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

SISTEMA DE BÚSQUEDA INTELIGENTE DE DIRECCIONES PARA EMPRESA DE DISTRIBUCIÓN POSTAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

DANIEL ESTEBAN PONCE MARIPANGUI

PROFESORA GUÍA:
ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN:
RODOLFO URRUTIA URIBE
MATÍAS TOBAR GIGOUX

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: INGENIERO CIVIL INDUSTRIAL

POR: DANIEL ESTEBAN PONCE MARIPANGUI

FECHA: MARZO 2021

PROFESORA GUÍA: ALEJANDRA PUENTE C.

SISTEMA DE BÚSQUEDA INTELIGENTE DE DIRECCIONES EN EMPRESA DE DISTRIBUCIÓN POSTAL

El rubro postal es una de las actividades más antiguas del mundo, su primer uso documentado se encuentra en Egipto en el 2400 a.C. cuando los faraones utilizaban mensajeros para enviar decretos por todos los territorios del Estado. Actualmente, el mercado postal experimenta variaciones a medida que las exigencias de los clientes van cambiando. A nivel mundial, el mercado crece alrededor de un 4% anual y Chile no se ha quedado atrás. Según datos del INE 2018 el segmento de paquetería (paquetes y encomiendas) en Chile crece a un 8,1% anual, mientras que el segmento postal (cartas y documentos) se contrae a un 7,5% anual. Esto revela que los clientes están cambiando el uso de los servicios de correo, dejando de enviar cartas y documentos y aumentando el envío de encomiendas.

El aumento en el flujo de correos, la competencia creciente y los clientes cada vez más exigentes han demandado a las empresas optimizar sus procesos para asegurar un buen nivel de servicio. Por esta razón, los servicios de correos han implementado el código postal para optimizar el proceso de clasificación y distribución, el cual disminuye el tiempo y la posibilidad de cometer errores en el destino de los envíos. El proceso de asignación de un código postal a una dirección es llamado normalización. Sin embargo, en empresas que no poseen un sistema automático de normalización, la asignación de códigos la realizan operadores manualmente, lo que conlleva un costo mensual de más de $26 millones de pesos. Este costo puede ser mayor si se considera el mayor tiempo del envío en el sistema y el reproceso en caso de la corrección de un envío erróneo.

En esta memoria, se propone como solución un sistema de búsqueda inteligente de direcciones postales que permita normalizar automáticamente grandes volúmenes de direcciones utilizando modelos de procesamiento del lenguaje natural.

La solución propuesta plantea un clasificador de direcciones para segmentar y etiquetar sus atributos (nombre de calle, número principal e información adicional). El clasificador, además, verifica si posee un número principal bien definido, si no lo posee la dirección se identifica como inválida y no se normaliza, debido que el código postal requiere reconocer un frente de cuadra para poder ser asignado. Con las direcciones etiquetadas, se aplica un modelo coincidencia de texto utilizando la distancia de Levenshtein y el ratio de similitud de Levenshtein, generando un ranking con las tres direcciones más similares en base al *puntaje ranking*, para luego calcular el *puntaje de selección.* Finalmente, la dirección es normalizada si se cumplen los criterios de asignación. El modelo de coincidencia de texto con mejor rendimiento elimina abreviaturas que enuncian un tipo de calle (por ejemplo: *PSJE* o *AVDA),* calcula el puntaje ranking utilizando el ratio de similitud de Levenshtein y busca la coincidencia comparando la cadena completa de texto. Finalmente, se genera categorías de confiabilidad según el *puntaje de selección* para disminuir el error de la normalización

Al normalizar envíos de Quilicura a través de coincidencia directa, se logra un 20% de normalización. Al aplicar la solución propuesta, se alcanza sobre un 90% de normalización con un error asociado menor a 5%. Al utilizar las categorías de confiabilidad, se alcanza sobre 80% de normalización con un error asociado menor a 1%.

Esta memoria se limita a normalizar las direcciones a nivel de comuna y en zonas urbanas. Además, se entrenan los modelos utilizando solo direcciones de la comuna de Quilicura y Santiago.

DEDICATORIA

A mi madre y a mi padre por estar siempre presentes….

AGRADECIMIENTOS

Para finalizar esta etapa universitaria quiero agradecer a todas las personas que me han acompañado, en mayor o menor medida, en este proceso.

Quiero agradecer a mi madre María Eliana, y a mi padre Marcos, por darme todas las oportunidades para llegar a este punto, por sus consejos y su apoyo incondicional, sin su ayuda mi presente sería muy diferente, muchas gracias a ambos. Asimismo, quiero agradecer a mi tía Cecilia, por su amor incondicional, por su cariño a todos quienes la rodean, por mirar la vida con alegría. A mi hermano Marcos y hermana Daniela, de quienes he aprendido cosas que no aprenderé tomando clases y que a pesar de nuestras diferencias hemos estado siempre apoyándonos en los momentos importantes.

También, quiero agradecer a Gladys Lobos por haber estado a mi lado durante este proceso, por darme ánimo y confiar en mí. Muchas gracias por los consejos, el cariño y la paciencia. Definitivamente la pandemia de Covid-19 hizo más difíciles las cosas, y es necesario decir que tu apoyo fue clave para no perder el norte.

No puedo dejar de agradecer a mis amigos y amigas de la universidad, en especial a Sebastián Silva y Gerardo Sepúlveda quienes estuvieron presente durante casi toda la carrera a mi lado y espero que lo sigan haciendo en un futuro. También, quiero agradecer al equipo del CEIN 2019, tremendo equipazo, agradecido de haber compartido y aprendido de cada uno de ustedes.

Además, quiero agradecer a Alejandra Puente (Profesora Guía), a Rodolfo Urrutia (Profesor Co-Guía) y Rodrigo Pizarro quienes me apoyaron durante el desarrollo del trabajo de memoria, aportándome con nuevos puntos de vistas, soluciones a problemas específicos y por su disposición a ayudarme durante todo el proceso.

Finalmente, agradezco a las personas que conforman el departamento de ingeniería industrial por escuchar a los estudiantes, empoderarlos y ayudar a generar cambios radicales que, de una manera u otra, marcan tendencia a nivel facultad. Durante los años que estuve estudiando siempre hubo problemas que solucionar y nunca dejaron de haber cambios, siempre se trabajó por hacer un mejor departamento del cual me enorgullezco de haber pertenecido.

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE FIGURAS

1 INTRODUCCIÓN

1.1 Caracterización de la Industria

1.1.1 Rubro

A medida que las civilizaciones fueron abarcando territorios más extensos, mantener una buena comunicación era esencial para su organización la cual se llevaba a cabo a través de mensajeros, siendo análogos al sistema postal moderno. De hecho, el primer uso documentado de mensajería organizada data del 2400 a.C en Egipto cuando los faraones utilizaban mensajeros para comunicar sus decretos por todos los territorios del Estado (Casanova, 2018). Este sistema ha evolucionado con el fin de transportar mensajes, documentos escritos, encomiendas y/o cualquier otro paquete desde un lugar a otro. En la actualidad, la función principal de correos se ha mantenido, pero ajustándose en cada momento a las exigencias de sus usuarios y a la tecnología.

El mercado mundial Postal, que incluye cartas y paquetería de los miembros de la Unión Postal Universal (UPU), creció a tasa promedio anual de 4,0% entre 2009 y 2015 (Ingresos en dólares), empujado principalmente por Asia-Pacífico. El 93% del mercado postal corresponde a países industrializados y Asia-Pacífico (Empresa de Correos de Chile; BBVA Asesorías Financieras S.A., 2017)

1.1.2 Industria Postal en Chile

En Chile, esta tendencia se ve atenuada en el negocio postal y amplificada en paquetería, particularmente en los últimos años, lo que evidencia un cambio en la composición del negocio.

La principal empresa de correos en el país es “Correos de Chile”, sin embargo, existen varias otras empresas en el mismo rubro: WSP y Chilepost en el segmento postal; ChileExpress, Starken, DHL, BlueExpress y otras en el segmento Express. En ambos negocios, el principal criterio de compra de los clientes es el nivel de servicio, y más precisamente, la entrega a tiempo.

1.1.3 Productos y/o Servicios

Actualmente los servicios postales se segmentan según el tipo de servicio que ofrecen:

- **Segmento Postal Nacional:** Corresponde principalmente al envío de cartas. Se caracteriza por ser, en su mayoría B2C (Business to Customer) y B2B (Business to Business). El 80% de los envíos corresponden a clientes institucionales y el 20% a clientes personas.

  El mercado chileno de correspondencia Postal corresponde a documentos de menos de 500 gramos con niveles de servicio que involucran entregas en 3 o más días, se caracteriza por ser un mercado muy concentrado, con importante participación de Correos de Chile. Otros proveedores de estos servicios incluyen Chilepost, WSP, Envía, Chile Parcels.

  Los drivers de crecimiento para el negocio postal están limitados. Se espera que la sustitución digital reemplace gradualmente parte de los envíos que las empresas realizan a sus clientes (Empresa de Correos de Chile; BBVA Asesorías Financieras S.A., 2017)

- **Segmento Internacional:** Corresponde al envío de cartas y paquetes desde y hacia otros países. Las empresas que participan en este segmento funcionan bajo los estándares de la Unión Postal Universal (UPU). Este tipo de envío es el que ha experimentado el mayor crecimiento en los últimos años debido al explosivo aumento del comercio electrónico transfronterizo, donde el principal actor es el mercado asiático.
- **Segmento Express Nacional:** Corresponde principalmente al envío de paquetería, que corresponde principalmente a la distribución de documentos expresos menores a 500 gramos (courier documento, mensajería) y paquetería en formato express o normal menor a 50 Kg. (encomienda, courier paquete, valija comercial, e-commerce). Estos servicios son ofrecidos a nivel nacional e internacional.

  Se caracteriza por tener un crecimiento acelerado debido al aumento del comercio electrónico, junto con esto, la mayor cobertura de internet en diferentes zonas ha planteado el desafío de llegar cada vez a zonas más extremas del país que comienzan a requerir del servicio de paquetería. Los clientes en este segmento se identifican por ser exigentes con los plazos de entrega, demandando mejores servicios logísticos de apoyo y distribución (Accenture, 2018).

Los principales actores de este mercado son: Chilexpress, Bluexpress, Turbus, Correos de Chile, WSP Express y Lit-Cargo. Este es un mercado más diversificado y es el que enfrenta el mayor aumento de demanda.

1.1.4 Clientes

El comercio por internet ha ido creciendo en los últimos años y con esto el envío de paquetes. "Antes de la pandemia de Covid-19, los envíos a domicilios en canales de e-commerce representaban entre un 60% y 70% de las ventas online y con clientes cada vez más exigentes que buscan recibir lo antes posible sus productos la industria de paquetería se vuelve un pilar fundamental" (Diario Financiero, 2020).

La velocidad y la eficiencia son los motores clave para las compras en línea, se espera que las plataformas de comercio electrónico sean capaces de entregar sus productos el mismo día o incluso dos horas posteriores a la compra en zonas urbanas. Según estudio de Adimark (2019), las condiciones básicas más importantes para los clientes con los que debe cumplir el comercio electrónico son: despachos rápidos (58%), políticas de devolución simple (48%) y monitoreos del estado de envío (48%) lo que afecta directamente a los servicios de paquetería exigiendo una mejor logística y calidad de servicio (BareInternational.cl, 2019).

Adicionalmente existe una tendencia de cambio en los tipos de envíos, con el despacho de documentos (segmento postal) en descenso y el envío de paquetes (segmento express) en aumento debido al e-commerce. El gráfico 1 muestra la evolución en la cantidad de envíos de documentos (cuantificado en número de documentos) versus los envíos de paquetería (en kilogramos) desde enero 2010 a diciembre 2018. En este periodo se produjo un aumento anual promedio de 8,1% en el segmento express y una contracción del 7,5% para el segmento postal. En pocas palabras, los clientes de servicios postales están cambiando su uso, reemplazando las cartas tradicionales por paquetes más grandes que implican nuevas exigencias para los sistemas de correspondencia.

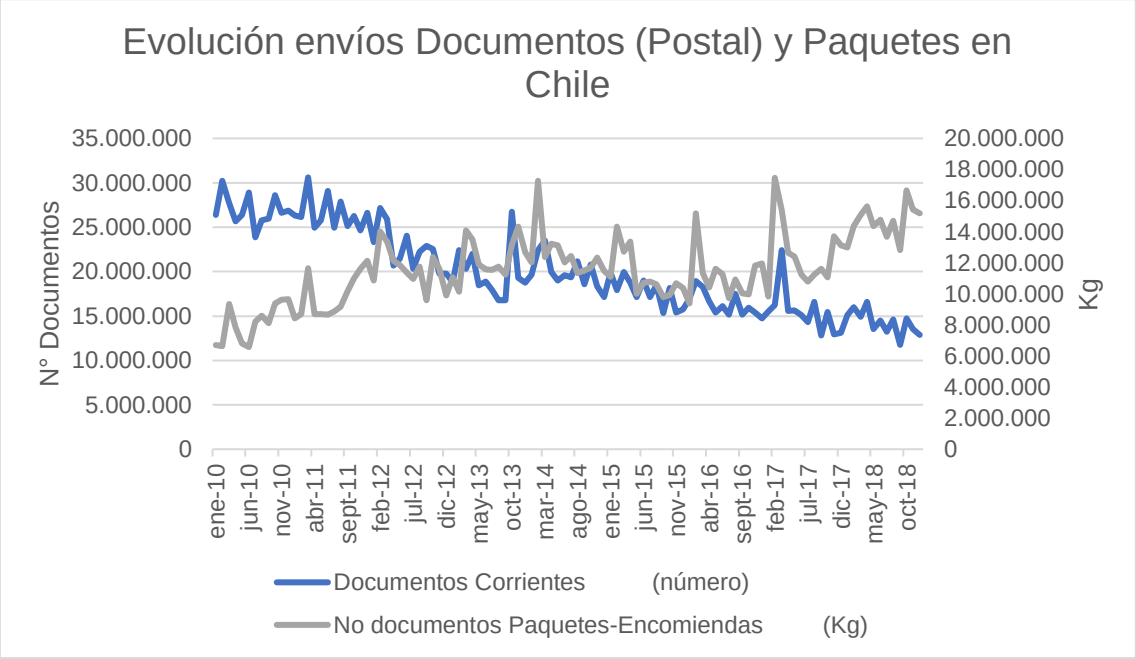

Figura 1. Evolución mercado postal en Chile. Fuente: INE 2018

1.1.5 Dimensionamiento de actividad realizada

Para poder dimensionar la actividad realizada por una empresa de correos se utilizará como ejemplo a Correos de Chile, debido a que es una compañía que publica su memoria anual según la Ley de Transparencia y por lo tanto existe información fidedigna al respecto. Cabe mencionar que es la empresa de correos más grande en Chile.

Correos de Chile es una empresa pública fundada en 1736, dedicada al servicio de correspondencia, giros postales y al mercado de envíos y encomiendas nacionales e internacionales, cumpliendo con las funciones de Servicio Postal Universal. En sus más de 270 años de existencia, Correos de Chile se ha perfeccionado, siendo reconocida dentro de las mejores empresas de correos a nivel internacional. Está presente en todo Chile, con 227 sucursales desde Arica hasta Puerto Natales incluyendo Isla de Pascua y Juan Fernández (Empresa de Correos de Chile, 2019).

Durante 2019, Correos de Chile percibió ingresos por ventas de MM$111.095, un 4.8% más que en 2018 lo cual no fue suficiente para cubrir los costos del periodo, lo cual se debe a las dificultades políticas que sacudieron a Chile durante 2019. Durante el mismo año se observó un aumento del 22% en envíos de paquetería respecto a 2018, lo que significa un total de 29 millones de envíos tanto nacional e internacional. Por otro lado, en 2019 se alcanzó el récord de 112.000 envíos de procesamiento diario en la planta de clasificación de paquetería, alcanzando el 91% de su capacidad teórica.

En cuanto a sus clientes, Correos de Chile los divide en institucionales, retail e internacionales:

Clientes Institucionales: En este grupo se identifican 3 tipos de clientes, las Grandes Cuentas postales (Bancos, AFP, ISAPRES, autopistas) con casi 160 millones de envíos, Estado (Ministerios, Servicios Públicos y Municipalidades) con casi 30 millones de envíos y el e-Commerce (Retail, grandes, medianos y pequeños e-Commerce y MarketPlace) con casi 1.5

millones de envíos anuales. En relación con 2018, los segmentos Grandes Cuentas y Estado sufrieron una baja del 13% y el segmento e-Commerce un aumento del 16%.

Clientes Retail: Aquí están los Clientes Nacionales de Paquetería y Documentos (clientes que ingresan a sucursal) con casi 8.3 millones de envíos anuales, Clientes de Casillas (clientes que poseen casilla de recepción, pueden ser personas naturales, instituciones o e-Commerce) con casi 30.000 envíos anuales y Clientes Migrantes (intensivo uso de giros y paquetería internacionales) con 13.700 envíos anuales. Este grupo de clientes experimentó una disminución de 10% en envíos respecto del año 2018.

Clientes Internacionales: En este grupo se encuentran los Clientes Privados (empresas que envían desde / hacia el extranjero) con casi 5.7 millones de envíos anuales y Clientes UPU (cliente de tipo normativo, bajo definiciones de la Unión Postal Universal) con más de 18 millones de envíos anuales. Este grupo de clientes experimentó un crecimiento de 17%, equivalente a 3,4 millones de envíos adicionales respecto del año 2018.

