UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

Profesor Patrocinante:

Dra. Rosa L. Figueroa I.

Tesis para optar al grado de
Magíster en Ciencias de la Ingeniería con Mención en Ingeniería Eléctrica

Desarrollo de algoritmos para la extracción de características y la Clasificación automática de la obesidad en registros médicos electrónicos con un enfoque jerárquico multiclase

Concepción, Abril de 2017

Christopher Alejandro Flores Jara

UNIVERSIDAD DE CONCEPCIÓN
Facultad de Ingeniería
Departamento de Ingeniería Eléctrica

Profesor Patrocinante:
Dra. Rosa L. Figueroa I.

Desarrollo de algoritmos para la extracción de características y la Clasificación automática de la obesidad en registros médicos electrónicos con un enfoque jerárquico multiclase

Christopher Alejandro Flores Jara

Tesis para optar al grado de
Magíster en Ciencias de la Ingeniería con mención en Ingeniería Eléctrica

Abril 2017

Resumen

La obesidad es una enfermedad crónica con un creciente impacto a nivel mundial. Se caracteriza por un aumento de grasa corporal que significa un riesgo para la salud de las personas. A menudo, la obesidad se asocia a otras enfermedades crónicas, denominadas comorbilidades, siendo las más frecuentes la hipertensión arterial, las dislipidemias y la diabetes *mellitus* tipo 2. El riesgo de sufrir estas comorbilidades es mayor a medida que aumenta el índice de masa corporal.

En este trabajo se presenta un método para identificar automáticamente la obesidad de los pacientes de un sistema de registros médicos electrónicos, utilizando como método de extracción de características el algoritmo de alineación local de *Smith-Waterman*. Se utilizó un conjunto de datos compuesto por 2610 registros médicos electrónicos de-identificados, obtenidos desde el Hospital Guillermo Grant Benavente de Concepción, los cuales fueron etiquetados manualmente para ser utilizados en dos problemas de clasificación. El primer problema consistió en la identificación de la presencia o ausencia de obesidad. El segundo problema de clasificación consistió en identificar los tipos de obesidad: moderada, severa, mórbida o no mencionada. Cada registro médico electrónico fue representado con el enfoque de bolsa de palabras, utilizando características extraídas en base a n-gramas y mediante el algoritmo de *Smith-Waterman*. Estas características fueron obtenidas a partir de la información textual disponible sobre la obesidad, sus principales comorbilidades y el índice de masa corporal.

Se utilizó un enfoque jerárquico y no jerárquico para clasificar los registros médicos electrónicos, entrenando y evaluando el desempeño de una máquina de soporte vectorial y de *Naïve Bayes*. En general, con la máquina de soporte vectorial se obtuvo un mejor desempeño que con *Naïve Bayes*, gracias a la utilización de características extraídas por el algoritmo de *Smith-Waterman*.

Dedicado a mis padres y a mi hermana, y en especial
a mi abuela Irma, cuyo recuerdo vivirá
por siempre en mi corazón.

Tabla de contenidos

Lista de tablas

Lista de figuras

Abreviaciones

Mayúsculas

OMS : Organización Mundial de la Salud.
ENS : Encuesta Nacional de Salud.
SVM : *Support Vector Machine*, Máquina de Soporte Vectorial.
NB : *Naïve Bayes*, Bayes Ingenuo.
MNB : *Naïve Bayes* Multinomial.
HGGB : Hospital Guillermo Grant Benavente.
UMLS : *Unified Medical Language System*, Sistema Unificado de Lenguaje Médico.
NLTK : *Natural Language Toolkit*, Conjunto de Herramientas de Lenguaje Natural.
NLP : *Natural Language Processing*, Procesamiento del Lenguaje Natural.
SW : *Smith-Waterman.*
IMC : Índice de Masa Corporal.
EMR : *Electronic Medical Records*, Registros Médicos Electrónicos.
RBF : *Radial Basis Function*, Función de Base Radial.
BOW : *Bag of Words*, Bolsa de Palabras.
OVO : *One Versus One*, Uno Contra Uno.
OVA : *One Versus All*, Uno Contra Todos.
AVA : *All Versus All*, Todos Contra Todos.
ECOC : *Error-Correcting Output Codes*, Códigos de Salida de Corrección de Errores.
IG : *Information Gain*, Ganancia de Información.
TF-IDF : *Term Frequency-Inverse Document Frequency*, Frecuencia de Término – Frecuencia Inversa de Documento.
FPR : *False Positive Rate*, Tasa de Falsos Positivos.
FNR : *False Negative Rate*, Tasa de Falsos Negativos.
TP : *True Positive*, Verdadero Positivo.
TN : *True Negative*, Verdadero Negativo.
FP : *False Positive*, Falso Positivo.
FN : *False Negative*, Falso Negativo.
MAP : *Maximum a Posteriori*, Máxima a *Posteriori.*
O : Obesidad.
NO : No Obesidad.
ONM : Obesidad No Mencionada.
OMO : Obesidad Mórbida.
OS : Obesidad Severa.
OM : Obesidad Moderada.
ACC : *Accuracy*, Precisión predictiva.
F1 : Valor-F1, *F-measure.*
PCU : Pérdida Cero/Uno, *0/1 loss.*

Minúsculas

i2b2 : *Informatics for Integrating Biology & the Bedside*, Informática para Integrar la Biología y la cabecera.

Capítulo 1. Introducción

1.1 Introducción General

La obesidad es una enfermedad crónica definida como un aumento de grasa corporal que pueda ser perjudicial para la salud de las personas [1]. A nivel mundial, la obesidad ha tenido un creciente impacto en la población, siendo considerada una epidemia por la Organización Mundial de la Salud (OMS). Según cifras de la OMS, más de 1.9 billones de adultos tenían sobrepeso el año 2014, de los cuales sobre 600 millones tenían obesidad [2-4]. En Chile, de acuerdo a la Encuesta Nacional de Salud (ENS) del año 2010, el 25% de la población adulta tenía obesidad, prevalencia que es mayor en las mujeres [3].

La obesidad es una enfermedad de difícil tratamiento, pues depende del estilo de vida y hábitos de quien la padece. Además, ésta suele estar asociada a otras enfermedades, denominadas comorbilidades, tales como la resistencia a la insulina, la hipertensión arterial, el colesterol alto, las enfermedades cardiovasculares, la depresión, entre otras [5, 6]. El riesgo de sufrir estas enfermedades es mayor conforme aumenta el índice de masa corporal (IMC) de los pacientes [7].

La informatización de los sistemas hospitalarios ha permitido generar una gran cantidad de información biomédica de tipo textual, lo que hace indispensable el desarrollo de nuevas herramientas tecnológicas que permitan organizar y crear conocimiento útil a partir de dichas fuentes [8]. Una de las formas de organizar la información de tipo textual es la categorización de textos, los cuales permiten asignar etiquetas o clases a los textos basados en su contenido utilizando un algoritmo de clasificación [9-11]. Para entrenar un modelo de clasificación es necesario extraer características (*tokens*) que permitan el aprendizaje de los algoritmos de clasificación implementados, siendo la forma más tradicional de extraer características el uso de secuencias de palabras (n-gramas) [12]. Idealmente, se espera que los textos que pertenecen a una misma categoría compartan características similares. Ante esta idea, se plantea el uso del algoritmo de *Smith-Waterman* (SW) como método de extracción de características. El algoritmo de SW fue desarrollado para la alineación local de secuencias biológicas, pero recientemente se ha extendido su uso en tareas de extracción y clasificación de textos [13, 14].

En este trabajo de tesis se propone un método de extracción de características junto con métodos de clasificación multiclase para analizar, con un enfoque jerárquico y no jerárquico, el problema de la identificación automática de la obesidad y sus tipos en un sistema de registros

médicos electrónicos de-identificados, obtenidos desde el Hospital Guillermo Grant Benavente (HGGB) de Concepción. Una de las principales contribuciones de esta tesis de Magíster en el proceso de clasificación de textos biomédicos es la incorporación del algoritmo de alineación local de SW como método de extracción de características (método utilizado generalmente en la alineación de secuencias biológicas). Se espera que las características extraídas mediante el algoritmo de SW permitan un mejor desempeño en los clasificadores implementados que el método tradicional de extracción de características basado en n-gramas.

1.2 Hipótesis de Trabajo

Un método de extracción de características basado en la alineación local de textos permitiría mejorar la clasificación del problema de la obesidad, en comparación con los métodos tradicionales de extracción basados en n-gramas.

1.3 Objetivos

1.3.1 Objetivo General

Diseñar e implementar, desde un enfoque multiclase y jerárquico, algoritmos para la clasificación del estado nutricional de un paciente mediante la utilización del algoritmo de *Smith-Waterman*, como método de extracción de características.

1.3.2 Objetivos Específicos

- Crear una base de datos *gold standard* utilizando registros médicos electrónicos que contengan información sobre la presencia y ausencia de obesidad.
- Realizar un análisis exploratorio y estadístico de los datos seleccionados, con el fin de conocer la población de estudio y densidad de datos.
- Implementar y comparar tres métodos de extracción de características basados en n-gramas y mediante el algoritmo de alineación local de *Smith-Waterman*.
- Realizar selección de características utilizando el método de ganancia de información.
- Ajustar los parámetros de los algoritmos de clasificación implementados.
- Clasificar y evaluar el desempeño de los algoritmos de clasificación implementados.

1.4 Alcances y Limitaciones

- Se utilizarán registros médicos electrónicos de-identificados del Hospital Guillermo Grant Benavente de Concepción, obtenidos entre los años 2011 y 2012. Se consideraron los campos estructurados y de texto libre escritos en lenguaje natural.
- Se consideraron dos problemas de clasificación para la identificación de obesidad: la presencia o ausencia de obesidad, y los tipos de obesidad (moderada, severa, mórbida y tipo no mencionado).

1.5 Temario

La organización de este informe se detalla a continuación:

- En el capítulo 1 se introduce el trabajo de tesis, se presenta la hipótesis, el objetivo general y los objetivos específicos, los alcances y las limitaciones.
- En el capítulo 2 se realiza una revisión bibliográfica de los trabajos relacionados al estudio de la obesidad y sus comorbilidades en registros médicos electrónicos, y se presentan aspectos teóricos relacionados a la clasificación supervisada.
- En el capítulo 3 se describen los registros médicos electrónicos utilizados para la identificación de la obesidad, y la metodología utilizada en la clasificación.
- En el capítulo 4 se presentan los resultados obtenidos del proceso de anotación y clasificación de los registros médicos electrónicos.
- En el capítulo 5 se presentan las conclusiones y discusiones de los resultados obtenidos del trabajo de tesis, junto con las mejoras que podrían realizarse en un trabajo futuro.
- En el capítulo 6 se presentan las publicaciones originadas a partir del trabajo de tesis.
- En el anexo se presentan los documentos que autorizan el uso de los registros médicos electrónicos de-identificados del HGGB de Concepción e información complementaria sobre el procesamiento de los datos utilizados.

Capítulo 2. Revisión Bibliográfica

En este capítulo se define la obesidad como enfermedad crónica y se muestran las principales comorbilidades asociadas a dicha enfermedad. Asimismo, se detallan los algoritmos de aprendizaje supervisado que se utilizarán en este trabajo, así como las principales herramientas derivadas del procesamiento del lenguaje natural descritas por la literatura científica y que aplican a la categorización o clasificación de textos.

2.1 Obesidad y comorbilidades

La obesidad es una enfermedad crónica definida como un aumento en la cantidad de grasa corporal que significa un riesgo para la salud de las personas [1]. Ésta es producida por un desbalance en el gasto energético, debido a factores genéticos o ambientales [1, 15]. Uno de los indicadores antropométricos más utilizados para evaluar cuantitativamente la obesidad es el IMC, definido como el cociente entre la masa, expresado en kilogramos, y el cuadrado de la altura, expresado en metros cuadrados:

$$\text{IMC} = \frac{masa}{altura^2} \ [kg/m^2] \tag{2.1}$$

En pacientes adultos, un IMC mayor o igual a 30 indica obesidad [7] (ver Tabla 2.1). Habitualmente, los pacientes con un IMC mayor o igual a 50 son incluidos en una nueva categoría denominada superobesidad [16].

TABLA 2.1 Estado nutricional de un individuo en función del IMC [7].

Estado nutricional	IMC
Bajo peso	<18.5
Peso normal	18.5-24.9
Sobrepeso	25-29.9
Obesidad Moderada	30-34.9
Obesidad Severa	35-39.9
Obesidad Mórbida	≥40

La obesidad se ha convertido en una epidemia mundial. Según cifras de la OMS, el 2014 más de 1.9 billones de adultos tenían sobrepeso, de los cuales 600 millones eran obesos [2]. Estas

cifras han convertido a la obesidad en uno de los problemas de salud pública más importantes en gran parte de los países del mundo, disminuyendo la esperanza de vida de sus habitantes y generando altos costos socio-económicos [3].

Chile no es ajeno al escenario mundial de la obesidad. De acuerdo a cifras de la Encuesta Nacional de Salud (ENS), el año 2010 el 25.1% de la población adulta tenía obesidad. En todos los grupos etarios, las mujeres tienen la mayor prevalencia de obesidad, tendencia que se acentúa en las personas de mayor edad (Fig. 2.1) [3].

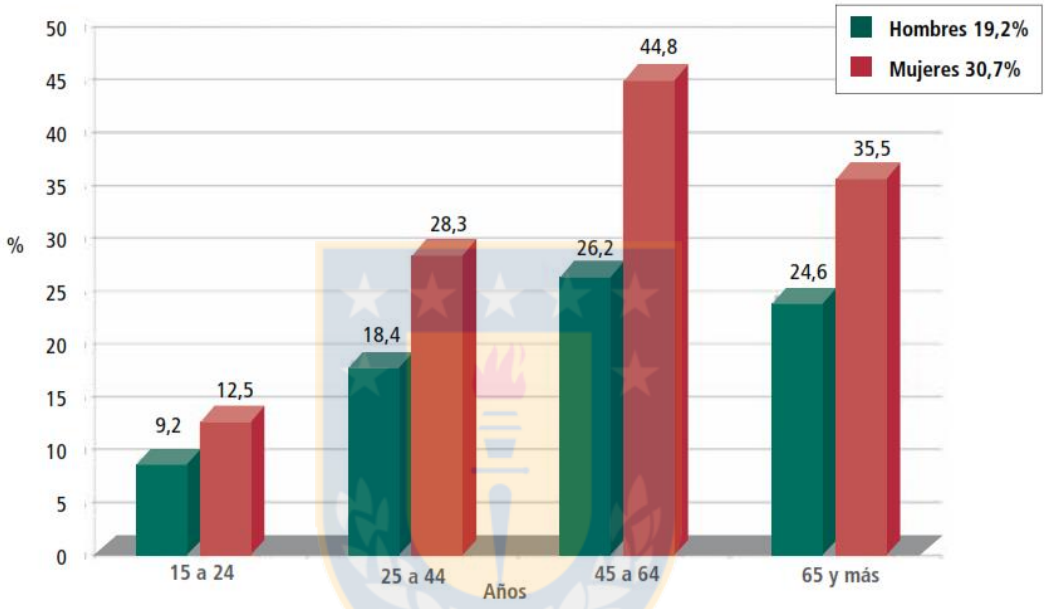


Fig. 2.1 Prevalencia de la obesidad en Chile por edad y sexo [3].

La obesidad es una enfermedad que impacta significativamente en la esperanza y calidad de vida de quienes la padecen, ya que suele estar asociada a otras enfermedades denominadas comorbilidades. Entre las principales comorbilidades asociadas a la obesidad se encuentran las enfermedades producidas por alteraciones metabólicas, como la diabetes *mellitus* 2, las dislipidemias y la hipertensión arterial. Otras comorbilidades tienen su origen en los efectos físico-mecánicos producidos por el exceso de peso, como la osteoartritis, la hipoventilación y la apnea del sueño. Otro grupo de enfermedades corresponden a las comorbilidades de tipo psicopatológicas que afectan la salud mental de los pacientes [17]. En la Tabla 2.2 se muestran las principales comorbilidades de la obesidad, agrupadas en enfermedades metabólicas, mecánicas y psicosociales.

TABLA 2.2 Principales comorbilidadades asociadas a la obesidad [17].

Metabólicas	Mecánicas	Psicosociales
Diabetes *mellitus* 2	Hipoventilación	Depresión
Dislipidemias	Apnea del sueño	Ansiedad
Hipertensión arterial	Miocardiopatía	Alteraciones conductuales
Cardiovasculares	Insuficiencia cardiaca	Mayor riesgo a las adicciones
Neoplasias	Osteoartrosis	Discriminación social
Colelitiasis		
Hígado graso		
Ovario poliquístico		

Los riesgos asociados a las comorbilidades de la obesidad aumentan conforme aumenta el IMC (ver Tabla 2.3) [18]. Si el IMC es menor a 25, el riesgo de sufrir comorbilidades es bajo, mientras que si el IMC es mayor a 40, el riesgo de padecer comorbilidades es extremadamente alto y el tratamiento médico indicado es la cirugía bariátrica.

TABLA 2.3 Riesgos para la salud según el IMC y las comorbilidades de la obesidad [18].

IMC	Riesgos	Comorbilidades
<25	Mínimo	Bajo
25-<27	Bajo	Moderado
25-<30	Moderado	Alto
30-<35	Alto	Muy alto
35-<40	Muy alto	Extremadamente alto
>40	Extremadamente alto	Extremadamente alto

2.2 Clasificación supervisada

La clasificación es un método de aprendizaje supervisado que busca construir modelos a partir de muestras de entrenamiento etiquetados para clasificar ejemplos de etiqueta desconocida [19]. El proceso de clasificación supervisada se muestra en la Fig. 2.2. Durante la etapa de entrenamiento se extraen características de los datos de entrada (etiquetados), los cuales le permiten a un algoritmo de aprendizaje supervisado construir el modelo de clasificación. Luego, en la etapa de predicción, el modelo de clasificación es utilizado para etiquetar nuevos datos (sin etiqueta).

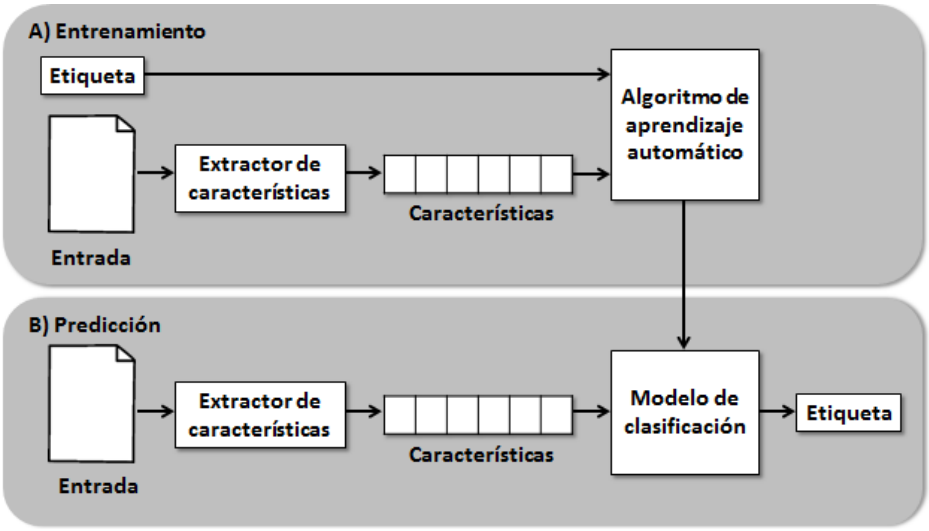


Fig. 2.2 Etapas de la clasificación supervisada [1]. A: entrenamiento. B: predicción.

2.2.1 Máquina de soporte vectorial

Una máquina de soporte vectorial (SVM) es un algoritmo de aprendizaje supervisado que se fundamenta en la teoría estadística de aprendizaje propuesta por Vladimir Vapnik en 1995. A partir de n datos de entrenamiento representados mediante la tupla (x_i, y_i), una SVM permite asignar una etiqueta $y_i \in \{-1,1\}$ a una muestra representada por el vector de atributos x_i. En el caso ideal de un problema de clasificación binario (dos clases) linealmente separable, existen infinitos hiperplanos (canónicos) que pueden separar ambas clases en los datos de entrenamiento (ver Fig. 2.3) [19].

