UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

Profesor Patrocinante:

Dra. Rosa L. Figueroa I.

Informe de Memoria de Título
para optar al título de:

Ingeniero Civil Biomédico

Extracción de Información en EMR para la Identificación de Obesidad mediante el Estudio de Comorbilidades Asociadas

Concepción, Abril de 2015

Christopher Alejandro Flores Jara

UNIVERSIDAD DE CONCEPCIÓN
Facultad de Ingeniería
Departamento de Ingeniería Eléctrica

Profesor Patrocinante:
Dra. Rosa L. Figueroa I.

Extracción de Información en EMR para la Identificación de Obesidad mediante el Estudio de Comorbilidades Asociadas

Christopher Alejandro Flores Jara

Informe de Memoria de Título
para optar al Título de

Ingeniero Civil Biomédico

Abril 2015

Resumen

El presente trabajo tuvo la finalidad de identificar obesidad y sus tipos mediante las principales enfermedades asociadas a ella, utilizando registros médicos electrónicos sin identificación de los pacientes, provenientes del Hospital Guillermo Grant Benavente de Concepción en 43 subespecialidades médicas.

El problema contempló el estudio de campos estructurados y de texto libre de cada registro médico para una posterior clasificación en dos niveles. El primer nivel correspondió a la identificación de obesidad junto con otros estados nutricionales como el bajo peso, normopeso y sobrepeso. Posteriormente, se procedió a clasificar los tipos de obesidad en las categorías moderada o leve (tipo I), severa (tipo II), mórbida (tipo III) y superobesidad (tipo IV).

La clasificación fue realizada en una colección de registros médicos (corpus principal) creada a partir de una recuperación de información de la base de datos, utilizando palabras claves de las categorías del primer nivel de clasificación. Posteriormente, un grupo de anotadores analizó cada registro médico recuperado para etiquetarlos de acuerdo a las clases de ambos niveles, creando un Gold Standard. Para la implementación de los clasificadores Support Vector Machine (SVM) y Naïve Bayes (NB), fue necesaria una representación numérica de cada registro del corpus principal utilizando términos relacionados a las comorbilidades de la obesidad en forma de unigrams y bigrams, a través de una matriz binaria, TF (Term Frequency) y TF-IDF (Term Frequency-Inverse Document Frequency).

El más alto desempeño en el primer nivel de clasificación fue el obtenido por SVM con una exactitud igual a 89,10%, utilizando bigrams con una matriz binaria. Por su parte, NB obtuvo en el mismo nivel un porcentaje de exactitud igual a 84,49%, utilizando igual segmentación y representación de la información que SVM. Analizando la clase “obesidad” en particular, el rendimiento de SVM es superior al alcanzado por NB, obteniéndose porcentajes de exactitud iguales a 89,22% y 85,07% respectivamente, en un enfoque basado en bigrams y usando una matriz binaria. En el segundo nivel de clasificación, nuevamente SVM logró el mejor desempeño, alcanzando una exactitud igual a 93,80% mediante bigrams y las matrices TF y TF-IDF para representar la información, mientras que el mejor desempeño de NB fue una exactitud de 82,17% con el mismo tipo de segmentación, pero con una matriz binaria.

Un sueño cumplido

Agradecimientos

A Dios por la posibilidad de estar escribiendo estas palabras

A mis padres y hermana por su educación e incondicional apoyo

A todos mis familiares y personas que me tratan como uno de los suyos

A los profesores que me enseñaron a lo largo de la vida

A los amigos por todos los momentos

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Nomenclatura

Funciones

$\Phi(x)$	: Transformación de kernel

Vectores

w(t,d)	: Vector de pesos de un término en un documento
$\vec{w}$	: Vector normal al hiperplano
(x_i, y_i)	: Tupla i-ésima del vector de atributos x y las clases y

Matrices

M_{mxn}	: Matriz de dimensiones m x n

Escalares

‖w‖	: Norma Euclídea
{max}	: Máximo valor
{argmax}	: Máximo valor de un argumento
log_{10}	: Logaritmo en base 10
$P(h/D)$	: Probabilidad condicionada de h dado D

Abreviaciones

Mayúsculas

ENS	: Encuesta Nacional de Salud
OECD	: Organization for Economic Cooperation and Development
NLP	: Natural Language Processing
EMR	: Electronic Medical Records
HGGB	: Hospital Guillermo Grant Benavente
IMC	: Índice de Masa Corporal
SVM	: Support Vector Machine
NB	: Naïve Bayes
TF	: Term Frequency
IDF	: Inverse Document Frequency
WEKA	: Waikato Environment for Knowledge Analysis
UMLS	: Unified Medical Language System
IR	: Information Retrieval
IE	: Information Extraction
ECOC	: Error Correcting Output Coding
PoS	: Part of Speech
BoW	: Bag of Words
MAP	: Máxima A Posteriori
ASCII	: American Standard Code for Information Interchange
TP	: True Positive
TN	: True Negative
FP	: False Positive
FN	: False Negative
HTA	: Hipertensión Arterial
DM	: Diabetes Mellitus
ERGE	: Enfermedad por Reflujo Gastroesofágico
SAHOS	: Síndrome de Apnea e Hipopnea Obstructiva del Sueño
CPAP	: Continuous Positive Airway Pressure
HDL	: High Density Lipoprotein
LDL	: Low Density Lipoprotein
ICC	: Insuficiencia Cardiaca Congestiva
IAM	: Infarto Agudo al Miocardio
TVP	: Trombosis Venosa Profunda

Minúsculas

i.2.b.2.	: Informatics for integrating Biology to the Bedside
acc.	: Accuracy

Capítulo 1. Introducción

1.1. Introducción General

Desde la década de los 70’ a la fecha, las mejoras en las condiciones de salud de la población en Chile aumentaron las expectativas de vida, disminuyendo las tasas de mortalidad infantil, desnutrición y enfermedades infecciosas. Sin embargo, las enfermedades crónicas no transmisibles sufrieron un alza considerable, teniendo como principales responsables a la obesidad y el sobrepeso [1]. Según la Encuesta Nacional de Salud (ENS), realizada el año 2010 a 8.900.000 personas por el Ministerio de Salud en Chile, el 19,2% de los hombres y el 30,7% de las mujeres mayores a 15 años tienen esta enfermedad crónica. Estas cifras sufren un alza importante si se considera el sobrepeso o preobesidad, alcanzando un 45,3% y 33,6% en hombres y mujeres, respectivamente. Además, la población adulta sobre los 45 años posee la mayor prevalencia de obesidad, lo que se complica con el surgimiento de otras enfermedades crónicas [2]. Si bien, desde el año 2009 al año 2011 Chile bajó de la posición 37 a la 35 en el indicador de obesidad a nivel mundial, ocupa el sexto lugar entre los países de la OECD (Organization for Economic Cooperation and Development) [3].

La obesidad no sólo está asociada a complicaciones físicas de salud, siendo las más frecuentes la hipertensión arterial, dislipidemias y diabetes mellitus tipo 2, que además aumentan el riesgo de padecer enfermedades cardiovasculares y trombosis cerebral, sino también a trastornos sicológicos como la depresión [4]. Esto último se debe a que la calidad de vida de estos pacientes se ve deteriorada al impedirles realizar sus actividades cotidianas de forma normal, o incluso, el poder movilizarse, debiendo permanecer postrados lo que complica aún más su condición basal.

Las políticas preventivas de la obesidad y sus comorbilidades necesitan mecanismos de evaluación del real impacto que tienen en la población. Esto último, junto con la inminente informatización de los recintos hospitalarios, siendo un claro ejemplo el reemplazo de los registros clásicos del historial clínico de los pacientes por los de carácter electrónico, hacen necesario el desarrollo de mecanismos automatizados de identificación de enfermedades, extrayendo información mediante el uso del Procesamiento del Lenguaje Natural (NLP, del inglés Natural

Language Processing) para la búsqueda de patrones ocultos en los textos dada la complejidad en que se presentan los datos.

Este trabajo de memoria de título plantea un método de identificación automática de la obesidad, estudiando sus principales comorbilidades descritas por la literatura. Como fuente de información se utilizaron registros médicos electrónicos (EMR, del inglés Electronic Medical Record) facilitados por el Hospital Regional Guillermo Grant Benavente (HGGB) de Concepción sin identificación de los pacientes para un posterior análisis estadístico.

1.2. Objetivos

1.2.1 Objetivo General

Desarrollar e implementar un sistema de identificación de obesidad mediante el estudio de sus comorbilidades en registros médicos electrónicos.

1.2.2 Objetivos Específicos

- Generar un corpus con enfermedades crónicas asociadas a la obesidad en distintas subespecialidades médicas
- Generar un sistema de identificación automática de la obesidad en EMR
- Generar estadísticas del corpus para las distintas subespecialidades médicas, relacionando las distintas enfermedades asociadas a la obesidad
- Evaluar los modelos de clasificación

1.3. Alcances y limitaciones

El trabajo se realizó con EMR sin la identificación de los pacientes para un resguardo de su privacidad, obtenidos entre los años 2011 y 2012 desde el HGGB de Concepción. Se dispuso de un set de datos de 43 subespecialidades médicas en los campos Diagnóstico, Anamnesis, Indicaciones Médicas, Hábitos y Antecedentes Mórbidos.

El estudio se realizó en función de la obesidad y sus comorbilidades asociadas descritas en este informe.

1.4. Temario y Metodología

La organización de este informe se detalla a continuación:

- En el Capítulo 2 se define y clasifica la obesidad en función del Índice de Masa Corporal (IMC) [5] y se presentan sus comorbilidades más frecuentes. Además, se describen las técnicas de clasificación supervisada SVM y NB, la validación cruzada, utilizada como método de selección del set de pruebas y entrenamiento, y el proceso de anotación para la generación de un Gold Standard.
- En el Capítulo 3 se detallan los materiales y métodos utilizados. Corresponde a la aplicación de los fundamentos descritos en el Marco Teórico para el procesamiento de los datos y su uso en la extracción de información. Además, se establecen las comorbilidades de la obesidad a utilizar en el estudio, y se describe el uso del software WEKA (Waikato Environment for Knowledge Analysis) [6], utilizado para la clasificación de los EMR.
- En el capítulo 4 se muestran los resultados obtenidos del análisis de los datos, del procesamiento de la información y de la clasificación efectuada en los EMR.
- En el capítulo 5 se presenta una discusión del trabajo a partir de los resultados obtenidos.
- En el capítulo 6 se dan a conocer las conclusiones del trabajo.
- En el capítulo 7 se presenta el trabajo futuro.
- En el capítulo 8 se da a conocer el trabajo presentado producto de esta memoria de título.

Capítulo 2. Revisión bibliográfica y marco teórico

En este capítulo se expondrán trabajos relativos al procesamiento natural de EMR para la extracción de información y clasificación de enfermedades de forma supervisada. Además, se detallarán aspectos de la obesidad como enfermedad crónica y sus principales comorbilidades, así como los fundamentos de la minería de textos y del aprendizaje automático aplicados a textos escritos en lenguaje natural y que dan lugar al tema desarrollado en este trabajo.

2.1. Revisión bibliográfica

La informatización de los sistemas hospitalarios ha traído consigo mejoras continuas en salud que benefician directamente a los pacientes. Un claro ejemplo es el progresivo cambio en la forma de registrar las interacciones entre el médico y el paciente mediante formatos digitales. Un registro médico electrónico permite tener el control histórico del estado de salud de los pacientes y poder acceder a ellos de manera instantánea desde distintas unidades en un centro de salud, resguardando la información personal. Sin embargo, no basta con disponer de una plataforma digital que facilite y agilice el acceso a los datos médicos, sino también procesarla, extrayendo información relevante de las enfermedades y hábitos de los pacientes para un posterior análisis estadístico o automatización de diagnósticos utilizando el NLP. El lenguaje natural es aquel que permite la comunicación e intercambio de información de forma oral o escrita. Su procesamiento es una disciplina de la inteligencia artificial que estudia las relaciones gramaticales para la creación de sistemas que permitan la comunicación hombre-máquina [7].

Es debido a lo anterior que surgen trabajos para el procesamiento de los textos en EMR considerando las distintas relaciones de las enfermedades, además del lenguaje propio del dominio en estudio. La información se presenta muchas veces de forma dispersa, por lo que la utilización de herramientas automáticas para su procesamiento permite establecer diversas relaciones para la toma de decisiones que de forma tradicional tiene un alto costo si sólo se considera el tiempo dedicado para tales propósitos.

2.1.1 Lenguaje médico

Los textos científicos se caracterizan por la objetividad a la hora de transmitir información, exponiendo los contenidos de forma lógica y clara para evitar ambigüedades de interpretación. El lenguaje médico no es ajeno a esta descripción, presentando un alto nivel de abstracción a la hora de presentar un informe [8]. Sin embargo, uno de los mayores problemas a la hora de procesar y extraer información de fichas clínicas es que el registro se realiza sin una estructura predeterminada, su actualización es en forma progresiva y los sistemas computacionales existentes son incapaces de atender las distintas variaciones léxicas. En el trabajo descrito por Camacho et al. [9] se hace un análisis del lenguaje médico presente en las fichas electrónicas, enfatizando en la gran cantidad de términos para expresar una misma condición en un paciente y cómo los sistemas computacionales las identifican de forma distinta, aunque para el personal médico tenga el mismo significado. Esta diversidad léxica es favorecida además por el uso de abreviaciones y la presencia de monosemia, es decir, la existencia de términos con significado único. Por otro lado, se describe el problema que resulta al identificar una condición en un paciente sujeto a la temporalidad en que ocurren los hechos.

Es fundamental para la realización de un estudio basado en el NLP una comprensión del tipo de texto y el dominio en que está inmerso. El lenguaje médico posee una gran diversidad léxica y no está ajena a los errores gramaticales en los textos libres o no estructurados.

Para los sistemas computacionales es más fácil comprender la existencia de un término en un texto por lo que las variaciones léxicas no hacen más que dificultar su aprendizaje. El ser humano es capaz de entender un mensaje con cierto grado de error o carente de elementos conectores debido a que durante su desarrollo ha aprendido reglas gramaticales que le permiten identificar una alteración en el mensaje y obviarlo. Sin embargo, los sistemas computacionales carecen, en un principio, de dicho conocimiento, por lo que un entrenamiento basado en reglas y patrones del lenguaje es fundamental para la obtención de un óptimo desempeño.

2.1.2 Procesamiento del corpus

El preprocesamiento de la información es fundamental para reducir la brecha existente entre el texto entregado y los sistemas de aprendizaje automático. Según el trabajo descrito por Pérez [10], se distinguen tres etapas para la obtención de un corpus.

-Selección: El proceso de extracción de conocimiento comienza con la selección del conjunto de documentos o corpus, previo estudio de sus características, para una posterior evaluación del modelo de minería de textos.

-Preprocesamiento: En esta etapa el corpus creado es sometido a técnicas del NLP para adecuar los textos a las capacidades de lectura de los sistemas computacionales. Mediante un análisis morfológico se busca identificar las raíces de las palabras y el reconocimiento de frases, etiquetándolas según el contexto en una categoría gramatical a través de dos formas:

- Etiquetación en base a reglas: Según el contexto y dependiendo de cada idioma, se asigna una etiqueta gramatical a la palabra analizada.
- Etiquetación estocástica: Mediante la probabilidad de ocurrencia de etiquetas. Se suele utilizar N –grams, que analiza la ocurrencia de las N-1 frecuencias antes a la estudiada.

De esta forma, se favorece la comprensión de las oraciones dentro del texto puesto que el sistema computacional tiene conocimiento previo del rol de cada término que varía según el contexto.

El preprocesamiento de los textos contempla además la eliminación de aquellas palabras que no entregan información relevante, como artículos, preposiciones, conjunciones y pronombres, conocidos como stopwords, excluyendo las de carácter negativo. Esto último es fundamental para disminuir el error en la identificación de enfermedades, puesto que la sola presencia de un término que describe una condición no es razón suficiente para atribuírsela a un paciente. En el trabajo descrito por Cheng [11] se propone un sistema de identificación de negaciones basadas en el algoritmo NegEx, utilizando expresiones regulares y una lista de las negaciones más frecuentes obtenidas desde la UMLS (Unified Medical Language System) debido a las ambigüedades presentes

en los textos médicos, en cuanto se presentan frases afirmativas, negativas e hipótesis de posibles enfermedades. Además, se discute la incorporación de codificaciones internacionales a los EMR escritos en lenguaje natural como una forma de reducir los errores en la identificación de enfermedades en los pacientes.

Otro aspecto fundamental en el preprocesamiento de los textos y recuperación de información es la segmentación, que consiste en separar un concepto de una oración o frase. La segmentación puede ser llevada a cabo mediante elementos espaciadores, signos de puntuación, tabulaciones y guiones [9]. El problema surge cuando un término es compuesto o no existe claridad en el límite de las sentencias, debiéndose recurrir al conocimiento en el dominio para establecer un marco de referencia.

En el preprocesamiento se procede además a la normalización de la información, que consiste en la homogenización de todo el texto de la colección de documentos que se procesarán como forma de favorecer la compatibilidad entre los sistemas computacionales y la información proporcionada para disminuir los costos asociados en el tiempo de ejecución y los errores en el desempeño de los algoritmos de aprendizaje automático. Los métodos más utilizados corresponden a la conversión de mayúsculas a minúsculas, el control de números, fechas y abreviaturas y la eliminación de signos ortográficos como tildes y diéresis [12] [13].

-Representación de los textos: Una vez aplicadas las técnicas del NLP, se procede a la representación del corpus mediante tokens, que corresponden a palabras, frases, conceptos o características. El enfoque Bolsa de palabras (Bag of Words) es el método más sencillo y consiste en representar la información por palabras sin importar el orden.

Una vez finalizada esta etapa, los textos poseen una organización estructurada y es posible aplicar en ellos técnicas de la minería de datos, sean modelos descriptivos o predictivos, para la extracción de información.

2.1.3 Recuperación y extracción de información

Tras seleccionar y procesar el corpus, es posible generar conocimiento a partir de los textos. En el trabajo descrito por Vilares [13], se analizan los sistemas de procesamiento automático de la información en cuanto a su extracción y recuperación. La Extracción de Información (IE, del inglés Information Extraction) tiene la función de detectar y presentar información relevante de los textos de manera automática para su almacenamiento y posterior uso. El problema de estos sistemas es la gran dependencia del dominio en la que están inmersos, haciendo difícil dar respuestas específicas formuladas por el usuario. Una solución a este problema son los sistemas denominados Búsqueda de respuesta y su aplicación más frecuente es su uso en los navegadores Web. Por su parte, los sistemas de Recuperación de Información (IR, del inglés Information Retrieval) tienen la función de identificar documentos mediante la consulta de un usuario a través términos claves. Por lo tanto, existe una dependencia entre IR e IE, donde el primero sirve como filtro de los documentos para una posterior extracción de información útil.

