Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Superiores
Unidad Morelia

Reconstrucción del historial clínico en pacientes mexicanos con Diabetes Mellitus Tipo 2 mediante el uso de métodos computacionales.

T E S I S

que para obtener el título de:

Licenciado en Tecnologías para la Información en Ciencias

presenta:

César Arcos González

DIRECTORA DE TESIS:

Dra. Marisol Flores Garrido

Morelia, Michoacán, Septiembre 2023

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES
UNIDAD MORELIA

MTRA. IVONNE RAMÍREZ WENCE
DIRECTORA
DIRECCIÓN GENERAL DE ADMINISTRACIÓN ESCOLAR
P R E S E N T E

Por medio de la presente me permito informar a usted que en la **sesión ordinaria 03** del **Comité Académico de la Licenciatura en Tecnologías para la Información en Ciencias** de la Escuela Nacional de Estudios Superiores (ENES), Unidad Morelia, celebrada el día **20 de abril de 2023**, se acordó poner a su consideración el siguiente jurado para la presentación del Trabajo Profesional del alumno **César Arcos González** de la Licenciatura en **Tecnologías para la Información en Ciencias,** con número de cuenta **41912679-5**, con el trabajo titulado: **"Reconstrucción del historial clínico en pacientes mexicanos con Diabetes Mellitus Tipo 2 mediante el uso de métodos computacionales",** bajo la dirección como tutora de la **Dra. Marisol Flores Garrido.**

El jurado queda integrado de la siguiente manera:

Presidente:	Dr. Víctor Hugo Anaya Muñoz
Vocal:	Dr. Luis Miguel García Velázquez
Secretario:	Dra. Marisol Flores Garrido
Suplente:	Dra. Anel Gómez García
Suplente:	Dra. Adriana Menchaca Méndez

Sin otro particular, quedo de usted.

A t e n t a m e n t e
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Morelia, Michoacán a 28 de septiembre de 2023.

DRA. YUNUEN TAPIA TORRES
SECRETARIA GENERAL

Agradecimientos Institucionales

Quiero expresar mi profundo agradecimiento a la Licenciatura en Tecnologías para la Información en Ciencias de la Universidad Nacional Autónoma de México por brindarme una formación académica de calidad que ha marcado un hito en mi vida. Esta invaluable experiencia educativa ha sido fundamental en mi desarrollo profesional y personal.
Asimismo, agradezco al programa UNAM-DGAPA-PAPIIT IA106620 por su apoyo y contribución a mi formación. Este respaldo ha enriquecido mi aprendizaje.
Quiero expresar también mi gratitud a los miembros del jurado por sus valiosas observaciones y por dedicar su tiempo y conocimientos en la evaluación de mi trabajo:

- Dr. Víctor Hugo Anaya Muñoz
- Dr. Luis Miguel García Velázquez
- Dra. Marisol Flores Garrido
- Dra. Anel Gómez García
- Dra. Adriana Menchaca Méndez

Agradecimientos Personales

Quiero expresar mi sincero agradecimiento a todas las personas que contribuyeron a la realización de esta tesis. En primer lugar, a mi familia por su constante apoyo y ánimo durante este proceso. Agradezco a mis amigos y colegas por sus ideas y debates enriquecedores que han mejorado este trabajo. También a mi asesora por su orientación y valiosas sugerencias.

Resumen

La construcción de un modelo de aprendizaje automático requiere una cantidad sustancial de datos de alta calidad. Este estudio se enfocó en la reconstrucción y mejora del historial clínico de pacientes mexicanos con diabetes mellitus tipo 2 en el estado de Michoacán a partir de datos tomados de notas médicas. El objetivo fue facilitar a futuras investigaciones la creación de modelos de aprendizaje automático y el uso efectivo de los datos para análisis. Nuestro proyecto implementó el método de Reconocimiento de Entidades Nombradas para extraer parámetros clínicos de las notas médicas y así enriquecer el historial clínico. Se empleó Pyspark para la manipulación de datos, aprovechando la computación en la nube para garantizar una plataforma robusta capaz de manejar grandes volúmenes de información. Adicionalmente, se llevó a cabo un análisis de los datos con el propósito de mejorar la precisión en la fecha de diagnóstico de la hipertensión, aplicando reglas desarrolladas en colaboración con el personal médico.

Abstract

The construction of a machine learning model requires a substantial amount of high-quality data. This study focused on the reconstruction and improvement of the clinical history of Mexican patients with type 2 diabetes mellitus in the state of Michoacán using data extracted from medical records. The objective was to facilitate future research in creating machine learning models and the effective use of data for analysis. Our project implemented the Named Entity Recognition method to extract clinical parameters from medical notes and thus enrich the clinical history. Pyspark was employed for data manipulation, leveraging cloud computing to ensure a robust platform capable of handling large volumes of information. Additionally, a data analysis was carried out to improve the accuracy of the hypertension diagnosis date, applying rules developed in collaboration with medical personnel.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La diabetes mellitus es una enfermedad crónica caracterizada por niveles elevados de glucosa en sangre, acompañada de un metabolismo alterado de grasas y proteínas. La glucosa en sangre aumenta porque no se puede metabolizar en las células, debido a la falta de producción de insulina por parte del páncreas o a la incapacidad de las células para utilizar eficazmente la insulina que se está produciendo. La diabetes es una enfermedad compleja que se presenta en diferentes tipos, cada uno con mecanismos particulares. Por ejemplo, en la diabetes tipo 1 el sistema inmunitario del organismo destruye las células que liberan insulina y llega a eliminar su producción, mientras que en la diabetes tipo 2 el cuerpo no puede utilizar la insulina que produce [14].

Esta enfermedad es de gran interés en nuestro país, pues México siempre ocupa los primeros lugares a nivel mundial de diabetes mellitus tipo 2. Tan solo en el primer trimestre de 2021 se registraron 3,831 casos de dicha enfermedad [15], por lo que no se espera que las estadísticas bajen a corto plazo. Aún más, la contingencia de salud causada por el virus SARS-CoV-2 creó una alarma en la población mexicana por la vulnerabilidad de las personas con diabetes mellitus [16].

Si las personas que padecen diabetes mellitus no logran estabilizar los niveles de glucosa, son propensos a presentar numerosas complicaciones graves que se prolongan en el tiempo. Algunas comienzan a los pocos meses de iniciarse la diabetes, aunque la mayoría suelen aparecer al cabo de algunos años y empeorar de forma gradual.

Una de las complicaciones que aparece con mayor frecuencia son las enfermedades car-

diovasculares, que constituyen la primera causa de defunción en el mundo y probablemente lo seguirán siendo, debido al aumento de su prevalencia en los países con menos recursos y al envejecimiento de la población.

Existen grandes expectativas sobre el potencial de la inteligencia artificial (IA) en el análisis y predicción de la diabetes mellitus [17]. Sin embargo, los modelos de aprendizaje automático requieren un volumen importante de datos para su entrenamiento.

En el caso del estado de Michoacán, el Instituto Mexicano del Seguro Social (IMSS) se presenta como una fuente importante de información sobre la diabetes mellitus tipo 2, pues la institución atiende anualmente a miles de personas diagnosticadas con la enfermedad. Además, desde el año 2003 el IMSS consiguió el desarrollo e implementación del Sistema de Información de Medicina Familiar (SIMF) en unidades de primer nivel de atención. Esto se traduce en el manejo de un Expediente Clínico Electrónico (ECE) para cada derechohabiente (ver Fig. 1.1) y representa una posibilidad importante de contar con datos organizados, a gran escala, en formato digital, que permitan comprender mejor la evolución de la enfermedad en pacientes mexicanos.

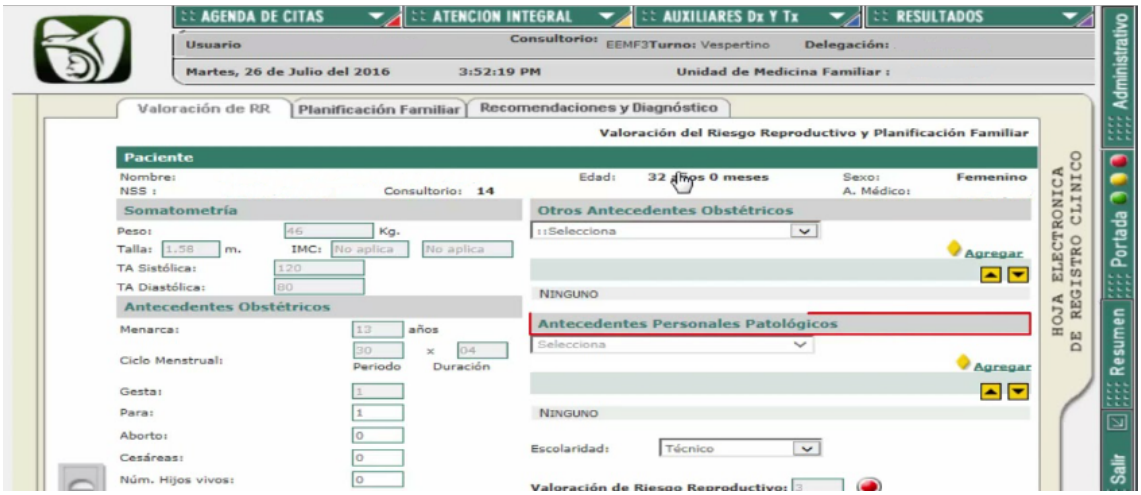


Figura 1.1: Imagen ilustrativa del Sistema de Informacioń de Medicina Familiar (SIMF), implementado en el Instituto Mexicano del Seguro Social desde 2003.

Este trabajo de tesis se inscribe en el marco del proyecto "Estudio longitudinal para el desarrollo de modelos predictivos de complicaciones crónicas de la diabetes mellitus tipo 2", que se lleva a cabo con apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT) a

través del Fondo Institucional de Fomento Regional para el Desarrollo Científico, Tecnológico y de Innovación (FORDECyT) de los Programas Nacionales Estratégicos (ProNacEs) en la convocatoria «2019-06 Proyectos de investigación e incidencia en ciencia de datos y salud: integración, procesamientos, análisis y visualización de datos de salud en México».

Tras ser aprobado por los comités de ética correspondientes, el grupo de trabajo del proyecto obtuvo acceso al SIMF del Estado de Michoacán, consiguiendo así datos correspondientes a 287,474 pacientes diabéticos. El objetivo principal en el proyecto es utilizar los datos para desarrollar e implementar modelos de aprendizaje automático que permitan analizar la evolución de complicaciones en pacientes diabéticos.

La base de datos del SIMF cuenta con una base de datos con millones de registros de consultas médicas a lo largo de más de una década. Los datos representan más de 80 Gigabytes de información, organizada en más de 80 tablas relacionales que integran el registro de notas médicas, agenda de citas, historia clínica, incapacidades, recetas, notas administrativas y notas de trabajo social, entre otras.

A pesar de contar con este volumen de información, los modelos de aprendizaje automático no son inmediatamente aplicables. Es necesario transformar los datos a un formato que resulte apropiado para las tareas planteadas de predicción de complicaciones de la diabetes. Aunque el preprocesamiento de información no siempre recibe atención en el contexto general de tareas de aprendizaje, representa un paso fundamental para cualquier análisis que se realice con este enfoque.

En esta tesis planteamos dos contribuciones principales. La primera consiste en crear un conjunto de datos estructurado que facilite el análisis de la información, especialmente mediante la implementación de algoritmos de aprendizaje automático. Esta tarea requiere trabajar con personal médico para explorar con detalle los datos del SIMF e identificar la información que podría ser relevante para el estudio de la diabetes y enfermedades asociadas.

Más aún, durante el desarrollo de esta tarea se observó que sólo una parte de la información de interés se encuentra en un formato tabular. Indicadores importantes para estudiar la enfermedad aparecen en notas médicas como la que se muestra en la Fig. 1.2. Para extraer estos indicadores de las miles de notas que forman parte del SIMF, plantea-

mos el uso de herramientas del estado del arte en el procesamiento del lenguaje natural. Específicamente, utilizamos Reconocimiento de Entidades Nombradas basado en el uso de redes neuronales artificiales.

Nuestro trabajo logra la extracción automática de parámetros a partir de notas médicas. Después, estos datos se integran con la información extraída directamente de tablas que componen el SIMF. Todos los datos, provenientes de notas médicas o de tablas, se limpian y ajustan a un formato estandarizado que permite reconstruir el historial clínico de pacientes y facilita diversos análisis, incluyendo aquellos que usan algoritmos de aprendizaje automático.

FEM DE 57 AÑOS DE EDAD, ES DIABETICA TIPO 2 DE 8 AÑOS DE EVOLUCION, HIPERTENSION ARTERIAL DE 2 AÑOS DE EVOLUCION, ACUDE A DOTACION MENSUAL DE MEDICAMENTO, NIEGA POLIS Y DATOS DE VASOESPASMO, NO MAREO, NO PALPITACIONES, NO DOLOR PRECORDIAL, NO PLENITUD GASTRICA TRANSGRESOR DIETETICO, NO TRANSGRESIONES MEDICAMENTOSAS, SIN SENSACION DE GUANTE O CALCETIN, NO MARCHA CLAUDICANTE, NO INTERNAMIENTOS EN MES PREVIO, SE REPORTA CON CEFALEA FRONTAL PREDOMINIO MATUTINO LEVE INTENSIDAD , ENERO 2018 GLUC 202 MG/DL. EF.- BUEN ESTADO GENERAL, NEUROLOGICAMENTE INTEGRA, CUELLO SIN IY, RSCS RITMICOS DE BUENA INTENSIDAD Y FRECUENCIA NO SOPLOS, CSPS BIEN VENTILADOS SIN AGREGADOS, ABDOMEN SIN PUNTOS DOLOROSOS, SIN DATOS DE IRRITACION PERITONEAL, GIORDANO BILATERAL NEGATIVO, EXTREMIDADES INTEGRAS NO EDEMA. PULSOS PRESENTES LLENADO CAPILAR CONSERVADO, NO HOMANS PLAN.- DIETA PARA DIABETICO DE 1800 KAL, BAJA EN SAL MENOS DE 2 GRS/DIA, CAMINATA DIARIA DURANTE 30 MINUTOS, SE ORIENTA SOBRE DATOS DE ALARMA, CUIDADOS DE PIES Y DEL ADULTO MAYOR, PREVENCION DE CAIDAS, SE EXPLICAN EFECTOS SECUNDARIOS DE MEDICAMENTOS, RECETA MANUAL FOLIO 17 B 8358793 PIOGLITAZONA 15 MG 1X2, SE SOL GLUC DE CONTROL, CITA ABIERTA A URGENCIAS EN CASO NECESARIO

Figura 1.2: Ejemplo de nota médica en el Expediente Clínico Electrónico (ECE) de una persona con diabetes.

Así, el producto principal de esta tesis es un conjunto de datos estructurados, con un tamaño aproximado de 10.62 GB, que contiene la información de 278,324 pacientes diabéticos en el estado de Michoacán y que incluye datos tomados de notas médicas.

La segunda contribución del trabajo fue motivada por el proceso de exploración de datos en el SIMF que se realizó para integrar el conjunto de datos. En esta exploración pudimos identificar que, en algunos casos, pacientes recibían medicamentos para cierta enfermedad antes de tener el diagnóstico en su expediente. Estos casos nos llevaron a preguntarnos si los expedientes clínicos tienen algunos errores en el registro de diagnósticos y, de ser el caso, cómo podríamos contribuir a corregirlos.

Para responder a estas preguntas, nos enfocamos en aquellos pacientes diabéticos que padecen también hipertensión. La hipertensión es una enfermedad importante en nuestro

contexto nacional y los datos con los que contamos permiten analizar los expedientes clínicos para identificar pacientes que la padecen.

De este modo, nuestro trabajo subraya la posibilidad de analizar los datos para ajustar errores de registro e ilustra esta tarea mediante la propuesta de una metodología para el caso de la hipertensión.

Nuestro trabajo, a través de la base de datos generada y de la propuesta para la corrección de posibles errores de registro en el diagnóstico de la hipertensión, busca contribuir a estudios que enriquezcan el conocimiento que se tiene de la diabetes en la población mexicana y, eventualmente, abonen a estrategias de monitoreo que mejoren el manejo de la enfermedad y mitiguen sus riesgos. Esto implicaría incluso reducir los costos de la diabetes mellitus en el sector salud, pues esta enfermedad es una de las más costosas [18].

1.1. Planteamiento del proyecto

1.1.1. Preguntas de investigación

En este trabajo se busca contestar las siguientes preguntas de investigación:

- ¿De qué forma puede extraerse automáticamente información de notas médicas para crear una base de datos estructurada, adecuada para el estudio de la diabetes en México?
- ¿Cómo integrar los registros generados de notas médicas con aquella información que se encuentra en formato tabular cuando hay conflicto, especialmente en las fechas del historial?
- ¿Cómo identificar y lidiar con inconsistencias en el registro del diagnóstico de hipertensión?
- ¿Cómo se puede establecer un procedimiento para reconstruir automáticamente historiales clínicos a partir de información en el SIMF, que permitan el análisis de la diabetes en México y de comorbilidades específicas de la enfermedad?

1.1.2. Objetivo General

Desarrollar una metodología que, mediante el uso de aprendizaje automático, permita la extracción, integración, limpieza y estructuración de información relacionada con complicaciones de la diabetes mellitus tipo 2 en México a partir del archivo histórico del SIMF en el estado de Michoacán, permitiendo así la reconstrucción del historial clínico de pacientes.

1.1.3. Objetivos específicos

1. Implementar un algoritmo de procesamiento de lenguaje natural para extraer de notas médicas atributos de interés para el análisis de la diabetes mellitus tipo 2.

2. Proponer un algoritmo que permita la limpieza e integración de indicadores detectados en el objetivo 1 a un formato estándar que permita reconstruir automáticamente el historial de pacientes diabéticos.

3. Mediante el análisis de la hipertensión, una complicación específica relacionada con la diabetes, identificar posibles errores en el registro del diagnóstico de pacientes y proponer estrategias para corregirlos.

1.2. Organización de este documento

Para describir el trabajo realizado en esta investigación, consideramos adecuado presentar cada una de las tres tareas principales (extracción de parámetros, integración del conjunto de datos, propuesta para ajustar diagnósticos de hipertensión) en un capítulo independiente que muestra las ideas, los resultados y la discusión relacionados con la tarea. Así, el resto de este documento está organizado de la siguiente manera.

Capítulo 2: diabetes mellitus tipo 2. En este capítulo se abordan conceptos fundamentales relacionados con la diabetes mellitus tipo 2 y se describen los parámetros clínicos asociados a la enfermedad con los que trabajaremos. También se mencionan las

principales complicaciones observadas en pacientes con diabetes, con especial atención a la hipertensión.

Capítulo 3: Redes neuronales artificiales. Este capítulo presenta los antecedentes de nuestro trabajo en relación a los métodos computacionales utilizados. Para ello se describe la arquitectura de redes neuronales artificiales relevantes en la metodología seguida para extraer parámetros clínicos en el proyecto.

Capítulo 4: Extracción de información. En este capítulo se presenta una explicación del Reconocimiento de Entidades Nombradas (NER, por sus siglas en inglés), el método utilizado para la extracción de parámetros clínicos. En este mismo capítulo se presentan los resultados obtenidos al utilizar esta metodología en nuestro proyecto, señalando el desempeño que se observó en la extracción de los parámetros clínicos deseados.

Capítulo 5: Creación de datos estructurados. Una vez completada la tarea de extracción de parámetros, se integró esta información con otros indicadores existentes en el SIMF para crear el conjunto de datos que representa la principal contribución de este trabajo. En este capitulo se describen las diferentes etapas en la estructuración de los datos, que - en el formato en que se presentan - pueden utilizarse en análisis y predicción de complicaciones de pacientes diabéticos.

Capítulo 6: Ajuste de diagnóstico de hipertensión. Al analizar el SIMF para determinar parámetros pertinentes en nuestro estudio, incluyendo aquellos obtenidos mediante NER, se observaron algunas inconsistencias en el diagnóstico de la hipertensión. En este capítulo se proponen algunas reglas para llevar a cabo un ajuste en la fecha de diagnóstico de la enfermedad. Asimismo, se presentan los resultados obtenidos al aplicar en nuestros datos el procedimiento propuesto.

Capítulo 7: Conclusiones. Este capítulo resume las contribuciones de este trabajo de tesis, así como algunas posibles líneas de trabajo futuro.

Para complementar el trabajo hemos incluido el Apéndice A, que busca documentar aspectos del conjunto de datos creado que podrían ser relevantes para futuros usuarios. Esta documentación utiliza como guía la propuesta de Gebru et al. [19] para un manejo responsable de datos.

Capítulo 2

Diabetes Mellitus tipo 2

La diabetes mellitus tipo 2 ha despertado gran interés en la comunidad científica debido a su alarmante prevalencia en diversas sociedades alrededor del mundo, incluyendo la de nuestro propio país. Esta condición metabólica afecta a millones de personas y su incidencia continúa en aumento, representando un desafío significativo para la salud pública y el bienestar de la población. Nuestro trabajo busca contribuir en el desarrollo de estrategias efectivas de prevención, diagnóstico y tratamiento de la enfermedad, que permitan mitigar los riesgos asociados y mejorar la calidad de vida de quienes la padecen.

Puesto que la enfermedad es el tema central de nuestro estudio, en este capítulo presentamos conceptos fundamentales relacionados con la diabetes mellitus tipo 2. En la siguientes secciones se caracteriza la enfermedad y se describe el procedimiento del Instituto Mexicano del Seguro Social (IMSS) para identificar y atender a pacientes diabéticos. Este procedimiento es de interés para nuestro trabajo, pues representa una de las bases para establecer los parámetros que deben extraerse del SIMF por su relevancia en el análisis de la enfermedad; la Sección 2.3 enlista los parámetros identificados. Finalmente, en la sección 2.4 se describen las principales complicaciones relacionadas con la diabetes, profundizando en la descripción de la hipertensión.

2.1. Diabetes mellitus tipo 2

La diabetes mellitus es una enfermedad crónica caracterizada por niveles elevados de glucosa en sangre, acompañada de un metabolismo alterado de grasas y proteínas. La glucosa en sangre aumenta porque no se puede metabolizar en las células, debido a la falta de producción de insulina por parte del páncreas o a la incapacidad de las células para utilizar eficazmente la insulina que se está produciendo.

Hay tres tipos principales de diabetes mellitus: (a) tipo 1, en la que el páncreas no produce insulina; (b) tipo 2, en el que las células del cuerpo son resistentes a la acción de la insulina que se está produciendo y con el tiempo la producción de insulina disminuye progresivamente; y (c) diabetes gestacional [1], que ocurre durante el embarazo y puede causar algunas complicaciones durante el embarazo y al nacer y aumenta el riesgo de diabetes tipo 2 en la madre y obesidad en la descendencia [21]. Para esta investigación nos enfocaremos principalmente en la diabetes mellitus tipo 2.

La diabetes mellitus tipo 2 es causada por una combinación de factores genéticos y de estilo de vida. Aunque los genes que predisponen a un individuo a la diabetes se consideran un factor esencial en el desarrollo de la enfermedad, la activación de una predisposición genética requiere la presencia de factores ambientales y de comportamiento, particularmente aquellos asociados con el estilo de vida. Los factores más importantes son el sobrepeso, la obesidad abdominal y la inactividad física. Las influencias intrauterinas y de la primera infancia también pueden influir [22].

La diabetes mellitus tipo 2, previamente conocida como Diabetes no insulinodependiente ó diabetes del adulto, representa el 90-95 % de todos los casos de Diabetes. Esta forma engloba a los individuos que tienen una deficiencia de insulina relativa y que presentan resistencia periférica a esta misma [1].

La insulina es una hormona que se produce en el páncreas e interviene en el aprovechamiento metabólico de los nutrientes, permitiendo que la glucosa ingrese a las células para ser utilizada como fuente de energía. Debido a esto, la insulina es importante para

[1] Además de la diabetes melitus, existe la diabetes insípida que es un trastorno caracterizado por la excreción de grandes cantidades de orina hipotónica [20]. Esta enfermedad no se aborda en nuestro trabajo.

mantener los niveles de glucosa estables. Si están demasiados altos o bajos, pueden causar problemas en la salud; esta anomalía en los niveles de glucosa se conoce como Hiperglicemia e Hipoglicemia, respectivamente.

- Hiperglicemia: La hiperglicemia es el término técnico para los niveles altos de glucosa en sangre (azúcar en sangre). El nivel alto de azúcar en sangre ocurre cuando el cuerpo tiene muy poca insulina o cuando el cuerpo no puede usar la insulina correctamente [23].

- Hipoglicemia: La hipoglicemia se presenta cuando el nivel de glucosa en la sangre (azúcar en la sangre) disminuye por debajo de lo necesario para proporcionar suficiente energía para las actividades de su cuerpo. A esto también se le llama baja glucosa en sangre [24].

Se recomienda a las personas que viven con diabetes que sus niveles de glucosa en sangre se mantengan en el rango de 70-99 mg/dl (3.9-5.5 mmol/L) [25]. Para mantener un nivel estable de glucosa se sugiere: hacer ejercicio regularmente, comer alimentos saludables, bajar el exceso de peso, controlar los niveles de estrés, dejar de fumar, tomar los medicamentos que sean necesarios. En la Tabla 2.1 se muestran los niveles de glucemia propuestos por la Norma Oficial Mexicana para el control de la diabetes en México.

Tabla 2.1: Tabla niveles de glucemia [11].

Glucemias	Bueno	Regular	Malo
Glucemia en ayunas (mg/dl)	<70	70-130	>140
Glucemia posprandial de 2h. (mg/dl)	<140	<200	>240

Si las personas que padecen de diabetes mellitus no logran estabilizar los niveles de glucosa, son propensas a presentar numerosas complicaciones graves que se prolongan en el tiempo. Algunas comienzan a los pocos meses de iniciarse la diabetes, aunque la mayoría suelen aparecer al cabo de algunos años y suelen empeorar de forma gradual.

2.2. Manejo de diabetes mellitus tipo 2 en el Instituto Mexicano del Seguro Social (IMSS)

El Instituto Mexicano del Seguro Social (IMSS) ha generado diversos procedimientos para manejar enfermedades comunes en el primer nivel de atención médica.