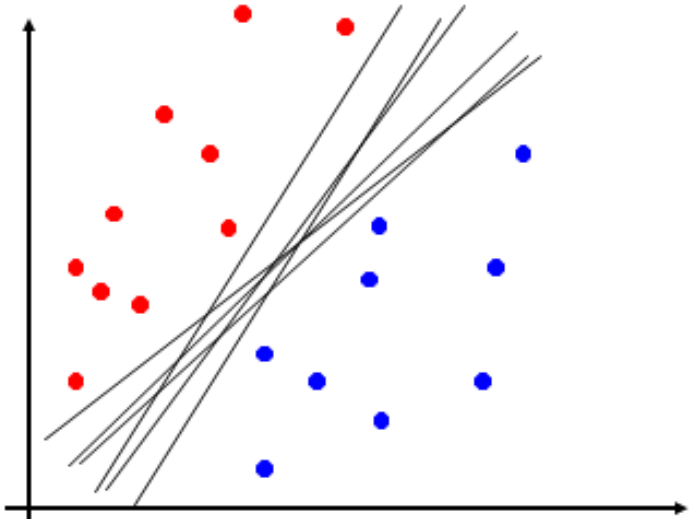


Fig. 2.3 Hiperplanos canónicos en un problema de clasificación binario ideal[2].

[1] Disponible en: http://www.nltk.org/book/ch06.html. Fecha de último acceso: Agosto de 2016

El objetivo de una SVM es encontrar un hiperplano óptimo que pueda separar ambas clases, maximizando la distancia (margen) existente entre los vectores de soporte, los cuales corresponden a puntos en el espacio de características (ver Fig. 2.4) [19].

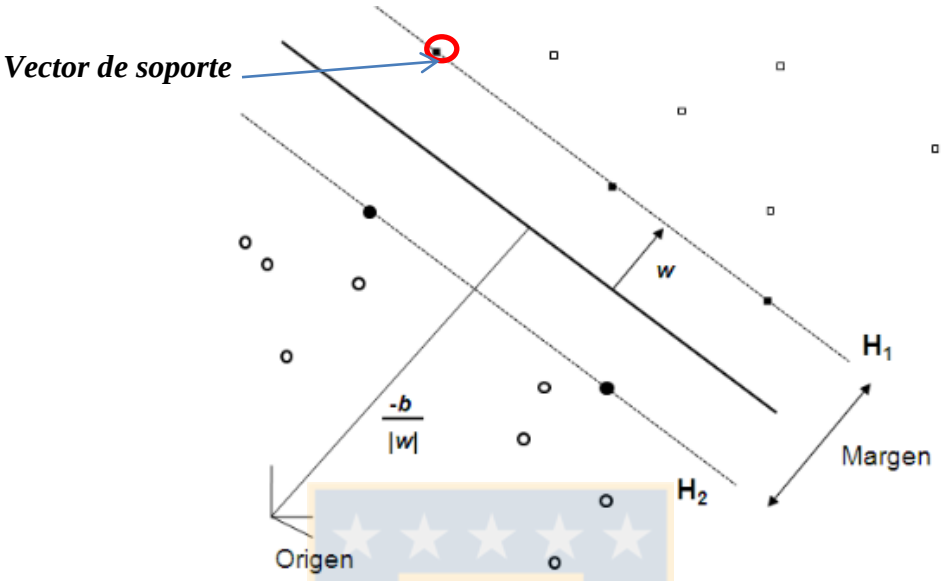


Fig. 2.4 Margen y vectores de soporte en una SVM [19].

Matemáticamente, sea w un vector perpendicular al hiperplano y b la separación al origen de las coordenadas, el hiperplano óptimo satisface la ecuación 2.2 [19]:

$$w \cdot x_i \pm b = 0 \tag{2.2}$$

Al finalizar la etapa de entrenamiento, los datos pueden ser separados linealmente por el hiperplano óptimo (ver Fig. 2.5). Este hiperplano se busca minimizando [19]:

$$\frac{1}{2}\left|\left|w^2\right|\right| \tag{2.3}$$

Sujeto a $\quad x_i \cdot w + b \geq +1 \;\; para\; y_i = +1$

$$x_i \cdot w + b \leq -1 \;\; para\; y_i = -1$$

[2] Disponible en: https://www.cs.utexas.edu/~mooney/cs391L/slides/svm.ppt. Fecha de último acceso: Febrero de 2017

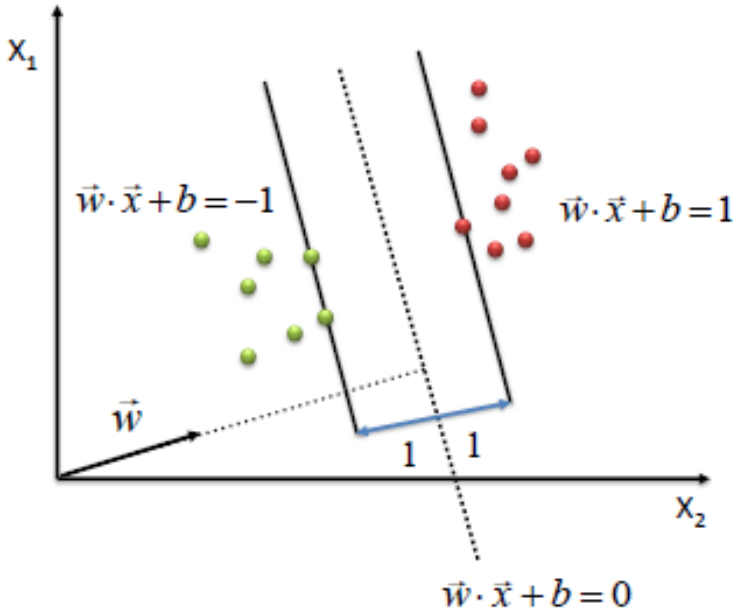


Fig. 2.5 Hiperplano óptimo tras la etapa de entrenamiento[3].

El problema de minimización de la ecuación 2.3 puede ser resuelto mediante el uso de multiplicadores de *Lagrange* (α_i) buscando maximizar [19]:

$$W(\alpha) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j \tag{2.4}$$

Sujeto a:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Con:

$$0 \leq a_i$$

En presencia de valores atípicos (*outliers*), la búsqueda del hiperplano óptimo debe considerar variables de holgura que minimicen el error de clasificación. De esta forma, la ecuación 2.3 es modificada según [19]:

$$x_i \cdot w + b \geq +1 - \xi_i \;\; para\; y_i = \; +1 \tag{2.5}$$

[3] Disponible en: http://chem-eng.utoronto.ca/~datamining/dmc/support_vector_machine.htm. Fecha de último acceso: Febrero de 2017

$$x_i \cdot w + b \leq -1 + \xi_i \quad para\ y_i = -1$$

Con este parámetro de holgura, el hiperplano óptimo se busca minimizando:

$$\frac{1}{2}\left||w^2|\right| + C\sum_{i=1}^{n}\xi_i \tag{2.6}$$

Lo que es equivalente a maximizar la ecuación 2.4

Sujeto a:

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

Con:

$$0 \leq a_i \leq C \tag{2.7}$$

Donde *C* es un parámetro de costo o ponderación que penaliza el error de clasificación en la etapa de entrenamiento. Finalmente, en el caso de un problema linealmente separable, la solución del hiperplano óptimo se obtiene según la ecuación 2.8, donde n_s es el número de vectores de soporte [19]:

$$w = \sum_{i=1}^{n_s}\alpha_i y_i\, x_i \tag{2.8}$$

Datos no linealmente separables

Cuando los datos no son linealmente separables, una SVM utiliza una función de transformación $\Phi\ (x_i)$ para mapear los datos de entrada en un espacio de características de mayor dimensión. Tal como se muestra en la Fig. 2.6, la función de transformación $\Phi(x_i)$ aumenta la dimensión de los datos de entrada desde R^2 a R^3, permitiendo separar los datos linealmente. Al no tener un conocimiento previo de la función $\Phi(x_i)$, resulta necesario utilizar una función *kernel* $K(x_i, x_j)$ [20]:

$$K(x_i, x_j) = (\Phi\ (x_i), \Phi(x_j)) \tag{2.9}$$

De esta forma, se puede construir una función de transformación sin utilizar explícitamente $\Phi(x_i)$. Finalmente, cuando el hiperplano *kernel* se lleva a las dimensiones del problema original, una SVM puede separar linealmente las clases [20].

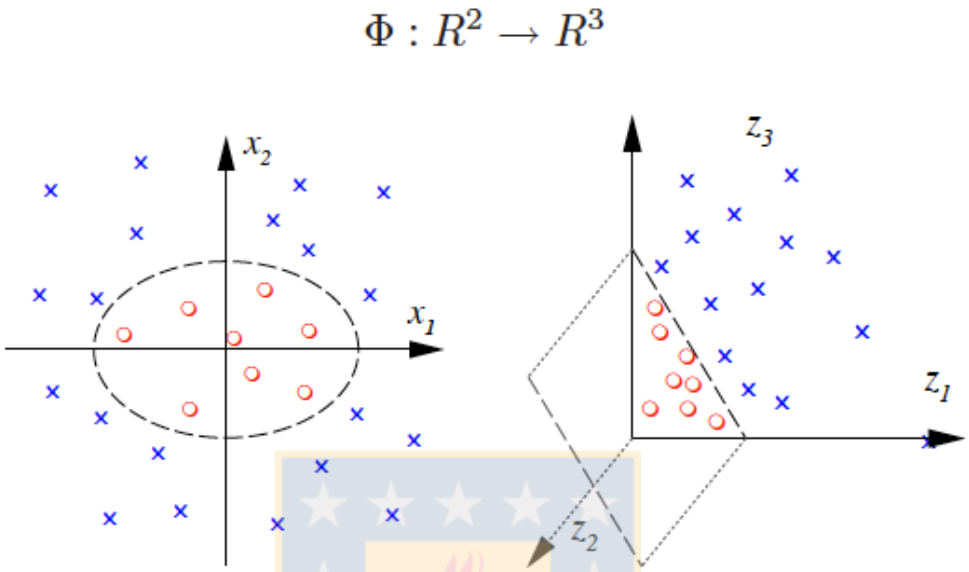


Fig. 2.6 Datos no separables linealmente y su transformación en el espacio de características[4].

Los *kernels* más usados son el tipo lineal:

$$K_L\left(x_i, x_j\right) = x_i \cdot x_j \tag{2.10}$$

y el *kernel* de función de base radial (RBF):

$$K_{RBF}\left(x_i, x_j\right) = e^{-\gamma(x_i - x_j)^2} \tag{2.11}$$

Otros *kernels* son el tipo sigmoideo, polinomial, gaussiano, entre otros [21].

[4] Disponible en: http://courses.cs.ut.ee/2011/graphmining/Main/KernelMethodsForGraphs. Fecha de último acceso: Febrero de 2017

Clasificación multiclase con una SVM

Una SVM es un clasificador diseñado para resolver problemas binarios. Por lo tanto, se deben utilizar estrategias que permitan su uso en problemas de índole multiclase (más de dos clases). Estas estrategias descomponen el problema multiclase original en un número determinado de clasificadores binarios, para luego clasificar los datos combinando todas las decisiones obtenidas (Fig. 2.7) [19]. Una de las técnicas más utilizadas, dada la facilidad de su implementación es la estrategia uno contra todos (OVA), la cual construye *n* clasificadores binarios igual a la cantidad de clases del problema de clasificación. En otra estrategia denominada uno contra uno (OVO), se construyen *n* clasificadores binarios según:

$$n = \frac{N\,(N-1)}{2} \tag{2.12}$$

Donde *N* es el número de clases del problema. Otras técnicas de descomposición binaria son los códigos correctores de error (ECOC), la estrategia todos contra todos (AVA), entre otras [19].

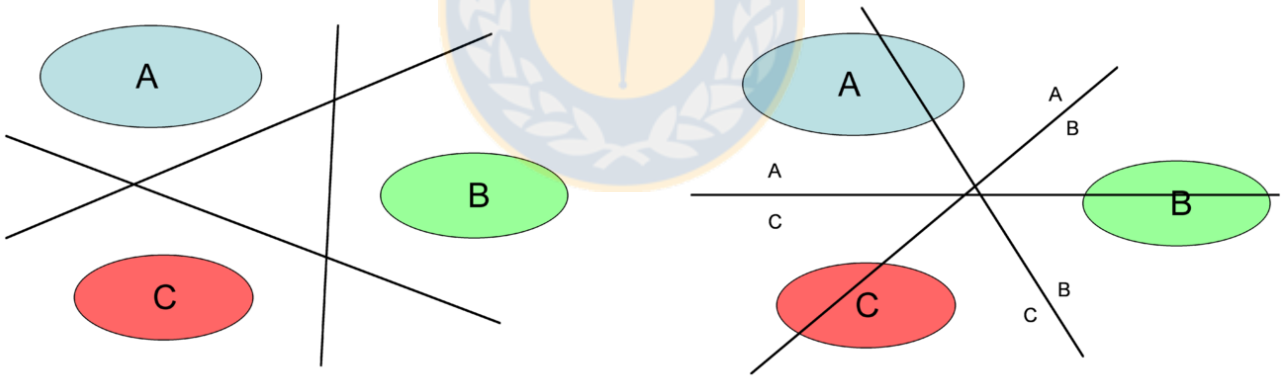


Fig. 2.7 Comparación de estrategias en clasificadores ante problemas multiclase[5]. De izquierda a derecha: uno contra todos, uno contra uno.

[5] Disponible en http://courses.media.mit.edu/2006fall/mas622j/Projects/aisen-project/. Fecha de último acceso: Octubre de 2016

2.2.2 Naïve Bayes

Naïve Bayes o Bayes ingenuo (NB) es un algoritmo de aprendizaje supervisado que se fundamenta en la independencia condicional de los atributos de un problema de clasificación, dada una clase, en una distribución normal de los atributos numéricos y en la inexistencia de variables ocultas que afecten el proceso de predicción. NB puede ser representado mediante un grafo dirigido cuyas aristas están orientadas desde las clases a los atributos, tal como se muestra en la Fig. 2.8 [19].

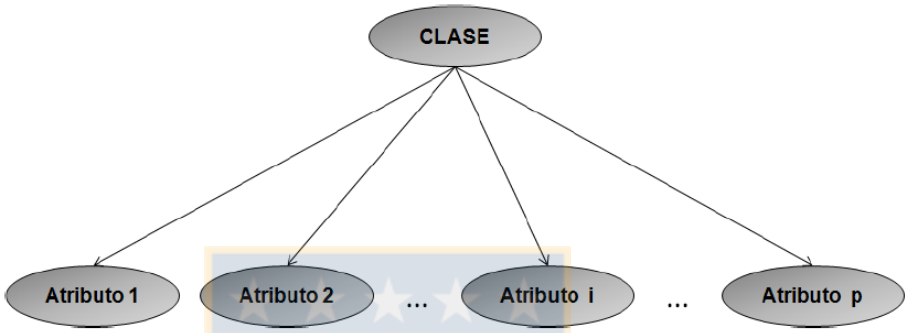


Fig. 2.8 Representación de NB mediante un grafo [19].

El aprendizaje bayesiano busca encontrar la hipótesis *h* más probable (hipótesis máxima a posteriori o MAP) a partir de un conjunto de datos de entrenamiento *D*, utilizando el teorema de Bayes [22]:

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)} \tag{2.13}$$

$$h_{MAP} = argmax_{h \in H}\, P(h/D) \tag{2.14}$$

Como *P(D)* es una constante independiente de *h*, la ecuación 2.14 puede ser escrita de la forma:

$$h_{MAP} = argmax_{h \in H}\, P(D/h)P(h) \tag{2.15}$$

Finalmente, asumiendo que los atributos son condicionalmente independientes [19], la aproximación de NB resulta [22]:

$$v_{NB} = argmax_{v_j \in V}\, P(v_j) \prod_{i=1}^{n} P(a_i/v_j) \tag{2.16}$$

Existen diversos modelos de NB dependiendo de la distribución de los datos del problema de clasificación. Algunos modelos son NB Gaussiano, NB Bernoulli, NB multinomial (MNB), entre otros [23] .

2.2.3 Selección de características

La selección de características consiste en determinar un subconjunto de características relevantes para mejorar el rendimiento de un clasificador [19]. Un proceso de selección de características típico se muestra en la Fig. 2.9 [24]. En primer lugar, se selecciona un subconjunto de datos candidatos desde el conjunto inicial. Posteriormente, con una función, se evalúa la calidad de los datos seleccionados. El proceso continúa hasta que un criterio de parada determina si continuar con la búsqueda de otro subconjunto o no. Finalmente, se evalúa si el subconjunto seleccionado es válido.

Existen tres métodos para la selección de características: métodos de filtros, envolturas (*wrapper*) y una combinación de ambos (híbrido) [19] [24]. Los métodos de filtros analizan los atributos sin utilizar el algoritmo de clasificación. Existen diversos métodos de filtros para la selección de características, tales como: métodos basados en ganancia de información (I*nformation gain*), relación de ganancia (*gain ratio*), selección basada en correlación (*Correlation-based Feature* S*election*), entre otros [19]. Por otro lado, en el método de envoltura se utiliza el algoritmo de clasificación creando distintos subconjuntos de datos, generalmente, mediante validación cruzada, para evaluar y seleccionar el subconjunto que tenga el mejor desempeño. Este método tiene un costo computacional mayor que el método de filtro dificultando su uso en conjuntos de datos extensos.

Por otro lado, existen diversos métodos de búsqueda del subconjunto de datos que pueden mejorar el rendimiento de un clasificador. Algunos métodos evalúan de forma individual cada uno de los atributos y los ordenan de acuerdo al tipo de evaluador utilizado (*Ranker*) o analizan subconjuntos (*greedy, Best First, Random Search, Genetic Search*, entre otros) [24].

En este trabajo se utilizará un método filtro de selección de características basado en ganancia de información, puesto que existe evidencia en la literatura de que este método tiene un buen desempeño en la clasificación de textos [25] [24].

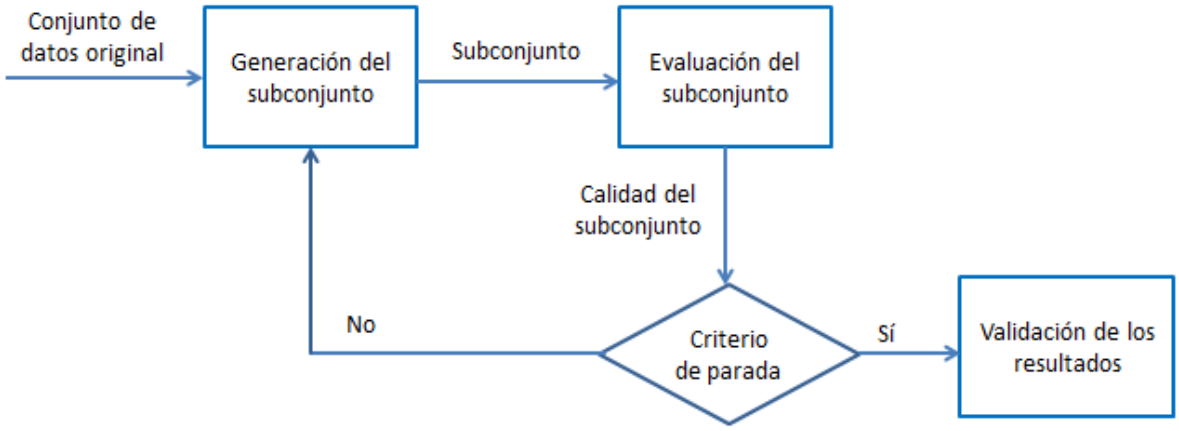


Fig. 2.9 Proceso de selección de características [24].