La IR, por lo tanto, tiene como función reducir el conjunto inicial de los documentos para una potencial necesidad, maximizando el número de documentos relevantes, en desmedro de los que no lo son. Las principales tareas de la recuperación de información y donde la minería de textos adquiere un rol fundamental son los siguientes:

- Recuperación ad hoc: Se dispone de un corpus de forma permanente y sólo las consultas varían para cada usuario, filtrándose los documentos de interés de forma específica e importando el orden en que son recuperados, como en los buscadores Web.
- Categorización o clasificación de documentos: Los documentos disponen de etiquetas fijadas de forma previa según el contenido de los textos, de forma manual o automática. En esta tarea, las consultas permanecen relativamente estables. El algoritmo de clasificación es fundamental para categorizar las instancias. Un claro ejemplo es la detección de correo no deseado.
- Clustering de documentos: A diferencia de la categorización de documentos, en tareas de clustering no se dispone de clases prefijadas por lo que sólo se procede a un agrupamiento de los textos de acuerdo a medidas de distancia. El objetivo es tener una alta similaridad intra-clusters y alta disimilaridad entre clusters.

- Segmentación de documentos: Consiste en la división de un documento en secciones según los temas tratados para resaltar los aspectos más importantes del texto.

Para favorecer dichas tareas, es frecuente la utilización de términos índices, que corresponden a palabras representativas de los textos. Si bien, estos últimos a menudo corresponden a una representación de todas las palabras de los textos, se obtiene un mejor desempeño si es que la indexación es propiciada por especialistas en el área en estudio. Por otro lado, dadas las variaciones léxicas [9], es frecuente la utilización de stemming. Este último procedimiento consiste en la reducción de una palabra a su raíz, eliminando sus terminaciones de manera que se produzca una reducción de la cantidad de términos similares en el conjunto de documentos.

Para la IR en fichas clínicas es fundamental tener un dominio del lenguaje médico [8] [9] para la elaboración de términos claves que permitan reducir el total de documentos con la finalidad de elaborar un corpus que sólo contenga información referente a enfermedades o hábitos en los pacientes. En términos simples, la IR reduce el problema de analizar una gran cantidad de documentos a una consulta a una base de datos, con la reducción de tiempo que eso implica. Para tales propósitos, es fundamental la realización de un adecuado preprocesamiento de los textos [10] debido a la alta dependencia que estos sistemas de recuperación tienen con las palabras y sus variaciones. Por su parte, los sistemas de IE, al requerir una mayor exactitud para la entrega de contenidos de los textos, necesitan un mayor enfoque en el NLP para un análisis de las distintas asociaciones de los textos. En caso de su uso en medicina, requieren además de una comprensión de las distintas interacciones que una enfermedad pueda tener.

2.1.4 Clasificación de textos médicos

Una vez obtenido el corpus a estudiar, es posible aplicar en ellos técnicas del aprendizaje automático, previa representación numérica de la información. Para tareas de clasificación de documentos es necesario, en primer lugar, definir las clases a identificar en los textos para su posterior anotación y clasificación.

En el trabajo descrito por Üzuner [14] se relata un desafío organizado por i2b2 (Informatics for Integrating Biology & the Bedside) para la creación de un sistema de IE en fichas clínicas para la identificación de obesidad y sus principales comorbilidades mediante el uso del NLP. En este trabajo, un grupo de médicos identificaron en los registros de alta médica las quince enfermedades más comunes de la obesidad como son el asma, enfermedad cardiovascular aterosclerótica, insuficiencia cardíaca congestiva, depresión, diabetes mellitus, cálculos biliares y colecistectomía, enfermedad por reflujo gastroesofágico, gota, hipercolesterolemia, hipertensión, hipertrigliceridemia, apnea obstructiva del sueño, osteoartritis, enfermedad vascular periférica e insuficiencia venosa.

Los clasificadores requieren de una representación numérica para su implementación, motivo por el cual la identificación de enfermedades debe estar sujeta a una transformación para su procesamiento a través de la presencia o no de palabras claves en los textos. Las clases definidas en este trabajo fueron:

- Presente: El paciente tiene la enfermedad
- Ausente: El paciente no tiene la enfermedad
- Cuestionable: El paciente puede tener la enfermedad
- No mencionada: La enfermedad no es mencionada en el registro médico

Un aspecto fundamental a considerar en el trabajo con EMR, es la total confidencialidad que debe existir por parte de los integrantes de un estudio con los datos de los pacientes, por tratarse de información sensible. En este caso, se utilizaron identificadores sintéticos, de forma de codificar cualquier alcance que se pudiera tener con los pacientes.

Todo proceso de clasificación supervisada requiere que las entidades estén debidamente etiquetadas [13]. Para tal propósito, dos médicos de manera independiente fueron los encargados de anotar cada registro médico para la etiquetación de las clases según las enfermedades antes descritas, incluida la obesidad. La etiquetación fue realizada mediante un enfoque textual e intuitivo.

- Juicios textuales: No requiere mayor análisis
- Juicios intuitivos: Para la clase No mencionada
- En caso de no haber mayor información, se clasifica como ausente
- Información adicional como exámenes y pruebas médicas, justifican la presencia de una enfermedad

El nivel de acuerdo de ambos evaluadores para cada enfermedad fue medido a través del índice de Kappa [15], obteniéndose el nivel más bajo en la tarea intuitiva. Finalmente, un tercer evaluador fue el encargado de resolver las discrepancias para la obtención de un Gold Standard. Este último no está ajeno al sesgo dado por la experiencia o especialidad del personal médico, sobre todo a la hora de realizar un juicio intuitivo. Esto implica que para un mismo caso dos personas especializadas en distintas áreas del conocimiento otorguen clasificaciones dispares. Esto último no es trivial, puesto que la evaluación de todo sistema supervisado depende necesariamente de la comparación con un estándar humano, y sobre todo en el ámbito de la medicina, se presentan ambigüedades en los textos que complican aún más su procesamiento y evaluación.

Un aspecto importante a considerar es la distribución de las clases, en este caso, la categoría Ausente fue la menos representada en todas las enfermedades. La medición de los resultados de los participantes fue realizada mediante Precision, Recall y F-Measure [16].

La identificación de enfermedades puede ser complementada con conceptos vinculados a ellas en función de las indicaciones que el personal médico realiza a los pacientes, propios de su lenguaje [9]. Para la tarea de identificación de las enfermedades de forma intuitiva, la incorporación de los medicamentos, tratamientos y los síntomas fue crucial para la obtención de los mejores resultados.

En el trabajo descrito por Solt et al. [17] se propone un método de clasificación semántica de los registros, basado en reglas según el contenido de los textos. De esta forma, fue posible la detección de cada una de las clases en función de sus etiquetas asociadas. Para tales efectos, se hizo uso de clasificadores binarios, denominados sistemas expertos, para la realización de etiquetas en base a la semántica del texto. Este tipo de sistemas de clasificación fue desarrollada previo al aprendizaje máquina y es dependiente del dominio en que es aplicado para la obtención de resultados óptimos, debiéndose crear el conjunto de reglas de forma manual, limitando su aplicación para otros enfoques. No obstante, un conocimiento del dominio en el tema a estudiar permite una automatización de estos sistemas expertos, como es el caso de una ontología médica. La principal ventaja de este mecanismo es el enfoque caja blanca, intrínseco de las reglas, que permiten un análisis posterior de las interacciones que se dieron lugar en el proceso de clasificación, a diferencia de sistemas sofisticados, que si bien pueden tener un mejor desempeño, dificultan una decodificación de los patrones que dieron lugar al aprendizaje durante el entrenamiento de los clasificadores. Este aspecto es fundamental en ciencias como la medicina, pues los esfuerzos se centran precisamente en dar respuesta a los factores que pueden desencadenar una enfermedad en particular.

El problema fue abordado en primer lugar, creando múltiples diccionarios con terminología médica, incorporando abreviaciones, sinónimos, equivalencias, errores ortográficos, sufijos y términos relacionados para las distintas enfermedades, además de analizar los registros médicos en sus distintas estructuras, como el diagnóstico y la evolución clínica del paciente. Junto a lo anterior, fueron descartados aquellos registros médicos que hacían mención de la condición en los antecedentes familiares, pero no en el paciente, como una forma de reducir los falsos positivos. Mediante la utilización del algoritmo NegEx [11], se procedió al reconocimiento de negaciones en frases obtenidas por segmentación del texto [9]. Para la clasificación de los documentos, se utilizó un algoritmo binario que otorgaba una o más etiquetas a un registro médico, de acuerdo a las características del texto y considerando las clases definida en el problema. Al producirse una etiquetación múltiple se procedió a un orden de asignación. La clase cuestionable tuvo la menor ponderación, en caso de presentarse en combinación con la presencia o ausencia de una enfermedad. De igual forma, una mención positiva tuvo prioridad sobre las de carácter negativo. Este método fue producto del análisis realizado al texto, según la distribución de las clases y considerando que lo que se busca precisamente es la identificación de enfermedades en cada registro médico.

Finalmente, en el trabajo se hace mención a que el sistema propuesto es escalable, por cuanto la cantidad de reglas creadas aumentan en función del número de registros médicos. Además tiene la ventaja de ser reutilizable, pues ya se dispone de una colección de elementos representativos por enfermedad las que fueron validadas por bibliotecas de conocimiento médico.

Otro enfoque para la clasificación de los registros médicos fue el descrito por Ambert y Cohen [18]. En este trabajo se plantea la necesidad de identificar los HotSpots, que corresponden a las características que mayor información entregan al sistema, una vez segmentado el texto [9] con conocimiento en el dominio. La técnica utilizada para la identificación de estas características fue en base a un ordenamiento según la ganancia de información, previa determinación de un umbral de corte en forma empírica. Mediante la construcción de términos claves en cada enfermedad, se procedió a cuantificar con una ventana de 100 caracteres alrededor de estos tokens los elementos más importantes para cada enfermedad, siendo representados en un vector binario. Al igual que en el trabajo publicado por Solt et al. [17], se procedió a identificar las negaciones existentes previo a la tokenización y vectorización.

La selección del set de pruebas y entrenamiento fue mediante el uso de la validación cruzada estratificada [16], utilizando el conjunto más pequeño como aprendizaje del clasificador. De esta forma, se buscó evaluar la contribución de pequeños tamaños de muestra en el sistema implementado.

Los sistemas de clasificación binarios deben ser adaptados para su utilización en problemas de índole multiclase. En este trabajo se utilizó el enfoque ECOC (Error Correcting Output Coding), que es una técnica de descomposición de problemas multiclase en forma binaria entre todos los subconjuntos de las clases posibles. Cada nuevo documento fue sometido a la decisión de cada clasificador binario. La decisión final está sujeta a la comparación resultante en cada clasificación de acuerdo a la mínima distancia con un vector generado, como la suma de las diferencias a nivel de bit entre cada clasificador binario. De esta forma, se reduce la probabilidad de generar errores al estar la decisión final sujeta al veredicto de múltiples clasificadores. Para esta última tarea se hizo uso de una máquina de soporte vectorial con kernel lineal, ajustando los pesos de forma manual debido a que se tiene conocimiento de la distribución de las clases [15].

Finalmente, en el trabajo se hace mención a las ambigüedades presentes en los textos, en particular, a la presencia de negaciones aunque no se trate de una propiamente tal, afectando el aprendizaje y evaluación de los clasificadores en la determinación de los falsos y verdaderos negativos.

Por otro lado, para la tarea de clasificación de las enfermedades de forma textual, los mejores resultados fueron obtenidos identificando medicamentos, tratamientos y síntomas en los textos, previo estudio del corpus y posterior creación de términos y frases más representativas.

En el trabajo descrito por Yang et al. [19] se propone un sistema de clasificación analizando la estructura de los registros médicos y la importancia de los textos registrados en cada una de ellas. Se procedió, en primer lugar, a clasificar cada texto contenido en los pacientes según el diagnóstico, antecedentes mórbidos, historia laboral, exámenes físicos de laboratorio, medicamentos u otros antecedentes, ponderándolos en base a la capacidad predictiva que pudieran tener a la hora de clasificar las distintas enfermedades durante el entrenamiento, según el número de etiquetas del Gold Standard y la cantidad de sentencias presentes en cada registro. De esta forma, se pretende comprender la importancia que los distintos tipos de textos tienen a la hora de describir una enfermedad, pues tal como se describió por Solt et al. [17], existen sentencias que pueden afectar negativamente el aprendizaje de los clasificadores durante el entrenamiento, como la mención a los antecedentes familiares o condiciones pasadas de los pacientes. Se propone además el uso de etiquetación de las unidades léxicas según el contexto, denominado POS (Part of Speech), como forma de comprender las relaciones entre los términos de los textos.

Las clasificaciones de tipo textual requieren del almacenamiento de una gran cantidad de términos del domino en estudio para su búsqueda en los registros médicos. En este trabajo se procedió a la recopilación de múltiples conceptos médicos para referirse a una enfermedad, sus sinónimos, subclases, tratamientos y síntomas asociados, según lo indicado a partir de recursos bibliográficos como la UMLS. El sistema se sustenta en base a la etiquetación de distintas sentencias, según su similitud con las almacenadas en los diccionarios. Por otra parte, al tratarse de una tarea de identificación explícita, estos sistemas deben permitir las similitudes parciales, ignorando el orden de las palabras si fuera necesario, sobre todo en las enfermedades de nombre compuesto o que posean distintas formas de expresión.

Finalmente, se propone ampliar el vocabulario de palabras claves, principalmente las abreviaturas utilizadas. Además se discute la reutilización del sistema implementado, pues si bien utiliza herramientas del aprendizaje automático, requiere del ajuste manual de ciertos parámetros durante la etapa de entrenamiento y un conocimiento médico para definir las características relevantes del sistema.

Otro enfoque para la tarea textual fue el trabajo descrito por Ware et al. [20] donde se procedió al uso de expresiones regulares para la búsqueda de palabras claves y un análisis del contexto en que las palabras estaban inmersas. Al igual que en los trabajos descritos por Yang et al. [19] y por Ambert y Cohen [18], se procedió a eliminar los textos que hacían mención a antecedentes familiares, por producir confusiones en los clasificadores, además de los textos que indicaban un comentario de los hábitos de los pacientes como el consumo de medicamentos y los textos de admisión al recinto hospitalario, que señalaban una hipótesis que posteriormente podía ser refutada. Se distinguieron en este trabajo dos tipos de conceptos: principales, asociados a los nombres de las enfermedades y secundarios, asociados a los tratamientos y fármacos. El enfoque estuvo basado en el estudio de los medicamentos de las distintas enfermedades, considerando los nombres comerciales, abreviaciones y genéricos, fundamentado en la alta relación esperada entre la indicación de un fármaco y una enfermedad, al igual que para un procedimiento o tratamiento. De la misma manera que en el trabajo publicado por Ambert y Cohen [18], se procedió a la realización de una búsqueda de las palabras claves y su vecindad, pero en este caso, delimitado por signos de puntuación.

Si bien los textos en su preprocesamiento pueden ser normalizados [9][12], en este trabajo el enfoque fundamentado en expresiones regulares permitió la identificación textual de las distintas enfermedades a pesar de las variaciones léxicas dadas por la presencia de letras mayúsculas y minúsculas, incluso en las abreviaturas médicas. Además, la identificación fue apoyada por datos numéricos asociados principalmente a la masa corporal, la talla y el IMC, para la búsqueda de obesidad o para estimar valores anormales en los exámenes de los pacientes. Esto último fue realizado para la tarea intuitiva principalmente.

De acuerdo a los trabajos descritos para la identificación de obesidad y sus comorbilidades, es necesario considerar términos claves y diversos aspectos médicos que describen una enfermedad,

no sólo de forma explícita, sino también indicando características inherentes a ella. Por tanto, la elaboración de distintos diccionarios con terminología médica es una herramienta fundamental como apoyo al NLP en este dominio.

2.2. Obesidad

La obesidad es una enfermedad crónica caracterizada por un aumento en la cantidad de grasa corporal en un individuo que signifique un riesgo para la salud. Este aumento de tejido adiposo puede ser el resultado del desbalance del gasto energético o un desorden del apetito debido a factores genéticos o ambientales [21].

Para una evaluación cuantitativa de la obesidad se utiliza el IMC, resultado del cociente entre la masa y el cuadrado de la talla de un individuo, para una clasificación en tres grados, tal como se muestra en la Tabla 2.1. Sin embargo, es común la incorporación de una nueva categoría, denominada superobesidad, para pacientes con un IMC mayor o igual a 50 [22].

TABLA 2.1 Estado nutricional en función del IMC

Estado	IMC [Kg/m^2]
Bajo peso	≤19,9
Normopeso	20-24,9
Sobrepeso	25-29,9
Obesidad grado I (Moderada)	30-34,9
Obesidad grado II (Severa)	35-39,9
Obesidad grado III (Mórbida)	≥40

El IMC puede ser complementado con el índice cintura/cadera para un tratamiento terapéutico. Si dicho valor es superior a 1 en varones y 0,9 en damas, se asocia a trastornos metabólicos y/o enfermedades cardiovasculares, en una distribución Androide, tal como se muestra en la Fig. 2.1.a. Por otro lado, si el índice cintura/cadera es inferior a 1 en hombres y 0,9 en mujeres, se está en presencia de una distribución Ginecoide, como se muestra en la Fig. 2.1.b, asociado a comorbilidades articulares y complicaciones en la irrigación venosa de los miembros inferiores [4].

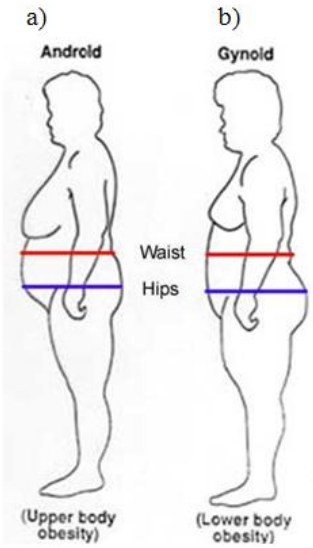

Fig. 2.1. Distribución de la grasa corporal. a) Androide**, b)** Ginecoide **[23]**

2.1.1 Comorbilidades de la obesidad

La obesidad es una enfermedad que tiene un efecto significativo en la esperanza y calidad de vida de quienes la padecen. Es considerada una enfermedad crónica de difícil tratamiento, pues su evolución depende directamente del estilo de vida y hábitos de los pacientes.

Las enfermedades asociadas a la obesidad se deben principalmente al efecto de la acumulación de tejido adiposo en el organismo. Sus complicaciones van desde un efecto mecánico en los músculos respiratorios, que disminuyen el volumen pulmonar; acumulación de placas de ateroma en los vasos sanguíneos, hasta efectos endocrinos y metabólicos [4]. En la Tabla 2.2 se muestra un resumen de las principales enfermedades asociadas a la obesidad.

TABLA 2.2 Consecuencias de la obesidad

Tipo	Enfermedad
Pulmonar	Enfermedad obstructiva crónica, asma bronquial, síndrome de hipoventilación pulmonar, apnea obstructiva del sueño
Síndrome metabólico	Diabetes mellitus tipo 2, dislipidemias
Cardiacas	Trombosis, ateroesclerosis, cardiopatías isquémicas, anginas de pecho, infartos agudos al miocardio, hipertensión arterial
Cáncer	Mama, útero, cérvix, próstata, riñón, colon, esófago, estómago, páncreas e hígado
Hígado	Hígado graso, esteatohepatitis y la cirrosis hepática
Ginecológicos	Anormalidades menstruales, infertilidad y síndrome de ovario poliquístico
Venosa crónica	Úlceras varicosas
Enfermedad Periodontal	Gingivitis, periodontitis

Los riesgos asociados a las comorbilidades aumentan a medida que aumenta el grado de obesidad. La Tabla 2.3 muestra la relación entre el grado de obesidad y el riesgo asociado a sus comorbilidades. Si el IMC es menor a 35, el riesgo de sufrir enfermedades cardiovasculares ateroesclerosis, diabetes tipo II e hipertensión arterial, ciertos tipos de cáncer y derrame cerebral, que corresponden a las principales causas de muerte en el mundo, es moderado. Los individuos con un IMC mayor a 35 poseen al menos una comorbilidad que es agravada con la obesidad, dando lugar al desarrollo de más enfermedades conforme se avanza en edad. La esperanza de vida es considerablemente afectada, así como la autoestima y la movilidad. Se recomienda un régimen alimentario para favorecer la disminución de peso, así como la práctica frecuente de actividad física en la medida que fuera posible. Finalmente, si el IMC es mayor a 40, es decir, individuos con obesidad mórbida o superobesidad, el riesgo de padecer comorbilidades es extremadamente alto y el tratamiento indicado para la reducción de peso es la cirugía bariátrica [24].

