanda durante el periodo mencionado. La Cámara de Comercio de Santiago (CCS) informó que el comercio electrónico en Chile experimentó una caída nominal del 18 % durante febrero del 2023, muy similar a lo registrado en enero del mismo año. Las principales causas de esta tendencia se relacionan con el repliegue de las compras online tras la reapertura del comercio físico post cuarentenas, las bases de comparación históricamente altas de los años

previos, la disminución de los ingresos de los hogares, la alta inflación y tasas de interés de política monetaria. A ello se le suma el hecho de que los consumidores habían vuelto a gastar una parte de sus recursos en servicios presenciales, como recreación, gastronomía y turismo, reduciendo las holguras disponibles para la compra de bienes físicos[3].

Desafíos para 2028

En la actualidad, la empresa cuenta con una participación de mercado que ronda el 15 %. Su valoración en el año 2022 alcanzó los 250 millones de dólares. Durante dicho año, la entidad gestionó la entrega de aproximadamente 12 millones de paquetes.

Considerando los retos a enfrentar, la empresa ha establecido metas ambiciosas para el año 2028. Entre ellas, destacan la aspiración de posicionarse como la principal entidad en servicio dentro de su sector y alcanzar una cuota de mercado del 30 %, desafíos que delimitan su visión estratégica a largo plazo.

1.2. Mercado y/o marco institucional

1.2.1. Actores

En el dinámico escenario de la industria del despacho en Chile, la competencia se ha intensificado, especialmente durante la pandemia, alimentada por el crecimiento exponencial del comercio electrónico. En 2021, Blue Express ha consolidado su posición como uno de los principales actores, alcanzando una significativa participación del 15 %, que junto a Correos de Chile, Starken y Chilexpress, constituye el grupo de los cuatro mayores actores. Este éxito coexiste en un entorno donde el número de nuevos participantes se ha duplicado, con más de 58 empresas emergentes dedicadas a la entrega a domicilio[4].

La competencia ha adquirido una diversidad creciente, evidenciada por el auge de las empresas emergentes que ahora contribuyen con el 20 % del tráfico de paquetería en el comercio electrónico. Más allá de las cuatro grandes, surge un grupo intermedio que incluye nombres como WSP, Shipit, DHL, FedEx y Urbano, compitiendo activamente por capturar una porción sustancial del mercado. Este escenario competitivo plantea interrogantes sobre la sostenibilidad a largo plazo, con lo que se podría dar lugar a ajustes, adquisiciones y alianzas en la industria[4].

A pesar de la presencia destacada de las cuatro principales empresas, se observa que las compañías emergentes, con especializaciones diversas, disputan un 20 % del mercado, desafiando la supremacía histórica de los actores más consolidados. La dinámica competitiva en la que tanto los gigantes establecidos como los nuevos participantes se enfrentan a un entorno empresarial en constante evolución refleja un paisaje fluido y desafiante para la industria del despacho en Chile.

En el siguiente gráfico, construido por KawesLab, se observan las cifras en millones de paquetes entregados, dejando de lado a las cuatro grandes empresas de la industria (Correos de Chile, Chilexpress, Blue Express y Starken). En 2022, la consultora registró 41,1 millones de paquetes entregados por las empresas pequeñas y medianas pertenecientes al rubro.

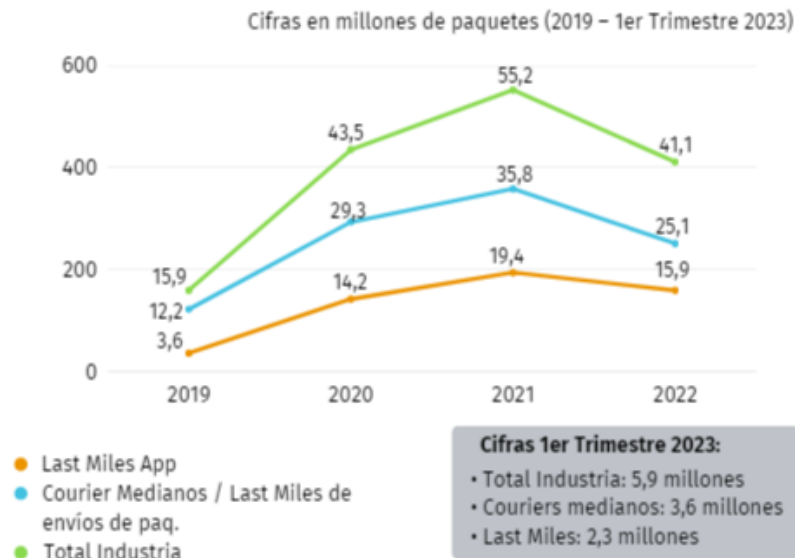


Figura 1.6: Paquetes entregados por empresas pequeñas y medianas de última milla (millones). Fuente: KawesLab.

1.2.2. Tendencias del mercado

El mercado de despacho en Chile experimenta transformaciones significativas, influenciadas por diversos factores, entre ellos el aumento de la inflación y la necesidad de ajustar los montos de renta acorde al valor de mercado. Según un estudio de Kawesqar Lab, en el 2020, el comercio electrónico en Chile experimentó un crecimiento excepcional del 50 %, un fenómeno que dio lugar al surgimiento de numerosas empresas de envíos y despacho para atender la creciente demanda logística. No obstante, las estimaciones de la Cámara de Comercio de Santiago (CCS) indican un ajuste en las proyecciones del e-commerce, reduciendo la expectativa del 20 % de las ventas totales para el 2023 a un 14 %. Este cambio en las previsiones afecta directamente a las compañías más pequeñas, ya que la depresión en el consumo complica el panorama de supervivencia para estas empresas emergentes.

El nuevo mapa logístico está dominado por las cuatro grandes empresas: Chilexpress, Correos de Chile, Blue Express y Starken. Este oligopolio se enfrenta a la creciente competencia de empresas emergentes especializadas que han revolucionado el mercado. La proliferación de estas compañías responde a la alta demanda del e-commerce, que ha aumentado significativamente la cantidad de compras y paquetes enviados a los hogares[4].

La adaptación de las empresas de despacho a las tendencias del mercado incluye estrategias como mejoras en la oferta online, optimización de períodos de entrega, despachos gratuitos y alternativas de envío. Además, la industria está atenta a las demandas de los consumidores, como la entrega en tiempos cada vez más cortos, lo que representa un desafío para el sector comercial y las empresas de transporte[5].

Las empresas líderes, como Starken, Blue Express y FedEx Express, anticipan un crecimiento continuo del negocio y apuestan por la tecnología para mejorar los tiempos de reparto. Este impulso tecnológico se refleja en el aumento del 30 % en el negocio de última milla en

2021, según cifras de la Cámara de Comercio de Santiago. En medio de este panorama, las empresas del sector están fortaleciendo sus áreas digitales y logísticas para mantener su relevancia y ganar terreno en un mercado en constante evolución. La innovación, adquisiciones estratégicas y alianzas se revelan como herramientas clave para enfrentar los desafíos y aprovechar las oportunidades emergentes en la industria del despacho en Chile[6].

1.2.3. Adquisiciones, alianzas y ampliaciones

En los últimos tiempos, las empresas de última milla han explorado distintos enfoques para fortalecer su presencia en el mercado, optando por adquisiciones estratégicas y alianzas con otras compañías, con el objetivo de optimizar su alcance territorial y reducir los tiempos a lo largo de toda la cadena logística.[5]

Blue Express, tras ser adquirida por Empresas Copec, experimentó una expansión significativa de su red, compuesta por 1.300 puntos, donde se gestionan paquetes para entregas y devoluciones de consumidores. Esta adquisición le permitió incorporar 681 estaciones de servicio y 422 tiendas de conveniencia, consolidando así su presencia en el mercado[5]

En el caso de Starken, la empresa inauguró su primera sucursal completamente automatizada, operativa las 24 horas del día y los siete días de la semana, destinada al envío y recepción de paquetería. A través de un acuerdo estratégico con Mall Plaza, estableció puntos de autoatención en centros comerciales de alto tráfico, superando los 260 puntos de venta a nivel nacional. La compañía tiene planes ambiciosos de inversión en infraestructura, incluyendo la apertura de nuevos centros regionales, sucursales de atención continua y un centro de clasificación automatizado para paquetería de menor tamaño[5].

FedEx Express, por su parte, respondió al crecimiento de la demanda en el sector de paquetes pequeños para clientes residenciales y pequeñas y medianas empresas. En 2022, ampliaron la superficie de su centro de distribución en Pudahuel de 7.600 a 14 mil metros cuadrados. La expansión incluye 53 puntos de envío, 880 vehículos, 29 puntos de distribución y emplea a más de 2.000 colaboradores, consolidando su posición en el mercado logístico[5].

Capítulo 2

Descripción del Proyecto

2.1. Información general de la empresa de despacho

2.1.1. Proceso general de despacho

El proceso general de despacho que siguen las empresas del rubro de logística y envío implica varias etapas. Este proceso se caracteriza por tener un enfoque logístico, y por su capacidad para gestionar grandes volúmenes de pedidos de forma eficiente y segura. De forma general, se identifican las siguientes etapas:

1. **Compra a través de un e-commerce**

El cliente final efectúa la compra, ya sea en un e-commerce integrado con la empresa de despacho, una empresa no integrada o un emprendimiento que lleva los pedidos a un punto de entrega o recogida de la empresa.

2. **Recepción de solicitud**

La empresa recibe solicitudes de envío de sus clientes. Estos clientes pueden ser personas naturales o empresas. La solicitud contiene información detallada sobre el paquete, como la dirección de origen y de destino, y otras instrucciones especiales.

3. **Generación de la orden**

Dependiendo del origen de la compra,

- E-commerce integrado: La plataforma genera automáticamente una orden de servicio en el sistema de la empresa.
- Empresa no integrada: La empresa genera manualmente una solicitud de despacho en el sistema de la empresa de despacho.
- Entrega en un punto: La entrega en el punto de recibimiento de la empresa se considera como la generación de la orden.

4. **Asignación de recursos**

Se asignan los recursos necesarios para gestionar los envíos, como vehículos de transporte, personal de almacén, entre otros. Esta asignación se realiza considerando la urgencia, la ruta de entrega y otros factores logísticos.

5. **Empaque y etiquetado**

Los paquetes se preparan para ser enviados, asegurándose de que estén correctamente

empaquetados y que estén etiquetados con información relevante, con los detalles de origen y destino. En algunos casos, se incluyen códigos de barra o números de orden, para facilitar el seguimiento.

6. Carga en vehículos

Los paquetes se cargan en los vehículos de transporte designados. Existen empresas que utilizan softwares que optimizan las rutas de entrega y minimizan los tiempos de transporte, pero también hay algunas que depositan esta responsabilidad en el conocimiento de los couriers o transportistas.

7. Transporte y monitoreo

Los vehículos transportan los paquetes según las rutas asignadas. Algunas empresas ofrecen servicio de monitoreo del estado del envío en tiempo real.

8. Llegada a centros de distribución

En algunos casos, los paquetes pasan por centros de distribución regionales, para una gestión más eficiente antes de continuar su ruta hacia el destino final.

9. Entrega a destinatario

Los transportistas entregan el paquete al destinatario, siguiendo las instrucciones de la solicitud de envío. Esto incluye firmas por parte del destinatario para confirmar la entrega del paquete, y otros métodos de verificación.

10. Confirmación y seguimiento

Las empresas solicitan confirmación de la entrega por parte del destinatario o cliente final.

11. Seguimiento post-entrega y atención al cliente

Después de la entrega del pedido, la empresa puede proporcionar herramientas para el seguimiento post-entrega y gestionar cualquier inconveniente o consulta del cliente asociada a la entrega.

2.1.2. Productos y servicios

En la empresa de despacho se ofrece una amplia gama de productos y servicios para satisfacer las necesidades de los clientes. De manera general, algunos de estos son:

- Envíos nacionales e internacionales: servicios de envío de paquetes y documentos a destinos nacionales e internacionales.
- Servicios de logística: soluciones de logística para empresas, incluyendo almacenamiento, distribución y transporte de mercancías.
- Soluciones para e-commerce: integraciones con plataformas de e-commerce para facilitar la gestión de envíos y devoluciones.
- Servicios de valor agregado: servicios adicionales como embalaje, seguro de envío y seguimiento en línea.

Diferenciándose en el nivel de responsabilidades que adquiere la empresa estudiada para con el cliente, se pueden identificar más específicamente los siguientes servicios ofrecidos:

Fullfilment

El servicio de Fulfillment permite a los comercios externalizar sus servicios, asegurando eficiencia en todos los procesos: se encarga de recibir los productos del cliente (o ir a buscarlos), almacenarlos y sincronizarlos con su e-commerce, preparar las órdenes y despachar al cliente final, además de mantener actualizado el inventario. Los sistemas se mantienen conectados y actualizados 24/7, para llevar todos los pedidos al día.



Figura 2.1: Servicio de Fullfilment. Fuente: Elaboración propia.

Envíos: Same Day - Next Day - Express

La empresa ofrece distintos tipos de envíos a sus clientes, en función de cuándo requiere el cliente final la entrega de su pedido. El servicio Same Day permite despachar los productos el mismo día en que se realiza la compra, con cobertura dentro de la Región Metropolitana. Cuenta con un 98% de efectividad en las entregas para el mismo día. Con el servicio Next Day, el cliente final recibe su pedido el día siguiente de haber realizado la compra. El envío Express es el estándar, con el cual el cliente final recibe su pedido entre 3 y 7 días después de haber realizado la compra.



Figura 2.2: Servicio de envíos Same Day. Fuente: Elaboración propia.

Integraciones e-commerce

Se ofrece la integración directa al sistema de un comercio, de manera personalizada, o a través de plugins/APIs, para plataformas digitales que agrupan vendedores y compradores. Se asegura un proceso de venta más fácil y rápido.

Puntos de recogida y envío

Los almacenes y locales comerciales en alianza con la empresa dan a sus clientes la comodidad de enviar y retirar encomiendas cerca de sus hogares. En estos puntos se recibe una encomienda, que es retirada por un courier, luego se consolida la carga en el centro de distribución y se prepara para la distribución al cliente final.



Figura 2.3: Servicio Puntos Pick-Up. Fuente: Elaboración propia.

Logística reversa

Se ofrece el servicio de devolución y asistencia dedicada para mejorar la experiencia post-venta del cliente final. Se notifican las distintas etapas para que el cliente pueda realizar un seguimiento: preparación, preparación del retiro, retiro, producto enviado y entregado.



Figura 2.4: Servicio Logística Reversa. Fuente: Elaboración propia.

2.1.3. Clientes

En cuanto a sus clientes[7], la empresa los divide en:

- PYMEs y emprendedores: se ofrecen soluciones para emprendedores y PYMEs que desean hacer envíos con la empresa.
- Clientes de e-commerce: se ofrecen integraciones con plataformas de e-commerce para facilitar la gestión de envíos.
- Clientes particulares: se ofrecen servicios de envío y recepción de paquetes para clientes particulares.

2.1.4. Orden de servicio

Una orden de servicio (OS) es un número único que se le asocia a un pedido o paquete para identificarlo. Una OS recibe, a lo largo de su ciclo en la empresa, lo que se conoce como “pinchazos”, que corresponden a siglas que se le asocian para saber en qué parte del proceso de entrega se encuentra. En el esquema a continuación se muestran los pinchazos que recibe una OS a lo largo de su ciclo en la empresa.

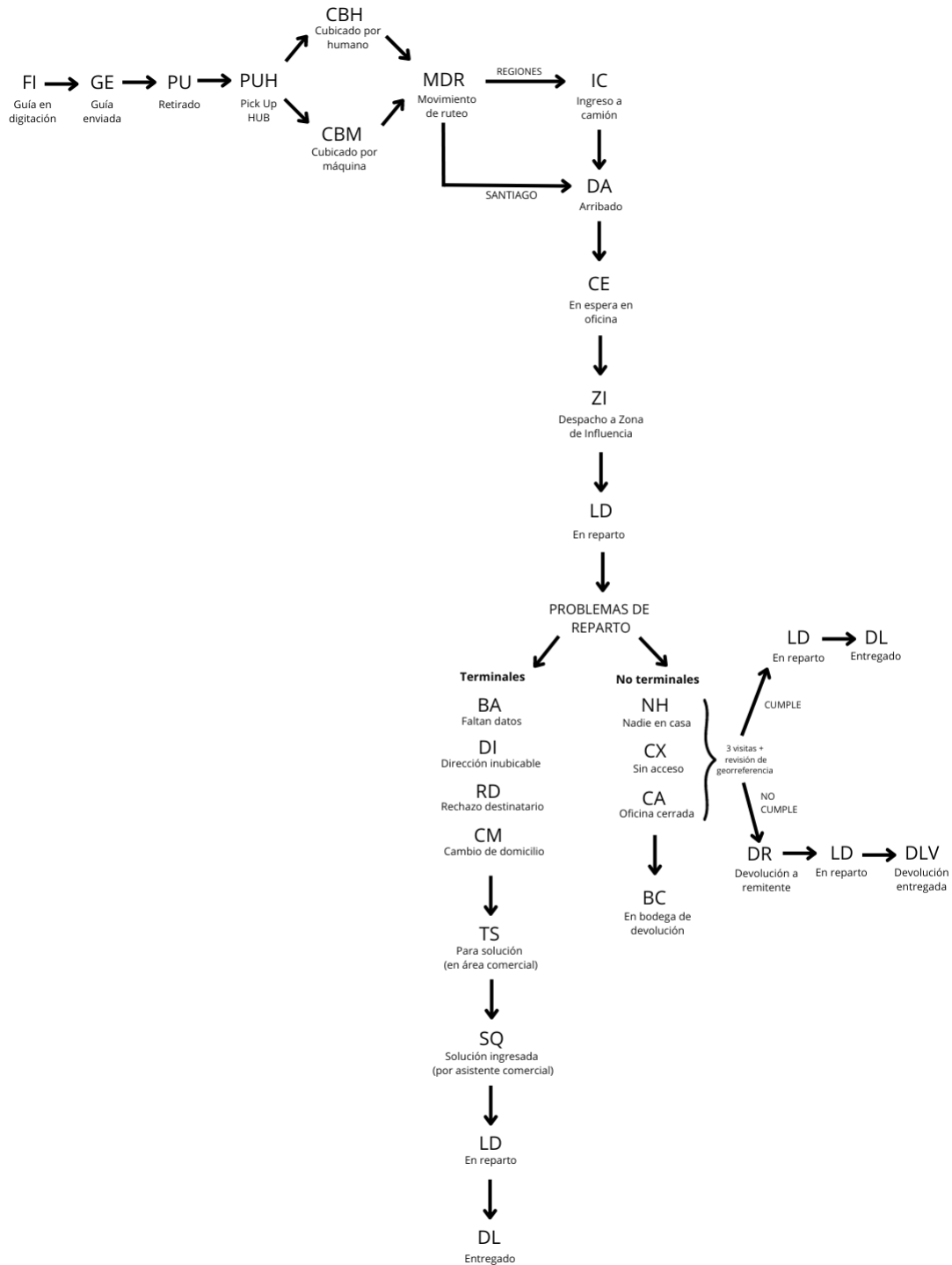


Figura 2.5: Esquema de pinchazos que puede tener una orden de servicio.
Fuente: Elaboración propia.

2.1.5. Proceso general de una dirección asociada a una orden de servicio

En la sección 2.1.1 se hizo referencia al proceso general que sigue un pedido dentro de una empresa de despachos. Para comprender el contexto del problema, es importante conocer las

etapas que atraviesa la dirección asociada a una orden de servicio cuando esta ingresa al sistema.

1. Emisión de la orden de servicio

La OS tiene asociada una dirección, que es ingresada por el destinatario del pedido. Si una dirección es declarada de forma incorrecta, implicará un problema para las siguientes etapas. Este problema se podría generar porque la dirección está incompleta, por errores de ortografía, entre otros.

2. Normalización

La dirección contenida en esa orden de servicio se "limpia", para que posteriormente pueda ser localizada. Dado el proceso actual de normalización, puede ocurrir que una dirección nueva se normalice de forma incorrecta. Esta es la etapa de mayor interés para este trabajo.

3. Georreferenciación

Una vez que la dirección es normalizada, se le asocian coordenadas (latitud y longitud) para ubicarla en el mapa. Esto se realiza mediante dos proveedores de georreferenciación: Google Maps y Here We Go App. En algunos casos, estos proveedores no localizan correctamente una dirección, ya sea porque no se encuentra en los registros, o porque se debe a algún error en las etapas anteriores.

4. Asignación de ruta

Dependiendo de la zona a la que pertenece el punto encontrado (coordenadas de latitud y longitud), se le asociará una ruta para que el pedido pueda ser despachado en la dirección.

5. Entrega exitosa

Si la ruta se asignó de manera correcta, es posible que el pedido se pueda entregar de forma exitosa en la nueva dirección que ingresó al sistema.



Figura 2.6: Proceso general de una nueva dirección que se registra en el diccionario de la empresa. Fuente: Elaboración propia.

2.2. Identificación del problema y oportunidad

2.2.1. Descripción del problema

En la organización, se ha detectado una problemática relacionada con el proceso de normalización de direcciones en su base de datos. Actualmente, se emplea un enfoque de programación convencional para normalizar las direcciones proporcionadas por los destinatarios. Este método es en algunos casos desventajoso, ya que elimina información relevante contenida en la dirección, lo que conlleva a complicaciones en su posterior georreferenciación. Por ejemplo, en una de las líneas del código actual se eliminan las palabras que contienen el prefijo “COND” (haciendo alusión a la abreviación de “condominio”, y el prefijo “AV” (abreviación de “avenida”) con el propósito de limpiar las direcciones. Es decir, si se tiene la dirección “AV LAS CONDES”, la salida de esa dirección sería “LAS”. En la Tabla 2.1 se muestran ejemplos de direcciones cuya mala calidad es producto del problema que se ha descrito.

Tabla 2.1: Ejemplos de direcciones incorrectamente normalizadas y sus consecuencias en georreferenciación. Fuente: Elaboración propia.

Ingreso ¹	Normalización ²	Georreferenciación ³	Distancia [m] ⁴
AV LAS CONDES 10250, LAS CONDES	LAS 10250, LAS	Las lavandulas, Las Condes, Región Me- tropolitana	3.790
AV EJERCITO LI- BERTADOR 688 DEPTO 704 ESQUI- NA DOMEYKO, SANTIAGO	EJERCITO LIBER TADOR 688 704, SANTIAGO, CHILE	Ejército Libertador 688, Puente Alto	18.020
Camino a Viluco 37,BUIN,REGION METROPOLITANA DE SANTIAGO, BUIN	VILUCO, BUIN, RE- GIÓN METROPOLI- TANA, CHILE	Camino La Glorieta, Quillota, Valparaíso	102.960

¹ La columna *Ingreso* contiene la dirección y comuna declaradas por el destinatario.

² La columna *Normalización* contiene la normalización que se realiza en el sistema de la empresa de la dirección declarada por el destinatario.

³ La columna *Georreferenciación* contiene la dirección correspondiente a las coordenadas en que el sistema encontró la dirección normalizada.