Ganancia de información

El método de selección de atributos basado en ganancia de información (IG) evalúa la importancia de un atributo según [19, 24]:

$$IG\ (clase, atributo) = H(clase) - H(clase|atributo) \tag{2.17}$$

Donde *H (clase)* es el valor de entropía de una clase y *H (clase/atributo)* es el valor de entropía mutua entre una clase y los atributos. Estas entropías se calculan según:

$$H(Y) = -\sum_{y \in Y} p(y) \cdot log_2\,(p(y)) \tag{2.18}$$

$$H(Y/X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y/x) \cdot log_2(p(y/x))$$

La entropía es una medida del grado de incertidumbre de una distribución de probabilidad [26]. En el caso de una distribución de probabilidad uniforme, es decir, en una distribución donde cada valor tiene la misma probabilidad de ocurrencia, la entropía es máxima (máxima incertidumbre). La Fig. 2.10 muestra la representación gráfica de la entropía de un problema de clasificación binaria. En este caso, la entropía máxima se obtiene con una probabilidad (p) igual a

0.5, esto significa, cuando la probabilidad de una instancia de pertenecer a una de las dos clases es la misma.

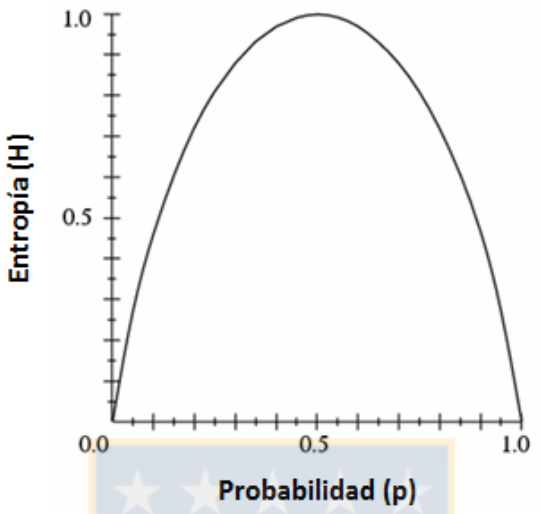


Fig. 2.10 Función de entropía para un problema de clasificación binario[6]

2.2.4 Evaluación de un modelo de clasificación

Para evaluar un modelo de clasificación supervisada, se debe disponer de datos de prueba y de entrenamiento independientes entre sí, con el objeto de evitar errores por sobreajuste (*overfitting*) [27]. Para seleccionar las muestras que conformarán el conjunto de entrenamiento y de pruebas se divide el conjunto de datos original mediante métodos de muestreo como el *hold-out y* la validación cruzada de *k* particiones (*k-folds*). El método *hold-out* es la forma más simple para seleccionar el conjunto de pruebas y de entrenamiento. Consiste en dividir el conjunto de datos original en un único conjunto de entrenamiento y otro conjunto de pruebas, generalmente un 66% y 34% de los datos originales, respectivamente [19]. La desventaja de este método es que el desempeño del algoritmo de clasificación dependerá de cuán distribuidas estén las clases en el conjunto de entrenamiento.

Por otra parte, en la validación cruzada se crean *k* subconjuntos de muestras mutualmente excluyentes en igual cantidad de iteraciones dejando como entrenamiento las *k-1* particiones y como pruebas, el resto de las particiones (Fig. 2.11). Si en cada una de las particiones existe

[6] Disponible: https://ai.vub.ac.be/sites/default/files/ch3.pdf. Fecha de ultimo acceso: Diciembre de 2016

aproximadamente la misma distribución de las clases, la validación cruzada se denomina estratificada. Otro tipo de validación cruzada consiste en utilizar una sola muestra de prueba y el resto como entrenamiento (validación cruzada dejando uno afuera o *leave one out*). Este método tiene un alto costo computacional y no es recomendado para conjuntos de datos extensos por la gran cantidad de iteraciones que se realizan [19]. En validación cruzada, el desempeño final de un clasificador se obtiene promediando cada una de las métricas obtenidas en cada iteración.

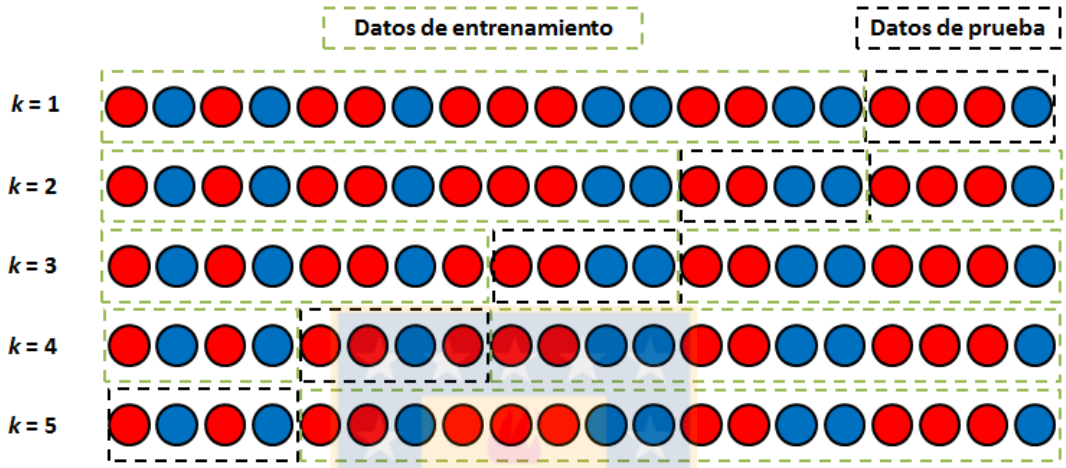


Fig. 2.11 Proceso de validación cruzada con 5 particiones [7].

Finalmente, para medir el desempeño de los métodos de clasificación supervisada se utilizan métricas estándares basadas en una matriz de confusión (Tabla 2.4). Ésta contiene los resultados de los verdaderos positivos (TP) y negativos (TN), correspondientes a los aciertos del clasificador, y los falsos positivos (FP) y negativos (FN), correspondientes a los errores del clasificador [19].

TABLA 2.4 Matriz de confusión para un problema de clasificación binario.

		Clase predicha	
		Positivo	**Negativo**
Clase real	**Positivo**	TP	FN
	Negativo	FP	TN

Fuente: Elaboración propia.

[7] Disponible en: http://s3lab.deusto.es/weka-de-los-datos-a-la-informacion-parte-1/. Fecha de último acceso: Febrero de 2017

2.2.5 Clasificación jerárquica

Muchos problemas de clasificación tienen un carácter jerárquico, es decir, existen clases nodos que tienen asociadas ramificaciones o subclases [28]. Ejemplos de clasificación jerárquica se pueden encontrar en la categorización de textos, etiquetado de imágenes, clasificación funcional de genes, entre otros [29].

La forma más simple de tratar un problema de clasificación jerárquico es ignorar tal jerarquía y transformar la estructura en un único nivel (*flat classification*) [30, 31]. Tal como se muestra en la Fig. 2.12, si la predicción del clasificador es 2.2.2 se asume indirectamente que también pertenece a las clases predecesoras 2.2 y 2.

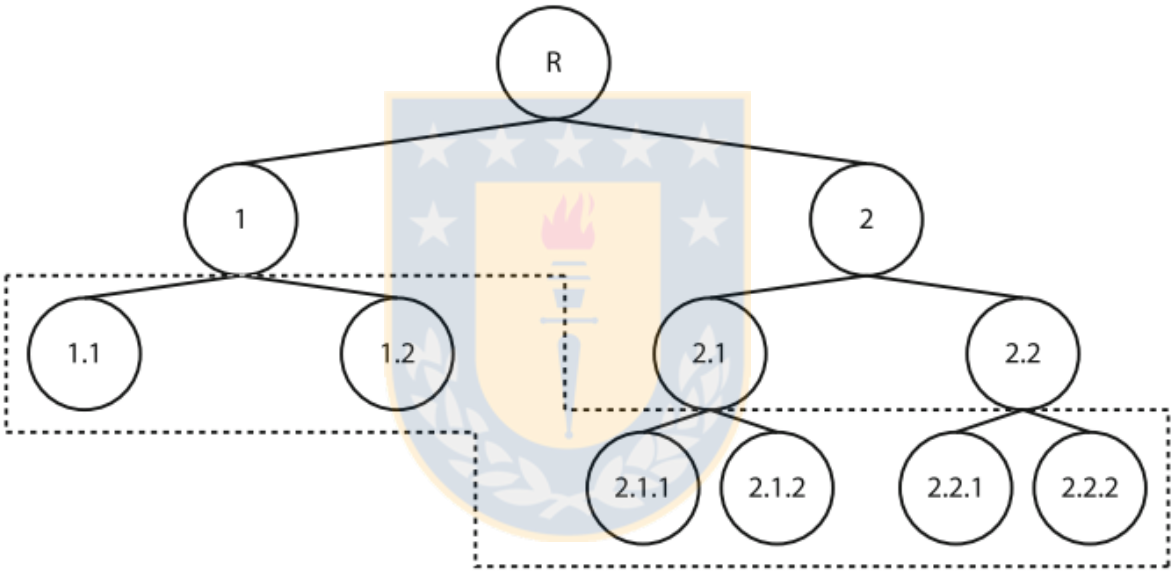


Fig. 2.12 Ejemplo de clasificación jerárquica en un único nivel [30].

La principal desventaja de este método es que no se consideran las relaciones existentes entre la clase nodo y sus descendientes. Sin embargo, si la cantidad de clases es reducida, los resultados entre un método con y sin jerarquía son similares [31].

Otros métodos de clasificación jerárquica construyen clasificadores a medida que se desciende localmente por los niveles de la jerarquía. Tal como se muestra en la Fig. 2.13, la clasificación comienza construyendo clasificadores en el nodo raíz (P1) que permitan predecir las clases 1 o 2 en el primer nivel de la jerarquía. Consiguientemente, se construyen clasificadores que permitan predecir las etiquetas descendientes a las clases 1 y 2 de forma independiente (P2 y P3). Asumiendo que fueron asignadas las clases 1.1, 2.1 y 2.2.2, la jerarquía resulta: 1→1.1 y 2→2.1, y 2→2.2.2. La

principal desventaja de este método es la propagación del error de clasificación desde los niveles superiores de la jerarquía hacia los nodos descendentes [31].

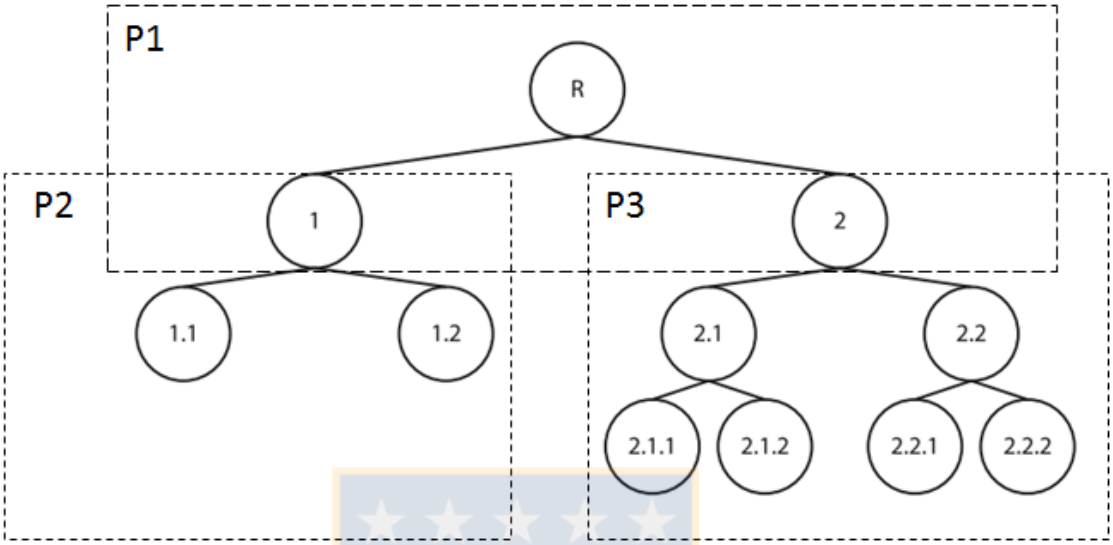


Fig. 2.13 Ejemplo de clasificación jerárquica con el método local [30].

Los métodos de clasificación jerárquica a menudo son descritos como un problema de clasificación de múltiples etiquetas (*multilabel*). A diferencia de la clasificación multiclase, en la que una instancia puede tener sólo una clase del conjunto de etiquetas posibles, la clasificación de múltiples etiquetas permite asignar más de una clase a un mismo ejemplo [30].

Los métodos de clasificación de múltiples etiquetas se dividen en métodos que adaptan un algoritmo de clasificación supervisado determinado (SVM, árboles de decisión, *k*-vecinos más cercanos, entre otros) y en métodos que transforman el problema original en problemas binarios (*binary relevance*) o multiclase (*label powerset*) [32, 33].

El método de transformación binaria es uno de los métodos más utilizados en problemas de clasificación de múltiples etiquetas por lo simple de su implementación. En este método se descompone el problema original en tantos conjuntos de datos como números de clases existan en el problema de clasificación, asociando a cada uno de ellos una clase del problema. En la Fig. 2.14.B se muestra la descomposición del problema de la Fig. 2.14.A en cuatro conjuntos de datos asociados a la presencia o ausencia (representado por ¬) de las etiquetas λ_1, λ_1, λ_1 y λ_4. La predicción de las clases de una nueva instancia está asociada a la unión de todas las etiquetas predichas por los clasificadores binarios. La principal desventaja de este método es que asume que las clases del

problema son independientes [29]. Ante este problema surge el método de transformación multiclase, que asocia a cada instancia un único conjunto de etiquetas con cada una de las clases asociadas a dicha instancia. En la Fig. 2.14.C. se muestra la transformación del problema de la Fig. 2.14.A donde cada combinación de etiquetas es tratada como una nueva meta-clase, para luego utilizar algún método de clasificación multiclase. Si bien este método considera la dependencia existente entre las etiquetas, su complejidad computacional es más alta [29].

A

Instancia	Etiquetas
1	{λ2,λ3}
2	{λ1}
3	{λ1,λ2, λ3}
4	{λ2,λ4}

B

Instancia	Etiqueta	Etiqueta	Etiqueta	Etiqueta
1	¬λ1	λ2	λ3	¬λ4
2	λ1	¬λ2	¬λ3	¬λ4
3	λ1	λ2	λ3	¬λ4
4	¬λ1	λ2	¬λ3	λ4

C

Instancia	Etiquetas
1	λ2,3
2	λ1
3	λ1, 2,3
4	λ2,4

Fig. 2.14 Clasificación de múltiples etiquetas. A: problema original. B: transformación binaria. C: transformación multiclase [29].

2.2.6 Categorización de textos

El desarrollo tecnológico de los últimos años ha permitido generar una gran cantidad de información en formato digital. Se estima que para el año 2020 habrán 35 *zettabytes* de información digital disponible, gran parte de ella, presentada de forma no estructurada escrita en lenguaje natural, por lo que se requieren nuevas tecnologías para extraer información de dichas fuentes y descubrir conocimiento relevante para tomar decisiones [27, 34]. Frente a esto, la categorización o clasificación de textos ha cobrado gran importancia, convirtiéndose en una de las principales técnicas para organizar la información digital disponible [35] .

Preprocesamiento de textos

El preprocesamiento es una etapa que permite preparar los textos para la extracción de características y posterior clasificación. En la colección de textos (*corpus*) se aplican métodos derivados del Procesamiento del Lenguaje Natural (NLP), tales como análisis morfológico y reconocimiento de frases. Dentro de los métodos de análisis morfológico se encuentra la lematización o *word stemming* para identificar el tronco o raíz de las palabras. En el reconocimiento de frases, se etiquetan las palabras en función del rol que cumplen en una oración o frase como sustantivos, adjetivos, verbos, entre otros (Partes del Habla) [27, 36]. Igualmente, en el preprocesamiento de los textos se incluyen tareas de identificación de los segmentos de texto que contienen la información más relevante para la clasificación (*hotspots*) [37].

Extracción y representación de textos

Una vez que los textos han sido preprocesados, se debe seleccionar la forma de representarlos mediante atributos o características (*tokens*). Estas características pueden ser extraídas mediante una secuencia de n-palabras, denominado n-gramas, donde n corresponde a la cantidad de palabras de la secuencia. Los n-gramas más utilizados en extracción de características son los unigramas (una palabra) y los bigramas (secuencia de dos palabras) [12].

El método más simple para representar los textos mediante sus características es la utilización de la bolsa de palabras (BOW), la cual considera todas las palabras del *corpus* sin importar su orden [35]. Otros métodos más complejos que consideran la semántica de las palabras permiten mejorar la relación conceptual de las características utilizadas para representar los textos [38].

Para reducir el tamaño de representación de los textos es común eliminar las *stopwords* o palabras vacías, las cuales tienen una gran frecuencia en el *corpus* (artículos, preposiciones, conjunciones, entre otros) [27]. Luego de la representación, los textos adquieren una estructura que permite aplicar en ellos métodos de clasificación supervisada para generar nuevo conocimiento [27] (Fig. 2.15).

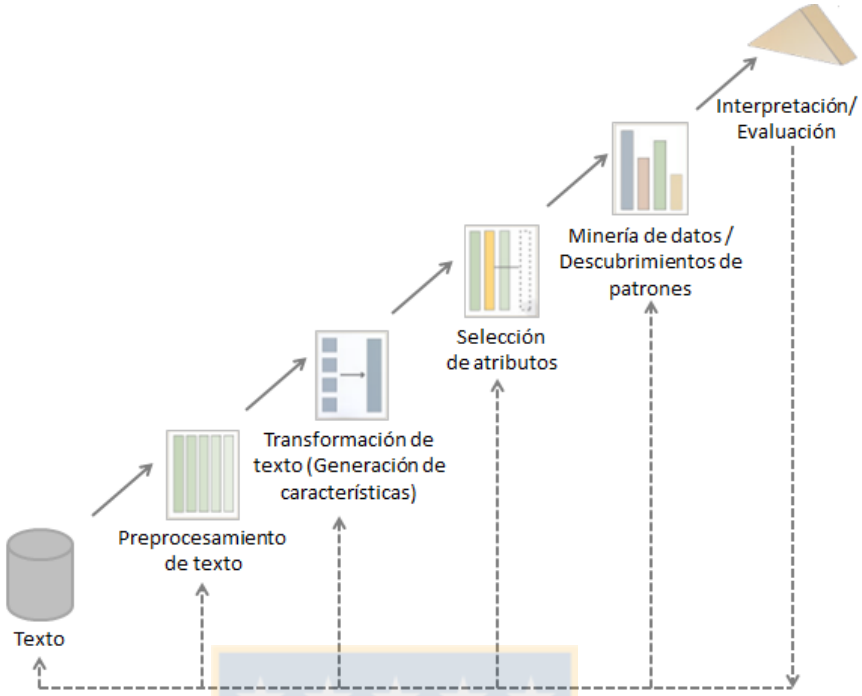


Fig. 2.15 Etapas para la clasificación de textos [8].

Creación de un gold standard para clasificación

La categorización de textos es un proceso de aprendizaje supervisado, por lo que se requieren ejemplos etiquetados (*gold standard*) por expertos humanos para entrenar los algoritmos de clasificación. Si no se dispone de un *gold standard*, los textos deben ser etiquetados en un proceso de anotación. Este proceso debe ser realizado por al menos dos anotadores especialistas en el contenido del *corpus* para evitar anotaciones subjetivas [39]. El proceso de anotación finaliza cuando un tercer sujeto evalúa los desacuerdos y decide la anotación final.

Para medir el nivel de acuerdo entre dos anotadores se utiliza una métrica de concordancia estadística denominada índice de *Kappa* de Cohen (k) que se calcula según [39, 40]:

$$k = \frac{Po - Pe}{1 - Pe} \tag{2.19}$$

Donde Po y Pe corresponden a las concordancias observadas entre los anotadores y a las concordancias atribuibles al azar, respectivamente. Po puede ser calculado según:

[8] Disponible en: http://www3.cs.stonybrook.edu/~cse634/presentations/TextMining.pdf. Fecha de último acceso: Agosto de 2016

$$Po = \frac{1}{N}\sum_{i=1}^{C} N_{ii} \tag{2.20}$$

Donde *N* corresponde a la cantidad de ejemplos anotados y *C* es el número de clases. Por otro lado, *Pe* puede ser calculado según:

$$Pe = \frac{1}{N^2}\sum_{i=1}^{C} N_{i.} \cdot N_i \tag{2.21}$$

Como se muestra en la Tabla 2.5, si el acuerdo entre los anotadores es total, $k = 1$. Si el porcentaje de acuerdo entre los anotadores es igual al porcentaje de acuerdo por el azar, $k = 0$ [39].

