Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Modelos predictivos aplicados a trayectorias de pacientes que padecen diabetes mellitus.

Trabajo Fin de Grado

Grado en Ingeniería Informática

Autor: Alexis Gil Calatayud

Tutor: Juan Miguel García Gómez

Curso 2019-2020

Agradecimientos

Primero que todo, me gustaría agradecer a mi tutor Juan Miguel García Gómez por guiarme, ayudarme y apoyarme siempre que lo he necesitado. También quería destacar su dedicación e implicación en la realización del presente trabajo.

Por supuesto a mi familia, por haberme brindado la magnífica oportunidad de estudiar este grado, por el apoyo constante recibido a lo largo de estos años y por constituir la base sólida de mis logros.

A mis amigos y compañeros por prestarme su ayuda y apoyo cuando la situación lo requería. Sin ellos hoy no estaría donde estoy.

También quiero acordarme de todos los profesores que he tenido en cada curso por todo lo aprendido. Gran parte de los conocimientos que tengo hoy en día son gracias a ellos.

En resumen, agradecer a todas aquellas personas que han aportado su granito de arena para ser lo que soy hoy.

Resumen

En la actualidad, cada vez es más extendido el uso de las nuevas tecnologías en las ciencias de la vida. En particular, una herramienta que se está empezando a desarrollar en medicina son los modelos predictivos. Esto es debido a una mayor facilidad de acceso a la tecnología que tienen los profesionales de la medicina y a los prometedores resultados del análisis masivo de datos biomédicos.

La diabetes mellitus es una enfermedad compleja, crónica y generalmente evolutiva. Debido a la gran cantidad de personas que padecen esta incurable enfermedad, la diabetes ha derivado en un problema serio de salud pública, siendo una de las principales causas de muerte en el mundo.

De la búsqueda de información paralela a la proporcionada por los médicos que los pacientes realizan surgen diferentes fuentes de información como por ejemplo plataformas de comunicación entre pacientes o incluso las redes sociales. En concreto, la diabetes se sitúa entre las primeras enfermedades con mayor presencia en redes sociales. Hoy en día, los profesionales de la salud y las entidades son conscientes de esta tendencia y se han adaptado divulgando contenido informativo y educacional entre las diferentes fuentes de internet.

La explotación de estas fuentes de datos no estructuradas de información relacional masiva permitiría el descubrimiento de factores compuestos asociados a la evolución de enfermedades de larga duración, como la diabetes, aportando información complementaria a las historias clínicas asistenciales.

El objetivo de este proyecto es diseñar y desarrollar un módulo que permita la generación de redes de términos médicos estandarizados a partir de las conversaciones de foros de pacientes con diabetes. Con dicha red, queremos caracterizar los elementos más relevantes de la trayectoria clínica de la diabetes desde el punto de vista de los pacientes, y por lo tanto complementaria a la ruta asistencial.

Nuestro modelo consiste, básicamente, en una asociación múltiple de los términos médicos más relevantes y ocurrentes extraídos automáticamente de redes sociales mediante análisis de lenguaje natural. Dicha relación se ilustra mediante un grafo ponderado de ocurrencias. Todo esto se basa en vivencias que personas con diabetes han experimentado a lo largo de su vida como paciente y que, más tarde, han detallado en diferentes foros para diabéticos.

El impacto del módulo desarrollado es analizar y predecir relaciones entre factores reportados por pacientes con diabetes, que pueden ser de interés para anticipar complicaciones en la trayectoria de pacientes con diabetes mellitus.

Palabras clave: diabetes mellitus, modelo predictivo, grafo ponderado.

Abstract

Nowadays, the use of new technologies in the life sciences is becoming more widespread. In particular, a tool that is beginning to be developed in medicine is predictive models. This is due to a greater ease of access to technology that medical professionals have and to the promising results of the massive analysis of biomedical data.

Diabetes mellitus is a complex, chronic and generally evolutionary disease. Due to the large number of people who suffer from this incurable disease, diabetes has led to a serious public health problem, being one of the leading causes of death in the world.

From the search for information parallel to that provided by the doctors that the patients carry out, different sources of information arise, such as communication platforms between patients or even social networks.

The exploitation of these unstructured data sources of massive relational information would allow the discovery of compound factors associated with the evolution of long-lasting diseases, such as diabetes, providing complementary information to healthcare medical histories.

The objective of this project is to design and develop a module that allows the generation of networks of standardized medical terms from the conversations of diabetes patients' forums. We want to characterize the most relevant elements of the clinical trajectory of diabetes from the point of view of patients.

Our model basically consists of a multiple association of the most relevant and occurring medical terms automatically extracted from social networks through natural language analysis. This relationship is illustrated by a weighted graph of occurrences. All this is based on experiences that people with diabetes have experienced throughout their lives as patients and that they have later detailed in different forums for diabetics.

The impact of the module developed is to analyse and predict relationships between factors reported by patients with diabetes, that may be of interest to anticipate complications in the trajectory of patients with diabetes mellitus.

Keywords: diabetes mellitus, predictive model, graph, correlation.

Resum

En l'actualitat, cada vegada és més extens l'ús de les noves tecnologies en les ciències de la vida. En particular, una ferramenta que s'està començant a desenvolupar en medicina són els models predictius. Açò és degut a una major facilitat d'accés a la tecnología que tenen els professionals de la medicina i als prometedors resultats de l'anàlisi massiu de dades biomèdiques.

La diabetes mellitus es una enfermetat complexa, crónica i generalment evolutiva. Degut a la gran cantitat de persones que pateixen aquesta incurable enfermetat, la diabetes ha esdevingut un seriós problema de salut pública, sent una de les principals causes de mort al món.

De la cerca d'informació paral·lela a la proporcionada pels metges que els pacients realitzen sorgeixen diferents fonts d'informació, com per exemple plataformes de comunicació entre pacients o fins i tot les reds socials. Concretament, la diabetes es situa entre les primeres enfermetats amb major presencia en reds socials. Hui per hui, els professionals de la salut i les entitats són conscients d'aquesta tendencia i s'han adaptat divulgant contingut informatiu i educatiu per les diferents fonts d'internet.

L'explotació d'aquestes fonts de dades no estructurades d'informació relacional massiva permetria el descobriment de factors compostos associats a l'evolució d'enfermetats de llarga duració, com la diabetes, aportant informació complementària a les històries clíniques assistencials.

L'objectiu d'aquest projecte és dissenyar i desenvolupar un mòdul que permeta la generació de xarxes de termes mèdics estandaritzats a partir de les converses de fòrums per a persones amb diabetes. Amb aquesta xarxa, volem caracteritzar els elements més rellevants de la trajectòria clínica de la diabetes des del punt de vista dels pacients, i per tant complemetària a la ruta asistencial.

El nostre model consisteix, bàsicament, en una associació múltiple dels termes mèdics més rellevants i ocurrents trets automàticament de xarxes socials mitjançant anàlisi de llenguatge natural. Aquesta relació s'escenifica mitjançant un graf ponderat d'ocurrències. Tot açò es basa en vivències que persones amb diabetes han experimentat al llarg de la seua vida com a pacient i que, més tard, han detallat en diferents fòrums per a diabètics.

L’impacte del mòdul desenvolupat es analitzar i predir relacions entre factors reportats per pacients amb diabetes, que poden ser d’interés per a anticipar complicacions en la trajectòria de pacients amb diabetes mellitus.

Paraules clau: diabetes mellitus, model predictiu, graf, correlació.

Tabla de contenidos

Índice de Figuras

Índice de Tablas

1. Introducción

Hoy en día la tecnología y la medicina van cada vez más unidas de la mano. Poco a poco se van introduciendo los sistemas software de apoyo a los profesionales de la salud en el día a día de la vida clínica y esto beneficia tanto al médico en cuestión como al paciente.

La diabetes mellitus es una enfermedad que afecta a una buena parte de la población mundial. Una gran parte de estas personas no es consciente o no presta suficiente atención a sus niveles de azúcar en sangre y, sin embargo, esta enfermedad está muy presente y es más común de lo que la población cree.

Los modelos predictivos utilizan la estadística, el aprendizaje automático y el Big Data como base sólida para predecir resultados de ciertos problemas o situaciones. Evidentemente, estas predicciones no son para nada al azar. Detrás de ellas hay un análisis previo de un repositorio de datos históricos.

La disciplina del análisis predictivo aplicado a la medicina en el campo multidisciplinar está empezando a despegar, puesto que es algo relativamente nuevo para este ámbito. Vistos los resultados positivos de la aplicación de la predicción en dicho campo, recientemente se están empezando a hacer importantes inversiones para consolidar estos modelos, pudiendo llegar a ser la base de la medicina en un futuro cercano.

1.1 Diabetes

La diabetes mellitus es una enfermedad crónica y metabólica en la cual el cuerpo humano presenta unos niveles muy altos de glucosa en sangre. Este glúcido, en particular, es un monosacárido, es decir, un elemento de composición simple que entra en el organismo a través de los alimentos que ingerimos y que es necesario y fundamental para el correcto funcionamiento de las células del organismo. Para que el cuerpo humano pueda asimilar este azúcar, necesita de una hormona producida por el páncreas llamada insulina [12].