1.2 Mercado y/o Marco Institucional

1.2.1 Actores

Los principales actores de la industria corresponden a las empresas de correos, las cuales se dividen según el servicio que ofrecen y entre ellos se consideran competidores. En la figura 2 se aprecian los principales actores de la industria postal.

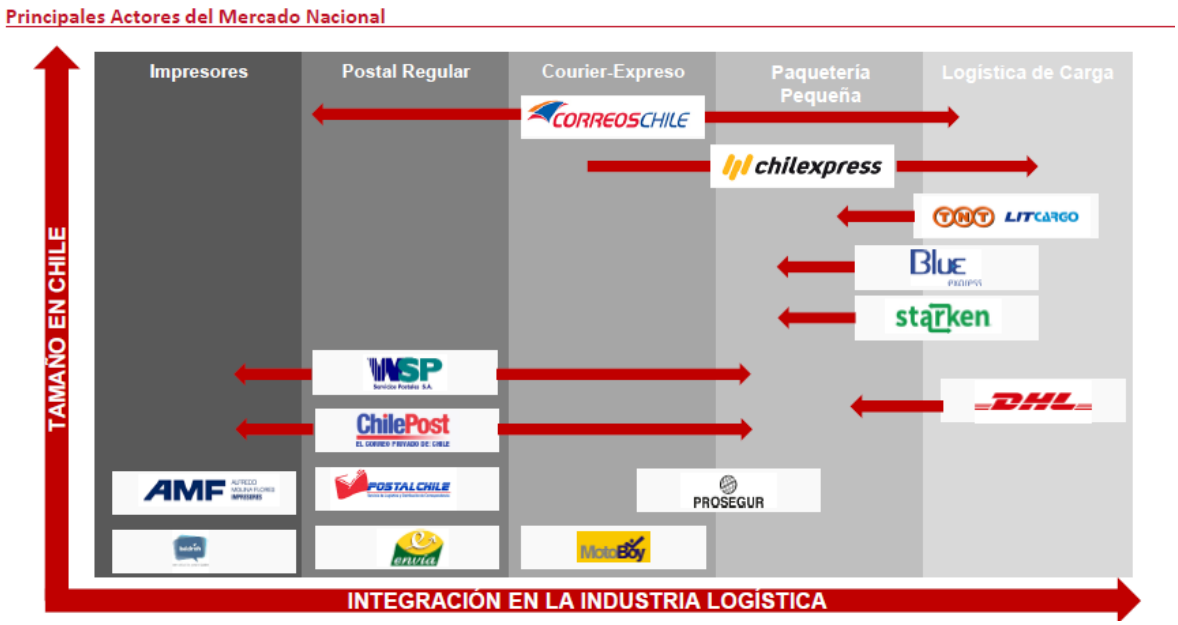

Figura 2: Mercado Postal Chileno. Fuente: Correos de Chile 2019

En cuanto a los proveedores, la empresa Correos de Chile informa que el 97% de sus proveedores corresponde a PyMES nacionales que representan un 79% del total gastado y el 3% restante a proveedores internacionales a los cuales se les destina un 21% del total gastado (Empresa de Correos de Chile, 2019). Por último, se menciona que ninguna compra supera el 10% del presupuesto.

El principal regulador del sector postal a nivel mundial es la Unión Postal Universal (UPU), corresponde a la segunda organización internacional más antigua del mundo, siendo fundada en 1874 con sede en Berna, Suiza.

Para mantener el territorio postal global, la UPU establece las reglas para el intercambio de correo internacional entre sus estados miembros. También proporciona asistencia técnica, asesorando sobre la mejora de la calidad de los servicios postales y estimulando el crecimiento del volumen de correo para ayudar al desarrollo socioeconómico de las naciones. La red postal pública es la red de distribución física más grande del mundo con más de 640,000 puntos postales. También es uno de los empleadores más grandes con unos 5,5 millones de empleados. En general, procesaron y entregaron 350 mil millones de artículos postales nacionales e internacionales y más de 6 mil millones de paquetes a nivel internacional en 2012. Muchas publicaciones también brindan a los clientes correo urgente, servicios financieros postales y una gama de servicios electrónicos (Universal Postal Union, 2013).

1.2.2 Marco Normativo

En Chile existen decretos específicos que regulan el funcionamiento de los servicios postales (Empresa de Correos de Chile, 2020). El 14 de febrero de 1957 se crea el Decreto N°394 que aprueba la creación del *Reglamento para el Servicio de Correspondencia*. Luego, el 4 de noviembre de 1960 se aprueba el Decreto N°5.037 denominado *Decreto Supremo que fija el texto definitivo de la Ley Orgánica del Servicios de Correos y Telégrafos.* Más tarde, el 10 de diciembre de 1980 se aprueba el Decreto N°203 con el cual se aprueba la *Política Nacional Postal* el cual funciona como marco de referencia y guía para el desarrollo de estos servicios dentro del país y con el exterior.

Además, existe la ley N°18.016, publicada el 5 de agosto de 1981 que autoriza al Estado para desarrollar actividades empresariales relacionadas con prestaciones telegráficas, y faculta al presidente de la República para transformar el Servicio de Correos y Telégrafos. Con esta ley se da paso a la creación de la *Empresa de Correos de Chile* a través del Decreto con Fuerza de Ley N° 10 publicado el 30 de enero de 1982 en donde la definen como "un organismo de administración autónoma del Estado, con patrimonio propio, que estará sujeta a la fiscalización de la Contraloría General de la República. Se regirá por las disposiciones del presente decreto con fuerza de ley y sus reglamentos y, en lo no previsto en ellos, por la legislación común. Sus relaciones con el Gobierno se efectuarán a través del Ministerio de Transportes y Telecomunicaciones."

1.2.3 Tendencias del Mercado

El 93% del mercado postal corresponde a países industrializados y Asía-Pacífico, siendo este último el que más impacto tuvo en el crecimiento del mercado debido al gran aumento de envíos desde China. En los últimos años se ha evidenciado un cambio en la composición del negocio de correos en el mundo, migrando del negocio postal al negocio de paquetería, tendencia que estaría influenciada por el aumento del comercio electrónico, el aumento en la accesibilidad de internet y mayor cobertura de lugares de entrega de los servicios de correos. Los envíos postales y de paquetería internacional presentaron un crecimiento del 31,6% en 2016 con respecto a 2015. A continuación, las figuras 3 y 4 muestran la tendencia en el tráfico de envíos en el mundo.

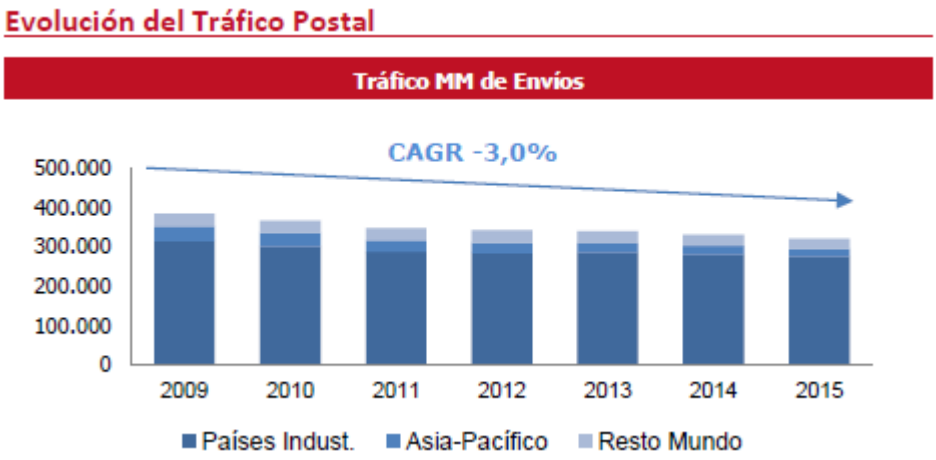

Figura 3. Fuente: Emisión Bonos Corporativos, Empresa Correos de Chile, septiembre 2017

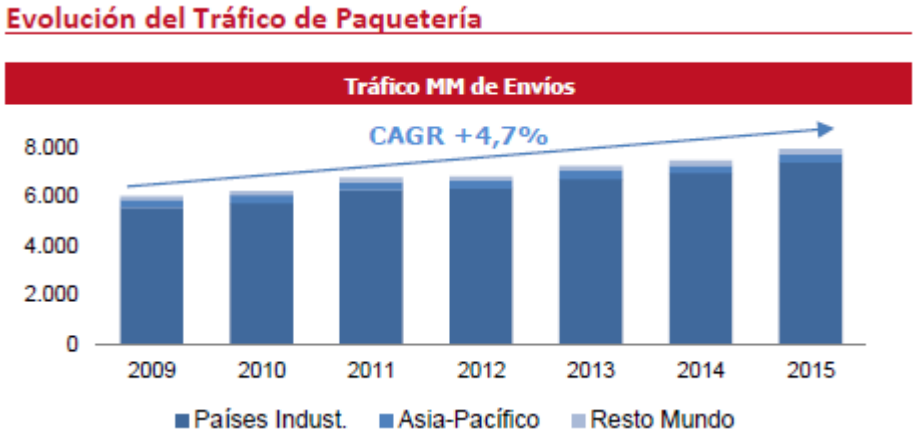

Figura 4. Fuente: Emisión Bonos Corporativos, Empresa de Correos de Chile, septiembre 2017

2 DESCRIPICIÓN DEL PROYECTO

En esta sección, se describen las características generales del presente proyecto. Se contextualiza el proceso postal, sus productos y servicios, clientes. En adición, se describe el problema, la oportunidad encontrada, sus causas y consecuencias. Por último, se explicitan los objetivos y alcances definidos.

2.1 Información general del rubro postal

2.1.1 Proceso Postal

Para entender dónde se sitúa el problema que se abordará en esta memoria es necesario mencionar de manera generalizada los procesos que existen en un servicio postal. A continuación, se detallan cinco partes que componen el sistema de correos (Asiain, 2017):

1. Retiro: esta etapa refiere a la recolección de los envíos, ya sea en sucursales, city box o ubicación del cliente, para llevarlos a su planta destino.

2. Admisión: esta etapa refiere a la admisión de envíos en donde son ingresados al sistema y se preparan para ir al paso de clasificación.

3. Clasificación: esta etapa refiere a la categorización de los envíos realizada en las plantas de clasificación, las cuales identifican los productos y los etiquetan para su entrega final.

4. Distribución: esta etapa refiere a la distribución de los envíos de cara al cliente final a través de los diferentes medios que posea cada empresa, ya sea carteros, móviles o retiro en sucursal.

5. Retorno de Información: esta etapa refiere a la captura de información al momento de entregar los envíos, registrando entre otros datos, la experiencia del cliente cuya opinión es sumamente relevante para la evaluación del proceso completo.

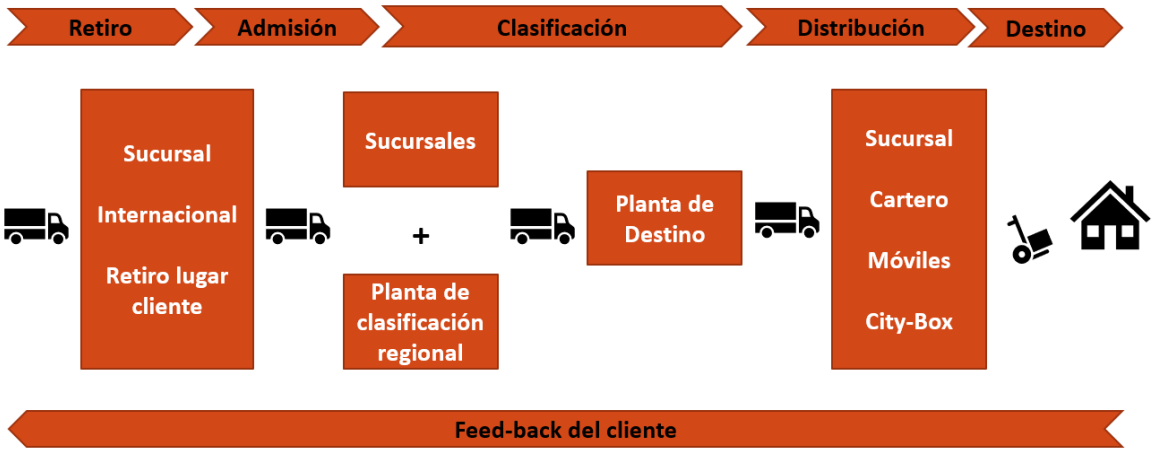

Figura 5: Proceso Postal. Fuente: Elaboración Propia

2.1.2 Productos y servicios

Un servicio de correo puede tener varios productos que responden a distintas necesidades de los clientes. Estos se diferencian según el tamaño del paquete, el tiempo de entrega, si tiene seguimiento o no, según el tipo de mercado (persona o empresa), si se comprueba la recepción (certificada) y si es nacional o internacional. Existen dos grandes tipos de servicios: paquetería y postal, la primera hace referencia a paquetes que utilicen un volumen mayor a una carta y el segundo está enfocado a cartas y documentos.

Como referencia, la empresa de Correos de Chile posee 14 productos distintos en el servicio de paquetería, de los cuales 7 están dedicados al mercado Personas y Empresas, 6 solo al mercado Empresas y 1 solo al mercado Personas. Por otro lado, Correos de Chile posee 9 productos en el servicio postal, de los cuales 5 están dedicados al mercado Personas y Empresas, 3 solo al mercado Empresas y 1 solo al mercado Personas. Parte importante de los productos están enfocados a empresas, ya que son las que más utilizan el servicio de correos.

El proceso específico a analizar corresponde al proceso de admisión y clasificación de la correspondencia y decidir el destino del correo. Esto se realiza a través de una clasificación de sus direcciones para poder definir si: se retira en sucursal y en cuál o si se entrega a domicilio y quién la reparte.

El proceso se complejiza debido a que las direcciones poseen distintos niveles de información lo cual afecta el tiempo de procesamiento. La manera más precisa de escribir una dirección es agregándole el código postal lo cual asegura que el paquete llegue a la cuadra correcta. Por esta razón, parte del proceso de clasificación consiste en asignar un código postal a la dirección entrante con lo que se busca estandarizar la dirección y asegurar su despacho rápido.

2.1.3 Cliente

Como se mencionó, existen varios servicios y productos distintos dependiendo del tipo de mercado al que apuntan. Por la misma razón, se pueden recopilar direcciones de distintas fuentes dependiendo de los remitentes, quienes no siempre conocen al destinatario. Es importante reconocer el origen de las direcciones, ya que, según el tipo de remitente se exigen diferentes niveles de información.

Los clientes que corresponden a empresas no conocen al destinatario y obtienen las direcciones de dos formas: el destinatario entrega su dirección a la empresa de forma voluntaria y con el nivel de detalle que exija dicha empresa; el cual no necesariamente está estandarizado, o la empresa obtiene la dirección del destinatario a través de una base CRM (Client Relationship Management) de clientes, de las cuales, el 90% de las veces no se encuentran estandarizadas. Este caso corresponde al enfoque principal de esta memoria, cuyo mercado es el más grande para las empresas postales con un movimiento anual de más de 190 millones de envíos anuales (Empresa de Correos de Chile, 2019).

Los clientes que corresponden a personas naturales conocen, la mayoría de las veces, al destinatario por lo que pueden entregar una dirección estandarizada y con código postal. Este tipo de clientes no representa un problema mayor para la estandarización de direcciones por lo que no será foco de estudio de esta memoria.

Por último, están los clientes internacionales quienes, no siempre conocen al destinatario, pero exigen el código postal. Este tipo de clientes entrega casi el 100% de sus direcciones estandarizadas lo cual asegura que el paquete llegue, al menos, a la cuadra correspondiente.

2.2 Identificación del problema y oportunidad

2.2.1 Descripción del Problema

El aumento en el flujo de correos, la competencia creciente y los clientes cada vez más exigentes han demandado a las empresas optimizar sus procesos para asegurar un buen nivel de servicio (Universal Postal Union, 2016). Por esta razón, con el objetivo de mejorar la precisión de la identificación de direcciones se utilizan los códigos postales, ya que de esta manera el proceso de clasificación se hace más expedito y disminuye la posibilidad de cometer el error de enviar el correo a un destino equivocado.

En Chile, el código postal posee siete dígitos con los cuales se identifica el lugar geográfico de una dirección. Los tres primeros dígitos corresponden a la comuna y los últimos cuatro dígitos al frente de manzana, por ejemplo: el código postal de Santiago es 832 0000. La empresa de Correos de Chile tiene registradas 114 comunas con nivel de detalle de frente de manzana (es decir, un 90% de las direcciones urbanas) con un código postal. El resto de las direcciones sólo tiene un código postal comunal (Elfenbaum, CCM - Comunidad Informática, 2020) .

El problema surge cuando las empresas no están digitalizadas o no poseen suficiente tecnología para lidiar de manera automática con direcciones que no poseen código postal, ya que esta asignación la lleva a cabo un operador, pudiendo demorar la entrega de paquetes entre 2 días a 1 semana, aumentando los costos y afectando el desempeño del servicio en toda la cadena logística de distribución. En algunas empresas que no poseen sistemas inteligentes de normalización de direcciones, el operador logra asignar un 20% de asignaciones exactas de código postal y a través de la relajación de algunas reglas de coincidencia puede lograr un 50% de asignaciones confiables.

Las direcciones postales tienen características particulares que las diferencian de otras cadenas de texto, además, están conformadas por una estructura implícita, comprendiendo elementos como “calle”, “ciudad” y “código postal”. Sin embargo, el orden de los atributos no es fijo y no todos los atributos están presentes en todas las instancias. En adición, no solo el formato de las direcciones difiere de país a país, sino que incluso varía entre regiones de un mismo país (Marrero, 2007). Estas variaciones complican el proceso de estandarización y asignación de un código postal.

