ULPGC
Universidad de
Las Palmas de
Gran Canaria

ESCUELA DE INGENIERÍA DE TELECOMUNICACIÓN Y ELECTRÓNICA

TRABAJO DE FIN DE MÁSTER

MEDvolution: Aplicación predictiva multiplataforma basada en un sistema de recomendación para la evolución del cuadro médico de un paciente

Titulación: Máster Universitario en Ingeniería de Telecomunicación

Autor: Yguanira del Pino Vega Vega

Tutor: Dr. Luis Hernández Acosta

Fecha: julio 2022

AGRADECIMIENTOS

"Transmitir mi más sincero agradecimiento a todos aquellos que me han ayudado a lo largo de esta etapa y han colaborado en esta investigación.

En primer lugar, a mi tutor, el Doctor Luis Hernández Acosta, por la confianza depositada en mí y su gran ayuda en la orientación, planificación, información y organización en este Trabajo de Fin de Máster.

En segundo lugar, a mi familia, mi madre, mis hermanos, tíos y primos. A mis amigos y compañeros, con los que he compartido historias y anécdotas, reído y llorado, que han estado a lo largo de toda mi carrera apoyándome en todo momento y animándome a seguir adelante.

También, expresar mi más sentido agradecimiento a aquellas personas que estaban y hoy ya no están entre nosotros. Que fueron mi fuente de motivación y esfuerzo. Las promesas realizadas, al igual que ellos, jamás serán olvidadas.

Desarrollar este estudio ha tenido un gran impacto en mi desarrollo personal y profesional y es por eso por lo que me gustaría agradecer a todas aquellas personas que me han apoyado durante este proceso.

A todos ellos, mil gracias."

RESUMEN

La humanidad se encuentra en la era de la información, su conocimiento abre fronteras, expande creencias e historias. Sin embargo, se genera tanta información a nivel global que estudiarla, analizarla u organizarla es una misión casi imposible. A causa de este problema, se desarrollaron en la última década múltiples tecnologías, entre ellas, los Sistemas de Recomendación. Su objetivo, facilitar el acceso a grandes volúmenes de información.

Concretamente, este proyecto se centra en la información médica o sanitaria. Esto se debe a la gran relevancia que ha tenido la pandemia de COVID-19. Esta pandemia demostró a nivel mundial la importancia de la información, del contraste de datos y, sobre todo, del papel tan importante que cobra el personal sanitario en la sociedad.

En este punto entra MEDvolution, una herramienta de ayuda para el día a día en los puestos de trabajo de los médicos. Esta aplicación garantiza un trato personalizado para el paciente aumentando el tiempo de respuesta del médico. Este incremento lo consigue aportándole la mayor cantidad de información posible al médico sin que este sienta la necesidad de emplear múltiples horas investigando o discutiendo con otros expertos.

MEDvolution emplea como tecnología central el uso de Sistemas de Recomendación basados en Redes Bayesianas. Con esta aplicación, el médico puede predecir el cuadro evolutivo de un paciente a partir de sus síntomas, comorbilidades y otras características; y actuar en base a unas recomendaciones que proporciona el sistema, procurando la mejor atención en el menor tiempo posible al paciente.

ABSTRACT

Humanity is in the information age, its knowledge opens frontiers, expands beliefs and stories. However, so much information is generated globally that studying, analysing or organising it is an almost impossible mission. Because of this problem, multiple technologies have been developed in the last decade, including Recommender Systems. Their objective is to facilitate access to large volumes of information.

Specifically, this project focuses on medical or health information. This is due to the great relevance of the COVID-19 pandemic. This pandemic demonstrated the worldwide importance of information, the data contrast and, above all, the important role played by health personnel in society.

This is where MEDvolution comes in, a tool to help doctors in their day-to-day work. This application guarantees personalised treatment for the patient, increasing the doctor's response time. This increase is achieved by providing the doctor with as much information as possible without him/her feeling the need to spend many hours researching or discussing with other experts.

MEDvolution employs the use of Bayesian Network-based Recommendation Systems as its core technology. With this application, the doctor can predict the evolution of a patient's condition based on symptoms, comorbidities and other characteristics, and act based on the recommendations provided by the system, providing the best care in the shortest possible time for the patient.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

ÍNDICE DE ABREVIATURAS

CN	*Credal Networks*
CNN	*Convolutional Neural Network*
CSRF	*Cross-Site Request Forgery*
CSV	*Comma Separated Values*
GAD	Gráfico Acíclico Dirigido
HCDSNS	Historia Clínica Digital en el Sistema Nacional de Salud
HCR	Historia Clínica Resumida
HTTP	*Hypertext Transfer Protocol*
ID	*Influence Diagrams*
IDE	*Integrated Development Environment*
INSS	Instituto Nacional de la Seguridad Social
IoT	*Internet of Things*
ITS	Infecciones de Transmisión Sexual
LFM	*Latent Factor Model*
LIMIDs	*LImited Memory Influence Diagrams*
MLP	*Multilayer Perceptron*
MN	*Markov Networks*
MOGA	*Multi Objective Genetic Algorithm*
NaN	*Not a Number*
OMS	Organización Mundial de la Salud
PRM	*Probabilistic Relational Models*
RB	Redes Bayesianas
RNN	*Recurrent Neural Network*
SCTID	*SNOMED CT IDentifier*
SNOMED	*Systematized Nomenclature of Medicine – Clinical Terms*
SQL	*Structured Query Language*
SR	Sistema de Recomendación
SRS	Sistema de Recomendación para la Salud
TIC	Tecnologías de la Información y la Comunicación
UCI	Unidad de Cuidados Intensivos
VCS	*Version Control System*
ZB	Zettabytes

CAPÍTULO I. INTRODUCCIÓN

En este capítulo se detalla la información correspondiente al contexto inicial de este TFM, sus objetivos y la estructura del documento con una breve reseña de los siguientes capítulos.

1. Contexto

Tras la primera década del siglo XXI, la convergencia de un gran número de tecnologías se consolidó durante los dos años siguientes, suponiendo la creación de un nuevo concepto que en la actualidad es ampliamente conocido, *Big Data*, que alude a grandes cantidades de datos o macrodatos. Entre las tecnologías precursoras del gran evento se encuentran redes sociales como Facebook, Instagram o WhatsApp; la expansión de Internet con Internet de las Cosas (del inglés *Internet of Things* - IoT), el aumento de ancho de banda debido a la fibra óptica y la computación en la nube (del *inglés Cloud Computing*), entre otras.

Los volúmenes de datos no cesan su crecimiento con los años, este fenómeno se aprecia en la Figura 1, donde se ilustra el crecimiento de datos en Zettabytes (ZB) por año. Por tanto, en sus inicios, la gestión y análisis de Big Data supuso una gran carga computacional que con las tecnologías previas no se podía realizar. Llegados a este punto, emergieron varias tecnologías basadas en inteligencia artificial para afrontar el problema planteado, entre ellas, el Sistema de Recomendación (SR, del inglés *Recommender System* - RS).

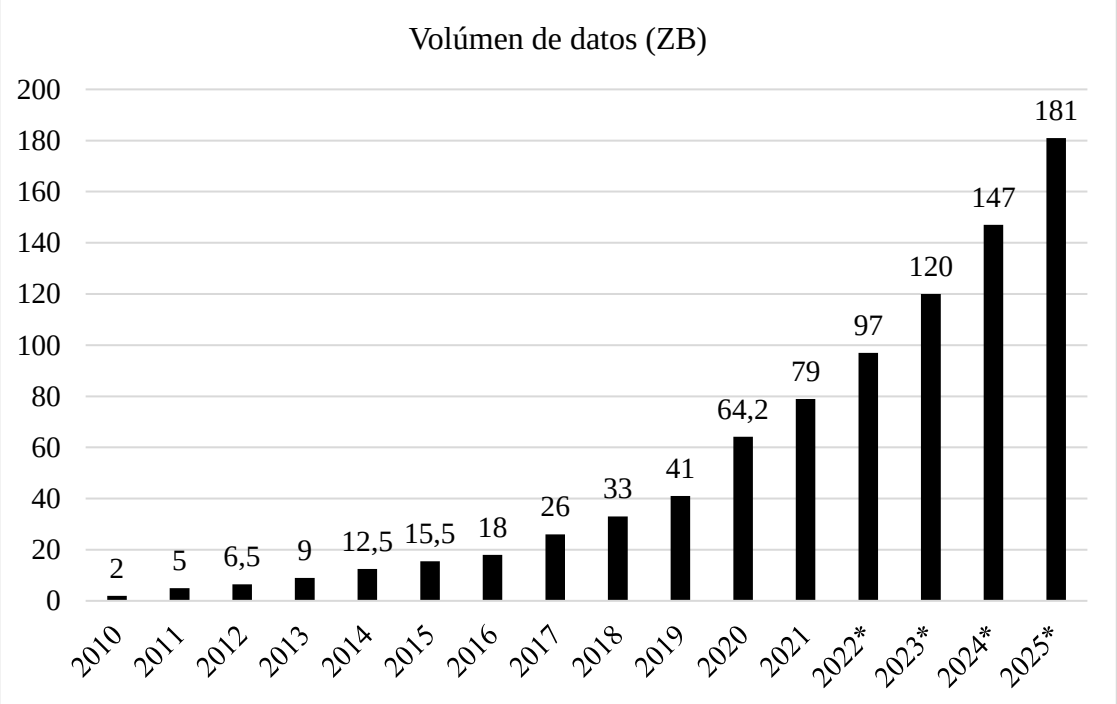

Figura 1. Evolución de Big Data. [1]

Los Sistemas de Recomendación son herramientas y técnicas *software* que proporcionan sugerencias que pueden ser de utilidad para un usuario. Se puede ver un ejemplo del impacto de estos sistemas en la Figura 2, donde sus sugerencias representan 2/3 de todas las películas de Netflix vistas y el 35% de las ventas de Amazon [2]. Por consiguiente, no es erróneo afirmar que estos sistemas, hoy en día, forman parte de la vida cotidiana de gran parte de la sociedad.

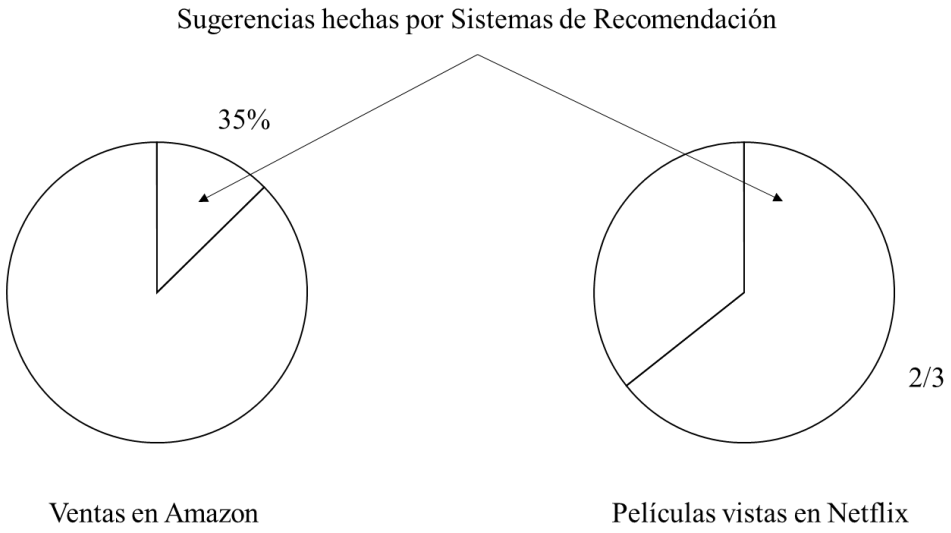

Figura 2. Ejemplo de impacto de los SR en empresas como Amazon o Netflix.

Desde su creación, estos sistemas han logrado un gran éxito en escenarios comercializados, siendo este ámbito su principal campo de implementación. Sin embargo, las sugerencias proporcionadas para el usuario tienen una gran variabilidad en función del marco de actuación en el que se desee implantar el sistema y los datos analizados, por lo que los SR actualmente se aplican en múltiples ámbitos diferentes al comercio como son ocio, salud y enseñanza, entre otros.

2. Antecedentes

En la prehistoria, la medicina consistía en la recolección de plantas que aliviaban o disminuían ciertas dolencias. Con la evolución de la humanidad y los diferentes avances tecnológicos, ese tipo de medicina se ha convertido en una ínfima parte de lo que es hoy en día la medicina moderna, compuesta por múltiples divisiones especializadas. Por

ejemplo, unos expertos en 2020 estipularon que el número de enfermedades raras en el mundo rondaba las 7.000 aunque creían que ese cifra era sólo la mitad de las que realmente existían [3]. Las cifras anteriores son sin tener en cuenta las enfermedades comunes que afectan a la población general ni la infinidad de tratamientos existentes.

En 2019, la Organización Mundial de la Salud (OMS, del inglés *World Health Organization* - WHO) realizó una publicación [4] respecto a las Infecciones de Transmisión Sexual (ITS) en la que indica cifras alarmantes como que cada día un millón de personas contraen una de estas infecciones o que anualmente se estima que unos 376 millones de personas contraen alguna de las cuatro principales ITS. Teniendo como premisa que cada una de esas personas visitó a un médico, se generó a nivel clínico la actualización de 376 millones de expedientes de pacientes en la categoría de enfermedades de transmisión sexual. Si a estos datos incorporamos la clasificación de la CIE-10 [5], que ordena las patologías en 21 macro-categorías con más de tres categorías, se generaron miles y miles de datos clínicos.

Previamente, en 2018, en la revista NEJM Catalyst del grupo NEJM se publicó un artículo [6] en el que sus expertos hacen una previsión de que para el año 2020 se incrementaría la generación de datos relacionados con el ámbito de la salud a más de 2,314 exabytes. Ahora, en 2022 y tras declararse en 2020 la pandemia de COVID-19 que propició el desarrollo exponencial de las plataformas de teletrabajo, teleenseñanza y telemedicina a nivel mundial, el cálculo de la previsión realizada es mucho menor que la cantidad real de datos generados.

Respecto al marco de la salud, el cual es el ámbito de actuación de este Trabajo de Fin de Máster, la enorme cantidad de datos clínicos existentes en diferentes sitios de Internet y plataformas de almacenamiento, y la sobrecarga de información médica obstaculiza a los profesionales médicos a la hora de tomar decisiones orientadas al paciente. Sin embargo, el tema del análisis de datos no es un concepto desconocido en este campo, ya que para mejorar la calidad de la atención médica, puede realizarse un filtrado de macrodatos y un posterior análisis de los datos resultantes. Algunas de las aplicaciones más empleadas en el campo médico que gestionan grandes volúmenes de datos se ilustran en la Figura 3.

Figura 3. Aplicaciones de Big Data en el ámbito de la salud. [6]

Relativo a la tecnología empleada en este proyecto, el SR que se emplea en el ámbito sanitario se denomina Sistema de Recomendación para la Salud (SRS, del inglés *Health Recommender System* - HRS). Los SRS sugieren información médica que está destinada a ser relevante para la evolución del tratamiento médico asociado a la historia del paciente [7]. Estos sistemas proporcionan al personal sanitario, pacientes e investigadores información filtrada según ciertos parámetros como pueden ser edad, tipo de enfermedad o fármacos empleados, entre otros; a partir de la comparativa y el análisis de los datos de miles de pacientes, facilitando la evolución positiva de un paciente frente a una enfermedad o la actuación de los médicos frente a la misma.

En el artículo [8] se plantean los escenarios más comunes de actuación de los SRS: recomendación de comida como pueden ser dietas, sustitutivos de comida; medicamentos para curar enfermedades o más enfocados a ciertas enfermedades como la diabetes o la migraña entre otras; predicción de riesgos asociados al estado de una enfermedad; recomendación de actividades físicas, y médicos para pacientes. Tras el análisis de las aplicaciones de los Sistemas de Recomendación y los datos manejados en los contextos mencionados se obtiene una conclusión clara, los SRS están más enfocados a pacientes y población general que a los propios profesionales sanitarios.

Realizando un análisis actual de las aplicaciones más populares y usadas a nivel mundial para el personal médico [9] se encuentran aplicaciones como epocrates [10], que ofrece

elementos educativos como artículos y vídeos, además de otros servicios de información respecto a medicamentos incluyendo interacciones dañinas de estos; UpToDate [11], una plataforma que ofrece guías prácticas, información médica, últimas investigaciones y noticias, aparte de la posibilidad de debatir o consultar cuestiones médicas con otros expertos; o MDCalc [12], que contiene una extensa base de datos médicos, guías y, sobre todo, destaca por sus calculadoras médicas, herramientas de apoyo cuyo objetivo es orientar en el diagnóstico, realizar cálculos clínicos, análisis cuantitativos o, incluso, ayudar con la tomas de decisiones.

Figura 4. Logos de las aplicaciones epocrates, UpToDate y MDCalc.

Algo que caracteriza a las aplicaciones mencionadas en el párrafo anterior es que ofrecen una gran cantidad de información médica clasificada a término general, o lo que es lo mismo, se ofrecen los datos de enfermedades: síntomas, pruebas diagnósticas y tratamientos más comunes entre la mayoría de la población que padece dicha enfermedad sin tener en cuenta a la minoría restante, a excepción de las particularidades que surgen en los diferentes foros de debate a partir del intercambio de información entre médicos. Por tanto, en el ámbito sanitario, los profesionales deben buscar información constantemente, contrastar, consultar y debatir con otros expertos respecto a un paciente o comparar historias clínicas.

Aparte de las acciones que deben realizar los médicos, hay que recordar que “cada persona es un mundo”, los síntomas de un paciente no tienen por qué coincidir con los de otro, no necesariamente el tratamiento suministrado a un paciente puede ser beneficioso para otro con la misma patología, ni una enfermedad evoluciona igual en un adulto que en un niño, llegando incluso a diferir entre adultos. Por consiguiente, las tareas de búsqueda de datos, comparativas de expedientes o consultas a otros profesionales, en función del caso, se puede demorar días e incluso semanas. Este hecho, en un contexto generalista, puede no suponer un problema, quizás una molestia, pero cuando un paciente está en riesgo de muerte, dicha temporalidad pasa a ser crítica.

3. Objetivos

Con este proyecto se quiere proporcionar una herramienta de apoyo a los médicos para optimizar su tiempo de respuesta y proporcionar el mejor resultado posible para el paciente. Por consiguiente, el objetivo principal de este Trabajo de Fin de Máster es la creación de una plataforma *software* para web y dispositivos móviles (principalmente tabletas) basada en un sistema de recomendación que ayude a los profesionales del ámbito sanitario a realizar una predicción evolutiva del cuadro médico de un paciente a partir de sus datos en estudio y del historial de los pacientes con la misma patología.

Para conseguir el objetivo principal del proyecto, se deben alcanzar una serie de objetivos específicos que describen el futuro desarrollo del proyecto:

1. Análisis de los sistemas *software* y plataformas existentes para almacenamiento de datos clínicos de pacientes con la intención de saber con exactitud qué datos son realmente los requeridos en la actualidad y poder utilizarlos luego en nuestro desarrollo para la creación del "perfil" de un paciente.
2. Análisis de los SR centrándose en aquellos aplicados en el ámbito de la salud.
3. Desarrollo de una infraestructura (*Back-end*) del SRS para la administración y posterior utilización de aquellos datos procedentes de historiales clínicos que sean necesarios para que el SRS desarrolle su funcionalidad.
4. Desarrollo de la interfaz de usuario (*Front-end*) del Sistema de Recomendación para la Salud que permita conectar a los profesionales del ámbito sanitario con el *Back-end* para recuperar y procesar los datos procedentes de los historiales clínicos necesarios para solicitar la predicción evolutiva acerca del cuadro médico de un paciente.
5. Implementación de la funcionalidad completa del SRS mediante la integración del *Front-end* y *Back-end* desarrollados en una aplicación web responsiva que facilite el acceso y uso del SRS a los profesionales del ámbito sanitario.

4. Estructura del documento

Tras la exposición de los objetivos que ocupan este proyecto, en este aparatado se introducen de forma breve los capítulos que preceden a este en el documento y que se orientan a la investigación y desarrollo de este Trabajo de Fin de Máster.

En los siguientes capítulos, se explica y detalla en el CAPÍTULO II el estado del arte referente a los Sistemas de Recomendación indicando qué es lo que se desarrolla, y en el CAPÍTULO III qué *software* se ha aplicado durante el desarrollo de este proyecto.

Desde el CAPÍTULO IV hasta el CAPÍTULO VII, se detalla el proceso de creación de cada una de las partes que componen el proyecto hasta su versión final que se analiza en el CAPÍTULO VIII, correspondiente a las conclusiones.

Las secciones restantes del documento, por orden, conciernen a la bibliografía empleada en este proyecto, la estimación monetaria para su realización y la documentación adicional necesaria para una mayor compresión y entendimiento del Trabajo de Fin de Máster redactado en este documento.

CAPÍTULO II. ESTADO DEL ARTE

La técnica *software* principal de este proyecto es un Sistema de Recomendación para la Salud y, por consiguiente, en este capítulo se detalla qué es este sistema, algunos ejemplos actuales y qué datos serían necesarios para su desarrollo. Finalmente, se detalla las principales características del sistema a desarrollar.

En la actualidad, hay un gran problema de acceso a la información debido a que Internet tiene un crecimiento exponencial, en otras palabras, hay tanta información en la red que es difícil encontrar la que se desea. Por ello, una de las soluciones a este conflicto son los Sistemas de Recomendación.

1. Sistema de Recomendación

Un SR es una herramienta y técnica *software* que proporciona sugerencias de elementos que probablemente sean de interés para un usuario en particular. "Elemento" es el término general utilizado para denotar lo que el sistema recomienda a los usuarios [13].

Por tanto, los SR ayudan a la toma de decisiones de un usuario proporcionándole varios elementos que pueden interesarle, lo que facilita su búsqueda. Un ejemplo actual de ellos se encuentra en la aplicación mundialmente conocida como Netflix y la cual proporciona al usuario sugerencias de películas, series o documentales en función de un porcentaje de similitud con otros elementos de la categoría que el usuario haya visualizado con anterioridad o que tenga en su lista para ver. Netflix no es el único ni el mayor ejemplo de aplicación de los SR, pero si marcó un antes y un después con el desarrollo de esta técnica *software*. Estos sistemas juegan un gran papel en aplicaciones como Spotify, LinkedIn o Facebook, también en grandes corporaciones como Amazon y YouTube. Por consiguiente, como se puede apreciar, los SR no están limitados a un solo campo o a una sola temática, son sistemas que pueden abarcar múltiples datos de diferentes ámbitos proporcionando miles de sugerencias basadas en distintos criterios.

1.1. Clasificación

En el apartado anterior, cuando se menciona que un SR es una técnica *software*, generalmente se hace referencia al tipo de algoritmo de filtrado de datos que ejecuta el sistema para proporcionar la sugerencia más acertada para el usuario. También se indicó que los SR no se restringen a un solo ámbito de actuación, estando presentes en multitud de aplicaciones diferentes, por lo que disponen de una clasificación de cuatro categorías [14]–[18] y sus respectivas subcategorías (Figura 5) en función de los datos que maneje el sistema.

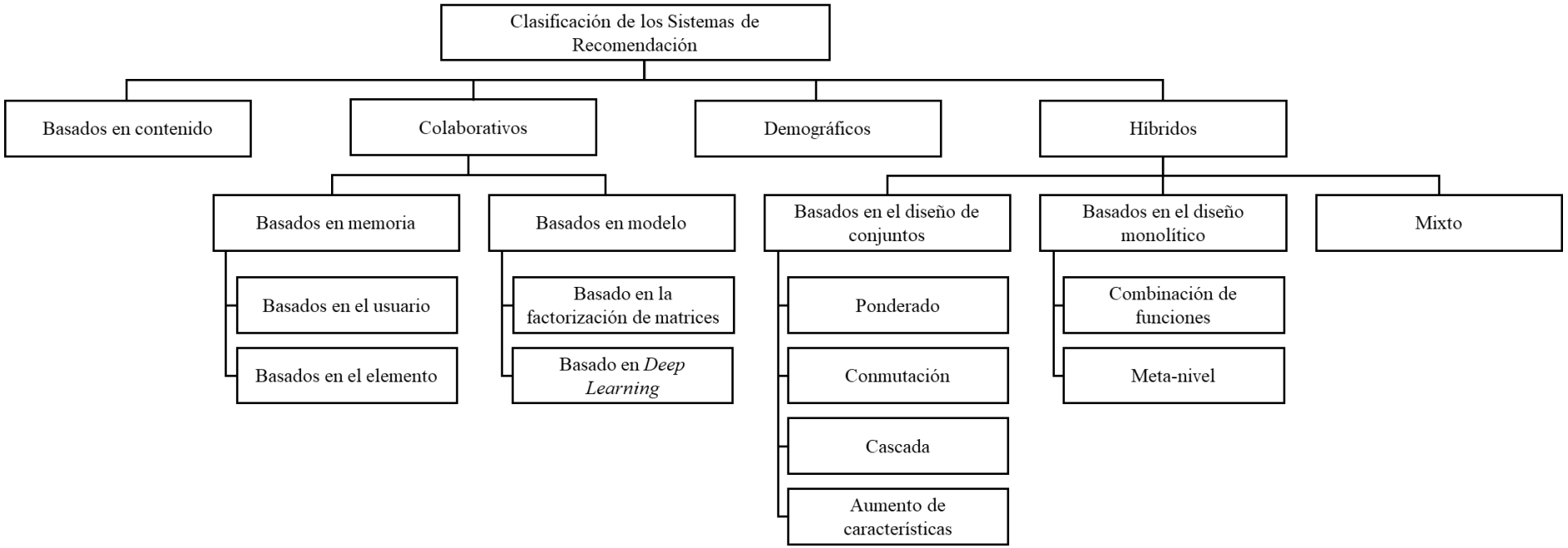

Figura 5. Estructura de la clasificación de los Sistemas de Recomendación

A continuación, se explicarán de forma breve cada una de las categorías existentes en la clasificación indicando además de su funcionamiento, las ventajas y desventajas existentes en sus algoritmos de filtrado y, por consiguiente, del sistema.

1.1.1. Sistemas de recomendación basados en el contenido

Los SR basados en el contenido se focalizan en el filtrado de información en base al perfil del usuario, las propiedades del elemento y el criterio de recomendación estipulado.

El criterio de recomendación estipulado consiste en conocer que elementos le han gustado al usuario a partir de la retroalimentación: explícita como puede ser una calificación, o implícita como el historial de navegación o el de compra de artículos. Teniendo en cuenta lo anterior, ese tipo de algoritmo de filtrado tiene un comportamiento como el mostrado en la Figura 6, a partir del perfil del usuario y los elementos que le han gustado con anterioridad, el sistema busca elementos con propiedades similares para recomendárselos al usuario.

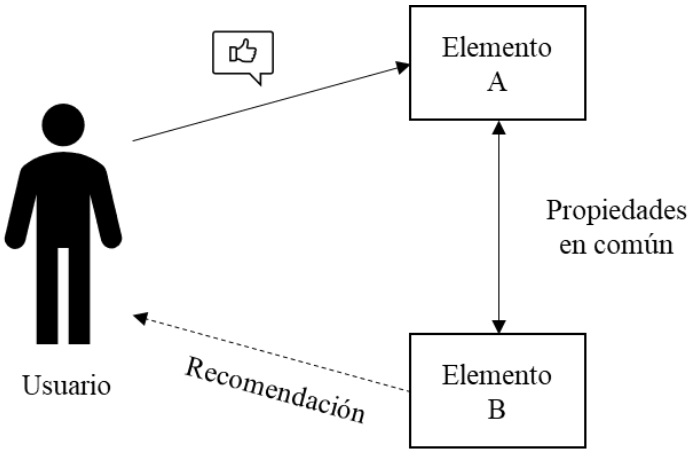

Figura 6. Ejemplo de una recomendación basada en contenido.

Un gran inconveniente de este tipo de sistemas es su dependencia de la información del contenido. Esta le afecta de dos formas diferenciadas: el análisis de contenido limitado y la especialización excesiva. La carencia de recomendaciones diversificadas recae en una sobre especialización, es decir, las recomendaciones serán de elementos muy similares, aportando unas nuevas características mínimas entre elementos.

1.1.2. Sistemas colaborativos de recomendación

A diferencia de los sistemas de recomendación basados en el contenido, los sistemas colaborativos de recomendación emplean filtrado colaborativo, es decir, utilizan la información obtenida de múltiples usuarios del sistema para proporcionar predicciones a un usuario en particular. Son los más empleados en la actualidad y un claro ejemplo de aplicación de este tipo de sistema se encuentra en las plataformas digitales de las empresas Amazon, YouTube y Netflix, integrado dentro de sus sistemas de búsqueda y recomendación de contenido.

Para que estos sistemas ejecuten las predicciones de forma correcta sus algoritmos requieren, a término general, tres tipos de entradas: una participación activa del usuario en el sistema, una forma de cuantificar el interés de los usuarios para el sistema y el acompañamiento de algoritmos que emparejen a usuarios con intereses similares.

Además, los algoritmos de los sistemas colaborativos de recomendación se clasifican en los dos tipos que se detallan a continuación.

1.1.1.2. Filtrado colaborativo basado en memoria

Este filtrado se categoriza en dos algoritmos en base al filtrado de información: usuario o elemento. Sin embargo, independientemente del algoritmo, en esta técnica el sistema de sugerencia a un usuario se basa en el comportamiento y preferencias de otros usuarios hacia los elementos.

- Algoritmo de recomendación basado en el usuario: el sistema, ilustrado en la Figura 7, identifica usuarios con preferencias similares al usuario al que se quiere realizar la recomendación. Por tanto, siguiendo con el ejemplo, se buscan usuarios existentes en el sistema a los que les guste los elementos A y B. Posteriormente, a partir de los elementos que más les gustan a estos, según el ratio de interés, que puede estar clasificado por número de visualizaciones, venta de producto, valoraciones, etc.; realizar una recomendación al usuario con un elemento que difiera entre él y los usuarios similares a él, en este caso el elemento C.

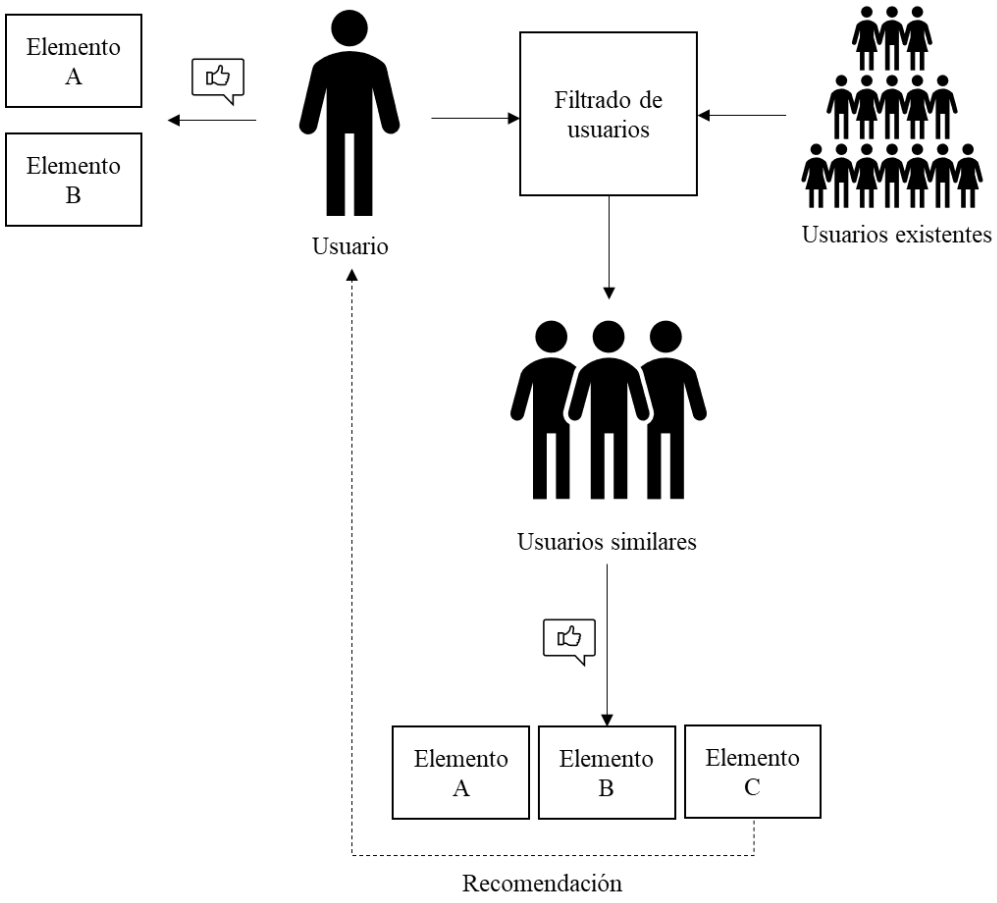

Figura 7. Ejemplo de algoritmo de recomendación basado en usuario.

– Algoritmo de recomendación basado en el elemento: la principal diferencia entre este algoritmo y el anterior es que este identifica usuarios a los que le gusta el mismo elemento que al usuario al que se realiza la recomendación, tal y como se representa en la Figura 8. Por tanto, la única característica que une a los usuarios es que les gusta un mismo elemento y la recomendación viene dada según el ratio de interés de los usuarios frente al resto de elementos, es decir, si a dos de los tres usuarios seleccionados también les gusta B y al tercero C, la recomendación será el elemento B.

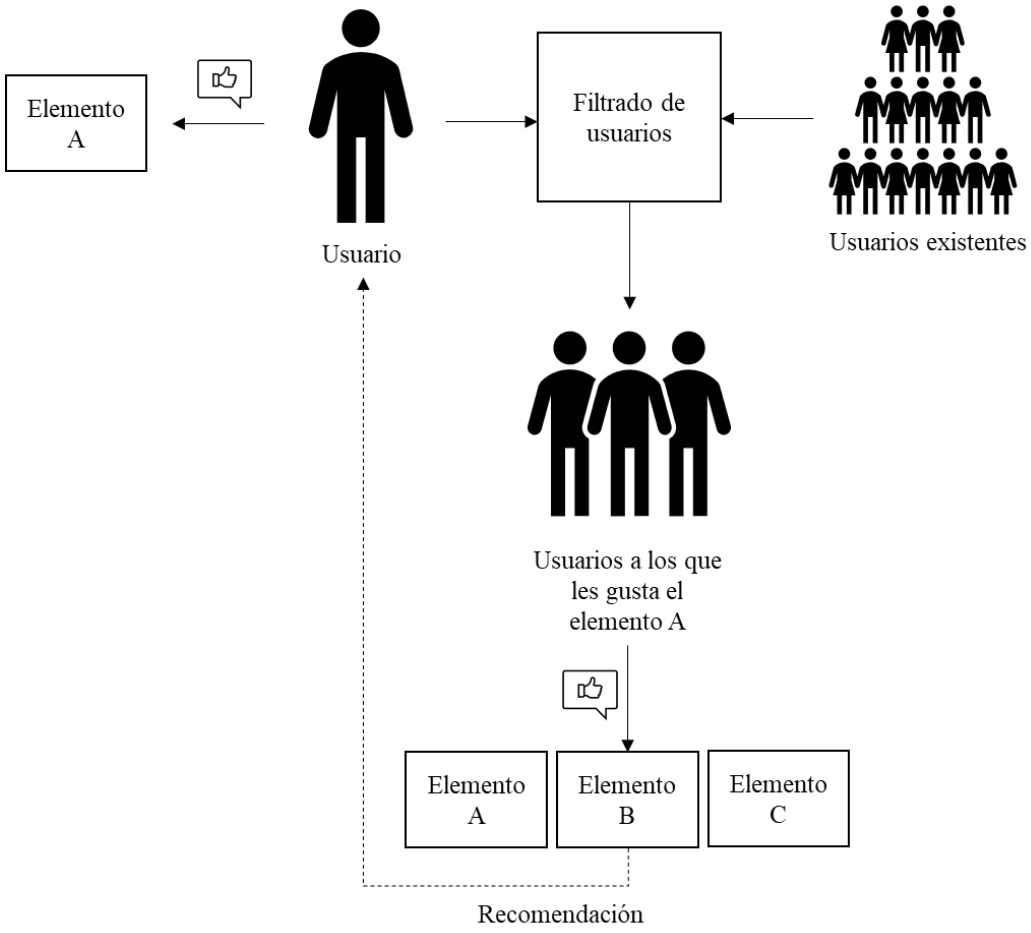

Figura 8. Ejemplo de algoritmo de recomendación basado en el usuario asado en elemento

Como se puede apreciar, entre ambos algoritmos sólo hay algunas diferencias respecto a la selección de usuarios y criterio de calificación para la recomendación, el procedimiento es similar. No obstante, el filtrado colaborativo basado en memoria presenta múltiples retos. Uno de ellos consiste en que, si las calificaciones proporcionadas por los usuarios son inferiores a los elementos del conjunto de datos, los valores de similitud calculados no serían veraces.