El algoritmo para la detección y manejo de la diabetes mellitus tipo 2 inicia situando al paciente en tres grupos definidos respecto a la edad y etnicidad. Posteriormente se realizan estudios de laboratorios con un interés en la glucosa plasmática en ayunas (GPA) y hemoglobina glicosilada (HbA1c). Con estos indicadores se decide qué programa se le ofrecerá. La Fig. 2.1 muestra los criterios de decisión establecidos en la Institución.

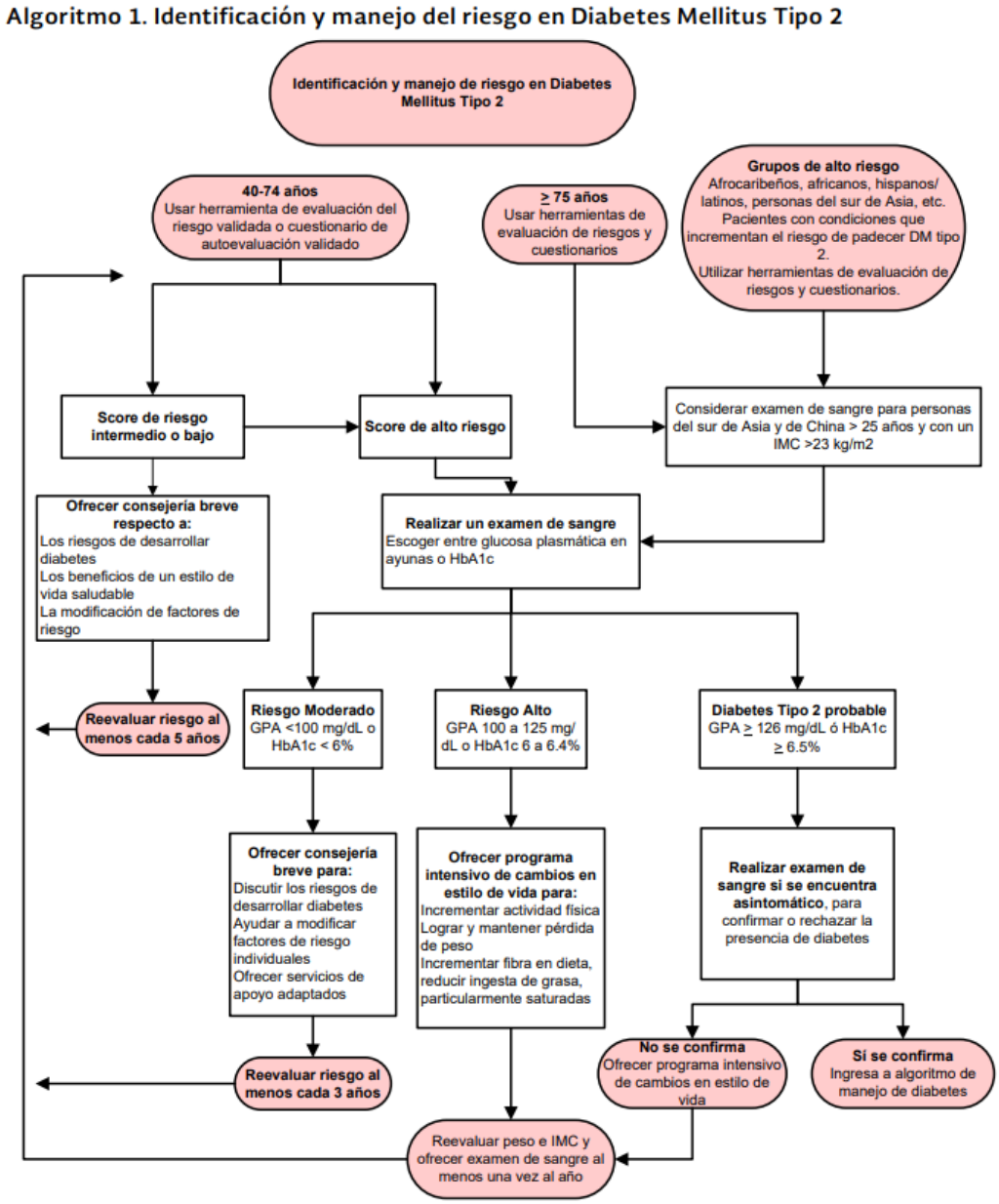


Figura 2.1: Identificación y manejo del riesgo en diabetes mellitus tipo 2 en el Instituto Mexicano del Seguro Social de acuerdo a la guía establecida para el primer nivel de atención [1].

Una vez identificado el paciente con diabetes mellitus tipo 2, se crea un tratamiento farmacológico adecuado para la persona. El algoritmo del Instituto Mexicano del Seguro Social (IMSS) divide en dos grupos a los pacientes: estables e inestables. Los pacientes inestables reciben insulina, mientras que los estables se dividen en 3 grupos de acuerdo a su hemoglobina glicosilada (HbA1c) para proporcionarles un conjunto de medicamentos específicos (ver Fig. 2.2). Finalmente, se realiza un seguimiento trimestral.

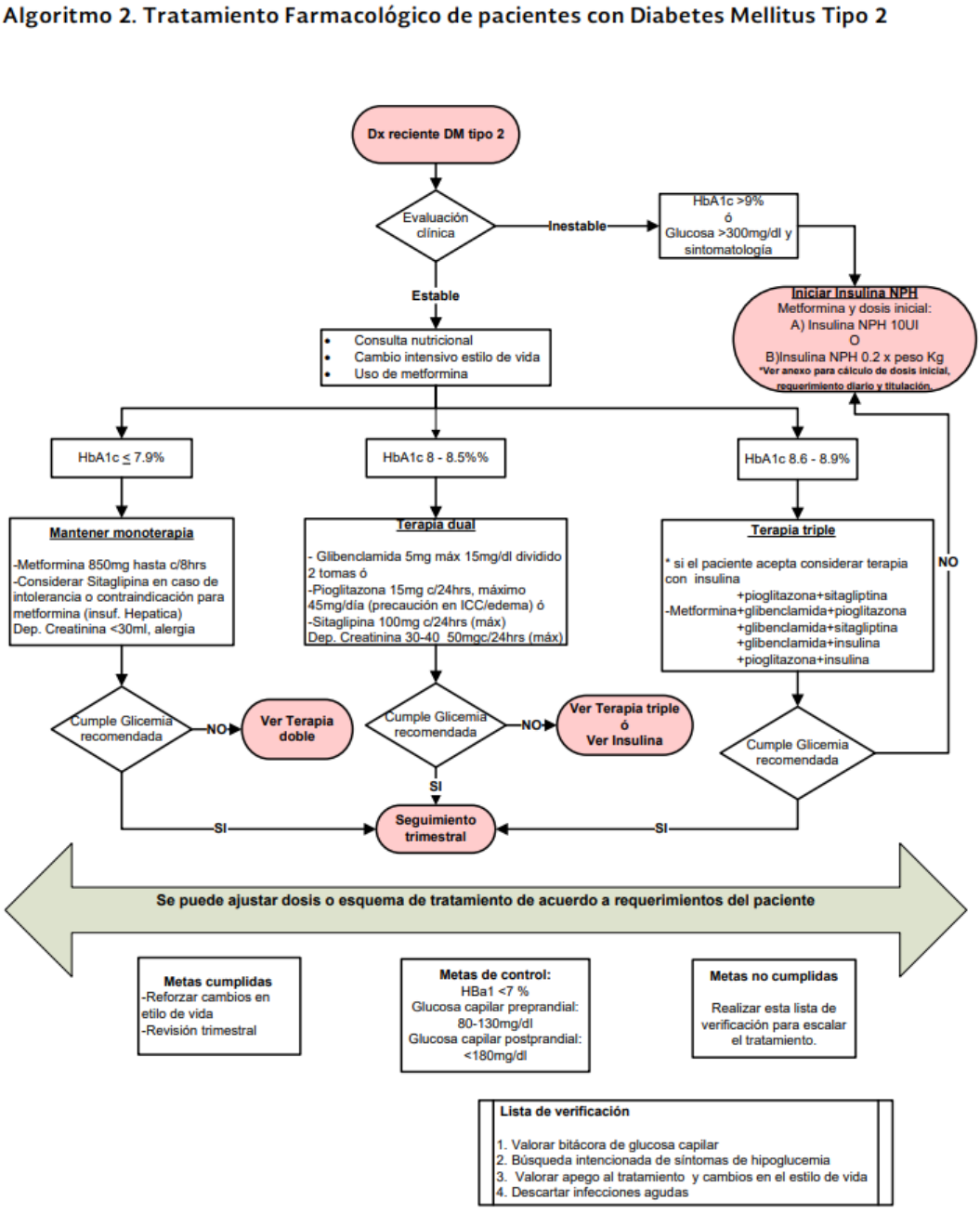


Figura 2.2: Tratamiento farmacológico de pacientes con diabetes mellitus tipo 2 en el Instituto Mexicano del Seguro Social, de acuerdo a la guía establecida para el primer nivel de atención [1].

2.3. Parámetros relacionados con la diabetes mellitus tipo 2

Los pacientes, diagnosticados y potenciales, de diabetes mellitus tipo 2 reciben la recomendación de realizar estudios clínicos para evitar la enfermedad y/o controlarla. En esta investigación se trabajó con personal médico para establecer los parámetros de interés en el análisis de la evolución de la enfermedad. Aunque el punto de partida está dado por los diferentes parámetros de laboratorio propuestos por el Instituto Mexicano del Seguro Social (IMSS), esta información se complementó con lo establecido por la *American Diabetes Association* (ADA) [11] y por autores que han utilizado un enfoque de inteligencia artificial en el estudio de la diabetes mellitus [26].

Los parámetros establecidos se describen a continuación:

- Glucosa: La glucosa es un monosacárido aldosa que está estrechamente relacionado con los procesos de fotosíntesis y respiración, y sirve como reserva de energía y combustible metabólico en la mayoría de los organismos. Como monómero y como parte de estructuras más complejas, como polisacáridos y glucósidos, la glucosa también juega un papel importante en los productos alimenticios modernos, particularmente en lo que respecta al sabor o la estructura [27].

- Colesterol: El colesterol (3-hidroxi-5,6 colesteno) es una molécula indispensable para la vida, desempeña funciones estructurales y metabólicas que son vitales para el ser humano. Se encuentra anclado estratégicamente en las membranas de cada célula, donde modula la fluidez, permeabilidad y, en consecuencia, su función. Esta regulación implica que el contenido en colesterol de las membranas modifica la actividad de las enzimas ancladas en ellas, así como la de algunas proteínas transportadoras y de receptores de membrana. El colesterol proviene de la dieta o es sintetizado por nuestras células (principalmente en los hepatocitos); es precursor de otras biomoléculas fisiológicamente importantes tales como las hormonas esteroideas (andrógenos, estrógenos, progestágenos, gluco y mineralcorticoides), ácidos biliares y la vitamina D [28].

- Triglicéridos: Los triglicéridos forman parte de las grasas del organismo y estructuralmente son moléculas anfipáticas, es decir, con un extremo hidrofóbico y un extremo hidrofílico, en consecuencia para viajar en el plasma deben transportarse de una forma que permita, que el extremo hidrofílico interactúe con la fase acuosa del plasma y que el extremo hidrofóbico no [29].

- HDL: Las lipoproteínas de alta densidad (*High Density Lipoprotein*- HDL), como indica su nombre, se caracterizan por ser las lipoproteínas que presentan mayor densidad (1.063 - 1.21 g/mL) pero menor tamaño (4-13 nm). Están constituidas en un 50 % por proteínas (35 % Apo AI, 10 % Apo AII y 5 % Apo C) y en un 50 % por lípidos (25 % fosfolípidos, 20 % colesterol esterificado, 5 % triglicéridos). Su función es transportar el colesterol desde los tejidos periféricos, incluyendo la pared arterial, hasta el hígado para su posterior excreción en forma de sales biliares, proceso conocido como transporte reverso de colesterol. También pueden transportar el colesterol a órganos endocrinos para la síntesis de hormonas esteroideas [30].

- LDL: Las lipoproteínas de baja densidad (d<1,063g/ml, >1,019g/ml) se caracterizan por su contenido en apoB-100 y tienen como componente lipídico mayoritario los ésteres de colesterol [31], su función es llevar a los tejidos el colesterol, que es captado por las células a través de receptores localizados en la membrana celular. Los niveles de colesterol captados no sólo regulan el número de receptores sino también la cantidad de colesterol producida por las células. Esto permite a las células controlar su nivel de colesterol. El colesterol transportado por las LDL se conoce como colesterol malo ya que, en las personas con niveles elevados de colesterol, las LDL se pueden acumular en las paredes de las arterias, donde pueden ser modificadas y participar en los procesos implicados en el desarrollo de la placa aterosclerótica [32].

- Presión arterial: La presión arterial es la fuerza que la sangre ejerce contra las paredes arteriales. Cuando el médico mide la presión arterial, el resultado se registra con dos números. El primer número, llamado presión arterial sistólica, es la presión causada cuando el corazón se contrae y empuja la sangre hacia afuera. El segundo número, llamado presión arterial diastólica, es la presión que ocurre cuando el corazón se

relaja y se llena de sangre [33].

- Tasa de filtración glomerular: La tasa de filtración glomerular (TFG) es un índice trascendente de la función renal global y uno de los parámetros más importantes de la fisiología humana. Es necesaria para diagnóstico, seguimiento de pacientes con deterioro de la función renal, chequeos epidemiológicos, ajuste de dosis de drogas nefrotóxicas o de eliminación renal, estadificación de la enfermedad renal crónica, etc [34].

- HbA1c: La hemoglobina glucosilada (HbA1c) es la más abundante de los componentes menores de la hemoglobina en los eritrocitos humanos (aproximadamente el 80 % de la HbA1). Así pues, se puede definir como la condensación de la glucosa en la porción N-terminal (grupo valinaterminal) de la cadena beta de la hemoglobina A, siendo por tanto su denominación química N-1-desoxifructosil-beta-Hb; de tal forma que el organismo se encuentra expuesto a la modificación de su hemoglobina por la adición de residuos de glucosa: a mayor glicemia, mayor adición de glucosa a la hemoglobina [35].

- Plaquetas: Las plaquetas son células sanguíneas fundamentales para la hemostasia y son las principales implicadas en alteraciones como la trombosis, trastornos hemorrágicos y en eventos trombóticos hereditarios o adquiridos. Con una estructura celular anucleada con forma discoide de aproximadamente 0.5 x 3.0 µm, tienen su origen de los megacariocitos a través de un proceso endomitótico [36].

- Creatinina: La creatinina es un producto metabólico no enzimático de la creatina y la fosfocreatina, que en condiciones normales se produce a una tasa constante desde el tejido muscular esquelético (alrededor de 2 % por día de la reserva total de creatina). Es una molécula pequeña y no circula unida a proteínas plasmáticas, por lo que se filtra libremente a nivel glomerular [37].

- Ácido úrico: El ácido úrico es el producto final del catabolismo de las purinas en humanos producido mediante la acción enzimática de la xantino óxidorreductasa (XOR)26. Esta enzima se descubrió en la leche y, desde un principio, se pensó que

podría participar activamente en la producción de especies reactivas del oxígeno (EROs). Existe en dos formas que son convertibles entre sí, la xantino oxidasa (XO) y la xantino deshidrogenasa (XDH) [38].

- Urea: La urea es un compuesto químico cristalino e incoloro; de fórmula CO(NH2). Se encuentra en mayor proporción en la orina, en el sudor y en la materia fecal. Es el principal producto terminal del metabolismo de las proteínas en los mamíferos, como los humanos. La urea se forma principalmente en el hígado como un producto final del metabolismo. El nitrógeno de la urea, que constituye el 80 % del nitrógeno en la orina, procede de la degradación de los diversos compuestos con nitrógeno, sobre todo de los aminoácidos de las proteínas en los alimentos [39].

Adicionalmente se agregaron los parámetros de peso, estatura, fecha de nacimiento y sexo.

Tabla 2.2: Rango normal establecido para distintos parámetros clínicos[1].

Parámetro	Rango Normal
Glucosa	70-130
Colesterol	<200
Triglicéridos	<150
Lipoproteínas de alta densidad (HDL)	>50
Lipoproteínas de baja densidad (LDL)	<100
Presión arterial	<120/80
Tasa de filtración glomerular	90 a 120 mL/min/1.73 m2
HbA1c	<7
Plaquetas	150 a 400 $\times$ 10^9/L
Creatinina	0.7 a 1.3 mg/dL
Ácido úrico	3.5 y 7.2 mg/dL
Urea	12-54 mg/dl

2.4. Complicaciones de la diabetes mellitus tipo 2

La diabetes mellitus tipo 2 es una enfermedad crónica y es común que afecte gradualmente diferentes funciones del cuerpo.

Para analizar la evolución de la enfermedad y la aparición de complicaciones asociadas, en esta investigación se consideraron las categorías usadas en la *International Classification of Diseases* (ICD) para la diabetes mellitus.

La ICD es una clasificación médica, listada por la Organización Mundial de la Salud (*World Health Organization*), que contiene códigos para enfermedades, signos y síntomas, hallazgos anormales, quejas, circunstancias sociales y causas externas de lesiones o enfermedades. En esta clasificación, las complicacioes relacionadas con la diabetes mellitus tipo 2 son las siguientes:

- Diabetes mellitus tipo 2 con coma: El coma hipoglucémico se define como un estado en el que el paciente no está despierto (o responde solo al dolor), con una concentración de glucosa en sangre de 2.72 mmol/L (49 mg/dL) o menos, y respuesta sintomática (un retorno de la conciencia) a la administración de glucosa intravenosa [40].

- Diabetes mellitus tipo 2 con cetoacidosis: Diabetes cetoacidosis es un trastorno metabólico que consta de tres anomalías concurrentes: glucosa en sangre alta, cuerpos cetónicos altos y acidosis metabólica. Estos trastornos individualmente pueden asociarse con otras enfermedades o eventos metabólicos [41].

- Diabetes mellitus tipo 2 con complicaciones renales: La definición y clasificación de la enfermedad renal crónica (ERC) ha evolucionado con el tiempo, pero las guías internacionales actuales definen esta condición como una función renal disminuida mostrada por una tasa de filtración glomerular (TFG) de menos de 60 ml/min por 1·73 m2, o marcadores de daño renal, o ambos, de al menos 3 meses de duración, independientemente de la causa subyacente. La diabetes y la hipertensión son las principales causas de ERC en todos los países de ingresos altos y medios, y también en muchos países de ingresos bajos [42].

- Diabetes mellitus tipo 2 con complicaciones oftálmicas: Las complicaciones oftálmicas de la hiperglucemia afectan de manera profunda la córnea y la retina. La córnea experimenta 4 veces más glucosa en la película lagrimal para diabéticos que

en las lágrimas de control. El setenta por ciento de los diabéticos sufren de complicaciones corneales denominadas colectivamente queratopatía diabética. La retinopatía diabética es la principal causa de pérdida visual en personas con diabetes, siendo responsable de la mayoría de los casos de disminución de la visión en esta población. Además, se destaca como la principal causa de ceguera en individuos mayores de 50 años.[43].

- Diabetes mellitus tipo 2 con complicaciones neurológicas: Las complicaciones neurológicas de la diabetes son el resultado del impacto de la hiperglucemia tanto en la función como en la estructura de los nervios. Los déficits de la conducción nerviosa, la resistencia al bloqueo de la conducción isquémica y la percepción alterada a los estímulos térmicos, táctiles y vibratorios son evidentes en la fase metabólica temprana de la enfermedad. La lesión estructural, que afecta a los axones y las células de Schwann, se hace evidente más tarde cuando se establece la neuropatía crónica [44].

- Diabetes mellitus tipo 2 con complicaciones circulatorias: Las complicaciones del pie siguen siendo una preocupación importante para pacientes con diabetes mellitus tipo 2. Las principales complicaciones del pie incluyen ulceración del pie, celulitis, absceso, gangrena húmeda, gangrena seca y fascitis necrotizante, con diferentes conceptos fisiopatológicos detrás de cada una de ellas. La gangrena se produce debido a la reducción del suministro de sangre en los tejidos corporales que conduce a la necrosis. Esta condición puede surgir debido a una lesión, infección u otras condiciones de salud, principalmente diabetes [45].

- Diabetes mellitus tipo 2 con complicaciones no especificadas: Las complicaciones no especificadas se agrupan dentro de los grupos relacionados con el diagnóstico *MS-DRG v39.0*, y corresponden a trasplante simultáneo de páncreas y riñón, trasplante de páncreas, trasplante simultáneo de páncreas y riñón con hemodiálisis, diabetes mellitus tipo 2 con múltiples complicaciones, diabetes mellitus tipo 2 con otras complicaciones especificadas, diabetes mellitus tipo 2 sin complicaciones, diabetes mellitus tipo 2 sin múltiples complicacionesc.

- Diabetes mellitus tipo 2 sin complicaciones: Las personas que tienen controlada la diabetes mellitus tipo 2 son aquellas que mantienen las concentraciones de azúcar en sangre, o de glucosa, dentro de unos márgenes saludables.

- Hipertensión: La hipertensión es un trastorno frecuente, crónico, relacionado con la edad, que a menudo conlleva complicaciones cardiovasculares y renales debilitantes. La hipertensión no está relacionada directamente con la diabetes, como se puede observar en la tabla 2.3, pero es muy probable desarrollar hipertensión teniendo diabetes mellitus, por lo cual, nos enfocaremos en esta complicación. Se dará más detalle en la sección de Hipertensión.

En la Tabla 2.3 se muestran las diferentes categorías de complicaciones y su respectivo código ICD (*International Classification of Disease* por sus siglas en ingles) . Es importante tener en cuenta que estos códigos ICD se utilizan para registrar y codificar las complicaciones de la diabetes mellitus tipo 2 en los sistemas de salud, lo que permite una mejor monitorización y tratamiento de la enfermedad.

Tabla 2.3: Diversos códigos ICD relacionados con la diabetes mellitus tipo 2 [12].

Complicación	Código ICD
Diabetes mellitus tipo 2 con coma	E11.0
Diabetes mellitus tipo 2 con cetoacidosis	E11.1
Diabetes mellitus tipo 2 con complicaciones renales	E11.2
Diabetes mellitus tipo 2 con complicaciones oftálmicas	E11.3
Diabetes mellitus tipo 2 con complicaciones neurológicas	E11.4
Diabetes mellitus tipo 2 con complicaciones circulatorias	E11.5
Diabetes mellitus tipo 2 con otras complicaciones especificadas	E11.6
Diabetes mellitus tipo 2 con múltiples complicaciones	E11.7
Diabetes mellitus tipo 2 con complicaciones no especificadas	E11.8
Diabetes mellitus tipo 2 sin complicaciones	E11.9
Hipertensión	I10X

Existen otros códigos ICD que, dependiendo del estudio, podrían ser relevantes para el análisis de las complicaciones de la diabetes mellitus tipo 2. Además, los códigos pueden variar según la versión de la Clasificación Internacional de Enfermedades que se esté utilizando. Por esta razón, fue fundamental trabajar a lo largo del proyecto con profesionales

de la salud familiarizados con la codificación y documentación de las complicaciones de la enfermedad.

También con base en la colaboración con personal médico, se decidió analizar con más detalle el caso de pacientes diagnosticados simultáneamente con diabetes e hipertensión. Esto obedece, en parte, a la frecuencia con la que se presenta este perfil.

En la siguiente sección se describe con mayor profundidad la hipertensión, destacando aspectos que fueron relevantes en nuestro estudio para identificar errores en el registro del diagnóstico de la enfermedad.

2.4.1. Hipertensión

La aparición concurrente de la hipertensión arterial en pacientes que padecen diabetes mellitus tipo 2 es un fenómeno de alta frecuencia, que en ocasiones plantea un dilema diagnóstico. Surge la interrogante de si la diabetes mellitus precede a la hipertensión arterial, o si es esta última la que potencialmente induce la aparición de la diabetes.

La complejidad de la interacción entre ambas enfermedades plantea desafíos en la determinación precisa del origen y secuencia de estas condiciones, lo cual resulta de suma importancia, ya que impacta directamente en la estrategia de tratamiento y en el enfoque clínico dirigido a estos pacientes.

Para entender el problema de la hipertensión, recordemos que la sangre impulsada por el corazón fluye a través del sistema circulatorio, específicamente por el sistema arterial, sometida a una presión denominada presión arterial (PA), o tensión arterial [46]. La presión sanguínea depende de muchos factores, como la cantidad de sangre que esté bombeando el corazón y el diámetro de las arterias a través de las cuales pasa la sangre. De hecho, se ha establecido que la presión arterial resulta de la interrelación de tres factores: gasto cardíaco, resistencia vascular periférica y volumen intravascular [47].

Medir la PA es fácil, tanto para el propio paciente como para cualquier persona, si se cuenta con un aparato adecuado. Clínicamente, los niveles de PA los expresamos en milímetros de mercurio (mmHg) pero la PA tiene en realidad dos componentes: la presión arterial sistólica (PAS), que viene determinada por el impulso cardíaco generado por las contracciones del ventrículo izquierdo y que comúnmente es denominada la alta; y la

presión arterial diastólica (PAD), la baja, que depende de las resistencias que oponen las arterias al paso de la sangre [46]. Así, la presión sanguínea se mide tanto en el momento en el que el corazón se contrae, llamado sístole, como en el momento en el que se relaja, llamado diástole [48].

Existen diferentes formas de medir la presión arterial dependiendo del paciente, algunos ejemplos son los siguientes:

- Método auscultatorio: El método auscultatorio consiste en colocar un estetoscopio sobre la arteria humeral y se mide con un esfigmomanómetro, el cual consta de un manguito o cojín acoplado y un tubo de caucho que a su vez se une a una perilla comprensible que se usa para insuflar el manguito [49].

- Método oscilométrico: El método oscilométrico tiene como principio el análisis de las oscilaciones de la pared arterial, según las condiciones de presión existentes por dentro y fuera de ella. La oscilación captada por el manguito será máxima cuando exista un equilibrio de presión por dentro y fuera de la arteria y esto coincide con la presión arterial media, el aparato capta esta oscilación máxima y determina la presión media y luego por un cálculo aritmético determina la máxima y mínima[50].