TABLA 2.3 Riesgos de la obesidad en función del IMC

IMC	Riesgo	Comorbilidades
<25	Mínimo	Bajo
[25-27)	Bajo	Moderado
[27-30)	Moderado	Alto
[30-35)	Alto	Muy alto
[35-40]	Muy alto	Extremadamente alto
>40	Extremadamente alto	Extremadamente alto

2.3. Clasificación supervisada

Los avances tecnológicos han favorecido el desarrollo de sistemas computacionales que han permitido incrementar las capacidades de almacenamiento y acceso a la información para un uso científico o comercial. En este sentido, la estadística aplicada a los sistemas informáticos ha posibilitado el surgimiento de disciplinas como el reconocimiento de patrones e inteligencia artificial, bioinformática y data mining.

El aprendizaje computacional es el encargado de generar hipótesis a partir de un número finito de casos para la clasificación o agrupamiento de datos. Si se dispone de las etiquetas a las que pertenecen estos datos, el aprendizaje se denomina supervisado. Este último, ya sea para resolver problemas binarios o multiclase, se sustenta en la capacidad de separar los datos durante una fase de entrenamiento, y utilizando esa información generada, poder clasificar datos no vistos [15]. Para realizar la clasificación, un algoritmo de aprendizaje supervisado requiere de los siguientes elementos:

- Características: Corresponden a los atributos que describen una instancia
- Clases: Corresponden a las categorías en que cada instancia es catalogada
- Algoritmo de aprendizaje: Sistema que permite discriminar en base a las características y clases datos no vistos

Un algoritmo de aprendizaje supervisado debe aprender conceptos durante el entrenamiento para la obtención de un modelo que le permita generalizar y clasificar datos durante la etapa de pruebas. Por último, todo sistema de aprendizaje automático debe ser evaluado, pues existen algoritmos de clasificación con mejor desempeño para una tarea en particular, considerando la distribución y las características de los datos. Las etapas de este aprendizaje se muestran en la Fig. 2.2.

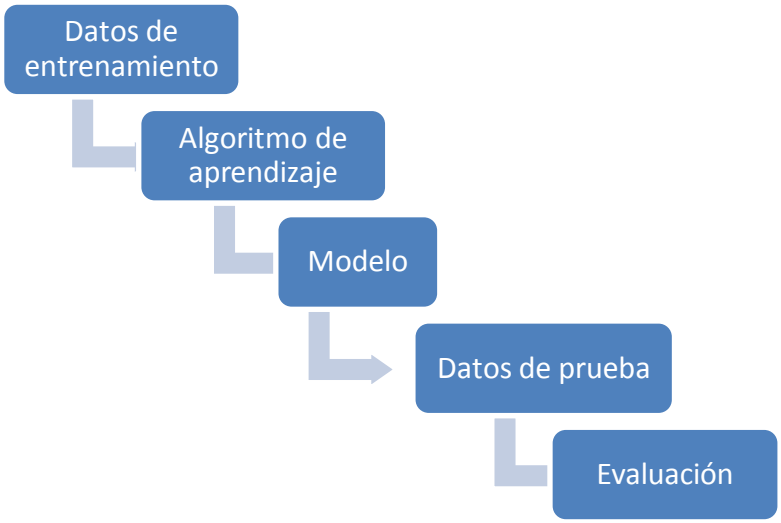

Fig. 2.2. Etapas de la clasificación supervisada

2.3.1 Support Vector Machine

Una máquina de soporte vectorial (SVM, del inglés Support Vector Machine) es un método de clasificación supervisada fundamentada en la teoría estadística del aprendizaje, propuesto por Vladimir Vapnik et al. en 1995. Dadas n tuplas (x_i, y_i) de entrada, donde X_i es el vector de atributos con $i \in \{1, \dots n\}$ e y_i las etiquetas de las clases con $y_i \in \{-1, 1\}$, se definen infinitos hiperplanos que pueden separar las clases [15], tal como se muestra en la Fig. 2.3.

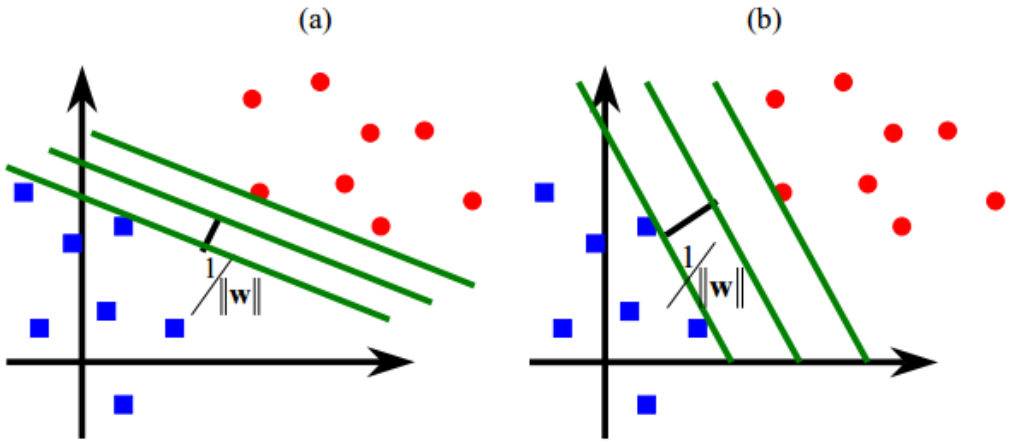

Fig. 2.3 Hiperplanos de separación de clases. De izquierda a derecha: Menor margen (a), mayor margen (b) [14]

El objetivo de una SVM es encontrar el hiperplano (canónico) que maximice el margen entre las clases. El hiperplano óptimo es el que satisface la ecuación 2.1

$$w \cdot x_i \pm \mathrm{b} = 0 \tag{2.1}$$

Donde,

w : Vector normal al hiperplano

$|b|/||w||$: Distancia perpendicular del hiperplano al origen

||w|| : Norma Euclídea de los pesos de los datos

Como resultado del entrenamiento, el hiperplano satisface las siguientes expresiones para ambas clases:

$$x_i \cdot w + b \geq 1, \qquad y_i = 1 \tag{2.2}$$

$$x_i \cdot w + b \leq -1, \qquad y_i = -1 \tag{2.3}$$

Ambos hiperplanos se forman mediante los vectores de soporte, que corresponden a puntos en el espacio que aseguran un margen máximo. En la Fig. 2.4 se aprecian ambos hiperplanos que se mantienen paralelos al canónico.

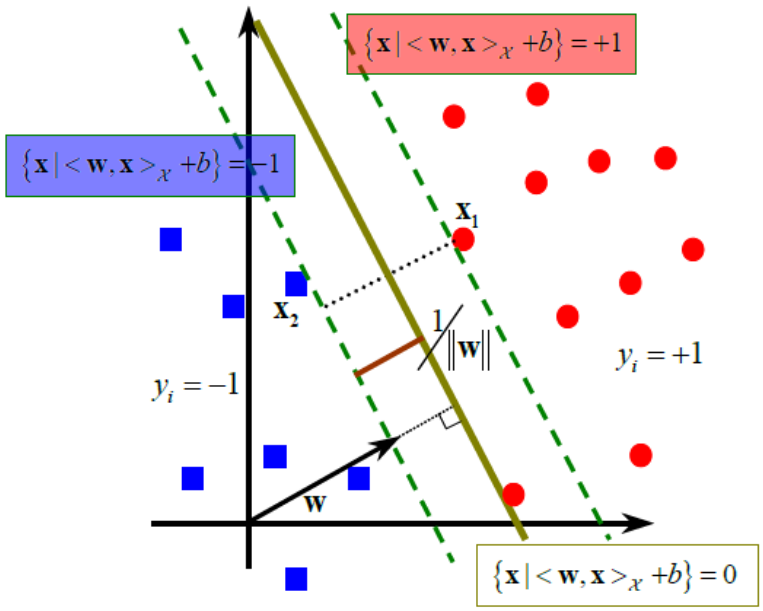

Fig. 2.4 Hiperplano canónico y margen entre ambas clases [25]

Dependiendo de la naturaleza de los datos, un problema de clasificación binaria puede obedecer a un tipo lineal o no lineal. En el primer caso, un hiperplano como el definido en la ecuación 2.1 es utilizado para separar ambas clases. Sin embargo, si un problema no es separable linealmente, el hiperplano de la ecuación 2.1 no es suficiente para discriminar ambas clases, encontrándose elementos de una categoría en otra.

SVM: Separación lineal

Un problema separable linealmente se muestra en la Fig. 2.4. El hiperplano canónico se define según la ecuación 2.1. La presencia de ruido en los datos, producto de su adquisición o simplemente outliers, pueden dificultar la obtención del hiperplano óptimo. Dichos elementos son ignorados y contenidos dentro del margen de separación.

SVM: Separación no lineal

Un problema no separable linealmente se muestra en la Fig. 2.5. Para tales efectos, se procede a la realización de una transformación no lineal, mediante una función Kernel Φ, para aumentar las dimensiones del espacio de entradas x, buscando encontrar una linealidad [25], definida como:

$$\Phi \rightarrow x \rightarrow x'$$
$$x \rightarrow \boldsymbol{x} \coloneqq \Phi(\mathrm{x}) \tag{2.4}$$

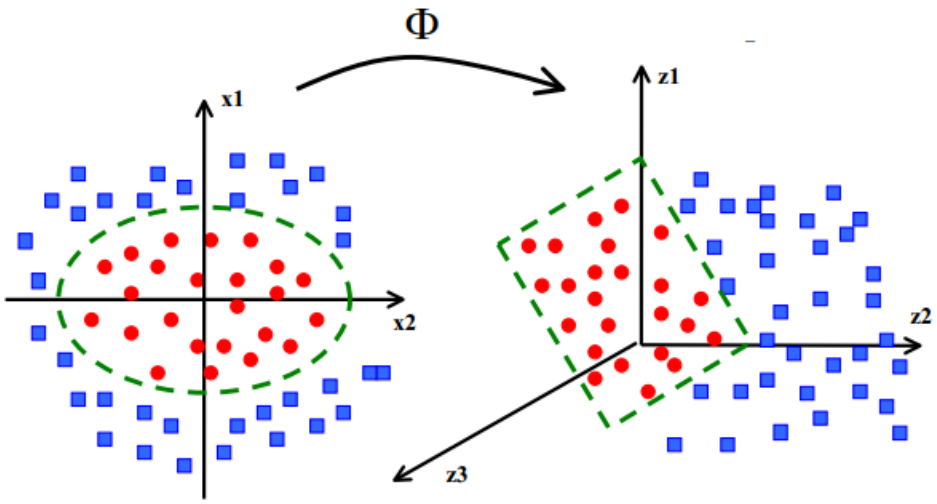

Fig. 2.5 SVM: Caso de separación no lineal y su transformación en el espacio de características [25]

Existen diversos Kernels para la transformación de las características a otra dimensión, siendo las más comunes, la función polinómica, pudiendo ser lineal o un grado mayor a 2; Gaussiana; Sigmoide, como las funciones tangentes parabólicas, logarítmicas y arcotangente [25].

Support Vector Machine Multiclase

No todos los problemas obedecen a un problema de separación de dos clases. Para tales casos, se disponen de técnicas basadas en el agrupamiento binario, construyendo n clasificadores según el método utilizado, para posteriormente unirlos y tomar una decisión final. Las técnicas más comunes son one against all (uno contra todos) y one against one (uno contra uno) [26].

Uno contra todos

Se procede a la creación de n clasificadores binarios como número de clases hayan, enfrentando cada una al resto. Así, cada hiperplano separa la clase i de los $n-1$ restantes. Finalmente, la decisión está sujeta a la categoría cuyo clasificador maximice el margen, según:

$$C_i = argmáx_{i=1\ldots n}(w_i \cdot x + b_i) \tag{2.5}$$

Uno contra uno

En esta modalidad se procede al enfrentamiento de cada clase entre sí, construyendo n clasificadores binarios según:

$$n = \frac{\mathrm{k}(\mathrm{k}-1)}{2} \tag{2.6}$$

Donde k corresponde a la cantidad de clases. De esta forma, se obtiene un hiperplano para cada uno de estos enfrentamientos binarios, añadiendo un voto a la clase seleccionada. La decisión final está sujeta a la clase que más votos tenga.

2.3.2 Naïve Bayes

Naïve Bayes (NB) es un clasificador supervisado fundamentado en las Probabilidades Máximas a Posteriori (MAP), en una distribución normal de los datos numéricos y en una independencia condicional de los atributos dada una clase en particular [15]. Es decir, este clasificador requiere del cálculo de las probabilidades de los atributos, dada una clase, considerando que estos últimos no tienen algún tipo de relación entre sí. NB puede ser representado mediante un grafo donde los arcos se dirigen desde las clases a los atributos, tal como se muestra en la Fig. 2.6.

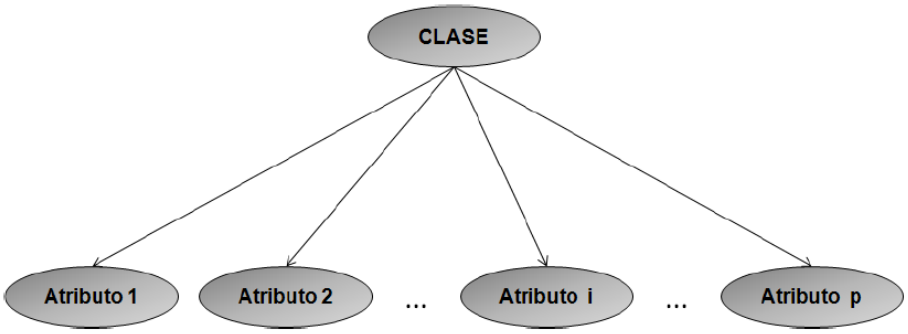

Fig. 2.6 Grafo que representa el clasificador NB [15]

El aprendizaje Bayesiano busca encontrar una hipótesis h entre todas las H posibles tras observar una serie de datos D [27] según:

$$\mathrm{h_{MAP}} = \operatorname{argmax} \mathrm{P(h/D)} \tag{2.7}$$

El Teorema de Bayes establece la probabilidad condicional de la hipótesis h dado D, como:

$$P(c) = \frac{P(D/h) \cdot P(h)}{P(D)} \tag{2.8}$$

Reemplazando la ecuación 2.7 en la ecuación 2.8 y considerando que $P(D)$ es independiente de h y por lo tanto constante, se tiene:

$$h_{MAP} = argmax_{h \in H} P(D/h) \cdot P(h) \tag{2.9}$$

Si además se supone que $P(h)$ es constante, es decir, todas las probabilidades de las hipótesis son iguales, se tiene la máxima verosimilitud (maximum likelihood):

$$h_{MAP} = argmax_{h \in H} P(D/h) \tag{2.10}$$

Si se aplica esta formulación matemática a un problema de clasificación, con C las clases equivalentes a las hipótesis H, y a_n los atributos equivalentes a los datos D, se tiene:

$$C_{MAP} = argmax_{c_i \in C} P(a_1, a_2, \ldots a_n / c_i) \cdot P(c_i) \tag{2.11}$$

Finalmente, suponiendo que los atributos son condicionalmente independientes, se define el clasificador Bayesiano según la ecuación 2.12.

$$C_{Naïve\ Bayes} = argmax_{c_i \in C} P(c_i) \cdot \prod_{j=1}^{n} P\left(\frac{a_j}{c_i}\right) \tag{2.12}$$

2.3.3 Set de entrenamiento y pruebas

Los algoritmos de clasificación supervisada construyen una función de decisión o reglas en función del aprendizaje obtenido durante la etapa de entrenamiento. Sin embargo, a la hora de evaluar este aprendizaje es necesario considerar el tamaño del conjunto de datos a utilizar. Si se usa todo el set de datos como entrenamiento no existirán instancias para ser evaluadas, o en caso de utilizar el mismo conjunto de datos, se producirá overfitting o sobreajuste, obteniéndose resultados muy optimistas con baja capacidad de generalización. El otro extremo es entrenar el modelo con pocos datos, muchas veces para reducir los tiempos de procesamiento, produciéndose underfitting [16]. Ambos efectos se muestran en la Fig. 2.7.