TABLA 2.5 Nivel de acuerdo según el índice de Kappa [40].

Kappa (k)	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

Desbalance de clases

Uno de los problemas recurrentes en la categorización de textos es el desbalance de clases. Un conjunto de datos está desbalanceado si las clases no están representadas equitativamente, afectando la predicción para la clase que es considerada positiva [41]. A medida que se aumenta la proporción de las instancias que representan a la clase positiva, el rendimiento del clasificador mejora [42]. De forma similar, si la cantidad de instancias positivas no son representativas en el conjunto de datos de entrenamiento, el desempeño empeora. Así, desbalances de clases tienen un impacto mayor si el clasificador utilizado es dependiente de la distribución de los datos como los algoritmos de clasificación probabilísticos o los basados en árboles de decisión [43].

Desbalances de clase también se producen cuando en problemas multiclase se utiliza la estrategia uno contra uno (OVO) para descomponer el problema original en una gran cantidad de clasificadores binarios [41].

Evaluación

La técnica más utilizada para la selección del conjunto de datos de entrenamiento y prueba es la validación cruzada [41]. En problemas con desbalance de clases, la validación cruzada estratificada ha sido utilizada para representar de mejor forma la distribución de las clases en cada conjunto de datos para analizar el desempeño de las clases minoritarias [44, 45].

La métrica más utilizada para evaluar el desempeño de un método de clasificación supervisada es la precisión predictiva (*accuracy*), que mide el porcentaje de aciertos de un clasificador. Sin embargo, en problemas con desbalance de clases, la precisión predictiva puede subestimar la evaluación de las clases minoritarias [46]. Otras métricas derivadas de la recuperación de información (*information retrieval*), se calculan para cada clase con el fin de analizar el efecto del desbalance de clases. Las más utilizadas son la precisión (*precision*), la exhaustividad (*recall*) y el valor-F1 (*F-measure* o *F1-score*). En recuperación de información textual, la precisión mide la fracción de los documentos recuperados que son relevantes para una determinada búsqueda, la exhaustividad mide la fracción de documentos relevantes que son exitosamente recuperados y el valor-F1 es la media armónica entre la precisión y la exhaustividad [41, 46]. La Fig. 2.16 muestra la relación entre precisión y exhaustividad a partir de los documentos relevantes y recuperados.

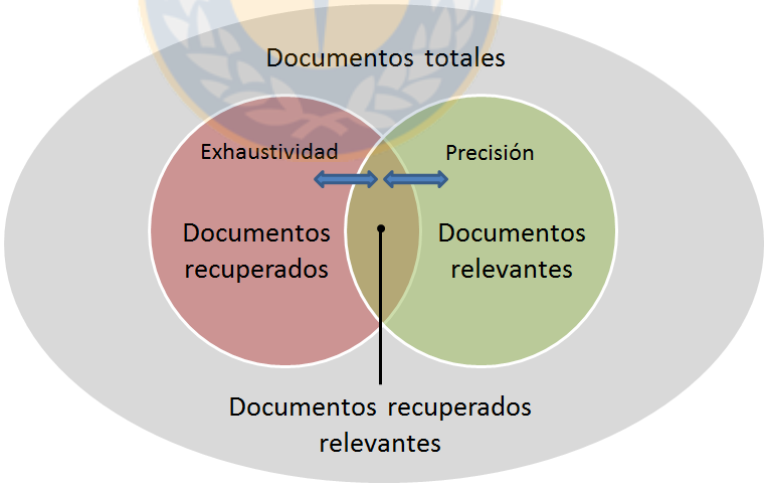


Fig. 2.16 Diagrama de Venn que representa la relación entre los documentos relevantes y recuperados[9].

[9] Disponible en: http://www.scielo.br/img/revistas/tinf/v24n2/a02f1.jpg. Fecha de último acceso: Noviembre de 2016

Algunas de las aplicaciones de la categorización de textos incluyen la organización de documentos, el filtrado de textos, la desambiguación del significado de las palabras, la categorización jerárquica de páginas web, entre otras [47].

2.2.6.1 Obesidad y comorbilidades en registros médicos electrónicos

La categorización automática de textos biomédicos ha despertado el interés de los investigadores en desarrollar herramientas que puedan organizar y crear conocimiento útil a partir de dicha información textual [8].

Desde que la obesidad se ha convertido en una pandemia mundial, se ha buscado desarrollar herramientas que faciliten el estudio, diagnóstico y tratamiento de esta enfermedad y sus comorbilidades utilizando registros médicos electrónicos [48-54]. Bordowitz *et al.* [55] implementaron un método para calcular automáticamente el IMC a partir de la información disponible de peso y altura, lo que permitió mejorar el tratamiento de la obesidad y el sobrepeso al complementar la información disponible sobre el estado nutricional de los pacientes. En el trabajo de Uzuner [9] se muestran los resultados de un desafío organizado por la *Informatics for Integrating Biology & the Bedside (i2b2)*, creado para evaluar métodos de identificación automática de la obesidad y sus comorbilidades en registros médicos electrónicos. Éstos fueron etiquetados por dos especialistas médicos a partir de información textual e intuitiva sobre la obesidad y quince de sus comorbilidades más frecuentes (ver Tabla 2.6). Para evaluar el desempeño de los métodos de clasificación se utilizó una medida ponderada del valor-F1, la cual asigna el mismo peso a cada clase del problema (macro-promedio). Yang *et al.* [11] y Solt [10] obtuvieron los mejores resultados en la identificación de las enfermedades. Yang *et al.* [11] utilizaron un conjunto de recursos semánticos, tales como sinónimos, conceptos, subconceptos, tratamientos y síntomas, obtenidos desde un sistema de terminología médica como es UMLS (*Unified Medical Language System*)[10]. Para clasificar los registros médicos electrónicos utilizaron diccionarios creados a partir de los recursos semánticos, junto a métodos basados en reglas y de aprendizaje automático como una SVM. En la clasificación textual, obtuvieron un macro-promedio del valor-F1 de un 81%, mientras que en la clasificación intuitiva obtuvieron un macro-promedio valor-F1 de 63%. Solt *et al.* [10]

[10] UMLS es un repositorio de terminología médica creado para facilitar la interoperabilidad en los sistemas de información médico. Disponible en: https://www.nlm.nih.gov/research/umls/. Fecha de ultimo acceso: Agosto de 2016

utilizaron un clasificador semántico basado en reglas, junto a términos claves de cada enfermedad, sinónimos, errores tipográficos más frecuentes, abreviaciones, entre otros. En la clasificación textual obtuvieron un macro-promedio del valor-F1 de 80%, mientras que en la clasificación intuitiva obtuvieron un macro-promedio del valor-F1 de 67%. Más recientemente, Murtaugh *et al.* [14] utilizaron expresiones regulares para extraer información sobre el estado nutricional de los pacientes, tales como el peso, la altura, el índice cintura-cadera y el IMC, obteniendo una precisión predictiva y valor-F1 sobre el 98%.

TABLA 2.6 Comorbilidades de la obesidad utilizadas en el desafío i2b2 [9].

Asma	Cálculo biliar/ colecistectomía	Hipertrigliceridemia
Aterosclerosis	Reflujo gastroesofágico	Apnea obstructiva del sueño
Insuficiencia cardiaca congestiva	Gota	Osteoartritis
Depresión	Hipercolesterolemia	Enfermedad vascular periférica
Diabetes *mellitus*	Hipertensión	Insuficiencia venosa

2.3 Expresiones regulares

Una expresión regular es una notación algebraica utilizada para representar cadenas de texto [56]. El objetivo de una expresión regular es representar todos los posibles lenguajes sobre un alfabeto Σ mediante operadores de composición y lenguajes primitivos [57]. Los operadores de composición son la unión, la concatenación, el cierre y el paréntesis. Los lenguajes primitivos son el lenguaje compuesto por la palabra vacía (λ), el lenguaje vacío (Ø) y el lenguaje formado por los elementos del alfabeto (aЄΣ). Por ejemplo, a partir de las expresiones regulares α y β, se pueden construir las siguientes expresiones regulares derivadas:

- α+β (operación de unión)
- α·β (operación de concatenación)
- α* v β* (operación de cierre)
- (α) (operación de paréntesis)

Las expresiones regulares pueden ser utilizadas para realizar búsqueda de información en textos mediante patrones predefinidos. Algunos ejemplos de búsqueda se muestran en la Tabla 2.7.

TABLA 2.7 Ejemplos de expresiones regulares.

Expresión regular	Significado	Ejemplo
imc\s+=\s+\d+	-imc seguido de = y uno o más números	el paciente tiene un **imc = 30** kg/m2
saturaci[óo]n\s+:\s+\d+[\.,]\d+	-saturación o saturacion seguido de : y un número decimal	porcentaje de **saturación : 97.6**%
[^\d+\s+]	-cualquier carácter excepto números y espacios	3 **meses y** 48 **semanas de vida**

Fuente: Elaboración propia.

Existen caracteres especiales, denominados metacaracteres, que cumplen un rol específico dentro de una expresión regular. Éstos son interpretados con un significado distinto a su representación normal. Por ejemplo, \d+ representa un dígito que puede ser repetido una o más veces. Algunos de los metacaracteres más utilizados se muestran en la Tabla 2.8.

TABLA 2.8 Principales metacaracteres utilizados en las expresiones regulares [56].

Operador	Significado
.	Cualquier carácter, excepto saltos de línea
^	Inicio de una cadena de texto
$	Final de una cadena de texto
[a-b]	Intervalo de caracteres entre a y b
|	Disyunción de caracteres
*	Repetición de caracteres cero o más veces
+	Repetición de caracteres una o más veces
?	Puede o no incluir una expresión regular previa
()	Subexpresión o grupo de caracteres
!	Exclusión de caracteres de un grupo
\	Permite escapar caracteres (\s, \d, \w, entre otros)
{n ,m}	Cantidad determinada de caracteres entre n y m
[^]	Negación de a caracteres

Las expresiones regulares tienen soporte en distintos lenguajes de programación incluyendo *Python* (librería *re*). Actualmente se utilizan en la práctica clínica para clasificar información de textos médicos con resultados comparables a los algoritmos de uso tradicional como SVM [13].

2.4 Algoritmo de Smith-Waterman

El algoritmo de Smith-Waterman (SW) es un método de programación dinámica que permite encontrar regiones de similitud local entre dos secuencias [58]. En 1981 fue propuesto por Temple Smith y Michael Waterman para el estudio de secuencias biológicas [59].

El algoritmo de SW asegura el descubrimiento de la alineación óptima local a través de un sistema de puntuaciones que favorece las coincidencias y penaliza las diferencias encontradas entre las secuencias analizadas [58]. Para esto, el algoritmo utiliza una matriz (H) donde se alinean las secuencias en una etapa de inicialización, llenado y rastreo, las cuales se describen a continuación.

Inicialización: En esta etapa, el algoritmo construye la matriz H de dimensiones $(n+1)x(m+1)$ y la inicializa según:

- $H(i,0) = 0,\ Si\ \ 0 < j < n$
- $H(0,j) = 0,\ Si\ 0 < j < m$

Donde n y m corresponden a las longitudes de las dos secuencias que serán alineadas. Por ejemplo, en la Fig. 2.17 se muestra la inicialización de la matriz H para las secuencias a = GTCCTAC y b = GTACGTATC.

SW	–	C	A	G	T	A	T	C	G	T
–	0	0	0	0	0	0	0	0	0	0
G	0									
T	0									
A	0									
C	0									
G	0									
T	0									
A	0									
T	0									
C	0									

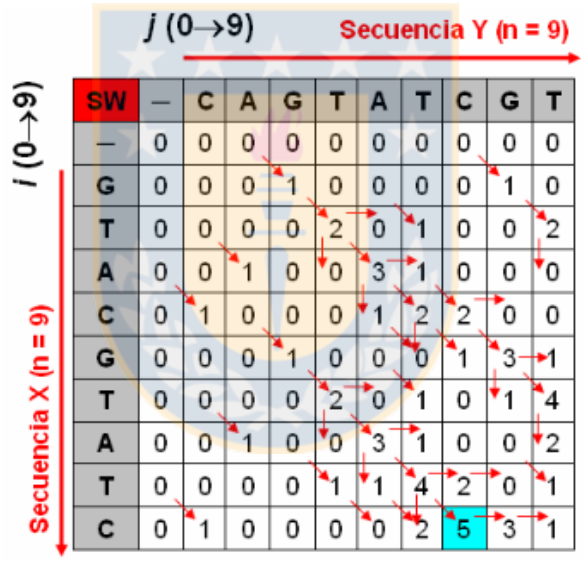
Fig. 2.17 Inicialización de la matriz de alineación dinámica de SW[11].

[11] Disponible en: http://www.ehu.eus/biofisica/juanma/bioinf/pdf/1_pairwise.pdf. Fecha de último acceso: Octubre de 2016

Llenado de la matriz: Se calcula el resto de los valores de la matriz *H* a partir de**:**

$$H(i,j) = max\begin{cases} 0 \\ H(i-1,j-1) + w(a_i, b_j) \\ H(i-1,j) + w(a_i, -) \\ H(i,j-1) + w(-, b_j) \end{cases} \quad ,Si\ 1 \le i \le m, 1 \le j \le n \tag{2.22}$$

Donde *w* es un factor que pondera la inserción, el borrado e igualdad/desigualdad de caracteres entre las secuencias. De la ecuación 2.22, el valor cero se utiliza para evitar valores de similitud negativos. Un desplazamiento diagonal *(i-1,j-1)* significa que el carácter a_i es igual o distinto al carácter b_j. Desde el punto de vista de la secuencia *a*, las dos últimas líneas de la ecuación 2.22 significan que se ha producido un borrado y una inserción, respectivamente. La Fig. 2.18 muestra el llenado de la matriz de la Fig 2.17.

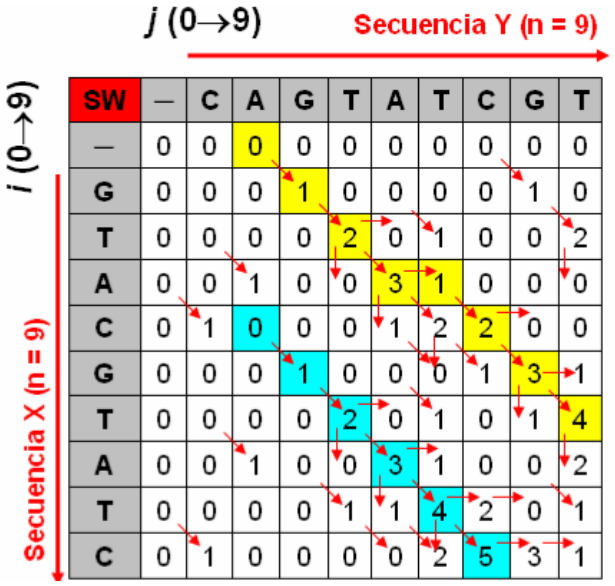


SW	–	C	A	G	T	A	T	C	G	T
–	0	0	0	0	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	1	0
T	0	0	0	0	2	0	1	0	0	2
A	0	0	1	0	0	3	1	0	0	0
C	0	1	0	0	0	1	2	2	0	0
G	0	0	0	1	0	0	0	1	3	1
T	0	0	0	0	2	0	1	0	1	4
A	0	0	1	0	0	3	1	0	0	2
T	0	0	0	0	1	1	4	2	0	1
C	0	1	0	0	0	0	2	5	3	1

Fig. 2.18 Llenado de la matriz de alineación dinámica de SW [12].

Rastreo (*backtracking*): En primer lugar, se localiza el valor máximo en la matriz *H*. Posteriormente, con desplazamientos hacia atrás respecto a la posición actual, se selecciona el valor máximo entre todas las direcciones posibles: *(i-1,j), (i,j-1)* o *(i-1, j-1)*. Los desplazamientos se realizan hasta que se encuentre una posición con valor cero en la matriz H (Fig. 2.19).

[12] Disponible en: http://www.ehu.eus/biofisica/juanma/bioinf/pdf/1_pairwise.pdf. Fecha de último acceso: Octubre de 2016



SW	–	C	A	G	T	A	T	C	G	T
–	0	0	0	0	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	1	0
T	0	0	0	0	2	0	1	0	0	2
A	0	0	1	0	0	3	1	0	0	0
C	0	1	0	0	0	1	2	2	0	0
G	0	0	0	1	0	0	0	1	3	1
T	0	0	0	0	2	0	1	0	1	4
A	0	0	1	0	0	3	1	0	0	2
T	0	0	0	0	1	1	4	2	0	1
C	0	1	0	0	0	0	2	5	3	1

Fig. 2.19 Alineamiento local de dos secuencias utilizando el algoritmo de SW. En turquesa la alineación óptima; en amarillo una alineación subóptima [13].

Finalmente, una vez trazada la ruta, se encuentra la alineación óptima local. En el ejemplo, la alineación óptima entre las secuencias a = GTCCTAC y b = GTACGTATC se muestra en la Fig. 2.20.

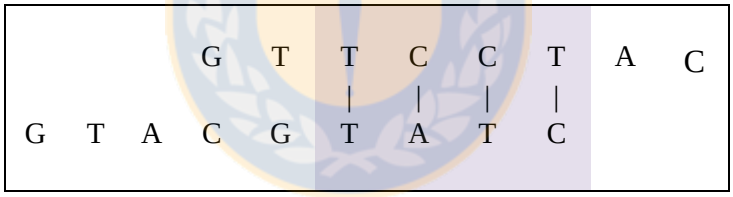


Fig. 2.20 Alineación óptima entre las secuencias GTCCTAC y GTACGTAT [14].

El algoritmo de SW ha sido recientemente utilizado en la clasificación de textos mediante el uso de expresiones regulares. Estas expresiones se crean a partir de los términos representativos de los textos, los cuales son extraídos mediante el algoritmo de SW [14].

[13] Disponible en: http://www.ehu.eus/biofisica/juanma/bioinf/pdf/1_pairwise.pdf. Fecha de último acceso: Octubre de 2016

[14] Disponible en: http://www.ehu.eus/biofisica/juanma/bioinf/pdf/1_pairwise.pdf. Fecha de último acceso: Octubre de 2016

2.5 Discusión

En la revisión bibliográfica realizada, se puede observar que los métodos utilizados en la identificación de la obesidad y sus comorbilidades se basan en la construcción de diccionarios con la terminología médica utilizada en los registros médicos electrónicos para referirse de forma explícita a las enfermedades, así como a los tratamientos e indicaciones médicas [9-11]. Por otro lado, uno de los principales indicadores de obesidad es el IMC. No obstante, este indicador no siempre está disponible en los registros médicos electrónicos o puede estar presente en forma de texto libre [55].

De acuerdo a lo señalado en la Tabla 2.3, existe una estrecha relación entre las comorbilidades de la obesidad y el IMC. A medida que el IMC aumenta, el riesgo de sufrir comorbilidades también aumenta. Por ende, se espera tener un mayor registro de comorbilidades y terminología médica asociada a estas enfermedades en pacientes con obesidad, lo que debería intensificarse en sus tipos más severos, que en pacientes con un IMC menor a 30.

El algoritmo de SW está siendo utilizado recientemente en categorización de textos, gracias a su capacidad de encontrar regiones de similitud locales entre dos secuencias de texto [13]. En particular, su uso en la generación automática de expresiones regulares ha despertado el interés de los investigadores para el desarrollo de aplicaciones clínicas que permitan generar conocimiento a partir de los registros médicos electrónicos [13, 14]. En el caso de la identificación de obesidad, se espera que este algoritmo permita encontrar los términos más representativos (*tokens)* asociados a cada clase para que le permitan, a un algoritmo de clasificación supervisada, distinguir claramente sobre la presencia o ausencia de obesidad, y diferenciar claramente los tipos de obesidad.

