UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
ESCUELA DE POSTGRADO Y EDUCACIÓN CONTINUA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ESTRUCTURACIÓN DE MEDICAMENTOS EN COMPRAS PÚBLICAS MEDIANTE EL USO DE PROCESAMIENTO DE LENGUAJE NATURAL

TESIS PARA OPTAR AL GRADO DE MAGISTER EN CIENCIA DE DATOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

CAMILA JACQUELINNE PULGAR FERNÁNDEZ

PROFESOR GUÍA:
MARCELO OlIVARES ACUÑA

PROFESORA CO-GUÍA:
JOCELYN DUNSTAN ESCUDERO

COMISIÓN:
SEBASTIÁN RÍOS PEREZ

SANTIAGO DE CHILE
2023

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGISTER EN CIENCIA DE DATOS,
RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL
POR: CAMILA JACQUELINNE PULGAR FERNÁNDEZ
FECHA: 2023
PROF. GUÍA: MARCELO OLIVARES ACUÑA

ESTRUCTURACIÓN DE MEDICAMENTOS EN COMPRAS PÚBLICAS MEDIANTE EL USO DE PROCESAMIENTO DE LENGUAJE NATURAL

Actualmente, en el mundo del *e-commerce*, el proceso de estructurar productos posee una gran relevancia. Contar con información estructurada representa una ventaja para las organizaciones, generando valor a partir de los estudios realizados. Por ejemplo, identificar los diferentes segmentos de clientes para distintos tipos de artículos. Debido a lo anterior, la tarea de estructurar información es enfrentada continuamente. Sin embargo, a diferencia de la gran cantidad de trabajos que existe en el área del *e-commerce*, en el área de los medicamentos no hay trabajos variados que enfrenten esta problemática.

En Chile existe una alta dispersión de precios de medicamentos en compras públicas, lo que genera repercusiones a la sociedad en general. Es por esto que se establece una solución a este problema, a través del monitoreo de precios, tarea que es realizada por la estructuración de información de medicamentos en compras públicas. Contar con un sistema que permita estructurar descripciones de medicamentos, podría simplificar todo el trabajo que conlleva la extracción manual de la información relevante de los medicamentos.

Debido a lo anterior, en el presente trabajo de tesis se enfrenta la necesidad de estructurar medicamentos, mediante la extracción de los valores de los atributos: forma farmacéutica, principio activo y concentración. Estos últimos se encuentran presentes en las descripciones de los fármacos, las cuales están escritas en texto libre. Por ejemplo, un medicamento cuya descripción es "Levotiroxina 100 mg x 90 cm", posee una forma farmacéutica con un valor de **comprimido**, un principio activo igual a **levotiroxina** y una concentración de **100 mg**.

Con el objetivo de facilitar la estructuración de medicamentos, se propone la creación de un algoritmo de estructuración. Para lograr lo antes mencionado, se utiliza una combinación de herramientas de Procesamiento de Lenguaje Natural (PLN) y de *Machine Learning* (ML), a lo largo de 3 subprocesos que son enfrentados utilizando métodos diferentes. Además, se añaden etapas de supervisión humana, entregando la opción de validar y ayudar al algoritmo a entregar valores correctos, generando a su vez un sistema semi-supervisado, el cual es capaz de estructurar un 75 % de los datos utilizados.

Para lograr lo mencionado anteriormente, se entrega un contexto de la base entregada por CENABAST, además de una revisión del estado del arte actual. Por otro lado, también se establece una serie de objetivos, además de una metodología que abarca las etapas, desde los análisis exploratorios de datos, hasta el establecimiento de métricas que ayudan a evaluar los resultados generados por la aplicación del algoritmo construido.

A mi familia, y a todos los que me permitieron concluir esta etapa de mi vida.
Por su total apoyo y por creer en mí.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

En Chile, existe una gran variación en los precios de medicamentos. Este hecho ha sido demostrado en estudios de los últimos años sobre precios en los fármacos [1]. Según monitoreos realizados por el Servicio Nacional del Consumidor (SERNAC), existen diferencias en los precios de medicamentos de hasta 34 veces, comparando el precio de un medicamento alternativo genérico con uno de marca [2]. Esta extensa dispersión ha sido comprobada por el estudio de la ONG ANADEUS [3], en donde se exponen registros sobre diferencias del gasto anual en medicamentos de uso crónico de hasta un 6533 %. Estas diferencias de precios son importantes en la medida que la población siga adquiriendo medicamentos de marca, puesto que según el SERNAC, existen diferencias de hasta $181 mil entre los precios de medicamentos originales de marca y bioequivalentes [4]. Finalmente, de acuerdo a los datos entregados por la Central de Abastecimiento del Sistema Nacional de Servicios de Salud (CENABAST), durante 2020 se detectaron diferencias porcentuales entre precio retail y precio CENABAST, de hasta un 687 % en los medicamentos más comprados por los adultos mayores [5].

A partir del contexto presentado anteriormente, surge la necesidad de disminuir la variación de los precios en medicamentos, a partir de la supervisión de estos. Una alternativa para monitorear la diferencia de precios entre fármacos, es la asociación de medicamentos con una misma característica, es decir, un mismo principio activo, forma farmacéutica y concentración, lo que permite observar la variabilidad de precios para un mismo medicamento. Lo anterior se logra a través de los datos de medicamentos estructurados, tarea que requiere una gran cantidad de recursos.

En el mundo actual, el uso de datos estructurados se ha transformado en un elemento crucial en la tarea de gestión de información. Poseer datos estructurados permite la realización de análisis profundos en los datos, a partir de los cuales se genera información relevante para cualquier organización. Actualmente, los *retailers* han sido los beneficiarios más destacados. Por ejemplo, se han generado mejores experiencias de compra para el cliente, a partir de filtros que permiten la obtención de productos con atributos similares.

Si bien el mundo del *e-commerce* es uno de los mayores beneficiarios de la investigación sobre estructuración de productos, no es el único campo que se ve favorecido de esta acción. En el área de la salud, y en específico en el campo de la compra de medicamentos, la estructuración de información puede generar grandes beneficios. Por ejemplo, la posibilidad de comparar los precios de productos con atributos similares. Como se dijo anteriormente, el

mercado de los medicamentos posee una alta dispersión de precios, por lo que se genera la necesidad de monitorearlos de forma continua, en especial cuando se trata de un mercado de compras públicas, tarea que es tratada en el presente trabajo. Para poder monitorearlos se establece una tarea de estructuración de medicamentos, a partir de la descripción de estos.

Para tener un buen entendimiento del siguiente trabajo de tesis, es necesario definir algunos conceptos previos que son utilizados de forma constante. El término de **estructuración** se refiere a un formato de organización, administración y almacenamiento de datos que generalmente se elige para un acceso eficiente a los datos [6]. En términos del presente trabajo, estructurar medicamentos hace referencia a organizar el texto de una descripción de medicamento en columnas fijas, establecidas con anterioridad a partir de las propiedades/atributos de los medicamentos. Por lo tanto, en el presente trabajo se lleva a cabo la estructuración de textos de medicamentos en 3 columnas: forma farmacéutica, principio activo y concentración. Esta organización permitiría posteriormente monitorear la variación de precios de los medicamentos, a partir de la agrupación de productos con una misma característica. Para la realización de esta tarea, se utilizan datos de medicamentos en el mercado de compras públicas, los cuales son entregados por la Central de Abastecimiento del Sistema Nacional de Servicios de Salud (CENABAST).

En Chile, las compras públicas en el sector de la salud presentan una tendencia creciente. Desde el año 2015, el sector de la salud en Chile representa más del 50 % del total que es comprado por el gobierno [7]. Además, el mercado de la salud está considerado dentro de los más importantes del mercado público. Este mercado posee el rol de no solo proveer medicamentos a la población, sino que provee también a hospitales, consultorios, servicios de salud y CENABAST, lo que corresponde al 89 % de gasto público en salud [8].

La Central de Abastecimiento del Sistema Nacional de Servicios de Salud, de aquí en adelante CENABAST, es una institución pública dependiente del Ministerio de Salud (Minsal). Se encarga de gestionar los procesos de compra mandatados por este, y por todas aquellas instituciones que se adscriban al Sistema Nacional de Servicios de Salud (SNSS), para el ejercicio de acciones de salud [9]. Desde el año 2012, CENABAST actúa como un intermediario en el proceso de compras públicas, además de decidir la canasta de productos que se pueden intermediar, acción en la que participan también representantes del Minsal, el Fondo Nacional de Salud (FONASA) y directores de los Servicios de Salud.

Los establecimientos de salud tienen, por su parte, la tarea de escoger un "arsenal farmacológico", el cual es un registro en donde se decide qué medicamentos y productos se dispondrán para ser utilizados por estos [10]. El arsenal debe ser compuesto por productos que están incluidos en el formulario único nacional, el cual corresponde a un listado de drogas científicamente escogidas, identificadas por su nombre genérico, y las formas farmacéuticas que correspondan para ser utilizadas con fines preventivos, de diagnóstico y terapéuticos. CENABAST, en conjunto con representantes del área, deben encargarse de crear este listado.

La Red Pública de Salud, entendida como todas las entidades que se encuentran adscritas al Sistema Nacional de Servicios de Salud (SNSS), tienen el deber de realizar adquisiciones de fármacos e insumos médicos. Esta adquisición puede realizarse utilizando 3 métodos: (1) Licitación Pública, (2) Convenio Marco o (3) Trato Directo. Por otro lado, estas entidades

también pueden elegir la presencia de un organismo de intervención, correspondiente a CENABAST [10].

La decisión de compra de insumos médicos, en general, es una tarea de suma importancia, tanto para el Estado como para la población. La buena y correcta compra de estos podría significar el ahorro de millones de pesos en compras públicas. Además, debido a la ley publicada durante el año 2020, la Ley Nº21.198 [11], conocida como "Ley CENABAST", se le entregó a esta institución la facultad de intermediar la compra de medicamentos e insumos con farmacias del tipo independientes. Esto implica que estas entidades compren y vendan a la población a un precio económico, debido al establecimiento de un precio máximo de venta.

Dado que muchas de las instituciones públicas deciden utilizar la intermediación entregada por CENABAST para la compra de medicamentos, el correcto manejo de las licitaciones podría implicar una significativa disminución de gastos públicos [12].

Actualmente, en el mercado público de medicamentos, existe una dificultad en la tarea de comparar las características entre los mismos productos. Es decir, no existe una clasificación de medicamentos que permitan hacer análisis econométricos que busquen comparar mercados iguales, o que permita capturar los efectos fijos entregados por cada medicamento. Esto se debe, a que no se tiene a la fecha, una forma de comparar si un medicamento es igual a otro, lo cual impide a su vez el estudio de productos similares.

El presente trabajo de tesis se desarrolla dentro de un proyecto FONDEF con CENABAST, donde el objetivo es añadir valor a los datos entregados por CENABAST, a través de los análisis futuros que puedan ser posibles gracias a la estructuración de texto libre. Esta estructuración se lleva a cabo mediante la utilización de Procesamiento de Lenguaje Natural, en adelante PLN y herramientas de Aprendizaje de máquinas, en adelante AM. Con el fin de lograr una buena estructuración de medicamentos, se establecen una serie de objetivos que buscan una implementación eficiente.

Se tiene por objetivo general el **desarrollar un algoritmo de estructuración a través de herramientas de PLN y AM, para lograr de forma eficiente, que al menos un 70 % de los medicamentos de la base de datos Vista Medicamentos, entregada por CENABAST, logre ser estructurado de forma correcta.**

Para lograr lo anterior, se establecen los siguientes objetivos específicos:

1. Realizar un análisis exploratorio de datos a la base de datos Vista Medicamentos.
2. Crear diccionarios limpios de fallas escriturales de cada uno de los atributos de los medicamentos.
3. Establecer métricas que permitan observar la mejora de los algoritmos.
4. Utilizar herramientas de Procesamiento de Lenguaje Natural (*Word2Vec*), y herramientas de Aprendizaje de máquinas (*Support Vector Machine*), para lograr una estructuración correcta de los medicamentos.

Se busca el diseño y entrenamiento de un algoritmo que logre automatizar la extracción de

valores de atributos en base a descripciones de medicamentos no estructuradas. Estas descripciones que se encuentran como texto libre, se denominan **descripción del medicamento**. Un ejemplo de una descripción del medicamento es "Levofloxacino 500 mg Comprimido". En lo que sigue del trabajo, cada vez que se quiera mencionar el proceso de estructuración de medicamentos, se está haciendo referencia al proceso de estructurar un texto libre presente en las descripciones de los medicamentos, en 3 columnas de atributos fijas.

El trabajo de estructuración automática no es una tarea fácilmente automatizable, puesto que se presentan una serie de dificultades que surgen al momento de realizar esta tarea:

- **La extracción de atributos y valores de atributo de un medicamento**: El algoritmo creado para lograr la estructuración, debe tener la capacidad de detectar tanto los atributos como los valores de atributo de un medicamento, una tarea que aumenta la dificultad del problema. Por ejemplo, se tiene un medicamento con la descripción "Levofloxacino 500 mg Comprimido" y el algoritmo de estructuración debe ser capaz de identificar que "Levofloxacino" corresponde al atributo **principio activo** y que "500 mg" corresponde a **concentración**.
- **Los valores de un atributo pueden tener diversas formas de escritura**: Dada la característica de texto libre que tienen las descripciones de los medicamentos, un valor de atributo puede ser escrito de diversas formas. Por ejemplo, para el caso del atributo **concentración**, un valor de atributo puede ser "100mg", pero este mismo valor puede ser escrito como "100 miligramos". Además, se deben considerar los errores ortográficos que puedan existir en las descripciones, dado que estos fueron escritos inicialmente por una persona, por lo que un valor de atributo puede tener varios valores asociados, pero que representan lo mismo. Por ejemplo, para el atributo **forma farmacéutica**, un valor de atributo corresponde a "comprimido", pero este puede estar escrito como "comprido". Es por esto que la solución al problema debe tener la capacidad de detectar las diversas formas de escribir un valor de atributo y homologarla a un mismo valor.

En el siguiente trabajo de investigación se presenta un marco teórico como contextualización de las compras públicas en Chile, además del estado del arte en el área del PLN para la clasificación de texto. También se presentan los objetivos, hipótesis, descripción de herramientas y datos que son utilizados a lo largo de la construcción del algoritmo de estructuración, así como también las fases metodológicas que permiten desarrollarlo. Por último, se presentan los resultados, discusiones y conclusiones del trabajo realizado.

Capítulo 2

Marco Teórico

En el siguiente capítulo se presenta una compilación de antecedentes necesarios para el correcto entendimiento del trabajo de tesis. Primeramente, se presenta el marco conceptual, describiendo de forma ligera los conceptos asociados al algoritmo de estructuración. De forma posterior, se documentan antecedentes académicos relacionados a trabajos similares al realizado, indagando por las principales áreas de aplicación. Finalmente, con el fin de entregar un contexto más profundo del desarrollo del trabajo, se entrega una recopilación de información relacionada con el mundo de las compras públicas en Chile, con un enfoque final en las compras públicas en medicamentos, área en la cual se desarrolla el proyecto.

2.1. Marco Conceptual

Para lograr comprender de mejor forma el trabajo realizado, se describen a continuación todos aquellos algoritmos y herramientas utilizadas a lo largo de la construcción del sistema de estructuración.

Todas las herramientas que son mencionadas a continuación, son recurrentemente utilizadas por los campos de Aprendizaje de máquinas y Procesamiento de Lenguaje Natural. El aprendizaje de máquinas, aprendizaje automático o *Machine Learning*, tiene un enfoque en el uso de datos y algoritmos, los que buscan imitar la forma en que los humanos aprenden, a la vez que mejoran gradualmente su precisión [13]. Por otro lado, el Procesamiento de Lenguaje Natural busca combinar la lingüística computacional (modelos basados en reglas del lenguaje humano) con modelos estadísticos, de aprendizaje profundo y aprendizaje automático. Además, tiene por principal objetivo el entregar a las máquinas la capacidad que tiene el ser humano de entender textos y palabras habladas [14].

2.1.1. Expresiones regulares

Usualmente conocidas como **regex** (por la contracción de *regular expression*), las expresiones regulares corresponden a un lenguaje utilizado para la manipulación y análisis de texto. Son muy utilizadas para realizar complejos trabajos de búsquedas y reemplazos en un texto determinado, además de ayudar en la validación de que los datos de texto estén bien construidos [15].

Una expresión regular corresponde teóricamente a una cadena que está conformada por

una combinación de caracteres normales y metacarácteres especiales (o metasecuencias). Los primeros tienen la característica de coincidir con ellos mismos, mientras que los metacarácteres son caracteres o secuencias de estos, que ayudan a representar ideas, tales como, la cantidad, ubicaciones o tipos de caracteres. Por ejemplo, [^**m-z**], el cual corresponde a un rango negativo de caracteres, entrega todos aquellos caracteres que no coinciden con los caracteres que están en el rango especificado, es decir, que no pertenezcan al rango de m a z [16].

La coincidencia de patrones consiste en encontrar una porción de un texto que coincide o describe una expresión regular. Se puede predecir una gran cantidad de coincidencias, teniendo en cuenta a su vez que:

1. **La coincidencia más temprana a la izquierda gana:** Las expresiones regulares se aplican comenzando por el primer carácter encontrado a la izquierda, y avanza hacia la derecha. Cada vez que el motor de expresiones regulares (correspondiente al código que busca el texto) encuentra una concordancia, este vuelve al inicio.

2. **Los cuantificadores estándar son voraces:** Los cuantificadores indican la cantidad de veces que se puede repetir algo, y los cuantificadores estándar buscan coincidir con los patrones tantas veces como sea posible. En ocasiones, se pueden llegar a conformar con menos del máximo de coincidencia, solo si es necesario para lograr el éxito.

2.1.2. *Word2Vec*

Corresponde a un algoritmo creado como herramienta para PLN. Entrega una implementación eficiente de las arquitecturas de *bag-of-words* y *skip-gram*, para calcular representaciones vectoriales de palabras, las cuales pueden ser utilizadas posteriormente para muchas aplicaciones de PLN [17].

Esta herramienta utiliza una red neuronal para aprender las relaciones entre las palabras, las cuales provienen de un corpus de texto que es utilizado como entrada, y genera los vectores de palabras como salida. Construye un vocabulario a partir de los datos del texto de entrada, el cual utiliza a modo de entrenamiento, y luego aprende la representación vectorial de las palabras. Estas pueden ser utilizadas posteriormente como una matriz de características para muchas aplicaciones de PLN o aprendizaje automático [18].

Word2Vec es utilizado generalmente como un *word embedding*, el cual incrusta cada palabra en un espacio continuo de baja dimensión, con el fin de representar palabras con vectores numéricos, los cuales transmiten información semántica y sintáctica (género, sinónimos, etc.) de las palabras [19].

2.1.3. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) o "máquinas de soporte vectorial" corresponde a un algoritmo de clasificación del tipo supervisado. Se basa en los principios de la teoría del aprendizaje estadístico y la optimización convexa. Su objetivo es separar clases de puntos, a través del establecimiento de un límite apropiado en el espacio de datos [20].

SVM busca de forma eficiente un buen hiperplano de separación, es decir, un hiperplano que tenga una buena medida de rendimiento de generalización, dentro de un espacio de

características de alta dimensión. Este hiperplano de separación se basa en el concepto de **margen**, el cual corresponde a la mínima distancia entre el hiperplano establecido y los puntos de datos más cercanos a este. Por lo tanto, se define el **hiperplano óptimo** como el máximo margen de separación entre dos clases [20].

2.1.4. *Naïve Bayes*

El modelo *Naïve Bayes* corresponde a un algoritmo de clasificación del tipo supervisado. Es una forma simple de clasificadores bayesianos, basado en el teorema de Bayes, los cuales asignan la clase más probable a una instancia determinada descrita por un conjunto de atributos. El nombre de *Naïve Bayes* se debe al supuesto que asume que todas las características son independientes entre sí [21].

El algoritmo de *Naïve Bayes* [22] utilizado en el presente trabajo, corresponde a:

$$P\left(C_i \mid \wedge v_j\right) = \frac{P\left(\wedge v_j \mid C_i\right) P\left(C_i\right)}{P\left(\wedge v_j\right)} \quad i = 1 \ldots n \tag{2.1}$$

Donde $P\left(C_i \mid \wedge v_j\right)$ corresponde a la probabilidad de que un ejemplo específico pertenezca a la clase C_i a partir de los valores de características dados, v_j. ($\wedge v_j$ denota la conjunción de todos los valores de característica en el ejemplo.)

El objetivo del clasificador *Naïve Bayes* es determinar la clase C_i con la mayor probabilidad condicional $P\left(C_i \mid \wedge v_j\right)$. Dado que el denominador $P\left(\wedge v_j\right)$ de la expresión anterior es constante para todas las clases C_i, el problema se reduce a encontrar la clase C_i con el valor máximo para el numerador. Por lo tanto, el clasificador *Naïve Bayes* asume la independencia de las características del ejemplo, de modo que:

$$P\left(\wedge v_j \mid C_i\right) = \prod_j P\left(v_j \mid C_i\right) \tag{2.2}$$

2.1.5. *Logistic regression*

Logistic regression, regresión logística o modelo logit, es un algoritmo que busca analizar la relación existente entre una variable dependiente categórica, y múltiples variables independientes. Busca estimar la probabilidad de que un evento ocurra, ajustando los datos a una curva logística. Cuando la variable dependiente no es binaria, es decir, es una variable que está compuesta por más de dos categorías, se puede emplear una regresión logística multinomial [23].

2.1.6. *Random forests*

Random forests es un conjunto de árboles de decisión independientes construidos aleatoriamente. Funciona sustancialmente mejor que los clasificadores de un solo árbol. Se utiliza un subconjunto aleatorio de atributos para la división de nodos, mientras crece cada árbol de decisión. Para cada árbol, normalmente se extrae un conjunto de arranque (con reemplazo) de los datos de entrenamiento originales, es decir, se selecciona una instancia de los datos de entrenamiento y se reemplaza nuevamente antes de dibujar la siguiente instancia [24].

2.1.7. Distancia de Levenshtein

Corresponde a una medida de distancia propuesta por Vladimir Levenshtein, denominada también como "distancia de edición". Entrega una medida de similitud entre dos cadenas de texto, a través del establecimiento del número de eliminaciones o inserciones de caracteres dentro de una cadena de caracteres X, para transformarla en otra cadena Y [25]. Mientras más pequeña sea la distancia de Levenshtein, más similitud hay entre dos cadenas de texto.

Por ejemplo, la distancia de Levenshtein entre las cadenas de texto "sodio cloruro" y "cloruro de sodio" es 2, puesto que solo se realizan 2 inserciones de caracteres (de), esto implica que la cadena es bastante similar mientras más cercana a 0 sea la distancia de Levenshtein. Por otro lado, la distancia entre las palabras "polivitaminico" y "cotrimoxazol" es de 11, lo que implica que ambas palabras son muy diferentes.

2.1.8. TF-IDF

Corresponde al acrónimo de *Term Frequency - Inverse Document Frequency*, o frecuencia de término – frecuencia inversa de documento. Mide la importancia de una palabra para un documento, dentro de una colección de documentos. Es utilizada en PLN para convertir palabras en vectores, incluyendo información semántica y ponderado a palabras poco comunes.

Utiliza dos métricas:

1. Frecuencia de término (TF): Corresponde a la cantidad de veces que aparece una palabra en un documento.

   Se calcula como:

$$TF = \frac{\text{Número de repeticiones de una palabra en un documento}}{\text{Cantidad de palabras en un documento}} \tag{2.3}$$

2. Frecuencia inversa de documento (IDF): Corresponde a una colección de documentos. Las palabras poco usuales tienen puntajes altos, mientras que las palabras más comunes poseen más altos.

   Se calcula como:

$$IDF = \log[\frac{\text{Número de documentos}}{\text{Número de documentos que contienen a la palabra}}] \tag{2.4}$$

Finalmente, para obtener el valor final de TD-IDF, se utiliza:

$$TD - IDF = TD \times IDF \tag{2.5}$$

2.2. Estado del Arte

Existe una gran variedad de literatura sobre la tarea de estructuración de productos, a partir de la extracción de atributos de los títulos o descripciones de estos. Además, ha sido explorada en una gran variedad de áreas, y enfrentada utilizando diversas técnicas. Dentro de

estas, la utilización de herramientas de PLN y algoritmos de *machine learning* toma especial relevancia.

El mundo de los *retailers* en general, es una de las áreas que más utiliza la extracción de atributos contenidos en títulos o descripciones de productos. Esto dado principalmente por la gran cantidad disponible de datos de ventas y transacciones. Es por ello que todos los trabajos presentados a continuación, son investigaciones realizadas en esta área particular. La mayoría de los *retailers* administran sus productos como bloques de entidades con específicos atributos que logren identificarlos (generalmente se presenta la marca, tamaño, color, etc). Esta utilización de los productos como un bloque de entidades dificulta, en ocasiones, la eficiencia de las aplicaciones que utilizan las grandes empresas para realizar estudios o predicciones. Además, también genera posibles dificultades en la construcción de filtros o aplicaciones de recomendaciones de productos similares al público.