- Método ultrasonográfico: Los equipos que incorporan esta técnica usan un transmisor de ultrasonido ubicado sobre la arterial braquial bajo el brazalete del esfigmomanómetro. Cuando el brazalete es desinflado, el movimiento de la pared arterial causado por la PA sistólica origina un cambio en la fase doppler en el reflejo de la onda del ultrasonido [51].

Una vez definida la presión arterial podemos pasar a la definición de hipertensión, que es una enfermedad caracterizada por un aumento de la presión en el interior de las arterias. Como consecuencia, las arterias se van dañando provocando enfermedades cardiovasculares.

Hay diferentes factores de riesgos para desarrollar hipertensión arterial. Algunos de ellos son [52]:

- Diabetes mellitus.

- Dieta no saludable.
- Inactividad física o sedentarismo.
- Obesidad.
- Ingesta alcohólica.
- Tabaquismo.
- Antecedentes familiares y genética.
- Edad avanzada.
- Etnia afrodescendiente

Como se puede observar, en los factores de riesgo aparece la diabetes mellitus, debido a que es muy común desarrollar hipertensión teniendo diabetes.

2.4.2. Diagnóstico y tratamiento de la hipertensión.

El IMSS cuenta con un algoritmo para la detección y tratamiento de la hipertensión arterial, que se describe en las figuras 2.3 y 2.4.

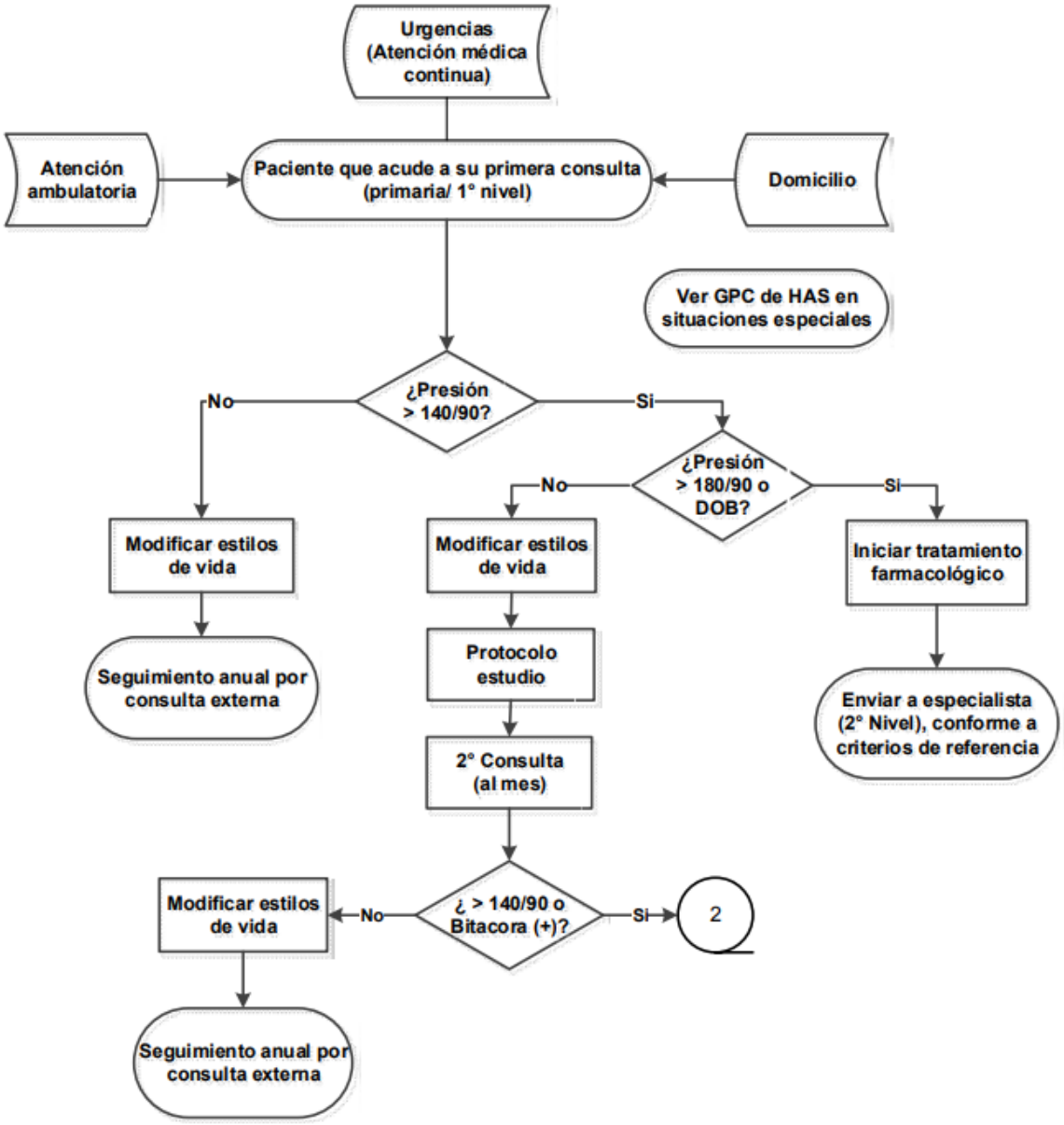


Figura 2.3: Detección, diagnóstico y tratamiento de la hipertensión arterial sistémica [2]

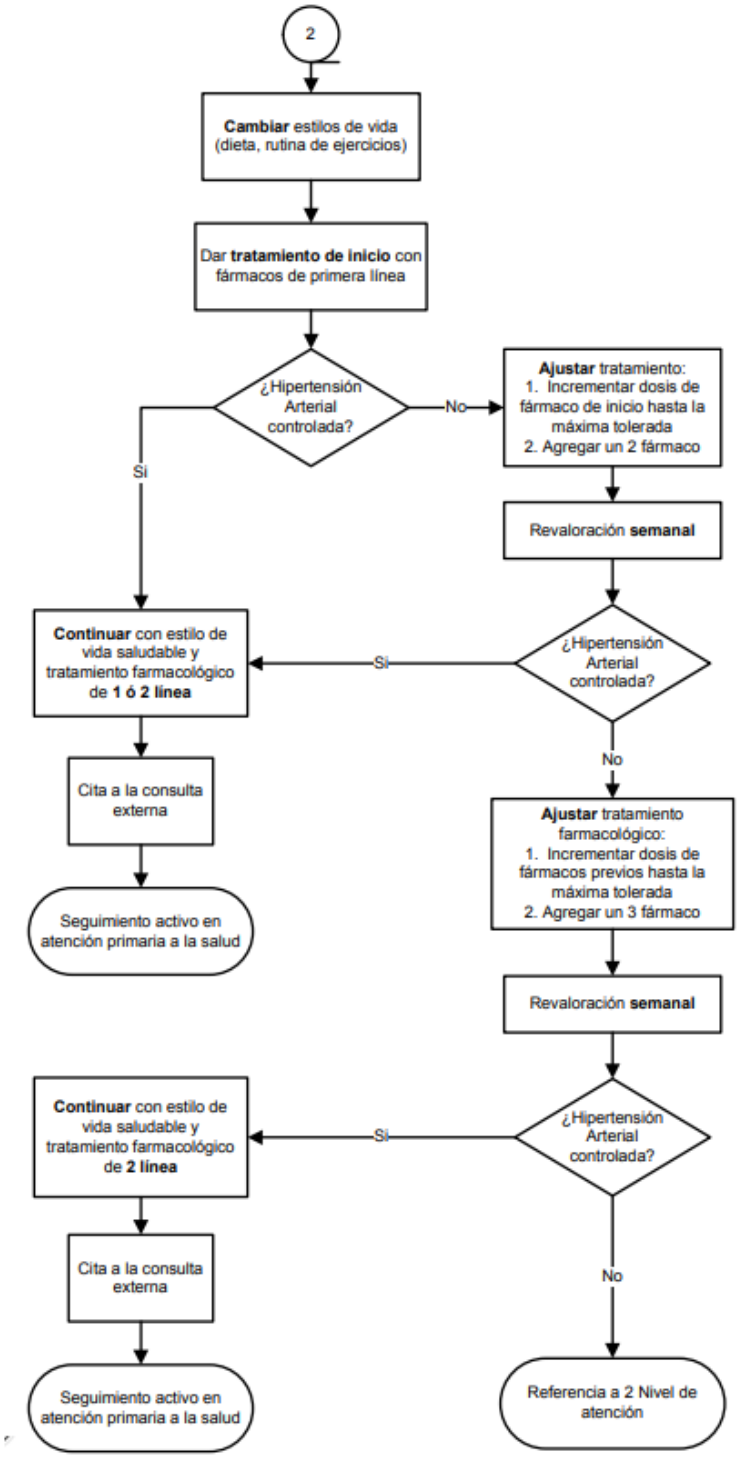


Figura 2.4: Detección, diagnóstico y tratamiento de la hipertensión arterial sistémica [2]

Para evitar la hipertensión es recomendable seguir una dieta saludable con menos sal, ejercitarse regularmente.

Las guías de práctica clínica del IMSS, señalan como segunda línea de acción para el tratamiento de la hipertensión arterial la prescripción de fármacos antihipertensivos que pueden ayudar a reducir la presión arterial, como los diuréticos, los inhibidores de la enzima convertidora de angiotensina (IECA), los antagonistas de los receptores de angiotensina II (ARA II), los bloqueadores de los canales de calcio y los betabloqueantes.

Cada tipo de medicamento funciona de manera diferente en el cuerpo y puede tener efectos secundarios específicos, por lo que los pacientes generalmente trabajan de cerca con su médico para encontrar el medicamento más adecuado para ellos y que proporciona los mayores beneficios con el menor riesgo posible.

2.4.3. Complicaciones de la hipertensión

Aunque la hipertensión es una enfermedad que puede tratarse, incluso con medicamentos puede haber valores altos en la presión arterial que podrían causar un derrame cerebral. Es importante analizar las diferentes complicaciones que puede causar la hipertensión, ya que unida con la diabetes mellitus tipo 2 puede causar mucho daño a las personas.

A continuación se listan las complicaciones más comunes de la hipertensión.

Daño en las arterias:

- Aneurismas: Existen diferentes tipos de aneurismas, una de las más peligrosa es la cerebral la cual es una protuberancia o dilatación en un vaso sanguíneo en el cerebro. En ocasiones, tiene el aspecto de una cereza que cuelga de un tallo . Un aneurisma cerebral puede presentar una pérdida o una rotura, y causar sangrado en el cerebro (accidente cerebrovascular hemorrágico) [53].

Daño al corazón:

- Arteriopatía coronaria: Es una enfermedad en la que el aporte de sangre al miocardio (músculo cardíaco) está bloqueado en parte o en su totalidad. Esta provoca un estrechamiento en una o más de estas arterias puede causar una interrupción del riego sanguíneo, lo que se manifiesta como dolor torácico (angina) o infarto de miocardio (ataque al corazón) [54].

- Insuficiencia cardiaca: Es una alteración de la función cardiaca, la cual causa ineptitud del corazón para hacer circular la cantidad óptima de sangre que se requiere en el organismo. Según avanza esta ineptitud pueden presentarse alteraciones como dilatación cardiaca, congestión venosa e hidropesía (acumulación anormal de líquido en algún cavidad o tejido del organismo). Estos fenómenos son de aparición tardía 1 significan que el corazón ha perdido su capacidad de realizar su trabajo aun en reposo [55].

Daño al cerebro:

- Ataque isquémico transitorio: El ataque isquémico transitorio es un déficit neurológico focal, súbito, de etiología isquémica en el que no existe daño neuronal permanente. Habitualmente no dura más de 60 min. Se caracteriza por tener una recuperación casi espontánea, no hay evidencia de lesión en los estudios imagenológicos como la resonancia magnética (RM) [56].

- Demencia vascular:El término de demencia vascular comprende a todas aquellas demencias secundarias a una o varias lesiones vasculares cerebrales, de cualquier etiología. Este término incluye las siguientes entidades;

    - Demencia multi-infarto: demencia secundaria a la repetición de infartos corticales en el territorio de arterias de calibre mediano o grande.
    - Demencia por infarto estratégico: demencia debida a un infarto en una localización tal que afecta a varias funciones cognitivas
    - Demencia vascular subcortical: acumulación de infartos lacunares o lesiones vasculares de la sustancia blanca periventricular y profunda por enfermedad de los vasos pequeños
    - Demencia post-ictus: cualquier tipo de demencia que se desarrolla después de un ictus
    - Demencia mixta: la combinación de distintas alteraciones (amiloidopatía,sinucleinopatía, taupatía y enfermedad vascular).

- Demencia por lesiones hemorrágicas [57].

Daño a los riñones:

- Glomeruloesclerosis (Cicatrización renal): La glomeruloesclerosis es el término usado para describir el tejido cicatrizado que se presenta dentro del riñón en las bolas pequeñas de los diminutos vasos sanguíneos llamados glomérulos. Los glomérulos ayudan a los riñones a filtrar la orina de la sangre [58].

- Insuficiencia renal aguda: La insuficiencia renal aguda (IRA) se define como la disminución en la capacidad que tienen los riñones para eliminar productos nitrogenados de desecho, instaurada en horas a días. La eliminación de productos de desecho no es la única función de estos órganos, quienes además desempeñan un papel imprescindible en la regulación del medio interno, manteniendo el equilibrio electrolítico y la volemia en unos márgenes muy estrechos [59].

Daño a los ojos:

- Coroidopatía: La coroides, a diferencia de los vasos retinianos, está comandada por el tono simpático. En la hipertensión arterial la coroides sufre fenómenos de isquemia, observándose lóbulos coroideos sin perfusión debido a necrosis fibrinoide de los vasos. Estas zonas de ausencia de perfusión coroidea producen focos de necrosis isquémica en el epitelio pigmentario retiniano suprayacente, que se denominan manchas de Elschnig [60].

- Retinopatía hipertensiva: La retinopatía hipertensiva es un daño en la retina ocasionado por la elevación abrupta de la presión arterial ya sea por causa primaria o secundaria. Produce desde disminución de la agudeza visual hasta ceguera (ver Fig. 2.5). Diferentes estudios han demostrado que la retinopatía hipertensiva se asocia a la subida de las cifras tensionales sin embargo también influye factores como la arterioesclerosis, la enfermedad de las arterias carótidas y edad avanzada [61].

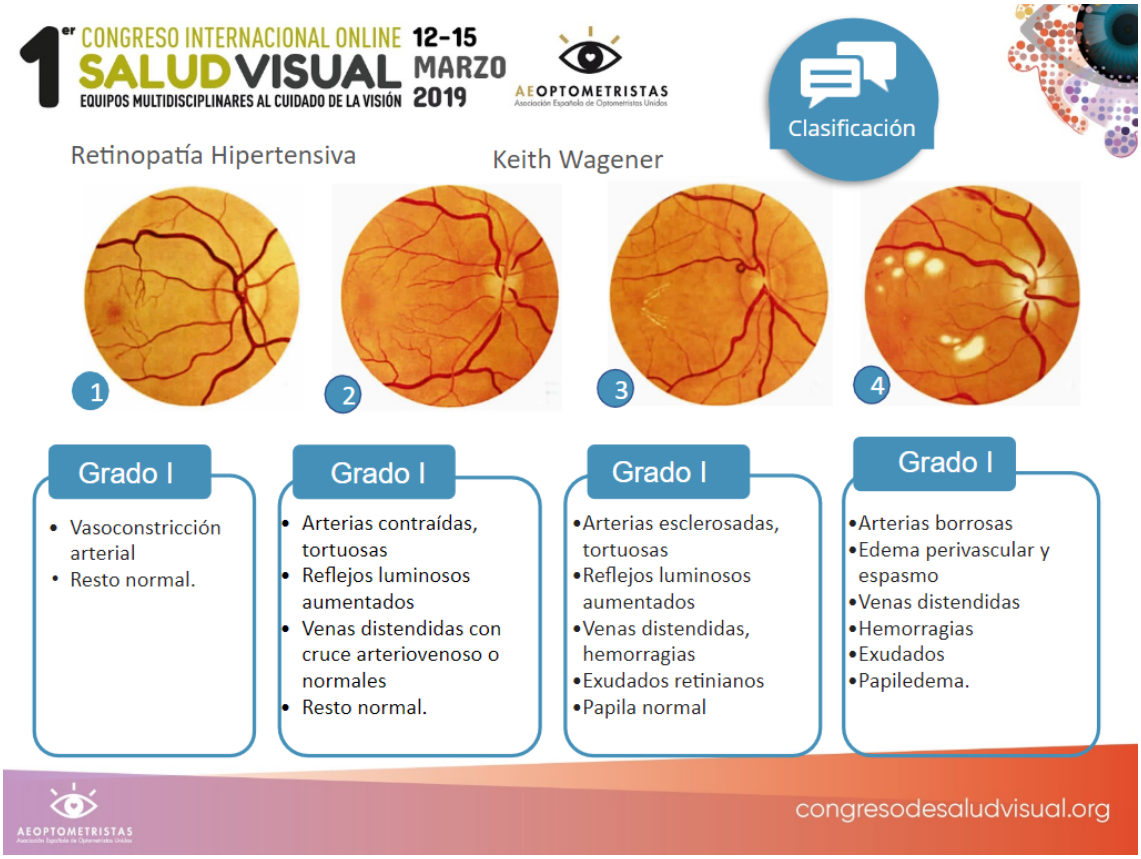


Figura 2.5: Grados de retinopatía hipertensiva [3]

Capítulo 3

Redes neuronales artificiales

El Sistema de Información de Medicina Familiar (SIMF) contiene datos en tablas relacionales que resultan de gran relevancia para el estudio de la diabetes. Sin embargo, una buena parte de la información sobre pacientes y sobre resultados de laboratorio se encuentra concentrado en notas médicas, es decir, en textos en formato libre, con tamaño e información variable.

Para recuperar esta información e integrarla a la conformación de historiales clínicos fue necesario implementar herramientas para el Reconocimiento de entidades nombradas (NER, por las siglas en inglés de *Named Entity Recognition*). Dichas herramientas se basan en redes neuronales artificiales.

En este capítulo se describen brevemente los fundamentos de las redes neuronales artificiales y de las arquitecturas relevantes para el NER.

3.1. Aprendizaje automático

Las redes neuronales forman parte del aprendizaje automático supervisado, una rama del aprendizaje automático (*machine learning*) que se ocupa de construir algoritmos que mejoran automáticamente con ejemplos del fenómeno que se desea analizar.

En años recientes se han desarrollado muchas aplicaciones exitosas de aprendizaje automático, que van desde programas que aprenden a detectar transacciones fraudulentas con tarjetas de crédito, hasta sistemas de filtrado de información que aprenden las preferencias

de lectura de los usuarios y vehículos que aprenden a circular por la vía pública [62].

Además del aprendizaje supervisado, el aprendizaje automático abarca el aprendizaje no supervisado y el aprendizaje por refuerzo, así como algunas formas híbridas como el aprendizaje semi-supervisado y el aprendizaje autosupervisado. Puesto que nuestro proyecto se centra en el uso de redes neuronales, nos enfocaremos en el aprendizaje supervisado.

El aprendizaje supervisado implica aprender un mapeo entre un conjunto de variables de entrada $X = (x_1, x_2, \ldots, x_N)$ y una variable de salida $Y = (y_1, y_2, \ldots, y_N)$ y, posteriormente, aplicar dicho mapeo para predecir el valor de Y para datos no vistos. El aprendizaje supervisado es la metodología más importante en el aprendizaje automático y también tiene una importancia central en el procesamiento de datos multimedia [63].

Para situar adecuadamente los algoritmos que se describen en este capítulo, podemos decir que el aprendizaje automático es un subcampo de la inteligencia artificial, el aprendizaje profundo es un subcampo del aprendizaje automático, y las redes neuronales son la parte esencial de los algoritmos de aprendizaje profundo (ver Fig. 3.1).

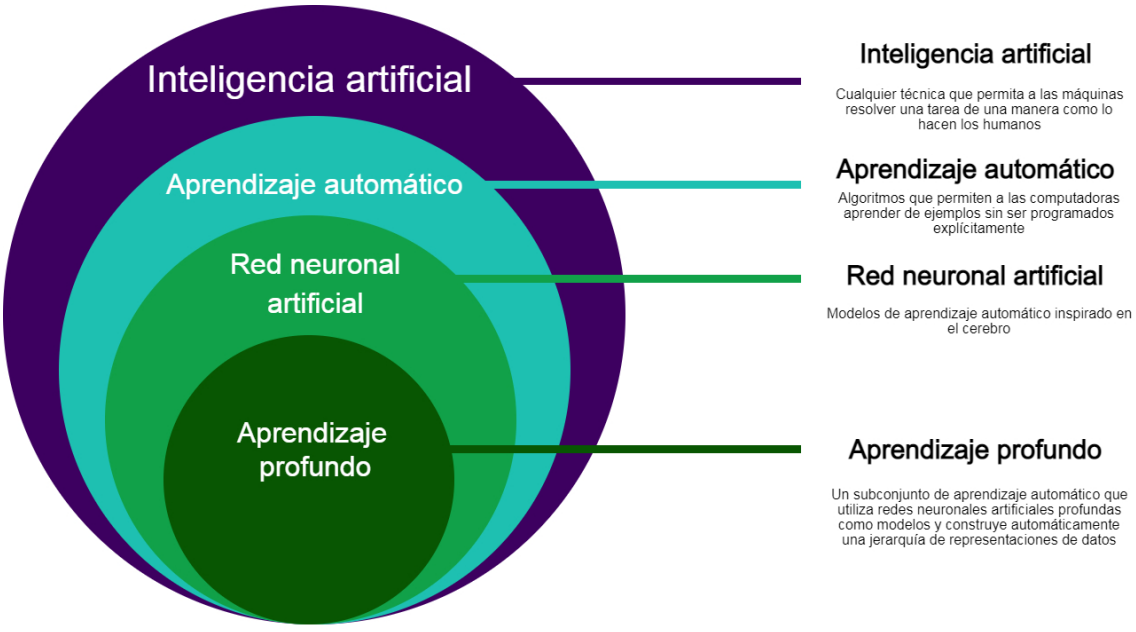


Figura 3.1: Subdisciplinas en inteligencia artificial [4].

3.2. Aprendizaje profundo

Reconocer patrones a partir de datos evita la necesidad de especificar formalmente todo el conocimiento que requiere un algoritmo para clasificar. La jerarquía de conceptos permite que la computadora capture patrones complicados a partir de otros más simples. Si dibujamos un gráfico que muestre cómo se construyen estos patrones uno encima del otro, el gráfico es profundo, con muchas capas, por esta razón el enfoque recibe el nombre de aprendizaje profundo [64].

El aprendizaje profundo permite que los modelos computacionales que se componen de múltiples capas de procesamiento aprendan representaciones de datos con múltiples niveles de abstracción. Estos métodos han mejorado drásticamente el estado del arte en el reconocimiento de voz, el reconocimiento de objetos visuales, la detección de objetos y muchos otros dominios, como el descubrimiento de fármacos y la genómica [65].

En esta sección se detallarán de forma general los algoritmos más conocidos del aprendizaje profundo, así como algunos conceptos fundamentales en su funcionamiento.

3.2.1. Redes neuronales profundas

Uno de los algoritmos más utilizados en el aprendizaje profundo son las redes neuronales profundas (*deep neural network* o DNN, por las siglas en inglés). Una DNN es una colección de neuronas (nodos de un algoritmo) organizadas en una secuencia de múltiples capas, donde las neuronas reciben como entrada las activaciones (resultados de cálculos) de neuronas de la capa anterior y realizan un cálculo simple cuyo resultado se envía a las neuronas de la siguiente capa (ver Fig. 3.2).

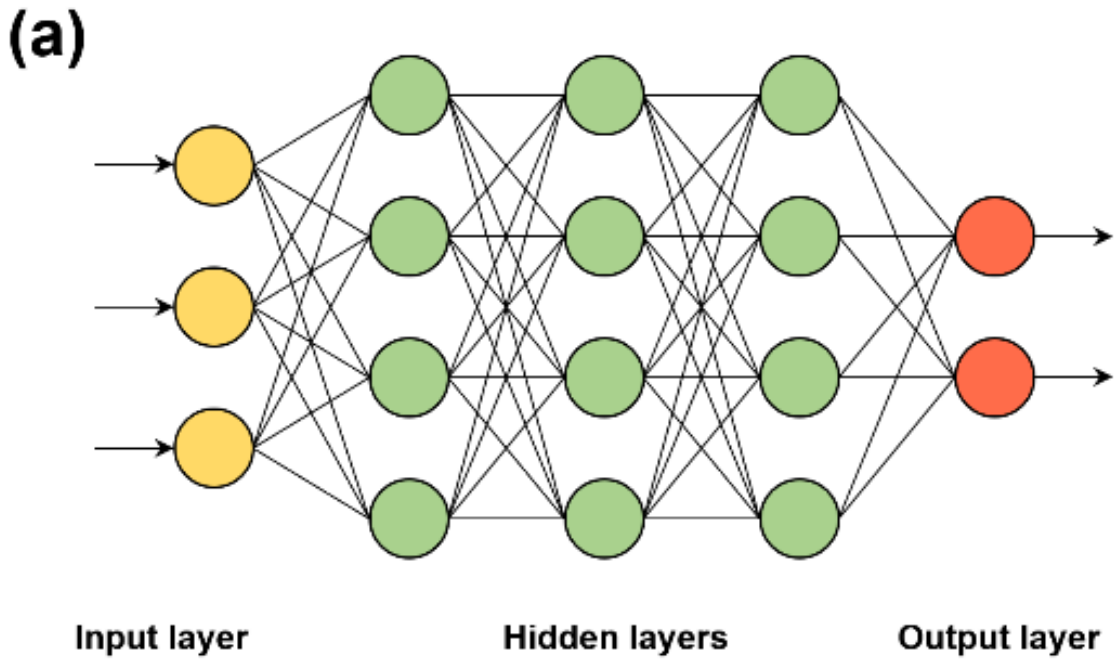


Figura 3.2: Ejemplo de una red neuronal profunda (deep neural network), organizada en una secuencia de múltiples capas (layers). Imagen tomada de [5].