2.1.1.2. Filtrado colaborativo basado en modelo

Debido al gran aumento de datos, el filtrado colaborativo basado en memoria consume una gran cantidad de recursos informáticos, lo que provoca una ralentización del sistema. En cambio, los algoritmos de recomendación de filtrado colaborativo basados en modelos proporcionan una gran velocidad de entrenamiento, reducción de espacio en memoria y

una mayor precisión. Esto se debe a que emplean técnicas de *machine learning* para hallar patrones en los datos.

A continuación, se presenta brevemente las dos técnicas empleadas en el filtrado colaborativo basado en modelo.

- Algoritmo de recomendación basado en factorización de matrices: es una de las técnicas con mayor éxito en los SR. Una de las formas de factorización de matrices empleada en este tipo de algoritmo se aprecia en la Figura 9 y se conoce como modelo de variable latente (del inglés *Latent Factor Model* - LFM).

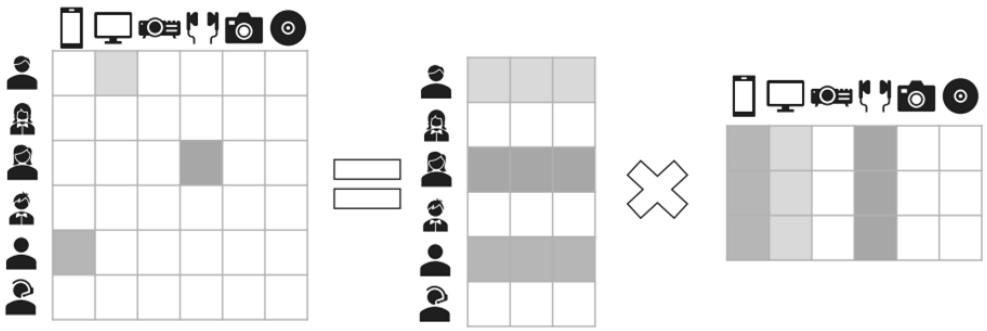

Figura 9. Factorización matricial basada en Modelo de Factor Latente. [14]

A grandes rasgos, las técnicas de factorización matricial empleadas en este algoritmo, como LFM, intentan inferir los datos desconocidos. Por consiguiente, son capaces de crear sugerencias nuevas a partir de los datos iniciales.

- Algoritmo de recomendación basado en *deep learning*: es una técnica que se basa en los métodos de aprendizaje de la inteligencia artificial. En su aplicación a los SR, su función consiste en disminuir la relación no lineal entre usuarios y elementos y aprender las características ocultas de usuarios y elementos. Algunos de los modelos de *deep learning* más extendidos dentro de los SR son Perceptrón Multicapa (del inglés *Multilayer Perceptron* - MLP), Red Neuronal Convolucional (del inglés *Convolutional Neural Network* - CNN), Red Neuronal Recurrente (del inglés *Recurrent Neural Network* - RNN) y Redes Bayesianas (RB, del inglés *Bayesian Networks* - BN).

Se debe tener en cuenta que estos algoritmos desarrollan un modelo para predecir la información desconocida entrenando la información conocida en la matriz de usuario-elemento, tarea que se dificulta si la matriz contiene escasos datos.

1.1.3. Sistemas de recomendación demográfica

Este tipo de sistemas son bastante característicos ya que a diferencia de los anteriores que emplean las preferencias del usuario, estos emplean la información personal del usuario para catalogarlo demográficamente, es decir, según diferentes parámetros como su ubicación geográfica, grupo social al que pertenece, edad, etc., y recomendar elementos en función de la clasificación del usuario y de usuarios con el mismo perfil demográfico, como se ejemplifica en la Figura 10.

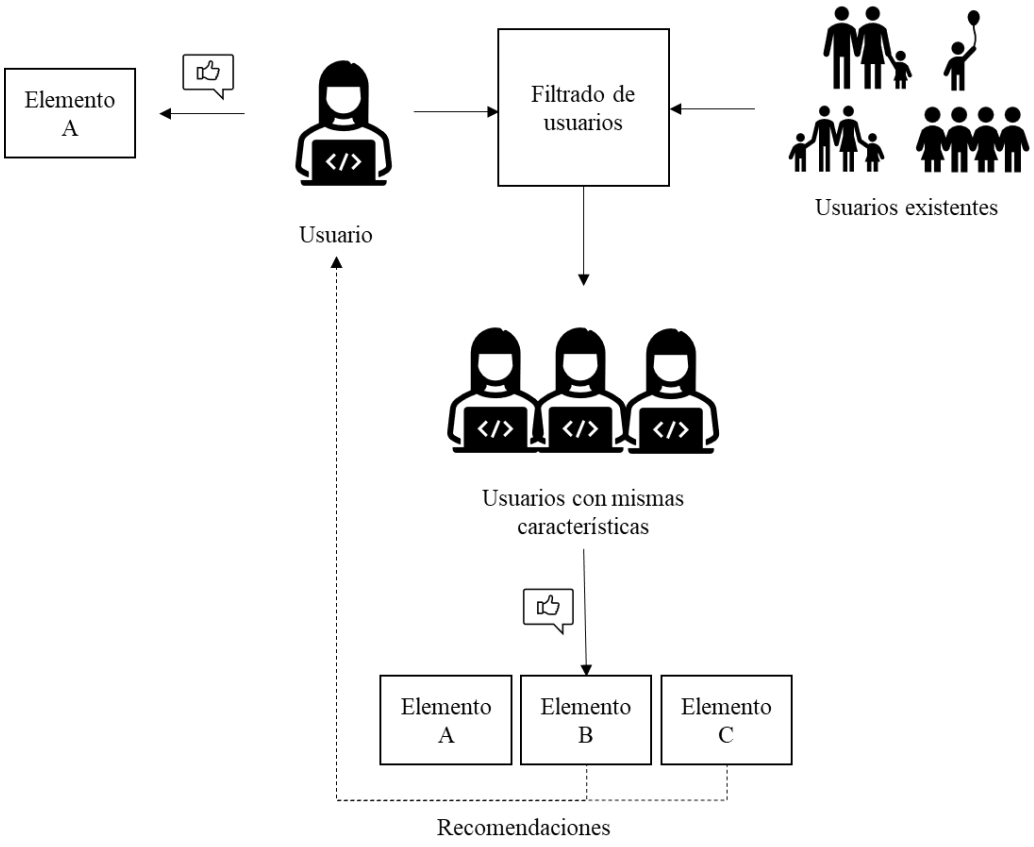

Figura 10. Ejemplo de sistema de recomendación demográfico.

1.1.4. Sistemas de recomendación híbridos

Como su propio nombre indica, estos sistemas contemplan el uso de varias técnicas de recomendación mejorando la predicción general de los SR. Se pueden utilizar diferentes

enfoques para construir estos sistemas como: utilizar varias técnicas de recomendación que creen múltiples escenarios que se pueden ir focalizando de más generalista a menos, dando lugar a una recomendación final al usuario lo más particular posible; o seleccionar un sistema de recomendación con recomendaciones más consistentes. Actualmente, existen tres formas de crear los sistemas de recomendación híbridos: mediante diseño de conjuntos, diseño monolítico y el sistema mixto.

La combinación, siguiendo unas reglas, de los resultados de múltiples algoritmos de recomendación para obtener un único resultado se denomina diseño de conjuntos. Según la regla que se utilice se clasifican en distintos métodos, entre los que destacan:

- Ponderado: consiste en la combinación lineal ponderada de los resultados de los algoritmos de recomendación.
- Conmutación: emplea la alternancia entre varios algoritmos de recomendación según las características de filtrado que se deseen.
- Cascada: este método concatena, según un orden de prioridad, varias técnicas de recomendación para optimizar los resultados de la anterior.

El diseño monolítico consta de la unión de múltiples estrategias de recomendación en un único algoritmo integrando múltiples funciones que contemplan diferentes fuentes de datos para realizar una recomendación.

Por último, el sistema mixto consiste en mostrar una lista de los resultados obtenidos por diferentes algoritmos de recomendación al usuario.

1.2. Ejemplos actuales de SRS

Ahora que se ha contextualizado los SR y descrito su funcionamiento a grandes rasgos, aunque los Sistemas de Recomendación para la Salud están en mayor medida orientados a los pacientes y público general, es cierto que existen sistemas enfocados a la ayuda de profesionales médicos. Un gran porcentaje de estos se enfocan a la recomendación de medicación efectiva frente a una enfermedad para un paciente, mientras que otros sí se focalizan en la ayuda a la predicción de enfermedades.

Actualmente, lo que más existe son Sistemas de Recomendación específicos para una enfermedad concreta que ayudan a los pacientes modificando sus hábitos y dietas. A

continuación, se comentarán de forma breve el funcionamiento de dos de los más recientes.

En primer lugar, se encuentra el SRS para el diagnóstico de cáncer de cuello uterino en mujeres [19]. La creación de este recomendador se basa en la premisa de que este cáncer es una de las mayores causas de muerte en mujeres a nivel mundial. El algoritmo que emplea es el Algoritmo Genético Multiobjetivo (del inglés *Multi Objective Genetic Algorithm* - MOGA) que se encarga de seleccionar la población con mayor predisposición a la enfermedad utilizando operadores de genética natural y así hallar características clave de la enfermedad.

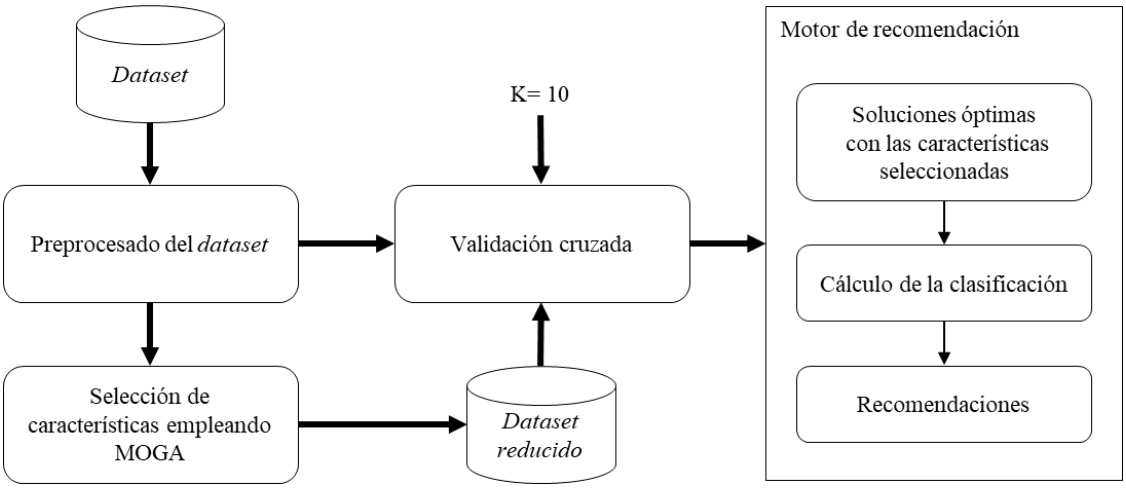

Figura 11. Arquitectura del SRS para el diagnóstico de cáncer de cuello uterino en mujeres.

En la arquitectura mostrada en la Figura 11, se puede observar como el sistema parte de un *dataset* con la información de pacientes que se procesa y valida hasta obtener las características causantes del cáncer y las cuales son los datos de entrada del Sistema de Recomendación. Posteriormente, el sistema clasifica las características y recomienda aquellas con mayor probabilidad, es decir, aquellas que el sistema estima ser clave para padecer la enfermedad y así conseguir el pronóstico del cáncer de cuello de útero en las mujeres.

En segundo lugar, destaca el Sistema de Recomendación para la Salud destinado a pacientes alojados en la Unidad de Cuidados Intensivos (UCI)[20]. El objetivo de este recomendador es predecir e informar al personal médico sobre los pacientes en estado crítico para actuar de forma inmediata reduciendo la tasa de mortalidad. Por tanto, es un

sistema en tiempo real y para cumplir dicha meta, el sistema se estructura como se muestra en la Figura 12.

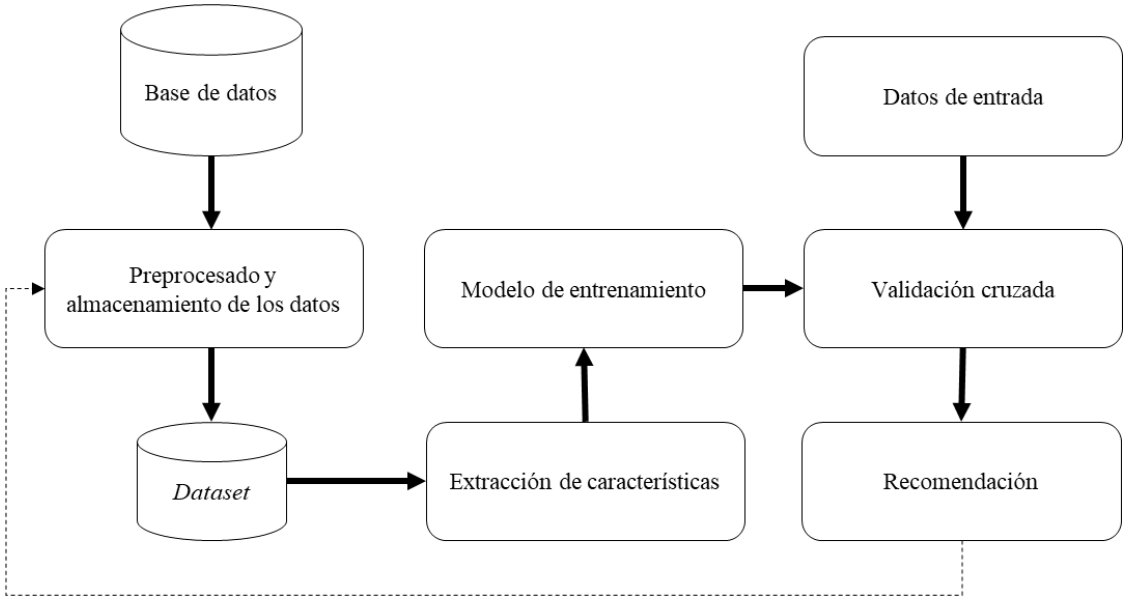

Figura 12. Arquitectura del SRS para la detección de pacientes en estado crítico en la UCI.

El sistema obtiene los datos de la base de datos central del hospital y del terminal de registro. Dichos datos se procesan y almacenan en un *dataset* del cual se extraen las características que se emplean para el modelo de entrenamiento del sistema. A continuación, se valida la salida del modelo desde el terminal de registro y se genera la recomendación final que retroalimenta los datos. Por consiguiente, al incorporar el propio resultado del sistema a los datos empleados para entrenamiento, han conseguido diseñar e implementar un sistema cíclico que, además, proporciona una recomendación cada minuto.

2. Datos

Dado que en este proyecto se pretende desarrollar un SRS, al igual que en los ejemplos anteriormente descritos, es necesario obtener datos sanitarios referentes a pacientes con los que poder realizar cálculos y análisis que sean útiles para los médicos. La mejor fuente donde obtener estos datos se denomina historia clínica.

La historia clínica es un documento legal que almacena todos los datos sanitarios referentes a la salud, asistencia y procesos médicos suministrados a un paciente. Se

origina la primera vez que una persona acude al médico. Es un documento identificativo, personal y único, de un paciente asociado a un hospital, un centro de atención primaria o un consultorio médico.

Una historia clínica se puede categorizar en tres tipos [21]: cronológica, es la estructura original del documento y está ampliamente extendida en los hospitales; orientada por problemas de salud, este tipo de documento tiene una estructura generalista y, por consiguiente, es la que se emplea en los centros de atención primaria, comúnmente conocidos como centros de salud; y de seguimiento concreto de una enfermedad, su estructura es focalizada, con preguntas concretas respecto a una enfermedad. Esta última categoría se emplea en unidades especializadas y con patologías concretas como puede ser la unidad de oncología.

Dependiendo de la categoría e independientemente de la floritura, la información [21], parcial o completa incluida en la historia clínica que aporta el conocimiento del estado veraz y actualizado de salud de un paciente es la siguiente:

- Documentación relativa a la hoja clínico-estadística.
- Autorización del ingreso.
- Informe de urgencia.
- Anamnesis y exploración física.
- Evolución.
- Órdenes médicas.
- Hoja de interconsulta.
- Informes de exploraciones complementarias.
- Consentimiento informado.
- Informe de anestesia.
- Informe de quirófano o registro del parto.
- Informe de anatomía patológica.
- Evolución y planificación de cuidados de enfermería.
- Aplicación terapéutica de enfermería.
- Gráfico de constantes.
- Informe clínico de alta.

La historia clínica de un paciente cuenta con todo su historial médico detallado en un único documento. Sin embargo, este informe cuenta con una gran restricción, sólo es accesible por el centro u organismo médico que la almacena.

La población rara vez permanece todo el ciclo de su vida en una misma ubicación geográfica, eventualmente motivos familiares, laborales o de ocio provocan el desplazamiento de una persona fuera de su zona residencial a otro punto geográfico, ya sea dentro de su mismo país o fuera de este. Asimismo, en dichos desplazamientos, las personas pueden requerir asistencia sanitaria debido a múltiples causas. El problema de este tipo de asistencia sanitaria, externa al centro médico habitual, es que el equipo médico que atiende no dispone del registro médico del paciente, o lo que es lo mismo, su historia clínica. Por tanto, no puede tratarlo de forma adecuada debido al desconocimiento de alérgenos, cirugías o patologías previas del paciente, más allá de la anamnesis recogida durante la consulta.

2.1. HCDSNS

Para solucionar el problema anteriormente planteado, a nivel nacional, en el año 2006 el Gobierno de España concretó el proyecto Historia Clínica Digital en el Sistema Nacional de Salud (HCDSNS) [22], finalmente implantado en el año 2010. Su objetivo es garantizar que los pacientes puedan ser atendidos en cualquier servicio del Sistema Nacional de Salud con la garantía de disponer de su información clínica previa. Este proyecto surgió a raíz de las facilidades de acceso a la información sanitaria que ofrecen las Tecnologías de la Información y la Comunicación (TIC), proporcionando al personal sanitario y al propio paciente la información clínica de este último en su comunidad autónoma y en todo el territorio nacional de forma telemática las 24 horas del día.

Ya que los datos del proyecto deben ser accesibles por profesionales de todas las comunidades autónomas se declaró un comité [23] constituido por 46 miembros de 27 sociedades y asociaciones sanitarias españolas y 12 expertos, con perfiles de tipo institucional como gestores de centros y expertos en admisión y documentación clínica, para crear un consenso de los informes sanitarios y los datos que estos debían incluir, con el objetivo de que fueran fácilmente interpretados por los profesionales sanitarios independiente de la terminología empleada durante sus años de formación.

De entre los informes definidos destaca uno, la Historia Clínica Resumida (HCR) del paciente. En él se estipuló el conjunto mínimo de datos (Figura 13) referente al paciente que debe conocer un profesional sanitario para proporcionarle una asistencia sanitaria de calidad independientemente de donde se encuentre dentro del territorio nacional.

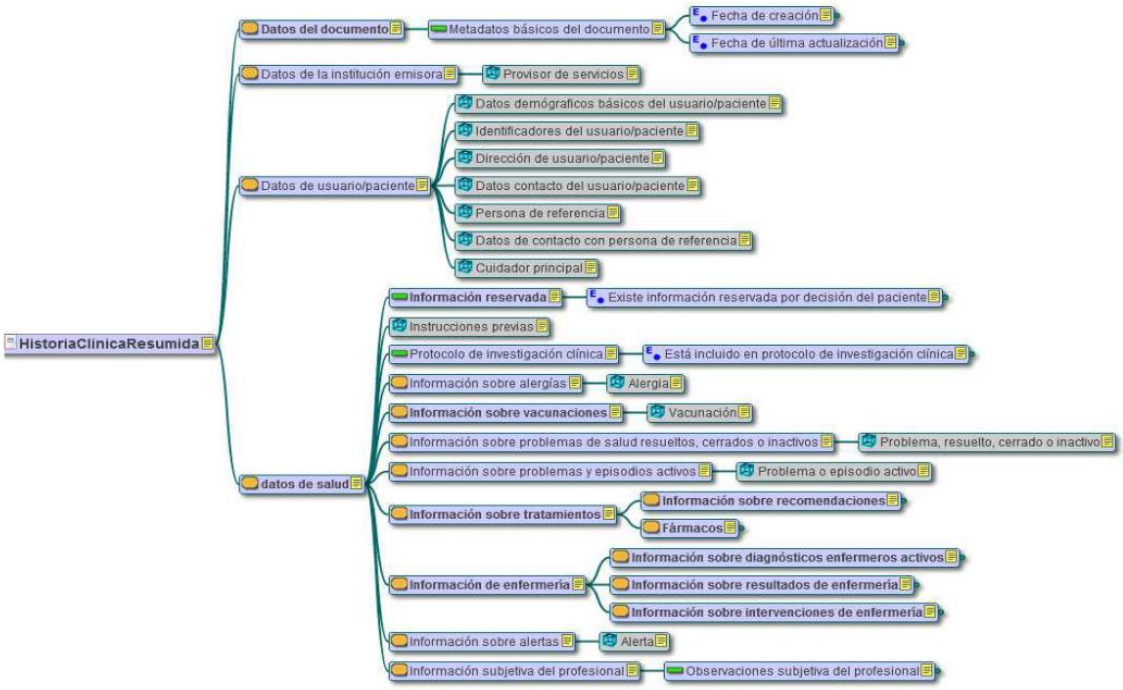

Figura 13. Arquetipo del informe HCR [24]

De todos los datos presentes en la HCR, a continuación, se detallan en la Tabla 1 aquellos referentes a la salud del paciente, indicando la variable, su descripción y la persona que tiene acceso a ella, siendo "P" el profesional sanitario y "C" el ciudadano al que pertenece la Historia Clínica.

VARIABLE	DESCRIPCIÓN	ACCESO
Existe información reservada por decisión del paciente	Informa la existencia de algún dato clínico que no figura en la HCR por decisión del propio paciente.	P / C
Existe documento de instrucciones previas	Informa la existencia de que HCR está disponible en el Registro de Últimas Voluntades.	P / C
Está incluido en protocolo de investigación clínica	Informa de la inclusión en un protocolo de investigación en la fecha de última actualización.	P / C
Alergias	Apartado donde se indica el tipo de alergia y la fecha de diagnóstico.	P / C

Vacunaciones		Su contenido es el tipo de dosis de vacuna y la fecha de administración.	P / C
Problemas Resueltos, Cerrados o Inactivos		Apartado donde se indica el tipo de problema resuelto, cerrado o inactivo.	P / C
Problemas y Episodios Activos		Listado donde figura la fecha de diagnóstico y el tipo de problema o episodio hasta la fecha de actualización.	P / C
Tratamiento	Recomendaciones	Se trata de recomendaciones terapéuticas que no incluyen fármacos.	P / C
	Fármacos	Prescripciones activas a la fecha de última actualización.	P / C
Diagnósticos Enfermeros activos		Los que figuren en la historia a la fecha de última actualización.	P / C
Resultados de Enfermería		Aquellos resultados seleccionados para identificar la evolución del paciente, como resultado de las intervenciones planificadas. Los que figuren en la historia a la fecha de actualización.	P / C
Alertas		Su contenido deben ser advertencias clave de carácter objetivo que por su especial trascendencia deban ser resaltadas para ser tenidas en cuenta por cualquier profesional que deba prestar atención.	P / C
Observaciones Subjetivas del Profesional		Valoraciones del profesional de interés para el manejo de los problemas de salud por otro profesional.	P

Tabla 1. Conjunto de datos mínimos de salud de la HCR de un paciente. [23]-[25]

En el último informe de la situación del sistema de HCDSNS [26], que data de enero de 2021, se indica que poco más de 42 millones de españoles tienen referencia o cuentan con uno o más documentos de HCDSNS, originando en la base de datos del sistema 53.745.847 referencias. Por tanto, ya que el sistema garantiza la interoperabilidad entre los sistemas de salud de cada comunidad española (estando disponible el informe de HCR en diecisiete de las dieciocho comunidades autónomas), es posible el acceso desde el sistema de una comunidad a las más de 53 millones de referencias. A nivel regional, la comunidad autónoma de Canarias dispone de los sistemas de salud Drago y Selene [27] para la recopilación de dicha información.

En el ámbito específico de Canarias, se denomina Sistema de Información Clínico asistencial de Atención Especializada al sistema comúnmente llamado Selene, sistema de información que gestiona los pacientes en centros hospitalarios, centros de Atención Especializada adscritos a éstos y centros de urgencias.

Para la atención primaria se desarrolló el sistema Drago con la finalidad de realizar una gestión administrativa y clínica. Lo conforman varios módulos que facilitan la gestión de agendas del personal sanitario, citas, listas de espera y solicitud de pruebas complementarias al ámbito de Atención Especializada, entre otros. Asimismo, provee información de las historias clínicas a los sistemas de Salud Pública, Salud Laboral, Fármaco vigilancia, Instituto Nacional de la Seguridad Social (INSS) y Selene.

3. Análisis y definición de propuesta

Una vez contextualizado y detallado el estado del arte referente a la tecnología base de este proyecto, a continuación, se definirá el sistema propuesto.

Se crearán dos Sistemas de Recomendación: uno que a partir de síntomas del paciente proporcione la enfermedad que puede tener, incorporando recomendaciones para aquellas cuya probabilidad sea superior al 20%; y otro que, para una enfermedad específica, a partir de parámetros de la enfermedad obtener probabilidades de predicción de otros parámetros definitivos. Estos últimos pueden ser, por ejemplo, la probabilidad de que el paciente sea transferido a la UCI. Por consiguiente, ambos sistemas propuestos serán SR híbridos: colaborativo y demográfico. Colaborativo dado que el criterio de predicción de probabilidades será a partir de usuarios similares en el sistema, y demográficos debido al carácter de los datos, ya que estos no son originados a partir de la utilización del sistema, sino que deben ser indicados por el usuario que lo emplee.

Respecto a los datos empleados en este proyecto, debido a no poder disponer de una colaboración con un centro sanitario que pueda proporcionar datos anónimos de las historias clínicas de los pacientes, se deben obtener datos de otras fuentes como bases de datos colaborativas, retos de Kaggle o *datasets* en Internet. Sin embargo, aunque no se pueda trabajar con datos procedentes del proyecto HCDSNS sí que se quiere dar la posibilidad de que el sistema sea compatible con él. Por tanto, ya que la historia clínica en España emplea un estándar terminológico específico, SNOMED (del inglés *Systematized Nomenclature of Medicine – Clinical Terms*), el sistema de datos utilizará el mismo, facilitando la posibilidad de integración futura con sistemas que ya cuenten con este estándar médico. Además, al incorporar este estándar al sistema no se limita a nivel nacional, sino que también se puede usar a nivel internacional.

3.1. SNOMED CT

SNOMED CT [28] es la terminología de atención clínica integral, multilingüe y codificada de mayor amplitud, precisión e importancia desarrollada en el mundo, estando presente en más de 80 países (Figura 14). Se caracteriza por su contenido científicamente validado y su adaptación e integración a otros estándares médicos internacionales.

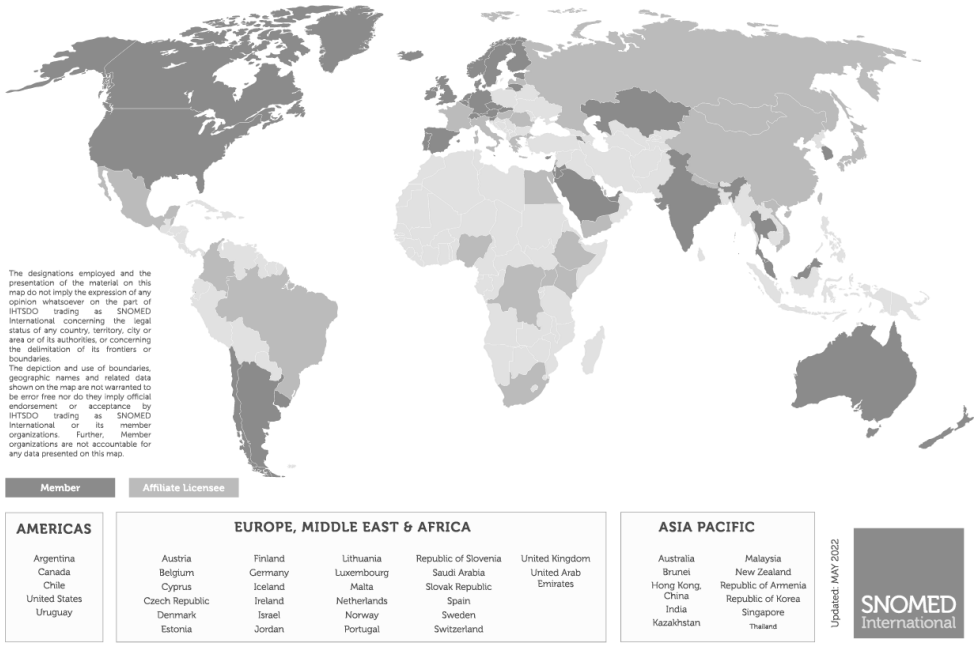

Figura 14. Miembros y licencias afiliadas a SNOMED.

Este estándar cuenta con 18 ediciones, mostradas en la Tabla 2, para diferentes países, siendo su edición internacional la que cuenta con más de 352.567 conceptos.

EDICIONES DE SNOMED CT		
Internacional	Español	Neozelandesa
Argentina	Estados Unidos	Noruega
Australiana	Estonio	Sueca
Belga	Finlandés	Suiza
Canadiense	Irlandés	Traducción común al francés
Danesa	Neerlandesa	Uruguaya

Tabla 2. Ediciones de SNOMEC CT. [29]

En España, el Ministerio de Sanidad es el encargado de suministrar y gestionar la licencia de afiliación de SNOMED CT [30] a dos ediciones: la internacional y la de español. Para este proyecto, se dispone de una licencia de un año con acceso a ambas ediciones y actualizaciones del estándar durante el período de vigencia de la licencia.

CAPÍTULO III. TECNOLOGÍAS *SOTFWARE*

En este capítulo se describe el *software* utilizado durante el desarrollo del trabajo.

A continuación, se realiza en cada apartado una descripción breve, resaltando las funcionalidades principales de los programas y librerías fundamentales empleadas durante este proyecto.

1. PyCharm

El programa PyCharm [31] es un entorno de desarrollo integrado (del inglés *Integrated Development Environment* - IDE) multiplataforma empleado para desarrollar código en el lenguaje de programación Python. Es de la empresa JetBrains y entre sus principales funcionalidades se encuentran: un editor de código inteligente, navegación inteligente por el código y refactorizaciones rápidas y seguras.

Aunque inicialmente fue desarrollado para trabajar con Python, actualmente es compatible con los lenguajes de programación: JavaScript, CoffeeScript, TypeScript, Cython, SQL, HTML/CSS; y los lenguajes de plantillas: AngularJS o Node.js, entre otros. Además, consta con una variedad de herramientas como son: un depurador integrado y ejecutor de pruebas, una terminal de comandos o integración con las principales bases de datos y sistema de control de versiones (del inglés *Version Control System* – VCS).

PyCharm consta de dos tipos de licencias: la *Community* que es gratuita y basada en un desarrollo puro en Python y la *Professional*, con múltiples características entre las que destaca el soporte de desarrollo web. Ofrece una gran compatibilidad con *frameworks* de trabajo de desarrollo web moderno como puede ser Django.

Figura 15. Logo de PyCharm.

2. Django

Django [32] es un *framework* de alto nivel para el lenguaje de programación Python que fomenta el desarrollo rápido. Fue lanzado en 2005, gratuito y de código abierto. Entre sus características destacan:

- Rapidez de desarrollo de aplicaciones. Incluye funciones para manejar tareas comunes de desarrollo web. Por ejemplo: autenticación de usuarios o administración de contenido.

- Seguridad. El sistema ofrece protecciones contra ataques como: *Clickjacking* (se engaña a un usuario para que clique en un elemento de la página web que es invisible o está disfrazado de otro elemento), falsificación de solicitudes entre sitios (del inglés *Cross-Site Request Forgery* - CSRF) o ejecución remota de código, entre otros.
- Escalabilidad. Su arquitectura permite incorporar *hardware* en cualquier nivel (servidores de bases de datos, servidores de almacenamiento en caché o servidores web/de aplicaciones) debido a que separa de forma clara los componentes.

Figura 16. Logo del marco web Django.

3. NumPy y pandas

NumPy [33] es un proyecto creado en 2005 de código abierto cuyo objetivo es facilitar la computación numérica con Python. Consiste en una librería que proporciona un objeto de matriz multidimensional, varios objetos derivados (como matrices y matrices enmascaradas) y múltiples métodos para realizar operaciones rápidas con matrices.

pandas [34] es una librería de código abierto de Python especializada en el análisis y manipulación de datos basados en los *arrays* de la librería de NumPy. Es rápida, potente, flexible y fácil de usar. Algunas de sus características son:

- Lectura y escritura de archivos en ficheros en formato CSV (del inglés *Comma Separated Values*), Excel y bases de datos SQL (del inglés *Structured Query Language*).
- Acceso a los datos mediante índices o nombres para filas y columnas.
- Métodos para reordenar, dividir y combinar conjuntos de datos.

Figura 17. Logos de las librerías de Python NumPy y andas.

4. pyAgrum

pyAgrum [35] es una librería científica desarrollada en C++ y Python dedicada a las RB y otros modelos gráficos probabilísticos.

Está basada en la librería aGrUM y proporciona una interfaz de alto nivel que permite crear, gestionar y realizar cálculos eficientes con Redes Bayesianas y: redes de Markov (del inglés *Markov Networks* - MN), diagramas de influencia (del inglés *Influence Diagrams* - ID) y diagramas de influencia de memoria limitada (del inglés *LImited Memory Influence Diagrams*- LIMIDs), redes de dependencia (del inglés Credal *Networks* - CN), RB dinámicas y modelos relacionales probabilísticos (del inglés *Probabilistic Relational Models* - PRM).

Figura 18. Logo de la librería pyAgrum.

5. Google Colab

Google Colab o *Colaboratory* [35] es un entorno colaborativo de Google que permite escribir y ejecutar código Python mediante *notebooks* de Colab. Asimismo, permite almacenar dichos cuadernos y trabajar con datos que tengas almacenados en el Drive y compartirlos con tu equipo de trabajo. El entorno permite incorporar texto enriquecido, enlaces e imágenes y modificar algunas prestaciones del equipo sobre el que se ejecuta el código.

Figura 19. Logo de Google Colab.

CAPÍTULO IV. *DATASETS*

En este capítulo se detallan los *datasets* seleccionados, su proceso de adaptación a los requerimientos del proyecto y su resultado final para servir de entrenamiento y validación de los SR a desarrollar.

