Instituto Tecnológico de Orizaba
División de Estudios de Posgrado e Investigación
Maestría en Sistemas Computacionales

TESIS

TÍTULO DEL PROYECTO: **Desarrollo de una aplicación para realizar la clasificación (predicción) de la aparición de padecimientos incapacitantes en pacientes con Diabetes Tipo II en la región de Orizaba.**

PRESENTADO POR:

I.S.C. Sahir Omar Burciaga Sosa M07011229

PARA OBTENER EL GRADO DE:

Maestro en Sistemas Computacionales

DIRECTOR DE TESIS:

MCE Beatriz A. Olivares Zepahua

Orizaba, Ver., México. 05 de Septiembre del 2013.

Índice General

Índice General ii
Índice de Tablas v
Índice de Figuras vii
Resumen viii
Abstract ix
Introducción x
Capítulo 1. Antecedentes 1
1.1 Marco teórico 1
1.1.1 Conceptos Médicos 1
1.1.1.1 Diabetes mellitus tipo II (DM) 1
1.1.1.2 Padecimientos incapacitantes en la diabetes mellitus tipo II 2
1.1.1.2.1 Retinopatía Diabética (RD) 2
1.1.1.2.2 Nefropatía diabética 3
1.1.1.2.3 Pie Diabético 4
1.1.3 Minería de Datos 5
1.1.3.1 Descubrimiento de conocimiento en base de datos 6
1.1.3.2 Clasificación y predicción 7
1.1.3.3 Modelo de Minería de Datos 8
1.1.3.4 Técnicas de Minería de Datos 8
1.1.3.4.1 Técnicas de Clasificación 8
1.1.3.4.2 Selección de Características 9
1.1.3.4.3 Balanceo de Clases 10
1.1.3.4.4 Evaluación de algoritmos 10
1.1.3.5 Herramienta de Minería de Datos Weka 11
1.1.4 Ontología 11
1.1.4.1 Lenguaje de Ontologías Web (*Web Ontology Language* OWL) 12
1.1.4.2 Editor de ontologías Protégé 13
1.1.4.3 Jena 13
1.2 Planteamiento del problema 14
1.3 Objetivo general y objetivos específicos 14
1.3.1 Objetivo General 15
1.3.2 Objetivos específicos 15

1.4 Justificación 15
Capítulo 2. Estado de la práctica 17
2.1 Trabajos relacionados 17
2.2 Análisis comparativo 25
2.3 Propuesta de solución 35
Capítulo 3. Aplicación de la metodología 37
Introducción 37
3.1 Arquitectura de la aplicación 37
Casos de uso 39
Diagrama de clases 40
3.2 Desarrollo de los modelos de minería de datos utilizando KDD 42
3.2.1 Selección de Datos 43
3.2.1.1 Población 43
3.2.1.1.1 Tamaño de la muestra 44
3.2.1.2 Técnica de recolección de datos 44
3.2.1.3 Aspectos éticos 44
3.2.1.4 Selección de Variables 44
3.2.2 Pre procesamiento 49
3.2.4 Minería de Datos e Interpretación de resultados 50
3.2.4.1 Minería de datos e interpretación de resultados para el modelo de nefropatía diabética 51
3.2.4.2 Minería de datos para el modelo de retinopatía diabética 54
3.2.4.3 Minería de datos para el modelo de pie diabético 59
3.3 Diseño de la ontología utilizando la metodología Methontology 62
3.3.1 Especificación 63
3.3.2 Conceptualización 63
3.3.2.1 Glosario de términos 64
3.3.2.2 Taxonomía de conceptos 68
3.3.2.3 Diagrama de relaciones binarias 69
3.3.2.4 Diccionario de conceptos 71
3.3.2.5 Relaciones binarias en detalle 71
3.3.2.6 Atributos de instancia en detalle 72
3.3.2.7 Reglas 73
3.3.2.8 Instancias 74

3.3.4 Implementación 74
Capítulo 4. Resultados 79
4.1 Caso de estudio: Predicción de nefropatía diabética en un paciente con DMTII 80
4.2 Caso de estudio: Predicción de retinopatía diabética en un paciente con DMTII 83
4.3 Caso de estudio: Predicción de pie diabético en un paciente con DMTII 86
4.4 Caso de estudio: Recomendación de tratamiento no farmacológico para un paciente con DMTII 89
4.5 Caso de estudio: Recomendación de tratamiento farmacológico para un paciente con DMTII 93
Capítulo 5. Conclusiones y Recomendaciones 95
5.1 Conclusiones 95
5.1 Recomendaciones 96
Referencias 98
ANEXO 1 103

Índice de Tablas

Tabla 1.1 Tipos de variables soportadas por el algoritmo C4.5. 9
Tabla 2.1 Tabla comparativa de artículos relacionados 26
Tabla 2.2 Propuesta de solución para el proyecto de tesis. 35
Tabla 3.1 Variables seleccionadas para la construcción de los modelos de MD. 46
Tabla 3.2 Distribución de la población 49
Tabla 3.3 Pre procesamiento de las columnas..... 50
Tabla 3.4 Distribución de la población para el modelo de nefropatía diabética 51
Tabla 3.5 Resultado de aplicar el algoritmo C4.5 sobre la población original con 32 atributos..... 52
Tabla 3.6 Resultado de aplicar el algoritmo C4.5 sobre la selección de características del evaluador CfsSubsetEval 53
Tabla 3.7 Resultados de aplicar el algoritmo C4.5 sobre la selección de características del evaluador *InfoGainAttributeEval*..... 53
Tabla 3.8 Resultados de aplicar el algoritmo C4.5 sobre una población de 6 atributos. 53
Tabla 3.9 Resultado de aplicar el algoritmo *Cost Sensitive* para mejorar la sensibilidad del algoritmo C4.5 sobre una población de 6 atributos 54
Tabla 3.10 Resultados de evaluar el modelo de nefropatía diabética elegido sobre la población original con 6 atributos..... 54
Tabla 3.11 Distribución de la población para el modelo de retinopatía diabética..... 55
Tabla 3.12 Resultado de aplicar el algoritmo *Cost Sensitive* para mejorar la sensibilidad del algoritmo C4.5 sobre una población de 13 atributos..... 56
Tabla 3.13 Resultados de aplicar el algoritmo C4.5 sobre una población sobre muestreada con 32 atributos..... 56
Tabla 3.14 Resultado de aplicar el algoritmo C4.5 sobre la selección de características del evaluador CfsSubsetEval 57
Tabla 3.15 Resultados de aplicar el algoritmo C4.5 sobre la selección de características del evaluador InfoGainAttributeEval..... 57
Tabla 3.16 Resultados de aplicar el algoritmo C4.5 sin poda para predecir retinopatía diabética sobre una población sobre muestreada de 12 atributos 58
Tabla 3.17 Resultados de aplicar el algoritmo *Cost Sensitive* para mejorar la especificidad del algoritmo C4.5 sin poda sobre una población sobre muestreada de 12 atributos..... 58
Tabla 3.18 Resultado de evaluar el modelo de retinopatía diabética sobre la población original con 12 atributos..... 58

Tabla 3.19 Distribución de la población para el modelo de pie diabético 59
Tabla 3.20 Resultados de aplicar el algoritmo *Cost Sensitive* para mejorar la sensibilidad del algoritmo C4.5 sobre una población con 11 atributos..... 60
Tabla 3.21 Resultados de aplicar el algoritmo C4.5 sobre una población sobre muestreada de 32 atributos..... 60
Tabla 3.22 Resultados de aplicar el algoritmo C4.5 sobre la selección de características del evaluador CfsSubsetEval..... 60
Tabla 3.23 Resultados de aplicar el algoritmo C4.5 para predecir pie diabético sobre la selección de características del evaluador *InfoGainAttributeEval*..... 61
Tabla 3.24 Resultado de aplicar el algoritmo *Cost Sensitive* para mejorar la especificidad del algoritmo C4.5 sin poda sobre una población con 15 atributos 61
Tabla 3.25 Resultados de evaluar el modelo de pie diabético sobre la población original de 15 atributos..... 62
Tabla 3.26 Número de términos de la ontología 65
Tabla 3.27 Glosario de términos 65
Tabla 3.28 Taxonomía de conceptos..... 68
Tabla 3.29 Diccionario de conceptos 71
Tabla 3.30 Detalle de relaciones binarias..... 71
Tabla 3.31Atributos de instancia en detalle 72
Tabla 3.32 Reglas de la ontología 73
Tabla 3.33 Instancias del dominio..... 74
Tabla 4.1 Resultado de la predicción de nefropatía diabética sobre 10 pacientes con DMTII. 82
Tabla 4.2 Resultado de la predicción de retinopatía diabética sobre 10 pacientes con DMTII. 85
Tabla 4.3 Resultado de la predicción de pie diabético sobre 10 pacientes con DMTII. 89

Índice de Figuras

Figura 1.1 Fases en el proceso de KDD. 6
Figura 3.1 Arquitectura de la aplicación. 38
Figura 3.2 Diagrama de casos de uso de la aplicación. 39
Figura 3.3 Diagrama de clases 40
Figura 3.4 Actividades realizadas en la fase de selección de datos 43
Figura 3.5 Fases de la metodología Methontology 62
Figura 3.6 Tareas de la fase de Conceptualización 64
Figura 3.7 Relación Paciente – Medición Sustancia 69
Figura 3.8 Relación Paciente – Hallazgo clínico 70
Figura 3.9 Relación Paciente – Farmacéutica producto biológico 70
Figura 3.10 Relación Paciente – Farmacéutica producto biológico - Intervención 70
Figura 3.11 Taxonomía de conceptos creada con Protégé. 75
Figura 3.12 Relaciones de la ontología creadas con la herramienta Protégé 76
Figura 3.13 Definición de atributos con la herramienta Protégé 77
Figura 3.14 Definición de individuos en la ontología con la herramienta Protégé 77
Figura 4.1 Inicio de sesión de usuarios 79
Figura 4.2 Pantalla de bienvenida de la aplicación 80
Figura 4.3 Pantalla de captura de variables para predecir nefropatía diabética. 81
Figura 4.4 Resultado de la predicción de nefropatía diabética 81
Figura 4.5 Parte del árbol de decisión del modelo de nefropatía diabética 82
Figura 4.6 Pantalla de captura de variables para predecir nefropatía diabética. 84
Figura 4.7 Resultado de la predicción de retinopatía diabética. 84
Figura 4.8 Parte del árbol de decisión del modelo de retinopatía diabética 85
Figura 4.9 Pantalla de captura de variables para predecir pie diabético 87
Figura 4.10 Resultado de la predicción de predecir pie diabético. 87
Figura 4.11 Parte del árbol de decisión del modelo de pie diabético 88
Figura 4.12 Pantalla de captura de variables para recomendar tratamiento no farmacológico. 90
Figura 4.13 Recomendaciones del tratamiento no farmacológico 92
Figura 4.14 Posibilidad de guardar la información en formato PDF 92
Figura 4.15 Pantalla de captura de variables para recomendar tratamiento farmacológico. 94
Figura 4.16 Posibles opciones de medicamentos sugeridos en el tratamiento farmacológico 94

Resumen

Las personas con diabetes son propensas a desarrollar complicaciones médicas, especialmente si no reciben el tratamiento adecuado de su enfermedad. Dentro de las complicaciones crónicas más comunes en los diabéticos se encuentran la retinopatía diabética, la nefropatía diabética y el pie diabético que son algunas de las principales causas de ceguera, daño renal y necesidad de amputación. Una de las dificultades en la detección de estos padecimientos es que muchas veces se presentan de manera asintomática, incluso en etapas avanzadas cuando las opciones de tratamiento ofrecen pocas garantías de mejorar la calidad de vida del paciente, por lo que la mejor vía para combatir estas enfermedades es la prevención.

La alta prevalencia y gravedad de estas complicaciones sugiere la necesidad de crear una aplicación que ofrezca a los médicos una alternativa computacional en la identificación y tratamiento de estas enfermedades. Es por eso que esta tesis propone el desarrollo de una aplicación Web que, utilizando Minería de Datos, realice la predicción de la aparición de complicaciones en pacientes con Diabetes Tipo II en la región de Orizaba, además de la generación de una Ontología que provea de cuidados preventivos y recomendaciones farmacológicas para evitar o retrasar la aparición de estas complicaciones.

Con el desarrollo de la aplicación, se les proporcionaría a los profesionales de la salud una alternativa computacional a los métodos convencionales existentes en la identificación de estas enfermedades y además se ayudaría en el proceso de toma de decisiones, como por ejemplo reforzar un diagnóstico médico o identificar pacientes en potencial riesgo de padecer estas complicaciones. De igual forma, con la herramienta el médico obtendría rápidamente recomendaciones personalizadas en el tratamiento farmacológico y no farmacológico del paciente, intentando así proporcionar un tratamiento integral para su enfermedad y mantenerlo en buen estado de salud para retrasar la aparición de complicaciones.

Abstract

People with diabetes are susceptible to develop medical complications, especially if they don’t receive properly treatment for their illness. Among the most common chronic complications in diabetics are the diabetic retinopathy, the diabetic nephropathy and the diabetic foot, which are a major cause of blindness, kidney damage and the need for amputation. One difficulty in detecting these disorders is that sometimes they occur without symptoms, even in advanced stages of the illness, when treatment options offer little assurance to improve the life quality of patients, so the best way to combat these diseases is prevention.

The high prevalence and severity of these complications suggest the need to create an application that offers doctors a computational alternative in the identification and treatment of these diseases. That's why this thesis proposes the development of a Web application that, using Data Mining, performs the occurrence prediction of complications in patients with Diabetes Type II in the region of Orizaba; in addition to that, generate an Ontology that provides preventive care and pharmacological recommendations to prevent or delay the occurrence of these complications.

With the development of this application, it would be provided to healthcare professionals a computational alternative to existing conventional methods in the identification of these diseases and also help in the process of decision making, such as reinforcing a medical diagnosis or identify patients at potential risk of these complications. Similarly, with this tool the doctor would quickly get customized pharmacological and non-pharmacological treatment of the patient, trying to provide a comprehensive treatment for their illness and keep him in good health to delay the onset of complications.

Introducción

La diabetes mellitus, o simplemente diabetes, es una enfermedad que representa un grave problema de salud en México y en el mundo. La Encuesta Nacional de Salud y Nutrición (ENSANUT) en el 2012 reveló que el 9% de la población mexicana adulta padece esta enfermedad [1] y que se estima que esta cantidad probablemente sea del doble debido a que muchas personas aún no son diagnosticadas. Por otro lado, el Instituto Nacional de Estadística, Geografía e Informática (INEGI), en el censo llevado a cabo en 2010, reveló que la diabetes es la responsable del 14% de las muertes en el país representando la segunda causa de mortalidad en México [2].

Dentro de los padecimientos incapacitantes más comunes en los diabéticos se encuentran las complicaciones micro vasculares: retinopatía diabética, nefropatía diabética y pie diabético. La nefropatía diabética (ND) es una de las complicaciones más temidas de la diabetes y es la responsable de la mayoría de los casos de insuficiencia renal crónica terminal. Por su parte la retinopatía diabética representa una de las principales causas de ceguera y el pie diabético una de las principales causas de amputación. El monto financiero que emplea el sector salud para controlar dichas complicaciones es muy alto, principalmente en el caso del manejo de nefropatía diabética

Los organismos de salud pública buscan constantemente la implementación de Tecnologías de Información y Comunicaciones (TICs) para facilitar la gestión, el seguimiento y el diagnóstico temprano de pacientes; entre esas tecnologías se encuentran técnicas de Minería de Datos que permiten analizar la información y encontrar patrones de comportamiento en los datos y Ontologías que permiten representar conocimiento dentro de un dominio en particular mediante un conjunto de conceptos y sus relaciones.

Bajo este enfoque, en este trabajo se propone el desarrollo de una aplicación Web que, utilizando técnicas de Minería de Datos, prediga la aparición de complicaciones en pacientes con Diabetes Tipo II en la región de Orizaba y mediante el uso de una ontología

se infieran recomendaciones para el tratamiento no farmacológico y farmacológico del paciente diabético con el fin de retrasar/prevenir la aparición de dichas complicaciones.

El resto del trabajo se organiza de la siguiente manera: en el capítulo 1 se muestran los elementos conceptuales que conforman los fundamentos base de la tesis, la problemática que originó su elaboración, los objetivos y justificación. En el capítulo 2 se realiza una recopilación y análisis de los trabajos relacionados y se describe la propuesta de solución para el problemática que se plantea. En el capítulo 3 se muestra la metodología seguida para el desarrollo del proyecto, la arquitectura de la aplicación, la construcción de los modelos de minería de datos y de la ontología. En el capítulo 4 se muestran los resultados obtenidos tras el desarrollo de la aplicación Web y finalmente en el capítulo 5 se presentan las conclusiones y recomendaciones.

Capítulo 1. Antecedentes

En este capítulo se encuentran los elementos conceptuales que permiten conocer los fundamentos base de la tesis; de igual manera, se menciona la problemática que originó su elaboración. Así como los objetivos y la justificación de la realización de la tesis.

1.1 Marco teórico

En este apartado se describen los conceptos básicos necesarios para comprender de una mejor manera el presente proyecto de tesis.

1.1.1 Conceptos Médicos

A continuación se describen los conceptos médicos relevantes que se involucran en este tema de tesis.

.

1.1.1.1 Diabetes mellitus tipo II (DM)

La diabetes mellitus es un conjunto de enfermedades que da como resultado defectos en la secreción de insulina, por lo que el cuerpo no regula la cantidad de azúcar en la sangre. Se caracteriza por la presencia de hiperglucemia, es decir, el aumento de los niveles de glucosa en sangre. Existen dos tipos de Diabetes: Tipo I que es la incapacidad del cuerpo para producir insulina y requiere que la persona se aplique inyecciones de insulina periódicamente y la Diabetes Mellitus o Tipo II que es una condición en la que el cuerpo no usa adecuadamente la insulina que produce, a veces se presenta combinada con un déficit en la secreción de la misma. En este tipo de diabetes se presentan dos situaciones: o el cuerpo no produce suficiente insulina o las células ignoran a ésta. Es el tipo más común en todo el mundo, representa entre el 90 y el 95% de la población diabética [3] [4] [5].

Algunos de los síntomas que se relacionan con la presencia de la diabetes son el aumento del apetito, poliuria, pérdida de peso, cansancio y sed, así como la existencia de antecedentes familiares de diabetes, presencia de obesidad o problemas cardiacos entre otros [3]. Dados síntomas tan variados, no existe un factor externo que garantice la presencia de la enfermedad, por lo que es necesario para su diagnóstico recurrir a una

evaluación clínica y análisis de laboratorio que consiste principalmente en la medición del nivel de glucosa en sangre [3]. Es por esto que el diagnóstico de la diabetes muchas veces se efectúa varios años después de la aparición de la enfermedad, cuando ésta se encuentra más avanzada y el peligro de complicaciones como ceguera o insuficiencia renal es mayor [5] [6].

1.1.1.2 Padecimientos incapacitantes en la diabetes mellitus tipo II

Las personas que padecen diabetes mellitus tienen una alta probabilidad de presentar diversas complicaciones que, a la larga, provocan daño en diversos órganos del cuerpo como son los ojos, riñones y piel. Las complicaciones de la diabetes son causa de gran parte de la morbilidad y mortalidad que acompañan a este trastorno y se dividen en micro vasculares (retinopatía, neuropatía y nefropatía) y macro vasculares (enfermedad vascular periférica y enfermedad vascular cerebral) [7]. El daño causado al organismo por la presencia de la hiperglucemia a lo largo del tiempo (15 a 20 años) es una de las principales causas de la aparición de estas enfermedades. Sin embargo, algunos pacientes, a pesar de mantener una diabetes prolongada, nunca sufren de neuropatía, retinopatía o retinopatía, pese a que su control de la glucemia no es diferente a aquéllos que sí desarrollan complicaciones, por lo que se sospecha que existe un factor genético que hace más vulnerable a algunas personas a desarrollar complicaciones. Varios estudios realizados demuestran que mantener un buen control de la glucemia disminuye la frecuencia, evolución y severidad de la complicaciones micro vasculares en personas con diabetes tipo I y II [7]. A continuación se mencionan algunas de las complicaciones más comunes de los diabéticos y en las cuales hará énfasis en este trabajo.

1.1.1.2.1 Retinopatía Diabética (RD)

La retinopatía diabética es una complicación ocular de la diabetes que se origina por el deterioro de los vasos sanguíneos que irrigan la retina del fondo del ojo. Los pacientes que presentan diabetes por un largo periodo de tiempo, aproximadamente 15 años, poseen un 60% de probabilidades de presentar vasos sanguíneos dañados y un 21% se presentar retinopatía al momento del diagnóstico. La retinopatía diabética es una de las causas más frecuente de ceguera entre los adultos [3].

La retinopatía diabética es asintomática, es decir, es posible padecer retinopatía diabética durante mucho tiempo sin presentar ningún síntoma. Por lo general no hay dolor ni pérdida de la visión hasta el momento en que ocurre un daño significativo y se producen complicaciones. Algunos de los síntomas de la retinopatía diabética incluyen visión doble o borrosa, dificultad para leer, manchas en la visión, pérdida parcial o total de la visión, dolor, presión o enrojecimiento constante del ojo [8].

La detección de la retinopatía diabética se hace con una fotografía de retina mediante una cámara no midriática o con una oftalmoscopia a través de pupila dilatada. Ambos procedimientos los realizan e interpretan personal con entrenamiento específico [9]. Dentro de los cuidados preventivos que un paciente diabético toma para reducir la probabilidad de pérdida de la visión originada por la retinopatía diabética y sus complicaciones se encuentran [10]: mantener los niveles de azúcar dentro de los límites mediante el consumo de una dieta saludable, el chequeo constante de la glucosa, el ejercicio físico regular, la toma de insulina o medicamentos en caso de ser necesario, control de presión arterial, no fumar ya que agrava muchos de los otros problemas de salud que enfrentan las personas con diabetes y el control de la dislipidemia (concentración anormal de lípidos en sangre).

1.1.1.2.2 Nefropatía diabética

La nefropatía diabética es una de las complicaciones más temidas de la diabetes. Tanto por el costo económico de su tratamiento como por el impacto en el bienestar del paciente diabético. La nefropatía diabética es la causa principal de insuficiencia renal crónica en México y en muchos otros países latinoamericanos y es responsable de la mayoría de los casos de insuficiencia renal crónica terminal. La nefropatía diabética se origina principalmente por la presencia de hiperglicemia sostenida por un largo tiempo (10 a 15 años) [11].

Los factores de riesgo asociados a progresión de la nefropatía diabética se dividen en dos, los modificables y los no modificables. Dentro de los no modificables se encuentran la edad, etnia, características genéticas, historia familiar de nefropatía, bajo peso al nacer,

disminución del número de nefronas (congénita o adquirida) y el tiempo de evolución de la DM [3]. Dentro de los modificables están el mal control de la glucemia o de la presión arterial, presencia de proteinuria, obstrucción, reflujo o infección del tracto urinario, uso de antiinflamatorios no esteroideos y otras nefrotoxinas, anemia, dislipidemia, tabaquismo y sobrepeso u obesidad [3].

Inicialmente la nefropatía diabética no tiene síntomas [12], éstos se desarrollan en etapas finales y se originan por la excreción de grandes cantidades de proteína en la orina o debido a insuficiencia renal. Algunos de los síntomas son: hinchazón, por lo general alrededor de los ojos en las mañanas; después hinchazón general del cuerpo como la hinchazón de las piernas, apariencia espumosa en la orina causada por la proteinuria, aumento de peso involuntario por acumulación de líquido, anorexia, náuseas y vómitos, sensación de malestar general, fatiga, dolor de cabeza e hipo frecuente.

La nefropatía está presente entre el 10 y el 25% de los pacientes con diabetes mellitus tipo II al momento del diagnóstico [12]. La detección de la nefropatía diabética se basa en la medición de la microalbuminuria, de la proteinuria y de la filtración glomerular. El control estricto de la glucemia así como de la presión arterial es indispensable para evitar la progresión del daño renal. La meta a alcanzar es de una hemoglobina glucosilada <7.0% y presión arterial<130/80. La reducción a <1.2 g/día del sodio en la dieta disminuye la presión arterial, pero la adherencia a esta dieta es difícil de lograr. Al igual que en la retinopatía, el hábito de fumar se asocia con mayores riesgos de desarrollar esta enfermedad. El sobrepeso y la obesidad también son factores de riesgo por lo que se recomienda mantener una dieta adecuada para controlar tanto el peso como los niveles de colesterol y triglicéridos. Estas medidas son más efectivas cuando se aplican en forma temprana con monitoreo de las metas establecidas para su cumplimiento [13].

1.1.1.2.3 Pie Diabético

El pie diabético es el trastorno ulcerativo de los pies que se produce por el daño en nervios y arterias del pie ocasionados por la hiperglucemia sostenida. Las lesiones, la mayoría de las veces, aparecen sin presencia de dolor, lo cual empeora la situación ya que el paciente

no acude a tiempo a un centro de salud. Un diagnóstico precoz y un tratamiento adecuado otorgan grandes posibilidades de mantener la integridad del pie evitando así su amputación [3]. Al menos 15% de los diabéticos presentará ulceraciones en el pie durante su vida. Se estima que 85% de los diabéticos que sufre amputaciones previamente ha padecido una úlcera [12].

Los factores de riesgo para desarrollar úlceras o amputación en pacientes con diabetes mellitus son: insuficiencia arterial, deformidad de pie con evidencias de sitios de presión, callosidad plantar severa, patología de uñas, historia previa de úlcera o amputación, neuropatía periférica, infección, enfermedad vascular periférica, calzado inapropiado [9] [12] [13]. Otros factores importantes de considerar son la duración de la diabetes mayor a 10 años, el mal control metabólico determinado por la hemoglobina glucosilada >7.0%; sexo masculino; presencia de retinopatía diabética, neuropatía diabética, tabaquismo o consumo de alcohol.

Los síntomas más frecuentes del pie diabético son pies fríos, dolor en reposo, piel seca, debilidad muscular, deformaciones y aparición de callos plantares entre otros. Para su diagnóstico se realiza un examen físico que incluye el estado de la piel (temperatura, color, resequedad, fisuras, eritema, zonas anormales de presión, edema), así como un examen vascular que incluye prueba de llenado capilar, presencia de pulsos tibial posterior, pedio, poplíteo y femoral, prueba de retorno venoso, uso del diapasón (128 Hertz), monofilamento y valoración del calzado [13].

Es posible evitar las lesiones del pie diabético siguiendo algunas medidas de autocuidado de los pies como lavado diario, cuidado de las uñas, uso de calzado apropiado y buena hidratación entre otros [13] .

1.1.3 Minería de Datos

En años recientes la Minería de Datos ha despertado un gran interés en el sector industrial y de negocios debido a la existencia y manejo de grandes cantidades de datos por parte de las empresas y a la necesidad de convertir estos datos en información valiosa. Mehmed

Kantardzic define a la Minería de Datos como "el proceso para encontrar información nueva, valiosa y no trivial en grandes volúmenes de datos" [14]. En otras palabras, la Minería de Datos permite extraer conocimiento útil y desconocido que se encuentra oculto en los datos.

La Minería de Datos se aplica en una gran variedad de tipos de repositorios como son base de datos relacionales, almacenes de datos, archivos planos, secuencias de datos y *World Wide Web* entre otros [15]. Unas de las tareas principales de la Minería de Datos son la predicción y la clasificación [15], más adelante se explica cada una de ellas.

1.1.3.1 Descubrimiento de conocimiento en base de datos

El Descubrimiento de Conocimiento en Base de Datos, mejor conocido por sus siglas en inglés KDD (*Knowledge Discovery from Data*), es el proceso de extracción de información no trivial, útil y previamente desconocida de los datos [16]. En ocasiones algunos autores utilizan el término de Minería de Datos y Descubrimiento de Conocimiento en Base de Datos como sinónimos, pero existe una clara diferencia entre ellos. KDD es un proceso que posee una seria de fases que permiten el descubrimiento de conocimiento y Minería de Datos se refiere a una determinada fase en este proceso [17].

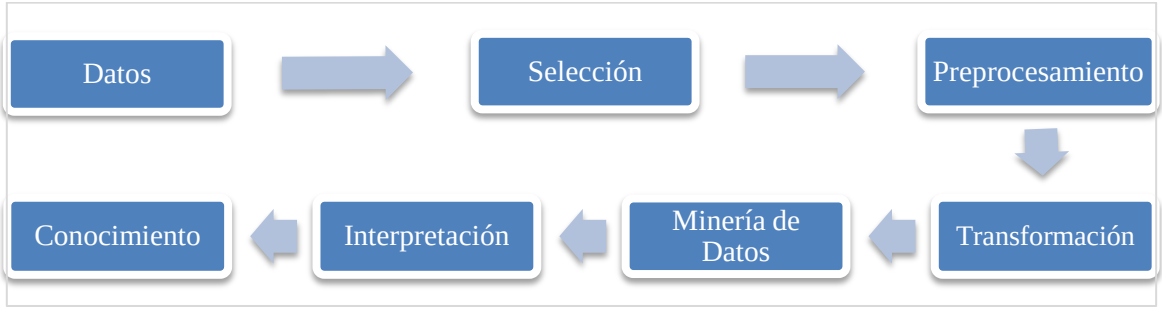


Figura 1.1 Fases en el proceso de KDD.

