Extracción de Información de Evoluciones Clínicas Digitales mediante técnicas de Machine Learning

Eckert Karina[1a], Montenegro Sergio Daniel[2b], López Forastier Nicolás[3c], Candia Gabriel José[1d]

[1]Departamento de Ingeniería y Ciencias de la Producción, Universidad Gastón Dachary
[2]Universidad Católica de las Misiones - Integrando Salud
[3]Servicio de Cirugía Cardiovascular, Hospital Escuela de Agudos Dr. Ramón Madariaga
Posadas, Misiones, Argentina

{[a]karinaeck, [b]drmontenegro, [c]bnforas5, [d]gabrieljcandia}@gmail.com

Abstract. Este trabajo demuestra el uso de un modelo de machine learning para extraer información referida a factores de riesgo cardiovascular de evoluciones clínicas desestructuradas redactadas en español. El mismo describe un procedimiento para el análisis de corpus y filtrado de evoluciones relevantes para entrenamiento y testeo del modelo. Los resultados muestran la efectividad de los recursos utilizados en extraer la información relevante y, a su vez, plantean una relación entre la complejidad de la información a extraer, y la cantidad de datos de ejemplo necesaria para alcanzar valores de performance elevados.

Keywords: Machine Learning, Extracción de Información, Factores de Riesgo Cardiovascular, IBM® Watson.

1 Introducción y problemática

Desde el surgimiento de la computación, se han ido desarrollando cada vez más y mejores sistemas informáticos con funcionalidades de procesamiento de texto [1]. Estas funcionalidades se han trasladado a distintas plataformas utilizadas como soporte a actividades profesionales específicas. Tal es el caso de la profesión médica, donde gran parte de la información manipulada está relacionada al estado y la evolución de factores clínicos de los pacientes, almacenados en registros denominados historias clínicas.

En la actualidad, es muy frecuente que las Instituciones de Salud usen Historias Clínicas Electrónicas como herramienta de registro médico. Sin embargo, mucha de la información médica registrada se encuentra en forma narrativa o de texto libre, imposibilitando en muchos casos su categorización y análisis para la toma de decisiones médicas. Este gran volumen de información no estructurada, sumada a la gran demanda de los servicios de salud, lleva a que los médicos pierdan tiempo leyendo largas evoluciones de texto libre, en lugar de tener una lista categorizada de problemas médicos relevantes, que permitirían además usarlo como input para herramientas de soporte para la toma de decisiones. Por tal motivo, es deseable que se pueda extraer y

analizar automáticamente información relevante para dicho profesional de la salud, tomando como base las evoluciones de un paciente determinado, reduciendo los tiempos destinados a la revisión de antecedentes.

Este trabajo presenta un estudio de la eficacia de un modelo de machine learning en extraer información de textos desestructurados, en forma de entidades y relaciones propias de un dominio médico definido por los factores de riesgo cardiovascular, que permiten, entre otras aplicaciones, identificar y predecir el riesgo que tiene un paciente de padecer una patología cardiovascular en los próximos años. El modelo es entrenado utilizando los recursos de IBM® Watson, en base a evoluciones clínicas digitalizadas, y luego se evalúan parámetros de performance del mismo. La principal contribución del trabajo es la aplicación de dicha herramienta al dominio mencionado, utilizando como fuente para el proceso, evoluciones en lenguaje español.

El trabajo se organiza de la siguiente manera: en la Sección 2 se introducen los procesos de extracción de información, la técnica de machine learning, las herramientas utilizadas y trabajos relacionados. La Sección 3 presenta el origen, calidad y cantidad de datos, para luego definir el dominio de estudio específico. La Sección 4 presenta un análisis de la estructura sintáctica en la que se encuentra la información relevante, y luego define un sistema de entidades y relaciones para su extracción. La Sección 5 presenta el procedimiento realizado para el entrenamiento del modelo de machine learning. La Sección 6 presenta las pruebas y resultados obtenidos. Finalmente, la Sección 7 presenta las conclusiones y trabajos futuros.

2 Conceptos preliminares

2.1 Extracción de información

La Extracción de Información (EI) refiere al uso de métodos computacionales que tienen por objetivo la identificación y recuperación de cierto tipo de información relevante, a partir de texto en lenguaje natural, mediante la ejecución de un procesamiento automático. Así, este proceso tiene por objetivo obtener ocurrencias de una (o varias) clases particulares de eventos, entidades, y las relaciones existentes entre dichas entidades [2, 3].

Los métodos de EI utilizan datos de entrada que consisten en una (o varias) colecciones de documentos denominados *corpus*, que pueden ser correos electrónicos, páginas web, artículos de noticias, documentos de investigación, entre otros tipos de colecciones. A partir de éstos se obtiene una representación de la información relevante contenida en dichas fuentes según ciertos criterios específicos. Estos criterios son dependientes del dominio, por lo que es necesario que un humano realice previamente la tarea de especificar los tipos de eventos, entidades o relaciones relevantes del mismo, de modo que la representación de la información obtenida refleje el contenido semántico de tales colecciones [3].

En este contexto, una entidad es la forma en la que puede categorizarse un objeto del mundo real. Así, una mención a una entidad es un ejemplo de “algo” de ese tipo, es decir, una ocurrencia en el texto que se asocia a dicha entidad. Por otro lado, una relación define una asociación binaria y ordenada entre dos entidades. En este caso,

para que exista una mención de relación, el texto debe definir explícitamente el enlace entre las dos entidades, y debe hacerlo dentro de una única frase [4].

La estructura de las piezas de información obtenidas mediante EI debe permitir que las mismas sean almacenadas en un medio informático que posibilite su eventual procesamiento y recuperación [3]. De esta manera, el proceso realizado sobre el texto en lenguaje natural debe ser capaz de ajustar los datos de salida a planillas o bases de datos con un formato bien definido, de modo que permita el posterior aprovechamiento de la información [5].