A pesar de que la extracción de atributos de texto libre es una tarea que puede ser implementada en cualquier área, hasta ahora el área que más la utiliza y donde se encuentran concentradas las investigaciones, es el *e-Commerce*. Ghani R. et al. [26] plantean el problema de extracción de pares atributo-valor como una doble tarea. Buscan tratar con dos tipos de atributos de forma separada, atributos tanto implícitos como explícitos, formulando dos problemas de clasificación. Los atributos implícitos hacen referencia a la semántica de las descripciones de los productos, mientras que los explícitos son los que se pueden reconocer a simple vista. Por ejemplo, para un producto del tipo vestuario, un atributo explícito sería el color, mientras que el implícito sería la tendencia que tiene. Los problemas de clasificación fueron específicos para cada tipo de atributo. En el caso de los implícitos, se etiquetaron aproximadamente 600 datos, los cuales fueron utilizados con un algoritmo de *Naïve Bayes*, mientras que los datos restantes fueron utilizados en un algoritmo de *Expectation-Maximization* para estimar los parámetros máximos a posteriori de un modelo generativo. Por otro lado, el tratamiento de los atributos explícitos se realiza mediante el algoritmo de *Naïve Bayes* y un algoritmo semisupervisado de vista múltiple (co-EM).

Otros enfoques para resolver el problema, los presenta Q. Wang et al. [27], quienes proponen el uso de *question answering (QA)*. En este estudio se construye un modelo de *QA* en donde se trata a cada atributo como una pregunta, identificando a su vez el posible intervalo de respuestas en donde se encuentre el correspondiente valor del atributo. Utilizando un modelo BERT (*Bidirectional Encoder Representations from Transformers* o Representación de Codificador Bidireccional de Transformadores) para todos los atributos, codifica el contexto y pregunta, implementando así un modelo escalable. Para aquellos casos en los que no se detecten valores para un atributo específico, se establece un algoritmo de clasificación binario, para predecir si hay una posible respuesta o no, en el contexto de la pregunta.

También se pueden observar otras alternativas pertenecientes al área de aprendizaje profundo. Autores como X. Ma y E. H. Hovy [28] proponen el uso de redes neuronales, entregando un enfoque de etiquetado abierto, basado principalmente en la comprensión de los atributos, para la extracción de los valores de atributos. Proponen una arquitectura compuesta por elementos de BiLSTM (*Bidirectional LSTM*) y de CRF (*Conditional random fields*), en donde modelan al atributo de forma semántica, capturando la interacción semántica con el título y generando la representación de este con CRF. Por otro lado, a diferencia del trabajo presenta-

do anteriormente, Huang et al. [29], abordan la problemática utilizando una arquitectura de red con la mezcla BiLSTM-CRF, sin embargo, consideran al atributo como una sola etiqueta.

Finalmente, respecto a los enfoques más recientes que se han implementado para resolver el problema en el área del *e-Commerce*, Karamanolakis G., Ma J. y Dong X. [30] buscan modelar la tarea de extracción de atributos como etiquetado de secuencias y lo resuelven a través de la utilización de modelos de aprendizaje profundo, tales como *Bidirectional LSTM* (BiLSTM), mejorando los resultados con la aplicación de *Conditional Random Fields* (CRF).

En el área de la salud se han implementado algunas herramientas de NLP con el fin de extraer atributos de textos clínicos, sin embargo, no hay hasta la fecha un trabajo que se dedique totalmente a la tarea de extraer atributos y sus valores, desde la descripción en texto libre de medicamentos. Si bien la tarea de estructuración que se realiza en este trabajo es un tema a tratar de forma constante en el mundo del *e-Commerce*, representado por la variedad de trabajos de investigación que existen, en el área de la salud no es un trabajo tan recurrente.

Mandhan S. et al. [31], presentan en su trabajo la extracción de atributos y valores numéricos de datos clínicos. Trabajo en el cual, para extraer atributos numéricos y sus valores desde los registros, realizan dos pasos principales. En una primera etapa se presenta la tarea de extracción de atributos y valores numéricos mediante el desarrollo de un modelo de reconocimiento de entidades nombradas (*Named Entity Recognition*, abreviado como NER), utilizando también librerías de *Stanford NLP*. En una segunda etapa, se busca asociar de forma correcta los atributos extraídos con sus valores. Esto logrado por medio de la aplicación de un módulo de extracción de relaciones en un *framework* de *Apache cTAKES*. Finalmente, se integra el modelo NER como componente del *framework* de *cTAKES*, para extraer finalmente las relaciones.

Du, M. et al. [32], por otro lado, trabajan con la historia clínica electrónica, la cual almacena la información de datos sociales, médicos y preventivos de un paciente en formato digital, lo que permite el uso de estos para la enseñanza e investigación científica. Este trabajo se enfoca en la extracción de atributos a partir de textos no estructurados de historia clínica electrónica. Para esto, se propone el uso de un modelo de red neuronal *end-to-end*, con el fin de extraer diferentes valores de atributo del texto no estructurado. Cada frase del texto es considerada una instancia, sobre la cual se utiliza en un inicio un *word embedding* (incrustación de palabras) preentrenado, logrando iniciar de mejor forma los modelos de redes neuronales. Posteriormente, se realizan pequeños ajustes en el texto para lograr una mejor obtención de la semántica de las palabras.

A modo de conclusión, a pesar de que existe una gran variedad de trabajos que utilizan herramientas de PLN y de *machine learning* para la estructuración de textos, no existen registros de trabajos similares en el área de la salud, en donde se estructuren textos libres de medicamentos. Es por lo anterior, que este presente trabajo adquiere valor, dado que permite la realización de estudios que requieren la estructuración de fármacos.

2.3. Antecedentes

En las presentes secciones, se entrega una descripción del contexto en el cual se desenvuelve el trabajo de tesis, detallando cómo funcionan las compras públicas en Chile, mencionando mecanismos y su proceso. Además, también se describe el área específica sobre la cual se trabaja, correspondiente a las compras públicas en el sector de la salud.

2.3.1. Compras públicas en Chile

El sistema de compras públicas presente en Chile, corresponde a un conjunto de instituciones, prácticas, normas y organismos públicos, cubiertos por la Ley N°19.886. La ley de Compras, entrada en vigencia en 2003, tiene por propósito fundamental el "Contribuir a la eficiencia y transparencia en la gestión pública en la adquisición de bienes y servicios, mediante la apertura del mercado de las compras públicas y la mejora de las capacidades de compra de los funcionarios encargados de las mismas". Consecuencia de lo anterior, se crea la Dirección de Compras y Contratación Pública (DCCP), Dirección ChileCompra en lo que sigue. Esta tiene la misión de "generar eficiencia en la contratación del Estado con altos estándares de probidad y transparencia, poniendo a disposición (...) la plataforma transaccional www.mercadopublico.cl, donde los organismos del Estado compran y los proveedores venden sus bienes y servicios". Además, también tiene como fin el asesoramiento y trabajo constante con los organismos, con el objetivo de realizar un uso eficiente de los recursos públicos. Finalmente, se debe mencionar que el compromiso de Dirección ChileCompra es "maximizar la eficiencia en las compras del Estado, entregando un servicio simple, resolutivo y confiable" [33].

Respecto a montos transados, estos han ido en aumento, salvo entre los periodos 2015-2016 en donde hubo una cifra constante de $6,8 billones, generados por 95.000 proveedores que transaron al menos una oferta durante los últimos 12 meses. Desde el año 2003, la incorporación de los organismos públicos ha ido en un aumento gradual, puesto que aumentó desde 350 instituciones a más de 650 durante el año 2016, incluyendo instituciones como Ministerios, intendencias, municipalidades, FF.AA., hospitales públicos, universidades estatales, entre otras [34].

Según el artículo 30 de la Ley N°19.886, las funciones que posee Dirección ChileCompra corresponden a "promover la máxima competencia posible en los actos de contratación de la Administración, desarrollando iniciativas para incorporar la mayor cantidad de oferentes". Además de lo anterior, también tiene el deber de organizar el mercado público y efectuar procesos de contratación pública para su correcto funcionamiento, operando a su vez, un portal de comercio electrónico que faculta a los proveedores a ingresar ofertas a un catálogo electrónico, a partir del cual los funcionarios públicos podrán tomar decisiones de compra [35].

La contratación pública corresponde al "proceso de identificación de necesidades, la decisión acerca de la persona, física o jurídica, más adecuada para cubrir estas necesidades y, por último, la comprobación de que el bien o prestación se entregan en el lugar correcto, en el momento oportuno, al mejor precio posible, y que todo ellos se hace con ecuanimidad y transparencia". Por otro lado, el objetivo fundamental de la contratación pública es la "entrega de bienes y la prestación de servicios necesarios para el cumplimiento de las funciones de la autoridad pública de una manera puntual, económica y eficiente" [36].

Los mecanismos de contratación pública de carácter eficaz, eficiente y enfocado en mantener un rendimiento económico, permiten al Estado suministrar servicios y bienes con una cobertura adecuada. Por lo tanto, corresponden a un pilar importante en el proceso de prestaciones de servicios, puesto que "una contratación pública bien gestionada (...) debe desempeñar un papel de primer orden en el fomento de la eficiencia del sector público" [35]. Además, la contratación pública representa un aspecto importante en la actividad económica del país, en especial si se hace un enfoque en la administración, puesto que la contratación pública permite asegurar de alguna forma el contar con los insumos necesarios que facultan el cumplimiento de los objetivos, además de contribuir en la gestión y distribución del dinero que es aportado por contribuyentes. Sumado a lo anterior, el proceso de compras públicas tiene por fin el satisfacer los requerimientos del Estado, tanto en bienes como en servicios, que permitan cubrir las necesidades públicas, cumpliendo al mismo tiempo el mandato de preservar el "bien común" [36].

En Chile, el proceso de compras públicas se basa en los principios de igualdad, transparencia, integridad y libre concurrencia de oferentes. Es denominado también un "procedimiento administrativo", encargado de regular un acto jurídico con carácter público, fijando al mismo tiempo las obligaciones y derechos entre aquellos que participen del procedimiento. El Estado, por su parte, tiene la obligación de adquirir bienes y servicios para incentivar el correcto desarrollo de sus funciones. Es por eso que se emplean una serie de mecanismos de compra, que permiten desarrollar un buen proceso de adquisiciones de bienes y servicios [35].

2.3.1.1. Mecanismos de compra

Según la Ley Nº19.886 del año 2003, correspondiente a la Ley de Compras públicas, todas las entidades que pertenecen a la Red Pública de Salud deben realizar la contratación de productos, servicios u otros, a través de los procedimientos de contratación pública establecidos en los artículos Nº7, Nº8 y Nº30 de la Ley Nº19.886. Estos se detallan de forma breve a continuación.

1. Convenio Marco:

   Corresponde a una modalidad de compra de bienes y servicios que ayuda a facilitar la ejecución de compras públicas, mediante un catálogo electrónico administrado por la Dirección ChileCompra. Establece los precios y condiciones de compra, además de disponer de una variedad de proveedores, los cuales son seleccionados a través de un concurso público [37].

   Respecto al proceso de compra a los proveedores mediante Convenio Marco, cada entidad pública debe consultar si el servicio o producto a adquirir, se encuentra disponible en la tienda, antes de llamar a una licitación pública o efectuar algún trato directo [37].

2. Licitación Pública:

   Corresponde a la segunda opción de compra pública, siendo solo en defecto de la primera opción. Posee un procedimiento administrativo de carácter concursal, en el cual la Subsecretaría realiza un llamado público, con el fin de convocar interesados para que formulen propuestas, de entre las cuales seleccionará y aceptará la más conveniente. Todo esto sujetándose a bases de licitación fijadas [38].

La licitación es adquirida por la persona o empresa que haya ofrecido las condiciones más convenientes, según los criterios de evaluación descritos en las bases. Además, por ley, los organismos están obligados a realizar licitaciones públicas por contrataciones que superen las 1.000 UTM [38].

3. Trato Directo:

   Corresponde a un mecanismo de compra excepcional, en donde las compras son menores a las 100 UTM y que implica la contratación de un solo proveedor, por lo que no requiere de un concurso público [39].

   Según el artículo N°10 de la Ley N°19.886, entre las diversas razones por las cuales se puede utilizar este mecanismo, se encuentra el escenario en el cual no se presentan interesados en una licitación pública. Además, pueden existir situaciones en las que solo exista un único proveedor del bien o servicio. También existen casos de emergencias presentadas con su fundamento correspondiente, entre otros escenarios [39].

En la Tabla 2.1 se pueden observar un resumen de los tipos de mecanismos de compra y sus características más importantes.

Tabla 2.1: Tipo de contratos y sus características. Fuente: Malgarini I. [8]

Tipo de Mecanismo	**Utilidad**	**Carácter**	**Participación**
Licitación pública	Norma general	Concursal	Cualquier persona
Trato directo	Por la naturaleza del negocio	Excepcional	Trato directo
Convenio marco	Adquisiciones recurrentes o compras estándares	Concursal	Cualquier persona

2.3.1.2. Proceso de compra

La normativa de compras públicas es un factor de importancia a la hora de realizar las compras públicas, puesto que indica al comprador la prioridad de elección del mecanismo por el cual adquiera un bien o servicio.

Según el artículo N°8 de la Ley N°19.886, las entidades que quieran adquirir bienes o servicios deben utilizar el mecanismo de Convenio Marco, licitados y adjudicados por la Dirección ChileCompra, sin importancia en el monto de las contrataciones. Además, los que serán publicados en un catálogo de Convenios Marco [40].

En el caso de que la contratación por Convenio Marco no proceda, por regla general, las entidades deberán establecer las adquisiciones de suministros y/o servicios a través de la licitación pública. Este escenario se presenta cuando la compra de bienes y/o servicios es superior a 1000UTM, o cuando el producto no se encuentra o no cumple con las condiciones mínimas requeridas.

Finalmente, en aquellos casos en los cuales no haya interesados en las licitaciones públicas, en casos en que exista solo un proveedor, o si el servicio es de naturaleza confidencial cuya difusión pudiese afectar la seguridad o el interés nacional, entre otras, las entidades podrán realizar las contrataciones mediante Trato Directo.

Figura 2.1: Prioridad de los mecanismos de compra. Fuente: Malgarini I. [8]

En la Figura 2.1 se observa un resumen de la prioridad establecida de los mecanismos de compra. Los cuales, buscan el cumplimiento de tres objetivos, que apuntan principalmente a lograr un mejor abastecimiento del sistema público:

1. Buscar la generación de un ahorro a través de procesos de compra que generen precios competitivos, lo que permita generar una reducción en el gasto público.

2. Buscar la minimización de costos de transacción asociados a las compras, a través de los mecanismos de compra, con el fin de generar un proceso ágil y de fácil utilización por parte de los compradores.

3. Establecer mecanismos de compra mucho más transparentes y transables, que permitan realizar un seguimiento posterior a cualquier compra de bienes y/o servicios realizados por las entidades públicas.

2.3.2. Compras públicas de medicamentos en Chile

2.3.2.1. Marco Regulatorio

Las leyes que regulan la compra pública en Chile, son aplicables también a la regulación de la compra y venta de medicamentos. El marco regulatorio del sector de los fármacos en Chile, y las principales características de los compradores fundamentales, incentivan a que la compra de estos productos tenga una mayor relevancia dentro del proceso de abastecimiento público.

A pesar de que actualmente hay una ley que regula de una manera general las compras públicas en Chile, existe una necesidad que de que haya regulaciones reglamentarias o administrativas, lo que permite una mayor flexibilidad a la hora de necesitar la introducción de cambios. El modelo normativo por el que se optó aplicar en Chile, consiste en regular todas aquellas materias relacionadas, dando una mayor preferencia a la aplicación de normas de segundo nivel jerárquico, es decir, el uso de normas reglamentarias, condiciones directivas y de uso.

En la Tabla 2.2 se presenta la estructura de las principales normativas en materia de contratación pública.

Tabla 2.2: Normativas en contratación pública. Fuente: Elaboración propia.

Nombre	Tipo de Norma
Ley Nº 19.886 de Bases sobre contratos administrativos de suministro y prestación de servicios.	Primer orden
Reglamento de la Ley Nº19.886	Segundo orden
Condiciones de uso del Sistema de Información	Tercer orden
Directivas de Compras Públicas	Tercer orden
Manual de Procedimientos de Adquisiciones	Tercer orden

Alineado con los valores públicos, el principal objetivo de esta legislación es la cooperación en la transparencia y eficiencia de la gestión pública, en el proceso de adquisición de bienes y servicios, además de tener un incentivo por mejorar la facultad de compra de los organismos públicos.

Posterior a la promulgación de la Ley de Compras, Ley Nº19.886, y el establecimiento de su respectivo reglamento, se instauró un sistema electrónico de información de contratación pública. Este tiene la característica de ser "de acceso público y gratuito" [41], además de estar a cargo de la Dirección ChileCompra. En el artículo Nº54 de la Ley antes mencionada, se decreta que las instituciones públicas tienen el deber de desarrollar todos sus procesos de compra utilizando únicamente el Sistema de Información de la Dirección. Procesos que incluyen todos aquellos documentos, actos y resoluciones relacionados de forma directa o indirecta con los procesos de compras públicas. Por otro lado, el reglamento también estipula que los organismos no podrán adjudicar contratos a aquellas ofertas que no fueron recibidas a través del Sistema de Información [41].

En lo que respecta al ámbito de la aplicación, la Ley de Compras, establece que esta será aplicada para todos "los contratos que celebre la Administración del Estado, a título oneroso, para el suministro de bienes muebles, y de los servicios que se requieran para el desarrollo de sus funciones" [41]. Por lo tanto, en el ámbito regulatorio, para la aplicabilidad de la ley de Compras se necesita el cumplimiento obligatorio de cuatro requisitos [41]:

1. Que la contratación sea celebrada por la Administración del Estado.
2. Que la contratación sea a título oneroso.
3. Que el objeto de la contratación sea el suministro de bienes muebles y servicios.
4. Que estos bienes y servicios sean necesarios para el desarrollo de la función pública.

2.3.2.2. Procedimiento de compra

En primer lugar, es necesario destacar que las necesidades de compras y decisiones sobre qué producto se comprará, es realizada de forma individual por cada uno de los establecimientos de salud. Lo anterior recibe el nombre de **arsenal farmacológico**, y corresponde "a un listado de medicamentos considerado indispensable para atender las patologías más frecuentes y prioritarias de la población beneficiaria" [42].

Cada hospital tiene la flexibilidad de establecer su propio arsenal farmacológico, eligiendo los medicamentos sobre el listado ya establecido previamente en el Formulario Nacional de Medicamentos (FN). Este, según la autoridad sanitaria, y basada en el poder legal que se le

otorga en el artículo N°100 del Código Sanitario, se define como la lista de todos aquellos productos que son indispensables para que el país tenga una correcta terapéutica, procurando además, las medidas necesarias para que la población, y los servicios que entregan atención de salud, se encuentren correctamente abastecidos de estos. [43] Dentro de cada establecimiento de salud, la tarea de establecer y velar por la correcta aplicación terapéutica del arsenal farmacológico, es realizada por el Comité de Farmacia y Terapéutica, la cual es la unidad técnica dentro de un hospital que define aquellos requerimientos locales de cada establecimiento.

Los establecimientos de salud, además de tener la tarea de decidir el arsenal farmacológico, también pueden escoger el mecanismo de compra para abastecerlo, cumpliendo a su vez con el Marco Regulatorio de compras públicas. Los mecanismos de compra directa más destacados, son las licitaciones públicas y el trato directo. Tal y como se presenta en la Figura 2.2, se observa que además de los mecanismos de compra tradicionales, la compra de medicamentos puede ser intermediada por CENABAST, en donde los compradores recurren a este servicio para adquirir los productos.

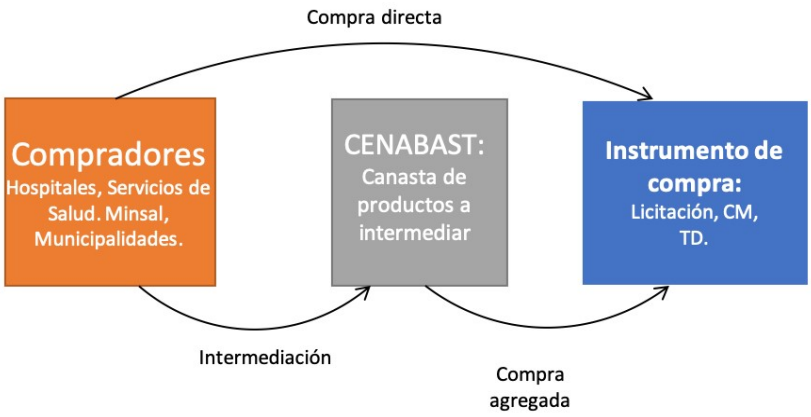


Figura 2.2: Mecanismos alternativos de compra de medicamentos en el sector público. Fuente: Malgarini I. [8]

Finalmente, se menciona que el procedimiento de compra de medicamentos en el sector público, puede ser considerado complejo, dada la cantidad de entidades involucradas desde la generación de un requerimiento de medicamentos, hasta la utilización de estos bajo ciertos controles.

En la Figura 2.3, se presentan las tres etapas básicas del proceso de adquisición de medicamentos, las cuales son:

1. Generación del requerimiento: Corresponde a la etapa en la que se determina qué productos se adquirirán, dependiendo de ciertos criterios y requerimientos generados por el personal de un establecimiento específico. A partir de esto se elabora un arsenal farmacológico único del establecimiento.

2. Elección de mecanismo de compra: Con el fin de adquirir el arsenal farmacológico establecido, es necesaria la elección del mecanismo de compra, dentro de los cuales se encuentra Convenio Marco, Licitaciones, Trato Directo y CENABAST.

3. Utilización del medicamento: Corresponde a la etapa final y contempla la utilización de los productos que son adquiridos y el control que debe tener el uso de estos. Por ejemplo, la existencia de protocolos de prescripción.

Figura 2.3: Etapas del proceso de compra de medicamentos en el sector público. Fuente: Elaboración propia.

Es importante mencionar, que el mecanismo de compras públicas más utilizado actualmente es la licitación pública, seguido por trato directo. También se destaca que los principales compradores corresponden a CENABAST y los hospitales, representando más del 80 % de las compras de medicamentos del Estado. En la Figura 2.4 se observan los cinco segmentos de compradores con mayor volumen en la participación en compras totales de medicamentos.

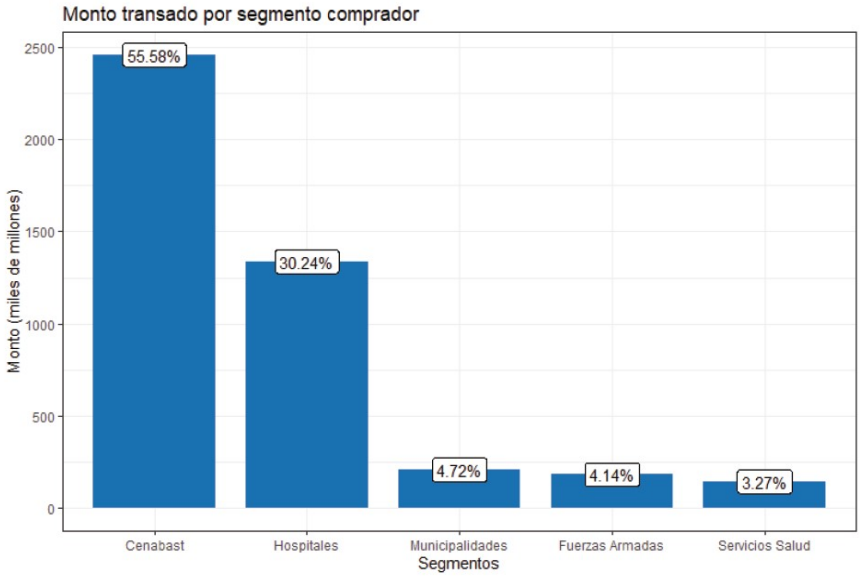


Figura 2.4: Participación en compras totales de medicamentos de los cinco segmentos de compradores con mayor volumen. Fuente: Malgarini I. [8]

2.3.2.3. CENABAST en las compras públicas de medicamentos

CENABAST es un organismo público que depende del Ministerio de Salud, y se encarga del abastecimiento de todos los Servicios de Salud, además de comprar los fármacos y dispositivos médicos, según una orden de consultorios y hospitales del sistema.

Desde el año 2012, CENABAST se focaliza principalmente en su rol de intermediario en las compras de medicamentos, sin tener la tarea de almacenar inventario o realizar la distribución de los productos a los diferentes establecimientos. Estos últimos son contratados de forma directa por los proveedores adjudicados. Las tareas otorgadas por su rol de intermediador, son realizadas sobre las necesidades de sus compradores, además de los planes y

programas del Ministerio de Salud.