En la Figura 1.1 se muestran las fases correspondientes al proceso de KDD. Como primer paso se identifican las fuentes de **Datos** de donde se obtendrá la información, entre las más comunes se encuentran las bases de datos, documentos, archivos o almacenes de datos entre otros. Posteriormente se realiza la **Selección** de las variables relevantes, es decir, si se

deseara conocer si una persona padece obesidad, el peso, la edad y la altura serían las variables a identificar. Después se hace un **Pre procesamiento** de los datos, en el cual se detecta y trata la presencia de valores atípicos, faltantes o perdidos y se eliminan datos erróneos e irrelevantes. Si es necesario se realiza la **Transformación**, que consiste principalmente en la aplicación de técnicas de reducción, discretización, normalización y escalado simple entre otras. En la fase de **Minería de Datos** se decide el algoritmo que se va a utilizar, ya sea descriptivo, predictivo o ambos. Después, en la evaluación e **Interpretación**, se analizan los resultados y, si los resultados no corresponden a la experiencia del experto, se regresa a las etapas anteriores para reajustar la información y realizar una nueva iteración. Finalmente, después de realizar todas las etapas, se obtiene la generación de nuevo **Conocimiento** [18].

1.1.3.2 Clasificación y predicción

La clasificación y predicción permiten analizar los datos y extraer modelos que ayudan a predecir el comportamiento de los datos. Por ejemplo, si en un banco desean predecir si un cliente con un perfil determinado está interesado en adquirir un préstamo, en esta situación la tarea de análisis de datos que se utilizaría es la de **clasificación** debido a que se construye un modelo para predecir una variable categórica clase como “Sí interesado” o “No interesado” las cuales se representan con valores discretos. Por el contrario, si en el banco quisieran predecir cuánto dinero otorgar a un cliente determinado para minimizar riesgos de sufrir una pérdida, la tarea de análisis sería una **predicción** donde se construiría un modelo para encontrar valores continuos [15]. En otras palabras, la clasificación se utiliza si se desean obtener valores discretos y la predicción en el caso de valores continuos; sin embargo, algunos autores e incluso herramientas como las de Microsoft, usan solamente el término predicción sin importar el tipo específico de valor a buscar.

1.1.3.3 Modelo de Minería de Datos

Cuando se habla de un modelo de Minería de Datos se refiere no solamente a un algoritmo, sino a un conjunto de datos, estadísticas y patrones que se utilizan para realizar las tareas de Minería de Datos como predicciones o clasificaciones [19]. Una estructura de Minería de Datos contiene la información que representa el origen de los datos. El modelo de Minería de Datos obtiene los datos a partir de esta estructura y posteriormente la analiza y procesa usando un algoritmo de Minería de Datos. Una vez que se concluye este proceso, se almacenan los resultados, patrones, metadatos y enlaces de la estructura de Minería de Datos en el modelo. De esta manera es posible construir diversos modelos con distintos algoritmos y estructura de datos obteniendo como resultado un conjunto diferente de metadatos y patrones que permitirán realizar predicciones con diversos resultados [19].

1.1.3.4 Técnicas de Minería de Datos

Existen diversas técnicas de Minería de Datos que se desarrollaron y probaron a través de los años en diferentes proyectos. A continuación se describen algunas de ellas.

1.1.3.4.1 Técnicas de Clasificación

Los algoritmos de minería de datos supervisados son aquellos que predicen el valor de un atributo clase con base en un conjunto de datos conocido, es decir, funcionan para la clasificación o para la predicción. A partir de los valores donde la clase se conoce se realiza una relación entre la clase y los otros valores y es por medio de estas relaciones que se realiza la predicción sobre una clase desconocida. A esto se le conoce como aprendizaje supervisado y se lleva a cabo en dos etapas: entrenamiento y prueba. El entrenamiento consiste en la construcción de un modelo usando algún algoritmo de Minería de Datos sobre un conjunto de datos con una clase conocida y las pruebas consisten en evaluar al modelo con el resto de los datos [20]

Los Árboles de Decisión constituyen uno de los algoritmos de minería de datos más populares debido a que representan el conocimiento en forma de árboles que resultan intuitivos y fáciles de comprender para las personas. Son como un diagrama de flujo en forma de estructura de árbol donde cada hoja representa un atributo y cada rama un camino

distinto para obtener la predicción. Generalmente son simples y rápidos y tienen buen nivel de precisión [15].

El algoritmo C4.5 genera arboles de decisión, admite atributos nominales, numéricos y nulos aunque la variable clase es obligatoriamente nominal. Utiliza la ganancia de información y la medición de entropía para seleccionar los atributos en cada nodo del árbol, además de utilizar la estrategia de poda para reducir los errores en la clasificación. La poda consiste en reemplazar un subárbol de un nodo por una hoja o una de sus ramas con la clasificación más común. En la Tabla 1.1 se observan algunas características del algoritmo C4.5 [21].

Tabla 1.1 Tipos de variables soportadas por el algoritmo C4.5.

Algoritmo	Nombre en WEKA	Nominal	Numérico	Valores Nulos
C4.5	J48	Sí	Sí	Sí

1.1.3.4.2 Selección de Características

Las técnicas de Selección de características permiten elegir las variables más relevantes en un conjunto de datos para usarlas posteriormente en la construcción de un modelo de Minería de Datos. Existen diversas técnicas de selección de características que permiten evaluar tanto el nivel predictivo de las variables de manera individual como de subconjuntos de atributos. El evaluador de características por ganancia de información (*InfoGainAttributeEval*) es un algoritmo que evalúa la capacidad de predicción individual de cada atributo mediante la medición de la ganancia de información con respecto a la variable clase, esta técnica se usa en conjunto con el método *Ranker* que ordena a los atributos de acuerdo al resultado obtenido en su evaluación [21]

El evaluador de subconjunto de características (*CfsSubsetEval*) es un algoritmo que evalúa a un subconjunto de atributos considerando la capacidad de predicción individual de cada uno junto con el grado de redundancia entre ellos, esta técnica se usa en conjunto con el método *BestFirst* que busca el espacio de subconjuntos de atributos por *hill climbing* codicioso y *backtracking* [21].

1.1.3.4.3 Balanceo de Clases

Un conjunto de datos no está balanceado si las clases no son aproximadamente equitativas, es decir, existe una clase con mayor número de instancias (clase mayoritaria) mientras que en la otra existen muy pocos elementos (clase minoritaria), esta situación genera un problema en las tareas de clasificación ya que la mayoría de los clasificadores tienden a trabajar con la clase mayoritaria e ignorar la clase minoritaria [22].

El algoritmo SMOTE (*Synthetic Minority Oversampling Technique*) es una técnica de sobre muestreo que genera instancias artificiales con el objetivo de balancear la muestra de datos, se basa en la regla del vecino más cercano donde para cada instancia minoritaria se buscan sus k instancias minoritarias vecinas y se crean N nuevas instancias entre la instancia original y sus k vecinas [22]

El algoritmo *Cost Sensitive* es una meta algoritmo que, a partir de una matriz de costos, permite indicarle a un algoritmo base (como el algoritmo C4.5) que asigne mayor importancia en clasificar una clase en particular [21].

1.1.3.4.4 Evaluación de algoritmos

Existen diversas técnicas que permiten probar la precisión de un modelo de Minería de Datos:

- Conjunto de entrenamiento: Consiste en utilizar todo el conjunto de entrenamiento para probar qué tan bien predice la variable clase el modelo.
- Porcentaje de división: Consiste en probar qué tan bien predice el modelo un cierto porcentaje de los datos del conjunto de entrenamiento.
- Suministrar Conjunto de pruebas: Consiste en entrenar al modelo de minería de datos con una población y suministrar otro conjunto de datos para realizar pruebas.
- Técnica de validación cruzada: Divide el conjunto de datos en k segmentos de tamaño semejante (normalmente 10, donde cada segmento es representativo de la población original) y se aplican k iteraciones de tal manera que, en cada iteración,

se selecciona 1 segmento para pruebas y el resto (k-1) para entrenamiento. A final se promedian los resultados que se obtuvieron en cada iteración [21].

1.1.3.5 Herramienta de Minería de Datos Weka

Weka es un software de Minería de Datos de código abierto desarrollado en Java en la Universidad de Waikato en Nueva Zelanda y tiene como objetivo proporcionar una amplia colección de algoritmos de aprendizaje automático y pre procesamiento de datos que sirva como herramientas a los investigadores y profesionales por igual [14]. Permite a los usuarios probar rápidamente y comparar diferentes técnicas de aprendizaje automático en nuevos conjuntos de datos. Su arquitectura modular y extensible permite sofisticados procesos de Minería de Datos que se construyen a partir de la amplia colección de algoritmos de aprendizaje y herramientas que proporciona; entre ellos se incluyen algoritmos para la regresión, clasificación, agrupamiento, la minería de reglas de asociación y la selección de atributos. Proporciona una API para Java lo cual permite realizar diversas tareas de minería de datos directamente en las aplicaciones.

1.1.4 Ontología

Una de las definiciones de ontologías más aceptada es la propuesta por Gruber en 1993, que define a una ontología como "una especificación explícita de una conceptualización". Posteriormente, en 1997, Brost define a una ontología como "una especificación formal de una conceptualización compartida" y en 1998 Studer fusiona estas dos definiciones e indica que "una ontología es una especificación formal y explícita de una conceptualización compartida". **Especificación formal** se refiere a que la ontología es legible para la computadora. **Conceptualización** es un modelo, una abstracción del mundo que se desea representar para algún propósito. **Explícita** se refiere a que el modelo está bien definido, para ello se hace uso de relaciones, funciones, restricciones, axiomas y conceptos. Y finalmente **compartida** ya que captura conocimiento aceptado por un grupo. Una ontología representa conocimiento dentro de un dominio en particular mediante un conjunto de conceptos y sus relaciones [20]. Una de las ventajas de utilizar ontologías es que resultan muy útiles para inferir conocimiento a partir de sus conceptos y sus relaciones. Por ejemplo supóngase que en una ontología existe el concepto "persona" y "abuelo" y se tiene que una

persona “x” es padre de otra persona “w” y a su vez “w” es padre de una persona “z” bajo este contexto una ontología es capaz de deducir que la persona “x” es abuelo de “z”. Los componentes principales de una ontología son los conceptos, relaciones, instancias y axiomas [5].

Concepto: representa un conjunto o grupo de entidades o “cosas” dentro de un dominio. Por ejemplo, dentro de un ámbito educativo, un profesor y un alumno son conceptos.

Relación: describe la interacción entre dos conceptos. Es posible presentarlos de dos maneras: taxonómicas que organizan los conceptos de forma jerárquica, o asociativas que relacionan a dos conceptos que no están en la misma jerarquía.

Instancia: representa un objeto determinado de un concepto. Por ejemplo una instancia del concepto Alumno es “Juan”.

Axioma: restringe los valores de los conceptos o instancias, permiten inferir nuevo conocimiento y ayudan a verificar la consistencia de la ontología.

1.1.4.1 Lenguaje de Ontologías Web (*Web Ontology Language* OWL)

Es el lenguaje de ontologías para la creación de ontologías o bases de conocimiento recomendado por el W3C. OWL se utiliza cuando la información contenida en los documentos necesita procesarse por las aplicaciones, en comparación con situaciones en las que el contenido se requiere presentar a seres humanos [6] . OWL se utiliza para representar de forma explícita el significado de los términos en vocabularios y las relaciones entre esos términos. OWL ofrece tres versiones con diferentes niveles de expresividad.

OWL Lite es útil para aquellos usuarios que necesitan sobre todo una jerarquía de clasificación y restricciones simples. Por ejemplo, mientras que soporta las restricciones de cardinalidad, sólo permite valores de cardinalidad de 0 o 1. OWL Lite proporciona una ruta

de migración rápida para tesauros y otras taxonomías. OWL Lite también tiene una complejidad formal menor que OWL DL.

OWL DL les sirve a usuarios que quieren la máxima expresividad conservando todas las conclusiones que garantizan que sea computable y que todos los cálculos terminarán en tiempo finito. OWL DL incluye todas las construcciones del lenguaje OWL, pero bajo ciertas restricciones. OWL DL se llama así debido a su correspondencia con lógicas descriptivas.

OWL Full está diseñado para los que quieren la máxima expresividad y la libertad sintáctica de RDF sin garantías computacionales. Es poco probable que cualquier software de razonamiento sea capaz de apoyar el razonamiento completo para todas las características de OWL Full.

1.1.4.2 Editor de ontologías Protégé

Protégé es una plataforma libre y de código abierto con un conjunto de herramientas para la construcción de modelos de dominio y aplicaciones para base de conocimiento con ontologías [23]. Contiene un gran conjunto de estructuras de modelado de conocimiento y acciones que apoyan la creación, visualización y manipulación de ontologías en diversos formatos de representación. Protégé permite la definición de clases, jerarquía de clases, variables, restricciones de valor variable, relaciones entre las clases y las propiedades de estas relaciones.

1.1.4.3 Jena

Jena que es un marco de trabajo de código abierto para construir aplicaciones Web semántica, incluye una API para Java que permite leer y procesar ontologías escritas en lenguaje OWL. Además, permite realizar consultas a una ontología en lenguaje SPARQL y cuenta con un motor de inferencias basado en reglas con lo que es posible inferir nuevo conocimiento a partir de una ontología [24].

1.2 Planteamiento del problema

En años recientes la incidencia de la diabetes en la población mexicana aumentó de forma alarmante [25] [6] [23] [26]. La diabetes es una enfermedad que no mata por sí misma pero que detona en otros problemas que afectan la calidad de vida del enfermo que, en determinados casos, desembocan en una muerte prematura.

Las personas con diabetes son muy propensas a desarrollar complicaciones, especialmente si no tienen un tratamiento adecuado de su enfermedad. Dentro de las complicaciones crónicas más comunes en los diabéticos son la retinopatía, la nefropatía y el pie diabético. El monto financiero que emplea el sector salud para controlar dichas complicaciones es muy alto [26], principalmente en el caso del manejo de nefropatía diabética.

Las complicaciones de la diabetes generalmente se presenten de forma asintomática, es decir, el paciente no siente dolor, malestar o síntoma que lo percate de la presencia de la enfermedad incluso en etapas avanzadas cuando el riesgo de insuficiencia, ceguera o amputación es mayor y las opciones de tratamiento ofrecen pocas garantías en mejorar la calidad de vida del paciente.

Por otro lado, los estudios formales reportados en México se centran en el análisis estadístico de la presencia de la enfermedad [25] [6] [23] y las complicaciones que de ella derivan, como son la gestión de tratamientos de largo plazo, la necesidad de unidades médicas especializadas, los costos de hospitalización y el aumento de pensiones por incapacidad entre otros [27]. Sin embargo, como se menciona en [28], en el área de la salud existe una gran cantidad de información relacionada con las enfermedades, en este caso la diabetes tipo II y los padecimientos de ella derivados, que con el uso de técnicas de minería de datos permitiría obtener predicciones sobre la aparición de posibles complicaciones y así realizar intervenciones tempranas para prevenir o retrasar su aparición.

1.3 Objetivo general y objetivos específicos

A continuación se presentan el objetivo general y los objetivos específicos del proyecto de innovación.

1.3.1 Objetivo General

Desarrollar una aplicación Web que, utilizando Minería de Datos, realice la clasificación (predicción) de la aparición de padecimientos incapacitantes en pacientes con Diabetes Tipo II en la región de Orizaba y la generación de una ontología que provea de cuidados preventivos y recomendaciones farmacológicas para retrasar/evitar su aparición.

1.3.2 Objetivos específicos

- Identificar, apoyado de un médico experto, las variables que se involucran en la aparición de padecimientos incapacitantes en pacientes con Diabetes Tipo II.
- Construir y validar el modelo de Minería de Datos entrenado, con el apoyo de un médico experto, siguiendo la metodología KDD.
- Construir, con el apoyo de un médico experto, la ontología de cuidados preventivos para retrasar/prevenir la aparición de padecimientos incapacitantes en pacientes con Diabetes Tipo II.
- Desarrollar el módulo para la gestión de datos y para la consulta de la clasificación de la aparición de padecimientos incapacitantes en pacientes con Diabetes Tipo II en la región de Orizaba.
- Desarrollar el módulo para la gestión y consulta de la ontología de cuidados preventivos para retrasar/prevenir la aparición de padecimientos incapacitantes en pacientes con Diabetes Tipo II.

1.4 Justificación

Debido a la gran prevalencia de la diabetes en México y al aumento en la incidencia de padecimientos como la ceguera o la insuficiencia renal derivadas de esta, la necesidad de crear una aplicación para la identificación y tratamiento preventivo de estas enfermedades es importante.

Con el desarrollo de esta aplicación, mediante el uso de técnicas de Minería de Datos, se ayudaría a la identificación de posibles complicaciones (retinopatía, nefropatía, pie diabético) en los pacientes con Diabetes Mellitus Tipo II y se les proporcionaría a los

profesionales de la salud una alternativa computacional a los métodos convencionales existentes en la identificación de estas enfermedades.

Por otra parte, la aplicación, mediante el uso ontologías, sería capaz de inferir conocimiento y sugerir recomendaciones para el tratamiento de un paciente en particular, con el objetivo de mantenerlo en buen estado de salud y así evitar o retrasar la aparición de complicaciones, con lo que se lograría disminuir el número de admisiones a hospitales por posibles complicaciones y se reducirían los costos en el manejo de estas enfermedades que representan una gran carga económica para el individuo y los sistemas de salud en el país.

De igual forma es importante señalar que, tomando como referencia el gran impacto que tiene la diabetes en México, es necesario hacer todos los esfuerzos posibles en cada uno de los diferentes sectores con el objetivo de minimizar el problema de representa esta enfermedad.

Capítulo 2. Estado de la práctica

Existen múltiples investigaciones donde se hace uso de técnicas de Minería de Datos y Ontologías para obtener información relevante que ayude a combatir la enfermedad de la diabetes que representa un problema de salud importante que afecta a una gran cantidad de personas en distintos países [5] y que repercute en la estabilidad de los sistemas de salud y en la sociedad en general.

En este capítulo se realiza una recopilación y análisis de los trabajos relacionados con el presente proyecto de tesis de manera que se conozcan las técnicas, tecnologías y formas de desarrollo que se llevan a cabo actualmente en este campo. En los siguientes apartados se presentan algunos resúmenes de los artículos más relevantes y una tabla comparativa de los trabajos analizados.

2.1 Trabajos relacionados

En [28] analizaron la aplicación de técnicas de clasificación de minería de datos, como árboles de decisión, redes neuronales y algoritmos basados en reglas, al enorme volumen de datos de salud; en particular se realizó un caso de estudio sobre un conjunto de datos médicos de niños con diabetes mellitus (tipo II) y diabetes insípida (tipo I). En el estudio se identificaron 9 atributos que son importantes para el diagnóstico de la diabetes como sexo, edad, duración de la enfermedad, historia previa, hipertensión, tipo de insulina utilizada, glucosa, hemoglobina y micro albuminuria. Al aplicar los métodos de minería de datos obtuvieron como resultado que las técnicas mencionadas son capaces de ayudar al diagnóstico de la enfermedad en nuevos pacientes.

Alcanzar un diagnóstico correcto es un problema importante en el análisis de la ciencia médica. Este tipo de dificultad se resuelve con la ayuda de algoritmos de inteligencia artificial. En [29] realizaron una comparación de diferentes técnicas de clasificación utilizando el software de código abierto WEKA. Los algoritmos o métodos probados son

Red de Bayes, Función de Base Radial, Árbol Podado, Aprendiz de Regla Individual Conjuntiva y Vecino más Cercano. Utilizaron un conjunto de datos de pacientes con cáncer de mama que consta de un total de 699 instancias. Sólo el 75% de los datos en general se emplearon como conjunto de entrenamiento y el resto se utilizó para comprobar la precisión de la clasificación. Como conclusión, después de comparar todos los algoritmos determinaron que el mejor fue el de Red de Bayes con una precisión del 89.71% y con el error promedio más bajo de .2140 en comparación de los demás. El Vecino más Cercano obtuvo el menor grado de precisión con un 84.57%.

En [30] aplicaron técnicas de minería de datos para la identificación de factores de riesgo asociados a la aparición de diabetes. Como técnica de minería de datos utilizaron un árbol de decisión. El árbol clasificó correctamente la mayoría de los casos no diabéticos con una precisión del 96.54%, por el contrario la tasa de precisión para los casos de diabetes son más bajos con un 69.47% pero aún se considera como suficiente para el objetivo de la aplicación. Al interpretar el árbol de decisión los resultados mostraron que la edad es el factor más importante asociado con la aparición de la diabetes, los individuos mayores de 50 años de edad tienen un riesgo mayor de contraer la enfermedad. El índice de masa corporal fue el siguiente factor más importante seguido de la relación cintura-cadera. El resto de los nodos en el árbol de decisión fueron interpretados de una manera similar.

Los médicos necesitan apoyarse de una metodología de predicción fiable para diagnosticar la diabetes. En [4] analizaron las técnicas de minería de datos para identificar cuáles son los más adecuados para la clasificación eficiente de la diabetes. El conjunto de datos de entrenamiento que utilizaron contiene 768 registros y cada uno de ellos cuenta con 8 atributos. Se aplicaron diferentes técnicas de clasificación como C-RT, CS-RT, C 4.5, ID3, K-NN, NAIVE-BAYES, PLS-DA, SVM Y RND TREE. De los algoritmos de clasificación anteriores RND TREE obtuvo el 100% de precisión pero debido a que el conjunto de reglas era enorme y sufría de ajustes en los datos optaron por el algoritmo C4 con un 91% de precisión. Posteriormente aplicaron técnicas de selección de características al conjunto de datos para obtener un subconjunto reducido de atributos clave y así mejorar el proceso de

clasificación. Sin embargo, al aplicar dicha técnica, la tasa de clasificación disminuyó a un 88%.

La Diabetes Mellitus es una enfermedad crónica que causa complicaciones graves a la salud incluyendo falla renal (riñón), enfermedades del corazón, derrame cerebral y ceguera. En [31] propusieron la aplicación de la técnica de clasificación Naive Bayes para predecir las probabilidades de que los pacientes diabéticos padezcan una enfermedad cardiaca. El conjunto de datos que utilizaron se recolectó del instituto de investigación líder en diabetes en Chennai, India que contiene cerca de 500 registros de pacientes. Los atributos de la diabetes que se identificaron en el sistema propuesto son: sexo, edad, herencia, peso, presión arterial, nivel de azúcar, nivel de glucosa, nivel de hemoglobina glucosilada y colesterol. La herramienta WEKA se utilizó para las tareas de minería de datos. Finalmente el modelo de Naive Bayes fue capaz de clasificar el 74% de los casos correctamente. Los resultados mostraron que el método propuesto se desempeña bien en comparación con otros similares en la literatura del área, teniendo en cuenta que los atributos tomados para el análisis no son indicadores directos de enfermedades del corazón.

En [32] compararon tres modelos de clasificación para la predicción de diabetes utilizando factores de riesgo comunes de la enfermedad. En el estudio se desarrollaron tres modelos: regresión logística, redes neuronales artificiales y árbol de decisión. Un total de 1487 personas participaron en el estudio. Se identificaron 12 variables importantes tras realizar una prueba de chi-cuadrado. Para evaluar los modelos se usaron tres criterios: precisión, sensibilidad y especificidad. La precisión se refiere a los casos clasificados correctamente, la sensibilidad a los casos positivos que se clasifican como tales y la especificidad mide los casos negativos que se clasifican como tales. Como resultado se obtuvo que el modelo de regresión logística logró una precisión de 76.13% con una sensibilidad del 79.59% y una especificidad del 72.74%. El modelo de redes neuronales tuvo una precisión de 73.23% con una sensibilidad del 82.18% y una especificidad del 64.49% y el árbol de decisión (C5.0) logró una precisión de 77.87% con una sensibilidad de 80.68% y una especificidad del 75.13%.

En [25] se estudió el proceso de construcción de modelos de predicción como regresión logística, árbol de decisión y el perceptrón multicapa para predecir la neuropatía periférica en pacientes con diabetes mellitus. El conjunto de datos fue conformado por 274 pacientes del hospital de enfermedades metabólicas en Tianjin. Se seleccionaron 30 variables tales como sexo, edad, hemoglobina, ocupación y estado civil, entre otras. Como resultados el árbol, el perceptrón multicapa y la regresión logística lograron un porcentaje de certeza del 88.63%, 85.36% y 88.02% respectivamente. El árbol presentó la capacidad de diagnóstico más alta. Después de aplicar el método de validación cruzada, el perceptrón multicapa realizó de manera óptima el trabajo seguido de la regresión logística y el árbol de decisión. Sin embargo, la regresión logística resultó la mejor elección desde el punto de vista de eficiencia y exactitud.

En [5] aplicaron la meta plasticidad artificial en el perceptrón multicapa (AMMLP) como técnica de minería de datos para el diagnóstico de la diabetes. Los datos pertenecen al Instituto Nacional de Diabetes y Enfermedades Digestivas y del Riñón ubicado en Estados Unidos y se componen de 768 casos. Los resultados obtenidos se compararon con el árbol de decisión, el clasificador bayesiano y otros algoritmos propuestos recientemente por otros investigadores, que se aplicaron a la misma base de datos. La robustez de los algoritmos se examinó con base en la precisión de la clasificación, el análisis de sensibilidad y la especificidad y la matriz de confusión. El AMMLP obtuvo una precisión del 89.53%, seguido del clasificador bayesiano con un 77.77% y por último el árbol de decisión con un 75.3%.

La retinopatía diabética (RD) es una de las complicaciones de la diabetes y es una de las causas más comunes de ceguera. Por desgracia, en muchos casos el paciente no es consciente de los síntomas hasta que es demasiado tarde para un tratamiento eficaz. En [33] se desarrolló un modelo de clasificación que ayuda al médico a identificar la prevalencia y severidad de la RD y a determinar la relación entre los factores de riesgo, prevalencia y severidad de la retinopatía diabética de los informes clínicos. Se recolectaron 3,450 historias clínicas del centro de atención de salud para diabéticos “Andhra Pradesh” y como los datos disponibles no se encontraban en formato estructurado se aplicó la técnica de

clasificación de minería de textos para predecir los factores de riesgo de la retinopatía diabética. El estudio mostró que una duración relativamente corta en la gestión de los casos, instituida antes de la aparición de la retinopatía clínicamente identificable, redujo significativamente el riesgo de desarrollar retinopatía en pacientes con diabetes tipo II.

Los pacientes de hemodiálisis (HD) son propensos a sufrir tratamientos de diálisis a largo plazo y la necesidad de ser hospitalizados. Si la tasa de hospitalización de un centro de hemodiálisis es elevada, la calidad del servicio es baja. Por lo tanto, la disminución de la tasa de hospitalización es un problema crucial para los centros de salud. En [34] se combinó la abstracción temporal (AT) con técnicas de minería de datos para analizar datos biológicos de pacientes con diálisis con el fin de desarrollar un sistema de apoyo de decisiones. El conjunto de datos se recolectó en el centro de hemodiálisis en Taiwán y contiene datos bioquímicos como albúmina, colesterol, triglicéridos y ácido úrico entre otros. En el estudio se aplicó minería de datos mediante reglas de asociación con un mínimo de múltiples soportes y árboles de decisión para analizar los datos de AT de los pacientes en HD y descubrir reglas para prevenir la hospitalización de los pacientes. Los resultados experimentales mostraron que los diferentes métodos de minería de datos se combinan eficazmente. Entre los patrones de hospitalización encontrados en este estudio, fue la albúmina el índice más importante para predecir la hospitalización de los pacientes. La evolución en el tiempo de la evolución bioquímica de análisis de sangre demostró tener una relevancia definitiva en la hospitalización.

En [35] se utilizó una función de entropía para identificar los factores clave relacionados con la hemodiálisis. Sobre el esquema propuesto se aplicó el algoritmo de agrupamiento k-media con los factores clave para categorizar a los pacientes. Se utilizaron técnicas de minería de datos para encontrar las reglas de asociación de cada grupo. Las reglas encontradas se emplearon para determinar si un paciente requiere de hemodiálisis. El experimento utilizó registros de exámenes de salud prestados por un hospital general en Taiwán. A partir de los resultados experimentales se encontró que si el nitrógeno ureico en sangre está en el intervalo (60 ± 1,5) y el Natrium está en el intervalo (140 ± 2,5), los pacientes tienen un alto riesgo de recibir una diálisis. El nitrógeno en la sangre resultó ser

un indicador fiable de alto riesgo pero el Natrium no quedo claramente definido. A la inversa, si el ácido úrico está en el intervalo (6,5 ± 0,25), los triglicéridos en el rango (159,75 ± 25) y el potasio en el intervalo (4,14 ± 0,25) los pacientes tienen bajo riesgo de recibir una diálisis.

La nefropatía diabética (ND) es una complicación frecuente y la principal causa de muerte en personas con diabetes. Para predecir con exactitud el inicio de la ND en [36] aplicaron diferentes técnicas de aprendizaje automático, como soporte de máquinas de vectores (SVM) y métodos de selección, a un conjunto de datos irregulares y desequilibrados de diabetes. La visualización de los factores de riesgo es otro importante objetivo para dar a los médicos información intuitiva en el patrón clínico de cada paciente. Se recolectaron los datos médicos de 292 pacientes con diabetes. Para predecir la aparición de la ND, compararon varios métodos de clasificación tales como regresión logística, SVM y SVM con un método de aprendizaje sensible al costo. También se aplicaron varios métodos de selección para eliminar características redundantes y mejorar el rendimiento de la clasificación. Los mejores resultados se obtuvieron al usar SVM lineal combinado con envoltorios o métodos selectos incrustados. Finalmente el método propuesto predijo la aparición de ND de 2 a 3 meses antes del diagnóstico real. Además, el sistema de visualización proporcionó a los médicos información intuitiva para el análisis de factor de riesgo que facilitan la planificación de estrategias de tratamientos eficaces y apropiados.