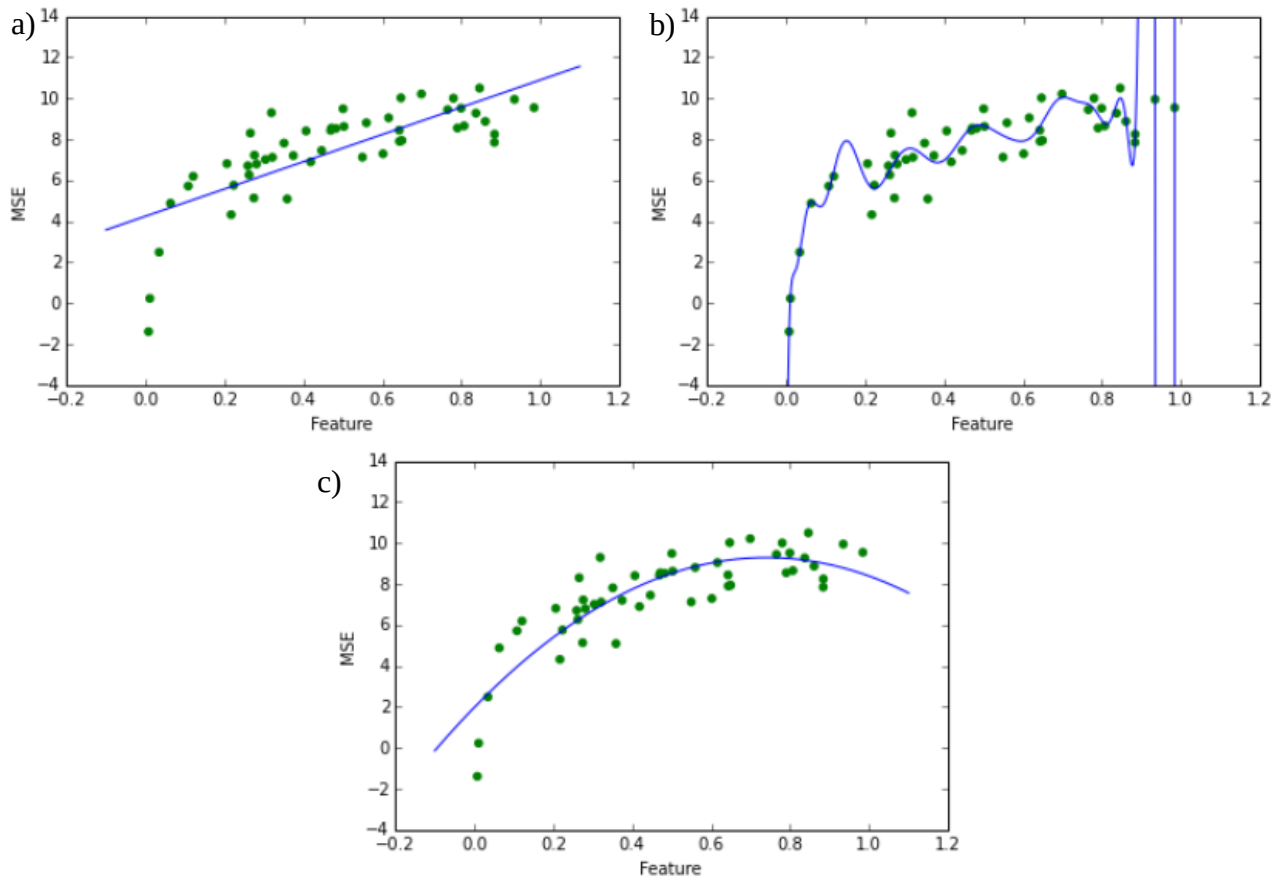

Fig. 2.7 Curva de Aprendizaje (Azul). a) Underfitting**. b)** Overfitting**. c)** Óptimo **[28]**

Validación cruzada: k-fold cross validation

En la validación cruzada se crean k particiones aproximadamente iguales y mutualmente excluyentes en igual número de iteraciones. La evaluación del algoritmo se realiza, en primer lugar, entrenando en k-1 particiones, dejando el resto como set de prueba tal como se muestra en la Fig. 2.8. En el caso de que se necesite disponer de una distribución de las clases en cada partición igual que los datos originales, se utiliza una validación denominada cruzada estratificada (stratified k-fold cross validation). Cuando el número de particiones es igual al número de datos, el set de pruebas estará compuesto por una única instancia y el de entrenamiento, por el número de instancias menos una. Esta forma del método se denomina validación cruzada dejando uno afuera (leave one-out cross validation). La validación cruzada tiene la ventaja de otros métodos de selección del set de pruebas y entrenamiento, como el tradicional Hold-out que establece una única división de cada conjunto en 1/3 y 2/3 respectivamente, contrarrestar una inadecuada distribución de las clases al evaluar el desempeño del clasificador como el promedio obtenido en cada iteración de selección. Además, la validación cruzada tiene la finalidad de evitar el sobreajuste (overfitting), producida cuando el clasificador no es capaz de generalizar durante el entrenamiento, ajustándose en forma específica a cada instancia [16] (ver Fig. 2.7.b).

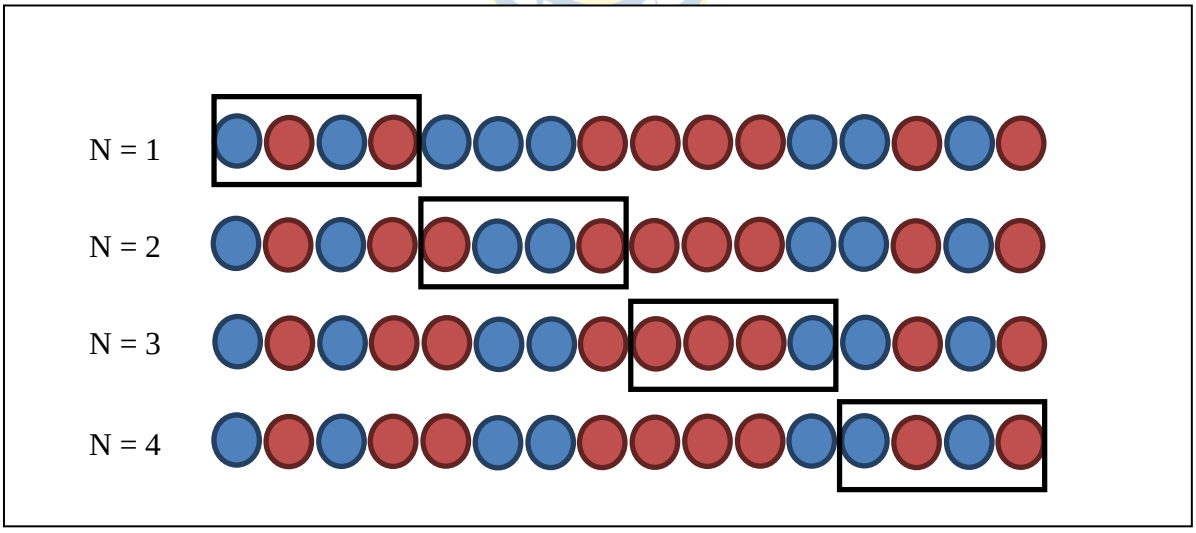

Fig. 2.8 Selección del set de datos mediante validación cruzada con k = 4. En cada iteración (N) se selecciona el set de entrenamiento (en rectángulo) y prueba

2.4. Proceso de anotación

La evaluación de los clasificadores supervisados está sujeta a una comparación con un Gold Standard creado por expertos humanos en el tema en un proceso de anotación. Al finalizar dicho proceso se debe medir el grado de acuerdo para una estimación de la significancia del estándar de forma de ser utilizando en la evaluación de los clasificadores supervisados.

Los anotadores encargados de clasificar los documentos de un corpus deben tener conocimiento del tema en cuestión, debiéndose realizar un entrenamiento previo. La evaluación debe ser realizada por al menos dos personas para evitar las anotaciones subjetivas. Además, cada decisión debe ser ajena a influencias de los demás anotadores. En caso de discrepancias, un tercer evaluador debe tomar la decisión final [29]. Para garantizar una correcta anotación, los evaluadores deben poseer un conocimiento a priori de las clases a identificar en los documentos, las que deben estar definidas de forma clara para evitar ambigüedades.

2.4.1 Índice de Kappa de Cohen

Para la evaluación del grado de acuerdo entre los evaluadores para dos o más categorías, se utiliza una medida de concordancia estadística denominada índice de Kappa (k). El índice de kappa de Cohen fue propuesto para problemas con dos evaluadores o clases binarias [15], cuyo valor es calculado según:

$$k = \frac{P_o - P_e}{1 - P_e} \tag{2.13}$$

Donde,

P_o: Proporción de concordancia acordada entre los observadores

P_e: Proporción de concordancia dada por el azar

Si se considera una distribución de clases, como la que se muestra en la Tabla 2.4, dados dos evaluadores A y B, cuyas clasificaciones se contrastan en forma vertical para un evaluador y en forma horizontal para el otro, P_o puede ser calculado mediante el cociente entre la suma de todos los aciertos y el número total de casos (N) según:

$$P_o = \frac{\sum_i^n \pi_{ii}}{N} \tag{2.14}$$

Por otro lado, P_e puede ser obtenido considerando el marginal de cada clase para ambos evaluadores, según:

$$P_e = \frac{\sum_i^n \pi_i^A \cdot \pi_i^B}{N^2} \tag{2.15}$$

TABLA 2.4 Distribución de clases según dos evaluadores para el cálculo de k

		A		
		Clase 1	⋯	**Clase n**
B	**Clase 1**	$\pi_{0,0}$	⋯	$\pi_{0,m}$
	⋮	⋮	⋱	$\pi_{0,n}$
	Clase n	$\pi_{n,0}$	⋯	$\pi_{n,m}$

Los distintos niveles de acuerdo se muestran en la Tabla 2.5, siendo k = 0 un total desacuerdo y k =1 un acuerdo total.

TABLA 2.5 Grado de acuerdo en función del índice de kappa

Kappa (κ)	**Grado de acuerdo**
<0,00	Sin acuerdo
0,00-0,20	Insignificante
0,21-0,40	Mediano
0,41-0,60	Moderado
0,61-0,80	Sustancial
0,81-1,00	Casi perfecto

Capítulo 3. Materiales y métodos

En este capítulo se describen los materiales utilizados en el presente trabajo, correspondientes a la base de datos con los EMR de los pacientes del HGGB de Concepción en las distintas subespecialidades médicas, haciendo una breve descripción de los campos contenidos para el estudio de la obesidad y de las herramientas utilizadas tanto para la anotación y clasificación de los registros. Además, se describe la metodología utilizada en el procesamiento de los datos según lo estipulado en el Capítulo 2.

3.1. Materiales

Se dispuso de una base de datos con 65.780 EMR sin la identificación de los pacientes, provenientes desde el HGGB de Concepción en 43 subespecialidades médicas. Cada registro médico está compuesto por campos estructurados y de texto libre, siendo cada uno de ellos considerados en este trabajo como un documento en particular. Los campos no estructurados corresponden a la interacción entre el médico y el paciente, donde se registran los aspectos más importantes del estado de salud de este último para un control histórico.

Los campos de texto libre son los siguientes:

- Anamnesis: Corresponde al registro de lo relatado por el paciente en la atención médica en cuanto a síntomas y conductas
- Diagnóstico: Corresponde al registro del médico que hace de la condición de salud del paciente, de acuerdo a las enfermedades que posee
- Indicaciones médicas: Corresponde a las recomendaciones que el personal clínico hace al paciente para tratar una condición, de acuerdo a su estado de salud

Por su parte, los campos estructurados corresponden al registro de hábitos o enfermedades en función de su presencia o ausencia en los pacientes, obedeciendo a una categoría. Estos campos son los siguientes:

- Hábitos: Corresponde a las conductas de los pacientes en función de su estilo de vida, que pueden dar lugar a enfermedades
- Antecedentes mórbidos: Corresponde a las enfermedades que los pacientes poseen al momento de la visita al médico

Es importante señalar que los EMR fueron facilitados por el Hospital sin la identificación de los pacientes, para un resguardo de su privacidad.

3.2. Metodología

Para dar cumplimiento a la identificación de la obesidad y sus tipos en los EMR, se definió una metodología de trabajo fundamentada en el Capítulo 2.

Para la realización de un estudio mediante el NLP, en primer lugar, se debe tener conocimiento del dominio y de las singularidades presentes en los textos. Como fue descrito en el trabajo de Camacho et al. [9], el lenguaje médico posee una gran diversidad léxica lo que dificulta su procesamiento. Para este estudio, se procedió a analizar los campos estructurados y de texto libre, con énfasis en la obesidad y sus comorbilidades. Para una simplificación del problema en cuanto a la cantidad de enfermedades asociadas a la obesidad a ser consideradas en este estudio, se utilizaron en primera instancia las señaladas en el trabajo publicado por Üzuner [14]. Sin embargo, este listado fue modificado luego del proceso de anotación, añadiendo las palabras claves que el grupo de evaluadores estimaron más frecuentes en los EMR que hacían mención a las comorbilidades de la obesidad.

Las clases fueron definidas en dos niveles de acuerdo a lo señalado en la Tabla 2.1. El primer nivel consistió en poder identificar el estado nutricional de los pacientes para determinar un bajo peso, normopeso, sobrepeso u obesidad. Según lo observado en los EMR, la palabra clave "obesidad" tiene una alta relación con que el paciente tuviese dicha enfermedad, prácticamente no existiendo negaciones para tal condición. Es por esta razón, que el enfoque del primer nivel fue en base al estado nutricional de los pacientes, de manera de añadir contraejemplos a la obesidad. Por su parte, el segundo nivel corresponde a una clasificación de los tipos de obesidad.

Una vez definidas las clases, se procedió al preprocesamiento de la base de datos de manera de homogeneizar cada texto contenido en los EMR, normalizándolos y eliminando stopwords.

Tras tener conocimiento del dominio de la obesidad y sus comorbilidades, se procedió al procesamiento de la información con la finalidad de generar un corpus principal, utilizado para la clasificación en los dos niveles, y otro basado en comorbilidades. El corpus principal fue creado producto de una primera recuperación de información, utilizando expresiones regulares para la búsqueda de palabras claves como el "IMC" y las asociadas al primer nivel de clasificación (normopeso, bajo peso, sobrepeso y obesidad), debido a que la identificación de los tipos de obesidad obedece a un subconjunto de la clase "obesidad". Por otro lado, el corpus de comorbilidades fue creado como un subconjunto del principal mediante un filtrado por palabras claves en el dominio de las enfermedades más comunes de la obesidad en una segunda recuperación de información. Este último corpus fue utilizado para la extracción de características, segmentando los textos según un enfoque basado en una o dos secuencias de palabras (tokens), unigrams o bigrams respectivamente. De esta forma, se buscó clasificar los EMR del corpus principal sólo en función de las comorbilidades de la obesidad. La Fig. 3.1 muestra el doble proceso de recuperación de información para la obtención de los EMR utilizados en la clasificación de los dos niveles (corpus principal) y la generación de documentos destinados a la extracción de características (corpus de comorbilidades).

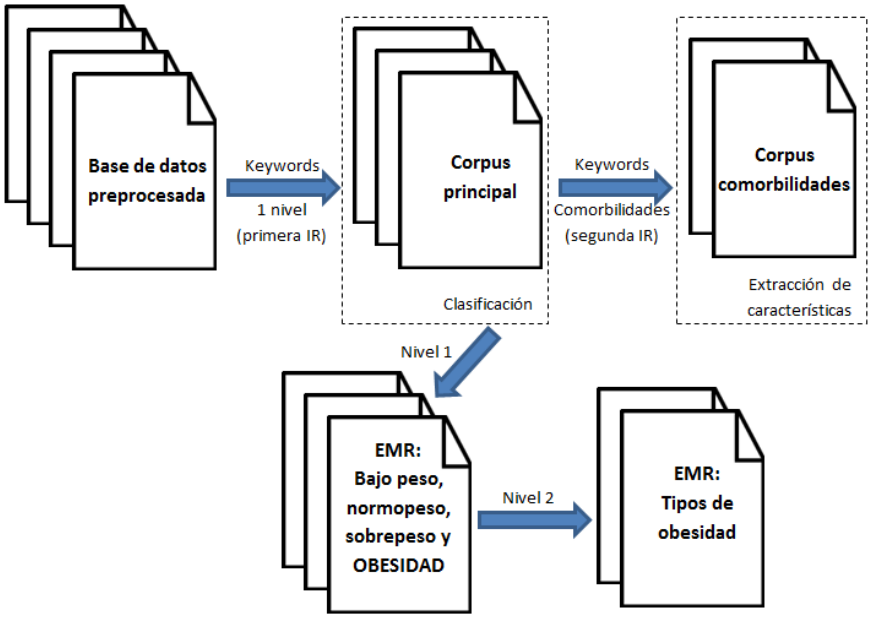

Fig. 3.1 Sistema implementado para la obtención del corpus principal (clasificación) y de las comorbilidades de la obesidad (extracción de características)

En un proceso de anotación fueron etiquetados todos los EMR del corpus principal según las categorías de los dos niveles para la creación de un Gold standard. Posteriormente, se realizó la representación de la información, mediante tres formas distintas: Un enfoque binario, a través de la presencia o ausencia de los tokens en los EMR del corpus principal; en función de la frecuencia de aparición de los tokens y finalmente, ponderando la frecuencia de aparición por su importancia en el conjunto total de documentos. A estas matrices se incorporaron las etiquetas del Gold Standard.

Todo lo descrito anteriormente fue programado en lenguaje Python. De igual forma, se desarrolló una interfaz gráfica para etiquetar cada registro del corpus principal en los dos niveles de clasificación para la generación del Gold Standard en el proceso de anotación.

Una vez obtenidas las matrices que representan numéricamente el corpus principal en ambos niveles, se procedió al uso del software WEKA [6] para la implementación de los clasificadores supervisados. Para tales efectos, se consideró el uso de SVM y NB para clasificar los EMR en ambos niveles. El desempeño de ambos clasificadores fue evaluado utilizando el Gold Standard.

El trabajo consideró que todos los campos disponibles en la base de datos fuesen de textos libres y estructurados para la clasificación, pues en ambas modalidades fue posible encontrar

información respecto a la obesidad y sus comorbilidades. Además, cada campo de un registro médico fue considerado un documento en particular.

3.3. Definición de las clases

Para realizar la identificación de obesidad es necesario definir, en primer lugar, las clases a utilizar. En este trabajo, el enfoque fue el de un problema multiclase, las que se sustentan en la Tabla 2.1. Se trabajó en dos niveles de clasificación. El primer nivel consistió en discriminar cada registro médico del corpus principal en función del estado nutricional de los pacientes, según tengan bajo peso, normopeso, sobrepeso u obesidad. Si bien el problema pudo abordarse en forma binaria pues los tres primeros estados nutricionales no corresponden a obesidad, el enfoque propuesto permite un mayor detalle y análisis, pudiéndose agrupar posteriormente. En forma adicional, se añadió la clase "sin información" de manera de eliminar los EMR que fueron recuperados, pero que no hacían referencia de alguna condición en el paciente, es decir, los falsos positivos. Por su parte, el segundo nivel correspondía a EMR en cuyos pacientes se hacía mención a algunos de los tipos de obesidad. Una representación gráfica de las clases se muestra en la Fig. 3.2.

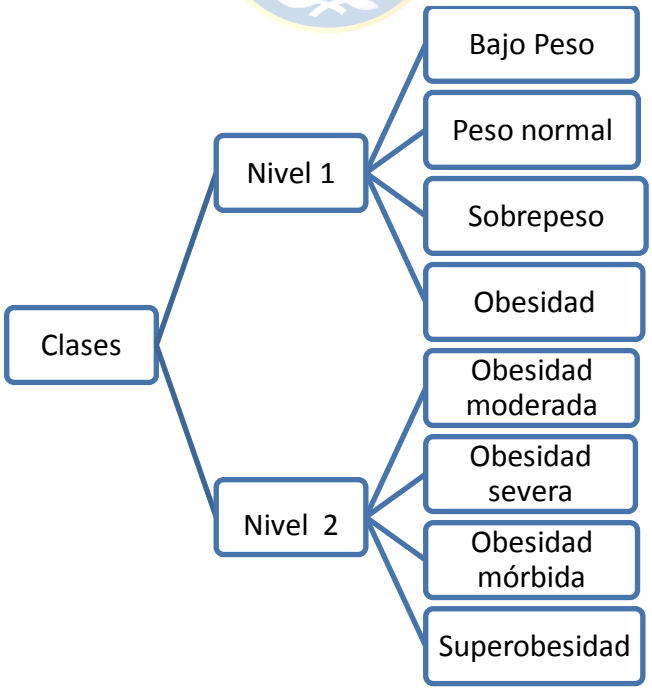

Fig. 3.2 Jerarquía de clases para ambos niveles de clasificación

Nivel 1: Estado nutricional (Identificación de obesidad)

- Bajo peso: Corresponden a los pacientes en cuyo registro médico se especifica un bajo peso
- Normopeso: Corresponden a los pacientes en cuyo registro médico se especifica un peso normal
- Sobrepeso: Corresponden a los pacientes en cuyo registro médico se especifica un sobrepeso o preobesidad
- Obesidad: Corresponden a los pacientes en cuyo registro médico se especifica obesidad

Nivel 2: Identificación de los tipos de obesidad

- Obesidad tipo I: Corresponde a los pacientes en cuyo registro médico se especifica una obesidad moderada o leve
- Obesidad tipo II: Corresponde a los pacientes en cuyo registro médico se especifica una obesidad severa
- Obesidad tipo III: Corresponde a los pacientes en cuyo registro médico se especifica una obesidad mórbida
- Obesidad tipo IV: Corresponde a los pacientes en cuyo registro médico se especifica una superobesidad

3.4. Preprocesamiento

El preprocesamiento de la información es fundamental para la reducción de la brecha existente entre el texto entregado y las capacidades de lectura de los sistemas computacionales. Para tales efectos, se emplearon los siguientes métodos, propios de la minería de textos, en cuanto a la normalización y a la eliminación de stopwords.