2.2.2 Declaración del Problema

El problema principal por tratar en esta memoria corresponde a la ***asignación de códigos postales a una cantidad significativa de direcciones no normalizadas provenientes de clientes tipo empresas***

2.2.3 Causas

La causa principal que genera este problema es la gran cantidad de variaciones que puede adquirir una misma dirección, ya que cada cliente puede dar las indicaciones que desee al momento de escribir un destino. Por ejemplo: *Avenida Libertador General Bernardo O’Higgins 1371, Santiago,*

Región Metropolitana puede tener variaciones como *Av Libertador Bernardo O'Higgins 1371, Santiago, RM* o *Alameda Libertador Bernardo O'Higins 1371, Stgo, RM* (el error en O'Higins es intencional para demostrar las posibles variabilidades). Estas variaciones pueden ser tanto en forma como en contenido.

Otras dos causas secundarias han sido identificadas. Por un lado, la ausencia de modelos inteligentes de análisis de datos impide estandarizar un mayor porcentaje de direcciones. Por otro lado, no es posible exigir un formato específico de direcciones a los clientes, ya que de exigirse se corre el riesgo de que el cliente cambie de proveedor de servicio y la calidad del servicio se vea afectada (Delgado, Martínez, & Covas, 2015).

2.2.4 Consecuencias

Como consecuencias del problema de normalización automática de direcciones se han identificado tres ámbitos que se verían afectados: debilitamiento en posición de mercado, disminución en la satisfacción de los clientes y aumento en los costos de distribución.

La digitalización de las operaciones fundamentales en la industria postal ofrece distintas soluciones con distintos beneficios. La modernización de aplicaciones centrales y de la arquitectura de datos permite crear soluciones de próxima generación y capacidades de análisis predictivo basadas en datos. Diseñar un marco integrado de información ayuda a aumentar la eficiencia al tiempo que aumenta la calidad del servicio y las operaciones dentro de áreas como la personalización y la atención al cliente. Por último, la implementación de algoritmos y análisis avanzados, como la estandarización de direcciones y algoritmos de pareo, ofrece oportunidades de optimización en toda la cadena de valor, desde la recopilación hasta el enrutamiento y la programación, al tiempo que mejora la concentración en el cliente. Esta última, es la oportunidad de mercado que se está perdiendo al no incorporar tecnologías de datos al emparejamiento de la dirección con un código postal, lo que da espacio para que la competencia tome ventaja (BCG, 2017).

La satisfacción del cliente de un servicio postal se ve afectada por cinco variables (Delgado, Martínez, & Covas, 2015):

- Falta de materiales necesarios para la prestación del servicio
- Demora en las respuestas ante reclamos
- Calidad de los envíos cuando llegan al destino final
- Demora en los tiempos de entrega
- Servicios no prestados por falta de información

Estas dos últimas variables son las que se ven afectadas por el problema principal mencionado anteriormente. Primero, no poseer un código postal aumenta los tiempos de reparto entre 2 a 7 días. Y segundo, no se puede exigir un formato y contenido a la dirección que entrega el cliente, ya que supondría una traba en la prestación del servicio.

Las direcciones que no poseen código postal pasan por un proceso de asignación de este, el cual es llevado a cabo por operadores en cada centro de distribución, lo que ralentiza la logística de entrega, manteniéndose más días en el sistema, y finalmente aumentando los costos de operación. Para cuantificar este costo extra, se propone considerar el costo asociado a los operadores que realizan

la normalización a mano, por ejemplo, Correos de Chile posee 21 plantas de clasificación a lo largo del país en las cuales trabajan 2 personas por planta asignando códigos postales. El costo de tener a cada operador en esta función es equivalente a su salario de $620.000 pesos (INE 2019), si la normalización se hiciera a través de un sistema inteligente se podría generar un ahorro de $26,04 millones de pesos mensuales en términos de salario. Es importante recordar que existen otros costos asociados a un mayor tiempo del envío en la empresa (afectando la calidad de servicio) y de reprocesos debido a la corrección de un envío incorrecto.

2.3 Objetivos del Proyecto

2.3.1 Objetivo general

Desarrollar un sistema de búsqueda inteligente de direcciones que permita aumentar el porcentaje de asignación de códigos postales de forma automática, utilizando modelos de procesamiento del lenguaje natural[1].

2.3.2 Objetivos específicos

- Decidir y configurar la estructura de direcciones a utilizar, con el fin de segmentar correctamente las direcciones
- Construir los diccionarios de prefijos de calle, abreviaturas y números para el preprocesamiento de cadenas de texto
- Diseñar enfoques de coincidencia de texto para las direcciones, considerando diferentes métodos de comparación y preprocesamiento de texto
- Encontrar y aplicar mejor enfoque de coincidencia de texto a direcciones de comuna de Santiago

2.4 Alcances

2.4.1 Normalización de direcciones

Existen varias herramientas de normalización de direcciones. Una de ellas es la coincidencia de texto que busca asociar una dirección correcta a través de la similitud de palabras. Otra corresponde a la georreferenciación utilizando tráfico histórico, que asocia la dirección a normalizar con direcciones idénticas entregadas anteriormente y sus coordenadas geográficas de entrega. También, se puede identificar el destinatario y asociar todas las direcciones donde ha recibido envíos.

Cada herramienta de normalización posee beneficios y dificultades, sin embargo, en esta memoria se trabaja la coincidencia de texto para asignar códigos postales. Este trabajo se realiza utilizando solo la información proveniente del texto de la dirección a normalizar y la base oficial de direcciones estandarizadas.

No se plantea como prioridad reconocer cada parte de la estructura o atributo de una dirección (algunos ejemplos de atributos de una dirección: calle, número principal, departamento, block,

[1] El procesamiento del lenguaje natural se centra en el análisis de las comunicaciones humanas y, en concreto de su lenguaje (iic.uam.es/inteligencia-artificial/procesamiento-del-lenguaje-natural/)

parcela, esquina, comuna, provincia, región, código postal), debido a que se limita a la asignación de un código postal, lo cual se realiza identificando el nombre de calle y número principal.

2.4.2 Datos a utilizar

Dado el tiempo acotado para desarrollar el trabajo de memoria, este se limita a utilizar las direcciones de Quilicura en primer lugar, para luego evaluar el modelo con mejor rendimiento en direcciones de la comuna de Santiago. La justificación recae en dos aspectos principales: el primero, la base oficial de direcciones posee 4.026.000 de registros, sin embargo, muchos nombres de calles se repiten entre comunas. Al contar todas las calles con nombres únicos se obtienen 66.000 registros distintos. El segundo corresponde al costo computacional de analizar direcciones para cada comuna, pues, a mayor cantidad de comunas analizadas mayor es el tiempo de procesamiento de los datos debido a las diferencias en la nomenclatura de direcciones en las distintas comunas del país. No obstante, la metodología propuesta es replicable a mayor nivel, permitiendo estandarizar potencialmente todas las direcciones de las 346 comunas de Chile.

Por otro lado, la base oficial de direcciones no considera números que comiencen con “0”, por ejemplo: “CALLE PIEDRA ROJA 063 CASA 56”, posee el número “063”, sin embargo, la coincidencia para esta dirección será con el número “63”. Esto se debe a que el número “063” no existe en la base oficial. Esta limitante, no afecta la capacidad de coincidir direcciones similares, pero sí de asignar códigos postales, debido a que este, depende del número de calle.

Con el objetivo de mejorar la coincidencia de direcciones, se utilizan diccionarios de abreviaturas que apoyan los procesos de clasificación y coincidencia de direcciones. Las abreviaturas consideradas son acotadas y se explicitan en las secciones 5 y 6. Es por esto, que el rendimiento del normalizador de direcciones puede variar dependiendo de la cantidad de abreviaturas que se consideren

2.4.3 Aplicabilidad del modelo

Este trabajo no supone la implementación del sistema de búsqueda inteligente de direcciones postales, debido a que en esta memoria no se trabaja directamente con una empresa del rubro postal. Por esta misma razón, las recomendaciones que surjan de este trabajo serán limitadas al rendimiento de los modelos aplicados y su efectividad para estandarizar direcciones chilenas.

3 MARCO CONCEPTUAL

3.1 Estandarización de direcciones postales

El proceso de *Address Standarization* (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 (Rivas, 2016). En este trabajo, se utilizará el concepto de dirección normalizada cuando se le haya asignado un código postal.

3.1.1 Clasificación de Texto

El proceso de *Text Classification* o clasificación de texto consiste en asignar automáticamente categorías o etiquetas predefinidas a texto libre, lo cual es muy importante para la recuperación de información y muchas otras aplicaciones (Wang & Xiao-Jing, 2005). Este proceso es parte del reconocimiento de entidades nombradas o NER por sus siglas en inglés (Named-Entity Recognition). Identificar cada parte de la dirección es necesario para aplicar nuevos algoritmos que permitan trabajar con la información de la cadena de texto. El rendimiento de este paso tiene un impacto directo en la capacidad de los algoritmos de coincidencia de texto para reconocer una dirección igual a otra, a pesar de las variaciones que puedan existir.

Existen varias técnicas que se pueden utilizar para la clasificación de texto, como:

Clasificación basada en reglas lógicas: los modelos de clasificación de texto basados en reglas aplican un conjunto de reglas escritas a mano y usan información contextual para asignar etiquetas a subconjuntos de palabras. Estos criterios a menudo se conocen como reglas de marco de contexto. Una de estas reglas podría ser: "Si una palabra es seguida por un número y hay un único número en la oración entonces, es un nombre de calle ". Este tipo de técnicas poseen un buen rendimiento cuando la estructura a clasificar es simple y con pocos atributos a categorizar. Debido a que, no se utiliza un conjunto de datos de entrenamiento, el modelo es considerado de aprendizaje no supervisado.

Modelos de Deep Learning: La clasificación de texto se ha beneficiado del reciente resurgimiento de las arquitecturas de Deep Learning, debido a su potencial para alcanzar una alta precisión con menos necesidad de características de ingeniería. Las dos arquitecturas principales de aprendizaje profundo utilizadas en la clasificación de texto son las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN). Por un lado, los algoritmos de Deep Learning requieren muchos más datos de entrenamiento que los algoritmos tradicionales de aprendizaje automático. Por otra parte, los algoritmos tradicionales de aprendizaje automático como Support Vector Machine (SVM) y Naive Bayes (NB) alcanzan un cierto umbral en el que agregar más datos de entrenamiento no mejora su precisión. Por el contrario, los clasificadores de Deep Learning continúan mejorando a medida que aumentan los datos disponibles.

Etiquetado estocástico (probabilístico): un enfoque estocástico incluye frecuencia, probabilidad o estadísticas. El enfoque estocástico más simple descubre la categoría más utilizada para una palabra específica en los datos de entrenamiento anotados y utiliza esta información para etiquetar esa palabra en el texto no anotado. Uno de estos métodos es calcular las probabilidades de varias secuencias de etiquetas que son posibles para una dirección y asignar las etiquetas de la secuencia

con la mayor probabilidad. Los modelos ocultos de Markov (HMM) y los Campos Aleatorios Condicionales (CRF) son enfoques probabilísticos para realizar la clasificación de texto.

3.1.2 Importancia de la clasificación de texto

Según IBM (Schneider, 2016), se estima que alrededor del 80% de toda la información no está estructurada, siendo el texto uno de los tipos más comunes de datos no estructurados. Debido a la naturaleza desordenada del texto, analizar, comprender, organizar y clasificar los datos de texto es difícil y requiere mucho tiempo, por lo que la mayoría de las empresas no pueden extraer valor de eso.

Aquí es donde interviene la clasificación de texto con aprendizaje automático. Mediante el uso de clasificadores de texto, las empresas pueden estructurar información comercial como correo electrónico, documentos legales, páginas web, conversaciones de chat y mensajes de redes sociales de una manera rápida y rentable. Esto permite a las empresas ahorrar tiempo al analizar datos de texto, ayudar a informar las decisiones comerciales y automatizar los procesos comerciales.

Algunas de las razones por las cuales las empresas están aprovechando la clasificación de texto con el aprendizaje automático son las siguientes:

Escalabilidad: Analizar y organizar el texto manualmente lleva tiempo. Es un proceso lento donde un humano necesita leer cada texto y decidir cómo estructurarlo. El aprendizaje automático cambia esto y permite analizar fácilmente millones de textos a una fracción del costo.

Análisis en tiempo real: Hay situaciones críticas que las empresas deben identificar lo antes posible y tomar medidas inmediatas (por ejemplo, crisis de relaciones públicas en las redes sociales). Los clasificadores de texto con aprendizaje automático pueden hacer precisiones exactas en tiempo real que permiten a las empresas identificar información crítica al instante y tomar medidas de inmediato.

Criterios consistentes: Los anotadores humanos cometen errores al clasificar datos de texto debido a distracciones, fatiga y aburrimiento. Se generan otros errores debido a criterios inconsistentes. Por el contrario, el aprendizaje automático aplica la misma lente y criterios a todos los datos, lo que permite a los humanos reducir los errores con los modelos de clasificación de texto centralizados.

3.1.3 Coincidencia de dirección

El proceso conocido en inglés como *Address Matching* o en español como *coincidencia de una dirección*, es un proceso que compara una dirección o una tabla de direcciones con los atributos de un conjunto de direcciones de referencia. El objetivo es determinar si una dirección en particular está dentro de un rango de direcciones asociado al conjunto de referencia. Si una dirección está dentro del rango de características de la dirección de referencia, es considerada como una coincidencia (match) y la localización es recuperada (Rivas, 2016)

El Address Matching es una aplicación de la coincidencia de cadenas de texto o *string matching.* El problema de la coincidencia de cadenas es que hay dos cadenas, una es el texto T [1, ..., n], es decir, la cadena principal dada y la otra es el patrón P [1, ..., m], es decir, la cadena dada P debe

coincidir con cadena principal T, donde los valores de m y n pueden ser distintos o iguales. La coincidencia de cadenas se usa de forma variable en aplicaciones de palabras reales como el esquema de base de datos, los sistemas de red y la coincidencia de direcciones.

Existen diversos algoritmos que permiten realizar string matching (Singla & Garg, 2012):

Algoritmo de coincidencia aproximada de cadenas (*fuzzy Matching*): es la técnica de encontrar cadenas de texto que coincidan aproximadamente con un patrón. El problema de la coincidencia aproximada de cadenas generalmente se divide en dos problemas secundarios: encontrar coincidencias aproximadas de subcadenas dentro de una cadena dada y encontrar cadenas de diccionario que coincidan aproximadamente con el patrón. Además, considera el largo de las cadenas para comparar su similitud. En Python existe la librería FuzzyWuzzy que realiza un cálculo de distancia con lógica difusa.

Algoritmo basado en la distancia de Levenshtein: la distancia de Levenshtein es una medida para medir la cantidad de diferencia entre dos secuencias (es decir, una distancia de edición). El término distancia de edición se usa a menudo para referirse específicamente a la distancia de Levenshtein. La distancia de Levenshtein entre dos cadenas se define como el número mínimo de ediciones necesarias para transformar una cadena en la otra, siendo las operaciones de edición permitidas la inserción, eliminación o sustitución de un solo carácter a la vez.

3.1.4 Código Postal

El sistema de código postal en Chile fue implementado por la empresa de correos estatal, Correos Chile. El código está conformado por siete dígitos, en donde los tres primeros números indican la comuna y los últimos cuatro indican la manzana. Por esta razón, personas que viven en una misma manzana comparten el mismo código postal.

Correos de Chile tiene registradas 114 comunas por frente de manzana (es decir, un 90% de las direcciones urbanas) con un código postal. El resto de las direcciones, por ejemplo, en zonas rurales sólo tiene un código postal comunal (Elfenbaum, CCM, 2020).

3.2 Distancia de Levenshtein

Uno de los algoritmos de coincidencia de textos 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. La distancia de edición hace referencia al número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. Se entiende por operación, una inserción, eliminación o sustitución de un caracter. Por ejemplo, asumiendo un costo unitario: la distancia de Levenshtein entre "casa" y "calle" es de 3 porque se necesitan al menos tres ediciones elementales para cambiar uno en el otro. El resultado de la distancia de Levenshtein corresponde a un número entero, sin embargo, es posible calcular el ratio de similitud de Levenshtein, que considera el largo de las cadenas de texto, y entrega un resultado entre 0 y 1.

3.2.1 Definición de la distancia de Levenshtein (o distancia mínima de edición)

Sean dos vectores $X = \{x_1, \dots, x_n\}$ y $Y = \{y_1, \dots, y_m\}$, se busca calcular la distancia de edición para transformar X en Y. Es necesario notar que D(X,Y) no es lo mismo que D(Y,X) si es que los costos de inserción y eliminación no son los mismos (Aggarwal, 2015). Sea X_i el segmento de los primeros i elementos de X, y sea Y_j el segmento de los primeros j elementos de Y. Sea $D_{X,Y}(i,j)$ el costo óptimo de edición entre los segmento mencionados anteriormente, el cual se define como:

$$D_{X,Y}(i,j) = min \begin{cases} D(i-1,j) + costo\ eliminación \\ D(i,j-1) + costo\ inserción \\ D(i-1,j-1) + I_{ij} * costo\ sustitución \end{cases}$$

$$I_{ij} \begin{cases} 0 & cuando\ elementos\ i-ésimo = j-ésimo \\ 1 & en\ caso\ contrario \end{cases}$$

Además, D(i,0) = costo de eliminar i elementos ∀ i. D(0,j) = costo de insertar j elementos ∀ j.