No se dispone de los datos de pacientes por parte de un centro sanitario, pero, a pesar de este hecho, se han podido recopilar a través de Internet. Concretamente, los datos empleados en este proyecto se han obtenido de la plataforma Kaggle [36]. Es una comunidad en línea subsidiaria de la empresa Google que se caracteriza por la gestión de datos científicos que emplean ingenieros y profesionales de *Deep Learning* y *Machine Learning* a nivel mundial.

1. *Datasets* originales

Para este Trabajo de Fin de Máster, se han seleccionado un total de cuatro *datasets* registrados en la plataforma Kaggle. Estos *datasets* han sido divididos en dos categorías según unos criterios que se detallan en los siguientes apartados.

1.1. *Datasets* generales

A la categoría "general" pertenecen los *datasets* que contienen información de síntomas de enfermedades y la enfermedad que padece el paciente. En este caso, se ha seleccionado un archivo, "*Trainning.csv*", del conjunto de archivos que conforman el *dataset* [37]. Comprende un total de 132 síntomas de enfermedades y 42 enfermedades diferentes. Se ha escogido debido a que los datos tienen una licencia de base de datos abierta y una puntuación de 8,24 a pesar de que no se indica la fuente de los datos.

La puntuación de Kaggle proviene de una clasificación denominada *Usability Rating* que se trata de un número de cero a diez que califica la facilidad de uso de un *dataset* basándose en factores como: nivel de documentación, disponibilidad de contenido público relacionado, *kernels* como referencias, tipos de archivo y cobertura de metadatos clave. Por tanto, a término general, al tener un 8,24 de puntuación es un *dataset* que satisface los estándares de calidad de la plataforma.

1.2. *Datasets* específicos

En esta sección se encuentran las categorías correspondientes a cada enfermedad. Debido a la amplia disponibilidad de información en Internet, el representante de esta sección para este proyecto es la enfermedad COVID-19, siendo su categoría "covid19" con un total de tres *datasets*:

- *Dataset* [38]. Este *dataset* contiene diferentes parámetros como enfermedades que padece el paciente, si ha sido transferido a la UCI, si ha sido intubado, etc. La fuente de los datos es una base de datos abiertos del Gobierno de México y tiene una puntuación de 9,71.
- *Dataset* [39]. Contiene edad y síntomas de pacientes. Tiene una puntuación 9,41 y los datos han sido obtenidos de un hospital.
- *Dataset* [40]. Incluye dos archivos, uno con síntomas y otro con parámetros, mayormente enfermedades del paciente simultáneas al COVID-19. A pesar de que tiene una puntuación de 4,71, los datos proceden de un artículo de investigación de *Machine Learning* en el que indican que la fuente de los datos son varias organizaciones a nivel mundial que publicaron los datos de forma abierta.

1.3. Resumen

En conjunto, con los cuatros *datasets* seleccionados se dispone de más de un millón de datos para este proyecto. En la Tabla 3 se puede ver un resumen genérico en el que las filas corresponden a pacientes y las columnas son síntomas y parámetros característicos de la enfermedad.

DATASET	**CATEGORÍA**	**FILAS**	**COLUMNAS**	**CONTENIDO PRINCIPAL**
[37]	general	4.920	134	Síntomas de enfermedades.
[38]	covid19	566.602	23	Parámetros.
[39]	covid19	2.575	6	Síntomas y un parámetro.
[40]	covid19	1.143	27	Síntomas y parámetros.

Tabla 3. Resumen de las características de los datasets.

No obstante, no se va a utilizar la totalidad de los datos presentes ya que se requiere realizar un filtrado de datos y formateo al estándar SNOMED CT, eliminando aquellos datos irrelevantes para este trabajo.

2. Formateo de los *datasets*

A continuación, con el objetivo de crear un *dataset* principal y único por categoría, se detalla paso por paso las operaciones realizadas a cada uno de los *datasets* para ajustarlos

a los requerimientos del sistema y de SNOMED. No obstante, para realizar el ajuste, a término general, para cada *dataset* se siguen los siguientes pasos:

1. Eliminar las columnas innecesarias.
2. Renombrar las columnas que lo requieran.
3. Ajustar datos de las columnas, si es necesario.
4. Traducir los nombres de columnas y sus posibles valores a SNOMED CT.
5. Eliminar columnas duplicadas si existen.
6. Incorporar el *dataset* formateado al principal de la categoría indicando los datos desconocidos con el valor "-1".

2.1. Traducción al estándar SNOMED CT

Antes de comenzar a detallar el proceso de formateo de cada *dataset* original hay que describir un paso clave del proyecto, ya que le aporta un alto grado de internacionalización y se repite para todos y cada uno de los *datasets* empleados, la traducción de los datos a SNOMED CT.

Indiferentemente del *dataset*, tras la ejecución de los tres primeros pasos indicados con anterioridad se obtiene una tabla con columnas que representan términos médicos, ya sean síntomas de enfermedad u otro tipo de parámetros referentes a una enfermedad. Asimismo, las columnas pueden tener valores que también representan términos médicos, por ejemplo, la columna que indica el tipo de enfermedad del paciente.

SNOMED CT se basa en la interconexión de tres tipos de componentes: conceptos, descripciones y relaciones. En el CAPÍTULO VII de este documento se describe con un mayor nivel de detalle cómo se estructura SNOMED CT y dónde está su potencial. Por ahora, sólo se va a indicar de forma breve los dos primeros de esos componentes debido a su nivel de relevancia para la tarea que acontece.

Los conceptos indican pensamientos médicos como "absceso" y son representados mediante un identificador numérico único (SCTID, del inglés *SNOMED CT IDentifier*), mientras que las descripciones son términos legibles. Por ejemplo, en la Figura 20 para SNOMED la enfermedad Influenza es un concepto cuyo identificador es "6142004". Las descripciones asociadas a ese identificador son formas de aludir a lo que representa. Puede ser la forma coloquial, tanto en español como en inglés ("Gripe" o "Flu", respectivamente) o más técnica ("Influenza"), entre otros.

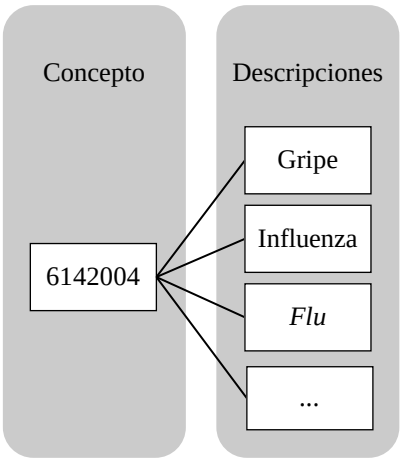

Figura 20. Ejemplo de relación entre concepto y descripciones en SNOMED CT.

Por tanto, el paso de traducción a SNOMED CT consiste en crear un glosario del que obtener, a partir del término médico presente en la tabla el SCTID, es decir, el concepto al que hace referencia. Para ello SNOMED ofrece una poderosa herramienta web [41] que consiste en un buscador en línea que permite buscar conceptos o descripciones en todas sus ediciones, mostrando sus características, relaciones y jerarquías con otros conceptos, entre otras funcionalidades.

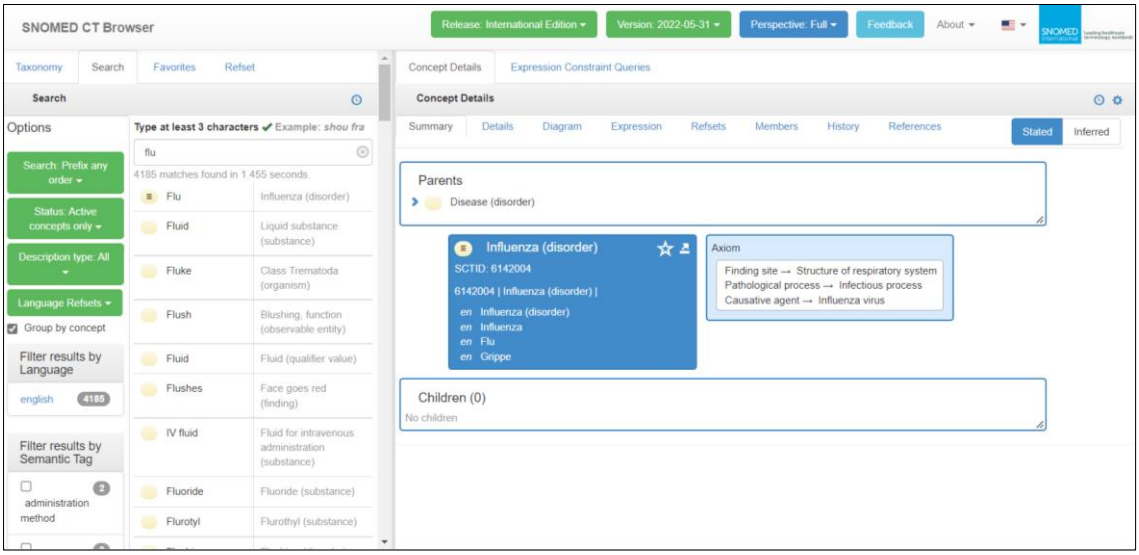

Figura 21. Herramienta de búsqueda de SNOMED CT.

Actualmente, el proyecto consta con un total de seis glosarios que abarcan diferentes ámbitos: síntomas, patologías, hábitos, procedimientos médicos, estados del paciente y otros. En este último glosario se encuentran múltiples términos como “género”, “desconocido” o “paciente”.

2.2. *Dataset* de síntomas de enfermedades

A partir de este apartado comienza la descripción de la reorganización específica realizada a los *datasets* originales, siendo el *datatset* [37] el primero. Pertenece a la categoría "general" y, por consiguiente, debe seguir el formato de que sus columnas sean síntomas de enfermedades, excepto una que consiste en la enfermedad que padece el paciente.

Para conseguir el formato deseado se realiza un análisis de las columnas del *dataset* y se eliminan columnas que no cumplen los requisitos (Tabla 4) por no ser síntomas de la enfermedad. En los casos de la columna "*fluid_overload.1*" se elimina porque está repetida y la de "*yellow_urine*" porque no se detalla con claridad el síntoma que representa.

COLUMNAS ELIMINADAS
extra_marital_contacts
family_history
fluid_overload.1
history_of_alcohol_consumption
receiving_blood_transfusion
receiving_unsterile_injections
Unnamed: 133
yellow_urine

Tabla 4. Lista de columnas eliminadas del datatset [37].

Por otra parte, muchas de las columnas hacen referencia a síntomas de forma coloquial o tienen fallos de escritura y, en algunos casos, crean confusión. Para solucionar esto, se realiza una investigación y consulta a personal sanitario para obtener una terminología clara, dando como resultado los nombres mostrados en la Tabla 5. Este error no sólo se limita a las columnas, tres valores de la columna inicialmente llamada "*prognosis*", que indica la enfermedad del paciente, cambian su valor: "*Bronchial Asthma*" se limita a "Asthma" para eliminar redundancia en el término, "*Osteoarthristis*" y "*(vertigo) Paroymsal Positional Vertigo*" se corrigen gramaticalmente a "*Osteoarthritis*" y "*Paroxysmal positional vertigo*".

NOMBRE DE COLUMNA	NUEVO NOMBRE DE COLUMNA
acute_liver_failure	*acute_hepatic_failure*
skin_rah	*skin_eruption*
nodal_skin_eruptions	*localized_skin_eruption*
spotting_ urination	*interrupted_urination*
patches_in_throat	*white_patches_in_oral_mucosa*
irregular_sugar_level	*irregular_glucose_level*
yellowish_skin	*yellow_skin*
pain_behind_the_eyes	*eye_pain*
diarrhoea	*diarrhea*
yellowing_of_eyes	*yellow_eyes*
swelling_of_stomach	*swollen_abdomen*
swelled_lymph_nodes	*swollen_lymph_node*
blurred_and_distorted_vision	*hazy_vision*
redness_of_eyes	*red_eyes*
sinus_pressure	*nasal_pressure*
bladder_discomfort	*bladder_pain*
congestion	*nasal_congestion*
pain_during_bowel_movements	*pain_during_defecation*
pain_in_anal_region	*anal_pain*
bloody_stool	*blood_in_feces*
irritation_in_anus	*anal_icth*
puffy_face_and_eyes	*swollen_face*
enlarged_thyroid	*thyroid_enlargement*
excessive_hunger	*insatiable_hunger*
swelling_joints	*swollen_joint*
movement_stiffness	*joint_stiffness*
spinning_movements	*dizziness*
foul_smell_of urine	*foul_smell_of_urine*
continuous_feel_of_urine	*urge_to_urinate*
passage_of_gases	*burping*
internal_itching	*burning_epigastric_pain*
toxic_look_(typhos)	*looks_ill*
altered_sensorium	*altered_mental_status*

belly_pain	*stomach_ache*
dischromic _patches	*discoloration_of_skin*
distention_of_abdomen	*swollen_abdomen*
prominent_veins_on_calf	*varicose_veins_on_calf*
pus_filled_pimples	*pimples*
scurring	*atrophic_scar*
silver_like_dusting	*silver_skin_spots*
yellow_crust_ooze	*crust_on_skin*
prognosis	*diagnosis*
breathlessness	*dyspnea*
brittle_nails	*trachyonychia*
bruising	*contusion*

Tabla 5. Columnas renombradas en el dataset [37].

También, el *dataset* cuenta con columnas que reflejan dos síntomas que no tienen por qué darse de forma simultánea. Por tanto, dichas columnas se transforman en dos con valores duplicados, tal y como se refleja en la Tabla 6.

COLUMNA INICIAL	COLUMNA NUEVA 1	COLUMNA NUEVA 2
cold_hands_and_feets	*cold_hands*	*cold_feet*
swollen_extremeties	*swollen_bilateral_upper_limbs*	*swollen_bilateral_lower_limbs*
drying_and_tingling_lips	*dry_lips*	*pins_and_needles*

Tabla 6. Columnas divididas en el dataset [37].

Tras el filtrado y correcciones realizadas se obtiene un *dataset* con un total de 4.920 filas y 129 columnas, de las cuales 128 son síntomas de enfermedades. Posteriormente, se adapta a SNOMED CT los términos correspondientes a los valores de la columna "*diagnosis*" y los nombres de las columnas. Finalmente, se incorpora de forma completa al *dataset* principal de la categoría.

2.3. *Dataset* de parámetros de COVID-19

Es el primero de los tres *datasets* incorporados en la categoría "covid19" y sólo consta de parámetros. El formato de los *datasets* específicos consiste en que sus columnas sean

parámetros relevantes para el diagnóstico de la enfermedad, excluyendo completamente síntomas de dicha patología.

Para el *dataset* [38], se ha eliminado las columnas contempladas en la Tabla 7 por no contener información útil para la predicción del cuadro médico de un paciente de COVID-19. Tanto el identificador de un paciente como la fecha de entrada en el hospital son irrelevantes para dicho propósito. En el caso de la columna "*other_disease*", sólo aporta conocimiento general de si el paciente presenta otro tipo de enfermedad simultánea al COVID-19, pero no indica qué tipo de enfermedad.

COLUMNAS ELIMINADAS
id
entry_date
other_disease

Tabla 7. Lista de columnas eliminadas del datatset [38].

En cuanto a los cambios de nombre en las columnas de este *dataset*, se debe a una mejor comprensión de lo que representan, dichos cambios se reflejan en la Tabla 8. Existe una excepción, la columna "*intubed*". El cambio de nombre de esta columna a "*not_intubed*" es debido a que en SNOMED CT no se contempla el estado del paciente como intubado, si no lo contrario, que no ha sido intubado.

NOMBRE DE COLUMNA	**NUEVO NOMBRE DE COLUMNA**
cardiovascular	*cardiovascular_disease*
icu	*icu_transfer*
inmsupr	*immunosuppression*
sex	*gender*
tobacco	*smoker*
*intubed**	*not_intubated*
covid_res	*diagnosis*

Tabla 8. Columnas renombradas en el dataset [38].

Respecto al ajuste de datos de las columnas (Tabla 9), aquellas columnas cuyos valores corresponden a 1, 2, 97, 98, 99 se han cambiado, siguiendo el orden, por: 1, indica que tiene, padece o sufre; 0, no tiene, padece o sufre; y "-1", que indica que el dato no existe debido a su desconocimiento. El valor NaN (del inglés *Not a Number*) se intercambia en

las columnas que contienen la fecha "9999-99-99". Para la columna "*not_intubated*" la asignación es inversa ya que los datos inicialmente indican si el paciente ha sido intubado. En el caso de las columnas "*gender*" y "*patient_type*", sus valores se cambian por su representación original, eliminando la asignación numérica del *dataset* original. En la columna "diagnosis", que inicialmente era "*covid_res*" se asigna "*covid_19*" a los resultados positivos y al resto "*unknown*" porque se desconoce si el paciente padece la enfermedad o no, o si puede ser otra patología.

NOMBRE DE COLUMNA	ANTIGUOS VALORES	NUEVOS VALORES
asthma	1, 2, 97, 98, 99	0, 1 o -1
cardiovascular_disease		
contact_other_covid		
copd		
diabetes		
hypertension		
icu_transfer		
immunosuppression		
obesity		
pneumonia		
pregnancy		
renal_chronic		
smoker		
*not_intubated**		
date_symptoms	yyyy-mm-dd o "9999-99-99"	yyyy-mm-dd o nan
date_died		
gender	1 o 2	"*female*" o "*male*"
patient_type	1 o 2	"*outpatient*" o "*inpatient*"
diagnosis	1, 2 o 3	"*covid19*" o "*unknown*"

Tabla 9. Lista de cambio de representación de valores en las columnas del dataset [38].

Por otro lado, a partir de la diferencia entre la fecha de los primeros síntomas y la fecha de fallecimiento, si existe, se obtiene los días de supervivencia del paciente frente a la enfermedad. Por consiguiente, se crea una nueva columna denominada "*survival_days*". En ella, aquellos pacientes que pasan con éxito la enfermedad se les atribuye el número

999 para indicarlo. A continuación, se eliminan las columnas "*date_died*" y "*date_symptoms*" que ya no son necesarias.

Tras la reestructuración de datos se obtiene un *dataset* con un total de 220.657 filas y 18 columnas, todas ellas parámetros importantes para la predicción del cuadro evolutivo de un paciente de COVID-19. Posteriormente, se adapta a SNOMED CT los términos correspondientes a los valores de las columnas "*gender*" y "*patient_type*" y los nombres de las columnas. Finalmente, se incorpora el *dataset* de forma completa al *dataset* principal de la categoría.

2.4. *Dataset* de síntomas de COVID-19

Para adaptar el formato del *dataset* [39] al requerido en el proyecto, en primer lugar, se ha cambiado de nombre a tres columnas (Tabla 10) para un mejor entendimiento.

NOMBRE DE COLUMNA	NUEVO NOMBRE DE COLUMNA
bodyPain	*body_pain*
infectionProb	*diagnosis*
runnyNose	*runny_nose*

Tabla 10. Columnas renombradas en el dataset [39].

En segundo lugar, se han reestructurado dos columnas:

- "*fever*". Esta columna representa la fiebre presentada por el paciente en la escala de temperatura Fahrenheit. Debido a que esta escala no se utiliza a nivel internacional, se ha optado por crear tres columnas que representan los tres grados médicos de fiebre [42]: febrícula (37.2 - 38°C / 99 - 100.4°F), fiebre (38-41°C / 100.4 - 105.8°F) e hipertermia (>41°C / >105.8°F). Por cada paciente, según el rango en el que se sitúe la fiebre registrada, se asigna con un "1" en la columna correspondiente al rango y un "0" en las dos restantes.
- "*diffBreath*". En el *dataset* original se distingue entre tres grados de dificultad para respirar: ninguna, moderada y severa. En la terminología médica, el grado "ninguna" hace referencia a la ausencia de síntoma, "moderada" es dificultad respiratoria y "severa" hace referencia al síntoma denominado disnea, que implica una sensación de opresión en el pecho que impide respirar profundamente. Por consiguiente, se crean dos columnas nuevas

"*difficult_breathing*" y "*dyspnea*", en las cuales se asigna un "1" cuando el paciente padece el síntoma, si no un "0".

Tras las operaciones realizadas, se eliminan las columnas originales. A continuación, en la columna "*diagnosis*" se asigna "*covid_19*" a los pacientes que tienen la patología y al resto "*unknown*" porque el paciente tiene síntomas, pero no coincide con COVID-19.

Después del formateo realizado, se obtiene un *dataset* con un total de 2.575 filas y 9 columnas. Posteriormente, se adapta a SNOMED CT los términos correspondientes a los nombres de las columnas. Por último, se dividen los datos en dos tablas: una con las correspondientes columnas de síntomas que se incorpora en el *dataset* principal de la categoría "general" y otra con sólo la columna de edad con los valores de los pacientes que sí tienen COVID-19, que se integra en el *dataset* principal de la categoría "covid19".

2.5. *Dataset* de síntomas y parámetros de COVID-19

Es el *dataset* [40] es el último seleccionado y contiene tanto información de síntomas como de parámetros relevantes para el diagnóstico de COVID-19. En primera instancia se unen los archivos "*symptoms.csv*" y "*comorbidity.csv*". Para adaptar su formato, se realiza un cambio de nombres en las columnas que faciliten el entendimiento de lo que representan, dichos cambios se reflejan en la Tabla 11.

NOMBRE DE COLUMNA	NUEVO NOMBRE DE COLUMNA
cardiovascular	*cardiovascular_disease*
cerebrovascular	*cerebrovascular_disease*
infectious	*infectious_disease*
kidney	*chronic_kidney_disease*
liver	*liver_disease*
lung	*chronic_lung_disease*
malignancy	*cancer*
myalgia	*muscle_pain*
neurodegeneretive	*neurodegenerative_disease*
outcomes	*deceased*
septic	*septic_shock*
surgical	*surgical_history*

Tabla 11. Columnas renombradas en el dataset [40].

Respecto a los valores de las columnas, el cambio se limita a la columna que representa el género del paciente, en la que se elimina la asignación numérica existente en el *dataset* original. Por otra parte, debido a que el *dataset* contempla también síntomas, se crea la columna de "*diagnosis*" indicando que todos los pacientes tienen COVID-19.

Tras la reorganización de datos se obtiene un total de 1.143 filas y 28 columnas. Posteriormente, se adapta a SNOMED CT los términos correspondientes a los valores de las columnas "*gender*" y "*diagnosis*" y los nombres de las columnas. Finalmente, por la estructura del *dataset*, se requiere una división de datos. Las columnas "*headache*", "*fever*", "*cough*", "*fatigue*", "*nausea*", "*diarrhea*", "*muscle_pain*", "*dyspnea*" y "*diagnosis*" se incorporan al *dataset* principal de la categoría "general", mientras que el resto al de la categoría "covid19".

3. *Datasets* principales

El objetivo de estos *datasets* es servir de entrenamiento y validación para los Sistemas de Recomendación a desarrollar en el proyecto. Por tanto, de acuerdo con ellos, el *dataset* de la categoría "general" comprende el conjunto de datos correspondiente a síntomas de pacientes y la respectiva enfermedad que padecen. Por otro lado, el *dataset* de la categoría "covid19" agrupa parámetros relevantes para el diagnóstico del cuadro médico de la enfermedad.

Como se refleja de forma general en la Figura 22, se obtienen tras un proceso repetitivo de formateo y unión de los *datasets* originales del proyecto. Una vez completado el proceso se obtienen dos *datasets* con las características indicadas en la Tabla 12. En conjunto, se han obtenido más de siete mil datos viables para su uso en el proyecto. En el ANEXO I, se detalla cada una de las variables que componen las columnas de los distintos *datasets*, indicando el término médico al que hacen referencia, descripción y posibles valores.

NOMBRE	CATEGORÍA	DESCRIPCIÓN	FILAS	COLUMNAS	TAMAÑO
160237006	general	Historial de síntomas y enfermedad de los pacientes.	8.638	127	7,40 MB
840539006	covid19	Parámetros de los pacientes de COVID -19.	223.071	28	48,41 MB

Tabla 12. Características de los datasets principales del proyecto.

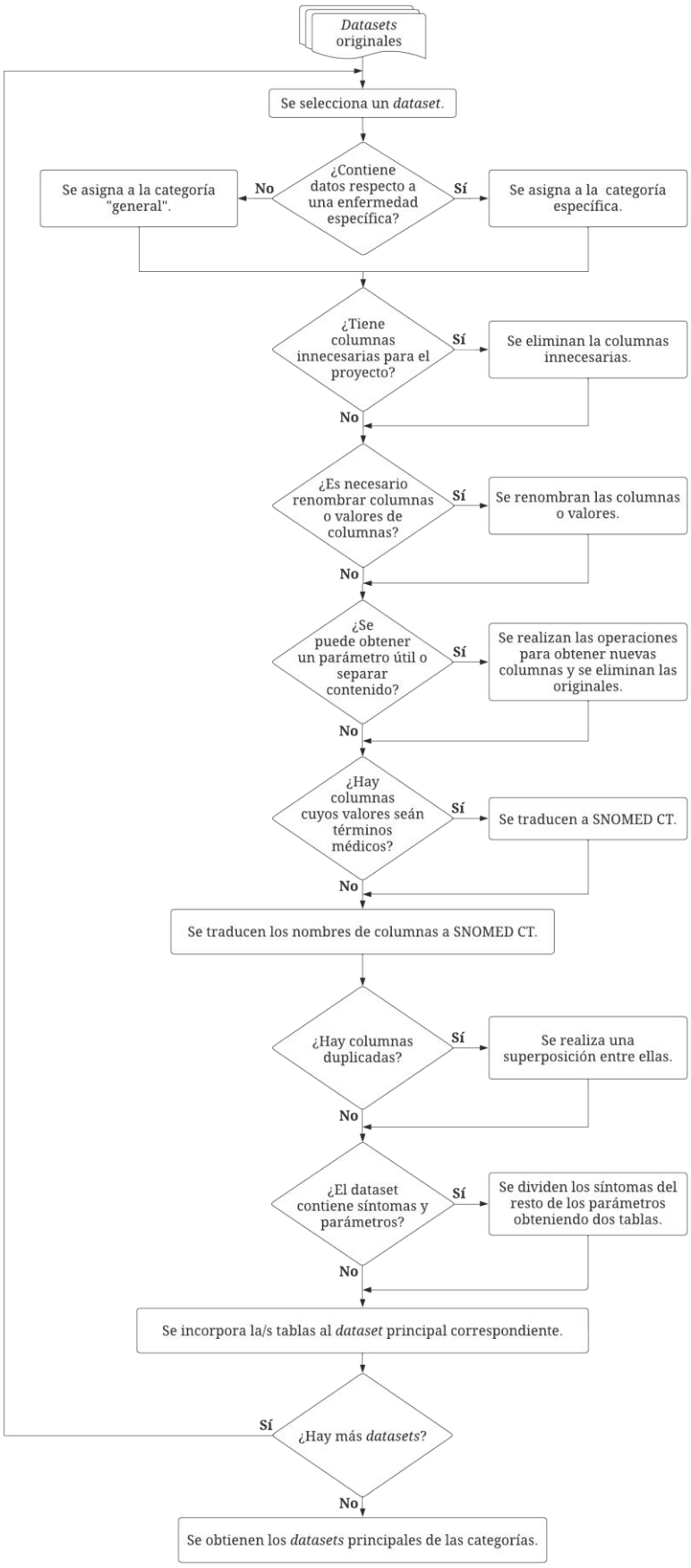

Figura 22. Diagrama de flujo del formateo de los datasets originales y obtención de los principales.

CAPÍTULO V. REDES BAYESIANAS

En este capítulo se detalla el proceso de obtención de los modelos de Redes Bayesianas empleados por los Sistemas de Recomendación del proyecto.

Para los Sistemas de Recomendación, las técnicas de aprendizaje se pueden categorizar en dos tipos: aquellas que usan un enfoque probabilístico y las que usan una predicción basada en una clasificación [43]. Por las características de este proyecto y los datos empleados se ha seleccionado un enfoque probabilístico empleando Redes Bayesianas.

1. Redes Bayesianas

Las Redes Bayesianas [44][45] son una clase específica de modelo gráfico acíclico probabilístico donde la dirección de sus nodos es orientada en un sentido, siendo el término específico para denominar este tipo de gráficos Gráfico Acíclico Dirigido (GAD, del inglés *Directed Acyclic Graph* - DAG). Respecto a sus nodos, estos representan variables aleatorias, mientras que los arcos entre ellos relaciones de dependencia. Como se ilustra en la Figura 23, este tipo de diseño imposibilita comenzar en otro nodo que no sea el inicial ni volver a un nodo anterior.

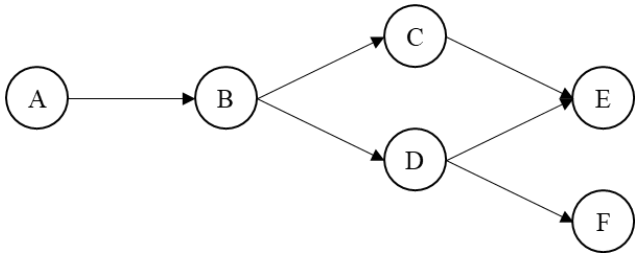

Figura 23. Ejemplo de RB sin probabilidades.[44]

Respecto a las probabilidades de una Red Bayesiana, se obtienen según una factorización particular de la distribución de probabilidad conjunta. Continuando con el ejemplo de la Figura 23, la probabilidad conjunta es la mostrada en la Ecuación 1. Es directamente proporcional al producto de la probabilidad inicial y todas las probabilidades del resto de nodos condicionadas al nodo previo.

$$P(A,B,C,D,E,F) = P(A) \cdot P(B|A) \cdot P(C|B) \cdot P(D|B) \cdot P(E|C) \cdot P(E|D) \cdot P(F|D)$$

Ecuación 1. Distribución de probabilidad conjunta total de la red de la Figura 23.

Si se extrapola a un término generalista, la función de distribución conjunta consiste en el producto de las probabilidades individuales condicionadas a las variables de los nodos anteriores, obteniendo:

$$P(X) = \prod_{i=1}^{n} P(X_i|parents\ (X_i))$$

Ecuación 2. Función de distribución conjunta de una RB.

Sin embargo, el mayor potencial de las Redes Bayesianas consiste en el fenómeno de inferencia, el cual consiste en un razonamiento probabilístico o propagación de probabilidades. Este método se basa en conocer el valor de algunas variables y, a partir de dichos valores, obtener la probabilidad actualizada o posterior del resto, debido a que, al conocer el valor definitivo de ciertas variables, las probabilidades iniciales del resto se modifican. Asimismo, si no se conocen valores con anterioridad, es decir, no se tienen evidencias, se obtienen las probabilidades a priori.

Para explicar mejor qué es el fenómeno de inferencia, se parte del ejemplo ilustrado en la Figura 24. Este ejemplo consiste en una Red Bayesiana con cuatro variables aleatorias que representan: el cielo está o estaba nublado, está o estaba lloviendo, el aspersor está o estaba funcionando y el césped está mojado.

Todas las variables del ejemplo tienen dos estados: sí o no, que responden a la pregunta realizada en cada nodo de la red. Asimismo, el gráfico de la Figura 24 cuenta con relaciones de dependencia entre nodos, indicadas mediante flechas, debido a que:

- Cuando el cielo está nublado el aspersor se apaga para ahorrar agua.
- Cuando llueve, casi la totalidad de las veces, el cielo está nublado.
- Para que el césped esté mojado debe haberle llegado agua de forma previa, ya sea por medio del aspersor o la lluvia.

Las afirmaciones realizadas consiguen que las probabilidades de cada nodo se modifiquen a causa de las relaciones que se crean entre nodos y el estado de los nodos anteriores. La única excepción, en este caso, es el nodo inicial, que como la igualdad en la probabilidad de sus estados indica, es 100% aleatorio.

Ahora que se dispone de las probabilidades de que los eventos sucedan, se conoce que hay una alta probabilidad de que el césped esté mojado independientemente del estado climático puesto que es difícil que un aspersor deje de funcionar correctamente con frecuencia. Sin embargo, si se conocen evidencias, las probabilidades cambian, eso es lo que se conoce como fenómeno de inferencia.

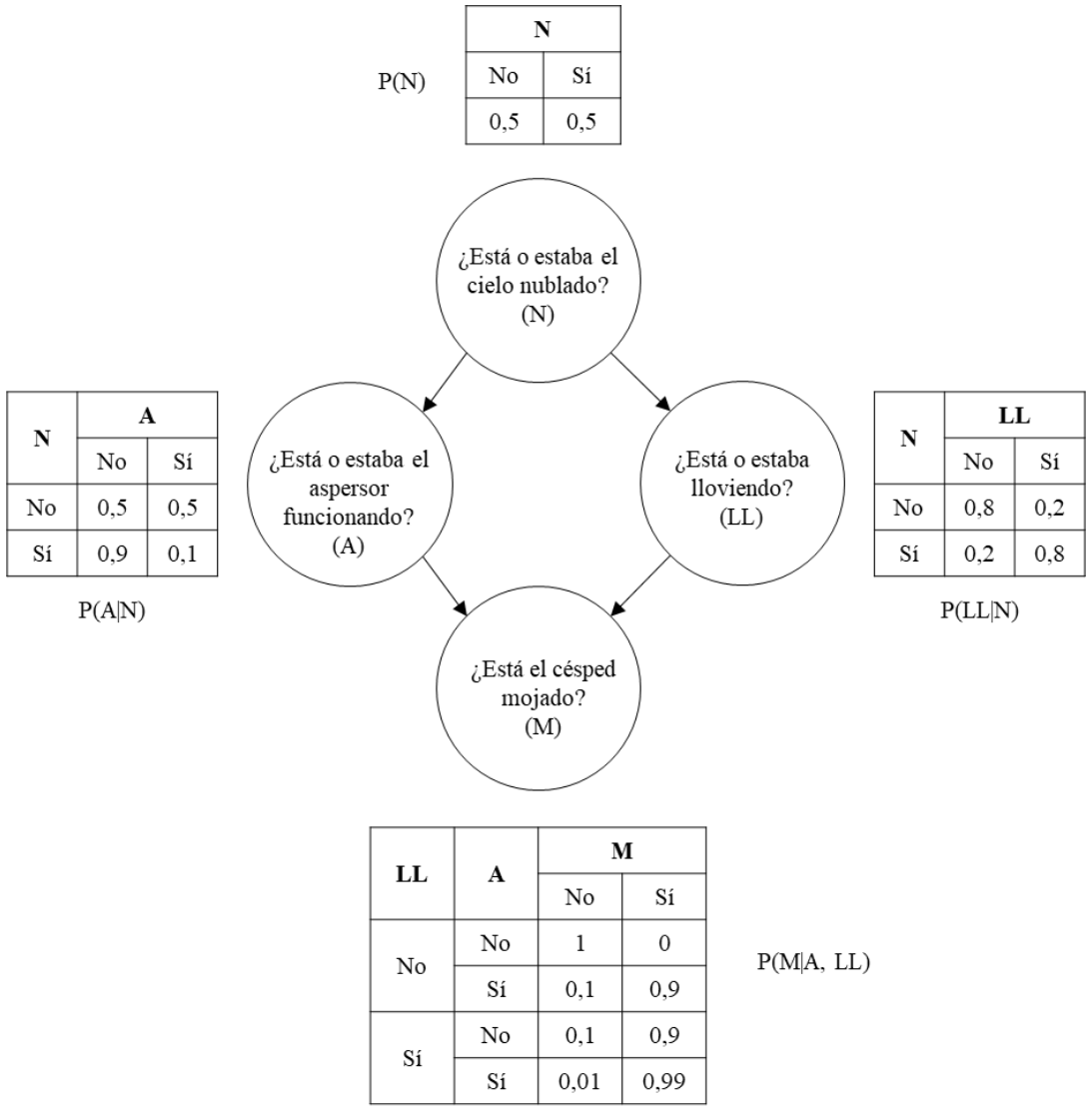

Figura 24. Ejemplo de RB con cuatro variables y sus probabilidades [46].