3.4.1 Normalización

El proceso de normalización tiene la función de reducir las variaciones léxicas presentes en los textos para aumentar la eficiencia de los sistemas computacionales. Fue necesario procesar cada término contenido en los EMR para aumentar el desempeño de los clasificadores supervisados. Los principales métodos utilizados se describen a continuación:

- Conversión de palabras a minúsculas: Esta conversión garantiza igualdad léxica entre todos los términos de los textos, sin afectar el significado que una palabra tiene en un contexto determinado.
- Saltos de línea y espacio: En los textos libres no existen restricciones para separar palabras u oraciones, en el uso de espacios y saltos de línea. El problema surge al existir un exceso de estos elementos que tienen un valor ASCII (American Standard Code for Information Interchange) y pueden ser interpretados como un término por los sistemas computacionales. Se procedió a eliminar estas desproporciones para la reconstrucción del texto, considerando un elemento espaciador o de salto de línea, según corresponda.
- Signos ortográficos: Mediante expresiones regulares se procedió a la eliminación de todo elemento que no correspondía a una letra, para la obtención de un texto plano. Dentro de los principales elementos se encuentran los signos de exclamación e interrogación, paréntesis, en el caso de las oraciones; o tildes y diéresis en el caso de las palabras.
- Datos numéricos: Para reducir las variaciones numéricas del IMC presente en los textos libres, se reemplazaron todos los valores de este indicador por el límite inferior mostrado en la Tabla 2.1. De esta forma, se favorece el aprendizaje de los clasificadores supervisados, ya que un número, por mínima variación decimal que tenga con otro, es interpretado de manera muy distinta por los clasificadores, aunque puedan corresponder a un mismo grado de obesidad o estado nutricional.

3.4.2 Stopwords

Existen palabras cuyo aporte es escaso en la extracción de información, como artículos, preposiciones, y conjunciones, además de tener las mayores frecuencias de aparición en los textos

[30]. En esta etapa se procede a eliminar dichos elementos sin afectar el contexto en las que están inmersas, excluyendo las negaciones. La detección de negaciones por un sistema de aprendizaje automático es fundamental para rechazar la presencia de un concepto, que por sí solo daría cuenta de una condición en particular. Estas palabras sólo aportan ruido a los sistemas y disminuyen la eficiencia de los clasificadores, pues aumentan significativamente las dimensiones del problema y por ende los tiempos de procesamiento. Mediante la Ley de Zipf es posible representar gráficamente las palabras con mayores frecuencias, y en base a ello, crear una lista de stopwords [30]. Sin embargo, si se tiene conocimiento del dominio se pueden definir ciertas palabras como stopwords e incluirlas en esa lista. El ideal es disponer sólo de palabras que aporten o estén en directa relación con el estudio que da lugar al procesamiento de los textos.

3.5. Procesamiento de la información

Los sistemas computacionales requieren de datos numéricos para su implementación. Por lo tanto, una vez preprocesados los textos, es necesario representarlos de forma que los algoritmos de clasificación supervisada puedan tener un entrenamiento adecuado y puedan ser evaluados. En esta sección se describen las etapas relativas a la recuperación de información y extracción de características bajo el dominio de las comorbilidades de la obesidad.

3.5.1 Recuperación de información

Las palabras claves permiten la identificación de alguna particularidad en los textos y su eficiencia aumenta si su elección se realiza con conocimiento del dominio en estudio. En este caso, se procedió, en primer lugar, a la identificación de las palabras que hacían mención a la obesidad y en contraparte, al sobrepeso, bajo peso y peso normal, incluyendo información del IMC de los textos libres, para determinar los registros que aportaban información al estudio. Por lo tanto, la creación de un corpus bajo este sistema, está directamente relacionado con la cantidad de palabras consideradas como claves. El problema surge cuando un término tiene diversas formas para ser expresado, como es el caso del lenguaje médico, por lo tanto, la estrategia utilizada fue la de analizar en un orden alfabético cada palabra contenida en los textos mediante expresiones regulares

y determinar cuáles hacían mención a alguna condición. Estas expresiones se fundamentaban en las raíces de las palabras, como una forma de reducir las variaciones léxicas registradas en los textos. Esta búsqueda de palabras claves se extendió además a la identificación de las distintas formas de expresar las comorbilidades asociadas a la obesidad. En este sentido, fue necesaria la creación de distintos diccionarios para el almacenamiento y corrección de las palabras que hacían mención en forma explícita a las comorbilidades de la obesidad o en forma indirecta, los tratamientos y recomendaciones que el personal médico indicaba a los pacientes. Además, se analizó el contexto en que estos términos estaban inmersos y mediante el uso de bibliotecas médicas [31] se determinó si la palabra tenía relación con alguna enfermedad. Las enfermedades utilizadas para este estudio son las descritas en [14] y las obtenidas en el proceso de anotación.

Además de las múltiples formas de expresar un término médico, que dificultan el procesamiento de los textos, se presentan los errores de tipo ortográfico. Los diccionarios creados con los términos claves de la obesidad y sus comorbilidades fueron utilizados además para corregir los errores de escritura agrupándolos todos en un término común para cada caso. De esta forma, se pretende favorecer el aprendizaje de los clasificadores supervisados.

Una vez definidas las palabras claves del sistema se procedió a la IR. Esta etapa tiene la función de reducir la cantidad de documentos según un determinado criterio de forma que la IE se reduzca a la consulta de un número menor de documentos [13]. En este trabajo se procedió a una doble IR para la creación del corpus principal, a partir de la base de datos preprocesada mediante términos claves del estado nutricional de los pacientes (ver Fig. 3.2), y un corpus de comorbilidades a partir de un filtrado de documentos recuperados del corpus principal, utilizando términos claves relacionados a las principales enfermedades de la obesidad.

- Corpus principal: Con los términos claves del primer nivel de clasificación se creó un corpus a partir de todos los EMR de la base de datos en una primera recuperación de información. Esta colección de documentos fue utilizada para la clasificación en los dos niveles. Para el primer nivel fueron utilizados todos los EMR recuperados para clasificarlos en algún estado nutricional (ver Fig. 3.2). Por otro lado, para el segundo nivel, sólo fueron clasificados los EMR de este corpus que correspondían a pacientes con obesidad y que hacían mención a algún tipo de ella.

- Corpus de comorbilidades: Esta colección de documentos fue creada como un subconjunto del corpus principal, filtrando los campos de los EMR mediante términos claves asociados a las comorbilidades de la obesidad, incluyendo sus tratamientos e indicaciones. Esta colección de documentos, producto de la segunda recuperación de información, fue utilizada para crear un diccionario de comorbilidades en forma de unigrams o bigrams correspondientes a las características del corpus principal en ambos niveles de clasificación.

La creación del corpus principal también permitió la realización de estudios relativos a la distribución de los registros por campo contenido en los EMR, así como las subespecialidades más influyentes en cuanto a la cantidad de registros donde se realizó la atención médica. Por lo tanto, un sistema basado en el NLP aplicado al sector salud, entrega los cimientos para la toma de decisiones en la creación de políticas públicas de carácter preventivo y facilitaría la asignación adecuada de recursos. Las cifras generadas a partir de este corpus son un complemento para el análisis que se generó a partir del sistema implementado.

3.5.2 Extracción de características

Para un mejor análisis de los textos, es fundamental poder estudiarlo mediante palabras representativas (tokens) o frases con sentido propio [9]. Del mismo modo, los sistemas de aprendizaje automático necesitan de una representación segmentada de los textos que favorezcan su funcionamiento. La extracción de características tiene la función de obtener términos representativos de los textos, lo que es favorecido si se tiene conocimiento del dominio en estudio.

El enfoque propuesto en este trabajo para la extracción de características fue basado en una segmentación utilizando N-grams a través una secuencia de palabras cuyo número de términos está dado por el valor de N. Para tales propósitos, se utilizaron unigrams y bigrams, es decir, el valor de N fue 1 y 2, respectivamente. De esta forma, se buscó estudiar la influencia de la disminución o aumento de la cantidad de dimensiones a través de estos tokens.

La segmentación se realizó en el corpus de comorbilidades para la obtención de tokens en este dominio y su posterior búsqueda en los textos de los EMR del corpus principal para ambos niveles de clasificación. En el segundo nivel, además se incluyeron unigrams y bigrams que describían algún tipo de obesidad y que fueron determinados a partir del estudio previo de los EMR, como el IMC.

3.6. Anotación

Los sistemas de aprendizaje supervisado requieren de un estándar para su evaluación. En este trabajo se procedió a una etiquetación manual de los EMR del corpus principal en ambos niveles de clasificación, previo conocimiento de las características del estudio y un compromiso de confidencialidad por parte de los anotadores, a pesar que los datos fueron facilitados por el Hospital sin identificación de los pacientes. Un correcto proceso de anotación garantiza una real medida del desempeño de los clasificadores, cuyas etapas se detallan a continuación.

3.6.1 Evaluadores

Para la generación del Gold Standard es necesario el criterio de evaluadores con conocimiento en el tema en estudio. Para este trabajo se solicitó la cooperación de dos anotadores quienes tuvieron la tarea de etiquetar cada registro médico del corpus principal, previo conocimiento de las clases de los dos niveles (ver Fig. 3.2). Para asegurar un correcto proceso de anotación, los evaluadores debían cumplir los siguientes requisitos:

- Confidencialidad: En el proceso de anotación cada evaluador adquirió un compromiso de confidencialidad para evitar divulgar la información contenida en los EMR
- Independencia: Cada evaluador debió trabajar de forma independiente, de manera de garantizar una real aproximación de las clases en cada texto.

Finalmente, para la obtención del Gold Standard, un tercer evaluador fue el encargado de resolver las discrepancias y de medir el grado de acuerdo entre los evaluadores. Este Gold Standard fue utilizado para evaluar y entrenar los algoritmos de clasificación supervisada.

3.6.2 Adquisición

Para la anotación de los evaluadores se creó una herramienta diseñada en el entorno gráfico QT 4 Designer y programado en Python, mostrada en la Fig. 3.3. El programa cumple las siguientes funciones:

1. Seguridad: El corpus principal, tras el preprocesamiento, fue encriptado para evitar el acceso al contenido de los EMR sin el uso de la herramienta de anotación. De esta forma, los anotadores sólo podían visualizar los distintos campos estructurados y de texto libre sólo en las sesiones de trabajo.
2. Almacenamiento: Los registros anotados y el progreso general fueron almacenados cada vez que se presentaba un nuevo caso. Así, los evaluadores tenían un control de la cantidad de registros anotados, los que podían clasificar en sesiones distintas.
3. Retroalimentación: Los evaluadores podían incorporar palabras claves relativas a la obesidad y sus comorbilidades como forma de retroalimentación al sistema implementado. Estas palabras claves fueron procesadas y utilizadas para la generación del corpus relativo a las comorbilidades de la obesidad.

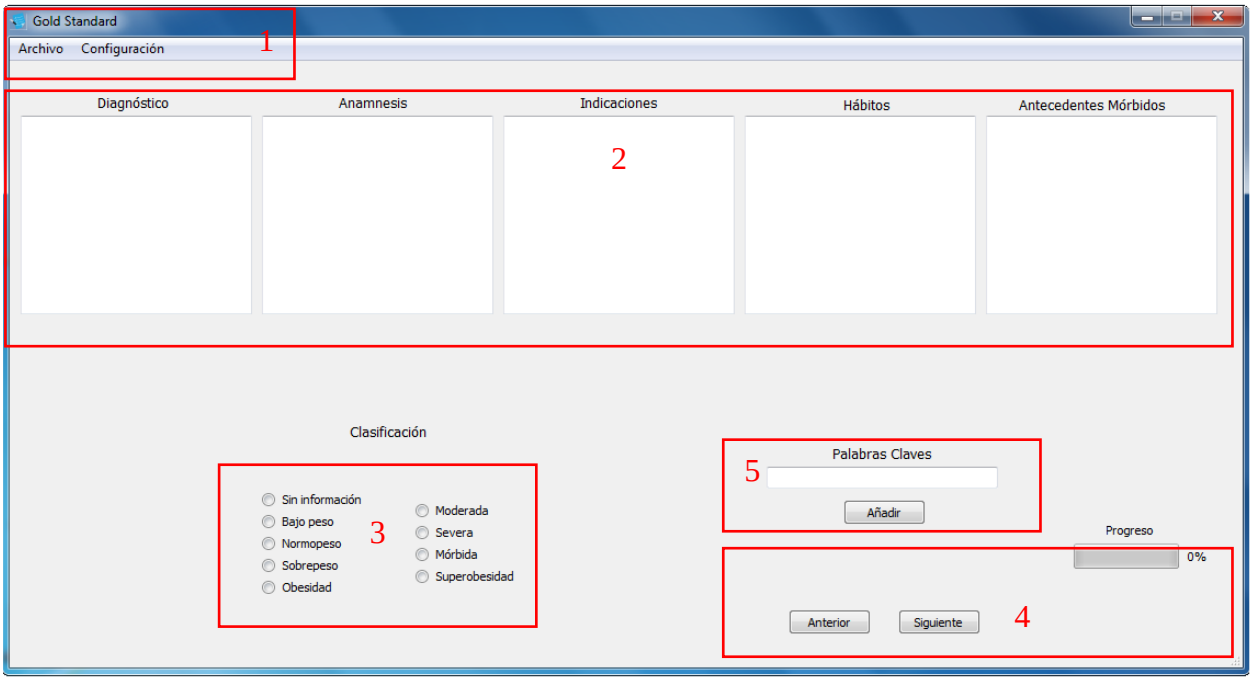

Fig. 3.3 Herramienta de anotación desarrollada. 1) Adquisición de la base de datos y obtención del Gold Standard final, **2)** Visualización de los registros médicos, **3)** Selección de las clases, **4)** Avance y progreso de la anotación, **5)** Ingreso de palabras claves

3.6.3 Nivel de Acuerdo

El grado de acuerdo de un Gold Standard es fundamental para una evaluación fidedigna de los sistemas de aprendizaje automático en su capacidad de identificar las distintas clases. Para una medición cuantitativa de los sistemas implementados, se procedió a utilizar el índice de Kappa de Cohen [15] para dos evaluadores en todas las categorías. Para la distribución de la cantidad de EMR etiquetados en cada clase, se procedió a la utilización de la Tabla 3.1.

TABLA 3.1 Distribución de frecuencias para cada clase según lo anotado por los evaluadores

		Evaluador 1				
		Sin información	**Clase 1**	**Clase 2**	**Clase 3**	**Clase 4**
Evaluador 2	**Sin información**	$f_{0,0}$	$f_{0,1}$	$f_{0,2}$	$f_{0,3}$	$f_{0,4}$
	Clase 1	$f_{1,0}$	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$
	Clase 2	$f_{2,0}$	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$
	Clase 3	$f_{3,0}$	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$
	Clase 4	$f_{4,0}$	$f_{4,1}$	$f_{4,2}$	$f_{4,3}$	$f_{4,4}$

3.7. Clasificación

La clasificación de los EMR para la identificación de la obesidad y sus tipos, requiere de una representación numérica de los textos para el aprendizaje de los algoritmos supervisados. Una vez construidas las matrices binarias, TF y TF-IDF y obtenido el Gold Standard, se procedió a implementar los clasificadores supervisados para luego ser evaluados.

3.7.1 Representación de la información

En data mining, y a diferencia de su homólogo en textos, se dispone de datos estructurados, numéricos, categóricos, reales o nominales, dispuestos a ser utilizados sin mayores procesamientos.

Los textos escritos en lenguaje natural deben ser adecuados y representados en una estructura numérica para que los sistemas computacionales puedan procesarlos. Para tales efectos, se dispuso de tres formas matriciales de representar el corpus principal, contabilizando en cada registro la aparición de los tokens obtenidos de la segmentación del corpus de comorbilidades de la obesidad.

El primer método utilizado para la representación de la información fue mediante un enfoque binario, es decir, considerando sólo la presencia o ausencia de los tokens en cada registro del corpus principal. De esta forma, un elemento que aparece más de una vez en un texto tiene la misma importancia que otro que sólo lo hace en una oportunidad.

En segundo lugar, para una comparación del modelo binario se procedió a considerar la frecuencia de aparición de cada término en el corpus, denominado TF (Term Frequency).

Los métodos antes descritos obedecen al enfoque BoW, donde el orden de los términos no tiene relevancia para tareas de clasificación de documentos, ni se hace uso de la gramática. Su representación es denominada Modelo de Espacio Vectorial [30].

Tanto el pesado booleano o TF no consideran la relevancia global de los términos en el total de documentos. El tercer método utilizado para la representación de la información fue TF-IDF (Term Frequency – Inverse Document Frequency) para ponderar cada token contenido en los registros según su frecuencia global en el corpus principal. De esta forma, un elemento que aparece en cada documento del corpus tendrá una menor relevancia que otro que sólo lo hace en menos ocasiones (IDF).

Matriz binaria

Corresponde al método más sencillo de representación de los documentos mediante los términos de la colección. Una vez obtenido el vocabulario, se procede a la búsqueda de la presencia o ausencia de ellos en el corpus registrando un 1 o un 0 en la matriz de representación, respectivamente. Por lo tanto, no existe ponderación alguna en función de la frecuencia de aparición de los términos en los EMR.

Matriz TF-IDF

TF-IDF es una medida de ponderación de los términos de un texto para determinar su relevancia. Esta técnica es utilizada tanto en IR como en minería de textos para la representación de los términos en un modelo de espacio vectorial [32].

Term Frequency (TF)

TF corresponde a la frecuencia de aparición de un término en un documento en particular. Intuitivamente, una palabra que aparece más veces en un documento pudiera ser más importante que otra con una frecuencia menor. Los valores calculados para cada término representativo son normalizados por la frecuencia máxima en el documento [30], según:

$$TF_{i,j} = \frac{f_{i,j}}{\max f_{i,j}} \tag{3.1}$$

Inverse Document Frequency (IDF)

Que un término tenga la mayor frecuencia de aparición en un documento en particular no es condición suficiente para ser el más relevante en el corpus. Bajo esta lógica, se procede a ponderar cada término t_i, en función de la cantidad total de documentos (Q) [30], según:

$$IDF_{t_i} = log_{10}(\frac{Q}{n_i + 1}) \tag{3.2}$$

Donde,

n_i: Número de documentos donde aparece el término t_i

Finalmente, TF-IDF es el producto de las ecuaciones 3.1 y 3.2, resultando el peso $w(t_i, d_j)$ de un término en un documento según:

$$w(t_i, d_j) = Tf_{t,d}\ x\ IDF_t \tag{3.3}$$

Para cada una de las representaciones antes descritas se procedió a la construcción de distintas matrices $M_{mxn,}$ donde m correspondía al número de EMR, incluidas las etiquetas y n el número de tokens, incluidas las clases para ambos niveles, tal como se muestra en la Tabla 3.2.