Con esta información es posible escribir un programa computacional recursivo que entregue el valor de la distancia de Levenshtein. La distancia de Levenshtein corresponde a un número entero el cual representa la cantidad de operaciones fundamentales que fueron necesarias para transformar la cadena X en Y o viceversa.

Más información acerca del cálculo de la distancia de Levenshtein se encuentra en el Anexo A

3.2.2 Distancia con lógica difusa – FuzzyWuzzy

La distancia de Levenshtein responde a la pregunta ¿qué tan similares son las cadenas X e Y? ya que, entrega como resultado la cantidad de operaciones realizadas para transformar X en Y. Por otro lado, la distancia con lógica difusa responde a la pregunta ¿son las cadenas X e Y lo mismo? Entregando como resultado un valor entre 0 y 1 ó un puntaje entre 0 y 100 (Carrera Arias, 2019).

Esta memoria se desarrollará utilizando la librería FuzzyWuzzy de Python, la cual funciona en base a lógica difusa. Posee varias métricas que varían dependiendo del enfoque de lógica difusa que se utilice. La principal métrica de esta librería corresponde al ratio similitud de Levenshtein, el cual considera el largo de las cadenas de texto a comparar y se formula de la siguiente manera:

$$Lev\_Ratio = \frac{(n+m) - D_{X,Y}(i,j)}{(n+m)}$$

Con *n, m > 0* siendo el largo del vector X e Y respectivamente.

La libraría FuzzyWuzzy posee variaciones al ratio de similitud de Levenshtein. Una de ellas corresponde al "*fuzz partial ratio"* el cual es capaz de detectar coincidencias en subcadenas. En otras palabras, si la cadena corta tiene una longitud k y la cadena más larga tiene longitud p, entonces el algoritmo busca la puntuación de la mejor coincidencia para la cadena de longitud k.

Es importante considerar que el lenguaje Python distingue entre letras mayúsculas y minúsculas, por lo que, es necesario realizar un preprocesamiento de las direcciones antes de aplicar los algoritmos de coincidencia de texto, por ejemplo, cambiar todas las letras a mayúsculas.

3.3 Clasificación de texto

Los modelos de clasificación de texto en base a reglas lógicas son útiles cuando las estructuras a etiquetar son relativamente simples. Por el contrario, los modelos probabilísticos son útiles cuando las estructuras a etiquetar son complejas y pueden variar entre cada observación. En (Tran, 2019) y (Wang, y otros, 2016) utilizan CRF para clasificar textos de direcciones utilizando entre 5 a 17 etiquetas, lo que muestra la complejidad en la estructura de las direcciones a clasificar.

En esta memoria se trabaja con direcciones a nivel de comuna, lo que significa que la comuna estará dada y la clasificación de direcciones solo se realiza utilizando tres etiquetas: Nombre de calle, número principal e información adicional.

3.3.1 Clasificador en base a reglas lógicas

En Chile, las direcciones postales poseen una estructura intrínseca común, primero se escribe el nombre de la calle luego el número principal y la información adicional que incluye referencias a departamentos, oficinas, torres, condominio, villa, entre otros. Esta estructura, de solo tres etiquetas, se mantiene estática, es decir sus campos no se permutan.

Estas características de las direcciones posibilitan construir un clasificador en base a reglas lógicas, en donde se toman como supuesto principal, que todas las direcciones poseen un número principal que identifica su posición en la calle. El supuesto se sustenta en la condición necesaria de ubicar la dirección en una cuadra específica para poder asignar un código postal, consecuentemente, resulta necesario la existencia de un número principal.

En virtud de conocer el rendimiento del clasificador, se incluirá un atributo que identifique si la dirección posee un número de calle definido. De la misma manera, se considera válida y correctamente clasificadas aquellas direcciones a las cuales se logra identificar un número principal.

3.3.2 Clasificador Probabilísticos

La complejidad en la estructura de algunos textos afecta negativamente el rendimiento de los clasificadores en base a reglas lógicas, debido a que pueden existir variaciones en el formato de escritura que invaliden algunas reglas. Por esta razón, los clasificadores probabilísticos son una alternativa viable para lidiar con textos complejos, que poseen una variedad significativa de etiquetas y que alcancen buenos resultados.

Los campos aleatorios condicionales (CRF por sus siglas en inglés) corresponden a un tipo de clasificador discriminativo, lo que significa que modelan el límite de decisión entre las diferentes clases (Chawla, 2017). Su principio subyacente es que aplican regresiones logísticas a las secuencias de entradas. Los CRF modelan la dependencia entre cada estado (o etiqueta) $y^i = \{y_1^i, \dots, y_T^i\}$ y el vector de entrada $x^i = \{x_1^i, \dots, x_T^i\}$ utilizando la probabilidad condicional $P(y|x)$. Debido a que se modelan datos secuenciales, se construye una función característica

$f_k(y_t, y_{t-1}, x_t)$ que posee información del contexto del dato en observación. Cada función característica se basa en la etiqueta anterior y la actual, comportándose de forma binaria (Sutton & McCallum, 2011). Para construir el campo condicional se debe asignar a cada función característica un vector de parámetros $\theta=\{\lambda_k\} \in R^K$, que van a ser entrenados con el siguiente algoritmo:

$$P(y|X;\lambda) = \frac{1}{Z(X)} exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\}$$

Para un instante t específico, la función normalización se define como:

$$Z(X) = \sum_{y} exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\}$$

Más información acerca de los CRF se encontrará en el anexo B

En (Tran, 2019) y (Wang, y otros, 2016) se utilizan los CRF para etiquetar segmentos de direcciones en diferentes campos (utilizando 5 o más etiquetas) obteniendo mejores resultados comparados con otros modelos de clasificación probabilísticos.

Dentro de los requisitos necesarios para la aplicación de un modelo CRF se encuentran los datos de entrenamiento, los cuales deben ser direcciones etiquetadas previamente. En este trabajo de memoria no se tienen datos etiquetados, por lo que el entrenamiento de un modelo CRF requiere de un paso extra, la creación de una base de entrenamiento. Por esta razón, se deja propuesto como trabajo futuro, la realización de un clasificador de texto para direcciones en base a un modelo CRF utilizando datos etiquetados y considerando estructuras de texto más complejas que las propuestas (Nombre de calle, número principal e información adicional).

4 METODOLOGÍA

Para la resolución del problema se utilizará la metodología CRISP-DM, la cual se divide en 6 etapas: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelamiento, evaluación e implementación. El trabajo de esta memoria abarcará, principalmente, las primeras 4 etapas de la metodología.

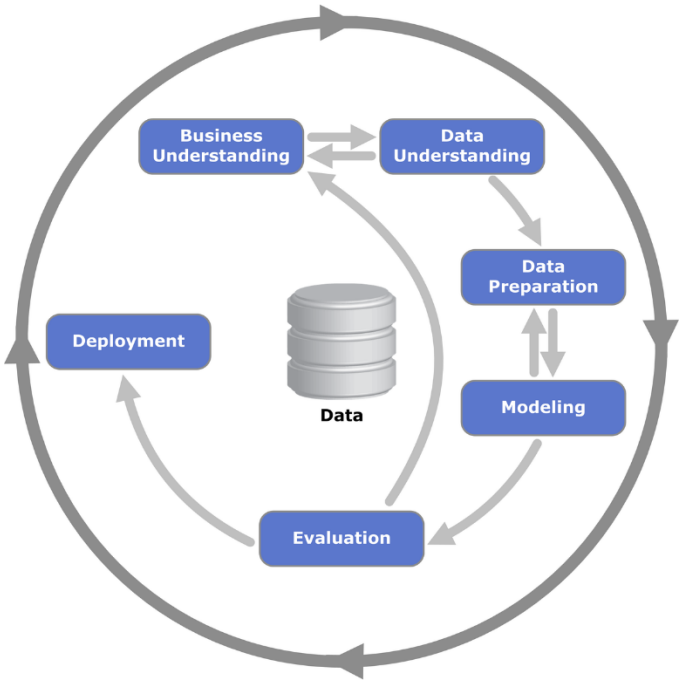

Figura 6: Metodología CRISP-DM

4.1 Entendimiento del Negocio

En esta sección se aborda el contexto del negocio con el objetivo de entregar un sentido a los datos que se analizan. Se revisa la acción de estandarizar direcciones y el origen de los datos con los que se trabaja.

4.1.1 Estandarización de direcciones

El proceso de *Address Standarization* (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 (Rivas, 2016). Este proceso es necesario en el rubro postal, adquiriendo real importancia en la etapa de clasificación, ya que aquí, se asigna un código postal a cada dirección. Esta tarea es llamada normalización de direcciones.

4.1.2 Mercado y Clientes

El mercado postal puede dividirse en 3 segmentos de clientes: Personas Naturales, Envíos Internacionales y Empresas (del medio local). Cada uno de estos casos representa una forma distinta de adquirir datos o direcciones. En el caso de personas naturales, se pide completar separadamente el nombre de calle, número, información adicional (departamento, block, villa, etc.) y, si es posible, el código postal. Con estos requisitos, este caso no representa el problema a resolver

en esta memoria. En los envíos internacionales es requisito agregar un código postal, por lo que este caso no aplica al objetivo que busca esta memoria. Por último, el segmento empresa, que corresponde al segmento más grande con más de 190 millones de envíos anuales solo en Chile (Empresa de Correos de Chile, 2019), no posee un nivel de exigencia que permita asignar automáticamente un código postal y con esto normalizar la dirección, ya que las direcciones vienen en una sola cadena de texto y no estandarizadas, por lo que este segmento es parte del problema a solucionar en esta memoria.

Por otro lado, según la base oficial de direcciones: a lo largo de Chile se registran 4.026.063 direcciones considerando la combinatoria del nombre de calle principal, número principal y comuna. Si solo se consideran nombres de calle y número principal diferentes, la cantidad de direcciones se reduce a 66.000, esto implica que muchas direcciones se repiten a lo largo del país. Por esta razón, es factible trabajar el problema a nivel de comuna.

4.2 Entendimiento de los datos

Para abordar el problema de normalización de direcciones se dispone de 2 bases de datos con direcciones: una base de datos se encuentra estandarizada con código postal a nivel de calle y número principal, la cual es llamada *base oficial de direcciones*. La otra base posee direcciones de clientes que no están normalizadas, por lo que no poseen código postal, a esta última se le llama *base de clientes.*

La base oficial contiene direcciones de Chile que poseen un código postal, lo que se traduce en 4.026.063 registros. Aquí encontramos información sobre: *dirección, tipo de edificación, id_direccion, id_codigopostal, id_comuna, nombre de comuna, nombre de provincia* y *nombre de región*. En el campo *dirección* se encuentra el nombre y número de la calle. Es importante tener en cuenta que, cada año se crean nuevas direcciones con la construcción de nuevos complejos inmobiliarios, lo que puede provocar una desactualización en la base oficial.

La base de clientes posee 10.480 direcciones totales, las cuales no están estandarizadas y poseen información del nombre de calle, número principal e información adicional (número de departamento, número de block y/o villa) en el campo *Dirección.* La información relacionada a la comuna se encuentra en el campo *Comuna,* distinto al de *Dirección,* lo que posibilita trabajar la normalización a nivel de comunas.

En primera instancia, el entrenamiento del modelo se reduce solo a la comuna de Quilicura, debido a que esta no posee direcciones rurales y sus direcciones están estandarizadas completamente, además, se disminuye la carga computacional para entrenar los modelos. Esta reducción aminora la cantidad de datos a trabajar en la base oficial a 49.950 y en la base de clientes a 1254 direcciones. Al realizar un emparejamiento directo entre la base de clientes y la base oficial de direcciones se encuentra un 20% de coincidencia. En una segunda instancia, el mejor modelo encontrado se probará con direcciones de la comuna de Santiago. La base oficial para la comuna de Santiago posee 57.397 direcciones y la base de cliente 807. Al realizar un emparejamiento directo entre ambas bases se encuentra un 18% de coincidencia.

En Chile se acostumbra a escribir las direcciones partiendo por el nombre de la calle, luego el número principal, información adicional, la comuna y región. Sin embargo, no existe una regla

definida sobre como escribir una dirección y pueden existir variaciones en la escritura al igual que diferentes estructuras (o campos de información en el texto de la dirección). Debido a que, se quiere asignar un código postal a la dirección se trabaja bajo el supuesto de que existe un número principal bien escrito que permite la normalización. Este supuesto es clave para distinguir entre una dirección correcta y candidata para la normalización o si requerirá una revisión más profunda antes de asignarle un código postal.

Una misma dirección se puede escribir de varias formas, omitiendo o 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. A continuación, se muestra la Tabla 1 con abreviaturas y prefijos de direcciones encontrados en la base de clientes:

Abreviaturas y prefijos de direcciones	
Pasaje	"PASAJE", "PJE.", "PJE", "PSJE.", "PSJE"
Avenida	"AVENIDA", "AVDA", "AVDA.", "AV.", "AV"
Calle	"CALLE"
Puente	"PTE", "PTE."
Puerto	"PTO.", "PTO"
Departamento	"DPTO", "DPTO.", "DEP.", "DEP", "DEPTO", "DPT", "DP"
Edificio o Block o Torre	"BLOCK", "BLOCK.", "EDIFICIO", "EDIF", "TORRE"
Casa	"CASA"
Condominio	"CONDOMINIO", "COND"
Villa	"VILLA", "V."

Tabla 1: Abreviaturas y prefijos de direcciones

4.3 Preparación de los Datos

El objetivo de esta memoria consiste, en pocas palabras, en asignar correctamente un código postal a una dirección entregada por un cliente. Para esta labor es necesario identificar satisfactoriamente las partes de cada dirección, específicamente el nombre de calle y el número principal para poder asignar el código postal. Además, al existir variaciones en las formas de escrituras se puede preprocesar los nombres de calle con el fin de modificar abreviaturas que puedan entregar más información o, al contrario, dificulten la normalización.

La segmentación y etiquetado para identificar cada parte de la dirección se realiza a través de un modelo clasificador en base a reglas lógicas. Se utiliza la información contextual de la dirección para identificar el número principal y, luego, el nombre de calle y la información adicional. Esto se ejecuta utilizando el lenguaje de programación Python, a través de sentencias condicionales (if, elif y else).

Con el nombre de calle identificado y etiquetado, se procede a buscar la dirección coincidente en la base oficial de direcciones. Sin embargo, este proceso puede ser mejorado si el nombre de calle es preprocesado modificando las abreviaturas que pueda contener. Debido a que existen abreviaturas generales y comunes en Chile, se construyen diccionarios que asisten este proceso. Estos diccionarios son utilizados en modo de filtros que pueden reemplazar, modificar o eliminar abreviaturas.

4.4 Modelamiento

Para la asignación de un código postal se necesita identificar la comuna, el nombre de la calle y el número principal de la dirección. Debido a que, los datos a trabajar poseen la comuna en un atributo separado de la dirección el problema se acota a identificar solamente el nombre de calle y número principal, para luego buscar una dirección coincidente en la base oficial que permita la normalización. La solución propuesta consiste en dos modelos, diferentes, complementarios entre sí, donde uno es utilizado para identificar la estructura de la dirección (clasificador de texto) y otro para la asignación del código postal.

4.4.1 Clasificador de Texto

El primer modelo en actuar corresponde al clasificador de direcciones en base a reglas lógicas. Se puede considerar que este modelo prepara las direcciones para luego ser normalizadas por el segundo modelo. Este modelo es de aprendizaje no supervisado, ya que no se entrena utilizando direcciones etiquetadas. Esto se debe a que, no se posee una base de entrenamiento de direcciones segmentadas y etiquetadas.

El clasificador considera que las direcciones están escritas por palabras y no por caracteres simplemente. Esta diferencia es crucial, debido a que, modela las direcciones siguiendo la misma lógica que utiliza una persona al escribir una dirección. Además, la información contextual se procesa en base a la lógica de la cultura popular de escritura de una dirección en Chile. Por lo tanto, al separar palabra por palabra, se identifica aquellas que corresponden a números con los cuales se identifica el nombre de calle principal.

4.4.2 Coincidencia de direcciones

El segundo modelo corresponde al asignador de códigos postales, el cual funciona buscando una dirección coincidente en la base oficial de direcciones. Se utiliza como base la distancia de Levenshtein, ya que, considera la cantidad de operaciones fundamentales (inserción, deleción y sustitución) que se necesitan para transformar una palabra en otra. Sin embargo, las direcciones poseen variaciones en sus escrituras que implican diferentes cantidades de palabras para referirse a una misma dirección, por ejemplo: *COMPAÑÍA 457*, en la comuna de Santiago, hace referencia a la calle *COMPAÑÍA DE JESÚS 457*, ya que, popularmente, se le menciona sólo “COMPAÑÍA” sin agregar las últimas dos palabras “DE JESÚS”. Por esta razón, es importante considerar una métrica que incluya la cantidad de palabras o el largo de la cadena de texto.

También, es importante considerar la forma en que se buscan direcciones coincidentes, ya sea, calculando la similitud palabra por palabra o utilizando la cadena completa de texto. Comparar palabra por palabra permite que, al existir direcciones que posean distinto número de palabras en la base oficial y en la base de clientes, estas no sean penalizadas de la misma manera que si se compara la cadena de texto completa. Por otro lado, al utilizar la dirección completa para buscar la coincidencia se puede ser más asertivo en direcciones que poseen palabras similares.