2.2 Machine Learning

Las técnicas de aprendizaje automático, o Machine Learning (ML), se definen como un conjunto de métodos que pueden detectar automáticamente patrones en los datos, y después utilizar esos patrones descubiertos para predecir datos futuros, o realizar otro tipo de toma de decisiones bajo incertidumbre [6]. También se define como el proceso (algoritmo) por el cual se estima un modelo que es compatible con un problema del mundo real, con cierto grado de probabilidad, obtenido a partir de un conjunto de datos (o muestra) generado mediante observaciones finitas en un ambiente ruidoso [7]. Existen diversos métodos de ML para el análisis de datos. A grandes rasgos, los mismos pueden clasificarse en dos grupos principales [6]:

a) Aprendizaje supervisado

También llamado predictivo, busca establecer un mapeo entre las entradas (o inputs) x_i, y las salidas (outputs) y_i, a partir de un conjunto de entradas-salidas dado, denominado conjunto de entrenamiento (training set). Dicho conjunto puede definirse como $D = \{(x_i, y_i)\}_{i=1}^{N}$, siendo N el número de ejemplos de entrenamiento. Por lo tanto, existe de manera predefinida el valor deseado de la salida y_i que se espera para cada valor de entrada de la forma x_i, donde este último puede definirse como un vector n-dimensional de valores denominados atributos, características o covariables. A su vez, esta categoría incluye los problemas de clasificación y regresión [6, 8].

b) Aprendizaje no supervisado

En este caso el conjunto de datos inicial es un grupo de entradas sobre las cuales se desea encontrar "patrones interesantes", por lo que los datos de entrenamiento en este caso se puede definir como $D = \{x_i\}_{i=1}^{N}$, sin contar con un conjunto de valores de salida deseados. Así, se trata de un problema menos definido donde no existe un conocimiento previo de los datos que indique qué tipo de patrones buscar [6, 8].

En este trabajo se utiliza el enfoque de aprendizaje supervisado, donde se cuenta con un conjunto de datos de entrada x_i, en formato de texto plano, para el cual diversos términos u oraciones deben ser manualmente asociados, mediante un proceso de etiquetado, a ciertas entidades y relaciones, conformando éstas al conjunto de datos de salida y_i. Luego, y a partir de este etiquetado, se realiza el entrenamiento de un modelo de ML que permita reconocer ocurrencias de tales entidades y relaciones tomando como fuente nuevos textos sin etiquetar.

2.3 IBM Watson

IBM® Watson es un sistema para el Procesamiento del Lenguaje Natural en profundidad (deep Natural Language Processing) que analiza y comprende las características del lenguaje humano [9]. El mismo provee servicios que buscan brindar a usuarios no técnicos la posibilidad de crear herramientas de etiquetado, o extracción de información, para cualquier tipo de datos en formato de texto [10]. En este sentido, Watson Knowledge Studio [11] es un entorno de trabajo disponible como servicio cloud SaaS (Software as a Service) desarrollado por IBM®, en el que los usuarios pueden cargar documentos de texto y etiquetarlos manualmente, permitiendo que luego sean utilizados para proveer extracción automática de información mediante el entrenamiento de un modelo de ML [10]. La extracción de términos y relaciones sobre conceptos específicos del dominio de aplicación se realiza utilizando Natural Language Understanding [12], otro servicio de IBM®, el cual explota el modelo de ML entrenado sobre textos proporcionados por el usuario u otro sistema informático.

2.4 Trabajos relacionados

Como antecedentes en la extracción de información médica de documentos desestructurados, se puede mencionar un trabajo realizado sobre el dominio definido por la enfermedad de las arterias coronarias [10]. También se ha estudiado la confección de resúmenes de historias clínicas con posibilidades de aplicación a la toma de decisiones médicas [13]. Otro trabajo se basa en la minería de datos para la detección temprana del riesgo cardiovascular a partir de campos estructurados y no estructurados. [14]. Hacia 2014 no había antecedentes de herramientas de Procesamiento del Lenguaje Natural para textos médicos escritos en español [15]. La relativa actualidad de los estudios mencionados da cuentas de que la explotación de datos médicos desestructurados se posiciona como una línea de investigación activa con gran relevancia para la industria de la salud [3].

3 Definición del dominio

Los datos consisten en 979 evoluciones referidas a la condición médica de distintos pacientes (de la provincia de Misiones), que fueron exportadas del sistema de historias clínicas electrónicas de Integrando Salud[1], en formato de planilla de cálculo. La misma cuenta con dos columnas: un identificador de la evolución (número secuencial único, iniciado en 1), y otra con el contenido de dicha evolución, en formato de texto plano. Estas evoluciones no tienen una relación a-priori entre sí, por lo que, para motivos de este estudio, fueron consideradas evoluciones independientes. Por este motivo, y con el fin de que la información extraída caracterice al paciente sobre quien trata la evolución, el marco sobre el cual se extrajo información consistió en evoluciones individuales donde, para cada evolución, se buscó definir la condición médica del paciente, para un dominio médico específico, según cómo dicha evolución haya descripto el estado del paciente en cuestión.

[1] Link al sitio oficial: https://www.integrandosalud.com/es-ar/

Dado que la cantidad de información presente en las evoluciones abarca un contenido muy diverso, comprendiendo aspectos tales como enfermedades y síntomas del paciente en el momento de la consulta, así como también apreciaciones de los expertos de la salud respecto de su estado previo y posibles diagnósticos, que a su vez pueden ser muy variados; es necesario en primer lugar especificar un dominio reducido comprendido por una cantidad limitada de tales características.