Finalmente, en cuanto a los algoritmos de clasificación jerárquica, se implementarán los métodos de transformación binaria y multiclase [32, 33], y se comparará su desempeño con un algoritmo jerárquico propuesto para esta tesis. De esta forma, se busca analizar la dependencia existente entre las clases obesidad y sus tipos, y cómo afecta ésta en el desempeño de los clasificadores implementados.

Capítulo 3. Materiales y Métodos

En este capítulo se describe el conjunto de datos y los métodos utilizados para clasificar los registros médicos electrónicos. Se detalla la metodología utilizada en el proceso de clasificación desde la etapa de preprocesamiento hasta la presentación de los distintos algoritmos de clasificación supervisada utilizados. Finalmente, se muestran las métricas utilizadas para la evaluación de los algoritmos utilizados.

3.1 Descripción del conjunto de datos

En esta tesis se utilizaron 66179 registros médicos electrónicos de-identificados provenientes del Hospital Guillermo Grant Benavente (HGGB) de Concepción. Estos registros médicos fueron recolectados entre el año 2011 y 2012 y contienen información de 46 especialidades médicas (ver Tabla C.1 del anexo). Como requerimiento de la unidad de investigación del HGGB fue necesario firmar una declaración de reserva y confidencialidad de la información, y una declaración de cumplimiento de las buenas prácticas clínicas en marco del proyecto FONDECYT "*Adaptive Selection of training set based on Active Learning*", código 11121463 (ver anexo A).

Cada registro médico electrónico contiene campos estructurados, donde se reportan la especialidad médica de la atención, sexo, factores de riesgo, hábitos y signos vitales de los pacientes, y de texto libre o no estructurado, donde se reportan exámenes físicos, historial médico, observaciones e indicaciones médicas.

3.2 Clasificación de los registros médicos electrónicos

Para la identificación de la obesidad en los registros médicos electrónicos, se utilizó la metodología que se muestra en la Fig. 3.1.

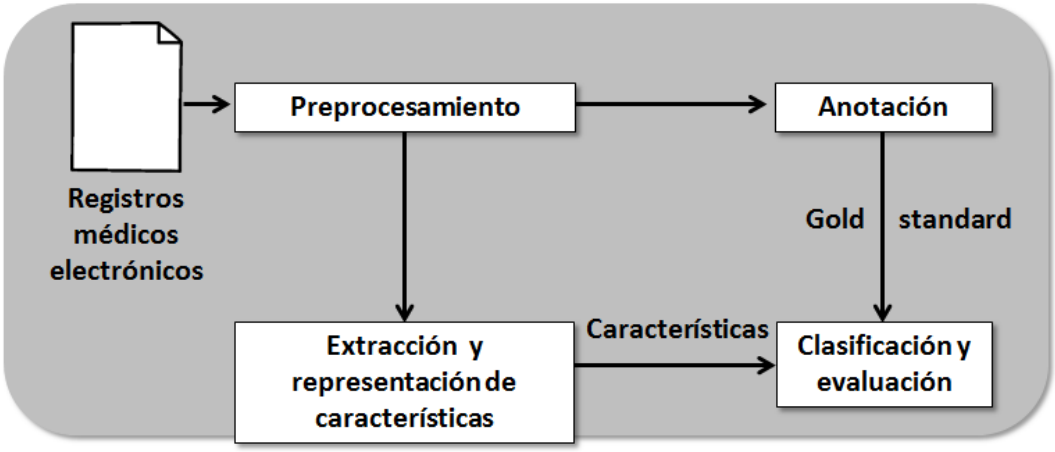


Fig. 3.1 Metodología utilizada en la identificación del estado nutricional de los pacientes en registros médicos electrónicos.

Fuente: Elaboración propia.

3.2.1 Preprocesamiento

Cada texto contenido en los registros médicos electrónicos fue procesado en cuatro etapas. La primera etapa consistió en normalizar cada texto, convirtiendo palabras a minúsculas y eliminando caracteres no alfanuméricos y las palabras vacías o *stopwords* (excluyendo las negaciones), las cuales fueron obtenidas desde la librería NLTK (*Natural Language ToolKit*) de *Python*[15]. En segundo lugar, se reemplazaron todos los valores de IMC por su valor mínimo, de acuerdo a las categorías de la Tabla 3.1, con la finalidad de favorecer la extracción de características de los valores pertenecientes a una misma categoría. La identificación del IMC en los registros fue realizada mediante expresiones regulares (por ejemplo, imc\s+\d+).

TABLA 3.1 IMC y valor mínimo utilizado en su normalización [7, 16].

Estado nutricional	IMC	Valor mínimo
Bajo peso	<18.5	0
Peso normal	18.5-24.9	18.5
Sobrepeso	25-29.9	25
Obesidad Moderada	30-34.9	30
Obesidad Severa	35-39.9	35
Obesidad Mórbida	40-49.9	40
Superobesidad	≥50	50

[15] Disponible en: www.nltk.org. Fecha de último acceso: Agosto de 2016

En tercer lugar, se creó un diccionario de comorbilidades asociadas a la obesidad, tomando como base las 15 enfermedades seleccionadas por Ozüner [9] (posteriormente, este diccionario fue ampliado en la etapa de anotación, añadiéndose dos enfermedades sugeridas por los anotadores: el hipotiroidismo y la enfermedad de *Cushing*). Al final de esta etapa, se obtuvo un diccionario de 507 términos conteniendo variaciones lingüísticas y clínicas de cada una de las 17 comorbilidades de la obesidad.

Finalmente, se creó un diccionario de palabras clave sobre el estado nutricional de los pacientes, considerando los términos de la Tabla 3.1 y el IMC, para filtrar los registros médicos electrónicos que no contenían estos términos mediante expresiones regulares. Además, se consideró que los registros médicos electrónicos recuperados fueran únicos y que correspondieran a pacientes adultos, de acuerdo a la especialidad médica donde se realizó la atención (ver Tabla C.1 del Anexo). Al finalizar esta etapa, se recuperaron 2701 registros médicos electrónicos con información relevante al estudio.

3.2.2 Anotación de los registros médicos electrónicos

Se definieron dos problemas de clasificación para la identificación de la obesidad y sus tipos. En el primer problema de clasificación las clases fueron obesidad (O) y ausencia de obesidad (NO), mientras que en el segundo problema las clases fueron: obesidad no mencionada (ONM), obesidad mórbida (OMO), obesidad severa (OS) y obesidad moderada (OM) [7, 16].

Posteriormente, se creó un *gold standard* para ambos problemas de clasificación. Dos estudiantes con conocimientos de las ciencias biomédicas revisaron y anotaron los 2701 registros médicos electrónicos recuperados en la etapa de preprocesamiento mediante la utilización de una herramienta de anotación diseñada en *QT-designer* y programada en *Python* (ver Fig. B.1 del Anexo). Éstos debían asignar una etiqueta correspondiente a alguna de las clases del primer problema de clasificación, y en caso de que el registro médico electrónico fuese etiquetado con la clase obesidad, debían asignar la etiqueta correspondiente al tipo de obesidad.

De igual forma, el grupo de anotadores debía entregar información de palabras claves sobre el estado nutricional de los pacientes y las comorbilidades de la obesidad para complementar los diccionarios creados en la etapa de preprocesamiento.

Al finalizar la anotación, se filtraron los registros médicos electrónicos que los anotadores consideraron falsos positivos, los cuales hacían mención a palabras claves del estado nutricional de

los pacientes, aunque no fuesen relevantes para este estudio (por ejemplo, “bajo peso molecular”). La cantidad de registros médicos electrónicos se redujo a 2610 para el primer problema de clasificación y a 2003 para el segundo problema de clasificación (subconjunto de registros del primer problema de clasificación).

Finalmente, un tercer anotador resolvió cualquier desacuerdo en la asignación de las etiquetas y midió el nivel de acuerdo entre ambos anotadores mediante el índice de *Kappa* de Cohen (k) [40].

3.2.3 Extracción y representación de características

El análisis exploratorio de los datos permitió establecer que en los campos de los registros médicos electrónicos, en los cuales se hace mención a las comorbilidades de la obesidad y el IMC, se puede encontrar información relevante para el estudio de la obesidad. Por esta razón, se filtraron los campos de los registros médicos que no contenían información sobre las comorbilidades de la obesidad y el IMC, utilizando las palabras claves del diccionario que contenía información sobre las 17 enfermedades más habituales de la obesidad. A partir de estos campos (de texto libre y estructurado) se procedió a extraer características, tokenizando cada uno de los textos mediante secuencias de un término (unigramas), de dos términos consecutivos (bigramas) y una combinación de ambos. Además, se utilizó el algoritmo de SW para extraer características, alineando de forma local pares de texto. La cantidad de alineaciones utilizadas por el algoritmo de SW está dada por [13]:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \tag{3.1}$$

Donde n es el número de textos, y k = 2. La cantidad de características obtenidas para cada problema de clasificación se muestra en la Tabla 3.2.

TABLA 3.2 Cantidad de características extraídas en cada problema de clasificación.

Característica	Cantidad	
	Primer problema	**Segundo problema**
Unigramas	5591	4362
Bigramas	22703	14030
Unigramas+bigramas	28294	18392
SW	9755	6221

Fuente: Elaboración propia.

Posteriormente, cada registro médico electrónico de cada problema de clasificación fue representado mediante el enfoque de bolsa de palabras (BOW) [35] utilizando las características (*tokens*) extraídas. Estas características fueron ponderadas mediante el método TF-IDF [60]:

$$TF - IDF(t, d, D) = TF(t, d) \cdot IDF(t, D) \tag{3.2}$$

Con:

$$IDF(t) = log_{10}(\frac{D}{d})$$

Donde TF es la frecuencia del término *t*, IDF es la frecuencia inversa del documento, *D* es el número de registros médicos electrónicos, d es el número de registros médicos electrónicos donde aparece el término *t*.

Finalmente, cada registro médico electrónico (EMR) es representado matricialmente M_{mxn}, donde cada fila de la matriz representa a un registro, mientras que cada columna corresponde al valor de cada característica ponderada por el método TF-IDF. Adicionalmente, se añadió una columna con la clase a la que pertenece cada registro para entrenar los algoritmos de clasificación (Fig. 3.2).

$$\left[\begin{array}{ccc|c} token_1 & \cdots & token_n & Clase \\ X_{1,1} & & X_{1,1} & Y_1 \\ \vdots & \ddots & \vdots & \vdots \\ X_{m,1} & \cdots & X_{m,n} & Y_m \end{array}\right]$$

Fig. 3.2 Representación matricial de los registros médicos electrónicos para fines de clasificación.

Fuente: Elaboración propia.

3.2.4 Clasificación y evaluación

En esta etapa se utilizaron dos enfoques para clasificar los registros médicos electrónicos: tratar cada problema de clasificación de forma independiente; e implementando algoritmos de clasificación jerárquica, debido a la dependencia existente entre la clase obesidad y sus tipos. En ambos enfoques se implementaron los clasificadores NB y una SVM utilizando la librería *Scikit-learn* de *Python* [61]. En el caso de NB, se implementó en ambos problemas de clasificación la variante *Naïve Bayes* multinomial (MNB), la cual es profusamente utilizada en la clasificación de

textos [62]. En el caso de SVM, en la identificación de los tipos de obesidad, se implementó el enfoque OVA, al tratarse de un problema multiclase.

En la clasificación jerárquica se implementaron tres métodos: transformación binaria, transformación multiclase, y un algoritmo que simula una clasificación jerárquica que utiliza los TP de la clase obesidad para entrenar y evaluar los tipos de obesidad (ver Fig. 3.3).

I .- Sean
P_1:{ UW, NW, OW, O}: Etiquetas para el primer problema de clasificación.
P_2:{ MO, SO, M, S}: Etiquetas para el segundo problema de clasificación.
Tr1, Tr2, Tt1, Tt2: Datos de entrenamiento y prueba para el primer y segundo problema de clasificación.
f2: Conjunto de características para el segundo problema de clasificación.
C1, C2: Clasificadores para el primer y segundo problema de clasificación.
TP: Verdadero positivo.
II.- Inicialización
Tr1 ← Datos etiquetados para C1
Tr2 ← 0
III.- Algoritmo
Entrenar C1 utilizando Tr1.
Para todo $x_i \in$ Tr1
  Si (etiqueta de x_i == 'O')
    Si (f2 está disponible para x_i)
      Obtener f2 para x_i
      Añadir x_i a Tr2
    Fin
  Fin
Fin
Evaluar C1 utilizandoTt1.
Para todo $y_i \in$ Tt1
  Si y_i es TP
    Si (f2 está disponible para y_i)
      Añadir y_i a Tr2
    Fin
  Fin
Fin
Entrenar C2 utilizando Tr2.
Evaluar C2 utilizando Tt2.

Fig. 3.3 Algoritmo de clasificación jerárquico propuesto.

Fuente: Elaboración propia.

Para evaluar el rendimiento de los algoritmos de clasificación, se utilizó validación cruzada con *k-fold* durante 10 veces para obtener una estimación confiable del error de clasificación [63, 64]. Se estableció un valor de $k = 5$, debido a que el conjunto de datos no es muy grande. En cada

ejecución, se promediaron las siguientes medidas de desempeño: la precisión predictiva (ACC), el valor-F1 (F1), la tasa de falsos positivos (FPR) y la tasa de falsos negativos (FNR), cuyas ecuaciones se muestran a continuación:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.3}$$

$$\mathrm{F1} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.4}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.5}$$

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

Para cada una de estas métricas se calculó un promedio ponderado, usando como peso el número de ejemplos por clase. De esta forma, se busca analizar el impacto de la distribución de las clases en cada problema de clasificación. Además, se realizó un *t-test* para comparar el desempeño de SVM y NB con un nivel de significancia igual a 0.05.

Adicionalmente, para la evaluación de los algoritmos de múltiples etiquetas de transformación binaria y multiclase, se calculó la métrica denominada pérdida 0/1 (PCU) que es utilizada para medir la dependencia existente entre las clases de una jerarquía [65]. Ésta es definida como:

$$PCU = \frac{1}{n} \sum_{i=1}^{n} [y_i \neq h_i] \tag{3.7}$$

Con y_i la clase de la instancia actual y h_i la clase predicha por el clasificador.

Capítulo 4. Resultados

En este capítulo se presentan los resultados obtenidos de las distintas etapas del proceso de clasificación de los registros médicos electrónicos. Se muestra un análisis exploratorio de los datos, el proceso de ajuste y la evaluación de los clasificadores SVM y NB.

4.1 Análisis exploratorio de los datos

Se realizó un análisis de los datos utilizados en cuanto al uso de los registros médicos electrónicos y estadísticas a partir del *gold standard* creado para la clasificación.

Se obtuvo un índice de *Kappa (k)* de 0.97 y 0.96 en el primer y segundo problema de clasificación, respectivamente. Estos resultados indican que el nivel de acuerdo entre ambos anotadores fue casi perfecto [40], por lo tanto, el *gold standard* creado puede ser considerado confiable para ser utilizado en los modelos de clasificación.

La Fig. 4.1 muestra la distribución de la cantidad de registros médicos electrónicos para cada problema de clasificación después del proceso de anotación. Existe un notorio desbalance de clases en ambos problemas de clasificación, favoreciendo a la clase obesidad y a los tipos no mencionados de obesidad.

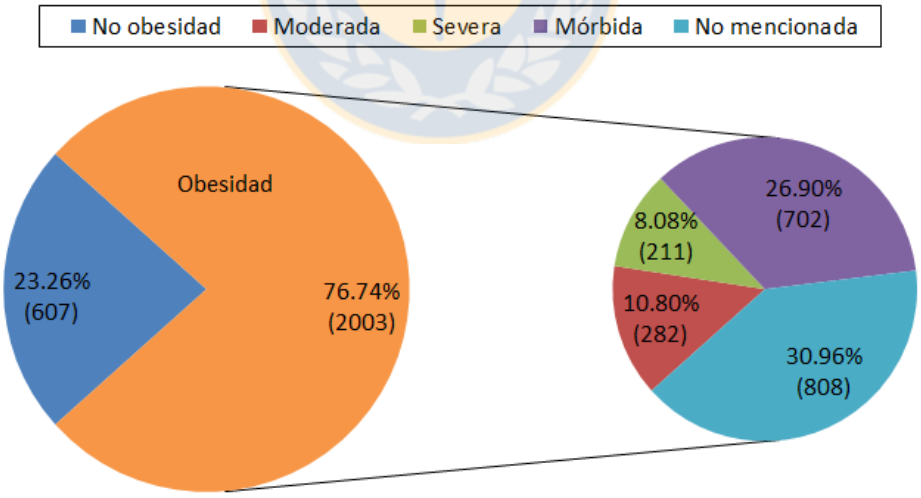


Fig. 4.1 Distribución de las clases en ambos problemas de clasificación.

Fuente: Elaboración propia.

El 83.82% de los registros médicos electrónicos, etiquetados con la clase obesidad, corresponden a mujeres (ver Fig. 4.2.A). Por otro lado, el 94.23% de los registros médicos anotados

donde se reportó sedentarismo en el campo de hábitos de los pacientes, corresponden a pacientes con obesidad. De estos registros, el 92.31% corresponde a mujeres (ver Fig. 4.2.B).

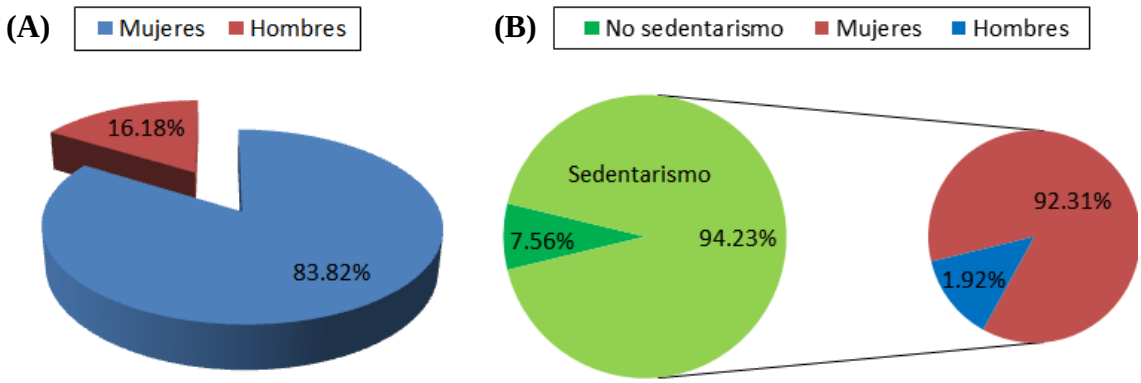


Fig. 4.2 Distribución de pacientes con obesidad por sexo y reporte de sedentarismo en el campo hábitos.

Fuente: Elaboración propia.

La Fig. 4.3 muestra que sólo el 3.94% del total los registros médicos electrónicos contienen información sobre la presencia o ausencia de obesidad, siendo los campos no estructurados los que mayor información aportaron en la recuperación de información.

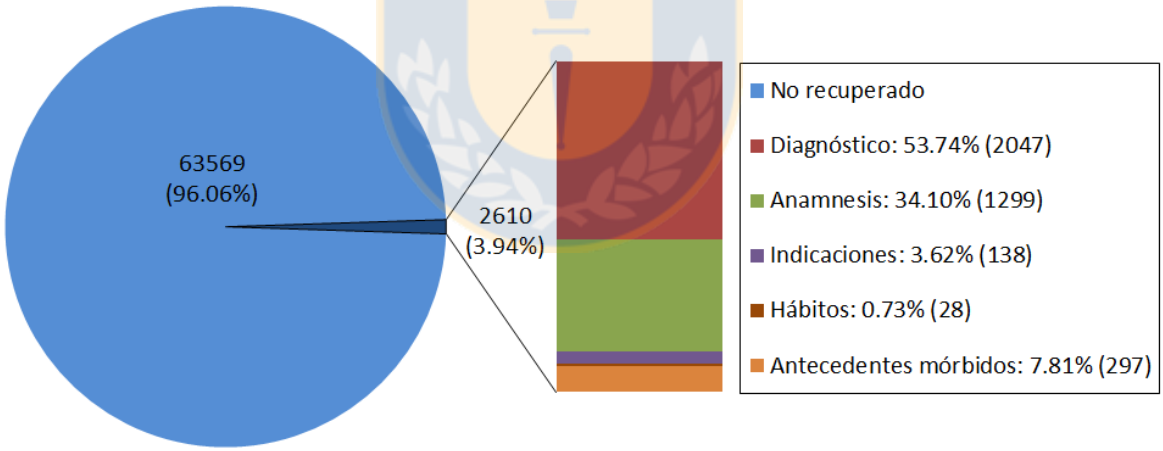


Fig. 4.3 Registros médicos electrónicos recuperados y distribución de los campos que contenían términos claves para la recuperación de información.