Una vez se conoce información previa sobre el suceso, para obtener las probabilidades propagadas de la red, se aplica la ecuación matemática correspondiente al Teorema de Bayes (Ecuación 3). Según este teorema, es posible obtener la probabilidad de un suceso (X_i) a partir de las evidencias (E) dado que es directamente proporcional al producto de la probabilidad de que el suceso ocurra y la probabilidad de que se den las evidencias habiendo ocurrido el suceso e inversamente proporcional a la probabilidad de las evidencias.

$$P(X_i|E) = \frac{P(X_i) \cdot P(E|X_i)}{P(E)}$$

Ecuación 3. Teorema de Bayes.

En la Figura 25 se muestran las probabilidades de la red tras descubrir que el aspersor no puede funcionar a causa de que se rompió la llave del agua del jardín. Por consiguiente, teniendo en cuenta la evidencia previa, la probabilidad de que tras unas horas o mañana el césped esté mojado es mucho menor a la inicial. Por tanto, independientemente de la Red Bayesina, el conocimiento previo de información que afecte al estado de uno o varios nodos modifica las probabilidades de sus nodos descendientes.

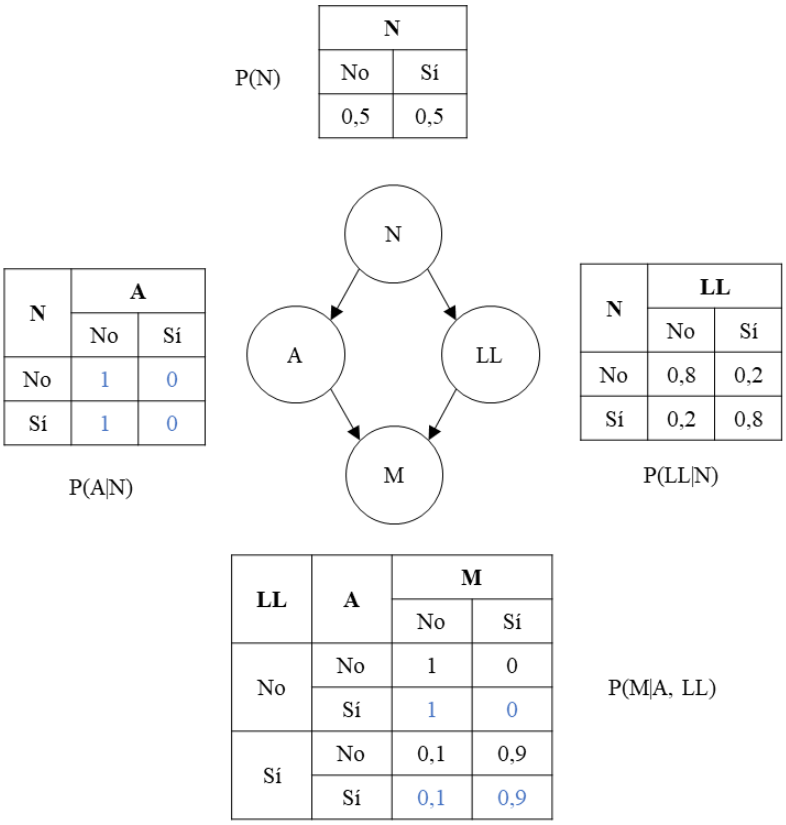

Figura 25. Fenómeno de inferencia en el ejemplo de la Figura 24.

2. Implementación de las Redes Bayesianas

Una vez detallado el marco teórico de las Redes Bayesianas, qué son y su funcionamiento, en este apartado se indica cómo se crea esta técnica de aprendizaje que utilizarán los Sistemas de Recomendación de este proyecto. Se han de desarrollar dos Redes Bayesianas distintas en función de las características de los *datasets* principales obtenidos en el CAPÍTULO IV.

De forma inicial, se ha decidido implementar en un módulo aparte estas redes con la librería gratuita de Python pyAgrum y la ayuda de la herramienta Google Colab. Esta

acción se realiza con el objetivo de realizar un estudio previo de cómo afectan los datos de entrenamiento y validación a la precisión de las RB antes de integrarlas en los Sistemas de Recomendación. Se pretende averiguar si los *datasets* principales obtenidos en el CAPÍTULO IV y que están descritos en el ANEXO I necesitan ser modificados o preprocesados antes de su empleo por las Redes Bayesianas para obtener la máxima precisión posible.

2.1. Red Bayesiana de la categoría "general"

El objetivo principal de la Red Bayesiana de esta categoría es obtener la probabilidad de que un paciente sufra una enfermedad a partir de una serie de síntomas. Para conseguirlo, se parte *dataset* principal de esta categoría, cuyo nombre es 160237006.

El *dataset* cuenta con un total de 42 enfermedades distribuidas en 120 casos cada una, excepto COVID-19 que cuenta con 2.414 casos. Asimismo, los casos en los que se desconoce la enfermedad del paciente suman un total de 1.304. Respecto a los síntomas de enfermedad que representan las columnas, en la Figura 26 se puede observar la frecuencia de aparición de estos para los múltiples pacientes, siendo los síntomas de la izquierda de la figura los que afectan a un mayor número de personas.

Inicialmente se divide el *dataset* en dos archivos, uno para entrenamiento y otro para validación de la red. El criterio para la división se basa en que los datos de validación sean el 25% del *dataset* inicial y que en ambos archivos exista la misma proporción de casos para conseguir mejores resultados. Tras la operación de división se obtienen dos archivos con las características mostradas en la Tabla 13.

ARCHIVO	TAMAÑO	FILAS	COLUMNAS	CASOS
Entrenamiento	6,63 MB	6.478	127	90 cada enfermedad Excepción: - COVID-19: 1.810 - Desconocida: 978
Validación	2,21 MB	2.160	127	30 cada enfermedad Excepción: - COVID-19: 604 - Desconocida: 326

Tabla 13. División del dataset 160237006 en datos de entrenamiento y validación.

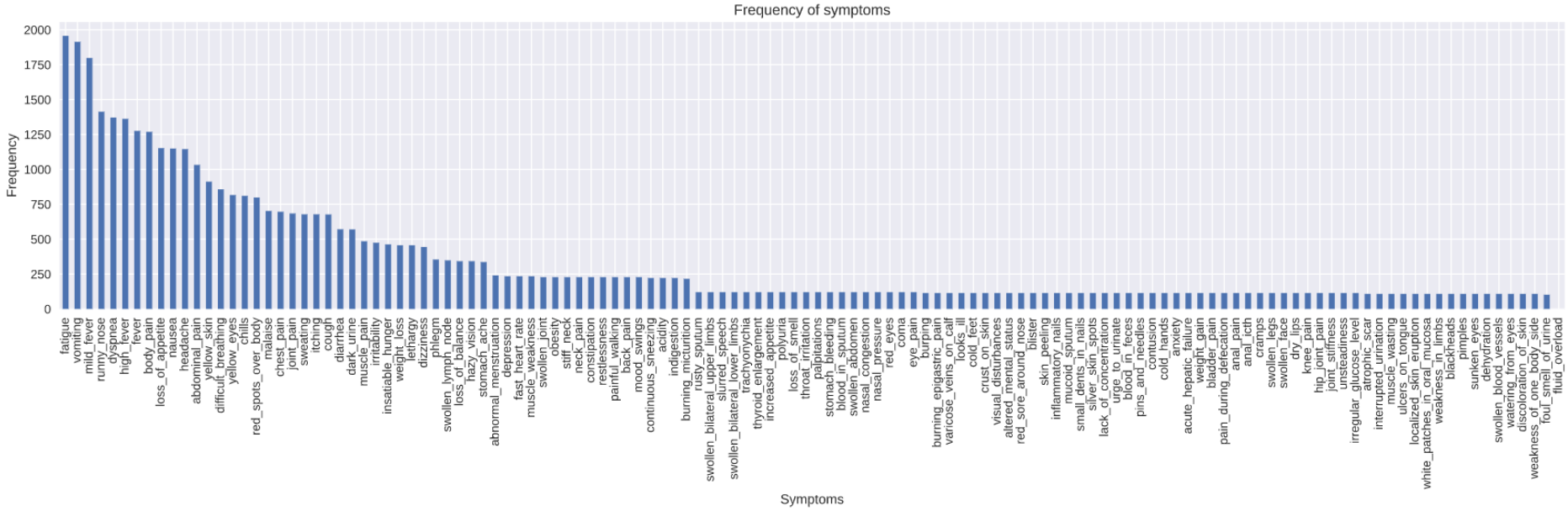

Figura 26. Frecuencia de los síntomas en el dataset 160237006.

Una vez obtenidos los archivos con los que se va a trabajar, para crear la red se emplea el método de pyAgrum *BayesNet()*. Este comando inicializa una Red Bayesiana vacía, es decir, sin nodos ni arcos.

A continuación, para crear la estructura de la red hay que crear los nodos y los arcos entre ellos. Para crear los nodos, en este caso, se emplean las funciones:

- *RangeVariable(aName, aDesc=‘’,minVal, maxVal)*. Crea un nodo cuyos valores son números enteros. En este método “*aName*” es el nombre del nodo, “*aDesc*” es su descripción y “*minVal” y “maxVal”* son sus valores mínimo y máximo respectivamente.
- *LabelizedVariable(aName, aDesc=‘’, labels)*. “*aName*” es el nombre del nodo, “*aDesc*” su descripción y “*labels*” es una lista con los posibles valores del nodo.

Para la creación de los nodos, se emplea el método *RangeVarible* con las columnas del archivo de entrenamiento correspondientes a los síntomas, siendo el nombre y descripción del nodo el nombre de la columna y los valores mínimo y máximo los valores cero y uno, que representan si se tiene o no dicho síntoma. Con *LabelizedVariable* se crea el nodo de diagnóstico cuyos valores son las 42 enfermedades indicadas en la Tabla 27 de este documento y la etiqueta de enfermedad desconocida.

Respecto a la relación entre nodos, lo ideal es que los 126 nodos correspondientes a los síntomas apunten al nodo de diagnóstico. Sin embargo, pyAgrum no soporta que un nodo tenga más de 15 nodos padre por lo que la estructura se realiza de forma inversa indicando los arcos con el comando *addArc(parentNode, node)* que dispone la Red Bayesiana. El resultado es la red de la Figura 27 con un total de 127 nodos y 126 arcos que relacionan el diagnóstico con los síntomas de una enfermedad.

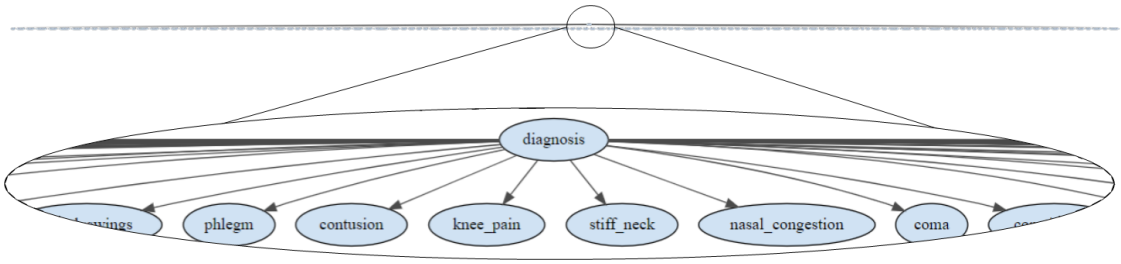

Figura 27. Estructura de la RB symptoms_BN.

El siguiente paso es que la red aprenda de los datos para obtener un modelo de predicción. pyAgrum tiene una característica propia y es que almacena todos los procesos

de aprendizaje de una Red Bayesiana en una clase simple denominada *BNLearner*. En ella se encuentra el algoritmo de aprendizaje y parámetros como la prioridad o las restricciones, entre otros. Para crear esta clase, pyAgrum dispone el método *BNLearner(filename, src, missingSymbols)*, donde "*filename*" indica la ruta al archivo de datos de entrenamiento, "*src"* corresponde con la Red Bayesiana para la que se quiere crear la clase, puede tener toda la estructura definida o sólo los nodos, y "*missingSymbols*" es una lista con los caracteres de texto que se han empleado para indicar datos desconocidos dentro del archivo, en este caso, "-1".

Por otro lado, antes de que la red aprenda, a causa de que se disponen de valores desconocidos, hay que indicarle a la clase *BNLearner* que emplee el algoritmo *Hope-Maximization* (HM) y la técnica *Laplace Smoothing* mediante los comandos *useEM(epsilon)* y *useAprioriSmoothing()*. El algoritmo HM [47] es una técnica iterativa general que estima un conjunto de variables que describen una distribución de probabilidad subyacente dada únicamente la parte observada de los datos completos producidos por la distribución, mientras que técnica *Laplace Smoothing* [48] evita que ninguna probabilidad sea completamente cero. En el caso de esta última técnica se ha decidido que el valor incorporado sea de 0,001.

Tras las operaciones realizadas, se obtiene una clase *BNLearner* con las características indicadas en la Figura 28. El algoritmo de aprendizaje por defecto de pyAgrum es el más empleado para el aprendizaje de Redes Bayesianas, el *Greedy Hill Climbing* [49]. Es una técnica iterativa que trata de encontrar una solución óptima, o cercana a la óptima, a un problema partiendo de un valor arbitrario.

```
Filename       : /content/drive/MyDrive/Colab Notebooks/MEDvolution_v2/symptoms_mvu_tr.csv
Size           : (6478,127)
Variables      : diagnosis[43], body_pain[2], difficult_breathing[2], dyspnea[2], fever[2], high_fever[2
Induced types  : False
Missing values : True
Algorithm      : Greedy Hill Climbing
Score          : BDeu
Prior          : Smoothing  (The BDeu score already contains a different 'implicit' apriori. Therefore,
Prior weight   : 1.000000
EM             : True
EM epsilon     : 0.001000
```

Figura 28. Características de la clase symtoms_BNLearner.

Con el objetivo de obtener el modelo de la red se emplea el método *learnParameters(dag)*, pasándole como parámetro de la función la estructura de la Red Bayesiana creada para que comience el entrenamiento de la red. Después de

aproximadamente 100 iteraciones, se completa el proceso de aprendizaje y se obtiene el modelo de la red con el que se va a trabajar. También se puede observar el nivel de entropía de la red (Figura 29), siendo el nodo “diagnosis” el que más entropía presenta con un 4,45 al tener más posibles valores.

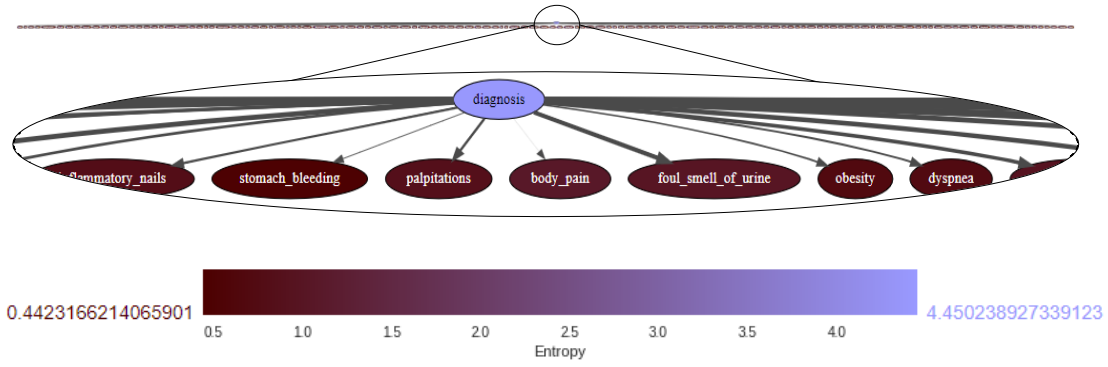

Figura 29. Entropía de la RB symptoms_BN.

Para evaluar el correcto funcionamiento de la red se ha diseñado la función *showPosterior*, detallada en el ANEXO II. Esta función crea la inferencia y genera una predicción para un parámetro de la red a partir de unas evidencias si las hay o no. Como ejemplo para la comprobación, se ha seleccionado un paciente que padece disnea, febrícula y fatiga. Tal y como se puede observar en la Figura 30, el sistema indica con un 55,69% de probabilidad que no conoce la enfermedad, sin embargo, también muestra que el paciente tiene un 41,17% de probabilidades de padecer la enfermedad llamada Tuberculosis.

```
showPosterior(symptoms_Model,evs={"dyspnea":1, "mild_fever":1, "fatigue":1},
              target='diagnosis')
```

```
unknown : 55.69 %
covid_19 : 1.78 %
tuberculosis : 41.17 %
```

Figura 30. Comprobación del funcionamiento del modelo de la red symptoms_BN.

Ahora que se sabe que el sistema funciona, queda comprobar la exactitud de las predicciones que realiza. Con este objetivo se ha diseñado la función *systemEvaluation*, la cual se describe en el ANEXO II y cuyo diagrama de flujo se puede apreciar en la Figura 31.

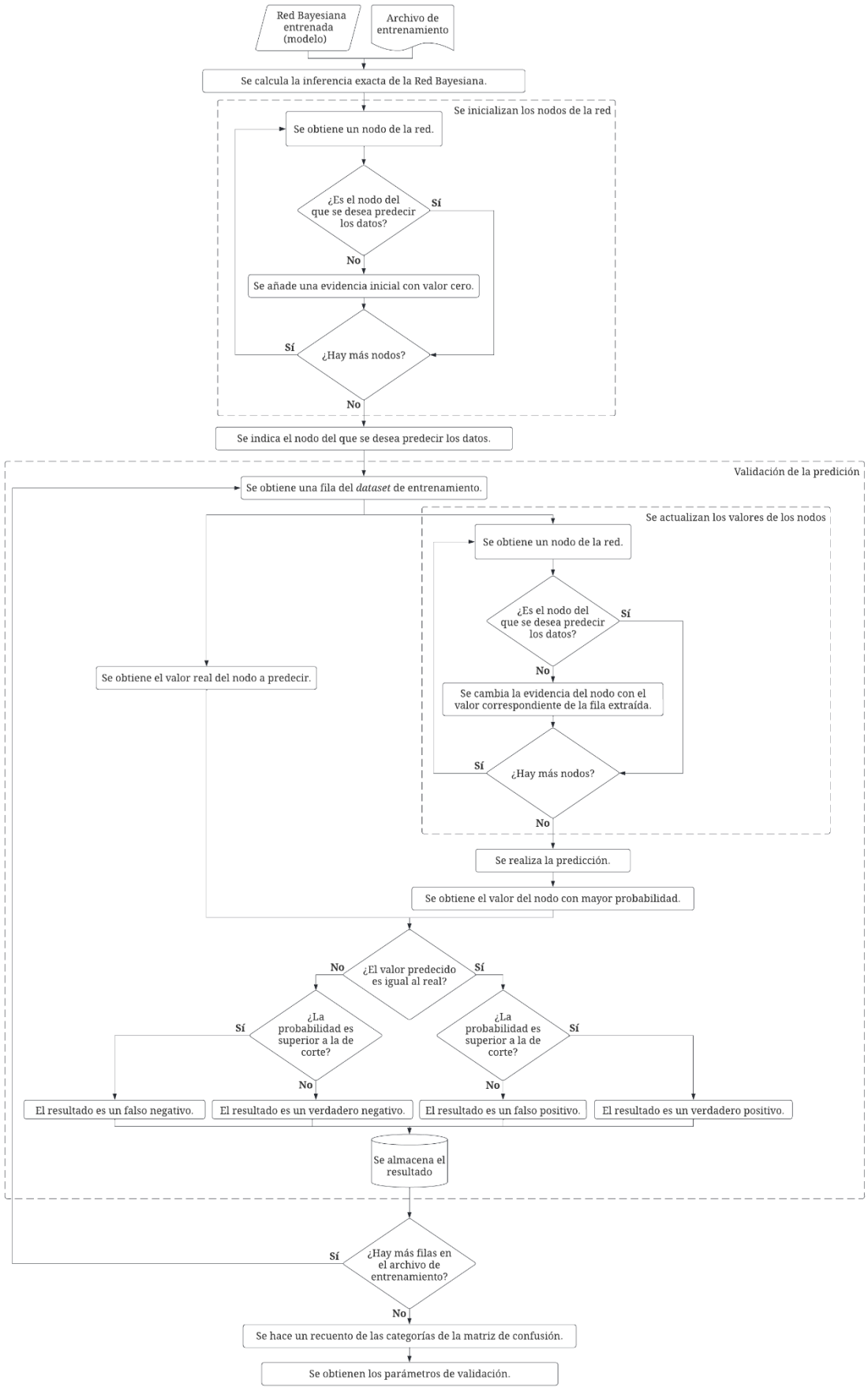

Figura 31. Diagrama de flujo de la función systemEvaluation.

En primer lugar, en el método *systemEvaluation* se obtienen los valores iniciales de los nodos de la red. Luego, por cada fila del *dataset* de validación se cambian los valores de los nodos por los existentes en el *dataset* (como si fueran las evidencias), se calcula la inferencia y se obtiene una predicción para un nodo de la red, en este caso "diagnosis". De la predicción se selecciona aquel valor con mayor probabilidad y se comprueba si el valor predicho es el mismo que se encuentra en la columna diagnosis del *dataset* y con cuanta exactitud para asignarle un lugar dentro de la matriz de confusión. Por último, el método hace un recuento de los datos de las diferentes categorías de la matriz: verdaderos y falsos positivos y, verdaderos y falsos negativos, y proporciona el valor de los siguientes parámetros de validación [50] de un modelo:

- Sensibilidad. Es la probabilidad de que el modelo prediga de forma correcta el valor existente en los datos.

$$Sensitivity = \frac{TP}{TP + FN}$$

Donde:

- TP (del inglés *True Positive*). Es el número de verdaderos positivos.
- FN (del inglés *False Negative*). Es el número de falsos negativos.

Ecuación 4. Sensibilidad de un sistema.

- Especificidad. Es la probabilidad de que el modelo prediga de forma correcta que el valor no es el existente en los datos.

$$Specificity = \frac{TN}{TN + FP}$$

Donde:

- TN (del inglés *True Negative*). Es el número de verdaderos negativos.
- FP (del inglés *False Positive*). Es el número de falsos positivos.

Ecuación 5. Especificidad de un sistema.

- Precisión. Es la relación de verdaderos positivos frente al total de positivos obtenidos, es decir, indica el grado de exactitud de predicción del modelo.

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

Ecuación 6. Precisión de un sistema.

El resultado de la ejecución de la función *systemEvaluation* se ilustra en la Figura 32. Como se puede ver, en la matriz de confusión del sistema se contaría con un ratio de 1.230 verdaderos positivos, 930 falsos positivos y, 0 verdaderos y falsos negativos. Estos datos concluyen que el sistema cuenta con una precisión en la predicción del 56,94%, un valor deficiente.

```
Confusion matrix, category count:
 - True positives:  1230
 - False positives:  930
 - True negatives:  0
 - False negatives:  0

Sensitivity = 1.000000
Specifity  = 0.000000
Precision  = 0.569444
```

Figura 32. Resultado de la validación de la Red Bayesiana symptoms_BN.

2.2. Red Bayesiana de la categoría “covid19”

El objetivo principal de esta Red Bayesiana es proporcionar datos que sean relevantes en la evolución del cuadro médico de un paciente de COVID-19. Para ello, se parte de un del *dataset* principal de esta categoría, cuyo nombre es 840539006.

Este *dataset* contiene 28 parámetros diferentes que se explican en la Tabla 28 situada en el ANEXO I. De todas las variables se han seleccionado cuatro que son de suma importancia en la evolución del paciente y cuya predicción indicará lo siguiente:

- “*icu_transfer*”. Indica si el paciente deberá o no ser transferido a la Unidad de Cuidados Intensivos del centro hospitalario donde se encuentre.
- “*not_intubated*”. Informa si el paciente no tendrá que ser intubado o sí debido a complicaciones en la enfermedad.
- “*deceased*” y “*survival_days*”. Proporcionan información respecto a la esperanza de vida del paciente, ya que las primeras cepas de COVID-19 cuentan con una alta tasa de mortalidad.

Se divide el *dataset* en dos archivos, uno para entrenamiento y otro para validación de la red. El criterio para la división se basa en que los datos de validación sean el 25% del *dataset* inicial y que en ambos archivos exista la misma proporción de personas fallecidas, no fallecidas y cuyo estado de supervivencia a la enfermedad es desconocido para conseguir mejores resultados. Tras la operación de división se obtienen dos archivos con las características mostradas en la Tabla 14.

ARCHIVO	TAMAÑO	FILAS	COLUMNAS	PACIENTES
Entrenamiento	38,81 MB	167.303	28	Fallecidos: 166.466 No fallecidos: 239 Desconocido: 618
Validación	12,94 MB	55.768	28	Fallecidos: 55.482 No fallecidos: 206 Desconocido: 80

Tabla 14. División del 840539006 en datos de entrenamiento y validación.

Ahora que se dispone de los archivos con los que se va a trabajar, se crea la red vacía con el método *BayesNet*. Posteriormente, se crean los nodos que van a conformar la estructura con las funciones *RangeVariable* y *LabelizedVariable*. Sin embargo, no es posible conocer la estructura de la red, por lo que no se pueden crear los arcos entre los nodos de forma manual.

pyAgrum da la posibilidad de crear los arcos de una red y, por ende, conocer la estructura de una Red Bayesiana junto con el resto de sus parámetros a partir del análisis de los datos de entrenamiento con la función *learnBN()*. No obstante, el único requisito para el uso de este método es que los datos de entrenamiento deben estar completos, es decir, no puede haber datos de columnas desconocidas.

En este caso, es imposible emplear la función *learnBN* debido a que hay datos desconocidos en el archivo de entrenamiento. Por tanto, al no poder crearse y analizarse la red a partir de los datos existentes, no es posible obtener unos resultados.

3. Análisis de los resultados

Tras la implementación de las Redes bayesianas y su entrenamiento con los *datasets* principales se han obtenido resultados de precisión deficientes. En el caso de la red de la categoría “general”, cuenta con un porcentaje de buenas predicciones de 56,94%, mientras que la red de la categoría “covid19” cuenta con datos inconcluyentes a causa de que no se ha podido ni crear correctamente la red.

Con el objetivo de mejorar estos resultados, se eliminan de los *datasets* los valores desconocidos y, en el caso del *dataset* 160237006 se suprimen las filas de pacientes cuya enfermedad se desconoce. Con estos cambios se pretende disminuir la incertidumbre al obtener probabilidades de predicción de los nodos de la red.

3.1. Modificaciones

Para la eliminación de los valores desconocidos en los *datasets* se ha optado por el método de imputación de valores debido a que eliminar las filas que contienen datos desconocidos no es una opción. Si se eliminaran las filas, casi la mitad de los datos recolectados inicialmente no existirían ya que los *datasets* principales son la unión de múltiples *datasets* diferentes que no tienen por qué tener las mismas características. Asimismo, en el caso del *dataset* de la categoría "general" también se ha optado por la eliminación de las filas de pacientes cuya enfermedad se desconoce para reducir el índice de incertidumbre del sistema.

El método de imputación empleado en ambos *datasets* es la sustitución por constante [51], que consiste en reemplazar los valores desconocidos por constantes cuyo valor viene determinado por razones teóricas o derivado de una investigación previa. A continuación, se indicarán las razones que justifican el reemplazo de los valores desconocidos en cada *dataset*.

Por un lado, para el *dataset* principal de la categoría "general", la suposición realizada es que los síntomas de los que se desconoce su valor son debido a que en el momento de la recogida de datos el paciente no los padecía. Por tanto, si no se recoge su valor es que no eran relevantes para el diagnóstico de la enfermedad en dicho momento. Entonces, a los síntomas cuyo valor es desconocido se les asigna el valor cero indicando que el paciente no los padecía.

Por otro lado, para el *dataset* principal de la categoría "covid19", las suposiciones realizadas son las siguientes:

- Cuando se desconoce el género del paciente, a partir de los datos disponibles no es posible asumir su estructura corporal. Por consiguiente, se ha optado por sustituir el valor "-1" por el identificador de SNOMED CT 394743007 que hace referencia a "género desconocido".
- Si el tipo de paciente es desconocido, se asume que el paciente es de ambulatorio, debido a que si fuera un paciente internado en un hospital este valor se especificaría.
- Aquellos pacientes de los cuales se desconoce si han fallecido o no, se asume que han pasado la enfermedad con éxito, ya que en caso contrario existiría documentación respecto a la defunción del paciente. Asimismo, en la columna del

tiempo de supervivencia, al asumir que el paciente ha pasado la enfermedad con éxito se asigna el valor 999, ya que si no fuera el caso se especificarían los días entre la fecha de aparición de los primeros síntomas y la de fallecimiento.

- Cuando se desconoce si el paciente ha sido intubado o no, inicialmente se asume que no ha sido intubado. En caso contrario, existiría constancia del procedimiento de intubación.
- Respecto al resto de columnas del *dataset*, los datos desconocidos se asumen como el valor negativo de esta, en este caso cero. Esto es a causa de que si el estado de embarazo del paciente o la presencia de uno o más trastornos, además de la enfermedad, entre otros, fueran relevantes para el diagnóstico de la enfermedad, en el momento de la recogida de datos se hubiera indicado.

El resultado de la aplicación del método de imputación ha supuesto la creación de cuatro variantes del *dataset* principal de la categoría "general" y dos variantes del de la categoría "covid19". Como se puede apreciar en la Tabla 15, el sufijo en el nombre de cada dataset nos indica si el archivo contiene datos desconocidos (mv, del inglés *missing values*) o la categoría de enfermedad desconocida (u, del inglés *unkown disease*).

NOMBRE	TAMAÑO	FILAS	COLUMNAS	DETALLE
160237006_mvu	8,77 MB	8.638	127	Contiene datos desconocidos y pacientes de los cuales se desconoce la enfermedad que padecen.
160237006_u				No contiene datos desconocidos, pero sí pacientes de los cuales se desconoce la enfermedad que padecen.
160237006_mv	7,45 MB	7.334		Contiene datos desconocidos, pero no pacientes de los cuales se desconoce la enfermedad que padecen.
160237006				No contiene datos desconocidos ni pacientes de los cuales se desconoce la enfermedad que padecen.
840539006_mv	49,97 MB	223.071	28	Contiene datos desconocidos
840539006				No contiene datos desconocidos.

Tabla 15. Características de las variantes de los datasets principales.

A continuación, se entrenarán las dos redes diseñadas con cada uno de los *datasets* de la Tabla 15 con el objetivo de comprobar si se obtienen mejores resultados.

3.2. Nuevos resultados

En esta sección se analizan los datos obtenidos tras el cambio realizado en los diferentes *datasets* principales y el entrenamiento de las Redes Bayesianas.

3.2.1. Red Bayesiana de la categoría “general”

Tras el proceso de entrenamiento de la red con los cuatro *datasets* diferentes, se observan mejoras ya que se ha comprobado la funcionalidad del modelo con el mismo ejemplo que en la Figura 30. Los múltiples resultados se muestran en la Figura 33 y como se puede observar, cambian drásticamente. Esto es consecuencia de la eliminación de la categoría de enfermedad desconocida y los parámetros desconocidos, que incorporaban demasiada incertidumbre en la red. Por tanto, se ha pasado de que el paciente tenga 41,17% de probabilidad de tener tuberculosis a más del 90%.

```
-> With missing values = -1 and unknown disease.

unknown : 55.69 %
covid_19 : 1.78 %
tuberculosis : 41.17 %


-> Without missing values and with unknown disease.

tuberculosis : 99.01 %


-> With missing values = -1 and without unknown disease.

covid_19 : 4.01 %
tuberculosis : 92.89 %


-> Without missing values and with unknown disease.

tuberculosis : 99.01 %
```

Figura 33. Comparación del funcionamiento de la RB con diferentes datasets de entrenamiento.

Respecto a la validación de la red entrenada con cada uno de los *datasets* se obtienen los resultados reflejados en la Tabla 16. Como se puede apreciar, efectivamente hay una mejora significativa en la precisión del modelo. Al eliminar las filas de pacientes con enfermedad desconocida y al eliminar los valores desconocidos, la red puede identificar cuáles son los síntomas característicos de cada enfermedad, elevando la precisión del sistema al 100% con una sensibilidad del 98,36%.

	VERDADEROS POSITIVOS	FALSOS POSITIVOS	VERDADEROS NEGATIVOS	FALSOS NEGATIVOS

160237006_mvu	1.230	930	0	0
160237006_u	1.657	114	177	212
160237006_mv	1.200	600	4	30
160237006	1.804	0	0	30

	SENSIBILIDAD	**ESPECIFICIDAD**	**PRECISIÓN**
160237006_mvu	1	0	0,5694
160237006_u	0,8866	0,6083	0,9356
160237006_mv	0,9756	0,0066	0,6667
160237006	0,9836	?	1

Tabla 16. Resultados de la validación de la RB symptoms_BN para los diferentes datasets.

3.2.2. Red Bayesiana de la categoría “covid19”

En cuanto a esta red, la eliminación de los valores desconocidos ha supuesto la obtención de la estructura y parámetros de la red a través de los datos. Dicha estructura obtenida junto con su entropía tras la ejecución del comando *learnBN* se muestra en la Figura 35.

Con relación al funcionamiento de la red, como ejemplo para la comprobación, se ha seleccionado un paciente de 61 años con COVID-19 y se quiere saber el tiempo de supervivencia a la enfermedad que se le estima. Tal y como se puede observar en la Figura 34, el sistema indica que el paciente tiene un 79,48 % de probabilidades de superar la enfermedad con éxito.

```
showPosterior(covid19Model,evs={'age':61}, target='survival_days')
```

```
4 : 1.06 %
5 : 1.26 %
6 : 1.28 %
7 : 1.45 %
8 : 1.43 %
9 : 1.36 %
10 : 1.31 %
11 : 1.11 %
12 : 1.13 %
999 : 79.48 %
```

Figura 34. Comprobación del funcionamiento del modelo de la red covid19_BN.

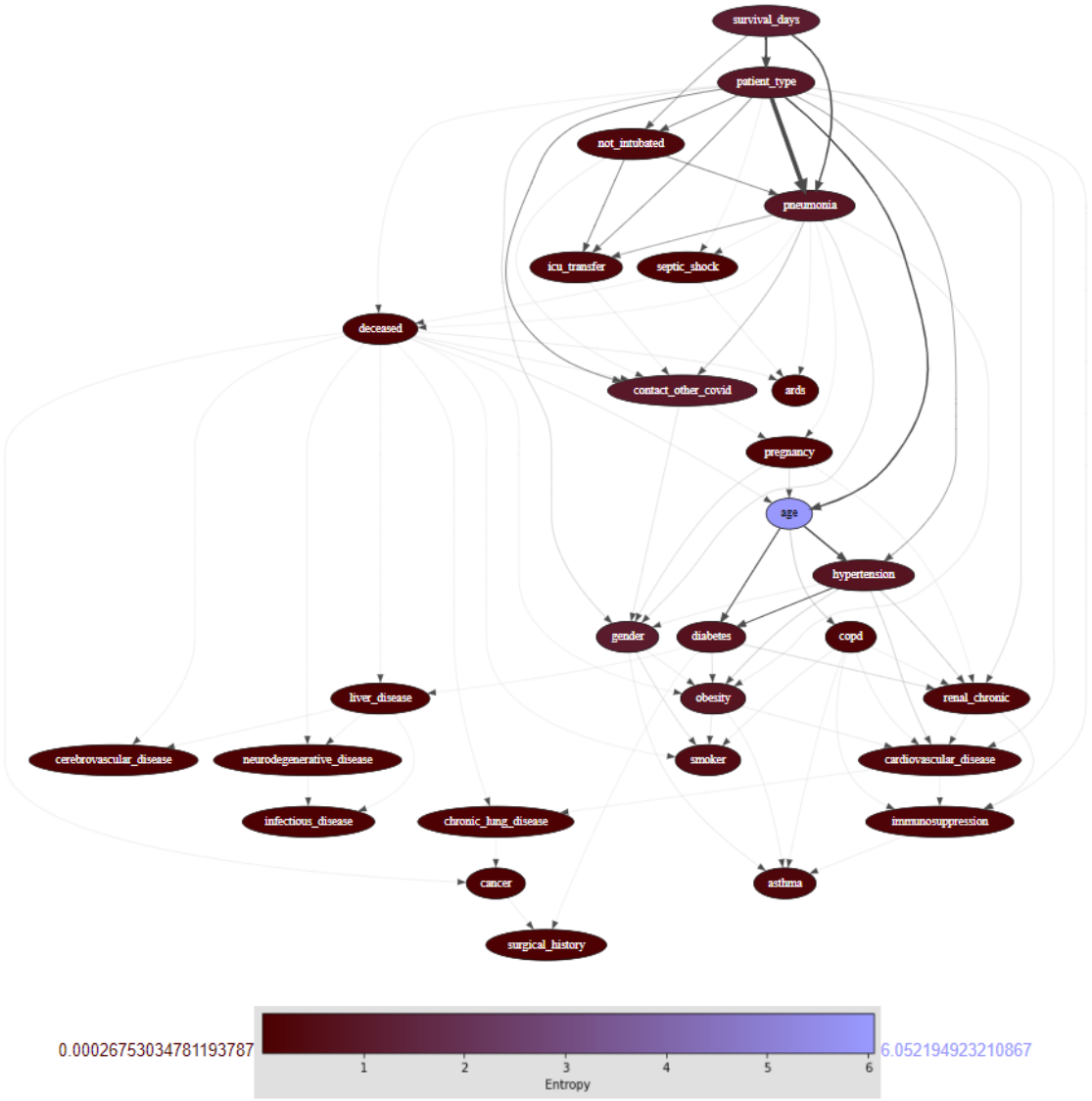

Figura 35. Entropía de la RB covid19_BN.