El páncreas, órgano situado en el abdomen (ver Figura 1), tiene como función principal la producción de hormonas. Una de las más importantes es la insulina. Dicha

hormona es la encargada de transportar la glucosa que entra en nuestro organismo hasta los músculos u otras células, para que pueda ser almacenada o utilizada en forma de energía.

Figura 1. Localización del páncreas [5].

Es en este punto donde distinguimos entre dos tipos de diabetes mellitus, dependiendo del comportamiento que adopta el cuerpo humano frente a esta hormona:

- En la diabetes tipo I, el páncreas no es capaz de producir insulina, de manera que la glucosa no entra en la célula y permanece en el flujo sanguíneo. A esto se le conoce como hiperglucemia. Aunque puede darse a cualquier edad, es más frecuente que se diagnostique en niños, adolescentes o adultos jóvenes y puede ser hereditaria [12].

- En la diabetes tipo II, el organismo no hace un uso adecuado de la insulina, es decir, las células que la procesan presentan resistencia a dicha hormona. Es la más común y aparece con mayor frecuencia en adultos mayores, personas con obesidad, personas cuyo historial familiar está relacionado con la enfermedad o personas sedentarias [12].

En términos menos formales, la insulina actúa como una puerta entre la glucosa presente en el corriente sanguíneo y las células. Si la puerta se encuentra cerrada, dicha hormona no puede pasar y, en consecuencia, no puede actuar. Por ello,

las personas con diabetes deben suministrar al cuerpo, de manera artificial, la dosis de insulina necesaria para regular el nivel de azúcar en la sangre.

1.1.1 Alcance de la enfermedad

Según el primer estudio mundial sobre la diabetes que realizó la Organización Mundial de la Salud (OMS) publicado en Abril de 2016 en el diario británico BBC [3], 1 de cada 11 personas en todo el mundo ya padece esta enfermedad. El estudio muestra una comparación del alcance de la enfermedad en 1980 y 2014, donde se puede observar que la cifra de diabéticos en todo el mundo se ha cuadruplicado. Más concretamente, ha pasado de 108 millones que había en 1980 a 422 millones en 2014.

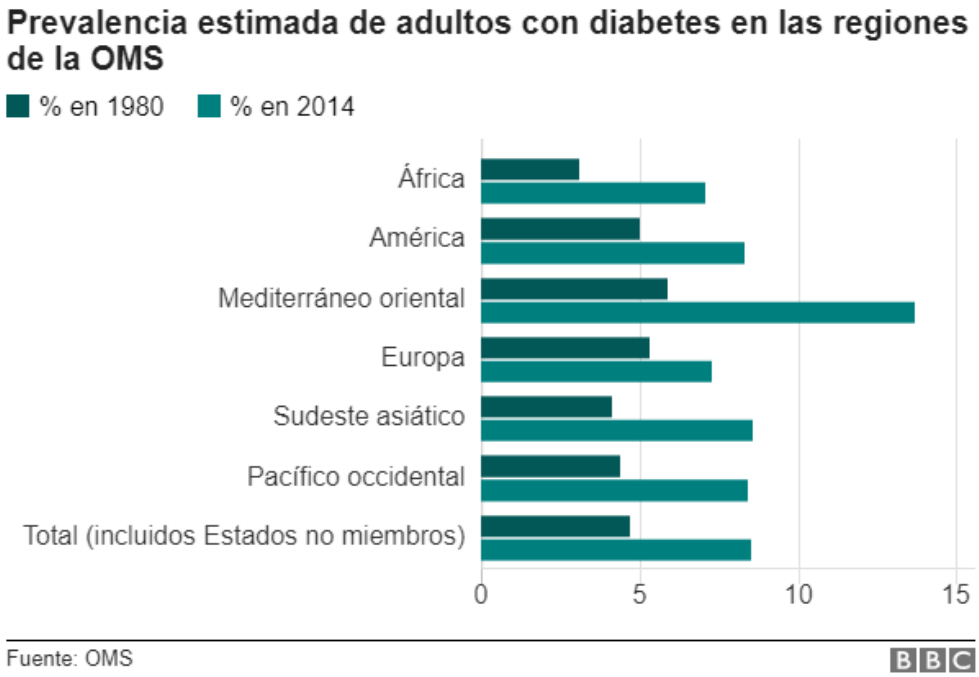

Figura 2. Diagrama del alcance de la enfermedad.

El director belga del Departamento de Gestión de enfermedades no transmisibles, discapacidad, violencia y prevención de enfermedades, Etienne Krug, aseguró que: “es una enfermedad silenciosa, pero su marcha está siendo implacable y tenemos que detenerla”. Varios científicos aseguran que la enfermedad aumenta más considerablemente en países de renta media y baja que en países cuya renta es elevada [3].

La diabetes se encuentra entre las diez primeras causas de muerte en el mundo, más concretamente se sitúa en el puesto número 8, causando 1,5 millones de muertes al año [3]. Es por ello que hay que prestar especial atención a los niveles altos de azúcar en sangre previos a un diagnóstico de la enfermedad.

1.1.2 Análisis de foros para diabéticos

En base a nuestra experiencia, podemos afirmar que los foros diabéticos están a la orden del día. Existen varios foros en los que la gente con diabetes participa activa y diariamente y de los cuales se nutre de información. Son muchos los *posts* que la gente publica y existen infinidad de hilos diferentes.

Hemos comprobado que, por ejemplo, pacientes que tienen más experiencia en esta enfermedad ayudan y aconsejan a los recién diagnosticados que buscan ayuda por esta vía. Los pacientes publican sus experiencias y es gracias a esto por lo que podemos realizar nuestro trabajo.

Nosotros, en particular, nos hemos centrado en uno de los foros más utilizados y que mayor índice de participación presenta [9]. También incluye noticias, publicaciones y guías de diferentes tópicos relacionados con la enfermedad.

1.2 Modelos predictivos

Un modelo predictivo es el resultado de un proceso de análisis de los datos históricos para, en base a estas observaciones, realizar una predicción sobre, normalmente, un suceso futuro. Estos modelos sientan su base en la parte de la estadística que utiliza la inferencia como procedimiento principal, es decir, a partir de una población estadística o muestra representativa llegar a determinar características, sacar conclusiones e inferir propiedades de todo el conjunto de la población [15].

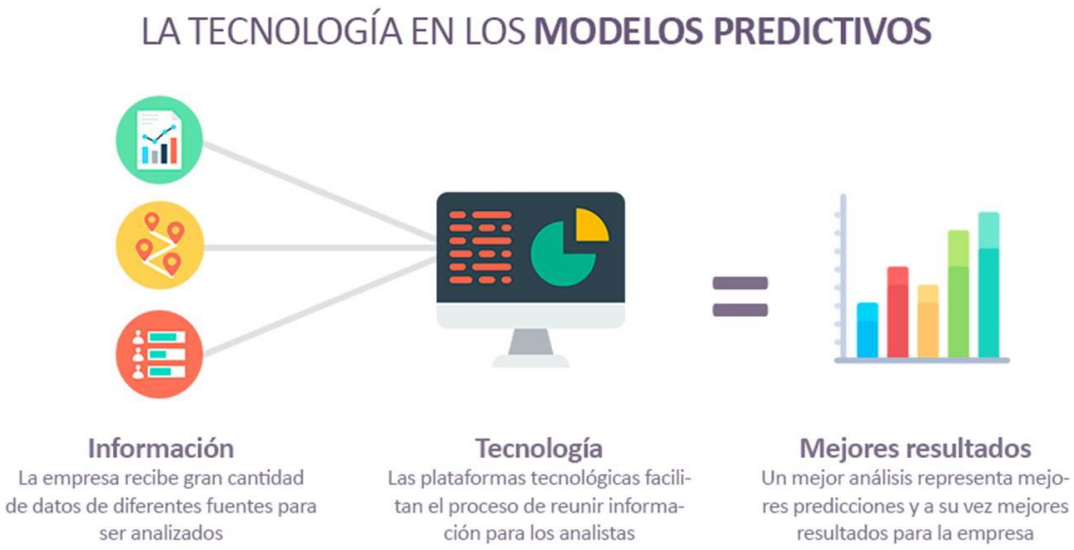

Figura 3. Esquema modelo predictivo [4].

Su objetivo es encontrar ciertos patrones y/o relaciones en grandes volúmenes de datos. Basan su técnica en observar y analizar eventos o sucesos del pasado para poder anticipar comportamientos o hechos futuros.

1.2.1 Impacto de los modelos predictivos

Como bien sabemos, muchos de los tratamientos iniciales que se les aplican a los pacientes no funcionan. En el caso de la diabetes, esto ocurre en un 43 % de los diagnosticados. Es por eso que hoy en día se está avanzando hacia un nuevo modelo de medicina moderna: la medicina de precisión. El Big Data y, por consiguiente, los modelos predictivos, representan una parte importante de la causa de la marcha hacia la nueva tendencia [1].