Fuente: Elaboración propia.

La Fig. 4.4 muestra la distribución de las diferentes especialidades médicas en los registros médicos etiquetados con obesidad. Endocrinología Adulto y Cirugía Adulto concentran el 53.87% de las atenciones médicas en obesidad.

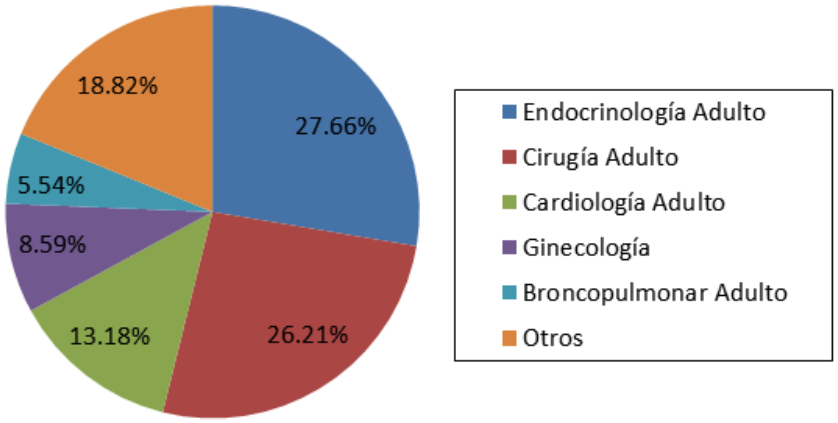


Fig. 4.4 Distribución de las especialidades médicas asociadas a la obesidad.

Fuente: Elaboración propia.

La Fig. 4.5 muestra la distribución de las principales comorbilidades de la obesidad, según lo reportado en el campo de antecedentes mórbidos de los pacientes (se muestran los porcentajes mayores a 1%). La diabetes *mellitus* y la hipertensión tienen la mayor prevalencia entre los pacientes con obesidad, con un porcentaje mayor al 20% en ambos casos.

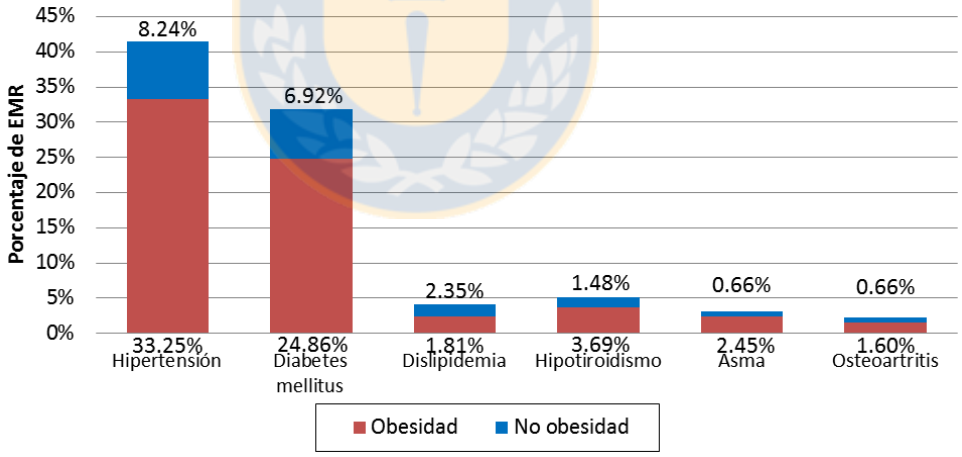


Fig. 4.5 Principales comorbilidades presentes en los registros médicos electrónicos etiquetados.

Fuente: Elaboración propia.

4.2 Experimentos de ajuste de los clasificadores

Una vez que cada registro médico electrónico fue representado mediante el método TF-IDF, se seleccionó el 20% de estos registros para ajustar los parámetros de los clasificadores NB y SVM. En el caso de NB se buscó ajustar el parámetro α (suavizamiento que evita probabilidades cero) y las probabilidades a *priori* (distribución de las clases). En SVM se buscó ajustar el parámetro *C* (penalización del error de clasificación durante el entrenamiento) y el tipo de *kernel* (se consideró un *kernel* lineal y uno de base radial, ampliamente utilizados en clasificación de textos [66]). En el caso del *kernel* de base radial (RBF) también se ajustó el parámetro γ (regulación del alcance de los vectores de soporte). En la Tabla 4.1 se muestra un resumen de los parámetros y los intervalos utilizados en el ajuste de cada clasificador.

TABLA 4.1 Parámetros finales ajustados para cada clasificador.

Clasificador	Parámetro	Opciones
SVM	*kernel*	lineal, RBF
	C	10^0, 10^1, 10^2, 10^3, balanceado
	$\gamma^{(*)}$	10^{-1},10^{-2},10^{-3},10^{-4}
NB	α	[0,1]
	prior	uniforme, omisión, balanceado

(*) Sólo para el *kernel* RBF

Fuente: Elaboración propia.

En el caso de SVM, un parámetro *C* balanceado ajusta automáticamente los pesos de forma inversamente proporcional a las frecuencias de las clases. Por otro lado, en el caso de NB, un *prior* uniforme utiliza una distribución de frecuencias equiprobable, una distribución a *priori* por omisión utiliza la distribución de frecuencias originales del problema, mientras que un *prior* balanceado utiliza una distribución inversamente proporcional a la frecuencia de las clases.

Los parámetros de cada clasificador fueron ajustados mediante una búsqueda exhaustiva (*Exhaustive Grid Search* [16]) utilizando la biblioteca *Scikit-learn* de *Python*. Se realizó una combinación de cada uno de los parámetros de la Tabla 4.1 para seleccionar el mejor modelo de clasificación mediante validación cruzada (considerando la cantidad de datos, se determinó un valor de $k = 5$). Como métrica de desempeño para seleccionar el mejor modelo de clasificación, se utilizó

[16] GridSearchCV, disponible en: http://scikit-learn.org/stable/modules/grid_search.html. Fecha de ultimo acceso: Noviembre de 2016

el promedio ponderado del valor-F1, utilizando como peso la cantidad de clases en cada problema de clasificación.

Una vez que los parámetros fueron ajustados, se procedió a realizar selección de características mediante un criterio basado en IG [24] con el uso del paquete *FSelector*[17], implementado en el lenguaje de programación *R*. Se seleccionó el mejor subconjunto para cada tipo de característica en ambos clasificadores, buscando el porcentaje que permitía el máximo promedio ponderado del valor-F1.

Las Fig. 4.6 y 4.7 muestran las curvas de desempeño de los clasificadores SVM y NB en función del porcentaje de características, seleccionado mediante el método de IG (se destaca el valor máximo). En la Fig. 4.6 se muestra que SVM obtiene siempre el mejor desempeño en ambos problemas de clasificación, utilizando características extraídas mediante el algoritmo de SW.

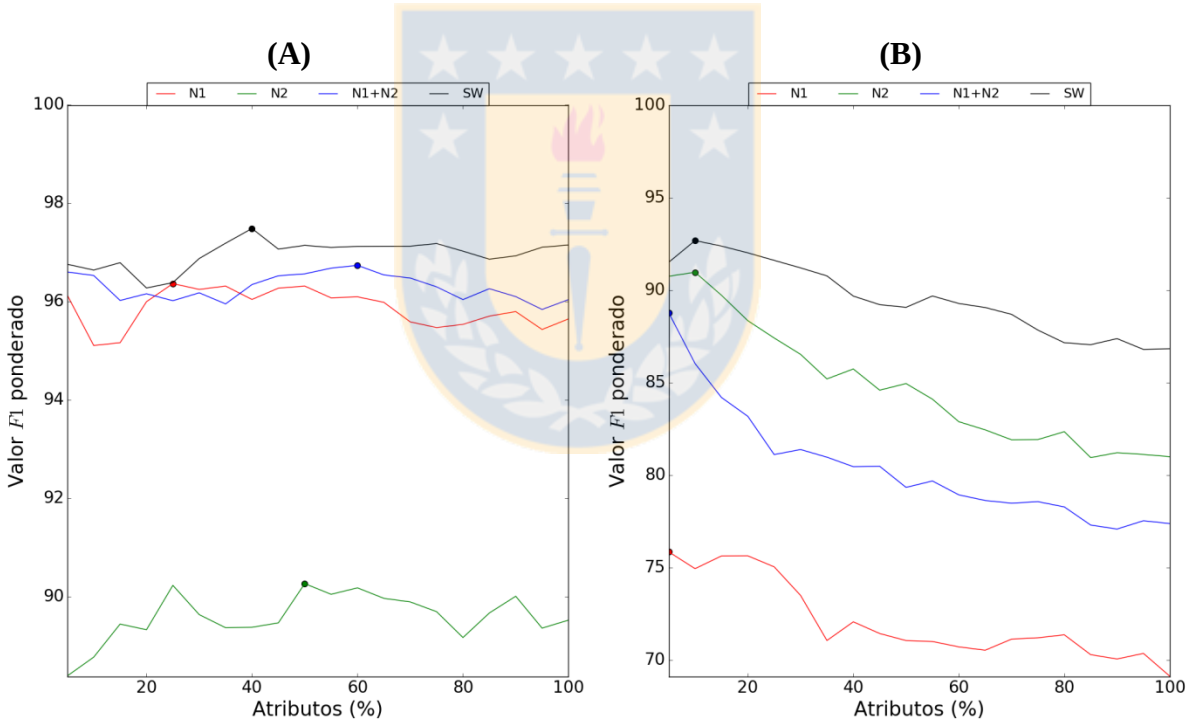


Fig. 4.6 Desempeño de SVM con el uso de diferentes características. A: primer problema de clasificación. B: segundo problema de clasificación.

Fuente: Elaboración propia.

[17] FSelector, disponible en: https://cran.r-project.org/web/packages/FSelector/FSelector.pdf. Fecha de último acceso: Noviembre de 2016

En la Fig. 4.7 se muestra que NB obtiene el máximo desempeño en ambos problemas de clasificación, utilizando características extraídas mediante el algoritmo de SW. En el segundo problema de clasificación, este desempeño se prolonga por toda la curva.

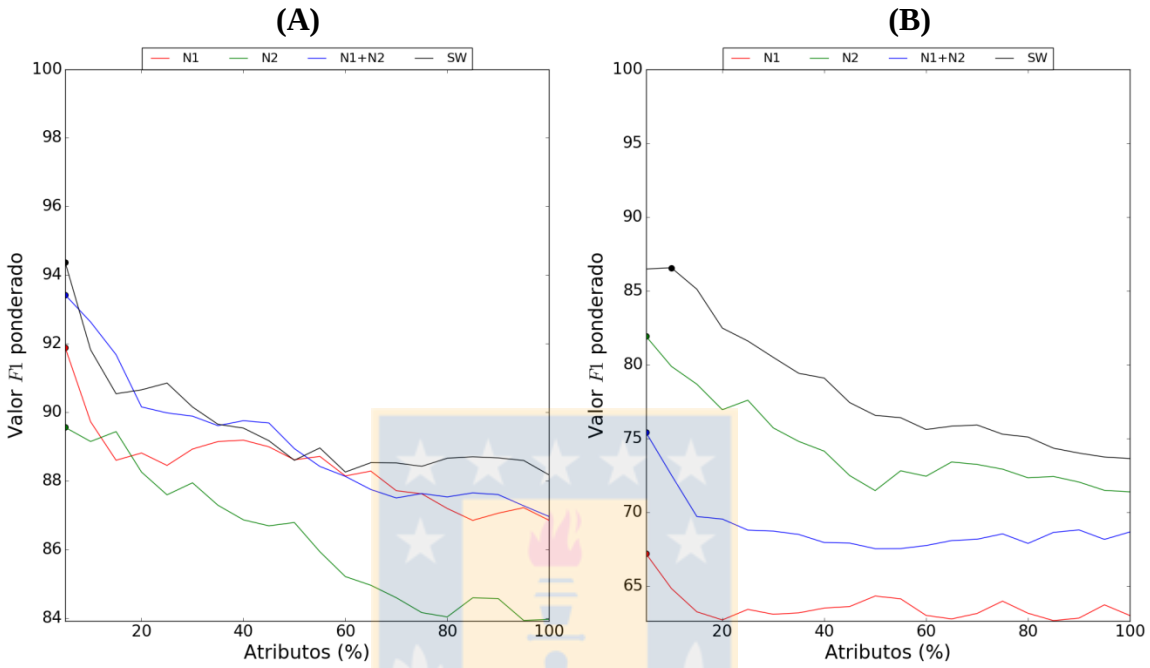


Fig. 4.7 Desempeño de NB con el uso de diferentes características. A: primer problema de clasificación. B: segundo problema de clasificación.

Fuente: Elaboración propia.

En la Tabla. 4.2 se muestra un resumen de los porcentajes de características seleccionados en cada problema de clasificación para ambos clasificadores. En todos los casos, el máximo valor-F1 ponderado se obtiene con el uso de características extraídas mediante el algoritmo de SW.

TABLA 4.2 Porcentaje de características seleccionado en cada problema de clasificación.

Problema de clasificación	Clasificador	Característica	Porcentaje seleccionado (%)	Máximo Valor-F1 ponderado (%)
Primer (obesidad/ ¬obesidad)	SVM	N1	25	96.37
		N2	50	90.27
		N1+N2	60	96.74
		SW	40	**97.49**
	NB	N1	5	75.86
		N2	5	90.96
		N1+N2	5	88.78
		SW	5	**92.68**
Segundo (tipos de obesidad)	SVM	N1	5	91.89
		N2	10	89.56
		N1+N2	5	93.43
		SW	10	**94.38**
	NB	N1	5	67.19
		N2	5	81.92
		N1+N2	5	75.40
		SW	10	**86.55**

Fuente: Elaboración propia.

Una vez establecido cada subconjunto de características, se realizó un nuevo ajuste de parámetros para cada clasificador y se evaluó su desempeño. Los parámetros finales ajustados para cada clasificador se muestran en la Tabla 4.3.

TABLA 4.3 Parámetros finales ajustados para cada clasificador.

Problema	Clasificador	Parámetros	Ajuste			
			N1	N2	N1+N2	SW
Primer (obesidad/ ¬obesidad)	SVM	*kernel*	RBF	Lineal	RBF	RBF
		C	balanceado	balanceado	balanceado	100
		γ	0.001	-	0.0001	0.001
	NB	α	0.8	1	0.9	0.1
		prior	omisión	Omission	omisión	omisión
Segundo (tipos de obesidad)	SVM	*kernel*	RBF	RBF	RBF	RBF
		C	balanceado	100	100	balanceado
		γ	0.01	0.001	0.001	0.0001
	NB	α	0.1	0.3	0.6	0.2
		prior	omisión	Omission	omisión	omisión

Fuente: Elaboración propia.

La selección de características permitió mejorar el desempeño de los clasificadores en ambos problemas de clasificación. Sin selección de características, el uso del algoritmo de SW permitió obtener los mejores desempeños en ambos clasificadores, excepto con NB en el uso de unigramas y

bigramas (N1+N2). Sin embargo, al realizar selección de características, el uso del algoritmo de SW permitió obtener los mejores desempeños en todos los casos, tal como se muestra en la Tabla 4.4.

TABLA 4.4 Desempeño de los clasificadores luego de la selección de características.

Problema	Clasificador	Característica	Valor-F1 ponderado sin selección (%)	Valor-F1 ponderado con selección (%)	Variación porcentual (%)
Primer (obesidad/ ¬obesidad)	SVM	N1	94.08	96.08	2.13
		N2	85.32	87.83	2.94
		N1+N2	93.13	94.11	1.05
		SW	**96.01**	**97.10**	1.14
	NB	N1	82.18	85.36	3.87
		N2	78.37	88.04	12.34
		N1+N2	**82.26**	88.52	7.61
		SW	79.69	**90.94**	14.12
Segundo (tipos de obesidad)	SVM	N1	61.58	75.65	22.85
		N2	73.14	89.09	21.81
		N1+N2	67.73	87.56	29.28
		SW	**78.70**	**90.77**	15.34
	NB	N1	54.22	62.71	15.66
		N2	60.24	76.87	27.61
		N1+N2	60.96	68.36	12.14
		SW	**63.66**	**77.68**	22.02

Fuente: Elaboración propia.

4.3 Clasificación

Las Tablas 4.5 y 4.6 muestran los resultados de la clasificación en los dos problemas de clasificación. En la Tabla 4.5 se observa que SVM obtiene un mejor desempeño que NB si se compara cada tipo de características en el primer problema de clasificación. En particular, con el uso de SW, SVM obtiene el mejor desempeño en términos de precisión predictiva (ACC), valor-F1 y en todos los promedios ponderados de las medidas de desempeño incluyendo las FPR y las FNR. En el caso de NB, los mejores desempeños se logran mediante el uso de bigramas y SW. En particular, con SW, NB logra un mejor desempeño del valor-F1 para la clase NO y el promedio ponderado de esta métrica. Por otro lado, se observa que la clase obesidad (O) tiene altos valores en la tasa de falsos positivos, mientras que la clase no obesidad (NO) tiene altos valores en la tasa de falsos negativos. Finalmente, en todos los casos se presentaron diferencias significativas entre el desempeño de ambos clasificadores ($p<0.05$).

TABLA 4.5 Desempeño de los clasificadores en el primer problema de clasificación con el enfoque no jerárquico.

Característica	Clasificador	Clase/Promedio	ACC (%)	F1 (%)	FPR (%)	FNR (%)
N1	SVM	NO	92.96	84.63	4.20	16.34
		O	92.96	95.42	16.34	4.20
		Promedio ponderado	92.96	92.90	13.51	7.03
	NB	NO	86.31	70.22	8.61	30.45
		O	86.31	91.08	30.45	8.61
		Promedio ponderado	86.31	86.21	25.35	13.71
N2	SVM	NO	94.56	87.78	2.31	15.77
		O	94.56	96.49	15.77	2.31
		Promedio ponderado	94.56	94.46	12.63	5.45
	NB	NO	91.36	80.00	3.48	25.59
		O	91.36	94.48	25.59	3.48
		Promedio ponderado	91.36	91.10	20.43	8.64
N1+N2	SVM	NO	96.04	91.09	**1.27**	12.76
		O	96.04	97.45	12.76	**1.27**
		Promedio ponderado	96.04	95.97	10.08	3.95
	NB	NO	88.37	75.36	8.10	23.28
		O	88.37	92.37	23.28	8.10
		Promedio ponderado	88.37	88.40	19.74	11.64
SW	SVM	NO	**97.10**	**93.64**	1.32	**8.09**
		O	**97.10**	**98.12**	**8.09**	1.32
		Promedio ponderado	**97.10**	**97.07**	**6.51**	**2.90**
	NB	NO	91.23	82.25	7.63	12.56
		O	91.23	94.15	12.56	7.63
		Promedio ponderado	91.23	91.37	11.41	8.78

Fuente: Elaboración propia.

La Tabla 4.6 muestra que SVM obtiene los mejores desempeños que NB en todos los tipos de características en términos de precisión predictiva (ACC) y valor-F1. En general, con el uso de SW, SVM obtiene los mejores desempeños en ACC, valor-F1, FPR y FNR y el mejor promedio ponderado de estas métricas. En el caso de NB, con el uso de SW se logra el mejor desempeño en términos de ACC y valor-F1. Se observa también que la clase obesidad no mencionada (ONM) tiene los valores más altos en la tasa de falsos positivos, mientras que la clase obesidad severa (OS) tiene los valores más altos en la tasa de falsos negativos.

TABLA 4.6 Desempeño de los clasificadores en el segundo problema de clasificación con el enfoque no jerárquico.