Las neuronas de la red implementan conjuntamente un mapeo no lineal complejo desde la entrada hasta la salida. Este mapeo se aprende de los datos adaptando los pesos de cada neurona usando una técnica llamada retropropagación [66]. Esta ajusta repetidamente los pesos de las conexiones en la red para minimizar una medida de la diferencia entre el vector de salida real de la red y el vector de salida deseado [67]. Este proceso se ilustra en la Fig. 3.3.

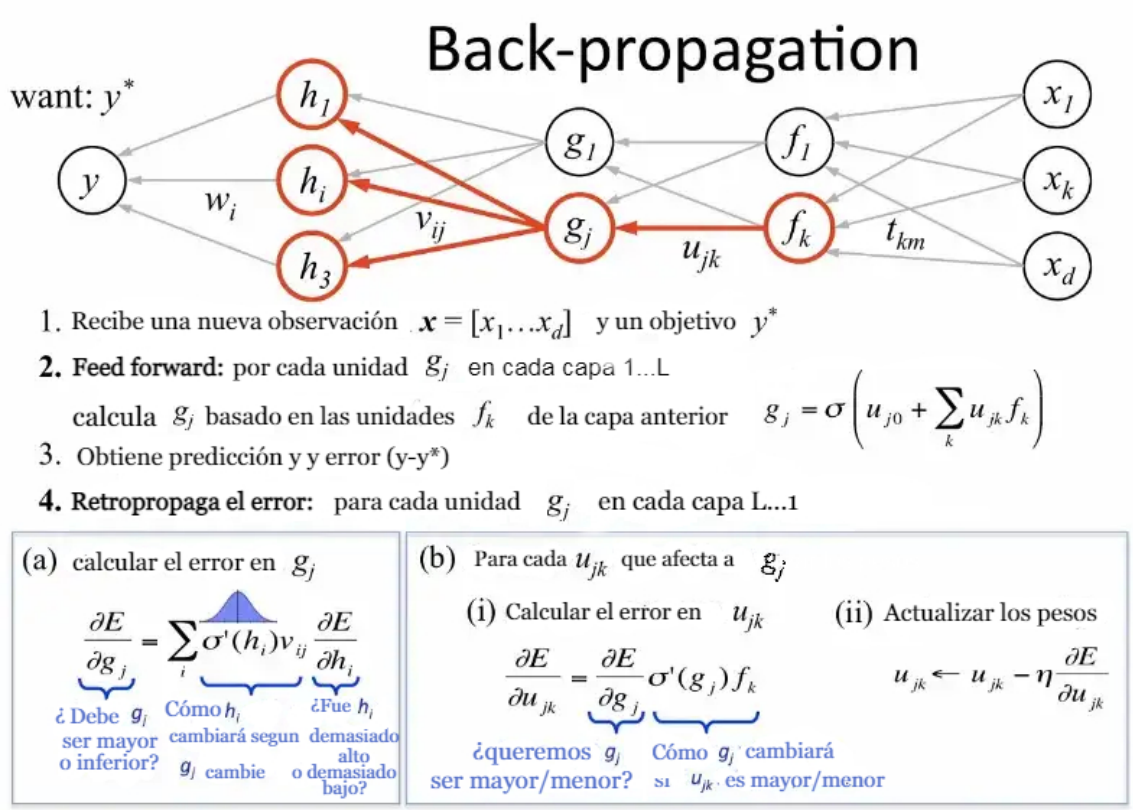


Figura 3.3: Método de retropropagación [6].

La primera fase de las DNN es una capa de entrada que, como su nombre indica, representa los datos de entrada del modelo. Estos pueden ser tan simples como datos escalares o datos más complejos como vectores, matrices o arreglos multidimensionales.

La segunda son las capas ocultas, que se encuentran entre la entrada y la salida del algoritmo. En estas capas la función aplica pesos a las entradas y las dirige a través de una función de activación como la salida. Las funciones de activación se utilizan especialmente en las neuronas artificiales para transformar una señal de entrada en una señal de salida que, a su vez, se alimenta como entrada a la siguiente capa de la pila. En una red neuronal artificial, calculamos la suma de productos de entradas y sus correspondientes pesos y finalmente aplicamos una función de activación para obtener la salida de esa capa en particular y suministrar como entrada a la siguiente capa [68]. En resumen, las capas ocultas realizan transformaciones no lineales de las entradas ingresadas a la red [69].

Las funciones de activación son esenciales en las redes neuronales porque permiten la introducción de la no linealidad necesaria para que la red pueda aprender y modelar

relaciones complejas en los datos. Cada una de estas funciones tiene diferentes propiedades y puede ser más adecuada para diferentes tipos de tareas y arquitecturas de red. Algunas de las funciones de activación más comunes son la función sigmoide, la función ReLU y la función tangente hiperbólica. La Fig. 3.4 muestra algunas de las funciones de activación más comunes en las redes neuronales artificiales.

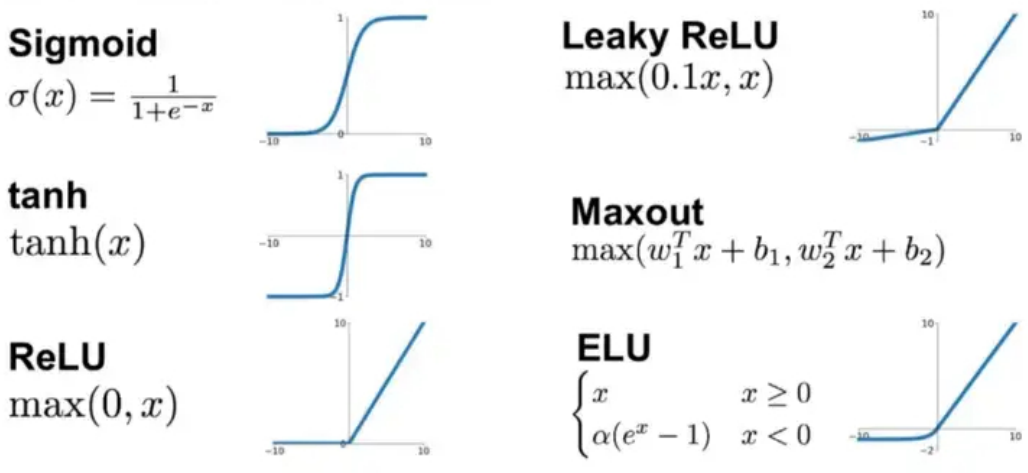


Figura 3.4: Funciones de activación más comunes en las redes neuronales artificiales. Imagen tomada de [7].

La última parte de la red neuronal es la capa de salida, que es la capa responsable de predecir los valores de interés. Esta puede predecir de uno a múltiples valores.

Existen diferentes tipos de DNN, definidos por características específicas en su arquitectura. Por ejemplo, un perceptrón multicapa es una DNN compuesta por capas completamente conectadas .

Un segundo tipo de DNN son las redes convolucionales, que se describen en la siguiente sección.

3.2.2. Redes neuronales convolucionales

Las redes neuronales convolucionales (CNN, por sus siglas en inglés) son algoritmos de aprendizaje profundo que toman imágenes de entrada y las convoluciona con filtros o núcleos para extraer características.

Una imagen $N \times N$ es convolucionada con un filtro $f \times f$ y esta operación de convolución aprende la misma característica en toda la imagen [70]. Por ejemplo, después de entrenar una red neuronal con ciertas imágenes de expresiones faciales de personas con emociones determinadas, la red neuronal será capaz de identificar las emociones representadas en una imagen similar.

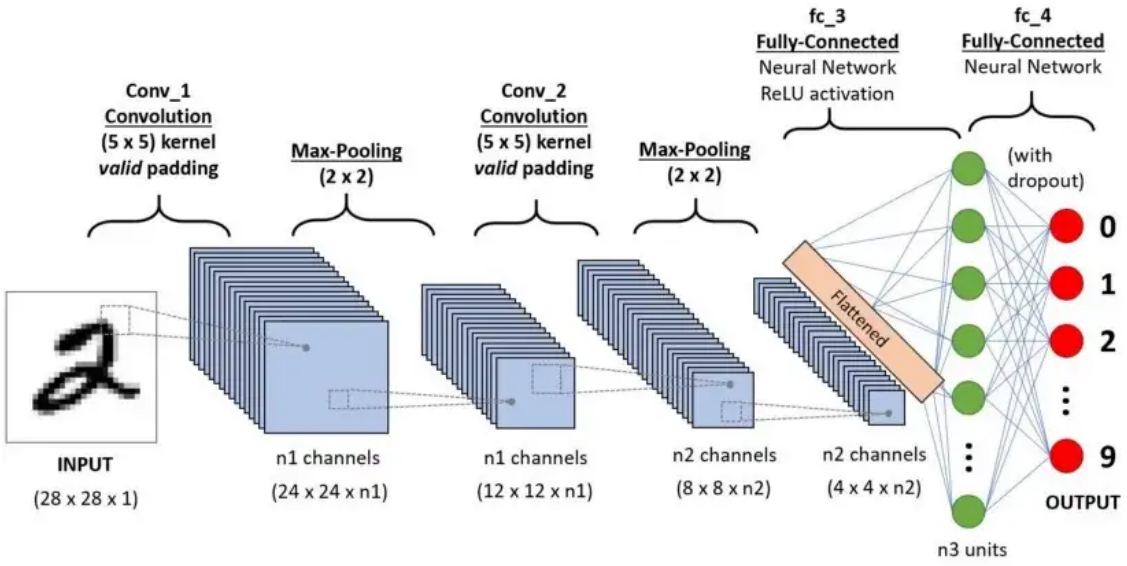


Figura 3.5: Proceso CNN[8]

Los datos de entrada de las CNN pueden ser cualquier imagen, pero esta será guardada como una matriz que representa el valor del píxel (intensidad de color) en cada posición. Por ejemplo, en el caso del formato RGB, cada imagen se compone de 3 matrices representando los valores en los canales RGB (red, green, blue).

La capa convolucional es una combinación matemática que hace un producto punto de dos funciones para producir una tercera. Esta operación se lleva a cabo mediante un detector de características (*feature detector*), también llamado kernel, que es un filtro que se utiliza para extraer las características de las imágenes. El filtro es esencialmente una

matriz que se mueve sobre los datos de entrada y que realiza el producto punto con la subregión de los datos de entrada para obtener una nueva matriz como salida.

El filtro se va moviendo de acuerdo al valor del *stride*. Stride se puede definir como la cantidad por la que se desplaza un filtro/kernel. El deslizamiento del filtro sobre la imagen de entrada no tiene por qué ser de solo una unidad en la dirección deseada; podemos controlar el movimiento deslizante con un número de nuestra elección, según el caso de uso. Los pasos más grandes a menudo ayudan a reducir el cálculo, generalizando el aprendizaje de funciones [71].

El siguiente paso de las redes convolucionales es el *pooling*, que se utiliza para reducir el tamaño del kernel sustituyendo (generalmente) una ventana de agrupación cuadrada de $q \times q$ escalares con un solo escalar, que caracteriza toda la ventana de agrupación. El escalar podría ser el valor máximo de la agrupación ventana (*Max-pooling*), la suma de los valores de la ventana de agrupación (*Sum-pooling*), u otro valor capaz para sintetizar el contenido de la ventana de agrupación [72].

Las capas anteriormente descritas se repiten hasta construir el modelo deseado. Antes de pasar a la última parte de las CNN es necesario hacer *flattening*, usado para convertir la matriz restante en un vector lineal que será el conjunto de entrada que le daremos a la red neuronal completamente conectada que nos dará la predicción. La Fig. 3.5 ilustra el proceso completo de una CNN.

3.2.3. Redes neuronales recurrentes

Un tercer tipo de DNN son las redes neuronales recurrentes (*recurrent neural networks* o RNN), diseñadas para aprender patrones secuenciales o variables en el tiempo.

Una red recurrente es una red neuronal con una conexión de retroalimentación. Los ejemplos incluyen BAM, Hopfield, máquina de Boltzmann y redes de retropropagación recurrente [73][74] . Las técnicas de redes neuronales recurrentes se han aplicado a una amplia variedad de problemas. Las redes neuronales simples parcialmente recurrentes se introdujeron a finales de 1980 por varios investigadores, incluidos Rumelhart, Hinton y Williams para aprender cadenas de caracteres [75].

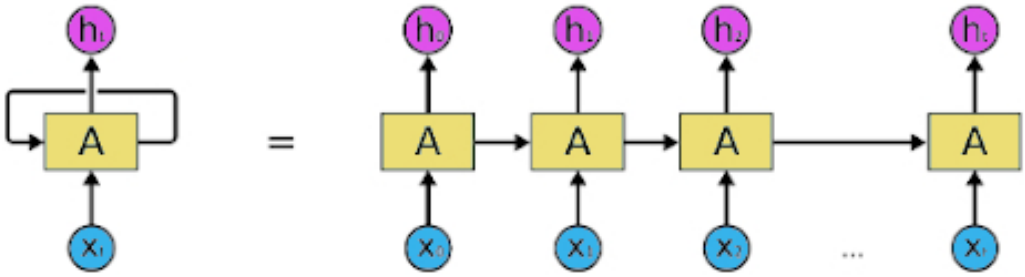


Figura 3.6: Imagen ilustrativa de una red neuronal recurrente [9].

Como se puede observar en la figuras 3.6 y 3.7, las RNN están relacionadas con secuencias y listas por lo que son muy usadas en modelado y generación de texto, reconocimiento de voz, generación de descripciones de imágenes, etc. No obstante las redes recurrentes tienen un problema: mientras se avanza en la red neuronal empieza a existir una brecha de información por lo que la RNNs no puede aprender el contexto de la información. Para solucionar este problema se implementó *Long Short Term Memory Networks* (LSTM).

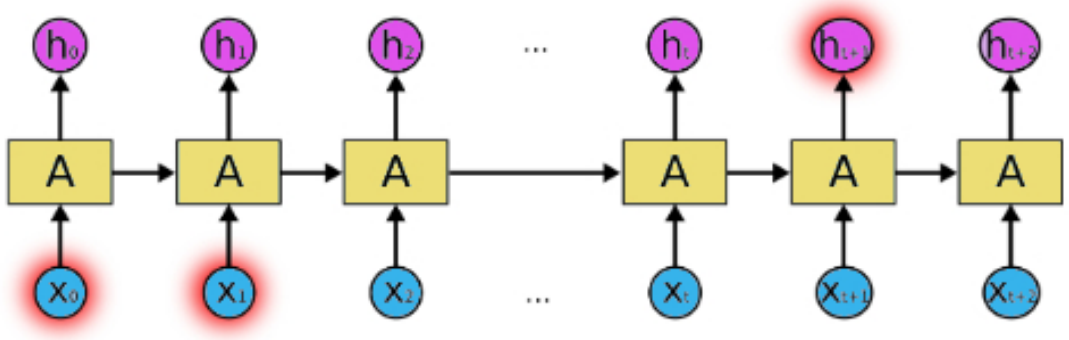


Figura 3.7: Secuencia RNN [9].

LSTM es la arquitectura RNN más común que recuerda valores arbitrarios en intervalos. Fue introducida por primera vez en 1997 por Hochreiter y Schmidhuber y funciona bien en hacer predicciones basadas en datos de series de tiempo, evitando el problema de dependencia a largo plazo que aquejaba a las RNN tradicionales [76].

LSTM también es muy adecuado para tareas de clasificación y procesamiento y se utiliza en Google Translate, Apple Siri y Amazon Alexa [72].

El componente crítico del LSTM es la celda de memoria y las puertas (incluida la puerta de olvido, pero también la puerta de entrada). El contenido de la celda de memoria es modulado por las puertas de entrada y las puertas de olvido. Suponiendo que ambas puertas estén cerradas, el contenido de la celda de memoria permanecerá sin modificar entre un paso de tiempo y el siguiente. La estructura de puerta permite que la información se retenga a lo largo de muchos pasos de tiempo y, en consecuencia, también permite que los gradientes fluyan a lo largo de muchos pasos de tiempo. Esto permite que el modelo LSTM supere el problema del gradiente de desaparición que ocurre con la mayoría de los modelos de redes neuronales recurrentes [77].

Los componentes más importantes de las redes LSTM son:

- Input gate: Decide qué nueva información vamos a almacenar en la celda.
- Forget gate: Decide qué información vamos a descartar en el estado de la celda.
- Output gate: La entrada y el estado anterior se activan como antes para generar la salida que se le atribuye al siguiente bloque LSTM.

La arquitectura de una celda LSTM se puede observar en la Fig. 3.8.

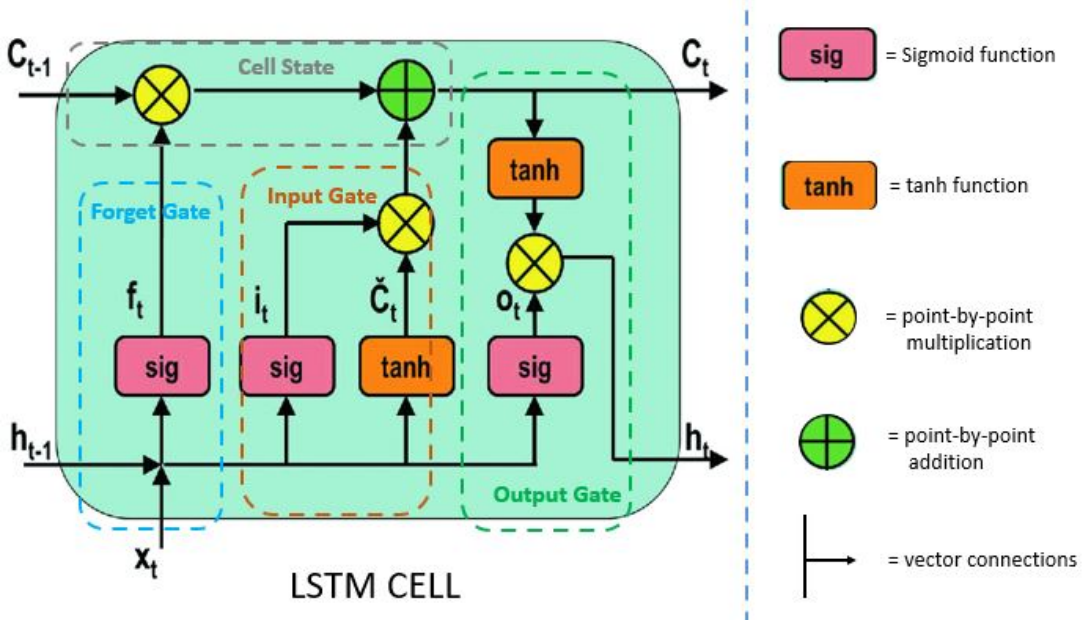


Figura 3.8: Imagen ilustrativa del trabajo interno de una celda LSTM. Imagen tomada de [10].

En el siguiente capítulo se describe cómo se usó una metodología, basada en los distintos tipos de redes neuronales artificiales descritos, para estructurar datos provenientes de notas médicas.

Capítulo 4

Extracción de información en notas médicas

Al analizar los datos del Sistema de Información de Medicina Familiar (SIMF) del IMSS, se identificó que las notas médicas contenían información de gran importancia para el diagnóstico, tratamiento y seguimiento de pacientes con diabetes mellitus tipo 2.

Un ejemplo de nota médica obtenida del expediente clínico electrónico de un paciente, puede verse en el Cuadro 4.1.

```
MASCULINO DE 65 AÑOS DE EDAD, EN SEGUIMIENTO POR DM 2/HAS/ ACUDE POR SUS
   MEDICAMENTOS DE CONTROL...SU TX ACTUAL LO REFIERE NIFEDIPINO 1X1, MET 0.5 X
   2, EL PACIENTE .COENTA QUE ESTA TOMANDO TELMISARTAN O LOSARTAN, PERO NO SABE
   BIEN, .SUS LAB DE CONTROL BH HB 15.6 HCT 43. 5, PLAQUETAS 207, LEUCO S7,
   NEUTROFILOS 50.1 VLDL 44. 4 COLESTEROL 152, HDL 37, LDL 70.6, TG 220, EGO
   NORMAL, .SE REFIERE ESTABLE, NO DATOS DE VASOESPASMO, NO DOLOR PRECORDIAL ,
   NO SX URINARIOS, ...IMC 30.56, SIN COMPROMISO CARDIOESPIRATORIO, RS CS
   RITMICO, ABDOMEN PERISTALSIS PRESENTE, EXTREMIDADES NO EDEMA, SIN LESIONES,
   BUEN LLENADO CAPILAR....ORIENTO DATOS D E ALARMA, REDUCCION DE PESO, DIETA
   DASH,REDUCCION DE SAL, CUIDADO DE LOS PIES, UÑAS, NO FUMAR NO CONSUMIR
   ALCOHOL, CITA ABIERTA A URGENCIAS. LAB DE GLUCOSA DE CONTROL YA QUE NO
   REPORTARON CREATININA Y GLUCOSA EN LOS LABORATORIOS, RECETAS POR TELMISARTAN
```

```
1X2, .INFORMO RIESGOS DE COMPLICCONES, POR NO TOMAR SUS MEDICAMENTOS.
```

Cuadro 4.1: Ejemplo de nota medica obtenida del Expediente Clínico Electrónico (ECE) de un paciente.

Sin embargo, analizar manualmente todas las notas médicas para extraer parámetros de interés es una tarea difícil, laboriosa para los profesionales de la salud y prácticamente inviable dado el volumen de notas presentes en nuestros datos. Para enfrentar esta tarea, se implementó un sistema automatizado de Extracción de Parámetros de notas médicas.

Este tipo de sistemas, en el contexto médico, han cobrado importancia en años recientes y representan una ejemplo del potencial del procesamiento del lenguaje natural en el ámbito de la salud [78, 79].

La componente fundamental en nuestro proyecto fue la tecnología de reconocimiento de entidades nombradas (NER, por sus siglas en inglés), que permite identificar y etiquetar automáticamente información relevante en las notas médicas, como nombres de medicamentos, dosis, frecuencia, entre otros.

En este capítulo describimos de manera general la estructura de NER y, posteriormente, explicamos cómo se utilizó en la base de datos del IMSS para extraer indicadores de salud importantes para el seguimiento de pacientes diabéticos.

4.1. Reconocimiento de entidades nombradas

Los humanos podemos identificar en un texto fácilmente distintas categorías semánticas, por ejemplo, cuándo se hace referencia a una persona, una institución, una fecha, una localización. Debido a esto podemos extraer información importante de un texto. Intentar recrear esto en la computación es el origen de la tarea de Reconocimiento de Entidades Nombradas o NER, por sus siglas en inglés.

NER es una técnica de procesamiento de lenguaje natural que identifica y categoriza información importante en texto. Una entidad nombrada (*named entity*) es un objeto del mundo real que pertenece a una categoría semántica, tal como países, personas, compañías, etc.

La Fig. 4.1 muestra un ejemplo de nota médica en la que se han resaltado distintos campos. La implementación del algoritmo de NER tiene como objetivo que identificar de manera automática el valor y la categoría de esta información de interés.

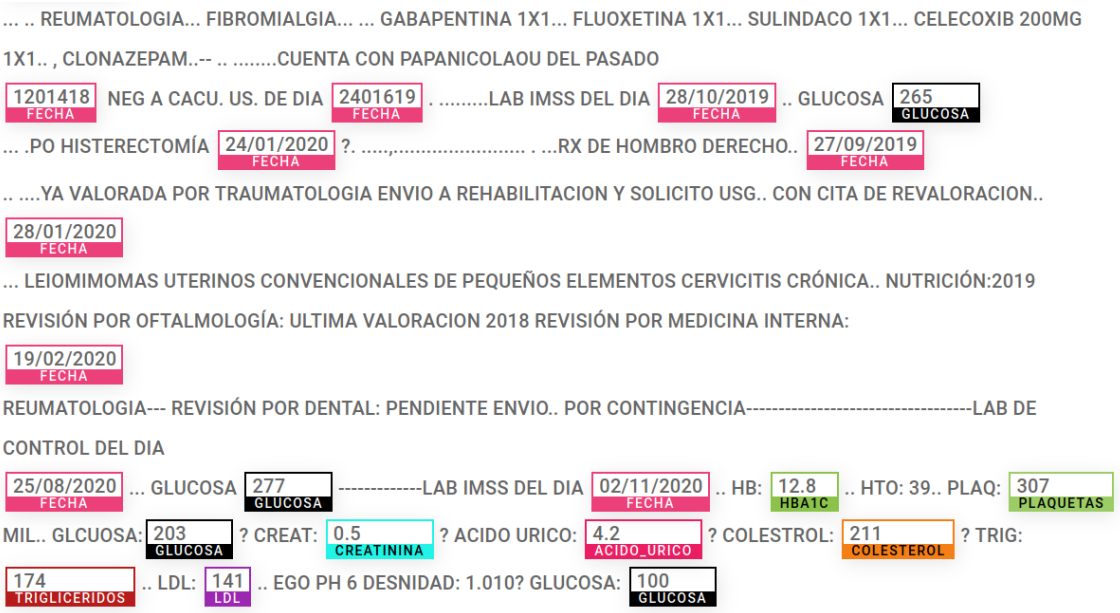


Figura 4.1: Ejemplo de una nota médica en la que se han etiquetado categorías y valores de interés. El reconocimiento de entidades nombradas (NER) busca la automatización de este proceso.

4.1.1. Fases de NER

El proceso general de NER inicia con una pila vacío, todas las palabras en el buffer y sin entidades. El siguiente paso define las acciones para cambiar de estado y, finalmente, predice la secuencia de acciones a tomar.

A continuación se describe con más detalle las cuatro fases que componen el proceso de NER: *Embed*, *Encode*, *Attend*, *Predict*. Puesto que en nuestro trabajo se utilizó la biblioteca SpaCy, de Python, nuestra descripción se enfoca en el funcionamiento del algoritmo de NER implementado como parte de esta biblioteca.

Embed

Las representaciones vectoriales de palabras (*word embeddings*) son un tipo de representación de palabras que permite que aquellas con un significado similar tengan una

representación similar.

Los embeddings son una técnica de aprendizaje automático que convierte datos no estructurados en vectores numéricos, lo que permite a los algoritmos de aprendizaje automático procesarlos de manera más efectiva.

Algunos de los algoritmos más conocidos para generar representaciones vectoriales de palabras son Embedding Layer, Word2Vec, GloVe, entre otros [80, 81].