Pacientes con enfermedades crónicas son casos de atención médica complejos que requieren de la interacción coordinada de múltiples profesionales. En [37] se presentó una ontología para el cuidado de pacientes con enfermedades crónicas y se implementaron dos procesos de personalización y una herramienta de apoyo a la decisión. El primer proceso de personalización adaptó el contenido de la ontología a las particularidades observadas en el registro sanitario de un paciente concreto proporcionando automáticamente una ontología personalizada que contiene sólo la información clínica que es relevante para los profesionales de la salud que sirve para administrar a ese paciente. El segundo proceso de personalización utilizó la ontología personalizada de un paciente para transformar automáticamente los planes de intervención describiendo los tratamientos generales de

atención sanitaria. Finalmente la ontología también se utilizó como la base de conocimiento de una herramienta de apoyo a las decisiones que ayuda a los profesionales de la salud a detectar circunstancias anómalas como diagnósticos equivocados, comorbilidades no observadas, falta de información, enfermedades relacionadas observadas o medidas preventivas.

La diabetes mellitus es un grave problema de salud en Tailandia. La gestión de pacientes con diabetes mellitus se convierte en una cuestión importante. En [38] se propuso una ontología basada en el desarrollo de un sistema de información clínica que permite a los proveedores de salud que no son expertos en la enfermedad de la diabetes sugerir a los pacientes las actividades esenciales para mejorar su calidad de vida y el logro de los objetivos del tratamiento de la diabetes mellitus. En el trabajo se centraron principalmente en el proceso de desarrollo de la ontología para la diabetes mellitus tipo II. Se describieron tres pasos principales, 1) dominio y el valor del ámbito, 2) adquisición de términos importantes, clases y conceptualización de jerarquía de clases y 3) casos de instancia. La ontología se diseñó con base en 1) la "Guía Práctica de Diabetes Mellitus de la Sociedad de Endocrinología de Tailandia" que es un documento que tiene el objetivo de guiar las decisiones y criterios con respecto al diagnóstico, manejo y tratamiento de la diabetes y 2) discusiones con médicos para verificar su exactitud.

La diabetes es un problema de salud importante en Arabia Saudita. Es la enfermedad endócrina más común en toda la población y grupos de edad. En [8] se realizó un análisis predictivo del tratamiento de la diabetes mediante la técnica de regresión de minería de datos. El software *Oracle Data Miner* se empleó como herramienta de minería de datos para predecir los modos de tratamiento de la diabetes. Los datos se obtuvieron de la Organización Mundial de la Salud. El conjunto de datos se estudió y se analizó para identificar la eficacia de los diferentes tipos de tratamiento para diferentes grupos de edad. Se identificaron cinco grupos de edad (15-24, 25-34, 35-44, 45-54, 55-64) y se consolidaron en 2 grupos que son: jóvenes y mayores. Como resultado, las predicciones sobre la eficacia de un tratamiento diferente para los dos grupos de edad fueron aclaradas. Control de la dieta, reducción de peso, el ejercicio y dejar de fumar son mutuamente

beneficiosos entre sí para el tratamiento de la diabetes. El tratamiento farmacológico de los pacientes en el grupo de edad joven se retrasó para evitar efectos secundarios. Por el contrario, los pacientes en el grupo de edad mayor se tuvieron que prescribir en el tratamiento farmacológico de inmediato, junto con otros tratamientos, ya que no hay otras alternativas disponibles.

Aunque los sistemas de conocimiento de medicina tradicional y moderna difieren, los estudios comparativos y enfoques aplicados son valiosos. Los síntomas del paciente y los tratamientos prescritos son campos importantes para vincular la ciencia de la medicina moderna con la medicina tradicional. Bajo ese contexto, en [39], construyeron una ontología para la medicina tradicional coreana que contiene información acerca de síntomas, enfermedades y tratamientos. Se utilizó Protégé para el diseño de la ontología y Jena para escribir los valores y propiedades de los medicamentos. Los datos se extrajeron textualmente de la Farmacopea Coreana (Administración de Drogas y Alimentos de Corea), libros de medicina tradicional y libros de texto adoptados por la mayoría de universidades coreanas. La información sobre el tratamiento se vinculó a la enfermedad en la ontología de la medicina tradicional así como a la enfermedad en la ontología de la ciencia médica moderna. El conocimiento de los tratamientos medicinales en la medicina tradicional resultaron útiles debido a su relación a las enfermedades y este conocimiento fue respaldado por el análisis de las relaciones entre enfermedades tradicionales y enfermedades modernas.

Los pacientes con diabetes tipo I toman insulina regularmente para impedir que sus niveles de glucosa se eleven en exceso. A la mayoría se les ofrece un programa de educación estructurada para enseñarles como auto gestionar su enfermedad mediante el cálculo de la cantidad de insulina para administrar en cada comida. Esta complejidad, combinada con el reciente crecimiento de los teléfonos inteligentes, llevó al desarrollo de una gran cantidad de aplicaciones para la gestión de la diabetes para dispositivos móviles. Estas aplicaciones permitieron a los pacientes diabéticos registrar lecturas de glucosa en la sangre, los carbohidratos que consume, la dosis de insulina, la actividad física realizada y otras actividades y observaciones. En [40] se desarrolló una ontología ligera que captura el tipo

de información registrada por tales aplicaciones con el fin de facilitar la interoperabilidad entre dichas aplicaciones. Se realizó una revisión sistemática de las aplicaciones para establecer una lista de términos que a capturar por la ontología y con esta lista definir clases y propiedades adecuados para luego investigar la forma en que se asignan a las normas existentes como HL7-RIM y OpenEHR.

En los párrafos anteriores se mencionan únicamente artículos publicados fuera de México, entre los publicados en el país sobre el tema se encuentran: [41] se construyó un cuestionario para medir el estilo de vida en pacientes diabéticos que consistía de 25 preguntas cerradas, distribuidas en dominios como: nutrición, actividad física, consumo de alcohol, entre otros. Para su construcción se usó: estadística descriptiva, confiabilidad alfa de Cronbach, coeficiente de correlación de Spearman, entre otros. En [42] se estudió el apego al tratamiento farmacológico en pacientes con diabetes tipo 2 mediante la medición del conteo de tabletas en el domicilio del paciente para determinar si el consumo es adecuado, la medición de hemoglobina glucosilada y la aplicación de un cuestionario para determinar factores relacionados al no apego del tratamiento, para el análisis de los datos emplearon estadística descriptiva y razón de momios. En [43] se estudió el deterioró en la calidad de vida de los pacientes diabéticos, mediante el uso de la Escala de Calidad de Vida Relacionada con la Salud (ESCARVIS), estadística descriptiva, *t student*, alfa de Cronbach, x^2 y prueba exacta de Fisher. En [44] se utilizaron técnicas de estadística descriptiva, correlación de Spearman, Alfa de Cronbach, entre otras para adaptar el cuestionario de calidad de vida de "Diabetes 39" hacia pacientes diabéticos mexicanos.

2.2 Análisis comparativo

Con base en los artículos recopilados en el apartado 2.1, se realizó una tabla comparativa con el objetivo de identificar las técnicas de desarrollo, tecnologías y enfermedades en las que actualmente otros investigadores trabajan. Para la construcción de la tabla se incluyeron los aspectos que se consideraron más relevantes para el análisis de este trabajo, como son los autores, el nombre y el objetivo del artículo, las técnicas o algoritmos que se emplearon para desarrollar el trabajo, el software utilizado y la enfermedad que se pretende combatir. La Tabla 2.1 muestra el análisis realizado del estado de la práctica.

Tabla 2.1 Tabla comparativa de artículos relacionados

Autores	Artículo	Objetivo	Técnicas	Software	Enfermedad
H. Kaur and S. Krishan Wasan [28]	*Empirical study on applications of data mining techniques in healthcare*	Analizar la aplicación de técnicas de clasificación de minería de datos sobre un conjunto de datos médicos de niños con diabetes.	Arboles de decisión, redes neuronales y basadas en reglas	No especifica.	Diabetes
M. Fauzi bin Othman and T. Moh Shan Yau [29]	*Comparison of different classification techniques using WEKA for breast cancer*	Investigar el desempeño de diferentes técnicas de clasificación sobre un conjunto de datos de cáncer de mama.	Red de Bayes, árbol podado, función de base radial, vecino más cercano y *single conjunctive rule learner*	WEKA	Cáncer de Mama
Koh, Hian Chye, and Gerald Tan [30]	*Data mining applications in healthcare*	Explorar las aplicaciones de minería de datos en el sector salud. Realiza un ejemplo que implica la identificación de factores de riesgo asociados a la aparición de diabetes.	Árbol de decisión	*SPSS's Clementine data Mining*	Diabetes

Autores	Artículo	Objetivo	Técnicas	Software	Enfermedad
K. Rajesh and V. Sangeetha [4]	*Application of Data Mining Methods and Techniques for Diabetes Diagnosis*	Analizar los métodos y técnicas de minería de datos para identificar cuáles son los más eficientes para la clasificación de la diabetes	C-RT, CS-RT, C 4.5, ID3, K-NN, LDA, NAIVE-BAYES, PLS-DA, SVM Y RND TREE	No especifica	Diabetes
G. Parthiban, A. Rajesh and S. K. Srivatsa [31]	*Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method*	Predecir las probabilidades de que los pacientes diabéticos padezcan una enfermedad cardiaca	Naive Bayes	WEKA	Enfermedad Cardiaca
Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang and Qing Liu [32]	*Comparison of three data mining models for predicting diabetes or prediabetes by risk factors*	Comparar tres modelos de clasificación para la predicción de diabetes utilizando factores de riesgo comunes de la enfermedad	Regresión logística, redes neuronales artificiales y árbol de decisión	*SPSS statistical Program*, *SPSS Modeler*	Diabetes
Li, Xin-yue, Jun, Zhuang, Zi-long, Cui and Liang-ping [45]	*Performance comparison between Logistic regression, decision trees, and*	Estudiar el proceso de construcción de modelos de predicción para predecir la neuropatía periférica en	Regresión logística, árbol de decisión y el perceptrón multicapa	No especifica	Neuropatía periférica

Autores	Artículo	Objetivo	Técnicas	Software	Enfermedad
	multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus	pacientes con diabetes mellitus			
Marcano-Cedeno and Andina [5]	*Data mining for the diagnosis of type 2 diabetes*	Aplicar la meta plasticidad artificial en el perceptrón multicapa (AMMLP) como técnica de minería de datos para el diagnóstico de la diabetes	Meta plasticidad artificial en el perceptrón multicapa (AMMLP), Bayes y árbol de decisión.	WEKA	Diabetes
S.S. Imambi and T. Sudha [33]	*Building Classification System to Predict Risk factors of Diabetic Retinopathy Using Text mining*	Identificar la prevalencia y severidad de la retinopatía diabética y determinar la relación entre los factores de riesgo, prevalencia y severidad de la retinopatía diabética.	Minería de Texto	No especifica	Retinopatía Diabética

Autores	Artículo	Objetivo	Técnicas	Software	Enfermedad
Jinn-Yi Yeh, Tai-Hsi Wu and Chuan-Wei Tsao [34]	*Using data mining techniques to predict hospitalization of hemodialysis patients*	Combinar la abstracción temporal (AT) con técnicas de minería de datos para el análisis de datos biológicos de pacientes con diálisis para desarrollar un sistema de apoyo de decisiones	Abstracción temporal, reglas de asociación y árboles de decisión	No especifica	Hemodiálisis
Tzu-Chuen Lu and Chun-Ya Tseng [35]	*Hemodialysis key features mining and patients clustering technologies*	Identificar los factores clave relacionados con la hemodiálisis	K-media, agregación, reglas de asociación	No especifica	Hemodiálisis
Baek Hwan Cho, Hwanjo Yu, Kwang-Won Kim, Tae Hyun Kim, In Young Kim, and Sun I. Kim [36]	*Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods*	Predecir la aparición de nefropatía diabética utilizando soporte de máquinas de vectores sobre un conjunto de datos irregulares y desequilibrados de diabetes	Regresión logística, soporte de máquinas de vectores	No especifica	Nefropatía Diabética

Autores	**Artículo**	**Objetivo**	**Técnicas**	**Software**	**Enfermedad**
D. Riaño, F. Real, J. A. López-Vallverdú, F. Campana, S. Ercolani, P. Mecocci and R. Annicchiarico [37]	*An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients.*	Desarrollar una ontología para el cuidado de pacientes con enfermedades crónicas	No aplica	Protégé	Enfermedades Crónicas
N. Chalortham, M. Buranarach & T. Supnithi [38]	*Ontology Development for Type II Diabetes Mellitus Clinical Support System*	Proponer una ontología para un sistema de información clínica que permita a usuarios no expertos en diabetes sugerir las actividades esenciales para mejorar la calidad de vida y el logro de los objetivos del tratamiento de la diabetes mellitus.	No aplica	*Hozo-ontology editor*	Diabetes
Aljumah, Gulam Ahamad, and Khubeb Siddiqui	*Application of Data Mining: Diabetes Health Care in Young*	Usar técnicas de minería de datos para descubrir los patrones que identifican el	Regresión y soporte de máquinas de vectores	*Oracle Data Miner*	Diabetes

Autores	Artículo	Objetivo	Técnicas	Software	Enfermedad
[46]	*and Old Patients*	mejor modo de tratamiento para la diabetes a través de diferentes grupos de edad.			
Sutton, Aldea, and Martin [39]	*Ontology of diabetes self-management*.	Desarrollar una ontología ligera que capture el tipo de información registrada por aplicaciones móviles dedicas a la gestión de pacientes con diabetes con el fin de que facilite la interoperabilidad entre éstas.	No aplica.	Protégé, Jena	Diversas
Jang, Hyunchul, et al. [40]	*Ontology for medicinal materials based on traditional Korean medicine*	Desarrollar una ontología para la medicina tradicional coreana que contiene información acerca de síntomas, tratamientos y medicinas para diversas enfermedades.	No aplica.	Protégé	Diabetes

Autores	Artículo	Objetivo	Técnicas	Software	Enfermedad
López, J. et al. [41]	Construcción y validación inicial de un instrumento para medir el estilo de vida en pacientes con diabetes mellitus tipo 2	Construir y validar un instrumento estadístico para medir el estilo de vida de pacientes de diabetes mellitus	No aplica (para efectos de minería de datos y/u ontologías)	No especifica	Diabetes Tipo II
Durán, B. et al. [42]	Apego al tratamiento farmacológico en pacientes con diagnóstico de diabetes mellitus tipo 2	Realizar estudio estadístico sobre el nivel de apego al tratamiento farmacológico en pacientes con Diabetes Tipo II	No aplica (para efectos de minería de datos y/u ontologías)	No especifica	Diabetes Tipo II
De los Ríos, J. et al. [43]	Calidad de vida en pacientes con diabetes mellitus tipo 2	Realizar un estudio respecto a la calidad de vida de los pacientes con Diabetes Tipo II	No aplica (para efectos de minería de datos y/u ontologías)	No especifica	Diabetes Tipo II
López, J. et al. [44]	Adaptación y	Realizar adecuaciones al	No aplica (para efectos	No	Diabetes Tipo

Autores	Artículo	Objetivo	Técnicas	Software	Enfermedad
	validación del instrumento de calidad de vida Diabetes 39 en pacientes mexicanos con diabetes mellitus tipo 2	instrumento existente para analizar la calidad de vida de pacientes con Diabetes Tipo II	de minería de datos y/u ontologías)	especifica	II

Como se observa en la tabla anterior, en [29] se compararon diversos métodos de minería de datos para el diagnóstico del cáncer de mama. En [33] se empleó minería de textos para identificar factores de riesgo relacionados con la retinopatía diabética. En [28] [4] [5], [30] , [32] se aplicaron diversas técnicas de minería de datos para el diagnóstico de la diabetes pero en todos los casos no se aborda el tema de los padecimientos asociados como la ceguera o la insuficiencia renal. En [34] y [35] se reportó la aplicación de Minería de Datos a la predicción de condiciones específicas de hemodiálisis (necesidad de hospitalización, diálisis) sin relacionarla directamente con variables de diabetes mellitus tipo II. En [31] se predijo la probabilidad de que los pacientes con diabetes mellitus tipo II padezcan una enfermedad cardiaca. En [25] y [36] se usó minería de datos para predecir oportunamente la aparición de nefropatía diabética en pacientes con diabetes mellitus tipo II. Por el lado de las ontologías en [37] se propuso una ontología para el cuidado de pacientes con enfermedades crónicas, en [39] una ontología con información de la medicina tradicional coreana, enfermedades, tratamiento y síntomas. En [38] se presentó una ontología para ayudar a los proveedores de salud con el tratamiento y seguimiento de pacientes con diabetes y finalmente en [40] una ontología que captura la información registrada por las aplicaciones móviles dedicas a la gestión de pacientes con diabetes. Todo esto permite confirmar las ventajas que proporciona el uso de Minería de Datos en tareas de clasificación y/o predictivas, así como la inferencia de conocimiento con el uso de ontologías, en el área de la salud en el mundo.

Por su parte, la mayoría de los artículos reportados en México que se analizaron utilizan estadística descriptiva para el análisis de los datos y se centran en identificar los factores que afectan la calidad de vida del paciente diabético y determinar si el apego a su tratamiento es el adecuado.

Por lo anterior, el objetivo de esta tesis resulta relevante ya que lo que busca es utilizar técnicas de minería de datos para el análisis de la información con las que no sólo se identifican los factores más relevantes para la aparición de complicaciones en pacientes diabéticos sino que además se crean modelos de comportamientos que se utilizan para predecir la aparición de estas enfermedades y se infiere un tratamiento tanto farmacológico

como no farmacológico mediante el uso de una ontología con el fin de mejorar su calidad de vida del paciente.

Después de realizar el análisis de los artículos relacionados se observaron los diferentes campos de acción donde se aplica la minería de datos, el uso de diferentes técnicas de análisis de datos y el manejo de herramientas de software, los cuales se consideraron para el desarrollo de este proyecto.

2.3 Propuesta de solución

Después de hacer un análisis de las tecnologías de información existentes en el mercado y comparar diferentes alternativas de solución considerando las tecnologías de Java y .NET se determinó utilizar la propuesta que se observa en la Tabla 2.2, en la cual se emplea JSF como tecnología para desarrollar la aplicación Web, PostgreSQL como gestor de base de datos, Jena como marco de trabajo para ontologías y WEKA como marco de trabajo para minería de datos. Dentro de las metodologías seleccionadas están Methontology para el desarrollo de ontologías y KDD para el desarrollo de minería de datos.

Tabla 2.2 Propuesta de solución para el proyecto de tesis.

Propuesta de solución	
Tecnología WEB	JSF
IDE desarrollo	NetBeans
Gestor de BD	PostgreSQL
Marco de trabajo ontología	Jena
Marco de trabajo minería	Weka
Metodología ontología	Methontology
Metodología minería	KDD

JSF se eligió sobre otras tecnologías porque tiene como respaldo a un lenguaje potente y muy utilizado en la actualidad que es Java y que ofrece muchas ventajas como ser multiplataforma, gratuito y que maneja APIs de minería de datos como es el caso de Weka.

Otra razón es porque JSF emplea el patrón arquitectónico MVC y es el estándar oficial de Java para el desarrollo de aplicaciones Web.

Como gestor de base de datos se decidió utilizar PostgreSQL debido a características como que garantiza ser 100% ACID en sus operaciones a diferencia de MySQL, que bajo ciertos motores de búsqueda, no garantiza la misma consistencia. Y además que al utilizar el lenguaje de Java resulta un poco más natural realizar la conexión de la aplicación a PostgreSQL que a SQL Server de Microsoft.

Como marco de trabajo de ontología, Jena fue el candidato seleccionado debido principalmente a que está diseñado para Java. En cuanto a la metodología para el desarrollo de la ontología se seleccionó *Methontology* porque se considera sencilla de aprender, con una buena documentación y se ha empleado en diversos proyectos con éxito.

En lo que respecta al marco de trabajo de minería de datos, Weka representa la mejor solución ya que es gratuito y posee una gran cantidad de algoritmos de minería de datos en comparación con *Analysis Services*. Otra razón es que se diseñó para utilizarse en Java. En cuanto a metodología de minería de datos se seleccionó KDD ya que es fácil de implementar y se menciona su uso en muchos artículos de minería de datos.

Capítulo 3. Aplicación de la metodología

En este capítulo se describe el seguimiento de la metodología para el desarrollo de la aplicación.

Introducción

Este estudio hace uso de técnicas de minería de datos para la elaboración de modelos que predigan la aparición de complicaciones en pacientes con Diabetes Mellitus Tipo II y también hace uso de una ontología que proporciona cuidados preventivos a pacientes diabéticos con el fin de intentar prevenir la aparición de complicaciones de la diabetes. A esta ontología se le asignó el nombre de Ontología de Cuidados Preventivos para Pacientes Diabéticos (OCPPD). Para consultar a los modelo de minería de datos y a la ontología OCPPD se desarrolló una aplicación Web bajo la tecnología de JSF.

Bajo el panorama de desarrollo descrito anteriormente, se decidió dividir este capítulo en tres partes: la primera en la que se define la arquitectura de la aplicación y se describen sus componentes, la segunda donde se menciona la metodología llevada a cabo para la construcción de los modelos de minería de datos y la tercera donde se describe la metodología llevada a cabo para la construcción de la ontología OCPPD.

3.1 Arquitectura de la aplicación

En la Figura 3.1 se muestra el esquema de arquitectura de la aplicación Web, como se observa está basada en el modelo arquitectónico MVC (Modelo Vista Controlador), el cual permite separar la lógica del negocio (modelo) de la presentación (vista) obteniendo así aplicaciones escalables y de fácil mantenimiento. Al tratarse de una aplicación Web, ésta se encuentra alojada en un servidor y un cliente accede a ella a través de un visualizador.

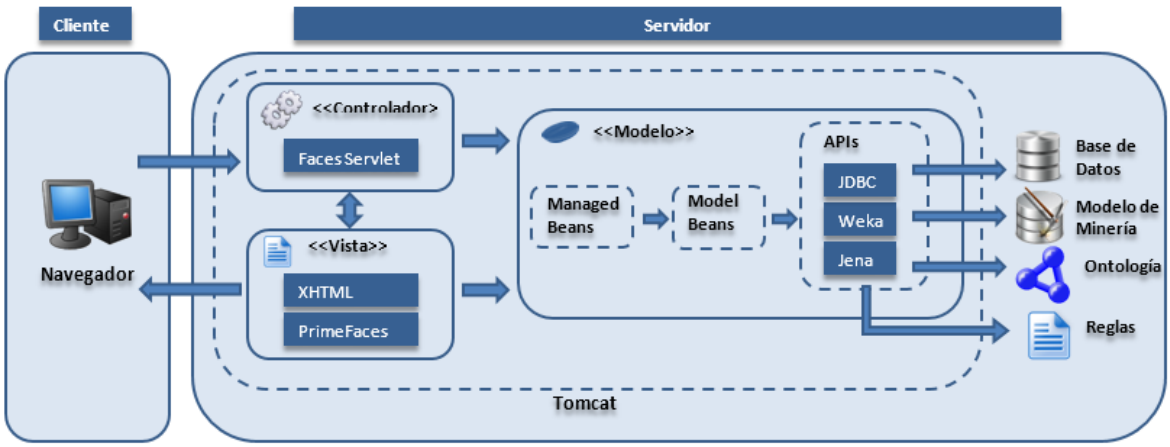


Figura 3.1 Arquitectura de la aplicación.

Los componentes que integran la arquitectura del sistema se describen a continuación:

Controlador: contiene el ***FacesServlet*** es decir el *servlet* de JSF cuya función principal es coordinar a la vista y al modelo dirigiendo el flujo de navegación y gestionando las diferentes peticiones de recursos.

Vista: está compuesta por archivos **XHTML** que representan la interfaz del usuario, su principal función es el despliegue y la captura de información. Para realizar dicha tarea se ayuda del uso de las etiquetas propias de JSF. También se incluye la biblioteca de **PrimeFaces** que es un componente para JSF que permite crear aplicaciones Web enriquecidas.

Modelo: en el modelo se encuentran los ***Managed Beans*** (*beans* administrados por el marco de trabajo) que representan la lógica de la aplicación. Permiten acceder a los componentes de la interfaz del usuario para su manipulación y permiten el paso de información a los ***Model Beans***. Por su parte los ***Model Beans*** representan a las clases que son importantes en el dominio del problema, en el caso de esta aplicación un paciente o un médico son ejemplo de *Model Beans*. Los ***Model Beans*** hacen uso de diferentes **APIs** como lo son Jena, Weka y JDBC. La API de **Jena** proporciona la funcionalidad necesaria para acceder a la **ontología** OCPPD y al archivo de **reglas**; la ontología está representada por un archivo con extensión `.owl` escrito en formato RDF/XML y las reglas por un archivo con extensión .rules escritas en el formato de Jena. La API de **Weka** permite utilizar los modelos de minería de datos que son los que se utilizan para hacer la predicción

de complicaciones en pacientes diabéticos. Los modelos de minería de datos se representan por archivos con extensión `.model`. Y finalmente la biblioteca de **JDBC** que permite conectarse al gestor de base de datos PostgreSQL y manipular su información.

Casos de uso

En la Figura 3.2 se muestra la funcionalidad del sistema a través de un diagrama de casos de uso. Como se aprecia en la imagen, sólo existe un actor: Médico, que representa a la persona que se encargará de manejar el sistema En total se identifican 11 casos de uso. Cuatro de ellos referentes a la gestión de los médicos (Registrar Médico, Consultar Médico, Eliminar Médico, Modificar Médico); dos a la autenticación del sistema (Ingresar al Sistema, Salir del Sistema); tres con respecto a la realización de las predicciones (Predecir Retinopatía Diabética, Predecir Nefropatía Diabética, Predecir Pie Diabético) que engloban todas las tareas necesarias para predecir la aparición de complicaciones de la diabetes; finalmente, dos a la obtención de tratamiento en un paciente en particular (Obtener Tratamiento No Farmacológico, Obtener Tratamiento Farmacológico).

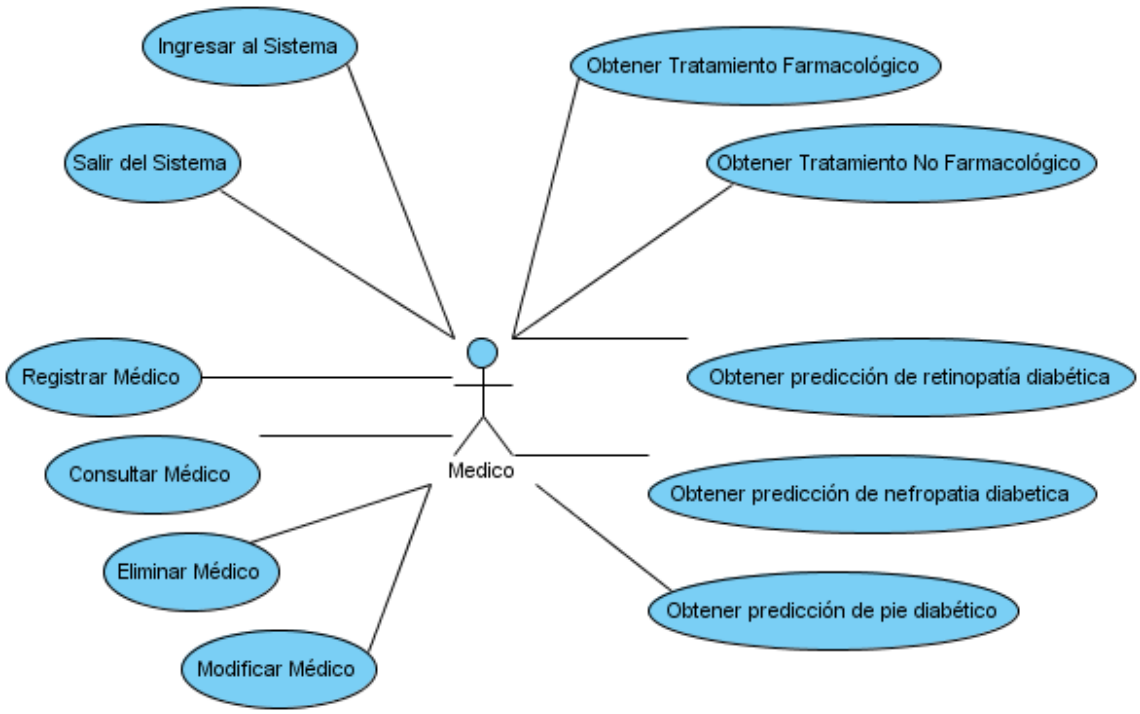


Figura 3.2 Diagrama de casos de uso de la aplicación.

Diagrama de clases

En la Figura 3.3 se describen las clases que integran al sistema mediante un diagrama de clases de alto nivel. Como se aprecia en la imagen, existen 6 paquetes en los que estarán contenidas todas las clases del sistema.