A partir del Decreto N°78 de 1980 del Ministerio de Salud, CENABAST también debe asumir la tarea de decidir la canasta de productos a intermediar. Decisión que es apoyada por la Comisión de Adquisiciones de CENABAST, la cual corresponde a una "instancia en la que representantes del sector Público de Salud, deciden el resultado de las licitaciones que encabeza la Central de Abastecimiento, respecto a aquellos productos que superan las 3 mil UF" [44]. Esta canasta de productos es definida de forma anual, en conjunto con la demanda agregada por cada producto en esta. Finalmente, esta lista de productos se distribuye a los establecimientos, con el fin de que los compradores soliciten los bienes a comprar a través de CENABAST.

Por otro lado, desde la perspectiva de la libre competencia, CENABAST tiene una función importante al momento de consolidar la demanda de los Servicios de Salud, además de repercutir como un actor económico en el mercado de medicamentos.

A pesar de que no hay reglas rigurosas sobre los productos que se debiesen adquirir de forma directa o compradas a través de CENABAST, la ley de presupuestos del 2017, señala en su glosa 2, del subtítulo 22, que al menos un 60 % del gasto en insumos médicos y medicamentos de establecimientos de salud o servicios médicos, deben ser adquiridos por medio de CENABAST, siempre y cuando los medicamentos o insumos médicos a comprar estén en la Canasta Esencial de Medicamentos (CEM) y que los precios obtenidos por CENABAST sean competitivos.

Capítulo 3

Marco Metodológico

En el siguiente capítulo, se presentan todos los objetivos e hipótesis que se establecen para la realización del trabajo de tesis. Posteriormente, se exponen las fases metodológicas que posee el presente trabajo, entregando un contexto de los datos utilizados, a través de un análisis exploratorio de datos, y realizando la limpieza, detección y corrección de datos correspondiente. Finalmente, se presentan las etapas (Figura 3.1) necesarias para la creación del algoritmo de estructuración, el cual se compone de tres subprocesos importantes: predicción de la forma farmacéutica, extracción del principio activo y la extracción de la concentración.

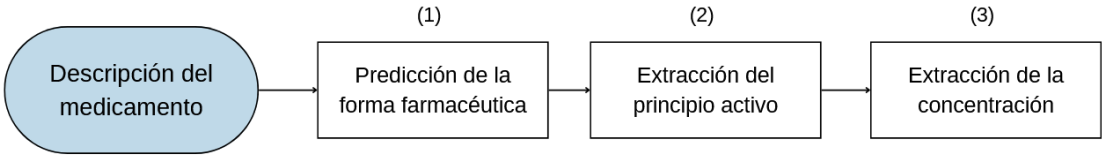


Figura 3.1: Diagrama del proceso de estructuración de medicametos.

Para tener un buen entendimiento de la siguiente sección, es necesario definir algunos conceptos previos que son utilizados de forma constante:

- **Descripción de un medicamento:** corresponde a un texto libre, es decir, que fue escrito por un humano, y que contiene información de un medicamento dado, por ejemplo, el principio activo de este.
- **Atributo de un medicamento:** corresponde a una característica que ayuda a describir un fármaco. Ejemplos de atributos son: principio activo, forma farmacéutica y concentración.
- **Valor de atributo:** corresponde a un valor específico que puede tomar un atributo determinado. Por ejemplo, para el atributo "principio activo", posibles valores pueden ser comprimido, ampolla y jarabe.
- **Principio activo:** es el ingrediente principal de un medicamento, responsable del efecto deseado.
- **Forma farmacéutica:** corresponde al medio por el que se adaptan los principios activos para constituir un medicamento.

- **Concentración:** corresponde a la cantidad de principio activo que contiene un medicamento.
- **Predicción:** se refiere a la utilización de un algoritmo para la estimación de una etiqueta, a partir de la información ingresada.
- **Extracción:** se refiere a la búsqueda y entrega de valores predeterminados, a partir de un texto específico.

Con el objetivo de entregar al lector una simplificación de la tarea a desarrollar, se presenta el siguiente diagrama. En la Figura 3.2 se observa el ejemplo de una descripción de medicamento, "Levotirozina 100 mg x 90 cm", a partir de la cual se deben extraer los valores de los atributos forma farmacéutica, principio activo y concentración. Respecto a los valores de cada atributo, estos se obtienen a lo largo del proceso de estructuración, el cual posee 3 subprocesos. En el primero se predice la forma farmacéutica del medicamento, correspondiente a comprimido. En el segundo se extrae el principio activo, que corresponde a Levotiroxina. Y en el tercero se extrae la concentración, correspondiente a 100 mg.

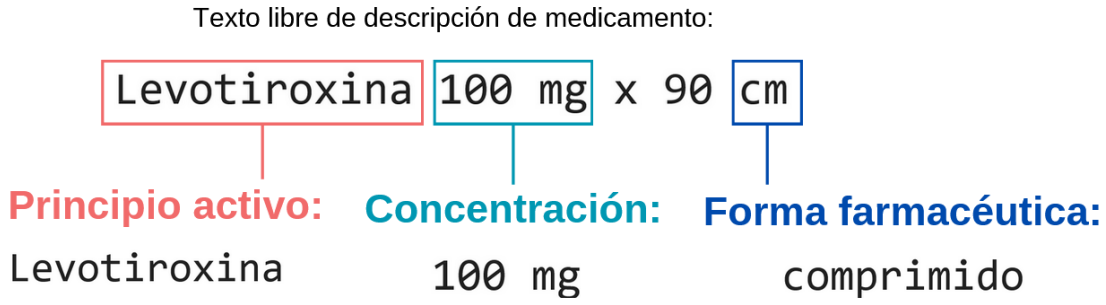


Figura 3.2: Diagrama de la estructuración de una descripción de medicamento.

3.1. Objetivos e Hipótesis

Con el fin de lograr una buena estructuración de medicamentos, se establece una serie de objetivos que buscan una implementación eficiente del algoritmo.

Se establece como objetivo general el aplicar un algoritmo de estructuración desarrollado a través de herramientas de PLN y AM, para lograr de forma eficiente que al menos un 70 % de los medicamentos de la base de datos Vista Medicamentos, logre ser estructurado de forma correcta.

Con el fin de lograr lo anterior, se establecen los siguientes objetivos específicos:

1. Realizar un análisis exploratorio de datos a la base de datos Vista Medicamentos.
2. Crear diccionarios limpios de fallas escriturales de cada uno de los atributos de los medicamentos.
3. Establecer métricas que permitan observar la mejora de los algoritmos.

4. Utilizar herramientas de Procesamiento de Lenguaje Natural (*Word2Vec*), y herramientas de Aprendizaje de máquinas (*Support Vector Machine*), para lograr una estructuración correcta de los medicamentos.

A partir de lo anterior, se establece como hipótesis del trabajo: la aplicación de un algoritmo desarrollado mediante herramientas computacionales, permite estructurar al menos un 70 % de la información de fármacos en compras públicas.

3.2. Fases Metodológicas

En la siguiente sección se presenta la metodología utilizada para el desarrollo del algoritmo de estructuración, comenzando desde la base de datos Vista Medicamentos, hasta la correcta estructuración de las descripciones de los fármacos. Para lo anterior, se establece una serie de fases que permiten desarrollar el trabajo de tesis, las cuales son:

1. Análisis Exploratorio de Datos (EDA).
2. Corrección y limpieza de las bases de datos.
3. Creación de diccionarios.
4. Creación del estructurador de medicamentos.
    a) Generación de subproceso de predicción de la Forma farmacéutica.
    b) Generación de subproceso de extracción del Principio activo.
    c) Generación de subproceso de extracción de la Concentración.
5. Establecimiento de métricas.

3.2.1. Análisis Exploratorio de los Datos

Para el correcto entendimiento de las etapas siguientes, se debe establecer un contexto de las bases de datos empleadas en la creación del algoritmo de estructuración. El algoritmo depende de tres bases de datos, sin embargo, solo una de ellas es de total importancia para la correcta creación del algoritmo: la base de datos **Vista Medicamentos**.

Con el objetivo de crear el algoritmo estructurador de medicamentos, es necesario contar con la información de medicamentos de diferentes fuentes, con el fin de tener una mayor diversidad en la información utilizada. Para lo anterior, se utilizan tres fuentes de datos, las cuales integran información relevante al algoritmo:

- Vista Medicamentos: Base de datos que contiene información sobre compras de medicamentos provenientes de convenios marco, licitaciones adjudicadas y trato directo.
- "Concentracion ISP" y "Forma farmaceutica ISP": Bases de datos que contienen información de medicamentos registrados por el Instituto de Salud Pública (ISP).

A continuación, se describe cada una de las bases de datos antes mencionadas.

3.2.1.1. Vista Medicamentos

Esta fuente de datos corresponde a un archivo de valores separados por comas (CSV), originalmente entregado por CENABAST. Vista Medicamentos contiene las órdenes de compra que fueron realizadas entre los años 2011 y 2020. Compras que provienen de Licitaciones adjudicadas, Convenios marco y Trato directo. El valor único en esta base de datos corresponde al código ZGEN, el cual agrupa los fármacos con un mismo principio activo, concentración y forma farmacéutica. Esta fuente de datos fue modificada posteriormente a pedido de CENABAST por una organización externa, quien realizó un previo etiquetado de datos, entregando para cada fármaco su respectiva forma farmacéutica, principio activo y concentración. Cabe destacar, que el método utilizado para la previa estructuración de Vista Medicamentos, es desconocido.

El objetivo de utilizar esta base de datos, es contar con información previamente etiquetada, para lograr entrenar los modelos que son utilizados en los procesos de extracción, además de contar con información de las etiquetas originales que permitan observar el desempeño del proceso de estructuración.

La fuente de datos está compuesta por 20 atributos y 2872330 registros, en los cuales hay 1385405 órdenes de compra. Dentro de las 20 columnas que posee la base, las más importantes corresponden a: "Descripcion comprador", "Descripcion proveedor", "Forma Farmaceutica", "Principio Activo" y "Concentracion". En la Tabla del Anexo A.1 se presenta una breve descripción de cada uno de los atributos presentes en la base de datos.

Es importante mencionar que, a pesar de que la base de datos posee una gran cantidad de atributos, los cuales podrían ser de bastante utilidad a la hora de realizar análisis econométricos sobre medicamentos en compras públicas, en el contexto del trabajo actual, solo algunos de estos son los importantes para la creación del algoritmo de estructuración:

- Descripción comprador y proveedor.
- Forma Farmacéutica.
- Principio Activo.
- Concentración.

De acuerdo con un análisis más detallado de cada una de las variables antes mencionadas, se debe destacar que solo los atributos "Descripción comprador" y "Descripción proveedor" poseen valores faltantes, con un porcentaje de faltantes, respecto al total de registros presentes en la fuente de datos, de 2,61 % y 17,13 % respectivamente. No obstante, la significativa falta de datos en uno de los atributos de interés, se corrige en la siguiente sección de **Corrección y Limpieza de los datos**.

Respecto a la variable "Forma Farmaceutica", se debe mencionar que es una variable del tipo categórica y está compuesta por 21 categorías, dentro de las cuales, las más relevantes son: **comprimido**, **ampolla**, **frasco** y **jarabe**. En la Figura 3.3 se observa un histograma para las categorías de la forma farmacéutica de los medicamentos, destacando a su vez que hay una gran brecha entre algunas categorías, en especial entre **comprimido** y **kit**.

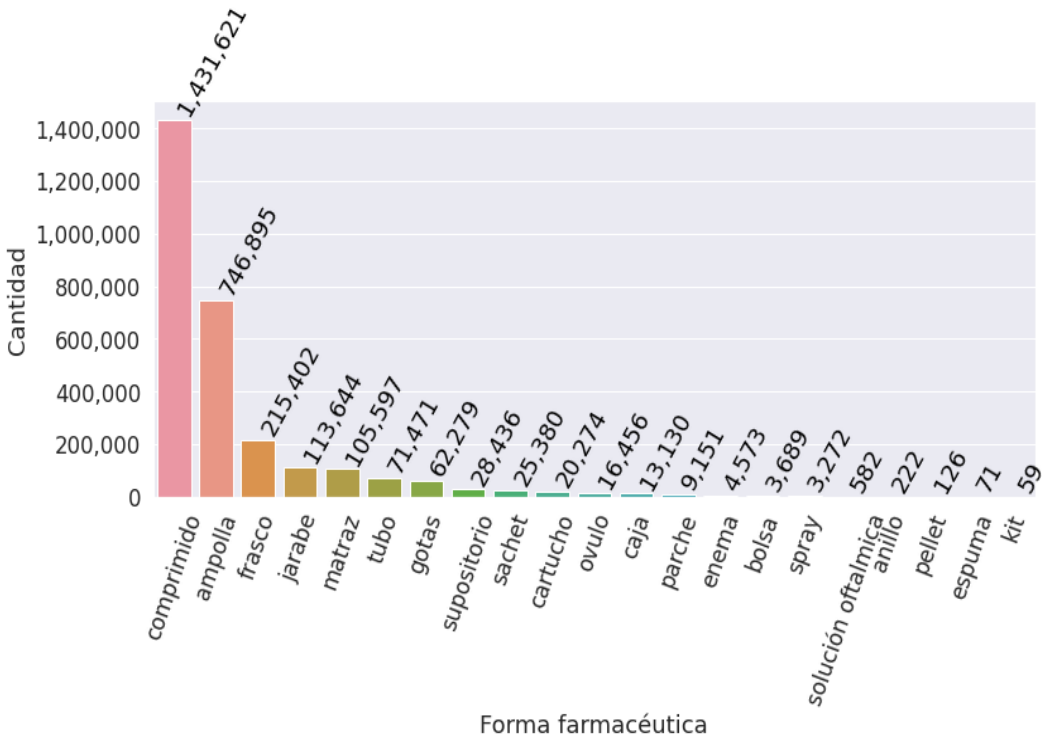


Figura 3.3: Cantidad de registros de órdenes de compra por cada forma farmacéutica.

En cuanto a la variable "Principio Activo", se caracteriza por ser del tipo categórica, sin embargo, a diferencia del atributo "Forma Farmaceutica", esta variable posee 1318 categorías, correspondientes al componente principal que posee cada medicamento. Dentro de los principios activos más destacados se encuentran el **cloruro de sodio**, **paracetamol** y **Diclofenaco**, presentes en el 2,93 %, 1,49 % y 1,1 % de los registros, respectivamente. Por otro lado, también se debe mencionar que algunos medicamentos están conformados por más de un principio activo. Como se puede observar en la Tabla 3.1, la mayoría de los medicamentos están compuestos de un solo principio activo, ya que un 92,02 % de los registros pertenece a esta categoría.

Tabla 3.1: Porcentaje de la cantidad de Principios activos en los registros de compras de medicamentos.

Número de principios activos en el medicamento	**Porcentaje [%]**
1	92,02
2	7,34
3	0,52
4	0,11
5	0,002

Finalmente, respecto a la variable "Concentración", se debe mencionar que este atributo está compuesto por un número y por una unidad de medida, las cuales pueden ser: mg/ml, ui,

mcg, % y g, entre otros. En la Tabla 3.2 se observa que el 78,41 % de los registros de la base de datos tienen una concentración con unidad de medida **mg**, es decir, casi la totalidad de los fármacos de la base de datos posee una unidad de medida de **miligramo**, puesto que la siguiente unidad de medida predominante de la base es **%**, la cual solo está presente un 11,55 %.

Tabla 3.2: Porcentaje de la cantidad de Principios activos en los registros de compras de medicamentos.

Unidad de medida	**Porcentaje [%]**
mg	78,417
%	11,556
ml	4,078
ui	2,989
meq	0,545
g	0,282
mcg	0,256

Pese a que solo hay 4 variables importantes para la creación del algoritmo de estructuración, en la base de datos Vista Medicamentos, también existen algunos atributos que ayudan a situar los datos dentro de un contexto. Ejemplo de esto son las variables "Orden de compra", "Segmento comprador", "Cantidad" y "Monto neto", de las cuales se presenta una breve descripción a continuación.

- **Orden de compra:** Código único que representa una compra. Está compuesto por una unidad de compra (ID que identifica a un cliente), un correlativo de compra (cuantas compras ha hecho el cliente en ese año), la modalidad por la cual se realizó la compra (Ej. Convenio marco) y el año en el que se realizó. La importancia de esta variable es la identificación del método de compra que se utiliza. En la Figura 3.4, se observa que hay una clara tendencia a la utilización de "Sin emisión automática". Mientras que el método menos seleccionado es el de "Compra Ágil".

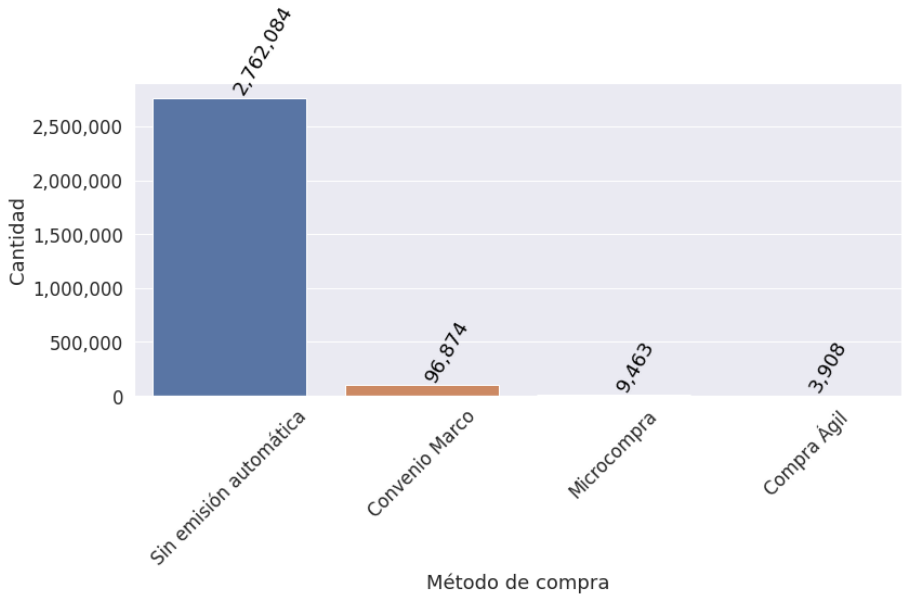


Figura 3.4: Cantidad de registros de órdenes de compra por cada Método de compra.

- **Segmento comprador:** Como su nombre lo especifica, corresponde al segmento al que pertenece la entidad compradora. En la Figura 3.5, se observa que la mayoría de los compradores son hospitales, representados por el 50 % de los registros, seguidos por las municipalidades, con un 29 % de los datos. También se debe destacar que solo un 0,1 % de los compradores pertenecen al segmento de "Organismos Públicos".

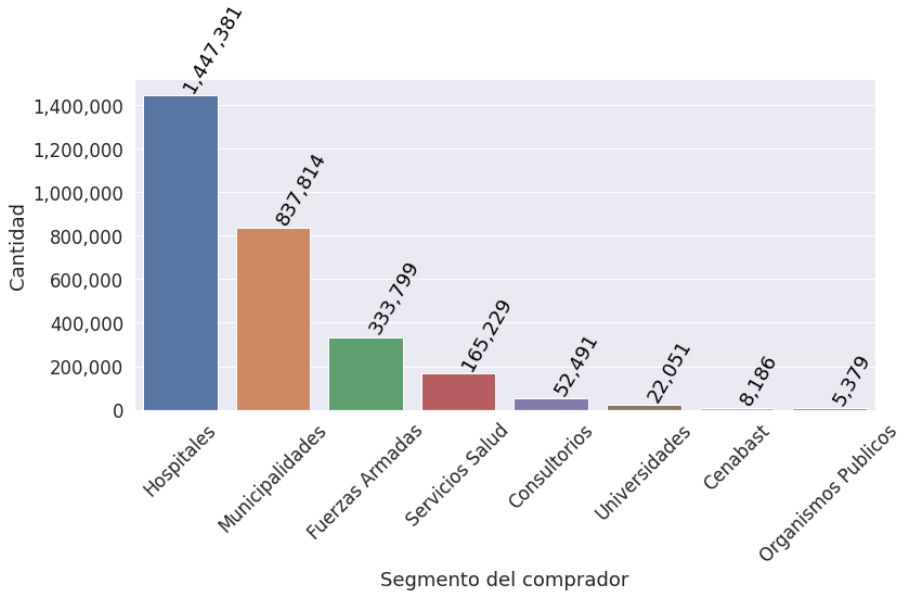
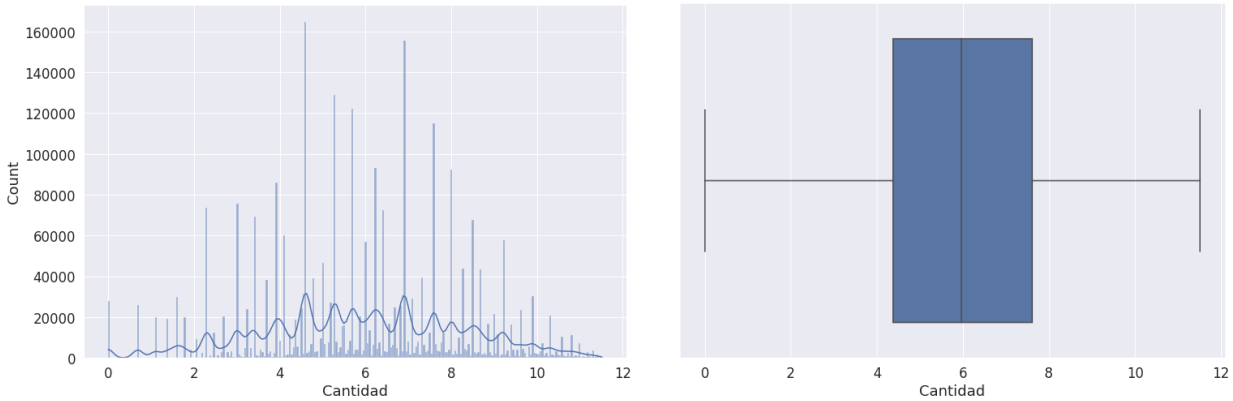


Figura 3.5: Segmentos a los cuales pertenecen los compradores.

- **Cantidad:** Corresponde a la cantidad de productos adquirida por cada uno de los registros de compra de fármacos presentes en la base de datos. En las Figuras 3.5(a) y 3.5(b) se presentan un histograma y un *Boxplot* de las cantidades normalizadas de

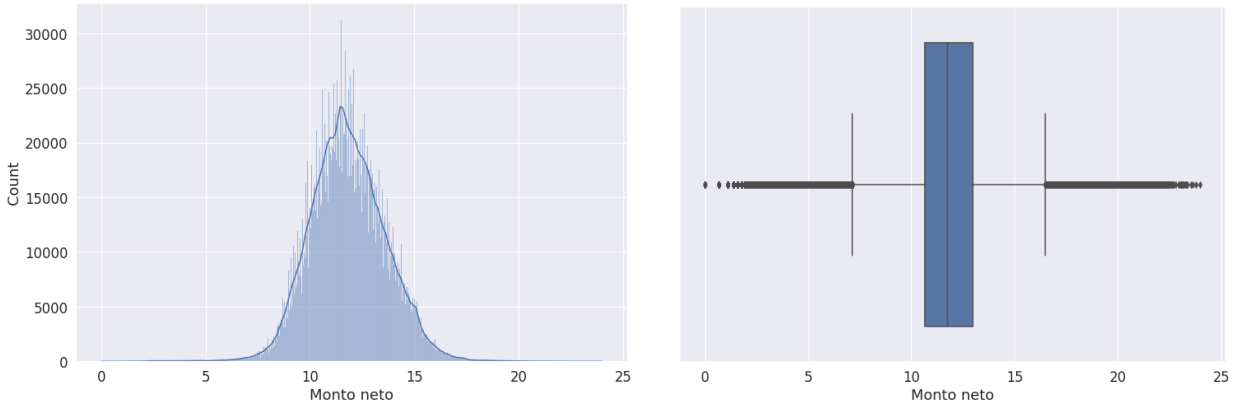
productos adquiridos. Se observa que el promedio de cantidad adquirida por compra es de aproximadamente 400 fármacos ($e^{5,9848}$).



(a) Histograma (b) *boxplot*

Figura 3.6: Cantidades de productos adquiridos por compra, datos normalizados.

- **Monto neto:** Se define como la multiplicación entre la cantidad de un producto adquirido y su precio único. En las Figuras 3.6(a) y 3.6(b) se presenta el histograma y el *Boxplot* del monto neto en compras de medicamentos. Se observa que el promedio aproximado de ambas gráficas es de 11,83, es decir, por cada compra registrada se presenta un monto neto a pagar de \$137.469 ($e^{11,8311}$).



(a) Histograma (b) *boxplot*

Figura 3.7: Monto neto por compra, datos normalizados.

Si bien las columnas descritas anteriormente no son significativas a la hora de cumplir con el objetivo del trabajo de tesis, sí ayudan a explicar de forma breve el comportamiento de los compradores de los medicamentos que se están tratando. Por ejemplo, el hecho de que gran parte de las compras sean hechas sin emisión automática, que haya una tendencia por comprar altas cantidades de productos a la vez, o que los hospitales son los principales usuarios de las adquisiciones de medicamentos, entrega una mayor importancia a los resultados que se puedan obtener de este algoritmo de estructuración, dado que el fin es contribuir a la población.