El resultado de la ejecución de la función *systemEvaluation* se ilustra en la Figura 36. Como se puede ver, los valores de precisión para las columnas son: 57,71% para la que indica si el paciente es o no intubado, 98,42% para la columna que representa si el paciente requiere ser transferido a la UCI, 99,89% la que indica la probabilidad del paciente de fallecer y 0% para el tiempo de supervivencia.

Exceptuando la primera y la última de las columnas, los valores obtenidos son bastante buenos. Sin embargo, hay que mejorar el valor de las restantes y por ello se propone eliminar la uniformidad en los valores de las columnas correspondientes a la edad y los días de supervivencia discretizándolos. Para lograrlo, se desarrollan las funciones *forAges*() y *forDays*(), cuyo objetivo es modificar el valor de la respectiva columna

incorporándola en un rango de diez escalones. Por ejemplo, si un paciente tiene 61 años, en la columna de edad se sustituye ese valor por “60-69”.

```
-> not_intubated:
Confusion matrix, category count:
 - True positives:  741
 - False positives:  543
 - True negatives:  824
 - False negatives:  53660

Sensitivity = 0.013621
Specifity  = 0.602780
Precision  = 0.577103

-> icu_transfer:
Confusion matrix, category count:
 - True positives:  53751
 - False positives:  861
 - True negatives:  579
 - False negatives:  577

Sensitivity = 0.989379
Specifity  = 0.402083
Precision  = 0.984234

-> deceased:
Confusion matrix, category count:
 - True positives:  55674
 - False positives:  63
 - True negatives:  14
 - False negatives:  17

Sensitivity = 0.999695
Specifity  = 0.181818
Precision  = 0.998870

-> survival_days:
Confusion matrix, category count:
 - True positives:  0
 - False positives:  55768
 - True negatives:  0
 - False negatives:  0

Specifity  = 0.000000
Precision  = 0.000000
```

Figura 36. Resultado de la validación de la Red Bayesiana covid19_BN.

Una vez modificadas las columnas de edad y días de supervivencia se obtiene un nuevo *dataset* cuyo nombre es 840539006_p, donde la “p” proviene del inglés *Preprocessed Data*. La red creada tras la ejecución del comando *BNlearn*, entrenando la red con el nuevo *dataset*, es la mostrada en la Figura 37. Aunque el nodo con mayor entropía sigue siendo el de la edad, la relación entre los nodos es diferente a la mostrada en la Figura 35.

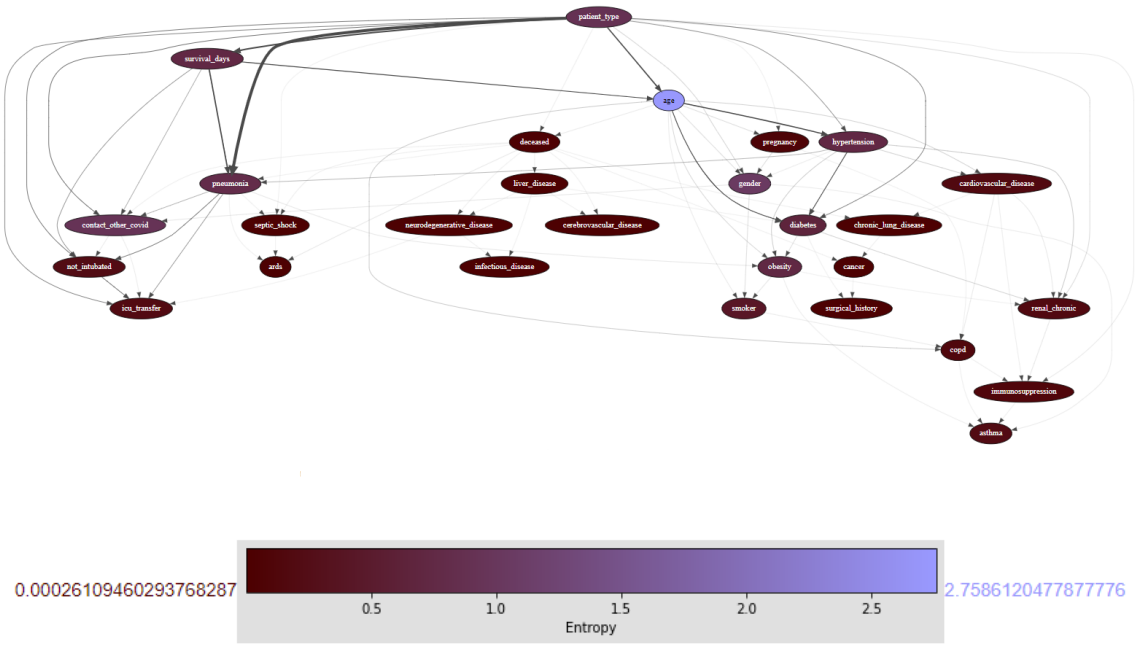

Figura 37.Entropía de la RB covid19_BN con datos preprocesados.

Después del proceso de entrenamiento con el *dataset* 840539006_p se comprueba el funcionamiento de la red con el mismo ejemplo que en la Figura 34, obteniendo mejores resultados. Estos se muestran en la Figura 38. Se pasa de tener diez resultados diferentes que hacían referencia a un día en concreto, a cuatro con una probabilidad mayor y que representan un rango de días.

```
-> Without missing values.

4 : 1.06 %
5 : 1.26 %
6 : 1.28 %
7 : 1.45 %
8 : 1.43 %
9 : 1.36 %
10 : 1.31 %
11 : 1.11 %
12 : 1.13 %
999 : 79.48 %

----------------------------------------------------------------------
-> Without missing values with processed data.

survive : 71.91 %
0-9 : 12.57 %
10-19 : 11.64 %
20-29 : 3.07 %
```

Figura 38. Comparación del funcionamiento de la RB con diferentes datasets de entrenamiento.

Aunque el sistema estime que el paciente de 61 años tiene altas probabilidades de superar el COVID-19 con éxito, también hay que considerar que estima que en los 20 primeros días desde el primer síntoma hay riesgo de muerte con más de un 10%. Este dato es muy útil a nivel médico para llevar una supervisión del paciente y prestar atención a su evolución.

Se realiza la validación de la red entrenada con cada uno de los datasets, obteniendo los resultados reflejados en la Tabla 17. Como se puede apreciar, efectivamente hay una mejora significativa en la precisión del modelo para el nodo "*survival_days*" debido a la discretización de sus valores. Asimismo, hay una mejora en la precisión del nodo "*not_intubated*" derivada de su relación con el nodo "*survival_days*", a consecuencia de que este último en la nueva estructura es uno de sus nodos previos.

Con este último *dataset*, los valores de precisión para las columnas son: 75,32% para la que indica si el paciente es o no intubado, 98,42% para la columna que representa si el paciente requiere ser transferido a la UCI, 99,89% para la que indica la probabilidad del paciente de fallecer y 88,21% para el tiempo de supervivencia.

	DATASET	VERDADEROS POSITIVOS	FALSOS POSITIVOS	VERDADEROS NEGATIVOS	FALSOS NEGATIVOS
not_intubated	840539006_mv	?	?	?	?
	840539006	741	543	824	53.660
	840539006_p	528	173	1.037	54.030
icu_transfer	840539006_mv	?	?	?	?
	840539006	53.751	861	579	577
	840539006_p	53.751	861	579	577
deseased	840539006_mv	?	?	?	?
	840539006	55.674	63	14	17
	840539006_p	55.685	61	3	19
survival_days	840539006_mv	?	?	?	?
	840539006	0	55.768	0	0
	840539006_p	49.195	6.573	0	0

COLUMNA	DATASET	VERDADEROS POSITIVOS	FALSOS POSITIVOS	VERDADEROS NEGATIVOS
not_intubated	840539006_mv	?	?	?
	840539006	0,0136	0,6028	0,5771
	840539006_p	0,0097	0,8570	0,7532
icu_transfer	840539006_mv	?	?	?
	840539006	0,9894	0,4021	0,9842
	840539006_p	0,9894	0,4021	0,9842
deseased	840539006_mv	?	?	?
	840539006	0,9997	0,1818	0,9989
	840539006_p	0,9997	0,0469	0,9989
survival_days	840539006_mv	?	?	?
	840539006	?	0	0
	840539006_p	1	0	0,8821

Tabla 17. Resultados de la validación de la RB covid19_BN para los diferentes dataset.

4. Resumen

En este capítulo se han obtenido, siguiendo el procedimiento ilustrado en la Figura 39, las dos Redes Bayesianas que serán el núcleo de los Sistemas de Recomendación para el proyecto.

Por un lado, está la RB de la categoría "general" que proporciona la probabilidad de que un paciente padezca una enfermedad a partir de unos síntomas. Por otro lado, la Red Bayesiana de la categoría "covid19" proporciona la probabilidad de que un paciente de COVID-19: tenga o no que ser intubado, tenga o no que ser transferido a la Unidad de Cuidados Intensivos o pueda fallecer a causa de la enfermedad y con cuantos días se estima.

A continuación, en la Tabla 18 se indican los parámetros de sensibilidad, especificad y precisión para las redes según la columna que se valide. Estos datos han sido el resultado de modificaciones realizadas a los *datasets* principales de CAPÍTULO IV con el objetivo de obtener los mejores resultados. El *dataset* principal final de la categoría "general" no presenta valores desconocidos y se han eliminado los pacientes de los cuales se desconoce la enfermedad que padecen. Para la categoría específica "covid19", el *dataset* final no presenta valores desconocidos y sus columnas que contenían valores uniformes se han discretizado mediante un preprocesado de datos.

***DATASET* DE ENTRENAMIENTO**	**COLUMNA**	**SENSIBILIDAD**	**ESPECIFICIDAD**	**PRECISIÓN**
160237006	*diagnosis*	98,36%	?	100%
840539006_p	*not_intubated*	0,97%	85,70%	75,32%
	icu_transfer	98,9%	40,21%	98,42%
	deceased	99,97%	4,69%	99,89%
	survival_days	100%	0%	88,21%

Tabla 18. Características finales de las RB symptoms_BN y covid19BN.

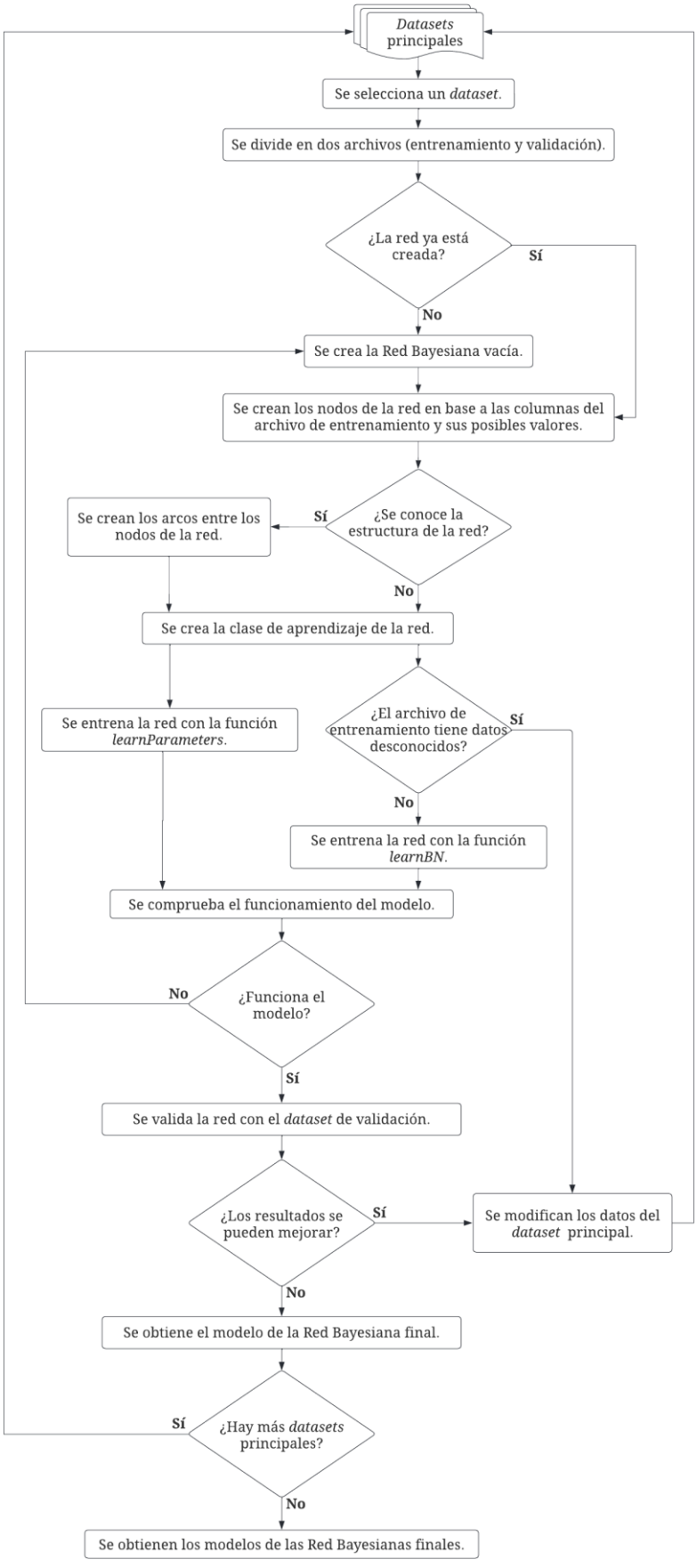

Figura 39. Proceso de obtención de los modelos de RB.

CAPÍTULO VI. RECOMENDADORES

En este capítulo se muestra la implementación completa de los dos Sistemas de Recomendación empleados en este proyecto.

El objetivo principal de este proyecto es ofrecer una herramienta de ayuda a los profesionales médicos para predecir la evolución de un paciente. Para ello, se propuso como eje central el desarrollo de dos Sistemas de Recomendación.

1. Implementación de los Sistemas de Recomendación

El desarrollo de un SR tiene dos grandes partes. La primera es la técnica de aprendizaje que utiliza y los datos con los que se entrena, este caso, las dos RB desarrolladas en el CAPÍTULO V. La segunda parte son los datos que se recomiendan al usuario del sistema.

1.1. Sistema de Recomendación de la categoría "general"

Para el sistema de esta categoría se propuso una recomendación para aquellas enfermedades que el paciente tiene más de un 20% de probabilidad para padecer.

Debido a la imposibilidad de tener datos de tratamientos de pacientes, se ha creado un *dataset* de recomendación para enfermedades. Este *dataset* cuenta con dos versiones, una inicial en la que se han extraído los datos para cada enfermedad de las correspondientes referencias; y una final, que es la que emplea el sistema, en la que se han mapeado los datos al estándar SNOMED CT con la ayuda del buscador proporcionado por SNOMED. Las características de ambas versiones se encuentran detalladas en el ANEXO III.

A modo de ejemplo, en la Tabla 19, se muestra un fragmento del *dataset* que refleja las tres columnas de tratamiento para la enfermedad denominada Asma. La parte superior refleja los datos obtenidos de *Mayo Clinic* [52], una entidad estadunidense sin ánimo de lucro dedicada a la práctica clínica, la educación y la investigación. La parte inferior de la tabla refleja el resultado después de mapear los datos superiores a SNOMED CT.

disease	**treatment1**	**treatment2**	**treatmet3**
Asthma	*Inhaled gas|Corticosteroid series|Montelukast|Zafirlukast|Product containing fluticasone and salmeterol|Theophylline|Salbutamol*	*Immunotherapy*	*Thermoplasty of bronchus*

disease	**treatment1**	**treatment2**	**treatmet3**
195967001	116177003|255877006|373728005|386880006|411106009|372810006|372897005	76334006	713348007

Tabla 19. Ejemplo de conversión de datos para el Asma a SNOMED CT.

Por consiguiente, la estructura final del Sistema de Recomendación de la categoría "general" es la reflejada en la Figura 40. En ella se puede apreciar las tres partes que componen un SR: los datos, la técnica de aprendizaje y la recomendación.

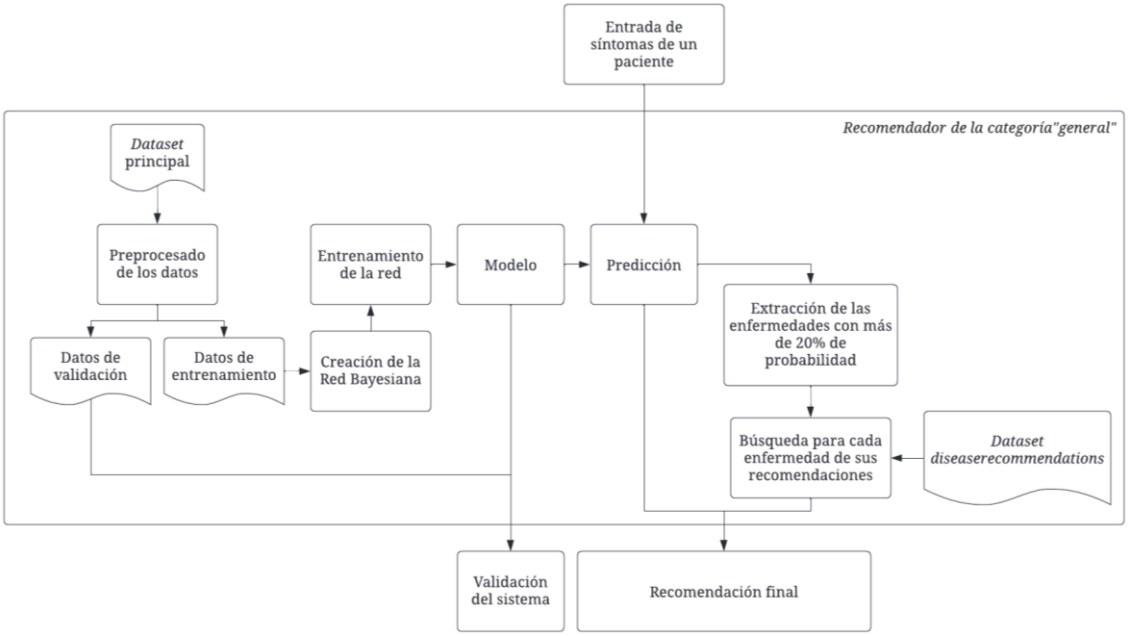

Figura 40. Estructura del recomendador de la categoría "general".

Respecto a su funcionamiento, siguiendo el ejemplo empleado desde un principio para la Red Bayesiana que forma parte de este SR, se solicita la recomendación para un paciente que padece febrícula, fatiga y disnea. El resultado de la petición se ilustra en la Figura 41.


```
{'predictionList': {'439401001': {'840539006': 1.01, '56717001': 98.99}}, '
recommendationList': {'439401001': {'56717001': {'diagnosticTest1': '84100007
421829000 5880005', 'diagnosticTest2': '424489006', 'diagnosticTest3': '396550006
', 'diagnosticTest4': '399208008|77477000', 'diagnosticTest5': '104173009', '
treatment1': '776412002|777438009|775855004|777364002|774747007|776544000', '
treatmentSideEffects': '16932000|79890006|18165001|39575007|424131007 420561004
131148009| 246636008', 'copingAndSupport': '390822007'}}}}
```


Figura 41. Recomendación para un paciente padece febrícula, fatiga y disnea.

La recomendación está compuesta por dos partes. La primera son las probabilidades obtenidas por la Red Bayesiana entrenada (verde) para el nodo de diagnosis (azul). La segunda parte es la de recomendación para las enfermedades que superen el 20% de probabilidad. En este caso, el sistema estima que el paciente padece tuberculosis (con SCTID: 56717001 y en naranja) con un 98,99% de probabilidad y, por consiguiente, ha extraído del *dataset diseaseRecommendations* los datos (amarillo) al respecto de esta enfermedad.

1.2. Sistema de Recomendación de la categoría “covid19”

El objetivo de este sistema es proporcionar información relevante para la evolución de un paciente de COVID-19. Por ello, de los datos disponibles se seleccionaron cuatro variables que cumplen dicha afirmación: “*not_intubated*”, “*icu_transfer*”, “*deceased*” y “*survival_days*”. Con ellas se puede saber si el paciente deberá o no ser intubado, si requiere una transferencia a la UCI o incluso si puede fallecer y en cuantos días.

Excepto la última de las variables, los valores que toman son cero o uno, que indican sí o no. Por consiguiente, las recomendaciones para estas variables se dan en el caso de que supere el resultado desfavorecedor de la variable el 20% de probabilidad. Cuando un paciente tiene más de un 20% de probabilidades de ser intubado (“*not_intubated = 0”*), de ser transferido a la Unidad de Cuidados Intensivos (“*icu_transfer = 1*”) o cuando el tiempo de supervivencia a la enfermedad resalta en un rango (“*survival_days = ‘10-19’*”) se recomienda al usuario mantener una vigilancia periódica sobre el paciente para tener el mayor margen de actuación posible. En el caso de que el paciente tenga una alta tasa de mortalidad se indica que el paciente mantenga a sus seres queridos cerca y asista a terapia psicológica.

Incorporando lo anterior, la estructura final del Sistema de Recomendación de la categoría “covid19” se ilustra en la Figura 42. En ella se puede apreciar las tres partes que componen un SR: los datos, la técnica de aprendizaje y la recomendación.

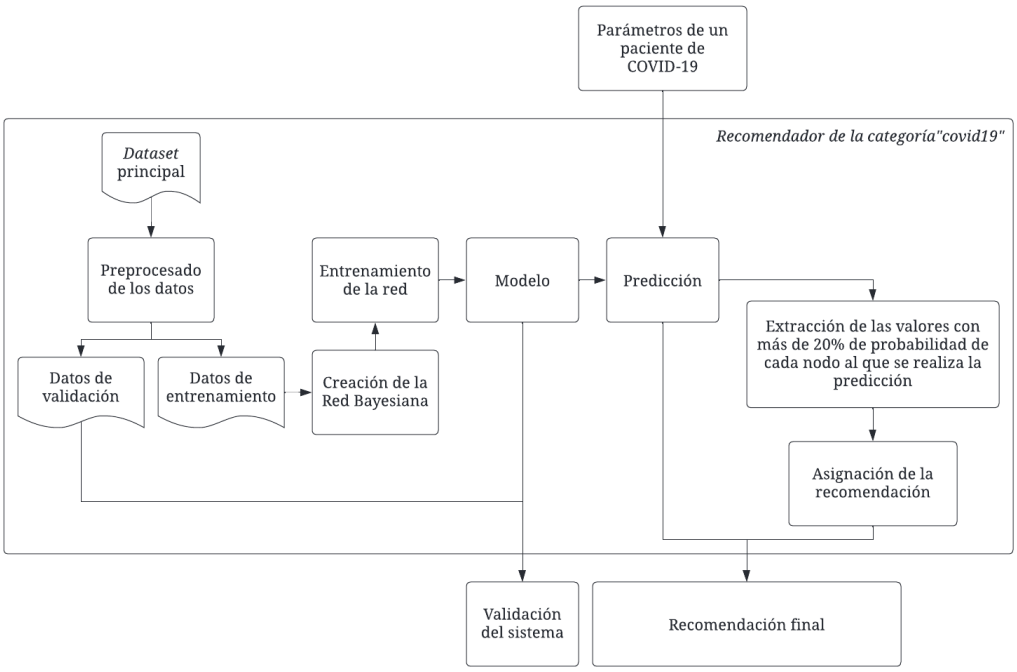

Figura 42. Estructura del recomendador de la categoría "covid19".

Respecto a su funcionamiento, siguiendo el ejemplo empleado desde un principio para la Red Bayesiana de esta categoría, se solicita la recomendación para un paciente de 61 años con COVID-19. El resultado de la petición se ilustra en la Figura 43.

{'predictionList': {'262007003': {'0': 2.94, '1': 97.06}, '397821002': {'0': 97.
39, '1': 2.61}, '419620001': {'0': 99.86}, '445320007': {'438949009': 87.84, '0-9
': 5.65, '10-19': 5.03, '20-29': 1.17}}, 'recommendationList': {}}

Figura 43. Recomendación para un paciente de 61 años con COVID-19.

La recomendación está compuesta por dos partes. La primera son las probabilidades obtenidas por la Red Bayesiana entrenada para cada nodo seleccionado. En este caso, la probabilidad de que el paciente no sea intubado (azul) es de 97,06% y de que no sea transferido a la UCI (verde) es de 99,39%. Respecto a la probabilidad de fallecimiento (naranja) del paciente es muy baja, con un 99,86% de probabilidad de supervivencia. Sin embargo, aunque es improbable que fallezca, las posibilidades nunca son cero, y el sistema estima que los rangos de días para confirmarlo son entre los veinte primeros días tras los síntomas con una probabilidad conjunta de más del 10%.

En cuanto a la segunda parte de la recomendación, no se indica nada debido a que ninguno de los nodos ha tenido una probabilidad superior al 20% en sus valores negativos.

2. Resumen

Se ha desarrollado he implementado la estructura de dos Sistemas de Recomendación, uno para la categoría “general” y otro para la categoría específica “covid19”.

Debido a que su estructura es similar, la recomendación sigue la misma organización. En primer lugar, se indica el apartado de “*predictionList*” en el que se incluyen las probabilidades obtenidas, con el modelo de la Red Bayesiana, de cada posible valor para los nodos de la red seleccionados para predecir. En la segunda parte, “*recommendationList*”, se incluyen las recomendaciones para el usuario correspondientes a los posibles valores de los nodos que han obtenido más de un 20% de probabilidad.

A continuación, en la Figura 44 se especifica un diagrama de flujo de la ejecución general que siguen ambos sistemas para obtener la recomendación final.

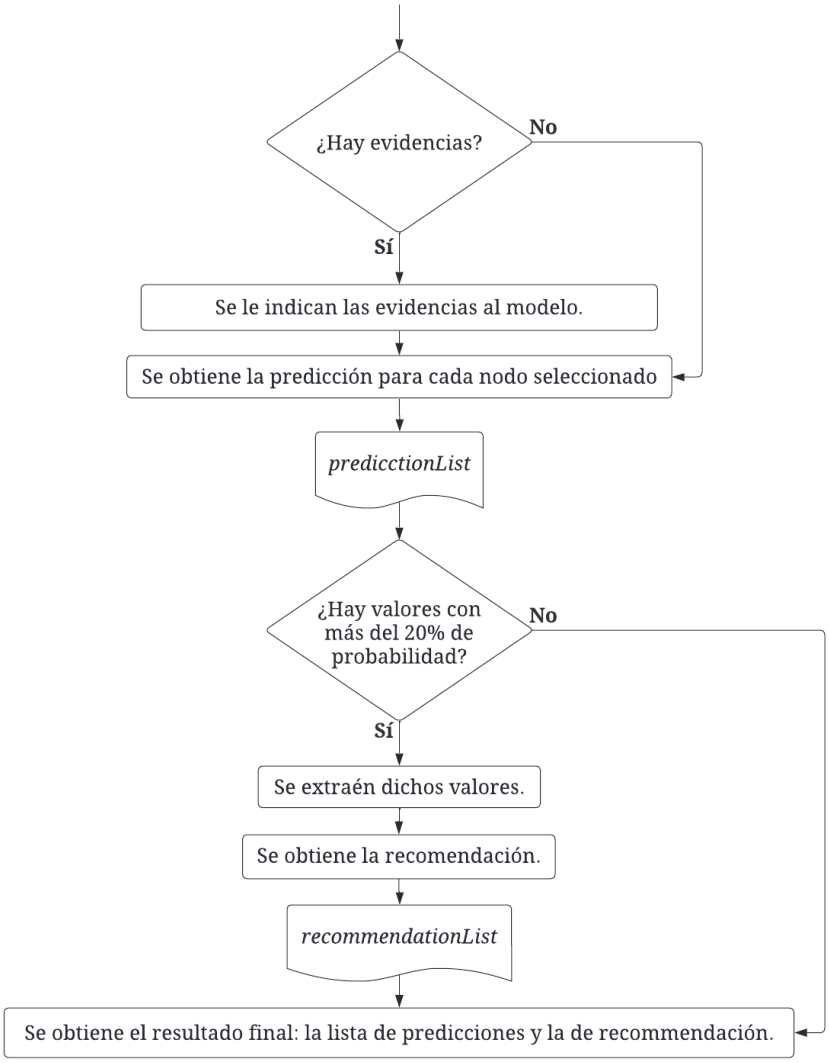

Figura 44. Diagrama de flujo general de la ejecución de los Sistemas de Recomendación implementados.

CAPÍTULO VII. MEDVOLUTION

En este capítulo se detalla la implementación y las características de la aplicación *sotfware* desarrollada en este proyecto.

MEDvolution es el nombre escogido para denominar la aplicación *software* desarrollada con el *framework* de Python Django, que permitirá al usuario hacer uso de los Sistemas de Recomendación creados en el CAPÍTULO VI. Su estructura está conformada por dos aplicaciones diferenciadas, tal y como se ilustra en la Figura 45, que intercambian información mediante el Protocolo de Transferencia de Hipertexto (del inglés *Hypertext Transfer Protocol* – HTTP)[53].

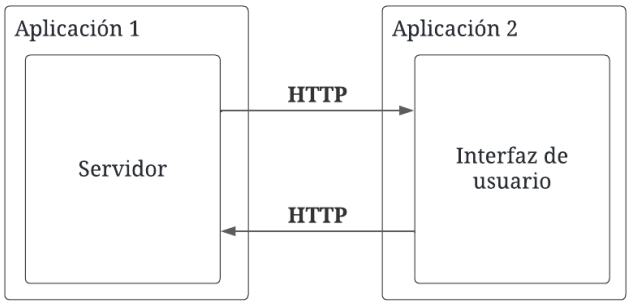

Figura 45. Estructura de MEDvolution.

1. Servidor

La aplicación 1 (Figura 46), denominada *server* en el código de implementación, es el núcleo de MEDvolución. En ella se almacenan todos los archivos referentes a la implementación de los Sistemas de Recomendación de este proyecto, desde los *datasets* originales hasta el *dataset* de *diseaseReccomendations*. Asimismo, se almacenan los archivos de SNOMED CT proporcionados por el Ministerio de Sanidad del Gobierno de España y aquellos archivos que gestionan estos datos, entre otros.

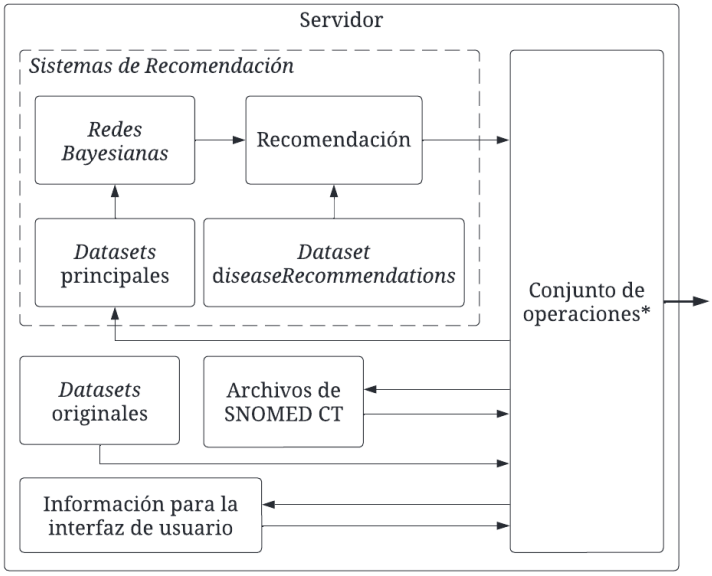

Figura 46. Estructura de la aplicación del servidor de MEDvolution.

En la Figura 46, se puede apreciar las diferentes partes que componen la aplicación, y una aproximación generalista de cómo interactúan entre ellas. El "Conjunto de operaciones" son los métodos y clases creados para el desarrollo y gestión de cada parte, así como del intercambio de información con la otra aplicación que compone MEDvolution.

A continuación, en la Figura 47 se muestra el proceso de ejecución del servidor cada vez que se inicializa. Cuenta con dos variables de estado globales que indican si hay nuevos archivos de SNOMED CT o nuevos *datasets* originales. En función de estas variables la ejecución del código varía.

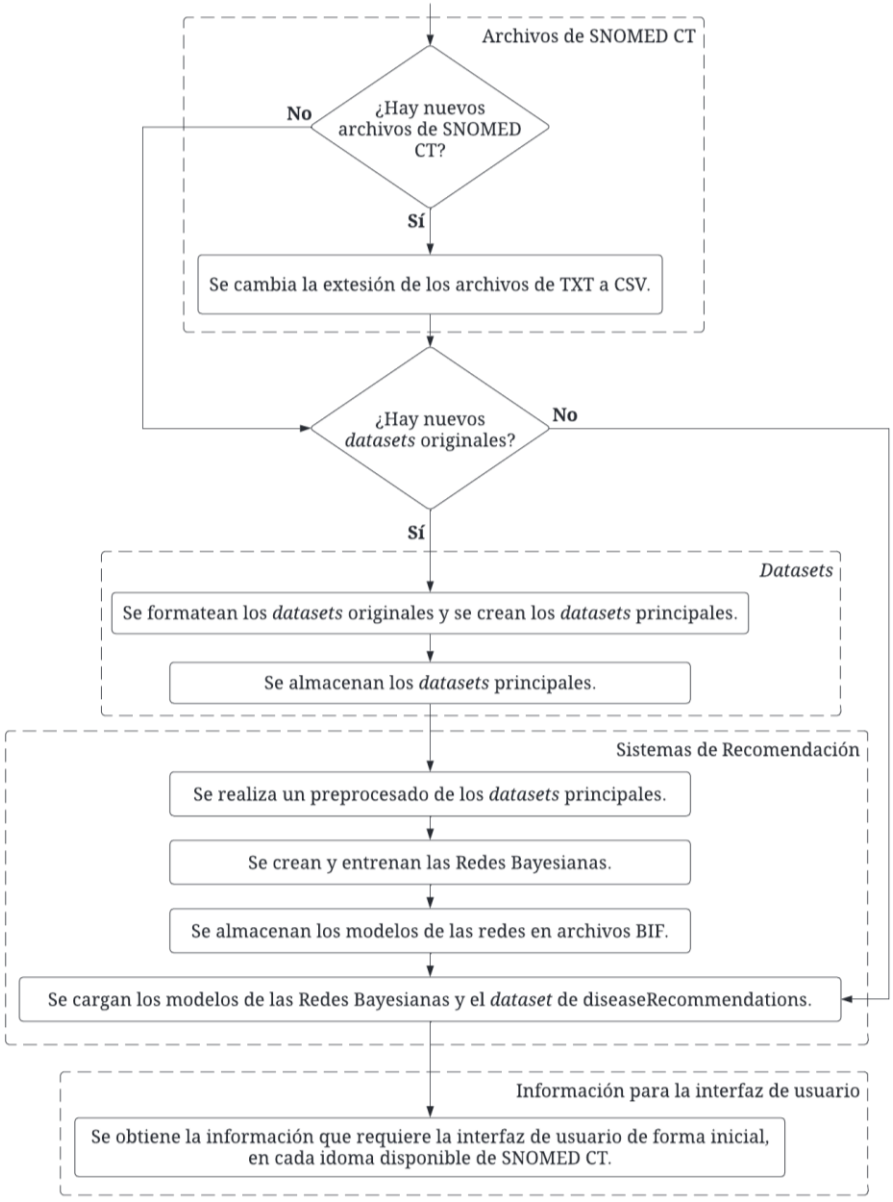

Figura 47. Diagrama de flujo de la inicialización del servidor.