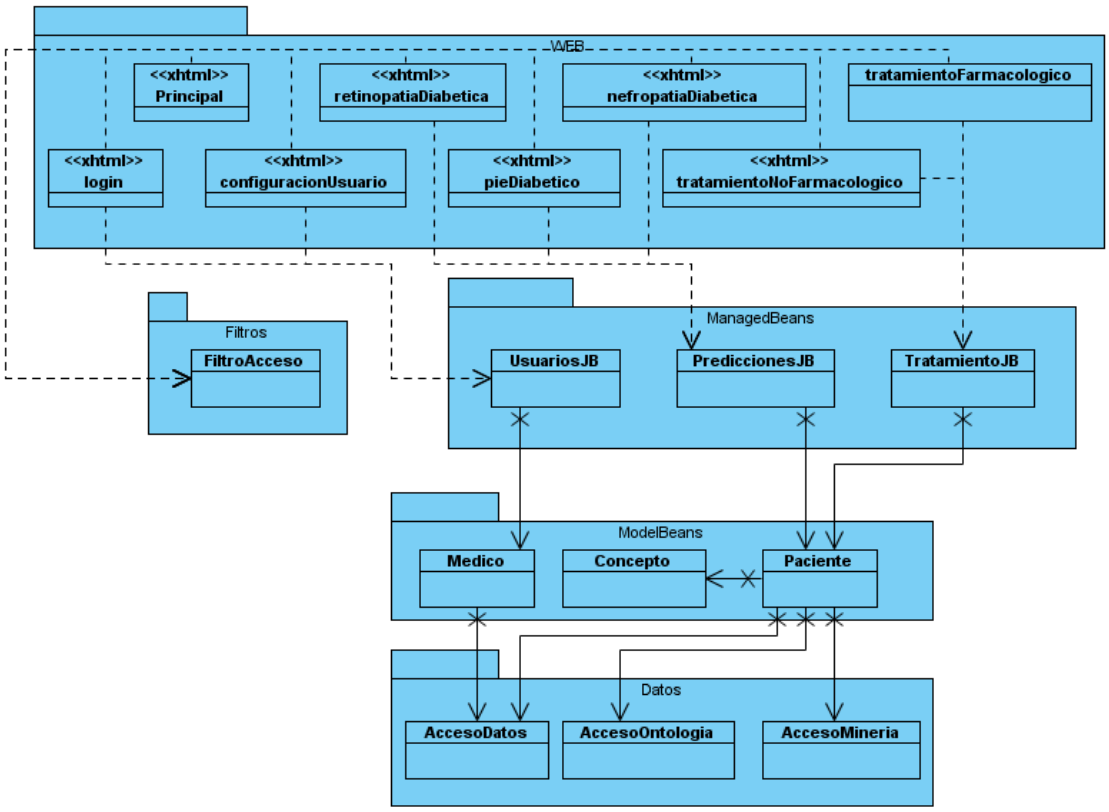


Figura 3.3 Diagrama de clases

WEB: representa a la Vista del sistema. Aquí se encuentran todas las páginas XHTML de la aplicación. Incluye a las páginas:

- **Login:** Representa la vista con los campos necesarios para permitir acceder al sistema.
- **Principal:** Representa la pantalla principal, en ella se encuentra el menú para dirigirse a todas las demás vistas.
- **ConfiguracionUsuario:** En esta vista se encuentran todos los campos para realizar las altas, bajas, modificaciones y consultas de los médicos.
- **RetinopatiaDiabetica:** Representa la vista que contiene los campos para capturar y mostrar los resultados de la predicción de retinopatía diabética.

- **NefropatiaDiabetica:** Representa la vista que contiene los campos para capturar y mostrar los resultados de la predicción de nefropatía diabética.
- **PieDiabetico:** Representa la vista que contiene los campos para capturar y mostrar los resultados de la predicción de pie diabético.
- **TratamientoFarmacologico:** Representa la vista que contiene los campos para capturar y mostrar el tratamiento farmacológico sugerido a un paciente.
- **TratamientoNoFarmacologico:** Representa la vista que contiene los campos para capturar y mostrar el tratamiento no farmacológico sugerido a un paciente.

Filtros: contiene a la clase **FiltroAcceso** que se comunica con el *servlet* de la aplicación y restringe a los usuarios el acceso a ciertos recursos.

ManagedBeans: aquí se encuentran todos los *beans* administrados que realizan la lógica del negocio, manipulan los componentes de las páginas de la vista y acceden a los *beans* de modelo. Incluye las clases:

- **PrediccionesJB:** Recibe los datos de la vista e invoca al modelo para que se realice la predicción de alguna complicación en un paciente.
- **CuidadosJB:** Recibe los datos de la vista e invoca al modelo para que se obtengan los cuidados preventivos para un paciente.
- **UsuariosJB:** captura los datos de la vista e invoca al modelo para que se registren, eliminen, consulten, modifiquen, inicien o cierren sesión los médicos.

ModelBeans: incluye las clases que son importantes en el dominio del problema.

- **Medico:** Representa a un médico y contiene los métodos para autenticar, cerrar sesión, registrar, modificar, consultar y eliminar a un médico del sistema.
- **Paciente:** Representa a un paciente y tiene los métodos necesarios para la realización de predicciones (nefropatía, retinopatía, pie diabético) y la obtención de tratamientos (farmacológico, no farmacológico).
- **Concepto:** Representa una clase de la Ontología.

Datos: incluye las clases para acceder a los repositorios de información.

- **AccesoDatos:** clase que contiene los métodos para acceder, manipular y consultar la base de datos.
- **AccesoOntologia:** clase que contiene los métodos para acceder, manipular y consultar la ontología de cuidados preventivos.
- **AccesoMineria:** clase que contiene los métodos para acceder y consultar los modelos de minería de datos.

Las páginas del paquete WEB se relacionan con las clases del paquete ManagedBeans lo que permite a las páginas Web invocar a los métodos correspondientes en los *beans* administrados, obtener sus resultados y presentarlos en la vista, por ejemplo: la página configuracionUsuario invoca al método listarUsuarios() perteneciente al *bean* administrado UsuarioJB. A su vez, las clases de los *beans* administrados contiene instancias de los *beans* de modelo e invocan los métodos correspondientes dependiendo de la petición originada en la vista, por ejemplo el método listarUsuarios() invoca al método listarMedicos() del *bean* de modelo Medico. Los *bean* de modelo se relacionan con el paquete Datos y contienen una instancia de este último paquete, por ejemplo la clase médico contiene una instancia de la clase AccesoDatos a través de la cual se realiza la conexión a la base de datos.

3.2 Desarrollo de los modelos de minería de datos utilizando KDD

En este apartado se describe la metodología que se utilizó para la construcción de los modelos de minería de datos. En total se elaboraron tres modelos. Uno para la predicción de retinopatía diabética, otro para nefropatía diabética y otro para pie diabético.

Para la construcción de estos modelos se contó con la colaboración del Hospital Regional de Río Blanco (HRRB). Un profesional de la salud del HRRB brindó su conocimiento y experiencia y fungió como asesor médico para el desarrollo de los modelos predictores. Por otro lado, también se obtuvieron algunas recomendaciones médicas por parte de un profesional de la salud de la Fundación Doctor Hernández Zurita, I.B.P. Del HRRB se obtuvo una población de pacientes con diagnóstico de diabetes mellitus tipo II que se utilizó para el entrenamiento de los modelos de minería de datos.

Para el análisis de los datos se empleó WEKA, que es una herramienta escrita en Java que contiene una colección de algoritmos de aprendizaje automático para tareas de Minería de Datos. Dentro de las tareas para las que se usó WEKA están el pre procesamiento de datos, clasificación, visualización, y selección de características.

Para el desarrollo de los modelos de minería de datos se siguió el proceso de descubrimiento de base de datos, mejor conocido por sus siglas en ingles KDD (*Knowledge Discovery from Data*).

En los temas siguientes se explica el desarrollo de los modelos de minería de datos a través de las diferentes fases de KDD.

3.2.1 Selección de Datos

En la Figura 3.4 se observan algunas de las actividades que se realizaron en la fase de selección de datos. Como primer paso se definieron las características de la población y el tamaño de la muestra, posteriormente se seleccionaron las variables que se involucran en la aparición de complicaciones en pacientes diabéticos, después se procedió a revisar todos los expedientes clínicos de los pacientes seleccionados, posteriormente la información se capturó digitalmente en un documento de Excel y, por último, se procedió a importar el documento de Excel a una base de datos para su manipulación en fases posteriores.

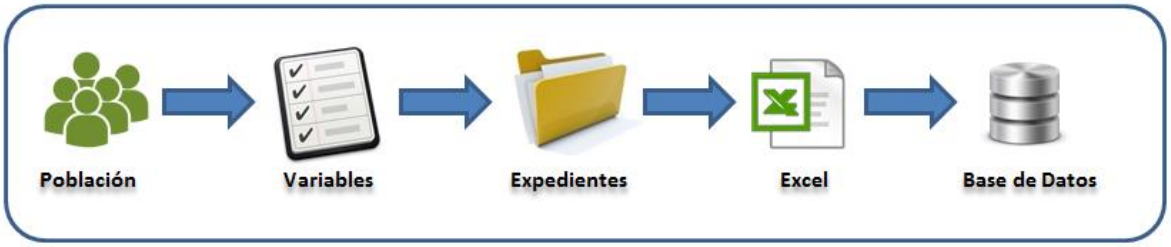


Figura 3.4 Actividades realizadas en la fase de selección de datos

3.2.1.1 Población

El universo de trabajo está conformado por pacientes con diagnóstico de Diabetes Mellitus Tipo II, en cualquier etapa de la enfermedad y con presencia de alguna de las siguientes complicaciones: retinopatía diabética, nefropatía diabética o pie diabético. También se

incluyeron pacientes con diagnóstico de Diabetes Mellitus Tipo II, en cualquier etapa de la enfermedad sin presencia de complicaciones, es decir, sin contar con un diagnóstico de retinopatía diabética, nefropatía diabética o pie diabético.

3.2.1.1.1 Tamaño de la muestra

La población está conformada por un total de 205 pacientes con Diabetes Mellitus Tipo 2 que acudieron a consulta externa, cirugía o urgencias en el HRRB en el periodo comprendido del año 2012 a 2014.

3.2.1.2 Técnica de recolección de datos

Se hizo uso de cédulas de recolección de datos para capturar la información de los expedientes clínicos de los pacientes diabéticos. Para agilizar el proceso de recopilación de información el llenado de las cédulas se hizo electrónicamente con la ayuda de Microsoft Excel. En el anexo 1 se observa la cédula de recolección de datos que se utilizó.

3.2.1.3 Aspectos éticos

Este trabajo se apega a lo establecido en el Reglamento de la Ley General de Salud en materia de investigación para la salud, Artículo 17. Es una investigación sin riesgo, no se producirá ninguna alteración física o psicológica en el paciente debido a que es un tipo de estudio retrospectivo, observacional e implica técnicas o métodos de investigación documental (expedientes clínicos). De igual forma, se garantiza la protección de datos personales de los pacientes debido a que su nombre, número de seguro, CURP o alguna otra forma de identificación no se requieren en ningún momento de la investigación.

3.2.1.4 Selección de Variables

Para la selección de las variables se consultaron diferentes guías de la práctica clínica elaboradas por instituciones de salud como el Instituto Mexicano del Seguro Social (IMSS) y la Asociación Latinoamericana de Diabetes (ALAD) entre otras, también se consultaron a tres médicos: un médico de una institución de salud pública, un médico de una institución benéfica y un médico de una institución de salud privada que aportaron sus conocimientos para la selección de variables, si bien la asesoría principal fue por parte del médico de salud

pública que se encargó de validar y corregir las variables seleccionadas, además de agregar algunas que consideraba de importancia para la realización del estudio. Un total de 62 variables se seleccionaron, en la Tabla 3.1 se muestran todas las variables involucradas en el estudio.

Tabla 3.1 Variables seleccionadas para la construcción de los modelos de MD.

Variable	Tipo	Escala	Calificación
Edad	Independiente	Cuantitativa	Años
Sexo	Independiente	Cualitativa	M, F
Alimentación	Independiente	Cualitativa	Buena, Mala, Regular
Duración diabetes	Independiente	Cuantitativa	Años
IMC	Independiente	Cuantitativa	
Tipo hipoglucemiante	Independiente	Cualitativa	Oral, Insulina
Tipo hipoglucemiante oral	Independiente	Cualitativa	Metformina, Glibenclamida, Glipizid, Glimepirid, Repaglinida, Nateglinida, Pioglitazon, Acarbosa, Exenatida, Sitagliptina, Vildagliptina
Tipo insulina:	Independiente	Cualitativa	Lispro, Aspart, Glulisina, NPH, Detemir, Glargina, Rápida
Hipertensión Arterial	Independiente	Cualitativa	Sí, No
Presión arterial sistólica Promedio	Independiente	Cuantitativa	mm/hg
Presión arterial sistólica Máxima	Independiente	Cuantitativa	mm/hg
Presión arterial diastólica Promedio	Independiente	Cuantitativa	mm/hg
Presión arterial diastólica Máxima	Independiente	Cuantitativa	mm/hg
No. Consultas oftalmólogo al año	Independiente	Cuantitativa	Número
No. Embarazos	Independiente	Cuantitativa	
Tabaco	Independiente	Cualitativa	Sí, No
Alcohol	Independiente	Cualitativa	Sí, No
HDL Promedio	Independiente	Cuantitativa	mg/dl
HDL Máxima	Independiente	Cuantitativa	mg/dl
LDL Promedio	Independiente	Cuantitativa	mg/dl

Variable	Tipo	Escala	Calificación
LDL Máxima	Independiente	Cuantitativa	mg/dl
Triglicéridos Promedio	Independiente	Cuantitativa	mg/dl
Triglicéridos Máxima	Independiente	Cuantitativa	mg/dl
Colesterol Promedio	Independiente	Cuantitativa	mg/dl
Colesterol Máxima	Independiente	Cuantitativa	mg/dl
Anemia	Independiente	Cualitativa	Sí, No
Hemoglobina Promedio	Independiente	Cuantitativa	g/dl
Hemoglobina Máxima	Independiente	Cuantitativa	g/dl
RDW Promedio	Independiente	Cuantitativa	%
RDW Máxima	Independiente	Cuantitativa	%
Plaquetas promedio	Independiente	Cuantitativa	mcL
Plaquetas máximo	Independiente	Cuantitativa	mcL
Leucocitos	Independiente	Cualitativa	mcL
Leucocitos	Independiente	Cualitativa	mcL
Glucosa Promedio	Independiente	Cuantitativa	mg/dl
Glucosa Máxima	Independiente	Cuantitativa	mg/dl
HbA1c Promedio	Independiente	Cuantitativa	%
HbA1c Máxima	Independiente	Cuantitativa	%
Nitrógeno Ureico promedio	Independiente	Cuantitativa	mg/dl
Nitrógeno Ureico máximo	Independiente	Cuantitativa	mg/dl
Creatinina sérica promedio	Independiente	Cuantitativa	mg/dl
Creatinina sérica máxima	Independiente	Cuantitativa	mg/dl
ALT Promedio	Independiente	Cuantitativa	ui/l
ALT Máxima	Independiente	Cuantitativa	ui/l
AST Promedio	Independiente	Cuantitativa	ui/l
AST Máxima	Independiente	Cuantitativa	ui/l
Proteína promedio	Independiente	Cuantitativa	mg/dl
Proteína máxima	Independiente	Cuantitativa	mg/dl
Proteína inicial	Independiente	Cuantitativa	mg/dl
Albumina promedio	Independiente	Cuantitativa	mg/dl
Albumina máxima	Independiente	Cuantitativa	mg/dl
Ácido Úrico promedio	Independiente	Cuantitativa	mg/dl
Ácido Úrico máximo	Independiente	Cuantitativa	mg/dl

Variable	Tipo	Escala	Calificación
Neuropatía Diabética	Independiente	Cualitativa	Sí, No
Antecedentes familiares nefropatía	Independiente	Cualitativa	Sí, No
Pie de Charcot	Independiente	Cualitativa	Sí, No
Insuficiencia arterial	Independiente	Cualitativa	Sí, No
Enfermedad vascular periférica	Independiente	Cualitativa	Sí, No
Hongo uñas	Independiente	Cualitativa	Sí, No
Retinopatía	Dependiente	Cualitativa	Sí, No
Nefropatía	Dependiente	Cualitativa	Sí, No
Pie Diabético	Dependiente	Cualitativa	Si, No

Una vez definidas las variables se procedió a obtener los datos a partir de los expedientes clínicos de los pacientes. Cabe mencionar que una de las mayores dificultades por la que se atravesó en el desarrollo de este proyecto fue la obtención de información real. Esto debido principalmente a que en el Hospital Regional de Río Blanco, como en la mayoría de los hospitales públicos en México, no cuenta con sistemas que implementen el expediente clínico electrónico (ECE), por lo que la recolección de datos se realizó de forma manual a partir de la revisión de expedientes físicos, lo que limitó el tamaño de la población disponible. La revisión de un expediente en algunas ocasiones llegó a tardar hasta una hora y aun así presentar información incompleta. Otras dificultad que se encontró es el cambio de personal médico, que ocasiona que existan diferentes formatos, estilos de redacción de notas médicas y pruebas de laboratorio. Por último, se presentaron muchos expedientes clínicos incompletos debido a que existe una gran cantidad de pacientes que acuden de forma irregular al médico. Todos estos factores originan una irregular e incompleta información clínica lo que se hace difícil obtener información valiosa y confiable de los datos.

Toda la información se capturó digitalmente en un documento de Excel haciendo uso de las cédulas de recolección de datos. Finalmente el documento de Excel se importó a una base de datos para su posterior manipulación. En la Tabla 3.2 se observa la distribución de la

población obtenida después de la captura de información de los expedientes clínicos. La mayor parte de la población padece de nefropatía diabética con un 45%, seguido de diabéticos sin ninguna complicación con un 40%. La presencia de retinopatía diabética es inferior con un 25% y de pie diabético apenas alcanza el 17%.

Tabla 3.2 Distribución de la población

	Número de Pacientes	Distribución
Diabetes Tipo II sin complicación	83	40%
Diabetes Tipo II + Pie diabético	34	17%
Diabetes Tipo II + Retinopatía	52	25%
Diabetes Tipo II + Nefropatía	92	45%
Total de Pacientes	**205**	

3.2.2 Pre procesamiento

En esta fase se dio tratamiento a los datos faltantes, erróneos e irrelevantes con la finalidad de que no causaran ruido en la construcción de los modelos de minería de datos. Se eliminaron todas aquellas variables que tuvieran más del 15% de datos faltantes, en la Tabla 3.3 se observan las variables que fueron eliminadas de acuerdo al criterio mencionado; de igual forma, el médico consideró factible que a algunas variables se les asignará el promedio de los valores para solucionar los datos perdidos, dichas variables se observan en el tercer renglón de la Tabla 3.3. Por otro lado, las variables que se observan en el segundo renglón de la Tabla 3.3 son aquéllas en las que el médico recomendó dejarlas con valores nulos al considerar que el promedio no era la mejor técnica para reemplazar estos valores debido a la naturaleza variable y aleatoria que estos atributos tienen. Finalmente 5 expedientes que poseían más del 7% de información faltante se eliminaron. Tras todo este pre procesamiento se obtuvo como resultado una población con 199 pacientes y 32 atributos los cuales se ilustran en el cuarto renglón de la Tabla 3.3 y estas variables son las que se utilizaron para entrenar los modelos.

Tabla 3.3 Pre procesamiento de las columnas

Variables	Descripción
No. Consultas Oftalmólogo, Insuficiencia Arterial, Antecedentes familiares nefropatía, Pie Charcot, Enfermedad Vascular Periférica Hongo uñas, HbA1c Promedio, HbA1c Máxima, LDL Promedio, LDL Máximo, HDL Promedio, HDL Máximo, Proteína Promedio, Proteína Máxima, Proteína Inicial, Albumina Máxima, Ácido Úrico Máximo, Albumina Promedio, Ácido Úrico Promedio, AST Promedio, AST Máxima, ALT Promedio, ALT Máxima, Triglicéridos Promedio, Triglicéridos Máximo, Colesterol Promedio, Colesterol Máximo, Alimentación, Hipoglucemiante Oral, Insulina	Variables Eliminadas
Tabaco, Alcohol, Tipo Hipoglucemiante, Nitrógeno Ureico Promedio, Nitrógeno Ureico Máximo, Creatinina Sérica Promedio Creatinina Sérica Máxima, Glucosa Promedio, Glucosa Máxima	Valor Nulo
RDW Máxima, Plaquetas Máxima, Leucocitos Máximo, Hemoglobina Máxima, RDW Promedio, Leucocitos Promedio, Plaquetas Promedio, Hemoglobina Promedio, No. Embarazos, IMC Duración Diabetes, Edad	Valor Promedio
Edad, Sexo, Duración Diabetes, IMC, Tipo Hipoglucemiante, Hipertensión, TAS Promedio, TAS Máxima, TAD Promedio, TAD Máxima, No Embarazos, Tabaco, Alcohol, Anemia, Hemoglobina Promedio, Hemoglobina Máxima, RDW Promedio, RDW Máximo, Plaquetas Promedio, Plaquetas máximo, Leucocitos promedio, Leucocitos máximo, Glucosa Promedio, Glucosa Máxima, Nitrógeno Ureico Promedio, Nitrógeno Ureico Máximo, Creatinina Sérica Promedio, Creatinina Sérica Máxima, Neuropatía Diabética, Nefropatía, Retinopatía, Pie Diabético	Variables Seleccionadas

3.2.4 Minería de Datos e Interpretación de resultados

El algoritmo de clasificación que se aplicó a los modelos de minería de datos fue el árbol de decisión C4.5; esta elección se hizo con base en los resultados positivos que este algoritmo dio en el análisis que se realizó en el capítulo 2, además de que se adapta perfectamente al conjunto de datos seleccionados ya que permite trabajar con valores nominales, numéricos y nulos. Cabe mencionar que se realizaron pruebas con otros algoritmos, como Naive-Bayes, pero los resultados fueron menos óptimos que en el caso de los árboles.

Para evaluar la precisión de los modelos, se tomaron en cuenta tres parámetros: precisión, sensibilidad y especificidad. La precisión se refiere al porcentaje de instancias que están correctamente clasificadas, la sensibilidad al porcentaje de predecir correctamente a un

paciente enfermo como enfermo y la especificidad al porcentaje de predecir a un paciente sano como tal.

Como técnica principal de pruebas se utilizó validación cruzada, ya que esta permite garantizar que el conjunto de datos que se utiliza para entrenar el modelo es independiente del conjunto de datos que se utiliza para probarlo. Para cada modelo en este estudio se utilizó validación cruzada con un valor de 10 iteraciones debido a que es uno de los valores típicos que se emplean en trabajos de minería [4], [31], [32], [34], [45], [5], [36] y es el valor por defecto de la herramienta Weka [21]. La razón por la que no se dividió la población en un conjunto para entrenamiento y un conjunto para pruebas es que se cuenta con una población pequeña para el número de variables (199 pacientes), por lo que no es recomendable su división, es decir, al seleccionar un 60% de la población para entrenamiento y un 40% para pruebas, por ese 40% de la muestra se estarían quitando características importantes de la población, esta es una situación muy común como se reportó en el estado del arte.

3.2.4.1 Minería de datos e interpretación de resultados para el modelo de nefropatía diabética

El conjunto de datos que se utilizó para la construcción del modelo de nefropatía diabética es el que se generó en la etapa de pre procesamiento que consta de 199 instancias y 32 atributos. La variable clase que se desea predecir en este conjunto de datos es la variable nefropatía diabética (Sí presenta/ No presenta). En la Tabla 3.4 se muestra la distribución de la población de nefropatía diabética; como se observa, la población en este modelo se encuentra bastante balanceada, contando con un 46.23% de pacientes con nefropatía y un 53.77% sin ésta.

Tabla 3.4 Distribución de la población para el modelo de nefropatía diabética

	Número de Pacientes	Distribución
Nefropatía diabética	92	46.23%
Sin nefropatía diabética	107	53.77%
Total	199	100%

Para la construcción de este modelo se realizaron diversas pruebas, a continuación se describen solo aquellas que mostraron los resultados más relevantes. Al menos que se indique lo contrario, todas las técnicas de minería de datos se utilizaron con sus parámetros por defecto.

Primeramente se aplicó el algoritmo C4.5 sobre la población descrita anteriormente, en la Tabla 3.5 se observa que se obtuvo una precisión del 83.91% con una sensibilidad de 78.3% y una especificidad de 88.8%. Dónde: la precisión significa que el 83.91% de la población se clasificó o se predijo correctamente, la sensibilidad significa que el modelo predijo correctamente al 78.3% de los pacientes que tienen nefropatía diabética y la especificidad significa que el modelo predijo correctamente al 88.8% de los pacientes que no tienen esta complicación. La sensibilidad aquí es un factor muy importante ya que el principal objetivo es identificar aquellos pacientes con altas probabilidades de padecer nefropatía diabética. Con los resultados obtenidos se observa que existe una diferencia aproximada del 10% entre la sensibilidad y especificidad.

Tabla 3.5 Resultado de aplicar el algoritmo C4.5 sobre la población original con 32 atributos

Datos	Algoritmo	Precisión	Sensibilidad	Especificidad
Población Original 32 atributos	C4.5	83.9196%	78.3%	88.8%

Algo que se busca al momento de realizar tareas de clasificación es encontrar las variables más relevantes para predecir la variable clase con el fin de mejorar la predicción y reducir el número de variables al mínimo necesario para realizar la predicción, como en el resultado anterior se empelaron los 32 atributos para construir el modelo se decidió volver a entrenarlo pero esta vez utilizando las variables más influyentes en la predicción por lo que se volvió a aplicar el algoritmo C4.5 pero sólo sobre las 9 variables seleccionadas por el evaluador `CfsSubsetEval`, con el método `BestFirst`, en la Tabla 3.6 se observa que se consiguió incrementar la precisión aproximadamente en un 3%, al igual que la sensibilidad en un 4.3% superando la barrera del 80%.

Tabla 3.6 Resultado de aplicar el algoritmo C4.5 sobre la selección de características del evaluador CfsSubsetEval

Datos	Algoritmos	Precisión	Sensibilidad	Especificidad
Población Original 9 atributos	C4.5 `CfsSubsetEval - BestFirst`	86.9347%	82.6%	90.7%

Continuando la búsqueda de las variables más influyentes se decidió cambiar de técnica de evaluación de características por lo que se aplicó el algoritmo C4.5 sobre las primeras 15 variables seleccionadas por el evaluador `InfoGainAttributeEval` con el método `Ranker`, en la Tabla 3.7 se observa que esta vez se obtuvieron mejores resultados con un aumento en la precisión del 4% logrando un total de 88.94% y un buen incremento en la sensibilidad de casi el 5% logrando un 87%, la especificidad se mantuvo igual que el resultado anterior.

Tabla 3.7 Resultados de aplicar el algoritmo C4.5 sobre la selección de características del evaluador *InfoGainAttributeEval*

Datos	Algoritmo	Precisión	Sensibilidad	Especificidad
Población Original 15 atributos	C4.5 *InfoGainAttributeEval - Ranker*	88.9447%	87%	90.7%

Intentando optimizar este resultado se optó por analizar la estructura del árbol resultante de la Tabla 3.7, se observaron las variables que se involucraron en la construcción del mismo, posteriormente se eliminaron aquéllas que no eran incluidas en su estructura y, bajo la recomendación del médico sobre la variables que consideraba más significativas, se aplicó nuevamente el algoritmo C4.5 realizando diversas pruebas con distintas combinaciones de variables, finalmente se obtuvo que el mejor resultado se generaba al utilizar el árbol C4.5 sobre un conjunto de 6 variables, en la Tabla 3.8 se muestra como se incrementó la predicción en aproximadamente un 1.5% logrando una precisión del 90.45 % y una sensibilidad del 89.1% y una especificidad del 91.6%.

Tabla 3.8 Resultados de aplicar el algoritmo C4.5 sobre una población de 6 atributos.

Datos	Algoritmo	Precisión	Sensibilidad	Especificidad
Población Original 6 atributos	C4.5	90.4523%	89.1%	91.6%

Bajo este panorama se decidió optimizar la sensibilidad del resultado anterior con la meta de alcanzar el 90% de sensibilidad, por lo que se utilizó el algoritmo *Cost Sensitive* tomando como base el algoritmo C4.5 y variando los parámetros de la matriz de costos para asignarle una ligera mayor importancia a clasificar correctamente la sensibilidad, finalmente en la Tabla 3.9 se muestra que se logró incrementar la sensibilidad en un 2.2% obteniendo un 91.3%, la precisión también se mejoró alcanzando un 91.95% y la especificidad en un 92.5%. Este último modelo fue el que se decidió utilizar para realizar la predicción de nefropatía diabética ya que fue el mejor resultado que se logró en este estudio.

Tabla 3.9 Resultado de aplicar el algoritmo *Cost Sensitive* para mejorar la sensibilidad del algoritmo C4.5 sobre una población de 6 atributos

Datos	Algoritmo	Precisión	Sensibilidad	Especificidad
Población Original 6 atributos	*Cost Sensitive* - C4.5	91.95%	91.30%	92.50%

Cabe recordar que este modelo utilizó para su entrenamiento y prueba la técnica de validación cruzada con 10 iteraciones por lo que al final se decidió probar este modelo sobre el conjunto de datos original descrito en la Tabla 3.4, en la Tabla 3.10 se observa que se obtuvieron buenos resultados logrando una precisión del 94.47%, una sensibilidad del 93.5% y una especificidad del 95.3%.