⁴ La columna *Distancia* en la tabla muestra la distancia en línea recta, según Google Maps, entre la dirección proporcionada por el destinatario y la dirección encontrada por la empresa.

La concepción de un modelo de machine learning mediante la aplicación de técnicas de aprendizaje supervisado con el propósito de normalizar direcciones podría conllevar una mejora sustancial en la precisión y eficacia del proceso de normalización de direcciones. A diferencia del enfoque de programación convencional, un modelo de machine learning tiene la capacidad de aprender a partir de los datos y ajustarse a contextos variables, lo que le habilita para la identificación y retención de información relevante para la correcta geolocalización de las direcciones.

Este avance promovería una mayor precisión en la georreferenciación y simplificaría el proceso de entrega, lo que se traduciría en un incremento de la satisfacción del cliente y una mejora en la eficiencia operativa de la empresa.

2.2.2. Oportunidad de mejora

La oportunidad que se identifica en el área de trabajo corresponde a mejorar el método de normalización de direcciones de la empresa, y reducir los errores causados por el proceso de normalización actual de la empresa, al eliminar información relevante de los datos, y por el ingreso incorrecto de direcciones por parte de los destinatarios.

2.2.3. Causas

En primer lugar, en muchos casos los consumidores son poco rigurosos al facilitar sus datos en formularios, sobre todo cuando lo hacen a través de la web. Esto puede llevar a que las direcciones no sean ingresadas correctamente, lo que dificulta su normalización, y en consecuencia, su georreferenciación.

En segundo lugar, uno de los elementos perjudiciales del proceso actual de normalización de direcciones es que se eliminan los siguientes caracteres: #, N^o, SN, INT., INTERIOR, PISO, LOCAL, CASA, POBLACIÓN, S/N, DPTO, DEPTO, OFICINA, DEPARTAMENTO, BLOCK, PASAJE, PSJE, PSJE., CONDOMINIO, SUCURSAL, CALLE N^o, UNDEFINED, SIN NUMERO, NUMERO, PARCELACION, (TEXTO), COND., ESQUINA, Ñ por I, Ñ± por N, Ñ© por E, Ñ3 por O.

En tercer lugar, no es posible exigir a los clientes cumplir con un formato específico de direcciones, ya que se correría el riesgo de que el cliente cambie de proveedor de servicio y la calidad del servicio se vea afectada[8].

2.2.4. Consecuencias

Cuando las direcciones no se normalizan correctamente, se asigna una ruta incorrecta a la orden de servicio. Esto genera una excepción conocida como mal ruteo (MR o MRH), que puede impedir o retrasar la entrega del pedido al cliente final, ya que se debe realizar el ruteo de forma manual o depender de los conocimientos de los courier. Además de requerir una gran cantidad de horas humanas destinadas a corrección de direcciones, esto afecta directamente el nivel de servicio de la empresa.

Entre las consecuencias principales que trae el problema de la normalización actual de direcciones, se identifican el debilitamiento en la posición de mercado, la disminución en la satisfacción de los clientes y el aumento en los costos de distribución.

Además, este problema afecta negativamente el objetivo de la empresa de ser la primera en servicio, ya que para lograrlo es necesario contar con el mejor nivel de servicio. Por lo tanto, es importante abordar y solucionar este problema, para mejorar la posición de la empresa en el mercado y aumentar la satisfacción de los clientes.

2.3. Objetivos del proyecto

2.3.1. Objetivo general

El objetivo general de este trabajo corresponde a “*desarrollar y evaluar un sistema de normalización de direcciones de calles, empleando un modelo supervisado de aprendizaje de máquinas, con el fin de compararlo con el modelo de normalización basado en programación tradicional actualmente utilizado por una empresa de despacho*”.

2.3.2. Objetivos específicos

- Levantar información sobre el proceso actual de normalización de direcciones, para identificar sus limitaciones y definir requerimientos de mejora.
- Desarrollar un modelo supervisado de aprendizaje de máquinas para normalizar direcciones.
- Realizar una evaluación comparativa de desempeño entre el modelo de aprendizaje de máquinas y el código de programación tradicional utilizado actualmente para normalizar direcciones, utilizando métricas de precisión, recall, y F1-score para cuantificar mejoras en la calidad de la normalización.
- Generar un plan de implementación y propuestas de mejora.

2.4. Alcances

2.4.1. Tipos de errores a abordar

El presente proyecto se enfoca principalmente en abordar ciertos aspectos del problema del sistema actual de normalización de direcciones de la empresa. Este sistema, por su antigüedad y empleo de un enfoque de programación convencional, genera una serie de complicaciones significativas en la precisa ubicación de direcciones.

El objetivo primordial es mostrar que es beneficioso capacitar una máquina para que realice con precisión la normalización de direcciones, preservando información relevante al recibir el nombre de la calle, avenida, pasaje u otro elemento identificativo, y evitando su eliminación. Además, se pretende corregir los errores de ortografía.

Para efectos de simplificar las etiquetas con que se entrenará el modelo, se considerará solamente normalizar el nombre de la calle/avenida/pasaje/otro.

Si bien se identifica como una de las causas del problema la poca rigurosidad de los destinatarios al declarar su dirección, la máquina no será entrenada para devolver una dirección correcta cuando lo que ingresa un destinatario corresponda a un lugar y no a una dirección. Por ejemplo, una dirección declarada como:

*SUCURSAL CORREOS CHILE LA FLORIDA ,
MALL LOS DOMINICOS ,
LOCAL INTIME PEDRO DE VALDIVIA ,*

no se considerará para ser incluida en la base de datos, ya que constituyen ejemplos de lugares, y no de direcciones.

Al ser las direcciones ingresadas por personas, están sujetas a los errores que estas cometen. Estos errores generan que las direcciones contengan información incompleta (falta el nombre de la calle, la numeración o la comuna) o información incorrecta (uso de abreviaturas, nombres de calle mal escritos, lugares en vez de direcciones o errores ortográficos).

Cabe mencionar que este tipo de errores no serán abordados en este trabajo, exceptuando el de errores ortográficos. Se pretende, por un lado, que la máquina aprenda cuando una dirección está mal escrita, y por otro lado, que la máquina no permita la pérdida de información relevante contenida en la dirección ingresada por el destinatario.

Para facilitar la comprensión, se ejemplificará a continuación. Si un destinatario declara la siguiente dirección:

JERONIMO DE ALDRETE

Lo que la máquina aprenderá como calle normalizada de manera correcta es:

GERONIMO DE ALDERETE

Y si el destinatario declara la siguiente calle:

VICUÑA MACKENNA

Lo que la máquina aprenderá como calle normalizada de manera correcta es:

VICUÑA MACKENNA

2.4.2. Validación de direcciones

Este trabajo pretende abordar principalmente el problema de la normalización de direcciones, a nivel del nombre de la calle. Lo más importante es que no se pierda información de las palabras que conforman la dirección. Debido al acotado tiempo con el que se cuenta, queda fuera del alcance de este trabajo la validación de la existencia de dichas calles. Esto se sugiere como futuros pasos de un próximo trabajo de título.

2.4.3. Avenidas y calles populares

Hay algunos casos de vías públicas a las cuales popularmente se les refiere con un nombre diferente al oficial o administrativo. Por ejemplo, la avenida Libertador Bernardo O'Higgins, se conoce entre la población como *Alameda*. Sin embargo, este no es su nombre formal. Estos casos quedan fuera del alcance de este trabajo, debido a la limitación de tiempo y la dificultad de conocer todos los casos que existen. Esto también se propone para abordar en trabajos futuros.

Tabla 2.2: Comparación entre nombres populares y oficiales de calles y vías.
Fuente: Elaboración propia.

Nombre popular	Nombre oficial
Alameda	Av. Libertador Bernardo O'Higgins
Gran Avenida	Av. José Miguel Carrera
Panamericana	Av. José Joaquín Prieto

2.4.4. Aplicabilidad del modelo

Con la realización de este trabajo de título se pretende obtener un modelo de machine learning que pueda recibir direcciones y entregar direcciones normalizadas de forma correcta. Sin embargo, queda fuera del alcance la etapa de “Deployment” de la metodología CRISP-DM, con lo cual, la realización de este trabajo no contempla implementar la automatización de la normalización de direcciones de esta empresa. Lo anterior, se debe principalmente a que se dispone de un periodo acotado de tiempo. Esta fase de la metodología será delegada al área de Tecnologías de Información de la empresa, para que el modelo pueda ser aplicado, y que se pueda hacer una revisión del diccionario completo de direcciones.

Capítulo 3

Marco Conceptual y Metodología

3.1. Estado del Arte

Se considera como parte del marco conceptual el trabajo de título “*Sistema de búsqueda inteligente de direcciones para empresa de distribución postal*”, de Daniel Ponce. En dicho trabajo se propuso como solución un sistema de búsqueda inteligente de direcciones postales que permitiese normalizar automáticamente grandes volúmenes de direcciones utilizando modelos de Procesamiento de Lenguaje Natural, con el fin de aumentar el porcentaje de asignación de códigos postales de forma automática.

La propuesta planteada por Ponce consiste en el desarrollo de un clasificador de direcciones diseñado para segmentar y etiquetar sus atributos clave: nombre de la calle, el número principal y la información adicional asociada. Este clasificador realiza una verificación previa para determinar si la dirección posee un número principal claramente definido, ya que en caso de no ser así, se considera inválida y no se somete al proceso de normalización. La justificación de esto se debe a la necesidad de que el código postal requiere un reconocimiento preciso del frente de cuadra para su asignación. Una vez etiquetadas las direcciones, se implementa un modelo de coincidencia de texto que emplea la distancia de Levenshtein y el ratio de similitud de Levenshtein. Este modelo genera un ranking de las tres direcciones más similares basándose en puntajes de similitud, los cuales luego se utilizan para calcular el puntaje de selección. La normalización de la dirección tiene lugar únicamente si se cumplen los criterios predefinidos de asignación.

Es importante señalar que el modelo de coincidencia de texto de mayor rendimiento es el que elimina abreviaturas asociadas a tipos de calles (por ejemplo, PSJE o AVDA), calcula el puntaje ranking mediante el ratio de similitud de Levenshtein y evalúa la coincidencia mediante la comparación de la cadena completa de texto. Para mejorar la precisión de la normalización, se establecen categorías de confiabilidad basadas en los puntajes de selección. Este trabajo se centró en la normalización a nivel de comuna y zonas urbanas, específicamente en las comunas de Santiago y Quilicura. Mediante el uso de categorías de confiabilidad, logró una tasa de normalización superior al 83 %, con un error de 0,64 %. En este trabajo no se abordó la georreferenciación de las direcciones normalizadas, limitándose principalmente a la asignación del código postal correspondiente según el algoritmo de coincidencia de texto aplicado a las direcciones existentes en la base de datos. Por lo tanto, la convergencia principal entre el trabajo presentado en esta memoria y el referido anteriormente radica en el

enfoque compartido hacia la normalización precisa de direcciones.

En el trabajo “*Address standardization using the natural language process for improving geocoding results*”[9] de Dilek Küçük Matci y Uğur Avdan, se pretende abordar el problema de que la aplicación efectiva de la geocodificación se ve limitada en ciertas áreas debido a diversas razones, entre las que se incluyen inconsistencias en los formatos de direcciones, sistemas de numeración inexactos, errores ortográficos, el uso generalizado de abreviaturas y la falta de datos referentes al proceso de geocodificación. En este trabajo se propone abordar estas problemáticas mediante un proceso integral de estandarización. Para lograrlo, se emplea un enfoque que descompone las direcciones, identifica posibles errores ortográficos y abreviaturas, y reorganiza las mismas mediante el Procesamiento del Lenguaje Natural (NLP). Utilizando direcciones de escuelas primarias en el distrito de Eskisehir como datos de prueba, se lleva a cabo inicialmente el proceso de geocodificación, utilizando tanto la API de geocodificación de Google como la API de geocodificación de ArcGIS. Posteriormente, las direcciones se “reformatean” en tres formas estándar mediante procesos de estandarización. Se geocodifican las direcciones reformateadas y se comparan los resultados obtenidos con los resultados de geocodificación no estandarizados. Los hallazgos muestran que la estandarización implementada conlleva una mejora significativa en la precisión de los resultados de geocodificación. Este método no solo contribuye a elevar la precisión del proceso de codificación geográfica, sino que también amplía su aplicabilidad en un contexto más extenso.

Otro trabajo estudiado fue “*Propuesta de guía metodológica para el manejo de problemas en la estandarización y calidad de datos de direcciones urbanas en Colombia*”[10], de Daniel Rivas. Este trabajo aborda la complejidad de las direcciones urbanas en Colombia, donde los factores geográficos y la generación de vías adicionales dificultan la estandarización. Se destaca la presencia común de problemas de calidad de datos, afectando la eficiencia organizacional. Se menciona que la falta de calidad en las direcciones impacta negativamente en la toma de decisiones y operaciones empresariales, y que aunque la problemática es reconocida, se evidencia una escasa documentación al respecto. En este trabajo se propone una guía metodológica para abordar los problemas de calidad de datos en direcciones, buscando reducir los impactos negativos en las organizaciones.

3.2. Normalización de direcciones

La normalización de direcciones es el proceso de actualizar e implementar un estándar o formato en los datos de direcciones[11]. Es importante para garantizar la integridad y precisión de los datos. Este proceso ayuda a corregir duplicados y anomalías en la base, evitar la creación y actualización de relaciones y dependencias no deseadas, prevenir borrados indeseados y optimizar el espacio de almacenamiento de datos[12].

La limpieza y preparación de datos en general implica identificar y corregir errores, inconsistencias y duplicados, para mejorar su calidad y utilidad. La normalización de datos es un proceso relacionado que implica transformar los valores de un formato inaceptable a un formato aceptable[13]. Se podría considerar la normalización de direcciones como un subconjunto de la limpieza y preparación de datos, ya que implica aplicar técnicas específicas para mejorar la calidad de los datos que contiene una dirección.

La normalización de direcciones es esencial para empresas de todos los sectores que trabajan con cantidades masivas de datos, ya que permite probar la existencia del domicilio, corregir la información errónea o desactualizada, identificar los registros duplicados y evitar la duplicidad de envíos, optimizar las operaciones, mejorar la segmentación de clientes, su experiencia y la reputación de la marca, entre otros.

3.3. Ciencia de Datos

Aunque existen múltiples definiciones para la ciencia de datos, para fines de este trabajo, parece apropiada la que entrega Data Science UC: “La ciencia de datos es una disciplina que, mediante la combinación de modelos matemáticos y estadísticos, la programación computacional y las técnicas de visualización de datos, permiten obtener el máximo valor de los datos para apoyar los procesos de tomas de decisiones”. [14]

La normalización de direcciones es un proceso importante para garantizar la integridad y precisión de los datos, ya que ayuda a corregir duplicados y anomalías en la base, evitando la creación y actualización de relaciones y dependencias no deseadas, previniendo borrados indeseados, y optimizando el espacio de almacenamiento. [12]. Así, al trabajar con ciencia de datos, se pueden desarrollar modelos y algoritmos que automaticen el proceso de normalización y mejoren su precisión.

3.3.1. Aprendizaje Automático

El Aprendizaje Automático o Machine Learning utiliza algoritmos matemáticos para permitir que las máquinas aprendan por sí mismas. Constituye una forma analítica de resolver problemas a través de la cual los algoritmos aprenden a identificar patrones de comportamiento, los clasifican y los utilizan para realizar predicciones [15]. A diferencia de la programación tradicional, en el aprendizaje automático no se entregan instrucciones, se entregan datos. Se utiliza la inteligencia artificial para que el programa encuentre patrones en un conjunto de datos masivos. Para este trabajo, se utilizarán algoritmos de aprendizaje automático para procesar datos y proponer correcciones a las direcciones erróneas, logrando obtener direcciones con una normalización óptima.

Los pasos generales para crear un modelo de Machine Learning son: recopilar y preparar los datos, elegir un algoritmo, entrenar el modelo utilizando los datos, para finalmente evaluar el modelo y utilizarlo. Existen tres tipos de aprendizaje de máquinas: supervisado, no supervisado y semi supervisado.

3.3.1.1. Aprendizaje supervisado

El aprendizaje supervisado se define por su uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifican datos o prevén resultados con precisión [16]. En el aprendizaje supervisado se conoce mucho sobre los datos de entrenamiento, lo que permite “asistir” a la máquina, mediante datos etiquetados y resultados correctos.

A medida que se introducen los datos en el modelo, se van ajustando sus ponderaciones, como parte de un proceso de validación cruzada. Se le puede informar a la máquina si su predicción fue precisa. Con el tiempo, la máquina aprende y adapta su modelo para mejorar el resultado. La máquina continúa ajustando el modelo con los datos de entrenamiento, y

uno lo ajusta según los nuevos datos. Además, el aprendizaje supervisado permite a las organizaciones resolver una amplia variedad de problemas del mundo real a escala[16]. Esto servirá para automatizar el proceso de normalizar direcciones en grandes conjuntos de datos. Se distinguen tres categorías dentro del aprendizaje supervisado:

- Regresión: el objetivo es predecir un valor numérico o continuo. La salida es una variable continua.
- Clasificación binaria: el objetivo es predecir entre dos clases o categorías. Por ejemplo: sí/no, 1 o 0.
- Clasificación multiclase: el modelo debe asignar la entrada a una de varias categorías (tres o más clases).

3.3.1.2. Aprendizaje no supervisado

En este tipo de aprendizaje, la máquina necesita muchos datos para hacer las observaciones por su cuenta. Cuantos más datos tenga, es más sencillo para la máquina observar y estudiar las tendencias que permitan definir una categoría. La máquina no conoce la respuesta correcta, sino que usa algoritmos para establecer las conexiones mediante el estudio y observación de los datos. Es decir, aprende y mejora mediante ensayo y error.

En el aprendizaje no supervisado, las categorías principales son:

- Clustering: se agrupan los datos en conjuntos homogéneos según una medida de similitud o distancia.
- Asociación: se descubren patrones de asociación o relaciones entre variables dentro de grandes conjuntos de datos.
- Reducción de dimensionalidad: reduce la cantidad de variables en un conjunto de datos, manteniendo la mayor cantidad posible importante de información.
- Generación de reglas: encuentra reglas interesantes o patrones en los datos.

3.3.1.3. Aprendizaje semi supervisado

En el aprendizaje semi-supervisado, se utiliza una combinación de datos etiquetados y no etiquetados para el entrenamiento del modelo. Cruza ambos tipos de aprendizaje, supervisado y no supervisado. Tiene sentido utilizar este tipo de aprendizaje cuando los otros dos métodos anteriores no resuelven el problema. Sirve cuando los datos son demasiados, y no es práctico aplicar el aprendizaje supervisado.

Las categorías principales son:

- Razonamiento inductivo: con los datos etiquetados, el programa usa razonamiento inductivo para expandir su propio vocabulario.
- Razonamiento transductivo: mejora el modelo usando conjeturas más finas sobre lo que habrá en los datos no etiquetados.

3.3.1.4. Aprendizaje por refuerzo

El aprendizaje por refuerzo se basa en la idea de que un agente toma decisiones en un entorno dinámico y recibe retroalimentación en forma de recompensas o penalizaciones. Se busca un modelo que clasifique con precisión distintos conjuntos de datos o que descubra clusters con sentidos. Hace que la máquina itere o repita continuamente para mejorar el resultado.

El objetivo del agente en el aprendizaje por refuerzo es aprender una política, es decir, una estrategia que le indique qué acción tomar en cada estado para maximizar la recompensa acumulada a lo largo del tiempo. El agente explora diferentes acciones en diversos estados y ajusta su política en función de las recompensas obtenidas.

Este enfoque es particularmente útil en situaciones donde el agente no tiene ejemplos etiquetados, pero puede aprender mediante la interacción con su entorno. Ejemplos de aplicaciones incluyen juegos, robótica, gestión de recursos y toma de decisiones en tiempo real.

3.3.2. Preparación de datos: conjuntos de entrenamiento y prueba

En el campo del aprendizaje automático, la correcta preparación de datos es fundamental para crear modelos efectivos y precisos. Esta preparación implica clasificar los datos en dos categorías: entrenamiento y prueba. El conjunto de entrenamiento permite al modelo discernir patrones y aprender de ellos, mientras que el conjunto de prueba evalúa su capacidad para aplicar estos aprendizajes de manera efectiva en escenarios que simulan condiciones reales, asegurando que los modelos estén preparados para interpretar datos nuevos de manera efectiva.

- Conjunto de datos de entrenamiento: cumple la función de instruir al modelo en la ejecución de tareas específicas o en la formulación de predicciones. Generalmente, estos datos están etiquetados, proporcionando las instancias de entrada y de salida esperadas, permitiendo que el modelo identifique estos aprendizajes.
- Conjunto de datos de prueba: se destina a la evaluación del rendimiento del modelo. Generalmente, estos datos no están etiquetados, lo que significa que se desconoce el resultado deseado, con lo cual permiten verificar la capacidad del modelo para extender sus predicciones más allá de su entrenamiento, ofreciendo una medida de generalización del modelo a nuevos datos.

En el ámbito de la ciencia de datos, es una práctica habitual asignar el 80 % de los datos al entrenamiento del modelo, y el 20 % restante a su evaluación.

3.3.3. Support Vector Machine

El modelo Support Vector Machine es un algoritmo de aprendizaje automático potente y popular que se utiliza para tareas de clasificación y regresión. Una de las ventajas clave de SVM es su capacidad para manejar datos de alta dimensión de manera efectiva, evitando el sobreajuste. SVM construye su solución en términos de un subconjunto de la entrada de entrenamiento (Awad & Khanna, 2015). Una vez que el modelo SVM ha sido entrenado utilizando el algoritmo SVM, puede utilizarse para predecir la clase o el valor de nuevos datos no etiquetados.

Hay dos algoritmos principales que se utilizan para entrenar máquinas de vectores de soporte: regresión de vectores de soporte y clasificación de vectores de soporte.

Support Vector Regression (SVR) es una variante de SVM que se utiliza para tareas de regresión. Su objetivo es encontrar un hiperplano que capture tantos puntos de entrenamiento dentro de un margen de error específico, minimizando al mismo tiempo la suma de errores.

La clasificación de vectores de soporte (SVC), por otro lado, se utiliza para tareas de clasificación. Su objetivo es encontrar el hiperplano que separe los puntos de datos pertenecientes a diferentes clases con el máximo margen posible, permitiendo al mismo tiempo un cierto grado de clasificación errónea. Para entrenar el modelo SVM, se debe elegir una función del kernel adecuada y ajustar los parámetros relevantes.

En el caso de un modelo de aprendizaje supervisado para normalizar direcciones, un SVC puede clasificar las direcciones en categorías predefinidas, basándose en características como la estructura y el formato de las direcciones. Esto podría ayudar a mejorar la precisión y eficacia en la normalización de direcciones, ya que el modelo SVC podría aprender a identificar patrones y relaciones en los datos para hacer predicciones más precisas. Además, un SVC puede ser efectivo en espacios de alta dimensión y puede adaptarse bien a situaciones donde el número de dimensiones es mayor que el número de muestras, lo que resulta útil en el caso de datos complejos o con muchas características, como las direcciones. Este modelo es robusto y preciso para problemas lineales y no lineales, es capaz de manejar múltiples clases de salida. Además, son pocos los parámetros que se tienen que ajustar.