La aplicación de, en nuestro caso, los modelos predictivos a la vida clínica diaria tanto de pacientes como de médicos es clave para aumentar la eficacia y eficiencia de las consultas. Con esto, podemos llegar a prever y predecir ciertas complicaciones o comportamientos de una persona que padece diabetes.

Por un lado y como efecto colateral, también mejora la calidad de la sanidad pública, puesto que las consultas duran menos, son más efectivas y se reduce el uso desaprovechado de tratamientos que al principio parece que van a resultar eficaces para un determinado paciente y, posteriormente, no lo son. Por otro lado, también se reduce, a largo plazo, los costes económicos, puesto que se ahorra en medicamentos y recursos que no resultan ser efectivos.

1.3 Aproximación del proyecto a Minería de Datos y Ciencia de Datos

Cuando hablamos de análisis de datos nos encontramos ante un conjunto de términos relacionados como *Data Mining* (Minería de Datos), *Big Data* (datos a gran escala) y *Data Science* (Ciencia de Datos) entre otros [21].

La ciencia de datos combina campos como el análisis descriptivo o predictivo, la estadística, el aprendizaje automático (*machine learning*) e incluso la minería de datos. Se asocia con la curiosidad puesto que a partir de un cierto volumen de datos, que puede ser grande o no (*Big Data*), intenta extraer patrones, propiedades o conclusiones haciendo uso de técnicas como pueden ser modelos predictivos [6].

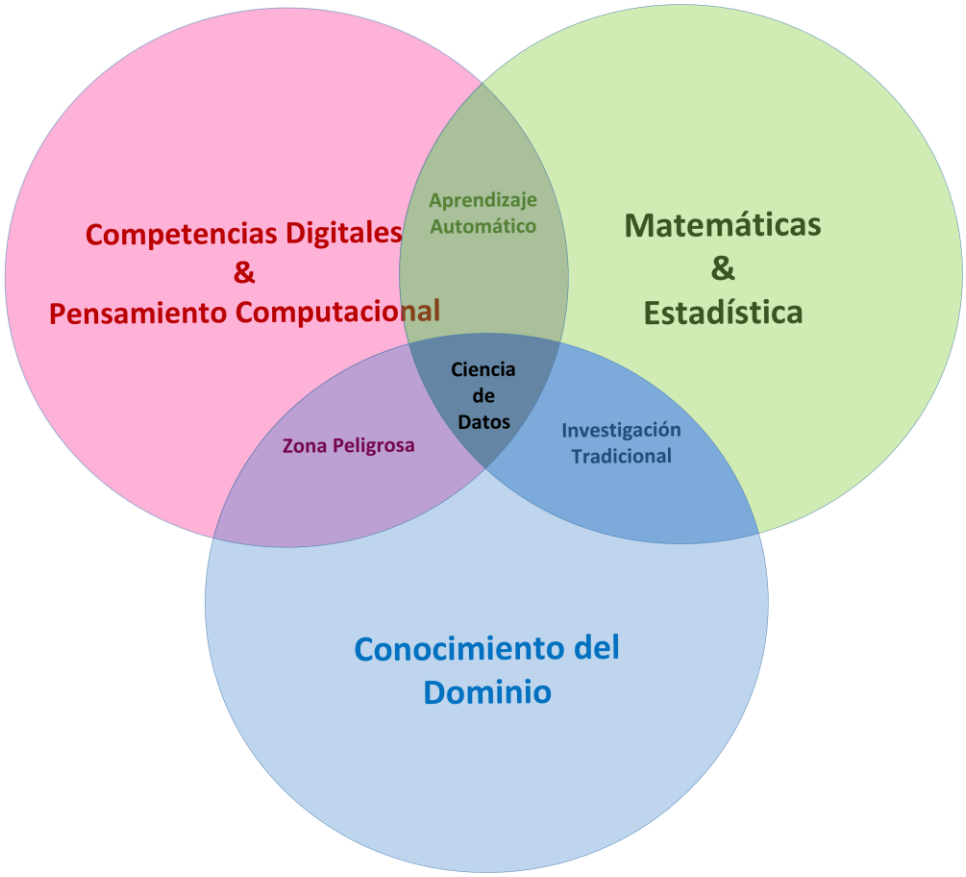

Figura 4. Ciencia de datos.

La minería de datos trata de encontrar patrones, anomalías o correlaciones en base a un gran volumen de datos organizado y estructurado en almacenes de datos la mayoría de las ocasiones. Este campo se enfoca desde un punto de vista empresarial, puesto que intenta satisfacer ciertos objetivos como pueden ser reducir costes de una determinada empresa o incrementar los beneficios de la misma [13].

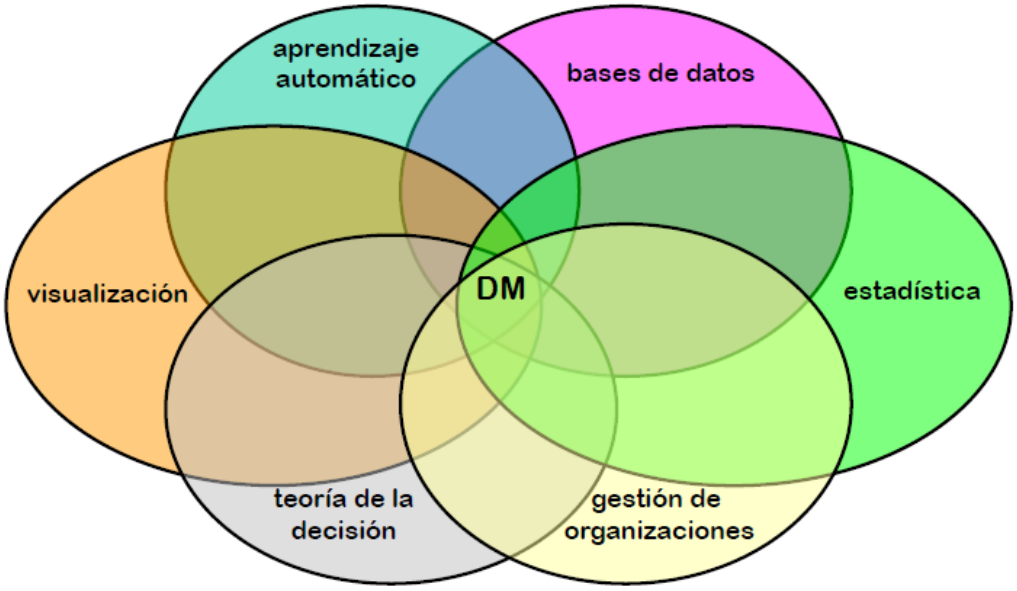

Figura 5. Minería de datos [22]***.***

Es posible tener muchos datos y no tener información. Los datos originales pueden ser insuficientes, incompletos, de baja calidad, erróneos, inadecuados, etc. También es posible encontrarlos desestructurados de tal forma que no sean fácilmente visibles y dificultando su entendimiento. Esto se da especialmente cuando la fuente de procedencia de los datos es de texto libre, como es el caso de este proyecto, puesto que la gente puede que no cuide su ortografía, cosa que complica el procesado y el análisis de los datos.

Después de aplicar a los datos iniciales procesos de limpieza para derivar en información, es decir, procesos que hacen visibles y entendibles los datos, el objetivo de estas técnicas es extraer conocimiento (modelos) a partir de dicha información. Los modelos, como en nuestro caso el modelo predictivo de red (grafo ponderado no dirigido) que se genera como resultado de todo el procedimiento, constituyen una herramienta de apoyo para la toma de decisiones [21].

Figura 6. Esquema minería de datos [22]**.**

Nuestro proyecto combina técnicas tanto de minería de datos como de ciencia de datos, puesto que extraemos información de gran cantidad de datos procedentes de fuentes de texto libre y transformamos esta información en conocimiento generando el modelo predictivo final pero sin un enfoque empresarial.

1.4 Redes sociales y análisis del lenguaje natural

Existe una tendencia ascendente en la comunidad de pacientes con diabetes y es la del uso de las nuevas vías de comunicación como son las redes sociales. Estas plataformas constituyen una herramienta de apoyo para estos pacientes puesto que pueden expresar sus vivencias y emociones sobre su patología. Además, haciendo un uso adecuado de las mismas, entidades y profesionales de la medicina pueden incluir contenido educativo e informativo, constituyendo así una fuente rica de información [8].

La Federación Española de Diabetes (FEDE) junto con Sanofi (grupo farmacéutico) organiza eventos con multitud de talleres de salud emocional para niños y adolescentes con diabetes. Los talleres cuentan con profesionales tanto de la enfermedad como de las redes sociales y el objetivo de estos es lograr que los jóvenes hagan un buen uso de estas plataformas para poder expresar sus sentimientos a la comunidad social. Un buen ejemplo de estos eventos es el SoyGeneraciónDiguan [8].