En este estudio, y con el objetivo de definir un dominio inicialmente acotado, se tuvieron en cuenta tres factores. Por un lado, abarcar ciertas condiciones que, en base a la evidencia existente en las evoluciones, permitan alcanzar una conclusión de mayor nivel de abstracción, tal como la posibilidad de inferir una enfermedad a partir de la identificación de una serie de síntomas. Además, dicha información debe ser mencionada en gran medida en las evoluciones, de modo que se cuente con ejemplos suficientemente frecuentes como para sustentar el entrenamiento y la posterior prueba de un modelo de ML. Finalmente, es deseable que el dominio abarcado sea relevante en el contexto médico actual, para el cual se espera que la aplicación del modelo entrenado pueda extraer información oportuna para el estudio de las enfermedades. En este sentido se busca satisfacer las necesidades de información típicas de los profesionales de la salud.

Teniendo en cuenta los factores mencionados, se comenzó realizando un primer análisis de las evoluciones, en el cual se identificó el tipo y la frecuencia de la información recopilada por los médicos en las mismas. Como resultado de este estudio, y debido a que contaba con los factores deseables, se definió como dominio médico al comprendido por los principales factores de riesgo cardiovascular. Este campo tiene una fundamental importancia en la determinación, entre otras, de las Enfermedades Crónicas No Transmisibles, las cuales representan más del 60% de las principales causas de muerte en Misiones, Argentina, así como en el resto del mundo [16]. Los factores de riesgo considerados en este estudio son: hipertensión arterial, diabetes, tabaquismo, colesterol y obesidad.

4 Análisis de datos y definición del Sistema de Tipos

A partir del primer análisis de las evoluciones que dio lugar a la definición del dominio de aplicación, fue necesario determinar la estructura y contexto semántico bajo el cual ocurrían las menciones a los factores de riesgo definidos. El conocimiento de dicha estructura fue requerido con el objetivo de establecer las entidades y relaciones que sirvieran de soporte al proceso de EI. Para ello, y para cada factor de riesgo, se realizó un estudio a mayor profundidad donde se definieron distintos términos de búsqueda factibles de retornar la información requerida. Estos términos fueron expresados en formato de expresiones regulares [17], de modo que permitiesen abarcar las posibles variaciones gramaticales de los términos de búsqueda utilizados, donde se consideró que cada expresión tiene asociada una o varias *entidades candidatas*. El empleo de estas expresiones en una búsqueda devolvió una cantidad determinada de ocurrencias, para las cuales se analizó y clasificó la información obtenida según su relación con el factor de riesgo en cuestión.

Luego, para cada expresión regular, se contabilizó la cantidad de ocurrencias obtenidas y su significado según el contexto, categorizándolas en positivas (ocurrencias

que coinciden con la existencia del factor de riesgo que se busca extraer), negativas (ocurrencias que refieren a dicho factor, pero como negación de la existencia del mismo), u otros (hace mención al factor, pero no indica si el paciente lo posee o no). También se contabilizaron los resultados donde la ocurrencia del término no se relacionaba con el factor de riesgo en cuestión.

Para ejemplificar la categorización se puede considerar el término de búsqueda "hiper*", relacionado al factor hipertensión arterial (HTA). En este caso, el asterisco es un comodín utilizado para indicar que, luego del término "hiper", puede haber una cantidad de caracteres arbitraria, y sin restricciones en cuanto al tipo de carácter. Entre ellos, el resultado "(…) fue visto por el cardiólogo el 29/03 quien le hizo ECG e informa paciente hipertenso (...)" es categorizado como Positivo, dado que indica que el paciente posee HTA. En cambio, el texto "(...) Si los registros son altos, vamos a empezar con medicación antihipertensiva (...)" se categoriza como Otro, dado que refiere al concepto HTA, pero no indica si el paciente lo tiene (o no). Finalmente, textos como "(…) ECG del 17/02/2010 que informa ritmo sinusal, regular, sin signos de hipertrofia (…)" y "(…) DX: - bocio multinodular hipercaptante (…)" son categorizados como No Relacionado, dado que no refieren al factor HTA. Para este término de búsqueda no se obtienen resultados negativos; sin embargo y para ejemplificar, se categorizaría como Negativo un texto como "paciente sin signos de hipertensión".

Dado que la categorización de la información obtenida como resultado de búsqueda de cada expresión regular requiere conocer el contexto en el que ocurren dichas menciones, se utilizó AntCoc [18] como software de soporte para el análisis de corpus. Dicho sistema permite utilizar expresiones regulares como término de búsqueda, retornando las ocurrencias en el corpus que coincidan con dicho término, además del contexto en el que se da cada una, consistiendo ello en una cierta cantidad de palabras antes y después de cada mención.

1	ni vomitos. Al ex. fisico presenta TA: 180/100. No refiere antec de HTA. Pulso: 84 x min. 2 R en 4 F NF, silencios impresionan libres. Sin
2	paciente de 73 años, con antec de HTA y aneurisma de aorta abdominal medicada con betabloqueantes y y enalapril.
3	ad fisica. -Medicación: no refiere. -Alergias: no refiere. -Antec fliares: HTA, DBT, cancer. Ex fis: Al ex. fisico TA: 150/90, FC: 116 x min.
4	Es muy nerviosa. nunca fue a ningún especialista. Tiene antecedentes de: -HTA: medicada con glioten 5 mg cada 12 hs. -Taural: 1 compr cada 12 hs -Clonaz
5	ro infiltración de la zona. Paciente de 54 años, con antecedentes de HTA, Psoriasis (+ de 20 años, 1978) y artritis psoriásica de larga data (+ de
6	itos marginales y signos leves de artrosis. Antecedentes: -Hipotiroidismo -HTA. -Alergia. Medicacion: - Paxon 50 -T4: 100 ug x dia. -Benadryl: 5 ml a ve
7	- L5. Actualmente sin dolor. Tomaba mecanyl Duo hasta hace 1 año. Refiere HTA. Medicada con atenolol 25 x dia. Vacunaciones: ATT hace 5 años Cirugias pr
8	sigmoides. TA: 180/100, FC: 80 x min, Peso: 80,5 kg, talla: 1,64, BMI: E: HTA mal controlada. no estuvo tomando su medicación hace 3 dias. P: Hago
9	mportancia refiere: -ACV izquierdo con hemiparesia braquiocrural derecha. -HTA- -ITU previa hace 5 años. No vacunado contra gripe ni neumonia. -Constipac
10	giere realizar estudios de mayor complejidad seguncuadro clinica. Antec: -HTA: medicada con atenolol 50 mg x. 2. Luego de ver al Dr.