5 MODELO CLASIFICADOR REGLAS LÓGICAS

Los clasificadores de texto basados en reglas lógicas aplican un conjunto de sentencias condicionales escritas a mano. Estas reglas instruyen al sistema a utilizar elementos semánticamente significativos de un texto para identificar categorías relevantes en función de su contenido (MonkeyLearn.com, 2019).

El modelo propuesto en esta memoria es desarrollado con el lenguaje de programación Python. Utiliza un conjunto de sentencias condicionales *if, else* y *elif* para verificar las reglas lógicas y se recorre las cadenas de texto utilizando ciclos *for*.

Se tiene como objetivo que el modelo separe las direcciones en la siguiente estructura: *Nombre de calle, número de calle, información adicional.* Como supuesto base se asume que las direcciones poseen un número principal, debido a que se quiere asignar un código postal. La lógica del clasificador se basa en encontrar el número principal dentro de las diferentes direcciones. En caso de no existir un número principal bien definido, la dirección se etiqueta como inválida y no se normaliza.

A continuación, en la tabla 2, se muestra el conjunto de direcciones de referencia para la construcción de las reglas lógicas del clasificador:

Dirección	Nombre Calle	Número de Calle	Info. Adicional
PJE. SENDA SANTA MARTA 0205 DEPTO. 21	PJE. SENDA SANTA MARTA	205	DEPTO. 21
CALLE 1 SUR 786	CALLE 1 SUR	786	
PJE 1 5183 POB. ALBORADA	PJE 1	5183	POB. ALBORADA
CALLE UNO SUR 786	CALLE UNO SUR	786	
ISMAEL BRICENO 551 23 14 PUCARA DE LAZANA	ISMAEL BRICENO	551	23 14 PUCARA DE LAZANA
ISMAEL BRICENO 1481 TORRE C DEPTO 32	ISMAEL BRICENO	1481	TORRE C DEPTO 32
PASAJE 4 40 VILLA HUELEN	PASAJE 4	40	VILLA HUELEN
1 ORIENTE 1985 DP 705	1 ORIENTE	1985	DP 705
1 ORIENTE 1985	1 ORIENTE	1985	
PASAJE VILLA MAYOR NORTE 357	PASAJE VILLA MAYOR NORTE	357	
PASAJE LA CASA PIEDRA 1360 DPTO 32	PASAJE LA CASA PIEDRA	1360	DPTO 32
AV LO CRUZAT 555 DEP 532	AV LO CRUZAT	555	DEP 532
PJE OLLAGUE 251 A A 11	PJE OLLAGUE	251	A A 11

Tabla 2. Direcciones de referencia para construcción de reglas de clasificación

En base a las direcciones de la Tabla 2, se procede de la siguiente manera: Primero el clasificador separa todos los caracteres de la cadena de texto por espacios en blanco, luego se separan los números de las palabras. Se guardan las palabras, los números y dos vectores con los índices de sus posiciones respectivamente.

Como regla general se asigna al campo *nombre de calle*, todas las palabras existentes hasta el primer número que aparezca, si es que existe un solo número.

Si hay más de un número, entonces, puede ser parte de un número secundario que indica el departamento, oficina o block, o puede que el primer número sea parte del nombre de la calle. En el caso de que existan dos números seguidos, se considera el primero como parte del nombre de calle y el segundo como principal (casos "pasaje 4 722" ó "calle 2 5859"). Si no hay dos números seguidos se considera el primero como número principal. Los números restantes se asignan a *información adicional*.

Para realizar una correcta clasificación, se trabaja con dos diccionarios: un diccionario con palabras que anteceden el nombre de calle, llamado *pre_calle* y otro con palabras que hacen referencia al tipo de edificación o lugar, llamado *tipo*. A continuación, se detalla el contenido de cada diccionario.

pre_calle: "PASAJE", "PJE.", "PJE", "PSJE.", "PSJE","AVENIDA", "AVDA", "AVDA.", "AV.", "AV","CALLE".

tipo: "BLOCK", "BLOCK.", "DPTO", "DPTO.", "DEP.", "DEP", "DPT", "DP", "DEPTO", "DEPTO.", "CASA", "VILLA", "TORRE", "COND","EDIF".

Estos diccionarios se utilizan para crear las sentencias condicionales, ya que entregan información semántica de las direcciones.

Las palabras y los números que no fueron asignados a *nombre de calle* o *número de calle* son asignados a la etiqueta *información adicional*.

Como último paso, el clasificador identifica si una dirección es válida o no. Para esto, en primer lugar, verifica si existe al menos un número asignado a *Número de calle* y, en caso de que no exista, se etiqueta como dirección inválida. En segundo lugar, se verifica si la dirección posee el indicativo "*S/N"*, el cual indica que no existe numeración principal para esa ubicación. En tercer lugar, se verifica que en el nombre de calle no posea indicativos del tipo de vivienda, ya que esto implica que el número identificado como principal puede deberse a una numeración interna del tipo de edificación, es decir si el nombre de calle posee palabras como "BLOCK" o "DPTO" se etiqueta como dirección inválida.

6 MODELAMIENTO – ASIGNACIÓN CÓDIGO POSTAL

La solución propuesta en esta memoria comprende dos etapas: la primera parte consiste en etiquetar las direcciones con *Nombre de Calle, Número Principal* e *información adicional* para cada subsegmento de texto de la dirección. La segunda etapa consiste en buscar direcciones coincidentes entre la base de clientes y la base oficial, a través de algoritmos de coincidencia de texto, y finalmente asignar el código postal correspondiente.

6.1 Coincidencia de textos

La distancia de Levenshtein es la métrica por excelencia para determinar qué tan similares son dos cadenas de texto, ya que cuenta la cantidad de operaciones fundamentales necesarias para llegar de un texto a otro. Sin embargo, esta métrica no considera el largo de los textos a comparar y aplicarla directamente puede arrojar resultados inesperados. Por otro lado, se tiene el ratio de similitud de Levenshtein el cual considera el largo de los textos, pero existe más de una forma para comparar cadenas de texto.

Para asegurar un mejor modelo de coincidencia de textos, se ha propuesto la siguiente metodología:

Filtros[2]

En primer lugar, se realiza un preprocesamiento de las cadenas de texto con el fin de modificar algunas abreviaturas y números que pueden afectar el cálculo de la distancia de Levenshtein y el ratio de similitud. A continuación, se muestran los diferentes valores que adquiere la distancia de Levenshtein para variaciones de una misma dirección:

Dirección Oficial	Dirección a Normalizar	Distancia de Levenshtein
PASAJE LOS QUELTEHUES	PASAJE LOS QUELTEHUES	0
PASAJE LOS QUELTEHUES	PSJE LOS QUELTEHUES	2
PASAJE LOS QUELTEHUES	PSJE. LOS QUELTEHUES	3
PASAJE LOS QUELTEHUES	PJE LOS QUELTEHUES	3
PASAJE LOS QUELTEHUES	PJE. LOS QUELTEHUES	4
PASAJE LOS QUELTEHUES	LOS QUELTEHUES	7

Tabla 3. Ejemplo cálculo de distancia de Levenshtein

Al ejemplo de la Tabla 3 se le debe considerar, además, que pueden existir errores tipográficos que pueden aumentar la distancia de Levenshtein. Es por esto, que se prueban tres formas diferentes de utilizar un filtro que modifica las abreviaturas del diccionario de palabras *pre_calle:* eliminando las abreviaturas, agregando la palabra completa y no utilizando el filtro. Esto se realiza sobre las direcciones de la base oficial y de la base de clientes.

[2] Al utilizar la palabra filtro se hace referencia a la modificación, reemplazo o eliminación de abreviaturas

Un caso especial de abreviaturas que no corresponde al diccionario *pre_calle* son las palabras *PUENTE* y *PUERTO*. Estas abreviaturas serán reemplazadas por sus palabras completas al utilizar el filtro de *pre_calle*. Esta práctica puede extenderse a otras abreviaturas, lo cual requiere la creación de un diccionario de abreviaturas comunes para la realización de reemplazos, esto se deja propuesto como trabajo futuro, ya que excede el alcance de esta memoria.

En adición, se ha observado que hay direcciones escritas por clientes del estilo *Calle 4*, siendo su nombre oficial *Calle Cuatro*. Por esta razón se utilizará otro filtro que reemplaza los números escritos con su símbolo numérico por su versión escrita. Este filtro es aplicado en todos los enfoques de coincidencia, ya que la distancia de Levenshtein generada al reemplazar el símbolo del número por su palabra es, a lo menos, el largo de la palabra, agregando ruido a los resultados.

Modelamiento del algoritmo de coincidencia

En segundo lugar, el modelo genera un ranking con las 3 direcciones más probables de ser la coincidencia correcta, a través de la distancia de Levenshtein, en una instancia y, en otra, utilizando la distancia con lógica difusa. Para el caso en donde el ranking se crea utilizando la distancia de Levenshtein se selecciona como mejores candidatos a las direcciones que posean la menor cantidad de operaciones fundamentales. Por otro lado, para el caso donde el ranking se genera utilizando el ratio de similitud de Levenshtein o lógica difusa, se seleccionan los candidatos con mayor puntaje. A este puntaje obtenido se le refiere como *puntaje de ranking.*

Existen dos formas posibles para aplicar la distancia a cada dirección, comparando palabra por palabra o comparando la cadena de texto completa. Se utilizan ambas formas para el cálculo del *puntaje de ranking.*

Con el ranking de 3 direcciones más similares se calcula nuevamente la distancia con lógica difusa obteniendo el *puntaje de selección.* Este puntaje se calcula en base a una variación del ratio de similitud de Levenshtein, el *fuzz partial ratio* (que busca coincidencia de la cadena más corta con subcadenas de la cadena más larga) y, en esta oportunidad, se compara la dirección filtrada del cliente con la dirección propuesta de la base oficial sin filtrar.

Se puede pensar, entonces, que el *puntaje de selección* entrega el mismo resultado que en el primer cálculo cuando se utiliza el ratio de similitud de Levenshtein, sin embargo, esta métrica varía debido a que busca la coincidencia de la cadena más corta con subcadenas de la cadena más larga. En las Tablas 4 y 5 se muestran algunos ejemplos en donde, al calcular el *puntaje de ranking* utilizando el ratio de similitud de Levenshtein y el *puntaje de selección* utilizando la variación del ratio de similitud se obtienen diferentes puntajes.

En la Tabla 4, se muestra la diferencia entre el *puntaje de ranking* y *puntaje de selección* cuando se varía el filtro a utilizar. Además, en este caso, la dirección posee distintas abreviaturas que anteceden al nombre de calle. En la Tabla 5, se observa la diferencia entre ambos puntajes cuando se compara la palabra “COMPANIA” con las direcciones “COMPANIA DE JESUS” y “PASAJE NUEVA COMPANIA”. Cabe mencionar que, la asignación en el ejemplo de la Tabla 5 se define por el calce perfecto entre el número de calle y número propuesto. Por último, en ambos ejemplos existen variaciones entre los *puntajes de ranking* y *puntajes de selección.*

Dirección Original	Nombre de Calle	Número de Calle	Calle Propuesta	Número Propuesto	Puntaje Ranking	Puntaje Selección
Filtro: Elimina abreviaturas de nombre de dirección						
AV PIEDRA ROJA 123 CASA 35	PIEDRA ROJA	123	PIEDRA ROJA	123	100	100
CALLE PIEDRA ROJA 063 CASA 56	PIEDRA ROJA	63	PIEDRA ROJA	63	100	100
PIEDRA ROJA 063 CASA 5	PIEDRA ROJA	63	PIEDRA ROJA	63	100	100
Filtro: No modifica abreviaturas de nombre de dirección						
AV PIEDRA ROJA 123 CASA 35	AV PIEDRA ROJA	123	PIEDRA ROJA	123	88	95
CALLE PIEDRA ROJA 063 CASA 56	CALLE PIEDRA ROJA	63	PIEDRA ROJA	63	79	90
PIEDRA ROJA 063 CASA 5	PIEDRA ROJA	63	PIEDRA ROJA	63	100	100
Filtro: Completa abreviaturas de nombre de dirección						
AV PIEDRA ROJA 123 CASA 35	AVENIDA PIEDRA ROJA	123	PIEDRA ROJA	123	73	90
CALLE PIEDRA ROJA 063 CASA 56	CALLE PIEDRA ROJA	63	PIEDRA ROJA	63	79	90
PIEDRA ROJA 063 CASA 5	PIEDRA ROJA	63	PIEDRA ROJA	63	100	100

Tabla 4. Ejemplo 1 diferencias entre Puntaje Ranking y Puntaje Selección

Dirección Original	Nombre de Calle	Número de Calle	Calle Propuesta	Número Propuesto	Puntaje Ranking	Puntaje Selección	Distancia Números	Propuesta
COMPANIA 1737 D-28 T-A	COMPANIA	1737	PASAJE NUEVA COMPANIA	425	73	90	1312	0
COMPANIA 1737 D-28 T-A	COMPANIA	1737	COPIAPO	1495	67	67	242	0
COMPANIA 1737 D-28 T-A	COMPANIA	1737	COMPANIA DE JESUS	1737	64	90	0	1

Tabla 5. Ejemplo 2 diferencias entre Puntaje Ranking y Puntaje Selección

6.2 Modelamiento – Criterios de asignación

Con los *puntajes de ranking* y *puntajes de selección* calculados se procede a la asignación del código postal. Se debe asegurar que se seleccione la dirección que más se asemeje según el puntaje de selección y luego, que el número propuesto corresponda a la misma cuadra del número principal de la dirección a normalizar, para lo cual se calcula la distancia entre el número principal y número propuesto.

La propuesta del número de calle supone que el número escrito en la dirección entregada por el cliente está correcto. Debido a que, se asigna el código postal correspondiente al frente de cuadra, la selección del número propuesto se realiza, en primer lugar, buscando el número idéntico. Si no se encuentra, se busca el número más cercano que posea la misma centena y paridad. Finalmente, si no se encuentra un número con las características anteriormente mencionadas, se selecciona el de menor diferencia numérica.

Para la asignación del código postal, el modelo identifica y etiqueta la dirección propuesta con el mayor puntaje de selección entre las triadas de direcciones propuestas para cada envío. Si solo hay una dirección con el mayor puntaje de selección, se le asigna a esta el código postal. En el caso de que existan dos o más direcciones con el mayor puntaje de selección, se escoge la alternativa que

posea una menor diferencia entre el número principal y el número propuesto. Finalmente, si ningún criterio de asignación se cumple, ya sea porque existen dos o más direcciones con los mismos puntajes de selección (y máximos en cada triada) y no coinciden con la menor diferencia entre números, la dirección se etiqueta como *no propuesta* y se aparta para una revisión manual.

6.3 Modelamiento – Enfoques a utilizar

Utilizando una combinación de diferentes filtros, algoritmos de coincidencias y métricas de distancia se proponen 12 enfoques de normalización de direcciones, los cuales serán puestos a prueba con direcciones de la comuna de Quilicura. El enfoque con mejor rendimiento será probado en direcciones de la comuna de Santiago.

Se tienen tres filtros, los cuales actúan en el preprocesamiento de los nombres de calles de cada dirección para realizar el cálculo del *puntaje de ranking*. Todos los enfoques utilizan el filtro de números que transforma los números arábigos a su versión escrita y el filtro de abreviaturas el cual elimina, completa o no modifica las abreviaturas correspondientes al diccionario *pre_calle.*

Los algoritmos de coincidencia a utilizar son dos: comparar palabra a palabra del nombre de calle y comparar la cadena texto completa del nombre de calle.

Las métricas de distancia que son consideradas son dos: la distancia de Levenshtein y el ratio de similitud de Levenshtein. La variación en la métrica de distancia es utilizada en el cálculo del *puntaje de selección*, ya que para el cálculo del *puntaje de ranking* solo se utilizará el ratio de similitud de Levenshtein.

Enfoque 1.1: Se aplica el filtro de número y las abreviaturas no se modifican. Se compara palabra a palabra. Se calcula el *puntaje de ranking* utilizando distancia de Levenshtein.

Enfoque 1.2: Se aplica el filtro de número y las abreviaturas no se modifican. Se compara palabra a palabra. Se calcula el *puntaje de ranking* utilizando ratio de similitud de Levenshtein

Enfoque 2.1: Se aplica el filtro de número y las abreviaturas se eliminan. Se compara palabra a palabra. Se calcula el *puntaje de ranking* utilizando distancia de Levenshtein.

Enfoque 2.2: Se aplica el filtro de número y las abreviaturas se eliminan. Se compara palabra a palabra. Se calcula el *puntaje de ranking* utilizando ratio de similitud de Levenshtein.

Enfoque 3.1: Se aplica el filtro de número y las abreviaturas se completan. Se compara palabra a palabra. Se calcula el *puntaje de ranking* utilizando distancia de Levenshtein.

Enfoque 3.2: Se aplica el filtro de número y las abreviaturas se completan. Se compara palabra a palabra. Se calcula el *puntaje de ranking* utilizando ratio de similitud de Levenshtein

Enfoque 4.1: Se aplica el filtro de número y las abreviaturas no se modifican. Se compara la cadena completa. Se calcula el *puntaje de ranking* utilizando distancia de Levenshtein.

Enfoque 4.2: Se aplica el filtro de número y las abreviaturas no se modifican. Se compara la cadena completa. Se calcula el *puntaje de ranking* utilizando ratio de similitud de Levenshtein

Enfoque 5.1: Se aplica el filtro de número y las abreviaturas se eliminan. Se compara la cadena completa. Se calcula el *puntaje de ranking* utilizando distancia de Levenshtein.