En el ANEXO IV se muestra las diferentes ejecuciones del servidor. Para el caso de si hay nuevos archivos, el servidor tarda en inicializarse en torno a diez minutos, mientras que para el caso de que no hay archivos nuevos, el servidor tarda aproximadamente cinco minutos. De todo el tiempo de ejecución inicial del servidor, independientemente del caso, hay unos cuatro minutos y medio que corresponden a la preparación de información para la interfaz de usuario.

Toda la información manejada en el servidor consiste en dígitos e identificadores de SNOMED CT. Si se envían a la interfaz de usuario es altamente probable que los usuarios de la aplicación no identifiquen dichos números. Por tanto, aquí entra en juego el gran potencial de SNOMED CT.

1.1. Empleo de SNOMED CT

SNOMED CT es un estándar médico internacional cuya estructura global se ilustra en la Figura 48. Como se puede apreciar, se muestran los tres componentes que forman el estándar y las interacciones entre ellos.

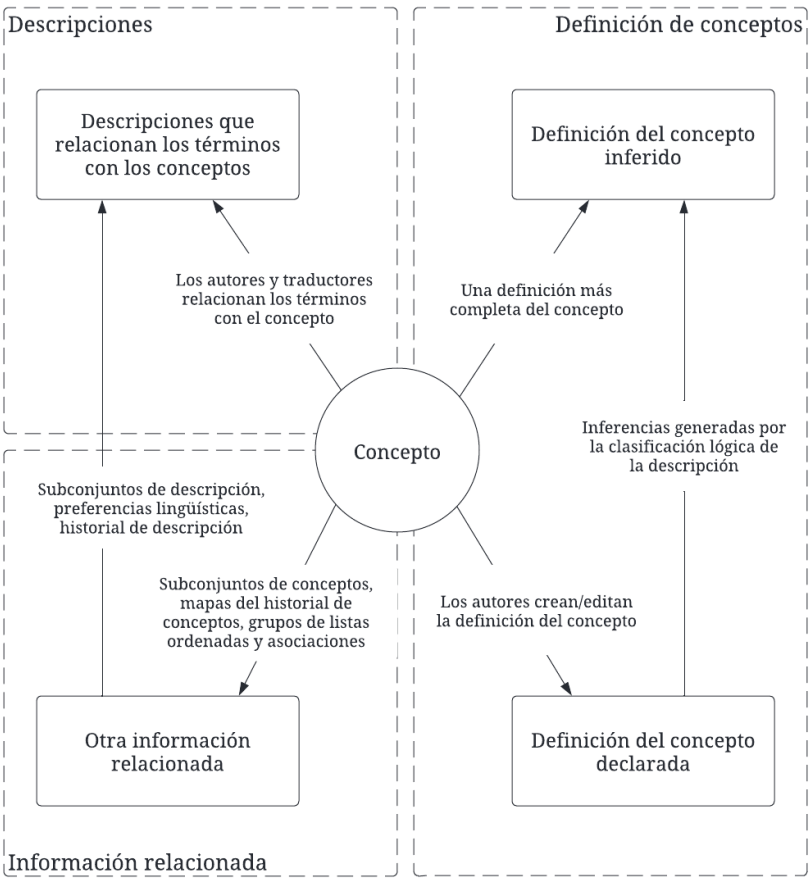

Figura 48. Gráfico de alto nivel de SNOMED CT [54].

De los tres componentes que conforman el modelo lógico del estándar (Figura 49), en este proyecto se emplean sólo dos: conceptos y descripciones. Los conceptos hacen referencia al "pensamiento médico" mientras que las descripciones definen el término médico.

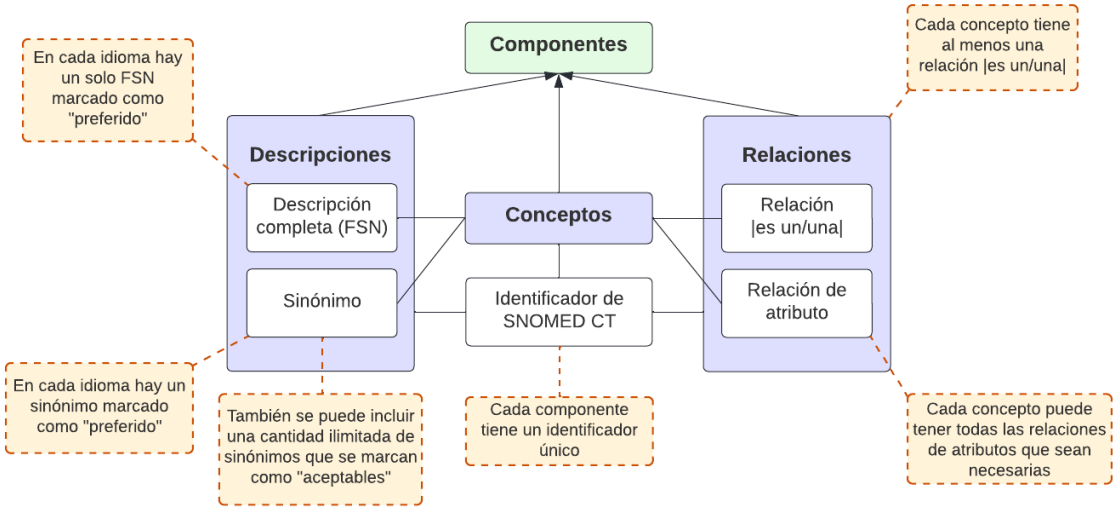

Figura 49. Modelo lógico de SNOMED CT [55].

El objetivo del empleo de SNOMED CT en el proyecto es enviar a la interfaz de usuario las descripciones de los conceptos que conforman las columnas de los *datasets* principales y que son las variables de entrada de los Sistemas de Recomendación. Y no sólo eso, también los resultados de la recomendación. SNOMED CT permite realizar esto gracias a su estructura relacional, tal y como se aprecia en la Figura 50.

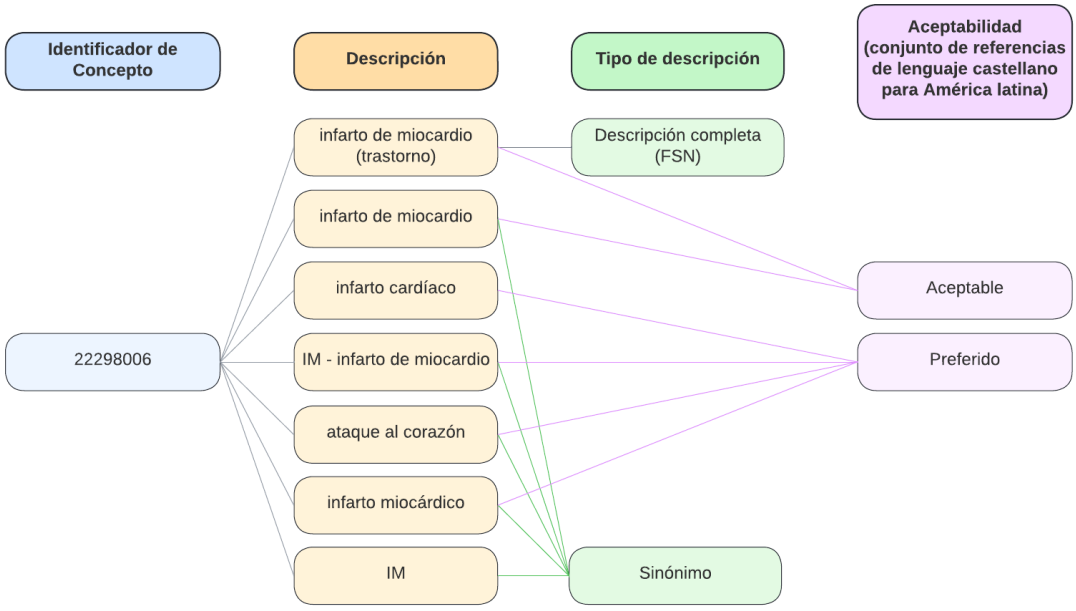

Figura 50. Ejemplo de descripciones de un concepto [55].

Para este proyecto, se ha obtenido una licencia del Ministerio de Sanidad del Gobierno de España para el uso de SNOMED CT, esta licencia permite el acceso a dos ediciones de SNOMED CT, la internacional y la española. De la edición internacional se dispone de la versión de enero de 2022, mientras que de la edición española la versión de junio de 2022.

Para cada una de las ediciones, se permite el acceso a tres tipos de publicaciones: "*full*", "*snapshot*" y "*delta*". En la primera, los archivos de esta contienen todas las versiones de todos los componentes y miembros del conjunto de referencia que se han publicado. La segunda contiene sólo la versión más reciente de cada componente y miembro del conjunto de referencia publicado. Y en la tercera, los archivos de publicación contienen sólo filas que representan versiones de componentes y miembros del conjunto de referencia creados desde la fecha de publicación anterior. Para este proyecto se han seleccionado la publicación "*full*" de ambas ediciones, siendo los archivos que las componen los ilustrados en la Tabla 20.

PREFIJO EN EL NOMBRE DEL ARCHIVO
*sct2_Concept_*
*sct2_Description_*
*sct2_Identifier_*
*sct2_Relationship_*
*sct2_RelationshipConcreteValues_*
sct2_sRefset_OWLExpression
*sct2_StatedRelationship_*
*sct2_TextDefinition_*

Tabla 20. Prefijos de los archivos de cada edición de SNOMED CT.

Concretamente, en este trabajo sólo se utiliza el archivo cuyo prefijo es "*sct2_Description_*" debido a que es el que almacena las descripciones para cada concepto. Por consiguiente, la información para la interfaz de usuario se obtiene inicialmente al ejecutar el servidor con ayuda de este archivo que se estructura siguiendo las características de la Tabla 21.

COLUMNA	DESCRIPCIÓN	POSIBLES VALORES
id	Identificador único para el miembro del conjunto de referencia.	Entero sin signo de 128 bits
effectiveTime	Indica la fecha u hora en la que se convirtió en parte de la versión actual.	Fecha u hora en formato YYYYMMDD hh:mm:ss
active	Indica el estado del miembro del conjunto de referencia, es decir si pertenece o no a la versión actual.	0 o 1
moduleId	Identifica el módulo jerárquico al que pertenece.	SCTID
conceptId	Indica el concepto al que hace referencia la descripción.	
languageCode	Indica el idioma en el que está escrita la descripción.	Cadena de texto (ISO-639-1 code)
typeId	Indica el tipo de descripción.	90000000013009 o 900000000000003001 (Sinónimo o Nombre totalmente especificado)
term	Es una cadena de texto que representa el concepto al que hace referencia la columna "*conceptId*".	Cadena de texto
caseSignificanceId	Indica si el texto del término puede modificarse cambiando los caracteres de mayúsculas a minúsculas (o viceversa).	SCTID

Tabla 21. Características del archivo "sct2_Description_".

Para obtener la información para la interfaz de usuario, se extraen de los *datasets* principales los nombres de las columnas (SCTID) correspondientes a las variables de entrada de los Sistemas de Recomendación. Luego, por cada SCTID, se obtiene su definición y sinónimos actuales del *dataset* "*sct2_Description_",* almacenándolos para cada idioma. En este proceso, el servidor emplea en torno a cuatro minutos y medio cada vez que se inicializa.

El mismo proceso de traducción se emplea para mandar la recomendación a la interfaz de usuario. Por ejemplo, en la parte superior de la Figura 51 se muestra la recomendación para un paciente que padece febrícula, disnea y fatiga. En la parte inferior de la figura se indica la traducción de dicha recomendación, siendo ahora fácilmente interpretable por el usuario.


```
 {'predictionList': {'439401001': {'840539006': 1.05, '56717001': 98.95}}, '
recommendationList': {'439401001': {'56717001': {'diagnosticTest1': '84100007
421829000 5880005', 'diagnosticTest2': '424489006', 'diagnosticTest3': '396550006
', 'diagnosticTest4': '399208008|77477000', 'diagnosticTest5': '104173009', '
treatment1': '776412002|777438009|775855004|777364002|774747007|776544000', '
treatmentSideEffects': '16932000|79890006|18165001|39575007|424131007 420561004
131148009| 246636008', 'copingAndSupport': '390822007'}}}}
```

```
 {'predictionList': {'Diagnóstico': {'synonyms': [], 'Enfermedad causada por
coronavirus 2 del síndrome respiratorio agudo severo': {'value': 1.05, 'synonyms
': ['Enfermedad causada por coronavirus del síndrome respiratorio agudo grave 2
', 'Enfermedad causada por sars-cov-2', 'Enfermedad causada por coronavirus del
síndrome respiratorio agudo severo 2', 'Enfermedad causada por covid-19', '
Enfermedad causada por coronavirus 2 del síndrome respiratorio agudo grave']}, '
Tuberculosis': {'value': 98.95, 'synonyms': ['Infección por mycobacterium
tuberculosis']}}}, 'recommendationList': {'Diagnóstico': {'Tuberculosis': {'
diagnosticTest1': 'Redacción de la historia clínica Y Examen físico', '
diagnosticTest2': 'Prueba de mantoux', 'diagnosticTest3': 'Prueba sanguínea', '
diagnosticTest4': 'Radiografía simple de tórax,Tomografía axial computarizada', '
diagnosticTest5': 'Cultivo microbiano de esputo', 'treatment1': 'Producto que
contiene isoniazida como único ingrediente,Producto que contiene rifampicina como
 único ingrediente,Producto que contiene etambutol como único ingrediente,
Producto que contiene pirazinamida como único ingrediente,Producto que contiene
bedaquilina como único ingrediente,Producto que contiene linezolid como único
ingrediente', 'treatmentSideEffects': 'Náuseas y vómitos,Pérdida del apetito,
Ictericia,Orina oscura,Facilidad para la formación de hematomas O Hemorragia,
Visión nebulosa', 'copingAndSupport': 'Terapias psicológicas'}}}}
```


Figura 51. Traducción de la recomendación.

2. Interfaz de usuario

Respecto a la segunda aplicación de MEDvolution, se corresponde a una interfaz web a través de la cual el usuario interactúa con los Sistemas de Recomendación creados. Consta de un total de cuatro páginas web accesibles en los idiomas inglés y español.

2.1. Página inicial

Esta es la página inicial (Figura 52) de la aplicación y es puramente informativa. En ella se describe el proyecto y sus características, indicando al usuario qué es lo que obtiene al usar los diferentes SR desarrollados.

2.2. Página de selección de evidencias

Esta página web se modifica en función del Sistema de Recomendación seleccionado, tal y como se puede apreciar en la Figura 53, que muestra los parámetros para el SR de la categoría “general”, y la Figura 54, que muestra los parámetros para el de la categoría “covid19”. Sin embargo, su estructura organizativa (*template*) se mantiene.

En la parte izquierda, consta de un filtro de búsqueda para los diferentes parámetros de entrada del Sistema de Recomendación, así como el método de selección del valor del parámetro, puede ser una casilla de verificación cuando sólo tiene dos valores posibles o un selector para cuando tiene múltiples. En la parte derecha, se indican las evidencias que se han seleccionado para la recomendación.

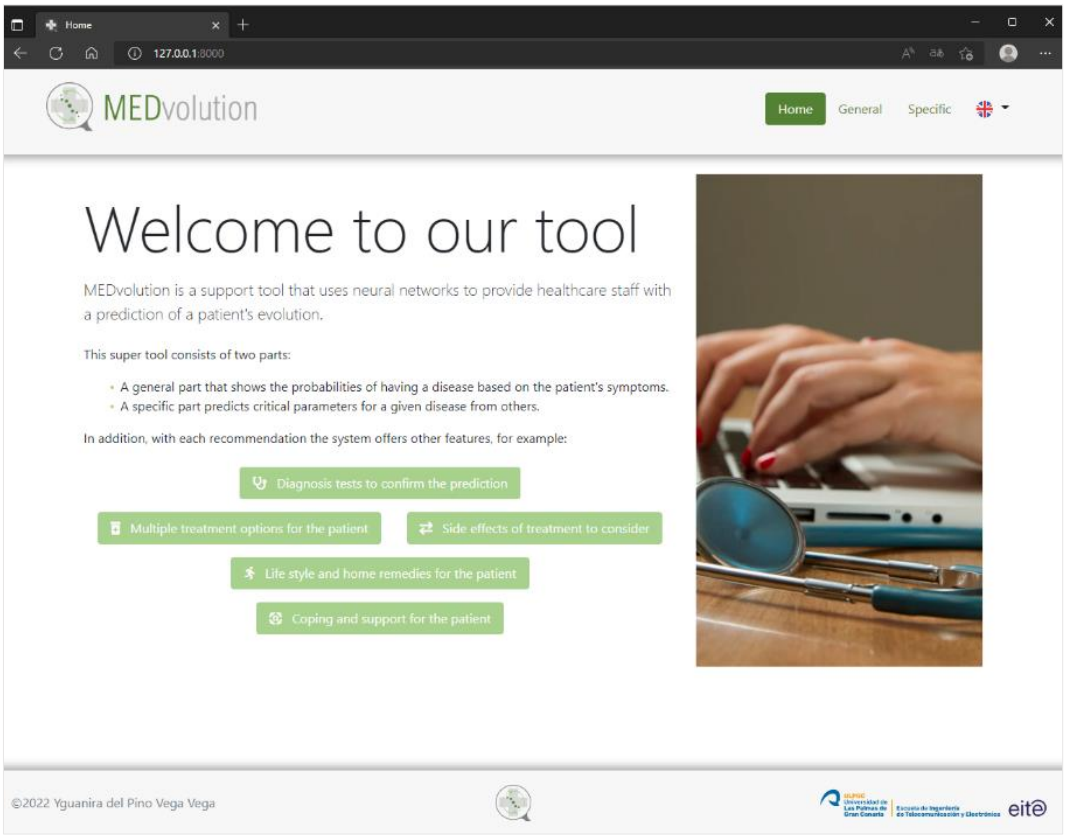

Figura 52. Página inicial de MEDvolution en inglés.

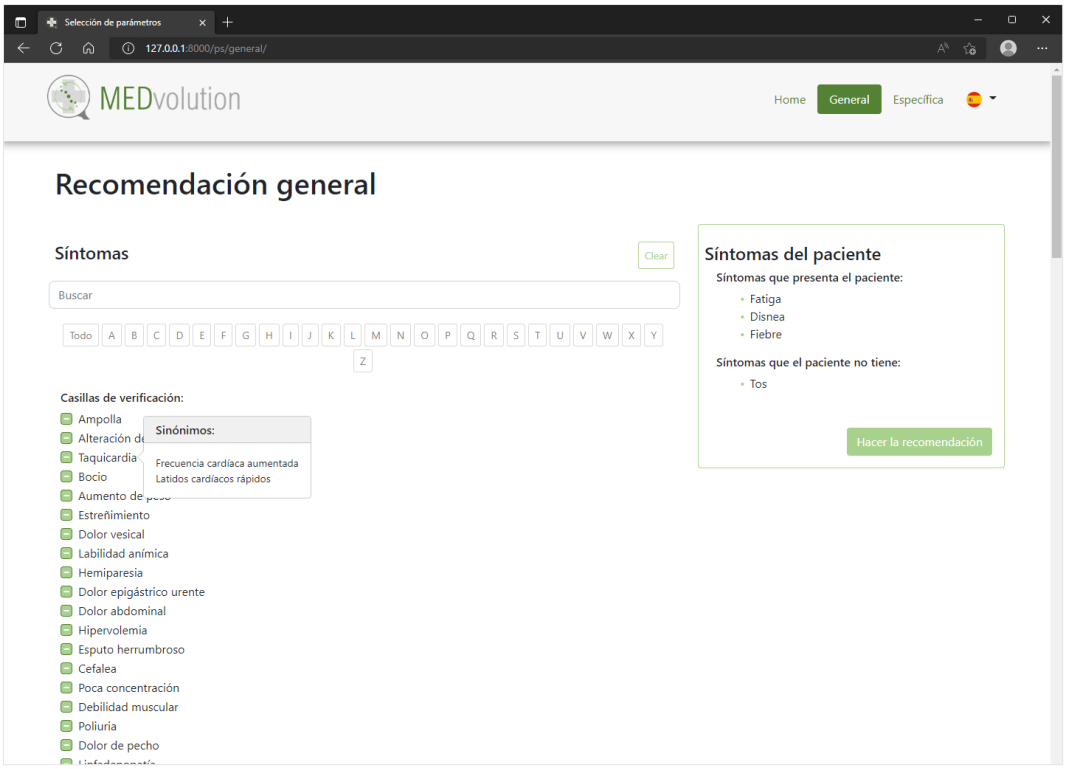

Figura 53. Página de selección de evidencias de MEDvolution para la categoría "general" en español.

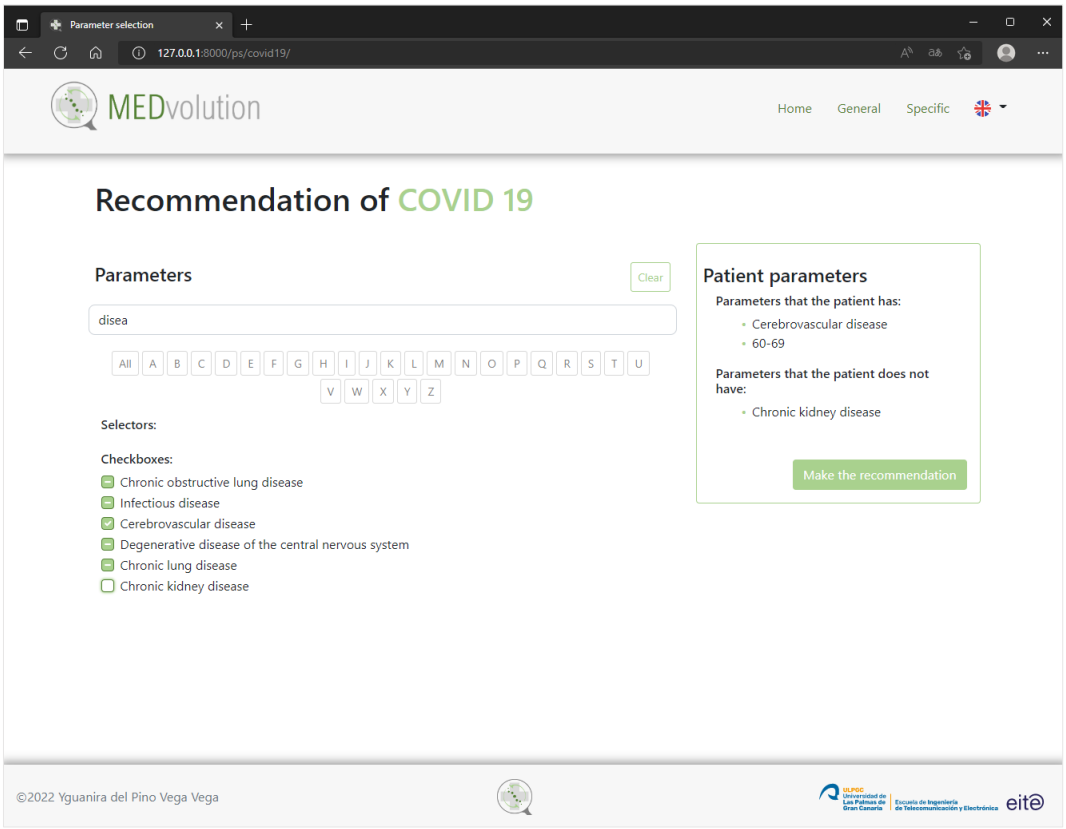

Figura 54. Página de selección de evidencias de MEDvolution para la categoría "covid19" en inglés.

2.3. Página de selección de enfermedades

En esta página se indica para que categoría especifica se va a realizar la recomendación. En este caso, como se muestra en la Figura 55, se dispone de una única categoría, la de "covid19". Asimismo, al pasar el cursor por encima de la categoría, se muestra otras formas de nombrarla. Este mismo sistema se emplea para los parámetros de entrada de los Sistemas de Recomendación en la página de selección de evidencias.

2.4. Página de recomendación

Esta es la última página que conforma la interfaz de usuario, en ella se muestra en la parte superior una o varias gráficas, correspondientes a las predicciones realizadas por el sistema de recomendación. En la parte inferior, se visualiza una recomendación para aquellos valores que superen el 20% de probabilidad. En las Figuras 53 y 54 se puede ver cómo cambia la página dependiendo de la categoría a la que pertenece el recomendador.

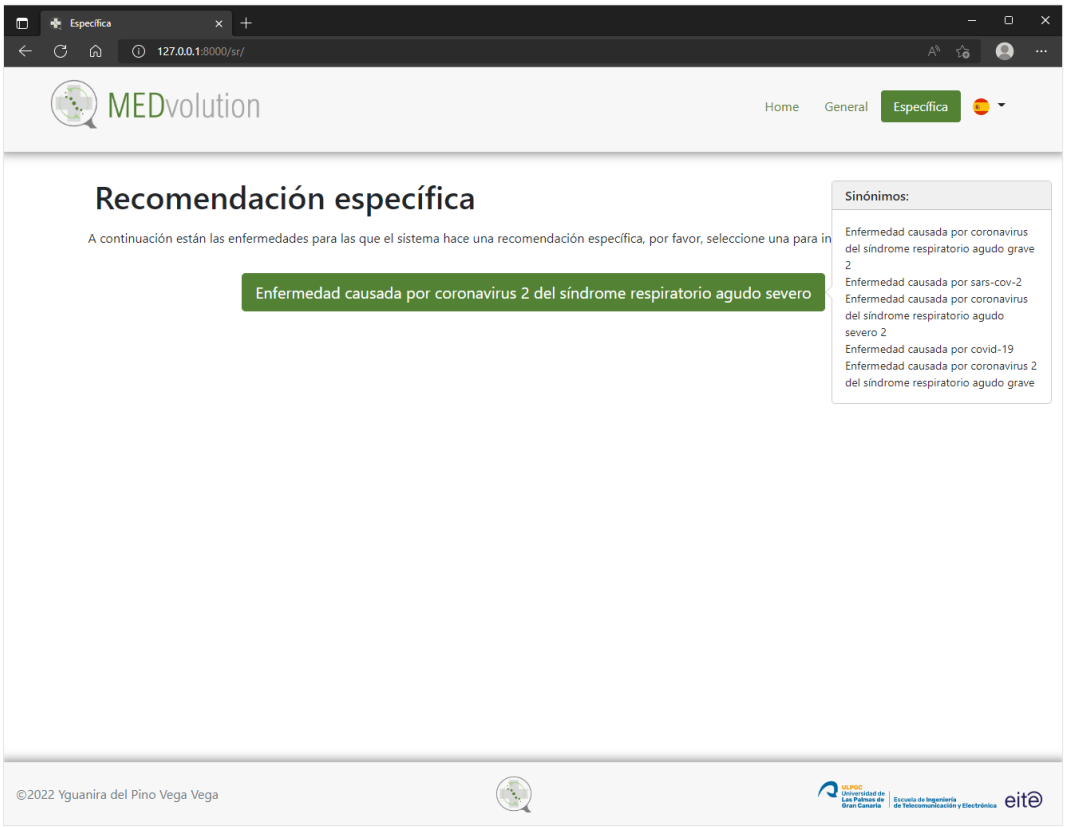

Figura 55. Página de selección de enfermedades de MEDvolution en español.

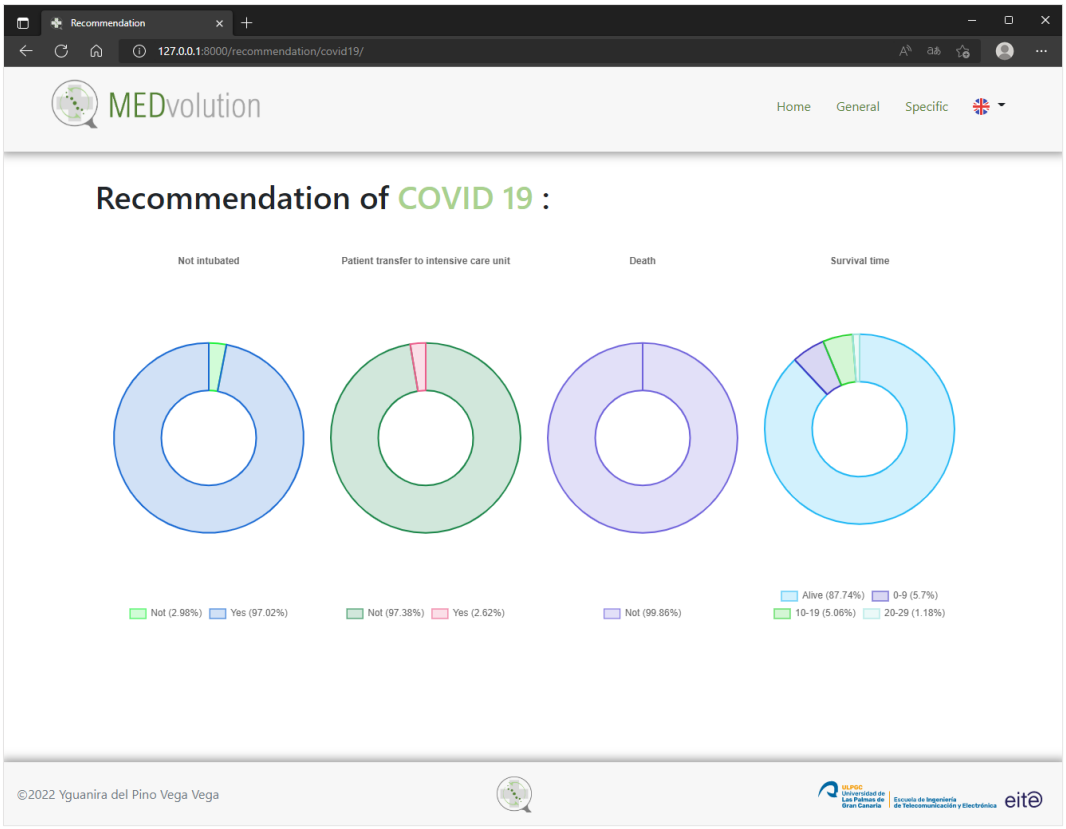

Figura 56. Página de recomendación de MEDvolution para la categoría "covid19" en inglés.

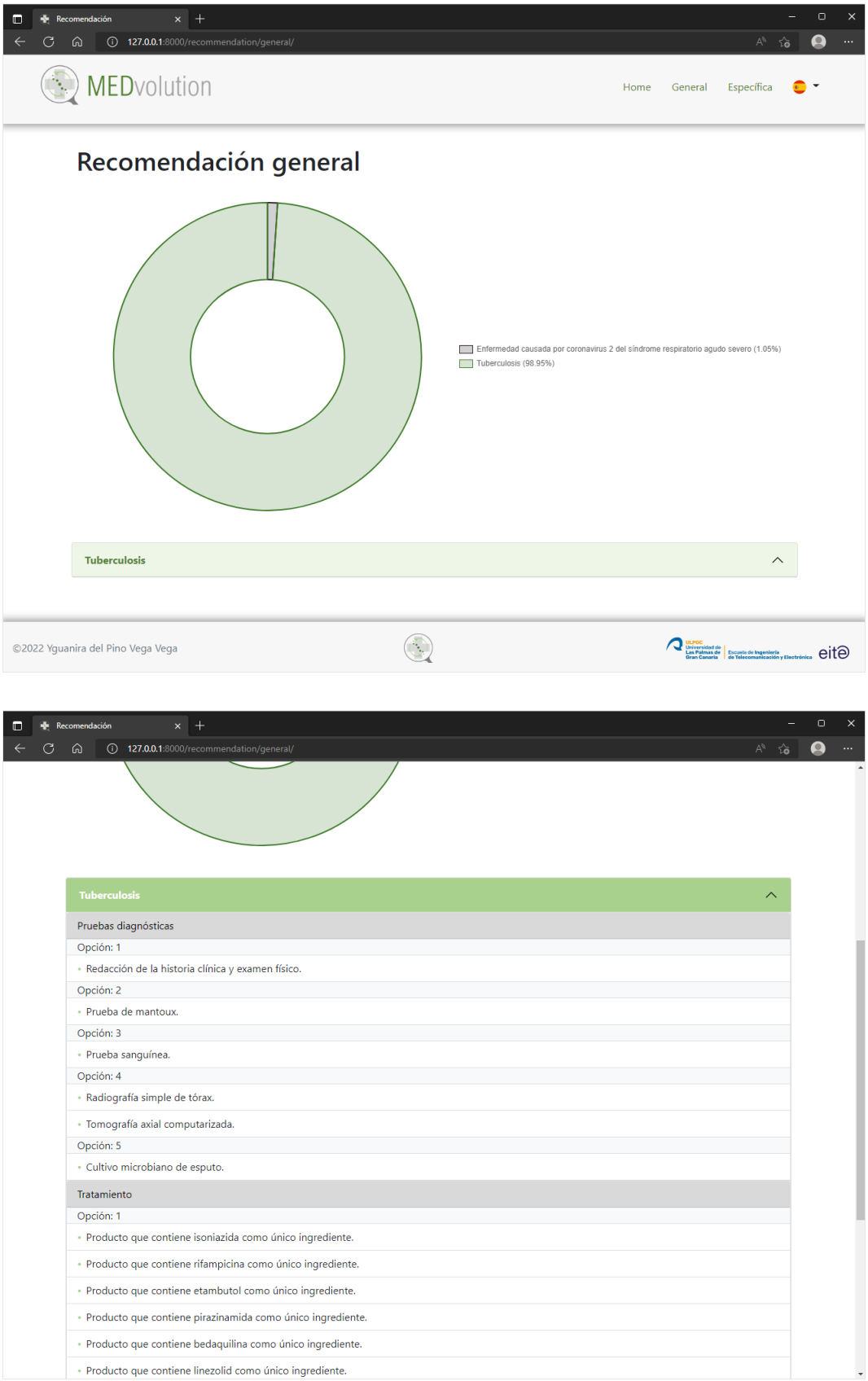

Figura 57. Página de recomendación de MEDvolution para la categoría "general" en español.

CAPÍTULO VIII. CONCLUSIONES

En este capítulo se hace un breve resumen el documento y se exponen las resoluciones finales del proyecto.

No es un secreto, ni un error, afirmar que hay ciertos ámbitos del conocimiento que se retroalimentan y que todo está intrínsecamente relacionado. Las grandes innovaciones de la historia son fruto de investigaciones, algunas para suplir un deseo, como el de volar o explorar nuevos mundos; y otras para solventar problemas cotidianos. Para realizar dichas investigaciones se deben crear objetos o fórmulas, y eso propicia el desarrollo de las tecnologías necesarias para ello. Por consiguiente, cuando el campo de la tecnología evoluciona a una nueva fase, otros campos, pero en especial el de la medicina, evolucionan.

1. Contexto inicial

Con el desarrollo de las tecnologías de la primera década del siglo XXI, los Sistemas de Recomendación surgieron para solucionar un problema que cada vez se hace más y más grande, la gestión de grandes volúmenes de datos.