Otro evento de esta índole es el *Diabetes On Tweet*, un encuentro digital en España entre tuiteros, blogueros, profesionales de la salud, expertos en la patología y pacientes que dio lugar al primer decálogo para hablar e informar sobre esta enfermedad en redes sociales, compuesto por 10 puntos (ver Figura 7). De estos puntos destacan, por ejemplo, el de remarcar siempre que se interviene sobre qué tipo de diabetes se está hablando o dar evidencia científica de todas las informaciones que se difunden [2].

Figura 7. Decálogo informativo [2].

Para analizar contenido elaborado mediante lenguaje natural se distinguen dos procesos. En primer lugar, comprender el mensaje que se transmite. El grado de complejidad de esto depende de la vía por la cual se ha comunicado dicho mensaje y en qué idioma. La comunicación, en redes sociales, puede producirse de forma escrita o hablada y en multitud de idiomas.

En segundo lugar, la generación de datos estructurados y manipulables a partir del mensaje inicial. En el presente proyecto se analiza contenido en lenguaje natural escrito para generar relaciones entre términos extraídos de dicho contenido. Esta disciplina en particular presenta un grado de dificultad mayor puesto que los diferentes mensajes que se analizan pueden ser más o menos sofisticados y pueden presentar errores.

1.5 Motivación

Según un estudio reciente publicado por la Federación Internacional de Diabetes (FID) en 2017, en España hay alrededor de 3.5 millones de adultos que padecen diabetes, es decir, una prevalencia del 10.4%. Entendemos adultos como personas entre 20 y 79 años de edad [18].

También se ha apreciado una tendencia demográfica ascendente de 2011 a 2013 en cuanto al número de personas que sufren esta enfermedad, pero que parece estabilizarse más tarde, tal y como podemos ver en la Figura 8.

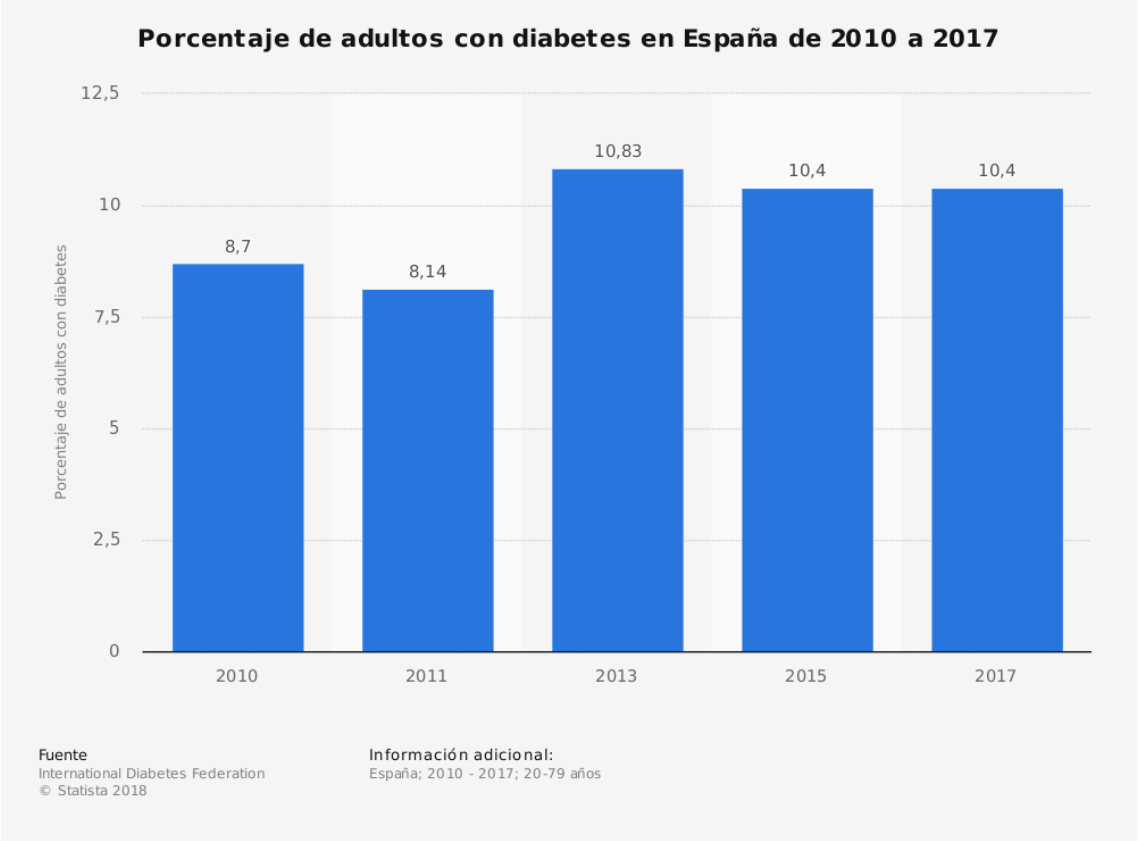

***Figura 8. Prevalencia nacional de adultos con diabetes* [11].**

1.6 Objetivos

El objetivo de este proyecto es diseñar y desarrollar un módulo que permita la generación de redes de términos médicos estandarizados a partir de las conversaciones de foros de pacientes con diabetes. Con dicha red, queremos caracterizar los elementos más relevantes de la trayectoria clínica de la diabetes

desde el punto de vista de los pacientes, lo que supone una vía complementaria a la ruta asistencial. Podemos distinguir varios subobjetivos:

- Extraer términos de fuentes heterogéneas de internet de información no estructurada reportada por personas con diabetes.
- Analizar cuáles de los términos médicos son de más relevancia y frecuentes en la comunidad de pacientes con diabetes.
- Generar una batería de términos médicos relacionados.
- Generar un grafo ponderado no dirigido que relacione a usuarios de dicha comunidad con los términos anteriores.

1.7 Estructura de la memoria

Tras esta introducción, pasaremos al segundo capítulo del presente proyecto. Es aquí donde se expone el estado del arte, comentando otros proyectos con finalidades similares.

Acto seguido, en la tercera parte, se detallan los materiales utilizados para el desarrollo del proyecto, donde se explica minuciosamente el caso de estudio realizado en este trabajo, es decir, se explica la parte de investigación realizada para el desarrollo del mismo.

El cuarto capítulo corresponde a la exposición de la metodología seguida en el proyecto. Se describe también la arquitectura del proyecto y el proceso aplicado a los datos para obtener información de los mismos. Además, se comenta la parte hardware y software del trabajo.

Seguidamente, en el quinto apartado, se comentan y se discuten los resultados intermedios que se han ido obteniendo a medida que se desarrollaba el proyecto y se analiza el resultado final. También se comentan las limitaciones del trabajo expuesto.

Para finalizar, se llega al capítulo de conclusiones y líneas futuras, donde se extrae una conclusión final y se describen posibles líneas de investigación sobre el presente proyecto.

2. Estado del arte

En este capítulo se recogen una serie de investigaciones científicas sobre la enfermedad que estamos tratando, constatando así los resultados positivos que tiene aplicar modelos predictivos en el mundo de la salud.

2.1 Evidencia científica

Estudios médicos realizados mediante herramientas predictivas han demostrado que la aplicación de estos modelos obtiene grandes resultados y que constituirán la base de la medicina en no mucho tiempo. Así lo ha demostrado un estudio realizado por el hospital comarcal de Alcañiz, Teruel [17].

Los objetivos de este estudio fueron evaluar el grado de control glucémico, lipídico y de tensión arterial en una muestra de pacientes con diabetes tipo II en consultas externas, además de identificar factores predictivos del control glucémico promedio durante el seguimiento del tratamiento de los pacientes [17].

Como modelo estadístico, se utilizó la regresión lineal simple y múltiple, donde la única diferencia entre ambas es que la simple maneja una variable independiente y la múltiple, como su nombre indica, dos o más. Lo que pretende este tipo de modelo es analizar la relación existente entre una variable dependiente (variable resultado o respuesta) y una o varias variables independientes (variables explicativas) [17].

La variable dependiente en este caso fue la hemoglobina glicosilada, de ahora en adelante HbA1c. Entre las variables independientes se encuentran el sexo, la edad y tiempo de evolución entre otras. Para este estudio, se consideraron significativas las relaciones con $p < 0,05$. Esto se refiere a que el resultado no proviene del azar [17].

Los resultados mostraron que más del 50% de los pacientes conseguían un control óptimo de los niveles de lípidos y de la tensión arterial pero, como podemos observar en la Figura 9, tan solo un 10% de estos conseguía un control ideal de HbA1c, siendo esta cifra HbA1c < 6,5% de media [17].