SpaCy usa un método que consiste en extraer 4 atributos de cada token. Un token es una palabra o un símbolo individual en un texto que se ha separado del resto del texto y se ha asignado un significado. SpaCy extrae: Norm (la forma normalizada de la cadena de texto), Prefix, Suffix y Shape (tamaño de la palabra) [82].

Una vez extraídos estos atributos, se crea una tabla usando la técnica de "*Bloom embeddings*", que es usada para almacenar representaciones distintas en una tabla compacta mediante el hash de cada entrada en varias filas de la tabla. Al representar cada entrada como la suma de varias filas, donde es poco probable que dos entradas choquen en varios hashes, la mayoría de las entradas acabarán con una representación distinta [83]. Por lo general, estas identificaciones son secuenciales, por lo que si un texto tiene un vocabulario de 100 palabras, sus palabras se asignan a un rango de 100 números [84]. Posteriormente, se concatenan los atributos para mezclarse en un perceptrón multicapa que consiste en una capa oculta y una unidad Maxout. Al final del proceso se obtiene un vector de 128 componentes por palabra.

Encode

Dada una secuencia de vectores de palabras, en el paso de Encode se calcula una representación llamada matriz de oraciones (*sentence matrix*), donde cada fila representa el significado de cada token en el contexto del resto de la oración [82].

Posteriormente, esta representación se pasa a una red neuronal convolucional para extraer una ventana de palabras (*windows of words*) en ambos lados de la palabra. El primer paso es pasar la matriz de oraciones por una red CNN de trigram, que obtiene una ventana en cada lado de la palabra; por consiguiente, pasamos de tener un vector de 128 dimensiones a uno de 384. Despúes, el vector se pasa a un perceptrón multicapa

para mapear el vector de nuevo a 128 dimensiones [84]. Este procedimiento se hace cuatro veces; en la primera iteración obtendremos un vector que es sensible a una palabra en ambos lados, en la siguiente iteración se está obteniendo información de dos palabras de cada lado, una vez que llevamos a cabo este proceso de apilado, obtenemos una ventana de cuatro palabras por cada lado.

La salida que se obtiene al final del proceso de Encode es la suma de las convoluciones y la entrada (input). Así, pasamos de una entrada de N vectores de 128 componentes y obtenemos una salida que es una matriz $N \times 128$. En otras palabras, codificamos vectores independientes del contexto en una matriz de oraciones sensible al contexto.

Attend

El paso de Attend reduce la representación matricial producida por el paso de Encode a un solo vector, que se puede pasar de entrada a una red neuronal estándar prealimentada (*standard feed-forward network*) para la predicción. La ventaja principal del mecanismo utilizado respecto a otras operaciones de reducción es que toma como entrada un vector de contexto auxiliar. Al reducir la matriz a un vector necesariamente se está perdiendo información; el vector de contexto es crucial porque dice qué información descartar y se adapta a la red que lo utiliza de entrada [82]. Las características que se toman en cuenta en este proceso pueden ser las entidades previas, que podrían estar muy atrás en el documento pero todavía estar condicionando el resultado.

Predict

El paso de predicción implica alimentar la representación ponderada del texto de entrada que se calculó durante el paso anterior (Attend) en una red neuronal que ha sido entrenada para predecir las etiquetas de cada token. La red generalmente usa una combinación de capas convolucionales y recurrentes para procesar la representación de entrada y hacer predicciones sobre las etiquetas.

Durante el entrenamiento, la red es alimentada con muchos ejemplos de texto etiquetados, donde cada token se anota con una etiqueta que indica si es parte de una entidad nombrada o no. La red está entrenada para minimizar la diferencia entre sus etiquetas

predichas y las etiquetas verdaderas de los ejemplos de entrenamiento.

En el momento de la inferencia, cuando se usa el algoritmo NER para hacer predicciones sobre texto nuevo, el paso de predicción implica pasar el texto de entrada a través de la red neuronal para obtener las etiquetas predichas para cada token. Las etiquetas pronosticadas luego se pueden usar para identificar entidades nombradas y sus límites en el texto.

En general, el paso de predicción es un componente importante del algoritmo SpaCy-NER, ya que permite que el algoritmo haga predicciones precisas sobre qué tokens en un texto dado son parte de entidades nombradas y cuáles no.

4.2. Uso de NER en notas médicas del IMSS

La metodología descrita en la sección anterior se utilizó para extraer parámetros importantes de las notas médicas registradas en el Sistema de Información de Medicina Familiar del IMSS para el caso de pacientes diagnosticados con diabetes.

El conjunto de datos abarcó 4,022,758 de notas médicas, con una extensión variable, redactadas en formato libre.

4.2.1. Bibliotecas utilizadas

La implementación de nuestro trabajo se hizo en el lenguaje de programación Python y se utilizaron principalmente las siguientes bibliotecas:

- SpaCy. Esta librería está orientada al procesamiento avanzado del lenguaje natural en Python y Cython. SpaCy se basa en las últimas investigaciones y está diseñada con el propósito de utilizarse en escenarios reales [85].

- PySpark. Debido a la cantidad de datos que se manejaron durante el proceso, se optó por utilizar PySpark, una interfaz para Apache Spark en Python. Esta biblioteca permite escribir aplicaciones Spark utilizando las API de Python, y proporciona el shell de PySpark para analizar datos de forma interactiva en un entorno distribuido. PySpark es compatible con la mayoría de las funciones de Spark, como Spark SQL, DataFrame, Streaming, MLlib (Machine Learning) y Spark Core [86].

- Pandas. Las pruebas iniciales de la investigación, utilizando sólo un fragmento de la base de datos, se llevaron a cabo utilizando Pandas. Esta librería, diseñada para manipulación y análisis de datos, permitió llevar a cabo una exploración de los datos y diseñar el proceso que, posteriormente, se implementó a nivel de la base de datos completa usando PySpark.

4.2.2. Etiquetado de párametros

El uso de NER para el reconocimiento de parámetros de nuestro interés requiere entrenar al algoritmo con notas médicas etiquetadas manualmente.

Debe considerarse que etiquetar los datos no es una tarea trivial. Identificar correctamente los parámetros de nuestro interés requiere entender el contexto e interpretar adecuadamente expresiones que algunas veces aparecen abreviadas, con lenguaje muy especializado o con variaciones de escritura.

La Tabla 4.1 muestra, en la columna izquierda, los parámetros que se plantearon como objetivo de la tarea de extracción de información. Esta lista se estableció con la guía del personal médico que participó en el proyecto, considerando aquellos indicadores que podrían estar correlacionados con la evolución de la diabetes y con la aparición de las diversas complicaciones.

La columna derecha de la misma tabla presenta una lista de variaciones comunes identificadas en las notas para hacer referencia los parámetros de interés.

Tabla 4.1: Parámetros considerados en la tarea de Reconocimiento de Entidades Nombradas. La columna derecha muestra variaciones de escritura detectadas para hacer referencia al parámetro. En la columna, las variaciones compuestas se representan con un guión; salvo el caso de las fechas, en las notas médicas no aparece esa marca.

Parámetro Médico	Variedad de escritura
FPG (Fasting Plasma Glucose)	fpg, FPG
HBA1C	HBA1C, hb, HB, hba1c, HbA1c, BH-CO-HB
HDL	HDL, hdl, HDL-COL
LDL	LDL, ldl, LDL-COLs
Colesterol	COLET, colesterol, COLESTEROL, COL, COLESTEROO, COL, COELST
Trigliceridos	TGC, Trigliceridos, TRIGLICERIDOS, TRIG, TG
Glucosa	GLU, GLUCOSA, glucosa, GLCU, GLUCOS, GLUCOSA-PP, GLUCEMIA
Presion Arterial	TA, T:A, PRESION, PRESION-ARTERIAL
Fecha	14/02/2020, 13-11-20, 15 DE OCT 2020, 300720, SEPTIEMBRE, NOV 18, OCT

Como se puede observar, todos los parámetros cuentan con variantes que los identifican en las notas médicas. La glucosa es uno de los indicadores que puede describirse con mayor diversidad de expresiones.

También la fecha puede estar registrada bajo formatos muy diversos; desde los más comunes, que comunican día, mes y año con números, hasta aquellos muy sencillo que sólo utilizan la abreviatura del mes.

Además, debe tomarse en cuenta que las notas médicas se redactan bajo fuertes restricciones de tiempo. Es muy común encontrar errores de dedo, algunas faltas ortográficas o, anotaciones adicionales. Todo esto agrega complejidad al problema.

Los siguientes son ejemplos tomados de notas reales:

- Fecha: 16/04/2020/2019
- Fecha: 30/10/20200
- Fecha: 100 DE NOV. 2020

- GLUCOSA 26/10/2020.. AYUNAS 114MG XDL, DESPUÉS DE COMER 235, 06NOV 2020 AYINAS 86MG XDL, DESPUÉS DE COMER 163

Con todas las consideraciones mencionadas, participantes del proyecto con formación profesional en medicina etiquetaron un total de 2000 notas para la fase de entrenamiento del algoritmo.

Para llevar a cabo el etiquetado de las notas, se utilizó el software Doccano.

Uso de Doccano

Para el entrenamiento del algoritmo se utilizan notas médicas etiquetadas, es decir, notas en las que se ha identificado previamente la columna de inicio y de fin de las entidades de interés.

En un primer intento, se usó un procesador de texto, que mostraba el inicio y final de cada carácter (número de columna), identificar ahí las entidades y registrar las columnas en una hoja de cálculo. Finalmente, este archivo se procesaba utilizando Python para adaptarlo al formato de SpaCy. Este proceso fue descartado, pues demoraba aproximadamente 15 minutos por nota médica.

La solución a este problema se encontró usando el software Doccano, una herramienta de código abierto para la anotación de textos por humanos. Este software ofrece diversas funciones de anotación que resultan útiles para diferentes tareas en el procesamiento de texto. Entre ellas, se incluye la capacidad de llevar a cabo la clasificación de texto, permitiendo etiquetar y categorizar distintos tipos de documentos o fragmentos de texto. Además, cuenta con herramientas para realizar el etiquetado de secuencias, lo que facilita la identificación y marcado de elementos específicos dentro de un texto, como entidades nombradas o partes de discurso. Asimismo, el software es capaz de manejar tareas más complejas de secuencia a secuencia, lo que implica procesar y generar secuencias de texto en función de ciertas reglas o patrones predefinidos. Por lo tanto, puede crear datos etiquetados para análisis de opiniones, reconocimiento de entidades nombradas, resúmenes de texto, etc. [87].

El etiquetado se hace mediante una interfaz del programa que facilita la tarea. Primero, se exportan los datos en formatos separados, en este caso una nota médica por

archivo de texto. Posteriormente, se especifican las entidades que se buscarán. Finalmente, las personas deben etiquetar los datos subrayando en cada nota el texto de interés y la categoría-entidad a la que corresponde. La Fig. 4.2 muestra un ejemplo de nota etiquetada usando Doccano.

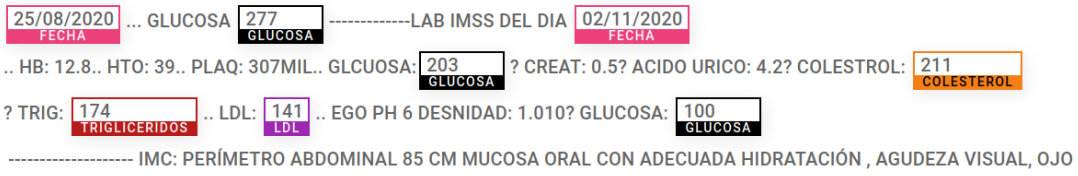


Figura 4.2: Ejemplo de la interfaz del software Doccano, utilizado para etiquetar notas médicas en la fase de entrenamiento del algoritmo de Reconocimiento de Entidades Nombradas (NER)

Al finalizar la tarea, los datos etiquetados se exportan en formato txt. Estos archivos pueden utilizarse para alimentar el algoritmo de NER de SpaCy.

4.2.3. Implementación de la metodología propuesta

Como se mencionó anteriormente, para la implementación de NER en las notas médicas se utilizó la biblioteca SpaCy. Esta biblioteca, escrita en Python y Cython, es de código abierto y se utiliza para el procesamiento avanzado del lenguaje natural. Los principales desarrolladores de SpaCy son Matthew Honnibal e Ines Montani y su código se encuentra publicado bajo la licencia MIT [88].

El formato de los datos que recibe *Spacy* consiste en una tupla con dos componentes por documento, en este caso, por nota médica. La primera componente consiste en el texto completo; la segunda contiene información sobre las entidades en el texto y consiste, a su vez, en una lista con una tupla de tres componentes por entidad: el número de carácter donde inicia el parámetro, donde termina y la identificación de la entidad. Un ejemplo de este formato se muestra en el Cuadro 4.2.

```
("MASCULINO DE 72 AÑOS, NIEGA ALERGIAS, TOXICOMANIAS NEGADAS, QX,, NEGADOS, TX,
   NEGADOS, HEMOTRANSFUSIONAL NEGADOS, PENSIONADO, ACUDE A CONTROL DE HAS DE 22
   AÑOS DE EVOLUCION Y OSTEOARTROSIS, CON DX HPB, TAMSULOSINA 1X1 , AL MOMENTO
   ESTABLE. DX DE ENF DE PARKINSON. TX. PRAMIPEXOL .5 MG- TAB Y MEDIA C12 HRS.
```

CLONAZEPAM MEDIA TBA AL DIA. ALGUNODATOS DEL PARKINSON. PRESENTA FUERZA MUSCULAR DISMINUIDO EN MPI PARA DORSIFLEXION, ... RX,. DE COLUMNA LUMBROSACRO CON RETROLISTESIS POSTERIOR L5S1,. CON GRANDES CAMBIOS DEGENRATIVOS,.. -.TX. DOPAADRENERGICO PRAMIPEXOL 1 MG C8 HRS,. GABAPENTINA 1 X2,. ASA MEDIA TAB AL DIA,. CON ELECTROMIOGRAFIA DE MSPS. IRM DE COLUMNA. CON DX ESTENOSIS OSEA DEL CANAL NEURAL ANIVEL DE L3 POR LO QUE FUE SIMETIDO A PROCESO QUIRURGICO.POSTOPERATORIO DE DESCOMPRESION NEURAL Y ESTABILIZACION YA ESTUVO EN REHABILITACION SE LE ENVIA A NATACION. VALORACION NEUROLOGICA EL APSADO 120319. SE LE DA DX DE PARKINSON RIGIDO AKINETICO.DIABETES MELLITUS DE RECIENTE DX. -- -------- LAB DEL DIA 050219 GLCU DE 116,. BH NL. CR.1 AU. 6, COLET 83,. TGC 58. EGO NL. TFG POR C-K 61.22 MIL-MIN.. S ELE AGREGA GABAPENTINA 1X1 Y SE AUMENTA DOPAMINERGICO. 1 1/2. C8 HRS, DXTX DE HACE 8 DIAS 125RECETA MANUAL POR GABAPENTINA 1X2.---------- IMC 23.21 MARCHA LENTA, PASOS CORTOS, BRADICINESIA, HIPOMIMIA, FACIAL, E HIPOFONIA LEVES, HIPOCINESIA X, RIGIDEX Y RUEDA DENTADA XX, , BRADICINSIA, X, FUERZA MUSCULAR EN EXTREMIDADES NORMAL. SIN DETERIORO COGNITIVO APARENTE,ORIENTADO EN LAS 3 ESFERAS, FUNCIONES,MENTALES SUPERIORES CONSERVADAS, BUENA COLORACION E HIDRATACION DE MUCOSAS Y TEGUMENTOS, OJOS SIMETRICOS,PUPILAS ISOCORICAS Y NORMORREFLECTICAS, REFLEJO CONSENSUAL CONSERVADO, SIN ALTERACIONES VISUALES, CAVIDAD BUCAL CON PLACA DENTOBACTERIANA, PULSOS CAROTIDEOS HOMOCROTOS Y SINCRONICOS CON EL RADIAL, SIN INGURGITACION YUGULAR, TRAQUEA CENTRAL MÓVIL, NO CRECIMIENTO DE LA TIROIDES, CS PS CON BUENA ENTRADA Y SALIDA DE AIRE SIN FENÓMENOS ESTETOACÚSTICOS AGREGADOS, RUIDOS CARDIACOS RITMICOS DE BUENA INTENCIDAD, NO GALOPE, SIN ARRITMIAS, NI CHOQUE DE PUNTA, , ABDOMEN BLANDO, PERIMETRO DE 100 CM, SIN REFLUJO HEPATOYUGULAR, NO HAY DOLOR A LA PALPACION BIMANUAL EN FOSAS RENALES, NI EN TRAYECTOS URETERALES, PULSOS FEMORALES PRESENTES DE BUENA INTENSIDAD Y AMPLITUD, RODILLAS PULSOS POPITLEOS PRESENTES , ROT CONSERVADOS, SIN EDEMA, SENSIBILIDAD SUPERFICIAL Y PROFUNDA DE CARCTERISTICAS NORMALES, MARCHA NORMAL, LLENADO CAPILAR DISTAL INMEDIATO, PULSOS TIBIALES Y DORSALES PRESENTES, PIEL NORMAL, SIN CAMBIOS DERMICOS, UÑAS DE LOS PIES SIN ALTERACIONES, CON LESIONES EN ARE AINTERDIGITAL. CON TRAYECTOS VENOSOS DILATADOS, CON DEFORMACION DE FALNAGES DISTALES DE DEDOS DE

```
 MANO-------------------------- DIETA, EJERCICIO DIARIO A TOLERANCIA,
CUIDADO DE LOS PIES, DIETA HIPOCALORICA, VARIADA CON HORARIO, DESAYUNO,
ALMUERZO, COMIDA, CENA Y COLCACION, EVITAR REFRESCOS", {'entities':
[(1013,1019,'FECHA'), ( 108,1031,'GLUCOSA' ),(1059,1061,'COLESTEROL'),
(1068,1070,'TRIGLICERIDOS') ] } )
```

Cuadro 4.2: Ejemplo de formato de datos de entrada usado por SpaCy. Este ejemplo corresponde a una instancia (nota médica) y contiene dos componentes principales: la nota y la información sobre entidades identificadas dentro de la nota.

El primer experimento para el uso de NER en nuestro conjunto de datos se realizó con 50 notas en el conjunto de entrenamiento y 20 iteraciones en el algoritmo. En esta prueba se evaluó la viabilidad de la implementación y el manejo de la misma con datos faltantes. La Tabla 4.3 muestra los resultados de este experimento. Como se puede observar, es posible reconocer la mayoría de los parámetros pero, en algunos casos, el texto identificado por el algoritmo es un segmento extenso de la nota (el caso de la glucemia en la Tabla).

Parámetro	Valor identificado por NER
GLUCEMIA	128., SU AGUDEZA VISUAL CONSERVADA , CARDIOPULMONAR SIN COMPROMISO APARENTE, ESTA ORINANDO DE ASPECTO NORMAL HAY EDEMA DE PIE DERECHO. SENSIBLIDAD DE PIES CONSERVADA.SUS RESULTADOS DE LAB DEL DIA 17
FECHA	2020
GLUCOSA	323
COLESTEROL	264
HDL	38
LDL	2143
TRIGLICERIDOS	414
FECHA	TA, 1000 DE GLUCOSA,

Tabla 4.3: Ejemplo de los resultados obtenidos en el primer experimento llevado a cabo con NER para identificar el valor de parámetros en una nota médica.

Posteriormente, se utilizaron 2000 notas etiquetadas por el personal médico. De este conjunto, se utilizaron 1581 notas para entrenar el modelo y 396 para evaluarlo. Un conjunto de 23 notas médicas se descartaron por no contener información relevante; en estos

casos, la persona que etiquetó la nota no encontró parámetros clínicos.

El modelo fue entrenado usando 20 iteraciones en Spacy.

4.2.4. Resultados y discusión

Utilizando la implementación descrita se logró rescatar 7,117,287 parámetros de notas médicas. El parámetro con mayor cantidad de apariciones identificadas fue el de *fechas* siendo las fechas (1,817,913 valores), seguido por glucosa (1,042,414 valores).

Para evaluar la calidad de los resultados se usaron diferentes métricas conocidas en aprendizaje automático supervisado. Todas estas métricas se basan en la proporción de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) identificados por el algoritmo. Las métricas se describen a continuación.

Exactitud (*accuracy*). En clasificación, es la proporción de predicciones correctas (verdadero positivo más verdadero negativo) dividida por el número total de predicciones hechas.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precisión. Mide la calidad de las predicciones positivas hechas por el modelo. Se refiere al número de verdaderos positivos divididos por el total de predicciones positivas. En otras palabras mide qué tan precisas son las predicciones positivas.

$$precision = \frac{TP}{(TP + FP)}$$

Sensibilidad (*recall*). Es una métrica que cuantifica el número de predicciones positivas correctas hechas de entre todas las predicciones positivas que podrían haberse hecho. A diferencia de la precisión, que solo comenta las predicciones positivas correctas de todas las predicciones positivas, la sensibilidad proporciona una indicación de las predicciones positivas perdidas [89].

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

F1-score. Es una medida que corresponde a la media armónica entre la Precisión y la Sensibilidad [90].

$$F_1\ score = 2\,\frac{Precision * Recall}{Precision + Recall}$$

En el caso de nuestros experimentos, decidir si una predicción fue acertada pasa por dos filtros. Primero, verificar si la categoría del parámetro (*label*) es la misma que en la predicción. Después, verificar el dato. En esta última evaluación consideramos una tolerancia de dos caracteres respecto al dato, como se ejemplifica en la Tabla 4.4.

Etiqueta Real	Valor Real	Etiqueta Predicción	Valor Predicción
Glucosa	135	Glucosa	135mg

Tabla 4.4: Ejemplo de una predicción considerada correcta, a pesar de diferir en dos caracteres respecto a la expresión original.

La tolerancia se estableció pensando en los valores que el algoritmo detecta junto con su unidad de medida.

Cuando ambos criterios se satisfacen, la categoría y el valor hasta dos caracteres de diferencia, consideramos que el resultado fue correcto (verdadero positivo).

Con estas consideraciones, se obtuvieron los resultados que se muestran en la Tabla 4.5. Conviene recordar que para esta evaluación se tomaron en cuenta 396 notas médicas (conjunto de prueba), lo que abarcó más de 6000 parámetros.

Parámetro	Exactitud	F1	Precisión	Sensibilidad
HIPERTENSION	0.368	0.538	0.606	0.484
FECHA	0.835	0.910	0.912	0.907
GLUCOSA	0.789	0.882	0.858	0.907
TRIGLICERIDOS	0.878	0.935	0.921	**0.950**
UREA	0.854	0.921	0.942	0.900
CREATININA	0.859	0.924	0.932	0.915
HDL	0.866	0.928	0.928	0.928
LDL	**0.933**	**0.965**	0.982	0.949
TASA DE FILTRACIÓN GLOMERULAR	0.47	0.643	0.827	0.526
AÑO DE DIAGNÓSTICO DIABETES	0.708	0.829	**1**	0.708
AÑO DE DIAGNÓSTICO HIPERTENSIÓN	0.711	0.831	1	0.711
INDICE DE MASA CORPORAL	0.7514	0.858	0.885	0.832
ACIDO URICO	0.907	0.951	0.960	0.943
COLESTEROL	0.832	0.908	0.918	0.899
PLAQUETAS	0.815	0.898	0.922	0.875
PESO	0.600	0.750	1	0.600
ALTURA	0.625	0.769	0.909	0.666
PRESION ARTERIAL	**0.257**	**0.409**	**0.529**	**0.333**
HBA1C	0.666	0.8	0.814	0.785

Tabla 4.5: Resultados obtenidos con NER en el conjunto de datos.

El promedio general de exactitud fue 0.722 , mientras que el promedio en la métrica F_1 fue de 0.823.

Considerando ambos indicadores (exactitud y F_1), el indicador mejor reconocido por el algoritmo fue LDL, con puntuaciones de 0.933 y 0.965, respectivamente. Esto podría deberse a que es menos propenso a tener errores ortográficos y no cuenta con abreviatura.

En contraste, el parámetro que tuvo más problemas para ser identificado fue presión arterial. Como se puede observar, por la puntuación de Precisión y Sensibilidad (.529 y .33, respectivamente), fue más común que no detectara valores que se encontraban en la nota a que reconociera incorrectamente valores. Esto podría ser consecuencia de la variedad de formas en que puede referenciarse este indicador.

Analizando de cerca los datos se pudo observar que algunos parámetros no fueron correctamente recuperados debido a que el algoritmo identificó caracteres vecinos como parte de la entidad. Por ejemplo, en el caso de fechas hubo respuestas como «02 AGOSTO 2018,,,,,,,MENCIONA», «28 MARZO Y 20 AGOSTO 2019» y «DICIEMBRE DICIEMBRE

2000». Estos problemas, en algunos casos, provienen de errores en la redacción de la propia nota («Diciembre diciembre 2000»).

De acuerdo a los resultados observados, consideramos que la implementación de este algoritmo fue exitosa. El uso de NER permitió recuperar información importante para el desarrollo de modelos de diabetes y de sus complicaciones en pacientes del estado de Michocán. El volumen de información (más de 4 millones de notas médicas) es importante y quizá, sin la automatización del proceso, estos datos no podrían tomarse en cuenta, o bien, hubieran demandado una cantidad importante de tiempo para su recuperación.

Capítulo 5

Creación de datos estructurados

La principal tarea planteada en este proyecto es crear un conjunto de datos que permita analizar la evolución de pacientes con diabetes mellitus tipo 2. Este conjunto debe integrar información de tablas relacionales existentes en el Sistema de Información de Medicina Familiar (SIMF) y la información recuperada de notas médicas siguiendo el procedimiento basado en NER, descrito en el capítulo anterior.

El corpus final, con datos estructurados, captura información de expedientes clínicos electrónicos de pacientes diabéticos y podrá usarse en el futuro para crear modelos que permitan estudiar la enfermedad y predecir oportunamente el riesgo de desarrollar complicaciones específicas.

En este capítulo se describirán las diferentes etapas realizadas para la creación del conjunto de datos. Además, se presentan detalles sobre la documentación de los datos y una muestra del conjunto de datos en su versión final.