La idea clave detrás de SVC es maximizar el margen entre clases. El margen se define como la distancia entre el hiperplano de decisión y los puntos más cercanos de cada clase, conocidos como vectores de soporte.

La función objetivo de un SVC se puede describir como:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i, \forall i \quad (3.1)$$

$$\xi_i \geq 0, \forall i$$

Donde:

- y_i es la etiqueta de clase del i -ésimo punto de datos.
- x_i es el i -ésimo punto de datos.
- ξ_i es la variable de holgura asociada al i -ésimo punto de datos.
- n es el número de muestras de entrenamiento.

Kernel lineal

En el contexto de las SVM, un kernel es una función que se utiliza para transformar los datos de entrada en un espacio dimensional más alto, donde pueda ser más fácil separarlos linealmente. Un kernel lineal es uno de los tipos más simples de kernel.

Un kernel lineal no transforma el espacio de entrada, sino que mantiene los datos en su espacio original, y encuentra un hiperplano que separa las clases. Matemáticamente, un kernel lineal es simplemente el producto escalar de dos vectores de entrada.

Cuando se asume que los datos son linealmente separables, o cuando el número de características es muy grande en relación con el número de muestras, los kernels lineales son útiles. En esos casos, un kernel lineal puede ser tan efectivo como uno más complejo, pero tiene la ventaja de que es más eficiente en términos de cálculo.

Al usar un kernel lineal en SVM, quiere decir que se está buscando un hiperplano en el espacio original de los datos, que pueda separar las clases de la manera más efectiva posible.

3.3.4. Procesamiento de Lenguaje Natural

El procesamiento de lenguaje natural (NLP por sus siglas en inglés) es una tecnología de Machine Learning que brinda a las computadoras la capacidad de interpretar, manipular y comprender el lenguaje humano[17]. NLP combina la lingüística computacional con modelos estadísticos, de Machine Learning y Deep Learning. Juntas, estas tecnologías permiten a los ordenadores procesar el lenguaje humano en forma de datos de texto o voz y “comprender” su significado completo, junto con la intención y el sentimiento del orador o escritor[18].

Sin embargo, las técnicas de NLP, como el uso de modelos de lenguaje preentrenados, pueden requerir grandes cantidades de datos específicos para obtener resultados óptimos. En el caso de la tarea de este trabajo, la disponibilidad de datos específicos podría ser limitada. Además, la normalización de nombres de calles con frecuencia implica reglas lingüísticas complejas y variaciones. SVM, al ser más flexible en términos de estructuras de reglas, puede adaptarse mejor a estas complejidades.

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) es una técnica utilizada en el procesamiento de lenguaje natural (NLP) y en la recuperación de información para reflejar la importancia de una palabra en un documento dentro de un conjunto de documentos (corpus). El valor TF-IDF aumenta proporcionalmente con el número de veces que una palabra aparece en un documento, pero se compensa con la frecuencia de la palabra en el corpus, lo que ayuda a controlar el hecho de que algunas palabras son generalmente más comunes que otras.

Term Frequency (TF) es la frecuencia con la que un término aparece en un documento. Se calcula como el número de veces que el término aparece en el documento dividido en el número total de términos del documento. Inverse Document Frequency (IDF) mide la importancia del término en todo el corpus. Se calcula como el logaritmo del número total de documentos en el corpus dividido en el número de documentos que contienen el término.

El TF-IDF se utiliza con frecuencia para transformar textos en un formato más manejable para los algoritmos de aprendizaje automático, y es particularmente útil en tareas como la búsqueda de información y clasificación de documentos.

3.3.5. Otros modelos de aprendizaje supervisado

Naive Bayes

El algoritmo de Naïve Bayes es un algoritmo de clasificación basado en el teorema de Bayes[19]. Este algoritmo es menos intenso a nivel computacional que otros algoritmos y, por tanto, resulta útil para generar rápidamente modelos de minería de datos para detectar relaciones entre columnas de entrada y columnas predecibles[20]. Por una parte, este algoritmo se puede utilizar para la exploración inicial de los datos y, posteriormente, aplicar los resultados para crear modelos de minería de datos adicionales con otros algoritmos más complejos y precisos desde el punto de vista computacional[20].

Por otra parte, el supuesto de independencia condicional en Naive Bayes podría no ser realista en el contexto de nombres de calles, donde la relación entre diferentes elementos de la dirección puede ser intrínsecamente dependiente. Un modelo con un algoritmo de Naive Bayes puede tener dificultades (como sufrir de sesgo o varianza) cuando se enfrenta a conjuntos de datos desbalanceados, lo cual es común en problemas de normalización de direcciones. Además, puede fallar si se encuentra con palabras nuevas que no están en el vocabulario, no tiene en cuenta las relaciones entre las palabras ni el orden de las mismas, y es muy sensible a la calidad y cantidad de los datos.

Redes Neuronales

Las Redes Neuronales Artificiales (RNA) son un tipo de algoritmo de aprendizaje automático inspirado en el funcionamiento del cerebro humano. Estas redes consisten en capas de nodos interconectados, conocidos como neuronas artificiales, que procesan información y aprenden a realizar tareas específicas a través del ajuste de los pesos y sesgos de sus conexiones.[21]

Sin embargo, las redes neuronales, especialmente en tareas de NLP, pueden ser computacionalmente intensivas, y requerir grandes cantidades de datos para su entrenamiento. Esto podría resultar poco práctico en escenarios con conjuntos de datos moderados. Además, las redes neuronales tienden a ser “cajas negras”, dificultando la interpretación de los resultados. En modelos de normalización de direcciones, SVM proporciona una alternativa más interpretable.

Árboles de decisión

Los árboles de decisión son modelos predictivos que se utilizan en aprendizaje supervisado tanto para tareas de clasificación como de regresión. Estos modelos funcionan con una estructura en forma de árbol, representando cada nodo una decisión (basada en una característica) y cada rama el resultado de esa decisión. Los árboles de decisión resultan bastante atractivos porque son modelos interpretables y sencillos de entender.

Sin embargo, los árboles de decisión son propensos al sobreajuste, y pueden volverse complejos si no se controlan de forma adecuada. SVM es un modelo que permite manejar de mejor forma problemas de no linealidad, y generalizar en conjuntos de datos más diversos, especialmente con kernel.

K-Vecinos más cercanos

El k-NN (K Nearest Neighbor, por sus siglas en inglés) es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación y regresión. Generalmente se usa como

un algoritmo de clasificación, partiendo de la suposición de que se pueden encontrar puntos similares cerca uno del otro. La idea central es asignar una etiqueta o valor numérico a una instancia basándose en las etiquetas o valores de las instancias vecinas más cercanas en el espacio de características[22].

k-NN tiende a sufrir la “maldición de la dimensionalidad”, lo que significa que no funciona bien con entradas de datos de alta dimensión. Después de que el algoritmo alcanza la cantidad óptima de funciones, las funciones adicionales aumentan la cantidad de errores de clasificación, especialmente cuando el tamaño de la muestra es más pequeño. En esa misma línea, k-NN puede requerir más datos para obtener resultados significativos. Este algoritmo puede ser propenso a sobreajustar si el conjunto de datos de entrenamiento es ruidoso o tiene variabilidad. Finalmente, kNN es sensible a la escala de las características debido a su cálculo de distancias euclidianas.

Tabla 3.1: Resumen y comparación de algoritmos de clasificación de aprendizaje supervisado comúnmente empleados. Fuente: Elaboración propia.

Algoritmo	Tipo de modelo	Características
Regresión Logística	Lineal	Adecuado para problemas lineales y fáciles de interpretar
SVM	Lineal y no lineal (con el uso de kernels)	Eficiente en espacios de alta dimensión, útil cuando hay límites de decisión complejos
Árboles de decisión	No lineal	Fácil de entender e interpretar, puede manejar tanto datos numéricos como categóricos
Naive Bayes	Probabilístico	Eficiente, asume independencia condicional entre características
k-NN	No paramétrico	Basado en la similitud entre instancias, simple pero puede ser computacionalmente costoso
RNA	No lineal	Puede aprender representaciones complejas, pero puede requerir grandes conjuntos de datos y poder computacional

3.3.6. Métricas de evaluación en machine learning

Matriz de confusión

Una matriz de confusión es una herramienta que permite analizar los resultados de cómo trabaja un algoritmo de aprendizaje supervisado. Siempre se presenta en forma de tabla, de manera que en cada columna aparece el número de predicciones de cada clase, mientras que en cada fila se muestra el número real de instancias de cada clase. Es decir, esta matriz pone en relación las predicciones realizadas por un algoritmo de aprendizaje supervisado y los resultados correctos que debería haber mostrado[23]. Dicho de otra forma, proporciona una vista detallada de los diferentes tipos de errores que el modelo comete, como falsos positivos y falsos negativos. Esto es de gran ayuda para la normalización de nombres de calles, donde los errores pueden tener un impacto significativo en la usabilidad de la información para georreferenciar. La matriz de confusión tiene la siguiente forma:

$$\begin{array}{cc} \text{VN} & \text{FP} \\ \text{FN} & \text{VP} \end{array}$$

Donde los elementos son:

- VN (Verdaderos Negativos): cantidad de direcciones que el modelo ha identificado correctamente como ya normalizadas. Es decir, son las direcciones que el modelo identificó correctamente como ya normalizadas y, por lo tanto, no requieren más procesamiento.
- FP (Falsos Positivos): cantidad de direcciones que el modelo ha normalizado incorrectamente. Es decir, son las direcciones que el modelo identificó erróneamente como necesitadas de normalización y, por lo tanto, normalizó incorrectamente.
- FN (Falsos Negativos): cantidad de direcciones que el modelo ha identificado incorrectamente como ya normalizadas. Es decir, direcciones que el modelo debería haber normalizado pero identificó erróneamente como ya normalizadas.
- VP (Verdaderos Positivos): cantidad de direcciones que el modelo normalizó correctamente. Es decir, las direcciones que el modelo identificó correctamente como necesitadas de normalización, y efectivamente las normalizó correctamente.

Sin embargo, en problemas con múltiples clases no se puede desempaquetar la matriz de confusión en los clásicos términos “verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos”, como se hace en problemas binarios. En estos casos, la matriz de confusión se puede utilizar para calcular las métricas específicas para cada clase, como la precisión, sensibilidad, F1 score, y otras métricas.

Precisión

La métrica de precisión es crucial en este contexto, ya que mide la proporción de nombres de calles normalizados correctamente en relación con el número total de predicciones. Una alta precisión asegura que la mayoría de las correcciones realizadas por el modelo son correctas, lo que es fundamental en aplicaciones como la normalización de nombres de calles para evitar cambios incorrectos en la data. La precisión se calcula de la siguiente forma:

$$Precisión = \frac{VP}{VP + FP} \quad (3.2)$$

Donde VP son los verdaderos positivos y FP son los falsos positivos.

Recall

El recall, también conocido como exhaustividad o sensibilidad, es una métrica que se utiliza para evaluar el rendimiento de un modelo de clasificación en Machine Learning. Esta métrica informa sobre la cantidad de verdaderos positivos que el modelo es capaz de identificar[24]. El recall se calcula como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos negativos. En otras palabras, el recall es la respuesta a la pregunta ¿qué porcentaje de los verdaderos positivos somos capaces de identificar?[24]. La fórmula para el recall es la siguiente:

$$Recall = \frac{VP}{VP + FN} \quad (3.3)$$

Donde VP son los verdaderos positivos, y FN son los falsos negativos. Es importante tener en cuenta que el recall y la precisión suelen estar en tensión. Es decir, mejorar el recall generalmente reduce la precisión y viceversa[25]. Esta métrica es importante para asegurar que el modelo pueda identificar la mayoría de las oportunidades de normalización de nombres de calles en el conjunto de datos, ya que, obtener un recall alto significa que el modelo no pasa por alto muchos casos que requieren corrección.

F1 Score

El F1 Score es la media armónica entre recall y precisión. Combina ambos en un solo valor, proporcionando una medida más completa para conocer el rendimiento del modelo. F1 Score toma valores entre 0 y 1, siendo 1 el mejor valor.

$$F1 = 2 * \frac{Precisión * Recall}{Precisión + Recall} \quad (3.4)$$

Jaccard Score

Jaccard Score es una métrica que sirve para evaluar el rendimiento en problemas de clasificación y recuperación de información. Se basa en la teoría de conjuntos, y su fórmula se expresa como la relación entre la intersección y la unión de dos conjuntos. El resultado varía entre 0 y 1, donde 0 significa que los conjuntos son disjuntos (es decir, que no hay elementos compartidos entre ellos) y 1 indica que los conjuntos son idénticos. En el contexto de clasificación, el Jaccard Score entrega una medida cuantitativa de la similitud entre dos conjuntos de etiquetas, y se utiliza en la evaluación de modelos de clasificación, con el fin de medir la precisión en la predicción de las etiquetas.

Dados los conjuntos A y B, el Jaccard Score (J) se calcula mediante la fórmula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.5)$$

Donde

- $|A \cap B|$ es el tamaño de la intersección entre los conjuntos A y B, es decir, el número de elementos compartidos por ambos conjuntos.
- $|A \cup B|$ es el tamaño de la unión de los conjuntos A y B, es decir, el número total de elementos en ambos conjuntos, sin considerar duplicados.

Distancia de Levenshtein

Uno de los algoritmos de coincidencia de texto más recomendado corresponde al de distancia de edición, en específico a la distancia de edición de Levenshtein, que permite minimizar los fallos de selección por causa de errores ortográficos[26]. La distancia de Levenshtein corresponde al número mínimo de operaciones (inserciones, eliminaciones o sustituciones de caracteres) necesarias para transformar una cadena en la otra.

La fórmula matemática para el cálculo de la distancia de Levenshtein entre dos cadenas a y b se define recurrentemente de la siguiente forma:

$$\text{lev}(a, b, i, j) = \begin{cases} \max(i, j) & \text{si } \min(i, j) = 0 \\ \min \left\{ \begin{array}{l} \text{lev}(a, b, i - 1, j) + 1, \\ \text{lev}(a, b, i, j - 1) + 1, \\ \text{lev}(a, b, i - 1, j - 1) + \text{costo_sustitucion}(a[i], b[j]) \end{array} \right\} & \text{de lo contrario} \end{cases} \quad (3.6)$$

Donde i y j son las longitudes actuales de las subcadenas a y b, y costo sustitución es una función que entrega 1 si los caracteres son diferentes y 0 si los caracteres son iguales.

El resultado que arroja es un valor entero no negativo, que indica cuán diferentes son las dos cadenas de texto. Si el valor es cero, las cadenas son idénticas, y un valor mayor significa que tienen mayores diferencias. Por ejemplo, para las cadenas de texto “AMERICO” y “AMÉRICO”, la distancia de Levenshtein es de 1, ya que se requiere solamente una operación de sustitución para transformar la primera cadena en la segunda. Esta distancia se puede calcular mediante un código de programación y hay funciones definidas en Python que lo hacen.

3.4. Interfaz de Programación de Aplicaciones

Las Interfaces de Programación de Aplicaciones, conocidas por sus siglas en inglés como API, constituyen un conjunto de reglas y protocolos imprescindibles en el ámbito contemporáneo del desarrollo de software. Estas interfaces facilitan la comunicación y el intercambio de datos entre sistemas y aplicaciones, permitiendo una integración armoniosa y una mejora de la funcionalidad.

En el contexto de las API, la palabra aplicación se refiere a cualquier software con una función distinta. La interfaz puede considerarse como un contrato de servicio entre dos aplicaciones. Este contrato define cómo se comunican entre sí mediante solicitudes y respuestas. La documentación de su API contiene información sobre cómo los desarrolladores deben estructurar esas solicitudes y respuestas.[27]

Las API proporcionan un enfoque estandarizado para la interacción entre aplicaciones, lo que permite a los desarrolladores aprovechar las funcionalidades y recursos ya existentes. Esta capacidad de las API para facilitar la cooperación entre aplicaciones es fundamental para la innovación y la eficiencia en el desarrollo de software.

La arquitectura de las API con frecuencia se explica en términos de cliente y servidor. La

aplicación que envía la solicitud se llama cliente, y la que envía la respuesta se llama servidor. Las integraciones de las API son componentes de software que actualizan automáticamente los datos entre los clientes y los servidores.

3.4.1. API REST

Las API REST son las más populares y flexibles que se encuentran en la web actualmente. El cliente envía las solicitudes al servidor como datos. El servidor utiliza esta entrada del cliente para iniciar funciones internas y devuelve los datos de salida al cliente. [27] La arquitectura REST también enfatiza la escalabilidad y la capacidad de caché, lo que la hace adecuada para su uso en la web moderna. El atributo más destacado de la API REST es su naturaleza sin estado. Esto implica que los servidores no retienen información del cliente entre las peticiones.

Los principales beneficios de las API de REST son:

- Se pueden utilizar para integrar nuevas aplicaciones con los sistemas de software existentes, es decir, se pueden utilizar las API con el código existente.
- En caso de ser necesario, se pueden realizar cambios en la API sin necesidad de reescribir todo el código.
- Son ampliables a distintas plataformas, lo cual puede aprovechar una empresa para dar acceso a sus bases de datos internas.
- Cualquier cambio que se haga en una de las partes del código no afecta a los sistemas que son enlazados por la API, ya que cada sistema está obligado a hacer los cambios internos necesarios.

3.5. Metodología CRISP-DM

La metodología que se utilizará para el desarrollo de este trabajo es la de CRISP-DM (Cross Industry Standard Process for Data Mining)[28], ya que es la que más se utiliza en la actualidad en la industria de la ciencia de datos. La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en varios niveles de abstracción[29].

Esta metodología consta de seis etapas: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelamiento, evaluación e implementación. La última etapa de esta metodología quedará fuera del alcance de esta memoria, por lo que se abordarán principalmente las cinco primeras etapas.

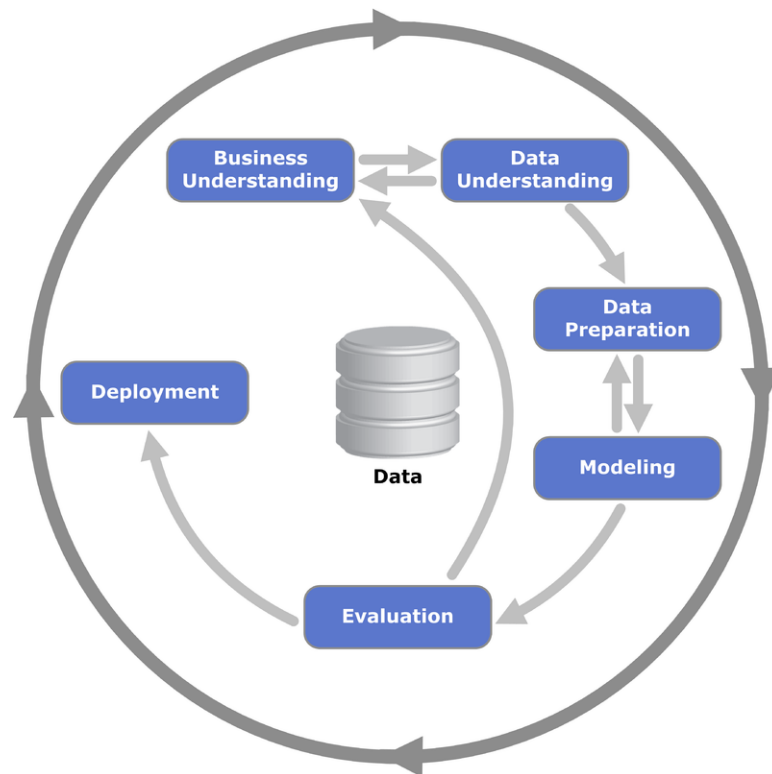


Figura 3.1: Metodología CRISP-DM. Fuente: Health Data Miner.

Capítulo 4

Desarrollo Metodológico

4.1. Comprensión del negocio

En esta sección se aborda el contexto del negocio, con el objetivo de dar sentido a los datos que se analizarán más adelante. Se revisa el concepto de estandarización de direcciones, y de dónde provienen los datos con los que se trabajan.

4.1.1. Contexto del negocio

Las empresas de despacho son una pieza clave dentro del comercio global y la logística. Estas compañías, que operan tanto a nivel nacional como internacional, se especializan en la entrega eficiente y precisa de una amplia gama de mercancías. El funcionamiento de una empresa que entrega servicios de despacho se basa en una red coordinada de actividades que incluyen la recolección, el almacenamiento, la gestión de inventarios, el transporte y la entrega final de productos a los consumidores o negocios. En su núcleo, las empresas de despacho dependen de sistemas logísticos y tecnología de la información para rastrear y gestionar los envíos a través de rutas y medios de transporte variados, como carretera, aire, entre otros. La eficiencia en este proceso no solo se mide en términos de velocidad, sino también en la precisión y fiabilidad en la entrega, lo cual es fundamental para mantener la confianza y satisfacción del cliente. La gestión eficaz de estas operaciones requiere de una coordinación impecable entre diferentes áreas y, la colaboración con socios externos como agentes de carga y proveedores de logística.

En este ecosistema, las direcciones precisas y su normalización son fundamentales para evitar retrasos, pérdidas y errores, lo que a su vez afecta la optimización de costos y la eficiencia operativa. Por lo tanto, las empresas de despacho continuamente buscan mejorar sus métodos y sistemas para garantizar una entrega rápida, segura y precisa, adaptándose a las demandas cambiantes del mercado y las expectativas de los clientes en un mundo cada vez más conectado.

La normalización de direcciones es particularmente importante para las empresas de envío, ya que permite una asignación de ruta y seguimiento precisos de los paquetes. Garantiza que se identifique el destino correcto, lo que reduce el riesgo de envíos perdidos o mal dirigidos. Además, la normalización de direcciones permite a las empresas de despacho optimizar sus rutas de entrega, ahorrando tiempo y costos de combustible. Al utilizar métodos de normalización de direcciones eficaces, las empresas de envío pueden garantizar que las direcciones

tengan el formato correcto y estén libres de errores o inconsistencias. Esto no sólo mejora la eficiencia operativa sino que también mejora la satisfacción del cliente al minimizar los errores de entrega y proporcionar información de seguimiento precisa.

No se puede subestimar la importancia de direcciones precisas en la industria de los envíos, ya que son cruciales para operaciones logísticas exitosas y eficientes.