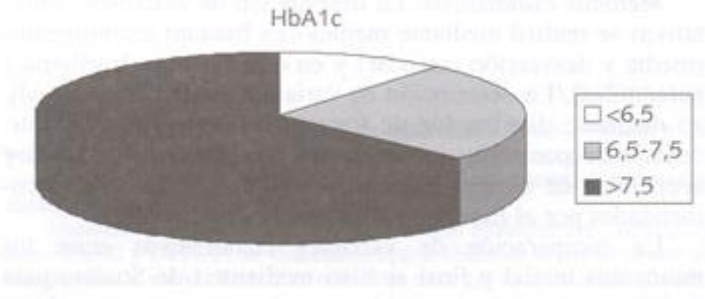

Figura 9. Diagrama de tarta de la variable HbA1c [17]**.**

Uno de los objetivos de este estudio durante el seguimiento del paciente es mantener el promedio de HbA1c en un nivel adecuado con el fin de prevenir posibles complicaciones que se pueden volver crónicas. La pérdida de reserva de insulina ha resultado uno de los factores predictivos clave en el deterioro del control glucémico [17].

Se puede concluir que gracias a establecer un modelo para detectar factores predictivos se ha podido observar qué variables independientes han sido las causantes de la dificultad de mantener un nivel promedio de control glucémico en personas con diabetes tipo 2.

Otro estudio que demuestra los buenos resultados que consiguen los modelos predictivos es el que ha realizado la universidad de Ciencias Médicas de Cienfuegos de Cuba [19].

El objetivo de este estudio es diseñar un modelo estadístico predictivo para el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II. El alto nivel de glucosa en sangre y otros factores que concurren con frecuencia en personas con diabetes provocan un daño en los vasos y nervios, dando paso así al pie diabético o pérdida de sensibilidad en el pie, que puede ocasionar heridas u otras dolencias [19].

Para el desarrollo del modelo, se realizó un estudio descriptivo con una muestra de pacientes con diabetes tipo II atendidos en la Clínica del Diabético de Cienfuegos del año 2011 al 2013, ambos incluidos. Se dividió el grupo de muestra en dos partes, una para crear el modelo con 795 pacientes y otra con 265 para evaluar la capacidad predictiva de dicho modelo [19].

Evidentemente, como variable dependiente se utilizó el padecimiento de pie diabético y como variables independientes la edad, el sexo, consumo de bebidas alcohólicas, consumo de café, hábito de fumar o hipertensión arterial entre otras. Aplicando diferentes pruebas y técnicas estadísticas a las variables se sacaron dos posibles modelos, uno mediante regresión logística y otro con árboles de decisión mediante el algoritmo CHAID [19].

Realizando un estudio comparativo de ambos modelos y teniendo en cuenta el de mayor porcentaje de pacientes bien clasificado (PT), de valor predictivo positivo (VPP) y de mayor valor de sensibilidad (Sb), se llegó a la conclusión de que el modelo que usa árboles de decisión garantiza una mayor capacidad predictiva respecto el otro modelo [19].

Modelo	PT	VPP	Sb
Regresión logística	0,792	0,838	0,893
Árbol de decisión	0,800	0,850	0,891

Tabla 1. Comparación de los modelos estadísticos [19].

Para el análisis de los resultados, se utiliza la curva característica de funcionamiento del receptor ROC (*Receiver Operating Characteristic*). Es una representación gráfica de la sensibilidad en función de la especifidad de un clasificador binario. Hay cuatro alternativas, como podemos observar en la Tabla 2:

	Valor en realidad	
Predicción	Verdaderos positivos (VP)	Falsos positivos (FP)
	Falsos negativos (FN)	Verdaderos negativos (VN)
Sensibilidad = VP/(VP + FN), Especifidad = VN/(FP + VN)		

Tabla 2. Tabla de contingencia o asociación entre dos variables [16].

Esta tabla tiene una interpretación y se puede extrapolar a diagnósticos. Por ejemplo, VP significa que se ha predicho un valor o diagnosticado determinada enfermedad y ese valor en la realidad ha resultado ser el predicho o la enfermedad finalmente diagnosticada ha sido la predicha, es decir, se ha acertado en la predicción. En FP se diagnostica la enfermedad en cuestión y el resultado muestra que el

paciente no presenta dicha enfermedad. El VN viene a decir que no se diagnostica o se predice que la persona no padece cierta enfermedad y las pruebas muestran que en realidad no la padece, por tanto, se acierta en la predicción. Por último, el FN significa que no se diagnostica la enfermedad pero el resultado final indica que el paciente si padece esta enfermedad, lo cual acarrea unas serias consecuencias negativas [16].

Como conclusión, se observó que ambos modelos presentan resultados similares y aceptables a la hora de clasificar pacientes, pero sometiendo estos resultados a la curva ROC se obtiene que el modelo de árbol presenta resultados ligeramente mejores.

3. Materiales

En este apartado se dará a conocer la procedencia del conjunto de datos utilizado para el desarrollo del proyecto, así como también se explicará en detalle el caso de estudio propuesto que constituye la base del trabajo.

3.1 Conjunto de datos

El conjunto de datos que manejamos a lo largo de todo el proyecto proviene directamente de fuentes heterogéneas, siempre en inglés, de internet. La fuente en la que nos hemos centrado es un foro que las personas con diabetes transitan con mucha frecuencia [9]. Es posible observar que los usuarios publican diariamente, por lo que la fuente está actualizada.

Puesto que la fuente que hemos escogido no proporciona API, para extraer toda la información que requeríamos hemos hecho uso de *web scraping*, una técnica muy conocida y muy utilizada cuando se quiere sacar información de sitios web. Profundizando un poco, te permite explorar un determinado sitio web y guardar la información que necesitas en la estructura de datos que más te convenga, para su posterior manipulación.

En este caso, para el desarrollo de la primera parte, la batería de términos, hemos elegido dos grandes hilos donde los usuarios publican diariamente los niveles de glucosa en sangre que presentan a lo largo del día y cuentan sus rutinas o sus sensaciones. Para recopilar todas las palabras que los usuarios han escrito, hemos hecho uso de la técnica anteriormente mencionada. Seguidamente, hemos usado la API de SNOMED para extraer los códigos de los términos en terminología médica. En este punto, es posible cambiar de fuente y seguir generando la batería de términos sin problemas.

Por otro lado, para la segunda parte, hemos escogido cuatro extensos hilos del mismo foro por cuestiones procedimentales. Aquí también se ha hecho uso de *web scraping* para extraer las palabras que los usuarios han publicado y, además, extraer también dichos usuarios. Puesto que esta parte es más tediosa, para realizar todo este proceso nos hemos tenido que fijar en el contenido HTML de estas páginas web. Cada una de estas 4 presenta una estructura igual entre ellas porque pertenecen al mismo

foro en cuestión, pero es diferente al resto de webs, por lo que será complicado extrapolar este procedimiento a otras páginas web.

3.2 Caso de estudio

Para la realización de este proyecto, creímos necesaria la creación de una amplia estructura de datos que incluyera términos médicos, puesto que a la hora de minar términos de las diferentes fuentes heterogéneas que ofrece internet fue necesario comprobar que los términos a incluir en el resultado final eran relevantes, es decir, tenían connotación médica y estaban relacionadas con la enfermedad.

A la hora de enfocar dicha necesidad, hemos desarrollado una batería de palabras basándonos en la terminología clínica que ofrece SNOMED – CT (*Systematized Nomenclature of Medicine – Clinical Terms*), un estándar codificado internacionalmente utilizado [20].

***Figura 10. SNOMED CT* [20].**

Profundizando un poco en esta herramienta, podemos observar que asigna una serie de números enteros o código a términos de lingüística médica. Puesto que este código actúa como identificador de la palabra, esto facilita a los profesionales de la salud una adecuada y precisa representación de la información, así como la posibilidad de analizar o comunicar datos clínicos de una forma más simple.

En la Figura 11, podemos observar un ejemplo de cómo funciona este sistema de nomenclatura. Se introduce en el buscador la palabra que nos interesa y si la tiene integrada en el sistema de nomenclatura médica muestra el identificador que tiene asociado dicha palabra. Hemos escogido la palabra “glucosa”, en inglés “*glucose*”, para mostrar un ejemplo. En la parte del “SCTID” aparece el código, en este caso, el identificador, que es único.

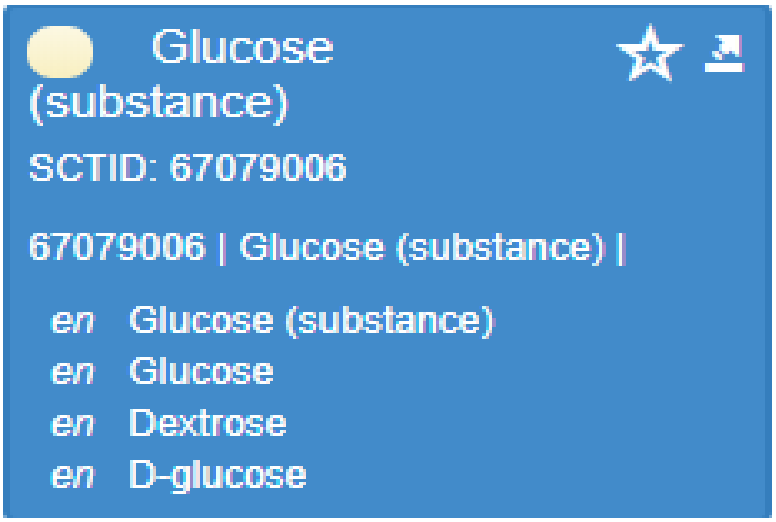

Figura 11. Identificador SNOMED CT [10].