Fig. 1. Primeros 10 resultados devueltos por AntCoc ante el término de búsqueda "HTA".

Para utilizar AntCoc es necesario exportar la información desde su formato original de planilla de cálculo, a uno (o varios) archivos de texto plano. Para este fin específico se desarrolló un script en el lenguaje de programación Python [19] que recorre cada una de las filas de la planilla, extrae la evolución asociada y, una vez completado el recorrido, exporta las evoluciones a archivos de texto plano de aproximadamente 1.000 palabras cada uno[2]. Estos archivos fueron luego incorporados a la herramienta para realizar el análisis del corpus. La **Fig. 1** presenta un ejemplo de la estructura de

[2] Definir un máximo de palabras por documento facilita ubicar el archivo de origen de cada ocurrencia para una búsqueda utilizando AntCoc.

los resultados devueltos por AntCoc. Como se observa, el término de búsqueda se encuentra en el centro de cada oración, y a los lados, el contexto de dicha ocurrencia.

Siguiendo el procedimiento para análisis de menciones especificado, y en base a un estudio inicial de las evoluciones, se definió una serie de expresiones regulares para cada factor de riesgo, y se categorizaron los resultados de búsqueda obtenidos utilizando AntCoc. La frecuencia para cada categoría indica la factibilidad de dichas expresiones regulares de retornar información relacionada a la categoría en cuestión y, consecuentemente, la representatividad que se obtendría de definir entidades relacionadas a las mismas. Así, esta métrica se utiliza para determinar si una categoría, asociada a cierta expresión regular, es válida como entidad del dominio. Cuanto mayor sea la frecuencia de alguna categoría en particular, mayor será la factibilidad de que dicha categoría sea aplicable en establecer una entidad que la relacione con el tipo de información referida en la expresión regular y, en consecuencia, con el factor de riesgo en cuestión. En este sentido, si bien se recomienda alcanzar 50 menciones de ejemplo para cada tipo de entidad y relación [20], en este trabajo se establece un mínimo de 8 menciones por categoría, dado que, para ciertos conceptos relevantes al dominio, la cantidad de menciones disponibles es reducida.

En la **Tabla 1** se detallan las expresiones regulares relacionadas al factor de riesgo hipertensión arterial, la frecuencia de las mismas para cada categoría, y su proporción en base al total de resultados por expresión regular. Como se muestra, para la expresión “HTA” se obtienen 74 ocurrencias positivas para las que cada evolución indica que el paciente en cuestión posee HTA; 1 ocurrencia indicando que el paciente no posee HTA, y 2 ocurrencias que refieren a HTA, sin indicar si el paciente la posee o no. Para esta expresión no hubo resultados no relacionados a HTA. En cuanto a la distribución de frecuencias para dicha expresión, se observa que el 95% de las mismas son positivas, lo cual implica que el término HTA está fuertemente relacionado con evoluciones donde el paciente posee HTA. Este análisis da pie a la definición de una entidad que extraiga menciones a HTA, cuya ocurrencia permite asumir que el paciente en cuestión posee dicho factor de riesgo.

Tabla 1. Categorización de resultados de búsqueda para expresiones regulares referidas a hipertensión arterial.

Hipertensión Arterial									
Expresión Regular	**Ocurrencias**					**Proporciones**			
	Relacionado			**No Rel.**	**Tot.**				
	Pos.	**Neg.**	**Otro**			**Pos./Tot.**	**Neg./Tot.**	**Otro/Tot.**	**NR/Tot.**
HTA	74	2	2	0	78	0,95	0,03	0,03	0,00
[^(a-zA-Z)]TA[?:]	159	0	0	0	159	1,00	0,00	0,00	0,00
arteria*	1	0	0	6	7	0,14	0,00	0,00	0,86
hiper*	3	0	1	40	44	0,07	0,00	0,02	0,91
presi?n*	5	0	2	26	33	0,15	0,00	0,06	0,79

De manera similar, para la segunda expresión regular se observa que el 100% de ocurrencias refieren al parámetro de búsqueda. En este caso, lo que se busca es el término “TA”, de tensión arterial (TA). Por lo tanto, la categoría “positiva” en este

caso indica que el texto retornado refiera al concepto “tensión arterial”. Así, la cantidad de resultados da pie a la definición de una entidad relacionada a TA. Además, dado que el término TA suele estar acompañado por los valores de tensión arterial diastólica (TAD) y tensión arterial sistólica (TAS)[3], relevantes para el caso de estudio, se puede definir una entidad para extraer dichos valores.

Para las demás expresiones regulares, por tratarse de términos muy genéricos donde, en el mejor de los casos, posee un 15% de ocurrencias positivas, pero sin alcanzar el mínimo de 8 menciones; y dado que la mayor proporción se encuentra en la categoría “no relacionado”, se considera no relevante la definición de entidades relacionadas a dichas expresiones, pues no retornaría información útil al dominio de estudio.

Según el análisis realizado para el factor de riesgo hipertensión arterial, podrían definirse 3 entidades que permitan capturar ocurrencias a HTA, TA y el valor asociado a la TA (TAD y TAS) en las evoluciones. Además, se puede definir una relación entre TA y su valor asociado de modo que el modelo de ML pueda ser entrenado para detectar estos patrones en nuevas evoluciones [21].