En esta industria, varios agentes desempeñan papeles importantes para garantizar un servicio eficiente y confiable. En primer lugar, están los operadores de entrada de datos (por ejemplo, quienes reciben pedidos en sucursales) y personal de atención al cliente, quienes son responsables de registrar con precisión las direcciones y detalles del envío. Los coordinadores de logística y planificadores de rutas juegan un papel fundamental en la asignación de rutas de entrega, pretendiendo que las mercancías se transporten de la manera más eficiente posible. Los conductores y repartidores son la cara visible, encargados de la entrega física de los paquetes a los destinatarios finales. Los analistas de datos y los equipos de TI se encargan del desarrollo y mantenimiento de sistemas de seguimiento y gestión de direcciones, velando por la precisión y la eficiencia en el procesamiento de los envíos. Los encargados de puntos y personal de manipulación de carga aseguran que los paquetes se almacenen, clasifiquen y manejen adecuadamente antes y después del transporte. Los equipos de ventas y marketing son importantes para entender las necesidades del cliente y promocionar los servicios de la empresa, mientras que los equipos de gestión estratégica y finanzas aseguran la viabilidad comercial y la sostenibilidad de las operaciones a largo plazo. Todos estos agentes (que se consideran como los más relevantes), trabajan en conjunto con los demás trabajadores para garantizar que la empresa de despacho pueda ofrecer servicios rápidos, seguros y precisos, manteniendo la satisfacción y confianza del cliente.

4.1.2. Estandarización de direcciones

El proceso de *Address Standardization* (estandarización de una dirección), es un proceso en donde se prepara la dirección en un formato conocido corrigiendo los errores de escritura para estructurar y especificar una forma normalizada de escribir la dirección[30]. Este procedimiento se torna indispensable en cualquier empresa dedicada a la distribución, siendo su relevancia evidente al asignar una ruta a un pedido próximo a ser despachado.

4.1.3. Mercado y normalización de direcciones

El mercado objetivo de la empresa de despacho se segmenta en tres categorías fundamentales: clientes empresariales, clientes individuales e industrias específicas. Cada una de estas categorías representa una fuente particular de obtención de datos, especialmente en lo que respecta a direcciones.

En determinadas circunstancias, las plataformas de comercio electrónico requieren que los compradores ingresen de manera organizada y separada los datos incluidos en una dirección, tales como el nombre de la calle, el número, la comuna y la información adicional. Además, algunas empresas cuentan con integraciones con aplicaciones como Google Maps. En estos casos, al intentar ingresar una dirección, se despliega una lista con las direcciones válidas asociadas a la aplicación, lo que obliga al cliente a seleccionar una dirección válida en el

mapa en lugar de ingresar información no validada. Por otro lado, existen comercios que, al completar los datos de despacho, simplemente solicitan el llenado de los campos obligatorios sin validar la dirección mediante un mapa. Esta práctica aumenta la probabilidad de que la dirección ingresada manualmente contenga errores ortográficos, asociaciones incorrectas de comunas, entre otros. La falta de rigurosidad en estos casos, sumado a que en la empresa no existe un nivel de exigencia que permita obtener datos limpios, se genera la necesidad de contar con un eficiente normalizador de direcciones.

4.1.4. Fecha de promesa

La fecha de promesa es el día específico en el que una empresa de despachos se compromete a entregar un paquete al cliente. Este compromiso se basa en distintos factores, incluyendo el tipo de servicio de envío, la distancia entre el origen y el destino del paquete y la capacidad operativa de la empresa. La fecha de promesa es clave en el nivel de servicio, ya que establece la expectativa para el cliente sobre cuándo recibirá su envío. Cumplir con la fecha de promesa significa entregar el pedido antes de o el mismo día prometido.

La capacidad de cumplir con la fecha de promesa es crucial para una empresa de despachos, ya que determina la confianza del cliente en la empresa, mientras que el incumplimiento puede llevar a la insatisfacción y pérdida de clientes.

4.1.5. Nivel de servicio

El nivel de servicio, en el contexto de una empresa de despachos, se define por la capacidad de cumplir con la entrega de órdenes dentro de las fechas de promesa acordadas con los clientes. Este criterio es fundamental para medir la eficacia y eficiencia de la empresa en su proceso logístico y en las expectativas de sus clientes.

Por un lado, el cumplimiento exitoso del nivel de servicio se evidencia cuando las órdenes son entregadas antes o durante la fecha prometida, garantizando la satisfacción del cliente y manteniendo el compromiso de entrega. Por otro lado, el incumplimiento de este criterio, que se refiere internamente como “fallo”, se produce cuando una orden no se entrega dentro del lapso temporal prometido. Esto se utiliza como un indicador crítico de rendimiento para identificar áreas de mejora. La empresa busca priorizar la minimización de estos fallos.

En un mercado competitivo, es esencial mantener un alto nivel de servicio para las empresas de despachos, ya que esto influye directamente en la retención de los clientes.

4.1.6. Impacto de las excepciones en el nivel de servicio

En primer lugar, cabe destacar que el fallo en el nivel de servicio no se debe solamente a las excepciones de mal ruteo. El fallo en el nivel de servicio indica la proporción de órdenes que no se entregaron en la fecha comprometida inicialmente al cliente, lo cual se puede deber no solo a un error relacionado a la dirección, sino a otros motivos como errores en la preparación de pedidos, gestión ineficiente de inventarios, capacidad de procesamiento limitada, entre otros. Las órdenes que corresponden al 5% de fallo en el nivel de servicio, tienen asociados distintos tipos de excepción: *mal ruteo (MR y MRH)*, *dirección inubicable (DI)*, *faltan datos (DA)* y *otros fallos*. En la empresa estudiada, solamente los pinchazos de excepción de mal

ruteo (MR y MRH) se consideran como fallo de georreferenciación, por lo que son los únicos de interés para este trabajo.

La Tabla 4.1 se construyó con cifras reales de la empresa objeto de estudio, recopiladas de las semanas 1 a la 19 del 2023. Se muestran las órdenes de servicio que tuvieron los pinchazos de excepción de mal ruteo (MR y MRH). Durante dicho periodo, que contempló 92 días hábiles, la demanda total fue de 2.759.629 órdenes de servicio, con un porcentaje de cumplimiento de un 95 %, y un 5 % de fallo. Es decir, aproximadamente 138 mil órdenes no cumplieron con la fecha de promesa de entrega.

En total, 41.286 órdenes tuvieron pinchazo de excepción de mal ruteo, representando aproximadamente un 1,5 % del total de órdenes. Cabe mencionar que, el hecho de que una orden tenga una excepción de mal ruteo no implica que no se cumpla con la fecha de promesa de entrega, y por ende, que afecte el nivel de servicio. En la tabla que se muestra a continuación, los porcentajes de cumplimiento y no cumplimiento (columnas *% Cumplimiento* y *% No Cumplimiento*), hacen referencia al porcentaje de órdenes con excepción de mal ruteo, que tuvieron excepción, y cumplieron (o no) con la fecha de promesa.

Tabla 4.1: Órdenes de servicio con pinchazos de excepción de mal ruteo durante las semanas 1-19 (2023). Fuente: Elaboración propia.

Excepción	Q OS	% de total OS	% Cumplimiento	% No Cumplimiento
MR	11.283	0,408859 %	61 %	39 %
MRH	30.003	1,087211 %	57 %	43 %

Asumiendo el mismo comportamiento de cumplimiento y no cumplimiento, se genera la Tabla 4.2, que muestra un pronóstico de los fallos que se generarían con una demanda de 40.000 órdenes de servicio diarias (la cual se proyecta prontamente alcanzar), considerando los datos extraídos de las semanas 1-19.

Tabla 4.2: Proyección de fallo diario debido a excepción de mal ruteo en órdenes de servicio, considerando una demanda diaria estimada de 40.000 unidades. Fuente: Elaboración propia.

Fallos Geo		Cumplimiento NDS		Fallo NDS		
Tipo	Q OS	Porcentaje	Q OS	Porcentaje	Q OS	Impacto NDS
MR	164	61 %	100	39 %	64	0,16 %
MRH	435	57 %	248	43 %	187	0,47 %

Para cuantificar el impacto en el nivel de servicio, se considera el porcentaje de órdenes con excepción de mal ruteo que efectivamente fallaron en la fecha de promesa, con respecto al total de órdenes diarias.

Una orden de servicio con excepción de mal ruteo podría tener, además, un pinchazo de BA (faltan datos) o DI (dirección incompleta). Si bien no se profundiza en ellos por quedar fuera del alcance de este trabajo, es relevante mencionar lo anterior, ya que, si se logran

corregir todos los MR y MRH de esa semana, el aumento en el nivel de servicio será como máximo de un 0,63 % ($0,16\% + 0,47\% = 0,63\%$), ya que la orden podría estar marcada, además, con una excepción del tipo DI o BA.

4.2. Comprensión de los datos

En esta fase, se recopilan los datos necesarios para llevar a cabo el proyecto y se realiza una exploración de los mismos.

4.2.1. Exploración inicial de los datos

Inicialmente, se explora la base de la empresa *reporte-mrh*, que contiene los datos de las órdenes de servicio que tuvieron pinchazo de excepción de mal ruteo. El análisis se realiza para los días entre el 19 de marzo y el 19 de abril, para las comunas de Santiago. En la figura a continuación se presenta la cantidad de MRH (mal ruteo), que hubo en cada comuna durante las fechas mencionadas anteriormente. En la columna *Cuenta de OS* figura la cantidad órdenes de servicio distintas. La comuna que destaca con una marcada incidencia de problemas en las rutas es, indiscutiblemente, Santiago.

Tabla 4.3: Distribución de Incidencias MRH por Comunas de Santiago entre el 19/03/2023 y el 19/04/2023. Fuente: Elaboración propia.

Comuna declarada por el destinatario	Cuenta de OS	Repetidos
BUIN	9	2
CALERA DE TANGO	1	4
CERRILLOS	8	3
CERRO NAVIA	1	1
COLINA	16	2
CONCHALI	1	1
EL BOSQUE	4	2
HUECHURABA	18	2
ISLA DE MAIPO	4	2
LA FLORIDA	27	2
LA PINTANA	9	1
LA REINA	6	1
LAMPA	5	1
LAS CONDES	64	1
LO ESPEJO	11	1
MACUL	2	2
MAIPU	27	1
NUNOA	17	1
PADRE HURTADO	2	1
PAINE	5	2
PENAFLORE	1	1
PENALOLEN	20	2
PROVIDENCIA	9	3
PUDAHUEL	3	1
PUENTE ALTO	12	1
QUILICURA	6	2
RECOLETA	10	1
RENCA	5	1
SAN BERNARDO	54	1
SAN JOAQUIN	2	2
SAN JOSE DE MAIPO	21	1
SAN MIGUEL	3	2
SAN RAMON	1	1
SANTIAGO	569	1
TILTIL	1	1
VITACURA	6	1
Total	960	1

No obstante, es esencial abordar un importante sesgo que afecta a esta estadística. La Tabla 4.3 incluye información sobre la comuna en la que declara residir el destinatario, y en el caso particular de Santiago, surgen inconvenientes significativos. El principal desafío radica en que el nombre de la comuna es idéntico al de la ciudad, lo que conduce a que muchas personas, al declarar su comuna, simplemente escriban “Santiago” sin especificar la comuna exacta a la que pertenecen. Esto se debe a que, aunque residen en la ciudad de Santiago, no necesariamente viven en la comuna de Santiago.

Entre las columnas que contiene la base estudiada se tienen distintos datos: dirección ingresada por el destinatario, comuna declarada por el destinatario, comuna a la cual el proveedor de georreferenciación asoció la dirección, coordenadas de latitud y longitud del lugar correspondiente, entre otros. Se identificaron patrones comunes y anomalías en las direcciones, como la frecuencia de errores ortográficos y la variabilidad en la estructura de las mismas.

4.2.2. Errores identificados en la declaración de dirección

En el ámbito de una empresa cuya operativa involucra la gestión de direcciones, la precisión en la declaración de éstas por parte de los destinatarios es fundamental. Sin embargo, se identifican diversos tipos de errores en estas declaraciones. Un error común se relaciona con la incorrecta especificación de la comuna: un destinatario puede errar al declarar la comuna correspondiente a su dirección, ya sea por descuido, desconocimiento u otros factores. Este error es particularmente problemático cuando una calle con el mismo nombre y numeración existe en varias comunas, lo cual puede llevar a confusiones significativas.

Además, existe la posibilidad de que, a pesar de que el destinatario indique correctamente la comuna, el sistema de georreferenciación empleado por la empresa no logre localizar adecuadamente la dirección. Esto sucede cuando la API de geolocalización asigna erróneamente las coordenadas de latitud y longitud, resultando en una comuna incorrectamente asociada a la dirección especificada.

Otro escenario problemático surge cuando el destinatario omite la comuna en su declaración. Esta omisión incrementa sustancialmente las probabilidades de asignaciones erróneas de coordenadas por parte del sistema de georreferenciación, pero es menos frecuente, ya que en la mayoría de las tiendas online es necesario llenar el campo de comuna de forma obligatoria.

Respecto a los componentes específicos de una dirección, como el nombre de la calle y la numeración, se encuentran múltiples fuentes de error. Estos pueden incluir faltas de ortografía, el uso de nombres no oficiales para calles o avenidas (por ejemplo, referirse a “Alameda” en lugar de “Avenida Libertador Bernardo O’Higgins”), la inclusión de intersecciones en lugar de la dirección exacta, o la mención de lugares genéricos en vez de direcciones específicas. En la Tabla 4.4 se muestra un caso para los últimos tres ejemplos mencionados.

Tabla 4.4: Ejemplos de direcciones declaradas de forma errónea. Fuente: Elaboración propia.

Problema en la dirección declarada	Dirección declarada por destinatario
Corresponde a un lugar	Correos Chile Sucursal La Florida Casilla 549
No corresponde al nombre oficial	Alameda 776 51 Dep51
Se declara una intersección de calles	Lo Martínez con Pasaje Tirandente 12333, 35 de Santa Rosa Porto Alegre

Otro error frecuente se presenta cuando los destinatarios proporcionan solamente el nombre de la calle sin la numeración correspondiente. Esta omisión complica significativamente la tarea de georreferenciación precisa. Además, se han identificado casos en los que los destinatarios incluyen información adicional irrelevante o confusa, como referencias a puntos de referencia locales o instrucciones de acceso, que no contribuyen a la precisión de la localización. También se observan inconsistencias en la estructura de las direcciones, como la inversión del orden de los elementos o la inclusión de abreviaturas no estandarizadas.

En suma, la variedad y complejidad de estos errores subrayan la importancia de implementar sistemas robustos y sofisticados de normalización y verificación de direcciones para

minimizar los impactos negativos en la logística y la satisfacción del cliente.

4.2.3. Problemas asociados a la normalización

Anteriormente, se identificaron diversas problemáticas asociadas a la normalización de direcciones, no atribuibles directamente a errores en el proceso de normalización en sí, sino a otras causas. A continuación, se presentan ejemplos concretos que ilustran errores en la normalización de direcciones, específicamente aquellos derivados de la pérdida de información crítica durante el proceso de normalización empleado actualmente.

Tabla 4.5: Ejemplos de errores en la normalización de direcciones. Fuente: Elaboración propia.

Dirección declarada por destinatario	Comuna declarada por destinatario	Normalización de la empresa	Comuna asociada por la empresa
FRAY CAMILO HENRIQUEZ 686 DEPTO 1711	SANTIAGO	FRAY+CAMILO+686++1711	PUENTE ALTO
AVENIDA INDEPENDENCIA 1050 612	SANTIAGO	INDEPENDENCIA ++1050	PUENTE ALTO

Es fundamental entender que la primera acción al registrar una nueva dirección en el sistema de la empresa es su normalización. Se ha observado que este proceso frecuentemente conlleva a la pérdida de elementos esenciales para una georreferenciación precisa. En los ejemplos de la Tabla 4.5, el mecanismo de programación actual omite información importante. Por ejemplo, en el primer caso, la dirección declarada “FRAY CAMILO HENRIQUEZ” en Santiago es erróneamente normalizada a “FRAY CAMILO”, una calle existente en Puente Alto.

Esta misma tendencia se repite en otros casos, como el segundo ejemplo, donde la omisión de la palabra “AVENIDA” en la normalización resulta en una dirección correspondiente a una comuna diferente. La columna *Comuna asociada por la empresa* refleja la comuna a la que pertenecen las coordenadas de latitud y longitud asignadas por la API geocode, evidenciando así las discrepancias generadas por el proceso de normalización.

Como se ha señalado previamente en la Sección 2.2.3, el proceso de normalización actual procede a la eliminación de caracteres de las direcciones, algunos de los cuales a menudo contienen información significativa para una georreferenciación precisa de dichas direcciones. Entre estos caracteres figuran términos como “AV”, “AVENIDA”, “COND”, y otros semejantes. Cuando estos términos aparecen como parte de una palabra en la dirección, la palabra en su totalidad es suprimida durante el proceso de normalización.

4.3. Preparación de los datos

En esta fase, se limpian y preparan los datos para su uso en la siguiente etapa. Se seleccionan las variables relevantes y se transforman los datos.

Se ha mencionado previamente que una misma dirección se puede escribir de varias formas, omitiendo u abreviando palabras, con más o menos información, utilizando variaciones en la estructura o formato de la dirección, con la aparición de errores tipográficos, ortográficos o gramaticales y con la repetición de palabras.[26]

La idea principal de este trabajo consiste, en palabras sencillas, en plantear un modelo que pueda entregar una dirección normalizada, a partir de la dirección entregada por un cliente, sin que se pierda información relevante, y corregir los errores ortográficos.

4.3.1. Construcción de la base de datos

Dado que se dispone de un periodo de tiempo acotado, para simplificar la complejidad se limita a trabajar con bases de datos que contienen solamente los nombres de calles/avenidas/pasajes u otro, excluyendo la numeración y la comuna. Inicialmente, se elaboró una base de datos simulada, incorporando nombres de calles correspondientes a una comuna específica. La creación de esta base de datos tenía un doble propósito. Primero, facilitar la comprensión mediante ejemplos y tablas en las etapas de esta sección, considerando la naturaleza confidencial de los datos reales de la empresa. La utilización de información ficticia permitió realizar una demostración más clara y comprensible de los procesos involucrados. Segundo, el entrenamiento de modelos de aprendizaje automático exige una cantidad considerable de datos. Considerando que se trabajó con 1.745 datos, el uso de direcciones provenientes de diversas comunas (como es el caso de las direcciones con malos ruteos) complica significativamente el proceso de aprendizaje de la máquina, una consideración particularmente relevante dada la limitación temporal del proyecto.

En primera instancia, se decide trabajar con la comuna de Cerrillos, dado que en esa comuna hay casos de calles que contienen los caracteres que se mencionaron en secciones anteriores (y que se eliminan en el proceso de normalización actual), como *COND*, *ESQ*, *INTERIOR*, entre otros.

Según información extraída de *código postal*[31], existen 218 calles en la comuna de Cerrillos. La lista de calles se encuentra en Anexo B.

Para entrenar el modelo de aprendizaje supervisado, se construye la base de datos en un archivo Excel, que cuenta con dos columnas. La primera columna contiene la dirección ingresada, mientras que la segunda columna contiene la etiqueta, que es la normalización correcta de la dirección que se ingresó como input.

La siguiente tabla ejemplifica cómo se ven algunas filas de la base de datos del modelo. En la primera columna, se tiene el input (dirección “ingresada”) y en la segunda columna el output, que correspondería a la etiqueta o “dirección normalizada”. Se muestran ejemplos de pares de direcciones ingresadas y sus correspondientes normalizaciones, tal como se han registrado en la base de datos construida para entrenar el modelo propuesto. Cada fila representa un caso de estudio, donde se incluye tanto la dirección tal como fue ingresada por el usuario (con posibles errores o variantes) como la dirección normalizada. Este extracto ilustra la diversidad y los tipos de correcciones que el modelo está diseñado para aprender y aplicar.

Tabla 4.6: Extracto de la base de datos para entrenamiento del modelo de aprendizaje supervisado. Fuente: Elaboración propia.

direccion_ingresada	direccion_normalizada
CONDE DE ORGAZ	CONDE DE ORGAZ
5 DE ABRIL	5 DE ABRIL
ONDURAS	HONDURAS
ACCESO CAMINO A LONQUEN,	ACCESO CAMINO A LONQUEN
NUEVA LAREDO	NUEVA LAREDO
SAN JOSE	SAN JOSE
AMASONAS	AMAZONAS
EL ALAMO	EL ALAMO

En los datos se elimina el prefijo de la dirección, mostrando solamente el nombre de la calle/pasaje/avenida. Además de simplificar el modelo, esto se hace para que posteriormente los resultados obtenidos sean comparables frente al modelo de normalización de la empresa, que, como se mencionó anteriormente, elimina dichos prefijos. De esta manera, habiendo eliminado esos casos (en que el método tradicional elimina esos sufijos), se podrá comparar con el modelo propuesto en cuanto a si se pierde o no información relevante.

Con el fin de evaluar la data real, se construyó una segunda base, esta vez compuesta por datos reales de la empresa, que proporcionó 1.660 direcciones asociadas a errores de ruteo. Estos datos provienen de una variedad de comunas y presentan una mayor diversidad en los nombres en comparación con la base de datos ficticia. Esta ampliación en la variedad de nombres conlleva una expansión en el espectro de datos a analizar. Consecuentemente, se anticipa que la precisión del modelo propuesto disminuirá al ser evaluado con esta base de datos, debido a su mayor complejidad y variabilidad.

4.4. Modelamiento

Para el desarrollo del modelo, se optó por una estrategia que integra las capacidades robustas de clasificación de las Máquinas de Soporte Vectorial (SVM) con las técnicas avanzadas de procesamiento de lenguaje natural (NLP), utilizando un kernel lineal. Esta elección se basó en la naturaleza intrínsecamente lingüística y no estructurada de las direcciones de texto. Mediante el uso de TF-IDF (Term Frequency-Inverse Document Frequency) dentro del NLP, se transformaron eficientemente estas direcciones en vectores numéricos, destacando la importancia relativa de las palabras, y reduciendo el ruido de términos comunes. Este paso es fundamental, ya que los modelos de aprendizaje automático como los SVC requieren que los datos estén en formato numérico para su procesamiento.

El algoritmo SVM, basado en el principio de minimización del riesgo estructural, encuentra un equilibrio entre la complejidad del modelado y la capacidad de aprendizaje mediante la extracción de información de muestra (Shi et al, 2023). Utilizar un método de aprendizaje supervisado con SVM simplifica el proceso de formación al convertirlo en un conjunto de ecuaciones lineales, evitando la programación cuadrática y proporcionando varias ventajas sobre los métodos tradicionales de aprendizaje automático. Una de estas ventajas es la capacidad del SVM para manejar valores atípicos, lo que lo hace resistente al ruido y mejora su

rendimiento en tareas de normalización de direcciones. Esta resiliencia al ruido es particularmente útil cuando se trata de datos del mundo real, donde las variaciones e inconsistencias son comunes.