Todo nuestro trabajo está basado en experiencias que personas con diabetes han publicado en el foro en el cual nos hemos centrado. El foro escogido no ha sido al azar, sino que nos hemos fijado en uno con un alto índice de interacción y a la orden del día.

Para el desarrollo de la batería de palabras, hemos aplicado un proceso que se conoce en la rama de minería de datos como ETL (*Extraction*, *Transformation*, *Load*), o lo que es lo mismo, extracción de datos, limpieza y transformación de estos y, por último, la fase de carga [21].

Se han extraído los textos que los usuarios de este foro han ido publicando. Una vez extraídos y recopilados, hemos transformado esta información que teníamos en forma de texto en una lista de palabras o *tokens*. Todas estas palabras se han sometido a un proceso de limpieza, es decir, se han eliminado los signos de puntuación que podían llevar ligados. Además, se han eliminado de esta lista las palabras que no son para nada interesantes, como por ejemplo los artículos o los pronombres. Este tipo de palabras se conocen en inglés como *stopwords*.

Llegados a este punto, hemos realizado un proceso de filtración de palabras haciendo uso de una API que ofrece SNOMED [7]. Una API, conocida también por sus siglas en inglés *Aplication Programming Interface*, es una interfaz de programación de aplicaciones que permite la comunicación entre componentes software.

Gracias al uso de esta API, con el filtrado de palabras hemos podido asignar los códigos SNOMED correspondientes anteriormente mencionados a los términos en terminología clínica. Esta relación entre término y código se ha almacenado en una

estructura de datos muy eficaz para el intercambio de datos conocida como JSON (*JavaScript Object Notation*). Esta última parte corresponde al proceso de carga de los datos.

Situando la batería de términos generada en el contexto de minería de datos, podemos catalogarla como un almacén de datos, puesto que tiene el mismo comportamiento o cumple con las mismas funciones que dicho almacén tiene desde un punto de vista empresarial. Son colecciones de datos cuyo objetivo principal es dar apoyo a los procesos de toma de decisiones [21].

El objetivo no es lo único que comparten estas dos estructuras. Además, presentan similitudes en cuanto a las propiedades se refiere. Nuestra batería y al igual que estos almacenes es integrada, es decir, incorpora datos recogidos de diferentes fuentes externas. Otra característica que comparten es que son no volátiles, esto es, solo es posible consultarlas y no modificarlas. No se puede insertar, eliminar o alterar datos de estas [21].

Adicionalmente, utilizando estadística general, hemos incluido un histograma de frecuencia en el cual se muestra qué términos relacionados con la diabetes son los que los usuarios dicen en más ocasiones.

4. Métodos

En este capítulo repasaremos la metodología empleada en el proyecto, así como la estructura de todo el procedimiento, la preparación y manipulación de los datos y las especificaciones técnicas.

4.1 Metodología

Por lo que respecta a la metodología del proyecto, se han utilizado diferentes librerías del lenguaje de programación utilizado como herramientas para llegar a las conclusiones finales. Como mencionaremos más adelante en el apartado 4.3, el lenguaje escogido ha sido Python, un lenguaje de alto nivel que destaca por ser altamente legible.

Nuestro cometido requiere tratar con diferentes páginas de internet. En base a esto, para abrir y leer la información de sitios web se ha utilizado el paquete *urllib.request*. A continuación, para extraer la información HTML de dichos sitios se hace uso de la librería *BeautifulSoup*, con la que podemos analizar, de los textos que extraemos, su estructura lógica con un *parser*.

Debido a que nuestros datos provienen de fuentes de texto libre en inglés, utilizamos la librería *nltk* (*Natural Language Toolkit*) para filtrar estos textos y quedarnos con los términos más relevantes. En concreto, se utiliza el paquete *stopwords* contenido en dicha librería, que contiene una serie de palabras en inglés que no se consideran útiles para seguir con el resto del procedimiento. Estas palabras pueden ser artículos, preposiciones, etc.

El desarrollo sigue con la eliminación de los signos de puntuación que pueden aparecer en el texto. Esto se consigue mediante el uso de la librería de expresiones regulares, que nos permite buscar en cualquier parte del texto los signos de puntuación y aplicarles una operación de eliminación. Una vez conseguido esto, se procede con la separación del texto por palabras.

Lo expuesto hasta ahora en este apartado es el procedimiento general que se sigue en las dos primeras partes del trabajo. En la primera, se genera un histograma de términos más frecuentes. Para ello, se utilizan vectores *numpy*. Esta librería constituye un excelente soporte a la hora de trabajar con matrices o vectores en Python. Es muy usada cuando se trabaja con funciones matemáticas.

Siguiendo con la primera fase, para interactuar con la API que ofrece SNOMED importamos JSON, puesto que nos permite extraer la información de la interfaz de esta plataforma con facilidad. Es una estructura de datos que cuenta con gran flexibilidad ya que las funcionalidades que proporciona son muy diversas. Esto se explica en detalle en el apartado 3.

Para mostrar las funciones y gráficas se utiliza *Matplotlib*, una librería específica para estas funcionalidades que opera en dos dimensiones. Los resultados son guardados en archivos en formato pdf, utilizando otra librería llamada *ReportLab*.

En la tercera y última parte del desarrollo se incluye la librería *networkx*, que nos permite construir y estructurar el resultado final, el grafo ponderado no dirigido. Mediante una composición de nodos y aristas que intersectan entre sí, hacemos visible el resultado de todo el proceso.

4.2 Arquitectura

Por lo que respecta a la arquitectura del proyecto, se ha estructurado en 3 partes, donde cada una de las cuales corresponde a un archivo Python que genera algún tipo de estructura de datos.

La primera parte es un proceso de minería de datos que genera como resultado intermedio un almacén de datos, que está formado por términos y por su correspondiente código en terminología médica. Puesto que este archivo es posteriormente utilizado en las otras dos partes, podríamos decir que la generación de este constituye la base del proyecto.

Partiendo de esta base, se llega al resultado de la segunda parte. Este resultado no es más que otra estructura de datos que va a ser utilizada en la tercera y última fase. Está formada por un conjunto de usuarios pertenecientes a una fuente de internet para gente con diabetes donde cada usuario tiene asociadas las palabras o términos que ha publicado en dicha fuente. También está contenida la frecuencia con la que los usuarios han usado cada palabra.

En la última parte y como resultado final del proyecto, se genera un modelo predictivo de red, en este caso un grafo ponderado no dirigido, donde se diferencia entre nodos usuario y nodos término, unidos por aristas ponderadas cuyo peso es la frecuencia anteriormente comentada.

Así pues, se ha establecido la arquitectura del proyecto de tal forma que los resultados intermedios sean visibles y claros para el objetivo del mismo.

4.3 Preproceso

Para que los resultados se ajustaran a los objetivos del presente proyecto y se enfocaran adecuadamente desde un punto de vista clínico, se ha aplicado un proceso de tratamiento a los datos originales previo a su manipulación. A esto se le conoce en minería de datos como ETL (*Extraction, Transformation, Load)* [21] y está incluido en la parte de preparación de los datos.

Figura 12. Proceso ETL (Extraction, Transformation, Load) [14]**.**

El objetivo de este proceso es obtener una vista minable, es decir, un conjunto de datos de interés para el problema concreto en un formato adecuado. Este conjunto proviene de unos datos originales que pueden ser repetitivos, inadecuados, erróneos, irrelevantes, faltantes, etc. Como se ha comentado en el punto 3.2, los datos se almacenan, en la primera parte del proyecto, en una batería de términos, que equivale en este caso a *data warehouse* [21].

Como se ha explicado anteriormente, se han extraído textos de usuarios de fuentes de internet. Estos textos han sido separados por palabras. Cada palabra se ha sometido a un proceso de eliminación de signos de puntuación que pudiera llevar asociados. Además, se ha comprobado que estas palabras no pertenecieran al conjunto de palabras que no aportan significado al texto conocido en inglés como *stopwords*. Un ejemplo de este conjunto puede ser un artículo o una preposición.

Una vez filtrados los datos, el preproceso concluye con la fase de carga de la información. En nuestro caso, se ha escogido el formato JSON, puesto que permite una fácil manipulación de los datos contenidos en esta estructura.

4.4 Equipamiento *hardware*

Para la visualización de la red de usuarios y términos final se ha hecho uso de una Unidad de Procesamiento Gráfico (GPU) AMD Radeon R7 M260. Las especificaciones de la misma se encuentran en la Tabla 3. Por otro lado, para el cómputo del resto de tareas y procesos se ha utilizado un ordenador TOSHIBA SATELLITE S50-B (Intel® Core™ i7-5500U, 8GB de RAM).

Parámetros	Especificaciones
Memoria	1 GB
Tipo de memoria	GDDR5 a 3.6 GHz
Interfaz	128 bits

Tabla 3. Especificaciones técnicas de la GPU utilizada.