3.2.1.2. "Concentracion ISP" y "Forma farmaceutica ISP"

Estas bases de datos corresponden a archivos separados por comas (CSV), y contienen información de medicamentos registrados por el Instituto de Salud Pública, en adelante ISP (https://registrosanitario.ispch.gob.cl/). Por ejemplo, información del tipo de envase del medicamento, la unidad de medida de su concentración y su principio activo. Esta página permite solicitar el registro sanitario de un producto farmacéutico que ha sido importado o fabricado en el país, y sobre el cual se realiza un proceso de evaluación y estudio sistemático de sus propiedades. Dentro de las características a evaluar se encuentran las del tipo farmacéuticas, clínicas, toxicológicas o farmacológicas. Se realiza lo anterior con el fin de verificar su calidad, eficacia y seguridad. Posterior al proceso de evaluación, se genera una inscripción de un rol especial con numeración correlativa, que habilita y autoriza la distribución y uso del producto en el país [45].

El objetivo de utilizar estas bases de datos es integrar más información respecto a las formas farmacéuticas, principio activo y concentración de un medicamento, datos que son utilizados posteriormente en la creación de diccionarios, los cuales permiten disminuir el tiempo de cómputo al momento de realizar la estructuración de medicamentos.

Se obtienen 2 fuentes de datos con diferente información, ambas con un mismo código llamado "Código ISP", el cual permite la creación de una nueva base a partir de la unión de las bases "Concentracion ISP" y "Forma farmaceutica ISP", con el objetivo de obtener una mayor información de los fármacos. Lo anterior se obtiene a partir de la creación de un diccionario de información, el cual complementa el diccionario creado a partir de la base Vista Medicamentos. Esto último es profundizado en la sección **Creación de diccionarios**.

De forma específica, la base de datos "Concentracion ISP" aporta información sobre: principio activo de un medicamento, su concentración y su correspondiente unidad de medida. Por otro lado, "Forma farmaceutica ISP" proporciona datos sobre el tipo de envase de un fármaco, su descripción, forma farmacéutica, condición de almacenamiento, periodo de eficacia y su contenido. Si bien, hay información relevante para el uso de los medicamentos, tal como la condición a la que se debe almacenar este, para la creación del estructurador solo son útiles algunas de las columnas de esta base de datos. En lo que sigue, se detallan únicamente las características de la nueva base de datos: "Información ISP", creada a partir de la intersección de las fuentes de datos mencionadas anteriormente. Además, es necesario mencionar que esta base de datos se asemeja a una base de datos estructurada y con medicamentos estructurados, lo que permite complementar la información aportada por Vista Medicamentos.

La base de datos "Información ISP" está compuesta por 5 atributos, los cuales se observan en un extracto de la base de datos de la Tabla 3.3, y se describen a continuación:

1. Envase: Describe el tipo de envase que tiene el fármaco, entregando características de este. Por ejemplo, **Ampolla de vidrio incoloro tipo I**.

2. Unidad de Medida del envase: Especifica de forma más breve el tipo de envase que tiene el medicamento, mayormente se define como la forma farmacéutica. Por ejemplo, **Frascos - Ampollas / 5ML**.

3. Principio activo: Define el principio activo de un medicamento, además, entrega la forma alternativa de escribir el principio activo de un medicamento. Por ejemplo, **Linezolida**.

4. Concentración: Corresponde a la parte numérica de la concentración de un medicamento, sin su unidad de medida. Por ejemplo, **250**.

5. Unidad de Medida de la concentración: Corresponde a la unidad de medida de la concentración de un medicamento. Por ejemplo, **mg/5 ml**.

Tabla 3.3: Extracto base de datos Información ISP.

Envase	Unidad de Medida del envase	Principio activo	Concentración	Unidad de Medida de la concentración
- Ampolla clase hidrolitica de vidrio tipo i incoloro con nave plastica de poliestireno de alto impacto.	G	atracurio besilato	25	mg/2,5 mL
- Frasco ampolla de vidrio incoloro tipo i etiquetado con tapon de goma sello de aluminio y tapa de polipropileno.	Frasco - ampolla /5ML	fluorouracilo	250	mg/5 mL
- Blister de lamina pvdc/pvc transparente incoloro y lamina de aluminio impresa.	Comprimidos recubiertos	hidroclorotiazida	12,5	mg

En la Tabla 3.3 se observa que, a diferencia de la base Vista Medicamentos, la base Información ISP no posee una columna de texto libre relacionada con la descripción de un fármaco. Sin embargo, sí se encuentra estructurada en 4 atributos similares a la base Vista Medicamentos: forma farmacéutica, principio activo y concentración (numeración + unidad de medida).

Esta fuente de datos posee 45300 registros y 5 atributos, los cuales fueron descritos anteriormente. Dentro de estas variables, se debe mencionar que estas poseen una cantidad despreciable de datos faltantes, puesto que en la columna de "Unidad Medida del envase", correspondiente a la variable con más datos nulos, solo un 3.87 % corresponde a ellos.

Respecto a la composición de los atributos, la variable "Unidad de Medida del envase", a diferencia de la variable "Forma Farmaceutica" de la base Vista Medicamentos, posee más categorías, aunque muchas de ellas no entregan una información muy exacta sobre la forma farmacéutica de un medicamento. A modo de comparación, "Forma Farmaceutica" de Vista Medicamentos está compuesta por 21 categorías, destacando que **comprimido** es la categoría más popular. Por otro lado, "Unidad de Medida del envase" posee 713 categorías, representadas a través de los valores únicos existentes en la columna, donde, al igual que en Vista Medicamentos, la categoría **comprimidos** es la más repetida. Aunque la cantidad de categorías presentes en la columna "Unidad de Medida del envase" sea extremadamente variada en comparación a la existente en Vista Medicamentos, es necesario mencionar que no todos los datos únicos representan una forma farmacéutica. Por lo tanto, se reducen las 713 alternativas a solo 19. El proceso de reducción se presenta en la siguiente sección de **Corrección y limpieza**.

En la Tabla 3.4 se presenta una comparativa entre algunas categorías de las columnas de "Unidad de Medida del envase" y "Forma Farmaceutica", de las bases Información ISP y Vista Medicamentos respectivamente. Se observa que muchas de las categorías son similares, y se añaden algunas que tienen por objetivo el especificar aún más el tipo de forma farmacéutica que tiene un medicamento. Por ejemplo, en "Forma Farmaceutica", solo la categoría de **ampolla** engloba a todos aquellos medicamentos que poseen una forma farmacéutica de ampollas, frasco ampollas y jeringas, mientras que en el atributo "Unidad de Medida del envase", existe una categoría específica para cada uno de ellos.

Tabla 3.4: Comparativa entre etiquetas de Información ISP y Vista Medicamentos.

Forma farmacéutica de Información ISP	**Forma farmacéutica de Vista Medicamentos**
Comprimidos	Comprimido
Comprimidos recubiertos	
Comprimidos masticables	
Grageas	
Cápsulas	
Frasco - ampolla	Ampolla
Ampollas	
Jeringas	
Tubos	Tubos
Cremas	
Pomos	

Por otro lado, en lo que respecta a la columna "Principio activo", posee 5298 diferentes valores únicos de principios activos, destacando que el que está más presente en la base de datos es el **paracetamol**. Además, se debe mencionar que, si bien pareciera que hay 5298 principios activos diferentes, esto no necesariamente es real, dado que la columna posee algunos principios activos escritos de formas diferentes. Por ejemplo, **Cloruro** y **Cloruro de sodio**. Lo anterior genera un desbalance en esta columna, ya que según el percentil 75, el 75 % de los principios activos registrados aparecen a lo más 7 veces, afirmación que no tendría sentido, dado que el principio activo más repetitivo es el **paracetamol**, el cual aparece 696 veces en la base, y existen otros principios activos que aparecen con frecuencia. Todo lo anterior indica que se debe realizar una limpieza en la columna, la cual es clave en la construcción del diccionario.

Finalmente, en relación con las columnas de "Concentración" y "Unidad de Medida de la concentración", ambas son parte importante de la concentración de un medicamento, puesto que la primera columna representa la numeración y la segunda corresponde a su unidad de medida. Respecto a la primera variable, se menciona que posee 5743 valores diferentes, y está compuesta de números enteros y decimales, destacando que los valores más repetidos corresponden a **10** y **100**. Por otro lado, también se debe mencionar que un 43,8 % de los valores son decimales y que posiblemente sean ruido en la base, ya que solo aparecen una sola vez.

Con respecto a la variable "Unidad de Medida de la concentración", se destaca que está compuesta de 246 valores diferentes, en donde el valor **mg** representa a un 59,45 % de los

registros. También se debe mencionar que posiblemente el 25,6 % de las unidades de medida presentes en la variable son ruido, ya que aparecen una única vez en toda la base de datos. Por ejemplo, **mg/comprimido**.

3.2.2. Corrección y pre-procesamiento de datos

Como se menciona en secciones anteriores, cada una de las bases de datos representa una parte importante en la creación del algoritmo estructurador. A pesar de que la base Vista Medicamentos sea significativamente más importante que las demás fuentes, esto no implica que las bases restantes sean despreciables y se pueda prescindir de ellas. Por lo anterior, es necesario que cada una de las fuentes de datos esté limpia y corregida para su correcta utilización.

Dado que la base Vista Medicamentos es la principal, es necesario realizar algunas modificaciones en ella, puesto que gran parte del algoritmo depende de qué tan bien estén los datos de esta base.

La primera corrección realizada corresponde a la eliminación de los datos nulos de cada una de las variables. Para los casos de las columnas "Descripcion comprador" y "Descripcion proveedor", el tener datos faltantes significa que no hay descripción de medicamento que se pueda estructurar, por lo que es irrelevante tener el resto de información de un registro. Considerando que la columna de "Descripcion comprador" solo tiene un 2,54 % de datos faltantes, y que la columna "Descripcion proveedor" posee un 17,27 % de datos faltantes en los registros, se utiliza esta última para completar la información faltante en "Descripcion comprador". En consecuencia de lo anterior, los datos faltantes de la variable de descripción de los medicamentos disminuyen de 2,54 % a 0,45 %. De esta manera, la base final queda con el 90 % de los registros originales.

Posterior a lo anterior, se continúa con solo 4 variables de la base original, las que corresponden a: "Descripcion comprador", "Forma Farmaceutica", "Principio Activo" y "Concentracion", y se eliminan los valores duplicados de la base, quedando con un 85 % de la base original.

Es importante recordar al lector que las descripciones de medicamentos utilizadas en el algoritmo están en texto libre, y que originalmente fueron escritas por una persona humana. Por lo tanto, poseen errores que en ocasiones dificulta la estructuración del texto. Por ejemplo, descripciones del tipo **según cotización presentada por el proveedor**, son bastante comunes en la fuente de datos, y dado que el objetivo del algoritmo es extraer los valores de atributo desde las descripciones, es imperante tener descripciones útiles. En la siguiente tabla se presentan aquellas descripciones de poca utilidad, junto a su frecuencia en la base.

Tabla 3.5: Ejemplo de descripciones irrelevantes y su frecuencia.

Descripción medicamento	Frecuencia en la base
SEGÚN RESOLUCION 4885/2015 QUE APRUEBA CONTRATACION DIRECTA PARA COMPRA DE MEDICAMENTOS	171
SEGÚN COTIZACIÓN PRESENTADA POR EL PROVEEDOR	116
COMPRA DE MEDICAMENTOS FARMACIA COMUNAL (JUNIO) LABORATORIO CHILE	102
CONTRATACION SUMINISTRO ARSENAL FARMACOLOGICO DE ATENCION PRIMARIA, PARA LOS ESTABLECIMIENTOS DE SALUD DEPENDIENTES DE LA MUNICIPALIDAD DE HUECHURABA	97
MEDICAMENTOS SEGUN RESOLUCION FUNDADA N. 667 DEL 27/02/2018	71
Proveniente de licitacion 3752-1-LE19 convenio suministros medicamentos Farmacia popular	70
...	...
SEGÚN CONTRATO SUMINISTROS 1175-279-LQ17	5
MEDICAMENTOS SEGUN CONVENIO SUMIINISTRO 2274-416-LP11	5

Se observa en la Tabla 3.5, que una cantidad de las descripciones de la base original posee texto libre que representa un aporte mínimo en el proceso de estructuración, puesto que, a pesar de no tener descripciones de medicamentos en el texto, sí tienen información sobre la forma farmacéutica, principio activo y concentración, aunque el registro no tenga una descripción real que se pueda estructurar.

Lo mencionado anteriormente representa uno de los errores más relevantes que posee la base de datos previamente estructurada de CENABAST. Esto último debido a que existe una cantidad no menor de descripciones que no poseen información de la forma farmacéutica, principio activo o concentración de un fármaco en el texto libre, pero que sí poseen etiquetas en cada una de las columnas de estos atributos. En la Tabla 3.6 se presenta un extracto de la base de datos Vista Medicamentos, en donde se observa la inconsistencia presente en algunos de los registros de la base.

Tabla 3.6: Ejemplo de descripciones irrelevantes, con información en los atributos de interés, datos de la base Vista medicamentos.

Descripción de Medicamento	Forma Farmaceutica	Principio Activo	Concentración
MEDICAMENTOS SEGUN CONVENIO SUMINISTRO 2274-416-LP11	Comprimido	Carbamazepina	200 mg
MEDICAMENTOS SEGUN CONVENIO SUMINISTRO 2274-416-LP11	Jarabe	Cefadroxilo	250 mg
MEDICAMENTOS FARMACIA COMUNAL (Lab. CHILE)	Jarabe	Aciclovir	200 mg
MEDICAMENTOS LAB. CHILE	Tubo	Aciclovir	5 %
listado de medicamentos	Frasco	Cloranfenicol	0.5 %
CONTRATO SUMINISTRO 2777-5-LQ19 SEGUN DETALLE	Jarabe	Claritromicina	250 mg
BIENESTAR SOCIAL 1RA ZONA NAVAL	Comprimido	Quetiapina	25 mg

A partir de lo presentado anteriormente, se eliminan de la base de datos, todos aquellos registros cuyas descripciones poseen una frecuencia superior a 4. Esto último permite mantener aquellas descripciones que están presentes con una frecuencia menor, pero que tienen información relevante de los atributos a estructurar.

Finalmente, en cuanto a las variables restantes de la fuente de datos Vista Medicamentos, las cuales son: “Forma Farmaceutica”, “Principio activo” y “Concentración”, debido a la ausencia de datos faltantes, no se realiza un proceso para prescindir de ellos. Sin embargo, sí se realiza una limpieza de texto, el cual incluye:

- Transformar todas las letras a minúsculas.
- Eliminar caracteres especiales.
- Eliminar tildes

Además, también se realiza un proceso similar sobre la descripción del medicamento (por ejemplo, eliminación de caracteres especiales), incluyendo una etapa en donde se divide una oración, separando cada palabra. Un ejemplo de este proceso se puede observar en la Figura 3.8.

```
Descripción del medicamento: ACIDO ASCORBICO CM 100 MG
División de la descripción: ['acido', 'ascorbico', 'cm', '100', 'mg']
```

Figura 3.8: Ejemplo de división de una descripción.

En lo que respecta a la base de datos “Información ISP”, dado que proviene de la unión de dos bases de datos (“Concentracion ISP” y “Forma farmaceutica ISP”), posee errores que deben ser solucionados. Primeramente, se tratan los valores faltantes, los cuales son eliminados, por lo que la base de datos se reduce a un 98 % del tamaño original.

Posteriormente, se trata el texto de cada una de las variables, realizando una limpieza a través de la eliminación de *stopwords* o “palabras vacías”, y extracción de números y símbolos. Una de las columnas con más problemas es “Unidad de Medida del envase”, cuyo contenido es similar a la presente en la columna “Forma farmacéutica” de la base Vista Medicamentos. Como se menciona en la sección **Análisis Exploratorio de Datos**, “Unidad de Medida del envase” posee 713 valores únicos, donde gran parte de ellos son efectivamente una forma farmacéutica, sin embargo, un porcentaje restante son valores que no implican necesariamente una forma farmacéutica. Dentro de este último grupo, hay 2 posibles alternativas para tratar este problema: (1) eliminar los datos que no coincidan con una forma farmacéutica, o (2) reemplazar los textos de esta variable con parte del texto presente en la columna “Envase”, que en ocasiones presenta la forma farmacéutica del medicamento. Para ejemplificar lo anterior, se presenta la siguiente Figura 3.9:

```
Envase: Ampolla de vidrio incoloro tipo I
Unidad de Medida del envase: dosis
Principio activo: fluorouracilo
Concentración: 250
Unidad de Medida de la concentración: mg/5 ml
```

Figura 3.9: Ejemplo de posible reemplazo, usando la columna Envase.

Se puede observar en el ejemplo, que el atributo “Unidad de Medida del envase” no posee el valor correcto, ya que **dosis** no es una forma farmacéutica. Sin embargo, se observa que en

la variable "Envase" está presente la palabra **ampolla**, la cual sí corresponde a una forma farmacéutica. Por lo tanto, se utiliza parte de esta variable para corregir el problema presente en "Unidad de Medida del envase". Si bien el proceso es útil en la mayoría de los casos en los que el valor presente en "Unidad de Medida del envase" no es una forma farmacéutica, en ocasiones no hay valores presentes en "Envase" que puedan ser de utilizadas, por lo tanto, en estos casos, se elimina el registro de la base.

En las Tablas 3.7 y 3.8, se presenta una comparación entre un extracto de la base pre-limpieza y un extracto de la base posterior a la limpieza. Se puede observar en la Tabla 3.8, el resultado de todas las modificaciones antes mencionadas, además de la extracción de las columnas más importantes para la siguiente sección: "Unidad de Medida del envase" (forma farmacéutica), "Principio activo" y "Concentración".

Tabla 3.7: Extracto de base Información ISP pre-limpieza.

Envase	Unidad de Medida del envase	Principio activo	Concentración	Unidad de Medida de la concentración
Ampolla de vidrio incoloro tipo I	AMPOLLAS	5-FLUOROURACILO	250	mg/5 mL
FRASCO AMPOLLA DE VIDRIO INCOLORO TIPO I	FRASCO - AMPOLLAS/5 ML	5-FLUOROURACILO	250	mg/5 mL
Blíster de PVC/ALU impreso	COMPRIMIDOS RECUBIERTOS	LINEZOLID	600,0	mg
Botella de vidrio ámbar tipo III con tapa PP	g	LINEZOLIDA	5,1000	g
Bolsa de co-poliéster, rotulada	mL	LINEZOLIDA	0,200	G

Tabla 3.8: Extracto de base Información ISP post-limpieza.

Unidad de Medida del envase (forma farmacéutica)	Principio activo	Concentración
ampolla	fluorouracilo	250 mg/5 ml
ampolla	fluorouracilo	250 mg/5 ml
comprimido	linezolid	600 mg
botella	linezolida	5,1 g
bolsa	linezolida	0,2 g

3.2.3. Creación de diccionarios

En términos del presente trabajo, un diccionario corresponde a un archivo que almacena información sobre formas farmacéuticas, principios activos y concentraciones de medicamentos, en la Tabla 3.9 se observa un ejemplo de diccionario, en el contexto trabajado. Es necesario mencionar que el objetivo de utilizar diccionarios durante el proceso de estructuración de datos de medicamentos, es disminuir el tiempo de extracción de valores de atributos desde las descripciones, delimitando las posibles respuestas del algoritmo para cada uno de los atributos.

Tabla 3.9: Ejemplo de diccionario utilizado en el proceso de estructuración de medicamentos.

Forma farmacéutica	**Principio activo**	**Concentración**
ampolla	f antihemof viii	1.000 ui
comprimido	cabergolina	0.5 mg
comprimido	donepezilo	23 mg
jarabe	prednisona	1 mg
ampolla	valproico acido	500 mg

Para obtener un diccionario general, se crea para base de datos un diccionario a partir de la agrupación de datos, agrupando por las columnas: "Forma farmacéutica", "Principio activo" y "Concentración". Esto finalmente resulta en las creaciones de 2 diccionarios con 793 y 2092 tripletas de datos, del diccionario de Vista Medicamentos y el diccionario de Información ISP respectivamente. En las Tablas 3.10 y 3.11 se observan extractos de cada uno de los diccionarios.

Tabla 3.10: Extracto de diccionario creado con datos de Información ISP.

Forma farmacéutica	**Principio activo**	**Concentración**
ampolla	fluorouracilo	250 mg/5 ml
ampolla	fluorouracilo	500 mg/5 ml
ampolla	acetato de sodio	222 mg/100 ml
ampolla	acetato de sodio trihidrato	30 g
ampolla	acido amidotrizoico	47,18 g/100 ml
ampolla	acido amidotrizoico	59,76 g/100 ml
ampolla	acido ascorbico	100 mg/ml

Tabla 3.11: Extracto de diccionario creado con datos de Vista Medicamentos.

Forma farmacéutica	**Principio activo**	**Concentración**
ampolla	aciclovir	250 mg
ampolla	aciclovir	500 mg
ampolla	adenosina	6 mg
ampolla	albumina humana	20 %
ampolla	alprostadil	500 mg
ampolla	amikacina	500 mg
ampolla	aminofilina	250 mg

Finalmente, se genera un diccionario general a través de la unión de los diccionarios creados con las bases de datos, con el fin de tener información que proviene de distintos orígenes,

proceso que se encuentra en el diagrama de la Figura 3.10. A causa de lo anterior, con la información complementada se tiene conocimiento de diversos nombres de principios activos y concentraciones, lo que permite una extracción más generalizada.

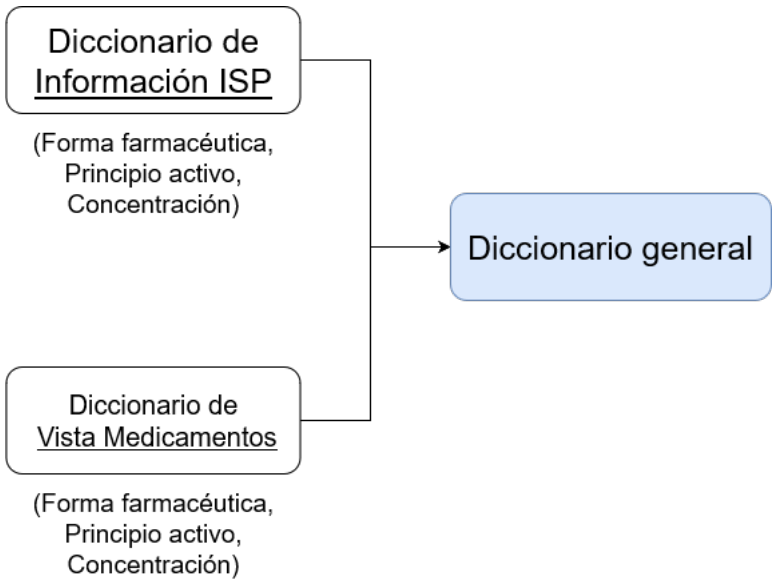


Figura 3.10: Proceso de unión de diccionarios, para la creación del diccionario general.

3.2.4. Creación del algoritmo estructurador de medicamentos

El proceso de creación del algoritmo está compuesto de 3 partes principales:

1. Subproceso de Predicción de Forma farmacéutica.
2. Subproceso de extracción del Principio activo.
3. Subproceso de extracción de la Concentración.

Cada uno de estos subprocesos tiene por objetivo el obtener cada uno de los atributos de una determinada descripción, la cual se encuentra en texto libre. Por ejemplo, para el caso de la descripción **Ranitidina 300 mg comprimidos**, el algoritmo debe ser capaz de detectar determinadas palabras según sea el proceso. En el primer proceso de predicción de Forma Farmacéutica, se espera que el estructurador identifique que la forma farmacéutica del medicamento, según su descripción, sea **comprimidos**, mientras que en el segundo proceso, debe ser capaz de extraer **Ranitidina**. Finalmente, en el último proceso, debe detectar que la Concentración es **300 mg**.

En la Figura 3.11 se presenta un diagrama general del proceso que recorre el algoritmo de estructuración, detallando específicamente los 3 subprocesos importantes del procedimiento. En lo que sigue se profundiza cada una de las etapas ya mencionadas.