Tabla 2. Sistema establecido para tipos de entidades.

Tipo de Entidad	Descripción
BMI	Entidad extraída cuando se menciona al BMI (Body Mass Index, Índice de Masa Corporal). Dicho valor mide el contenido de grasa corporal en relación a la estatura y el peso. Permite establecer si el paciente posee obesidad. Ej.: “BMI: 32”.
BMI_VAL	Entidad extraída representando el valor del BMI, asociado a la ocurrencia de la entidad BMI. Ej.: para “BMI: 32”, “32” es el valor asociado a BMI_VAL.
COL	Entidad extraída cuando se menciona que el paciente posee colesterol. Ej.: “toma medicación para el colesterol”.
COL_T	Entidad extraída cuando se menciona colesterol total (“Col T”). Ej.: “lab del 23/10/2009 que informa: Col T: 155”.
COL_T_VAL	Entidad extraída representando el valor de colesterol total, asociado a la ocurrencia de la entidad COL_T. Ej.: en “Col T: 155”, “155” es el valor asociado a COL_T_VAL.
DBT_OTRO	Entidad extraída cuando la mención refiere a diabetes neutra (que no implica DBT positiva ni negativa). Se utiliza para distinguir esta categoría respecto de la entidad DBT_POS. Ej.: “Solicito lab de rutina para descartar DBT”.
DBT_POS	Entidad extraída cuando se menciona que el paciente posee diabetes. Ej.: “Como antecedentes de importancia presenta: -DBT II”.
DEJAR_DE_FUMAR	Entidad extraída cuando se menciona que el paciente expresa el deseo de dejar de fumar. Es un indicio de que fuma, o bien, de que es un ex fumador. Ej.: "Quiere dejar de fumar".
EX_FUMADOR	Entidad extraída cuando se menciona que el paciente es un ex fumador. Ej.: “Ex fumador: dejó hace más de 5 años”.
FUMA	Entidad extraída cuando se menciona que el paciente fuma. Ej.: “fuma”, “está fumando”.
GANAS_DE_FUMAR	Entidad extraída cuando se menciona que el paciente siente ganas de fumar. Es un indicio de que fuma, o bien, de que es un ex fumador. Ej.: “Me dice que siente ganas de fumar a la siesta”.

[3] Por ejemplo, “TA: 180/100”. En algunos casos se expresan varios parámetros de TAS y TAD, como ser: “TA: 120/80, 130/80, 110/80, 120/80”.

Tipo de Entidad	Descripción
HTA	Entidad extraída cuando se menciona que el paciente tiene hipertensión arterial. Ej.: "HTA desde hace 5 años medicada".
NO_FUMA	Entidad extraída cuando se menciona que el paciente no fuma. Ej.: "fuma: niega", "Fuma: no", "No fuma", "sin fumar", "nunca fumo", "Fuma: nunca"
PESO	Entidad extraída cuando se menciona el peso del paciente. Ej.: "peso: 75kg".
PESO_VAL	Entidad extraída representando el peso del paciente, asociado a la ocurrencia de la entidad PESO. Ej.: para "peso: 75kg", "75" es el valor asociado a PESO_VAL.
TA	Entidad extraída cuando se menciona el término TA (tensión arterial). Ej.: "Trae notas de TA de 140/80".
TA_VAL	Entidad extraída representando el valor de tensión arterial, asociado a la ocurrencia de la entidad TA. Ej.: para "TA: 120/80", el valor de TA_VAL es "120/80". De dicho valor se deduce que "120" corresponde a TAS, y "80" a TAD.
TALLA	Entidad extraída cuando se menciona la talla del paciente. Ej.: "talla: 1,69".
TALLA_VAL	Entidad extraída representando la talla del paciente, asociada a la ocurrencia de la entidad TALLA. Ej.: para "talla: 1,69", "1,69" es el valor asociado a TALLA_VAL.
TBQ	Entidad extraída cuando se mencionan antecedentes de tabaquismo. Ej.: "Ex TBQ", "Fuma: ex TBQ en la adolescencia".

Siguiendo el análisis llevado a cabo para la definición de entidades y relaciones asociadas a hipertensión arterial, se repitió el estudio para los demás factores de riesgo. A partir de los mismos se definió el sistema de tipos de entidades y relaciones. La **Tabla 2** presenta los tipos de entidades definidos, mientras que la **Tabla 3** presenta los tipos de relaciones. Ambas conforman el Sistema de Tipos (Type System). Cada entidad/relación busca extraer información específica, relevante al dominio de aplicación.

Tabla 3. Sistema establecido para tipos de relaciones.

Tipo de Relación	Primer Tipo de Entidad	Segundo Tipo de Entidad
BMI_VAL	BMI	BMI_VAL
COL_T_VAL	COL	COL_T_VAL
PESO_VAL	PESO	PESO_VAL
TA_VAL	TA	TA_VAL
TALLA_VAL	TALLA	TALLA_VAL

5 Entrenamiento del Modelo de Machine Learning

Watson Knowledge Studio se basa en el aprendizaje supervisado para entrenar un modelo de ML. Este entrenamiento se realiza siguiendo una serie de pasos definidos en el flujo de trabajo de la propia herramienta, según los presenta la **Fig. 2**. Las tareas principales son la definición de las entidades y relaciones del dominio, la carga de documentos para etiquetado, el etiquetado de dichos documentos, y el entrenamiento (y prueba) del modelo en base a los documentos etiquetados. Todos los pasos se realizan utilizando la interface web provista por el propio servicio.