4.4.1. Justificación de la elección del algoritmo

Existen múltiples estudios que demuestran la eficacia de las máquinas de vectores de soporte en tareas de normalización de direcciones. La elección de un modelo supervisado, específicamente de máquinas de vectores de soporte para la normalización de direcciones, se justifica en función de los siguientes factores, estudiados en el trabajo de Bruzzone y Persello[32]:

- Alta precisión empírica y capacidad de generalización: las máquinas de vectores de soporte logran una alta precisión en las tareas de clasificación, lo que las hace adecuadas para la normalización de direcciones, donde la clasificación precisa es esencial.
- Robustez ante el fenómeno de Hughes: las máquinas de vectores de soporte ofrecen robustez ante valores atípicos y datos ruidosos, lo cual es particularmente importante en la normalización de direcciones, ya que es común encontrar variaciones y errores en la entrada de direcciones.
- Convexidad de la función de costos: la función de costos utilizada en las máquinas de vectores de soporte es convexa, lo cual garantiza que el modelo encuentre la solución óptima global en lugar de quedarse atascado en los óptimos locales, lo que lleva a una normalización de direcciones más precisa.
- Dispersión de la solución: las máquinas de vectores de soporte tienden a utilizar un pequeño subconjunto de puntos de datos de entrenamiento (vectores de soporte), lo que da como resultado un modelo más eficiente e interpretable para la normalización de direcciones, lo cual es ventajoso en términos de eficiencia computacional y reducción de la complejidad.
- Capacidad para abordar problemas no lineales con el truco del núcleo: las máquinas de vectores de soporte pueden manejar relaciones complejas y no lineales entre características de entrada y etiquetas de salida mediante el uso de funciones del núcleo, como la función de base radial o núcleos polinomiales. Esto permite que las máquinas de vectores de soporte manejen de manera efectiva la naturaleza no lineal de la normalización de direcciones, donde los patrones y las relaciones entre los componentes de las direcciones pueden variar mucho.

En el trabajo “*Un estudio comparativo de la normalización de direcciones utilizando máquinas de vectores de soporte y redes neuronales artificiales*”[33], que comparó las SVM y las redes neuronales artificiales para la normalización de direcciones, se descubrió que las SVM lograron mayor precisión en la clasificación y normalización de direcciones en comparación con las redes neuronales artificiales.

El estudio titulado “*Selección óptima de parámetros para máquinas de vectores de soporte*”[34], realizado por Schittkowski, se centró en el uso de SVM para tareas de clasificación supervisadas. Este estudio demostró las ventajas de este algoritmo en el manejo de datos

linealmente separables y no separables. También se destacaron mejoras y extensiones recientes a la formulación original, lo que indica la versatilidad y adaptabilidad de las SVM para diferentes tareas de clasificación.

Integración con técnicas de NLP

La eficacia de las SVM en la normalización de direcciones se ve potenciada significativamente cuando se combinan con técnicas de procesamiento de lenguaje natural (NLP). El uso de métodos de NLP como TF-IDF para la extracción de características permite transformar direcciones de texto en vectores numéricos manejables, optimizando así el rendimiento de las SVM en la clasificación de estas direcciones normalizadas.

Las técnicas de NLP facilitan el manejo de la complejidad y la variabilidad inherentes a los datos textuales de las direcciones. Al aplicar NLP, se puede extraer información valiosa de los patrones lingüísticos y estructurales presentes en las direcciones, lo cual es crucial para una clasificación precisa y efectiva mediante las SVM.

La combinación de NLP y SVM es particularmente efectiva en el contexto de las direcciones, un tipo de dato que a menudo varía en formato y estructura. Esta adaptabilidad es esencial para abordar con éxito la normalización de direcciones de distintas regiones y formatos.

4.4.2. Modelo propuesto

4.4.2.1. Descripción general del modelo propuesto

El modelo de normalización de direcciones propuesto utiliza principalmente un algoritmo de SVM, con técnicas de TF-IDF y kernel lineal, algoritmos clave del ámbito del procesamiento de lenguaje natural (NLP) y del aprendizaje automático supervisado.

El código desarrollado recibe las direcciones ingresadas (que pueden variar en formato, contener errores, ser incompletas, etc.), las transforma en una representación vectorial mediante TF-IDF, y luego utiliza un modelo de SVM entrenado para predecir y entregar la dirección en un formato normalizado. Así, se puede explicar dividiendo el código en cuatro etapas:

1. Vectorización (TF-IDF): El código comienza transformando las direcciones ingresadas en representaciones vectoriales mediante el uso de TF-IDF. Esto permite convertir el texto de las direcciones en un conjunto de vectores numéricos, donde cada número representa la importancia de una palabra dentro de una dirección particular en comparación con un conjunto de direcciones. Lo anterior prepara los datos para ser procesados por el modelo SVM.

2. Modelo SVM para normalización: Se entrena un modelo SVM con estos vectores como entrada (X) y las versiones normalizadas de las direcciones como las etiquetas objetivo (y). El entrenamiento ajusta los parámetros del modelo, de manera que aprende a asociar cada vector, que representa una dirección ingresada, con su correspondiente dirección normalizada.

3. Predicción: Una vez entrenado, el modelo puede tomar la representación vectorial de una nueva dirección ingresada, que no ha sido vista previamente, y predecir su forma normalizada. Esto no corresponde a una clasificación binaria (de si la dirección está limpia o sucia), sino un mapeo de la dirección ingresada a su versión normalizada.

4. Entrega de la dirección normalizada: El resultado final del código es una dirección normalizada para cada dirección ingresada. El modelo utiliza las representaciones vectoriales de las direcciones, para predecir cómo debería ser su normalización, basándose en los ejemplos que aprendió durante el entrenamiento.

4.4.2.2. Bases de datos

El modelo propuesto se puso a prueba utilizando dos conjuntos de datos distintos: uno ficticio y uno real, ambos conteniendo 1.745 registros. La base de datos ficticia se construyó con dos objetivos. En primer lugar, con el fin de hacer más explicativo el presente capítulo, mientras se respeta la confidencialidad de los datos de la empresa. En segundo lugar, se diseñó utilizando las calles de una comuna específica, en contraste con la base de datos real con la que se contaba, ya que en ella las direcciones eran mucho más variadas, provenientes de diversas comunas, y con una menor repetición de nombres de calles. Esta característica de la base de datos real presentaba un desafío para el aprendizaje efectivo del modelo, debido a su mayor heterogeneidad y la amplia dispersión geográfica de las direcciones. La construcción de una base de datos ficticia con calles de una sola comuna permitió una mayor consistencia en los datos, facilitando el aprendizaje supervisado del modelo y optimizando su capacidad

para generalizar, a partir de un conjunto de datos más controlado y homogéneo. Al emplear la base de datos real, se enfrentó al modelo a un escenario más complejo y diversificado.

4.4.2.3. Código

El modelo propuesto es de aprendizaje supervisado, ya que se entrenó utilizando direcciones etiquetadas. Para construir este modelo se utiliza el lenguaje de Python.

En un principio, se importan las librerías necesarias, incluyendo *pandas* para manejar el conjunto de datos, *TfidfVectorizer* para convertir las direcciones en vectores numéricos, *SVC* para el clasificador SVM, *train_test_split* para dividir los datos en conjuntos de entrenamiento y prueba, y finalmente *sklearn* y *distance*, para el cálculo de métricas del modelo. Se carga la base de datos, que contiene las dos columnas mencionadas anteriormente: *direccion_ingresada* y *direccion_normalizada*.

```
[ ] #Importar librerías
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split

[ ] #Cargar archivos en entorno de colab
from google.colab import files
uploaded = files.upload()

[ ] #Leer la base de datos
df = pd.read_excel('cerrillos.xlsx')
```

Figura 4.1: Fragmento del código para la normalización de direcciones. Fuente: Elaboración propia.

Las columnas de la base de datos se separan en variables dependientes e independientes. Las direcciones ingresadas (X) son las variables independientes, mientras que las direcciones normalizadas (y) son las variables dependientes.

A continuación, se crea el vectorizador de palabras (TF-IDF). Esto ayuda a identificar características clave que diferencian una dirección de otra, facilitando al modelo SVM la tarea de clasificación. Así, se utiliza *TfidfVectorizer* para convertir las direcciones en vectores numéricos, considerando la importancia de cada palabra en relación con todas las direcciones. Luego, se aplica el vectorizador a las direcciones ingresadas para obtener vectores numéricos.

```
[ ] #Separar columnas en variable dependiente e independiente
X = df['direccion_ingresada']
y = df['direccion_normalizada']

[ ] #Crear el vectorizador de palabras para transformar las direcciones en vectores
#numéricos con pesos según su frecuencia e importancia
vectorizer.fit(X)
vectorizer = TfidfVectorizer()

[ ] #Ajustar el vectorizador con el vocabulario de las direcciones ingresadas
X = X.fillna(' ')
vectorizer.fit(X)

TfidfVectorizer
TfidfVectorizer()

[ ] #Transformar las direcciones ingresadas en vectores numéricos
X_vect = vectorizer.transform(X)
```

Figura 4.2: Fragmento del código para la normalización de direcciones. Fuente: Elaboración propia.

Se divide la base de datos en conjuntos de entrenamiento y prueba, dedicando un 80% para entrenamiento y un 20% para prueba. Esto se realiza mediante *train_test_split*, función de Python que permite hacer una división de un conjunto de datos en dos bloques, de entrenamiento y prueba de un modelo.

Luego, se crea un clasificador SVM con un kernel lineal, según el enfoque lineal para la clasificación. El clasificador se entrena luego con los vectores numéricos y las direcciones normalizadas correspondientes.

Finalmente, se emplea el modelo entrenado para predecir las direcciones normalizadas en el conjunto de prueba.

```
[ ] #Dividir los datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_vect, y, train_size=0.8, test_size=0.2, random_state=42)

[ ] #Crear un clasificador SVM con kernel lineal
svm = SVC(kernel='linear')

[ ] #Entrenar el clasificador con los vectores numéricos y las direcciones normalizadas
svm.fit(X_train, y_train)

SVC
SVC(kernel='linear')

[ ] #Predecir las direcciones normalizadas para un conjunto de prueba
y_pred = svm.predict(X_test)
```

Figura 4.3: Fragmento del código en Python para la normalización de direcciones. Fuente: Elaboración propia.

Se importan las librerías necesarias para el cálculo de métricas clave para evaluar el rendimiento del modelo. Se calculan las métricas de precisión en la clasificación, la capacidad de identificar correctamente direcciones relevantes (recall), el balance entre precisión y recall

(F1 score), la similitud entre las predicciones y las etiquetas verdaderas (Jaccard score) y la distancia promedio de Levenshtein entre las predicciones y las verdaderas direcciones.

```
[ ] # Se importan las librerías necesarias para el cálculo de métricas
!pip install distance
from sklearn.metrics import precision_recall_fscore_support
from sklearn.metrics import jaccard_score
from distance import levenshtein

# Calcular las métricas de recall, F1 score y precision
precision, recall, f1, _ = precision_recall_fscore_support(y_test, y_pred, average='weighted')
print(f'Precision: {precision}')
print(f'Recall: {recall}')
print(f'F1 Score: {f1}')

# Calcular la métrica de Jaccard
jaccard = jaccard_score(y_test, y_pred, average='weighted')
print(f'Jaccard Score: {jaccard}')

# Calcular la distancia de Levenshtein promedio
distancias = [levenshtein(pred, true) for pred, true in zip(y_pred, y_test)]
distancia_promedio = sum(distancias) / len(distancias)
print(f'Distancia de Levenshtein Promedio: {distancia_promedio}')
```

Figura 4.4: Fragmento del código en Python para la normalización de direcciones (métricas de evaluación). Fuente: Elaboración propia.

A continuación, se ilustra cómo se presentan las predicciones del modelo SVM. Se imprimen las direcciones normalizadas predichas, permitiendo una evaluación inmediata y visual de la salida del modelo.

```
[22] #Imprimir las predicciones
print(y_pred)

['ARMONIA' 'VISTA HERMOSA' 'MEXICO' 'ELICURA' 'DIPUTADO ANGEL FANTUZZI'
'MANUTARA' 'LA HERRADURA' 'AEROPUERTO' 'MORELOS' 'CARRIEL SUR'
'DIR MARIA TRONCOSO CARRASCO' 'CAMINO A MELIPILLA' 'FELIX MARGOZ'
'RIO LOA' 'RIO MAYO' 'CULTURA MOLLE' 'BRASIL' 'CHAMIZA' 'EL PINAR'
'RIO CLARO' 'LOS FUNDADORES' 'PILOTO ACEVEDO' 'MARGARITA GARCIA TORO'
'PRINCIPAL' '5 DE ABRIL' 'LOS GERANIOS' 'AEROPUERTO'
'MONSEÑOR JOSE MARIA CARO' 'CALLE NUEVA LAREDO' 'LOS CERRILLOS' 'MORELOS'
'LAS GALAXIAS' 'HUELEN' 'HECTOR ORREGO' 'MANUTARA' 'RIO CLARO'
'LOS AZTECAS' 'RAPA NUI' '29 DE JUNIO' 'CALLE UNO' 'MEXICO' 'GUAYANA'
'REINOSA' 'AMAZONAS' 'LOS FUNDADORES' 'DEL FERROCARRIL'
'AV CENTRAL CARDENAL RAUL SILVA HENRIQUEZ' 'ALUMINIO'
'PEDRO AGUIRRE CERDA INTERIOR' 'ERNESTO CEA' 'PILOTO JACQUES LAGAS'
'LOS COPIHUES' 'ALUMINIO' 'LA HERRADURA' 'LA UNION' 'OSA MAYOR'
'CARDENAL JOSE MARIA CARO' 'LAS HORTENSIAS' 'COMERCIO' 'EL ATOMO'
'EL UNIVERSO' 'DEUTERIO' 'EL ESFUERZO' 'LAS DALIAS' 'SIETE' 'RIO MANSO'
'CARDENAL JUAN FRANCISCO FRESNO' 'MORELOS' 'PADRE MANUEL LACUNZA'
'INES AGUILA LEDEZMA' 'PILOTO ACEVEDO' 'PEDRO LAGOS PALACIOS'
'LOS COPIHUES' 'CAMINO A LONQUEN' 'ERNESTO CEA' 'ALUMINIO' 'SIETE'
'MI ESPERANZA' 'DEPARTAMENTAL' 'EL ATOMO' 'SUIZA' 'DOMINGO COLON'
'INES AGUILA LEDEZMA' 'PRESIDENTE SALVADOR ALLENDE' 'DEPARTAMENTAL'
'CORNETA CRISPIN REYES' 'NUEVA UNO' 'LAS VIOLETAS' 'LOS CLAVELES'
'AEROPUERTO' 'LOS JAZMINES' 'EL UNIVERSO' 'LOS ROSALES' 'RIO MAYO'
'MONSEÑOR JOSE MARIA CARO' 'COLO-COLO' 'CAMINO A MELIPILLA'
'DIVINA COMEDIA' 'CONCEJALA BERTA CARVAJAL' 'PADRE HURTADO' 'ALUMINIO'
'CIUDAD MADERO' 'MONSEÑOR JOSE MARIA CARO' 'LA PINTA' 'CINCO'
'CIUDAD MADERO' 'UNOS DE AGOSTO' 'RIO TOLTEN']
```

Figura 4.5: Visualización de resultados del modelo en Python. Fuente: Elaboración propia.

4.5. Evaluación

Se realizó una evaluación de los modelos, con el fin de concluir si es posible obtener mejores resultados utilizando un modelo de aprendizaje supervisado. El código de programación tradicional de la empresa se pudo estudiar y poner a prueba (ya que se dispone de este), pero por motivos de confidencialidad no se muestra explícitamente en este trabajo. Sin embargo, se proporciona información que permite entender su mecanismo general de operación.

Es importante destacar que ambos modelos fueron evaluados utilizando los dos conjuntos de datos diseñados para este propósito: uno que utiliza datos ficticios y otro basado en datos reales. Esto permitió una comparación de los rendimientos de cada modelo en ambos contextos.

4.5.1. Modelo de normalización actualmente utilizado por la empresa

4.5.1.1. Procesos relevantes modelo actual de normalización

En cuanto al actual modelo de normalización, se identificaron dos procesos relevantes dentro del código de programación:

Limpieza de caracteres o palabras

Se eliminan de una nueva dirección ingresada por un destinatario las siguientes palabras o caracteres:

#, N°, SN, INT., PISO, LOCAL, CASA, POBLACION, S/N, DPTO, DEPTO, OFICINA, DEPARTAMENTO, BLOCK, PASAJE, PSJE, PSJE., CONDOMINIO, SUCURSAL, CALLE, NO, UNDEFINED, SIN NUMERO, NUMERO, PARCELACION, (TEXTO), COND., COND, ESQUINA, Æ por I, Å± por N, Å© por E, Å3 por O, Åo por U, Åj por A

Cortar según palabras

Cuando se encuentra alguno de los siguientes, se elimina lo que les sigue en la dirección:

PISO, LOCAL, SN, CASA, POBLACION, S/N, DPTO, DEPTO, OFICINA,
DEPARTAMENTO, BLOCK, N°, INTERIOR

Estos procesos eliminan información relevante de las direcciones en algunos casos. En la Tabla 4.7 se presentan ejemplos de direcciones y cómo fueron normalizadas por el código de programación tradicional utilizado actualmente por la empresa. Cada fila muestra una dirección original (*direccion*) y su correspondiente versión normalizada generada por el modelo (*direccion_limpia*).

Tabla 4.7: Ejemplos de resultados de normalización de direcciones según el modelo de programación tradicional actual. Fuente: Elaboración propia.

direccion	direccion_limpia
CONDE DE ORGAZ	ORGAZ
5 DE ABRIL	ABRIL
ONDURAS	ONDURAS
ACCESO CAMINO A LONQUEN, NUEVA LAREDO	ACCESO CAMINO LONQUEN LAREDO
SAN JOSE	JOSE
AMASONAS	AMASONAS
EL ALAMO	ALAMO

A partir de la Tabla 4.7 se pueden identificar casos en que se elimina información relevante para la posterior georreferenciación. En su normalización, el código eliminó de las direcciones los números y las palabras “CONDE”, “DE”, “A”, “NUEVA”, “SAN” y “EL”. Además, las faltas de ortografía no son corregidas.

4.5.1.2. Descripción del código de programación tradicional

Como se ha mencionado, el modelo utilizando por la empresa es un código de programación convencional. El código realiza varias operaciones relacionadas con la normalización y limpieza de direcciones. A continuación, se describen de forma general las funciones principales que realiza el código.

Normalización de direcciones

En la primera parte del código, se normalizan los caracteres especiales en una cadena de dirección, llamada *txtDireccionBuscar.Text*. Se realizan múltiples llamadas al método *Replace* para sustituir caracteres con tilde, diacríticos y especiales por sus equivalentes en español sin tilde. Esto se hace para asegurar consistencia y facilitar la comparación entre direcciones.

Validación de caracteres

La función *validarCaracteres* realiza una validación más extensa de los caracteres presentes en una cadena. Itera sobre cada caracter, comprobando si es imprimible, espacio en blanco, signo de puntuación o alfanumérico. Se utilizan múltiples condiciones para determinar si un caracter cumple con las condiciones establecidos, y se devuelve un valor booleano, indicando si la cadena pasa la validación.

Limpieza de direcciones

La función *Limpiar_Direcciones* procesa una dirección específica. Se eliminan palabras clave (como “AVENIDA”, “PISO”, “CASA”, entre otros) y se eliminan espacios en blanco al principio y al final de la dirección. Luego, se realiza un procesamiento adicional para extraer información relevante de la dirección, como el número de piso o la comuna. La idea de esto es obtener como resultado final una dirección normalizada y limpia.

4.5.1.3. Métricas

Para el método que utiliza un código de programación tradicional fue necesario crear una métrica específica y adaptada para la precisión. Esta necesidad surgió debido a una distinción

fundamental entre los modelos de aprendizaje automático y los de programación tradicional. En el contexto de un modelo tradicional, no existe la posibilidad de incorporar directamente una línea de código que proporcione de manera automática la precisión del modelo, una funcionalidad típicamente presente en los modelos de machine learning. Por lo tanto, se diseñó una métrica personalizada para evaluar el rendimiento del modelo de programación implementado.

Inicialmente, la base de datos proporcionada al modelo tradicional para su evaluación consiste únicamente en una columna idéntica a la columna *direccion_ingresada* de la base de datos del modelo supervisado. Se incorporaron líneas de código adicionales para generar un archivo Excel que contuviera las direcciones después de haber aplicado todo el proceso del código, representadas en la columna *direccion_limpia*.

Para medir la precisión del modelo, se introdujo una tercera columna al descargar el Excel: *direccion_correcta*. Se añadió una cuarta columna, *indicador*, que devuelve un 1 si la dirección en *direccion_limpia* está correctamente normalizada, en comparación con *direccion_correcta*, es decir, si no se ha perdido información relevante o si no contiene errores ortográficos. En caso contrario, si la dirección perdió información relevante durante el proceso de normalización del código tradicional o si presenta errores ortográficos, se asigna un 0.

Indicador de precisión

Para la asignar el indicador, se utilizó en la columna *indicador*, en la cual mediante una fórmula de Excel se asignan los valores 1 o 0:

- 0: si *direccion_limpia* no está correctamente normalizada al compararla con *direccion_correcta*.
- 1: si *direccion_limpia* está correctamente normalizada al compararla con *direccion_correcta*.

En la Tabla 4.8 se muestra un extracto de la tabla Excel construida para obtener los resultados de la métrica de precisión.

Tabla 4.8: Evaluación del modelo de programación tradicional utilizando una métrica propia. Fuente: Elaboración propia.

<i>direccion_ingresada</i>	<i>direccion_limpia</i>	<i>direccion_correcta</i>	<i>indicador</i>
VEINTIUNO DE MAYO	VEINTIUNO MAYO	VEINTIUNO DE MAYO	0
AEROPUERTO RODELILLO,	AEROPUERTO RODELILLO,	AEROPUERTO RODELILLO	1
AV. LAS VIOLETAS	LAS VIOLETAS	LAS VIOLETAS	1
BELO HORISONTE	BELO HORISONTE	BELO HORIZONTE	0
BERTA CARBAJAL	BERTA CARBAJAL	BERTA CARVAJAL	0
COMERSIO	COMERSIO	COMERCIO	0
FELIZ MARGOS	FELIZ MARGOS	FELIX MARGOZ	0
FRANSIA	FRANSIA	FRANCIA	0
LAS GALACCIAS	LAS GALACCIAS	LAS GALAXIAS	0
NVA UNO	-	NUEVA UNO	0
DEPARTAMENTAL	-	DEPARTAMENTAL	0

La Tabla 4.8 incluye cuatro columnas: *direccion_ingresada* (dirección ingresada como input al modelo), *direccion_limpia* (normalización que realiza el modelo a la dirección ingresada), *direccion_correcta* (la normalización correcta de la dirección ingresada) y *indicador* (métrica personalizada que evalúa la precisión de la normalización, donde 1 indica una normalización exitosa y 0 una normalización de mala calidad).

Así, para calcular la precisión del modelo utilizado por la empresa, se determinó la proporción de direcciones normalizadas correctamente con respecto al total de direcciones ingresadas. La utilización de esta métrica permitió realizar una evaluación de la efectividad del modelo en la normalización de direcciones.

Tabla 4.9: Precisión obtenida para el modelo de programación tradicional, utilizando una base de datos ficticia y una base de datos real. Fuente: Elaboración propia.