Enfoque 5.2: Se aplica el filtro de número y las abreviaturas se eliminan. Se compara la cadena completa. Se calcula el *puntaje de ranking* utilizando ratio de similitud de Levenshtein

Enfoque 6.1: Se aplica el filtro de número y las abreviaturas se completan. Se compara la cadena completa. Se calcula el *puntaje de ranking* utilizando distancia de Levenshtein.

Enfoque 6.2: Se aplica el filtro de número y las abreviaturas se completan. Se compara la cadena completa. Se calcula el *puntaje de ranking* utilizando ratio de similitud de Levenshtein

6.4 Modelamiento – Resultados

Debido a que el modelo propuesto corresponde a un modelo de aprendizaje no supervisado no se puede obtener una métrica de desempeño basado en alguna comparación, por lo que, es necesario inspeccionar los resultados y validar si las asignaciones de código fueron correctas. Por esta razón, se inspeccionarán los resultados manualmente para cada enfoque propuesto validando las direcciones con una correcta normalización y las que no.

Para validar la asignación correcta de direcciones se revisa cada envío asegurando que cada propuesta coincida con la dirección a normalizar. Como punto de referencia, se utiliza Google maps buscando reconocer calles distintas que posee nombres similares que pudiesen afectar la asignación del código postal. Luego se contabiliza la cantidad de direcciones validadas manual y correctamente asignadas y se compara con la cantidad de direcciones propuestas totales. Es necesario recordar que, solo se admite una propuesta de código postal por cada envío.

Finalmente se selecciona el enfoque con menor error en la asignación de direcciones y se aplica en envíos para la comuna de Santiago. Con los resultados del mejor enfoque para las comunas de Santiago y Quilicura se proponen categorías de confiabilidad según el puntaje de selección para determinar propuestas *seguras*, *probables* y que *requieren revisión*.

7 RESULTADOS

En este capítulo, se reportan los resultados obtenidos durante el proceso de normalización propuesto en esta memoria. Primero, se comentan resultados en relación con la coincidencia directa, la cual corresponde al estado base del porcentaje de normalización. En segundo lugar, se muestran los resultados para el clasificador de texto en base a reglas lógicas. En tercer lugar, se ilustra el rendimiento del modelo de asignación de códigos postales en cada uno de sus enfoques, para luego analizar la recomendación de categorías de confiabilidad. En cuarto lugar, se declaran los resultados generales de la normalización, dimensionando la cantidad de envíos que a los cuales se asigna un código postal o no, respecto del total de envíos de clientes. Por último, se notifica el costo computacional de cada modelo utilizado.

7.1 Coincidencia directa

Al realizar la coincidencia directa entre la base de clientes y la base oficial se logra normalizar 251 de 1254 direcciones de clientes en Quilicura, es decir, sin la ayuda de ningún algoritmo, se le asigna un código postal al 20% de los envíos de esa comuna. Por otro lado, para los envíos de clientes de la comuna de Santiago se encuentra una coincidencia directa a 145 de 807 direcciones, es decir, sin la ayuda de ningún algoritmo se le asigna código postal al 18% de los envíos.

Los 1003 envíos de clientes de Quilicura que no lograron la coincidencia directa son normalizados con el modelo propuesto en esta memoria. Los 662 envíos de clientes de Santiago que no lograron la coincidencia directa son normalizados con el modelo de mejor rendimiento en base a los resultados de la comuna de Quilicura.

7.2 Clasificador en base a reglas lógicas

El clasificador en base a reglas lógicas de direcciones propuesto en esta memoria posee 3 etiquetas: Nombre de calle, número de calle e información adicional. Además de las etiquetas, el clasificador verifica si una dirección posee un número principal definido y si no lo posee, la identifica como inválida. Esto es posible debido a que, las reglas utilizadas se basan en el supuesto de que existe un número principal, lo cual es clave para poder asignar un código postal.

De los 1003 envíos de clientes de Quilicura a normalizar, se identificaron 932 direcciones válidas o con un número principal bien definido, mientras que 71 direcciones fueron identificadas como inválidas. Por otro lado, de los 662 envíos de clientes de Santiago, se identificaron 618 direcciones válidas y 44 inválidas.

Al analizar las direcciones que fueron etiquetadas como inválidas se encontró que la totalidad no poseía un número principal claro. Las Tablas 6 y 7 muestran ejemplos de direcciones correcta e incorrectamente etiquetadas.

De la Tabla 6, se observa que todas las direcciones que posee un número principal claro, es decir, un número luego del nombre de calle principal son etiquetadas correctamente.

Dirección	Nombre de Calle	Número	Info. Adicional
PJE. SENDA SANTA MARTA 0205 DEPTO. 21	PJE. SENDA SANTA MARTA	205	DEPTO. 21
CALLE 1 SUR 786	CALLE 1 SUR	786	
PJE 1 5183 POB. ALBORADA	PJE 1	5183	POB. ALBORADA
CALLE UNO SUR 786	CALLE UNO SUR	786	
ISMAEL BRICENO 551 23 14 PUCARA DE LAZANA	ISMAEL BRICENO	551	23 14 PUCARA DE LAZANA
ISMAEL BRICENO 1481 TORRE C DEPTO 32	ISMAEL BRICENO	1481	TORRE C DEPTO 32
PASAJE 4 40 VILLA HUELEN	PASAJE 4	40	VILLA HUELEN
1 ORIENTE 1985 DP 705	1 ORIENTE	1985	DP 705
1 ORIENTE 1985	1 ORIENTE	1985	
PASAJE VILLA MAYOR NORTE 357	PASAJE VILLA MAYOR NORTE	357	
PASAJE LA CASA PIEDRA 1360 DPTO 32	PASAJE LA CASA PIEDRA	1360	DPTO 32
AV LO CRUZAT 555 DEP 532	AV LO CRUZAT	555	DEP 532
PJE OLLAGUE 251 A A 11	PJE OLLAGUE	251	A A 11

Tabla 6. Ejemplo direcciones correctamente etiquetadas

Dirección	Nombre de Calle	Número	Info. Adicional
PARINACOTA BLOCK 515 DPTO B-11	PARINACOTA BLOCK	515	DPTO B-11
PJE 5 BLOCK 752 DEPTO 106	PJE	752	BLOCK 752 DEPTO 106
CALLE 1 BL 23 DEPTO 3	CALLE	23	BL 23 DEPTO 3
LAS VIOLETAS BLOCK 565 DPTO. A-22	LAS VIOLETAS BLOCK	565	DPTO. A-22
NUESTRA SENORA DEL CAMREN BLOCK 588 DEPTO 34	NUESTRA SENORA DEL CAMREN BLOCK	588	DEPTO 34
PARINACOTA S/N BLOCK 560 DEPTO 24 A	PARINACOTA S/N BLOCK	560	DEPTO 24 A
AV LO MARCOLETA S/N DEP A-23 BLOCK 0610	AV LO MARCOLETA S/N DEP A-23 BLOCK	610	
SAN MARTIN CON CHACABUCO 636 QUILICURA	SAN MARTIN CON CHACABUCO	636	QUILICURA
ISMAEL BRICENO, PASAJE 1 BLOCK 22 DEPTO 1	ISMAEL BRICENO, PASAJE	22	BLOCK 22 DEPTO 1

Tabla 7. Ejemplo direcciones incorrectamente etiquetadas

En la tabla 7, se puede observar que las direcciones no poseen un número principal claro, ya que la palabra “BLOCK” pertenece al diccionario que indica un tipo de vivienda y no como un posible número de calle. La estructura de direcciones que hacen referencia a esquinas, por ejemplo: “SAN MARTIN CON CHACABUCO” no están consideradas como una estructura válida en este clasificador. Por último, a las direcciones sin número o “S/N” no se les puede asignar un código postal, ya que este funciona en base a un número principal.

7.3 Asignación código postal

Para medir los resultados de la normalización se revisa manualmente la asignación de códigos en cada enfoque propuesto. Se contabiliza la cantidad de direcciones que se lograron normalizar y a las que no se logró asignar un código postal único. Además, se contabiliza la cantidad de direcciones normalizadas correctamente según una validación manual, es decir, aquellas

direcciones que se les asignó un código postal correspondiente a la misma calle de la dirección en cuestión.

Luego de obtener los resultados generales para cada enfoque, se selecciona el mejor para revisar los errores según el *puntaje de selección* con el fin de proponer las categorías de confiabilidad de normalización *segura*, *probable* y que *requiere revisión*.

7.3.1 Rendimiento general de enfoques de coincidencia

En esta sección se encuentran los resultados generales para cada enfoque de coincidencia, es decir se reporta la cantidad de direcciones normalizadas según el modelo y las direcciones normalizadas correctas según validación manual. En la Tabla 8 se encuentran los resultados de envíos a la comuna de Quilicura para cada uno de los enfoques propuestos:

Enfoque	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	5.1	5.2	6.1	6.2
Direcciones Totales a normalizar	1003	1003	1003	1003	1003	1003	1003	1003	1003	1003	1003	1003
Direcciones Válidas	932	932	932	932	932	932	932	932	932	932	932	932
Direcciones Inválidas	71	71	71	71	71	71	71	71	71	71	71	71
Normalizadas	898	927	920	930	898	923	932	907	920	926	932	925
Normalizadas correctamente	499	623	822	769	623	656	711	830	859	885	669	854
No normalizadas	34	5	12	2	34	9	0	25	12	6	0	7
% error validación manual	46,46%	33,15%	11,80%	17,49%	33,15%	29,61%	23,71%	10,94%	7,83%	5,04%	28,22%	8,37%

Tabla 8. Resultados generales de enfoques de coincidencia en envíos de Quilicura

La cantidad de direcciones "Normalizadas" corresponde al número de envíos a los cuales se les asignó un código postal único, es decir, cumplieron con los criterios de asignación. La suma de direcciones "Normalizadas" y "No Normalizadas" es igual a la cantidad de "Direcciones Válidas". Las direcciones "Normalizadas Correctamente" corresponden a aquellas que se validaron manualmente.

Se inicia la normalización a partir de las 1003 direcciones que no lograron la coincidencia directa. 71 direcciones fueron encontradas inválidas o mal escritas, lo que es igual para todos los enfoques debido a que el clasificador no cambia según el enfoque a utilizar. Con las 932 direcciones válidas o bien escritas se inicia la asignación de códigos postales. El % (porcentaje) de error según validación manual se calcula como: $\frac{Normalizadas\ Correctamente}{Direcciones\ Válidas}$. Esto, se debe a que se busca determinar qué tan confiable es el enfoque para la normalización. Además, un mismo denominador común facilita la comparación entre modelos.

		Puntaje Ranking con distancia de Levenshtein (Enfoque X.1)			Puntaje Ranking con Ratio Similitud Levenshtein (Enfoque X.2)		
	Enfoque	**Normalizadas Correctamente según revisión manual**	**No Normalizadas**	**% error normalización**	**Normalizadas Correctamente según revisión manual**	**No Normalizadas**	**% error normalización**
Comparación palabra por palabra	**1**	499	34	46,46%	623	5	33,15%
	2	822	12	11,80%	769	2	17,49%
	3	623	34	33,15%	656	9	29,61%
Comparación cadena completa	**4**	711	0	23,71%	830	25	10,94%
	5	859	12	7,83%	**885**	**6**	**5,04%**
	6	669	12	28,22%	854	7	8,37%

Tabla 9. Resultados Enfoques de coincidencia

En general, todos los modelos asignaron correctamente más del 50% de las direcciones a normalizar. Sin embargo, al observar la Tabla 9, se puede notar que los modelos que utilizan una comparación palabra por palabra reportaron resultados consistentemente peores que los enfoques que, utilizando los mismos filtros (enfoques 1 y 4, 2 y 5, 3 y 6 utilizan filtros iguales), comparan la cadena completa de texto. Además, para los enfoques 4, 5 y 6, calcular el *puntaje de ranking* utilizando el ratio de similitud de Levenshtein reporta mejores resultados que si se calcula utilizando la distancia de Levenshtein.

Si bien, los enfoques 4.1 y 6.1 lograron una asignación única para las 932 direcciones, poseen un mayor % de error según validación manual. Esto implica que un enfoque que logra asignar códigos postales a todos los envíos no necesariamente es aquel de mejor rendimiento.

El enfoque 5.2, que busca coincidir cadenas de texto completas, genera el puntaje de ranking en base al ratio de similitud de Levenshtein y obtuvo el menor % de error según validación manual con un 5,04%, comparando las 885 direcciones normalizadas correctamente sobre las 932 direcciones válidas. Las 6 direcciones que no se lograron normalizar se consideran dentro del error según validación manual, con el fin de poder comparar los diferentes enfoques.

Las 6 direcciones no normalizadas corresponden a un mismo nombre de calle, es decir, son 6 envíos a la misma dirección. Este caso es explicitado en la Tabla 10 y pueden ser fácilmente normalizadas por un operador utilizando las propuestas entregadas por el modelo.

Dirección Original	Nombre de Calle	Número de Calle	Calle Propuesta	Numero Propuesto	Puntaje Ranking	Distancia números	Puntaje Selección	Max P. Selección	CP propuesto
LOS NAUQUES 785	LOS NAUQUES	785	LOS NONQUES	785	82	0	82	0	8731494
LOS NAUQUES 785	LOS NAUQUES	785	PASAJE LOS MAQUIS	46	76	739	86	1	8700018
LOS NAUQUES 785	LOS NAUQUES	785	PASAJE LOS QUENES	243	76	542	86	1	8720953

Tabla 10. Dirección no normalizada en envíos de Quilicura

El problema que se observa en la Tabla 10 corresponde a la normalización de la dirección "*LOS NAUQUES 785"*. En primer lugar, esta dirección no está registrada en Google Maps, sin embargo, si existe "*Los NONQUES 785"*, por ende, es posible que la dirección "*LOS NAUQUES"* esté mal

escrita. En segundo lugar, el enfoque 5.2 elimina los prefijos de calle, como "*PASAJE"*, lo que resulta en un puntaje ranking mayor para "*LOS NONQUES"*, pero al calcular el puntaje de selección se considera la palabra "*PASAJE"* lo que aumenta el ratio de similitud de Levenshtein debido al largo de las cadenas. En tercer lugar, existen dos puntajes de selección más altos. Y por último, la menor distancia entre el número de calle y número propuesto no coincide con ninguno de los dos candidatos con mayor puntaje de selección, por lo que no se logra asignar un código postal.

Enfoque	**5.2**
Direcciones Totales a normalizar	662
Direcciones Válidas	618
Direcciones Inválidas	44
Normalizadas	617
Normalizadas correctamente	580
No normalizadas	1
% error validación manual	6,15%

Tabla 11. Resultado general del mejor enfoque en envíos de Santiago

Debido a que el enfoque 5.2 reporta los mejores resultados, fue aplicado a envíos a la comuna de Santiago y se obtuvieron los resultados reportados en la tabla 11.

Los resultados del enfoque 5.2 en envíos a la comuna de Santiago son similares con los resultados en envíos a la comuna de Quilicura en cuanto al porcentaje de error según validación manual, ya que se reporta un 6,15% y 5,04% respectivamente. Además, solo un envío no se logra normalizar, el cual se encuentra especificado en la tabla 12.

El problema que se observa en la tabla 12 corresponde a una dirección perteneciente a la comuna de Lampa y no de Santiago, por lo que se puede intuir que la dirección fue incorrectamente incluida dentro de los envíos a la comuna de Santiago. Este error escapa de los alcances del modelo propuesto.

Dirección Original	**Nombre de Calle**	**Número de Calle**	**Calle Propuesta**	**Numero Propuesto**	**Puntaje Ranking**	**Distancia números**	**Puntaje Selección**	**CP propuesto**
GENERAL SAN MARTIN NORTE 305 LAMPA	GENERAL SAN MARTIN NORTE	305	NUEVA SAN MARTIN	1490	70	1185	86	8340513
GENERAL SAN MARTIN NORTE 305 LAMPA	GENERAL SAN MARTIN NORTE	305	GENERAL MITRE	1905	65	1600	86	8361157
GENERAL SAN MARTIN NORTE 305 LAMPA	GENERAL SAN MARTIN NORTE	305	SAN MARTIN INTERIOR	14	65	291	70	8371067

Tabla 12. Dirección no normalizada en envíos de Santiago

7.3.2 Generación de categorías de confiabilidad de asignaciones

Debido a la existencia de errores en la normalización con el modelo, se ha propuesto la generación de categorías de confiabilidad con el propósito de asegurar el mínimo error posible. Para esto, se ha realizado una apertura de los resultados según el puntaje de selección. Para los envíos a la comuna de Quilicura, la apertura de resultados se muestra en la Tabla 13.

La Tabla 13, posee la cantidad de direcciones normalizadas correctamente según revisión manual, normalizadas por el modelo, error porcentual, error absoluto y el porcentaje de direcciones normalizadas correctas sobre el total de normalizadas (926 en Quilicura). Las filas resaltadas en negrita corresponden a puntajes en donde existe un aumento importante en el error absoluto de la normalización.