La definición de entidades y relaciones se realizó siguiendo los tipos establecidos en la Sección 4. Los documentos para etiquetado fueron obtenidos utilizando un script desarrollado en Python que importa las evoluciones de la planilla de datos original y las filtra en un conjunto representativo, que considera las menciones asociadas a las

entidades y relaciones bajo estudio. El filtrado se realiza utilizando las expresiones regulares definidas en el análisis de menciones. En dicho proceso se busca un mínimo de deseable de 70 menciones por entidad, de las cuales 50 menciones (o el 70%, según la cantidad disponible) son utilizadas para entrenamiento, y las 20 menciones restantes son utilizadas para pruebas. Para ambos conjuntos, las evoluciones relacionadas a entrenamiento y pruebas son exportados a documentos de texto plano, cada uno conteniendo aproximadamente 1.000 palabras según lo recomendado para realizar el etiquetado [20]. A su vez, las proporciones recomendadas para los conjuntos de Entrenamiento/Pruebas/Blind son de 70/23/7 respectivamente [22]. En este caso no se utilizó el conjunto Blind[4] dado que se considera que la muestra de evoluciones es representativa para los tipos de entidades/relaciones definidos. Así, el porcentaje asociado al mismo pasa a formar parte del conjunto de pruebas. En total, 12 documentos sumando 16.754 palabras fueron asignados al conjunto de entrenamiento, y 6 documentos con 7.132 palabras al de pruebas.

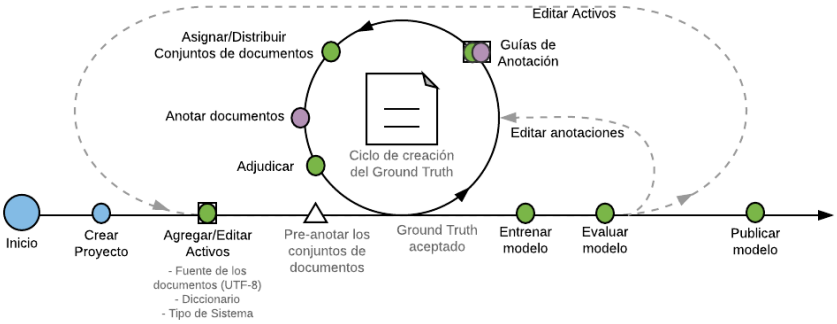

Fig. 2. Flujo de Trabajo de la creación de un modelo de Machine Learning [23, 24].

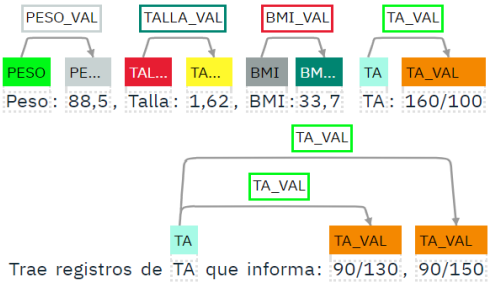

Fig. 3. Etiquetado de documentos.

Obtenidos los conjuntos de entrenamiento y prueba, los mismos fueron incorporados a la plataforma para llevar a cabo el etiquetado. El mismo consistió en revisar los documentos de cada conjunto y seleccionar manualmente las menciones referidas a las entidades y relaciones definidas, marcándolas con el tipo correspondiente. La **Fig.**

[4] "Blind Set" puede traducirse como "conjunto ciego", y se utiliza para testear el sistema de manera periódica, luego de que se hayan ejecutado varias iteraciones de testeo y mejora [22].

3 muestra un ejemplo de dicho procedimiento, donde están distintos términos de una evolución, y para cada uno existe una etiqueta asociada a la entidad a la que pertenece. Entre cada par de entidades se observan etiquetas correspondientes a las relaciones establecidas en el sistema de tipos.

Finalizado el etiquetado de los documentos, se realizó en primer lugar el entrenamiento del modelo de ML utilizando el conjunto de documentos para entrenamiento, y luego se realizaron las pruebas del modelo sobre el conjunto definido para tal fin.

6 Pruebas y resultados

Las pruebas evaluaron la capacidad del modelo de ML entrenado para extraer correctamente la información requerida en base a las entidades y relaciones definidas en el dominio. Así, se estudió su performance utilizando el modelo entrenado sobre el conjunto de documentos de prueba. Los parámetros considerados fueron Precisión, que mide cuántos de los ítems identificados por el sistema fueron correctamente identificados; Recall, que mide cuántos de los ítems que deberían ser identificados, fueron realmente identificados (cuanto mayor es este valor, mejor es el sistema en cuanto a que no ignora ítems correctos); y F1 score, que combina ambos parámetros para proveer una única medida de performance. Las mismas se calculan según se describe a continuación [10, 25]:

$$Precisión = \frac{anotaciones\ correctamente\ detectadas}{todas\ las\ anotaciones\ detectadas} \quad (1)$$

$$Recall = \frac{anotaciones\ correctamente\ detectadas}{cantidad\ de\ anotaciones\ que\ deberían\ haber\ sido\ detectadas} \quad (2)$$

$$F1\ score = \frac{2 \times precisión \times recall}{precisión + recall} \quad (3)$$

La **Tabla 4** presenta los parámetros de performance para las entidades del modelo, mientras que la **Tabla 5** presenta la performance de las relaciones; ambas ordenadas en base a la columna F1.

Tabla 4. Performance del modelo de ML para extracción de entidades.