TABLA 3.2 Matriz implementada para cada tipo de representación para su utilización en los clasificadores supervisados

Registro	**Token$_1$**	⋯	**Token$_n$**	**Nivel 1**	**Nivel 2**
R_1	$f_{0,0}$	⋯	$f_{0,n-2}$	$C1_{1-4}$	$C2_{1-4}$
⋮	⋮	⋱	⋮	⋮	⋮
R_m	$f_{m-1,0}$	⋯	$f_{m,n-2}$	⋮	⋮

Según sea el caso, es posible eliminar alguna de las dos últimas columnas para trabajar con las clases en el nivel 1 o 2.

3.7.2 Implementación de los clasificadores

Para la clasificación de los EMR del corpus principal se utilizó el software WEKA, de código libre y desarrollado en Java por la Universidad de Waikato, Nueva Zelanda [6]. El software permite realizar reglas de asociación, clustering y clasificación. Además, incorpora filtros para los datos de entrada y su visualización al igual que para los resultados obtenidos.

Para la selección del set de pruebas y entrenamiento se utilizó una validación cruzada con K = 10 [16]. Con este método de selección se pretende obtener una mejora en el aprendizaje de los clasificadores supervisados, evitando problemas de representación de las clases.

Los clasificadores utilizados para cada nivel fueron los descritos en el Marco Teórico, SVM y NB, previa segmentación del corpus de comorbilidades en unigrams y bigrams para la extracción de características, como se muestra en la Fig. 3.4.

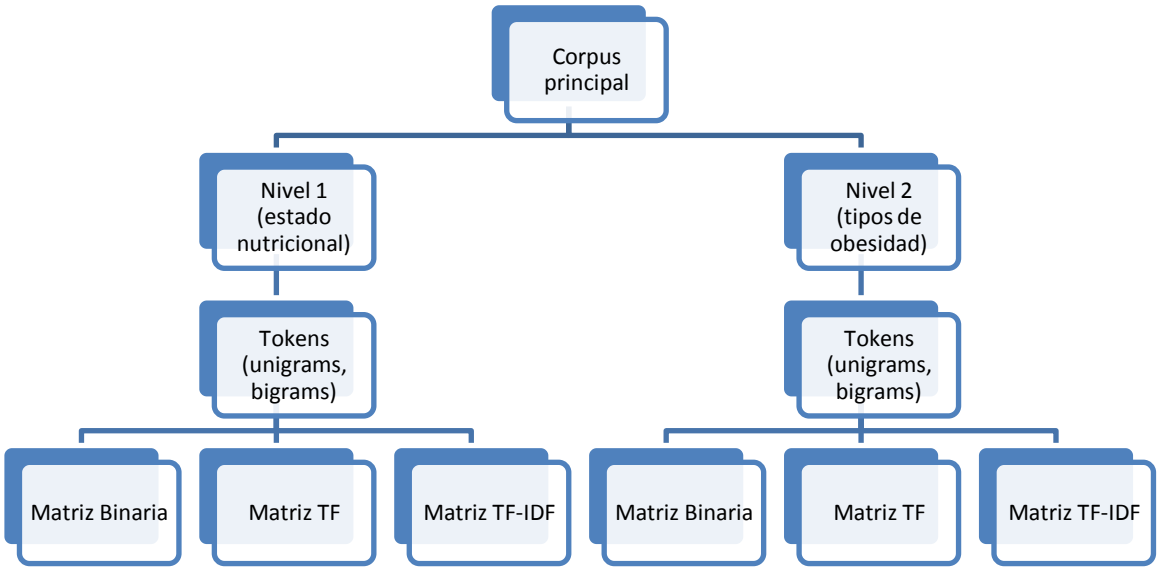

Fig. 3.4 Niveles de clasificación y representación matricial del corpus principal en función de los tokens obtenidos por segmentación del corpus de comorbilidades. Este esquema se aplica en ambos clasificadores supervisados

3.7.3 Evaluación

La evaluación de un algoritmo de clasificación supervisada busca medir la capacidad predictiva del modelo generado durante el aprendizaje en la etapa de entrenamiento, comparando la clasificación efectuada en la fase de pruebas con las etiquetas obtenidas en la anotación. Sin embargo, es posible incluir otros indicadores como el tiempo de ejecución para la construcción y uso del modelo, y la capacidad para el manejo de ruido o valores atípicos.

En este trabajo, se procedió a evaluar la capacidad predictiva de los clasificadores SVM y NB mediante la matriz de confusión (ver Tabla 3.3). Esta última es una representación matricial de los verdaderos positivos (TP, del inglés True Positive) y negativos (TN, del inglés True Negative), correspondientes a los aciertos del clasificador, y los falsos positivos (FP, del inglés False Positive) y negativos (FN, del inglés False Negative), que corresponden a errores de clasificación [15].

TABLA 3.3 Matriz de confusión

		Clasificación	
		Positivo	Negativo
Real	Positivo	TP	FN
	Negativo	FP	TN

Mediante la matriz de confusión es posible el cálculo de diversos indicadores para la evaluación del desempeño, destacando el Accuracy (Exactitud), Precision y Recall.

- **Accuracy**: Corresponde a la relación entre la cantidad de clasificaciones correctas y el total de casos. Esta tasa de aciertos es definido según:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.4}$$

- **Precision**: Corresponde a una medida de la exactitud para una estimación más acabada de los aciertos del clasificador, ponderándola con los falsos positivos. Es definido según:

$$Precision = \frac{TP}{TP + FP} \tag{3.5}$$

- **Recall**: Corresponde a una medida de tasa de aciertos positivos también denominada sensibilidad. Se define según:

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

Si bien estas mediciones obedecen a problemas de clasificación binaria, se extiende su uso en problemas de índole multiclase mediante una matriz de contingencia, que corresponde a una extensión de la matriz de confusión en cuya diagonal se indican los TP. Para los demás valores, se analiza cada clase contrarrestándola con el resto.

Capítulo 4. Resultados

En este capítulo se describen los resultados obtenidos en el procesamiento de los datos que dieron lugar a la generación del corpus utilizado para clasificación. Además, se detallan las particularidades encontradas en los textos que dieron lugar a la generación del corpus relativo a las comorbilidades asociadas a la obesidad. Finalmente, se muestran los resultados obtenidos tras la utilización de SVM y NB para una posterior comparación de sus rendimientos.

4.1. Análisis de los datos

El primer paso para el procesamiento de los datos es la realización de un estudio de las singularidades presentes en los textos. El primer aspecto a considerar es la similitud en el uso de los distintos tipos de campos por parte del personal médico. En ciertas ocasiones, los campos estructurados Hábitos y Antecedentes Mórbidos son utilizados para el ingreso de texto libre con las implicancias de tipo ortográfico que ello conlleva. El campo de texto libre Diagnóstico está compuesto por una estructura donde se hace mención a un diagnóstico principal y secundario. En ciertas ocasiones Diagnóstico tiene las características de un campo estructurado al sólo incorporar el nombre de una enfermedad, es decir, no existe un orden estricto por parte del personal médico para registrar los antecedentes de una visita médica.

El lenguaje médico tiene aspectos que lo caracterizan y diferencian del resto [9]. Los textos libres en esta base de datos no son extensos, reflejando un alto grado de abstracción por parte de su redactor. De igual forma, estos textos contienen información de distintas visitas al hospital de los pacientes para evaluar su evolución médica, lo que puede generar ambigüedades de interpretación dada la temporalidad en que ocurrieron los hechos.

Finalmente, otro aspecto a considerar es la presencia de datos incompletos en todos los campos de la base de datos, concentrándose muchas veces la información en una sola sección.

4.2. Recuperación de información

Tras procesar todos los datos y generar el corpus principal a través de la primera recuperación de información mediante palabras claves relativas al estado nutricional de los pacientes, se obtuvo 2462 EMR, representando sólo un 3,74% del total, tal como se muestra en la Fig. 4.1. Los documentos recuperados en esta primera etapa provienen en su mayoría del campo Diagnóstico, representando un 60,94% del total, es decir, el mayor procesamiento fue efectuado en los textos libres de los EMR.

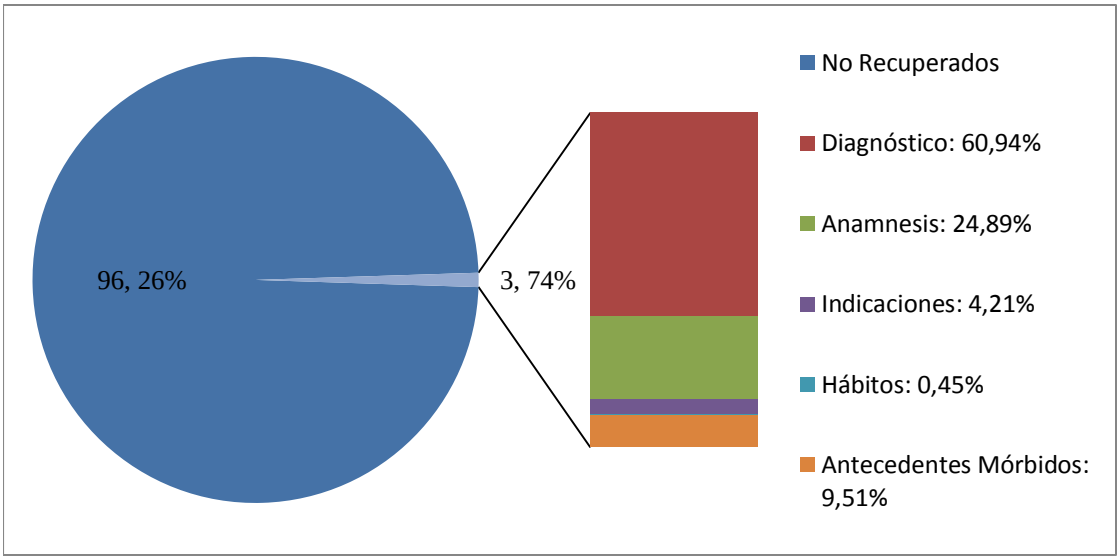

Fig. 4.1 Recuperación de información y distribución de la cantidad de documentos por campo en los EMR

Del total de EMR recuperados que hacían mención de la obesidad en los pacientes, sólo el 21,52% corresponde al género masculino. Esta cifra va en dirección con la entregada por la Encuesta Nacional de Salud del año 2010 [2], donde el sexo femenino es quien posee las cifras más altas de obesidad en todos los grupos etarios de la población Chilena.

El segundo proceso de recuperación, para la creación del corpus de comorbilidades de la obesidad, contuvo 4761 documentos, representando un 54,89 % del total. Para tales efectos, se procedió a la utilización de términos claves (ver Tabla 4.1), obtenidos mediante un estudio de los documentos recuperados y los términos reportados en el proceso de anotación.

TABLA 4.1 TÉRMINOS CLAVES ASOCIADOS A LAS COMORBILIDADES DE LA OBESIDAD

Enfermedad	Términos relacionados
Hipertensión arterial	HTA, Nifedipino, Furosemida, Atenolol (ATN), Enalapril, Losartán, Propanolol, Amlodipino, Olmesartan, Metildopa
Diabetes Mellitus	DM, Síndrome metabólico; Diabetes Tipo 1, DM 1, DMIR (insulino requirente, resistente), IR (insulino resistente), Insulina, NPH (Neutral Protamine Hagedorn); Diabetes tipo 2, DM 2, MTF (Metformina); Diabetes gestacional
Dislipidemia	DLP, hiperlipidemia, hipertrigliceridemia, hipercolesterolemia, HDL, LDL, Atorvastatina (ATV)
Asma	Salbutamol, Salmetero, Teofilina
Depresión	Sertralina, Amitriptilina, Fluoxetina, Paroxetina
Cálculos biliares	Colelitiasis
Osteoartritis	Meloxicam
Enfermedad por reflujo gastroesofágico	ERGE, Omeprazol, Ranitidina, Lanzopralol
Gota	Alopurinol
Apnea obstructiva del sueño	SAHOS, SAOS, CPAP
Hipotirodismo	T3 (Triyodotironina), T4 (Tiroxina), THS, Levotiroxina
Enfermedades cardiovasculares	ICC (Insuficiencia Cardiaca Congestiva), Carvedilol, Espironolactona, estenosis aórtica, enfermedad vascular periférica; enfermedad cardiovascular aterosclerótica ateroesclerosis, Aspirina (AAS), IAM (Infarto Agudo al Miocardio), Sten, angioplastía; Insuficiencia venosa, várices (VSEXT), TVP (Trombosis Venosa Profunda), Clopidogrel (Plávix) Digoxina, CV (Cardiovascular), TACO (Tratamiento Anticoagulante), doppler venoso, soporte elástico
Cushing	Cortisol

Por otro lado, se procedió a analizar las subespecialidades bajo el dominio de la obesidad en función de las comorbilidades asociadas y descritas este trabajo. La Fig. 4.2 muestra precisamente la distribución antes mencionada y generada a partir de una extracción textual de las enfermedades. Las subespecialidades con mayor influencia en las comorbilidades de la obesidad son Cardiología Adulto, Broncopulmonar Adulto y Cirugía Adulto, donde es posible visualizar las mayores frecuencias de las enfermedades diabetes mellitus, hipertensión arterial, hipotiroidismo, dislipidemia en el primer caso, insuficiencia cardiaca y enfermedad vascular, se añaden a las anteriores para la subespecialidad Broncopulmonar Adulto y colelitiasis para el caso de Cirugía Adulto. Las demás enfermedades se presentan en menor frecuencia como gota y cushing, no registrándose casos de enfermedad aterosclerótica.

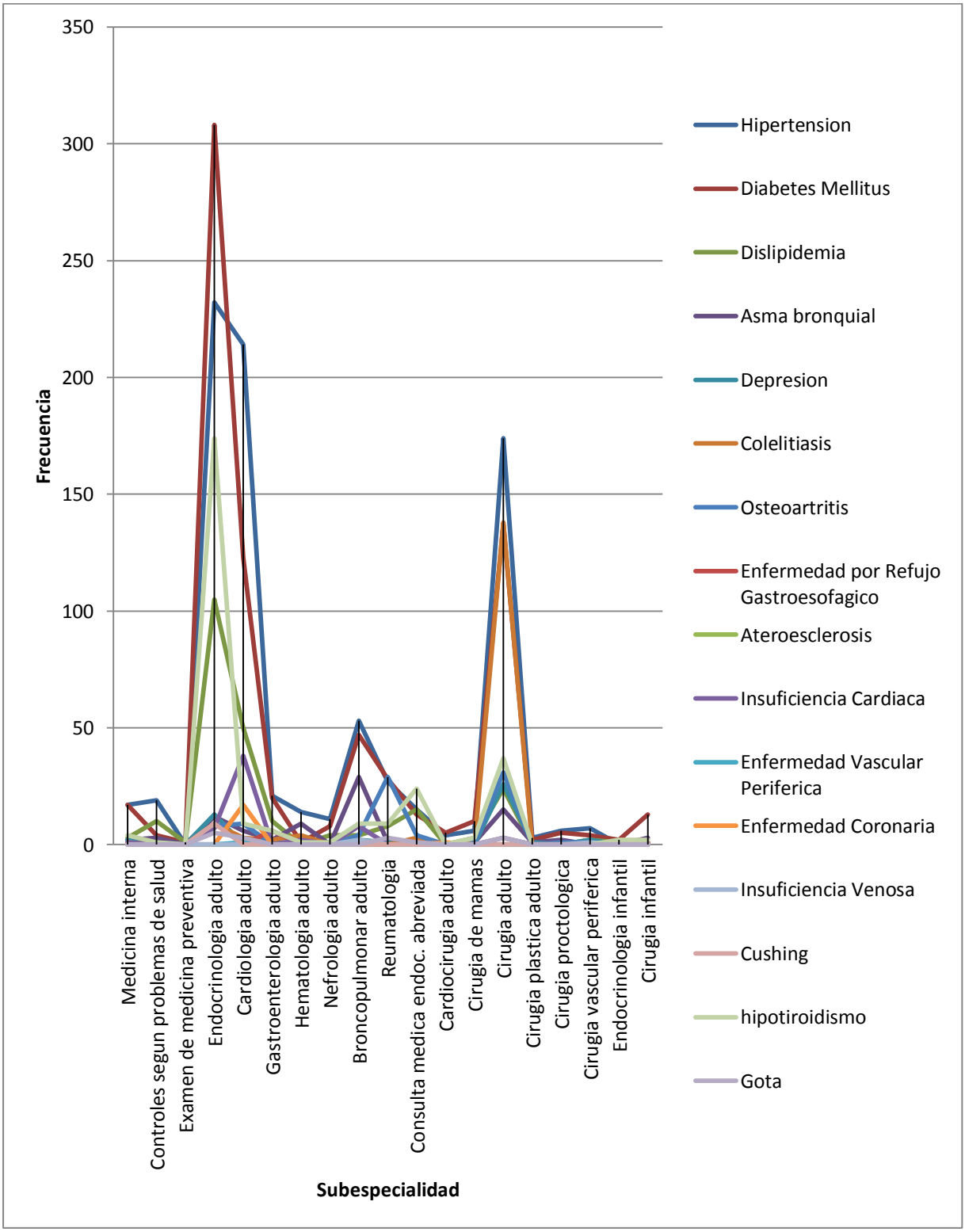

Fig. 4.2 Distribución de las comorbilidades de la obesidad por subespecialidad médica

4.3. Extracción de características

En base a los documentos contenidos en el corpus de comorbilidades de la obesidad, se procedió a segmentar los textos mediante unigrams y bigrams. Para el primer nivel de clasificación, utilizando unigrams se obtuvo 1069 tokens, mientras que para la segmentación basada en bigrams, el número aumenta a 1071. Para la creación de estos términos representativos fueron excluidas las palabras que hacían mención explícita a la obesidad y el sobrepeso, de forma de establecer una relación directa entre las enfermedades asociadas a ella y los estados nutricionales de los pacientes.

Para el segundo nivel de clasificación, se añadieron además los términos relacionados a los tipos de obesidad, tanto para unigrams y bigrams. Estas sentencias fueron determinadas a partir del estudio de los textos, debiéndose incorporar representaciones numéricas. Los tokens obtenidos en este nivel, para el caso de unigrams, fue 1093; mientras que para el caso de bigrams, el número de elementos representativos fue 1119.

4.4. Anotación

Una vez recuperados 2462 EMR para la obtención del corpus principal, se procedió a etiquetar cada instancia según el nivel de clasificación. En este estudio se incorporaron todos los campos de los EMR, donde cada evaluador debió etiquetarlos de forma textual o intuitiva utilizando palabras claves asociadas a las clases, el IMC o denominaciones alternativas para cada condición médica en ambos niveles de clasificación.

Junto con el grupo de evaluadores se procedió a la elaboración del Gold Standard. El número total de EMR debidamente etiquetados para el primer nivel de clasificación fue 2412, debido a que fueron excluidas las instancias, que si bien mencionaban algún término clave en los textos no hacían mención de alguna condición en los pacientes, sino en los antecedentes familiares u obedecían a un estado de salud pasado sin un detalle actual del mismo. Estos falsos positivos fueron determinados gracias a la etiqueta “sin información” que los evaluadores disponían en el software de anotación.

. La cantidad de EMR utilizados para la identificación de obesidad, junto a los demás estados nutricionales de los pacientes, se muestran en la Fig. 4.3 donde es posible distinguir una gran cantidad de instancias asociadas a la enfermedad crónica en estudio, en desmedro de las demás clases.