5.1. Información extraída de tablas del SIMF

Para crear la versión final de nuestro conjunto de datos se combinaron los resultados obtenidos con tablas del SIMF que contienen datos estructurados con información de:

- valores de parámetros clínicos relacionados con la diabetes mellitus,
- fechas de consultas,

- datos de somatometría,
- medicamentos,
- diagnósticos clínicos.

Las tablas que se tomaron en cuenta para extraer esta información se muestra en la Tabla 5.1.

Tabla 5.1: Tablas del Sistema de Información de Medicina Familiar que se utilizaron para crear nuestro conjunto de datos final.

Tabla	Valor	Tipo de información
corhis_somatometria	Peso	Datos de somatometría
corhis_somatometria	Altura	Datos de somatometría
corhis_somatometria	Presión arterial	Datos de somatometría
exphis_hc_diabetes	Glucosa	Valores de parámetros clínicos
exphis_hc_diabetes	Colesterol	Valores de parámetros clínicos
exphis_hc_diabetes	Trigliceridos	Valores de parámetros clínicos
exphis_hc_diabetes	HDL	Valores de parámetros clínicos
exphis_hc_diabetes	LDL	Valores de parámetros clínicos
exphis_hc_diabetes	Hba1c	Valores de parámetros clínicos
expcat_med_farmacia	Nombres de medicamento	Medicamentos
correl_receta_med	Codigo de medicamentos	Medicamentos

Consideramos que, en el futuro, quienes usen el conjunto de datos para implementar modelos podrán fácilmente filtrar la información de acuerdo a sus objetivos.

5.1.1. Preprocesamiento de los datos

Para combinar adecuadamente las tablas de interés se llevo a cabo el preprocesamiento que se describe a continuación. Todas estas tareas se llevaron a cabo en una instancia de *Google Cloud Computing* utilizando un tipo de máquina e2-standard-8 con 8 vCPU, 32GB memoria y 100 GB ssd de almacenamiento con sistema operativo ubuntu 18.

ID. La identificación (ID) es importante para diferenciar a quién pertenecen los datos. En nuestro caso, importa tanto el ID de paciente así como el ID de la consulta médica. Los datos originales presentan un problema, pues algunos ID de consulta médica se repiten con diferentes pacientes (ver Tabla 5.2). Para evitar confusión al momento de hacer consultas y segmentar los datos, se decidió hacer una combinación de ID-paciente e ID-consulta.

Tabla 5.2: Ejemplo de registros encontrados que vinculan el ID de paciente con el ID de una consulta médica. Como puede verse, los ID de consultas médicas no son únicos.

Id paciente	Id consulta
1	55
2	55
3	44
4	66

Recetas médicas. Para combinar la tabla con información de recetas médicas, se concatenaron las tablas con información de recetas médicas con los códigos correspondientes a los diferentes medicamentos. Con esto se obtuvo el nombre del medicamento de la tabla catálogo medicamentos (expcat_med_farmacia). Además, se hizo la conversión del nombre de los medicamentos a mayúsculas.

Diagnósticos. En las tablas relacionadas con los diagnósticos médicos lo primero que se realizó fue combinar la tabla de catálogo de diagnósticos (NOMBRE-EXACTO DE TABLA) con la tabla de diagnósticos clínicos (NOMBRE-EXACTO-DE TABLA).

Un segundo paso para procesar datos relativos al diagnóstico consistió en combinar los diagnósticos relativos al mismo paciente y consulta. En la estructura original de la tabla cada diagnóstico representa un nuevo registro (fila), como se muestra en la Tabla 5.3. Los datos se procesaron para combinarlos de forma tal que cada fila corresponda a una consulta médica, como se muestra en la Tabla 5.4.

Tabla 5.3: Ejemplo de instancias en la tabla de diagnósticos antes del procesamiento realizado. Diferentes diagnósticos corresponden a diferentes filas, incluso si se relacionan con la misma consulta médica.

Paciente	Consulta	Diagnóstico
155	989	E119
155	989	I10X
155	997	E119
152	9974	E118

Tabla 5.4: Ejemplo de procesamiento relativo a los diagnósticos. Los diagnósticos que recibe un paciente en una misma consulta médica se agrupan para constituir una sola instancia.

Paciente	Consulta	Diagnóstico
155	989	E119,I10X
155	997	E119
152	9974	E118

Glucosa y presión arterial. Al combinar las tablas relacionadas con glucosa y presión arterial notamos que algunas de ellas tenían definidas en sus columnas glucosa preprandial y posprandial, mientras que otras tablas solo tenían una columna de glucosa. Algo similar se observó con la presion arterial sistólica y diastólica. Por esta razón, al combinar las tablas se decidió crear solo una columna con un separador entre los datos, por ejemplo: 140|240.

Datos de nota médica. Los datos que se obtuvieron mediante el algoritmo de NER requirieron algunos filtros importantes. Uno de los problemas más comunes que identificamos fue que en ocasiones el algoritmo detectaba una cadena de texto más grande de lo esperado.

El siguiente es un ejemplo de valor reportado por el algoritmo para la categoría *fecha*: ‘‘`31 07 14 GLUCOSA 130 DESCONTROLADA, PACIENTE QUE SE RECOMIENDA APEGO A SU TRATAMIENTO, ALIMENTACION Y A REALIZAR EJERCICIO DIARIO (CAMINATA) POR LO MENOS MEDIA HORA. PACIENTE CON IMC 31.2 OBESIDAD GDO II, CINTURA 103 CM, FC 70 X MIN, FR 18 X MIN, TEMP 36 GC, CONCIENTE, ORIENTADO, TRANQUILO`’’.

En el ejemplo puede observarse que el algoritmo detectó la fecha (`31/07/14`) pero no se detuvo al finalizar la información correspondiente a la categoría.

Para este tipo de casos establecimos un procedimiento que consistió en tomar la moda de longitud de caracteres de cada columna, considerar una longitud de 5 caracteres adicionales a la moda. Cualquier respuesta con una longitud mayor, se eliminó y remplazó con un indicador de valor faltante.

Para el caso particular de las fechas consideramos una tolerancia de 12 caracteres respecto a la moda, considerando las diferentes formas de representar las fechas («11/11/2015» o «15 DE DICIEMBRE 2020», entre otras muchas).

Indicador de fuente. Finalmente, al conjunto de datos se agregó una columna para indicar de qué tabla proviene cada instancia (fila) de los datos. Consideramos que esto permite filtrar información, tener más orden en el conjunto de datos y brindar una base para evaluar el tipo y la fuente de incertidumbre que podría tener un dato; por ejemplo, un dato tomado directamente de una tabla del SIMF podría tener un error de registro, mientras que un dato extraído de la nota médica podría tener, adicionalmente, un error debido al proceso implementado de NER.

5.2. Integración y limpieza del conjunto final

Después de procesar la información, combinando y creando atributos descritos en la sección anterior, se conformó el conjunto de datos que consideramos el producto principal de este trabajo.

El proceso completo que se siguió se muestra en la Fig. 5.1, mediante un diagrama de flujo.

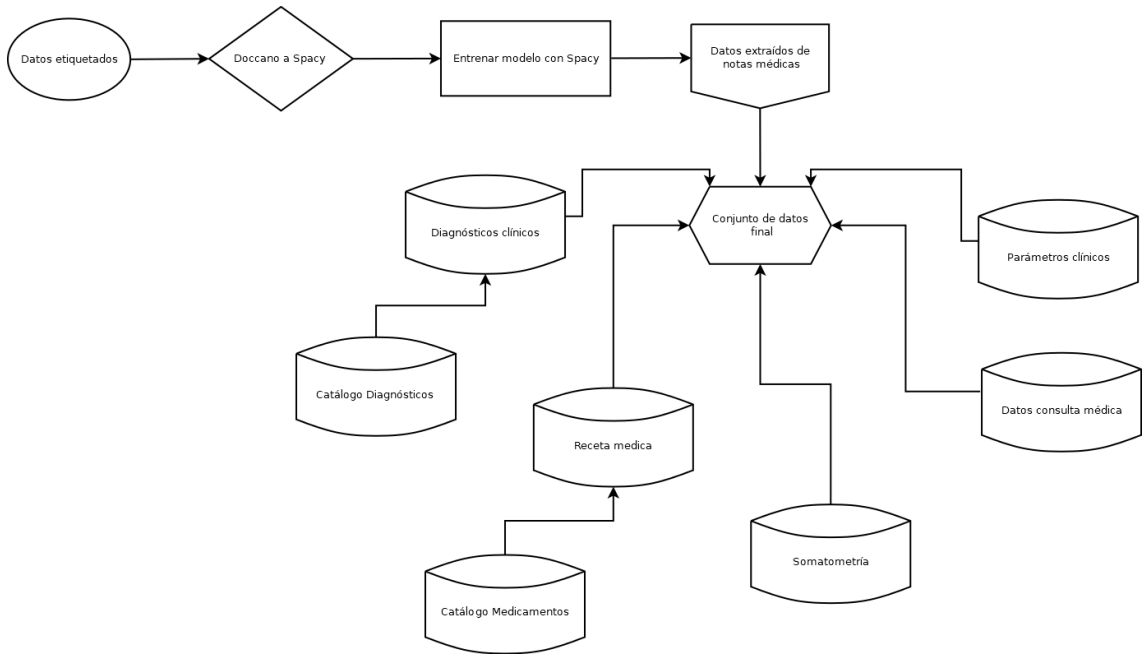


Figura 5.1: Diagrama que representa el proceso planteado para integrar información de diversas fuentes y construir el conjunto de datos planteado en esta investigación.

Finalmente, se implementaron algoritmos orientados a homologar datos provenientes

de distinta fuente. Esta tarea se describe con más detalle a continuación.

5.2.1. Integración de datos

Combinar datos de diferente fuente planteó el reto de integrar formatos distintos para cierta información.

En el caso de los datos faltantes, encontramos que diferentes conjuntos de datos utilizaban representaciones distintas: nan, 0, NaN, null, NULL, entre otros. Particularmente problemático nos parece el caso del 0, pues, aunque por el contexto es fácil deducir en cuáles columnas representa un valor faltante, no es una representación que facilite la tarea de filtrar y analizar información.

Para solucionar este problema se determinó remplazar las diferentes representaciones con el valor NaN[1].

Adicionalmente, se eliminaron los registros (filas) que sólo tenían datos faltantes vinculados a una fecha.

Integración de fechas

La gran variedad de formatos de fechas identificadas por el algoritmo de NER hicieron necesario procesar las fechas para unificarlas.

Decidimos utilizar como estándar el formato (2007-04-01 00:00:00). Para convertir todas las fechas fue necesario (1) agrupar fechas con formato parecido, (2) establecer reglas de conversión apropiadas para cada grupo.

De acuerdo a la estructura identificada, establecimos los siguientes grupos:

- Grupo 1: Fechas en las que el mes está escrito con nombre completo o con una abreviatura fácilmente reconocible. Por ejemplo: 12 DE ENERO DEL 2010, 09 DE AGO DEL 2013, 02 MAYO 20.

- Grupo 2: Fechas que sólo contienen números y espacios. Por ejemplo; 01 12 20, 05 12 2015

[1]En Python, NaN (de *Not a Number* es un valor de punto flotante especial que comúnmente se usa para representar información faltante.

- Grupo 3: Fechas que contienen solo números sin espacio. Por ejemplo: 150620, 150620, 20072007
- Grupo 4: Fechas que contienen solo números y guión. Por ejemplo: 19-07-07, 22-11-14
- Grupo 5: Fechas que contienen solo números y barra oblicua. Por ejemplo: 21/09/09

Para procesar al Grupo 1, (1) se eliminaron las preposiciones de la cadena de texto, (2) se creó un arreglo con las abreviaturas de los meses

$$([' ENE',' FEB',' MAR',' ABR',' MAY',' JUN',' JUL',' AGO',' SEPT',' OCT',' NOV',' DIC']).$$

Esto permitió identificar meses abreviados pero, también, reconocer instancias en las que la escritura del mes tenía algún error de dedo (por ejemplo, octobre), y (3) se usó un diccionario para mapear los meses a su respectivo número.

En el caso de las fechas que solo contenían meses y años (como «SEPTIEMBRE DEL 2011») se decidió completar la fecha con el primer día del mes.

La estructura general del algoritmo que se utilizó se muestra en el Algoritmo 1.

Algoritmo 1 Algoritmo para homologar el formato de las fechas en nuestro conjunto de datos.

```
Crear arreglo de meses
Crear diccionario meses a número
for fecha in fechas do
    if fecha es dato faltante then
        Indicar que es NaN y marcar que no fue procesado
        continue
    end if
    Cambiar fecha a mayúscula
    Eliminar espacios extras al inicio y al final del string
    Eliminar múltiples espacios
    if La fecha contiene solo números y guiones then
        if Fecha solo contiene números sin espacios then
            Separar la fecha en día, mes y año
            Formatear fecha usando datetime
            Marcar si fue procesado o no
        else
            Formatear fecha con guiones usando datetime
            Marcar si fue procesado o no
        end if
    else if La fecha contiene solo números y barra oblicua then
        Formatear fecha con barra oblicua usando datetime
        Marcar si fue procesado o no
    else if La fecha contiene solo números y espacios then
        Formatear fecha con espacios usando datetime
        Marcar si fue procesado o no
    else if La fecha esta escrita con palabras then
        Separar las palabras
        Eliminar preposiciones y contracciones
        for substring en fechas do
            Pasar a string los substring numéricas
            Identificar el mes escrito con letras
            Usar diccionario para convertirlo en su valor numérico
        end for
        if L thena fecha no incluye día
            Agregar el día 01 a la fecha
            Formatear fecha con datetime
        end if
    end if
end for
```

Por supuesto, las reglas que se implementaron toman en cuenta sólo los formatos más comunes, pues no era posible revisar las más de 2,280,664 fechas extraídas de las notas médicas.

Al terminar el proceso de homologación de fechas se logró ajustar 1,032,014 instancias; esto corresponde al 57 % de las instancias, considerando que 490,424 de las instancias totales aparecen como valores faltantes.

Para almacenar las fechas estandarizadas se creó una columna llamada `fechas procesadas`. La decisión de conservar la fecha en formato original se tomó considerando que, tratándose de un estudio de evolución de enfermedades en el tiempo, podría haber modelos que requieran una mayor precisión que la que se alcanza con nuestra propuesta en la tarea de manejar las fechas. En esos casos, ofrecer el formato original permitiría implementar un algoritmo propio para homologación de fechas.

5.3. Versión final del conjunto de datos

Al finalizar el trabajo, se obtuvo un conjunto de datos estructurados con un tamaño de 10.62 GB. Esta tabla incluye 9,577,839 filas referentes a 278,324 pacientes.

Las columnas o atributos que se consideran en la versión final, así como el tipo de datos al que corresponden, se muestran en la Tabla 5.5.

Tabla 5.5: Atributos y tipo de datos que se consideran en la versión final del conjunto de datos.

Atributo	Tipo
Cx_curp	Cadena de texto
Nota_medica	Cadena de texto
Glucosa	Cadena de texto
Colesterol	Numérico
Trigliceridos	Numérico
HDL	Numérico
LDL	Numérico
Fecha	Cadena de texto
Presion_arterial	Cadena de texto
Hipertension	Cadena de texto
Plaquetas	Numérico
Creatinina	Numérico
Acido_urico	Numérico
Urea	Numérico
Peso	Numérico
Altura	Numérico
TFG	Numérico
IMC	Numérico
año_de_diagnostico_diabetes	Fecha
año_de_diagnostico_hipertension	Fecha
Fechas_procesadas	Fecha
Bander_fechas_procesadas	Numérico
Fuente	Cadena de texto
In_consulta	Cadena de texto
Fecha_nacimiento	Fecha
Sexo	Cadena de texto
Medicamentos	Cadena de texto
Codigos_cie	Cadena de texto
Diagnosticos	Cadena de texto
Fecha_consulta	Fecha

En el conjunto de datos hay una cantidad variable de filas que se relacionan con un mismo paciente. Es posible analizar la información de cada paciente extrayendo la subtabla que le corresponde, mediante el valor de ID. En cada una de estas subtablas veríamos los atributos que se describieron anteriormente y cada fila correspondería a la información sobre el paciente generada en una visita al IMSS (consulta médica o laboratorio).

Puesto que no en todas las visitas se genera información sobre todos los atributos considerados en el conjunto de datos, existen muchos datos faltantes.

Se puede consultar una muestra de los datos en el siguiente enlace: https://docs.google.com/spreadsheets/d/1cgYY_KsEIOodEX0pVMuUno3UQ64Y5pezuzIksgUC9V4/edit#gid=0

5.4. Documentación del conjunto de datos

Este proyecto está orientado a la construcción del conjunto de datos que pueda servir a futuras investigaciones sobre la diabetes y su evolución, especialmente en pacientes mexicanos.

Puesto que el producto principal de la investigación es el conjunto de datos estructurado, nos parece muy importante documentar el procedimiento que se siguió para la construcción de los mismos y proveer información a futuros usuarios que les permita evaluar las posibilidades y limitaciones de estos datos.

Existen diferentes iniciativas que reconocen la importancia de los datos en los procesos y toma de decisiones. Por citar un ejemplo, existe *Data Ops*, un método para administrar automáticamente todo el ciclo de vida de los datos, desde la identificación, la limpieza y la integración hasta el análisis y la generación de informes. En este caso, el objetivo principal es la maximización del valor empresarial de los datos y la metodología toma prestadas prácticas comprobadas de DevOps en el ciclo de vida del desarrollo de software [91].

En este trabajo decidimos orientar la documentación de los datos a la reflexión crítica sobre los mismos y tomamos como guía la estructura propuesta en el artículo *Datasheets for datasets* por Timnit Gebru y otros colaboradores, incluyendo a Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III y Kate

Crawford [19].

La propuesta de Gebru *et al.* busca mejorar la transparencia, calidad y responsabilidad en la creación y uso de conjuntos de datos en el ámbito de la inteligencia artificial y el aprendizaje automático. La iniciativa es una reivindicación de las epistemologías críticas que buscan desafiar la hegemonía del conocimiento producido por los sistemas tecnológicos y promover un enfoque más reflexivo, responsable y ético en la producción de conocimiento.

Se planea que la documentación generada a través de *Datasheets for datasets* se distribuya junto con el conjunto de datos, con la intención de brindar información a futuros usuarios sobre la calidad y la relevancia del conjunto de datos en tareas específicas planteadas. El ejercicio de responder a las preguntas sugeridas por Gebru *et al.* puede observarse en el Apéndice A.

Finalmente, establecer el conjunto de datos que se busca estructurar mediante esta tesis, incluyendo los datos recuperados de notas médicas, requirió un proceso de exploración detallada de los datos en el SIMF. Esto permitió identificar algunas inconsistencias, o *errores de registro*, en la fecha de diagnóstico de hipertensión que aparece en el historial de pacientes. El siguiente capítulo describe el problema y presenta una propuesta para contribuir al ajuste de este tipo de errores.

Capítulo 6

Ajuste de diagnóstico de hipertensión

Trabajar en la creación de datos estructurados, como se describió en los capítulos anteriores, llevó a una exploración extensa de las diferentes tablas que conforman el Sistema de Información de Medicina Familiar (SIMF). En particular, se recurrió a diagramas y otros métodos de visualización para identificar tablas que contuvieran valores relevantes para el historial clínico del paciente.

La tabla de diagnósticos se estableció como información de máxima importancia, pues permite establecer enfermedades y el tiempo transcurrido entre la detección de cada una de ellas, en el caso de pacientes con más de un diagnóstico.

Para examinar la tabla de diagnósticos, se utilizó un diagrama como el que se muestra en la Fig. 6.1. El diagrama se elabora para cada paciente. En él, cada columna representa una fecha en la que el paciente tuvo una consulta médica de acuerdo a su expediente; es posible interpretar la figura como una línea de tiempo ordenada de izquierda a derecha. Cada fila en el arreglo representa un diagnóstico especificado en la clasificación de enfermedades relacionadas con la diabetes mellitus tipo 2 descrita anteriormente (Tabla 2.3). Finalmente, se representa con color rojo (-1) que el diagnóstico de esa fila (i) no aparece registrado en esa consulta (j) y con color verde (1) que sí está registrado, es decir, definimos un elemento a_{ij} de nuestra tabla como

$$a_{ij} = \begin{cases} 1 & \text{paciente tiene diagnóstico } i \text{ en tiempo } j \\ -1 & \text{paciente no tiene diagnóstico } i \text{ en tiempo } j \end{cases}$$

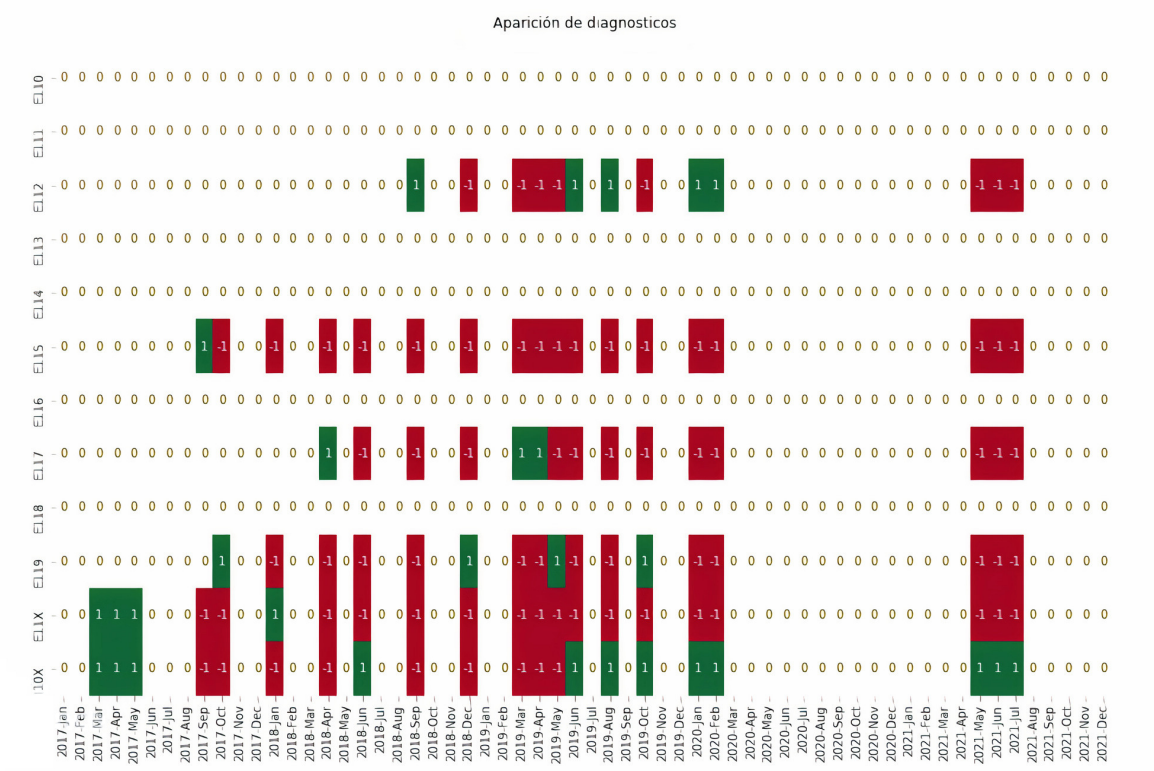


Figura 6.1: Línea de tiempo de diagnósticos para un paciente en el conjunto de datos. Cada columna representa la fecha de una consulta médica, cada fila representa una enfermedad y cada elemento ij indica si el paciente tiene (color verde) o no (color rojo) registrado el diagnóstico de la enfermedad i en la fecha j.

Existen secuencias en el diagnóstico de complicaciones que parecen improbables. Aunque es posible que algunos diagnósticos aparezcan de manera intermitente, la frecuencia con la que se observan estos casos en el archivo permite aventurar que podría tratarse de un problema sistemático en el registro de los diagnósticos durante la creación del expediente clínico, y no en el diagnóstico mismo.

Tener mayor precisión en la fecha de la aparición de complicaciones de la diabetes puede ser fundamental para implementar un modelo que refleje adecuadamente la evolución de la enfermedad. Con esta consideración, decidimos ajustar el diagnóstico de la hipertensión en los casos en donde, con base en otros elementos del expediente, se identificó un error de registro.

En este capítulo se muestran algunos casos que motivan nuestra propuesta de ajustar la fecha de los diagnósticos, se describen las reglas que establecimos en colaboración con personal médico para llevar a cabo el ajuste, y se presentan los resultados obtenidos al

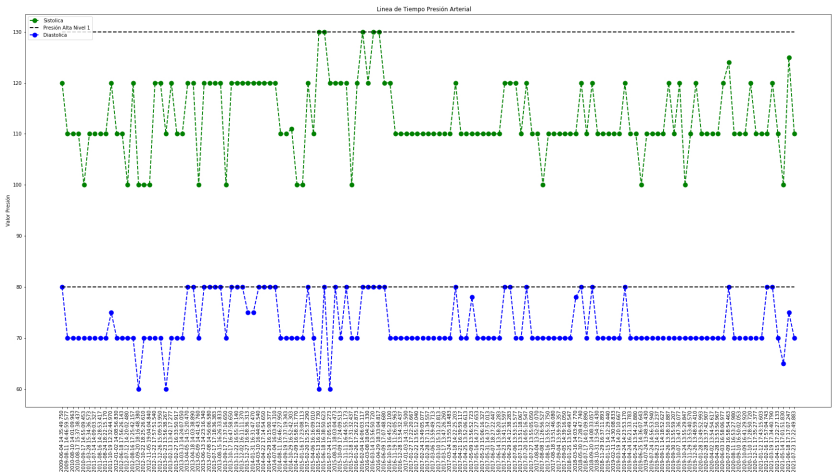
aplicar las reglas a nuestro conjunto de datos.