Tabla 3.10 Resultados de evaluar el modelo de nefropatía diabética elegido sobre la población original con 6 atributos

Datos	Precisión	Sensibilidad	Especificidad
Población Original 6 atributos	94.47%	93.50%	95.30%

3.2.4.2 Minería de datos para el modelo de retinopatía diabética

El conjunto de datos que se utilizó para la construcción del modelo de retinopatía diabética es el que se generó en la etapa de pre procesamiento que consta de 199 instancias y 32

atributos. La variable clase que se desea predecir en este conjunto de datos es la variable retinopatía diabética. En la
Tabla **3.11** se muestra la distribución de la población de retinopatía, como se observa sólo el 23.62% de la población presenta esta complicación y el resto no, por lo que la variable clase de retinopatía se encuentra bastante desbalanceada.

Tabla 3.11 Distribución de la población para el modelo de retinopatía diabética

	Número de Pacientes	Distribución
Retinopatía diabética	47	23.62%
Sin retinopatía diabética	152	76.38%
Total	199	100%

Como en el caso anterior, para la construcción de este modelo se realizaron diversas pruebas y, a menos que se indique lo contrario, todas las técnicas de minería de datos se utilizaron con sus parámetros por defecto.

Para el modelo de retinopatía diabética se siguieron los mismos pasos y se emplearon las mismas técnicas que para el modelo de nefropatía diabética, es decir, se utilizó el algoritmo C4.5 para entrenar el modelo, se utilizaron las técnicas de selección de características *InfoGainAttributeEval* y *CfsSubsetEval* para seleccionar las variables más relevantes, se usó el meta clasificador *Cost Sensitive* para dar mayor importancia a la predicción de la clase minoritaria, se analizó la estructura del árbol y se siguieron las recomendaciones del médico sobre las variables que consideraba más significativas, en la Tabla 3.12 se observa que, tras realizar todas esas pruebas, el mejor resultado se obtuvo al aplicar el algoritmo *Cost Sensitive* tomando como base el algoritmo C4.5 y variando los parámetros de la matriz de costos para asignarle una mayor importancia a clasificar correctamente la presencia de la enfermedad, se obtuvo una precisión del 84.42%, una sensibilidad del 72.30% y una especificidad de 88.20%. A pesar de que se emplearon las mismas técnicas de minería de datos esta vez no se obtuvieron tan buenos resultados como en el caso del modelo de nefropatía diabética, sólo se obtuvo el 72.3% de sensibilidad a pesar de utilizar el algoritmo *Cost Sensitive* para mejorarla. La razón por la que se obtuvo esta sensibilidad baja es a que

la población se encuentra desbalanceada, es decir, sólo el 23.62% de pacientes tienen retinopatía diabética y el 76.38% no la padecen. En tareas de clasificación, cuando una población se encuentra desbalanceada, el algoritmo tiende a clasificar correctamente las instancias de la clase mayoritaria e incorrectamente las instancias de la clase minoritaria tal y como ocurrió en este modelo.

Tabla 3.12 Resultado de aplicar el algoritmo *Cost Sensitive* para mejorar la sensibilidad del algoritmo C4.5 sobre una población de 13 atributos

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población Original 13 atributos	*Cost Sensitive* - C4.5	84.42%	72.30%	88.20%

Para combatir esta problemática del desbalanceo de clases se aplicó la técnica SMOTE (*Synthetic Minority Oversampling Technique*); con esta técnica se hizo un sobre muestreo para la clase minoritaria que solo tenía 47 valores y se igualó con la clase mayoritaria que posee 152 valores, con lo que se obtuvo una población de 304 pacientes donde 152 presentan retinopatía y los otros 152 no la presentan. Posteriormente se aplicó el algoritmo C4.5 sobre la población balanceada con sus 32 variables, en la Tabla 3.13 se observa como mejoraron los resultados, sobre todo el porcentaje de sensibilidad que era el que más interesaba mejorar, presentó un incremento de aproximadamente el 15.2% y se obtuvo una exactitud del 87.5%.

Tabla 3.13 Resultados de aplicar el algoritmo C4.5 sobre una población sobre muestreada con 32 atributos

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 32 atributos	C4.5	87.5%	87.5%	87.5%

Como en el resultado anterior se emplearon los 32 atributos para construir el modelo se decidió volver a entrenarlo pero esta vez utilizando las variables más influyentes en la predicción, por lo que aplicó el algoritmo C4.5 sobre las 12 variables seleccionadas por el evaluador *CfsSubsetEval* con el método *BestFirst*, en la Tabla 3.14 se observa que no se

consiguieron mejores resultados al emplear esta técnica, sino que, por el contrario, tanto la exactitud como la sensibilidad y especificidad disminuyeron..

Tabla 3.14 Resultado de aplicar el algoritmo C4.5 sobre la selección de características del evaluador CfsSubsetEval

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 12 Atributos	C4.5 CfsSubsetEval - BestFirst	85.5263%	86.2%	84.9%

Continuando la búsqueda de las variables más influyentes se decidió cambiar de técnica de evaluación de características por lo que se aplicó el algoritmo C4.5 pero sobre las primeras 24 variables seleccionadas por la técnica *InfoGainAttributeEval* con el método *Ranker*, en la Tabla 3.15 se observa que tampoco se obtuvieron mejores resultados al emplear esta técnica comparado con los resultados conseguidos en la Tabla 3.13.

Tabla 3.15 Resultados de aplicar el algoritmo C4.5 sobre la selección de características del evaluador InfoGainAttributeEval

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 24 atributos	C4.5 InfoGainAttributeEval - Ranker	86.8421%	84.9%	88.8%

Dado que las técnicas de selección de características no mejoraron los resultados se optó por analizar la estructura del árbol resultante de Tabla 3.13, se observaron las variables que se involucraron en la construcción del mismo, posteriormente se eliminaron aquéllas que no eran incluidas en su estructura y, bajo la recomendación del médico sobre las variables que consideraba más significativas, se aplicó nuevamente el algoritmo C4.5 realizando diversas pruebas con distintas combinaciones de variables, finalmente se obtuvo que el mejor resultado se generó al utilizar el árbol C4.5 con el parámetro configurado sin poda sobre un conjunto de 12 variables, en la Tabla 3.16 se muestra un incremento en la predicción de aproximadamente un 2.6% logrando una precisión del 90.13 % y mejorando notablemente la sensibilidad con un incremento del 6.5% logrando un 91.4%.

Tabla 3.16 Resultados de aplicar el algoritmo C4.5 sin poda para predecir retinopatía diabética sobre una población sobre muestreada de 12 atributos

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 12 atributos	C4.5 sin poda	90.1316%	91.4%	88.80%

Posteriormente se hicieron diversas pruebas aplicando el algoritmo *Cost Sensitive* tomando como base el algoritmo C4.5 y variando los parámetros de la matriz de costos para asignarle una ligera mayor importancia a clasificar correctamente la especificidad, finalmente en la Tabla 3.17 se muestra que se logró producir un incremento de aproximadamente un .7% en la sensibilidad y especificidad quedando con una precisión de 90.78%. Este último modelo fue el que se decidió utilizar para realizar la predicción de retinopatía diabética ya que fue el mejor resultado que se logró en este estudio.

Tabla 3.17 Resultados de aplicar el algoritmo *Cost Sensitive* para mejorar la especificidad del algoritmo C4.5 sin poda sobre una población sobre muestreada de 12 atributos

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 12 atributos	*Cost Sensitive* - C4.5 sin poda	90.78%	92.10%	89.50%

Es importante enfatizar que este modelo utilizó para su entrenamiento y prueba la técnica de validación cruzada con 10 iteraciones sobre una población donde se sobre muestreó la clase minoritaria obteniendo una población de 304 instancias por lo que al final se decidió probar este modelo sobre el conjunto de datos original descrito en la Tabla 3.11 que contiene sólo las 199 instancias originales, en la Tabla 3.18 se observa que se obtuvieron buenos resultados con una precisión de casi el 96%, una sensibilidad del 93.6% y una especificidad del 96.7%.

Tabla 3.18 Resultado de evaluar el modelo de retinopatía diabética sobre la población original con 12 atributos

Datos	Exactitud	Sensibilidad	Especificidad
Población Original 12 atributos	95.98%	93.60%	96.70%

3.2.4.3 Minería de datos para el modelo de pie diabético

El conjunto de datos que se utilizó para la construcción del modelo de pie diabético es el que se generó en la etapa de pre procesamiento que consta de 199 instancias y 32 atributos. La variable clase que se desea predecir en este conjunto de datos es la variable pie diabético. En la Tabla 3.19 se muestra la distribución de la población de pie diabético, como se observa se encuentra altamente desbalanceada ya que sólo el 17.1% de la población presenta esta complicación.

Tabla 3.19 Distribución de la población para el modelo de pie diabético

	Número de Pacientes	Distribución
Pie diabético	34	17.1%
Sin pie diabético	165	82.9%
Total	199	100%

Para el modelo de pie diabético se siguieron los mismos pasos y se emplearon las mismas técnicas que para el modelo de nefropatía diabética es decir, se utilizó el algoritmo C4.5 para entrenar el modelo, se utilizaron las técnicas de selección de características *InfoGainAttributeEval* y *CfsSubsetEval* para seleccionar las variables más relevantes, se usó el meta clasificador *Cost Sensitive* para dar mayor importancia a la predicción de la clase minoritaria, se analizó la estructura del árbol y se siguieron las recomendaciones del médico sobre las variables que consideraba más significativas, en la Tabla 3.20 se observa que tras realizar todas esas pruebas el mejor resultado se obtuvo el aplicar algoritmo *Cost Sensitive* tomando como base el algoritmo C4.5 y variando los parámetros de la matriz de costos para asignarle una mayor importancia a clasificar correctamente la sensibilidad, se obtuvo una precisión del 78.89% una sensibilidad del 67.6% y una especificidad del 81.2%. Como se aprecia en la tabla, se obtuvo una sensibilidad muy pobre debido al alto desbalanceo de clase puesto que los pacientes que padecen esta complicación apenas representan el 17.1%.

Tabla 3.20 Resultados de aplicar el algoritmo *Cost Sensitive* para mejorar la sensibilidad del algoritmo C4.5 sobre una población con 11 atributos

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población Original 11 atributos	*Cost Sensitive* - C4.5	78.89%	67.60%	81.20%

Como ya se discutió en el proceso de minería de datos del modelo de retinopatía diabética, para combatir esta problemática del desbalanceo de clases se aplicó la técnica de sobre muestreo SMOTE, con esta técnica se sobre muestreó la clase minoritaria que solo tenía 34 valores y se igualó con la clase mayoritaria que posee 165 instancias, con lo que se obtuvo una población de 330 pacientes donde 165 presentan pie diabético y los otros 165 no lo presentan. Posteriormente se aplicó el algoritmo C4.5 sobre la población balanceada con todas sus 32 variables, en la Tabla 3.21 se observa como mejoraron los resultados, sobre todo el porcentaje de sensibilidad que era el que más interesaba mejorar, presentó un incremento de aproximadamente el 17.2%, igual se aprecia que el porcentaje de la precisión superó el 80% quedando con un 85.45% y la especificidad con un 86.1%.

Tabla 3.21 Resultados de aplicar el algoritmo C4.5 sobre una población sobre muestreada de 32 atributos

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 32 atributos	C4.5	85.4545%	84.8%	86.1%

Como en el resultado anterior se emplearon los 32 atributos para construir el modelo, también se decidió reentrenarlo utilizando las variables más influyentes en la predicción por lo que se volvió a aplicar el algoritmo C4.5 pero ahora sólo sobre las 15 variables seleccionadas por el evaluador *CfsSubsetEval* con el método *BestFirst,* en la Tabla 3.22 se observa que no se consiguieron mejores resultados al emplear esta técnica.

Tabla 3.22 Resultados de aplicar el algoritmo C4.5 sobre la selección de características del evaluador CfsSubsetEval

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 15 atributos	C4.5 CfsSubsetEval - BestFirst	81.8182%	78.2%	85.5%

Continuando con la búsqueda de las variables más influyentes se decidió cambiar de técnica de evaluación de características, por lo que se aplicó el algoritmo C4.5 sólo sobre las primeras 28 variables seleccionadas por la técnica *InfoGainAttributeEval* con el método *Ranker*, como se observa la

Tabla **3.23** se consiguieron resultados similares a los obtenidos en la Tabla 3.21 pero con porcentajes ligeramente más bajos en la exactitud y sensibilidad y ligeramente superiores en la especificidad.

Tabla 3.23 Resultados de aplicar el algoritmo C4.5 para predecir pie diabético sobre la selección de características del evaluador *InfoGainAttributeEval*

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 28 atributos	C4.5 *InfoGainAttributeEval - Ranker*	85.1515%	83%	87.3%

Dado que las técnicas de selección de características no mejoraron los resultados, se optó por analizar la estructura del árbol resultante de la Tabla 3.21, se estudiaron las variables que se involucraron en la construcción del mismo, posteriormente se eliminaron aquéllas que no eran incluidas en su estructura y, bajo las recomendaciones del médico sobre la variables que consideraba más significativas, se aplicó el algoritmo C4.5 y se realizaron diversas pruebas con distintas combinaciones de variables, finalmente se obtuvo que el mejor resultado se generaba al utilizar aplicando el algoritmo *Cost Sensitive* tomando como base el algoritmo C4.5 con el parámetro configurado sin poda sobre un conjunto de 12 variables y variando los parámetros de la matriz de costos para asignarle una ligera mayor importancia a la especificidad, en la Tabla 3.24 se muestra como se mejoró la predicción en aproximadamente un 2.13% logrando una precisión del 87.58 %, una sensibilidad del 86.7% y una especificidad del 88.5%. Este último modelo fue el que se decidió utilizar para realizar la predicción de pie diabético ya que presenta los mejores índices en las tres áreas.

Tabla 3.24 Resultado de aplicar el algoritmo *Cost Sensitive* para mejorar la especificidad del algoritmo C4.5 sin poda sobre una población con 15 atributos

Datos	Algoritmo	Exactitud	Sensibilidad	Especificidad
Población SMOTE 15 atributos	*Cost Sensitive* - C4.5 sin	87. 58 %	86.70%	88.5%

	poda			

Cabe hacer mención de que este modelo utilizó para su entrenamiento y prueba la técnica de validación cruzada con 10 iteraciones sobre una población donde se sobre muestreó la clase minoritaria obteniendo una población de 330 instancias, dado lo cual se aplicó este modelo sobre el conjunto de datos descrito en la Tabla 3.19 que contiene sólo las 199 instancias originales, en la Tabla 3.25 se observa que se obtuvo una buena precisión con un 92.96% pero una baja sensibilidad del 76.5% y una especificidad del 96.7%. Este bajo porcentaje de sensibilidad se relaciona al hecho de que sólo se cuenta con 34 pacientes con pie diabético y, a pesar de que se hizo el sobre muestreo para construir un mejor modelo, las características originales de estos 34 pacientes no son suficientes para armar patrones de comportamiento que predigan a la mayoría de pacientes que presentan esta complicación.

Tabla 3.25 Resultados de evaluar el modelo de pie diabético sobre la población original de 15 atributos.

Datos	Exactitud	Sensibilidad	Especificidad
Población Original 15 atributos	92.96%	76.50%	96.40%

3.3 Diseño de la ontología utilizando la metodología Methontology

Para el desarrollo de la ontología OCPPD se utilizó la metodología Methontology, la cual contiene una serie de pasos y actividades que sirven como guía para la construcción de ontologías. En la Figura 3.5 se observan las cinco fases que componen a Methontology: Especificación, Conceptualización, Formalización, Implementación y Mantenimiento. Al igual que con los modelos de minería de datos, un médico del HRRB brindó su conocimiento para el desarrollo y validación de la ontología OCPPD. De igual forma se obtuvo asesoría sobre la construcción y la inferencia de conocimiento en las ontologías durante las estancias realizadas en la Universidad Carlos III de Madrid, en Leganés, Madrid, España.

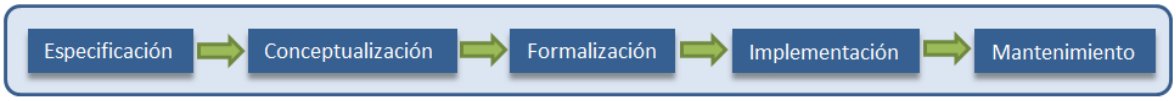


Figura 3.5 Fases de la metodología Methontology

Para el desarrollo de la ontología se utilizó Protégé, que es una herramienta para la creación, visualización y manipulación de ontologías. Protégé se empleó para la definición de clases, jerarquía de clases, propiedades, atributos e individuos entre otras cosas. También se usó un razonador genérico de reglas que viene implementado en la API de Jena, lo que permitió inferir el tratamiento para un paciente en particular.

3.3.1 Especificación

En esta fase se describieron los elementos que integran al documento de especificación de requerimientos de la ontología de cuidados preventivos.

Dominio: Cuidados para prevenir complicaciones en los pacientes con Diabetes Mellitus Tipo II

Propósito: El propósito de OCPPD es el proporcionar un vocabulario y estructura para utilizarse en el ámbito médico para proveer de cuidados preventivos tanto generales como personalizados al paciente diabético con el objetivo de retrasar/prevenir la aparición de complicaciones.

Alcance: El alcance de la ontología comprende todos los conceptos, atributos, relaciones, restricciones, individuos que son necesarios para que la ontología sea capaz de proveer de cuidados preventivos hacia el paciente diabético.

Fuentes de conocimiento:

(a) Médico experto

(b) Literatura médica (Guía IMSS, Guía ALAD, otras)

(c) Vocabulario (SNOMED-CT)

3.3.2 Conceptualización

En la Figura 3.6 se observan las tareas que se realizaron en la fase de Conceptualización. Cada una de estas tareas ayudaron a la construcción del modelo conceptual: términos, taxonomía, relaciones, axiomas, reglas.

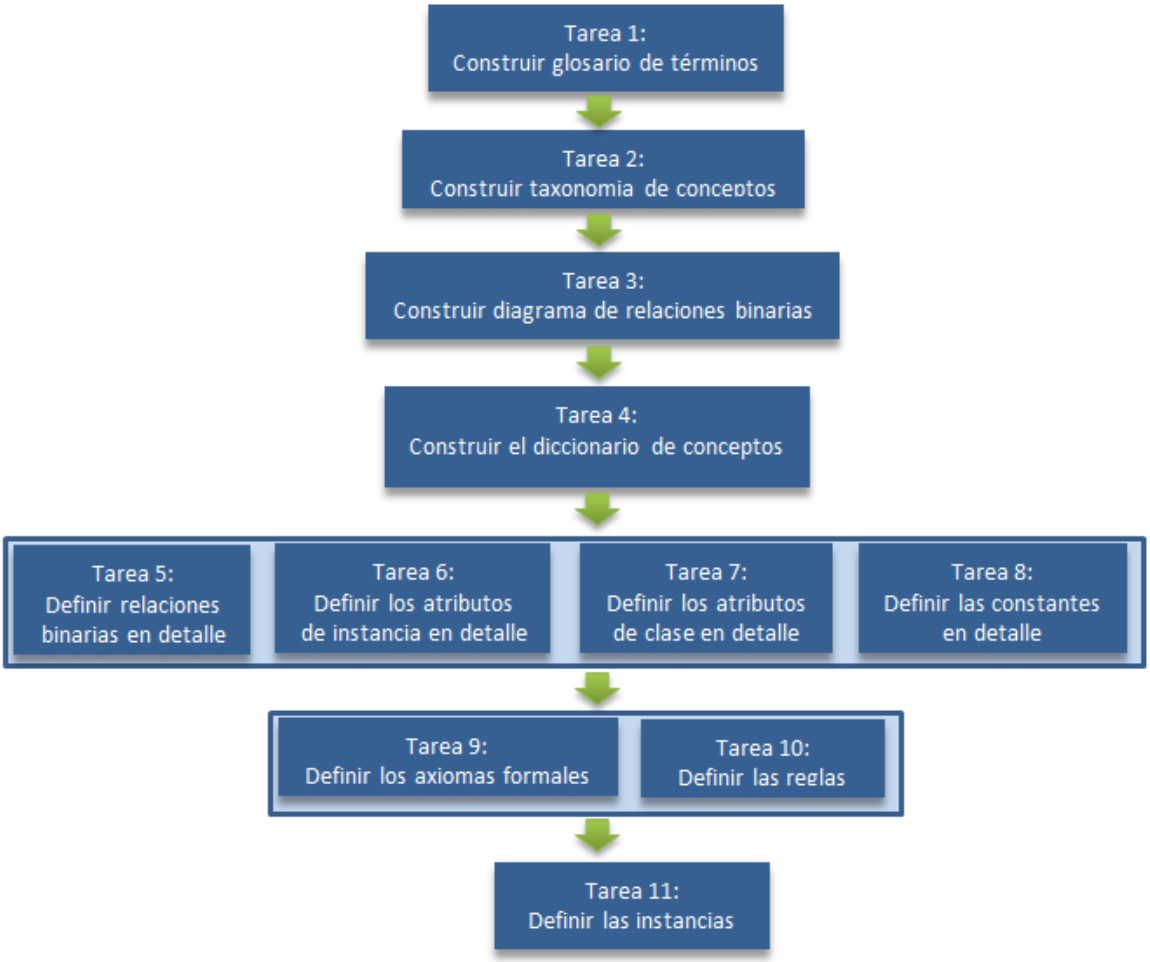


Figura 3.6 Tareas de la fase de Conceptualización

3.3.2.1 Glosario de términos

En esta tarea se identificaron todos los términos (conceptos, relaciones, instancias y atributos entre otros) importantes del dominio. Para la selección de los términos se contó con la asesoría de un médico del HRRB, además se realizó una revisión de diferentes artículos y guías de la práctica clínica referentes a la prevención de complicaciones micro vasculares en pacientes diabéticos. La terminología SNOMED-CT se utilizó como referencia para la construcción de la estructura jerárquica de la Ontología OCPPD.

SNOMED-CT es el vocabulario de mayor amplitud e importancia en el área biomédica, cuenta con cerca de un millón de términos asociados, por lo que básicamente cualquier concepto referente al ámbito médico es posible encontrarlo en este vocabulario. Como el principal objetivo de la ontología OCPPD es el de ser utilizada para inferir conocimiento, se decidió no incluir la definición, los sinónimos y acrónimos de los conceptos en esta ontología, en lugar de eso cada concepto de la ontología OCCPPD hace referencia al ID del vocabulario SNOMED-CT, por lo que si se requiere, se consulte la definición, sinónimos o acrónimos de los conceptos en dicho vocabulario. Esta decisión se tomó dado que se desea mantener la ontología OCCPPD lo más sencilla y ligera posible debido a que para realizar la inferencia de conocimiento la ontología se carga en memoria y, de ser muy grande, el tiempo en realizar el razonamiento será mayor. En la Tabla 3.26 se observa un resumen de la cantidad de términos utilizados en la presente ontología. Un total de 270 términos fueron necesarios para la construcción de esta, siendo entre ellos los conceptos y sus instancias los más numerosos. En la Tabla 3.27 se observa una parte del glosario de términos resultante.

Tabla 3.26 Número de términos de la ontología

Términos	Cantidad
Conceptos	75
Atributos	37
Relaciones	5
Instancias	153
TOTAL	270

Tabla 3.27 Glosario de términos

Nombre	Descripción	Tipo
Paciente	Una persona que recibe atención médica.	Concepto
Hallazgo Clínico	Observaciones médicas hechas al paciente.	Concepto
Enfermedad	Cualquier condición que afecte la salud del paciente.	Concepto
Monitorización del Estado Diabético	Tratamiento actual del paciente diabético.	Concepto

Nombre	Descripción	Tipo
Hallazgo en la Historia Clínica y Observaciones	Observaciones en la historia clínica del paciente.	Concepto
Farmacéutica Producto Biológico	Sustancia que se utiliza para el tratamiento o prevención de una enfermedad.	Concepto
Agente Hipoglucémico	Fármaco que disminuye la glucosa en la sangre.	Concepto
Hipoglucemiante Oral	Agente hipoglucémico que se administra por vía oral.	Concepto
Medición de Sustancia	Medir una sustancia.	Concepto
Tratamiento Farmacológico	Uso de fármacos para el tratamiento o prevención de una enfermedad	Concepto
Tratamiento No Farmacológico	Tratamiento o prevención de una enfermedad sin uso de fármacos	Concepto
Alimentación	Recomendación nutricional para mejorar la salud y el bienestar de la persona	Concepto
Ejercicio Físico	Recomendación de alguna actividad física para mejorar la salud y el bienestar de la persona	Concepto
Posee	Un paciente posee una medición de sustancia.	Relación
Requiere	Un paciente requiere de una intervención médica o de un fármaco.	Relación
Tiene	Un paciente tiene hallazgos clínicos.	Relación
Utiliza	Un paciente utiliza un fármaco.	Relación
Altura	Estatura de una persona expresada en centímetros.	Atributo
Peso	Peso de una persona expresada en centímetros.	Atributo
Edad	Tiempo transcurrido desde el nacimiento de una persona.	Atributo
Calorías	Cantidad de calorías necesarias en la alimentación diaria.	Atributo
Recomendación	Consejo médico para mejorar la salud.	
Diabetes Mellitus Tipo II	Un tipo de diabetes mellitus que se caracteriza por el aumento de los niveles de glucosa en sangre.	Instancia
Hipertensión Arterial	Enfermedad que se caracteriza por la elevación persistente de la presión arterial.	Instancia
IMC_Normal	Índice de Masa Corporal dentro de los parámetros normales.	Instancia

Nombre	Descripción	Tipo
Estilo de Vida Sedentario	Estilo de vida sedentario que no realiza ejercicio físico.	Instancia
Sexo Femenino	Sexo biológico femenino.	Instancia
Diabetes en Tratamiento con Dieta	Diabetes que se encuentra en tratamiento con un régimen alimenticio.	Instancia
Glucosa Alta	Glucosa en sangre con niveles altos.	Instancia
Triglicéridos Normal	Triglicéridos en sangre dentro de los parámetros normales.	Instancia
Colesterol Normal	Colesterol en sangre dentro de los parámetros normales.	Instancia
HO Biguanidas	Hipoglucemiante oral perteneciente a la familia de las biguanidas.	Instancia
HO Sulfonilureas	Hipoglucemiante oral perteneciente a la familia de las sulfonilureas.	Instancia
IBA Detemir	Insulina basal análoga	Instancia
Dieta Limitar Colesterol	Recomendación alimenticia que sugiere limitar el consumo de alimentos con alto contenido de colesterol.	Instancia
Dieta Restringir Sodio Hipertenso	Recomendación alimenticia que sugiere restringir el consumo de sodio en los alimentos.	Instancia
Ejercicio Aeróbico	Recomendación de un tipo de actividad física que exige la demanda continúa de oxígeno al cuerpo.	Instancia
Ejercicio Suave	Recomendación de ejercicio físico a una intensidad suave, donde el ritmo cardiaco requerido es bajo.	Instancia
Ejercicio Contraindicado	Recomendación de no realizar ejercicio.	Instancia
No Fumar	Recomendación de no inhalar tabaco u otra sustancia nociva para la salud.	Instancia

3.3.2.2 Taxonomía de conceptos

En esta tarea se estableció la jerarquía entre los conceptos del dominio. El vocabulario SNOMED-CT se utilizó como base para la construcción de la jerarquía. En la Tabla 3.28 se expone la jerarquía con algunos de los conceptos más relevantes de la ontología. Como se observa, existen 5 conceptos raíz de donde se desglosa toda la jerarquía, estos son: *Contexto social, farmacéutica producto biológico, hallazgo clínico, procedimiento* e *intervención.* Con este tipo de jerarquía se deduce que un *Paciente* es una *Persona en el entorno de la salud* y que también es una *Persona* y que a su vez también es un *Contexto social*. Con esta clasificación se generan relaciones como por ejemplo un *Paciente tiene_un Hallazgo clínico* donde hallazgo clínico es una enfermedad o un dato general como *Sexo biológico* o *Edad*. De igual manera, esta estructura permite realizar consultas si así se requiriera, como por ejemplo consultar todos los fármacos existentes o incluso ser más específicos y sólo consultar todos los fármacos que son hipoglucemiantes orales.

Tabla 3.28 Taxonomía de conceptos

Contexto social
Persona
Persona en el entorno de la salud
Paciente
Farmacéutica producto biológico
Preparación antidiabética
Agente hipoglucémico
Hipoglucemiante oral
Insulina
Hallazgo Clínico
Enfermedad
Trastorno por sitio del cuerpo
Trastorno del sistema corporal
Trastorno del sistema cardiovascular
Trastorno del sistema endocrino
Hallazgo en la historia clínica y observaciones
Hallazgo en el estado clínico general
Sexo biológico
Hallazgo general de la observación del paciente
Edad
Procedimiento
Procedimiento por método
Procedimiento de evaluación
Medición
Medición de sustancia

Medición de carbohidratos
Medición de glucosa
Medición de lípidos
Triglicéridos
Intervención
Tratamiento farmacológico
Control glucosa
Tratamiento no farmacológico
Alimentación
Ejercicio físico

3.3.2.3 Diagrama de relaciones binarias

El diagrama de relaciones binarias permite identificar las relaciones entre conceptos. En total se generaron 4 relaciones en la ontología.

En la Figura 3.7 se observa la relación que existe entre un *paciente* y una *medición de sustancia*, donde una medición de sustancia es una medición de colesterol o *una medición de triglicéridos* o de *glucosa* sólo por mencionar algunas. Bajo este panorama es posible representar a un paciente "Pedro" que posee una *medición de glucosa* "Glucosa alta" y además posee una medición de colesterol "Colesterol Alto".