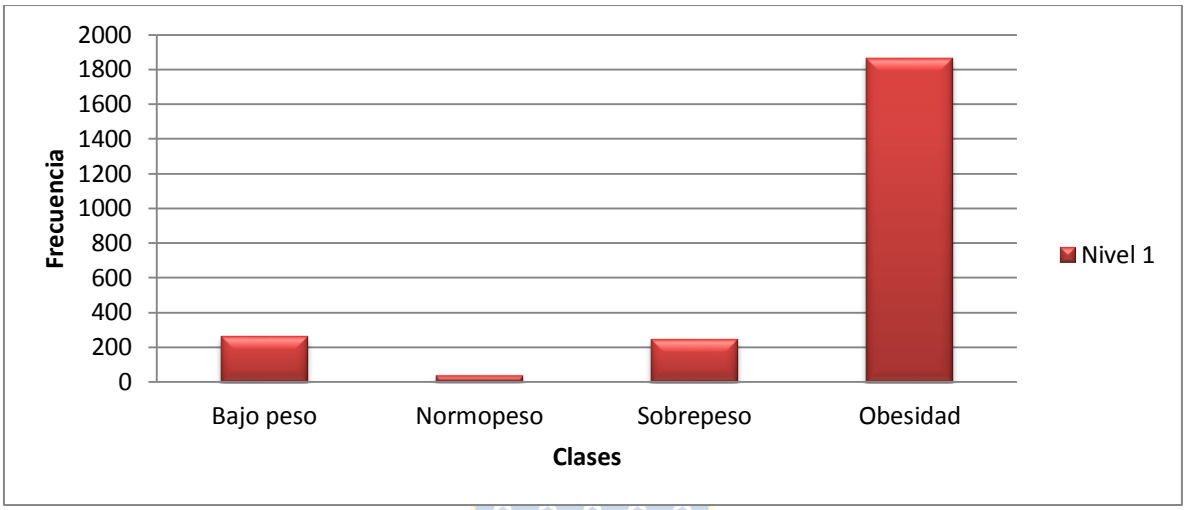

Fig. 4.3 Distribución de las clases para el primer nivel de clasificación

Por su parte, para el segundo nivel de clasificación, fueron debidamente etiquetados 920 EMR que hacían mención a algún tipo de obesidad, representando un 38,14% del total. Estas cifras denotan la baja importancia que se le asigna al registro de la obesidad de los pacientes, por lo que el uso de información adicional, como sus comorbilidades, pudiera ser un buen indicador del impacto que esta enfermedad tiene en la población.

La distribución de las clases del segundo nivel de clasificación se muestra en la Fig. 4.4, donde es posible identificar que la gran cantidad de EMR que hacían mención a un tipo de obesidad corresponden a pacientes con obesidad mórbida.

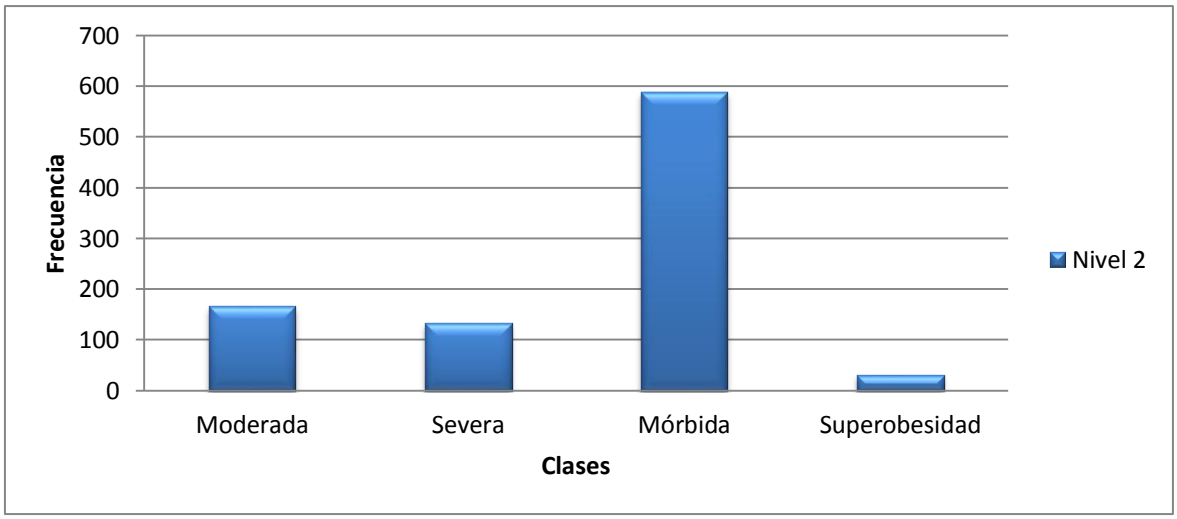

Fig. 4.4 Distribución de las clases para el segundo nivel de clasificación

Para ambos niveles de clasificación, el Gold Standard fue determinado en función de los acuerdos para cada instancia, evaluándose los registros donde no los hubo.

Finalmente, se obtuvo un índice de Kappa para ambos niveles de clasificación. El valor obtenido en el primer nivel fue igual a 0,97, mientras que para la clasificación de los tipos de obesidad, el índice obtenido fue 0,96. Por lo tanto, según la Tabla 2.5, el grado de acuerdo entre los evaluadores corresponde a un acuerdo casi perfecto.

4.5. Clasificación

Los resultados obtenidos de los clasificadores implementados se describen en esta sección. En primer lugar se muestran los aciertos globales de NB y SVM para un posterior análisis por categoría, incorporando medidas de Precision y Recall. Para realizar esta clasificación, se seleccionó el set de pruebas y entrenamiento mediante una validación cruzada con un valor de k igual a 10. De esta forma, se consideró todo el conjunto de datos, redefiniéndolos cada vez que se iteraba sobre ellos, buscando evitar problemas de underfitting y overfitting [16] por una inadecuada distribución de las clases o selección de la cantidad de los EMR para pruebas y entrenamiento.

4.5.1 Nivel 1: Estado nutricional (Identificación de obesidad)

Los aciertos obtenidos por el clasificador NB se muestran en la Tabla 4.2 para las representaciones Binaria, TF y TF-IDF utilizando unigrams y bigrams. Es posible visualizar que el mejor desempeño para una segmentación con unigrams se obtiene con el uso de una representación matricial TF o TF-IDF alcanzando un 84,41% de exactitud para este nivel. Por otro lado, los mejores resultados utilizando bigrams se obtienen con el uso de la matriz binaria con un 84,49%.

TABLA 4.2 PORCENTAJE DE ACIERTOS EN EL PRIMER NIVEL DE CLASIFICACIÓN MEDIANTE NB

Segmentación (N)	Representación matricial	Accuracy (%)
Unigram	Binaria	83,67
	TF	84,41
	TF-IDF	84,41
Bigram	Binaria	84,49
	TF	84,16
	TF-IDF	84,16

El porcentaje de aciertos obtenidos por SVM para este nivel de clasificación se muestran en la Tabla 4.3. El mejor desempeño se obtuvo al utilizar una representación binaria de la información con un 87,65% de exactitud para una segmentación basada en unigrams. De igual forma, mediante el uso de bigrams se obtuvo el mejor desempeño del clasificador, alcanzando una exactitud de 89,10% con una matriz binaria.

TABLA 4.3 PORCENTAJE DE ACIERTOS EN EL PRIMER NIVEL DE CLASIFICACIÓN MEDIANTE SVM

Segmentación (N)	Representación matricial	Accuracy (%)
Unigram	Binaria	87,65
	TF	86,02
	TF-IDF	86,02
Bigram	Binaria	89,10
	TF	88,47
	TF-IDF	88,47

4.5.2 Nivel 2: Tipos de obesidad

Para el segundo nivel de clasificación, los aciertos obtenidos mediante NB se muestran en la Tabla 4.4. Al utilizar unigrams, los mejores desempeños de este clasificador se obtienen con las representaciones TF y TF-IDF con un 72,39%. Por otro lado, al utilizar bigrams los mejores resultados se obtienen con el uso de una representación binaria de la información, alcanzado una exactitud de 82,17%.

TABLA 4.4 PORCENTAJE DE ACIERTOS EN EL SEGUNDO NIVEL DE CLASIFICACIÓN MEDIANTE NB

Segmentación (N)	Representación matricial	Accuracy (%)
Unigram	Binaria	72,28
	TF	72,39
	TF-IDF	72,39
Bigram	Binaria	82,17
	TF	81,30
	TF-IDF	81,30

En este mismo nivel de clasificación, los aciertos obtenidos con SVM se muestran en la Tabla 4.5. Para el caso de una segmentación con unigrams, el mejor desempeño fue alcanzado con una representación binaria de la información con un 80,76%. Por su parte, mediante el uso de bigrams, los mejores desempeños fueron obtenidos utilizando las representaciones TF y TF-IDF con una exactitud igual a 93,80%.

TABLA 4.5 PORCENTAJE DE ACIERTOS EN EL SEGUNDO NIVEL DE CLASIFICACIÓN MEDIANTE SVM

Segmentación (N)	Representación matricial	Accuracy (%)
Unigram	Binaria	80,76
	TF	80,33
	TF-IDF	80,33
Bigram	Binaria	93,70
	TF	93,80
	TF-IDF	93,80

4.5.3 Resumen de los resultados

Tras aplicar distintos tipos de segmentación en el corpus de comorbilidades para una posterior representación en los textos de cada registro médico del corpus principal en ambos niveles de clasificación, se tiene que los mejores resultados en el primer nivel de clasificación corresponden al uso de SVM, utilizando una representación binaria de la información con una segmentación basada en bigrams, como se muestra en la Tabla 4.6. La capacidad de este clasificador para identificar en los textos las comorbilidades de la obesidad se ve favorecida al utilizar bigrams, pues muchas de estas enfermedades se presentan en los textos con una secuencia de más de una palabra. Por su parte, si bien NB obtuvo un rendimiento inferior, su desempeño lo sitúa como una buena alternativa, no importando mayormente el tipo de segmentación o representación matricial utilizada.

TABLA 4.6 COMPARACIÓN DE EXACTITUD EN EL NIVEL 1

Clasificador	Segmentación	Representación	Accuracy (%)
SVM	Bigram	Binaria	89,10
NB	Bigram	Binaria	84,49
	Unigram	TF	84,41
	Unigram	TF-IDF	84,41
	Bigram	TF	84,16
	Bigram	TF-IDF	84,16

Para un análisis más minucioso del sistema implementado, se procedió a analizar cada clase de la Tabla 4.6 en sus respectivas matrices de confusión.

En la Tabla 4.7 se muestra la matriz de confusión de la categoría “bajo peso” donde se aprecia que SVM supera en un 0,38% al mejor resultado de exactitud obtenido por NB. Los altos porcentajes de exactitud alcanzados se explican por la gran cantidad de verdaderos negativos presentes en ambos clasificadores. De los 264 EMR que pertenecen a esta categoría (ver Fig. 4.3), SVM obtuvo una mayor cantidad de verdaderos y falsos positivos, y menor cantidad de falsos negativos que NB lo que se traduce en un menor valor de Precision y mayor Recall.

TABLA 4.7 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE BAJO PESO

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	Binaria	258	2145	3	6	99,63	98,85	97,73
NB	Bigram	Binaria	248	2146	2	16	99,25	99,20	93,94
	Unigram	TF	233	2144	4	31	98,55	98,31	88,26
	Unigram	TF-IDF	233	2144	4	31	98,55	98,31	88,26
	Bigram	TF	228	2148	0	36	98,51	100,00	86,36
	Bigram	TF-IDF	228	2148	0	36	98,51	100,00	86,36

Por otro lado, la Tabla 4.8 muestra la matriz de confusión de la clase "normopeso", que además posee la menor representación de los EMR en este nivel (ver Fig. 4.3). SVM tiene un mejor desempeño que NB si se comparan los mejores porcentajes de exactitud, superándolo en un 1,57%. La alta exactitud obtenida en ambos clasificadores se debe a la gran cantidad de verdaderos negativos detectados. Si sólo se consideran los 37 EMR asociados a esta categoría y los verdaderos positivos detectados, SVM posee un rendimiento muy superior a NB quien prácticamente es incapaz de identificar "normopeso". Esto último queda en evidencia en los bajos porcentajes de Precision y Recall que NB obtuvo en esta categoría.

TABLA 4.8 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE NORMOPESO

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	Binaria	29	2371	4	8	99,50	87,88	78,38
NB	Unigram	TF	4	2358	17	33	97,93	19,05	10,81
	Unigram	TF-IDF	4	2358	17	33	97,93	19,05	10,81
	Bigram	TF	1	2356	19	36	97,72	5,00	2,70
	Bigram	TF-IDF	1	2356	19	36	97,72	5,00	2,70
	Bigram	Binaria	1	2327	48	36	96,52	2,04	2,70

En la clase "sobrepeso" cuya matriz de confusión se muestra en la Tabla 4.9, SVM supera en un 1,7% de exactitud al mejor resultado obtenido por NB. Si bien NB obtuvo una mayor cantidad de verdaderos positivos, SVM lo supera en verdaderos negativos cuya cantidad supera con creces los 247 EMR etiquetados en esta categoría. La baja cantidad de verdaderos positivos y la gran cantidad de falsos positivos y negativos en ambos clasificadores queda expuesto en los porcentajes de Precision y Recall observados.

TABLA 4.9 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE SOBREPESO

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	Binaria	67	2100	65	180	89,84	50,76	27,12
NB	Bigram	Binaria	97	2029	136	150	88,14	41,63	39,27
	Bigram	TF	87	2025	140	160	87,56	38,33	35,22
	Bigram	TF-IDF	87	2025	140	160	87,56	38,33	35,22
	Unigram	TF	85	2023	142	162	87,40	37,44	34,41
	Unigram	TF-IDF	85	2023	142	162	87,40	37,44	34,41

Finalmente, en la Tabla 4.10 se muestra la matriz de confusión para la clase "obesidad "de la comparación de exactitud de la Tabla 4.6. Al clasificar "obesidad" se tiene un comportamiento similar a los aciertos del sistema global, pero los valores de exactitud obtenidos para cada representación matricial en ambas segmentaciones son superiores en los dos clasificadores. Como ha ocurrido en las categorías anteriores, nuevamente SVM obtiene un mejor porcentaje de exactitud que NB, superando a su mejor desempeño en un 4,15%. La principal diferencia entre ambos clasificadores corresponde a la cantidad de verdaderos positivos detectados en esta categoría, que posee la mayor representación de EMR con 1864 instancias, además de la cantidad de falsos negativos donde SVM amplía su brecha con NB a 4,35% si se comparan los porcentajes de Recall más altos.

La correcta detección de los verdaderos negativos en las sentencias es fundamental para clasificar de forma correcta los documentos, lo que se acentúa si no se dispone de negaciones explícitas de las condiciones estudiadas en este nivel. Por otro lado, las enfermedades asociadas a la obesidad no son de su exclusividad, por lo que pueden estar presentes en otras categorías afectando el desempeño de los clasificadores durante su aprendizaje.

TABLA 4.10 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE OBESIDAD

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	Binaria	1795	357	191	69	89,22	90,38	96,30
NB	Bigram	Binaria	1692	360	188	172	85,07	90,00	90,77
	Unigram	TF	1714	335	213	150	84,95	88,95	91,95
	Unigram	TF-IDF	1714	335	213	150	84,95	88,95	91,95
	Bigram	TF	1714	325	223	150	84,54	88,49	91,95
	Bigram	TF-IDF	1714	325	223	150	84,54	88,49	91,95

En el segundo nivel de clasificación, tal como se muestra en la Tabla 4.11, el uso de segmentación basada en bigrams es factor fundamental para la identificación de los tipos de obesidad. Al igual que las enfermedades, la clasificación de la obesidad está sujeta a términos compuestos por palabras y números, como por ejemplo el IMC, por lo que la transformación de estos valores en los textos al límite inferior de cada grado de obesidad (ver Tabla 2.1) favoreció el desempeño de ambos clasificadores al reducir las variaciones generadas principalmente por los números decimales.

TABLA 4.11 COMPARACIÓN DE EXACTITUD EN EL NIVEL 2

Clasificador	Segmentación	Representación	Accuracy (%)
SVM	Bigram	TF	93,80
	Bigram	TF-IDF	93,80
	Bigram	Binaria	93,70
NB	Bigram	Binaria	82,17

De igual forma que en el primer nivel de clasificación, se procedió a analizar en forma detallada los mejores resultados de exactitud de la Tabla 4.11 para todas las categorías.

En la Tabla 4.12 se muestra la matriz de confusión de la clasificación de "obesidad moderada" donde, al igual que en el sistema global, SVM tiene el mejor desempeño en la clasificación. A diferencia de NB, SVM obtuvo una mayor cantidad de verdaderos positivos y negativos y una menor cantidad de falsos positivos y negativos, lo que favoreció la diferencia de exactitud de 5,55% existente entre los mejores desempeños de ambos clasificadores y a un mayor porcentaje de Precision y Recall a favor de SVM.

TABLA 4.12 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE OBESIDAD MODERADA

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	TF	154	745	9	12	97,72	94,48	92,77
	Bigram	TF-IDF	154	745	9	12	97,72	94,48	92,77
	Bigram	Binaria	151	748	6	15	97,71	96,18	90,96
NB	Bigram	Binaria	130	718	36	36	92,17	78,31	78,31

En la Tabla 4.13 se muestra el detalle de la matriz de confusión de la clasificación de "obesidad severa" donde nuevamente SVM supera en exactitud a NB en un 6,74%. Al igual que en

la clase anterior, SVM posee una mayor cantidad de verdaderos negativos y positivos, lo que favorece su exactitud, y una menor cantidad de falsos positivos y negativos, lo que implica un mejor porcentaje de Precision y Recall que su símil NB.

TABLA 4.13 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE OBESIDAD SEVERA

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	TF	118	780	7	15	97,61	94,40	88,72
	Bigram	TF-IDF	118	780	7	15	97,61	94,40	88,72
	Bigram	Binaria	120	778	9	13	97,61	93,02	90,23
NB	Bigram	Binaria	84	752	35	49	90,87	70,59	63,16

La Tabla 4.14 muestra en detalle el desempeño de SVM y NB para la clasificación de "obesidad mórbida", que además posee la mayor cantidad de EMR etiquetados en este nivel. SVM obtuvo el mejor desempeño de exactitud, superando con creces a NB en un 10%. Tal como ha ocurrido en las categorías anteriores en este nivel, la supremacía de SVM tiene una alta relación a una mayor cantidad de verdaderos negativos y positivos y una menor cantidad de falsos positivos y negativos, siendo superior no sólo en exactitud, sino también en Precision y Recall.

TABLA 4.14 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE OBESIDAD MÓRBIDA

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	TF	579	292	39	10	94,67	93,69	98,30
	Bigram	TF-IDF	579	292	39	10	94,67	93,69	98,30
	Bigram	Binaria	581	290	41	8	94,67	93,41	98,64
NB	Bigram	Binaria	541	238	93	48	84,67	85,33	91,85

Finalmente, en la Tabla 4.15 se muestra la matriz de confusión de la clasificación de "superobesidad", categoría que posee la menor representación en este nivel. La brecha de exactitud existente entre el mayor desempeño de SVM y NB se redujo a 0,98% debido a la gran cantidad de verdaderos negativos detectados entre ambos sistemas, es decir, a la capacidad de identificar correctamente aquellos EMR que no pertenecen a esta categoría. Sin embargo, si se analiza la cantidad de instancias correctamente clasificadas, SVM supera con creces a NB. Si bien este último clasificador posee una Precision del 100% y ningún falso positivo, su Recall es muy inferior al obtenido por SVM, es decir, es incapaz de detectar correctamente la totalidad de las instancias

debidamente etiquetadas como “superobesidad”, lo que se fundamenta en su baja cantidad de verdaderos positivos (1 instancia de 32) y alta cantidad de falsos negativos.