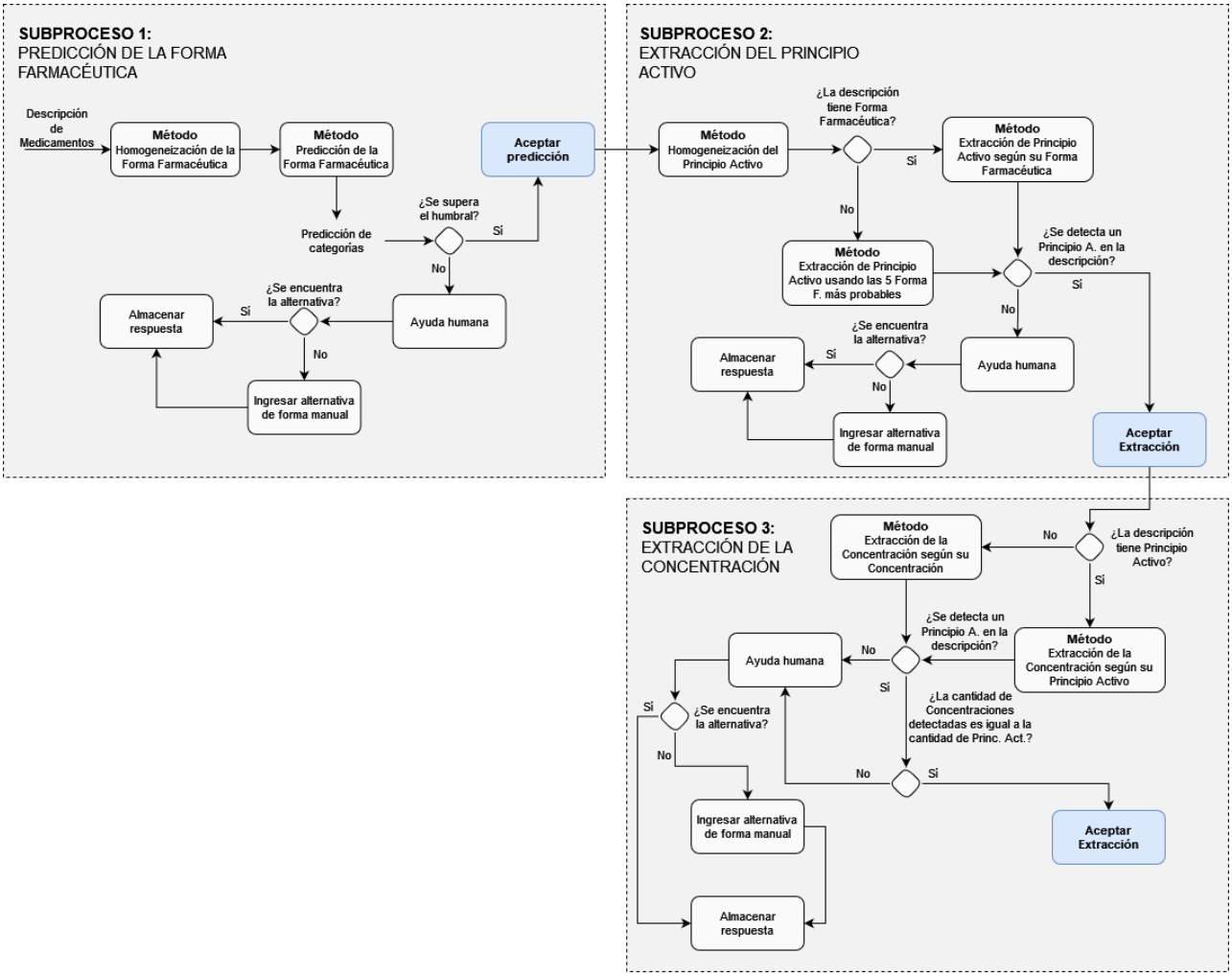


Figura 3.11: Diagrama con los 3 subprocesos del algoritmo de estructuración.

3.2.4.1. Predicción de la Forma farmacéutica

El subproceso de predicción de la forma farmacéutica, el cual se puede observar en la Figura 3.13, corresponde a la primera etapa de la estructuración, y se encarga de la predicción de la forma farmacéutica de las descripciones de medicamentos. El método utilizado para obtener las formas farmacéuticas del texto libre, es la clasificación.

Para escoger el algoritmo de clasificación, se utiliza la técnica de *cross validation*, dividiendo la muestra de datos en 5 grupos. Se aplica la técnica con 4 algoritmos de clasificación multiclase, los cuales son:

1. *Random forest.*
2. *Support vector machine.*
3. *Naïve Bayes.*
4. *Logistic regression.*

En la Figura 3.12, se presenta el promedio de la métrica *accuracy* para cada evaluación, observando que en promedio, el algoritmo *Support vector machine* presenta un mejor desem-

peño, con un *accuracy* promedio de 0.9542.

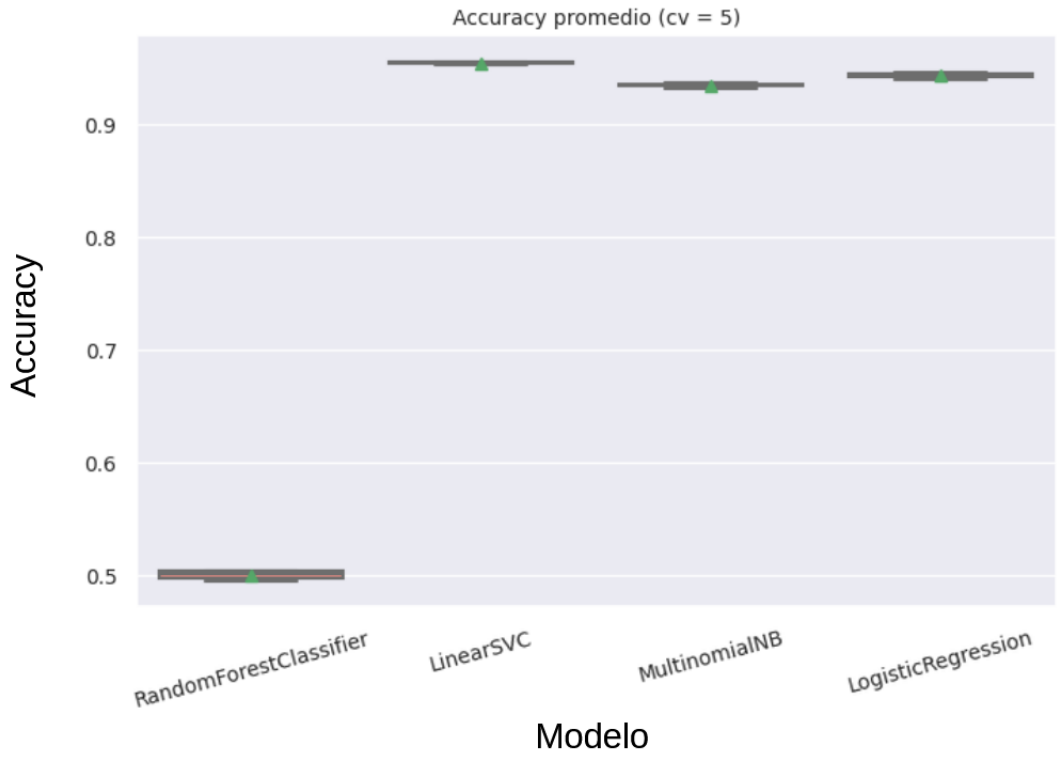


Figura 3.12: Promedio de la métrica *accuracy*, *cross validation* con 5 grupos.

En la Tabla 3.12 se presenta un resumen de los valores promedios de la métrica *accuracy*, para cada uno de los modelos aplicados.

Tabla 3.12: Resumen de *accuracy* promedio para cada modelo utilizado.

Algoritmo de clasificación	***Accuracy* promedio**
Support Vector Machine	0.954267
Logistic Regression	0.943300
Naïve Bayes	0.934483
Random Forest	0.500083

A partir de lo anterior, se utiliza el algoritmo de *Support Vector Machine* (SVM), debido al desempeño presentado. Además, se establece como etiqueta las categorías de la variable "Forma Farmaceutica" de la base Vista Medicamentos.

Se le debe recordar al lector, que las descripciones de los medicamentos están en texto libre, es decir, fueron escritas por un humano, por lo que no están libres de contener errores escriturales que puedan dificultar la detección de etiquetas de la forma farmacéutica. Por lo tanto, antes de realizar la clasificación, se realiza un proceso de "homogeneización de etiquetas", cuyo fin es ayudar al algoritmo de clasificación a detectar las etiquetas en las diversas descripciones de fármacos.

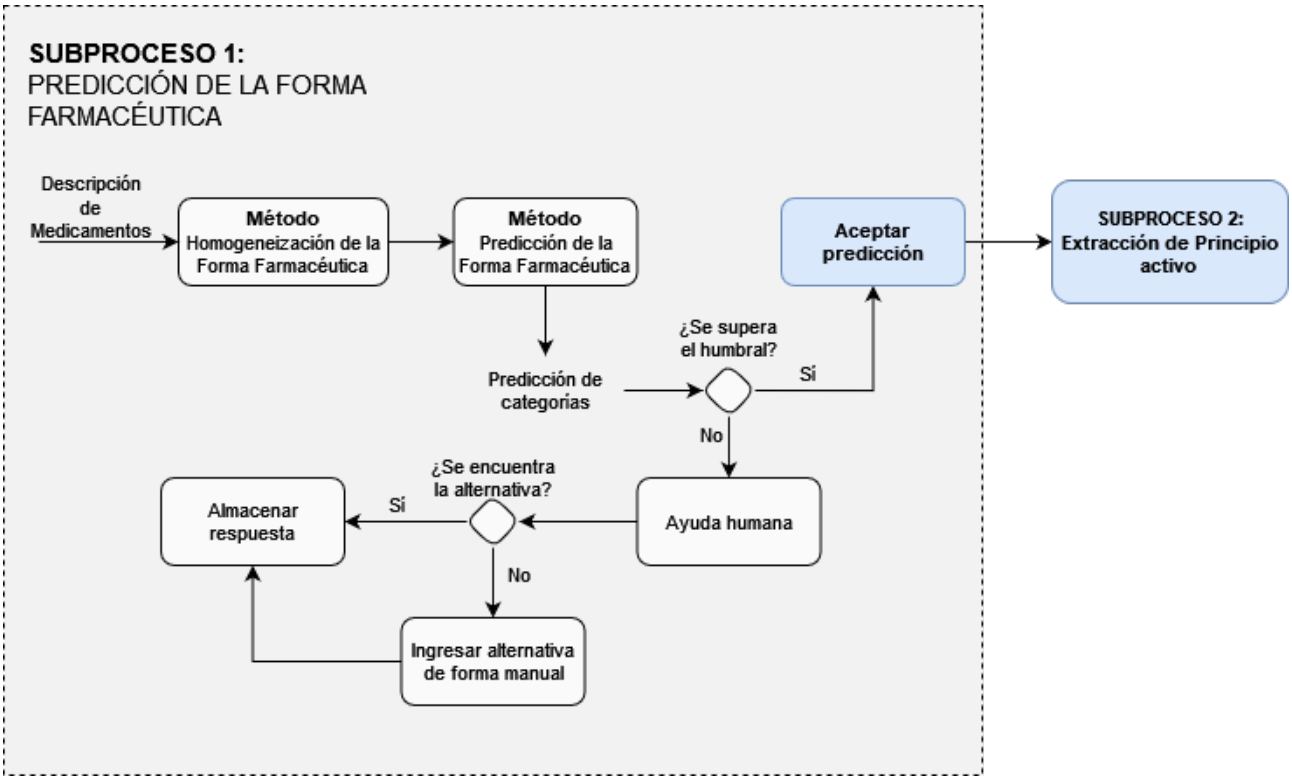


Figura 3.13: Diagrama del Subproceso 1: Predicción de la Forma farmacéutica.

Para lograr lo mencionado anteriormente, se utiliza *Word2Vec*. Este algoritmo es explicado en la sección de **Marco Conceptual**, y su propósito en este subproceso es el de leer los contextos de las palabras en cada descripción, y seleccionar aquellas palabras que son usadas en un mismo ambiente. Lo anterior se realiza con el objetivo de identificar las formas similares de escribir una determinada etiqueta, para que luego estas sean reemplazadas por la etiqueta de interés. Por ejemplo, una de las formas de escribir **ampolla** es **amp**, por lo que al aplicar *Word2Vec*, se debería identificar la similitud de ambas palabras, dado que son utilizadas en el mismo contexto, y, por lo tanto, reemplazar la palabra **amp** por **ampolla**, ayudando así al algoritmo de clasificación a detectar la etiqueta correcta.

En base a lo anterior, se obtienen todas las palabras "similares" a cada una de las etiquetas de la variable "Forma Farmaceutica", es decir, palabras que posiblemente puedan referirse a la original. En la Figura 3.14 se presenta un ejemplo de la obtención de palabras similares a **comprimido**.

```
[('comprmido', 0.756127119064331),
 ('compromido', 0.7549158930778503),
 ('comprido', 0.749249279499054),
 ('comrpimido', 0.7257900834083557),
 ('comprimidos', 0.722469687461853),
 ('comprimdo', 0.7054644823074341),
 ('mgcomprimido', 0.7052221298217773),
 ('cm', 0.7029447555541992),
 ('recubierto', 0.7013177871704102),
 ('cmprimido', 0.6983975172042847),
 ('omprimido', 0.6917062401771545),
 ('compimido', 0.6773164868354797),
 ('compromidos', 0.6747901439666748),
 ('comprimodos', 0.6732450723648071),
 ('capsula', 0.6663558483123779),
 ('cjxcm', 0.6633626818656921),
 ('cmmg', 0.6612696647644043),
 ('comprimodo', 0.6556822061538696),
 ('birranurada', 0.6540371775627136),
 ('comprimdos', 0.6517592668533325)]
```

Figura 3.14: Top 20 de palabras similares a **comprimido**, junto a la similitud de coseno, a partir de la aplicación de *Word2Vec*.

Se puede observar, que *Word2Vec* permite identificar palabras que derivan de la etiqueta **comprimido**. Por ejemplo, la palabra **comprimidos**. Sin embargo, también se logra identificar que se es capaz de reconocer las palabras mal escritas que se refieren a la misma etiqueta, y que no necesariamente son abreviaciones. Ejemplo de lo anterior, en casos de reconocimiento de la etiqueta con errores escriturales, se encuentran las palabras **comprmido**, **comprido**, **comrpimido**, entre otras. Mientras que un ejemplo de detección de abreviaciones, sería la palabra **cm**.

En la Figura 3.14 también se puede observar el puntaje de "similitud de coseno", el cual implica que entre más cercano a 1, los vectores de las palabras son más similares, mientras que un puntaje cercano a 0 denota que no existe similitud de las palabras. Este puntaje es utilizado posteriormente como un parámetro en el algoritmo de clasificación, y se define como:

Definición 3.2.1. Umbral de similitud: Corresponde al valor entregado por la similitud de coseno. Representa una medida de la similitud existente entre dos vectores de palabras, denotando que tan semejantes son.

Si bien el método de *Word2Vec* ayuda a detectar palabras que son utilizadas en los mismos contextos en donde son usadas las formas farmacéuticas, no para todas las formas farmacéuticas se logran detectar las palabras similares, es decir, palabras que sean capaces de ser reemplazadas por una etiqueta de forma farmacéutica. Como ejemplo de lo mencionado, se presenta la Figura 3.15, en donde se observan las palabras que *Word2Vec* detecta que son similares a **spray**. Según *Word2Vec*, las palabras "xylocaina" y "dimecaina" podrían ser reemplazadas por **spray**, lo cual no es correcto, puesto que, si bien son utilizadas en similares oraciones, no hay una completa exactitud de que sea la forma farmacéutica correcta de esa descripción específica, lo que podría desembocar en un modelo de predicción que clasifique de forma equivocada.

```
[('nasal', 0.79347825050354),
 ('iliadin', 0.7233459949493408),
 ('dimecaina', 0.7218639254570007),
 ('furoato', 0.7123663425445557),
 ('oximetazolina', 0.7080665230751038),
 ('xylocaina', 0.7031542062759399),
 ('oximetasolina', 0.7015194892883301),
 ('intranasal', 0.7003747224807739),
 ('xylocaína', 0.689795732498169),
 ('xilocaina', 0.6790933012962341),
 ('oxilin', 0.6778051853179932),
 ('aerosol', 0.6746944189071655),
 ('fluorato', 0.6687718629837036),
 ('fuorato', 0.6629132628440857),
 ('solspray', 0.6576741933822632),
 ('spary', 0.6549530029296875),
 ('solnasal', 0.6511277556419373),
 ('rinoval', 0.650933027267456),
 ('mometasona', 0.6479118466377258),
 ('spr', 0.6472674012184143)]
```

Figura 3.15: Top 20 de palabras similares a **spray**, junto a la similitud de coseno, a partir de la aplicación de *Word2Vec*.

De acuerdo con lo definido anteriormente, se utiliza el "umbral de similitud" con el objetivo de limitar hasta qué palabra conviene reemplazar, a partir del establecimiento de un umbral específico. Esto último permite homogeneizar las etiquetas de formas farmacéuticas.

Tomando en consideración lo planteado en párrafos anteriores, se busca resolver el problema a través de un método de "comparación". Primero, se realiza un entrenamiento inicial del modelo de clasificación SVM, denominado "Modelo base", con el fin de establecer métricas iniciales con las cuales comparar los siguientes resultados. Posteriormente, para cada una de las etiquetas pertenecientes a la columna de "Forma Farmaceutica", se lleva a cabo la homogeneización utilizando *Word2Vec*, es decir, se realiza el reemplazo de todas aquellas palabras que son consideradas similares a la etiqueta de forma farmacéutica estudiada. A continuación, se realiza un nuevo entrenamiento, comparando las nuevas métricas con los valores anteriormente registrados. En caso de que el nuevo entrenamiento del modelo entregue mejores resultados, se actualizan las métricas y la base de datos con el cambio de forma farmacéutica realizado. Sin embargo, si los nuevos resultados son bajos en comparación con las métricas de base, el algoritmo rechaza la sustitución de la etiqueta de "Forma Farmaceutica" y procede a homogeneizar la siguiente etiqueta en la lista. El proceso anteriormente descrito se puede observar en la Figura 3.16.

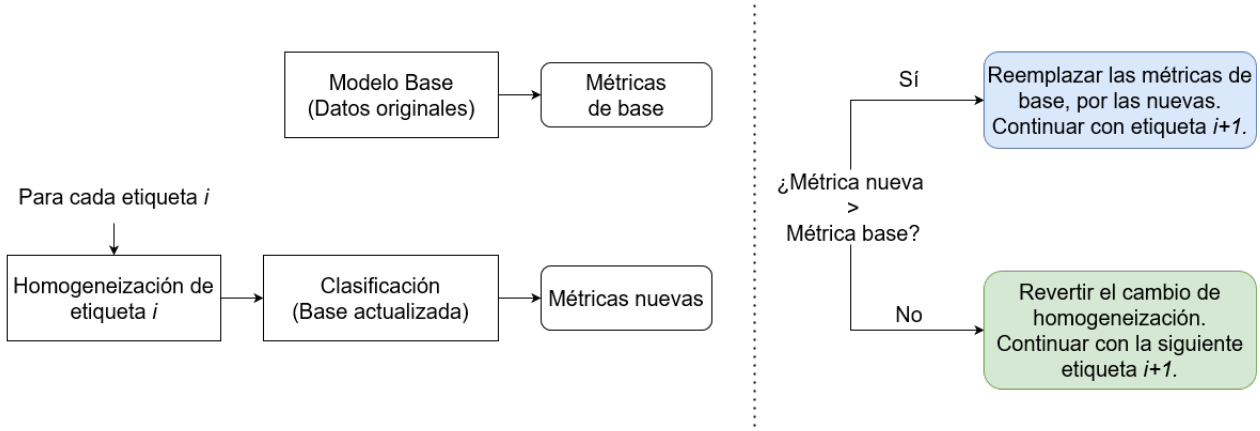


Figura 3.16: Etapa de homogeneización y clasificación.

Respecto al proceso de homogeneización, a modo de ejemplo, en la Tabla 3.13 se presentan algunas comparaciones entre una descripción antes y después de la homogeneización de algunas etiquetas de forma farmacéutica, correspondientes a: **comprimido**, **ampolla**, **frasco** y **gotas**.

Tabla 3.13: Ejemplo de homogeneización de una forma farmacéutica.

Descripción del medicamento original	**Descripción del medicamento post homogeneización**
diazepan 10 mg cm	diazepan 10 mg comprimido
gemcitabina fa 200 mg/10 ml solucion inyectable	gemcitabina fa 200 mg/10 ml solucion ampolla
10 cajas x 100 amp diazepam 10 mg / 2 ml	10 cajas x 100 ampolla diazepam 10 mg / 2 ml
aluminio hidroxido gel 6 % fc 150 ml	aluminio hidroxido gel 6 % frasco 150 ml
tramadol cl orhit 10 mg gtas	tramadol cl orhit 10 mg gotas
viadil comp gts x 15 ml	viadil comp gotas x 15 ml

En lo referente al proceso de clasificación utilizando SVM, se debe mencionar que los datos se dividen en los grupos de entrenamiento y prueba, con un porcentaje de 70 % y 30 % respectivamente. Se utiliza además TF-IDF, algoritmo explicado en la sección **Marco Conceptual**, para transformar el texto libre de los medicamentos de las bases de entrenamiento y prueba, en una representación significativa de números, la cual puede ser utilizada en el algoritmo para realizar la predicción de etiquetas. Por otro lado, se debe mencionar que los parámetros utilizados por el algoritmo son los establecidos por defecto en la librería de `sklearn`, por lo que utilizar herramientas de búsqueda de parámetros óptimos, puede ser un foco de trabajo futuro para mejorar el algoritmo estructurador.

Es necesario mencionar que durante el proceso de predicción de la forma farmacéutica, se utilizan las probabilidades de cada clase. Esto último implica que al momento de clasificar una determinada descripción de medicamento, en las 15 etiquetas de forma farmacéutica, además de extraer la probabilidad con la que la etiqueta ganadora es seleccionada, también se extrae la información de la probabilidad de las etiquetas restantes. Por ejemplo, al predecir la clase de forma farmacéutica para la descripción “ciprofloxacino comp 500 mg”, la etiqueta predicha corresponde a “comprimido”, con una probabilidad de 0,98283, es decir, la descripción tiene una probabilidad de 98,2 % de ser clasificada en la clase comprimido. Esto último también implica que la forma farmacéutica “comprimido”, tiene una probabilidad de

98,2 % de que se encuentre en la descripción de medicamento. Por otro lado, también se mantienen las probabilidades de que las demás etiquetas de forma farmacéutica estén presentes en el texto de medicamento. A partir de lo anterior, se establece un parámetro encargado de aceptar o rechazar las predicciones realizadas por el modelo de clasificación. Este parámetro se define como:

Definición 3.2.2. Umbral de probabilidad: Corresponde a un parámetro que determina la probabilidad mínima que debe tener una etiqueta para pertenecer a una descripción.

En la Tabla 3.14, se presentan descripciones de medicamentos, en conjunto con las 5 etiquetas de forma farmacéutica más probables, para un determinado fármaco.

Tabla 3.14: Ejemplos de descripciones de medicamentos con sus 5 etiquetas de forma farmacéutica más probables.

Descripción del medicamento	**Etiqueta**	**Probabilidad**
acido ascorbico cm 100 mg	comprimido	0,9879
	jarabe	0,0071
	supositorio	0,0035
	gotas	0,0012
	parche	0,00026
penicilina sodica 1.000.000 fco amp	ampolla	0,9202
	frasco	0,0207
	tubo	0,0078
	jarabe	0,0077
	comprimido	0.0072
dermabiotico ungto 15g	tubo	0,3913
	comprimido	0,2723
	ampolla	0,1699
	frasco	0,0326
	sachet	0,0268

Como se observa en el primer ejemplo de la tabla anterior, la etiqueta de forma farmacéutica más probable corresponde a **comprimido**, es decir, al momento de realizar las predicciones de forma farmacéutica, a la descripción "acido ascorbico cm 100 mg" se le sería asignada la etiqueta de comprimido. Se pueden observar también el resto de las etiquetas posibles, ordenadas a partir de sus probabilidades de pertenecer a la descripción. También se observa en la tabla un segundo ejemplo, en el cual la etiqueta de forma farmacéutica más probable para la descripción "penicilina sodica 1.000.000 fco amp" corresponde a **ampolla**. Si bien el porcentaje de la etiqueta más probable en esta descripción de medicamento, no es tan elevada como en el primer ejemplo, si se observa una clara preferencia por una etiqueta de forma farmacéutica específica. A pesar de que en ambos ejemplos mencionados se tiene una etiqueta destacable por su alta probabilidad, no en todos los casos se presenta el mismo escenario. Con el objetivo de comparar ambas circunstancias, se observa el último ejemplo

en la Tabla 3.14, en donde la mayor probabilidad correspondiente a 0,3913 establece que la etiqueta del atributo de la descripción es **tubo**, lo cual es correcto, dado que el algoritmo reconoce que esta etiqueta está relacionada con el término "ungto" (ungüento).

A partir de los ejemplos presentados, se puede concluir que hay situaciones en las cuales, para una determinada descripción de medicamento, existe una etiqueta predominante en base a su probabilidad. Sin embargo, si bien en la mayoría de las ocasiones hay una gran brecha entre la primera y segunda etiqueta con mayor probabilidad, esta probabilidad no siempre es lo suficientemente alta como para determinar si una descripción de medicamento tiene o no una forma farmacéutica específica. Debido a lo anterior, se utiliza el "umbral de probabilidad". Por defecto en el algoritmo de estructuración este parámetro está fijado en **0,9**, es decir, la etiqueta seleccionada tiene una probabilidad del 90 % de estar presente en un medicamento.