N°	Tipo de Entidad	F1	Precisión	Recall	#Menciones Entrenamiento	#Menciones Prueba
1	BMI	1.00	1.00	1.00	16	8
2	BMI_VAL	1.00	1.00	1.00	17	9
3	COL	1.00	1.00	1.00	8	3
4	COL_T_VAL	1.00	1.00	1.00	53	25
5	DEJAR_DE_FUMAR	1.00	1.00	1.00	42	15
6	TA	1.00	1.00	1.00	66	25
7	TALLA	1.00	1.00	1.00	23	11
8	TBQ	1.00	1.00	1.00	4	3
9	COL_T	0.96	0.96	0.96	105	49
10	TA_VAL	0.96	0.92	1.00	96	24
11	TALLA_VAL	0.95	1.00	0.90	23	10
12	PESO	0.87	0.91	0.83	23	10
13	PESO_VAL	0.80	0.89	0.73	24	12
14	FUMA	0.71	0.67	0.75	12	8
15	HTA	0.53	0.53	0.53	26	8
16	DBT_POS	0.50	0.40	0.67	14	4
17	EX_FUMADOR	0.33	0.40	0.29	35	14

N°	Tipo de Entidad	F1	Precisión	Recall	#Menciones Entrenamiento	#Menciones Prueba
18	GANAS_DE_FUMAR	0.29	0.33	0.25	15	12
19	DBT_OTRO	0.00	0.00	0.00	8	6
20	NO_FUMA	0.00	0.00	0.00	16	7

Como se observa, los primeros 13 tipos de entidades del modelo poseen parámetros de F1 score que van de 1 a 0.80, luego de los cuales dicho parámetro comienza a decaer, hasta llegar a un valor de 0 para las entidades DBT_OTRO y NO_FUMA. Paralelamente puede notarse una disminución, aunque no igualmente escalonada, de la cantidad de menciones de entrenamiento y prueba para cada tipo de entidad. La menor cantidad de menciones de entrenamiento se asocia a la entidad TBQ, con solamente 4 menciones; mientras que la menor cantidad de menciones de pruebas las poseen las entidades TBQ y COL. De las mismas, 4 entidades se encuentran dentro del rango de baja performance definido por la herramienta, siendo estas EX_FUMADOR, GANAS_DE_FUMAR, DBT_OTRO y NO_FUMA. Los demás tipos de entidades se encuentran por encima de dicho rango.

Tabla 5. Performance del modelo de ML para extracción de relaciones.

N°	Tipo de Relación	F1	Precisión	Recall	#Menciones Entrenamiento	#Menciones Prueba
1	BMI_VAL	1.00	1.00	1.00	30	16
2	TA_VAL	1.00	1.00	1.00	186	48
3	COL_T_VAL	0.96	0.96	0.96	158	74
4	TALLA_VAL	0.95	1.00	0.90	46	20
5	PESO_VAL	0.70	0.78	0.64	46	21

En cuanto al sistema de relaciones se observa que todas poseen un valor F1 score de entre 1 y 0.70, encontrándose también por encima del rango de baja performance.

7 Conclusiones

Este trabajo presenta un procedimiento para la definición de un sistema de entidades y relaciones específico del dominio médico, establecido con el fin de entrenar un modelo de ML que sea capaz de extraer información relacionada a factores de riesgo cardiovascular referidos en evoluciones médicas mediante el uso de IBM Watson™ Knowledge Studio. Para ello se definió una serie de expresiones regulares con las cuales se realizó un filtrado de las evoluciones relevantes para el etiquetado de documentos de entrenamiento y prueba. Luego del etiquetado se entrenó un modelo de ML mediante el conjunto de documentos de entrenamiento, y se evaluó la performance del mismo aplicándolo sobre el conjunto de documentos de prueba.

De los resultados obtenidos, se observa que las entidades para las que las menciones suelen seguir un patrón más estructurado, y no tan narrativo, por ejemplo, términos como “BMI”, “TA”, y “TBQ”, los parámetros de performance son mayores; mientras que dichos parámetros disminuyen cuando la narrativa es más variable en cuanto a la definición del factor de riesgo, como cuando se indica que un paciente es un ex fumador, o que no fuma. Éstos últimos poseen estructuras mucho más variables, por lo que sería necesaria una mayor cantidad de ejemplos etiquetados para aumentar la performance asociada.

Otra cuestión que se observa es que, si bien lo recomendado es contar con 50 menciones para entrenamiento y 20 para pruebas por cada entidad, ello solo pudo ser satisfecho para algunos de los tipos de entidades. Los 4 tipos de entidades con una cantidad de menciones de entrenamiento de 50 o más tuvieron un valor de F1 Score de al menos 0.96. Por otro lado, los tipos de entidades con parámetros de performance de 0 a 0.50 tienen no más de 35 menciones y, en consecuencia, no satisfaciendo lo recomendado. Ello indica una tendencia a que la falta de ejemplos afecte negativamente la performance del modelo para tales entidades.

Un aspecto relevante de esta aplicación es que brinda la posibilidad de realizar estudios estadísticos en el área, extrayendo automáticamente información relevante de una gran cantidad de evoluciones digitalizadas. Ello da lugar a la explotación de dicha información, para la cual pueden definirse tableros de control utilizando minería de datos, sirviendo así como soporte a la toma de decisiones médicas.

Como trabajo futuro se propone profundizar la aplicación con una cantidad de evoluciones más numerosa, de modo que se cuente con una mayor cuantía de ejemplos para entrenamiento y prueba. Además, sería deseable un estudio más detallado de la estructura de las evoluciones, ya que ello permitiría definir entidades más específicas, que puedan capturar la información relevante con términos más estructurados según son redactados por los profesionales de la salud. Por ejemplo, se podrían tomar entidades que para este trabajo fueron considerada unitarias, y descomponerlas en dos o más entidades relacionadas entre sí, de modo que la información relevante sea más específica y, por lo tanto, el modelo requiera menor cantidad de ejemplos para alcanzar niveles de performance elevados para dichas entidades.

Referencias

1 T. Haigh, «Remembering the Office of the Future: The Origins of Word Processing and Office Automation,» *IEEE Annals of the History of Computing*, vol. 28, nº 4, pp. 6-31, 2006.

2 D. C. Wimalasuriya y D. Dou, «Ontology-based information extraction: An introduction and a survey of current approaches,» *Journal of Information Science*, vol. 36, nº 3, pp. 306-323, 2010.