BBDD Proporcionada	Precisión
Ficticia	59 %
Real	61 %

4.5.2. Modelo de normalización propuesto

La validación del rendimiento del modelo SVM, aplicado a la normalización de direcciones y basado en aprendizaje supervisado, se llevó a cabo mediante una serie de métricas esenciales comúnmente empleadas en la evaluación de modelos de machine learning. Esto permitió conocer su eficacia en el procesamiento y ajuste de datos. A continuación se muestra un cuadro comparativo de las métricas obtenidas para el modelo, en los dos casos estudiados (base de datos ficticia y real).

Tabla 4.10: Métricas de rendimiento obtenidas para el modelo de normalización propuesto, con base de datos ficticia y real. Fuente: Elaboración propia.

Métrica	BBDD Ficticia	BBDD Real
Precisión	97 %	63 %
Recall	95 %	63 %
F1	95 %	57 %
Jaccard Score	93 %	54 %
Distancia de Levenshtein	0,35	1,93

La precisión (accuracy) se calculó comparando el número de predicciones correctas del modelo con el total de predicciones realizadas, ofreciendo una visión general de su eficacia. La sensibilidad (recall), que mide la proporción de casos positivos reales correctamente identificados, permitió evaluar la capacidad del modelo de detectar direcciones relevantes. El puntaje F1, promedio armónico de precisión y sensibilidad, proporcionó un balance entre ambas, reflejando la precisión general del modelo. El puntaje de Jaccard, que evalúa la similitud entre las predicciones del modelo y los valores verdaderos, ayudó a comprender la proximidad de las predicciones a las etiquetas reales. Finalmente, la distancia de Levenshtein,

que mide la diferencia entre las direcciones normalizadas por el modelo y las ideales, ofreció una perspectiva específica sobre la precisión en la normalización.

Estas métricas ayudan a comprender la capacidad del modelo para manejar con precisión la tarea de normalización de direcciones en un entorno de aprendizaje supervisado.

4.5.3. Evaluación entre modelos

En la comparativa metodológica realizada en este estudio, se evaluaron dos modelos: programación tradicional y modelo de aprendizaje supervisado, para la normalización de direcciones en una empresa de despacho. En ambos modelos se emplearon dos conjuntos de datos: uno ficticio, centrado en una comuna específica, y otro real, con datos de diversas comunas. La base de datos ficticia fue diseñada para simular condiciones realistas, incluyendo repeticiones de nombres de calles, facilitando así el aprendizaje supervisado del modelo. La base de datos real contenía registros de malos ruteos de direcciones de diversas comunas, para evaluar su efectividad en un entorno más complejo y diverso.

La métrica de evaluación para el modelo de programación tradicional, diseñada específicamente para este estudio, asignaba un valor de 1 a las direcciones correctamente normalizadas (es decir, aquellas que coincidían con la versión “idealmente normalizada”) y un valor de 0 a las direcciones con errores, incluyendo errores ortográficos o de otro tipo.

Para el modelo SVM, los resultados indican un alto grado de eficacia en el contexto controlado de la base de datos ficticia. La elevada precisión y sensibilidad reflejan la capacidad del modelo para normalizar correctamente las direcciones y reconocer patrones consistentes en los datos. Al utilizar la data real, los resultados muestran una disminución en la eficacia del modelo al enfrentarse a datos más heterogéneos y complejos. Esta disminución podría atribuirse a la diversidad de patrones en las direcciones de diferentes comunas, lo que presenta un mayor desafío para el modelo.

La diferencia en el rendimiento del modelo propuesto, entre la base de datos ficticia y la real, resalta la importancia de la diversidad de datos en el entrenamiento de modelos de aprendizaje supervisado. Mientras que el modelo demostró ser altamente efectivo en un entorno controlado, su aplicación en un contexto más variado presentó limitaciones significativas.

Esta observación sugiere la necesidad de incorporar un conjunto con una mayor cantidad de datos, más diverso y representativo durante la fase de entrenamiento, para mejorar la robustez y la generalización del modelo en aplicaciones prácticas. Además, estos hallazgos proporcionan una base para futuras investigaciones en la mejora de métodos de normalización de direcciones, especialmente en entornos complejos y heterogéneos.

Aplicando el modelo de programación tradicional a la base de datos ficticia, que contenía direcciones de una sola comuna, se obtuvo una precisión del 59 %. Este resultado refleja ciertas limitaciones del modelo al manejar datos controlados y relativamente homogéneos. Cuando se empleó este mismo modelo con la base de datos real, caracterizada por una mayor diversidad de direcciones procedentes de distintas comunas, la precisión experimentó un leve aumento, alcanzando un 61 %. Este ligero incremento sugiere que el modelo de programación

tradicional posee cierta adaptabilidad al enfrentarse a datos más variados y complejos.

El modelo de aprendizaje supervisado propuesto en el presente trabajo demostró una capacidad superior al ser evaluado en las mismas condiciones. Con la base de datos ficticias, este modelo alcanzó una precisión del 97 %, lo cual indicaría una excelente capacidad de normalización al aplicarse en un entorno de datos simplificado. Sin embargo, utilizando la base de datos real, su precisión se redujo a un 63 %. Aunque esta cifra es superior a la alcanzada por el modelo tradicional, su descenso destaca el desafío de la heterogeneidad y complejidad de los datos reales para el proceso de normalización.

Esta observación sugiere que, aunque los modelos de aprendizaje automático pueden superar en rendimiento a los enfoques tradicionales en entornos controlados, la programación tradicional puede ofrecer ciertas ventajas en términos de estabilidad y rendimiento consistente en escenarios más variados. Mientras que el modelo de aprendizaje supervisado muestra una eficacia sobresaliente en condiciones controladas y experimenta un rendimiento decreciente (pero aún competente) en escenarios más complejos, el modelo de programación tradicional posee una adaptabilidad, mejorando ligeramente su precisión en contextos de mayor diversidad. Esto destaca la importancia de considerar la naturaleza de los datos al elegir la metodología de normalización de direcciones más adecuada.

Al comparar ambos modelos, es importante destacar que la diferencia en la precisión no solo se deriva de las técnicas de modelado, sino también de cómo cada modelo maneja las peculiaridades y desafíos inherentes a los datos de direcciones. El modelo tradicional, basado en reglas y algoritmos predefinidos, puede ser limitado en su capacidad de adaptarse a variaciones inesperadas en los datos. En contraste, el modelo propuesto de aprendizaje supervisado, puede capturar y aprender de estas variaciones.

Aunque los resultados obtenidos para el modelo de machine learning son relativamente buenos, es crucial reconocer las limitaciones del estudio actual. El uso de una base de datos ficticia plantea preguntas sobre la aplicabilidad de los modelos en situaciones del mundo real. Además, la exclusión de elementos como la numeración y la comuna puede limitar la utilidad de los modelos en aplicaciones prácticas. Para futuras investigaciones, sería beneficioso incorporar una base de datos real más grande y más diversa, lo que permitiría evaluar la efectividad de los modelos en un espectro más amplio de casos. También sería útil explorar la adaptabilidad de los modelos a diferentes idiomas y convenciones de direcciones, lo que aumentaría su aplicabilidad en un contexto global.

Basándose en los resultados de la evaluación, se recomienda que para aplicaciones donde la precisión es crítica y los datos de direcciones son variados y complejos, el modelo de machine learning sería la opción preferida. Sin embargo, en situaciones donde la eficiencia y la simplicidad son más importantes, y los datos son más uniformes, el modelo de programación tradicional podría ser suficiente. Además, es importante considerar la integración de estos modelos en sistemas existentes, evaluando factores como la facilidad de implementación, el mantenimiento y la escalabilidad. La elección final debería basarse en un equilibrio entre la precisión, la complejidad y la practicidad en el contexto específico de la empresa de despacho.

4.6. Despliegue

En la última etapa de despliegue dentro de la metodología CRISP-DM, aunque este trabajo no ejecutará la implementación del modelo de aprendizaje supervisado desarrollado, se ofrecen recomendaciones estratégicas para la empresa, considerando las limitaciones y áreas de mejora identificadas. Según los resultados de precisión para el modelo supervisado, es importante un entrenamiento más diversificado y representativo para mejorar la robustez y la capacidad de generalización.

En el contexto actual de la empresa, que se basa en un método de programación tradicional para la normalización de direcciones, es crucial considerar su postura sobre la inversión en nuevas tecnologías. La empresa ha expresado su interés en evitar gastos considerables en software avanzado de normalización de direcciones, optando en cambio por maximizar el uso del gran equipo de TI ya existente. Así, buscan optimizar los recursos disponibles sin incurrir en costos adicionales significativos.

A pesar de que el sistema actual de normalización de direcciones es algo anticuado, la empresa mantiene un nivel de servicio alto, con una eficiencia aproximada del 95 %. En este marco, se propone un enfoque que mejore la eficiencia sin grandes inversiones. La iniciativa de este proyecto surge desde las necesidades identificadas en el área de control de distribución, la cual ha observado una cantidad considerable de horas dedicadas a la corrección manual de direcciones. Esto sugiere que una mejora en el método de normalización podría reducir la necesidad de intervención manual en este aspecto.

Actualmente hay una empleada que está encargada de las correcciones manuales de direcciones mediante un software de mantenimiento de la empresa. Este trabajo, que incluye tanto acciones preventivas como correctivas, enfrenta el desafío de abordar solo un 30 % de las direcciones que necesitan corrección, de un promedio diario de 1.400. Esta limitación en la capacidad de corrección manual no solo subraya la necesidad de optimizar el método de normalización de direcciones sino también la importancia de considerar soluciones tecnológicas que puedan aliviar esta carga y aumentar la eficiencia del proceso, garantizando así la mejora continua de la calidad del servicio y la satisfacción del cliente.

4.6.1. Recomendaciones para la mejora e implementación del modelo propuesto

En este trabajo se propone una estrategia para la revisión del diccionario de direcciones, compuesto por aproximadamente 40 millones de registros, y la implementación de un modelo de aprendizaje supervisado. Luego de evaluar los modelos de normalización de direcciones y de identificar limitaciones clave en el modelo de aprendizaje supervisado desarrollado, surge la necesidad de abordar desafíos en términos de generalización y adaptabilidad a datos complejos y heterogéneos. Considerando lo anterior, se propone la exploración del aprendizaje por refuerzo como una vía potencialmente prometedora para el perfeccionamiento del modelo. A continuación, se detallan las recomendaciones estratégicas para la integración, así como otras mejoras y ajustes.

1. Integración del Aprendizaje por Refuerzo: Se recomienda incorporar técnicas de aprendizaje por refuerzo para permitir que el modelo de aprendizaje supervisado mejore

continuamente, a través de la interacción con el entorno de datos. Esto permitirá al modelo adaptarse y optimizar sus estrategias de normalización de direcciones basadas en un feedback directo, mejorando su precisión y capacidad de manejo de datos variados y complejos.

2. Ampliación y diversificación del conjunto de datos: Para apoyar la implementación del aprendizaje por refuerzo, y mejorar la capacidad de generalización del modelo, es necesario enriquecer el conjunto de datos de entrenamiento con una mayor diversidad y cantidad de ejemplos, incluyendo direcciones de múltiples comunas, y que consideren variaciones complejas.

3. Desarrollo y optimización continua: Mediante el uso de aprendizaje por refuerzo, el modelo debería someterse a un proceso de desarrollo y optimización continuo. Esto implica ajustes periódicos, basados en el rendimiento del modelo en escenarios reales, asegurando que las estrategias de normalización sean efectivas y actualizadas.

4. Implementación de una API para integración efectiva: Para facilitar la aplicación práctica del modelo mejorado en el sistema existente de la empresa, se propone el desarrollo de una API robusta. Esta interfaz permitirá una integración fluida del modelo con los sistemas operativos de la empresa, apoyando la normalización de direcciones en tiempo real.

5. Evaluación y ajustes basados en feedback: Finalmente, se sugiere establecer un ciclo de feedback operativo, una vez que el modelo y la API estén implementados. Este proceso de evaluación continua permitirá identificar áreas de mejora, mediante el ajuste del modelo y de la API según sea necesario para optimizar su rendimiento.

Al adoptar estas recomendaciones, la empresa abordará las limitaciones identificadas en el modelo de aprendizaje supervisado propuesto, y se podrá posicionar a la vanguardia de la innovación tecnológica en cuanto a la normalización de direcciones. La integración del aprendizaje por refuerzo es prometedora para mejorar la precisión y eficiencia en la normalización de direcciones.

4.6.1.1. Carta Gantt

Para implementar exitosamente el modelo de aprendizaje supervisado, revisar integralmente el diccionario de direcciones de la empresa e integrar las tecnologías de aprendizaje por refuerzo, se requiere una planificación metódica. Se propone una carta Gantt que resume el enfoque recomendado para el proyecto, definiendo las etapas clave, partiendo por la evaluación inicial y la selección de herramientas, luego el desarrollo y optimización del modelo, y finalmente, su integración y evaluación en el entorno operativo de la empresa.

La implementación se estructura en un proyecto de 12 meses de duración, desglosado en las siguientes fases:

- **Inicio y planificación**

Durante los primeros dos meses, el proyecto comienza con la evaluación detallada del proyecto, donde se deben definir los objetivos, el alcance y los recursos necesarios. Además, se deben seleccionar las herramientas necesarias, enfocándose en plataformas de procesamiento de datos a gran escala y herramientas de aprendizaje por refuerzo. Parale-

lamente, se debe comenzar con la revisión inicial del diccionario de direcciones, mediante procesos automatizados que permitan una primera limpieza y análisis de los datos existentes.

- **Desarrollo y enriquecimiento de datos**

En los meses tres y cuatro, se continúa con la revisión avanzada del diccionario empleando algoritmos de aprendizaje por refuerzo para identificar y corregir errores más complejos. Paralelamente, se deberá ampliar el conjunto de datos de entrenamiento con nuevos datos para aumentar su diversidad, y se procederá con el desarrollo preliminar del modelo de aprendizaje supervisado, utilizando el conjunto de datos enriquecido para construir y realizar pruebas iniciales.

- **Optimización del modelo y pruebas iniciales**

Los meses cinco y seis se dedican a optimizar el modelo y realizar las pruebas iniciales, el modelo será mejorado con técnicas de aprendizaje por refuerzo, basándose en los resultados preliminares y las recompensas obtenidas. Se llevarán a cabo pruebas de validación cruzada para asegurar la robustez y capacidad de generalización del modelo. Además, se iniciará el desarrollo de una API destinada a facilitar la integración del modelo con los sistemas internos de la empresa.

- **Integración y pruebas del sistema**

En los meses siete y ocho, correspondientes a la fase de integración y pruebas del sistema, se concluirá el desarrollo de la API y se deberá integrar completamente con los sistemas existentes, realizando pruebas de integración, para verificar su funcionalidad y rendimiento. Se recomienda mantener la revisión del diccionario de direcciones como un proceso continuo, mejorando el diccionario basado en el feedback recogido del sistema.

- **Implementación y evaluación operativa**

En los meses nueve y diez, el modelo y la API serán desplegados de forma controlada en el entorno de producción de la empresa. Se establecerá un sistema de monitoreo, para recoger feedback operativo, el cual es crucial para los ajustes futuros.

- **Ajustes finales y evaluación del proyecto**

En los últimos dos meses, se realizarán los ajustes finales al modelo y la API, basándose en el feedback recopilado. Se llevará a cabo una evaluación final, para revisar el cumplimiento de los objetivos y documentar lo aprendido y lo relevante. Además, se planificará el mantenimiento y la actualización continua del modelo y el diccionario, asegurando su adaptabilidad a largo plazo.

ETAPA	TAREA / SEMANA	MES 1			MES 2			MES 3			MES 4			MES 5			MES 6			MES 7			MES 8			MES 9			MES 10			MES 11			MES 12																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48				
Inicio y planificación	1. Evaluación del proyecto	█	█	█	█																																																
	2. Selección de herramientas y tecnologías			█	█																																																
	3. Inicio de la revisión del diccionario de direcciones				█	█	█	█																																													
Desarrollo y enriquecimiento de datos	4. Continuación de la revisión del diccionario								█	█																																											
	5. Ampliación del conjunto de datos de entrenamiento										█	█	█																																								
	6. Desarrollo preliminar del modelo de aprendizaje												█	█	█	█																																					
Optimización del modelo y pruebas iniciales	7. Optimización del modelo con RL																█	█																																			
	8. Pruebas de validación cruzada																		█																																		
	9. Desarrollo de la API para integración																				█	█	█	█																													
Integración y pruebas de sistema	10. Finalización del desarrollo de la API																											█	█																								
	11. Pruebas de integración del modelo y la API																																																				
	12. Revisión continua del diccionario de direcciones																																																				
Implementación y evaluación operativa	13. Implementación gradual en el entorno de producción																																																				
	14. Monitoreo y recolección de feedback																																																				
Ajustes finales y evaluación del proyecto	15. Ajustes basados en el feedback																																																				
	16. Evaluación final del proyecto																																																				
	17. Planificación de mantenimiento y actualización																																																				

Figura 4.6: Propuesta de carta Gantt de doce meses de duración para el proyecto. Fuente: Elaboración propia.

Cada etapa de este proyecto requiere una estrecha colaboración entre diferentes equipos y departamentos, asegurando así que el modelo se desarrolle, pruebe e implemente de manera efectiva y eficiente. La asignación de responsabilidades claras es fundamental para el éxito del proyecto, permitiendo una transición suave hacia un método de normalización de direcciones más avanzado y automatizado.

4.6.2. Recomendaciones exploración chatbot IA

En el marco del despliegue de la metodología CRISP-DM para el presente trabajo, se ha puesto el enfoque principal en la exploración de modelos de aprendizaje automático para la normalización de direcciones. Sin embargo, el dinamismo y las innovaciones en el campo de la inteligencia artificial (IA) impulsan a considerar las alternativas tecnológicas emergentes, dado su creciente protagonismo y la promesa que ofrece en múltiples campos. Particularmente, su aplicación en procesos complejos, como la normalización de direcciones, emerge como un campo muy prometedor. Dada su capacidad para procesar y entender el lenguaje natural con cada vez más eficacia y precisión, se propone investigar la implementación de chatbots de IA para la normalización de direcciones. Las innovaciones en el campo de la IA, especialmente los chatbots, presentan una oportunidad única no solo por su eficiencia en procesamiento del lenguaje natural, sino también por su capacidad para integrarse sin problemas mediante API en los sistemas existentes de la empresa.

En la Sección 4.6.2.1 se muestra un experimento detallado de este estudio, que ilustra la superioridad de un chatbot de IA frente a los métodos tradicionales de normalización de direcciones, mostrando una mejora significativa en la precisión de la georreferenciación de direcciones normalizadas por este. La facilidad con la que estas tecnologías pueden ser implementadas, a través de una integración a una API, sin la necesidad de una comprensión profunda o modificaciones de los sistemas de software internos de la empresa, destaca su valor práctico.

Los resultados del experimento demuestran claramente la eficacia del chatbot de inteligencia artificial en la normalización de direcciones en comparación con el método de programación tradicional. En el primer escenario, se le pidió al chatbot normalizar las direcciones y luego estas fueron georreferenciadas utilizando una API de geolocalización. Esto permitió evaluar cuántas direcciones pudieron ser correctamente ubicadas a partir de la normalización proporcionada por el chatbot. Simultáneamente, se hizo el experimento con el método de programación tradicional, identificando cuántas direcciones lograron una georreferenciación precisa, independientemente de su calidad de normalización.

Los resultados muestran una diferencia significativa: un 79 % de las direcciones normalizadas por el chatbot de IA lograron una georreferenciación correcta, en contraste con el método de programación tradicional, que solo alcanzó un 18 %. Esta mejora de 61 puntos porcentuales demuestra la capacidad del chatbot de IA para optimizar el proceso de normalización de direcciones, lo que resulta en una mayor precisión en la georreferenciación.

El experimento realizado destaca el potencial de la IA para mejorar significativamente la precisión y eficiencia en la normalización de direcciones, lo que sugiere que su exploración más profunda podría abrir nuevas vías para optimizar los procesos logísticos y de georreferenciación.

Por lo tanto, se recomienda fuertemente a la empresa considerar la adopción e integración de chatbots de IA en el proceso de normalización de direcciones, como una alternativa al desarrollo de un modelo de aprendizaje supervisado. Esta recomendación no solo se basa en la búsqueda de una mayor eficiencia y precisión sino también en la flexibilidad y escalabilidad que estos sistemas pueden ofrecer. Integrar un chatbot de IA mediante una API podría resultar en una normalización de direcciones continua, automatizada y altamente precisa, lo cual es de gran valor para optimizar los procesos logísticos y de georreferenciación.

Además, esta propuesta se alinea con la tendencia hacia soluciones tecnológicas avanzadas, ofreciendo a la empresa una vía para mantenerse a la vanguardia en un mercado competitivo y dinámico. Se sugiere llevar a cabo una evaluación detallada de los requerimientos técnicos y logísticos para la implementación de esta tecnología, considerando los beneficios a largo plazo que la integración de un chatbot de IA puede representar para la mejora continua de los procesos internos de la empresa.

4.6.2.1. Experimento realizado

En la búsqueda de un entendimiento más profundo sobre el objetivo de este trabajo, se realizó un experimento para evaluar la eficacia de diferentes metodologías en la normalización de direcciones, una tarea crucial para la optimización de procesos en una empresa de despachos. El método actual de la empresa se basa en programación tradicional, mientras que la propuesta se centra en un modelo de aprendizaje supervisado utilizando aprendizaje de máquinas. Este modelo, aunque inicialmente limitado a la normalización de nombres de calles, plantea un enfoque más refinado y adaptable comparado con el método tradicional. Con el fin de explorar otra tecnología, y demostrar que es necesario implementar nuevos métodos, se llevó a cabo un experimento utilizando un chatbot de inteligencia artificial, particularmente el de Bing.

El experimento consistió en analizar los casos de malos ruteos ocurridos durante ocho días. Para ello, se crearon columnas de datos, una que refleja la eficacia del método de normalización actual en la obtención de una georreferenciación precisa (“MRH geo correcta normalización actual”), y otra que muestra los resultados obtenidos mediante la normalización a través de un chat de inteligencia artificial (“MRH geo correcta normalización IA”). Cabe destacar que los resultados obtenidos con el chat de inteligencia artificial fueron significativamente positivos. En la Figura 4.7 se muestra el mensaje entregado al chatbot para explicarle lo que debía hacer, y en la Figura 4.8 se observa la salida (normalización) que arrojó para algunas de las direcciones que se le consultó.

Para el experimento se seleccionó un conjunto de direcciones correspondientes a malos ruteos generados a lo largo de ocho días, sumando un total de 1.042 casos. El propósito era discernir si las imprecisiones eran atribuibles a fallas en la georreferenciación o en la propia normalización. Para ello, se empleó un enfoque detallado. En primer lugar, para estudiar el caso de la normalización actual, se tomaron las coordenadas geográficas ya asociadas a cada dirección en el sistema, y se verificó en Google Maps si estaban correctamente asociadas a la dirección. Aquellas que se encontraban precisamente localizadas se marcaron con un 1, mientras que las erróneas recibieron un 0.