Finalmente, como convergencia de lo mencionado anteriormente, el umbral de probabilidad es importante a la hora de predecir la forma farmacéutica, para un determinado medicamento, puesto que toda etiqueta con una probabilidad sobre 0,9, es aceptada como predicción de forma farmacéutica de un fármaco. Por otro lado, aquellas etiquetas con una probabilidad menor son descartadas, y en su reemplazo el algoritmo genera una lista de las 5 etiquetas con mayores probabilidades de pertenecer a una determinada descripción. Acción que es relevante para la siguiente etapa de supervisión humana en este subproceso.

3.2.4.1.1. Ayuda humana en el subproceso

Con el objetivo de entregar mejores resultados a la hora de estructurar los medicamentos, se añade un proceso de supervisión humana. Este proceso utiliza las respuestas generadas por el algoritmo, y entrega alternativas a una persona que esté especializada en conceptos de medicamentos, para que seleccione la alternativa de forma farmacéutica que más probabilidad tenga de pertenecer a la descripción.

Este proceso recoge los resultados del proceso de clasificación anterior, y utiliza el umbral establecido, para filtrar aquellas etiquetas que tienen baja probabilidad de representar al medicamento. Posteriormente, en aquellos casos en los que la probabilidad no cumple el mínimo establecido, se utiliza la lista de las K=7 etiquetas con mayores probabilidades de pertenecer a la descripción de medicamento, con el fin de entregarlas como alternativas a la persona con conocimientos en el área.

En la Figura 3.17 se presenta un ejemplo de cómo se realizan las preguntas, en los casos en los que no se detecta una etiqueta con una alta probabilidad.

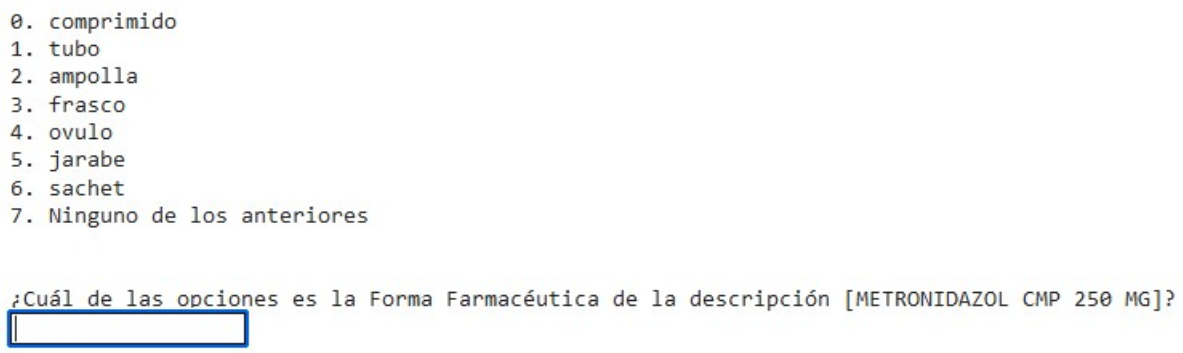


Figura 3.17: Ejemplo de supervisión humana en el subproceso.

Por otro lado, también se consideran aquellos casos en los que una determinada descripción no tiene una forma farmacéutica. En estos escenarios, se entrega la alternativa al humano de ingresar su propia respuesta, puesto que en ocasiones un medicamento no tiene una forma farmacéutica identificada. En caso de que no se pueda identificar la etiqueta correcta, se da la opción de ingresar un valor predeterminado llamado `NA`, el cual representa la ausencia de la información. En la Figura 3.18 se ejemplifica este escenario.

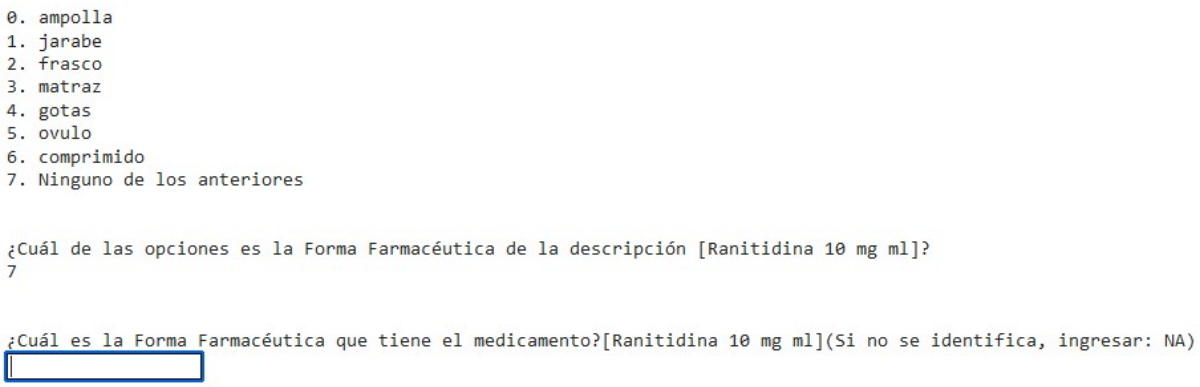


Figura 3.18: Ejemplo de supervisión humana en el subproceso, segundo escenario.

3.2.4.2. Extracción del Principio activo

Corresponde a la segunda etapa en el proceso de estructuración, y se encarga de la extracción del principio activo de las descripciones de medicamentos. A diferencia del primer subproceso, para la extracción de principios activos no se utiliza la clasificación, puesto que como se presenta en la sección **Análisis de Datos**, esta columna presenta 1318 etiquetas, lo que hace difícil la aplicación del método de clasificación. Además, a diferencia del primer subproceso, en esta etapa se realiza una extracción literal del valor de atributo desde la descripción del medicamento, utilizando una técnica de similitud entre palabras, y el diccionario creado con los valores previamente estructurados de la base Vista Medicamentos. Este último proceso se menciona en la sección **Creación de los diccionarios**.

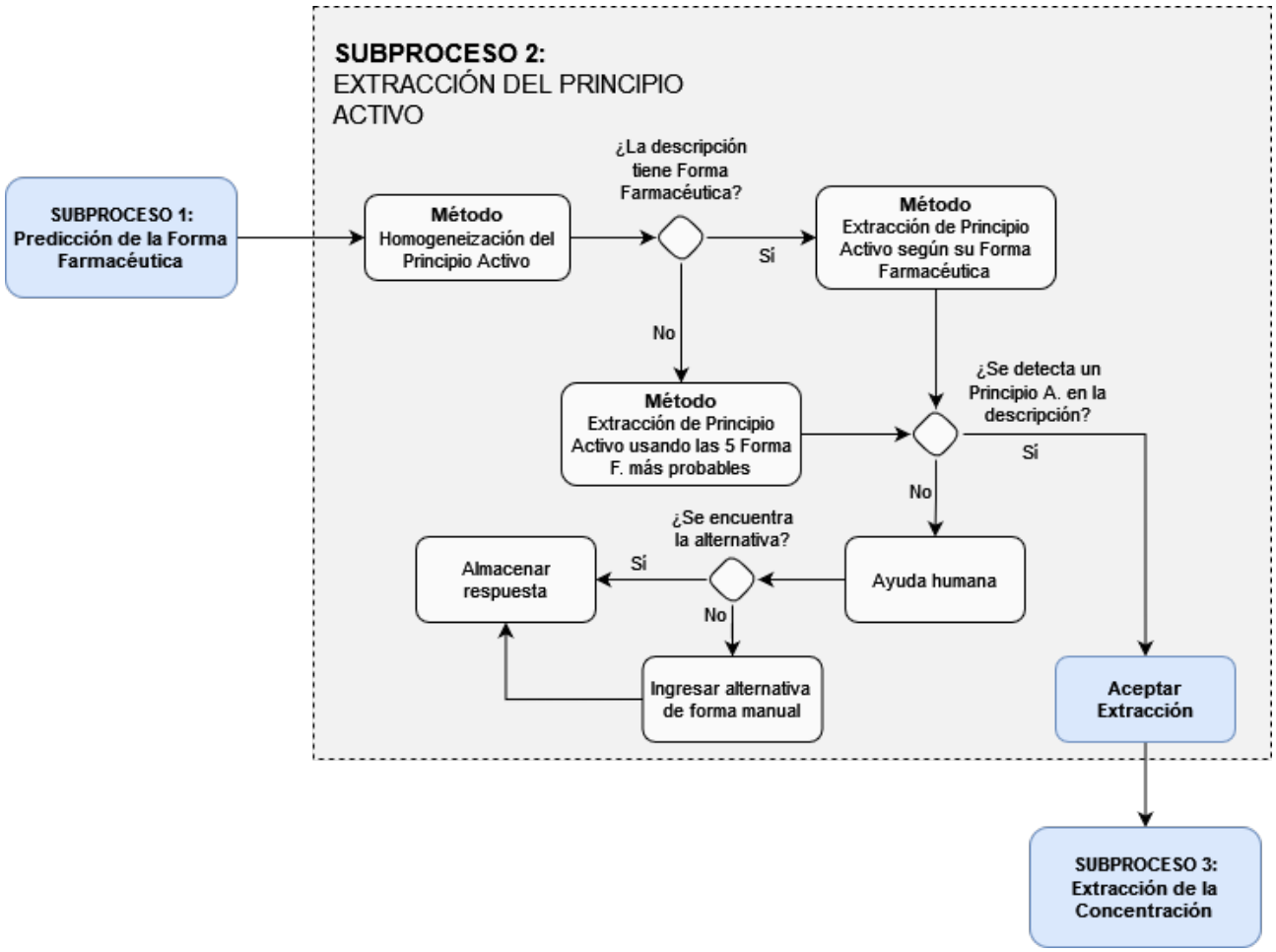


Figura 3.19: Diagrama del Subproceso 2: Extracción del Principio activo.

Al igual que el primer subproceso, en la etapa de extracción del principio activo también existe una fase de homogeneización, aunque diferente a la realizada en el subproceso de predicción de la forma farmacéutica. Lo anterior se debe a las características que presentan los datos de la columna "Principio Activo". Dentro del grupo de los principios activos, existen algunos de estos que están presentes con un nombre alternativo en la descripción del medicamento, valores de atributo que difieren de los presentes en la columna de "Principio Activo", los cuales fueron utilizados para la creación del diccionario. Dado que el algoritmo realiza una extracción textual en base a los principios activos presentes en el diccionario, al momento de extraer el valor de atributo de ciertas descripciones se podrían generar valores nulos, puesto que no se identificarían principios activos que estén presentes en el diccionario, implicando así la generación de una necesidad de homogeneización de las etiquetas.

Primeramente, el algoritmo realiza un reemplazo de algunos nombres alternativos, con el objetivo de homogeneizar los valores. En la Tabla 3.15 se presentan algunos ejemplos de sustitución entre un nombre alternativo y el nombre del principio activo correcto.

Tabla 3.15: Principios activos y sus nombres alternativos presentes en las descripciones.

Principio activo	**Nombre alternativo**
levofloxacino	auxxil
levotiroxina	eutirox
acenocumarol	neo sintrom
zidovudina	retrovir
paracetamol	supracalm
vitamina c	acido ascorbico
vitamina k	fitomenadiona

Posteriormente, ocurre un proceso de "similitud" entre textos del diccionario creado en la sección de **Creación de los diccionarios** y las descripciones de los medicamentos. Para este último proceso se utiliza la distancia de *Levenshtein*, presentada en la sección del **Marco Conceptual**, la cual permite generar una medida de semejanza a partir de la distancia entre cada descripción de medicamento y cada uno de los principios activos presentes en el diccionario, los cuales se comparan de forma individual a cada descripción. El algoritmo retorna aquel principio activo con menor distancia a la descripción, es decir, entrega el principio activo con mayor similitud a la descripción del fármaco. Este proceso se encuentra ejemplificado en la Tabla 3.16, en donde se presentan los principios activos más cercanos a la descripción, destacando **amoxicilina** como la etiqueta correcta. Todo lo descrito anteriormente implica que el valor de atributo encontrado tenga una alta probabilidad de estar presente en la descripción del medicamento, puesto que si no se encuentra, la distancia tiende a aumentar.

Tabla 3.16: Ejemplo de descripción de medicamento con sus 5 etiquetas de principio activo más cercanas a la descripción.

Descripción del medicamento	**Etiqueta**	**Distancia**
antibiotico amoxicilina 500mg. (vencimiento sup. 1 año.)	amoxicilina	0
	amox	2
	ampicilina	2
	cloxacilina	3
	doxiciclina	3

Es necesario mencionar que al igual que el subproceso de predicción de la forma farmacéutica, este subproceso también posee un parámetro que permite filtrar aquellos valores de atributos útiles. Este parámetro establece una distancia de similitud máxima entre un principio activo y una descripción del medicamento, puesto que entre mayor sea la distancia, más diferentes serán los textos, disminuyendo así la probabilidad de que la etiqueta de principio activo sea la correcta. A partir de lo anterior, todos aquellos valores de atributo que sean cercanos a la descripción, es decir, que tengan una distancia de 0 o 1, serán aceptados por el algoritmo. Por otro lado, para aquellas descripciones que poseen distancias mayores, es decir, sin principios activos, se genera una lista de los N principios activos con menores distancias, la cual es utilizada en la etapa de supervisión humana de este subproceso.

3.2.4.2.1. Supervisión humana en el subproceso

Con el objetivo de entregar mejores resultados durante el proceso de estructuración, se añade una etapa de supervisión humana en el subproceso. A partir de la "lista de los N principios activos con menores distancias" generada en el proceso anterior, se entrega la opción de completar aquellos valores de atributo que el algoritmo no logra detectar. Al igual que la etapa de supervisión humana en el subproceso de predicción de la forma farmacéutica, se requiere la intervención de una persona con conocimientos del área de medicamentos.

En la Figura 3.20 se presenta un ejemplo de ayuda humana en la extracción del principio activo, entregando 10 alternativas, derivadas de las distancias. Por otro lado, también se entrega la opción de ingresar la respuesta de forma manual, si se presenta el escenario en el cual la respuesta correcta no se encuentra dentro de las opciones presentadas. Finalmente, en caso de que el valor de atributo no se encuentre en la descripción, se presenta la opción de ingresar un valor nulo.

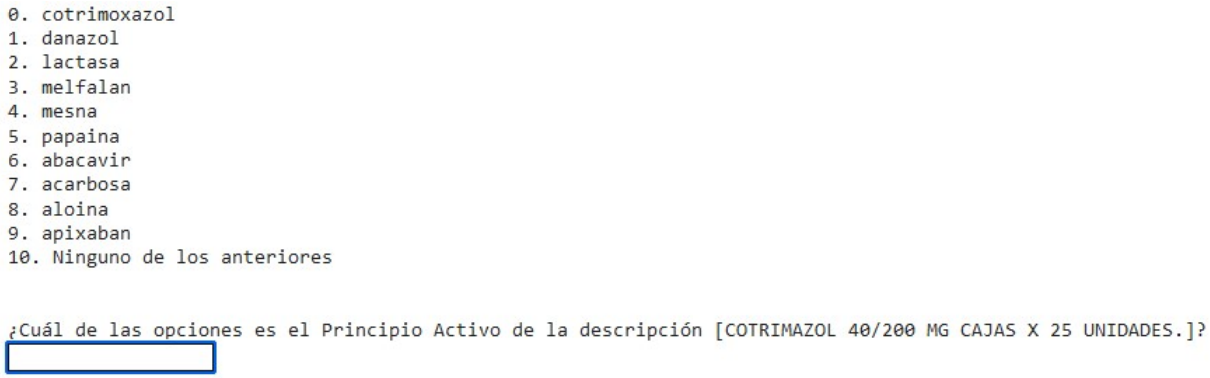


Figura 3.20: Ejemplo de supervisión humana en el subproceso.

3.2.4.3. Extracción de la Concentración

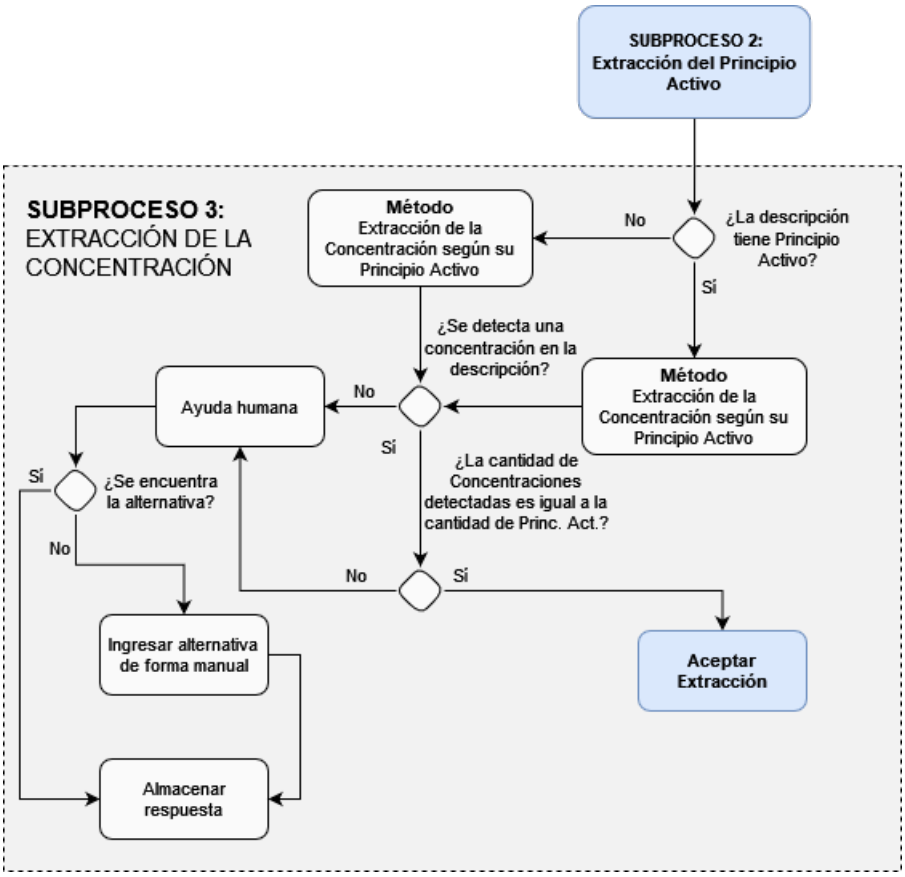


Figura 3.21: Diagrama del Subproceso 3: Extracción de la Concentración.

Corresponde a la tercera y última parte del proceso de estructuración de medicamentos. Tiene por objetivo la extracción de las concentraciones desde las descripciones de los fármacos. Respecto a la herramienta utilizada para realizar el proceso, esta corresponde a las expresiones regulares, las cuales son mencionadas en la sección de **Marco Conceptual**.

A diferencia de los dos primeros subprocesos, durante esta etapa no existe una homogeneización de los valores de atributos, por lo tanto, corresponde a un proceso más conciso que los anteriores.

Es imperante recordar que la concentración está compuesta por dos valores: el número y su unidad de medida. A partir de lo anterior, se utiliza una serie de expresiones regulares con el objetivo de extraer la concentración, a partir de su respectiva unidad de medida. Por ejemplo, se extrae el número **500** junto a la unidad **mg**, conformando finalmente la concentración **500 mg**. En el Código 3.1 se presenta una de las expresiones regulares utilizadas para extraer las concentraciones de las descripciones.

Código 3.1: Ejemplo de una expresión regular para extraer la concentración.

```
units = ['mgs' , 'mg', 'ml' , 'mcg', 'grs' ,'gr',  'g' , 'meq', 'cc', 'ui', '\%' ]
regex = '|'.join(units)
text = re.findall(r'(?<!\d|\.)\d+(?:\.\d+)?\s*?(?:' + regex + ')(?!\w)', text)
```

3.2.4.3.1. Supervisión humana en el subproceso

A diferencia de los dos subprocesos anteriores, durante la etapa de extracción de la concentración no se generan listas de valores posibles, en los casos en los cuales el algoritmo no identifique el valor de atributo correcto. Debido a lo anterior, durante el procedimiento de la supervisión humana se realiza la tarea de verificación si la extracción realizada está correcta.

El proceso de verificación implica la realización de algunas tareas como:

- Comprobar si la cantidad de concentraciones extraídas es igual a la cantidad de principios activos en el fármaco.
- En casos en los cuales existan menos concentraciones que principios activos, identificar cuál es la concentración faltante.
- En los casos en los cuales no se identifica la concentración, entregar el valor identificado.

En la Figura 3.22 se presenta un ejemplo de la supervisión humana en este subproceso.

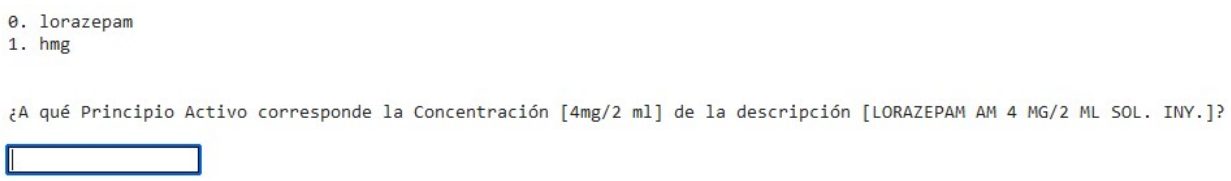


Figura 3.22: Ejemplo de supervisión humana en el subproceso.

3.2.5. Métricas de evaluación

Durante el proceso de estructuración de los medicamentos, se utiliza una serie de métricas que permiten tanto la búsqueda de mejores resultados, como la visualización de la calidad y efectividad del algoritmo de estructuración. Dado esto último, se establecen diversas métricas dependiendo del subproceso, y también métricas que ayudan a evaluar el desempeño del estructurador de forma general.

A partir de lo mencionado anteriormente, se dividen los tipos de métricas según sea el proceso:

- **Predicción de forma farmacéutica:** Debido al método de clasificación utilizado en el subproceso, se establecen métricas usuales que ayudan a identificar la evolución del desempeño del modelo, a partir de las modificaciones del umbral de similitud, el cual ayuda en el proceso de homogeneización. Debido a esto último, se utilizan las siguientes métricas para observar ka evolución del modelo de clasificación:

- ***Accuracy:*** Entrega un porcentaje de los casos en los cuales el modelo ha predicho los valores de forma correcta.

  Se calcula como:

$$Accuracy = \frac{Predicciones\ correctas}{Total\ de\ predicciones}$$

- ***Precision:*** Entrega un valor que representa la fracción de predicciones pertenecientes a la clase positiva, que son verdaderamente positivas, dentro de una etiqueta.

  Para cada clase, se calcula como:

$$Precision = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos}$$

  En donde los *Verdaderos positivos* corresponden a aquellas ocasiones en las que el algoritmo correctamente predice que una descripción tiene una determinada forma farmacéutica. Mientras que los *Falsos positivos* corresponden a los casos en los cuales el algoritmo predice que una descripción tiene una forma farmacéutica igual a la clase que se está analizando, cuando esta en verdad posee otra forma farmacéutica.

- ***Recall:*** Señala la capacidad del clasificador para detectar clases positivas. Entrega un valor que representa la fracción de predicciones positivas de la clase, que en verdad son positivas.

  Para cada clase, se calcula como:

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos}$$

  En donde los *Verdaderos positivos* corresponden a aquellas ocasiones en las que el algoritmo correctamente predice que una descripción tiene una determinada forma farmacéutica. Mientras que los *Falsos negativos* corresponden a los casos en los cuales el algoritmo predice que una descripción tiene una forma farmacéutica distinta a la que se está analizando.

- ***F1-Score:*** Presenta un valor que combina la *precision* y la *recall*, entregando un puntaje balanceado entre dos métricas importantes.

  Se calcula como:

$$F1 - Score = \frac{2(P * R)}{P + R}$$

Es necesario mencionar que la métrica establecida por defecto corresponde a *F1-Score*, debido a su capacidad de balancear dos métricas relevantes: *Recall* y *Precision*. La medida de desempeño seleccionada ayuda a evaluar cada uno de los modelos generados posterior a un reemplazo de etiqueta, procedimiento que permite la homogeneización. Por otro lado, debido a que las métricas de *recall* y *precision* son calculadas para cada una de las 21 etiquetas del atributo "Forma Farmaceutica", se utiliza el promedio para evaluar si la homogeneización de una etiqueta es conveniente para el modelo.

Debido a que las métricas mencionadas anteriormente solo muestra el desempeño del modelo con el 30 % de los datos, correspondiente a los datos de test, se define una métrica general para evaluar la capacidad del modelo de predecir correctamente la "Forma Farmaceutica" para todos los datos utilizados, la cual se define como precisión.

$$Precisión = \frac{Cantidad\ de\ etiquetas\ correctas}{Total\ de\ etiquetas}$$

- **Extracción del principio activo:** A diferencia del procedimiento anterior, para el subproceso de extracción del principio activo solo se utiliza una métrica: la precisión.