6.1. Diagnóstico de hipertensión en diversos ECE

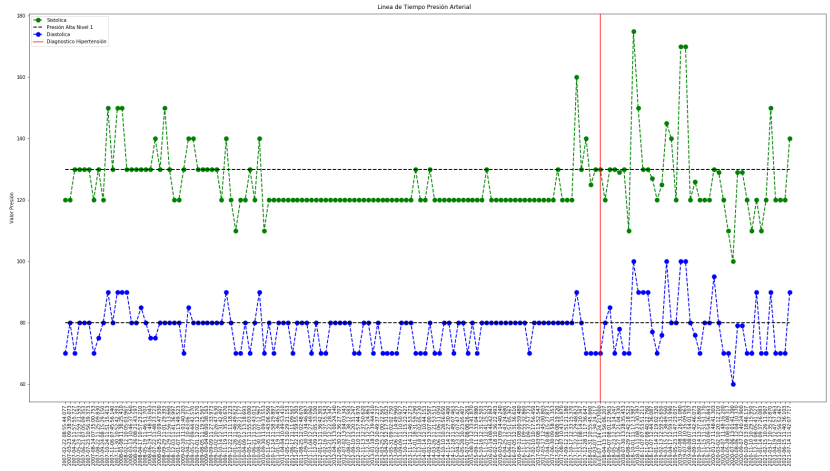
Para explorar el desarrollo de hipertensión en pacientes diabéticos se recurrió a la visualización de datos y se elaboraron gráficas que muestran los valores de presión arterial sistólica y diastólica, el límite de presión arterial y la fecha en que por primera vez se registra el diagnóstico de *hipertensión* en el ECE del paciente en cuestión.

La Fig. 6.2 muestra el caso de tres pacientes en la base de datos. Los valores de presión arterial en cada consulta médica están señalados en color verde (presión sistólica) y azul (presión diastólica). Las líneas horizontales, punteadas, señalan el límite considerado normal para cada tipo de presión.

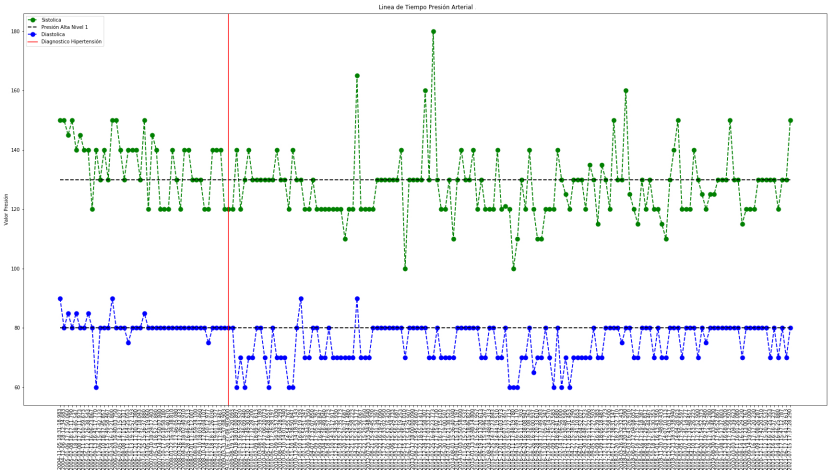
Estos casos nos parecen representativos de las situaciones en las que se sospecharía, o no, un error en la fecha registrada para el diagnóstico de hipertensión.



(a)



(b)



(c)

Figura 6.2: Valores de presión arterial sistólica (verde) y diastólica (azul) registrados en el historial clínico de tres pacientes. Cada valor corresponde a una visita médica. Las líneas horizontales señalan el límite considerado normal para cada tipo de presión. La primera persona no tiene problemas de presión. En los otros dos casos, la línea vertical roja señala el momento en que, de acuerdo al registro, los pacientes son diagnosticados con hipertensión.

La primera gráfica, Fig. 6.2a corresponde a una persona que no tiene problemas de presión. Todos los valores de la presión se mantienen por debajo de las líneas horizontales que marcan el límite considerado normal (130 para la presión sistólica y 80 para la presión diastólica).

En el caso de las Fig. 6.2b y 6.2c tenemos pacientes con hipertensión. La línea vertical roja señala el primer momento en que son diagnosticados con la enfermedad de acuerdo a lo registrados en su expediente clínico. En ambos casos puede observarse que los valores de presión están claramente por encima del límite desde mucho antes de recibir el diagnóstico.

En el caso de la Fig. 6.2b observamos valores muy altos, después valores controlados, después el diagnóstico de hipertensión y, finalmente, valores otra vez descontrolados.

En este caso podemos especular que se trata de alguien con problemas iniciales de presión, que logra controlarlos con algún factor (estilo de vida, medicamento) y que es hasta que nuevamente tiene problemas importantes que el personal médico registra su diagnóstico en el expediente.

Al revisar con mayor detalle este caso, se observó que la/el paciente estaba tomando medicamentos para hipertensión a partir del momento en que la presión aparece controlada. Esto sugiere que no se trata de un caso en el que la hipertensión haya pasado desapercibida para el personal médico, sino uno en el que hubo un error de omisión en el registro del diagnóstico.

Analizar el historial clínico de una muestra de pacientes revela casos similares a los que se han descrito en esta sección. Esto nos llevó a la pregunta sobre errores en el diagnóstico vs. errores en el registro del diagnóstico.

Nuestro objetivo en esta parte de la investigación consistió en identificar, mediante un algoritmo, casos que requieran un ajuste por cualquiera de estas dos causas (error en diagnóstico o en registro).

6.2. Ajuste de diagnóstico de hipertensión

Con base en nuestra exploración, consideramos que los pacientes pueden estar en cualquiera de las cuatro situaciones descritas a continuación:

1. Pacientes con presión arterial dentro del rango normal

2. Pacientes con problemas de presión cuya fecha de diagnóstico de hipertensión coincide con el inicio de consumo de medicamentos relacionados con la enfermedad

3. Pacientes con problemas de presión que reciben medicamentos para hipertensión *antes* de tener un diagnóstico registrado de la enfermedad

4. Pacientes con problemas de presión, es decir, con valores fuera del rango considerado normal y que, sin embargo, no aparecen registrados como hipertensos ni reciben medicamentos relacionados hasta tiempo después.

Nuestro objetivo es identificar los dos últimos casos y ajustar la fecha del diagnóstico para estos pacientes.

El primero de los casos de nuestro interés depende de analizar el historial de medicamentos prescritos para cada paciente, y examinar si alguno de ellos se relaciona con la hipertensión.

El segundo, depende de un análisis de los valores de la presión arterial que tiene cada paciente y de los criterios que utiliza el IMSS en el primer nivel de atención para diagnosticar la hipertensión.

En las siguientes secciones discutimos cada uno de estos casos de interés.

6.2.1. Diagnóstico de hipertensión

Los rangos para establecer lo que se considera una presión arterial normal dependen del país y de la institución. En esta investigación se utilizó la clasificación de la presión arterial (PA) usada por el Instituto Mexicano del Seguro Social (IMSS), que consiste en siete categorías que van desde *Óptima* hasta *Hipertensión sistólica aislada*. Las categorías establecidas se muestran en la Tabla 6.1.

Tabla 6.1: Rangos establecidos para la presión arterial en el Instituto Mexicano del Seguro Social [13]

Categoría	Sistólica (mm Hg)	Diastólica (mm Hg)
Óptima	<120	<80
Normal	120-129	80-84
Normal Alta	130-139	85-89
Hipertensión grado 1	140-159	90-99
Hipertensión grado 2	160-179	100-109
Hipertensión grado 3	>=180	>=110
Hipertensión sistólica aislada	>=140	<90

Con este marco de referencia, la hipertensión se puede detectar mantenido un nivel de la presión arterial sistodiastólica igual o superior a 140/90 mmHg, respectivamente, tomada en condiciones apropiadas en por lo menos tres lecturas – de preferencia en tres días diferentes o cuando la presión arterial inicial sea muy elevada y/o cuando el paciente presenta cifras normales bajo tratamiento antihipertensivo [92].

En nuestro proyecto decidimos tomar como valores límite de referencia 130/85, considerando que personas con esta PA son candidatos importantes a desarrollar la enfermedad.

6.2.2. Medicamentos en el tratamiento de la hipertensión

Consideramos que el consumo de medicamentos para controlar la presión arterial es un indicador muy importante de que la hipertensión se detectó. Quizá porque los medicamentos en el IMSS están sujetos a control de inventario, es más probable que se registren éstos al dignóstico mismo de la enfermedad de de manera explícita.

Basándonos en los criterios del personal médico participante en el proyecto, seleccionamos medicamentos para la hipertensión en las siguientes categorías:

- Antiarrítmicos.
- Calcio antagonistas.
- IECA.
- Diuréticos tiazídicos.

- Diuréticos de asa.
- ARA II.
- Beta-Bloqueadores
- Alfa-1 Bloqueadores

Los medicamentos específicos considerados en cada categoría se muestran en la Tabla 6.2.

Tabla 6.2: Medicamentos para el tratamiento de hipertensión arterial sistémica y su categoría

Medicamento	Clase
Amiodarona	Antiarrítmicos
Amlodipino	Calcio antagonistas
Captopril	IECA
Clortalidona	Diuréticos tiazídicos
Enalapril	IECA
Lisinopril	IECA
Ramipril	IECA
Furosemida	Diuréticos de asa
Hidroclorotiazida	Diuréticos tiazídicos
Losartan	ARA II
Metoprolol	Beta-Bloqueadores
Nifedipino	Calcio antagonistas
Prazocina	Alfa-1 Bloqueadores
Propafenona	Antiarrítmicos
Propranolol	Beta-Bloqueadores
Telmisartan	ARA II
Verapamilo	Calcio antagonistas

6.2.3. Reglas propuestas para identificar hipertensión

Con base en la información presentada anteriormente, propusimos las siguientes reglas para ajustar el diagnóstico de pacientes hipertensos en casos de errores de registro o de omisión médica.

Siguiendo los valores registrados en el expediente clínico electrónico de pacientes, sugerimos que alguien es hipertenso si:

Regla 1. Durante la cita médica le prescribieron medicamentos relacionados con la hipertensión, de acuerdo a la lista presentada en la Tabla 6.2.

Regla 2. Se registraron 3 valores consecutivos fuera del rango normal de la presión arterial, con una ventana de tiempo entre consultas médicas que no exceda 1 año.

Regla 3. Se registaron valores fuera del rango normal en al menos 50 % de sus consultas médicas, siempre y cuando la cantidad de consultas registradas sea igual o mayor al promedio general anual de todos los pacientes registrados en la base de datos.

Las dos primeras reglas propuestas se basan directamente en la información descrita en las secciones 6.2.1 y 6.2.2.

La tercera está también basada en los rangos considerados normales para la presión arterial, excepto que en este caso se relaja la condición de que los valores de la presión sean altos tres veces consecutivas. A cambio, importa la frecuencia de visitas al IMSS de cada paciente. Esta restricción obedece al hecho de que la cantidad de consultas médicas puede variar enormemente de un paciente a otro. Para determinar que tenemos suficiente información, pedimos que la cantidad de revisiones médicas sea igual o mayor al promedio general.

Las reglas se siguen conforme al orden propuesto. Es decir, en el caso de pacientes que reciben medicamentos (Regla 1) pero tenían mediciones de la presión fuera del rango de lo normal desde tiempo atrás (Regla 2), hemos decidido utilizar la fecha que se infiere con la Regla 1 por considerarla un indicio más sólido, con la información que tenemos, de un diagnóstico avalado por el personal médico.

La estructura general del código que se utilizó para esta implementación se muestra en los Algoritmos 2, 3 y 4.

Algoritmo 2 Estructura general Regla 1

Unir tablas exprel_diabetes_diagcie, corcat_CIE,corhis_consulta y seleccionar las columnas pertinentes
Tomar fechas más antiguas correspondientes a ambas enfermedades
Calcular días entre estas dos fechas
Filtrar solo fechas mayores a 0
Agrupar por paciente y fecha de consulta
Agrupar códigos de medicamentos relacionados con la hipertensión en sus diferentes presentaciones en una lista
Unir pacientes con los medicamentos de cada consulta
Filtrar datos de los pacientes antes del primer diagnóstico de diabetes mellitus tipo 2
for *paciente* en *pacientes* **do**
    for *consulta* en *paciente* **do**
        if paciente tiene uno de los medicamentos marcados **then**
            $medicamento\_marca \leftarrow 1$
        end if
    end for
end for
Tomar la menor fecha de los pacientes marcados

Algoritmo 3 Estructura general Regla 2

Tomar id restantes después de aplicar la regla de medicamentos
Transformar el conjunto de datos para agrupar los valores de presión arterial del paciente
for id en $id\_faltantes$ **do**
    Extraer información del paciente
    Cambiar el formato de la fecha a datetime
    Eliminar los valores faltantes de presión arterial sistólica y diastólica
    Eliminar consultas repetidas
    $Paciente\_hipertension\_fecha \leftarrow PrimerdiagnosticoHipertension$
    Filtrar para tener la información del paciente sólo antes del primer diagnóstico
    for $valor\_presion\_arterial$ intervalos de 3 valores consecutivos **do**
        cond1 $\leftarrow$ $valor\_presion\_arterial \geq$ limite presión
        cond2 $\leftarrow$ fecha entre consultas $\leq$ 1 año
        if cond1 and cond2 **then**
            Guardar la última fecha de este intervalo
        end if
    end for
end for

Algoritmo 4 Estructura general Regla 3

Tomar los id restantes después de aplicar la Regla 2
Transformar el conjunto de datos para contar cuántas veces tiene registrado la presión arterial en el año
Con los datos del paso anterior extraer el promedio de consultas médicas.
for id en $id\_faltantes$ **do**
    Extraer información del paciente
    Cambiar el formato de la fecha a datetime
    Eliminar los valores faltantes de presión arterial sistólica y diastólica
    Eliminar consultas repetidas
    $Paciente\_hipertension\_fecha \leftarrow$ Primer diagnóstico hipertensión
    Filtrar para tener la información del paciente solo antes del primer diagnóstico
    $Citas\_en\_un\_ano \leftarrow$ cantidad de citas en un año
    $Citas\_50\_porciento \leftarrow$ Dividimos entre dos $Citas\_en\_un\_ano$
    $presionAlta \leftarrow$ Cantidad de veces que presión sobrepasó el límite
    if $presionAlta \geq Citas\_50\_porciento$ **then**
        $cantidadConsultas \leftarrow$ Número de consultas de ese año
        if $cantidadConsultas \geq$ Promedio anual **then**
            Guardar los datos de la última consulta con presión arterial alta
        end if
    end if
end for

6.3. Resultados

Las reglas descritas en la sección anterior se aplicaron en todo el conjunto de datos de pacientes con diabetes mellitus tipo 2 e hipertensión, que abarcó 15,510 pacientes.

Al término de la tarea, encontramos 7726 diagnósticos de hipertensión ajustados, de los cuales

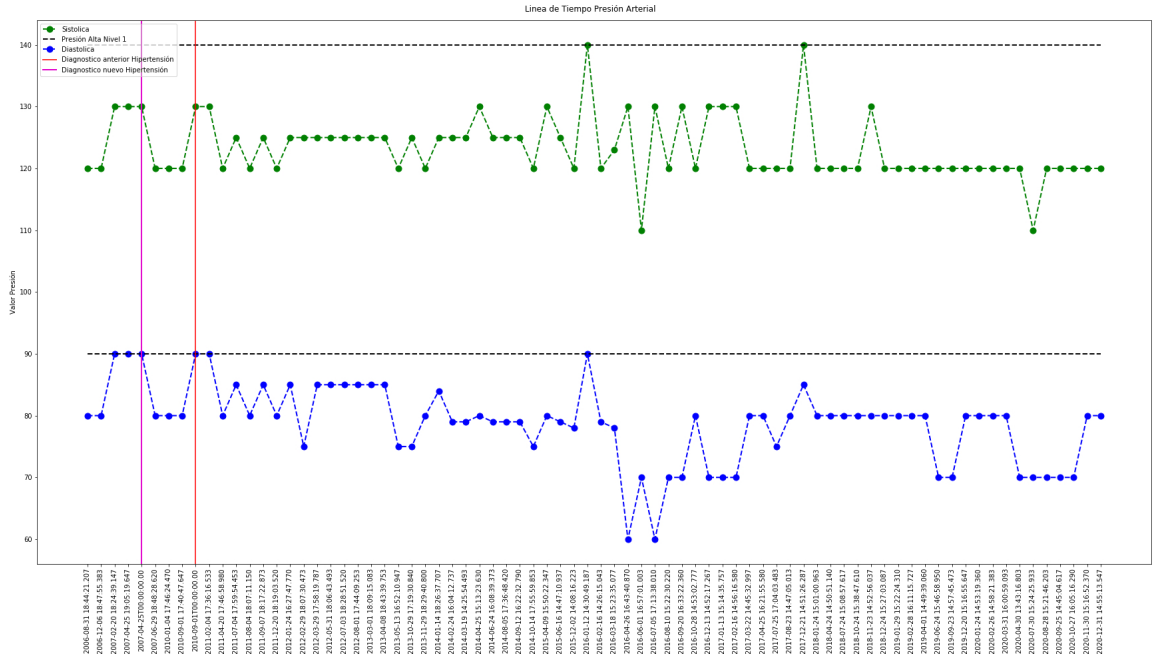
- 7626 se ajustaron siguiendo la Regla 1,
- 64 se ajustaron de acuerdo a la Regla 2,
- 35 correspondieron a la Regla 3.

Que la mayor parte de los diagnósticos ajustados se basen en los medicamentos resulta conveniente para nuestro estudio, pues hay un mayor nivel de certeza en que el error es de registro. Quizá por una limitación de tiempo, el personal médico receta un medicamento y no considera necesario escribir de manera explícita el diagnóstico, probablemente justo porque ya hizo una receta que alude al problema.

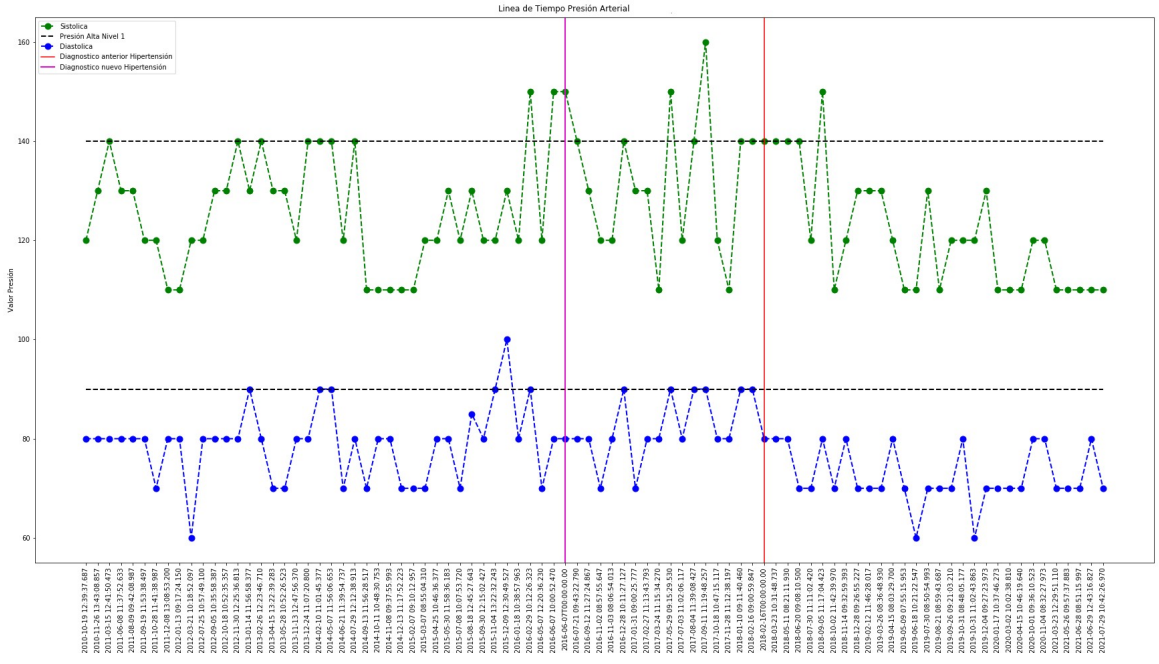
En el caso de la Regla 2 estamos ante casos de personas, 0.02 % de la población total, cuyo diagnóstico pasó desapercibido, pues con base en sus valores y de acuerdo a los criterios establecidos por el propio IMSS, deberían haber sido considerados hipertensos.

La Regla 3 abarcó menos instancias. Esto no es sorprendente tomando en cuenta que imponer la restricción de la cantidad de visitas médicas reduce la cantidad de pacientes considerados. Sin embargo, creemos que seguir esa restricción es un enfoque cauteloso ante la cantidad tan desigual de consultas médicas que los pacientes pueden tener.

La Fig. 6.3 muestra dos ejemplos de diagnósticos ajustados con nuestras reglas. La línea roja corresponde al momento del diagnóstico registrado por primera vez en el expediente, mientras que la línea rosa señala el diagnóstico ajustado.



(a)



(b)

Figura 6.3: Ejemplo del ajuste del momento del diagnóstico de hipertensión que se realizó en los historiales clínicos, siguiendo reglas propuestas en este trabajo. La línea vertical roja corresponde a la fecha registrada de aparición de la enfermedad; la línea rosa señala la fecha que puede inferirse a partir del uso de medicamentos o de los valores de presión sistólica y diastólica del paciente.

Resulta imposible establecer, usando solamente los valores de presión, la evolución de la enfermedad en cada paciente pero consideramos que hacer este ajuste puede impactar la forma en que el modelo refleje el desarrollo real de las complicaciones en pacientes con diabetes mellitus tipo 2.

Capítulo 7

Conclusiones

En este trabajó se abordó el problema de crear un conjunto de datos estructurados a partir de información en el Sistema de Información de Medicina Familiar (SIMF) del Instituto Mexicano del Seguro Social (IMSS). La información describe el historial clínico de pacientes con diabetes mellitus tipo 2 y la aparición gradual de diferentes complicaciones de salud.

La creación de este conjunto de datos es muy importante para poder hacer análisis de la evolución de la diabetes en el contexto nacional, pues la base de datos del SIMF es una de las más grandes relacionadas con la enfermedad.

Información importante se encontraba en notas médicas, en forma de un texto en lenguaje natural. El modelo de Reconocimiento de entidades nombradas (NER) resultó muy valioso para poder extraer información.

A pesar de su utilidad, NER tiene limitaciones. Por ser una técnica de aprendizaje automático, mejora conforme los datos son de mejor calidad. En el caso de las notas médicas, entrenar el algoritmo no es sencillo y plantea retos como reconocer diversas formas de escribir los parámetros clínicos, lidiar con faltas de ortografía y la estructura de la nota médica dada por cada especialista. Mejorar el modelo requeriría tener un formato estándar para las notas médicas.

Durante el proceso de extraer parámetros se identificó un problema en el registro del diagnóstico de enfermedades. Es común que, aún cuando el personal médico diagnostica una enfermedad y prescribe medicamentos apropiados, el campo en el SIMF correspon-

diente al diagnóstico aparezca sin información. El problema podría atribuirse al tiempo reducido que se asigna a las consultas, a que el registro se considere redundante a la luz de los medicamentos prescritos y de otros indicadores en el expediente, o a otros factores desconocidos para nosotros.

La falta de una fecha explícita y confiable que señale la aparición de una complicación puede afectar la eficacia de modelos para analizar la evolución de la salud de pacientes diabéticos. En el trabajo nos centramos en el caso de la hipertensión y propusimos un conjunto de reglas que podrían utilizarse para ajustar el registro del diagnóstico de pacientes diabéticos hipertensos. Consideramos que esta metodología es una contribución importante del proyecto, a la par de la base de datos generada.

La exploración realizada subraya la necesidad de continuar los esfuerzos orientados a mantener y mejorar el SIMF, incorporando estrategias que faciliten la limpieza y mejoren la precisión de los registros que se generan en el IMSS. Esto podrían contribuir a reducir datos faltantes en los expedientes clínicos electrónicos. Adicionalmente, se podría mejorar el sistema de almacenamiento utilizando un *data warehouse*[1] o un *data lake*[2] para que los investigadores puedan extraer el conocimiento de estos datos.

Al margen de los resultados de investigación, este trabajo resaltó el papel fundamental que, en problemas de la vida real, tiene la limpieza y preprocesamiento de datos, el trabajo en equipo y la colaboración con personas de diversas disciplinas. El trabajo interdisciplinario es fundamental para tener una perspectiva más amplia del problema y de las diversas técnicas y herramientas que se pueden usar para procesar grandes cantidades de datos.

Trabajo a futuro

En este trabajo se mostró el uso de técnicas computacionales que permiten recuperar información y mejorar bases de datos que contienen una gran cantidad de datos faltantes. También se propusieron metodologías para mejorar diagnósticos clínicos. Tanto la recuperación de información estructurada como el ajuste de diagnósticos (y ambas combinadas)

[1]Un *data warehouse* es un sistema que agrega datos de diferentes fuentes en un único almacén de datos central para admitir el análisis de datos,la inteligencia artificial (IA), etc [93].

[2]Un *data lake* es un repositorio centralizado que ingiere y almacena grandes volúmenes de datos en su forma original [94].

son importantes para mejorar el historial clínico electrónico de pacientes.

Para mejorar este trabajo se podría automatizar el proceso de extracción de parámetros clínicos usando un orquestador de datos, como *Airflow*[3], que esté conectado a la base de datos para que, cuando ingresan una nota médica, este pudiera extraer los atributos y guardarlos en una tabla.

Otro punto a considerar es la posibilidad de extender el ajuste del registro de diagnósticos a las diversas enfermedades vinculadas con la diabetes mellitus, como: complicaciones renales, complicaciones oftálmicas, complicaciones neurológicas, complicaciones circulatorias.

Finalmente, sería recomendable promover un estándar para escribir fechas o parámetros clínicos en las notas médicas. De este modo, se podría reducir la incertidumbre en los datos y algoritmos de IA que se utilicen en el futuro podrían tener un mejor desempeño.

[3]Apache Airflow es una plataforma de código abierto para desarrollar, programar y monitorear flujos de trabajo orientados a lotes [95].

Apéndice A

Base de datos generada

A.1. Motivación

Este conjunto de datos fue creado por César Arcos González, estudiante de la Licenciatura en Tecnologías para la Información en Ciencias en la Escuela Nacional de Estudios Superiores unidad Morelia de la Universidad Nacional Autónoma de México.