En segundo lugar, para evaluar la aplicación de inteligencia artificial, se introdujeron las direcciones a un chatbot de IA (Bing), y se le pidió que normalizara las direcciones entregando la calle, numeración y comuna. Luego, utilizando la API geocode, se obtuvieron las coordenadas de latitud y longitud correspondientes. Se buscaron estas coordenadas en Google Maps para determinar si la georreferenciación, basada en la dirección normalizada por IA, era acertada. Siguiendo el mismo sistema de clasificación, las localizaciones correctas se clasificaron con un 1, y las incorrectas con un 0. Es relevante destacar que para calificar una georreferenciación como correcta, se estableció un margen de tolerancia de 200 metros.

Este proceso permitió construir el cuadro comparativo que se presenta a continuación.

Tabla 4.11: Análisis comparativo de normalización tradicional y normalización mediante chatbot de IA. Fuente: Elaboración propia.

Fecha	MRH geo correcta normalización actual	MRH geo correcta normalización IA	Total MRH validados	% MRH geo correcta normalización actual	% MRH geo normalización IA
22-mar	21	103	133	16 %	77 %
23-mar	20	123	166	12 %	74 %
24-mar	59	117	137	43 %	85 %
25-mar	20	78	104	19 %	75 %
26-mar	7	107	128	5 %	84 %
27-mar	20	117	131	15 %	89 %
28-mar	16	84	116	14 %	72 %
29-mar	20	95	127	16 %	75 %
Total	183	824	1042	18 %	79 %

Así, en la columna *MRH geo correcta normalización actual* se muestra la cantidad de malos ruteos generados que tenían coordenadas de latitud y longitud correctamente asociadas, cuyo

problema de ruta no se debía a una mala georreferenciación. Para hacer el comparativo, en la columna *MRH geo correcta normalización IA*, se muestra la cantidad de direcciones que habiendo sido normalizadas mediante el chat de bing, y posteriormente georreferenciadas mediante la API Geocode, alcanzaron una correcta localización. Así, con respecto al total de direcciones correspondientes a mal ruteo validadas, se pudo generar las últimas dos columnas, con el porcentaje de direcciones que alcanzan una buena georreferenciación mediante el método de normalización actual, y mediante el método de normalización a través de IA.

La normalización mediante el chatbot de IA mostró un impacto de mejora en la georreferenciación de un 61% en comparación con el método tradicional de normalización. Esta evidencia respalda la hipótesis de que la implementación de técnicas más modernas y automatizadas en la normalización de direcciones puede ser una vía prometedora para mejorar la eficiencia y precisión en la logística de despachos. Los resultados del experimento son relevantes para este estudio, ya que demuestran que las metodologías avanzadas, incluyendo la inteligencia artificial y potencialmente el modelo propuesto, pueden superar significativamente el rendimiento de los métodos de normalización tradicionales.

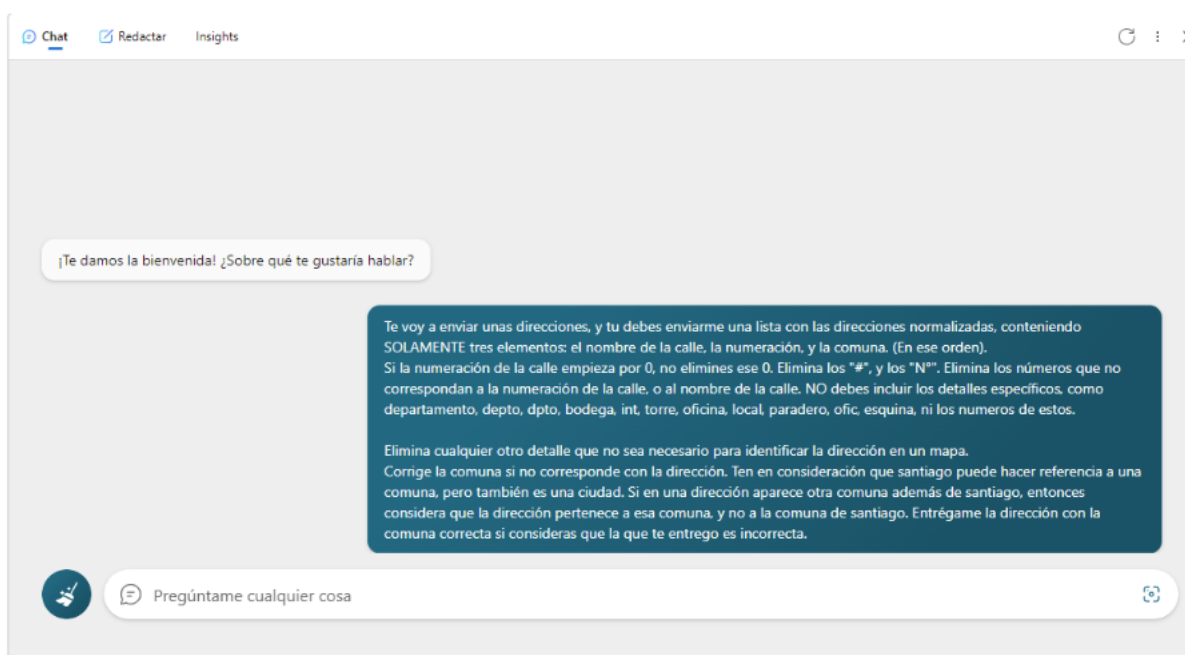


Figura 4.7: Captura de pantalla mostrando un mensaje enviado al chatbot Bing, en el que se solicita la normalización de direcciones con indicaciones específicas para incluir calle, numeración y comuna, entre otras. Fuente: Elaboración propia.

Direcciones a normalizar (mensaje enviado al chat)

AV. MANUEL RODRIGUEZ NORTE 867, DEPTO 707, SANTIAGO
EYZAGUIRRE, 1140, SECTOR 3 DEPTO D 204, SANTIAGO
AVENIDA SANTA ROSA 634 DPTO 416 DEPTO 416, SANTIAGO
AMUNATEGUI 810 DPTO 508, SANTIAGO
ARTURO PRAT 636 508, SANTIAGO
AV SANTA ROSA SANTIAGO CENTRO 170 DEPARTAMENTO 1707, SANTIAGO
EYSAGUIRRE 766 DEPA 615 DEJAR EN CONSEJERÍA, SANTIAGO
CATEDRAL 1310 DPTO 1507, SANTIAGO
ARTURO PRATT 644 DEPTO 706C ENTREGAR EN 644, SANTIAGO
ELEUTERIO RAMÍREZ 1024 APTO 2207A, SANTIAGO
DIECIOCHO 620, DEPTO 525 AL FRENTE DEL SUPERMERCADO EL LIDER, SANTIAGO
NUEVA SAN MARTIN, 1490, 1807, SANTIAGO
EYZAGUIRRE 771 DPTO 424, SANTIAGO
NUEVA SAN MARTÍN 1490 514, SANTIAGO
ELEUTERIO RAMÍREZ 1024 DEPARTAMENTO 1701A, SANTIAGO

Direcciones normalizadas (mensaje recibido)

Aquí está la lista de direcciones normalizadas:

- Manuel Rodriguez Norte 867, Santiago
- Eyzaguirre 1140, Santiago
- Avenida Santa Rosa 634, Santiago
- Amunátegui 810, Santiago
- Arturo Prat 636, Santiago
- Avenida Santa Rosa 170, Santiago
- Eyzaguirre 766, Santiago
- Catedral 1310, Santiago
- Arturo Pratt 644, Santiago
- Eleuterio Ramirez 1024, Santiago
- Dieciocho 620, Santiago
- Nueva San Martín 1490, Santiago
- Eyzaguirre 771, Santiago
- Nueva San Martín 1490, Santiago
- Eleuterio Ramirez 1024, Santiago

Figura 4.8: Ejemplos de interacción con el chatbot Bing: a la izquierda, las direcciones enviadas como input; a la derecha, las respuestas del chatbot con las direcciones normalizadas. La imagen ilustra la capacidad del chatbot para procesar y normalizar las direcciones según las especificaciones dadas. Fuente: Elaboración propia.

Capítulo 5

Resultados

En esta sección se presentan los resultados obtenidos de la evaluación de los dos modelos expuestos para la normalización de direcciones: uno basado en programación tradicional y otro empleando técnicas de aprendizaje automático, específicamente un modelo de clasificación de SVM en combinación con NLP. La evaluación de ambos modelos se llevó a cabo en dos fases distintas, utilizando inicialmente una base de datos ficticia y luego una base de datos real.

La primera fase de evaluación utilizó una base de datos ficticia, diseñada específicamente para este estudio, que contenía 1745 registros correspondientes a la comuna de Cerrillos.

Posteriormente, en la segunda fase, se empleó una base de datos real proporcionada por la empresa, compuesta por 1660 registros que incluían direcciones de distintas comunas de Santiago. Esta base de datos representaba un entorno más complejo y variado, ofreciendo una perspectiva más realista del rendimiento de los modelos en condiciones prácticas y diversificadas.

La comparación de los resultados obtenidos en ambas bases de datos permite una evaluación comprensiva de la eficacia y adaptabilidad de los modelos en diferentes contextos, brindando insights sobre su aplicabilidad en escenarios de normalización de direcciones en el mundo real.

5.1. Modelo de programación tradicional

La metodología empleada implicó la evaluación de direcciones normalizadas mediante una métrica propia, considerando la comparación con un conjunto de etiquetas correctas.

Aplicando el modelo de programación tradicional a la base de datos ficticia, que contenía direcciones de una sola comuna, se obtuvo una precisión del 59 %. Este resultado sugiere que, aunque el modelo fue capaz de normalizar correctamente una proporción significativa de las direcciones, aún presentó limitaciones en el manejo de los datos controlados y relativamente homogéneos de este conjunto.

Al utilizar el mismo modelo con la base de datos real, que incluye una mayor diversidad de direcciones de distintas comunas, se observó una ligera mejora en la precisión, alcanzando el

61 %. Este incremento, aunque marginal, indica una capacidad del modelo de programación tradicional para adaptarse a un entorno de datos más variado y complejo.

5.2. Modelo de machine learning propuesto

El modelo propuesto, que resulta de integración de un modelo de clasificación de SVM con un enfoque en el procesamiento lingüístico, se respalda mediante métricas exhaustivas, incluyendo precisión, recall, F1 score, Jaccard, matriz de confusión y distancia de Levenshtein promedio.

5.2.1. Entrenamiento con base de datos ficticia

Utilizando la base de datos ficticia se obtuvieron los siguientes resultados en la evaluación del modelo de aprendizaje de máquinas:

Tabla 5.1: Métricas del modelo propuesto utilizando base de datos ficticia.
Fuente: Elaboración propia.

Métrica	Resultado
Accuracy	97 %
Recall	95 %
F1	95 %
Jaccard Score	93 %
Distancia de Levenshtein	0,35

Los resultados obtenidos en las métricas reflejan un alto nivel de eficacia en la normalización de direcciones. Una precisión del 97 % indica que el modelo fue capaz de hacer predicciones correctas en 97 de cada 100 casos, lo que muestra una alta fiabilidad en su capacidad para identificar si una dirección está normalizada adecuadamente o no.

Un recall del 95 % revela que el modelo identificó correctamente el 95 % de las direcciones que efectivamente necesitaban ser normalizadas. Esto significa que el modelo fue muy eficaz en detectar las direcciones relevantes, asegurándose de que casi todas las direcciones que requerían normalización fueran reconocidas.

Un puntaje F1 del 95 % es particularmente bueno, ya que este valor combina la precisión y la sensibilidad. Un puntaje tan alto indica que hay un excelente equilibrio entre la capacidad del modelo para detectar direcciones que necesitan normalización y su habilidad para limitar las falsas alarmas, es decir, no identificar erróneamente direcciones como problemáticas cuando no lo son.

El puntaje de Jaccard del 93 % sugiere que las predicciones del modelo y las etiquetas correctas coinciden en un 93 %, lo que es una tasa de acuerdo muy alta. Este resultado indica que hay una gran similitud entre las decisiones del modelo y la realidad, reflejando una alta precisión en la clasificación de las direcciones.

Finalmente, una distancia de Levenshtein de 0,35 indica que, en promedio, se necesitan hacer muy pocos cambios (menos de medio cambio por dirección, en promedio) para que las direcciones normalizadas por el modelo coincidan con las direcciones ideales. Esto significa que las direcciones generadas por el modelo están muy cerca de ser perfectas, con solo ajustes mínimos necesarios. Esto se debe a que se está entrenando el modelo con aprendizaje supervisado, es decir, se le están entregando las etiquetas de cómo debe normalizar las direcciones.

En conjunto, estos resultados indican que el modelo es altamente eficiente y preciso en la tarea de normalización de direcciones, con un rendimiento sobresaliente en varias métricas importantes.

5.2.2. Entrenamiento con base de datos real

Al aplicar el modelo propuesto a la base de datos real, que incluye información más diversa y compleja de distintas comunas, se observó una disminución significativa en el rendimiento.

Tabla 5.2: Métricas del modelo propuesto utilizando base de datos real.
Fuente: Elaboración propia.

Métrica	Resultado
Accuracy	63 %
Recall	63 %
F1	57 %
Jaccard Score	54 %
Distancia de Levenshtein	1,93

Una precisión del 63 % indica que, en general, el modelo hizo predicciones correctas en aproximadamente 63 de cada 100 casos. Es decir, más de la mitad de las veces, el modelo pudo identificar correctamente si una dirección era normalizada de manera adecuada o no. El recall, también del 63 %, significa que el modelo fue capaz de identificar correctamente el 63 % de las direcciones que realmente necesitaban ser normalizadas. Es una forma de medir cuántas direcciones “relevantes” no se pasaron por alto.

El puntaje F1 del 57 % es una combinación de precisión y sensibilidad, proporcionando una medida de la precisión y la capacidad del modelo para no perder casos importantes. Este valor sugiere un equilibrio moderado entre la capacidad del modelo para detectar direcciones que necesitan normalización y su habilidad para evitar falsas alarmas.

Un puntaje de Jaccard del 54 % implica que, de todas las direcciones que el modelo identificó como normalizadas correctamente o no, aproximadamente 54 % de esas decisiones coincidieron con la realidad. Este puntaje mide la similitud entre las predicciones del modelo y las etiquetas correctas.

Finalmente, una distancia de Levenshtein de 1,93 indica que, en promedio, se necesitarían alrededor de dos cambios (inserciones, eliminaciones o sustituciones de caracteres) para corregir las direcciones normalizadas por el modelo y hacerlas coincidir con las direcciones ideales.

Sería ideal una puntuación más baja en esta métrica, ya que indicaría menor discrepancia entre las direcciones generadas por el modelo y las direcciones perfectamente normalizadas.

En resumen, estos resultados muestran un modelo con una capacidad moderada para normalizar direcciones, con un desempeño más fuerte en algunas áreas que en otras. Mientras que logra identificar correctamente una cantidad significativa de direcciones que necesitan normalización, aún hay margen de mejora en términos de precisión y en la exactitud de las direcciones normalizadas.

5.3. Comparación de resultados

Con el propósito de mejorar la comprensión de las diferencias en la normalización de direcciones por los modelos, se presentan ejemplos en la Tabla 5.3 que ilustran casos en que estos modelos aplican la normalización obteniendo distintos resultados.

Tabla 5.3: Ejemplos de normalización de calles por modelos de programación tradicional y propuesto. Fuente: Elaboración propia.

Nombre Calle	Normalización Programación Tradicional	Normalización Modelo Propuesto
21 DE MAYO	MAYO	21 DE MAYO
CALETERA AV. PEDRO AGUIRRE CERDA	CALETERA PEDRO AGUIRRE CERDA	CALETERA AV. PEDRO AGUIRRE CERDA
CALLE UNO	-	UNO
AV LOS JASMINES	LOS JASMINES	LOS JAZMINES
LOS PRECIDENTES	LOS PRECIDENTES	LOS PRESIDENTES
CONDE DE ORGAZ	DE ORGAZ	CONDE DE ORGAZ
CARDENAL JUAN FRANCISCO FRESNO	CARDENAL JUAN FRANCISCO	CARDENAL JUAN FRANCISCO FRESNO

En la Tabla 5.4 se muestra una comparación de los porcentajes de precisión que obtuvo cada modelo, según la base de datos utilizada.

Tabla 5.4: Comparación de precisión entre modelos de programación tradicional y propuesto según la base de datos aplicada. Fuente: Elaboración propia.

Base de datos	Modelo	Precisión
Ficticia	Programación tradicional	59 %
	Propuesto	97 %
Real	Programación tradicional	61 %
	Propuesto	63 %

La evaluación comparativa realizada de los resultados obtenidos para ambos modelos, utilizando dos bases de datos distintas (una ficticia y otra real) entrega una perspectiva crítica sobre su rendimiento en la normalización de direcciones. El modelo de aprendizaje supervisado alcanzó una precisión del 97 % con la base de datos ficticia, lo que sugiere una adaptación eficaz al conjunto de datos específico y controlado. El modelo de programación logró una precisión del 59 % bajo las mismas condiciones. Al aplicar los modelos a la base de datos real, la precisión del modelo de aprendizaje supervisado disminuyó a un 63 %, mientras que el modelo tradicional mostró una ligera mejora, alcanzando el 61 %.

Si bien el modelo de aprendizaje supervisado mostró un rendimiento bastante más alto en el entorno controlado de la base de datos ficticia, la disminución en su precisión al enfrentarse

a los datos de la base real, más variados y complejos, plantea preguntas sobre su aplicabilidad. Esta observación es crucial, ya que, aunque el modelo de aprendizaje supervisado mantuvo una ventaja sobre el método de programación tradicional para ambos escenarios, las diferencias en la precisión no son tan pronunciadas como para garantizar su implementación en un entorno real sin consideraciones adicionales.

Además, el contraste en la precisión del modelo supervisado sugiere la presencia de sobreajuste. Este es un fenómeno común cuando un modelo se ajusta demasiado a las características del conjunto de entrenamiento, comprometiendo su capacidad de generalización a datos nuevos. Es importante identificar esta tendencia, ya que supondría equilibrar la complejidad del modelo y la diversidad de los datos de entrenamiento, para mejorar su aplicabilidad, lo cual puede ser un punto a trabajar para una futura optimización.

5.4. Limitaciones y áreas de mejora

Este estudio ha permitido una comparación detallada entre dos modelos de normalización de direcciones: uno basado en programación tradicional y otro en aprendizaje supervisado. Si bien ambos enfoques ofrecen ventajas específicas, también presentan desafíos significativos en la gestión de errores de normalización de direcciones, destacando áreas cruciales para el desarrollo futuro.

Para el modelo de programación tradicional, la menor precisión observada en la base de datos ficticia sugiere una eficacia limitada en la identificación y corrección de errores, particularmente en entornos homogéneos. Este hallazgo apunta hacia la necesidad de mejorar su algoritmo para abarcar una gama más amplia de errores, lo cual podría incrementar sustancialmente su aplicabilidad en diversos escenarios.

Por otro lado, el modelo de aprendizaje supervisado, a pesar de demostrar una eficiencia notable en un entorno controlado, enfrenta desafíos al ser aplicado a datos reales, más complejos. Este contraste revela la importancia de un entrenamiento más diversificado que incorpore ejemplos realistas de errores de normalización, y sugiere la adaptación del modelo para mejorar su flexibilidad frente a variaciones y anomalías en los datos. La notable disminución en la precisión al aplicar el modelo a la base de datos real plantea preocupaciones de sobreajuste, subrayando la necesidad de implementar estrategias de validación más rigurosas y ajustes en el modelo y/o proceso de entrenamiento para mejorar su generalización.

Una limitación crítica identificada en este estudio es la relación prácticamente inexistente entre el departamento de TI y las demás áreas de la empresa. Esta desconexión plantea desafíos significativos para la integración de soluciones tecnológicas avanzadas, como los modelos de aprendizaje supervisado y las propuestas basadas en el aprendizaje por refuerzo. La falta de colaboración y comunicación efectiva entre TI y las otras áreas limita la capacidad de implementar y adaptar estas soluciones de manera que se alineen con las necesidades operativas y estratégicas de la empresa. Mejorar esta relación es fundamental para el éxito de futuras iniciativas tecnológicas.

Además, este estudio se centró exclusivamente en la normalización de nombres de calles, omitiendo elementos críticos como la numeración y la comuna. Esta simplificación, aunque

útil para fines experimentales, limita la aplicabilidad de los modelos en situaciones prácticas. La inclusión de datos reales más amplios y la consideración de todos los elementos relevantes en futuras investigaciones podrían mejorar significativamente la aplicabilidad y generalización de los modelos.

La incorporación del aprendizaje por refuerzo ofrece una vía prometedora para superar algunas de estas limitaciones, permitiendo a los modelos mejorar continuamente su rendimiento a través de la interacción con el entorno. Sin embargo, la implementación efectiva de estas tecnologías avanzadas requiere una relación más integrada y colaborativa entre el departamento de TI y el resto de la empresa.

En conclusión, la mejora de la comunicación y colaboración entre TI y las demás áreas de la empresa emerge como un área crítica de mejora para facilitar la integración efectiva de soluciones avanzadas de normalización de direcciones. Junto con la expansión de los conjuntos de datos utilizados y la exploración de nuevas tecnologías como el aprendizaje por refuerzo, esta mejora en la relación interdepartamental es esencial para avanzar hacia una mayor eficacia operativa y satisfacción del cliente en la era digital.

Capítulo 6

Conclusiones

En este estudio, se evaluaron dos modelos para la normalización de direcciones: un modelo basado en programación tradicional, actualmente en uso por la empresa de despacho estudiada, y un modelo supervisado de machine learning, que combina Máquinas de Soporte Vectorial (SVM) con Procesamiento Natural del Lenguaje (NLP). La evaluación se realizó tanto con una base de datos ficticia, centrada en la comuna de Cerrillos, como con una base de datos real, conteniendo direcciones de distintas comunas de Santiago.

El modelo de programación tradicional, aplicando un enfoque convencional, mostró eficacia, pero con una precisión variable: 59 % en la base de datos ficticia y un ligero aumento a 61 % en la base de datos real. En contraste, el modelo propuesto exhibió una precisión excepcional del 97 % en la base de datos ficticia, pero esta disminuyó significativamente a 63 % en la base de datos real. Este contraste subraya cómo los modelos pueden comportarse de manera diferente en entornos controlados en comparación con situaciones más complejas y variadas, resaltando la importancia de probarlos en contextos que se asemejen a escenarios reales.

Los resultados obtenidos mostraron que, aunque el modelo de programación tradicional, actualmente utilizado por la empresa, mantuvo un nivel de precisión constante, con un 59 % en la base de datos ficticia y un incremento marginal al 61 % en la base de datos real, el modelo de aprendizaje supervisado exhibió una notable del 97 % en el entorno ficticio, que disminuyó significativamente a 63 % al ser aplicado al conjunto de datos real. Este descenso en la precisión del modelo de aprendizaje supervisado en entornos reales pone de relieve las limitaciones inherentes al modelo propuesto y destaca la necesidad de un enfoque más adaptable que pueda enfrentar la diversidad y complejidad de los datos en escenarios auténticos.