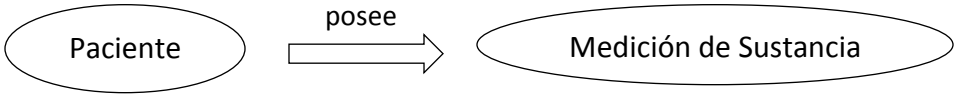


Figura 3.7 Relación Paciente – Medición Sustancia

En la Figura 3.8 se muestra la relación entre un *paciente* y un *hallazgo clínico*, un hallazgo clínico es una enfermedad, o un dato en la historia clínica del paciente: como la edad o el sexo biológico o *el hábito de hacer ejercicio* por mencionar algunos. Bajo este esquema continuando con el paciente "Pedro", además de sus mediciones de sustancia se le indica que tiene un *hallazgo clínico* "Hipertensión Arterial" además de otro *hallazgo clínico* "Estilo de vida sedentario" y otro *hallazgo clínico* "Diabetes en tratamiento con dieta".

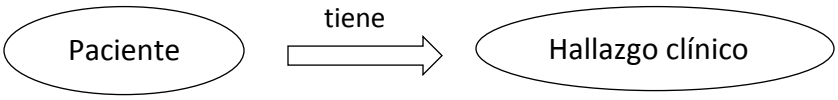


Figura 3.8 Relación Paciente – Hallazgo clínico

En la Figura 3.9 se ilustra la relación entre un *paciente* y un *producto farmacéutico*, un producto farmacéutico es un *hipoglucemiante oral* o una *insulina*. En este contexto, al paciente "Pedro" se le agrega que utiliza un *producto farmacéutico*, específicamente un *hipoglucemiante oral* "Biguanidas" y además utiliza otro *producto farmacéutico* esta vez una *insulina* "NPH".

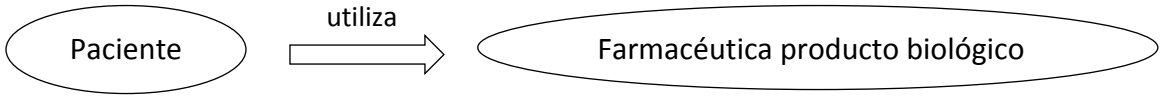


Figura 3.9 Relación Paciente – Farmacéutica producto biológico

En la Figura 3.10 se tiene una de las relaciones más importantes de la ontología ya que es la relación que se va a inferir automáticamente. Esta relación se da entre un paciente con un producto farmacéutico o entre un paciente con una intervención que se refiere al tratamiento farmacológico o no farmacológico del paciente. Continuando con el ejemplo de "Pedro" y dadas todas las relaciones descritas anteriormente se identifica que requiere de una *intervención*, un *tratamiento no farmacológico* de *alimentación* "Limitar grasas saturadas" debido a que posee el "Colesterol Alto". De igual forma se identifica que requiere de otra *intervención*, un *tratamiento no farmacológico* de *ejercicio físico* "Caminata" debido a que posee un "Estilo de vida sedentario".

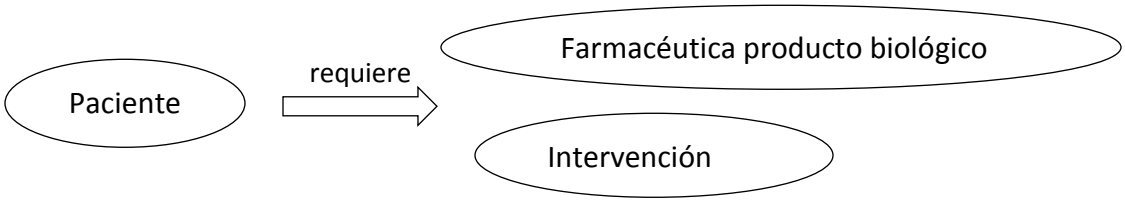


Figura 3.10 Relación Paciente – Farmacéutica producto biológico - Intervención

3.3.2.4 Diccionario de conceptos

En la Tabla 3.29 se observa el diccionario de conceptos donde se enlistan los conceptos con sus relaciones correspondientes. Estas se obtienen de los diagramas de relaciones binarias descritos anteriormente. Como se observa, el concepto paciente se involucra en todas las relaciones, esto es debido a que, con base en su historia clínica, los fármacos que consume y los resultados de las pruebas de laboratorio, se le recomienda su tratamiento.

Tabla 3.29 Diccionario de conceptos

Concepto	Relaciones
Paciente	- posee - requiere - tiene - utiliza
Medición de sustancia	- posee
Intervención	- requiere
Farmacéutica producto biológico	- requiere - utiliza
Hallazgo clínico	- tiene

3.3.2.5 Relaciones binarias en detalle

En esta tarea se identificaron los detalles de las relaciones binarias como dominio, rango y cardinalidad. En la Tabla 3.30 se observa el detalle de relaciones binarias, la cardinalidad entre un *paciente* y una *medición de sustancia* se interpreta de la siguiente manera: uno o muchos pacientes poseen ninguna o muchas mediciones de sustancia.

Tabla 3.30 Detalle de relaciones binarias

Relación	Dominio	Rango	Cardinalidad
posee	Paciente	Medición de sustancia	1:* .. 0:*
requiere	Paciente	Intervención	1:*.. 1:*
requiere	Paciente	Farmacéutica producto biológico	1:*.. 0:*
tiene	Paciente	Hallazgo clínico	1:* .. 1:*
utiliza	Paciente	Farmacéutica producto biológico	1:* .. 0:*

3.3.2.6 Atributos de instancia en detalle

En esta tarea se identificaron todos los atributos que posee una instancia de un concepto. Cada atributo cuenta con tipo de valor, rango de valores y cardinalidad. En la Tabla 3.31 se muestran algunos de los atributos de instancia más relevantes, la gran mayoría pertenecen al concepto *paciente* como *peso*, *altura*, resultados de laboratorio (*glucosa*, *colesterol*), recomendaciones alimenticias (*calorías*, *carbohidratos*), entre otros. Pero también conceptos como *hipoglucemiante oral* tienen sus atributos como *efectos secundarios* o *compuestos*. Otro atributo importante es el de *recomendación* que pertenece al concepto *intervención*, en este atributo se describe la recomendación que se le da al paciente, por ejemplo en la instancia *caminata* que pertenece al concepto *ejercicio físico -> tratamiento no farmacológico -> intervención* (ver jerarquía de conceptos) la recomendación correspondiente es "comenzar con caminata durante 10 minutos al día…".

Tabla 3.31Atributos de instancia en detalle

Atributo	Concepto	Tipo de Valor	Rango de Valores	Cardinalidad
Altura	Paciente	Numérico	Flotante	1:1
Calorías	Paciente	Numérico	Entero	1:1
Carbohidratos	Paciente	Numérico	Entero	1:1
Compuestos	Hipoglucemiante Oral	Cadena de Caracteres	Letras	1:1
Colesterol LDL	Paciente	Numérico	Flotante	0:1
Edad	Paciente	Numérico	Entero	1:1
Efectos Secundarios	Hipoglucemiante Oral	Cadena de Caracteres	Letras	1:1
Glucosa	Paciente	Numérico	Entero	0:1
Peso	Paciente	Numérico	Flotante	1:1
Peso ideal	Paciente	Numérico	Flotante	1:1
Recomendación	Intervención	Cadena de Caracteres	Letras	1:1
Valor Máximo Umbral	Medición Sustancia	Numérico	Flotante	1:1
Valor Máximo Umbral	Medición Sustancia	Numérico	Flotante	1:1

3.3.2.7 Reglas

El uso de reglas permite realizar el razonamiento automático de la ontología. Las reglas fueron creadas siguiendo el formato de "Jena Rules" debido a que el motor de inferencia que se utiliza pertenece a la API de Jena. Se emplearon tanto reglas con encadenamiento hacia adelante como reglas con encadenamiento hacia atrás (el razonador genérico de Jena permite trabajar híbridamente con ambos tipos). Un total de 146 reglas se emplearon para brindarle a la ontología la capacidad de realizar inferencias. En la Tabla 3.32 se muestran algunas de estas reglas junto con su descripción.

Tabla 3.32 Reglas de la ontología

Nombre	**rule_r70_ejercicio_sedentario**
Descripción	Recomienda realizar ejercicio como caminata a un paciente que cumple con las siguientes condiciones: padece diabetes mellitus tipo II, no tiene cuerpos cetónicos positivos, no tiene la glucosa baja y tiene un estilo de vida sedentario

Expresión	`(?i ont:tiene ont:Diabetes_Mellitus_Tipo_II)`
	`noValue(?i ont:posee`
	`ont:Cuerpos_Cetonicos_Positivos)`
	`noValue(?i ont:posee ont:Glucosa_Baja)`
	`(?i ont:tiene`
	`ont:Ejercicio_Estilo_de_Vida_Sedentario) ->`
	`(?i ont:requiere ont:Ejercicio_Caminata)`

Conceptos	Hallazgo Clínico, Procedimiento, Intervención
Nombre	**rule_r84_tratamiento_monoterapia_glucosa**
Descripción	Provee de tratamiento con monoterapia oral a un paciente que cumple con las siguientes condiciones: padece diabetes mellitus tipo II, tiene glucosa en descontrol, sin registro de medición de HbA1c, con una glucosa superior o igual a 131, con diabetes en tratamiento con dieta y que no esté bajo tratamiento de insulina.

Expresión	`(?i ont:requiere ont:TF_CG_con_Monoterapia_Oral)`
	`<-`
	`(?i ont:tiene ont:Diabetes_Mellitus_Tipo_II)`
	`noValue(?i, ont:HbA1c ?h)`
	`(?i ont:Glucosa ?glu)`
	`ge(?glu, 131)`
	`(?i ont:tiene`

	ont:Diabetes_en_Tratamiento_con_Dieta) noValue(?i, ont:variable ont:Requiere_Insulina_Basal_1)
Conceptos	Hallazgo Clínico, Procedimiento, Intervención

3.3.2.8 Instancias

Un total de 153 instancias se identificaron en esta tarea. En la Tabla 3.33 se muestran algunas de estas, con su nombre, nombre del concepto al que pertenece y los valores de sus atributos de instancia, en caso de que el campo de valores este vacío significa que esa instancia no cuenta con atributos.

Tabla 3.33 Instancias del dominio

Nombre	Concepto	Valores
Paciente X	Paciente	**Altura** = 164, **Peso** = 59, **IMC** = 22, **Edad** = 24, **Glucosa** = 90, **Triglicéridos** = 200, **Calorías** = 2200, **Carbohidratos** = 55, **Grasas** = 25, **Proteínas** = 20,
HO Biguanidas	Hipoglucemiante Oral	**Compuestos** = Metformina, **Contraindicaciones** = Insuficiencia Renal – Insuficiencia Respiratoria – Insuficiencia hepática -Alcoholismo, **Costo** = Bajo , **Efectos Secundarios** = Gastrointestinales, **Ventajas** = No hipoglucemia - no aumento de peso -reduce riesgo cardiovascular
IBA_Glargina	Insulina	**Pico de acción** = Ninguno, **Inicio de acción** = 2-4 horas, **Duración de la acción** = 20-24 horas
Glucosa Alta	Glucosa	**Riesgo Salud** = true, **Valor Máximo Umbral**=250, **Valor Mínimo Umbral**=191
Dieta evitar alcohol	Alimentación	**Descripción** = Alcohol, **Recomendación** = Evitar consumir bebidas alcohólicas
Ejercicio poco	Actividad Física	
Hipertensión Arterial	Trastorno del Sistema Cardiovascular	
Diabetes en tratamiento con dieta	Monitorización del estado diabético	

3.3.4 Implementación

En esta parte se realizó la fase de implementación mediante el uso de la herramienta Protégé. Esta herramienta permitió crear la ontología y guardarla en formato RDF/XML en

un archivo con extensión .owl. Para el desarrollo de esta fase se tomaron como base los resultados obtenidos en la fase de conceptualización.

Como primer punto, se desarrolló la taxonomía de los conceptos tomando como base lo definido en el apartado 3.3.2.2. En la Figura 3.11 se observa una parte de la taxonomía elaborada con la herramienta Protégé.

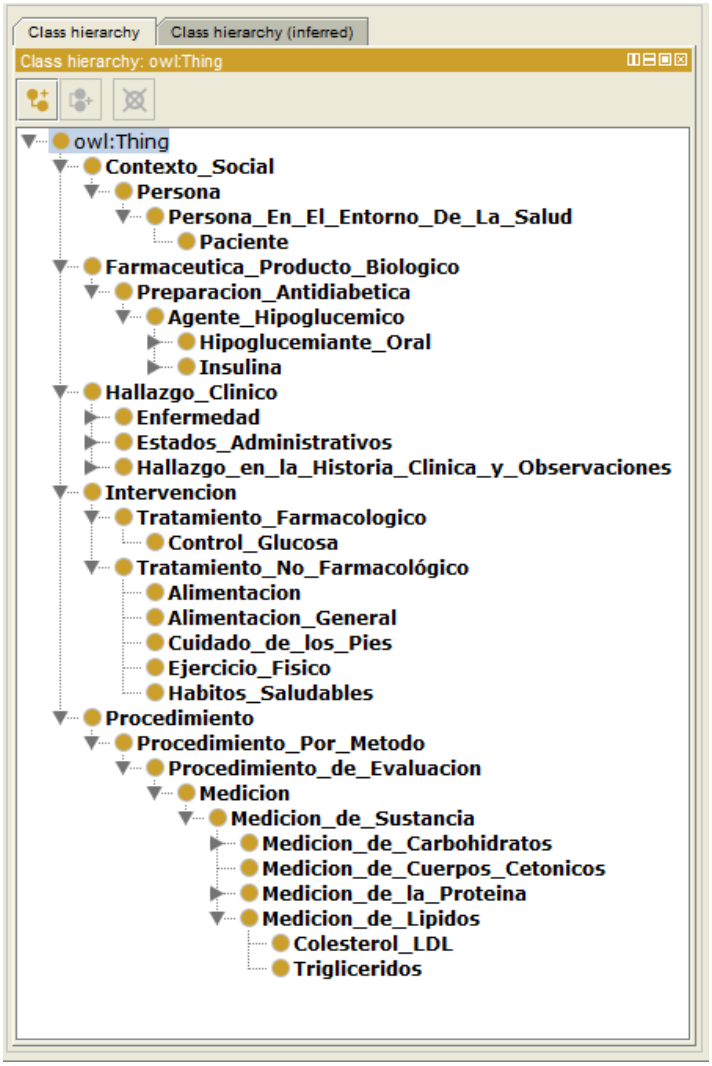


Figura 3.11 Taxonomía de conceptos creada con Protégé.

El siguiente paso fue crear las relaciones binarias descritas en la sección 3.3.2.5. Se asignaron el dominio y el rango sobre las relaciones creadas. En la Figura 3.12 se ilustra una parte de las relaciones de la ontología.

Posteriormente se crearon los atributos de las clases con base en lo descrito en la sección 3.3.2.6, para cada uno de ellos, se definió el tipo de dato o rango que manejan. En la Figura 3.13 se muestra una parte del listado de los atributos de la ontología.

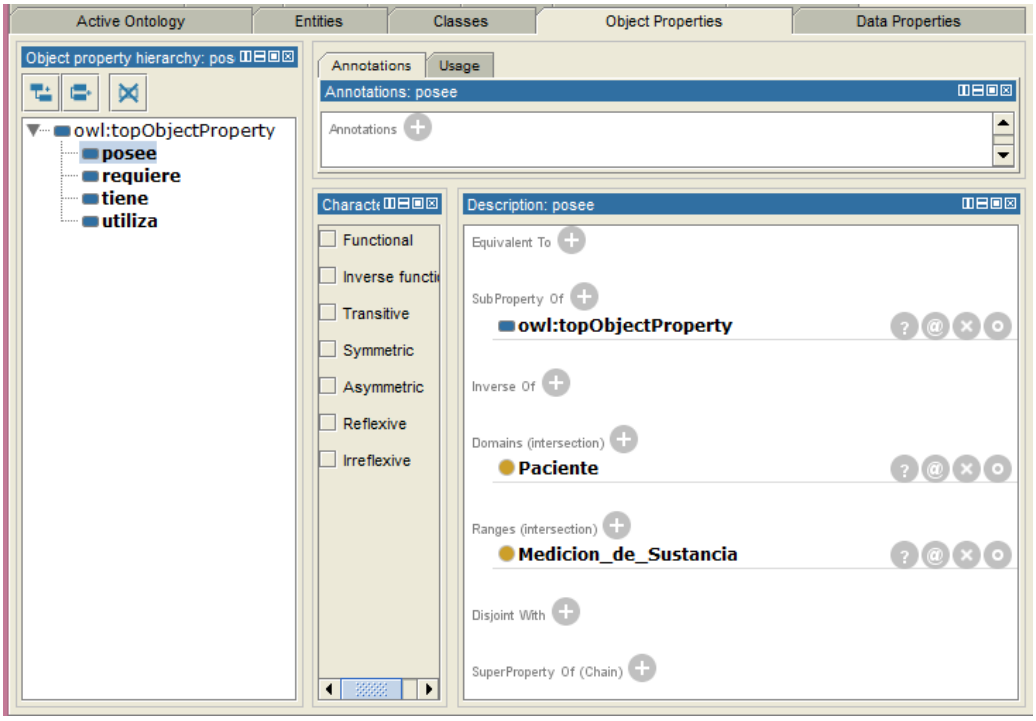


Figura 3.12 Relaciones de la ontología creadas con la herramienta Protégé

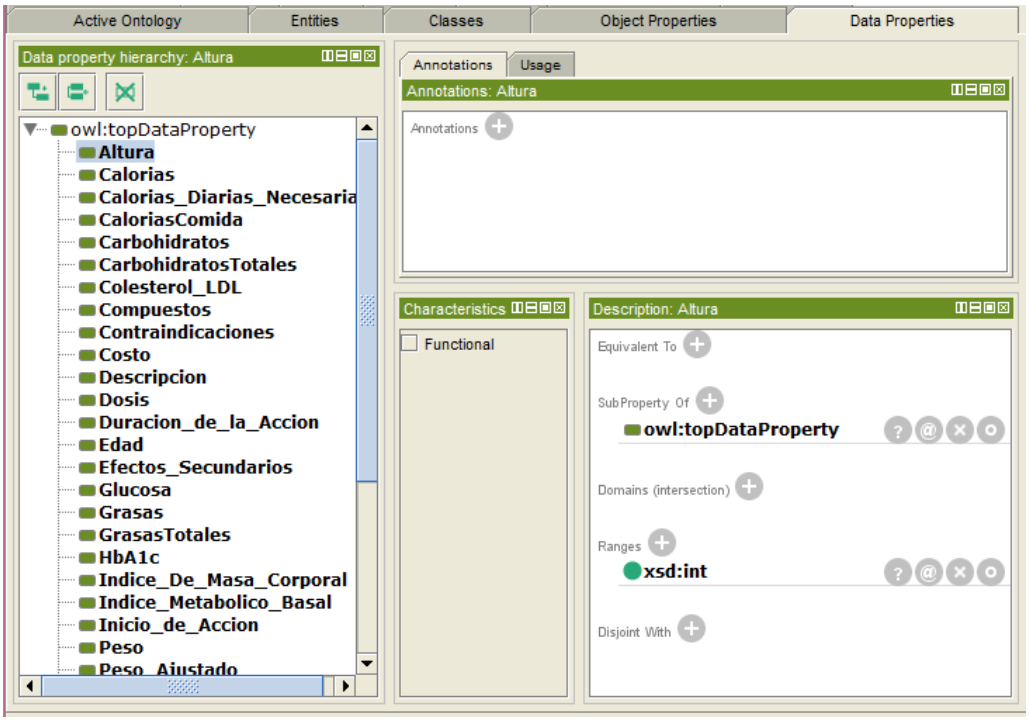


Figura 3.13 Definición de atributos con la herramienta Protégé

Después se procedió a crear las instancias del dominio que permanecen constantes en la ontología ya que aquéllas que pertenecen al concepto paciente se crean programáticamente desde la aplicación Web con base en las características propias del paciente. En la Figura 3.14 se observa un ejemplo de una instancia del concepto alimentación con sus respectivos atributos.

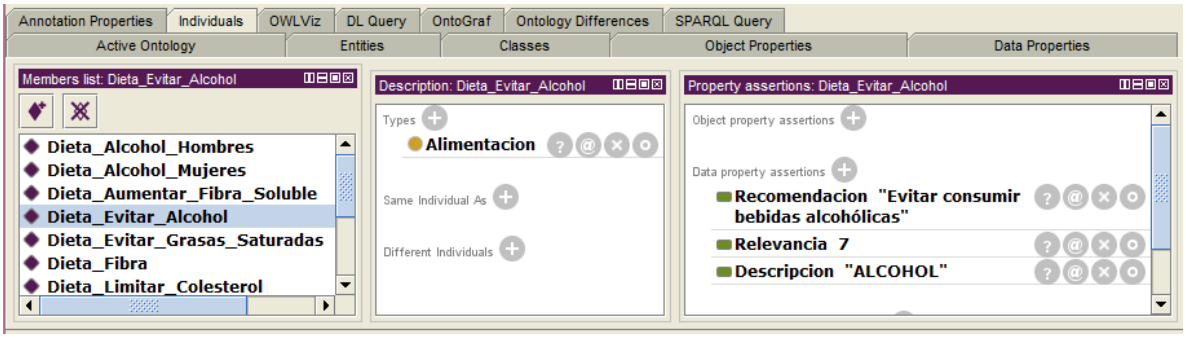


Figura 3.14 Definición de individuos en la ontología con la herramienta Protégé

Dentro de la ontología no es posible definir las reglas de inferencia por lo que todas las reglas identificadas en el apartado 3.3.2.7 se escribieron en un archivo de texto con extensión .rules.

Con esto finalmente se concluyó el desarrollo de la ontología y se obtuvieron como resultado dos archivos: uno que representa la ontología (`ocppd.owl`) y otro que representa las reglas de inferencia (`reglas.rules`). Ambos se integraron a la aplicación Web tal y como se describe en la arquitectura del sistema en la sección 3.1 de este documento.

Capítulo 4. Resultados

Como resultado de este trabajo de tesis se creó una aplicación Web que lleva por nombre "Sistema de Predicción y Prevención de Complicaciones de la Diabetes" o, en su forma abreviada, "SPPCD".

SPPCD es una aplicación que se basa en el uso de técnicas de minería de datos y ontologías para proporcionar a los profesionales de la salud una alternativa computacional con la capacidad de identificar complicaciones micro vasculares (nefropatía diabética, retinopatía diabética y pie diabético) en pacientes con Diabetes Mellitus Tipo II. También ofrece la posibilidad de sugerir el tratamiento no farmacológico y farmacológico para un paciente en particular con el fin de retrasar la aparición de las complicaciones descritas anteriormente.

En el presente capítulo se muestran 5 casos de estudio con el objetivo de mostrar los resultados obtenidos con la aplicación. Cabe mencionar que para utilizar la aplicación primeramente es necesario iniciar sesión en el sistema tal y como se ilustra en la Figura 4.1. Posteriormente, desde la pantalla de bienvenida de la aplicación (Figura 4.2) se accede a toda la funcionalidad del sistema, como la predicción de nefropatía, retinopatía, pie diabético o las recomendaciones de tratamiento farmacológico y no farmacológico.

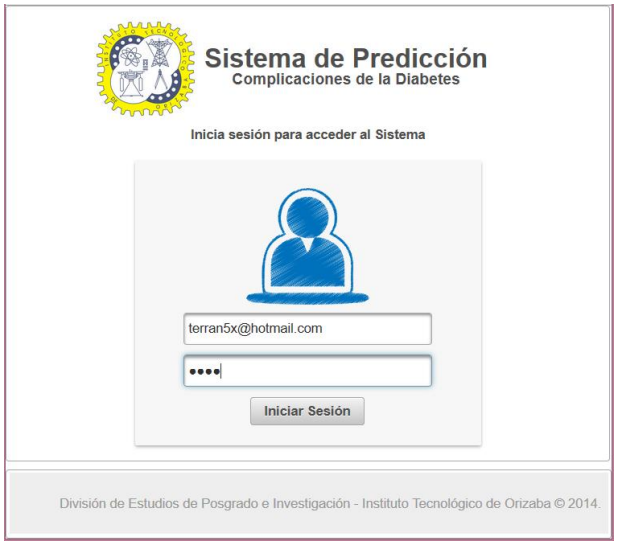


Figura 4.1 Inicio de sesión de usuarios

Figura 4.2 Pantalla de bienvenida de la aplicación

4.1 Caso de estudio: Predicción de nefropatía diabética en un paciente con DMTII

Se seleccionó un expediente de un paciente con diagnóstico de DMTII que acudió al módulo de consulta externa del HRRB, el paciente cuenta en su historia clínica con las siguientes características: un tiempo de 7 años de evolución desde el diagnóstico de la diabetes, es hipertenso, tiene una tensión arterial sistólica máxima de 160, es anémico, posee una creatinina sérica máxima de 1.3 y tiene retinopatía diabética. Se utilizó la aplicación SPPCD con el objetivo de predecir si el paciente tiene riesgo de padecer nefropatía diabética dado que esta herramienta posee un modelo de predicción que se construyó con base en casos reales de pacientes que padecen esta complicación.

A continuación se describe la funcionalidad de la aplicación para este caso de estudio en particular. Como primer paso desde la pantalla de bienvenida de la aplicación (Figura 4.3) se selecciona la opción “Predecir Nefropatía Diabética” posteriormente se capturan los valores de todas las variables requeridas del paciente en los campos correspondientes y se pulsa el botón “Predecir Nefropatía”, tal y como se muestra en la Figura 4.3. Acto seguido el sistema consulta al modelo de nefropatía diabética y predice que si existe riesgo de que el paciente presente nefropatía diabética con un 83.27% de probabilidad, este resultado se observa en la Figura 4.4.

Figura 4.3 Pantalla de captura de variables para predecir nefropatía diabética.

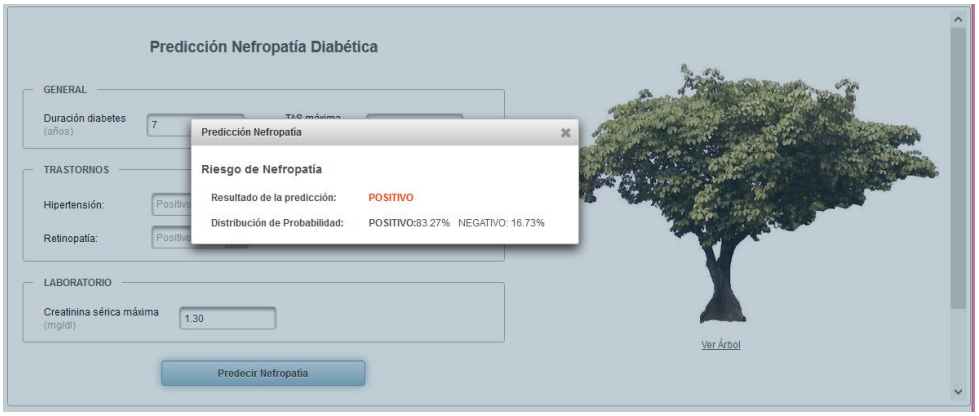


Figura 4.4 Resultado de la predicción de nefropatía diabética

Otra posibilidad es realizar la predicción manualmente desde el árbol de decisión que se muestra en la Figura 4.5 donde el nodo raíz se representa por la variable creatinina sérica máxima, el paciente al contar con un valor de 1.3 baja por la opción izquierda del árbol hacia el nodo retinopatía diabética, como el paciente tiene esta complicación se baja por la opción derecha hacia el nodo anemia y dado que es anémico se baja por la opción derecha del árbol para por fin llegar a la hoja que contiene la predicción de que si tiene riesgo de presentar retinopatía diabética.