  Esta métrica se define como:

  $$Precisión = \frac{Extracciones\ correctas\ de\ principio\ activo}{Total\ de\ extracciones}$$

  En donde *Extracciones correctas de principio activo* corresponde a las etiquetas de principio activo encontradas por el algoritmo, que coinciden con las etiquetas originales presentes en la base Vista Medicamentos. Por otro lado, *Total de extracciones* corresponde al total de descripciones presentes en la base de datos, sobre la cual se aplica el estructurador.

- **Extracción de la concentración:** Al igual que el procedimiento anterior, para el subproceso de extracción de la concentración, solo se utiliza una métrica: la precisión.

  Esta métrica está definida como:

  $$Precisión = \frac{Extracciones\ correctas\ de\ la\ concentración}{Total\ de\ extracciones}$$

  En donde *Extracciones correctas de la concentración* corresponde a las etiquetas de concentración que son encontradas por el algoritmo, y que coinciden con las etiquetas originales que se encuentran en la base utilizada. Por otro lado, *Total de extracciones* corresponde al total de descripciones presentes en la base de datos, sobre la cual se aplica el algoritmo.

- **Algoritmo de estructuración:** Con el fin de evaluar el desempeño del algoritmo estructurador de medicamentos, se construyen dos métricas de precisión, las cuales son definidas a continuación.

  - **Métrica 1:** Señala la cantidad de etiquetas correctas totales sobre la cantidad total de descripciones, es decir, entrega una razón de las etiquetas extraídas de forma correcta de la descripción, utilizando el estructurador.

    $$Precisión = \frac{Cantidad\ de\ etiquetas\ correctas}{Total\ de\ etiquetas}$$

  - **Métrica 2:** Presenta la cantidad promedio de etiquetas correctas dentro de una descripción de medicamento, es decir, señala el número de atributos cuyos valores coinciden con la etiqueta original.

$$Precisión\ promedio = \frac{\sum coincidencias\ de\ etiquetas}{Total\ de\ descripciones}$$

En donde *coincidencias de etiquetas* se define como la cantidad de etiquetas que coinciden dentro de una misma descripción, es decir, señala la cantidad de coincidencias entre los valores de atributo extraídos por el algoritmo, y los valores originales.

Para cada atributo i se define:

$$Coincidencia_i = \begin{cases} 1, & \text{Si Etiqueta extraída por el algoritmo = Etiqueta original} \\ 0, & \text{Si no lo es} \end{cases}$$

Donde $i \in \{\text{forma farmacéutica}, \text{principio activo}, \text{concentración}\}$.

Finalmente, a modo de ejemplo para las *coincidencias de etiquetas*, para la descripción AMPICILINA + SULBACTAM/UNASYN 1.5 G FCO AMP, se observa en la Tabla 3.17 y Tabla 3.18, que tanto para la etiqueta de forma farmacéutica como para el principio activo hay una coincidencia de etiquetas, mientras que en el caso de la concentración, observado en la Tabla 3.19, no hay coincidencia, como se puede ver a continuación.

Tabla 3.17: Ejemplo de coincidencia de etiquetas, subproceso de predicción de la forma farmacéutica.

Etiqueta de forma farmacéutica original	**Etiqueta de forma farmacéutica predicha**	**Coincidencia**
ampolla	ampolla	1

Tabla 3.18: Ejemplo de coincidencia de etiquetas, subproceso de extracción del principio activo.

Etiqueta de principio activo original	**Etiqueta de principio activo extraída**	**Coincidencia**
ampicilina + sulbactam	ampicilina + sulbactam	1

Tabla 3.19: Ejemplo de coincidencia de etiquetas, subproceso de extracción de la concentración.

Etiqueta de concentración original	**Etiqueta de concentración extraída**	**Coincidencia**
1.5 g	1000/500 mg	0

Por lo tanto, se obtiene que 2 de 3 atributos tienen etiquetas correctas, según la etiqueta original, lo que implica que la cantidad de "coincidencias de etiquetas" en este ejemplo sería 2. Para cada descripción de medicamento se calcula este parámetro, lo que permite posteriormente el cálculo del promedio de la precisión, la que señala el promedio de etiquetas correctas de una descripción.

Capítulo 4

Resultados y Discusión

En el presente capítulo se entregan los resultados obtenidos posterior a la aplicación del algoritmo de estructuración, aplicado en un entorno ofrecido por `Google Colaboratory`, el cual ofrece una `memoria RAM` de `12 GB` y un `disco virtual` de `100 GB`. Los resultados a presentar incluyen el tiempo de ejecución de cada subproceso y sus respectivas métricas, además de presentar el desempeño del estructurador a nivel general. Se destacan también los parámetros que influyen en la calidad de los resultados, presentando una comparación entre estos.

Para cada uno de los subprocesos, se presentan los resultados de métricas y tiempos de ejecución, comparando también estos mismos con los obtenidos al cambiar ciertos parámetros, además de su correspondiente análisis en torno a los resultados obtenidos. Es necesario mencionar que no se utilizan todos los datos de la base Vista Medicamentos, dada la cantidad de recursos que estos requieren para ser analizados, por lo tanto, el parámetro de N datos representa el tamaño de la base utilizada, y es relevante a la hora de entregar los resultados.

4.1. Subproceso de predicción de la Forma farmacéutica

En las siguientes tablas se presentan los resultados de este subproceso, dividiendo cada tabla según la cantidad de datos de la base a utilizar, y presentando para cada set de datos, sus correspondientes métricas a partir del cambio de un determinado parámetro.

Respecto a los parámetros utilizados para la obtención de resultados, estos corresponden a:

- Número de datos: Parámetro que señala la cantidad de descripciones a estructurar.
- Umbral de similitud (entre palabras): Parámetro utilizado en la homogeneización de etiquetas.
- Umbral de probabilidad: Parámetro utilizado para descartar ciertas predicciones que no tienen probabilidades altas de pertenecer a una descripción de medicamentos.

En la Tabla 4.1 se presentan los resultados obtenidos para un set de datos de tamaño N = 5000, con un umbral de probabilidad de 0,85 y un umbral de similitud entre 0,7 y 0,95.

Tabla 4.1: Resultados de estructuración del atributo Forma farmacéutica, con N = 5000, umbral de probabilidad de 0,85 y umbral de similitud variante entre 0,7 y 0,95.

N = 5000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,7	66,633	0,890	0,890	0,891	0,885	0,86654
Umbral = 0,8	68,239	0,9	0,9	0,896	0,894	0,87495
Umbral = 0,85	65,531	0,890	0,890	0,891	0,883	0,88113
Umbral = 0,9	67,103	0,892	0,892	0,894	0,885	0,85859
Umbral = 0,95	64,676	0,897	0,897	0,898	0,890	0,85814

En la Tabla 4.2 se presentan los resultados obtenidos para un set de datos de tamaño N = 100000, con un umbral de probabilidad de 0,85 y un umbral de similitud entre 0,7 y 0,95.

Tabla 4.2: Resultados de estructuración del atributo Forma farmacéutica, con N = 100000, umbral de probabilidad de 0,85 y umbral de similitud variante entre 0,7 y 0,95.

N = 100000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,7	4072,609	0,967	0,960	0,961	0,962	0,950902
Umbral = 0,8	3956,945	0,962	0,963	0,957	0,954	0,95695
Umbral = 0,85	3850,375	0,953	0,959	0,96	0,965	0,95466
Umbral = 0,9	4159,012	0,966	0,956	0,97	0,951	0,95104
Umbral = 0,95	4004,023	0,962	0,962	0,961	0,966	0,95054

Para observar el resto de tablas que resumen los resultados según número de datos, se recomienda al lector revisar el Anexo B.

A partir de los resultados presentados en las Tablas 4.1, 4.2, B.1,B.2 y B.3, se observa que a medida que aumenta el N de los datos utilizados, hay una clara tendencia del modelo a tener mejores resultados, así como también aumenta el tiempo de ejecución, pasando de un mínimo de 65,53 segundos a un máximo de 4159,01 segundos.

Respecto a los resultados entregados por las métricas *Accuracy, Recall, Precision* y *F1-Score*, se observa que existe un aumento de los valores de estas a medida que aumenta el tamaño del set de datos. Sin embargo, no hay mayor variación de los valores, o una preferencia clara por un valor al momento de variar el umbral de similitud, puesto que las métricas tienen específicos parámetros que generan altos valores, pero que no se mantienen para el siguiente análisis, evitando así fijar una conclusión específica. Sin embargo, lo anterior podría implicar una independencia del parámetro en el modelo, al momento en que este varía.

Pese a que las métricas anteriores no presentan una tendencia específica o grandes diferencias entre el cambio de un umbral a otro, sí se puede observar que la métrica precisión muestra una leve inclinación hacia un parámetro de umbral específico.

En la Figura 4.1 se observa que a pesar de que no hay una gran variación entre la precisión de un umbral y otro, sí se destaca que en el umbral = 0,85 hay un breve aumento de los valores, para los 5 escenarios presentados. A partir de lo anterior, se destaca que para el escenario de N = 100000, umbral de similitud = 0,85 y umbral de probabilidad = 0,85, un 95,47 % de las

descripciones de fármacos poseen una forma farmacéutica correcta. Por otro lado, se debe recordar que el desempeño de las primeras 4 métricas muestra el desempeño del modelo solo en el 30 % de los datos utilizados y del umbral de similitud, mientras que la métrica de precisión depende de la base utilizada en su totalidad, y del umbral de probabilidad, representando un valor final de que tan bien se estructuran de forma general, las descripciones de los datos en el primer subproceso. Lo anterior puede generar una mayor importancia sobre esta métrica, destacando que esta sí posee una preferencia por un valor de umbral específico.

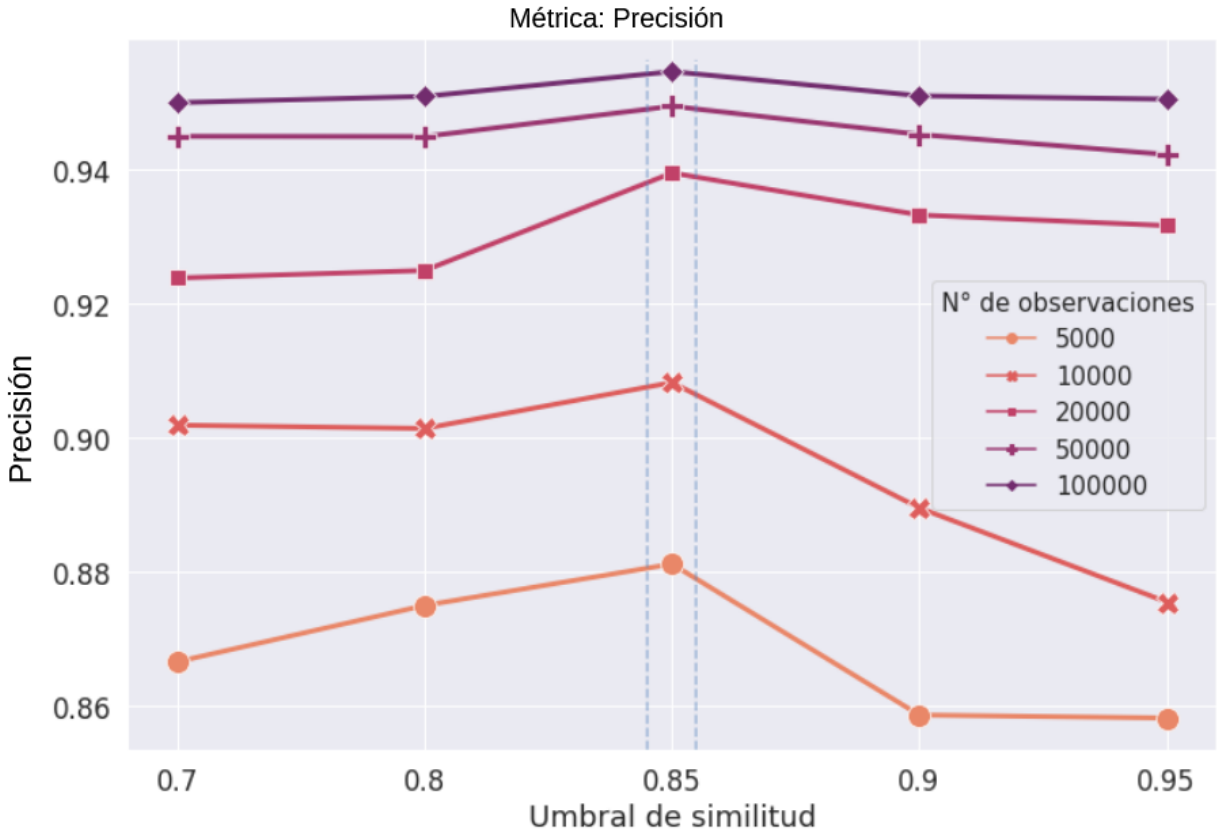


Figura 4.1: Desempeño de la métrica **precisión** durante el proceso.

A partir de lo anterior, se establece que el mejor parámetro de umbral de similitud corresponde a 0,85, puesto que, si bien los resultados de la mayoría de las métricas no son muy favorables a este valor, sí genera mejores resultados que el resto de umbrales analizados. Esto último se debe a la métrica de precisión, la cual entrega una mejor representación del desempeño del algoritmo en esta etapa del estructurador.

Posteriormente, fijando el parámetro **umbral de similitud = 0,85**, se analiza el siguiente parámetro, el correspondiente al umbral de probabilidad, encargado de decidir si una etiqueta es aceptada o no.

En la Tabla 4.3 se presentan los resultados obtenidos para un set de datos de tamaño N = 5000, con un umbral de similitud de 0,85 y un umbral de probabilidad entre 0,6 y 0,95

Tabla 4.3: Resultados de estructuración del atributo Forma farmacéutica, con N = 5000, umbral de similitud de 0,85 y umbral de probabiliad variante entre 0,6 y 0,95.

N = 5000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,6	71,611	0,74	0,74	0,68187	0,80	0,81164
Umbral = 0,7	63,625	0,74	0,74	0,73272	0,70067	0,74644
Umbral = 0,8	63,771	0,75333	0,75333	0,74335	0,71287	0,71587
Umbral = 0,85	64,768	0,74	0,74	0,73403	0,69524	0,68824
Umbral = 0,9	64,734	0,76	0,76	0,74046	0,71873	0,64989
Umbral = 0,95	65,044	0,78	0,78	0,75828	0,73302	0,60611

En la Tabla 4.4 se presentan los resultados obtenidos para un set de datos de tamaño N = 100000, con un umbral de similitud de 0,85 y un umbral de probabilidad entre 0,6 y 0,95

Tabla 4.4: Resultados de estructuración del atributo Forma farmacéutica, con N= 100000, umbral de similitud de 0,85 y umbral de probabilidad variante entre 0,6 y 0,95.

N = 100000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,6	4025,299	0,96	0,956	0,957	0,952	0,96225
Umbral = 0,7	3962,603	0,951	0,958	0,953	0,953	0,96003
Umbral = 0,8	3994,559	0,957	0,955	0,954	0,955	0,95791
Umbral = 0,85	4163,471	0,957	0,958	0,95	0,956	0,94868
Umbral = 0,9	4129,801	0,95	0,956	0,954	0,956	0,94091
Umbral = 0,95	3998,988	0,958	0,959	0,957	0,955	0,93809

Para observar el resto de tablas que resumen los resultados según número de datos, se recomienda al lector revisar el Anexo C.

A partir de los resultados presentados en las Tablas 4.3, C.1, C.2, C.3 y 4.4, se puede observar que existe un aumento en los valores mientras mayor sea el tamaño del set de datos utilizado. Por otro lado, respecto a las métricas de *Accuracy, Recall, Precision* y *F1-Score*, se puede observar también que, al igual que en los resultados analizados anteriormente, no hay una tendencia de los datos hacia algún parámetro de probabilidad en específico. Lo mencionado anteriormente tiene sentido, puesto que el umbral de probabilidad, a diferencia del umbral de similitud, es utilizado posterior a la aplicación de los modelos de clasificación, por lo tanto, su valor es totalmente independiente de las métricas derivadas del modelo de aprendizaje de máquinas, el cual depende del umbral de similitud.

Respecto a la métrica de precisión, esta sí presenta valores relevantes, y como se observa en la Figura 4.2, hay una clara disminución del desempeño del algoritmo mientras más grande sea el umbral de probabilidad. Lo anterior implica que el umbral de probabilidad = 0,6, entrega los mejores resultados en el modelo, puesto que señala que un 96,23 % de los datos son estructurados con su forma farmacéutica correcta. Por otro lado, un umbral de probabilidad de 0,6 implica que la probabilidad que debe tener una etiqueta de pertenecer a una descripción de medicamento, debe ser de al menos un 60 %.

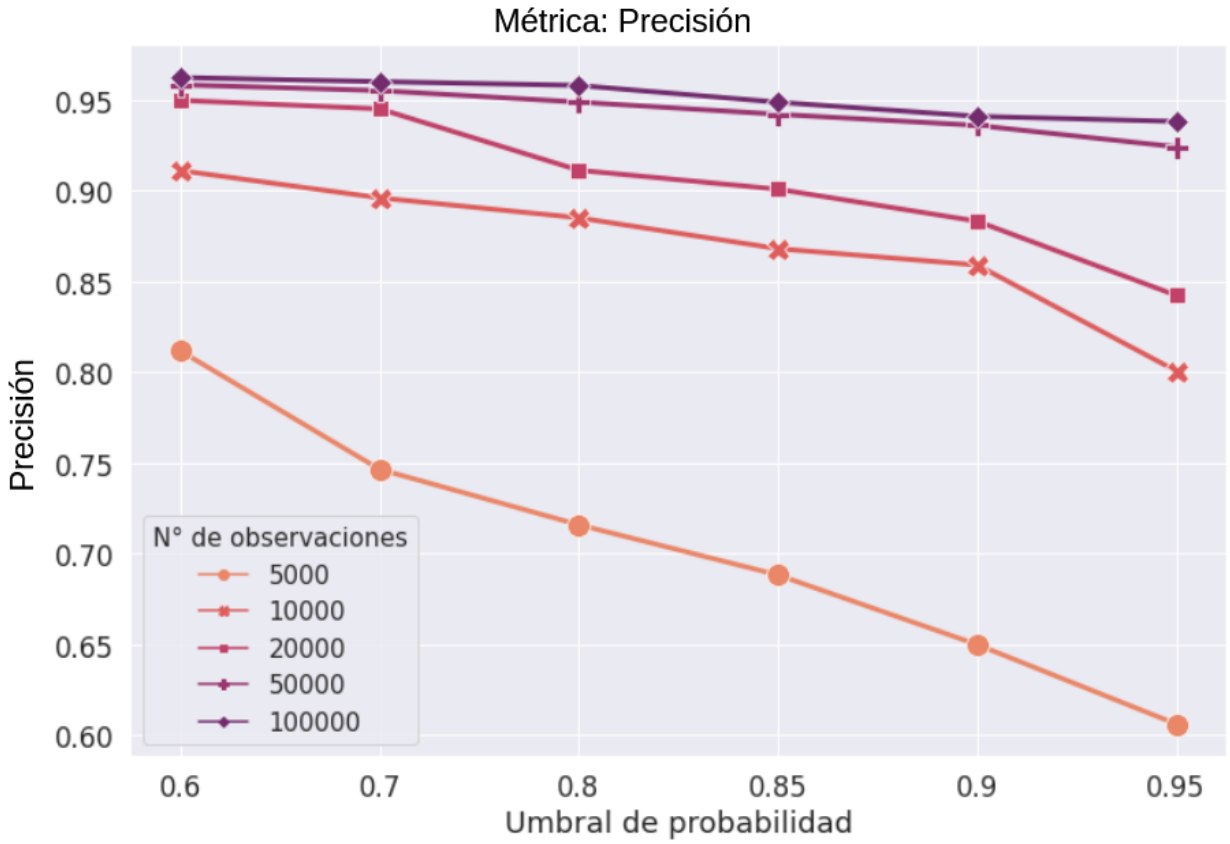


Figura 4.2: Desempeño de la métrica **precisión** durante el proceso.

Finalmente, se concluye que, según los valores entregados por la métrica de precisión, los mejores parámetros para el modelo son:

- N = 100000
- Umbral de similitud = 0,85
- Umbral de probabilidad = 0,6

Estos valores se establecen por defecto en la evaluación del siguiente subproceso.

4.2. Subproceso de extracción del Principio activo

En la Tabla 4.5 se presentan los resultados de este subproceso, entregando los valores de la métrica precisión y el tiempo de ejecución, según el tamaño N de los datos.

Tabla 4.5: Resultados de estructuración del atributo Principio activo, con un umbral de probabilidad = 0,6 y un umbral de similitud = 0,85.

	Tiempo de ejecución [Seg]	**Precisión**
N = 10000	401,14150	0,67654
N = 20000	1532,25019	0,71234
N = 40000	1916,995	0,73131
N = 60000	2132,077	0,74239
N = 100000	3892,593	0,7514

Se puede observar en la Tabla 4.5 que la Precisión va en aumento con la cantidad de datos presentes en el set de datos utilizado. Una precisión de 0,7514 implica que el 75 % de las descripciones fueron estructuradas con un principio activo correcto. Dado que los resultados no son los esperados, se realiza una revisión de las descripciones sin un principio activo correcto. A partir de lo anterior se obtiene que las descripciones sin una etiqueta correcta:

- **No poseen un principio activo que sea detectable por el algoritmo**, es decir, no hay un valor en la descripción de medicamento que coincida con algún principio activo del diccionario. Esto último tendría diversos posibles orígenes, como por ejemplo, la existencia de un nombre alternativo del principio activo que no esté presente en el diccionario, o inexistencia del valor de atributo en la descripción. A modo de ejemplo, se presentan en la Tabla 4.6 algunas descripciones con esta característica en común, en donde se observa además la etiqueta correcta.

Tabla 4.6: Ejemplo de algunas descripciones con etiqueta faltante de principio activo.

Descripción del medicamento	**Etiqueta correcta de principio activo**
bion 3 mini xaja x 30 comp	polivitaminico
cheltin fc 1 caja 30 comprimidos	fierro vitaminico
timarol 100mg 10ml	tramadol
sulfametoxazol 400 + trimetropina 80 mg comprimidos	cotrimoxazol

- **Las herramientas utilizadas en el algoritmo para medir distancias no detecta similitudes de palabras**, es decir, hay diversas formas de escribir un principio activo, por ejemplo, usando su abreviación. Debido a que el subproceso de extracción de principio activo funciona en base a las distancias de palabras, depende totalmente de la capacidad de la herramienta para reconocer palabras similares, pero no necesariamente iguales. A modo de ejemplo, se presentan en la Tabla 4.7 algunas descripciones de medicamentos que poseen esta característica, observando además su etiqueta correcta.

Tabla 4.7: Ejemplo de algunas descripciones con etiqueta faltante de principio activo.

Descripción del medicamento	Etiqueta correcta de principio activo
s.p.sodio clor 0,9 % am 250ml x 30 comp	sodio cloruro
ibandronato 150 mg comprimidos recubierto idena	ibandronico acido
cotrimozaxol comp 800+160 mg	cotrimoxazol
cefacidroxilo 500mg	cefadroxilo

4.3. Subproceso de extracción de la Concentración

En la Tabla 4.8 se presentan los resultados de este subproceso, entregando los valores de la métrica precisión y el tiempo de ejecución, según el tamaño N de los datos.

Tabla 4.8: Resultados de estructuración del atributo Concentración, con un umbral de probabilidad = 0,6 y un umbral de similitud = 0,85.

	Tiempo de ejecución [Seg]	Precisión
N = 10000	1,536	0,4974
N = 20000	2,449	0,501
N = 40000	5,143	0,5254
N = 60000	7,792	0,54021
N = 100000	13,652	0,54329

Finalmente, respecto al último subproceso del algoritmo, se puede observar en la Tabla 4.8 que, al igual que todos los subprocesos anteriores, la métrica de precisión tiene un aumento en su valor luego del incremento del tamaño set de datos utilizado. Sin embargo, se puede observar que el desempeño de este subproceso no es destacable, puesto que con un set de N = 100000 datos, solo un 54,3 % de las descripciones son estructuradas de forma correcta.

A pesar de lo anterior, es necesario mencionar que el desempeño del algoritmo en este subproceso no se debe totalmente a la construcción de este. A partir de un análisis de aquellas descripciones con un valor de concentración incorrecto, se obtiene que:

- **Las etiquetas originales están incorrectas**, es decir, la etiqueta presente en la columna original de concentración de la base Vista Medicamentos, posee valores que no están presentes en la descripción del medicamento, o no coinciden con los que se extraen desde la descripción. Esto último genera la disminución de la evaluación de desempeño del algoritmo en este subproceso, puesto que la métrica utilizada se basa en la comparación de la etiqueta obtenida con la etiqueta original. En la Tabla 4.9 se presentan algunos ejemplos de descripciones que generan incongruencias al momento de comparar las etiquetas originales de la base, con las etiquetas generadas por el algoritmo.

Tabla 4.9: Ejemplo de algunas descripciones con etiqueta incorrecta de concentración.