Aunque el modelo de aprendizaje supervisado mostró ser superior en términos de precisión bajo condiciones ideales, los resultados no fueron tan significativos al trasladar su aplicación a un contexto más variado y representativo. Esto sugiere que, mientras los modelos de aprendizaje automático ofrecen un camino prometedor hacia la mejora de la normalización de direcciones, es crucial adoptar enfoques más avanzados y tecnologías emergentes para superar las barreras actuales.

Se sugiere que los resultados podrían ser mejores con una base de datos real si se dispusiera de datos específicos de una comuna en particular o de un conjunto de datos más amplio.

Es importante considerar que la evaluación se realizó con una cantidad similar de datos en ambas bases, lo cual implica que una es considerablemente más diversa que la otra. Si ambas bases de datos presentaran una mayor diversidad, es probable que los resultados favorecerían más al modelo propuesto. Esta posibilidad destaca la importancia de probar los modelos en entornos de datos variados para evaluar plenamente su eficacia y potencial.

Además, se propone la continuación del desarrollo de modelos de aprendizaje supervisado, potencialmente enriquecidos con técnicas de aprendizaje por refuerzo, y su encapsulación dentro de una API para facilitar la integración con los sistemas existentes. Tal enfoque permitiría no solo una adaptación y mejora continuas basadas en feedback real sino también la posibilidad de aplicar estos modelos a bases de datos más amplias y representativas, ampliando significativamente su aplicabilidad y eficacia.

En vista de los hallazgos obtenidos, se recomienda fuertemente la exploración de tecnologías innovadoras, como los chatbots de IA, que podrían ser implementados a través de una API diseñada específicamente para este propósito. Integrar tal sistema no solo podría mejorar la precisión en la normalización de direcciones, sino también proporcionar una solución dinámica y escalable que se ajuste continuamente a las variaciones y actualizaciones en los datos de direcciones.

Concluyendo, este estudio ha puesto de manifiesto tanto el potencial como las limitaciones de los modelos actuales de normalización de direcciones. Mientras avanzamos, es esencial reconocer que la adopción de nuevas tecnologías y la adaptación a los cambiantes paisajes de datos serán cruciales para superar los desafíos operacionales y logísticos que enfrentan las empresas de despacho, impactando positivamente en la calidad del servicio al cliente y en la eficiencia operativa general.

6.1. Costo promedio de un error

En cuanto a los costos, conforme se expone en el estudio realizado por Beetrack, múltiples desafíos en la logística de distribución pueden impactar significativamente en las empresas dedicadas al servicio de despacho. Estos desafíos abarcan desde retrasos en la recolección y entrega de mercancías, hasta fallas operativas en la red logística de la empresa transportista, incluyendo también factores imprevistos como congestión vehicular, cambios no programados en rutas de distribución y errores en la entrega de los paquetes. Tales contratiempos no solo generan una incertidumbre operacional en el proceso de gestión y entrega de despachos, sino que también pueden repercutir negativamente en la percepción del cliente final respecto a la calidad del servicio y su experiencia de compra [35]. Un error común y costoso en este ámbito es la entrega fallida debido a direcciones incorrectas, lo que conlleva a la empresa a incurrir en gastos adicionales para el reenvío correcto del paquete.

Dentro del contexto chileno, según un estudio publicado en el diario La Tercera[36], realizado por Retail Compass, se tienen los siguientes costos promedios mínimos y máximos de envíos de productos para la categoría de productos medio y pequeños.

Tabla 6.1: Costo promedio de envíos de productos pequeños y medianos.
Fuente: Diario La Tercera.

	Precio mínimo	Precio máximo	Promedio
Productos medios	\$300	\$15.700	\$8.000
Productos pequeños	\$700	\$9.100	\$4.900
		Promedio general	\$6.450

De esta tabla se puede considerar el promedio simple de cada categoría y con esto obtener un promedio general para un envío realizado en la región metropolitana el cual es de \$ 6.450. Considerando aspectos como los costos adicionales de transporte, incluyendo gastos de combustible y horas hombre. Bajo esta premisa, y tomando en cuenta los siguientes supuestos:

- Un costo promedio por error de 6.450 pesos chilenos.
- Un volumen anual de procesamiento de 1 millón de nuevas direcciones, entendiendolas como direcciones que no estaban registradas en el robot de direcciones de la empresa.
- Una tasa de error del 39 % para el modelo de programación tradicional para normalizar direcciones.
- Una tasa de error del 37 % para el modelo de machine learning para normalizar direcciones.

Se procede a estimar la cantidad anual de errores para cada modelo:

$$1.000.000 \times 39 \% = 390.000 \text{ (errores utilizando el modelo tradicional)}$$

$$1.000.000 \times 37 \% = 370.000 \text{ (errores utilizando el modelo propuesto)}$$

El costo total estimado asociado a los errores para cada uno de los modelos sería:

$$390.000 \times \$6.450 = 2.515.500.000 \text{ CLP (modelo tradicional)}$$

$$370.000 \times \$6.450 = 2.386.500.000 \text{ CLP (modelo propuesto)}$$

De este modo, la implementación del modelo de machine learning implicaría un ahorro estimado de:

$$2.515.500.000 - 2.386.500.000 = 129.000.000 \text{ CLP}$$

Sin considerar la revisión del diccionario, sino solamente el millón de nuevas direcciones. Para una evaluación más precisa del impacto económico, resultaría idóneo contar con acceso a los datos históricos de la empresa, analizando detalladamente la frecuencia y naturaleza de los errores de despacho, así como los costos asociados reales.

6.2. Trabajos futuros

Aunque los resultados obtenidos para el modelo de machine learning desarrollado en el presente trabajo son relativamente buenos, es crucial reconocer las limitaciones del estudio

actual. El uso de una base de datos ficticia plantea preguntas sobre la aplicabilidad de los modelos en situaciones del mundo real. Además, la exclusión de elementos como la numeración y la comuna puede limitar la utilidad de los modelos en aplicaciones prácticas. Para futuras investigaciones, sería beneficioso incorporar una base de datos real más grande y más diversa, lo que permitiría evaluar la efectividad de los modelos en un espectro más amplio de casos. Además, se reconoce la limitación de este estudio en cuanto a la simplificación de la normalización únicamente a nombres de calles, avenidas y pasajes, sin considerar numeración y comunas, que fue adoptada por eficiencia temporal. Se destaca la importancia de futuras investigaciones que aborden este proceso de manera más integral, incorporando datos reales y considerando una variedad más amplia de elementos en las direcciones.

Para la sección de “Trabajos futuros” de este estudio sobre la normalización de direcciones, se presentan dos vías principales de investigación y desarrollo, considerando las tecnologías emergentes y la continuación del enfoque actual:

1. Innovación con Chatbots de IA

Una recomendación clave para trabajos futuros es la exploración y desarrollo de una API que integre un chatbot de IA para la normalización de direcciones. Esta tecnología avanzada, alineada con las últimas tendencias en inteligencia artificial, ofrece una prometedora solución para procesar y entender el lenguaje natural de manera eficiente. La propuesta específica consiste en desarrollar una API que funcione como un contenedor para invocar un chatbot de IA dedicado a la tarea de normalización. Esta API de normalización de direcciones basada en IA se podría luego integrar dentro de una API más amplia, facilitando su incorporación en el sistema actual de la empresa sin la necesidad de modificaciones extensas. Esta estrategia no solo representa un avance tecnológico sino que también promete mejorar la eficacia y eficiencia en la normalización de direcciones, adaptándose a nuevas variaciones y estructuras en los datos de direcciones.

2. Continuación del enfoque actual con mejoras

En paralelo, para aquellos interesados en continuar con el enfoque de aprendizaje supervisado adoptado en este trabajo, se recomienda la implementación de modelos que utilicen bases de datos más grandes y representativas, integrando técnicas de aprendizaje por refuerzo. El aprendizaje por refuerzo podría permitir a estos modelos mejorar continuamente su rendimiento a través de la interacción y el feedback del entorno, optimizando la toma de decisiones para la normalización de direcciones. Además, se propone encapsular el modelo de aprendizaje supervisado mejorado dentro de una API, lo que facilitaría su integración y uso en los sistemas existentes de la empresa. Esta aproximación ofrece un balance entre la innovación tecnológica y la aplicación práctica, asegurando que el modelo se mantenga relevante y efectivo frente a los desafíos emergentes.

Bibliografía

- [1] Organización Mundial del Comercio, “Comercio electrónico, comercio y la pandemia del COVID-19,” 2020, https://www.wto.org/spanish/tratop_s/covid19_s/ecommerce_report_s.pdf.
- [2] Hoel, B., “La COVID-19 cambia para siempre el modo en que compramos por internet,” 2020, <https://news.un.org/es/story/2020/10/1482072>.
- [3] Olivares, B., “Se extiende el ciclo negativo del e-commerce en Chile: Ventas anotan contracción de 18 % en febrero,” Emol, 2023.
- [4] Baeza, M., “¿Una burbuja logística? Crece la feroz competencia en la industria de envíos y despachos,” Diario Financiero, 2022.
- [5] Zúñiga, T., “Empresas de última milla se preparan para repunte en segundo semestre con mejoras en logística y tiempos de entrega,” Diario Financiero, 2023.
- [6] Baeza, M. Espinoza, M., “Copec entra al negocio de la última milla tras compra de Blue Express y buscará ampliar su cobertura,” Diario Financiero, 2022.
- [7] Blue Express, “Centro de ayuda,” 2023, <https://www.blue.cl/centro-de-ayuda/>.
- [8] Delgado, N., Martínez, G., Covas, D., “Procedimiento para la mejora del servicio de envíos de mensajería DHL express, perteneciente a la empresa de correos Cienfuegos,” Visión de futuro, 2014.
- [9] Küçük, D. Avdan, U., “Address standardization using the natural language process for improving geocoding results,” Computers, Environment and Urban Systems, 2018.
- [10] Rivas, D., “Propuesta de guía metodológica para el manejo de problemas en la estandarización y calidad de datos de direcciones urbanas en colombia,” 2016, <https://repository.upb.edu.co/handle/20.500.11912/2839>.
- [11] Data Ladder, “Guía rápida para la normalización y verificación de direcciones,” 2022, <https://dataladder.com/es/guia-rapida-para-la-normalizacion-y-verificacion-de-direcciones/>.
- [12] Muñoz, A., “¿por qué es importante la normalización de bases de datos?,” 2023, <https://blog.saleslayer.com/es/por-que-es-importante-la-normalizacion-de-base-de-datos>.
- [13] Ladder, D., “Guía de normalización de datos: Tipos, beneficios y proceso,” 2022, <https://dataladder.com/es/guia-de-normalizacion-de-datos-tipos-beneficios-y-proceso/>.
- [14] Bruce, P., Estadísticas prácticas para científicos de datos: 50 conceptos esenciales. O’Reilly Media, 2017.
- [15] Unir, “El aprendizaje automático: objetivos y ejemplos,” 2022, <https://www.unir.net/ingenieria/revista/machine-learning/>.

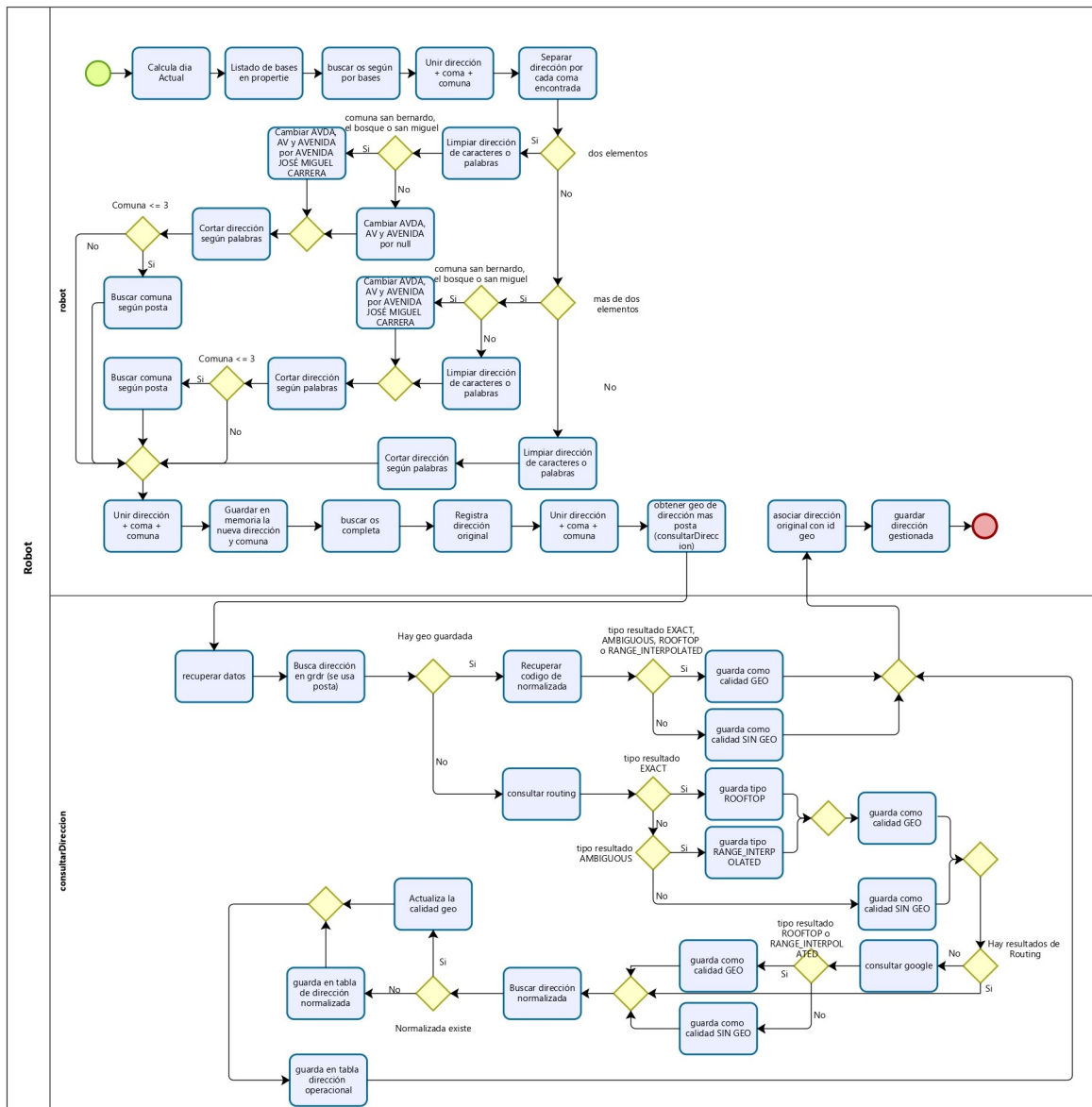
- [16] IBM, “¿Qué es el aprendizaje supervisado?,” 2023, <https://www.ibm.com/es-es/topics/supervised-learning>.
- [17] Services, A. W., “¿qué es el procesamiento de lenguaje natural?,” 2023, <https://aws.amazon.com/es/what-is/nlp/>.
- [18] IBM, “Procesamiento del lenguaje natural,” 2023, <https://www.ibm.com/es-es/topics/natural-language-processing>.
- [19] Román, V., “Algoritmos naive bayes: Fundamentos e implementación,” 2019,
- [20] Microsoft, “Algoritmo Naive Bayes de Microsoft,” 2023, <https://learn.microsoft.com/es-es/analysis-services/data-mining/microsoft-naive-bayes-algorithm?view=asallproducts-allversions>.
- [21] IBM, “¿Qué son las redes neuronales?,” 2021, <https://www.ibm.com/es-es/topics/neural-networks>.
- [22] IBM, “¿Qué es KNN?,” 2023, <https://www.ibm.com/mx-es/topics/knn>.
- [23] IMF, “¿qué es la matriz de confusión y cómo interpretarla?,” 2021, <https://blogs.imf-fomacion.com/blog/tecnologia/matriz-confusion-como-interpretarla-202106/>.
- [24] Heras, J., “Precision, recall, f1, y accuracy en clasificación,” 2020, <https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/>.
- [25] Google for Developers, “Classification: Precision and recall,” 2023, <https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall>.
- [26] Ponce, D., “Sistema de búsqueda inteligente de direcciones para empresa de distribución postal,” 2023, <https://repositorio.uchile.cl/bitstream/handle/2250/181650/Sistema-de-busqueda-inteligente-de-direcciones-para-empresa-de-Distribucion-Postal.pdf?sequence=1>.
- [27] Amazon Web Services, “¿qué es una interfaz de programación de aplicaciones (api)?,”
- [28] Caslop, P., “Manual crisp-dm de ibm spss modeler,” 2017, https://www.academia.edu/34994436/Manual_CRISP_DM_de_IBM_SPSS_Modeler.
- [29] Perez, M., “Minería de datos. la metodología crisp-dm de ibm,” 2014.
- [30] Ranzijn, B., “A geocoding algorithm based on a comparative study of address matching techniques. erasmus universiteit rotterdam,” 2013, <https://thesis.eur.nl/pub/14891/Ranzijn.pdf>.
- [31] Código Postal Co, “Calles de Cerrillos - Chile,” 2023, <https://codigo-postal.co/chile/santiago/calles-de-cerrillos/>.
- [32] Bruzzone, L Persello, C., “A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples.,” 2009, <https://scite.ai/reports/10.1109/tgrs.2008.2011983>.
- [33] Madhu, B., Rahman, M, Mukherjee, A., Islam, Z., Roy, R. Ershad, L., “A comparative study of support vector machine and artificial neural network for option price prediction,” 2021, [doi:10.4236/jcc.2021.95006](https://doi.org/10.4236/jcc.2021.95006).
- [34] Schittkowski, K., “Optimal parameter selection in support vector machines,” Journal of Industrial and Management Optimization, 2005.
- [35] Beetrack, “6 problemas de distribución logística de productos [última milla].”, <https://>

[//www.beetrack.com/es/blog/logistica-de-distribucion/](http://www.beetrack.com/es/blog/logistica-de-distribucion/).

- [36] Aravena, S., “Estudio revela que tiempo promedio de envío de productos crece en torno a 40

Anexos

Anexo A. Proceso de normalización actual



Anexo B. Calles de la comuna de Cerrillos

- 1 de Julio
- 12 de Febrero
- 14 de Octubre
- 15 de Agosto
- 16 de Julio
- 19 de Septiembre
- 21 de Mayo
- 24 de Diciembre
- 25 de Septiembre
- 5 de Abril
- Acapulco
- Acceso Camino a Lonquén
- Acceso Camino A Melipilla
- Aeropuerto
- Aeropuerto Melinka
- Aeropuerto Rodelillo
- Alelí
- Aluminio
- América Indígena
- Armonía
- Arzobispo Errázuriz
- Avenida 2
- Avenida 3
- Avenida 4
- Avenida 5
- Avenida 7
- Avenida 8
- Avenida Américo Vespucio
- Avenida Camino a Melipilla
- Avenida Central Cardenal Raúl Silva Henríquez
- Avenida del Ferrocarril
- Avenida Departamental
- Avenida Esquina Blanca
- Avenida Fernández Albano
- Avenida General Velázquez
- Avenida Lo Errázuriz
- Avenida Los Cerrillos
- Avenida Padre Hurtado
- Avenida Pedro Aguirre Cerda
- Avenida Presidente Salvador Allende
- Avenida Suiza
- Belo Horizonte
- Berilio
- Brasil
- Buzeta
- Caletera Avenida Pedro Aguirre Cerda
- Calle 1
- Calle Nueva Laredo
- Camino a Lonquén
- Camino a Melipilla
- Cardenal José María Caro
- Cardenal Juan Francisco Fresno
- Cardenal Raúl Silva Henríquez
- Carriel Sur
- Caupolicán
- Chamiza
- Ciudad Madero
- Colo-Colo
- Colombia
- Comercio
- Concejala Berta Carrvajal
- Conde De Orgaz
- Corneta Crispin Reyes
- Costa Rica
- Cristóbal Colón
- Deuterio
- Diputado Angel Fantuzzi
- Dir María Troncoso Carrasco
- Dirigente Víctor Cifuentes
- Divina Comedia

- Doctor Carlos Aranda
- Doctor Santibáñez
- Doctor Vargas Salcedo
- Domingo Colón
- Domínica
- Ejército Libertador
- El Álamo
- El Almendro
- El Aromo
- El Atomo
- El Esfuerzo
- El Pinar
- El Sendero
- El Universo
- Elicura
- Ernesto Cea
- España
- Estonia
- Farellones
- Félix Margoz
- Fernández Albano
- Florida
- Francia
- Fresia
- Gabriela Mistral
- Galileo
- Groelandia
- Guadalajara
- Guayana
- Haití
- Hector Orrego
- Héroes de Maipú
- Honduras
- Huelen
- Inés Aguila Ledezma
- Ingeniero Juan Ceron Reyes
- Jamaica
- José Miguel Carrera
- Julio Verne
- La Herradura
- La Pinta
- La Primavera
- La Unión
- La Ventura
- Las Dalias
- Las Encinas
- Las Galaxias
- Las Hortensias
- Las Hualtatas
- Las Palmas
- Las Paulonias
- Las Rosas
- Las Violetas
- Lincoyán
- Lituania
- Lo Errázuriz
- Los Almendros
- Los Aromos
- Los Aztecas
- Los Castaños
- Los Ceibos
- Los Clarines
- Los Clasicos
- Los Claveles
- Los Copihues
- Los Fundadores
- Los Geranios
- Los Jazmines
- Los Lirios
- Los Pinos
- Los Presidentes
- Los Rosales
- Los Vencedores
- Lucerna
- Manquehue
- Manutara
- Margarita García Toro
- Martín de Solier
- Mérida
- México
- Mi Esperanza
- Miguel Socías
- Miguel Zaror
- Monseñor José María Caro
- Monte Carmelo
- Morella

- Morelos
- Nicaragua
- Nueva 1
- Obispo Javier Vásquez
- Ongolmo
- Osa Mayor
- Pablo Suárez
- Padre Diego Sandoval
- Padre Manuel Lacunza
- Paicaví
- Paraguay
- Pasaje 29 De Junio
- Pasaje Amazonas
- Pasaje Atacameños
- Pasaje Colombia
- Pasaje Cultura Molle
- Pasaje El Verano
- Pasaje Los Conquistadores
- Pasaje Onas
- Paso Bajo Nivel Camino a Lonquén
- Patricio Acevedo
- Pedro Aguirre Cerda Interior
- Pedro Lagos Palacios
- Peulla
- Piloto Acevedo
- Piloto Jacques Lagas
- Piloto Lazo
- Plaza Norte
- Polonia
- Portugal
- Primo de Rivera
- Principal
- Progreso
- Puangue
- Rapa Nui
- Reinosa
- Renacimiento
- Río Claro
- Río Grande
- Río Loa
- Río Manso
- Río Mayo
- Río Toltén
- Román Salinas
- Rosa Ester Rodríguez
- Rubén Darío
- Salida Camino a Lonquén
- San Andrés
- San José
- San Juan
- Santa Ana
- Santa María
- Teatro Metro
- Transversal
- Trinidad y Tobago
- Ucrania
- Uruguay
- Valdivieso
- Venezuela
- Vista Clara
- Vista Hermosa
- Yucatán