Característica	Clasificador	Clase/Promedio	ACC (%)	F1 (%)	FPR (%)	FNR (%)
N1	SVM	OM	91.15	67.48	4.71	33.89
		OS	90.09	46.04	4.15	58.79
		OMO	91.36	86.93	3.55	18.00
		ONM	89.27	87.77	15.08	**4.32**
		Promedio ponderado	90.35	80.24	8.43	18.99
	NB	OM	90.86	63.09	3.56	43.32
		OS	87.58	40.95	7.03	58.45(*)
		OMO	83.39	77.34	15.41	18.79
		ONM	83.16	78.98	13.65	21.53
		Promedio ponderado	84.79	72.18	12.15	27.51
N2	SVM	OM	95.54	82.96	1.62	21.86
		OS	**94.43**	68.55	**1.35**	41.28
		OMO	**92.27**	88.44	3.53	15.49
		ONM	88.25	86.73	16.57	4.70
		Promedio ponderado	91.33	84.89	8.30	14.74
	NB	OM	92.55	69.93	2.52	37.56
		OS	91.17	58.31	5.13	40.47(*)
		OMO	87.62	81.83	8.06	20.30
		ONM	84.21	81.47	17.18	13.75
		Promedio ponderado	87.31	77.54	10.66	22.20
N1+N2	SVM	OM	95.28	81.63	**1.56**	24.03
		OS	93.97	67.57	2.14	39.19
		OMO	93.86	**91.01**	3.49	**11.04**
		ONM	89.58	87.95	13.66	5.64
		Promedio ponderado	92.34	85.99	7.18	13.64
	NB	OM	91.51	65.78	3.16	41.01
		OS	88.17	50.08	8.16	42.90
		OMO	83.27	77.20	15.57	18.88
		ONM	82.03	76.83	12.40	26.17
		Promedio ponderado	84.44	72.60	11.77	27.46
SW	SVM	OM	**95.71**	**83.82**	1.69	**20.08**
		OS	94.39	**71.10**	2.34	**33.34**
		OMO	93.22	89.98	**3.13**	13.47
		ONM	**89.99**	**88.36**	**13.00**	5.54
		Promedio ponderado	**92.39**	**86.48**	**6.83**	**13.28**
	NB	OM	93.53	77.47	4.23	20.21(*)
		OS	91.66	61.30	5.02	36.44
		OMO	88.79	83.29	6.41	20.05
		ONM	87.84	85.18	11.41	13.22
		Promedio ponderado	89.37	80.93	7.98	19.04

(*)No estadísticamente significativo (p>0.05)

Fuente: Elaboración propia.

Para la implementación de métodos de clasificación con un enfoque jerárquico, sólo se utilizaron características extraídas mediante el algoritmo de SW, en consideración de los buenos resultados que se indicaron anteriormente. En el caso de los clasificadores de múltiples etiquetas, se combinaron las características de ambos problemas de clasificación y se realizó una transformación binaria de las clases, según la presencia o ausencia de cada etiqueta asociada al registro médico. Se implementó el método de transformación binaria, el método de transformación multiclase y el algoritmo de clasificación jerárquica propuesto en la metodología del presente trabajo (Fig. 3.3), cuyos resultados se muestran en las Tablas 4.7, 4.8 y 4.9, respectivamente.

La Tabla 4.7 muestra los resultados de la clasificación jerárquica con el método de transformación binaria. En términos generales, con SVM se obtiene un mejor desempeño que con NB, excepto la FNR en las clases OM y OMO, pero no es estadísticamente significativo ($p>0.05$). Por otro lado, se obtuvo una pérdida 0/1 (PCU) de 16.31% con SVM y un 32.84% con NB.

TABLA 4.7 Desempeño de los clasificadores utilizando el método de transformación binaria.

Problema	Clasificador	Clase/Promedio	ACC (%)	F1 (%)	FPR (%)	FNR (%)
Primer (obesidad/¬obesidad)	SVM	NO	**97.20**	**93.87**	**1.33**	**7.63**
		O	**97.21**	**98.19**	**7.61**	**1.32**
		Promedio ponderado	**97.21**	**97.18**	**6.14**	**2.79**
	NB	NO	91.21	82.00	7.29	13.73
		O	91.21	94.16	13.73	7.29
		Promedio ponderado	91.21	91.32	12.23	8.79
Segundo (tipos de obesidad)	SVM	OM	**97.08**	**85.30**	**0.72**	20.98
		OS	**96.16**	**72.58**	**1.02**	**35.96**
		OMO	**94.52**	**89.36**	**2.37**	13.93
		ONM	**92.02**	**87.30**	**6.51**	**11.25**
		Promedio ponderado	**94.04**	**86.20**	**3.67**	**16.15**
	NB	OM	92.77	70.07	5.55	**20.86**$^{(*)}$
		OS	92.36	55.35	4.77	40.59
		OMO	88.87	80.61	10.18	**13.73**$^{(*)}$
		ONM	88.43	81.77	9.57	15.97
		Promedio ponderado	89.61	76.94	8.71	18.46

$^{(*)}$No estadísticamente significativo ($p>0.05$)

Fuente: Elaboración propia.

La Tabla 4.8 muestra los resultados de la clasificación jerárquica con el método de transformación multiclase. SVM obtiene un mejor desempeño que NB en todos los casos. Con este método, las PCU con este método disminuyeron con respecto al método de transformación binaria, obteniéndose para SVM un 10.77% y para NB un 20.44%.

TABLA 4.8 Desempeño de los clasificadores utilizando el método de transformación multiclase.

Problema	Clasificador	Clase/Promedio	ACC (%)	F1 (%)	FPR (%)	FNR (%)
Primer (obesidad/¬obesidad)	SVM	NO	**97.37**	**94.36**	**1.76**	**5.47**
		O	**97.37**	**98.28**	**5.47**	**1.76**
		Promedio ponderado	**97.37**	**97.37**	**4.60**	**2.63**
	NB	NO	92.74	84.43	4.77	15.43
		O	92.74	95.25	15.43	4.77
		Promedio ponderado	92.74	92.73	12.94	7.26
Segundo (tipos de obesidad)	SVM	OM	**97.41**	**87.22**	**0.75**	**17.69**
		OS	**96.12**	**74.52**	**1.69**	**28.75**
		OMO	**94.79**	**90.02**	**2.59**	**12.31**
		ONM	**92.77**	**88.91**	**7.68**	**6.26**
		Promedio ponderado	**94.48**	**87.55**	**4.29**	**12.35**
	NB	OM	93.66	69.31	3.11	32.94
		OS	93.36	58.42	3.63	41.37
		OMO	91.06	82.87	5.12	19.32
		ONM	88.30	81.73	10.03	15.36
		Promedio ponderado	90.55	77.94	6.66	21.95

Fuente: Elaboración propia.

Los resultados del algoritmo de clasificación jerárquica propuesto (ver Fig. 3.3) se muestran en la Tabla 4.9. Este algoritmo sólo afecta el desempeño de SVM y de NB en el segundo problema de clasificación. Se observa que con este método, SVM obtiene un mejor desempeño que el indicado en la Tabla 4.6. Por otro lado, el algoritmo jerárquico propuesto permitió mejorar el desempeño de NB en términos de ACC y el promedio ponderado de todas las métricas, excepto la FPR. Finalmente, se observa de la Tabla 4.9 que SVM obtiene mejores desempeños que NB.

TABLA 4.9 Desempeño de los clasificadores con el algoritmo jerárquico propuesto.

Problema	Clasificador	Clase/Promedio	ACC (%)	F1 (%)	FPR (%)	FNR (%)
Primer (obesidad/¬obesidad)	SVM	NO	**97.10**	**93.64**	**1.32**	**8.09**
		O	**97.10**	**98.12**	**8.09**	**1.32**
		Promedio ponderado	**97.10**	**97.07**	**6.51**	**2.90**
	NB	NO	91.23	82.25	7.63	12.56
		O	91.23	94.15	12.56	7.63
		Promedio ponderado	91.23	91.37	11.41	8.78
Segundo (tipos de obesidad)	SVM	OM	**99.62**	**95.54**	**0.34**	**1.37**
		OS	**99.46**	**89.09**	**0.34**	**8.11**
		OMO	**99.05**	**96.18**	**0.46**	**4.43**
		ONM	**98.46**	**99.04**	**3.40**	**1.10**
		Promedio ponderado	**98.93**	**96.50**	**1.62**	**3.04**
	NB	OM	97.74	77.96	2.24	2.90
		OS	97.42	59.78	2.17	19.19
		OMO	94.61	78.17	2.74	23.67
		ONM	90.66	94.10	16.31	7.68
		Promedio ponderado	93.75	82.64	8.09	13.82

Fuente: Elaboración propia.

4.3.1 Resumen de los resultados de la clasificación

En términos generales, el algoritmo de SW permitió obtener los mejores desempeños en ambos clasificadores. En los casos particulares donde SW tuvo un rendimiento inferior, la diferencia con respecto al mejor desempeño fue leve. Sin embargo, en todos los casos, el promedio ponderado de las métricas fue superior con el uso de SW.

La Tabla 4.10 muestra un resumen del desempeño de los clasificadores en el primer problema de clasificación con el uso de SW. En todas las métricas, SVM obtiene un mejor desempeño que NB.

TABLA 4.10 Resumen del desempeño de los clasificadores en el primer problema de clasificación con el enfoque no jerárquico.

Característica	Clasificador	Clase/Promedio	ACC (%)	F1 (%)	FPR (%)	FNR (%)
SW	SVM	NO	**97.10**	**93.64**	**1.32**	**8.09**
		O	**97.10**	**98.12**	**8.09**	**1.32**
		Promedio ponderado	**97.10**	**97.07**	**6.51**	**2.90**
	NB	NO	91.23	82.25	7.63	12.56
		O	91.23	94.15	12.56	7.63
		Promedio ponderado	91.23	91.37	11.41	8.78

Fuente: Elaboración propia.

La Tabla 4.11 muestra un resumen del desempeño de los clasificadores en el segundo problema de clasificación con el uso de SW. SVM obtiene en todas las métricas, excepto en la FPR de la clase OMN, un mejor desempeño que NB.

TABLA 4.11 Resumen del desempeño de los clasificadores en el segundo problema de clasificación con el enfoque no jerárquico.

Característica	Clasificador	Clase/Promedio	ACC (%)	F1 (%)	FPR (%)	FNR (%)
SW	SVM	OM	**95.71**	**83.82**	**1.69**	**20.08**
		OS	**94.39**	**71.10**	**2.34**	**33.34**
		OMO	**93.22**	**89.98**	**3.13**	**13.47**
		ONM	**89.99**	**88.36**	13.00	**5.54**
		Promedio ponderado	**92.39**	**86.48**	**6.83**	**13.28**
	NB	OM	93.53	77.47	4.23	20.21(*)
		OS	91.66	61.30	5.02	36.44
		OMO	88.79	83.29	6.41	20.05
		ONM	87.84	85.18	**11.41**	13.22
		Promedio ponderado	89.37	80.93	7.98	19.04

(*)No estadísticamente significativo (p>0.05)

Fuente: Elaboración propia.

En términos generales, el método de transformación multiclase permitió obtener mejores desempeños en los clasificadores, disminuyendo la PCU considerablemente respecto al método de transformación binaria. En cuanto al método de clasificación jerárquico propuesto, éste obtuvo un mejor desempeño global en el segundo problema de clasificación que el método de transformación multiclase.

Capítulo 5. Conclusiones

5.1 Sumario

En este trabajo se ha desarrollado un método de identificación automática de la obesidad y sus tipos en registros médicos electrónicos de-identificados. Se utilizaron dos enfoques en el problema de clasificación: tratar cada problema de clasificación de forma independiente y considerando una relación jerárquica entre la clase obesidad y sus tipos.

Se estudiaron las principales comorbilidades asociadas a la obesidad y el uso del IMC como indicador antropométrico del estado nutricional de los pacientes.

Para identificar la obesidad y sus tipos, se utilizaron dos algoritmos de clasificación supervisada: SVM y NB. En cada caso, se extrajeron características de los textos de los registros médicos electrónicos utilizando n-gramas (unigramas, bigramas y una combinación de ambos) y mediante el algoritmo de SW. Los resultados indican que el uso de características extraídas mediante el algoritmo de SW permitió mejorar el desempeño de los clasificadores implementados en la identificación de obesidad y sus tipos.

5.2 Conclusiones y discusiones

En el análisis exploratorio de los datos no se encontraron términos explícitos para negar la obesidad en los pacientes, por lo que se tuvo que utilizar terminología asociada a las diferentes categorías del IMC, tales como bajo peso, peso normal y sobrepeso, para considerar contraejemplos a la obesidad. Sin embargo, sólo un 3.94% de los 66179 registros médicos electrónicos de-identificados fueron relevantes para este estudio.

De acuerdo a los registros médicos electrónicos obtenidos tras el proceso de anotación, las mujeres tienen una mayor prevalencia a la obesidad con un 83.82% del total de los casos. Este porcentaje refleja la tendencia nacional de la prevalencia de obesidad en la población adulta, en la cual el 30.7% de las mujeres tiene esta enfermedad, mientras que los hombres un 19.2% [3]. Una causa que podría explicar la prevalencia de obesidad, reportada en este trabajo, podría ser la mayor

tendencia que las mujeres tienen, por sobre los hombres, en visitar centros de salud para tratar sus enfermedades.

Se emplearon dos enfoques en el proceso de clasificación. El primer enfoque consistió en tratar los dos problemas de clasificación de forma independiente, mientras que en el segundo enfoque se implementaron métodos de clasificación de múltiples etiquetas y un algoritmo que trató ambos problemas de clasificación de forma jerárquica, considerando la clase obesidad como el nodo padre de la jerarquía. Estos métodos son un enfoque más real si se desea implementar un método de identificación de obesidad y sus tipos de forma integrada en un sistema de registros médicos electrónicos de un hospital, permitiéndole al personal clínico obtener información de forma automática de las principales causas de la obesidad.

En ambos enfoques de clasificación se observaron altos porcentajes de precisión predictiva. Esto se debe a la gran cantidad de verdaderos negativos presentes en ambos problemas de clasificación, producto del desbalance de clases observado (ver Fig. 4.1).

Las comorbilidades mencionadas en este trabajo no son exclusivas de la obesidad. Esto pudo haber afectado el aprendizaje de los clasificadores supervisados SVM y NB al utilizar características extraídas de textos que hacían mención a las comorbilidades de la obesidad y el IMC. En el primer problema de clasificación, existen altos valores en la tasa de falsos positivos para la clase obesidad, lo que pudo afectar la clasificación de no obesidad, la cual presentó altos valores en la tasa de verdaderos negativos. En el segundo problema de clasificación, la clase obesidad no mencionada presenta altos valores en la tasa de falsos positivos, mientras que la clase obesidad severa presenta un alto valor en la tasa de verdaderos negativos. Mediante el uso del algoritmo de SW, se pudo disminuir los valores en la tasa de falsos negativos en ambos clasificadores.

Se observaron ambigüedades en el uso de la clase superobesidad en los registros médicos electrónicos en pacientes con obesidad mórbida y superobesidad. Por este motivo, se decidió combinar ambas clases para favorecer el desempeño de los clasificadores.

El desempeño de los clasificadores depende de las características seleccionadas para representar la información. La selección de características, con el método de IG, mejoró el desempeño de los clasificadores en los experimentos de ajuste realizados (ver Tabla 4.4).

En general, SVM obtuvo un mejor desempeño que NB en ambos problemas de clasificación, siendo favorecido por el uso del algoritmo de SW, el cual permitió representar de mejor forma las características presentes en los textos que a través del uso de n-gramas. En ambos enfoques, SVM

obtuvo los mejores promedios ponderados de cada métrica de desempeño utilizando características extraídas mediante el algoritmo de SW.

Si bien el algoritmo de SW mejoró el rendimiento de los clasificadores implementados, tiene un costo computacional mayor al uso de n-gramas en la extracción de características (ver Tabla D.1 del Anexo).

5.3 Trabajo futuro

Como trabajo futuro se pretende implementar un clasificador de textos basado en expresiones regulares, utilizando como método de extracción de características el algoritmo de SW, gracias a los buenos resultados obtenidos en este trabajo. Sin embargo, una tarea pendiente en el algoritmo de SW será reducir su tiempo de ejecución (ver Tabla D.1 del Anexo).

En cuanto a la clasificación jerárquica, el algoritmo propuesto tuvo un mejor desempeño en el segundo problema de clasificación. Por su parte, el algoritmo de transformación multiclase, obtuvo un mejor rendimiento en el primer problema de clasificación, pero con una PCU sobre el 10%. Al combinar ambos enfoques se podría implementar un método de clasificación jerárquico que mejore el rendimiento de ambos problemas de clasificación.

Capítulo 6. Publicaciones del trabajo de tesis

Producto del desarrollo del trabajo de tesis, se generaron las siguientes publicaciones:

Conferencia

- R. Figueroa, **C. Flores**, R. Cid, "Extracción de Información en Registros Médicos Electrónicos para el Estudio de la Obesidad", en *7th Biomedical Engineering Conference*, Concepción, Chile, 2014.

- R. Figueroa, **C. Flores**, "Obesidad en Chile: Un caso de estudio en registros médicos electrónicos", en Primer Simposio de Informática Médica en Salud (ISChile), Santiago, Chile, 2015 (**premiado como el mejor trabajo científico del evento por la Asociación Chilena de Informática en Salud ACHISA**).

- R. Figueroa, **C. Flores**, "Extracting Information from Electronic Medical Records to Identify Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures", en Ambient Intelligence for Health (AmIHealth), Puerto Varas, Chile, 2015, pp. 37-46.

- R. Figueroa, **C. Flores**, "Using the Smith-Waterman Algorithm to Extract Features in the Classification of Obesity Status", en 19th International Conference on Health Informatics and Technology, Rio de Janeiro, Brasil, 2017

Revista (ISI)

- R. Figueroa, **C. Flores**, "Extracting Information from Electronic Medical Records to Identify Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures", *Journal of Medical Systems*, vol. 40, no. 8, pp. 1-9, 2016

Bibliografía

[1] T. Abdalla, D. S. Abdulateef, F. Rahim *et al.*, "Prevalence of Overweight and Obesity among First Year Primary School Children in Sulaymaniyah City/Iraq", *Global Journal for Reseearch Analysis,* vol. 4, no. 6, pp. 92-95, 2015.

[2] A. Chuku, U. Onyeonoro, A. Ukegbu *et al.*, "Body Mass Index, Prevalence and Predictors of Obesity in Urban and Rural Communities in Abia State South Eastern Nigeria", *Journal of Diabetes & Metabolism,* vol. 6, no. 570, 2015.

[3] E. Atalah, "Epidemiología De La Obesidad En Chile", *Rev. Med. Clin. Condes.,* vol. 23, no. 2, pp. 117-123, 2012.

[4] V. Álvarez, "Obesidad: La Epidemia Del Siglo Xxi", *Rev. Med. Clin. Condes.,* vol. 14, no. 3, 2003.

[5] P. Soca, y A. Peña, "Consecuencias De La Obesidad", *Centro Nacional de Información de Ciencias Médicas,* vol. 20, no. 4, 2009.

[6] S. Markowitz, M. Friedman, y S. Arent, "Understanding the Relation between Obesity and Depression: Causal Mechanisms and Implications for Treatment", *Clin. Pssychol. Sci. Pract.,* vol. 15, no. 1, pp. 1-20, 2008.

[7] M. Moreno, "Definición Y Clasificación De La Obesidad", *Rev. Med. Cli. Condes.,* vol. 23, no. 2, pp. 124-128, 2012.

[8] A. M. Cohen, y W. R. Hersh, "A Survey of Current Work in Biomedical Text Mining", *Briefing in bioinformatics,* vol. 6, no. 1, pp. 57-71, 2004.

[9] Ó. Uzuner, "Recognizing Obesity and Comorbidities in Sparse Data", *J. Am. Med. Inform. Assoc.,* vol. 16, no. 4, pp. 561-570, 2009.

[10] I. Solt, D. Tikk, V. Gál *et al.*, "Semantic Classification of Diseases in Discharge Summaries Using a Context-Aware Rule-Based Classifier", *J. Am. Med. Inform. Assoc. ,* vol. 16, no. 4, pp. 580–584, 2009.