En el campo de la medicina, el volumen de datos existentes en múltiples plataformas de Internet ocasiona que los profesionales deban buscar información constantemente, contrastar, consultar y debatir con otros expertos respecto a un paciente o comparar historias clínicas. Por consiguiente, con el proyecto desarrollado en este documento, se quiere proporcionar una herramienta de apoyo a los médicos para optimizar su tiempo de respuesta y proporcionar el mejor resultado posible para el paciente. Para conseguir esta meta, el objetivo principal de este Trabajo de Fin de Máster es la creación de una plataforma *software* para web y dispositivos móviles basada en un sistema de recomendación que ayude a los profesionales del ámbito sanitario a realizar una predicción evolutiva del cuadro médico de un paciente a partir de sus datos en estudio y del historial de los pacientes con la misma patología.

Para conseguir el objetivo principal del proyecto, se debían alcanzar una serie de objetivos específicos que describían el futuro desarrollo del proyecto:

1. Análisis de los sistemas *software* y plataformas existentes para almacenamiento de datos clínicos de pacientes con la intención de saber con exactitud qué datos son realmente los requeridos en la actualidad y poder utilizarlos luego en nuestro desarrollo para la creación del "perfil" de un paciente.
2. Análisis de los SR centrándose en aquellos aplicados en el ámbito de la salud.

3. Desarrollo de una infraestructura (*Back-end*) del SRS para la administración y posterior utilización de aquellos datos procedentes de historiales clínicos que sean necesarios para que el SRS desarrolle su funcionalidad.
4. Desarrollo de la interfaz de usuario (*Front-end*) del Sistema de Recomendación para la Salud que permita conectar a los profesionales del ámbito sanitario con el *Back-end* para recuperar y procesar los datos procedentes de los historiales clínicos necesarios para solicitar la predicción evolutiva acerca del cuadro médico de un paciente.
5. Implementación de la funcionalidad completa del SRS mediante la integración del *Front-end* y *Back-end* desarrollados en una aplicación web responsiva que facilite el acceso y uso del SRS a los profesionales del ámbito sanitario.

2. Valoración de los objetivos

Una vez finalizado el proyecto, se aprecia que efectivamente se ha logrado diseñar de forma exitosa una herramienta de apoyo para el personal sanitario sin fronteras de idioma gracias al estándar internacional SNOMED CT. En la Tabla 22, se detalla cómo se han conseguido dichos objetivos.

OBJETIVO	JUSTIFICACIÓN
1	En el CAPÍTULO II se realiza una extensa y exhaustiva investigación referente a qué información se emplea para el diagnóstico de una enfermedad y donde se almacena; y en cuanto a los Sistemas de Recomendación existentes y los últimos desarrollados.
2	
3	Este objetivo parte en el CAPÍTULO IV con la selección de los datos que se emplean en el proyecto y finaliza en el CAPÍTULO VI con el desarrollo de los Sistemas de Recomendación que son el núcleo del proyecto.
4	En el CAPÍTULO VII se muestra la aplicación final desarrollada en este proyecto con sus dos partes, el servidor donde se encuentran los Sistemas de Recomendación y la interfaz de usuario que permite la interacción con ellos.
5	

Tabla 22. Conclusiones de los objetivos del proyecto.

Con la realización de este Trabajo de Fin de Máster queda demostrado el gran potencial de ayuda que tienen los Sistemas de Recomendación para los profesionales sanitarios, y, por ende, para sus pacientes.

3. Líneas futuras

Respecto al futuro de este proyecto, se propone, en primer lugar, solicitar la colaboración de uno o varios centros médicos u hospitalarios que proporcionen información detallada respecto a los síntomas de sus pacientes, enfermedades y tratamientos con el fin de elevar la precisión de los Sistemas de Recomendación. Además, colaborar con profesionales sanitarios para conocer qué precisan ellos en la recomendación y así, mejorar su calidad.

Por último, realizar un estudio en profundidad del estándar SNOMED CT para investigar si se pueden implementar nuevas funcionalidades o mejorar las existentes con el objetivo de que alcancen todo su potencial, es decir, que utilicen todos los componentes presentes en el estándar.

BIBLIOGRAFÍA

En esta sección se detallan las referencias consultadas durante la realización de este Trabajo de Fin de Máster.

[1] Statista, «Total data volume worldwide 2010-2025 | Statista», *Statista*, 2021. [En línea]. Disponible en: https://www.statista.com/statistics/871513/worldwide-data-created/. [Accedido: 06-ene-2022]

[2] J. Su, Y. Guan, Y. Li, W. Chen, H. Lv, y Y. Yan, «Do recommender systems function in the health domain: a system review», *arXiv Prepr. arXiv2007.13058*, n.º 92, 2020 [En línea]. Disponible en: https://arxiv.org/abs/2007.13058

[3] «¿Cuántas enfermedades raras existen?» [En línea]. Disponible en: https://www.nationalgeographic.com.es/ciencia/cuantas-enfermedades-raras-existen_15167. [Accedido: 26-may-2022]

[4] World Health Organization, «Infecciones de transmisión sexual», 14-jun-2019. [En línea]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis). [Accedido: 01-oct-2021]

[5] M. de Sanidad y C. Bienestar Social, «CIE • 10 • ES Clasificación Internacional de Enfermedades-10.ª Revisión Sistema de Clasificación de Procedimientos MINISTERIO DE SANIDAD, CONSUMO Y BIENESTAR SOCIAL MINISTERIO DE LA PRESIDENCIA, RELACIONES CON LAS CORTES E IGUALDAD BOLETÍN OFICIAL DEL EST», 2020 [En línea]. Disponible en: http://eciemaps.mcsb.gob.es/ecieMaps/errata/errata.html. [Accedido: 01-oct-2021]

[6] NEJM Catalyst, «Healthcare big data and the promise of value-based care», *NEJM Catal.*, vol. 4, n.º 1, pp. 1-7, 2018 [En línea]. Disponible en: https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0290. [Accedido: 19-sep-2021]

[7] J. A. Batsis *et al.*, «Development and Usability Assessment of a Connected Resistance Exercise Band Application for Strength-Monitoring.», *World Acad. Sci. Eng. Technol.*, vol. 13, n.º 5, pp. 340-348, 2019, doi: 10.5281/zenodo.

[8] T. N. T. Tran, A. Felfernig, y N. Tintarev, «Humanized Recommender Systems: State-of-the-art and Research Issues», *ACM Trans. Interact. Intell. Syst.*, vol. 11, n.º 2, pp. 171-201, 2021, doi: 10.1145/3446906.

[9] G. Davidson, «The top 10 medical apps for doctors | Blog de JotForm», *JotForm blog*, 2021. [En línea]. Disponible en: https://www.jotform.com/blog/medical-apps/. [Accedido: 02-oct-2021]

[10] «| Epocrates». [En línea]. Disponible en: https://www.epocrates.com/features. [Accedido: 02-oct-2021]

[11] «How We Help | UpToDate | Wolters Kluwer». [En línea]. Disponible en: https://www.wolterskluwer.com/en/solutions/uptodate/how-we-help. [Accedido: 02-oct-2021]

[12] «About Us - MDCalc». [En línea]. Disponible en: https://www.mdcalc.com/about-us. [Accedido: 02-oct-2021]

[13] F. Ricci, B. Shapira, y L. Rokach, «Recommender systems: Introduction and challenges», en *Recommender Systems Handbook, Second Edition*, 2015, pp. 1-34.

[14] W. Yue, Z. Wang, J. Zhang, y X. Liu, «An Overview of Recommendation Techniques and Their Applications in Healthcare», *IEEE/CAA Journal of Automatica Sinica*, vol. 8, n.º 4. pp. 701-717, 2021.

[15] T. N. T. Tran, A. Felfernig, C. Trattner, y A. Holzinger, «Recommender systems in the healthcare domain: state-of-the-art and research issues», *J. Intell. Inf. Syst.*, vol. 57, n.º 1, pp. 171-201, 2021, doi: 10.1007/s10844-020-00633-6. [En línea]. Disponible en: https://doi.org/10.1007/s10844-020-00633-6. [Accedido: 08-oct-2021]

[16] S. N. Mohanty, J. M. Chatterjee, S. Jain, A. A. Elngar, y P. Gupta, *Recommender System with Machine Learning and Artificial Intelligence*. 2020.

[17] L. Carina Pereira Brandão y A. Gomes da Silva, «Polytechnic Institute of Coimbra Higher Institute of Accounting and Administration of Coimbra Wavelet-Based Cancer Drug Recommender System», 2020.

[18] M. Kamran y A. Javed, «A Survey of Recommender Systems and their Application in Healthcare», *Tech. Journal, Univ. Eng. Technol. Taxila, Pakistan*, vol. 20, n.º Iv, pp. 111-119, 2015.

[19] M. Kuanr, P. Mohapatra, y J. Piri, «Health Recommender System for Cervical Cancer Prognosis in Women», *Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021*, pp. 673-679, ene. 2021, doi: 10.1109/ICICT50816.2021.9358540.

[20] A. A. Neloy, M. Shafayat Oshman, M. M. Islam, M. J. Hossain, y Z. Bin Zahir, «Content-Based Health Recommender System for ICU Patient», *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11909 LNAI, n.º November, pp. 229-237, 2019, doi: 10.1007/978-3-030-33709-4_20.

[21] Frederic Llordachs Marqués, «Qué es la historia clínica de un paciente y para qué sirve», *Qué es la historia clínica de un paciente y para qué sirve*, 2020. [En línea]. Disponible en: https://clinic-cloud.com/blog/la-historia-clinica-paciente-sirve/. [Accedido: 06-oct-2021]

[22] Ministerio de Sanidad Consumo y Bienestar Social, «Ministerio de Sanidad, Consumo y Bienestar Social - Profesionales - Historia Clínica Digital del Sistema Nacional de Salud», 2019. [En línea]. Disponible en: https://www.mscbs.gob.es/profesionales/hcdsns/contenidoDoc/home.htm. [Accedido: 28-sep-2021]

[23] Agencia de Calidad del Sistema Nacional de Salud, «Historia Clínica Digital en el Sistema Nacional de Salud Conjunto Mínimo de Datos», pp. 1-68, 2008.

[24] U. P. de Valencia, VeraTech, y Indizen, «Lista de arquetipos (Recusos de Modelado Clínico)», 2014 [En línea]. Disponible en: https://www.mscbs.gob.es/profesionales/hcdsns/areaRecursosSem/Lista_arquetipos_CMDIC_v1_20022014.pdf

[25] BOE 2019_LEY 5/2010, «Boletín Oficial del Estado», *Boletín Of. del Estado*, pp. 26798-26800, 2021.

[26] G. de España, «Sistema HCDSNS Historia Clínica Digital del Sistema Nacional de Salud. Informe de Situación 31 de Enero de 2021», 2021.

[27] Gobierno de Canarias, «Sistemas electromédicos y de información», *Servicio Canario de Salud*. [En línea]. Disponible en: https://www3.gobiernodecanarias.org/sanidad/scs/contenidoGenerico.jsp?idCarp

eta=3ccc5a31-7075-11e8-920b-97c0ce05d9e2&idDocument=eed8b445-706f-11e8-920b-97c0ce05d9e2. [Accedido: 07-oct-2021]

[28] SNOMED International, «SNOMED - Home | SNOMED International», 2019. [En línea]. Disponible en: https://www.snomed.org/. [Accedido: 30-may-2022]

[29] SNOMED International, «SNOMED - Members», 2022. [En línea]. Disponible en: https://www.snomed.org/our-stakeholders/members. [Accedido: 30-may-2022]

[30] «Ministerio de Sanidad - Profesionales - ÁREA DE DESCARGA DE SNOMED CT». [En línea]. Disponible en: https://www.sanidad.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/areaDescarga.htm. [Accedido: 30-may-2022]

[31] JETBRAINS, «PyCharm: el IDE de Python para desarrolladores profesionales, por JetBrains», 2022. [En línea]. Disponible en: https://www.jetbrains.com/es-es/pycharm/. [Accedido: 02-jun-2022]

[32] Django Software Foundation, «The Web framework for perfectionists with deadlines | Django», *Django Software Foundation*. 2018 [En línea]. Disponible en: https://www.djangoproject.com/. [Accedido: 02-jun-2022]

[33] NumPy Developers, «NumPy documentation — NumPy v1.22 Manual», 2022. [En línea]. Disponible en: https://numpy.org/doc/stable/index.html. [Accedido: 02-jun-2022]

[34] «pandas documentation — pandas 1.4.2 documentation». [En línea]. Disponible en: https://pandas.pydata.org/docs/index.html. [Accedido: 02-jun-2022]

[35] «Introduction to pyAgrum — pyAgrum 1.1.0 documentation». [En línea]. Disponible en: https://pyagrum.readthedocs.io/en/1.1.0/. [Accedido: 02-jun-2022]

[36] Kaggle, «Kaggle: Your Machine Learning and Data Science Community», *Kaggle*, 2019. [En línea]. Disponible en: https://www.kaggle.com/. [Accedido: 04-jun-2022]

[37] KAUSHIL268, «Disease Prediction Using Machine Learning | Kaggle», 2020. [En línea]. Disponible en: https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning?select=Training.csv. [Accedido: 04-jun-2022]

[38] T. Mukherjee, «COVID-19 patient pre-condition dataset | Kaggle», 2020. [En línea]. Disponible en: https://www.kaggle.com/datasets/tanmoyx/covid19-patient-precondition-dataset?select=covid.csv. [Accedido: 04-jun-2022]

[39] T. ALAM, «Covid-19 patients symptom dataset | Kaggle», 2021. [En línea]. Disponible en: https://www.kaggle.com/datasets/takbiralam/covid19-symptoms-dataset. [Accedido: 04-jun-2022]

[40] M. M. Ahamad, «Comorbidities and Symptoms of COVID-19 patient's | Kaggle», 2021. [En línea]. Disponible en: https://www.kaggle.com/datasets/martuza/comorbidities-and-symptoms-of-covid19-patients?select=symptoms.csv. [Accedido: 04-jun-2022]

[41] «International SNOMED CT Browser». [En línea]. Disponible en: https://browser.ihtsdotools.org/? [Accedido: 06-jun-2022]

[42] C. Schmidler, «100, 101 or 102 Degree Fever? Adult Guide to High Temperatures», *Health Pages*, 2021. [En línea]. Disponible en: https://www.healthpages.org/health-a-z/fever-adults/. [Accedido: 05-jun-2022]

[43] X. Amatriain, A. Jaimes*, N. Oliver, y J. M. Pujol, «Data Mining Methods for Recommender Systems», *Recomm. Syst. Handb.*, pp. 39-71, 2011, doi: 10.1007/978-0-387-85820-3_2. [En línea]. Disponible en: https://link.springer.com/chapter/10.1007/978-0-387-85820-3_2. [Accedido: 29-jun-2022]

[44] T. A. Stephenson, «An Introduction to Bayesian Network Theory and Usage». IDIAP, 2000 [En línea]. Disponible en: https://infoscience.epfl.ch/record/82584. [Accedido: 27-oct-2021]

[45] L. E. Sucar y M. Tonantzintla, «Redes Bayesianas», *Aprendiz. Automático conceptos básicos y Av.*, vol. 77, p. 100, 2006.

[46] «Tutorial pyAgrum — pyAgrum 1.1.1 documentation». [En línea]. Disponible en: https://pyagrum.readthedocs.io/en/1.1.1/notebooks/01-Tutorial.html. [Accedido: 15-jun-2022]

[47] «Algoritmo EM», *RUA. Repositorio Institucional de la Universidad de Alicante*. [En línea]. Disponible en: https://rua.ua.es/dspace/bitstream/10045/9887/11/Gallardo-Lopez-Domingo_10.pdf. [Accedido: 17-jun-2022]

[48] «Laplace Smoothing - Coding Ninjas CodeStudio». [En línea]. Disponible en: https://www.codingninjas.com/codestudio/library/laplace-smoothing. [Accedido: 17-jun-2022]

[49] C. Soria, H. Da Silva, y A. E. Martin, «Una propuesta para diversificar el campo de aplicación de un algoritmo Hill Climbing», *Inf. Científicos Técnicos - UNPA*, vol. 9, n.º 1, pp. 102-114, 2017, doi: 10.22305/ict-unpa.v9i1.239.

[50] A. Rojatkar, «Precision, recall, sensitivity and specificity», *OpenGenud IQ: Computing Expertise & Legacy*. [En línea]. Disponible en: https://iq.opengenus.org/precision-recall-sensitivity-specificity/. [Accedido: 19-jun-2022]

[51] «2.3 Valores faltantes», *Universidad de Valencia*. [En línea]. Disponible en: https://www.uv.es/webgid/Descriptiva/23_valores_faltantes.html. [Accedido: 19-jun-2022]

[52] Mayo Clinic, «Mayo Clinic Mayo Clinic», *Mayo Clin.*, pp. 1-10, 2019 [En línea]. Disponible en: https://www.mayoclinic.org/es-es. [Accedido: 25-jun-2022]

[53] E. Etecé, «HTTP - Concepto, para qué sirve y cómo funciona», 2021. [En línea]. Disponible en: https://concepto.de/http/. [Accedido: 27-jun-2022]

[54] SNOMED, «SNOMED CT Release File Specifications - Release File Specification - SNOMED Confluence», 2022. [En línea]. Disponible en: https://confluence.ihtsdotools.org/display/DOCRELFMT?preview=/26837190/142123219/SNOMED CT Release File Specifications-v25-20220202_165741.pdf. [Accedido: 28-jun-2022]

[55] «5. Modelo lógico de SNOMED CT - SNOMED CT Starter Guide (ES) -

SNOMED Confluence». [En línea]. Disponible en: https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=61154172. [Accedido: 28-jun-2022]

[56] «COIT | Colegio Oficial Ingenieros de Telecomunicación». [En línea]. Disponible en: https://www.coit.es/. [Accedido: 29-jun-2022]

[57] «COITT | Colegio Oficial de Ingenieros Técnicos de Telecomunicación». [En línea]. Disponible en: https://www.telecos.zone/. [Accedido: 29-jun-2022]

PRESUPUESTO

En esta sección del documento se detalla la estimación monetaria para la realización del proyecto.

1. Desglose del presupuesto

El ámbito de este proyecto corresponde al Colegio Oficial de Ingenieros de Telecomunicación (COIT)[56]. Sin embargo, este organismo no interviene en la fijación de honorarios que establezcan los ingenieros de telecomunicación habilitados con sus contratantes, sólo se encarga del visado de trabajos profesionales. Por consiguiente, no tiene documentación respecto al desglose del presupuesto de un proyecto.

Con el objetivo de solucionar este problema, se ha decidido realizar la estimación monetaria del coste del proyecto redactado en este documento en base a la adaptación de las pautas establecidas por el Colegio Oficial de Ingenieros Técnicos de Telecomunicación (COITT)[57] hasta el año 2008. Dichas pautas establecen que para realizar el presupuesto de un proyecto se deben abarcar los siguientes apartados:

- Recursos materiales.
- Trabajo tarifado por tiempo empleado.
- Costes asociados a la redacción del documento.
- Derechos del visado del COIT.
- Gastos de tramitación y envío.
- Aplicación de impuestos.

2. Recursos materiales

Respecto a la estimación económica de un proyecto se encuentran los recursos materiales empleados durante el diseño, desarrollo, validación y ejecución de este. Estos recursos a su vez se subdividen en recursos *hardware* y *software*.

El cálculo del coste de amortización se estipula en un período de cuatro años. Suponiendo un sistema de amortización lineal en el que el material inmovilizado se deprecia de forma constante durante el período de tiempo evaluado, se obtiene que dicho coste de amortización es directamente proporcional a la diferencia de precio entre el valor del material inmovilizado en la fecha de su compra y el valor residual actual. Asimismo, es inversamente proporcional a los años de vida útil del material.

$$Coste\ de\ amortización = \frac{Valor\ de\ adquisición - Valor\ residual}{Años\ de\ vida\ útil}$$

Ecuación 7. Coste de amortización.

A causa de que el presente proyecto tiene una duración de trescientas horas distribuidas aproximadamente en cuatro meses, y siendo este período inferior a cuatro años, el coste de amortización son los derivados de los cuatro meses en los que se ha desarrollado el proyecto.

2.1. Recursos *software*

Las herramientas *software* empleadas para el desarrollo de este Trabajo de Fin de Máster son: PyCharm, Google Colab y el paquete de ofimática Microsoft Office 365.

PyCharm pertenece a la empresa JetBrains. Esta empresa cuenta con una oferta especial en la que los estudiantes y el personal académico pueden usar todas las herramientas de JetBrains de forma gratuita tras verificar el dominio de correo de la universidad o escuela universitaria. Google Colab es una herramienta completamente gratuita de Google y, respecto al programa Microsoft Office 365, se dispone de licencia para su uso al pertenecer a la comunidad universitaria de la ULPGC, por lo que no implica un coste de amortización asociado.

En resumen, debido a la condición de estudiante universitario, el coste total de los recursos *software* es de cero euros (0€).

2.2. Recursos *hardware*

Para el desarrollo del proyecto sólo se ha requerido el uso de un único dispositivo *hardware*, un ordenador portátil HP Omen 15-DC0004NS.

RECURSO	VALOR DE ADQUISICIÓN	VALOR RESIDUAL	COSTE DE AMORTIZACIÓN
Portátil HP Omen	756,06 €	230 €	43,84 €
		TOTAL	43,84 €

Tabla 23. Coste de amortización de los recursos hardware.

A partir de la

Tabla 23 se deduce que el coste total de recursos *hardware* utilizados en el proyecto es de cuarenta y tres euros con ochenta y cuatro céntimos (43,84€).

3. Trabajo tarifado por tiempo empleado

El desarrollo completo del proyecto ha supuesto una duración total de 300 horas, 4 meses. En este período se incluye desde la búsqueda y análisis de datos iniciales hasta la elaboración de la documentación.

Siguiendo las pautas establecidas por el COITT, el importe percibido por las horas trabajadas de un Ingeniero de Telecomunicaciones se calcula con la siguiente expresión:

$$H = C_t \cdot 74{,}88 \cdot H_n + 96{,}72 \cdot H_e$$

Ecuación 8. Cálculo de honorarios.

Donde:

- C_t indica un factor de corrección en función del número de horas trabajadas.
- H_n indica las horas trabajas dentro de la jornada laboral.
- H_e indica las horas trabajadas fuera de la jornada laboral.

Asimismo, según lo establecido por el COITT, el coeficiente de corrección tiene un valor variable en función del número de horas empleadas de acuerdo con la Tabla 24. Teniendo en cuenta que no se ha trabajado fuera del horario laboral, el factor de corrección C_t tiene un valor igual a 0,6, debido a que el proyecto tiene una duración comprendida entre las 180 y 360 horas.

HORAS EMPLEADAS (H_N)	**FACTOR DE CORRECCIÓN (C_T)**
Hasta 36	1
36 - 72	0,9
72 - 108	0,8
108 - 144	0,7
144 - 180	0,65
180 - 360	0,6
360 - 540	0,55
540 - 720	0,5
720 - 1080	0,45
Más de 1080	0,4

Tabla 24. Factor de corrección en función de las horas trabajadas.

Aplicando el factor de corrección y las horas empleadas a la Ecuación 8, se obtiene el siguiente resultado:

$$H = 0{,}6 \cdot 74{,}88 \cdot 300 + 96{,}72 \cdot 0 = 13.478{,}4\ €$$

Ecuación 9. Estimación de los honorarios.

Los honorarios totales por tiempo dedicado libres de impuestos ascienden a una cuantía de trece mil cuatrocientos setenta y ocho euros y cuarenta céntimos (13.478,40 €).

4. Costes derivados de la redacción del documento

El importe de la redacción del proyecto se calcula a partir de la Ecuación 10.

$$R = 0{,}07 \cdot P \cdot C_n$$

Ecuación 10. Coste de la redacción del documento.

Donde:

- P es el presupuesto del proyecto.
- C_n es el coeficiente de ponderación en función del presupuesto.

El coeficiente C_n está determinado por el presupuesto del proyecto, el que, hasta el momento, asciende a trece mil quinientos veintidós euros con veinticuatro céntimos (13.522,24€), incluyendo los honorarios y el costo de los recursos. Teniendo en consideración esta cifra, el COITT establece que para un presupuesto menor de 30.050 €, el coeficiente de ponderación tiene un valor de 1.

$$R = 0{,}07 \cdot 13.522{,}24 \cdot 1 = 946{,}56\ €$$

Ecuación 11. Estimación del coste de redacción.

Por lo tanto, el coste libre de impuestos derivado de la redacción del proyecto asciende a novecientos cuarenta y seis euros con cincuenta y seis céntimos (946,56 €).

5. Derechos de visado COIT

El Colegio Oficial de Ingenieros de Telecomunicación tiene un documento informativo de 2021 respecto a las tarifas de visado para los múltiples proyectos de diferentes ámbitos

que puede realizar un ingeniero de telecomunicación. En él establece que el cálculo de la tarifa de visado para un proyecto de aplicación telemática como es el desarrollado de este documento se calcula mediante la Ecuación 12.

$$V = 0{,}0035 \cdot P \cdot C$$

Ecuación 12. Tarifa de visado COIT.

Donde:

- P es el presupuesto de ejecución material.
- C es el coeficiente reductor en función del presupuesto del trabajo.

El presupuesto de ejecución material lo conforman el coste de los recursos, mano de obra y redacción del documento oficial del proyecto, al no haber costes asociados a material fungible por no ser necesaria la impresión del documento para su evaluación. Por consiguiente, como se puede ver a continuación, el precio de ejecución material del proyecto es de catorce mil cuatrocientos sesenta y ocho con ochenta céntimos (14.468,8€).

$$P = 43{,}84 + 13.478{,}4 + 946{,}56 = 14.468{,}8€$$

Ecuación 13. Estimación del presupuesto de ejecución material.

Respecto al coeficiente reductor, el COIT tiene una tabla que especifica los coeficientes reductores por tramos de presupuestos. Para presupuestos inferiores de 30.050€, como es en este caso, se establece un coeficiente reductor cuyo valor es 1. Aplicando los valores en la Ecuación 12 se obtiene el precio final del visado COIT para este proyecto.

$$V = 0{,}0035 \cdot 14.468{,}8 \cdot 1 = 50{,}64\ €$$

Ecuación 14. Estimación del coste del visado COIT.

El coste de los derechos de visado del proyecto asciende a cincuenta euros con sesenta y cuatro céntimos (50,64€).

6. Gastos de tramitación y envío

Los gatos de tramitación y envío de la documentación están estipulados en seis euros y un céntimo (6,01 €).

7. Aplicación de impuestos

Como se ilustra en la Tabla 25, previo a la aplicación del Impuesto General Indirecto Canario, que corresponde al 7%, el coste del proyecto es de catorce mil quinientos veinticinco euros con cuarenta y cinco céntimos (14.525,45 €).

RECURSO	COSTE
Recursos materiales	43,84 €
Trabajo tarifado por tiempo empleado	13.478,4 €
Costes asociados a la redacción del documento	946,56 €
Derechos de visado del COIT	50,64 €
Gastos de tramitación y envío	6,01 €
Subtotal	14.525,45 €
Aplicación de impuestos (IGIC 7%)	1.016,78 €
TOTAL	15.542,23 €

Tabla 25. Coste total del proyecto.

El presupuesto total de este Trabajo de Fin de Máster, tal y como se muestra en la tabla, asciende a la cuantía de quince mil quinientos cuarenta y dos euros y veintitrés céntimos (15.542,23 €).

Las Palmas de Gran Canaria a 9 de julio de 2022

Firma:

Yguanira del Pino Vega Vega

ANEXOS

En esta última sección del documento se incorpora la documentación adicional necesaria para una mayor compresión y entendimiento del proyecto.

1. *Datasets* principales

En este anexo, se describe el contenido de los *datasets* empleados para el entrenamiento y validación de los sistemas de recomendación. Para cada *dataset* se detalla cada una de las columnas que componen la tabla indicando:

- Nombre. Identificador de SNOMED CT que identifica el concepto al que hace referencia la columna.
- Término médico. Descripción legible del concepto de la columna.
- Descripción. Explicación de lo que representa la columna.
- Valores. Posibles valores que puede tomar la columna.

1.1. Categoría "general"

A continuación, se detalla el contenido del *dataset* principal de la categoría (Tabla 26) que, inicialmente, contine una tabla con un total de 8.638 filas y 127 columnas. Las filas corresponden con datos de pacientes y de las columnas, 126 corresponden a síntomas de enfermedades y la restante indica la patología que sufre el paciente. En total, hay casos de 42 patologías diferentes.

NOMBRE	TÉRMINO MÉDICO	DESCRIPCIÓN	VALORES
60728008	Abdomen distendido	Indica si el paciente percibe el abdomen lleno y apretado.	0: No 1: Sí -1: Desconocido
178960009	Acidez gástrica	Indica si el paciente sufre de dolor el pecho, una sensación de ardor ascendente, concretamente en la zona detrás del esternón.	
63102001	Alteración visual	Indica si el paciente presenta una deformación en sentido de la vista.	
12241791000119109	Ambos ojos rojos	Indica si el paciente presenta una coloración ocular rojiza.	
732984005	Ampolla	Indica si el paciente tiene una o varias lesiones palpables e inflamadas, llenas de líquido linfático y otros fluidos corporales en la piel.	

44169009	Anosmia	Indica si el paciente sufre una pérdida de olfato.	0: No 1: Sí -1: Desconocido
48694002	Ansiedad	Indica si el paciente padece episodios de preocupación excesiva.	
299308007	Articulación de la cadera dolorosa al movimiento	Indica si el paciente padece molestias en la cadera al realizar movimientos como caminar.	
88092000	Atrofia muscular	Indica si el paciente presenta disminución en el tamaño de los músculos.	
8943002	Aumento de peso	Indica si el paciente ha subido de peso.	
72405004	Aumento del apetito	Indica si el paciente presenta un aumento del deseo de comer.	
3716002	Bocio	Indica si el paciente presenta un aumento del tamaño de la glándula tiroides.	
55300003	Calambre	Indica si el paciente tiene contracciones repentinas e involuntarias de uno o más músculos.	
25064002	Cefalea	Indica si el paciente presenta dolor de cabeza de forma recurrente.	
81680005	Cervicalgia	Indica si el paciente sufre una molestia que afecta a la nuca y las vértebras cervicales.	
249294004	Chorro miccional interrumpido	Indica si el paciente presenta una orina intermitente.	
409766009	Cicatriz atrófica	Indica si el paciente tiene cicatrices que se sitúan por debajo de la superficie de la piel circundante de la cara.	
371632003	Coma	Indica si el paciente presenta un estado de pérdida de conocimiento prolongada.	
247467008	Comedón	Indica si el paciente presenta una lesión cutánea eruptiva primaria del acné.	
68235000	Congestión nasal	Indica si el paciente tiene la nariz taponada.	
125667009	Contusión	Indica si el paciente presenta una lesión en la piel la piel producida por una fuerza externa.	
271765003	Costra cutánea	Indica si el paciente presenta una cáscara producto del proceso de cicatrización.	
161882006	Cuello rígido	Indica si el paciente sufre una molestia en cualquiera de las estructuras del cuello (músculos, nervios o vértebras, entre otras).	
26544005	Debilidad muscular	Indica si el paciente sufre una pérdida de la fuerza muscular.	
713514005	Debilidad muscular de miembro	Indica si el paciente sufre de pérdida de fuerza en alguna extremidad corporal.	

271767006	Descamación cutánea	Indica si el paciente sufre un desprendimiento superficial, asintomático y espontáneo de la piel.	0: No 1: Sí -1: Desconocido
89362005	Descenso de peso	Indica si el paciente ha bajado de peso.	
249275009	Deseo de orinar	Indica si el paciente padece un aumento en el deseo de miccionar.	
34095006	Deshidratación	Indica si el paciente padece un exceso de pérdida de líquido corporal.	
62315008	Diarrea	Indica si el paciente sufre de evacuación intestinal con heces flojas y líquidas.	
230145002	Dificultad para respirar	Indica si el paciente presenta dificultades respiratorias.	
248493003	Dilatación de vaso sanguíneo	Indica si el paciente sufre un ensanchamiento de los conductos sanguíneos.	
3253007	Discromía cutánea	Indica si el paciente presenta modificaciones concretas o difusas en el color de la piel por exceso, defecto o ausencia de diversos pigmentos.	
267036007	Disnea	Indica si el paciente tiene una sensación de opresión en el pecho que le impide respirar profundamente.	
21522001	Dolor abdominal	Indica si el paciente presenta molestia en el abdomen.	
68653001	Dolor anal	Indica si el paciente presenta una molestia en zona del ano y el recto.	
57676002	Dolor articular	Indica si el paciente sufre molestias en una o varias articulaciones del cuerpo.	
225595004	Dolor asociado con la defecación	Indica si el paciente sufre molestias durante la evacuación intestinal.	
161891005	Dolor de espalda	Indica si el paciente padece molestia en la parte posterior del torso.	
271681002	Dolor de estómago	Indica si el paciente padece de molesticas estomacales.	
1003722009	Dolor de la región de rodilla	Indica si el paciente sufre una molestia en la rodilla.	
29857009	Dolor de pecho	Indica si el paciente presenta una molestia en la zona pectoral.	
41652007	Dolor en el ojo	Indica si el paciente padece molestia ocular.	
21005005	Dolor epigástrico urente	Indica si el paciente sufre una molestia severa en la parte superior del abdomen.	
426899007	Dolor provocado por caminar	Indica si el paciente presenta molestias al desplazarse.	
58250006	Dolor urente durante la micción	Indica si el paciente padece de molestias al orinar.	

15803009	Dolor vesical	Indica si el paciente tiene molestias en la vejiga.	0: No 1: Sí -1: Desconocido
405729008	Enterorragia	Indica si el paciente presenta una hemorragia en el intestino debido a que sus heces tienen una tonalidad rojiza o negruzca.	
816994007	Eritema de piel nasal	Indica si el paciente sufre una coloración rojiza en la piel de la nariz.	
307233002	Erosión gástrica hemorrágica	Indica si el paciente presenta una lesión de la mucosa gástrica con sangrado.	
271834000	Eructo	Indica si el paciente tiene episodios de evacuación de aire por el orificio bucal.	
271807003	Erupción cutánea	Indica si en paciente presenta inflamación en la piel.	
724877007	Erupción cutánea localizada	Indica si en paciente presenta inflamación en la piel en una zona específica del cuerpo.	
43724002	Escalofríos	Indica si el paciente presenta escalofríos, sensación de contracciones y relajaciones musculares rápidas.	
270031000	Esputo con sangre	Indica si el paciente tiene expectoraciones con presencia de sangre.	
24816000	Esputo herrumbroso	Indica que el paciente tiene una secreción pulmonar purulenta con restos hemáticos.	
64503007	Esputo mucoide	Indica que el paciente tiene una secreción pulmonar con mocos.	
419284004	Estado mental alterado	Indica si el paciente presenta un estado mental confuso, aparentemente distraída y actuando de forma extraña.	
76067001	Estornudo	Indica si el paciente tiene estornudo o tiene episodios de estornudos.	
14760008	Estreñimiento	Indica si el paciente presenta una evacuación intestinal difícil o dolorosa con heces duras o secas.	
84229001	Fatiga	Indica si el paciente presenta falta de energía y motivación.	
386661006	Fiebre	Indica si el paciente tiene una temperatura corporal entre 38 – 41 °C.	
304213008	Fiebre baja	Indica si el paciente tiene una temperatura corporal entre 37,2 – 38 °C.	
103605005	Grano	Indica si el paciente presenta pequeñas cavidades superficiales en la piel llenas con pus.	
289195008	Habla incomprensible	Indica si el paciente tiene dificultad para efectuar una correcta comunicación por voz.	