Descripción del medicamento	**Etiqueta original de concentración**	**Etiqueta generada por el algoritmo**
carbetocin 100mg/1ml, ampolla	0100 mg	100mg/1ml
sevorane 250ml.	no def	250ml
trihexiteridilo 20 mg comprimidos	2mg	20mg
cotrimoxazol 80 mg ampolla	80/400mg	80mg

- **No hay valores de concentración en la descripción del medicamento, pero sí existe un valor en la etiqueta original**. Debido a la dependencia del algoritmo a la descripción, en los casos en los que no existe una concentración en la descripción, el estructurador retorna un valor faltante. Sin embargo, en ocasiones, a pesar de no existir un valor del atributo de concentración en la descripción, sí existe un valor en la etiqueta de concentración original, lo que genera una negativa al momento de comparar la etiqueta original, con la etiqueta obtenida por el algoritmo. En la Tabla 4.10 se observan algunas descripciones sin concentración, además de las etiquetas a comparar.

Tabla 4.10: Ejemplo de algunas descripciones con etiqueta incorrecta de concentración.

Descripción del medicamento	**Etiqueta correcta de concentración**	**Etiqueta generada por el algoritmo**
cloririo de gentamicina en frasco gotario.	0,3 %	NaN
cloramfenicol sol oftalmologica fcos	0,5 %	NaN
ranitidina amp.	50mg	NaN
gentamicina betam ung oft	0,3 %	NaN

Es necesario recordar al lector que la metodología utilizada para evaluar el desempeño del algoritmo consiste en comparar las etiquetas originales de cada atributo, con los resultados generados. Sin embargo, en el caso específico del último subproceso, comparar las etiquetas originales, con los resultados del algoritmo, no es una opción viable si se quiere conocer el verdadero desempeño del estructurador.

A partir de lo anterior, y con el objetivo de observar de mejor forma el desempeño del algoritmo para extraer la concentración de las descripciones de medicamentos, se realiza un proceso de reetiquetado de los datos, para el atributo de concentración. Este proceso tiene como fin el observar el verdadero desempeño del modelo, a partir de la comparación de las etiquetas generadas por el algoritmo de estructuración, y etiquetas originales correctas. Para esto, se reetiquetan de forma manual,3000 datos de concentración, a partir de las descripciones de medicamentos, disminuyendo así el problema de tener datos incorrectos desde el origen. Posterior a la modificación de etiquetas, se calculan nuevamente las métricas.

Tabla 4.11: Resultados de estructuración del atributo Concentración, con etiquetas reetiquetadas, con un umbral de probabilidad = 0,6 y un umbral de similitud = 0,85.

	Tiempo de ejecución [Seg]	**Precisión**
N = 1000	0,366	0,78604
N = 2000	0,405	0,78801
N = 3000	0,494	0,79154

En la Tabla 4.11 se puede observar un gran aumento de la métrica precisión, si se compara con la métrica obtenida anteriormente, en donde solo un 54,3 % de las etiquetas generadas por el algoritmo están correctas. Finalmente, realizando un análisis sobre las concentraciones que no fueron correctamente extraídas, se observa que:

- **No existe una identificación previa de algunas unidades de medida que acompañan a la concentración**. Debido a la herramienta utilizada en este subproceso, se requiere una previa identificación de las unidades de medida que aparecen en las descripciones de medicamentos. Por lo tanto, si no se identifica una unidad de medida, el algoritmo no será capaz de extraer la concentración correcta. En la Tabla 4.12 se presentan algunas descripciones de medicamento que tienen esta característica.

Tabla 4.12: Ejemplo de algunas descripciones con etiqueta incorrecta de concentración.

Descripción del medicamento	**Etiqueta correcta de concentración**	**Etiqueta generada por el algoritmo**
sulfato ferroso 125mg/ml 30 ml	125mg/ml 30 ml	30ml
nistatina tubo 100.000ui/15g	100.000ui/15g	15g
sodio cloruro al 0,9 % 500 cc	0,9 % 500 cc	0,9 %

4.4. Proceso de estructuración general

En la Tabla 4.13 se presentan los resultados generados por el algoritmo de estructuración al momento de realizar la estructuración de texto libre. Se entregan las métricas 1 y 2, definidas anteriormente, además del tiempo de ejecución, según el tamaño N de los datos.

Tabla 4.13: Resultados de estructuración del algoritmo de estructuración, con un umbral de probabilidad = 0,6 y un umbral de similitud = 0,85.

	Tiempo de ejecución [Seg]	**Métrica 1** (Precisión)	**Métrica 2** (Precisión promedio)
N = 20000	1284,825	70,354	2,04
N = 40000	3375,955	70,322	2,1
N = 60000	6793,28	73,028	2,15
N = 80000	8661,06	73,906	2,22
N = 100000	12836,949	74,934	2,24
N = 120000	16734,59270	75,945	2,5

Respecto a los resultados generales obtenidos del algoritmo de estructuración, se puede observar que, para el caso de la métrica precisión, existe un leve crecimiento de su magnitud a medida que aumenta el N utilizado. Dado que la métrica 1 corresponde a un *ratio* entre la cantidad de etiquetas que el algoritmo genera de forma correcta, y la cantidad total de etiquetas, se observa que un 75,9 % de las etiquetas totales obtenidas a partir de la estructuración de texto coinciden con la etiqueta original. Respecto al valor más alto de la métrica 1, posiblemente el resultado se debe a la baja capacidad de estructuración del último subproceso, puesto que al menos la mitad de las etiquetas generadas son incorrectas, lo que a su vez disminuye el desempeño del algoritmo en general.

En lo que respecta a la métrica de "Promedio de coincidencias de etiquetas", se observa que hay una total predominancia por un promedio de 2, lo que significa que en promedio, para cada una de las descripciones de medicamentos, solo 2 de 3 valores de atributo coinciden. Esto se debe principalmente a que el atributo con menos coincidencias sea la concentración.

A partir de lo antes mencionado, se debe recordar al lector que en la subsección previa se establece la idea de un reetiquetado de datos de forma manual, lo que permite observar el verdadero desempeño del subproceso. Siguiendo la idea antes planteada, se obtienen las métricas de desempeño del modelo para la estructuración de los 3000 datos reetiquetados, los que poseen una etiqueta correcta del atributo concentración.

Tabla 4.14: Resultados de estructuración del algoritmo de estructuración, con etiquetas reetiquetadas, con un umbral de probabilidad = 0,6 y un umbral de similitud = 0,85.

	Tiempo de ejecución [Seg]	**Métrica 1** (Precisión)	**Métrica 2** (Precisión promedio)
N = 1000	98,421	73,665	2,70
N = 2000	121,384	75,721	2,77
N = 3000	140,827	79,159	2,98

Se observa en la Tabla 4.14, que para N = 3000 datos, las métricas mejoran en comparación con las mejores métricas presentes en la Tabla 4.13. A partir de esto último, se puede

concluir que la corrección de las etiquetas incorrectas en el último subproceso, genera un mejor desempeño de algoritmo, por lo que al incrementar el N de datos utilizados puede causar una mejora en las métricas expuestas.

Capítulo 5

Conclusión y Trabajo Futuro

Actualmente, en Chile las investigaciones enfocadas en la estructuración de datos en el área de la salud es escasa. Si bien la problemática de transformar datos no estructurados a datos estructurados puede resolverse mediante el uso de personas, la cantidad de recursos humanos y monetarios, con el transcurso del tiempo, serán demandantes.

En el sector público de la salud, el contar con una herramienta que permita estructurar fármacos es de vital relevancia, puesto que permitiría el desarrollo de investigaciones en pos de mejorar las compras públicas. Por ejemplo, el ahorro que representaría el adquirir medicamentos a precios razonables. El algoritmo de estructuración facilita y promueve el desarrollo de este tipo de investigaciones, las cuales requieren de datos estructurados que no se pueden obtener fácilmente.

Respecto al beneficio generado por este trabajo de tesis, se encuentra la automatización de una tarea que usualmente se realiza de forma manual, disminuyendo los gastos de recursos que esto representa. Por otro lado, la aplicación más importante que se realiza utilizando el algoritmo, es la estructuración de datos que son relevantes para la búsqueda de mejores alternativas dentro del mundo de las compras públicas en Chile.

A pesar de que los resultados finales cumplieran con los objetivos, aún hay oportunidades de mejora en el algoritmo realizado. Aun cuando se logra generar valor a partir de la estructuración realizada, y se logra establecer un puntapié inicial para el desarrollo de un algoritmo más exacto, este proyecto tiene un camino futuro que recorrer, con el fin de entregar resultados correctos.

A partir de lo mencionado anteriormente, se proyecta un trabajo futuro sobre el algoritmo que permita mejorar el desempeño del algoritmo a partir del enfrentamiento de las diversas problemáticas encontradas en los datos. Dentro de estas se encuentran:

- Analizar el desbalance existente en las clases del atributo Forma Farmacéutica, dada la brecha que existe entre las etiquetas con mayor y menor frecuencia.
- Buscar mejores herramientas que permitan medir las distancias entre palabras de mejor forma. Permitiendo así identificar aquellos principios activos que son escritos de distintas formas. Además de enfrentar la problemática de tener múltiples principios activos en una única descripción.

- Identificar correctamente cuáles son todas las unidades de medida presentes en las concentraciones de la columna Concentración.

Finalmente, se establece también como trabajo futuro, el desarrollo de un sistema estructurador disponible al público, que permita a su vez la adquisición de datos para mejorar el desempeño del algoritmo a través de la adquisición de información nueva. Por ejemplo, los distintos nombres alternativos de cada principio activo, permitiendo al algoritmo realizar un reemplazo automático de valores.

Bibliografía

[1] Unidad de Monitoreo de Mercados SERNAC, "Medicamentos: Comportamiento de precios en un año de pandemia 2020- 2021," 2021, https://www.sernac.cl/portal/619/articles-64748_archivo_01.pdf.

[2] Centro Nacional de Farmacoeconomia (CENAFAR), "Medicamentos en chile: Revisión de la evidencia del mercado nacional de fármacos," 2013, https://www.minsal.cl/wp-content/uploads/2015/09/EstudioMedicamentos-22012014A.pdf.

[3] ANADEUS, "Estudio acerca de la diferencia de precios entre las distintas marcas de medicamentos y los medicamentos genéricos.," 2012.

[4] Servicio Nacional del Consumidor, "Hasta $181 mil de diferencia presentan los precios entre medicamentos originales de marca y bioequivalentes," 2019, https://www.sernac.cl/portal/604/w3-article-57125.html.

[5] Verbanaz, S. A., "Estructura del mercado de medicamentos en chile y gasto de bolsillo en salud en la ocde," Mayo 2022, https://obtienearchivo.bcn.cl/obtienearchivo?id=repositorio/10221/33130/1/Informe_final.pdf.

[6] Javatpoint, "An introduction to data structures.", https://www.javatpoint.com/data-structure-introduction.

[7] Mora, J., "Radiografía a las compras del sector salud: Basado en análisis de datos de compras públicas de las instituciones dependientes del ministerio de salud," Marzo 2019, https://observatoriofiscal.cl/archivos/documento/radiografC3\%ADaalascomprasdelsectorsalud_compressed.pdf.

[8] Malgarini, I., "Análisis econométrico de las compras públicas de medicamentos en chile," Master's thesis, Universidad de Chile, 2022.

[9] "Cenabast," 2022, https://www.cenabast.cl.

[10] Dirección de presupuestos, "Análisis del gasto y mecanismos de compra de medicamentos del sistema nacional de servicios de salud.," 2017, https://www.dipres.gob.cl/598/articles-168764_doc_pdf.pdf.

[11] Cenabast, "Ley cenabast.," 2022, https://www.cenabast.cl/ley-cenabast-remedios-mas-baratos-y-de-calidad/.

[12] Cenabast, "Cenabast registró compras históricas el 2021.," 2022, https://www.cenabast.cl/cenabast-registro-compras-historicas-el-2021/.

[13] "What is machine learning?.", https://www.ibm.com/cloud/learn/machine-learning.

[14] "What is natural language processing?.", https://www.ibm.com/cloud/learn/natural-language-processing.

[15] Stubblebine, T., "Regular Expression Pocket Reference". 1005 Gravenstein Highway North, Sebastopol, CA 95472.: O'Reilly, 2007.

[16] IBM, "Meta characters in regular expressions." https://www.ibm.com/docs/en/rational-clearquest/9.0.1?topic=tags-meta-characters-in-regular-expressions. Retrieved November 15, 2022.

[17] Mikolov, T., Chen, K., Corrado, G., y Dean, J., "Efficient estimation of word representations in vector space," 2013, doi:10.48550/ARXIV.1301.3781.

[18] "Google code archive - long-term storage for google code project hosting..", https://code.google.com/archive/p/word2vec/.

[19] Li, Y., Xu, L., Tian, F., Jiang, L., Zhong, X., y Chen, E., "Word embedding revisited: A new representation learning and explicit matrix factorization perspective," en International Joint Conference on Artificial Intelligence, 2015.

[20] Mammone, A., Turchi, M., y Cristianini, N., "Support vector machines," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 1, no. 3, pp. 283–289, 2009.

[21] Mooney, R. J., "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning," en Conference on Empirical Methods in Natural Language Processing, 1996, https://aclanthology.org/W96-0208.

[22] Ng, H. T., "Exemplar-based word sense disambiguation: Some recent improvements," CoRR, vol. cmp-lg/9706010, 1997, http://arxiv.org/abs/cmp-lg/9706010.

[23] Park, H. A., "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain," Journal of Korean Academy of Nursing, vol. 43, p. 154—164, 2013, doi:10.4040/jkan.2013.43.2.154.

[24] Breiman, L., "Random forests," Machine Learning, vol. 42, p. 5–32, 2001, https://doi.org/10.1023/A:1010933404324.

[25] Levenshtein, V. I., "Binary codes capable of correcting deletions, insertions and reversals.," Soviet Physics Doklady, vol. 10, pp. 707–710, 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[26] Ghani, R., Probst, K., Liu, Y., Krema, M., y Fano, A., "Text mining for product attribute extraction," SIGKDD Explor. Newsl., vol. 8, p. 41–48, 2006, doi:10.1145/1147234.1147241.

[27] Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., Yu, Z., y Elsas, J., "Learning to extract attribute value from product via question answering: A multitask approach," en Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining, KDD '20, (New York, NY, USA), p. 47–55, Association for Computing Machinery, 2020, doi:10.1145/3394486.3403047.

[28] Xu, H., Wang, W., Mao, X., Jiang, X., y Lan, M., "Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title," en Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, (Florence, Italy), pp. 5214–5223, Association for Computational Linguistics, 2019, doi:10.18653/v1/P19-1514.

[29] Huang, Z., Xu, W., y Yu, K., "Bidirectional LSTM-CRF models for sequence tagging," CoRR, vol. abs/1508.01991, 2015, http://arxiv.org/abs/1508.01991.

[30] Karamanolakis, G., Ma, J., y Dong, X. L., "Txtract: Taxonomy-aware knowledge extraction for thousands of product categories," CoRR, vol. abs/2004.13852, 2020, https://arxiv.org/abs/2004.13852.

[31] Mandhan, S., Sarath, P. R., y Niwa, Y., "Numerical atrribute extraction from clinical texts," 2015, doi:10.13140/RG.2.1.4763.3365.

[32] Du, M., Pang, M., y Xu, B., "Multi-task learning for attribute extraction from unstructured electronic medical records," en Semantic Technology (Wang, X., Lisi, F. A., Xiao, G., y Botoeva, E., eds.), (Singapore), pp. 117–128, Springer Singapore, 2020.

[33] "Dirección de compras y contratación pública.", https://www.chileatiende.gob.cl/instituciones/AE011.

[34] ChileCompra, "Chile, evaluación del sistema de compras públicas.," 2017, https://www.chilecompra.cl/wp-content/uploads/2017/11/MAPS-final-2017.pdf.

[35] Chile transparente, "Marco regulatorio de la competitividad de las compras públicas en chile.," 2018, https://www.chiletransparente.cl/wp-content/files_mf1545236124MarcoRegulatoriodiciembre.pdf.

[36] Organización para la Cooperación y el Desarrollo Económicos, "Recomendación del consejo sobre contratación pública.," 2015, https://www.oecd.org/gov/ethics/OCDE-Recomendacion-sobre-Contratacion-Publica-ES.pdf.

[37] "Qué es un convenio marco.", https://www.mercadopublico.cl/Home/Contenidos/QueEsCM.

[38] "Qué es una licitación.", https://www.mercadopublico.cl/Home/Contenidos/QueEsLicitacion.

[39] "Qué es un trato directo.", https://www.mercadopublico.cl/Home/Contenidos/QueEsTratoDirecto.

[40] ChileCompra, "Normativa de compras públicas: Ley nº19.886 y su reglamento.," Octubre 2016, https://www.chilecompra.cl/wp-content/uploads/2018/03/reglamento2016-octubre.pdf.

[41] Fiscalía Nacional Económica, "Estudio de mercado sobre compras públicas," Agosto 2020, https://www.fne.gob.cl/wp-content/uploads/2020/08/informe_preliminar_EM05_2020.pdf.

[42] Arriagada, L., "Geriatrización del arsenal farmacológico.," 2019, https://www.minsal.cl/wp-content/uploads/2019/07/Q.F.-Leonardo-Arriagada-Geriatrizaci\%C3\%B3n-del-arsenal-Farmacol\%C3\%B3gico.pdf.

[43] Ministerio de Salud, "Manual de selección de medicamentos: Metodología para la selección de medicamentos del formulario nacional y arsenales farmacoterapéuticos de los establecimientos de salud.," 2010, https://www.minsal.cl/sites/default/files/files/Manual\%20Seleccin\%20de\%20Medicamentos\%20FInal\%20con\%20Dise\%C3\%B1io.pdf.

[44] Cenabast, "Las definiciones de la comisión de adquisiciones en la adjudicación de productos.", https://www.cenabast.cl/las-definiciones-de-la-comision-de-adquisiciones-en-la-adjudicacion-de-productos/.

[45] "Registro sanitario de productos farmacéuticos," 2022, https://www.ispch.gob.cl/ana

med/medicamentos/registro-sanitario-de-productos-farmaceuticos/.

Anexos

Anexo A. Análisis exploratorio de datos

Tabla A.1: Descripción de columnas de la base Vista Medicamentos.

Columna	**Descripción**
Orden de compra	Código que representa una compra.
Licitación	ID que se presenta solo en casos en que compra se haya realizado mediante Licitación. Permite un cruce con la base de datos de mercado público.
Año	Corresponde al año en el que se realizó la orden de compra.
Mes	Corresponde al mes en el que se realizó la orden de compra.
Región	Corresponde a la región en la que se encuentra ubicado el comprador.
Segmento comprador	Segmento al cual pertenece la entidad compradora. Ej. Municipalidades.
Comprador	Nombre de la entidad compradora. Ej. Hospital Santo Tomas de Limache.
Proveedor	Nombre de la entidad que provee el producto comprado. Ej. Laboratorio Chile Corp.
ZGEN	Código que busca facilitar y estandarizar todos los requerimientos de la red de hospitales en Mercado Público.
Principio Activo	Corresponde al ingrediente principal del medicamento.
Concentración	Corresponde a la cantidad de fármaco disuelto en una determinada cantidad de disolución.
Forma Farmacéutica	Corresponde al formato por el que se adaptan los principios activos, permitiendo así la administración de una sustancia en el organismo.
Producto	Corresponde al producto adquirido en la compra.
Cantidad	Señala la cantidad de producto adquirido.
Precio unitario	Precio de un producto individual.
Monto neto	Monto total pagado, corresponde a la multiplicación del precio unitario por la cantidad de producto adquirido.
Descripción comprador	Descripción del medicamento, según el ente comprador.
Descripción proveedor	Descripción del medicamento, según el proveedor.

Anexo B. Subproceso de extracción del principio activo: Con el parámetro umbral de similitud variable, y umbral de probabilidad fijo en 0,85.

En la Tabla B.1 se presentan los resultados obtenidos para un set de datos de tamaño N = 10000, con un umbral de probabilidad de 0,85 y un umbral de similitud entre 0,7 y 0,95.

Tabla B.1: Resultados de estructuración del atributo Forma farmacéutica, con N = 10000, umbral de probabilidad de 0,85 y umbral de similitud variante entre 0,7 y 0,95.

N = 10000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,7	201,406	0,9	0,896	0,902	0,896	0,90187
Umbral = 0,8	204,645	0,901	0,907	0,908	0,904	0,90137
Umbral = 0,85	199,009	0,916	0,91	0,902	0,905	0,90824
Umbral = 0,9	203,912	0,926	0,920	0,921	0,918	0,88951
Umbral = 0,95	199,648	0,912	0,906	0,904	0,897	0,87547

En la Tabla B.2 se presentan los resultados obtenidos para un set de datos de tamaño N = 20000, con un umbral de probabilidad de 0,85 y un umbral de similitud entre 0,7 y 0,95.

Tabla B.2: Resultados de estructuración del atributo Forma farmacéutica, con N = 20000, umbral de probabilidad de 0,85 y umbral de similitud variante entre 0,7 y 0,95.

N = 20000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,7	579,597	0,933	0,925	0,934	0,931	0,92384
Umbral = 0,8	584,495	0,94	0,941	0,938	0,931	0,92496
Umbral = 0,85	598,59	0,932	0,935	0,933	0,93	0,93953
Umbral = 0,9	592,694	0,93	0,936	0,937	0,934	0,93325
Umbral = 0,95	613,69	0,931	0,932	0,931	0,928	0,93166

En la Tabla B.3 se presentan los resultados obtenidos para un set de datos de tamaño N = 50000, con un umbral de probabilidad de 0,85 y un umbral de similitud entre 0,7 y 0,95.

Tabla B.3: Resultados de estructuración del atributo Forma farmacéutica, con N = 50000, umbral de probabilidad de 0,85 y umbral de similitud variante entre 0,7 y 0,95.

N = 50000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,7	2334,613	0,948	0,95	0,947	0,946	0,94500
Umbral = 0,8	2348,204	0,950	0,951	0,950	0,949	0,94496
Umbral = 0,85	2492,282	0,946	0,947	0,946	0,945	0,94949
Umbral = 0,9	2352,061	0,951	0,952	0,952	0,951	0,94527
Umbral = 0,95	2402,752	0,942	0,94	0,949	0,947	0,94228

Anexo C. Subproceso de extracción del principio activo: Con el parámetro umbral de probabilidad variable, y umbral de similitud fijo en 0,85.

En la Tabla C.1 se presentan los resultados obtenidos para un set de datos de tamaño N = 10000, con un umbral de similitud de 0,85 y un umbral de probabilidad entre 0,6 y 0,95.

Tabla C.1: Resultados de estructuración del atributo Forma farmacéutica, con N = 10000, umbral de similitud de 0,85 y umbral de probabiliad variante entre 0,6 y 0,95.

N = 10000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,6	194,894	0,893	0,894	0,881	0,898	0,911
Umbral = 0,7	199,729	0,894	0,88	0,872	0,867	0,896
Umbral = 0,8	207,933	0,891	0,897	0,835	0,828	0,885
Umbral = 0,85	205,009	0,85	0,893	0,863	0,824	0,868
Umbral = 0,9	199,823	0,896	0,901	0,846	0,83	0,859
Umbral = 0,95	202,394	0,872	0,888	0,882	0,839	0,8006

En la Tabla C.2 se presentan los resultados obtenidos para un set de datos de tamaño N = 20000, con un umbral de similitud de 0,85 y un umbral de probabilidad entre 0,6 y 0,95.

Tabla C.2: Resultados de estructuración del atributo Forma farmacéutica, con N = 20000, umbral de similitud de 0,85 y umbral de probabilidad variante entre 0,6 y 0,95.

N = 20000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,6	580,190	0,914	0,914	0,912	0,90936	0,94971
Umbral = 0,7	593,720	0,918	0,9181	0,917	0,91355	0,94501
Umbral = 0,8	584,655	0,906	0,9067	0,904	0,90249	0,91120
Umbral = 0,85	618,149	0,916	0,9165	0,915	0,89329	0,90075
Umbral = 0,9	590,963	0,911	0,9113	0,91	0,886	0,88312
Umbral = 0,95	581,928	0,918	0,91	0,917	0,862	0,84226

En la Tabla C.3 se presentan los resultados obtenidos para un set de datos de tamaño N = 50000, con un umbral de similitud de 0,85 y un umbral de probabilidad entre 0,6 y 0,95.

Tabla C.3: Resultados de estructuración del atributo Forma farmacéutica, con N= 50000, umbral de similitud de 0,85 y umbral de probabilidad variante entre 0,6 y 0,95.

N = 50000	**Tiempo de ejecución [Seg]**	***Accuracy***	***Recall***	***Precision***	***F1-Score***	**Precisión**
Umbral = 0,6	2252,995	0,951	0,95	0,95	0,944	0,95829
Umbral = 0,7	2496,658	0,948	0,951	0,945	0,945	0,95508
Umbral = 0,8	2225,933	0,94	0,952	0,949	0,934	0,94874
Umbral = 0,85	2330,221	0,949	0,955	0,946	0,93	0,94204
Umbral = 0,9	2329,802	0,945	0,95	0,95	0,942	0,93591
Umbral = 0,95	2249,873	0,949	0,957	0,949	0,948	0,92409