3 S. Singh, «Natural Language Processing for Information Extraction,» 2018.

4 IBM Watson™, «Establecimiento de un sistema de tipos,» IBM Watson™, 19 Julio 2018. [En línea]. Available: https://cloud.ibm.com/docs/services/watson-knowledge-studio?topic=watson-knowledge-studio-typesystem#typesystem. [Último acceso: 29 Abril 2019].

5 Q. Zhu y X. Cheng, «The Opportunities and Challenges of Information Extraction,» de *2008 International Symposium on Intelligent Information Technology Application Workshops (IITAW)*, Shanghai, 2008.

6 K. P. Murphy, Machine learning: a probabilistic perspective, Cambridge, Inglaterra: The MIT Press, 2012.

7 J. Wang y Q. Tao, «Machine Learning: The State of the Art,» *IEEE Intelligent Systems*, vol. 23, nº 6, pp. 49-55, 2008.

8 M. G. Pecht y M. Kang, «Machine Learning: Fundamentals,» de *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*, Chichester, Inglaterra, John

Wiley and Sons Ltd, 2018, pp. 85-109.

9 R. High, «IBM Redbooks Content,» 12 Diciembre 2012. [En línea]. Available: http://www.redbooks.ibm.com/abstracts/redp4955.html?Open. [Último acceso: 30 Abril 2019].

10 L. Tonin, «Annotating Mentions of Coronary Artery Disease in Medical Reports,» 8 Abril 2017. [En línea]. Available: http://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1087619&dswid=-2755. [Último acceso: 30 Abril 2019].

11 IBM®, «Watson Knowledge Studio,» IBM, 15 Octubre 2017. [En línea]. Available: https://www.ibm.com/watson/services/knowledge-studio/. [Último acceso: 30 11 2018].

12 IBM®, «Natural Language Understanding,» IBM, [En línea]. Available: https://www.ibm.com/watson/services/natural-language-understanding/. [Último acceso: 30 11 2018].

13 M. Devarakonda, D. Zhang, C.-H. Tsou y M. Bornea, «Problem-oriented patient record summary: An early report on a Watson application,» *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 281-286, 2014.

14 C. I. Molina Espinoza, «Detección temprana de riesgo cardiovascular usando text mining en los campos de texto no estructurado del registro clínico electrónico,» 2014. [En línea]. Available: http://repositorio.uchile.cl/handle/2250/130804. [Último acceso: 12 Julio 2019].

15 R. Costumero, Á. García-Pedrero, C. Gonzalo-Martín, E. Menasalvas y S. Millan, «Text Analysis and Information Extraction from Spanish Written Documents,» *Springer International Publishing Switzerland 2014*, p. 188–197, 2014.

16 Ministerio de Salud y Desarrollo Social, «4° Encuesta Nacional de Factores de Riesgo,» 2018.

17 J. Goyvaerts, «Regular-Expressions.info,» Just Great Software, 30 Mayo 2016. [En línea]. Available: https://www.regular-expressions.info/index.html. [Último acceso: 2 Mayo 2019].

18 L. Anthony, «Laurence Anthony Website,» 5 Febrero 2019. [En línea]. Available: http://www.laurenceanthony.net/software/antconc/. [Último acceso: 2 Mayo 2019].

19 Python Software Foundation, «About Python™ | Python.org,» [En línea]. Available: https://www.python.org/about. [Último acceso: 23 11 2018].

20 IBM Watson™ Knowledge Studio, «Adding documents for annotation,» 5 Enero 2018. [En línea]. Available: https://cloud.ibm.com/docs/services/knowledge-studio?topic=knowledge-studio-documents-for-annotation&locale=en-us. [Último acceso: 2 Mayo 2019].

21 IBM Watson™ Knowledge Studio, «Annotating documents,» 14 Agosto 2017. [En línea]. Available: https://cloud.ibm.com/docs/services/knowledge-studio?topic=knowledge-studio-user-guide&locale=en-us. [Último acceso: 2 Mayo 2019].

22 IBM Watson™ Knowledge Studio, «Making machine-learning model improvements,» 14 Agosto 2017. [En línea]. Available: https://cloud.ibm.com/docs/services/knowledge-studio?topic=knowledge-studio-improve-ml&locale=es. [Último acceso: 2 Mayo 2019].

23 IBM Watson™ Knowledge Studio, «Machine learning model creation workflow,» IBM®, 19 Julio 2018. [En línea]. Available: https://cloud.ibm.com/docs/watson-knowledge-studio?topic=watson-knowledge-studio-ml_annotator&locale=en. [Último acceso: 2 Mayo 2019].

24 G. Candia y G. Alegre, «Uso de IBM Watson en el Establecimiento de Rankings de Documentos para la Toma de Decisiones Estratégicas,» de *Congreso Nacional de Ingeniería en Informática/Sistemas de Información - 6ta Edición*, Mar del Plata, 2018.

25 D. Maynard, W. Peters y Y. Li, «Metrics for evaluation of ontology-based information extraction,» de *CEUR Workshop Proceedings*, Edimburgo, 2006.

26 D. A. Ferrucci, «Introduction to “This is Watson”,» *IBM Journal of Research and Development,* vol. 56, nº 3.4, pp. 1-15, 2012.

27 I. W. Researcher, «The DeepQA Research Team,» IBM Watson, 1 Agosto 2016. [En línea]. Available: https://researcher.watson.ibm.com/researcher/view_group.php?id=2099. [Último acceso: 30 Abril 2019].

28 IBM®, «Enterprise-ready AI,» [En línea]. Available: https://www.ibm.com/watson/about/index.html. [Último acceso: 30 11 2018].

29 J. B. Penié, «LA HISTORIA CLÍNICA: DOCUMENTO CIENTÍFICO,» *Ateneo 2000,* vol. 1, nº 1, 2000.