20022000	Hemiparesia	Indica si el paciente presenta una reducción de la fuerza motora o parálisis parcial en las extremidades superior e inferior de un lado del cuerpo.	0: No 1: Sí -1: Desconocido
409702008	Hiperpirexia	Indica si el paciente tiene una temperatura corporal superior a 41 °C.	
21639008	Hipervolemia	Indica si el paciente presenta un aumento irregular del volumen de plasma en el organismo.	
246975001	Ictericia conjuntival	Indica si el paciente presenta una decoloración ocular amarillenta.	
282299006	Incapaz de mantener el equilibrio	Indica si el paciente presenta problemas de equilibrio corporal.	
162031009	Indigestión	Indica si el paciente sufre una sensación de incomodidad o ardor en la parte superior del abdomen.	
271713000	Inestabilidad general	Indica si el paciente presenta episodios de mareo junto a una pérdida de equilibrio.	
162221009	Inquietud motriz	Indica si el paciente presenta un aumento de movimientos o inquietud.	
197270009	Insuficiencia hepática aguda	Indica si el paciente sufre una pérdida acelerada de la función del hígado.	
162400007	Irritación de garganta	Indica si el paciente sufre inflamación de la faringe.	
18963009	Labilidad anímica	Indica si el paciente presenta inestabilidad emocional.	
248181001	Labios secos	Indica si el paciente presenta una deshidratación y agrietamiento de la zona labial.	
214264003	Letargo	Indica si el paciente sufre un estado de somnolencia prolongada.	
30746006	Linfadenopatía	Indica si el paciente presenta un aumento en el tamaño de los ganglios.	
367391008	Malestar general	Indica si el paciente sufre una sensación de incomodidad o molestia que afecta a todo el cuerpo.	
271584002	Manos frías	Indica si el paciente presenta una temperatura corporal baja en sus manos.	
404640003	Mareo	Indica si el paciente sufre una sensación de desvanecimiento, atontamiento, debilidad o inestabilidad.	
68962001	Mialgia	Indica si el paciente padece dolor muscular.	
52024008	Moco de la vía respiratoria superior	Indica si el paciente presenta mocos en la cavidad nasal.	
422587007	Náuseas	Indica si el paciente tiene sensación de vomitar.	

102660008	Nivel anormal de glucosa	Indica si el paciente tiene un nivel de azúcar irregular en sangre.	0: No 1: Sí -1: Desconocido
414916001	Obesidad	Indica si el paciente tiene una cantidad excesiva de grasa corporal.	
420103007	Ojo lloroso	Indica si el paciente sufre un lagrimeo excesivo y constante.	
246923005	Ojos hundidos	Indica si el paciente sufre una alteración ocular que propicia que sea menos prominente de lo normal.	
445506004	Oniquia	Indica si el paciente presenta una inflamación del lecho ungueal producida por hongos o bacterias.	
690701000119101	Orina con olor fétido	Indica si el paciente tiene una micción olorosa.	
167232003	Orina oscura/concentrada	Indica si el paciente tiene una micción con un alto nivel de opacidad.	
80313002	Palpitaciones	Indica si el paciente sufre sensaciones desagradables en los latidos cardíacos propios que se perciben como si el corazón estuviera latiendo con violencia o acelerado.	
249403003	Parches blanquecinos en la mucosa oral	Indica si el paciente presenta lesiones de color blanco en la parte interna de la boca.	
302772006	Parece enfermo	Indica si el paciente presenta una apariencia enfermiza.	
79890006	Pérdida del apetito	Indica si el paciente presenta una disminución del deseo de comer.	
225549006	Piel amarilla	Indica si el paciente presenta un tono de piel amarillenta.	
271585001	Pies fríos	Indica si el paciente presenta un descenso de temperatura en los pies.	
72768000	Placa eritematosa	Indica si el paciente presenta vesículas flácidas y ampollas sobre base eritematosa.	
26329005	Poca concentración	Indica que el paciente sufre una pérdida de atención al realizar una actividad.	
28442001	Poliuria	Indica si el paciente sufre un exceso de producción de orina.	
251365002	Presión nasal	Indica si el paciente presenta sensación de opresión en las cavidades nasales.	
418290006	Prurito	Indica si el paciente presenta un hormigueo peculiar o irritación incómoda de la piel que produce el deseo de rascar la zona afectada.	
90446007	Prurito del ano	Indica si el paciente presenta picor en la zona anal.	
84445001	Rigidez articular	Indica si el paciente presenta una sensación de movilidad limitada o dificultosa en una o varias articulaciones.	

64531003	Rinorrea	Indica si el paciente presenta un flujo excesivo de líquido por la nariz.	0: No 1: Sí -1: Desconocido
62507009	Sensación de hormigueo	Indica si el paciente tiene una sensación comparable a hormigas recorriendo su piel, causada por el frío, una contusión o la compresión sobre un nervio.	
55929007	Sentimiento de irritabilidad	Indica si el paciente tiene un estado emocional irascible, con un temperamento explosivo y se molesta o enoja fácilmente.	
249472009	Siempre con hambre	Indica si el paciente presenta un deseo excesivo de ingesta de alimentos.	
279044000	Síndrome de dolor corporal total	Indica si el paciente sufre de molestia corporal completa.	
248480008	Síntoma de tumefacción de la pierna	Indica si el paciente presenta una inflamación en la pierna.	
415691001	Sudoración	Indica si el paciente sufre un exceso de transpiración.	
3424008	Taquicardia	Indica si el paciente presenta una frecuencia cardíaca acelerada.	
49727002	Tos	Indica si el paciente presenta tos o episodios de tos.	
69192004	Traquioniquia	Indica si el paciente presenta un aumento de las estriaciones longitudinales, depresiones y puntilleo de la lámina ungueal.	
35489007	Trastorno depresivo	Indica si el paciente presenta un estado de ánimo deprimido y/o pérdida interés en actividades que antes se disfrutaban.	
386804004	Trastorno menstrual	Indica si el paciente presenta un período menstrual anormal.	
271771009	Tumefacción articular	Indica si el paciente presenta una inflamación en una o varias articulaciones del cuerpo.	
15633761000119102	Tumefacción de ambas extremidades superiores	Indica si el paciente sufre una inflamación de las extremidades corporales superiores.	
762898005	Tumefacción de ambos miembros inferiores	Indica si el paciente presenta una inflamación de las extremidades corporales inferiores.	
278528006	Tumefacción facial	Indica si el paciente presenta una inflamación en zona de la cara.	
66123000	Úlcera de la lengua	Indica si el paciente presenta llagas pequeñas y superficiales en la lengua.	
89704006	Uñas punteadas	Indica si el paciente tiene pequeñas depresiones en la superficie de las uñas.	

234057004	Venas varicosas perforantes de la pantorrilla	Indica si el paciente presenta venas retorcidas y agrandadas, que conectan al sistema venoso superficial con el sistema venoso profundo, en la zona de la pantorrilla.	0: No 1: Sí -1: Desconocido
246636008	Visión borrosa	Indica si el paciente sufre una disminución de la agudeza visual.	
422400008	Vómitos	Indica si el paciente ha vomitado o presenta episodios de vómito, expulsión del contenido estomacal.	
439401001	Diagnóstico	Indica la enfermedad que padece el paciente.	SCTID (Tabla 27)

Tabla 26. Características del dataset 160237006.

SCTID	TÉRMINO MÉDICO
11381005	Acné
3723001	Artritis
396275006	Artrosis
195967001	Asma
33688009	Colestasis
840539006	COVID-19
38362002	Dengue
609328004	Disposición alérgica
235595009	Enfermedad por reflujo gastroesofágico
387800004	Espondilosis cervical
4834000	Fiebre tifoidea
25374005	Gastroenteritis
235855004	Hemorroides interna-externa

SCTID	TÉRMINO MÉDICO
235875008	Hepatitis alcohólica
40468003	Hepatitis viral tipo A
66071002	Hepatitis viral tipo B
50711007	Hepatitis viral tipo C
707341005	Hepatitis viral tipo D
7111000119109	Hepatitis viral tipo E
34486009	Hipertiroidismo
302866003	Hipoglucemia
40930008	Hipotiroidismo
18165001	Ictericia
48277006	Impétigo
22298006	Infarto de miocardio
68566005	Infección urinaria

3218000	Micosis
233604007	Neumonía
61462000	Paludismo
128188000	Parálisis cerebral
9014002	Psoriasis
62014003	Reacción adversa causada por fármaco
82272006	Resfriado común
62479008	SIDA (Síndrome de Inmuno Deficiencia Adquirida)
405751000	Tipo de diabetes

37796009	Migraña
3219008	Tipo Y/O categoría de enfermedad desconocida
38341003	Trastorno hipertensivo arterial sistémico
56717001	Tuberculosis
13200003	Úlcera péptica
38907003	Varicela
128060009	Várices venosas
111541001	Vértigo posicional paroxístico benigno

Tabla 27. Posibles valores de la columna 439401001 del dataset 160237006.

1.2. Categoría “covid19”

En la Tabla 28, se describe el contenido del *dataset* principal de la categoría específica “covid19” que, inicialmente, contine una tabla con un total de 223.071 filas y 28 columnas. Las filas corresponden con datos de pacientes y las columnas corresponden con parámetros relevantes para la evolución del cuadro médico de un paciente.

SCTID	TÉRMINO MÉDICO	DESCRIPCIÓN	VALORES
161615003	Antecedente de cirugía	Indica si al paciente se le ha realizado algún tipo de cirugía con anterioridad.	0: No 1: Sí -1: Desconocido
195967001	Asma	Indica si el paciente padece Asma durante el tiempo que sufre COVID-19.	
86049000	Cáncer	Indica si el paciente padece algún tipo de cáncer durante el tiempo que sufre COVID-19.	
397669002	Edad	Indica la edad del paciente.	Rango numérico (0 - 120)

289908002	Embarazo	Indica si el paciente se encuentra en un estado de gestación.	0: No 1: Sí -1: Desconocido
62914000	Enfermedad cerebrovascular	Indica si el paciente padece alguna patología que afecte a la vasculatura cerebral durante el tiempo que sufre COVID-19.	
80690008	Enfermedad degenerativa del sistema nervioso central	Indica si el paciente padece alguna patología que afecte de forma progresiva al sistema nervioso durante el tiempo que sufre COVID-19.	
40733004	Enfermedad infecciosa	Indica si el paciente padece alguna patología infecciosa durante el tiempo que sufre COVID-19.	
413839001	Enfermedad pulmonar crónica	Indica si el paciente padece alguna patología que afecte a los pulmones de por vida durante el tiempo que sufre COVID-19.	
13645005	Enfermedad Pulmonar Obstructiva Crónica	Indica si el paciente sufre EPOC durante el tiempo que sufre COVID-19.	
709044004	Enfermedad renal crónica	Indica si el paciente padece alguna patología que afecte a los riñones de por vida durante el tiempo que sufre COVID-19.	
840546002	Exposición a COVID-19	Indica si el paciente ha estado en contacto con otro paciente de COVID-19.	
77176002	Fumador	Indica si el paciente tiene el hábito de fumar.	
263495000	Género	Indica la estructura corporal del paciente.	SCTID (Tabla 29) -1: Desconocido
38013005	Inmunosupresión	Indica si el paciente sufre una supresión o disminución de las reacciones inmunitarias debido a la administración de fármacos.	0: No 1: Sí -1: Desconocido
419620001	Muerte	Indica si el paciente ha fallecido.	
233604007	Neumonía	Indica si el paciente padece neumonía durante el tiempo que sufre COVID-19.	
262007003	No intubado	Indica si el paciente no ha sido sometido al proceso de intubación.	
414916001	Obesidad	Indica si el paciente padece obesidad durante el tiempo que sufre COVID-19.	
116154003	Paciente	Indica donde ha sido tratado el paciente.	SCTID (Tabla 29) -1: Desconocido

76571007	Shock séptico	Indica si el paciente padece hipotensión arterial peligrosa durante el tiempo que sufre COVID-19.	0: No 1: Sí -1: Desconocido
67782005	Síndrome de Distrés Respiratorio Agudo	Indica si el paciente padece SDRA durante el tiempo que sufre COVID-19.	
445320007	Tiempo de supervivencia	Indica los días que transcurren desde los primeros síntomas de COVID-19 en el paciente hasta su fallecimiento.	Rango numérico (0 - 120)
405751000	Tipo de diabetes	Indica si el paciente padece algún tipo de diabetes durante el tiempo que sufre COVID-19.	0: No 1: Sí -1: Desconocido
397821002	Traslado del paciente a la UCI	Indica si el paciente ha sido trasladado a la Unidad de Cuidados Intensivos.	
49601007	Trastorno del aparato cardiovascular	Indica si el paciente padece alguna patología que afecte al sistema cardiovascular durante el tiempo que sufre COVID-19.	
235856003	Trastorno hepático	Indica si el paciente padece alguna patología que afecte al hígado durante el tiempo que sufre COVID-19.	
38341003	Trastorno hipertensivo arterial sistémico	Indica si el paciente padece alguna hipertensión durante el tiempo que sufre COVID-19.	

Tabla 28. Características del dataset 840539006.

	SCTID	TÉRMINO MÉDICO
263495000	10052007	Estructura corporal masculina.
	1086007	Estructura corporal femenina.
116154003	373864002	Paciente de ambulatorio.
	416800000	Paciente internado en hospital.

Tabla 29. Posibles valores de las columnas 263495000 y 116154003 del dataset 840539006.

2. Funciones adicionales

En este anexo se incluyen las funciones realizadas en Python para la comprobación del funcionamiento y validación de una Red Bayesiana.

En total son dos funciones, la primera de ellas es *showPosterior*, que obtiene las probabilidades de un nodo teniendo evidencias previas o no. La segunda es *systemEvaluation*, obtiene los parámetros de validación de una Red Bayesiana: sensibilidad, especificidad y precisión.

```
def showPosterior(model: gum.BayesNet, target: str, evs: dict, cutoff_prob=1.0) -> dict[str, int]:
    """ Generates the prediction for the indicated parameter using the Bayesian network model passed by
    parameters and the evs.

    :param model: Bayesian network trained.
    :param target: The term for which the prediction is to be made.
    :param evs: Evidences for prediction.
    :param cutoff_prob: Lower cut-off probability.
    :return: Prediction, list of possible values for the indicated prediction term whose probability in percent is
    higher than the indicated cut-off.
    """

    # The exact inference of the Bayesian network passed by parameter is calculated.
    ie = gum.LazyPropagation(model)

    # The inference is created without evidence.
    ie.makeInference()

    # Evidence is incorporated into the inference.
    ie.setEvidence((evs, {})[evs is None])

    # A reference to the posterior probability of the node is obtained.
    items = ie.posterior(target).variablesSequence()[0]

    # An array is created where the prediction data will be stored.
    prediction = {}

    # For each of the probabilities obtained, it is checked whether it is higher than the cut-off probability passed by
    # parameter and stored as prediction.
    for index in range(len(items.labels())):
        prob = round(ie.posterior(target)[index] * 100, 2)
        item = items.label(index)
        if prob > cutoff_prob:
            prediction[item] = prob

    # The prediction data is returned.
    return prediction
```

Figura 58. Función showPosterior.

```
def systemEvaluation(model: gum.BayesNet, testdf: str, target: str, auc: object = 0.5) -> None:
    """ This function obtains the validation parameters (sensitivity, specificity and accuracy) of a Bayesian Network
    and displays them in the terminal.

    :param model: Trained Bayesian network.
    :param testdf: Path to the file containing the data to be used to validate the network.
    :param target: Network node for which the validation parameters are to be obtained.
    :param auc: Probability cut-off for the confusion matrix.
    """

    def init_belief(ei: gum.LazyPropagation) -> None:
        """ It is responsible for initialising all nodes in the network with an evidence.

        :param ei: Exact inference of the Bayesian network, network with a priori probabilities.
        """

        # Each of the nodes that make up the Bayesian network of the inference
        # engine is traversed and for those different from the node to be predicted,
        # an initial evidence is added with value '0'.
        for var in ei.BN().names():
            if var != target:
                ei.addEvidence(var, 0)

    def update_beliefs(ei: gum.LazyPropagation, bn: gum.BayesNet, row: object) -> None:
        """ It is responsible for set the evidences to all nodes in the network except for the node to be predicted.

        :param ei: Exact inference of the Bayesian network, network with a priori probabilities.
        :param bn: Trained Bayesian network.
        :param row: Row to be predicted.
        """

        # Each of the nodes that make up the Bayesian network passed by parameter is traversed.
        for var in bn.names():

            # If the node is the one for which the prediction is to be made, skip to
            # the next one.
            if var == target:
                continue

            try:

                # The value of the row corresponding to the node is obtained.
                idx = bn.variable(var).index(str(row.to_dict()[var]))

                # The evidence is updated with the new value.
                ei.chgEvidence(var, idx)

            except gum.NotFound:
                # This can happend when value is missing is the test base.
                pass

        # Calculations are performed to obtain the predictions for the selected node.
        ei.makeInference()

    def is_well_predicted(ei: gum.LazyPropagation, bn: gum.BayesNet, row: object) -> str:
        """ This function is in charge of predicting the value of the node for each row of the dataset and checking if
        the predicted value is equal to the existing one in the dataset and with what probability.

        :param ei: Exact inference of the Bayesian network, network with a priori probabilities.
        :param bn: Trained Bayesian network.
        :param row: Row to be predicted.
        :return: The category within the confusion matrix in which the result obtained fits is returned.
        """
```

```
    # A variable is created to store the predicted value.
    global outcome_predicted

    # The correct prediction value is obtained.
    outcome = str(row.to_dict()[target])

    # Evidence is updated for each row of the test dataset.
    update_beliefs(ei, bn, row)

    # The posterior of a node is calculated and returned.
    marginal = ei.posterior(target)

    # The predicted values are obtained.
    outcomes_predicted = marginal.variablesSequence()[0]

    # A variable is created to store the most likely prediction.
    max_prob = 0.0

    # Each predicted value is run through to obtain its probability.If the probability of the predicted value is
    # greater than the maximum, it is updated with the new values.
    for idx in range(len(outcomes_predicted.labels())):
        prob = marginal[idx]
        if prob > max_prob:
            max_prob = prob
            outcome_predicted = str(outcomes_predicted.label(idx))

    # If the predicted value is correct or not, it is classified as positive or negative. Furthermore, depending on
    # the probability threshold and whether it is positive or negative it will be classified as true or false.
    if outcome == outcome_predicted:
        if marginal.toarray()[1] < auc:
            return "True Positive"
        else:
            return "False Negative"
    else:
        if marginal.toarray()[1] >= auc:
            return "True Negative"
        else:
            return "False Positive"

# The exact inference of the Bayesian network passed by parameter is calculated.
ei = gum.LazyPropagation(model)

# The values of the network nodes are initialised.
init_belief(ei)

# The node for which the prediction is to be made is indicated.
ei.addTarget(target)

# The validation data is obtained from the file.
df_te = pd.read_csv(testdf)

# For each row of the test dataset its prediction is calculated and the
# classification of the result is obtained.
result = df_te.apply(lambda row: is_well_predicted(ei, model, row), axis=1)

# The values of each category in the confusion matrix are counted and displayed.
tp = sum(result.map(lambda x: 1 if x == "True Positive" else 0))
fp = sum(result.map(lambda x: 1 if x == "False Positive" else 0))
tn = sum(result.map(lambda x: 1 if x == "True Negative" else 0))
fn = sum(result.map(lambda x: 1 if x == "False Negative" else 0))

print("Confusion matrix, category count:")
print(" - True positives: ", tp)
```

```
print(" - False positives: ", fp)
print(" - True negatives: ", tn)
print(" - False negatives: ", fn, "\n")

# The evaluation parameters are calculated and displayed.
if (tp + fn) != 0:
    print("Sensitivity = {:f}".format(tp / (tp + fn)))
if (tn + fp) != 0:
    print("Specifity = {:f}".format(tn / (tn + fp)))
if (tp + fp) != 0:
    print("Precision = {:f}".format(tp / (tp + fp)))
```

Figura 59. Función systemEvaluation.

3. *Dataset* de recomendación de enfermedades

En este anexo se describe el contenido de las dos versiones del *dataset* de recomendación de enfermedades desarrollado para mostrar los resultados del Sistema de Recomendación de la categoría "general".

COLUMNA	DESCRIPCIÓN	VALORES
disease	Nombre de la enfermedad.	Tabla 27
snomedID	Identificador de SNOMED CT para la enfermedad.	
url	Enlace a la página de referencia para obtener los datos de la enfermedad.	Enlace a la página web de la que se han extraído los datos del resto de columnas.
diagnosticTest1	Múltiples pruebas para confirmar el padecimiento de la enfermedad.	Nombre de pruebas médicas.
diagnosticTest2		
diagnosticTest3		
diagnosticTest4		
diagnosticTest5		
treatment1	Diferentes tratamientos para la enfermedad, ordenados según el nivel de gravedad de esta, siendo el tercero el más radical.	Nombre de tratamientos, procedimientos, medicinas, etc.
treatment2		
treatment3		
treatment side effects	Lista de efectos adversos de los diferentes tratamientos.	Síntomas
alternative medicine	Lista con procedimientos, actividades o tratamientos que emplean medicina natural.	Nombre de tratamientos, procedimientos, medicinas, etc.
life style and home remedies	Lista de acciones, remedios, actividades que puede hacer el paciente en su vida cotidiana para mejorar su estado.	Nombre actividades, remedios, etc.

coping and support	Lista de organizaciones o grupos de apoyo para pacientes que sufren la enfermedad.	Nombre de organizaciones o grupos de apoyo.

Tabla 30. Primera versión de las características del dataset diseaseRecommendations.

COLUMNA	DESCRIPCIÓN	VALORES
disease	Identificador de SNOMED CT de la enfermedad.	Valor/es SCTID separados mediante "|" o nulo ("").
diagnosticTest1	Múltiples pruebas para confirmar el padecimiento de la enfermedad.	
diagnosticTest2		
diagnosticTest3		
diagnosticTest4		
diagnosticTest5		
treatment1	Diferentes tratamientos para la enfermedad, ordenados según el nivel de gravedad de esta, siendo el tercero el más radical.	
treatment2		
treatment3		
treatmentSideEffects	Lista de efectos adversos de los diferentes tratamientos.	
alternativeMedicine	Lista con procedimientos, actividades o tratamientos que emplean medicina natural.	
lifeStyleAndHomeRemedies	Lista de acciones, remedios, actividades que puede hacer el paciente en su vida cotidiana para mejorar su estado.	
copingAndSupport	Lista de organizaciones o grupos de apoyo para pacientes que sufren la enfermedad.	

Tabla 31. Segunda versión de las características del dataset diseaseRecommendations.

4. Ejecución del servidor

Es este anexo se incluye el proceso de ejecución inicial de la aplicación del servidor de MEDvolution, tanto para cuando hay archivos nuevos como para cuando no.

En el caso de que haya archivos nuevos, la traza de ejecución es la siguiente:

```
===> Archivos de SNOMED CT.
Inicializando los archivos de SNOMED CT:
   ==> Cambio de extensión de los archivos originales de txt a csv.
          => Se han encontrado archivos con extensión txt de los siguientes idiomas:
                 -> Idioma: en (8 files)
                 -> Idioma: es (8 files)
          => Se estan cambiando los archivos de extensión.
          => Cambio completado con éxito.
   ==> Obteniendo los datasets.
          => Se han encontrado archivos en los siguientes idiomas:
                 -> Idioma: en (8 files)
                 -> Idioma: es (8 files)
          => Se están actualizando las variables que almacenaran los datasets.
          => Se han almacenado los datasets con éxito.

===> Datasets.
Inicializando los datasets:
   ==> Obtención de los datasets originales.
          => Categorías de datasets disponibles: 2.
                 -> Categoría: covid19 (3 datasets).
                 -> Categoría: general (1 datasets).
   ==> Estructuración de los datasets originales.
          => Formateo de los datasets de la categoría 'covid19'.
                 -> Aplicando formato al dataset 1.
                       - Formateo general realizado con éxito (Filas: 220657 | Columnas: 18).
                       - Preparando el dataset para su formateo al estándar SNOMED CT:
                            · Adaptación al estándar SNOMED CT realizada con éxito.
                       - Actualizando los dataframes principales con los datos del dataset:
                            · El dataframe principal de la categoría 'covid19' ha sido actualizado (Filas: 220657 | Columnas: 18)
                 -> Aplicando formato al dataset 2.
                       - Formateo general realizado con éxito (Filas: 1271 | Columnas: 9).
                       - Preparando el dataset para su formateo al estándar SNOMED CT:
                            · Adaptación al estándar SNOMED CT realizada con éxito.
                       - Actualizando los dataframes principales con los datos del dataset:
                            · El dataframe principal de la categoría 'covid19' ha sido actualizado (Filas: 221928 | Columnas: 18)
                            · El dataframe principal de la categoría 'general' ha sido actualizado (Filas: 1271 | Columnas: 8).
                 -> Aplicando formato al dataset 3.
                       - Formateo general realizado con éxito (Filas: 1143 | Columnas: 28).
                       - Preparando el dataset para su formateo al estándar SNOMED CT:
                            · Adaptación al estándar SNOMED CT realizada con éxito.
                       - Actualizando los dataframes principales con los datos del dataset:
                            · El dataframe principal de la categoría 'general' ha sido actualizado (Filas: 2414 | Columnas: 14).
                            · El dataframe principal de la categoría 'covid19' ha sido actualizado (Filas: 223071 | Columnas: 28)
          => Formateo de los datasets de la categoría 'general'.
                 -> Aplicando formato al dataset 1.
                       - Formateo general realizado con éxito (Filas: 4920 | Columnas: 129).
                       - Preparando el dataset para su formateo al estándar SNOMED CT:
                            · Adaptación al estándar SNOMED CT realizada con éxito.
                       - Actualizando los dataframes principales con los datos del dataset :
                            · El dataframe principal de la categoría 'general' ha sido actualizado (Filas: 7334 | Columnas: 127).
   ==> Guardado de los datasets que seran de entrenamiento para los sistemas.
          => El dataset principal '160237006' ha sido almacenado, ocupa 7.400134 MB y contiene 7334 filas y 7334 columnas.
          => El dataset principal '840539006' ha sido almacenado, ocupa 48.406535 MB y contiene 223071 filas y 223071 columnas.
```

```
===> Sistemas de recomendacion.
Inicialización de los Sistemas de Recomendación:
   ==> Inicialización de las Redes Bayesianas.
         => Realizando preprocesado de los datasets.
               -> Preprocesado realizado con éxito.
         => Iniciando Redes Bayesianas.
               -> Iniciando Red Bayesiana para los datos de síntomas de pacientes con su enfermedad correspondiente.
                     - Obteniendo los datos de entrenamiento y validación.
                          · Datos obtenidos y almacenados con éxito:
                              - Archivo de entrenamiento 160237006_tr ocupa 5.632 MB y contiene 5500 filas y 127 columnas.
                              - Archivo de validación 160237006_te ocupa 1.878016 MB y contiene 1834 filas y 127 columnas.
                     - Creando la Red Bayesiana.
                     - Entrenando el sistema.
                          · Sistema entrenado y modelo almacenado con éxito.
                     - Validando el sistema.
                            -  Confusion matrix, category count:
                                  True positives:  1804
                                  False positives:  0
                                  True negatives:  0
                                  False negatives:  30

                            -  Sensitivity = 0.983642
                            -  Precision = 1.000000
                     - Se eliminan los archivos de entrenamiento y validación.
                          · Archivos eliminados con éxito.
               -> Iniciando Red Bayesiana para los datos de pacientes con COVID-19.
                     - Obteniendo los datos de entrenamiento y validación.
                          · Datos obtenidos y almacenados con éxito:
                              - Archivo de entrenamiento 840539006_tr ocupa 38.814296 MB y contiene 167303 filas y 28 columnas.
                              - Archivo de validación 840539006_te ocupa 12.938176 MB y contiene 55768 filas y 28 columnas.
                     - Entrenando sistema.
                          · Sistema entrenado y modelo almacenado con éxito.
                     - Validando el sistema.

                          · Nodo '262007003':
                            -  Confusion matrix, category count:
                                  True positives:  0
                                  False positives:  0
                                  True negatives:  1565
                                  False negatives:  54203

                            -  Sensitivity = 0.000000
                            -  Specifity = 1.000000

                          · Nodo '397821002':
                            -  Confusion matrix, category count:
                                  True positives:  54330
                                  False positives:  1438
                                  True negatives:  0
                                  False negatives:  0

                            -  Sensitivity = 1.000000
                            -  Specifity = 0.000000
                            -  Precision = 0.974215

                          · Nodo '419620001':
                            -  Confusion matrix, category count:
                                  True positives:  55688
                                  False positives:  80
                                  True negatives:  0
                                  False negatives:  0

                            -  Sensitivity = 1.000000
                            -  Specifity = 0.000000
                            -  Precision = 0.998565

                          · Nodo '445320007':
                            -  Confusion matrix, category count:
                                  True positives:  49160
                                  False positives:  6608
                                  True negatives:  0
                                  False negatives:  0
```

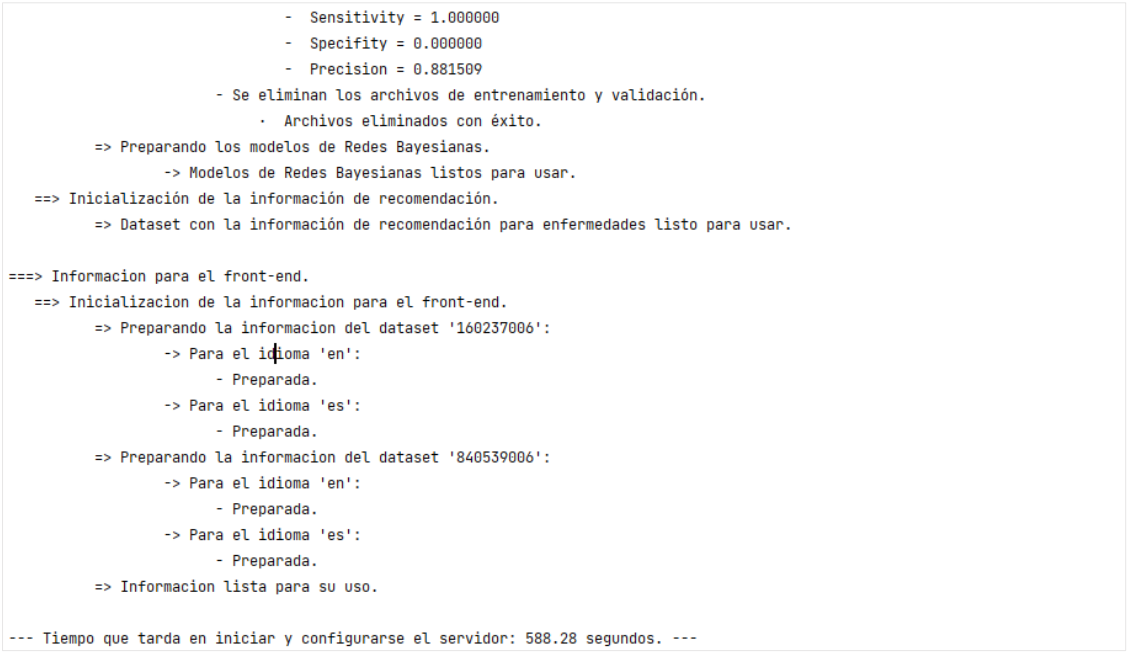

Figura 60. Ejecución del servidor con archivos nuevos.

Para cuando el servidor no cuenta con nuevos archivos, la traza de ejecución es la siguiente:

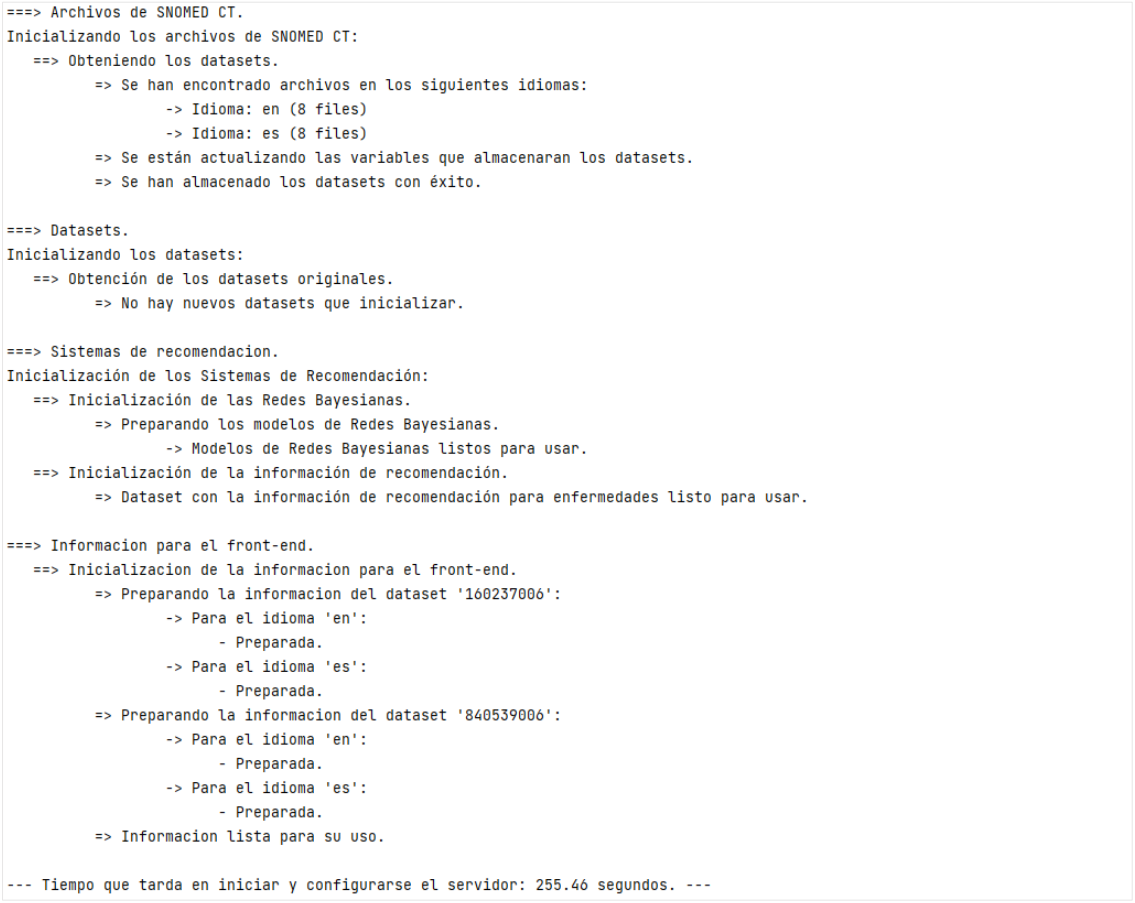

Figura 61. Ejecución del servidor sin archivos nuevos.

5. Pliego de condiciones

En este anexo se especifican las condiciones bajo las que se ha desarrollado este Trabajo de Fin de Máster, detallándose los requisitos *hardware* y *software* para el desarrollo y ejecución de este.

5.1. Requerimientos *software*

En lo que respecta a las características y especificaciones *software* empleadas, son las siguientes:

- Microsoft Windows 11 Home.
- PyCharm Professional 2021 3.3.
- Microsoft Office 2019.

5.2. Requerimientos *hardware*

Para este proyecto sólo se ha empleado un dispositivo hardware, un ordenador portátil con las especificaciones técnicas indicadas en la Tabla 32.

MODELO	HP OMEN 15-DC0004NS
FABRICANTE	HP
PROCESADOR	Intel® Core™ i5-8300H (frecuencia de base de 2,3 GHz, hasta 4 GHz con tecnología Intel® Turbo Boost)
MEMORIA RAM	SDRAM DDR4-2666 de 16 GB (1 x 16 GB)
CONTROLADOR GRÁFICO	1 TB 7200 rpm SATA + 256 GB PCIe® NVMe™ M.2 SSD
DISCO DURO	NVIDIA® GeForce® GTX 1050 (GDDR5 de 4 GB dedicados)

Tabla 32. Especificaciones del ordenador portátil empleado en el proyecto.