Puntajes Selección	Normalizadas Correctamente según revisión manual	Normalizadas	Error	Error Absoluto	% del total de Propuestas
100	235	235	0,00%	0	25,38%
98	236	236	0,00%	0	25,49%
97	248	248	0,00%	0	26,78%
96	252	252	0,00%	0	27,21%
95	393	393	0,00%	0	42,44%
94	395	395	0,00%	0	42,66%
92	403	403	0,00%	0	43,52%
91	404	405	0,25%	1	43,63%
90	**774**	**781**	**0,90%**	**7**	**83,59%**
89	781	788	0,89%	7	84,34%
88	786	794	1,01%	8	84,88%
87	787	795	1,01%	8	84,99%
86	**850**	**874**	**2,75%**	**24**	**91,79%**
83	854	878	2,73%	24	92,22%
82	855	879	2,73%	24	92,33%
81	858	882	2,72%	24	92,66%
80	860	886	2,93%	26	92,87%
79	865	891	2,92%	26	93,41%
78	874	900	2,89%	26	94,38%
77	877	903	2,88%	26	94,71%

Tabla 13. Apertura de resultados según Puntaje Selección en enfoque 5.2 en envíos de Quilicura

Las categorías de confiabilidad se construyen en base a la cantidad de errores que poseen en cada nivel de puntaje de selección. La categoría de *normalización segura* se propone para aquellos envíos con puntaje de selección mayor o igual a 91, ya que al nivel 90 existe un aumento en el error. En esta categoría, entra un 43,63% de las direcciones normalizadas correctamente. Aunque al nivel 91 se encuentra un solo error este corresponde a una dirección que no existe en la base oficial. Este error se muestra en la Tabla 14.

Dirección Original	Nombre de Calle	Número De calle	Calle Propuesta	Numero Propuesto	Puntaje Ranking	Distancia números	Puntaje Selección	CP propuesto	Propuesta
JARDIN DE MARTE NORTE 582	JARDIN DE MARTE NORTE	582	JARDIN DE MARTE ORIENTE	439	91	143	91	8722148	1
JARDIN DE MARTE NORTE 582	JARDIN DE MARTE NORTE	582	JARDIN DE MARTE SUR	582	85	0	85	8722138	0
JARDIN DE MARTE NORTE 582	JARDIN DE MARTE NORTE	582	JARDIN DE MARTE PONIENTE	437	84	145	84	8722164	0

Tabla 14. Error en categoría normalización *segura* (Quilicura)

Al buscar la dirección "JARDIN DE MARTE NORTE" en la base oficial no se encuentran registros. Una explicación para este error es que la calle pertenezca a un lote nuevo de casas y calles, ya que existe la dirección con las terminaciones "ORIENTE", "SUR" y "PONIENTE". Esto puede implicar que existen otros errores en la normalización, debido a que, no existe la calle buscada en la base oficial.

La categoría *normalización probable* se propone para envíos con puntaje de selección menor a 91 y mayor o igual a 87. Esto se debe a que el error asociado es cercano al 1%, además, al puntaje de selección 86 el error absoluto sube al triple. Por otro lado, al considerar normalizados los envíos en esta categoría se llega al 84,99% de normalización correcta.

Algunos errores repetidos en esta categoría se muestran en la tabla 15. El primero, la palabra "O'HIGGINS" pertenece al nombre de más de una dirección, siendo estas "PASAJE O'HIGGINS", "AMBROSIO O'HIGGINS" y "AVENIDA BERNARDO O'HIGGINS" (no fue propuesta por el modelo). Esta dirección puede ser considerada como errónea, debido a la ambigüedad que posee la palabra "O'HIGGINS", lo cual dificulta su normalización. Sin embargo, puede ser asignada discriminando por su número propuesto. El segundo error en la tabla pertenece a la dirección "AV LAS TORRES NORTE", la cual al ser buscada en la base oficial no fue encontrada. Este último error, es similar al encontrado en la categoría *normalización segura,* en donde la dirección puede pertenecer a un lote relativamente nuevo de casas.

Dirección Original	Nombre de Calle	Número De calle	Calle Propuesta	Numero Propuesto	Puntaje Ranking	Distancia números	Puntaje Selección	CP propuesto	Propuesta
O HIGGINS 365	O HIGGINS	365	PASAJE O"HIGGINS	287	89	78	90	8720300	0
O HIGGINS 365	O HIGGINS	365	AMBROSIO O"HIGGINS	383	67	18	90	8700430	1
O HIGGINS 365	O HIGGINS	365	PASAJE LOS GEORGIANOS	245	61	120	57	8720205	0
AV LAS TORRES NORTE 242	LAS TORRES NORTE	242	AVENIDA LAS TORRES ORIENTE	116	88	126	86	8732451	0
AV LAS TORRES NORTE 242	LAS TORRES NORTE	242	LAS TORRES ORIENTE	540	88	298	88	8700464	1
AV LAS TORRES NORTE 242	LAS TORRES NORTE	242	AVENIDA LAS TORRES SUR	197	80	45	73	8722189	0

Tabla 15. Errores en categoría normalización *probable*

Dirección Original	Nombre de Calle	Número de Calle	Calle Propuesta	Número Propuesto	Puntaje Ranking	Distancia números	Puntaje Selección	CP propuesto	Propuesta
PASAJE CORDOBA 0422	CORDOBA	422	PASAJE CORDOVA	422	86	0	77	8721011	0
PASAJE CORDOBA 0422	CORDOBA	422	CORDOVA	319	86	103	86	8722026	1
PASAJE CORDOBA 0422	CORDOBA	422	AVENIDA COLORADO	340	67	82	51	8730613	0
DE LA TRILLA 516	DE LA TRILLA	516	PASAJE DEL POTRILLO	289	75	227	68	8720265	0
DE LA TRILLA 516	DE LA TRILLA	516	PASAJE DE LA ERMITA	712	75	196	86	8722211	1
DE LA TRILLA 516	DE LA TRILLA	516	CALLE DEL TRIGAL	511	73	5	57	8721953	0
PANAMERICANA NORTE 8550	PANAMERI-CANA NORTE	8550	ANDALUCIA NORTE	1322	67	7228	67	8701589	0
PANAMERICANA NORTE 8550	PANAMERI-CANA NORTE	8550	ALCALA NORTE	1363	67	7187	86	8701554	1
PANAMERICANA NORTE 8550	PANAMERI-CANA NORTE	8550	PASAJE MONTERA NORTE	1587	65	6963	68	8701486	0

Tabla 16. Ejemplos de errores en categoría de normalización *requiere revisión*

Por último, la categoría *normalización requiere revisión*, se propone para envíos con puntaje de selección menor a 87. Algunos errores encontrados en esta categoría se muestran en la Tabla 16. Aquí se encuentran todas las direcciones normalizadas que no pertenecen a las categorías *segura* y *probable.*

Los tres ejemplos de errores en la Tabla 16 corresponde a envíos normalizados con 86 puntos de selección. El primer error corresponde a la dirección "PASAJE CORDOBA", sin embargo, se asigna un código perteneciente a la calle "CORDOVA". Aunque a dirección correcta corresponde a "PASAJE CORDOVA", el puntaje de selección se calcula comparando *Nombre de Calle* con *Calle Propuesta*, por lo que la palabra "PASAJE" agrega ruido a la comparación, lo que disminuye el puntaje de selección. El número propuesto en la alternativa de calle "CORDOVA" no posee la misma centena ni paridad que el número buscado, por lo tanto, no se encontró en la base oficial y se asignó el número más cercano, lo que puede dar indicios de direcciones incorrectamente normalizadas. Para el segundo y tercer error, no se encontró la dirección buscada en la base de datos oficial.

Debido a que, el enfoque 5.2 se aplicó a envíos en la comuna de Santiago, la apertura de estos resultados se muestra en la Tabla 17. Si se aplican las mismas categorías de confiabilidad se observa que propuestas de *normalización segura* y *normalización probable* no poseen errores (0% error). Y si se aceptan todas las propuestas en estas dos categorías se alcanza el 87,03% de las direcciones normalizadas correctamente (537 de 617).

Puntajes Selección	Normalizadas Correctamente según revisión manual	Normalizadas	error	error absoluto	% del total de propuestas
100	273	273	0,00%	0	44,25%
97	275	275	0,00%	0	44,57%
96	276	276	0,00%	0	44,73%
95	311	311	0,00%	0	50,41%
94	313	313	0,00%	0	50,73%
92	325	325	0,00%	0	52,67%
90	**534**	**534**	**0,00%**	**0**	**86,55%**
88	535	535	0,00%	0	86,71%
87	537	537	0,00%	0	87,03%
86	**554**	**554**	**0,00%**	**0**	**89,79%**
83	557	557	0,00%	0	90,28%
82	579	579	0,00%	0	93,84%
77	579	582	0,52%	3	93,84%
67	579	585	1,03%	6	93,84%
65	579	590	1,86%	11	93,84%

Tabla 17. Apertura de resultados según Puntaje Selección en enfoque 5.2 en envíos de Santiago

Para comparar los resultados de las categorías de confiabilidad para envíos de Quilicura y Santiago se construye la Tabla 18. En ella se encuentra el error en la normalización, y la cantidad de direcciones que pertenecen a cada categoría. Se puede observar que el modelo entrega una normalización perfecta para envíos de Santiago si se considera solo las categorías *segura* y *probable*, mientras que para envíos de Quilicura solo se alcanza 1,01% de error.

	Envíos Quilicura			Envíos Santiago		
	Normalización Segura	Normalización Probable	Total	Normalización Segura	Normalización Probable	Total
Dir. Válidas	932	932	932	618	618	618
Normalizadas	926	926	926	617	617	617
Normalizadas en categoría	405	390	795	325	212	537
Normalizadas Correctas en categoría	404	383	787	325	212	537
Error	0,25%	1,79%	1,01%	0,00%	0,00%	0,00%

Tabla 18. Resultados categorías de confiabilidad para envíos de Quilicura y Santiago

Se observa que, para envíos de Quilicura, las categorías de confiabilidad *segura* y *probable* poseen errores. Estas asignaciones incorrectas se deben a que el envío dirección no pudo ser encontrado en la base oficial de direcciones. Una explicación para estos errores corresponde a una

desactualización de la base oficial, es decir, se han creado nuevas calles o lotes inmobiliarios que no han sido incluidos en esta base de datos. Este fenómeno se puede observar con mayor frecuencia en comunas que están en crecimiento como lo es Quilicura. Por otro lado, la comuna de Santiago no está en crecimiento ni aumentan la cantidad de calles, así mismo, los envíos con destino a la misma comuna no poseen errores en las categorías *segura* y *probable*.

7.4 Resultados generales

Debido a que, se busca aumentar el porcentaje de normalización a través de modelos de procesamiento del lenguaje natural, se reportan los resultados generales considerando el total de envíos de clientes para Quilicura y Santiago desde que se realiza la coincidencia directa hasta la aplicación de categorías de confiabilidad.

	Quilicura	**Porcentaje (#/1254)**	**Santiago**	**Porcentaje (#/807)**
Direcciones totales	1254	100,00%	807	100,00%
Coincidencia directa	251	20,02%	145	17,97%
Direcciones totales a normalizar	1003	79,98%	662	82,03%
Direcciones válidas	932	74,32%	618	76,58%
Direcciones Invalidas	71	5,66%	44	5,45%
Normalizadas	926	73,84%	617	76,46%
No normalizadas	6	0,48%	1	0,12%
Correctamente Normalizadas	885	70,57%	580	71,87%
Normalizadas *seguras* + *probables*	795	63,40%	537	66,54%
Error correctamente normalizadas	41	3,27%	37	4,58%
Error normalizadas seguras + probables	8	0,64%	0	0,00%
Total Normalizadas	(926 + 251)	93,86%	(617+145)	94,40%
Total Normalizadas *seguras* + *probables*	(795+251)	83,41%	(537+145)	84,51%

Tabla 19. Resultados generales considerando todos los envíos

De la Tabla 19, se observa que la coincidencia directa para envíos de Quilicura alcanza un 20,02% y para envíos de Santiago 17,97%. Por lo tanto, se necesita normalizar el 79,98% y 82,03% de los envíos de Quilicura y Santiago, respectivamente.

Al aplicar el modelo clasificador, se obtiene que un 5,66% y 5,45% de los envíos de Quilicura y Santiago, respectivamente, poseen una dirección inválida, en otras palabras, no se identifica un

número de calle específico. Consecuentemente, solo a un 74,32% y 76,58% de las direcciones de Quilicura y Santiago fueron trabajadas por el modelo de coincidencia de direcciones para su normalización.

Debido a la elección del enfoque 5.2 como mejor modelo, un 0,48% y 0,12% de direcciones de Quilicura y Santiago respectivamente, no se logran normalizar, ya que no cumplieron los criterios de asignación. Esto implica que la normalización alcanza un 73,84% (926) y 76,46% (617) para Quilicura y Santiago (envíos que fueron intervenidos por el modelo de coincidencia de direcciones). De las 926 direcciones normalizadas de Quilicura, se validan correctas 885, es decir, posee un error de 4,42%. De las 617 direcciones normalizadas de Santiago, se validan correctas 580, es decir que posee un error de 6,00%.

Si se consideran todas las direcciones normalizadas y le agrega las direcciones por coincidencia directa, se alcanza la normalización de 93,86% y 94,4% de los envíos totales para Quilicura y Santiago, respectivamente. Con un error de 3,27% para Quilicura y 4,58% para Santiago

Por otro lado, si se acepta sólo la asignación de código postal de aquellas direcciones que lograron las categorías *segura y probable*, se encuentra que, de los envíos totales (considerando coincidencia directa), se logra 83,41% y 84,51% de normalización con un error de 0,64% y 0,00% para Quilicura y Santiago, respectivamente.

7.5 Costo computacional

Debido a que, los modelos funcionan en base a información que se encuentra en formato de texto su procesamiento resulta ser costoso en términos de tiempo. En la tabla 20, se encuentran los tiempos promedios para el clasificador de texto en la base oficial y base de clientes, como también para los modelos de asignación de código postal en cada uno de sus enfoques.

Modelo	**Tiempo Promedio**
Clasificador Reglas Lógicas Base Cliente (Quilicura)	1,07 seg ± 0,06 seg
Clasificador Reglas Lógicas Base Oficial (Quilicura)	33,27 seg ± 0,91 seg
Clasificador Reglas Lógicas Base Cliente (Santiago)	0,43 seg ± 0,04 seg
Clasificador Reglas Lógicas Base Oficial (Santiago)	42,14 seg ± 1,02 seg
Enfoque 1.1	7 min 48 seg ± 1min 7,04 seg
Enfoque 1.2	14 min 15 seg ± 1min 52,65 seg
Enfoque 2.1	6 min 34 seg ± 57,11 seg
Enfoque 2.2	14 min 6 seg ± 1min 48,65 seg
Enfoque 3.1	6 min 48 seg ± 1 min 1,09 seg
Enfoque 3.2	15 min 15 seg ± 1min 58,97 seg
Enfoque 4.1	5 min 36 seg ± 44,11 seg
Enfoque 4.2	9 min 25seg ± 1min 15,43 seg
Enfoque 5.1	5 min 31 seg ± 42,85 seg
Enfoque 5.2	8 min 58 seg ± 1min 12,75 seg
Enfoque 6.1	5 min 23 seg ± 40,21 seg
Enfoque 6.2	8 min 34 seg ± 1min 11,02 seg
Enfoque 5.2 (Santiago)	4 min 37 seg ± 37,46 seg

Tabla 20. Costo Computacional

En la Tabla 19, se puede observar que el clasificador de texto no demora más de un minuto en ser ejecutado. Sin embargo, este tiempo es alto considerando que la base oficial de direcciones de Santiago posee un poco más de 57 mil datos y el clasificador demora sobre 42 segundos en etiquetarlas. Esto se debe a que, para cada dirección, se comprueba si cumple con las reglas lógicas o no y luego se etiqueta. Pueden existir modelos probabilísticos que disminuyan este tiempo, sin embargo, al ser un modelo de aprendizaje no supervisado se privilegia la utilización de información incluida en las cadenas de texto.

Del mismo modo, los enfoques de coincidencia comparan cada dirección con todas las direcciones de la base oficial aumentando el tiempo de modelamiento. En la tabla 19 se observa que el enfoque 5.2 (de mejor rendimiento) posee un tiempo promedio de 8 minutos y 58 segundos, mientras que el enfoque más lento corresponde al 3.2 con 15 minutos y 15 segundo y el más rápido al 6.1 con 5 minutos y 23 segundos.

8 CONCLUSIONES Y TRABAJO FUTURO

Para realizar la normalización de direcciones se utilizaron dos modelos de procesamiento del lenguaje natural: un clasificador de texto en base a reglas lógicas y un modelo de coincidencia de textos para la asignación del código postal. Antes de la utilización de los modelos se lograba normalizar un 20% de direcciones a través de coincidencia directa. Después de la utilización de los modelos se logra sobre 90% de direcciones normalizadas considerando un error menor al 5%. Se concluye que utilizando modelos de procesamiento del lenguaje natural se consigue aumentar automáticamente la normalización de direcciones.

Previo a la utilización de un modelo de normalización de direcciones, se logra asignar códigos postales a 251 de 1254 envíos de Quilicura y 145 de 807 envíos de Santiago, a través de una coincidencia directa. Los envíos restantes son normalizados utilizando los modelos propuestos en esta memoria.

El clasificador en base a reglas lógicas segmenta y etiqueta las direcciones en *nombre de calle, número de calle* e *información adicional.* Al ser un modelo de aprendizaje no supervisado, no se obtiene una métrica asociada al error en la clasificación, sin embargo, identifica direcciones que se escapan de la estructura propuesta o que no se logra identificar un número de calle. Para direcciones de Quilicura se encontraron 71 direcciones inválidas de 1003, mientras que para Santiago se encontraron 44 de 662. Estas direcciones son excluidas de la normalización, ya que aumentan el error del modelo y pueden ser normalizadas o revisadas por un operador.