El conjunto se creó con recursos públicos, en el marco del proyecto *"Estudio longitudinal para el desarrollo de modelos predictivos de complicaciones crónicas de la diabetes mellitus tipo 2"*, que se llevó a cabo con apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT) a través del Fondo Institucional de Fomento Regional para el Desarrollo Científico, Tecnológico y de Innovación (FORDECyT) de los Programas Nacionales Estratégicos (ProNacEs) en la convocatoria «2019-06 Proyectos de investigación e incidencia en ciencia de datos y salud: integración, procesamientos, análisis y visualización de datos de salud en México». El proyecto fue dirigido por la Dra. Anel Gomez García, adscrita al Centro de Investigación Biomédica de Michoacán.

El propósito en la construcción de este conjunto de datos es estructurar la información relacionada con la diabetes mellitus contenida en el Sistema de Información de Medicina Familiar (SIMF) del Instituto Mexicano del Seguro Social (IMSS), y presentarla en un formato que facilite la realización de trabajos de investigación, especialmente aquellos que tienen un enfoque de modelación.

A.1.1. Composición

Las instancias que componen el conjunto de datos representan registros electrónicos de salud de pacientes con diabetes mellitus tipo 2 en el estado de Michoacán, México.

Cada instancia corresponde a un registro en el expediente de algún paciente. En total, existen 9,577,839 instancias en el conjunto de datos, que representan todas las instancias posibles de acuerdo a la información encontrada en el SIMF al momento de la realización del proyecto.

Cada instancia consta de:

- Id: identificación del registro.
- Cx_curp: id del paciente.
- Glucosa: valor de glucosa separada por | el cual el primer valor es la preprandial y la segunda postprandial.
- Colesterol: valor del colesterol.
- Triglicéridos: valor de los triglicéridos.
- Hdl: valor de hdl.
- Ldl:valor de ldl.
- Fecha: fecha encontrada en la nota médica.
- Presion_arterial: valor de la presión arterial.
- Hba1c: valor de hba1c.
- Hipertension: valor binario se es que se encontró en la nota médica.
- Plaquetas: calor de las plaquetas.
- Creatinina: valor de la creatinina.
- Acido_urico: valor del ácido úrico.

- Urea: valor de la urea.
- Peso: valor del peso del paciente.
- Altura: valor de la altura del paciente.
- Tfg: valor del tfg.
- Imc: índice de masa corporal,
- Año_de_diagnostico_diabetes: año de diagnóstico de la diabetes encontrado en la nota médica.
- Año_de_diagnostico_hipertensión: año de diagnostico de la hipertension encontrado en la nota medica.
- Fechas_procesadas: fechas procesadas para tener un valor fecha estandar.
- Bandera_fechas_procesadas: bandera con valor binario la cual indica si la fecha fue procesada.
- Fuente: la fuente del registro puede provenir del algoritmo o de una tabla de la base de datos.
- In_consulta: id de la consulta.
- Fecha_nacimiento: fecha de nacimiento del paciente.
- Sexo: sexo del paciente.
- Medicamentos: medicamentos recetados al paciente en la consulta medica.
- Codigos_cie: codigos cie de los diagnosticos marcados en la consulta medica.
- Diagnosticos: diagnósticos marcados en la consulta medica.
- Fecha_consulta: fecha de la consulta.

Los datos faltantes en este conjunto de datos corresponden a datos faltantes en la base de datos original, del SIMF. No se excluyó información de instancias individuales.

Las etiquetas que identifican a cada paciente fueron asignadas por este grupo, para eliminar la aparición de datos personales pero conservar un ID que permitiera reconstruir el historial clínico de cada persona.

Esto permite hacer explícitas las relaciones entre instancias (registros médicos), pues aquellas que corresponden al mismo paciente pueden agregarse usando el ID.

El conjunto se creó combinando datos que provienen de tablas relacionales en el SIMF y datos extraídos de notas médicas a través de un algoritmo basado en redes neuronales artificiales. En consecuencia, podria existir redundancia en algunos datos en algunos indicadores. Además, puesto que la incertidumbre asociada a los datos varía según la fuente (registro en el SIMF o algoritmo de reconocimiento de entidades nombradas en notas médicas), hemos considerado apropiado incluir un atributo que señala el origen de cada registro. Se sugiere verificar esta información y tomar en cuenta que podría haber errores en el conjunto de datos que replican errores existentes en el SIMF.

El conjunto de datos es autónomo y no se vincula con recursos externos. Además, ha sido procesado para eliminar datos que podrían identificar individuos. Tampoco existe información que podría ser ofensiva, amenazante o causal de algún daño a quienes utilicen el conjunto de datos.

En los datos es posible identificar subplobaciones por edad, por género y por las variables: peso, altura, año de diagnóstico diabetes, año de diagnóstico hipertensión y diagnósticos.

Los datos no revelan información sobre orígenes étnicos o raciales, orientaciones sexuales, creencias religiosas, opiniones políticas o afiliaciones sindicales, ubicaciones, datos financieros, datos biométricos o genéticos, formas de identificación gubernamental, números de seguridad socialo historial criminal, pero contiene datos relacionados con el historial clínico de pacientes con diabetes mellitus tipo 2. Estos datos podrían considerarse confidenciales, ya que involucran información sobre la salud de las personas, incluidos detalles médicos, tratamientos, resultados de exámenes, medicamentos recetados y otros aspectos relevantes para el manejo de la enfermedad. Deben manejarse con la sensibilidad y

consideraciones de conjuntos de datos similares.

A.1.2. Proceso de recolección

El conjunto de datos se construyó a partir de la información contenida en el SIMF y compartida con el equipo de investigación en el marco de un proyecto de investigación. En consecuencia, los datos provienen del registro de consultas médicas y exámenes de laboratorios hechos en el IMSS. Se desconoce si las personas representadas dieron su consentimiento para la recopilación y el uso de sus datos.

La selección de parámetros a incluír se realizó al interior del equipo de trabajo, en colaboración cercana con personal médico que determinó indicadores relevantes para el estudio de la diabetes mellitus tipo 2 y sus complicaciones.

En el caso de datos recuperados de notas médicas, se utilizó un algoritmo de reconocimiento de entidades nombradas (NER) basado en redes neuronales artificiales perteneciente a la biblioteca SpaCy. Las notas que se usaron para entrenar el algoritmo fueron etiquetadas por personal médico contratado para ese fin mediante el proyecto.

Los registros del SIMF corresponden a información recopilada directamente por personal médico del IMSS durante el periodo del 05-10-2006 al 05-08-2021.

El uso de esta información con fines de investigación se aprobó de acuerdo a la revisión CONBIOÉTICA-09-CEI-009-20160601.

A.1.3. Preprocesamiento/limpieza/etiquetado

Al momento de la integración de datos provenientes de tablas del SIMF y de notas médicas se llevaron a cabo algunas tareas de limpieza. Para evidenciar estos procesos (1) se etiquetaron los datos para identificar de qué tabla fueron extraídos, y (2) se agregó una bandera para señalar qué fechas fueron procesadas para homologar el formato en el que aparecen.

En cuanto a la limpieza y preprocesado de los datos, esta se realizó principalmente en datos extraídos de notas médicas que están marcados con la etiqueta NER en la columna "fuente". Uno de los campos que se procesó fue la columna fecha, en la que se aplicaron

expresiones regulares para convertir las fechas a formatos *datetime*. Adicionalmente se aplicó un filtro para eliminar datos identificados por el algoritmo NER que tenían una longitud anómala respecto a su campo correspondiente; con esto se buscó filtrar ruido que, por error, hubiera devuelto el algoritmo.

Los datos sin procesar se encuentran respaldados por el personal académico a cargo del proyecto.

Todas las tareas de limpieza y procesamiento se llevaron a cabo utlizando Python. El código está disponible a petición.

A.1.4. Usos

El conjunto de datos se ha utilizado a esta fecha, agosto de 2023, en tareas de predicción de hipertensión en pacientes diabéticos, utilizando un enfoque de aprendizaje automático.

Los datos podrían usarse, en general, para el desarrollo e implementación de modelos matemáticos y computacionales relacionados con la diabetes mellitus tipo 2 en pacientes mexicanos.

El conjunto no incluye datos nutricionales ni odontológicos. Sólo se consideraron datos clínicos, siguiendo el criterio del personal médico participante en el proyecto.

A.1.5. Distribución

Se espera que estos datos se pongan a disposición de la comunidad científica, cuando concluya el proyecto. Por el momento se desconoce la forma de distribución y los términos de uso aplicables.

A.1.6. Mantenimiento

El conjunto de datos se encuentra alojado en un servidor propiedad de IMSS. En este momento no se tiene contemplada ninguna actualización de los datos.

Para cualquier duda o aclaración, se puede contactar directamente al creador del conjunto de datos (cesar99@gmail.com) o a la coordinadora del proyecto (anel.gomez@imss.gob.mx).

Bibliografía

[1] Instituto Mexicano del Seguro Social. *Diagnóstico y Tratamiento Farmacológico de la Diabetes Mellitus Tipo 2 en el Primer Nivel de Atención. Guía de Evidencias y Recomendaciones: Guía de Práctica Clínica.* México, 2018.

[2] A Viniegra Osorio, C Sierra Soria, E Fabela Pérez, J Barbosa, M Medina González, M Rolón Montaño, and R Castaño Guerra. *Diagnóstico y Tratamiento de la Hipertensión Arterial en el Primer Nivel de Atención.* Instituto Mexicano del Seguro Social, Cuauhtémoc, Mexico, 2011.

[3] El manejo optométrico del paciente con Hipertensión Arterial | Optometristas.org, . URL https://optometristas.org/noticias/el-manejo-optometrico-del-paciente-con-hipertension-arterial.

[4] Artificial intelligence, enough of the hype! what is it? https://community.hpe.com/t5/hpe-blog-uk-ireland-middle-east/artificial-intelligence-enough-of-the-hype-what-is-it/ba-p/7046672, May 2019.

[5] Jieun Baek and Yosoon Choi. Deep neural network for predicting ore production by truck-haulage systems in open-pit mines. *Applied Sciences*, 10:1657, 03 2020. doi: 10.3390/app10051657.

[6] Jorge Leonel. Backpropagation - Jorge Leonel, 10 2018. URL https://medium.com/@jorgesleonel/backpropagation-cc81e9c772fd.

[7] Shruti Jadon. Introduction to Different Activation Functions for

Deep Learning, 2 2022. URL https://medium.com/@shrutijadon/survey-on-activation-functions-for-deep-learning-9689331ba092.

[8] Sumit Saha. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, 11 2022. URL https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[9] Understanding LSTM networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/, .

[10] Gaurav Singhal. Introduction to LSTM Units in RNN, 9 2020. URL https://www.pluralsight.com/guides/introduction-to-lstm-units-in-rnn.

[11] *NOM-015-SSA2- 1994. Para la prevención, tratamiento y control de la diabetes mellitus en la atención primaria*, volume 13. American Diabetes Association, .

[12] 2023 ICD-10-CM Codes E11*: Type 2 diabetes mellitus, . URL https://www.icd10data.com/ICD10CM/Codes/E00-E89/E08-E13/E11-.

[13] A Viniegra Osorio, C Sierra Soria, E Fabela Pérez, J Barbosa, M Medina González, M Rolón Montaño, and R Castaño Guerra. *Diagnóstico y Tratamiento de la Hipertensión Arterial en el Primer Nivel de Atención*. Instituto Mexicano del Seguro Social, Cuauhtémoc, Mexico, 2011.

[14] Diabetes: diferencias entre tipo 1 y tipo 2 | Cigna, . URL https://www.cigna.com/es-us/knowledge-center/hw/diabetes-uq1217abc.

[15] Secretaría de Salud. Diabetes mellitus tipo 2 hospitalaria 2021. https://www.gob.mx/salud/documentos/diabetes-mellitus-tipo-2-hospitalaria-2021.

[16] Marcos M Lima-Martínez, Carlos Carrera Boada, Marialaura D Madera-Silva, Waleskha Marín, and Miguel Contreras. COVID-19 and diabetes: A bidirectional relationship. *Clin Investig Arterioscler*, 33(3):151–157, October 2020.

[17] Akihiro Nomura, Masahiro Noguchi, Mitsuhiro Kometani, Kenji Furukawa, and Takashi Yoneda. Artificial intelligence in current diabetes management and prediction. *Curr Diab Rep*, 21(12):61, December 2021.

[18] B Jönsson and CODE-2 Advisory Board. Revealing the cost of type II diabetes in europe. *Diabetologia*, 45(7):S5–12, July 2002.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[20] Mirjam Christ-Crain, Daniel G Bichet, Wiebke K Fenske, Morris B Goldman, Soren Rittig, Joseph G Verbalis, and Alan S Verkman. Diabetes insipidus. *Nature reviews Disease primers*, 5(1):54, 2019.

[21] Gojka Roglic. WHO global report on diabetes: A summary. *Int. J. Noncommun. Dis.*, 1(1):3, 2016.

[22] K G M M Alberti, P Zimmet, and J Shaw. International diabetes federation: a consensus on type 2 diabetes prevention. *Diabet. Med.*, 24(5):451–463, May 2007.

[23] Hyperglycemia (high blood glucose). https://www.diabetes.org/healthy-living/medication-treatments/blood-glucose-testing-and-control/hyperglycemia, .

[24] Diabetes: Hipoglicemia e Hiperglicemia. *Northwestern Memorial Hospital*, .

[25] A Sosa, J C Celis, and A C Burgos. Conteo de hidratos de carbono como herramienta para el control de los niveles de glucosa. *Desarrollo científico de enfermería*, 20(8): 243–248, 2012.

[26] Hafiz Farooq Ahmad, Hamid Mukhtar, Hesham Alaqail, Mohamed Seliaman, and Abdulaziz Alhumam. Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Appl. Sci. (Basel)*, 11(3):1173, January 2021.

[27] A L Galant, R C Kaufman, and J D Wilson. Glucose: Detection and analysis. *Food Chem.*, 188:149–160, December 2015.

[28] Maldonado Saavedra, O Ramírez Sánchez, I García Sánchez, J R Reyes, and G M Bolaina. Colesterol: Función biológica e implicaciones médicas. *Revista mexicana de ciencias farmacéuticas*, 43:7–22, 2012.

[29] C I Ponte. Redescubriendo los triglicéridos como factor de riesgo cardiovascular. *Avances Cardiol*, 29(4):367–367, 2009.

[30] J E F Alfonso and I D S Ariza. Elevando el colesterol HDL: cuál es la mejor estrategia? *Revista da associação médica brasileira*, 54:369–376, 2008.

[31] Teresa L. Errico, Xiangyu Chen, Jesús M. Martin Campos, Josep Julve, Joan Carles Escolà-Gil, and Francisco Blanco-Vaca. Mecanismos básicos: estructura, función y metabolismo de las lipoproteínas plasm. *Clínica e Investigación en Arteriosclerosis*, 25(2):98–103, 2013. ISSN 02149168. doi: 10.1016/j.arteri.2013.05.003. URL https://doi.org/10.1016/j.arteri.2013.05.003.

[32] C M Miguel. *Libro de la salud cardiovascular del Hospital Clínico San Carlos y la Fundación BBVA*. Fundación BBVA, 2007.

[33] La presión arterial alta. https://www.nia.nih.gov/espanol/presion-arterial-alta, .

[34] J Pérez Loredo, C A Lavorato, and A L Negri. NUMEROSOS MÉTODOS DE MEDICIÓN (parte i). revista de nefrologia. *Revista de nefrologia, dialisis y trasplante*, 35:153–164, 2015.

[35] Mariela Bracho-Nava, Victoria Stepenka-Alvarez, Maribel Sindas-Villasmil, Yoleida Rivas de Casal, María Bozo de González, and Anyelo Duran-Mojica. HEMOGLOBINA GLICOSILADA O HEMOGLOBINA GLICADA, ¿cuál DE LAS DOS? *Saber (Cumana)*, 27(4):521–529, 2015.

[36] B Gómez-Gómez, F L Rodríguez-Weber, and E J Díaz-Greene. Fisiología plaquetaria, agregometría plaquetaria y su utilidad clínica. *Med Int Mex*, 34(2):244–263, 2018.

[37] Juan Pablo Huidobro E, Rodrigo Tagle, and Ana María Guzmán. Estimation of glomerular filtration rate with creatinine. *Rev. Med. Chil.*, 146(3):344–350, March 2018.

[38] Hernán Alcaíno, Douglas Greig, Pablo Castro, Hugo Verdejo, Rosemarie Mellado, Lorena García, Guillermo Díaz-Araya, Clara Quiroga, Mario Chiong, and Sergio Lavandero. Ácido úrico: Una molécula con acciones paradójicas en la insuficiencia cardiaca. *Rev. Med. Chil.*, 139(4):505–515, April 2011.

[39] Wikipedia contributors. Urea. https://es.wikipedia.org/w/index.php?title=Urea&oldid=149012369.

[40] H Ben-Ami, P Nagachandran, A Mendelson, and Y Edoute. Drug-induced hypoglycemic coma in 102 diabetic patients. *Arch. Intern. Med.*, 159(3):281–284, February 1999.

[41] Abbas E. Kitabchi and Barry M. Wall. Diabetic ketoacidosis. *Medical Clinics of North America*, 79(1):9–37, 1995. ISSN 0025-7125. doi: https://doi.org/10.1016/S0025-7125(16)30082-7. URL https://www.sciencedirect.com/science/article/pii/S0025712516300827.

[42] Angela C Webster, Evi V Nagler, Rachael L Morton, and Philip Masson. Chronic kidney disease. *Lancet*, 389(10075):1238–1252, March 2017.

[43] Gerard A Lutty. Effects of diabetes on the eye. *Invest. Ophthalmol. Vis. Sci.*, 54(14): ORSF81–7, December 2013.

[44] Andrew P Mizisin, G Diane Shelton, Monica L Burgers, Henry C Powell, and Paul A Cuddon. Neurological complications associated with spontaneously occurring feline diabetes mellitus. *J. Neuropathol. Exp. Neurol.*, 61(10):872–884, October 2002.

[45] Abdullah Al Wahbi. Autoamputation of diabetic toe with dry gangrene: a myth or a fact? *Diabetes Metab. Syndr. Obes.*, 11:255–264, June 2018.

[46] Jesús Honorato Pérez, Andrés Purroy Unanua, P.U. Andrés, and H.P. Jesús. *Hipertensión arterial*. Alianza Editorial, Madrid, Spain, 2002.

[47] Julio Álvarez, Francisco Aguilar, and Empar Lurbe. La medida de la presión arterial en niños y adolescentes: Elemento clave en la evaluación de la hipertensión arterial. *Anales de Pediatría*, 96(6):536.e1–536.e7, 2022.

[48] Presión arterial - Videos de salud: MedlinePlus Enciclopedia Médica, . URL https://medlineplus.gov/spanish/ency/anatomyvideos/000013.htm#:~:text=La%20presi%C3%B3n%20sist%C3%B3lica%20se%20mide%20cuando%20el%20los,cuando%20el%20los%20ventr%C3%ADculos%20del%20coraz%C3%B3n%20se%20relajan.

[49] Tania Molina Jiménez and Blandina Bernal Morales. Método Auscultatorio. *QUÍMICA FARMACÉUTICA BIOLÓGICA GUÍA DE PRÁCTICAS DE MORFOFISIOLOGÍA*.

[50] Waldo Rodríguez-Valera Yoel Flores-Alés Andrés J. Stanchi Nestor Oscar Rejas-López Juan Antúnez-Sánchez, Guillermo Ramírez-Sánchez. La formación on line desde el aula virtual veterinaria: resultados y experiencias. *REDVET. Revista Electrónica de Veterinaria*, 2008. URL https://www.redalyc.org/articulo.oa?id=63617117001.

[51] J P González-Rivas and La Medición De La Presión Arterial En La Optimizando. OPTIMIZANDO LA MEDICIÓN DE LA PRESIÓN ARTERIAL EN LA CONSULTA. *Revista Venezolana de Endocrinología y Metabolismo*, 14(3):179–186, 2016.

[52] Ministerio de salud pública – el ministerio de salud pública ejerce la rectoría del sistema nacional de salud a fin de garantizar el derecho a la salud del pueblo ecuatoriano. http://salud.gob.ec, .

[53] Aneurisma cerebral - Síntomas y causas - Mayo Clinic, 6 2022. URL

https://www.mayoclinic.org/es-es/diseases-conditions/brain-aneurysm/symptoms-causes/syc-20361483.

[54] Ranya Sweis and Arif Jivan. Introducción a la arteriopatía coronaria (coronariopatía), 2 2022. URL https://www.msdmanuals.com/es-mx/hogar/trastornos-del-coraz%C3%B3n-y-los-vasos-sangu%C3%ADneos/arteriopat%C3%ADa-coronaria-coronariopat%C3%ADa/introducci%C3%B3n-a-la-arteriopat%C3%ADa-coronaria-coronariopat%C3%ADa.

[55] Estela Guadalupe Moreno Betanzo, Javier Velasco Ruiz, and Hilse Alcocer Tapia. *Insuficiencia cardiaca.* 1985. URL http://pbidi.unam.mx:8080/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat02029a&AN=tes.TES01000035677&lang=es&site=eds-live.

[56] E Sánchez Chávez and A. S. Salazar Rosales. *Manejo del estrés a nivel odontológico en pacientes con antecedentes de ataque isquémico transitorio (AIT) tratados con antiagregantes plaquetarios.* 20220.

[57] Instituto Mexicano del Seguro Social. *Guía de Práctica Clínica Diagnóstico y tratamiento de la demencia vascular en el adulto en los tres niveles de atención.* 3 2017.

[58] default - Stanford Medicine Children's Health, . URL https://www.stanfordchildrens.org/es/topic/default?id=glomerulosclerosis-85-P04574.

[59] FJ Gaínza de los Ríos and JM López Gómez. Insuficiencia Renal Aguda. *Nefrología al día*, (2659-2606):https://www.nefrologiaaldia.org/317.

[60] Ricardo Flores Chávez and Elizabeth Calderón Taboada. *La hemodiálisis como factor de riesgo para la progresión de la retinopatía hipertensiva en pacientes con enfermedad renal crónica.* 2014. URL http://pbidi.unam.mx:8080/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat02029a&AN=tes.TES01000716988&lang=es&site=eds-live.

[61] Daniela Gasca Cuello, Jennifer Paola Martínez Parra, Juan sebastián Gómez Gordillo, Susan Lizeth Delgado Contreras, and Ricardo Andres Fuentes Martinez. Manifestaciones de la retinopatía hipertensiva y de la retinopatía diabética en población adulta. *Scientific and Educational Medical Journal*, 1(1):64–72, dic. 2020. URL https://www.medicaljournal.com.co/index.php/mj/article/view/15.

[62] Tom Michael Mitchell. *Machine Learning*. McGraw-Hill Education, New York, United States, 1997.

[63] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia*, pages 21–49. Springer, 2008.

[64] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Amsterdam University Press, Amsterdam, Netherlands, 2016.

[65] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

[66] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73: 1–15, 2018. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2017.10.011. URL https://www.sciencedirect.com/science/article/pii/S1051200417302385.

[67] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 10 1986. doi: 10.1038/323533a0. URL http://dx.doi.org/10.1038/323533a0.

[68] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *towards data science*, 6(12):310–316, 2017.

[69] DeepAI. Hidden Layer, 6 2020. URL https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning.

[70] Rahul Chauhan, Kamal Kumar Ghanshala, and RC Joshi. Convolutional neural network (cnn) for image detection and recognition. In *2018 First International Con-*

ference on Secure Cyber Computing and Communication (ICSCCC), pages 278–282. IEEE, 2018.

[71] Nikhil Ketkar and Jojo Moolayil. *Convolutional Neural Networks*, pages 197–242. Apress, Berkeley, CA, 2021. ISBN 978-1-4842-5364-9. doi: 10.1007/978-1-4842-5364-9_6. URL https://doi.org/10.1007/978-1-4842-5364-9_6.

[72] Witold Pedrycz and Shyi-Ming Chen. *Deep Learning: Concepts and Architectures*. Springer Publishing, New York, United States, 2019.

[73] Bart Kosko. Bidirectional associative memories. *IEEE Transactions on Systems, man, and Cybernetics*, 18(1):49–60, 1988.

[74] Fernando J Pineda. Recurrent backpropagation. *Backpropagation: Theory, architectures, and applications*, page 99, 1995.

[75] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001.

[76] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[77] Josh (Consultor) Patterson and Adam Gibson. *Deep learning : a practitioner's approach.* O'Reilly Media, 2017. ISBN 9781491914250. URL http://pbidi.unam.mx:8080/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat02025a&AN=lib.MX001001979816&lang=es&site=eds-live.

[78] Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, and Kazuhiko Ohe. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192, 2009.

[79] Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. Comprehend medical: a named entity recognition and relationship extraction web service.

In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851. IEEE, 2019.

[80] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[81] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[82] Explosion. SPACY'S ENTITY RECOGNITION MODEL: incremental parsing with bloom embeddings & residual CNNs, November 2017.

[83] Explosion. GitHub - Explosion/Floret: FastText + Bloom embeddings for compact, full-coverage vectors with SPACY. URL https://github.com/explosion/floret.

[84] Compact word vectors with Bloom embeddings · Explosion, 4 2022. URL https://explosion.ai/blog/bloom-embeddings.

[85] SpaCy: industrial-strength natural language processing (NLP) in python, .

[86] PySpark documentation — PySpark 3.3.2 documentation. https://spark.apache.org/docs/latest/api/python/, .

[87] doccano: Open source annotation tool for machine learning practitioners, .

[88] Wikipedia contributors. SpaCy, 1 2023. URL https://en.wikipedia.org/wiki/SpaCy.

[89] J. Brownlee. How to calculate precision, recall, and f-measure for imbalanced classification, 2020. URL https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/.

[90] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. volume Vol. 4304, pages 1015–1021, 01 2006. ISBN 978-3-540-49787-5. doi: 10.1007/11941439_114.

[91] Julian Ereth. Dataops-towards a definition. *LWDA*, 2191:104–112, 2018.

[92] J Sellén Crombet. Hipertensión arterial: diagnóstico, tratamiento y control. *Plaza de la Revolución*, 2008.

[93] What is a data warehouse? | IBM, . URL https://www.ibm.com/topics/data-warehouse.

[94] What is a data lake? https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake/, .

[95] What is AirflowTM? — Airflow documentation. URL https://airflow.apache.org/docs/apache-airflow/stable/index.html.