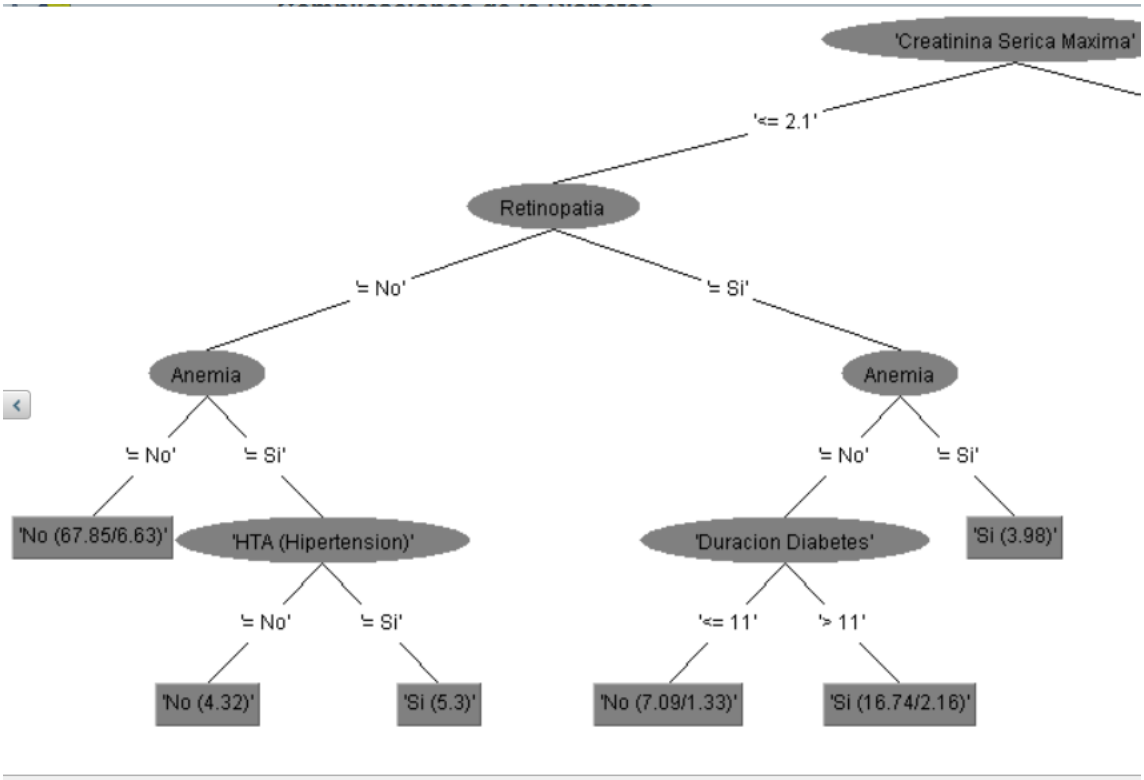


Figura 4.5 Parte del árbol de decisión del modelo de nefropatía diabética

En la Tabla 4.1 se muestran los resultados de utilizar la herramienta para predecir nefropatía diabética sobre 10 expedientes de pacientes con diagnóstico de DMTII. Los expedientes se eligieron de forma aleatoria con el método `random` del lenguaje Java. Como se observa el algoritmo predijo correctamente los 10 casos probados.

Tabla 4.1 Resultado de la predicción de nefropatía diabética sobre 10 pacientes con DMTII.

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Duración Diabetes	11	9	15	2	11	4	1	15	20	6
Hipertensión	Si	No	No	No	Si	Si	Si	No	Si	Si
Tensión Arterial Sistólica Máxima	110	110	120	130	110	140	140	130	135	140
Anemia	No	No	No	No	No	No	Si	No	Si	No
Creatinina Sérica Máxima	0.6	1.1	1.1	0.7	0.6	1.1	5.7	1.5	13.1	3.1
Retinopatía	No	No	No	No	No	No	Si	Si	No	No
Predicción	**No**	**No**	**No**	**No**	**No**	**No**	**Si**	**Si**	**Si**	**Si**
Real	No	No	No	No	No	No	Si	Si	Si	Si
Probabilidad	89.1%	89.1%	89.1%	89.1%	89.1%	89.1%	96.9%	83.1%	96.9%	96.9%

4.2 Caso de estudio: Predicción de retinopatía diabética en un paciente con DMTII

Se seleccionó un expediente de un paciente con diagnóstico de DMTII que acudió al módulo de consulta externa del HRRB, el paciente cuenta en su historia clínica con las siguientes características: sexo femenino, un tiempo de 10 años de evolución desde el diagnóstico de la diabetes, utiliza para el tratamiento de la diabetes hipoglucemiantes orales, es hipertensa, tiene una tensión arterial sistólica promedio de 150, cuenta con 8 embarazos, posee una hemoglobina máxima de 14.9, una glucosa promedio de 247, una glucosa máxima de 247, una creatinina sérica máxima de 1.4, no tiene neuropatía diabética ni insuficiencia renal. Se utilizó la aplicación SPPCD con el objetivo de predecir si el paciente tiene riesgo de padecer retinopatía diabética dado que esta herramienta posee un modelo de predicción que se construyó con base en casos reales de pacientes que padecen esta complicación.

Como primer paso desde la pantalla de bienvenida de la aplicación (Figura 4.2) se selecciona la opción “Predecir Retinopatía Diabética” posteriormente se capturan los valores de todas las variables requeridas del paciente en los campos correspondientes y se pulsa el botón “Predecir Retinopatía”, tal y como se muestra en la Figura 4.6. Acto seguido el sistema consulta al modelo de retinopatía diabética y predice que no existe riesgo de que el paciente presente retinopatía diabética con un 91.21% de probabilidad, este resultado se muestra en la Figura 4.7.

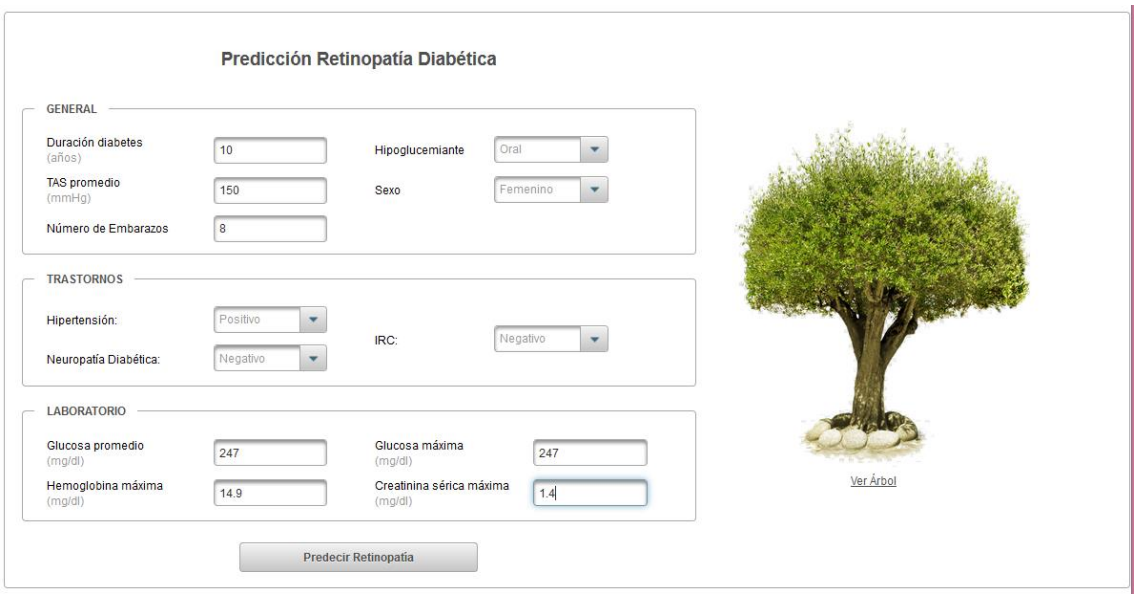


Figura 4.6 Pantalla de captura de variables para predecir nefropatía diabética.

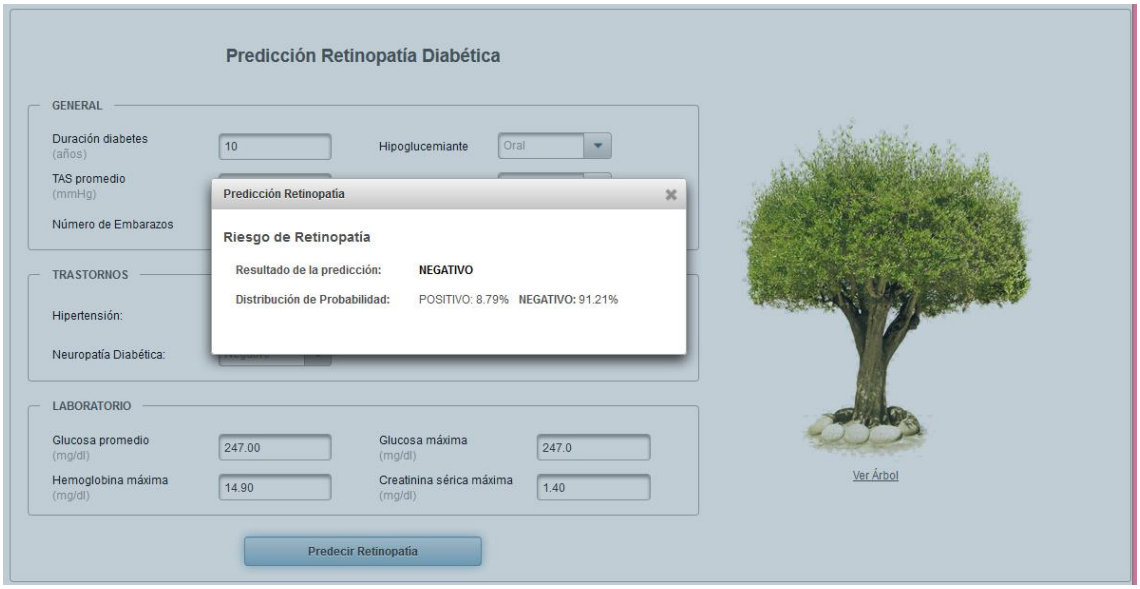


Figura 4.7 Resultado de la predicción de retinopatía diabética.

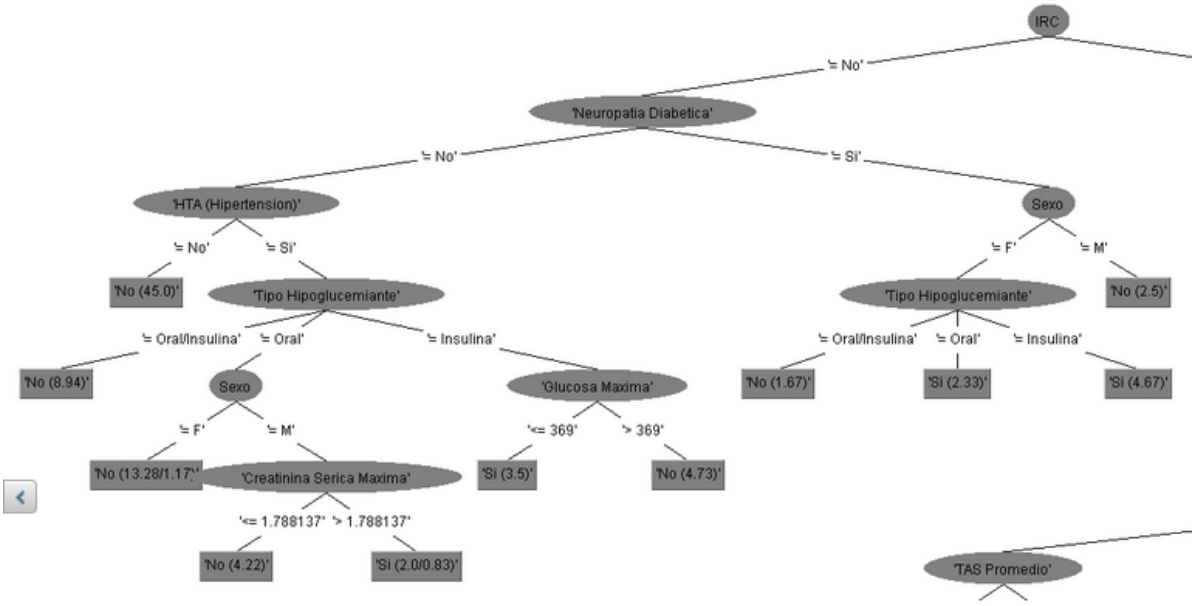


Figura 4.8 Parte del árbol de decisión del modelo de retinopatía diabética

Otra posibilidad es realizar la predicción manualmente desde el árbol de decisión que se muestra en la Figura 4.8, donde el nodo raíz representa a la variable insuficiencia renal, el paciente al no contar con esta complicación baja por la opción izquierda del árbol hacia el nodo neuropatía diabética y como el paciente no tiene esta complicación se baja por la opción izquierda hacia el nodo hipertensión, al ser hipertenso se baja por la opción derecha hacia el nodo tipo de hipoglucemiante, el paciente consume hipoglucemiantes orales por lo que se baja por la opción central hacia el nodo sexo y al ser femenino se baja por la opción izquierda para por fin llegar a la hoja del árbol que contiene la predicción de que no tiene riesgo de presentar retinopatía diabética.

En la Tabla 4.2 se muestra los resultados de utilizar la herramienta para predecir retinopatía diabética sobre 10 expedientes de pacientes con diagnóstico de DMTII.

Tabla 4.2 Resultado de la predicción de retinopatía diabética sobre 10 pacientes con DMTII.

#	1	2	3	4	5	6	7	8	9	10
Sexo	F	M	M	F	M	M	F	F	F	M
Duración Diabetes	10	1	3	30	4	4	12	6	20	20
Tipo Hipoglucemiante	Insulina	Oral Insulina	Oral	Insulina	Oral	Oral	Oral Insulina	Oral	Oral	Oral

#	1	2	3	4	5	6	7	8	9	10
Hipertensión	No	No	Si	Si	No	No	Si	No	Si	Si
TAS Promedio	125	120	125	143.33	110	67	120	134	134.43	138.67
No. Embarazos	3	0	0	10	0	0	4	7	11	0
Hemoglobina Máxima	9.6	14	17.6	11.4	13.4	13.2	11.8	8.6	10.9	11.42
Glucosa Promedio	218	123	305.7	240	239.13	183	180.2	116	134.43	138.67
Glucosa Máxima	492	123	702	266	375	221	274	134	205	232
Creatinina Sérica Máxima	4	0.7	1.4	0.9	0.6	0.9	6.7	1.1	5.3	2.2
Neuropatía Diabética	No	No	No	No	No	No	No	No	No	No
Insuficiencia Renal	Si	No	No	No	No	No	Si	No	Si	Si
Predicción	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**Si**	**Si**
Real	No	No	No	No	No	No	No	No	Si	Si
Probabilidad	81.3%	97.9%	90.9%	85.8%	97.9%	97.9%	81.3%	97.9%	94.5%	92.7%

4.3 Caso de estudio: Predicción de pie diabético en un paciente con DMTII

Se seleccionó un expediente de un paciente con diagnóstico de DMTII que acudió al módulo de consulta externa del HRRB, el paciente cuenta en su historia clínica con las siguientes características: sexo femenino, un tiempo de 7 años de evolución desde el diagnóstico de la diabetes, utiliza para el tratamiento de la diabetes hipoglucemiantes orales, no es hipertensa, tiene una tensión arterial diastólica promedio de 72,fuma, no es anémica, tiene una hemoglobina promedio de 12.33, una hemoglobina máxima de 13.8, un RDW promedio de 14.17, un nitrógeno ureico promedio de 22, un nitrógeno ureico máximo de 45, una creatinina sérica promedio de .73 y no posee insuficiencia renal. Se utilizó la aplicación SPPCD con el objetivo de predecir si el paciente tiene riesgo de padecer pie diabético dado que esta herramienta posee un modelo de predicción que se construyó con base en casos reales de pacientes que padecen esta complicación.

Como primer paso, desde la pantalla de bienvenida de la aplicación (Figura 4.2) se selecciona la opción “Predecir Pie Diabético”, posteriormente se capturan los valores de todas las variables requeridas del paciente en los campos correspondientes y se pulsa el botón “Predecir Pie Diabético”, tal y como se muestra en la Figura 4.9. Acto seguido el sistema consulta al modelo de pie diabético y predice que no existe riesgo de que el paciente presente pie diabético con un 100% de probabilidad, este resultado se muestra en la Figura 4.10.

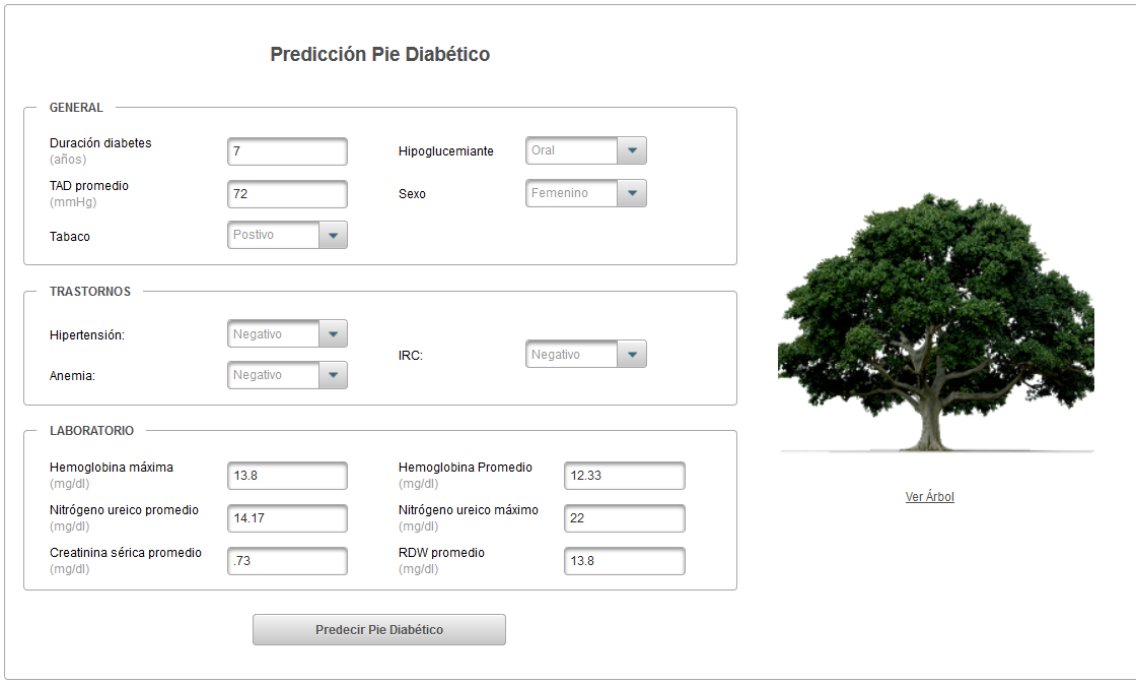


Figura 4.9 Pantalla de captura de variables para predecir pie diabético.

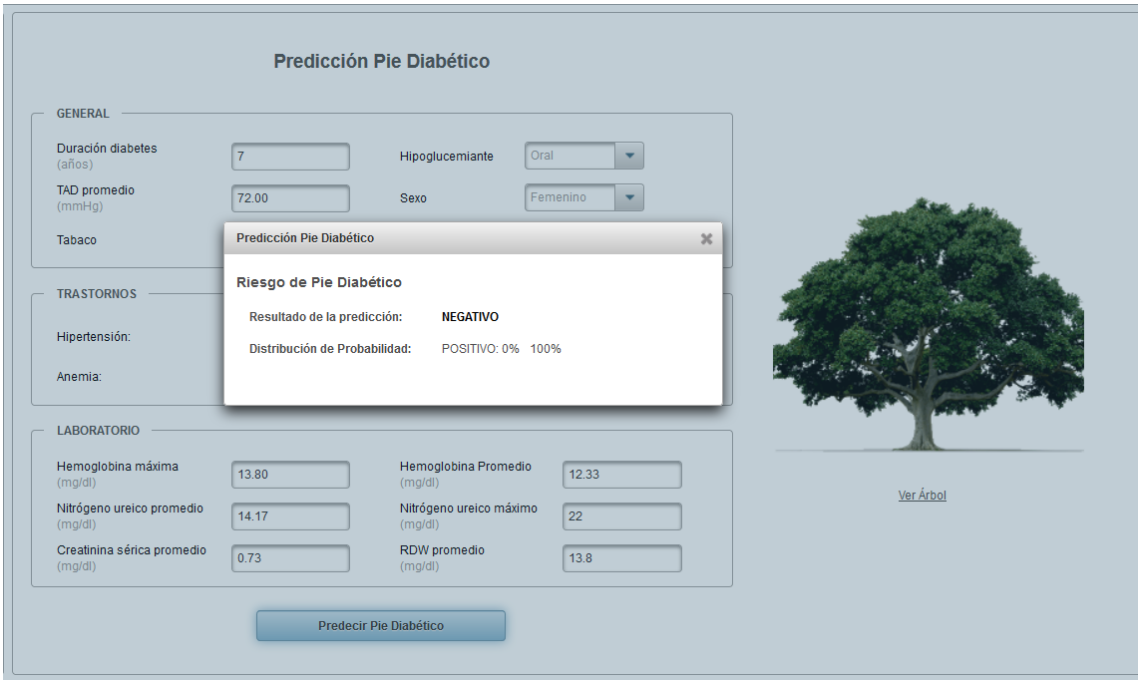


Figura 4.10 Resultado de la predicción de predecir pie diabético.

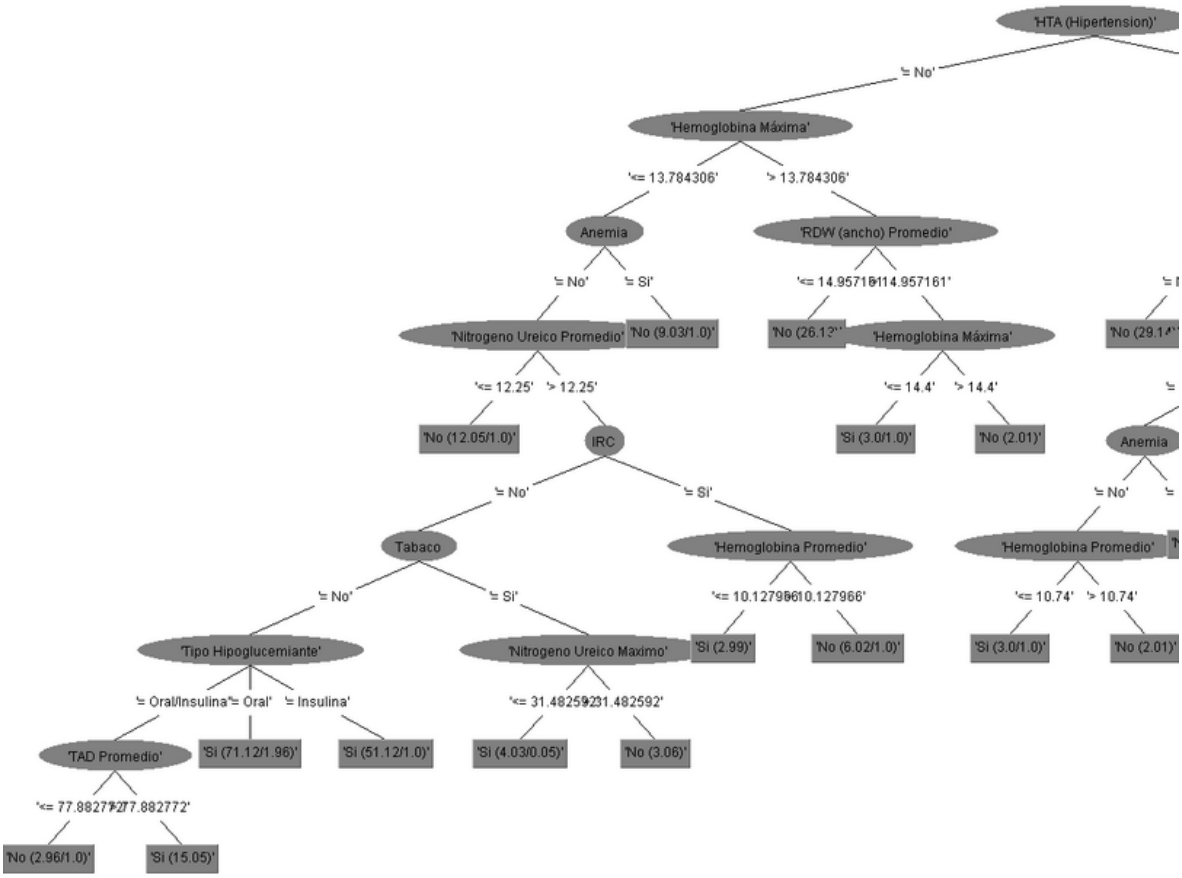


Figura 4.11 Parte del árbol de decisión del modelo de pie diabético

Otra posibilidad es realizar la predicción manualmente desde el árbol de decisión que se muestra en la Figura 4.11 donde el nodo raíz se representa por la variable hipertensión como el paciente no es hipertenso se baja por la opción izquierda del árbol hacia el nodo hemoglobina máxima que tiene un valor de 13.8 por lo que nuevamente se baja por la opción derecha hacia el nodo RDW que tiene un valor de 13.8 por lo que se baja por la opción izquierda para por fin llegar a la hoja del árbol que contiene la predicción de que no tiene riesgo de presentar pie diabético.

En la Tabla 4.3 se muestra los resultados de utilizar la herramienta para predecir pie diabético sobre 10 expedientes de pacientes con diagnóstico de DMTII. Como se observa el algoritmo predijo correctamente 9 de los 10 casos probados.

Tabla 4.3 Resultado de la predicción de pie diabético sobre 10 pacientes con DMTII.

#	1	2	3	4	5	6	7	8	9	10
Sexo	M	M	M	F	M	M	F	M	M	F
Duración Diabetes	10	18	15	7	10	13	11	20	18	22
Tipo Hipoglucemiante	Oral	Oral	Oral	Oral	Oral Insulina	Oral	Oral Insulina	Insulina	Insulina	Oral Insulina
Hipertensión	Si	No	No	Si	Si	Si	Si	Si	Si	Si
TAD Promedio	70	72.86	78.75	106.67	85	76.67	91	78.75	84	80
Tabaco	No	Si	No	No	Si	Si	No	Si	No	No
Anemia	Si	No	No	No	Si	Si	Si	Si	Si	No
Hemoglobina Promedio	8.8	9.5	13.6	12.4	10.73	6.93	9.27	7.85	8.8	9.28
Hemoglobina Máxima	10.1	9.8	13.6	12.4	11.6	8.3	11.1	9.4	9.4	10.8
RDW Promedio	15.25	12.4	14.8	11.7	14.9	15.7	14.37	14.3	13.37	14.41
Nitrógeno Ureico Promedio	72.5	9	16	32.4	34.25	62.33	67	106.8	126.33	45.79
Nitrógeno Ureico Máximo	96	11	17	58	54	93	46	120	141	148
Creatinina Sérica Promedio	4.3	0.65	9.95	1.13	1.84	1.83	3.07	12.38	16.07	3.71
Insuficiencia Renal	Si	No	No	No	Si	Si	Si	Si	Si	Si
Predicción	**Si**	**No**	**Si**	**No**	**No**	**No**	**No**	**No**	**No**	**No**
Real	Si	No	Si	No	No	No	No	No	No	**Si**
Probabilidad	81.8%	85.8%	95.9%	96.8%	89.5%	83.5%	72.4%	82.4%	82.4%	59.9%

4.4 Caso de estudio: Recomendación de tratamiento no farmacológico para un paciente con DMTII

Se seleccionó un expediente de un paciente con diagnóstico de DMTII que acudió al módulo de consulta externa del HRRB, el paciente cuenta en su historia clínica con las siguientes características: sexo femenino, una edad de 46 años, 148 cm de altura, 71 kg de peso, un estilo de vida sedentario, consumo de alcohol positivo, sin tratamiento actual para controla la diabetes, glucosa de 133, cetonas negativo, un colesterol 177 y triglicéridos de 248. Se utilizó la aplicación SPPCD para sugerir un tratamiento no farmacológico para paciente con el objetivo de mantenerlo en buen estado de salud y evitar la aparición de complicaciones. Estas recomendaciones se hacen a partir de una ontología que infiere los cuidados que son adecuados para un paciente en particular con base en diversas guías de la Salud de instituciones como ALAD, IMSSS, AAEC y ADA. A continuación se describe la funcionalidad de la aplicación para este caso de estudio en particular. Como primer paso

desde la pantalla de bienvenida de la aplicación (Figura 4.2) se selecciona la opción "Tratamiento no Farmacológico" posteriormente se capturan los valores de todas las variables requeridas del paciente en los campos correspondientes y se pulsa el botón "Tratamiento", tal y como se muestra en la Figura 4.12.

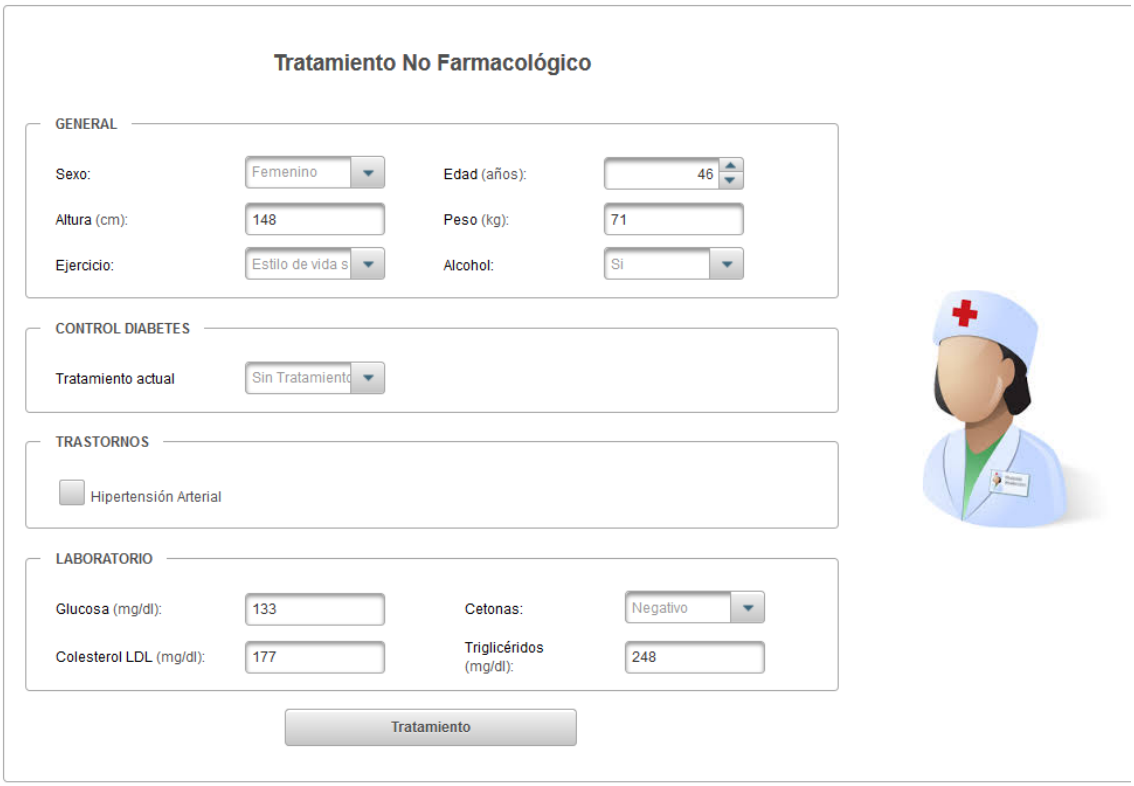


Figura 4.12 Pantalla de captura de variables para recomendar tratamiento no farmacológico.