Para el modelo de asignación de códigos postales se estudia el rendimiento de 12 enfoques de coincidencia, los cuales varían según la métrica de similitud (distancia de Levenshtein o ratio de similitud de Levenshtein), el filtro utilizado en el preprocesamiento de los nombres de calle (elimina, completa o no modifica las abreviaturas) y el algoritmo de coincidencia (comparar palabra por palabra o cadena completa de texto). El modelo, en primer lugar, aplica un filtro para preprocesar el nombre de calle, luego genera un ranking de las tres direcciones más similares utilizando el puntaje de ranking. Por último, se calcula el puntaje de selección con el cual se verifican los criterios de asignación y se normaliza el envío. Los criterios de asignación son: se elige la dirección con máximo puntaje de selección, si hay más de uno se discrimina utilizando el número de calle, eligiendo el número más cercano. El número de calle propuesto se elige buscando el número idéntico al número de calle, si no se encuentra, se busca otro que posea la misma paridad y centena, y en caso de no cumplir ninguna de las anteriores, se asigna el número más cercano.

Los enfoques de coincidencia que utilizan la comparación de cadena completa obtuvieron mejores resultados que aquellos que comparan palabra por palabra. Además, los enfoques que utilizan el ratio de similitud de Levenshtein para calcular el puntaje ranking obtuvieron mejores resultados que aquellos que utilizaban la distancia de Levenshtein. El mejor rendimiento de los modelos de coincidencia lo obtuvo el enfoque 5.2, el cual utiliza el filtro que elimina las abreviaturas, genera el puntaje ranking en base al ratio de similitud de Levenshtein y compara la cadena completa de texto. De 932 envíos válidos de Quilicura, se logró normalizar 926 de los cuales, a través de una validación manual, se comprobó que 885 códigos postales fueron asignados correctamente obteniendo un error de 5,04%. De 618 envíos válidos de Santiago, se logró normalizar 617 de los cuales, a través de una validación manual, se comprobó que 580 códigos postales fueron asignados correctamente obteniendo un error de 6,15%.

Con el objetivo de disminuir el error en el proceso de normalización se propone utilizar categorías de confiabilidad en base al puntaje de selección. Se define la categoría normalización *segura* para aquellos envíos con puntajes de selección entre 100 y 91 (incluido). La categoría normalización *probable* para envíos con puntajes de selección entre 90 y 87 y, finalmente, la categoría normalización *requiere revisión* para envíos con puntaje de selección igual o menor que 86.

Para los envíos de Quilicura: la categoría normalización *segura* posee solo un envío con asignación errónea, el cual se debe a que no se encuentra la dirección buscada en la base oficial y se considera que ese error no es generado por el modelo. En la categoría normalización *probable* se encuentran 7 asignaciones incorrectas, lo que conlleva un 1,01% de error. En la categoría normalización *requiere revisión* se encuentran los demás errores. Por otro lado, para envíos de Santiago, las categorías normalización *segura* y *probable* no poseen asignaciones incorrectas, en otras palabras, poseen error de 0,00%.

Previo a la utilización de modelos de procesamiento de lenguaje natural para normalización de direcciones, se logra asignar códigos postales por coincidencia directa a un 20,02% de los envíos de Quilicura y 17,97% de los envíos de Santiago. Al agregar los envíos normalizados utilizando modelos inteligentes, se alcanza un 93,86% de normalización para Quilicura y 94,40% para Santiago con un error asociado de 3,27% y 4,58% respectivamente. Si solo se aceptan los envíos normalizados pertenecientes a las categorías *segura* y *probable*, se alcanza un 83,41% de normalización para Quilicura y 84,51% para Santiago con un error asociado de 0,64% y 0,00% respectivamente.

El costo asociado al proceso de normalización de direcciones se dimensiona en base al sueldo de los operadores que realizan esta labor, el cual se estima; considerando todas las plantas de clasificación, en más de $26 millones de pesos (CLP) mensuales. Sin embargo, pueden existir mayores costos asociados a un mayor tiempo del envío en el sistema y al reproceso de envíos que no pudieron ser entregados. Con la solución propuesta en esta memoria, se alcanza a normalizar más del 80% de las direcciones con un error menor al 1%. Esto permitiría disminuir de dos a un operador, lo que significa una rebaja de al menos $13 millones de pesos mensuales.

Se deja como propuesto incluir nuevas estructuras de direcciones para el clasificador de texto, con el objetivo de disminuir la cantidad de direcciones inválidas. Esto implica, reconocer otras formas de escritura de direcciones e identificar el nombre y número de calle correctamente. La existencia del clasificador en base a reglas lógicas facilita la creación de una base de entrenamiento para modelo probabilísticos de clasificación de texto, aumentando las posibilidades de reconocer estructuras más complejas.

También, se deja propuesto para el modelo de asignación de código postal incluir números que comiencen con “0” y diferenciarlos de aquellos números idénticos que no posean “0” al inicio, por ejemplo: “63” y “063”. Realizar este perfeccionamiento implica una mejor normalización debido a que el “0” afecta el código postal que se asigna. En adición, se propone identificar los envíos normalizados incorrectamente utilizando la información de la columna número propuesto, debido a que, si no cumple con los criterios de asignación de número (calce perfecto, o misma centena y paridad) puede que no haya encontrado la coincidencia correcta para la normalización. Con esto, se puede facilitar el trabajo que realiza el operador y aumentar el porcentaje de normalización.

Los diccionarios de abreviaturas utilizados en los modelos de clasificación y coincidencia de texto son acotados a las direcciones estudiadas. Por esta razón, al extender el modelo a nuevas comunas se recomienda actualizar y mejorar estos diccionarios, ya que pueden tener un impacto importante al momento de comparar direcciones.

Por último, se deja como propuesto explorar otras herramientas para el proceso de normalización y se recomienda complementarlas con la solución propuesta en esta memoria. Algunas otras herramientas son la asignación de código postal según datos georreferenciados, en otras palabras, se utilizan datos históricos, obtenidos desde los equipos PDA (Personal Digital Assistant) utilizados por los carteros, para asociar la dirección a normalizar con direcciones idénticas entregadas anteriormente y sus coordenadas geográficas de entrega, con estos datos se puede triangular la posición de entrega y asignar un código postal. Asimismo, a partir de información histórica potencialmente almacenable por las compañías postales, se puede identificar el destinatario y asociar todas las direcciones donde ha recibido envíos, luego buscar la dirección más probable y normalizarla.

La herramienta utilizada para la normalización en este proyecto se basa en la similitud entre palabras. El beneficio de este enfoque es que sólo se necesita la dirección del destinatario y una base estandarizada de las direcciones de la comuna de destino. Por otro lado, se requiere contar con conocimiento de la estructura y escritura de las direcciones a normalizar. Dicho esto, el proceso de asignación de códigos postales en base a modelos de procesamiento del lenguaje natural reporta buenos resultados alcanzando a normalizar más del 80% de las direcciones con un error cercano a 0% y a normalizar más del 90% de las direcciones con un error menor al 5%. Con este modelo se logra aumentar considerablemente el porcentaje de normalización de manera automática.

9 BIBLIOGRAFÍA

Accenture. (2018). *How Could Last Mile Delivery Evolve to Sustainably Meet Customer Expectations?*

Aggarwal, C. C. (2015). Edit Distance. En *Data Mining* (págs. 82-84). New York: Springer.

Asiain, F. (2017). *Herramienta de Simulación para Evaluación de Rendimiento de Cuarteles de Correos de Chile.* Santiago.

BareInternational.cl. (28 de Enero de 2019). Obtenido de 2019: Tendencias E-Commerce En Chile y La Experiencia de Clientes: https://www.bareinternational.cl/tendencias-2019-de-ecommerce-en-chile/

BCG. (10 de Octubre de 2017). Obtenido de Boston Consulting Group Web Site: https://www.bcg.com/industries/transportation-travel-tourism/center-digital-transportation/postal-parcel

Carrera Arias, F. J. (Febrero de 2019). *Datacamp.com*. Obtenido de https://www.datacamp.com/community/tutorials/fuzzy-string-python

Casanova, F. (2018). *Historia de Nuestra Historia*. Obtenido de https://hdnh.es/historia-y-evolucion-del-correo-postal/#:~:text=El%20primer%20uso%20documentado%20de,y%20data%20del%20255%20a.C.

Chawla, R. (7 de Agosto de 2017). *Medium*. Obtenido de Overview of Conditional Random Fields: https://medium.com/ml2vec/overview-of-conditional-random-fields-68a2a20fa541

Delgado, N., Martínez, G., & Covas, D. (2015). Procedimiento Para La Mejora Del Servicio De Envíos De Mensajería DHL Express, Perteneciente A La Empresa De Correos Cienfuegos. *Revista Científica "Visión de Futuro", 19*(1), 103-120. Obtenido de https://revistacientifica.fce.unam.edu.ar/index.php?option=com_content&view=article&id=380&Itemid=83

Diario Financiero. (18 de Mayo de 2020). *Despachos a Domiciol Aumentarán su Participación en Ventas Online al 85% en 2021*. Obtenido de https://www.df.cl/noticias/empresas/industria/despachos-a-domicilio-aumentaran-su-participacion-en-ventas-online-al/2020-05-18/210250.html

Elfenbaum, D. (6 de Marzo de 2020). *CCM*. Obtenido de https://es.ccm.net/faq/10424-codigo-postal-de-chile

Empresa de Correos de Chile. (2019). *Memoria Anual 2019.* Santiago.

Empresa de Correos de Chile. (30 de Mayo de 2020). *Marco Normativo*. Obtenido de Correos Transparente: https://correostransparente.correos.cl/marco-normativo.html

Empresa de Correos de Chile; BBVA Asesorías Financieras S.A. (2017). *Prospecto Comercial Bono Serie A.* Santiago, Chile.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

Marrero, R. (2007). *Estandarización de Direcciones Postales de Clientes de ETECSA Villa Clara (Trabajo de Diploma).* Trabajo de Diploma, Universidad Central "Marta Abreu" de las Villas.

MonkeyLearn.com. (2019). *MonkeyLearn*. Obtenido de https://monkeylearn.com/text-classification/#:~:text=Rule%2Dbased%20approaches%20classify%20text,categories%20based%20on%20its%20content.&text=Rule%2Dbased%20systems%20are%20human,can%20be%20improved%20over%20time.

Rivas, D. (2016). *PROPUESTA DE GUÍA METODOLÓGICA PARA EL MANEJO DE PROBLEMAS EN LA ESTANDARIZACION Y CALIDAD DE DATOS DE DIRECCIONES URBANAS EN COLOMBIA (Tesis de Magister).* Tesis de Magister, Universidad Pontificia Bolivariana, Medellín.

Schneider, C. (25 de Mayo de 2016). *IBM*. Obtenido de The biggest data challenges that you might not even know you have: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

Singla, N., & Garg, D. (6 de Enero de 2012). String Matching Algorithms and their Applicability in various Applications. *International Journal of Soft Computing and Engineering (IJSCE), I*(6), 218-222.

Sutton, C., & McCallum, A. (2011). An Introdution to Conditional Random Fields. *Foundation and Trends in Machine Learning, 4*(4), 267-373. doi:10.1561/2200000013

Tran, V.-T. (2019). A Novel Conditional Random Fields Aided Fuzzy Matching in Vietnamese Address Standardization. *The Tenth International Symposium on Information and Communication Technology* (pág. 6). Hanoi: ACM. doi:10.1145/3368926.3369687

Universal Postal Union. (2013). *UPU Annual Report 2013.* Berne.

Universal Postal Union. (2016). *Research On Postal Markets.* Berne.

Wang, M., Haberland, V., Yeo, A., Martin, A., Howroyd, J., & Bishop, J. (2016). A probabilistic parser using Conditional Random Fields and Stochastic Regular Grammar. IEEE.

Wang, Y., & Xiao-Jing, W. (2005). A new approach to feature selection in text classification. *International Conference on Machine Learning and Cybernetics*, *6*, págs. 3814-3819. Guangzhou. doi:10.1109/ICMLC.2005.1527604

ANEXO A

Distancia de Levenshtein – Extensión

Uno de los algoritmos más recomendado corresponde al de distancia de edición, en específico a la distancia de edición Levenshtein, que permite minimizar los fallos de selección por causa de errores ortográficos. La distancia de edición hace referencia al número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. Se entiende por operación, bien una inserción, eliminación o sustitución de un caracter. Por ejemplo, asumiendo un costo unitario: la distancia de Levenshtein entre "casa" y "calle" es de 3 porque se necesitan al menos tres ediciones elementales para cambiar uno en el otro. Existen modificaciones al modelo que permiten otorgar diferentes costos a cada edición y a cada carácter basada en contexto de la data de entrenamiento, los cuales son considerados modelos estocásticos comúnmente usados en correctores ortográficos de texto.

Sean dos vectores $X = \{x_1, \dots, x_n\}$ y $Y = \{y_1, \dots, y_m\}$, se busca calcular la distancia de edición para transformar X en Y. Es necesario notar que D(X,Y) no es lo mismo que D(Y,X) si es que los costos de inserción y eliminación no son los mismos (Aggarwal, 2015). Sea X_i el segmento de los primeros i elementos de X, y sea Y_j el segmento de los primeros j elementos de Y. Sea D(i,j) el costo óptimo de edición entre los segmento mencionados anteriormente, el cual se define como:

$$D(i,j) = min \begin{cases} D(i-1,j) + costo\ eliminación \\ D(i,j-1) + costo\ inserción \\ D(i-1,j-1) + I_{ij} * costo\ sustitución \end{cases}$$

$$I_{ij} \begin{cases} 0 & cuando\ elementos\ i-ésimo = j-ésimo \\ 1 & en\ caso\ contrario \end{cases}$$

Además, D(i,0) = costo de eliminar i elementos ∀ i. D(0,j) = costo de insertar j elementos ∀ j.

Con esta información es posible escribir un programa computacional recursivo que entregue el valor de la distancia de Levenshtein.

ANEXO B

Campos Condicionales Aleatorios - Extensión

Los campos aleatorios condicionales corresponden a un tipo de clasificador discriminativo, lo que significa que modelan el límite de decisión entre las diferentes clases (Chawla, 2017). Su principio subyacente es que aplican regresiones logísticas a las secuencias de entradas.

Los CRF modelan la dependencia entre cada estado (o etiqueta) $y^i = \{y_1^i, \dots, y_T^i\}$ y el vector de entrada $x^i = \{x_1^i, \dots, x_T^i\}$ utilizando la probabilidad condicional $P(y|x)$. Para esto primero es necesario modelar la distribución condicional de la siguiente manera:

$$\hat{y} = arg\max_{y} P(y|x)$$

Componentes

Como los datos a modelar son secuenciales, es necesario tomar el contexto previo al punto que se quiere predecir. Por esta razón, se utiliza una función característica que posee varios valores de entrada, estos son:

1. Un conjunto de vectores de entrada X
2. La posición t de la data que se quiere predecir
3. La etiqueta y_{t-1} en el punto t-1 de X
4. La etiqueta y_t en el punto t de X

Y se define la función característica como: $f_k(y_t, y_{t-1}, x_t)$

El propósito de la función características es representar las particularidades de la secuencia en un punto específico. Un ejemplo de esto es el reconocimiento de las partes del discurso utilizando CRFs, es decir:

$f_k(y_t, y_{t-1}, x_t) =$ 1 si y_{t-1} es un sustantivo, e y_t es un verbo. En otro caso es 0.

Similarmente, $f_k(y_t, y_{t-1}, x_t) =$ 1 si y_{t-1} es un verbo, e y_t es un adverbio. En otro caso es 0.

Arquitectura

Cada función característica se basa en la etiqueta anterior y la actual, comportándose de forma binaria (Sutton & McCallum, 2011). Para construir el campo condicional se debe asignar a cada función característica un vector de parámetros $\theta=\{\lambda_k\} \in R^K$, que van a ser entrenados con el siguiente algoritmo:

$$P(y|X;\lambda) = \frac{1}{Z(X)} exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\}$$

Para un instante t específico, la función normalización se define como:

$$Z(X) = \sum_{y} exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\}$$

Para calcular los parámetros se utiliza la estimación de máxima verosimilitud. Para aplicar esta técnica primero se aplica el logaritmo negativo a la distribución:

$$l(\theta) = -log \prod_{i=1}^{N} P(y^i|x^i)$$

$$l(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_t^i, y_{t-1}^i, x_t^i) + \sum_{i=1}^{N} log\ Z(x^i)$$

Como se está trabajando con el negativo del logaritmo, se buscará el mínimo (argmin). Para esto se deriva parcialmente respecto a λ_k y se obtiene:

$$\frac{\partial l(\theta)}{\partial \lambda_k} = \frac{-1}{N}\sum_{i=1}^{N} F_k(y^i, x^i) + \sum_{i=1}^{N} p(y^i|x^i)F_k(y^i, x^i)$$

Con:

$$F_k(y^i, x^i) = \sum_{t}^{T} f_k(y_t, y_{t-1}, x_t)$$

Por último, se utiliza el método de Descenso por Gradiente iterativamente para cada derivada parcial hasta que los valores converjan. De esta manera se obtiene la ecuación de actualización del Descenso por Gradiente para un CRF:

$$\hat{\lambda} = \lambda + \alpha[\sum_{i=1}^{N} F_k(y^i, x^i) + \sum_{i=1}^{N} p(y^i|x^i)F_k(y^i, x^i)\ ]$$

Resumiendo, para utilizar CRFs primero se define una función característica. Se definen parámetros aleatorios (lambda) que se calculan a través del método de Descenso por Gradiente iterativamente hasta que los valores convergen. Si bien los CRFs son similares a la Regresión Logística, dado que se basan en la distribución de probabilidad condicional, pero el algoritmo se extiende al aplicar la función característica como datos de entrada secuencial.