4.5 Equipamiento *software*

La elaboración de este proyecto se ha llevado a cabo utilizando Python como lenguaje de programación. Más concretamente, se ha utilizado la versión 3.6 del mismo. La herramienta utilizada para la creación del código fuente ha sido PyCharm, un entorno de desarrollo integrado muy eficiente en el desarrollo software.

Figura 13. Logo PyCharm.

A lo largo de todo el procedimiento, se han utilizado varias librerías que este lenguaje de alto nivel ofrece. Entre las librerías utilizadas, cabe destacar un par de ellas, que han resultado ser de vital utilidad y constituyen la base del proyecto. Para el web scraping se ha utilizado *BeautifulSoup*, que permite extraer datos de archivos HTML y XML. Para la representación de la red o grafo final hemos utilizado *networkx*, que permite estudiar, crear y manipular redes complejas.

Otra librería que nos ha resultado muy útil y es muy conocida en el entorno Python es *Matplotlib*. Se trata de una librería que permite mostrar funciones y gráficas en dos dimensiones (2D). Con su uso, conseguimos mostrar tanto la red resultante como el histograma de frecuencia.

5. Resultados y discusiones

En este capítulo se analizarán y se discutirán tanto los resultados que se han ido obteniendo a lo largo del proyecto como el resultado final. Además, se comenta alguna limitación que ha surgido a la hora de realizarlo.

5.1 Resultados intermedios

En este proyecto, a partir de la manipulación de los datos iniciales que disponemos, tratamos de llegar a una solución final. Para ello, debemos seguir un proceso que genera resultados intermedios, y no por ser de esta índole son de menos relevancia.

El caso de estudio que hemos propuesto, genera una extensa batería de términos médicos asociados a su correspondiente código estructurado en un diccionario Python, muy fácilmente manipulable y visible. Esta estructura se almacena en un archivo JSON ya que va a ser utilizada posteriormente y es un objeto muy fácil de manejar. En la Figura 14 se observa que está ordenado alfabéticamente para una mayor eficiencia.

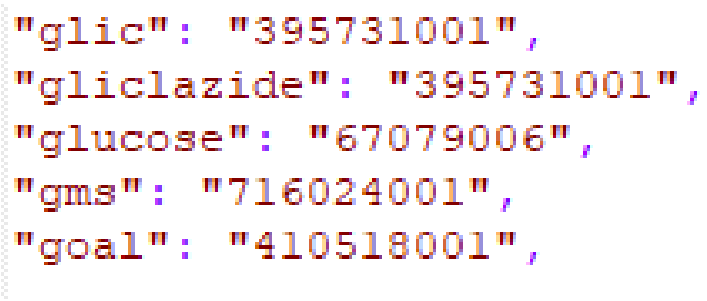

Figura 14. Muestra de la batería de términos.

Además de esto, se genera un histograma de frecuencia (ver Figura 15) en el que se puede visualizar cuáles son los términos que los usuarios más repiten, y por tanto, cuáles son de más relevancia.

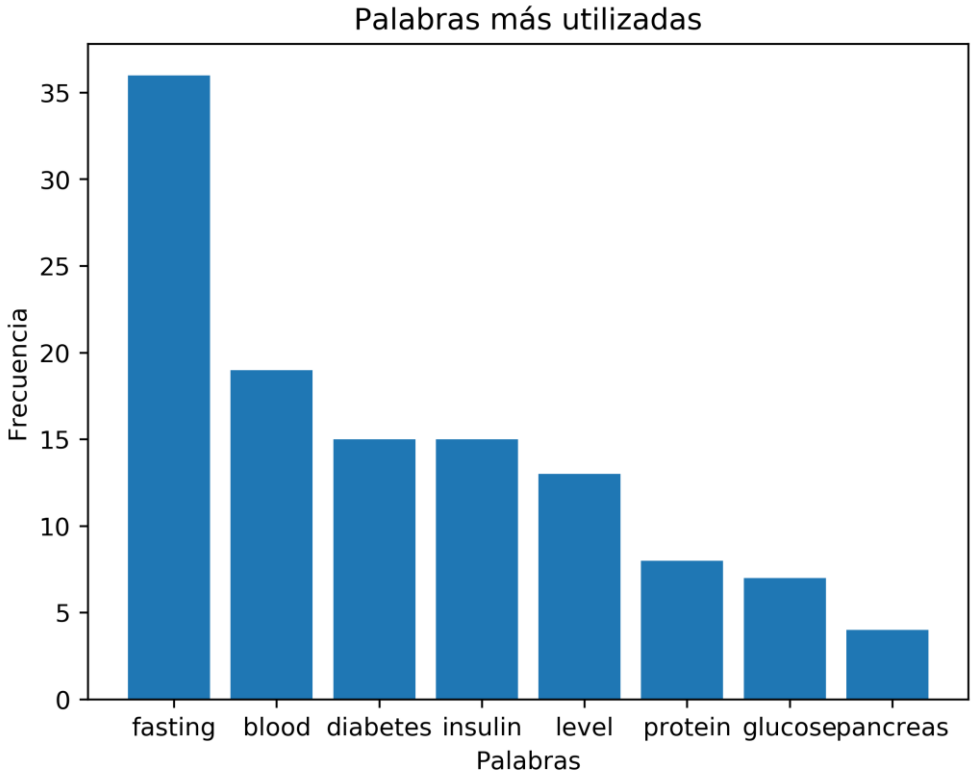

Figura 15. Histograma de frecuencia.

En la segunda parte del proceso, se genera otro diccionario Python que, a su vez, contiene otro diccionario Python. Esta estructura es muy útil y para nada compleja, pues te permite preparar los datos para su posterior uso de una manera cómoda y simple. En dicha estructura, almacenamos todos y cada uno de los usuarios extraídos de la fuente de datos en cuestión y los asociamos a todos los términos que ha expresado y la frecuencia con la que los ha dicho. Todo esto, al igual que en la primera parte, se guarda en un archivo JSON para su posterior uso en la generación del resultado final.

En la Figura 16, se pueden observar los dos diccionarios que se comentan. El externo, tiene como claves todos y cada uno de los usuarios extraídos y como valor otro diccionario. Este diccionario interno tiene como claves las palabras que los usuarios han publicado y como valor la frecuencia con la que las han dicho.

```
{"Shawn14564": {"fasting": 4, "blood": 6, "exercise": 6,
 "JoKalsbeek": {"diet": 3, "Good": 2, "weight": 6,
 "Antje77": {"diabetes": 3, "much": 2, "comment": 1,
```


Figura 16. Muestra de los datos estructurados.

Así pues, en la inmensa mayoría de proyectos, para conseguir un resultado que se acerque lo más posible al óptimo, es necesario establecer, al principio, unos subobjetivos o soluciones parciales que se vayan cumpliendo a medida que se avanza en el desarrollo del proyecto.

5.2 Análisis del resultado final

Con todo lo anteriormente mencionado montado y estructurado, ya es posible lanzarse a generar el resultado final, la red de usuarios y términos. La red representa una relación ponderada entre usuarios y términos. El peso de una arista, en nuestro grafo no dirigido ponderado, es la frecuencia, es decir, el número de veces que un usuario ha dicho el término en cuestión.

Es por esto que es fundamental distinguir entre nodos usuario y nodos término. Como podemos observar en la Figura 17, hemos pintado los nodos usuarios de color azul y los nodos término de color verde. Esta imagen ofrece una muestra del resultado final del trabajo.

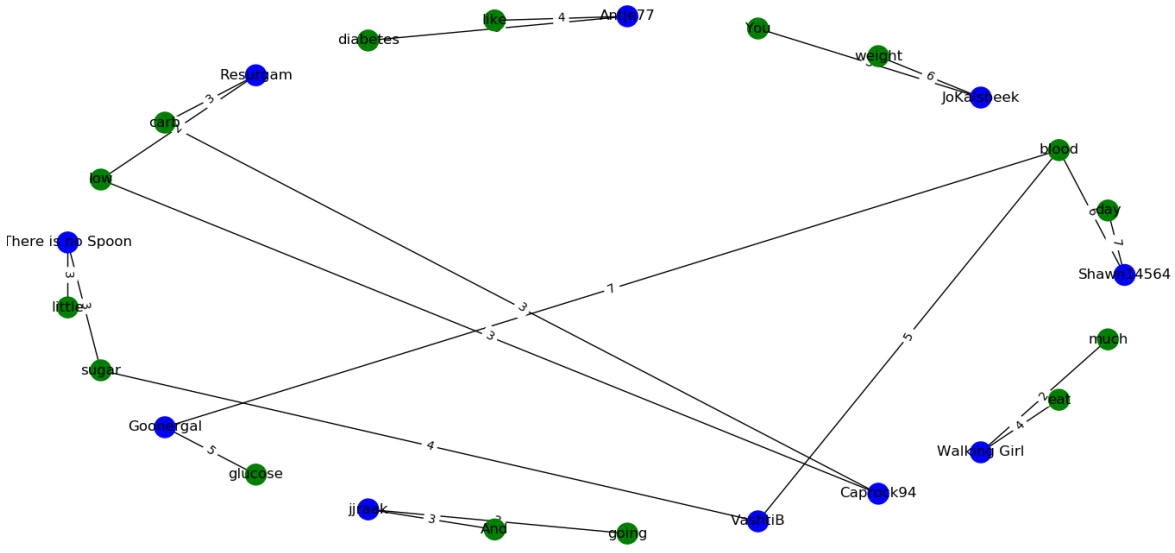

Figura 17. Grafo ponderado no dirigido.