Acto seguido el sistema consulta a la ontología OCCPD y genera las recomendaciones del tratamiento no farmacológico, en la Figura 4.13 se observa que estas recomendaciones se encuentran agrupadas en pestañas, a continuación se describe de manera muy breve el contenido de cada una de ellas:

- **Observaciones:** Son observaciones muy generales donde se muestra el cálculo del Índice de Masa Corporal del paciente y su edad.
- **Laboratorio:** En este apartado se muestran los resultados de las pruebas de laboratorio con sus valores de referencia correspondientes y se indica en color rojo

si estos se encuentran en valores anormales. En este caso se observa que tanto los triglicéridos, como la glucosa y colesterol se encuentran en valores fuera de los normales.

- **Calorías totales diarias:** Se muestras el cálculo de calorías y macronutrientes (carbohidratos, proteínas y grasas) recomendadas para el consumo diario del paciente, dado que esta persona tiene sobrepeso se le recomienda una dieta baja en calorías.
- **Distribución de nutrientes por cada comida:** Es exactamente lo mismo que la sección anterior solo que se indica las calorías y distribución de macronutrientes para cada comida del día (desayuno, comida y cena)
- **Alimentación:** En este apartado se enlistan las recomendaciones alimenticias, en este caso como el paciente tiene triglicéridos altos se le recomienda evitar consumir alcohol, igual se le recomienda evitar alimentos con alto contenido de colesterol dado que lo tiene alto, al igual que evitar las grasas saturadas, entre otras cosas.
- **Ejercicio:** Contiene todas las recomendaciones referentes a la actividad física del paciente dado que tiene un estilo de vida sedentario se le recomienda realizar caminata todos los días a una intensidad suave durante 10 minutos y aumentar 5 minutos por semana hasta llegar a una meta de 30 minutos, posteriormente se recomienda aumentar la intensidad a moderada en combinación con el ejercicio aeróbico de su preferencia (natación, bicicleta, baile, caminata), entre otras recomendaciones.
- **Cuidados de los pies:** Aquí se enlistan recomendaciones generales para mantener sanos los pies, como su correcto aseo, el corte de la uñas, el uso de calzado entre otras cosas.
- **Hábitos:** Se enlistan hábitos saludables como no fumar.
- **Otras recomendaciones alimenticias:** Finalmente aquí se presenta una lista de recomendaciones alimenticias como la norma mexicana del plato del bien comer, que bebidas tomar (jarra del buen beber), como cocinar los alimentos, que alimentos evitar o limitar su consumo, entre otras cosas.

En la Figura 4.14 se observa la exportación de todas las recomendaciones en formato PDF.

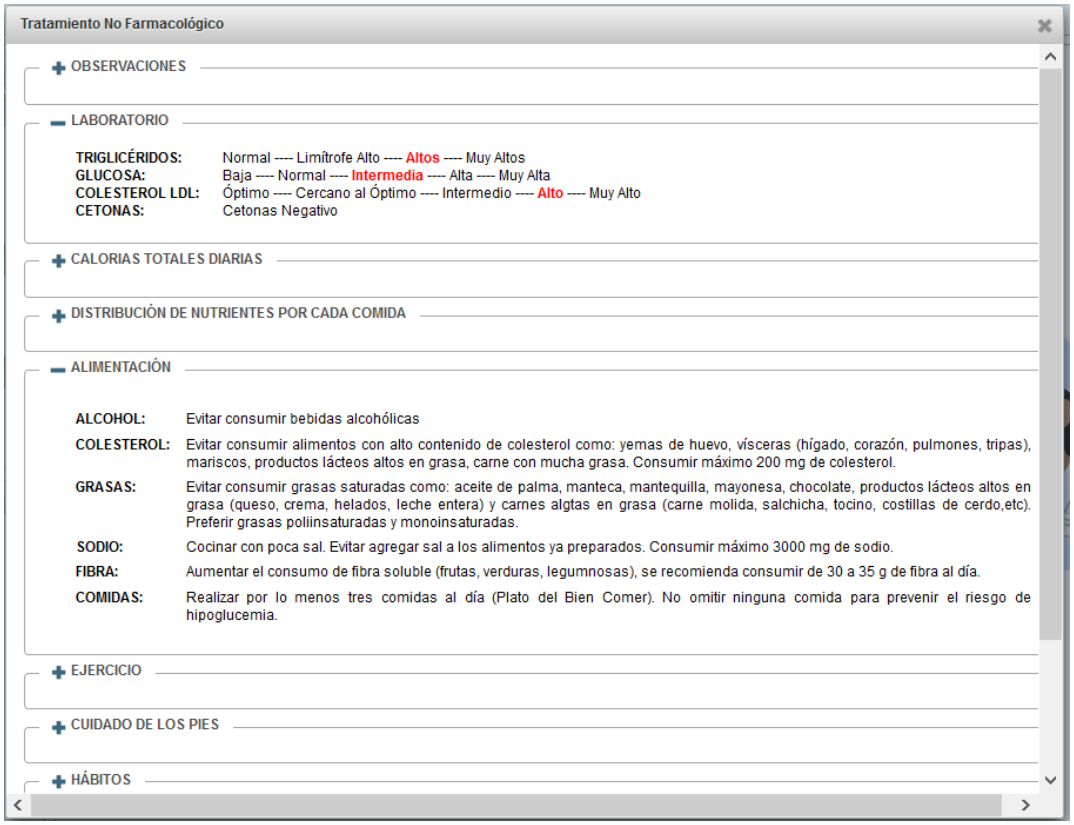


Figura 4.13 Recomendaciones del tratamiento no farmacológico

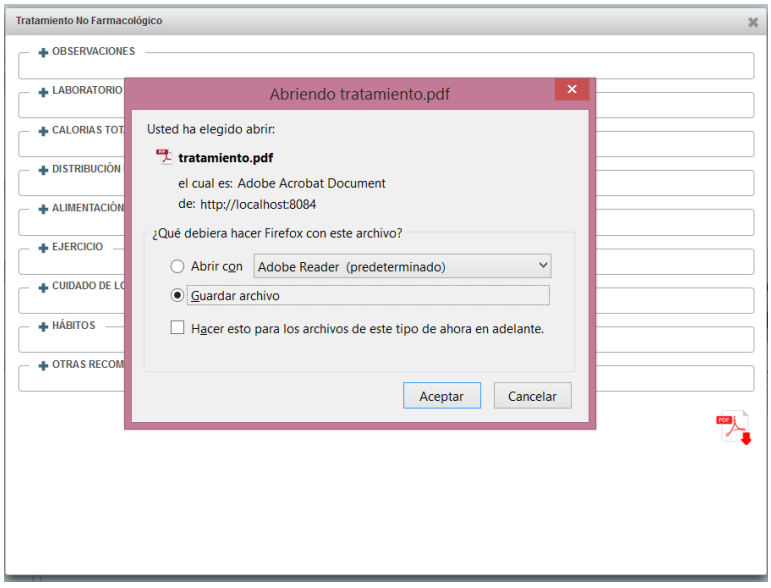


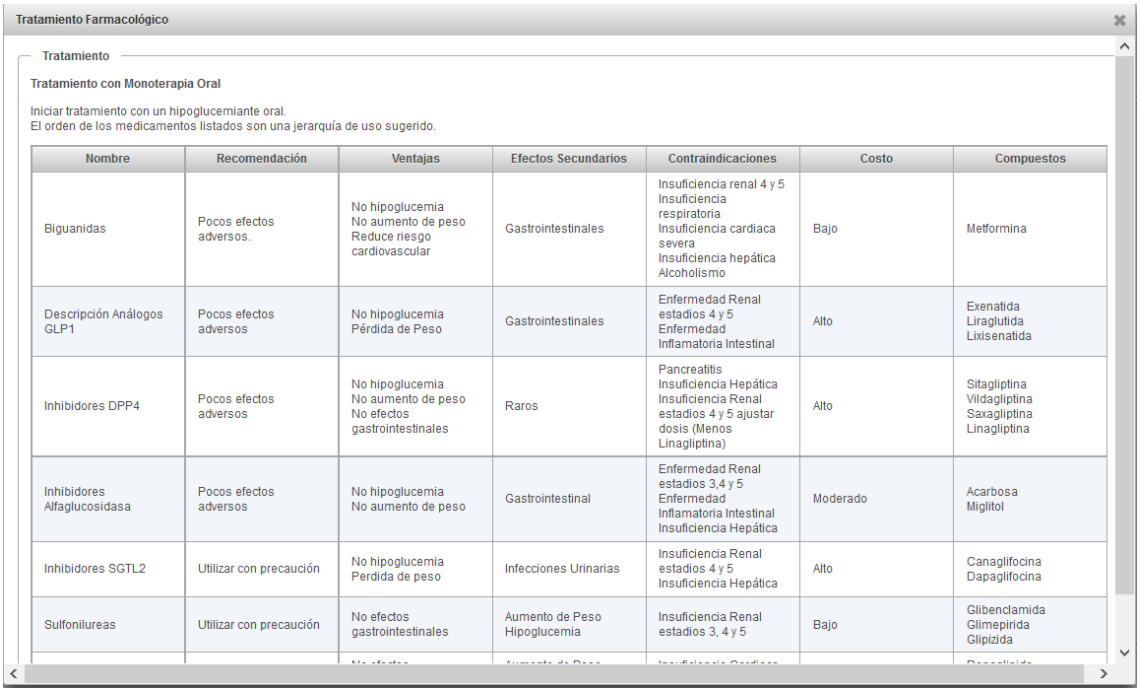
Figura 4.14 Posibilidad de guardar la información en formato PDF

4.5 Caso de estudio: Recomendación de tratamiento farmacológico para un paciente con DMTII

Se seleccionó un expediente de un paciente con diagnóstico de DMTII que acudió al módulo de consulta externa del HRRB, el paciente cuenta en su historia clínica con las siguientes características: tratamiento actual con dieta para controlar la diabetes y una glucosa de 215. Se utilizó la aplicación SPPCD para sugerir un tratamiento farmacológico para el paciente con el objetivo de mantenerlo en buen estado de salud y evitar la aparición de complicaciones. Estas recomendaciones se hacen a partir de una ontología que infiere los cuidados adecuados para un paciente en particular.

Como primer paso, desde la pantalla de bienvenida de la aplicación (Figura 4.2) se selecciona la opción “Tratamiento Farmacológico” posteriormente se capturan los valores de todas las variables requeridas en los campos correspondientes y se pulsa el botón “Tratamiento”, tal y como se muestra en la Figura 4.15. Acto seguido el sistema consulta a la ontología OCCPD y genera las recomendaciones de los posibles medicamentos que se le podrían recetar al paciente, estos se enlistan en una tabla ordenados jerárquicamente del mas al menos recomendado con base en el criterio sugerido por la Asociación Americana de Endocrinólogos Clínicos, en la tabla se ilustran algunas de sus características como efectos secundarios, contraindicaciones, costos, ventajas así el médico selecciona el medicamento más adecuado que tenga disponible en el hospital y de acuerdo a las necesidades del paciente. Por ejemplo, en este caso, la mejor opción es recetarle al paciente metformina pero en caso de que en el hospital no tuvieran disponible este medicamento se le podría recetar la segunda opción que es la familia de los análogos glp1 tomando en cuenta que se tendría que evaluar la situación económica del paciente dado que el precio de este medicamento es alto en comparación de otros. En la Figura 4.16, se muestra lo descrito anteriormente.

Figura 4.15 Pantalla de captura de variables para recomendar tratamiento farmacológico.



Nombre	Recomendación	Ventajas	Efectos Secundarios	Contraindicaciones	Costo	Compuestos
Biguanidas	Pocos efectos adversos.	No hipoglucemia No aumento de peso Reduce riesgo cardiovascular	Gastrointestinales	Insuficiencia renal 4 y 5 Insuficiencia respiratoria Insuficiencia cardiaca severa Insuficiencia hepática Alcoholismo	Bajo	Metformina
Descripción Análogos GLP1	Pocos efectos adversos	No hipoglucemia Pérdida de Peso	Gastrointestinales	Enfermedad Renal estadios 4 y 5 Enfermedad Inflamatoria Intestinal	Alto	Exenatida Liraglutida Lixisenatida
Inhibidores DPP4	Pocos efectos adversos	No hipoglucemia No aumento de peso No efectos gastrointestinales	Raros	Pancreatitis Insuficiencia Hepática Insuficiencia Renal estadios 4 y 5 ajustar dosis (Menos Linagliptina)	Alto	Sitagliptina Vildagliptina Saxagliptina Linagliptina
Inhibidores Alfaglucosidasa	Pocos efectos adversos	No hipoglucemia No aumento de peso	Gastrointestinal	Enfermedad Renal estadios 3,4 y 5 Enfermedad Inflamatoria Intestinal Insuficiencia Hepática	Moderado	Acarbosa Miglitol
Inhibidores SGTL2	Utilizar con precaución	No hipoglucemia Perdida de peso	Infecciones Urinarias	Insuficiencia Renal estadios 4 y 5 Insuficiencia Hepática	Alto	Canaglifocina Dapaglifocina
Sulfonilureas	Utilizar con precaución	No efectos gastrointestinales	Aumento de Peso Hipoglucemia	Insuficiencia Renal estadios 3, 4 y 5	Bajo	Glibenclamida Glimepirida Glipizida

Figura 4.16 Posibles opciones de medicamentos sugeridos en el tratamiento farmacológico

Es importante resaltar que los casos de estudio ilustrados en este capítulo y varios más se realizaron con el asesor médico del Hospital Regional de Río Blanco, dando éste su aval para las predicciones y sugerencias obtenidas con la aplicación Web.

Capítulo 5. Conclusiones y Recomendaciones

5.1 Conclusiones

Es importante valorar el impacto de las Tecnologías de Información y Comunicación en el sector de la salud. En este trabajo se ofrece una alternativa computacional en la identificación de complicaciones micro vasculares (retinopatía diabética, nefropatía diabética y pie diabético) en pacientes con Diabetes Mellitus Tipo II.

Con la aplicación de las diferentes técnicas de Minería de Datos se permitió analizar diversas variables que se relacionan con la aparición de estas complicaciones que de otra forma sería difícil de evaluar por un profesional de la salud. Además hizo posible generar modelos de comportamiento basados en las características propias de pacientes que ya padecen estas complicaciones con lo que se ofrece una solución alternativa a los métodos convencionales para diagnosticar estas enfermedades.

Con el desarrollo de esta aplicación se espera ayudar a los profesionales de la salud en el proceso de toma de decisiones, como por ejemplo reforzar un diagnostico o identificar pacientes en potencial riesgo de padecer estas enfermedades.

En general se obtuvieron buenos resultados al probar la precisión de los modelos sobre el conjunto completo de datos. El modelo de nefropatía diabética logró una precisión del 94.47%, una sensibilidad del 93.5% y una especificidad del 95.3% lo que se considera bastante aceptable; el modelo de retinopatía obtuvo una precisión del 95.98%, una sensibilidad del 93.6% y una especificidad del 96.7% lo que también se considera bastante aceptable y el modelo de pie diabético obtuvo una precisión del 92.96% y una especificidad del 96.40% aunque la sensibilidad se quedó con un porcentaje bajo (76.5%) comparado con los resultados anteriores, pero, considerando las características de la población y que fue un resultado superior al 75%, no se considera un mal porcentaje.

Por otro lado, en este trabajo también se creó una alternativa computacional mediante el uso de ontologías que ayuda a los médicos en el planteamiento del tratamiento

farmacológico y no farmacológico del paciente diabético con riesgo de presentar alguna complicación. La herramienta proporciona recomendaciones para el tratamiento no farmacológico cubriendo sus cuatro pilares fundamentales que son: alimentación, ejercicio, hábitos saludables y cuidado de los pies, ya que muchas veces se le resta importancia a este tratamiento o simplemente no se cuenta con el tiempo suficiente en la consulta para atender todas estas recomendaciones. La aplicación también proporciona de recomendaciones generales sobre el tratamiento farmacológico más adecuado para un paciente así como una pequeña descripción de cada medicamento aunque cabe mencionar que inicialmente no se tenía contemplado esta funcionalidad por lo que aún quedan muchas mejoras por realizar. El principal objetivo del uso de esta ontología es intentar mantener en óptimas condiciones de salud al paciente para evitar o retrasar la aparición de complicaciones.

El asesor médico del Hospital Regional de Rio Blanco consideró buenos los resultados obtenidos por los modelos de minería de datos y encontró de mucha utilidad el uso de la ontología para sugerir recomendaciones en el tratamiento de pacientes diabéticos por lo que se invitó a instalar esta aplicación en el servidor del hospital para utilizarse en el área de Consulta Externa.

5.1 Recomendaciones

Lamentablemente la mayoría de los hospitales públicos en México no cuenta con sistemas que implementen el expediente clínico electrónico (ECE), lo que genera un gran problema en lo que respecta a la obtención de la población en proyectos como esta tesis, ya que la recolección de datos se realiza de forma manual a partir de la revisión de expedientes físicos, lo que ralentiza y limita el tamaño de la muestra, además de que se genera la posibilidad de cometer errores en la captura de información. Por lo que, como trabajo a futuro, es recomendable obtener una población de pacientes, suficientemente grande para dividirla en un conjunto de entrenamiento y un conjunto de pruebas, de una institución que implemente el ECE y reentrenar a los modelos; además sería muy interesante estudiar el comportamiento de las variables importantes que en este trabajo se eliminaron debido a que pocos pacientes contaban con los estudios de laboratorio correspondientes como es el caso

de: hemoglobina glucosilada, colesterol HDL, colesterol LDL, triglicéridos y micro albuminuria, entre otros.

Por otro lado, sería recomendable complementar la ontología OCCPD para que proporcione recomendaciones para el tratamiento farmacológico de la hipertensión y la dislipidemia que son afecciones comunes en los pacientes diabéticos. Además de profundizar en las recomendaciones para que el sistema sea capaz de recomendar dosificaciones, verificar existencias, contraindicaciones y efectos secundarios de cada fármaco.

Referencias

[1] ENSANUT. ENSASUT. [Online]. http://ensanut.insp.mx/

[2] Instituto Nacional de Estadística y Geografía. INEGI. [Online]. http://www3.inegi.org.mx/sistemas/biinegi/default.aspx

[3] Organización Panamericana de la Salud / Panamá, "Guía para la atención integral de las personas con diabetes mellitus," *Panamá : Organización Panamericana de la Salud*, 2009.

[4] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, no. 3, 2012, ISSN: 2277-3754.

[5] Marcano Cedeno, Alexis , and Diego Andina, "Data mining for the diagnosis of type 2 diabetes," in *In World Automation Congress (WAC)*, 2012, ISSN 2154-4824.

[6] Gustavo Olaiz Fernández, Rosalba Rojas, Carlos A. Aguilar Salinas, Juan Rauda, and Salvador Villalpando, "Diabetes mellitus en adultos mexicanos. Resultados de la Encuesta Nacional de Salud 2000," *Salud Publica Mex*, vol. 49, no. s3, 2007.

[7] Fan L. Longo et al., *Harrison principios de medicina interna*.: Mc Graw Hill, 2012.

[8] National Eye Instiute. [Online]. http://www.nei.nih.gov/health/diabetic/diabeticretino.pdf

[9] Ministerio de Salud, "Guía Clínica Diabetes Mellitus Tipo 2," 2010, ISBN: 978-956-8823-48-1.

[10] México: SecretarÍa de SALUD, "Guía de Práctica Clínica, Diagnóstico y Tratamiento de Retinopatía Diabética," 2009.

[11] CONSENSO ALAD. Prevención, diagnóstico y tratamiento temprano. [Online]. http://www.alad-latinoamerica.org/DOCConsenso/PREVENCION%20DE%20NEFROPATIA.pdf

[12] Asociación Latinoamericana de Diabetes, "Guías Alad de diagnóstico y tratamiento de diabetes mellitus tipo 2".

[13] IMSS, "Diagnóstico y Tratamiento de la Diabetes Mellitus tipo 2," 2009.

[14] Mehmed Kantarzdzic, *Data Mining: Concepts, Models, Methods, and Algorithms (2nd ed.)*.: Wiley-IEEE Press, 2011.

[15] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques (3rd ed.)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[16] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus, "Knowledge discovery in databases: An overview," *AI magazine*, vol. 13, no. 3, p. 57, 1992.

[17] Usama Fayyad, Piatetsky-Shapiro Gregory, and Smyth Padharaic, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[18] César Pérez López, *Minería de datos: técnicas y herramientas*.: Ediciones Paraninfo, 2007.

[19] Microsoft Corporation. (2013, Agosto) MSDN Library. [Online]. http://msdn.microsoft.com/es-es/library/cc645779.aspx#feedback

[20] Sholom M. Weiss and Nitin Indurkhya, *Predictive Data Mining: A Practical Guide*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1998.

[21] Remco R. Bouckaert et al., "Weka Manual," *The university of Waikato*, 2014.

[22] Nancy Pérez Castro, "Preprocesamiento de datos termográficos por medio de técnicas de balanceo de clases y análisis de cumulus (Clustering)".

[23] Jorge Escobedo de la Peñá and Beatriz Rico Verdín, "Incidencia y letalidad de las complicaciones agudas y crónicas de la diabetes mellitus en México; Incidence and fatlity rates of short and long-term diabetes mellitus complications in Mexico," *Salud pública Méx*, vol. 8, no. 4, pp. 236-42, 1996.

[24] Apache. jena. [Online]. http://jena.apache.org/documentation/inference/

[25] Secretaría de Salud, "Programa de Acción Específico 2007 – 2012," *Unidades de Especialidades Médicas UNEMES*, 2007.

[26] Salvador Villalpando et al., "Prevalence and distribution of type 2 diabetes mellitus in Mexican adult population: a probabilistic survey," *salud pública de méxico*, vol. 52, pp. s19-s26, 2010.

[27] Armando Arredondo and Esteban de Icaza, "Costos de la diabetes en América Latina: evidencias del caso Mexicano," *Value Health*, vol. 14, no. 5, pp. s85-8, 2011.

[28] Kaur Harleen and Krishan Wasan Siri, "Empirical study on applications of data mining techniques in healthcare," *Journal of Computer Science*, vol. 2, no. 2, pp. 194-200, 2006, ISSN 1549-3636.

[29] Mohd Fauzi Bin Othman and Thomas Moh Shan You, "Comparison of different classification techniques using WEKA for breast cancer," *In 3rd Kuala Lumpur International Conference on Biomedical Engineering*, pp. 520-523, 2007, ISBN 978-3-540-68016-1.

[30] Hian Chye Koh and Gerald Tan, "Data mining applications in healthcare," *Journal of Healthcare Information Management*, vol. 19, no. 2, p. 65, 2011.

[31] G. Parthiban, A. Rajesh, and S.K. Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method," *International Journal of Computer Applications*, vol. 24, no. 3, pp. 7-11, 2011.

[32] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93-9, 2013, ISSN 1607-551X.

[33] S.S. Imambi and T. Sudha, "Building Classification System to Predict Risk factors of Diabetic Retinopathy Using Text mining," *International Journal*, vol. 2, no. 7, pp. 2309-2312, 2010, ISSN : 0975-3397.

[34] Jinn-Yi Yeh, Tai-Hsi Wu, and Chuan-Wei Tsao, "Using data mining techniques to predict hospitalization of hemodialysis patients," *Decis. Support Syst*, vol. 50, no. 2, pp. 439-448, January 2011, ISSN 0167-9236.

[35] Tzu-Chuen Lu and Chun-Ya Tseng, "Hemodialysis key features mining and patients clustering technologies," *Adv. Artif. Neu. Sys*, p. 6, 2012.

[36] Baek Hwan Cho et al., "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods," *Artif. Intell. Med*, vol. 42, no. 1, pp. 37-53, 2008.

[37] David Riaño et al., "An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients," *J. of Biomedical Informatics*, vol. 45, no. 3, pp. 429-446, 2012, ISSN 1532-0464.

[38] N. Chalortham, M. Burranarach, and T. Supnithi, "Information and Creativity Support Systems," in *Ontology Development for Type II Diabetes Mellitus Clinical Support System*, 2009.

[39] David Sutton, Arantza Aldea, and Clare Martin, "An ontology of diabetes self management," *Proceedings of the first international workshop on Managing interoperability and complexity in health systems*, 2012.

[40] Hyunchul Jang, "Ontology for medicinal materials based on traditional Korean medicine," *Bioinformatics*, vol. 26, no. 18, pp. 2539-2360, 2010.

[41] Juan Manuel López Carmona, Cuauhtémoc Rául Ariza Andraca, José Raymundo Rodriguez Moctezuma, and Catarina Munguía Miranda, "Construcción y validación inicial de un instrumento para medir el estilo de vida en pacientes con diabetes mellitus tipo 2," *Salud pública Méx*, vol. 45, no. 4, pp. 259-267, 2003.

[42] Blanca Rosa Durán Varela, Blanca Rivera Chavira, and Ernesto Franco Gallegos, "Apego al tratamiento farmacológico en pacientes con diagnóstico de diabetes mellitus tipo 2," *Salud Pública Mex.*, vol. 43, no. 3, pp. 233-236, 2001.

[43] José Lauro de los Rios Castillo, Juan José Sánchez Sosa, Pedro Barrios Santiago, and Verónica Guerrero Sustaita, "Calidad de vida en pacientes con diabetes mellitus tipo 2," *Revista Médica del IMSS*, vol. 42, no. 2, 2004.

[44] Juan Manuel López Carmona and Raymundo Rodriguez Moctezuma, "Adaptación y validación del instrumento de calidad de vida Diabetes 39 en pacientes mexicanos con diabetes mellitus tipo 2," *Salud Pública Méx*, vol. 48, no. 3, pp. 200-208, 2006.

[45] LI Chang-ping et al., "Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus," *Chinese Medical Journal*, vol. 125, no. 5, pp. 851-857.

[46] Abdullah A. Aljumah, Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui, "Application of Data Mining: Diabetes Health Care in Young and Old Patients," in *Journal of King Saud University-Computer and Information Sciences*, 2012, ISSN 1319-1578.

ANEXO 1

Cédula de recolección de datos

Fecha:			**No. de Registro:**
Edad:	**Sexo: M [ ] F [ ]**		**Duración diabetes:**
IMC:		**Tipo hipoglucemiante: Oral [ ] Insulina [ ]**	
Tipo hipoglucemiante oral: **Metformina [ ] Glibenclamida [ ] Glipizid [ ] Glimepirid [ ] Repaglinida [ ]** **Nateglinida [ ] Pioglitazon [ ] Acarbosa [ ] Exenatida [ ] Sitagliptina [ ] Vildagliptina [ ]**			
Tipo insulina: **Lispro [ ] Aspart [ ] Glulisina [ ] NPH [] Detemir [ ] Glargina [ ] Rápida [ ]**			
Glucemia Prom:	**Glucemia Max:**	**HbA1c Prom:**	**HbA1c Max:**
Tabaco: Si [ ] No[ ]		**Alcohol: Si [ ] No [ ]**	
HDL Prom:	**HDL Max:**	**LDL Prom:**	**LDL Max:**
LDL Prom:	**LDL Max:**	**Triglicéridos Prom:**	**Triglicéridos Max:**
Colesterol Prom:	**Colesterol Max:**	**T.A. sistólica Prom:**	**T.A. sistólica Max:**
T.A. diastólica Prom:	**T.A. diastólica Max:**	**Creatinina Sérica Pro**	**Creatinina Sérica Ma**
ALT Prom:	**ALT Max:**	**AST Prom:**	**AST Max:**
Hemoglobina:			
No. consultas al oftalmólogo al año:		**No. Embarazos:**	
Conteo glóbulos blancos promedio:		**Conteo glóbulos blancos máximo:**	
Albuminuria Prom:		**Albuminuria Max:**	
Recuento Plaquetas Prom:		**Recuento Plaquetas Max:**	
Nitrógeno Ureico Prom:		**Nitrógeno Ureico Max:**	
Ácido Úrico Prom:		**Ácido Úrico Max:**	
Proteinuria Prom:	**Proteinuria Max:**		**Proteinuria Inicial:**
Antecedentes familiares nefropatía: Si [ ] No [ ]		**Presenta Retinopatía:**	
Insuficiencia arterial: Si[ ] No[ ]		**Pie de Charcot: Si [ ] No [ ]**	
Enfermedad vascular periférica Si [ ] No [ ]		**Hongo uñas: Si [ ] No [ ]**	
Antecedentes ulceras o amputación: Si[ ] No [ ]		**Infección: Si[ ] No [ ]**	
Complicaciones: **Retinopatía[ ] Nefropatía[ ] Pie Diabético[ ] Ninguna[ ]**			