[11] H. Yang, I. Spasic, J. K. JA *et al.*, "A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries.", *J. Am. Med. Inform. Assoc. ,* vol. 16, no. 4, pp. 596-600, 2009.

[12] I. A. Braga, M. C. Monard, y E. T. Matsubara, "Combining Unigrams and Bigrams in Semi-Supervised Text Classification", en 14th Portuguese Conference on Artificial Intelligence Aveiro, Portugal, 2009, pp. 489-500.

[13] D. D. A. Bui, y Q. Zeng-Treitler, "Learning Regular Expressions for Clinical Text Classification", *J. Am. Med. Inform. Assoc.,* vol. 21, no. 5, pp. 850-857, 2015.

[14] M. Murtaugh, B. Gibson, D. Redd *et al.*, "Regular Expression-Based Learning to Extract Bodyweight Values from Clinical Notes", *J. Biomed. Inform.,* vol. 54, pp. 186-190, 2015.

[15] M. Tejero, "Genética De La Obesidad", *Boletín médico del Hospital Infantil de México,* vol. 65, no. 6, 2008.

[16] M. Alanis, W. Goodnight, E. Hill *et al.*, "Maternal Super-Obesity (Body Mass Index ≥50) and Adverse Pregnancy Outcomes", *Obstetric Anesthesia Digest,* vol. 89, no. 7, pp. 924-930, 2011.

[17] A. Arteaga, "El Sobrepeso Y La Obesidad Como Un Problema De Salud", *Rev. Med. Clin. Condes.,* vol. 23, no. 2, pp. 145-153, 2012.

[18] A. Baltasar, *Obesidad Y Cirugía: Cómo Dejar De Ser Obeso,* segunda ed., Madrid, España: Arán Ediciones, 2001.

[19] L. Ornella, "Códigos Correctores De Error En Problemas De Clasificación Multiclase De Datos De Marcadores Moleculares", Tesis de Doctorado, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, Argentina, 2010.

[20] E. M. Segovia, "Pump Operating Points Classification with the Help of Support Vector Machines", Tesis de Magíster, Departamento de Mecánica de Fluidos y Turbomaquinaria, technische universität kaiserslautern, Alemania, 2013.

[21] H.-T. Lin, y C.-J. Lin, "A Study on Sigmoid Kernels for Svm and the Training of Non-Psd Kernels by Smo-Type Methods", *Neural Computation*, pp. 1-32, 2003.

[22] T. Mitchell, *Machine Learning*: McGraw-Hill Science, 1997.

[23] V. Metsis, I. Androutsopoulos, y G. Paliouras, "Spam Filtering with Naive Bayes – Which Naive Bayes? ", en Third Conference on Email and Anti-Spam, California, Estados Unidos, 2006.

[24] M. A. Amrita, "Performance Analysis of Different Feature Selection Methods in Intrusion Detection", *International Journal of Scientific & Technology Research*, vol. 2, no. 6, pp. 225-231, 2013.

[25] C. Lee, y G. Geunbae, "Information Gain and Divergence-Based Feature Selection for Machine Learning-Based Text Categorization ", *Information Processing and Management*, vol. 42, pp. 155-165, 2006.

[26] A. Krause, y C. Guestrin, "Near-Optimal Nonmyopic Value of Information in Graphical Models", en Twenty-First Conference on Uncertainty in Artificial Intelligence, Edinburgh, Scotland, 2005.

[27] C. Pérez, "Evaluación De Reglas De Asociación En Text Mining Utilizando Métricas Semánticas Y Estructurales ", Tesis de Magíster, Departamento de Informática y Ciencias de la Computación, Universidad de Concepción, 2010.

[28] S. Dumais, y H. Chen, "Hierarchical Classification of Web Content", en 23rd annual international ACM SIGIR conference on Research and development in information retrieval, New York, Estados Unidos, 2000, pp. 256-263

[29] M. S. Sorower, *A Literature Survey on Algorithms for Multi-Label Learning*, Oregon State University, Corvallis, 2010.

[30] C. Silla, y A. Freitas, "A Survey of Hierarchical Classification across Different Application Domains", *Data Mining and Knowledge Discovery*, vol. 22, pp. 31-72, 2011.

[31] A. P. Santos, y F. Rodrigues, "Multi-Label Hierarchical Text Classification Using the Acm Taxonomy ", en 14th Portuguese Conference on Artificial Intelligence, Aveiro, Portugal, 2009, pp. 553-564.

[32] A. Carvalho, y A. Freitas, "A Tutorial on Multi-Label Classification Techniques".

[33] M.-L. Zhang, y Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms", *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819-1837, 2014.

[34] J. Camargo, J. Camargo, y L. Aguilar, "Conociendo Big Data", *Revista Facultad de Ingeniería*, vol. 24, no. 38, pp. 63-67, 2015.

[35] T. Joachims, *Learning to Classify Text Using Support Vector Machines*, New York: Engineering and Computer Science, 2002.

[36] C. Ramasubramanian, y R. Ramya, "Effective Pre-Processing Activities in Text Mining Using Improved Porter's Stemming Algorithm", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 12, pp. 4536-4538, 2013.

[37] K. H. Ambert, y A. M. Cohen, "A System for 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.

[38] W. Zhanga, T. Yoshidab, y X. Tangc, "A Comparative Study of Tf*Idf, Lsi and Multi-Words for Text Classification", *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[39] F. d. Borja, "Metodología, Construcción Y Explotación De Corpus Anotados Semántica Y Anafóricamente", Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, España, 2007.

[40] A. Viera, y J. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic.", *Fam. Med.*, vol. 37, no. 5, pp. 360-363, 2005.

[41] G. Forman, y M. Scholz, "Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement", *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 49-57, 2010.

[42] L. K. McKnight, A. Wilcox, y G. Hripcsak, "The Effect of Sample Size and Outcome Prevalence on Supervised Machine Learning of Narrative Data", en Proceedings of the AMIA Symposium. American Medical Informatics Association, 2002, pp. 519-522.

[43] M. Maragoudakis, K. Kermanidis, A. Garbis *et al.*, "Dealing with Imbalanced Data Using Bayesian Techniques ", en 5th international conference on language resources and evaluation, Genoa, Italia, 2006, pp. 1045-1050.

[44] X. Liu, y Z. Zhou, "The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study", en Sixth International Conference on Data Mining Hong Kong, China, 2006, pp. 970-974.

[45] X.-Y. Liu, J. Wu, y Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning", *IEEE transactions on systems, man, and cybernetics*, vol. 39, no. 2, pp. 539-550, 2009.

[46] X. Guo, Y. Yin, C. Dong *et al.*, "On the Class Imbalance Problem", en Fourth International Conference on Natural Computation IEEE, 2008, pp. 192-201.

[47] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1-47, 2002.

[48] U. Ergüin, "The Classification of Obesity Disease in Logistic Regression and Neural Network Methods", *Journal of Medical Systems*, vol. 33, no. 1, pp. 67-72, 2009.

[49] G. Wood, X. Chu, C. Manney *et al.*, "An Electronic Health Record-Enabled Obesity Database.", *BMC Med. Inform. Decis. Mak.*, vol. 12, no. 1, pp. 1-8, 2012.

[50] C. Ayash, S. Simon, R. Marshall *et al.*, "Evaluating the Impact of Point- of-Care Decision Support Tools in Improving Diagnosis of Obese Children in Primary Care.", *Obesity (Silver Spring)*, vol. 21, no. 3, pp. 576-582, 2013.

[51] A. Smith, Á. Skow, J. Bodurtha *et al.*, "Health Information Technology in Screening and Treatment of Child Obesity: A Systematic Review.", *Pediatrics*, vol. 131, no. 3, pp. e894-902, 2013.

[52] J. Cochran, y A. Baus, "Developing Interventions for Overweight and Obese Children Using Electronic Health Records Data.", *J . Nurs. Inform.*, vol. 19, no. 1, pp. 1-9, 2015.

[53] S. Heydari, S. Avatollahi, y N. Zare, "Comparison of Artificial Neural Networks with Logistic Regression for Detection of Obesity.", *Journal of Medical Systems*, vol. 36, no. 4, pp. 2449-2454, 2012.

[54] M. Kuebler, E. Yom-Tov, D. Pelleg *et al.*, "When Overweight Is the Normal Weight: An Examination of Obesity Using a Social Media Internet Database", *PloS One*, vol. 8, no. 9, pp. 1-8, 2013.

[55] R. Bordowitz, K. Morland, y D. Reich, "The Use of an Electronic Medical Record to Improve Documentation and Treatment of Obesity", *Fam. Med.*, vol. 39, no. 4, pp. 274-279, 2007.

[56] D. Jurafsky, *An Introduction to Natural Language Processing, Linguistic and Speech Recognition*, 2 ed.: Prentice Hall, 1999.

[57] J. Font, "Generación De Sistemas Basados En Reglas Mediante Programación Genética", Tesis de Magíster, Facultad de informática, Universidad Politécnica de Madrid, Madrid, España, 2008.

[58] A. Carrión, "Análisis Y Caracterización De Trabajos Blast Para La Planificación Eficiente En Entornos Grid Y Supercomputación", Tesis de Magíster, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, España, 2010.

[59] T. F. Smith, y M. S. Waterman, "Identification of Common Molecular Subsequences", *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195-197, 1981.

[60] B. G. Gebre, M. Zampieri, P. Wittenburg *et al.*, "Improving Native Language Identification with Tf-Idf Weighting", en Eighth Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, Georgia, 2013, pp. 216-223.

[61] L. Buitinck, G. Louppe, l. Mathieu Blonde *et al.*, "Api Design for Machine Learning Software: Experiences from the Scikit-Learn Project", en European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases, Praga, República Checa, 2013.

[62] A. M. Kibriya, E. Frank, B. Pfahringer *et al.*, "Multinomial Naive Bayes for Text Categorization Revisited", en Australasian Joint Conference on Artificial Intelligence, Berlin, Alemania, 2004, pp. 488-499

[63] G. Vanwinckelen, y H. Blockeel, "On Estimating Model Accuracy with Repeated Cross-Validation.", en BeneLearn and PMLS 2012, Bélgica, 2012.

[64] I. H. Witten, E. Frank, y M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, tercera ed.: Series in Data Management Systems, 2011.

[65] K. Dembczynski, W. Waegeman, W. Cheng *et al.*, "On Label Dependence in Multi-Label Classication", *Machine Learning*, vol. 88, no. 1, 2012.

[66] W. Zhang, T. Yoshida, X. Tang *et al.*, "Text Classification Based on Multi-Word with Support Vector Machine", *Knowledge-Bases Systems*, vol. 21, no. 8, pp. 879-886, 2008.

Anexo A. Documentos de aprobación para el uso de los registros médicos electrónicos del HGGB

A continuación se presentan los documentos que autorizan el uso de los registros médicos electrónicos de-identificados del HGGB de Concepción en marco del proyecto FONDECYT "*Adaptive Selection of training set based on Active Learning*", código 11121463.

DECLARACIÓN DE RESERVA Y CONFIDENCIALIDAD DE LA INFORMACIÓN.

El colaborador del proyecto que subscribe declara que es consciente de la importancia de sus responsabilidades en cuanto a no poner en peligro la integridad, disponibilidad y confidencialidad de la información que se maneja en el proyecto, especialmente aquella información proveniente de bases de datos del Hospital Clínico Regional Dr. Guillermo Grant Benavente sea está sensible o no. También declara que no guardará copias ni divulgará información sin previa autorización aún después de que finalicé su compromiso con el proyecto.

El colaborador del proyecto también declara haber tomado conocimiento y aceptado las normas señaladas en el documento de buenas prácticas clínicas.

NOMBRE
Alumno Memorista
Ingeniería Civil Biomédica
Depto. Ingeniería Eléctrica
Facultad Ingeniería
Universidad Concepción.

DECLARACIÓN DE CUMPLIMIENTO DE LAS

BUENAS PRÁCTICAS CLÍNICAS

El Investigador que suscribe declara que sus actuaciones están en pleno acuerdo con la Declaración de Helsinki (1964 y sus modificaciones de 1975, 1983, 1989, 1996, 2000 y 2004), con las normas de la "Buenas Prácticas Clínicas (GCP) establecidas por la Organización Mundial de la Salud (OMS, WHO, 1996), la ICH Harmonized Tripartite Guidelines for Good Clinical Practice (1996), por la Normas Éticas Internacionales para la investigaciones biomédicas con sujetos humanos (Organización Panamericana de la Salud y por el Consejo de Organizaciones Internacionales de las Ciencias Médicas (CIOMS, 1996), las Operacional Guidelines for Ethics Comités that Review Biomedical Research (WHO 2000) y por las Regulaciones Nacionales (LEY Nº 19.628 sobre “Protección de la vida privada o protección de datos de carácter personal”, Publicada en el Diario Oficial de 28 de agosto de 1999, Norma Técnica N° 57 del 04 de Junio del 2001, del Ministerio de Salud del Gobierno de Chile, Regulación de la Ejecución de Ensayos Clínicos que utilizan Productos Farmacéuticos en Seres Humanos; Ley Nº 20120 y su Reglamento sobre “La Investigación Científica en el Ser Humano, Su Genoma, Y Prohíbe La Clonación Humana” y la Ley 20584, de título “Regula los derechos y deberes que tienen las personas en relación con acciones vinculadas a su atención en salud”, promulgada el 13-04-2012 y en vigencia desde el 31.10.2012., párrafo Nº7).

Nombre completo
Cargo/ Profesión
Fecha

Anexo B. Herramienta de anotación

Para etiquetar manualmente los registros médicos electrónicos recuperados, se implementó una herramienta de anotación, la cual fue diseñada en *QT4 Designer*[18] y programada en *Python*, configurable para distintos problemas de clasificación (Fig. B.1). Este programa requiere que los registros médicos electrónicos estén encriptados para mantener la confidencialidad en el proceso de anotación.

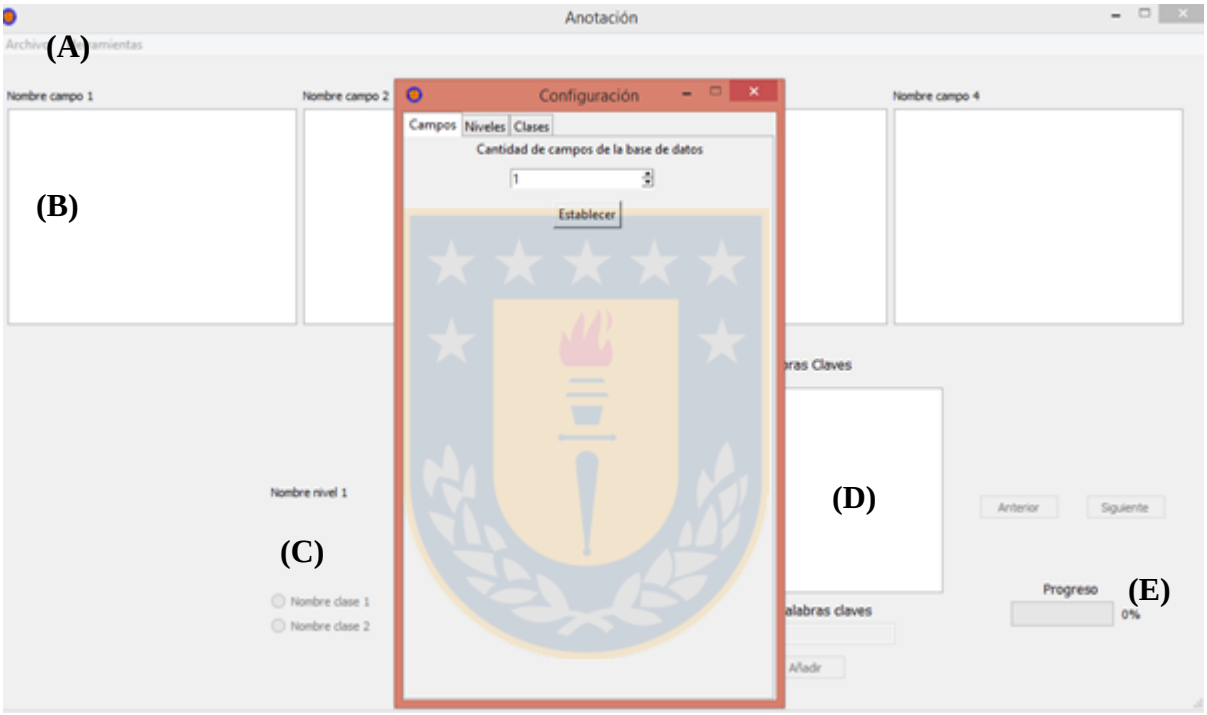


Fig. B.1 Herramienta de anotación. A: configuración. B: campos de los registros médicos electrónicos. C: clases de los problemas de clasificación. D: ingreso de palabras claves. E: progreso general del proceso de anotación

Fuente: Elaboración propia.

[18] *QT4 Designer* es una herramienta que permite el diseño de interfaces gráficas mediante *widgets*. Disponible en: http://doc.qt.io/qt-4.8/designer-manual.html. Fecha de ultimo acceso: Agosto de 2016

Anexo C. Especialidades médicas reportadas en los registros médicos electrónicos

La Tabla C.1 muestra las 46 especialidades médicas de la base de datos utilizada de los registros médicos electrónicos del Hospital Guillermo Grant Benavente de Concepción. Se destacan las **especialidades médicas** de los registros médicos recuperados para fines de clasificación.

TABLA C.1 Especialidades médicas reportadas en los registros médicos electrónicos del HGGB.

ALIVIO DEL DOLOR Y CUIDADOS PALIATIVOS	ENDOCRINOLOGÍA INFANTIL
ALTO RIESGO OBSTÉTRICO	**EXAMEN DE MEDICINA PREVENTIVA**
BRONCOPULMONAR ADULTO	**GASTROENTEROLOGÍA ADULTO**
BRONCOPULMONAR INFANTIL	GASTROENTEROLOGÍA INFANTIL
CARDIOCIRUGÍA ADULTO	GENÉTICA INFANTIL
CARDIOLOGÍA ADULTO	**GINECOLOGÍA**
CARDIOLOGÍA INFANTIL	**HEMATOLOGÍA ADULTO**
CIRUGÍA ABDOMINAL ADULTO	MEDICINA FÍSICA Y REHABILITACIÓN INFANTIL
CIRUGÍA ADULTO	**MEDICINA INTERNA**
CIRUGÍA DE MAMAS	**NEFROLOGÍA ADULTO**
CIRUGÍA INFANTIL	NEFROLOGÍA INFANTIL
CIRUGÍA PLÁSTICA ADULTO	NEONATOLOGÍA
CIRUGÍA PROCTOLÓGICA	**NEUROCIRUGÍA ADULTO**
CIRUGÍA TÓRAX ADULTO	NEUROLOGÍA ADULTO
CIRUGÍA VASCULAR PERIFÉRICA	NUTRICIÓN ADULTO
CONSULTA MÉDICA ABREVIADA	NUTRICIÓN INFANTIL
CONSULTA MÉDICA CARDIOL. ABREVIADA	**ONCOLOGÍA ADULTO**
CONSULTA MÉDICA ENDOC. ABREVIADA	**OTORRINOLARINGOLOGÍA**
CONSULTA MÉDICA REUMAT. ABREVIADA	PEDIATRÍA
CONSULTAS MÉDICAS	**REUMATOLOGÍA**
CONTROLES SEGÚN PROBLEMAS DE SALUD	SALUD OCUPACIONAL
DERMATOLOGÍA ADULTO	SUBESPECIALIDAD
ENDOCRINOLOGÍA ADULTO	**UROLOGÍA ADULTO**

Fuente: Elaboración propia.

Anexo D. Tiempo de ejecución de los principales procesamientos

La Tabla D.1 muestra los tiempos de ejecución de los principales procesamientos realizados para la clasificación de los registros médicos electrónicos. Se destacan los tiempos que superan los 60 minutos de duración.

TABLA D.1 Tiempo de ejecución de las principales etapas de procesamiento.

Procesamiento	Tiempo (minutos)
Preprocesamiento	4.81
Tokenización SW	**351.60**
Tokenización ngrams	11.60
Representación TF-IDF	3.77
Sintonización de parámetros	**328.81**
Selección de características	56.74
Clasificación SVM	52,38
Clasificación NB	6.60

Fuente: Elaboración propia.