Gracias a esta estructura, es más fácil que un profesional de la salud llegue a predecir complicaciones en la trayectoria de un paciente. Pongamos un ejemplo: en una determinada consulta, un paciente con diabetes presenta unos síntomas que hasta ese momento no había presentado y dice seguir un comportamiento habitual. El profesional de la salud en cuestión, puede basarse en nuestro sistema y ver si algún usuario de la red ha experimentado lo mismo, es decir, puede comprobar si ha expresado estos síntomas, con qué frecuencia e incluso puede que el usuario haya comentado lo que a él le funcionó en dicho caso.

5.3 Limitaciones

Se había iniciado el proyecto con la idea de corregir las palabras que los usuarios podían introducir incorrectamente. La realidad es que esto no se ha podido realizar porque la librería que Python proporciona para esto, nos ha dado muchos problemas para instalarla. La librería en cuestión es *hunspell* y al parecer no se puede usar en Windows.

Cabe destacar que la API de SNOMED CT incluye un corrector. Si introduces una palabra incorrectamente en el buscador, es muy probable que la corrija automáticamente.

Por otro lado, hay muchos usuarios y gran cantidad de términos relacionados a diferentes usuarios. Debido a la enorme cantidad de nodos y aristas que intersectan entre sí, la visibilidad del grafo final es compleja. Bastaría con ampliar la parte de la imagen que se quisiera consultar en un momento determinado.

5.4 Discusiones

Como se ha comentado previamente, los modelos predictivos se basan en el análisis de sucesos pasados para intentar anticiparse a un evento futuro, y esto es justo en lo que se basa nuestra investigación.

A la vista de los resultados obtenidos, se puede observar el potencial de un modelo predictivo de red que parte de unos simples comentarios de personas con diabetes. Manipulando, filtrando y relacionando estos datos y, a su vez, aplicando

estadística general, llegamos a una muy buena solución final, que cumple con la función de modelo predictivo que nos hemos propuesto desde un principio.

La importancia o el significado de nuestro resultado final van más allá de la predicción de posibles complicaciones en trayectorias de pacientes con diabetes. Se puede llegar a mejorar la calidad de las consultas incorporando sistemas de este tipo tanto en la sanidad pública como en la privada. También se puede reducir el tiempo de dichas consultas, lo que supondría una reducción del embotellamiento que, como sabemos, se produce en el mundo sanitario.

Cuanto más grande sea el volumen de datos del que disponemos para analizar, mayor será nuestra red de usuarios y términos, y por tanto, se podrá tomar una decisión más precisa. Es por esto que hemos escogido fuentes de datos muy extensas, para que, cuanto mayor cantidad y variedad de datos, más precisa sea la decisión y menor ratio de error haya.

También es lógico pensar que el paciente que es atendido por un profesional médico que se apoya en nuestro sistema tiene una mayor confianza, puesto que nuestro proyecto se basa en vivencias que otros pacientes han experimentado y superado.

6. Conclusiones y líneas futuras

En este capítulo vamos a exponer las conclusiones finales obtenidas de la realización del presente proyecto así como algunas posibles líneas futuras de trabajo para seguir desarrollando y para mejorar el mismo.

6.1 Conclusiones

En el presente trabajo se ha investigado la posibilidad de crear un modelo predictivo de red para una determinada enfermedad a partir de fuentes heterogéneas de internet. Se han aplicado diferentes métodos, estrategias y técnicas para llegar a la mejor solución posible. Enmarcando estos modelos en un contexto sanitario, observamos que están deviniendo en una fuerte herramienta de trabajo en la que basarse en la toma de decisiones.

La medicina de precisión, es decir, la predicción aplicada en el mundo de la salud es un campo que está empezando a despegar. Muchas universidades y equipos de investigación ya han empezado a realizar una exhaustiva investigación para sacar el máximo rendimiento a estos modelos, puesto que han comprobado los exitosos resultados que conlleva aplicarlos.

A la vista de los resultados obtenidos, a continuación, se exponen las conclusiones principales del presente trabajo de fin de grado:

- Se ha extraído con éxito términos médicos procedentes de fuentes para personas con diabetes.
- Se ha generado un histograma de frecuencia mostrando los términos más relevantes y más frecuentes.
- Se ha creado una extensa batería de términos médicos asociados a su código en terminología clínica.
- Se ha generado un grafo ponderado no dirigido que relaciona los usuarios y estos términos donde la ponderación entre cada usuario y cada término corresponde al número de veces que un usuario ha dicho un término.

6.2 Líneas futuras

Por lo general, cualquier proyecto que incluya una parte de investigación, trata de despejar incógnitas y resolver preguntas del tema tratado pero, a su vez, también genera nuevas ideas, nuevas líneas de trabajo e incluso nuevas preguntas.

Una posible e interesante línea de trabajo relacionada con el presente trabajo es la consideración de incluir conceptos, además de solo palabras, tanto en la batería de términos generada como en el resultado final. Entendemos concepto en este caso como un conjunto de más de una palabra que tiene significado distinto al que tienen las palabras que forman dicho conjunto por separado. Un ejemplo puede ser *fasting blood glucose*, que significa nivel de glucosa en sangre en ayunas.

Otra posible aunque más compleja línea de investigación es estar actualizando nuestro almacén de datos continuamente cada cierto tiempo. Puesto que contamos con limitación de tiempo esto no nos es posible incluirlo en el presente proyecto.

7. Referencias

[1] CIO España. (2014). Obtenido de https://www.ciospain.es/sanidad/los-modelos-predictivos-impulsan-la-medicina

[2] Fundación para la diabetes. (2015). Obtenido de https://www.fundaciondiabetes.org

[3] 1 de cada 11 personas en el mundo ya tiene diabetes, advierte la OMS. (2016). BBC Mundo.

[4] Imágenes de Artical Blog. (2017). Obtenido de https://blog.artical.xyz

[5] Imágenes de El Español. (2017). Obtenido de https://www.elespanol.com/ciencia/salud

[6] Piperlab Blog. (2017). Obtenido de https://piperlab.es/

[7] API SNOMED CT. (2019). Obtenido de https://snowstorm-fhir.snomedtools.org/fhir/ValueSet/$expand?url=http://snomed.info/sct?fhir_vs&count=100&offset=0&filter=glucose&_format=json

[8] El Día. (2019). Obtenido de https://www.eldia.es

[9] Global Diabetes Community. (2019). Obtenido de https://www.diabetes.co.uk

[10] Imágenes de SNOMED CT. (2019). Obtenido de https://browser.ihtsdotools.org

[11] Imágenes de Statista. (2019). Obtenido de https://es.statista.com

[12] MedlinePlus. (2019). Obtenido de https://medlineplus.gov

[13] Software y Soluciones de Analítica. (2019). Obtenido de https://www.sas.com/es_es/insights/analytics/data-mining.html

[14] Troyanx Soluciones Informáticas. (2019). Obtenido de http://troyanx.com/Hefesto/proceso_etl.html

[15] Arimetrics. (2019). Documentación modelos predictivos. Obtenido de https://www.arimetrics.com/glosario-digital/modelo-predictivo

[16] Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. Revista chilena de infectología, 29(2), 138-141.

[17] Gimeno Orna, J. A., Boned Juliani, B., Lou Arnal, L. M., & Castro Alonso, F. J. (2003). Factores relacionados con el control glucémico de pacientes con diabetes tipo 2. An. Med. Interna, 20(3), 122-126.

[18] International Diabetes Federation. (2017). Diabetes Atlas.

[19] López Fernández, R., Yanes Seijo, R., Suárez Surí, P. R., Avello Martínez, R., Gutiérrez Escobar, M., & Alvarado Flores, R. M. (2016). Modelo estadístico predictivo para el padecimiento de pie diabético en pacientes con. Universidad de Ciencias Médicas de Cienfuegos, 14(1), 30-40.

[20] Ministerio De Sanidad, Consumo y Bienestar. (2019). SNOMED CT . Obtenido de https://www.mscbs.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/quees.htm

[21] Universidad Politécnica de Valencia. (2019). Business Intelligence. Obtenido de https://poliformat.upv.es/access/content/group/CFP_509_23504/BI/BI_AD_t1_t2.pdf

[22] Universidad Politécnica de Valencia. (2019). Minería de datos. Obtenido de https://poliformat.upv.es/access/content/group/CFP_509_23504/DM/BI_DM_t3.pdf