TABLA 4.15 COMPARACIÓN DE RESULTADOS PARA LA IDENTIFICACIÓN DE SUPEROBESIDAD

Clasificador	Segmentación	Matriz	TP	TN	FP	FN	Acc. (%)	Precision (%)	Recall (%)
SVM	Bigram	TF	12	886	2	20	97,61	85,71	37,50
	Bigram	TF-IDF	12	886	2	20	97,61	85,71	37,50
	Bigram	Binaria	10	886	2	22	97,39	83,33	31,25
NB	Bigram	Binaria	1	888	0	31	96,63	100,00	3,13

Capítulo 5. Discusión

Durante el desarrollo del presente trabajo se utilizaron EMR obtenidos desde el HGGB de Concepción sin la identificación de los pacientes, para la búsqueda de obesidad y sus tipos, estudiando sus principales comorbilidades asociadas [14] y haciendo uso de clasificadores supervisados para tales propósitos.

Pese a las ventajas que un sistema de registro digital pudiera tener, como el acceso simultáneo desde distintas unidades en un centro de salud o la instantaneidad de la información, es necesario generar sistemas automatizados de identificación de patrones en los EMR, como enfermedades o hábitos, para poder establecer relaciones que, debido al tamaño de los documentos, son difíciles de realizar en forma manual. Por lo tanto, una vez incorporado un registro electrónico para el almacenamiento de la información de los pacientes, es imperante la necesidad de generar sistemas basados en el NLP para la extracción de información de forma automática.

El trabajo realizado requirió de múltiples procesamientos manuales para adecuar los textos escritos en lenguaje natural. Lo anterior se debe a los errores cometidos por parte del personal médico para registrar la información de los pacientes, debiéndose crear múltiples diccionarios para la corrección de los términos representativos en cada nivel de clasificación, además de las comorbilidades asociadas a la obesidad. La reducción de las variaciones léxicas es fundamental para una correcta representación de la información y el aprendizaje de los clasificadores supervisados.

Los textos relativos a la obesidad carecen de negaciones explícitas de esta condición en los pacientes, razón por la cual fue necesaria la incorporación de contraejemplos en función del estado nutricional de los pacientes (ver Tabla 2.1). De todas formas, la cantidad de EMR recuperados (ver Fig. 4.1) es escasa, porcentaje que es aún menor si sólo se consideran los registros que tienen relación con la obesidad o más aún, sus tipos. Sólo un 3,74% de los EMR fueron recuperados en primera instancia para la creación del corpus principal. Este problema se acentúa, ante la escasez de documentos disponibles de forma pública para fines académicos o los que existen están escritos en inglés. El estricto resguardo que las instituciones de salud tienen con la información de sus pacientes

no hace más que agravar esta situación, por lo que este tipo de instancias es fundamental para la contribución de sistemas que vayan en beneficio de la población.

Los campos de texto libre, o no estructurados, corresponden al registro de la comunicación médico-paciente o a las observaciones que el personal clínico hace en cada atención. Sin embargo, fue posible encontrar en los campos estructurados, como hábitos o antecedentes mórbidos, la presencia de textos libres, por lo que fue necesaria su inclusión en el trabajo. Además estos últimos disponían de información explícita de las principales comorbilidades de la obesidad en los pacientes.

La mayor cantidad de EMR con pacientes que padecen de obesidad corresponden al género femenino. Por otro lado, las principales enfermedades asociadas a la obesidad (ver Fig. 4.2) son las de tipo cardiovascular, principalmente la hipertensión arterial y endocrinas, como la diabetes y el hipotirodismo. Estos antecedentes concuerdan con el estudio realizado por la ENS [2], entregándose además un mayor detalle del impacto de otras enfermedades por subespecialidad médica. Estas estadísticas pueden servir para la creación de políticas públicas de carácter preventivo de la obesidad y sus comorbilidades, o la correcta asignación de recursos en un centro de salud.

En cuanto a la distribución de las clases, en el primer nivel de clasificación, que correspondía a la identificación de obesidad junto con los demás estados nutricionales de los pacientes, se aprecia (ver Fig. 4.3) una gran cantidad de EMR etiquetados con la clase “obesidad” en comparación con las demás categorías. Esto influyó en los resultados obtenidos por ambos clasificadores, siendo el de mejor desempeño SVM, debido a que NB se ve influenciado por las probabilidades (a posteriori) de ocurrencia de las clases. Esto último es más notorio en la identificación de la obesidad en particular, obteniéndose porcentajes de exactitud superiores a los sistemas globales en ambos clasificadores. Sin embargo, el no considerar el término “obesidad” como elemento representativo del corpus principal, tras la segmentación de la colección de documentos de comorbilidades, deja entrever que los clasificadores tuvieron la capacidad de encontrar patrones asociados a las enfermedades de la obesidad durante su aprendizaje. Caso similar ocurre en el segundo nivel de clasificación (ver Fig. 4.4) donde las categorías “Mórbida” y “Superobesidad” tienen la más alta y baja representación en este nivel y es precisamente SVM quien posee el mejor desempeño.

En el primer nivel de clasificación, SVM alcanza un 89,10% de aciertos superando en un 4,61% a NB, brecha porcentual que alcanza un 4,15% de exactitud si se analiza la clase "obesidad" en particular (ver Tabla 4.10). De igual forma, tanto el Precision y Recall son superiores en SVM, destacando en la capacidad de detectar los verdaderos positivos. La cercanía existente entre obesidad y sobrepeso (también conocida como preobesidad), afectó en mayor proporción la detección de este último, obteniéndose porcentajes de Precision y Recall muy inferiores (ver Tabla 4.9) si se comparan con los resultados de la identificación de "obesidad" (ver Tabla 4.10). De igual forma ocurre con la clase "normopeso" en el desempeño de NB (ver Tabla 4.8), dejando entrever que las comorbilidades de la obesidad no son de su exclusividad, pudiendo estar presentes en otros estados nutricionales. Por otro lado, en "bajo peso" el rendimiento alcanzado por ambos clasificadores es incluso superior a la identificación de obesidad (ver Tabla 4.7) debido a la gran cantidad de verdaderos negativos detectados, favorecidos por la distribución de las clases, y a una menor influencia de las comorbilidades de la obesidad como era de esperarse en esta categoría. Un aspecto a destacar en este nivel de clasificación es la similitud de resultados obtenidos por NB en los distintos tipos de segmentación y representaciones matriciales (ver Tabla 4.2), lo que no ocurre con SVM, donde el uso de la matriz binaria y bigrams generó una diferencia de 3,08% con el desempeño más bajo alcanzado (ver Tabla 4.3).

Por otro lado, en la clasificación de los tipos de obesidad, nuevamente SVM supera con creces a NB en un 11,63%. Si se analiza cada clase en particular, los porcentajes de exactitud, Precision y Recall son superiores en SVM, salvo en "superobesidad", donde si bien NB obtiene un 100% en Precision (ver Tabla 4.15) su Recall es aproximadamente 12 veces inferior a SVM. Estos resultados fueron obtenidos mediante una segmentación basada en bigrams, lo que favoreció el reconocimiento de enfermedades y grados de obesidad al tratarse de terminología compuesta por más de una palabra o una combinación con números. Un aspecto a destacar en este nivel es el uso de la matriz binaria, obteniéndose los mejores resultados con NB (ver Tabla 4.4). Si bien con SVM esto último no fue así, la diferencia entre la exactitud alcanzada por TF y TF-IDF y la representación binaria con este clasificador es de un 0,1% (ver Tabla 4.5). Esta tendencia deja entrever que la sola presencia de un término en un texto es razón suficiente para determinar alguna condición médica en un paciente debido a la naturaleza de los textos que se caracterizan por su brevedad.

Capítulo 6. Conclusión

El presente trabajo de extracción e identificación de información fue producto de una convergencia entre disciplinas asociadas a la minería de textos y el conocimiento del dominio de la obesidad y sus comorbilidades. El enfoque propuesto planteó el desafío de poder relacionar las principales enfermedades de la obesidad para su identificación en los EMR.

El uso de sistemas basados en el NLP no sólo permite la generación de mecanismos capaces de automatizar diagnósticos para la identificación de enfermedades, en este caso la obesidad y sus tipos, sino también tiene la potencialidad de generar múltiples estadísticas de los documentos para la toma de decisiones. Un ejemplo de esto último es la gráfica mostrada en la Fig. 4.2, donde es posible identificar las subespecialidades más propensas a tener pacientes con ciertas enfermedades. Por lo tanto, estos procesamientos son una herramienta eficiente para un análisis interno de un centro de salud, para una correcta distribución de los recursos humanos, tecnológicos y económicos.

Si bien en la identificación de obesidad (primer nivel de clasificación) hubo un notorio desbalance de clases a favor de esta categoría (ver Fig. 4.3), la exclusión de palabras claves explícitas sobre esta enfermedad y el sobrepeso como términos representativos del corpus principal, favoreció que el aprendizaje de los clasificadores supervisados fuera sólo a partir de las comorbilidades de la obesidad. De esta forma, se propuso un sistema que pueda complementar la escasa información que se registra de la obesidad en los EMR (ver Fig. 4.1).

En cuanto a los resultados, SVM obtuvo el mejor desempeño en todas las tareas de clasificación, siendo favorecido por el uso de bigrams. El mayor desempeño con SVM fue obtenido para la clasificación de los tipos de obesidad con una exactitud de 93,80%. Además, todas las clases de este nivel obtuvieron porcentajes de exactitud sobre el 90% con SVM, siendo favorecidas por el uso de bigrams y representaciones matriciales dependientes de la frecuencia de aparición de los tokens en los textos como el TF y TF-IDF. Incluso, en la categoría de menor representación en este nivel, como es la clase “superobesidad”, SVM obtuvo un rendimiento muy superior a NB si se comparan los respectivos Precision y Recall y la cantidad de verdaderos positivos (ver Tabla 4.15). No obstante a lo anterior, el rendimiento alcanzado por SVM mediante el uso de matrices binarias

para representar la información es muy cercano al obtenido por TF y TF-IDF, por lo que su implementación junto con un enfoque basado en bigrams, es una buena alternativa por el procesamiento computacional que implica obtener dichas matrices. Por otro lado, el uso de datos numéricos como elementos representativos favoreció la disminución de ambigüedades de este nivel, debido a que en ciertas ocasiones se hacía mención a un tipo de obesidad, y al mismo tiempo, a un valor de IMC que no correspondía a tal condición.

Por otro lado, en el primer nivel de clasificación, la máxima exactitud obtenida en el sistema global fue alcanzada por SVM con un 89,10%, mientras que para la identificación de obesidad la cifra fue 89,22% (ver Tabla 4.10). La diferencia con NB se amplía si se comparan los porcentajes de Precision y Recall. Sin embargo, el costo computacional de NB para procesar la información es muy inferior al realizado por SVM, convirtiéndolo en una buena alternativa debido al desempeño obtenido.

Capítulo 7. Trabajo futuro

El trabajo propuesto implicó múltiples procesamientos de información con base médica, cuyas etapas pueden ser optimizadas para una mejora continua. El rendimiento de los clasificadores depende de la cantidad de atributos seleccionados como elementos representativos de los documentos, por lo que se pretende analizar cuáles son los que tienen mayor relevancia en el desempeño de los clasificadores supervisados.

Dada la complejidad del lenguaje médico, se hace necesaria la convergencia de múltiples unidades del sector para una comprensión más acabada de las múltiples relaciones que tienen las enfermedades y sus tratamientos, y como una forma de evaluar el real impacto que los sistemas de aprendizaje automático pudieran tener.

Los corpus creados tanto para su clasificación como para la extracción de términos asociados a las enfermedades están sujetos a modificaciones conforme se dispongan nuevos EMR. Una vez obtenida esta información adicional, se puede someter a prueba la eficacia de los clasificadores ante la presencia de variaciones presentes en los textos.

Capítulo 8. Presentación del trabajo

Parte de este trabajo fue presentado en la International Student Conference Chile/ 7th Biomedical Engineering Conference, realizada en la Universidad de Concepción, Chile, durante el 3 y 4 de noviembre del 2014. En este trabajo se presenta el preprocesamiento de los EMR para la generación de diversas estadísticas de la obesidad y sus tipos. A continuación se detallan algunos aspectos del paper enviado.

Título: Extracción de Información en Registros Médicos Electrónicos para el Estudio de la Obesidad

Autores: Christopher Flores, Rodolfo Cid, Rosa Figueroa

Institución: Universidad de Concepción

Este trabajo fue aceptado y presentado en modalidad oral

Bibliografía

[1] Oportunidades y desafíos de innovación (2010). Fecha de último acceso: Marzo del 2015. [En línea]. Disponible en: http://www.fundacionchile.com/bio-detalle-biblioteca-area/detalle-biblioteca-area.index/3148/chile-saludable-oportunidades-y-desafios-de-innovacion

[2] Indicadores de obesidad en la población Chilena (2010). Fecha de último acceso: Marzo del 2015. [En línea]. Disponible en: http://www.sochob.cl/pdf/encuesta_nacional_salud_20092010_obesidad.pdf

[3] Panorama de salud 2013, informe OECD sobre Chile y comparación con países miembros (2013). Fecha de último acceso: Marzo del 2015. [En línea]. Disponible en: http://web.minsal.cl/sites/default/files/INFORME%20OCDE_2013_21%2011_final.pdf

[4] P. Enrique y A. Miño, "Consecuencias de la obesidad", *Acimed*, Vol. 20, no. 4, pp. 84-92, 2009

[5] M. Moreno, "Definición y Clasificación de la Obesidad", *Rev. Méd. Clín. Condes*, Vol. 23, no. 2, pp. 124-127, 2012

[6] I. Witten y E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, Estados Unidos: Ed. Morgan Kauffman, 2005, pp. 365-423

[7] M. Martí *et al.*, *Tecnologías del Lenguaje*, 1st ed. Barcelona, España: Ed. UOC, 2003, pp. 9-12

[8] P. Álvarez, "La sintaxis del lenguaje de los textos científicos. Los tipos oracionales y los giros de participio y gerundio: Estudio de un corpus ruso-español de textos médicos", Tesis de Magister, Facultad de Traducción e Interpretación, Universitat Autónoma de Barcelona, España, 2008

[9] J. Camacho, S. Moreno, F. Suárez-Obando, J.C. Puyano y C. Gómez-Restrepo, "El procesamiento de lenguaje natural y su relación con la investigación de salud mental", *Revista Colombiana de Psiquiatría*, Vol. 42, no. 2, pp. 227-233, 2013

[10] C. Pérez, "Evaluación de reglas de asociación en text mining utilizando métricas semánticas y estructurales", Tesis de Magister, Departamento de Informática y Ciencias de la Computación, Universidad de Concepción, Chile, 2010

[11] J. Cheng, "Diseño e implementación de un algoritmo para la detección de negación de textos clínicos en español", Tesis de grado, Facultad de Informática, Universidad Politécnica de Madrid, España, 2014

[12] M. Vallez y R. Pedraza, "El Procesamiento natural del lenguaje en la recuperación de información textual y áreas afines", *Anuario académico sobre documentación digital y comunicación interactiva*, no. 5, 2014. Fecha de último acceso: Marzo del 2015. [En línea]. Disponible en: http://www.upf.edu/hipertextnet/numero-5/pln.html

[13] J. Vilares, "Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español", *Repositorio institucional de la Universidad de Alicante*, no. 36, pp. 57-58, 2006

[14] O. Üzuner, "Recognizing obesity and comorbidities in sparse data", *J. Am. Med. Inform. Assoc.*, Vol. 16, no. 4, pp. 561-570, 2009

[15] L. Ornella, "Códigos correctores de error en problemas de clasificación multiclase de datos de marcadores moleculares", Tesis Doctoral, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, Argentina, 2010

[16] J. Olmo, "Minería de datos mediante programación automática con colonias de hormigas", Tesis Doctoral, Departamento de Informática y análisis Numérico, Universidad de Córdoba, España, 2013

[17] I. Solt, D. Tikk, V. Gál y Z. Kardkovács, "Semantic classification of diseases in discharge summaries using a context-aware rulebased classifier", *J. Am. Med. Inform. Assoc.*, Vol. 16, no. 4, pp. 580-584, 2009

[18] K. Ambert y A. M. Cohen, "A system for a classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection", *J. Am. Med. Inform. Assoc.*, Vol. 16, no. 4, pp. 590-595, 2009

[19] H. Yang, I. Spasic, JA. Keane y G. Nenadic, "A text mining aproach to the prediction of disease status from clinical discharge summaries", *J. Am. Med. Inform. Assoc.*, Vol. 16, no. 4, pp. 596-600, 2009

[20] H. Ware, C. J. Mullet y V. Jagannathan, "Natural Language Processing framework to assess clinical conditions", *J. Am. Med. Inform. Assoc.*, Vol. 16, no. 4, pp. 585-589, 2009

[21] B. Moreno, S. Monereo y J. Álvarez, *Obesidad, la epidemia del siglo XXI*, 2nd ed. Madrid, España: Ed. Diaz de Santos, 2000, pp. 1-3

[22] J. Félix, *Metabolismo, nutrición y shock*, 4th ed. Colombia: Ed. Médica Panamericana, 2006, pp. 35-39

[23] Janet P. Wallace (2012). What is overweight & obesity?. Fecha de último acceso: Marzo del 2015. [En línea]. Disponible en: http://www.iub.edu/~k536/facts.html

[24] A. Baltasar, *Obesidad y cirugía, cómo dejar de ser obeso*, 2nd ed. Madrid, España: Ed. Arán Ediciones, 2001, pp. 35-46

[25] E. García, "Boosting support vector machines", Tesis de Magister, Departamento de Ingeniería Eléctrica y Electrónica, Universidad de los Andes, Colombia, 2005

[26] A. Zubiaga, V. Fresno y R. Martínez, "Comparativa de aproximaciones a SVM Semisupervisado multiclase para clasificación de páginas web", *Sociedad Española para el Procesamiento del Lenguaje Natural*, pp. 63-70, 2009

[27] T. Mitchell, *Machine Learning*, Ed. McGraw-Hill Science, 1997, pp. 154-157

[28] D. Bridge (2014). Underfitting and overfitting. Fecha de último acceso: Marzo del 2015. [En línea]. Disponible en: http://www.cs.ucc.ie/~dgb/courses/ai2/08-fitting.pdf

[29] F. de Borja Navarro, "Metodología, construcción y explotación de corpus anotados semántica y anafóricamente", Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universitat d' Alacant, 2007

[30] F. Bravo (2013). Procesamiento de texto y modelo vectorial. Fecha de último acceso: Diciembre del 2014. [En línea]. Disponible en: http://www.cs.waikato.ac.nz/~fjb11/clases/irintro.pdf

[31] Biblioteca Nacional de Medicina de EEUU. Fecha de último acceso: Diciembre del 2014. [En línea]. Disponible en: http://www.nlm.nih.gov/

[32] B. Gebrekidan, M. Zampieri, P. Wittenburg y T. Heskes, "Improving native language with TF-IDF weighing", in *8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*, Atlanta, USA, 2013, pp. 216-233